中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

FAST 数据处理系统
可视化分析与自动化管理技术研究

作者姓名：钱旭冉

指导教师：朱明 研究员

中国科学院国家天文台

学位类别：理学博士

学科专业：天文技术与方法

培养单位：中国科学院国家天文台

2018 年 06 月

Data Visualization and Task Management for FAST Data Processing Pipeline

A thesis submitted to
The University of Chinese Academy of Sciences
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy of The Methodology and Technology of Astronomy

By
Penny Xuran Qian
Supervisor: Professor Ming Zhu

National Astronomical Observatories, Chinese Academy of Sciences

June, 2018

中国科学院大学
研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：_____ 日期：_____

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：_____ 导师签名：_____ 日期：_____

摘　要

FAST 是当前世界上最大的单天线射电望远镜。国家科教领导小组审议确定的国家九大科技基础设施之一，FAST 将开展多项巡天任务，引领射电天文学的国际前沿。FAST 预计覆盖频率范围 70MHz 到 3GHz 进行多波段观测，产生多达 60PB 的数据，这些 “大数据” 无疑对于后续的数据处理带来了极大的挑战，其处理不仅需要一套高自动化的软件环境，将原始数据流转换为便于存储和理解的规范化数据格式，更应当结合新的数据分析技术和展示手段来充分挖掘数据中的科学价值。

可视化分析是对天文高维度异构数据进行挖掘和理解的重要手段。当前，绝大多数的数据可视化软件并非为探索高维度数据而设计，而且它们也并非针对天文数据处理所需要的特定工作流而开发，对开源数据可视化软件 Glue 的开发与设计正是填补了这方面的空缺，我们使用了 “视图互联” 的可视化分析方法来帮助天文学家同时探索多个数据集之间的关系，此外，通过设计并开发 Glue 软件的多样化的三维交互功能，不仅满足了高维度天文数据探索的需要，更是将视图互联延伸到三维空间，这也是可视化分析领域的创新。针对 FAST 工程所产生的巨大数量的、高维度的、异构的数据，我们通过详尽的文献和技术调研，基于高度可扩展的 Glue 软件平台并结合快速数据搜索及云端计算等技术手段，提出了 “无缝化”FAST 可视化分析模块的设计框架以及一系列可行性技术方案，期望为 FAST 提供一个高效的数据处理工作流。

FAST 数据中的科学含义也可以通过可视化的方式进行传达。当天文学家完成了数据处理和解读时，将对数据进行可视化渲染，将 FAST 发现的科学成果栩栩如生地展示出来，是对科学数据进行理解的重要一环。我们通过探索现有的高维度天文数据展示的新技术，以实例给出具体实现方法及其展示结果并对其进行分析。通过设计与开发免费开源三维和动画软件 Blender 扩展插件，实现对高维度天文数据实时的三维展示。同时，我们还探索并利用立体显影及虚拟现实等技术，实现对 FAST 数据和成果的可行并新颖的 “三维” 展示，进一步拓展数据的交互和展示范围，进行对全天图数据的展示。

FAST 要求自动化跟踪所有脉冲星搜索过程和数据文件，并记录脉冲星候选和参数。通过结合数值模拟的任务管理经验，我们设计并开发了基于 Python 的轻量级流水线自动调度工具 SiMon 来完成自动流水线并行调度。我们通过决策树模型，实现自动调度任务流并实时监控运行状态及归档记录，不仅能够根据当前系统的计算资源，自动选择一种能在最短时间完成处理工作的调度算法来执行数据处理流水线，还能够在任务中断时自动进行回滚重启的操作。此外，我们利用规范化的程序架构设计，使得该工具具备高度可扩展性和实用性，用户能够根据任务管理需求和特点进行自定义智能监控功能的开发，实现真正满足天文研究中对批量任

务管理的需求。

FAST 的数据处理将更加无缝化。传统上，天文学家在数据处理时，需要首先把相关数据下载到其本地计算机上并加以处理。鉴于 FAST 的巨大数据量，任何的本地计算机都不具备足够的存储容量，因此这种传统的数据处理方式在面对 FAST 的数据挑战上是不现实的。我们展望了使用云计算辅以虚拟天文台技术作为解决这些问题的关键步骤，使整个数据处理工作流更加无缝化。毫无疑问，这将是未来 FAST 软件工程的重点。

关键词： 天文数据处理，数据可视化，自动化任务管理，虚拟现实，天体物理，机器学习，云计算

Abstract

FAST is the largest single-dish radio telescope in the world. As a Chinese mega-science project, FAST is expected to unlock the state-of-the-art of radio astronomy through its highly ambitious surveys. The project will cover the bandwidth from 70MHz to 3GHz, which will generate 60 PB of data in total. These "big data" poses an outstanding challenge for the subsequent data processing. The storage and processing of such amount of data require serious software engineering efforts, while new analysis and visualization techniques are expected to be carried as to fully discover the scientific meaning beneath the data.

First, visualization is an essential method of analysis and understand the high-dimensional and multivariate astronomical data. Currently, most data visualization packages are not designed for visual exploration of high-dimensional datasets, neither are they optimized for the specific workflows involved in astronomical data processing. As such, open source visual analysis software Glue is developed to fill these gaps. Glue employ a "linked-view" system to help astronomers exploring the relations within and across related datasets. General purpose viewers, such as 1D histogram, 2D scatter plot, 3D scatter plot, and 3D volumetric rendering, are built-in with Glue. Special purpose viewers, which is needed for visualizing dedicated astronomical datasets, can be implemented within the Glue framework as plugins. The support of 3D visualizations and selections enable Glue to extend the "linked-view" diagram from 2D space to 3D, which is a novel contribution to the visualization field. Through its Publish/Subscribe pattern and highly modular code architecture, new functions such as remote access and communication with third-party software can be implemented with reasonable effort. Thus Glue will be key for the FAST analysis module and provide further service for FAST research.

Second, the stories behind the data need to be visualized and conveyed. When astronomers finish their analysis of the FAST data, it is important to "tell the stories" by state-of-the-art visualization technology. This project makes use of Blender, a free and open-source tool to create attractive 3D real-time display for high dimensional astronomy data. Besides, several advanced visualization technique are explored and applied on exploring data set with more than two dimensions, including stereoscopic display and virtual reality technique. All these pioneer work will directly benefit the dissemination and comprehension of FAST data products, which will in turn inspire scientific discoveries for FAST.

Third, the data needs to be processed in as efficiently as possible. Due to the massive data rates being generated by FAST, it is critical to implement highly efficient real-time data processing pipelines. A lightweight python-based tool SiMon is developed, aiming to schedule observational data processing pipelines with minimum human supervision. Astronomers can specify a list of data processing jobs, and then SiMon will be able to automatically find a way to make full use of available computational resources to complete these jobs with maximum efficiency.

Last but not least, the data is becoming seamless. Traditionally, astronomers download the data to their local computers to perform data analysis. This workflow, however, is no longer practical because the enormous sizes of FAST data render downloading and local storing impossible. Cloud computing technologies, in combination with the Virtual Observatory technology, offers elegant solutions to these challenges. Should these technologies be deployed to the FAST project, the workflow of data process will become seamless, as the logistic of data is automatically taken care of by the cloud. Astronomers will find the data available at their disposal with a single click. Undoubtedly, this will be the future of the FAST software efforts.

Keywords: Astronomy Data Processing, Data Visualization, Task Management, Virtual Reality, Astrophysics, Machine Learning, Cloud Computing

目　录

图形列表

表格列表

第一章　引言

1.1　FAST 工程概况和科学目标

500 米口径球面射电天文望远镜 (FAST)[1]是中国建成的目前世界上最大的单口径射电望远镜，如图 1.1所示，其主动反射面口径达到 500 米，在观测时有效照明口径可达到 300 米。自 1994 年中国科学家提出利用贵州洼地建造 FAST 射电望远镜的工程方案，直至 2007 年 7 月，FAST 作为国家重大科学工程得到国家发改委正式批准立项[10]，通过前后约十年的工程建设和调试，FAST 总工程于 2016 年竣工并于同年 9 月第一次试观测。在经过不到一年的调试阶段后，2017 年 8 月，FAST 首次新发现两颗脉冲星，编号分别为 J1859-0131(又名 FP1-FAST pulsar #1) 和 J1931-01(又名 FP2)[2]，这也开启了 FAST 多科学目标同时扫描巡天 (The Commensal Radio Astronomy FAST Survey, CRAFTS) 的观测模式。

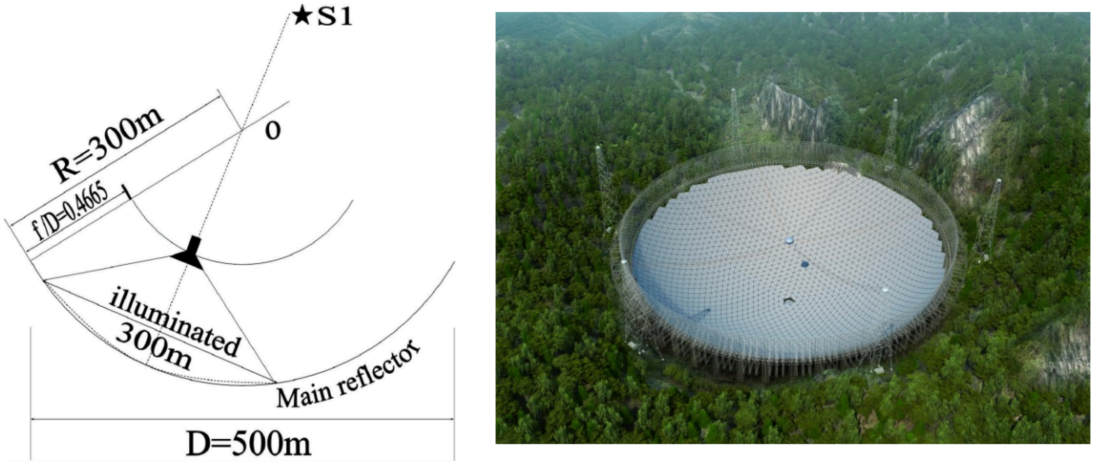

图 1.1: FAST 的结构几何图（左图）和三维模型图（右图）。图片来源 R. Nan, et al.(2011)[1]Fig. 1

FAST 工程有着三项自主创新。除却天然的贵州喀斯特洼地构成的台址以及轻型索拖动机构与并联机器人来实现接收机高精度定位和跟踪，FAST 能够看到更多更远的天体的一大创新点便是利用主动反射面技术在地面改变球差。[1] 反射面负责接收电磁波并反射到焦点，以便通过接收机对这些电波进行接收和记录。FAST 在观测时，会将一个口径 500 米、球面曲率半径 300 米，球冠张角 110-120° 的球面，通过主动变形形成有效照明口径 300 米、焦距约 138 的旋转抛物面[3]，将电磁波汇

[1]http://fast.bao.ac.cn/en/en/

[2]http://crafts.bao.ac.cn/pulsar/pulsar_media/

[3]http://fast.bao.ac.cn/showNews.php?Action=Cur&ID=4

聚在焦点上，实现了用传统望远镜的接收技术实现宽频带观测。这也使得 FAST 能观测的最大天顶角（Zenith Angle）达到 40^o，比上一代与被评为人类 20 世纪十大工程之首的美国 Arecibo 300 米望远镜多出了近一倍。

FAST 最初的科学动机是利用其优越的灵敏度和高观测速度对低频射电波段的信号进行研究。在 R. Nan, et al.(2011)[1] 和 D. Li, et al.(2013)[11] 都指出，FAST 科学目标主要有以下几点：

- 银河系中性氢巡天。中性氢 (氢原子 HI) 气体的 21-cm(频率 1.42GHz) 谱线是氢原子基态的一条自旋翻转禁戒跃迁的谱线，其跃迁的概率极低，但是由于中性氢在宇宙中有着大量的分布，所以我们可以很容易能观测到这条谱线。现有的银河系中性氢寻天主要集中在银盘内，例如 Very Large Array(VLA)[4]、Dominion Radio Astrophysical Observatory(DRAO)[5]、Arecibo 和 Parkes 等，而 FAST 优越的观测范围，使得 FAST 在高银维天区有着较大的优势，不仅能够寻找和统计高银纬的氢云，还更大的补足了银河系内氢云的分布，通过高银纬氢云的完整统计，可以示综银河系盘外结构，研究 ISM 总体演化，这对深入研究恒星行程的物理过程具有重要作用。此外，利用氢的窄线自吸收 (HI Narrow Self-Absorption features，HINSA) 可以示综分子云中的原子丰度，这对于研究恒星形成十分重要[12]。

- 银河系外中性氢巡天。为了利用中性氢构建星系演化的图像，需要将中性氢的观测距离延伸到 $z \leq 0.3$。目前，对中性氢云进行巡天的最大红移为 $z \sim 0.2$。然而利用 FAST 的高灵敏探测器和多波束，巡天红移预计可以达到 $0.3 \leq z \leq 0.7$，并可以观测到该红移范围处星系团核心的中性氢。另外，得益于 FAST 的高灵敏度，FAST 可以用于观测矮星系。作为恒星和星系形成的基本材料，中性氢巡天可以协助我们理解星系的形成机制[13] 及动力学（如旋转曲线／暗物质晕等）结构。

- 脉冲星观测。脉冲星是中子星的一种，也就是恒星演化到末期，经由引力坍缩发生超新星爆炸后，形成的密度极大的天体。脉冲星能够以极其精确的周期发出脉冲信号。这个机制可以用于星际航线的导航、引力波探测、探测 ISM 等。FAST 预期发现 4000 个河内脉冲星，并尝试搜索第一个系外脉冲星[14]，在 FAST 这样巨大的脉冲星巡天样本中也许会找到目前尚未发现而可能存在的新品种，例如中子星-黑洞双星。

- 主导国际甚长基线干涉测量网 (VLBI)。一个望远镜的角分辨率 θ 由所观测频率的波长（λ）和望远镜的口径（D）决定：$\theta = \lambda/D$。因此，为了达到同样

[4]http://www.vla.nrao.edu/

[5]https://www.nrc-cnrc.gc.ca/eng/solutions/facilities/drao.html

的分辨率，波长越长，则望远镜所需的口径越大。由于射电波段的波长远远长于可见光，射电望远镜的口径也必须具有巨大的口径。口径越大，维持反射面的抛物面结构就越困难。VLBI 正是为了解决这个问题应运而生的。这个网络利用干涉的原理，用多个天文望远镜同时观测一个天体，模拟一个大小相当于望远镜之间最大间隔距离的巨型望远镜的观测效果。FAST 作为目前世界上最大口径的射电望远镜，加入 VLBI 网络将产生引领性的作用。

其中，宇宙中性氢 HI 巡天和脉冲星搜索被看作是 FAST 的重点科学课题和早期科研目标中最重要的两个方面[15]，这两个方面也将是 FAST 数据处理系统的重点所在。

1.2 FAST 数据处理系统的概念和框架

数据处理系统可以说是对科学研究工作流的一个具体实现，与传统的天文研究工作流相似的是，现代天文研究也是基于对数据的获取与处理，通过一系列工具和处理手段来理解数据并获取最终结果，对应数据量和复杂性的增长，相应的工具和处理手段也丰富起来，在 S. G. Djorgovski & R. Williams（2005）[16] 就对现代天文工作流概括如下：

- 数据收集：收集由不同观测设备生成的原始数据流。仪器效应能够被移除，并以专业特定的方式进行校准，通常会使用一些数据处理流水线（Data Reduction Pipeline）。
- 数据归档：包括对原始和处理的数据、元数据和派生数据产品的存储和归档，这也包括了优化数据库架构、索引、可搜索性、互操作性和数据融合等问题。尽管还有这部分还有比较大的探索空间，但从目前看来这些挑战已经被多方面研究，并取得很大进展。
- 数据分析：包括聚类分析、自动分类、异常值或异常搜索、模式识别、多元相关搜索和科学可视化，它们通常在测量属性或图像的一些高维参数空间中。这也是目前的关键技术挑战。
- 数据理解：将分析结果转化为实际知识。这里的问题本质上是更偏向方法论。我们需要学习如何通过增加数据量、复杂性和质量以及计算机领域技术的进步来提出新类型的问题，这也是科学创造力所在。

同样的，理想中的 FAST 数据处理系统不仅仅是传统意义上的对望远镜收集的数据进行 “清理” 的 Data Reduction Pipeline，而更应是一个能含括上述四个阶段的数据产品生产线：原材料是望远镜观测所返回的观测状态、谱线等采集的各

种数据，生产线提供了各种工具针对这些数据进行筛选、清洗、分析、挖掘等一系列“处理过程”，并输出规范化天文数据结构的数据，这些数据经由专业软件展示和分析，最终帮助天文学家理解所探寻的科学问题。而这些数据按不同需求进行包装和展示，也可以用来吸引不同的用户如大众科普、科学交流等。可以说，这种使得 FAST 研究人员通过输入原材料并参与处理过程，就可以获得用于结果展示和应用的数据产品的框架，才是在现代天文研究背景下 FAST 数据处理系统最理想的模式，如图 1.2FAST 数据处理系统的概念框架模型图所示。

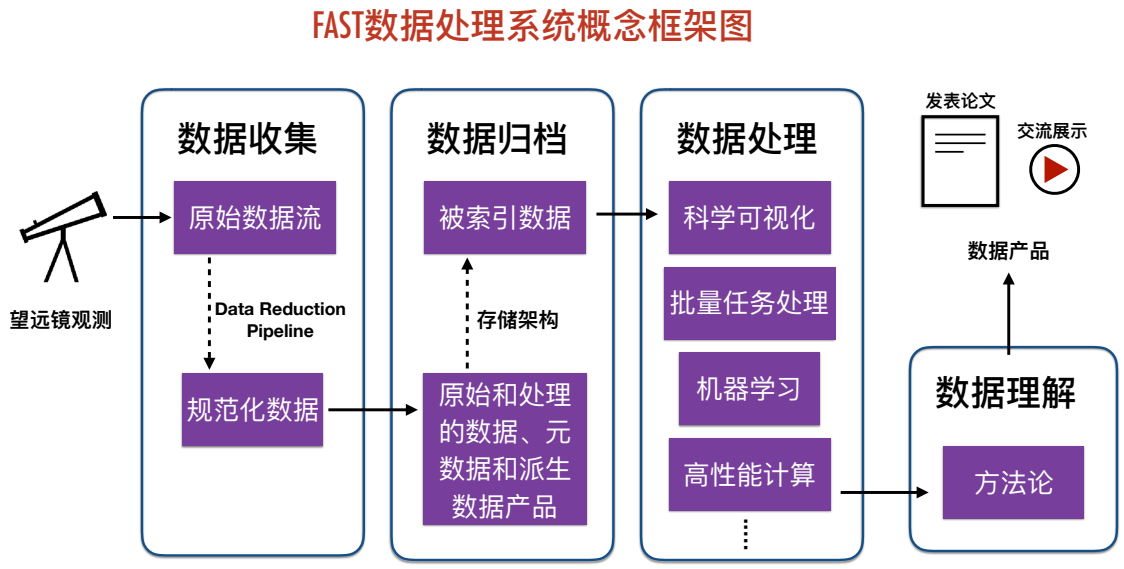

图 1.2: FAST 数据处理系统概念框架图。

这个“无缝化”数据处理系统理念与虚拟天文台构想（Virtual Observatory Concept) 紧密关联。虚拟天文台（VO）构想是将先进的计算机技术把现有和未来的丰富天文数据的资源无缝地融合在一起，并借助面向网络的各种软硬件资源为天文学家及公众提供前所未有的天文研究和教育服务。而很多组织和项目也在朝着这个目标进行科研与开发，例如国际虚拟天文台联盟 (The International Virtual Observatory Alliance, IVOA)[17]，一个新兴的、开放的并基于网络的分布式研究环境，针对于具有大量和复杂数据集的天文学研究。也因此，FAST 数据处理系统预期是汇集了数据存档和服务，以及数据挖掘和分析工具的完整体，与 IVOA 相似，它既是技术支持，但由科学驱动，旨为天文学家和计算机科学（CS）和信息科技专业人员和统计人员之间的协作提供了极好的机会。

同样的思想也与 Harvard-Smithsonian Center for Astrophysics 的无缝天文项目组[6]紧密相连。从图 1.3可以看出，无缝天文的思想便是将研究人员置于整个研究

[6]https://projects.iq.harvard.edu/seamlessastronomy

过程的中心，所有的工具和技术都可以简单获得并直接使用，无缝天文学试图让文学和数据之间的联系变得更加无缝和隐形，这样以来研究人员可以花更多的时间思考科学，而不用去寻找信息和工具。

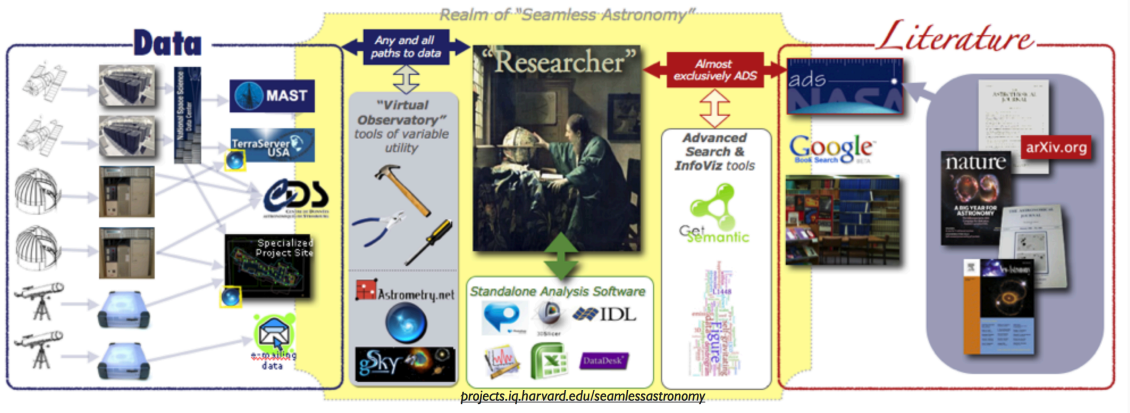

图 1.3: Harvard-Smithsonian Center for Astrophysics 的无缝天文示意图，展示了研究人员如何坐在文学和数据之间，从每个角度获取信息，整合自己的分析工具，然后生成新的出版物和结果并反馈到这些数据来源。来源：Harvard 无缝天文项目页面

1.3 数据收集与归档

数据收集和归档是 FAST 数据处理系统框架下的两个基础组成部分，但同时也面临着挑战。规划中的 FAST 巡天观测，包括 19 波束进行的脉冲星快速扫描以及河外巡天对宇宙射电源的搜索，其每小时的数据预计可达到 5TB, 每次 8 小时的巡天观测产生的数据就是 40T。第一批全面的脉冲星和中性氢巡天预计可在 18 月内完成。在不损失观测精度的情况下，约需 60PB 的存储空间。这些海量数据对数据收集和归档提出了巨大挑战。

面对相同的挑战，国际上很多巡天项目利用高自动化的数据处理流程及完备的数据服务从而取得了成功。例如国际上最成功的天文巡天项目之一 SDSS，其成功很大程度上是归功于其数据开放政策与非常完备的天文数据库及相关的数据服务，其大部分的论文的研究工作都是利用 SDSS 存档数据进行的。同样的还有利用单口径望远镜 Arecibo 进行的 The Arecibo Legacy Fast ALFA (ALFALFA) 巡天[18] 项目，其通过一系列自动化 IDL 数据处理流水线，自动化将每小时约 1.2GB 的巡天数据由原始 FITS 结构转化为易于处理（单个 600s 漂移扫描）的 IDL 格式，这大大简化了数据存储以及后期分析处理的复杂度。

FAST 巡天数据处理流水线目标是负责科学、高效的处理数据并具备一定的自动化分析功能。以 FAST 中性氢巡天为例，计划中的数据处理流水线基本流程应

包括数据转换与传输、自动信号处理、人工视觉检测和三维成图并科学存储四个阶段，如图 1.4所示。

1. 数据转换与传输：包括将望远镜观测得到的数据转换成自定义的数据结构（例如 ALFALFA 使用的 Drift Structure）并进行存储传输。

2. 自动信号处理：为消除仪器本身身的干扰及宇宙背景信号的影响，自动化对数据进行一系列信号处理包括对整个观测单元的数据进行行噪声校准、带宽通量校准及基线定标，并包括相应的程序崩溃处理机制，要求输出的处理结果不仅保留了小尺度特征例如河外星系信号，而且也包含大尺度结构如高速云团（HVCs）和星系间中性氢结构（Galactic HI）。

3. 人工视觉检测：包括一系列交互式可视化界面，使得数据处理人员能够对瑕疵数据重新进行信号校准、带通校准和定基线，并对每个 Position-Velocity 图进行行 RFI 标定，且能够利用自动信号提取算法获得一系列候选信号源。

4. 三维成图及科学存储：在对数据进行第二遍处理（包括用用连续谱源进行重校准）后并且处理结果符合质量评估要求后，利用 Homogeneous Resolution Kernel 对数据进行平滑并将采样数据重组为三维数据单元 (Data Cube) 并进行科学存储。

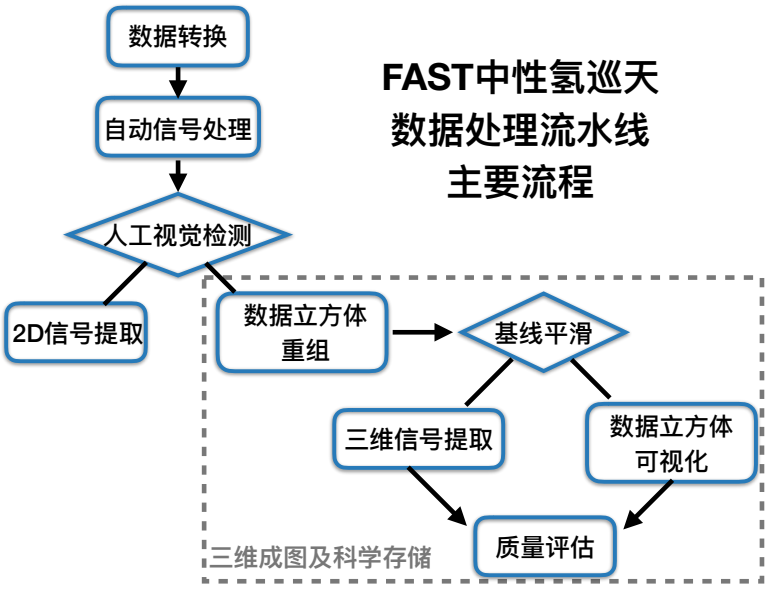

图 1.4: FAST 中性氢巡天数据处理流水线主要流程图。

目前也已有一批可用的软件工具包用于完成以上数据处理步骤，如 The Arecibo Legacy Fast ALFA (ALFALFA) Survey 所开发的一套针对处理河外中性氢巡天的

IDL 数据流水线[7]；如 GILDAS 软件[8]，其被用在 IRAM 30 米望远镜和 NOrthern Extended Millimeter Array（NOEMA，除去 VLBI 观测），能够针对毫米波或亚毫米波的射电信号进行数据处理等。此外，Livedata 和 Gridzilla 为多波束单镜数据数理系统做带宽校准和成图软件系统，被用于 The Arecibo Ultra-Deep Survey (AUDS)[19] 以及 The Arecibo Galaxy Environment Survey (AGES) 的数据处理[9]。除去这些功能较完全的数据处理软件，程序编写也能够完成数据处理流程的部分功能，例如通过 MATLAB 对基线进行平滑[10]，以及天文 Python 工具包 AstroPy 中包含的各类数据进行类型转换和重写的函数模块[20] 等。这些软件工具包的应用能够在 FAST 的数据处理流水线的开发实现中发挥作用。

对已收集及经过预处理的 FAST 数据，一种较优的归档方式是根据天文数据的特点选择数据的记录方式和存储格式。传统上，大部分望远镜归档及采集数据是通过将每个观测数据存成一个 FITS 文件[21,22]，相应的描述观测参数的信息（如位置坐标等）作为元数据（Metadata）并打包存放到文件头。这种方式对单个天体的数据存储和传输都是很有效的，但对大规模巡天数据的存储并不是最优化的。FAST 巡天时通过 19 波束进行的脉冲星快速扫描，或河外巡天对宇宙射电源的搜索，每小时的数据可达到 5TB, 每次 8 小时的巡天观测产生的数据就是 40T。以普通的磁盘阵列 1GB/秒的速度读或写一次就要约 11 多小时，无法满足实际需要。我们考虑到一种可行性方案为，将大天区的数据先分割成成千上万个 FITS 文件来记录、存储、传输，而当需要处理整体数据时再重新读回每个文件，根据文件头的信息确定各文件数据的位置，然后对数据进行拼接、叠加（积分）等操作。这些运算在小数据情况下利用高性能计算机的快速的处理能力可以在短时间内完成，满足实际需要。并且这种基于位置分割的天文数据新的存储方法，能够尽可能减少无必要的数据读写和传输，将是 FAST 数据归档与存储的另一个突破点。

1.4 数据分析与理解

数据分析和理解是科学研究的最关键步骤，也是本论文的重点研究部分。在经历数据采集和归档后，那些不太适合直接进行分析的数据，例如单数据量达到 TB 级别的数据块，已经能够满足普通个人计算机和分析工具进一步进行处理的条件。而事实上，大多数数据和数据特征并不能被人类直接理解，这也是因为数据复杂性，如数据高维度和数据间关系复杂等原因。类似的，FAST 的观测将同时产生不

[7] http://caborojo.astro.cornell.edu/alfalfalog/idldocs/

[8] http://www.iram.fr/IRAMFR/GILDAS/

[9] http://www.naic.edu/ ages/livedata.html

[10] http://www.mathworks.com/matlabcentral/fileexchange/24916-baseline-fit?focused=5133659&tab=function&requestedDomain=www.mathworks.com

同波段的数据，这就要求 FAST 数据处理系统使用或开发更先进的数据分析方法，能够针对数据复杂性和高维度特征进行分析和展示。

科学可视化是现代天文数据分析的一个重要手段。科学可视化是将数值信息转换成图像展示，从而帮助理解那些采取错综复杂而又往往规模庞大的方程、数字等等形式所呈现的科学概念或结果。科学可视化的概念从上世纪 80 年代提出[23,24]。它不仅仅用于数据展示，数据交互处理、包括质量化、量化和比较阶段的成图和分析，对深层次的理解科学问题也至关重要。尽管展示手段和数据类型不同，可视化能够清晰地给出数据维度上的分布趋势（如 Histogram)、维度与维度间的关联（如 Scatter Plot）以及维度信息在空间分布上展现的特点（如三维体渲染）等。

传统的天文展示方法在高维度数据上比较局限。在早期的研究时代，天文学家使用组合光谱、2D 等值线图像和位置速度构建高维度数据的三维 “感觉”，如 R. L. Snell, et al. (1980)[25]。其后，一系列 Channal Maps（如 H. G. Arce, et al. (2011)[26] 中的 Fig.6) 被用作显示 “数据立方体” 中的特征区域，进一步的，天文学家使用二维图像加上一维动作条（即 Channel Movies）来展示三维数据立方体，这迫使研究人员需要记住在其他 channel 的内容从而对数据结构在大脑中对数据的三维结构进行重建。近年来，一些新颖的可视化方法被用作展示高维度数据，例如 H. G. Arce, et al. (2011)[26] 中的 Fig.30，其通过使用 Schematic Picture 巧妙的展示了 Perseus 中各个壳层结构的尺寸、质量、动量、能量分布与位置关系等参数。这些方式虽然一定程度上给予用户多于二维的信息，但是并不能够提供真正的空间三维信息，并且缺乏足够的交互性。

射电天文领域第一次研究三维可视化的适用性可追溯到上世纪 90 年代（R. P. Norris (1994)[27]），当时已经明确三维可以更好的理解射电天文数据的高维度。三维技术的主要优点是更容易视觉识别数据的全部结构，包括延伸到多个通道（channel）的微小特征。Norris 提出一个关键点是，对于数据检查，质量化呈现结果可行的，但是交互式和量化假设检验需要定量可视化。在过去二十年里，几乎没有任何新的三维可视化开发了用于三维射电天文数据的工具。在 90 年代中期，Oosterloo (1995)[28] 调研了使用体绘制技术用于射电天文可视化，他分析了有关 Ray Casting 算法[29] 的功能以及与相关问题，并指出一般的三维可视化的优点和缺点。然而，由于计算机资源的匮乏，他并没有开发出一个实时的三维交互软件包。

现代天文领域已有不少天文工具支持数据可视化。但实际上，对能够可交互式对高维度数据进行探索与分析的软件仍非常匮乏[7]。尽管当前有很多软件可以绘制常用的直方图、散点图等，但针对天文研究而专门开发的相关软件却不多。天文研究对可视化分析的需求，包括经常需要用到多种视图如 3D 立体渲染、3D 散点图、绘制天文的光谱谱线，这些视图需要具备数据拟合、数据着色、对数/线性/指数坐标轴、误差棒等，数据拟合可能需要用到特色的数学或物理模型，多个数据集之间可能需要相互关联、交叉证认。只有少数发表的天文学可视化作品展示了互动

的需要并提供定量可视化提出解决方案，例如 Amati, et al.（2003）[30]，J. Ahrens, et al.（2006）[31] 和 D. Li, et al.（2008）[32]。因此，对 FAST 可用的可视化分析模块的设计及技术研究，不仅需要良好的软件平台基础，也需要新颖的可视化技术的融入。这部分内容将在本论文进行重点展开。

此外，自动化处理也成为海量数据下，数据处理系统必不可少的一个方面。由于 FAST 每天将对大天区、多目标源、多频端目标进行观测，其不可避免地将产生巨大的原始数据。这些原始观测数据必须实时地进行处理（on-the-fly reduction），并且将经过初步自动处理过的、经过索引的数据存入天文数据量，天文学家方能开展研究工作。从技术角度而言，这是一个并发调用数据处理流水线的典型场景。假如 FAST 观测的天区为 $(W \times H)$ 的矩形区域，而数据处理流水线每次能处理 $(xtimesy)$ 的矩形区域，那么需要重复调用流水线 $(W/x \times H/y)$ 次，方能完成对该天区数据的处理（假设子天区之间的数据相互独立）。假设服务器能并行处理 c 个任务，并有 $(W/x \times H/y) \gg c$，那么数据池中的任务将需要分多批次分别处理。在一个批次中，不同的天区计算的实际不一定相同，因此有可能有个别任务提前完成，相应的 CPU 处于空闲等待的状态，直到同批次的其他计算全部完成。这些等待实际无疑降低了数据实际处理的效率。

为了解决以上问题，我们对计算机领域的并发作业调度软件进行了调研，并试图将这些软件用在 FAST 并发数据处理中。遗憾的是，这些软件大多都是针对计算机领域或商业领域定制开发的软件，如服务器负载均衡软件。我们尚未找到针对天文计算的特点专门开发的软件。为了填补这个空缺，我们开发了 SiMon (Simulation Monitor) 轻量级作业调度工具。该工具使用了非阻塞式决策树算法，智能地根据作业优先级及当前服务器的可用资源来调度数据处理流水线，并保证每个流水线按要求完成。这不仅大大降低了天文学家的工作量，而且也最大程度地缩短了所需的计算时间，从而保证了 FAST 数据分析与理解的正常运行。

1.5 论文结构

因此，本博士课题主要围绕 FAST 数据处理系统框架中的数据分析与理解模块，其主要内容包括：

1）分析天文学家对 FAST 数据可视化和视觉探索的需求，提出了对 FAST 数据可视化分析模块的框架设计与开发流程，其不仅是对 FAST 处理系统分析与展示部分的完善，也是对天文数据可视化的补充和创新；

2）探索性地提出了借鉴其它非天文领域的最新技术来对 FAST 成果数据进行展示，以期望满足 FAST 对三维数据展示的需求，这也是多学科合作的新型科研模式的一次具体实践；

3）设计并开发了批量任务管理工具 SiMon，自动化管理批量任务，预计对

FAST 脉冲星搜索过程和数据文件以及记录脉冲星候选和参数进行跟踪及自动化，最大程度地利用可用的计算资源来达到用时最少的目的。

4）在大数据背景下展望了云计算和虚拟天文台技术在 FAST 数据处理流程中的重要作用。

论文的结构组织如下：FAST 可视化分析模块的设计与实现（章节 二），FAST 数据产品展示的三维展示（章节 三），批量任务管理工具的设计与实现（章节 四）。最后，章节 五将对本论文进行总结及未来展望。

第二章　可视化分析模块的设计与实现

中性氢 (HI) 巡天是 FAST 的重要科学目标之一，其研究需要完善的数据可视化技术来展现三维中性氢气体空间分布。此外，对特定区域的多波段数据的分析也要求对不同观测数据进行比较，这带来了数据结构复杂和高维度两大难点。而开源工具包 Glue，最初由哈佛史密松天文台（CfA）团队开发，是一个针对高维度数据可视化与分析的软件。在其创新性融入了视图互联（linked-view）技术的基础上，我们设计并实现了三维展示和选取的功能和更多的视图模块。所有这些特点都使得 Glue 成为 FAST 可视化分析模块的最佳软件平台。考虑到 FAST 观测天区广以及数据量巨大的特点，数据存储很可能部署在远程数据中心和云平台，因此模块设计中也会整合地图册搜索功能和基于云端的数据接口，并支持自动化数据的获取、可视化及分析。

本章将首先阐述 Glue 软件的核心技术和创新点，以及三维可视化功能模块的设计与工作流。本章还对 FAST 数据可视化分析模块进行需求分析并提出一系列可行性方案，所有工作将应用到 FAST 的数据处理系统并作为 FAST 数据分析与理解的重要途径。

2.1　Glue：基于视图互联的多维度天文数据分析软件

2.1.1　背景介绍

“大数据” 为天文学研究带来的全所未有的契机，但也伴随着对数据分析与探索的巨大挑战[33,34]。除了单个数据集大小（Volume）从 MB 到 GB 甚至 TB 的延伸[35]，数据集的高维度也对数据分析造成了困扰，如何去理解每个维度对应的含义，以及发现不同维度之间的关联，往往是回答某个科学问题的关键。尽管机器学习确实有着强大的处理大数据的能力，并能够通过训练极大的加快科学研究工作流[36]，但这也仅仅是解决大数据挑战的一种可能方式。大多数情况下，天文学家往往需要交互式的探索数据集自身和数据集之间复杂的联系。探索这些异构的、高维的、庞大的数据集是当前数据可视化领域迫切需要解决的问题。

在 Goodman(2012)[2] 中提到了对高维度天文数据处理分析的一个必要方法是通过 “linked view”，也就是视图互联，创建几个简单的视图并将它们链接在一起，当用户与其中一个视图交互时，例如选取了一个兴趣区域，其它视图也会随之更新并显示交互所选取的区域[37]，Gresh et al. (2000)[38]、Tukey (1977)[39] 和 Wong & Bergeron (1997)[40] 等研究中也指出这种交互视图能够非常有效的对高维度数据进行可视化可分析。关于视图互联的示意图如图 2.1所示，四个视图分别使用了四种

不同类型的展示方法，且视图与视图间的数据之间存在着联系，在一个视图中选取的部分也会在其它视图中被高亮显示出来（例如图中的红色部分）。一个高效的“linked view”系统，不仅能够深入探索单个数据集及其内部属性之间的联系，可以方便得出不同数据集关于同一个数据源的比较关系，还能够可视化展示这些数据和数据间的联系，并给与用户极大的自由使得他们能够灵活选取兴趣区域并进行进一步分析。

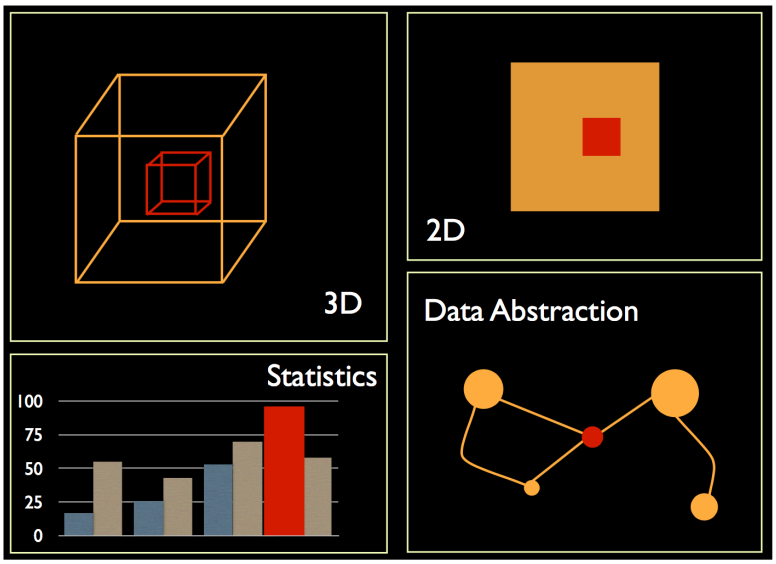

图 2.1: 该示意图展示“linked-view”的具体含义。本图来源 A. Goodman(2012)[2] 的 Fig. 3，本图由 Michelle Borkin 创作。

鉴于对高维度复杂数据集探索的需求并受到“linked-view”思想的启发，我们与 CfA 团队合作，参与开发了一个新的开源软件包名为 Glue[1]。Glue 旨在服务科学家探索数据集内部属性之间的关系，以及不同数据集之间的联系。通过灵活的可视化交互及展示和友好的图形用户界面，使科学家轻松地进行数据集的多维链接可视化，交互式或编程性的来选取数据子集，并且可以看到这些选取的子集数据在不同的可视化视图中（例如图形、图表、诊断图表等）展示。Glue 同时具备可操作的图形用户界面（GUI）和可交互的代码输入终端。GUI 相比于早起单调的命令行界面，对于用户来说在视觉上更易于接受，使用起来也更加自然，在使用 GUI 时，通常更容易将重点放在科学问题上; 这促进了对数据的更多创造性探索。GUI 由较不精确的交互手段 - 鼠标移动、按键等进行介导。虽然原则上可以记录这些交互操作并创建“可重复”的工作流，但是生成的日志通常不是人类可阅读的内容。相比之下，程序化交互比使用 GUI 更具有可重复性、易理解性和可扩展性，但它对于非专业编程人员而言往往很难使用[41]。因此，Glue 这样的混合 GUI 和

[1]http://glueviz.org/en/stable/

代码编写的用户界面集合了两者的优点，将 GUI 的流动性与编写代码的精度和可再生性结合起来，从而能够实现更敏捷的数据探索。

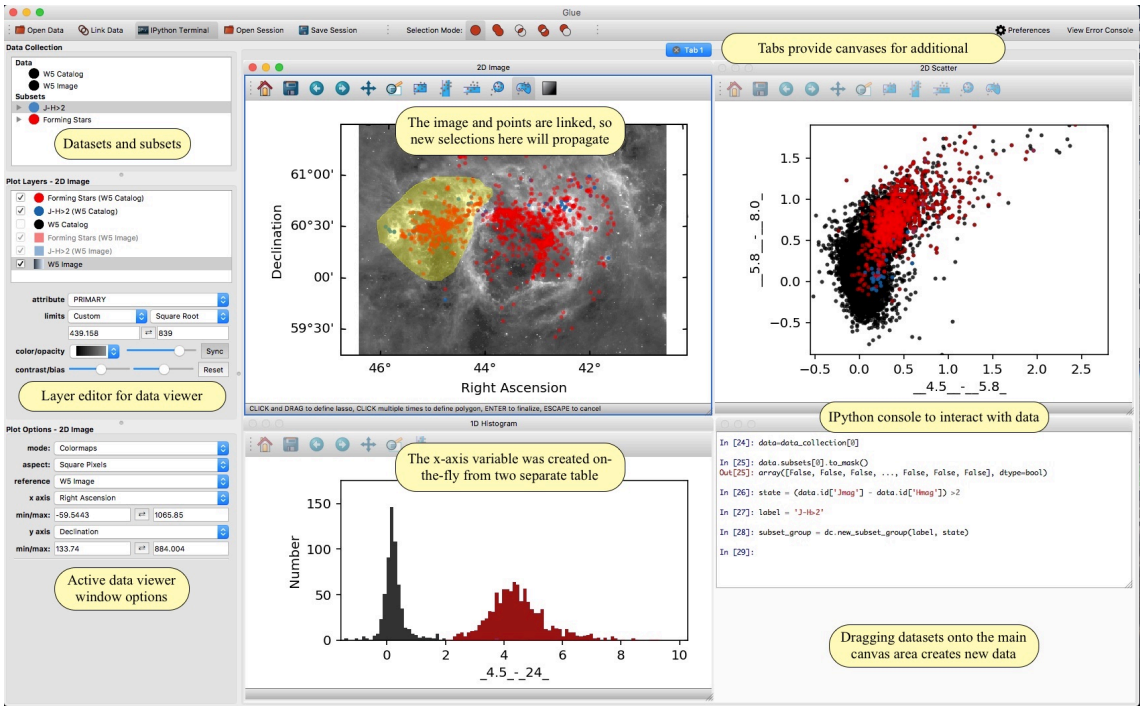

图 2.2: 该图展示了 Glue 的主要应用程序窗口，黄色背景文字框给出了各个模块的注释。该界面最主要部分是一个画布（Canvas）区域，用户可以在其中添加各种数据视图，每个 Tab 实际上都是一个新的窗口画布，以创造一个最适合于特定问题的可视化环境。左侧是一个侧边栏，包含所有加载的数据集列表、已创建的数据视图及一系列显示设定。加载的数据集可以是不同的类型或格式，它们不需要合并（Merge）就可以使用 Glue 自带的数据互联功能进行关联。

我们通过图 2.2给出了 Glue 软件的标准界面和基本工作流。Glue 的标准界面分布提供了多样化的显示选项，如散点图、直方图和二维图像显示等，用户具有很大的灵活性来指定特定的可视化方式。不少可视化软件用户界面较为复杂（例如 Blender[2]、Adobe Photoshop[3]等），用户通常需要一定学习时间才能正常使用。在这点上，Glue 根据数据集类型给出合理的绘图猜测，并允许 “拖放和绘图” 方案，以便用户可以用最小的努力获得最佳可视化效果。一个典型 Glue 工作流，正如 Beaumont(2014)[42] 中描述的：首先，加载数据并对相关属性进行关联；其后，数据集将被拖放到主画布上以创建对应数据类型的视图；最后，通过选取等交互操作来探索数据中不同的区域，在一个视图中进行的选取将实时传播到所有其他的链接视图，这也是对 “brushing and linking” 范式的实例化，即 “将一个可视化

[2]https://www.blender.org/

[3]http://www.adobe.com/products/photoshop.html

视图中的交互操作自动映射到其它可视化视图中”[4]。默认情况下，完整的数据集显示为灰色，而用户可以自定义选择突出显示颜色（参见图 2.2中的红色和蓝色高亮显示）。来自不同来源的数据在一起组合在一起，将 “brushing and linking” 由单数据集扩展到跨多个数据集。

2.1.2 数据互联与视图互联

Glue 的名字也蕴含了它最重要的两个特征：首先，该程序允许用户在其探索可视化环境中 “粘合” 不同的数据集，而并不需要合并文件，不同数据之间的属性可以被共享（例如同样表示维度，在地图文件 A 中的 “纬度” 与文件 B 中的 “lat” 可以通过建立逻辑关系进行关联）或通过简单的转换进行链接（如 “时间” 的数学表达式乘以 15）。这种 “在线处理” (On-the-fly Processing) 方式也就是数据互联。其次，在视图之间进行交互式或算法性地显示，通过选择数个数据子集并在不同视图之间显示出关联，有效地 “实时” 将信息显示在一起，也就是视图互联。

为了最有效且直接的实现数据互联与视图互联，我们设计并实现了针对天文研究需求的发布/接收（Publish/Subscribe）架构，在这个架构中，发布者发出消息而不用提前知晓特定的接收者信息，而接收者仅获取其感兴趣的信息而不用知晓发布者或其他接收者信息。这种基于信息传递的范式的优势在于其高自定义程度，能够直接保证可视化展示中对数据或视图产生改变时仍具备的高度同步性。并且也更易于新型可视化方法的实现或新数据类型的支持，这对扩展软件应用到天文以外的领域也有益处。

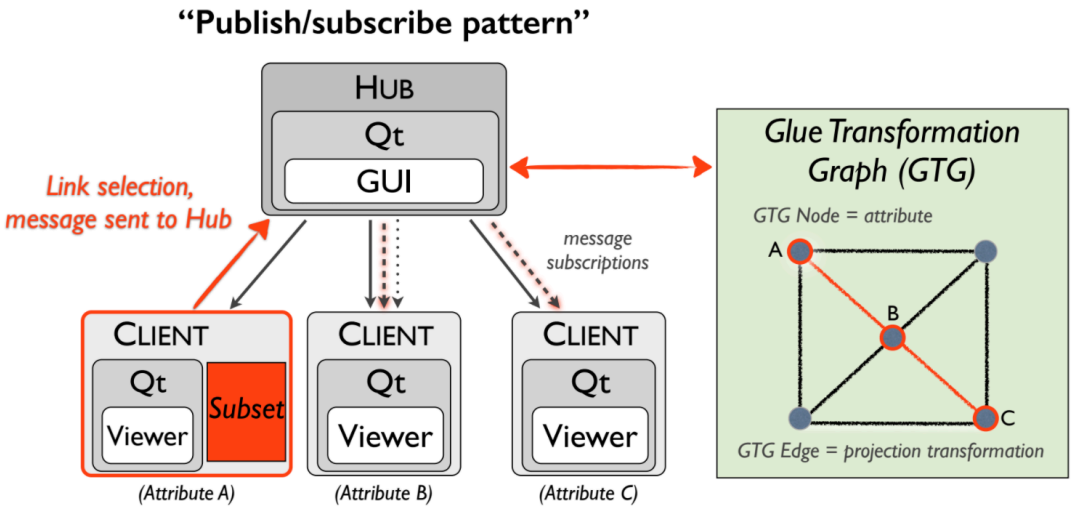

图 2.3: Glue 软件针对数据互联与视图互联的发布/接收 (Publish/Subscribe) 设计架构示意图，图片来源 Borkin, Qian, et al. (In prep.)。

[4]http://www.infovis-wiki.net/index.php?title=Linking_and_Brushing

我们通过图 2.3给出发布/接收 (Publish/Subscribe) 架构的示意图，图中 “终端 (Hub)” 表示发布者，“客户端 (Clients)” 表示接收者，每个客户端都代表一种特定的可视化展示方法 (Visualization Encoding)。运行时，每个客户端只接收与其特定的可视化方法相关的信息或数据，而终端则负责收集所有客户端的信息并记录状态改变情况，并在合适的时机广播相关信息。

视图互联发生在当一个客户端中的一个数据子集被选取，并且该数据也在其他相关的客户端中被通过高亮等视觉方式标出。如图 2.3中左下角客户端所示，当用户通过 QT 视图界面选取了其一个数据子集，该被选取子集被存储为一个子集对象 (Subset)。这个状态改变将被发送到终端，终端相应的将该信息发送到所有的客户端，相关的客户端会接收该信息并做出响应（例如在其所在视图中高亮该子集数据为不同颜色显示等）。

由于客户端可能包含多个数据集，因此需要定义被选取的数据集的属性（例如变量或数值等）与其它客户端中该数据集的属性的联系，当该联系信息被终端接收，如图 2.3所示，终端便会在 “Glue Transformation Graph(GTG)” 中找寻属性之间已定义的联系式。GTG 中的节点代表所有加载的数据的属性，连接节点的边是量化方程式来定义节点间的关系，又或者说是映射方程。这些映射方程或者是针对两个空间向量系统或者非空间向量系统。这种映射也可以在已连接的数据集中进行延伸，例如图中所示，属性 A 与属性 B 相关联，属性 B 与属性 C 相关联，那么当一个数据子集通过属性 A 被选取，那么这个选取将在属性 C 上也体现出视觉高亮效果。以上便是对单数据集的视图互联结合示意图进行的方法剖析。

而需要能够在不同数据集中进行视图互联，不同数据集间的数据互联也是必不可少的。也就是说，除非多数据集间的属性关联也在 GTG 中通过映射方程进行定义，否则通过上述方式进行视图互联并不能实现。因此，在 GTG 中添加相应的映射方程是实现数据互联的关键。边的添加能够通过系统中包含的一系列预定义映射方程（例如坐标系转换），用户通过 Python 编程也能够自定义互联方式。节点的定义也可以通过利用已存在的属性做出数学变换来获得，例如，i 和 j 是一个数据集中的属性，而一个新的属性 k 可以是 $i*100$ 或 $i+j$。而这种定义新节点不仅能够提供用户一个新的属性用于可视化及分析，在很多情况下也使得数据互联更为直观。

2.1.3 三维视图互联的设计与实现

我们在 CfA 组织了对 20 位天文学家的一系列问券调研及数次小型访问，旨在通过问答形式来准确获得天文学家进行数据分析与可视化的工作流程，了解现使用软件的不足，以及获取使用此类分析工具时所期待完成的任务类型。调研对象主体在年龄、性别和职业成熟度上均有不同，调研对象主体也在天文专业领域上

有所不同，包括侧重于观测真实数据、通过程序获取数值模拟数据以及两者结合。我们将这些信息概括为一系列针对所有天文领域研究都适用的核心任务类型分类，处理信息的方式是首先将其进行分级并归入合适的类别归属中（例如数据分析任务类别、数据类型、现有软件缺点等），其次整理所有类别和主题并将其量化表示，例如使用亲和图法（Affinity Diagraming）。

根据 Amar, et al. (2005)[43] 任务分类法，以及收集的调研数据，我们将天文学家进行数据分析时所要完成的任务归类为四大类：关联（Correlate，即比较多个数据集合）；找寻兴趣区域（Find anomalies，即找寻突出部分、不连续部分等）；聚类（Cluster，即聚类并找寻数据所呈趋势）；提取（Derive，即进行计算）。这些分析类任务均可归为探索（Explore）这个大主题下，尽管天文学家认为在研究中已经知晓所需探索的命题和所需计算的量值，调研发现其主要的操作仍然是探索数据来获得预期外的发现并得出合适的命题假设。而随着天文数据复杂性的增大，三维分析功能更有效更直接的探索天文高维度数据。调研发现，三维可视化能够特征化数据分布（Characterize distribution）及找寻兴趣区域，而三维选取则是过滤（Filter，即剔除非兴趣数据）、提取（Derive）和关联（Correlate）等高维度数据分析任务不可缺少的部分。以下我们便结合需求分析对三维视图互联功能的设计与实现进行详细阐述：

2.1.3.1 三维可视化

三维视图互联的首先要素是满足天文研究需求的多样化三维可视化方式。我们通过调研发现，在 Hassan & Fluke (2011)[7] 就曾结合天文数据的本质与高维度可视化方法对天文三维可视化方法做了一个归总如下：散点图（Scattered points，数据由一系列的坐标信息及数据属性信息构成）、规则类网格（Structured grid，数据值由一系列规则三维网格单元指定，网格单元与 Cartesian 坐标系对应，例如天文中常见的三维体渲染）、非规则类网格（Unstructured grid，数据值由二维或三维形状的边缘信息指定，例如三维等值面）和自适应类网格（Adaptive grid，数据值由一个多样化分辨率的规则网格指定，对于兴趣区域则提供较高分辨率）。而从天文技术领域比较权威的国际虚拟天文台联盟产品[5]及 The Astronomical Data Analysis Software and Systems（ADASS）会议发表的论文（调研结果可参考我们整理的 Google 表格[6]）概括得出，虽然使用的名称不同，但是天文领域主要三维展示方法是直接体绘制、等值面和三维散点，这三种方式足够满足常用天文数据类型的展示需求[44]。

我们选择的主要绘图库是 Python 开放源码库 Vispy[7]，最主要原因是其高性能

[5]https://mast.stsci.edu/portal/Mashup/Clients/Mast/portal.html

[6]https://docs.google.com/spreadsheets/d/1VhCRXdI_hshuBdMEfG8SML3gGjlriou6z4xedhikKC8/edit#gid=0

[7]http://vispy.org

三维交互式功能，通过 OpenGL 库来对目前的图形处理单元（GPU）的计算性能进行充分利用，对超大规模数据集也有很好的支持[45]。单数据集在视图中的三维显示与数据属性及渲染算法直接相关，由天文数据格式转换为图像处理格式由天文专业软件包 AstroPy[8]完成。AstroPy 是针对天文数据读取及运算开发的 Python 库，其包含了多样的功能函数能够将天文常用数据类型例如 FITS 读取并存储为普通数值数组的形式，而渲染算法在 Vispy 库已经封装，通过参数的传入及渲染对象的实例化便可以获得 GPU 端渲染的计算结果。而在一个视图中对多个数据集进行三维显示，便需要在计算渲染结果的同时运用到数据链接，通过定义不同数据集间的坐标属性映射关系，对同一个坐标位置上多个数据集的值进行计算（例如对 RGBA 四个通道分别相加取平均值），从而获得最终该位置坐标显示值。为了充分利用 GPU 的运算能力，我们重新定义了 Python 接口使得多个数据集输入能够直接传递到 GPU 端，将复杂计算任务移到运算功能更强大的 GPU 上进行，而软件端可以直接获取渲染最终结果并进行展示。

我们通过图 2.4给出通过 Glue 多样化三维可视化功能展示银河系纤维结构 (Filament) 的示例。纤维结构被定义为非常长而细且在速度尺度上保持连续性的分子云，在远红外（如 500μm）二维图像上呈现明显的吸收特征，被认为与银河系悬臂结构特征分布有着紧密的联系[46]，也因此对定义银河系的三维结构至关重要。在已知的十五个纤维结构候选体中，Filament5 是三维结构特征最为突出的一个[47]，由图 2.4所示。我们通过三维可视化来展示 Filament5 的三种不同成分的结构分布特征，并通过多数据集在单视图中的显示给出了该区域的一个更为全面的信息。图 a 和 b 分别显示的是通过等值面和体绘制对高分辨率 C18O 的渲染。图 c 表示较弥散的 13CO（红色）和较稠密的 C18O（蓝色）在该区域的分布特征。图 d 除却体绘制显示 13CO 和 C18O 外，还叠加了 BGPS_HCO Catalog[3]（红点）和 HOPS Catalog[4]（蓝点）用于比对该区域中已知的分子云核的位置与分布，也辅助确认 Filament 结构的轮廓。

2.1.3.2 三维选取与交互

三维显示相比于二维显示最突出的优点是其可以减少深度带来的遮挡。三维交互操作来对三维空间展示的对象的各个角度进行观察，例如使用默认鼠标左键拖动来旋转可视化立方体，以及鼠标滚轮放大或缩小到多维数据集中心周围等，使得在二维屏幕上观察第三维度信息非常直观[48]，这在二维显示中是无法直接实现的。三维选取的实现对于扩展视图互联到三维空间是至关重要的，但其实现并不简单，主要是因为将二维屏幕上执行的选取操作投影到特定三维区域并没有一种固定范式，且没有一种选取方案能够适合所有的用户需求，不同的研究数据和任

[8]http://www.astropy.org

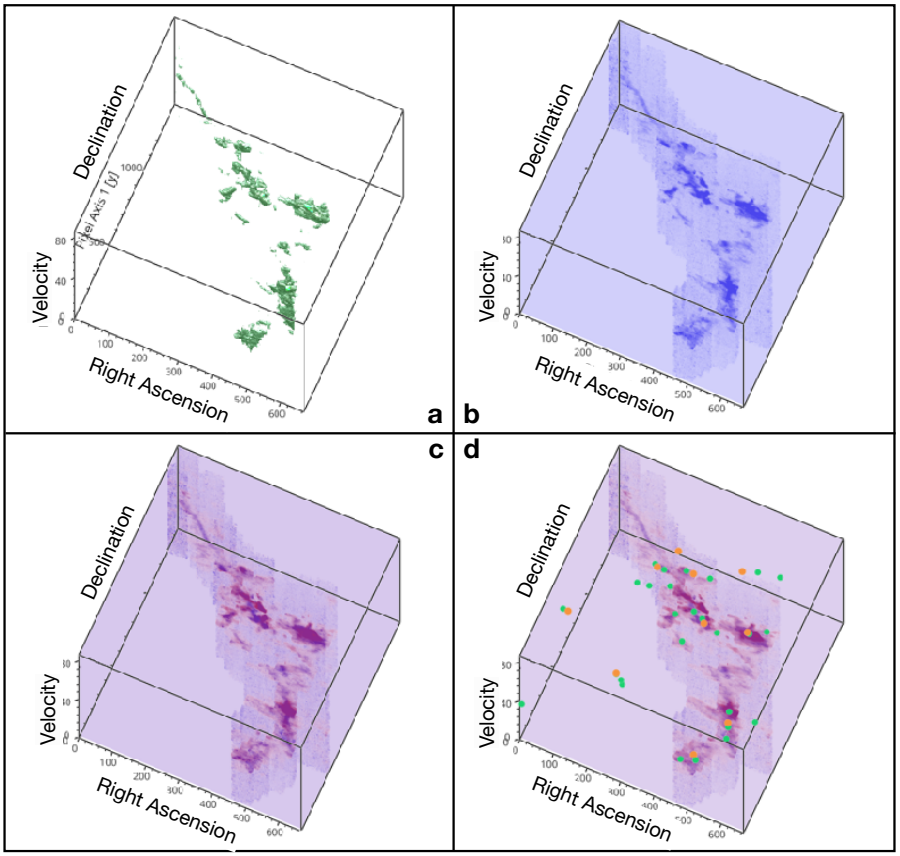

图 2.4: 通过 Glue 多样的三维展示手段来显示银河系内 Filament5 纤维结构。图 a 和 b 分别显示的是通过等值面和体绘制对高分辨率 C18O 的渲染；图 c 中红色表示较疏离的 13CO，蓝色是比较紧密的 C18O，两种成分通过位置关系重叠在一个视图中；图 d 中除却体绘制显示 13CO 和 C18O 外，还叠加了 BGPS_HCO Catalog[3]（红点）和 HOPS Catalog[4]（蓝点）。

务需要不同的选择机制，例如基于面的三维选取方法就不适用于无定型边界的三维体绘制渲染。

通过调研现有的能够符合天文数据特点的三维选取方法，如图 2.5所示，详细文档可参考我们归总的 Google 文档[9]，考虑到通用性，即一个选取方式能够使用多种数据类型及数据表达方式，易用性，即对输入设备和数据本身没有额外要求，以及可行性，即在工程实现上较易开发，这三个原则，我们为 Glue 的三维选取设计了以下两种方案：

基本选取：我们对其定义为通过用户绘制的选取形状和所处的投影方向来定义三维选取区域，如图 2.6的顶部面板所示。为了进一步精确化选取区域，我们还

[9]https://docs.google.com/a/cfa.harvard.edu/document/d/1cEn-8OnXNlxdO3E7ZuyO1QbM90lr-bsfpCyK3a60ECY/edit?usp=sharing

Implemented in Glue	Name	Description (definition terms)	Data Representations	Data Struture - change name?	Citations	Sketch Icon
TRUE	3D Ray-casting based approaches	Selection done by a ray casting line, at a distance, easy understandable	3d scatter plot, Isosurface(code work)	scatter points, unstructured grid	2, 34, 37, 40, 54	
TRUE	Simple Shape Based Primitives on 2D	2D shape extruded to 3D	3d scatter plot, volume rendering, isosurface(code work needed)	scatter points, structured grid	37,	
FALSE	Simple Shape Based Primitives on 3D	3D shapes	3d scatter plot, volume rendering, isosurface	scatter points, structured grid, unstructured grid	35, 46	
TRUE	**Extendable Selection**	Selection region changed live by input(non-intermediate selection)	3d scatter plot, volume rendering	scatter points, structured grid		
FALSE	Selection for specific structures (eg. 3D Streamline)	Aim at special 3D representations or scenes	isosurface	unstructured grid	28, 31, 58	
FALSE	Immersive Virtual Environment Selection	Dedicated input device	3d scatter plot, volume rendering, isosurface	scatter points, structured grid, unstructured grid	49, 10, 5, 41, 9, 23, 6, 8	
FALSE	Structure-aware Selection	Deduce selection intention, approximate in the first place	3d scatter plot, volume rendering	scatter points, structured grid	14, 15 (in 2D) 56, 39, 43	

图 2.5: 该表总结了我们对天文和计算机领域中的三维选取方法的文献调研结果。表中第一列给出该方法是否在 Glue 中被实现，第二列对可用的三维选取方法进行了归类，并在第三列对其进行了说明，第四列和第五列分别描述了可适用的数据表现手段与数据类型，第六列是参考文献来源，编号可对应文献列表，最后一列给出该选取方法的简易示意图以便理解。

设计并实现了逻辑迭代多次选取功能，即用户通过多次选取，新的选取区域将以用户选择的逻辑关系叠加到原选取区域上，例如，在选择 “和” 逻辑之后，最终选择区域被更新为所有选择区域的重叠部分，通过多次 “和” 选取操作，用户能够完成复杂的默认选取形状外的选取，如图 2.6中的底部面板所示。

智能选取：与基本选择方案不同，智能选取方案并没有固定的选取形状，最终选取区域由数据本身或算法辅助定义下的用户操作决定。考虑到通常情况下，用户所选取的区域一般都包含兴趣区域的中心或视觉可见的密度分布集中的部分，需要较强的主观性判断，因此我们设计选取区域的起始点由用户鼠标点选获得。紧接着，如何扩展该起始点到更大范围的一块区域，就需要对数据本身及用户行为进行考虑。本节将围绕图 2.7对智能三维选取的两种情况，即预先定义结构下的三维选取和算法引导下的三维选取分别展开讨论：

预定义结构下的选取

一般情况下，如果数据集本身存在附加约束条件，例如预先定义的层次结构，且用户将该约束条件与数据本身一起加载到 Glue 中，那么我们可以认为，这个约束条件是用户希望对选取进行规范的一个标准。也正如图 2.7上部分所示，这个时候的三维选取将基于整个数据集内的每个结构部分相应地提取选择。

我们也通过图 2.8阐释这类选取的工作流程和三维显示效果，图中使用的是 L1448 恒星形成区的 CO 成分[49] 的三个维度分别为 Position-Position-Velocity（PPV）数据立方体，该数据包含着预计算得到的 Dendrogram 树形结构，它是等值面的变化拓扑作为轮廓等级的函数的抽象[50]，通过层级表明结构分布关系，图左侧是 Glue 的数据列表及显示设定面板，中部是 PPV 数据立方体的三维体绘

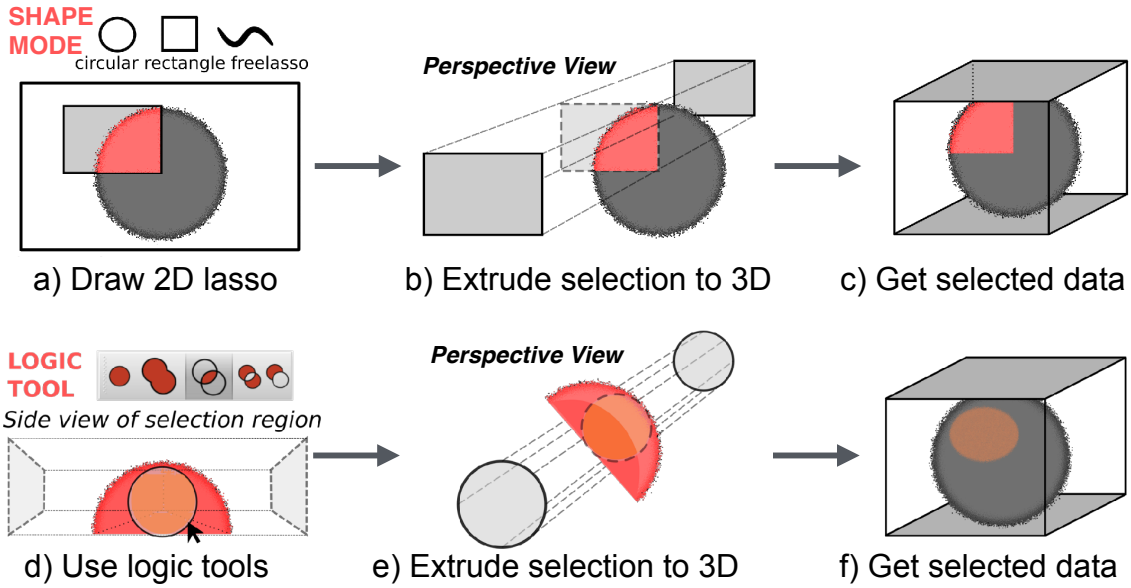

图 2.6: 该图详细的解释了用户通过在二维屏幕上选取的区域如何投射到三维数据集中并显示。上列图 abc 解释了拉索选取如何投影到三维：首先用户通过鼠标在二维屏幕上选取一个兴趣区域，这个选取的区域会根据相应形状投影到三维并映射出对应的三维数据集，Glue 会将新选取的三维数据集存储为一个子数据集并显示为不同的颜色。而下列图 def 介绍了三维选取与逻辑模式并用以获得更精确的选取区域的过程：Glue 提供了五中逻辑模式分别是：重新选取，逻辑或，逻辑与，异或，反蕴含。通过选取不同的逻辑模式并进行多次选取，就可以得到精确前一次的选取区域。如图 f 所示，选取逻辑与并进行两次圆形选取，可以得到一个自定义的椭圆区域。

制显示，图中右上角是树形图 Dendrogram 的二维显示，右下角是对 PPV 数据立方体的通量属性的直方图统计。我们通过视图区自带的选取按钮，点选树形图的枝干，通过 Glue 的 Publish/Subscribe 范式，选取的操作被终端 Hub 接收并广播，相应的三维视图区和直方图视图区对相对应的特定数据子集做出高亮显示相应。这类预定义结构的方法也能够实现对体数据这类无定型结构的数据进行量化选取。

算法引导选取

在预定义数据结构没有被指定的条件下，我们则需要对用户希望选取的区域做出一个猜测，并根据该猜测进行算法的开发和操作的引导，正如图 2.7的下部分所示。为了保证猜测的合理性，我们结合上文提到的调研分析的结果，对两种常用三维展示方法即三维体绘制和三维散点图分别进行讨论。

对于三维体绘制，数据展示的特点便是结构无定型化，整个数据立方体并没有确定的方式将其分割，因此，用户通过视觉观察发现的特征区域，无法立时由用户将其定义并返回给软件。而这些特征区域的特点便是在显示上往往与其邻近区

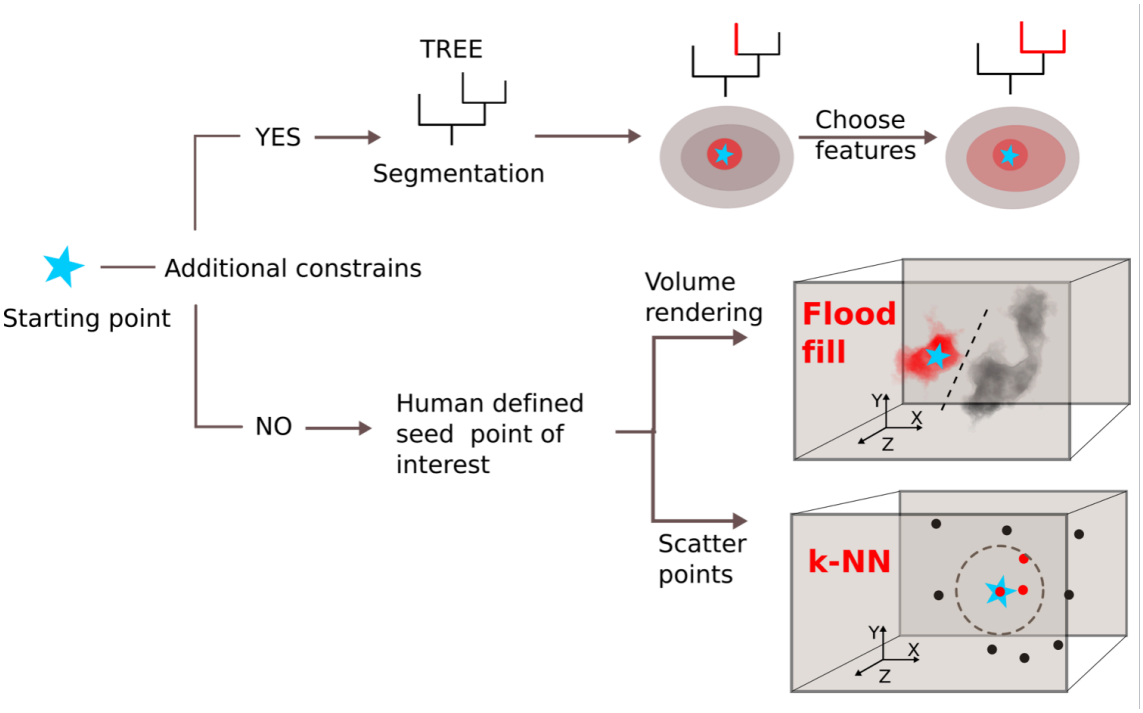

图 2.7: Glue 智能三维选取的设计方案概念图。

域有所区分，显示为不同的颜色或灰度差异，这些差异能够被程序量化，并通过设定阈值的方式将这种显示上的区分定义出来。填充算法是指定不规则区域内部像素填充为填充色的过程，常用的填充算法包括种子算法、扫描线填充和边缘填充，这些算法通过不同的填充方式完成对某个区域的颜色填充。对于我们的设计需求，即填充由种子点开始进行扩展及无需确定填充边界这两个条件，我们选取了种子算法（Floodfill Algorithm）作为引导算法。

种子算法或 Floodfill 算法，其思路类似洪水从一个区域扩散到所有能到达的区域而得名。该算法往往从多边形区域的一个内点开始，由该点沿一个方向延伸（深度优先）或沿多个方向延伸（广度优先），并不断迭代这个过程，直到所有的延伸都达到了设定的阈值，如下列伪代码所示。这个过程中，延伸方向的顺序并不影响最终结果，而阈值的大小将直接决定所扩散的区域的范围。

```
function Floodfill(x, y, z, threshold)
    if note[x][y][z].value > threshold:
        mark note[x][y][z] as visited
    if note[x][y][z].value <= threshold:
        return
```

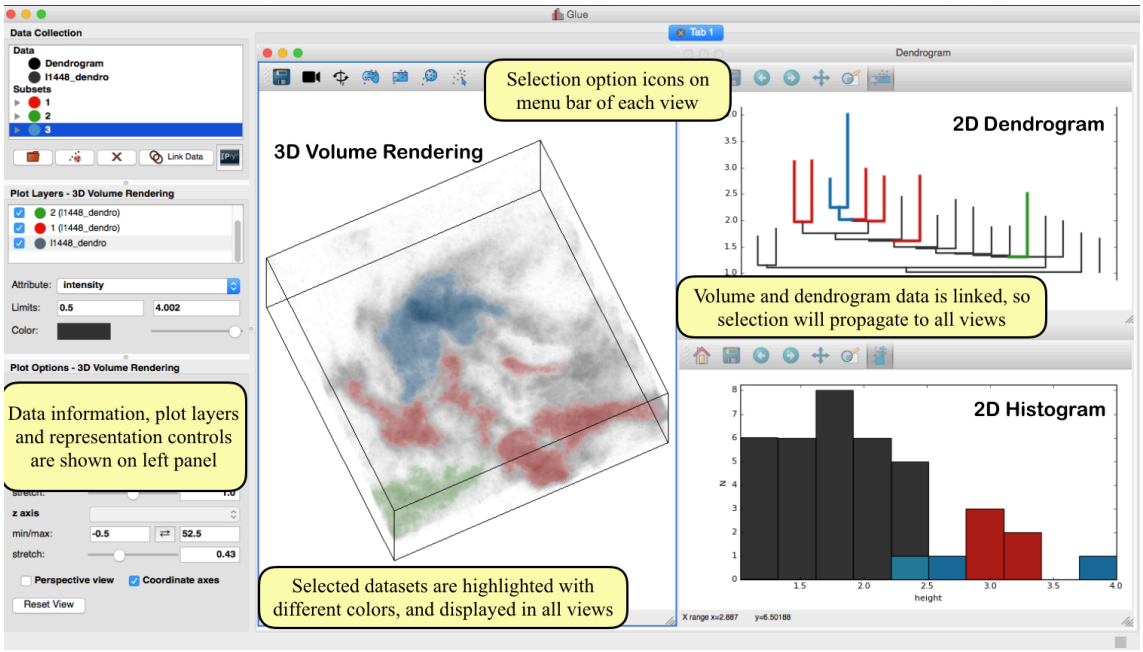

图 2.8: 该图显示了预定义结构（Dendrogram 树形图）下的三维选取与展示。

```
Floodfill(x-1, y, z, threshold)
Floodfill(x, y-1, z, threshold)
Floodfill(x, y+1, z, threshold)
Floodfill(x+1, y, z, threshold)
Floodfill(x, y, z-1, threshold)
Floodfill(x, y, z+1, threshold)
```

由于数据的动态范围千差万别，而鼠标移动的范围是有限的（最大距离为屏幕对角线距离），所以需要对动态范围进行正则化处理。假定数据集是由 n 个离散的数据点 $(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n)$ 组成的，其中 $\mathbf{p}_i$ 为 k-维向量。则其动态范围为 $[\min(\mathbf{p}_i), \min(\mathbf{p}_i)]$。正则化的作用是把该动态范围自适应的压缩到指定的区间 $[a, b]$ 中。这往往需要用一个指数的映射[51] 来完成：

$$\sigma : \mathbb{R}^K \to [0,1]^K$$
$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

正则化映射范围 $[a, b]$ 的选取需要根据数据渲染的需求而定，主要取决于用户屏幕的尺寸、分辨率等。同时，为了优化用户体验，我们应当尽量使得用户用最少的鼠标移动达到精准选取的目的。经过我们代入不同的参数数值，发现当正则化空间范围为 $[1.01, 101]$ 时，用户具备最佳交互体验，核心计算步骤如下：

$$t = m/s$$
$$t = 1 + 10^{(4t-2)}$$

其中，m 是鼠标移动距离，s 是屏幕对角线，t 表示阈值，且公式中的 t 正则化到范围 $[1.01, 101]$。

对于三维散点图，其数据特点便是每个散点都是一个独立的部分，而用户的选取可能是针对某一个散点对其属性进行查看，或者一群在空间分布特别的点群做一个统计学分析。对单个物体的点选已被验证能够通过基于光线投射（Ray Casting）的三维选取方法实现[52,53]，其基本原理是通过发射一条光线，根据该光线的位置信息来判断与其相交的物体。实现中，我们通过判断屏幕上的鼠标点选位置与数据点在屏幕上的显示位置之间的距离，将三维选取简化为二维视觉判断。虽然该方法的不足在于如果多个三维点在二维显示上有重叠，那么尽管在三维中它们的位置相差甚远，算法还是将它们都归类为同一个选取子集，从而要求用户对视角进行调整并重新选取，未来的改进中，我们会将添加对额外的条件，如比较子集中点源的深度信息来进一步限制选取范围。

对点群 (Scatter Points) 的三维选取，我们希望沿用类似体绘制 (Volume Rendering) 选取的种子类算法，也就是选取区域由种子点起始并发散。一方面是考虑到选取的单个点源可以作为种子，从而能够同时实现对单点和点群的选取，另一方面是用户对想要选取的点群与种子之间往往存在着某些特征上的共同点。而这类根据已知样本（例如种子）的某些特征，判断一个新的样本（例如想要选取的点群）属于哪种已知的样本类的问题，正是机器学习 (Machine Learning) 中的分类问题。通过对特征的定量描述并标签，我们便可以在监督下归类出选取的子集。

常用的分类 (Classification) 与回归 (Regression) 算法中，k-最近邻算法（k-Nearest Neighbor, k-NN)[54] 最为满足我们的需求：首先，k-NN 算法为非参数化算法 (Non-parametric Algorithm)，即对目标数据集的分布特征没有要求，该算法对数据集不做显性的理论假设（例如高斯分布等），这就避免了对数据认知不足时做出错误建模的危险。该特性在天文数据选取非常重要，因为天文学家往往是在处理数据的过程中学习数据本身，而难以在数据分析之前就对数据有足够的认知并正确建模。其次，k-NN 算法是基于实例（Instance-based）的学习或者是局部近似（Local approximation），且将所有计算推迟到分类之后的惰性学习 (Lazy learning)，也就是说，对选取集合以外的数据的计算会被延迟到数据被归入当前选取中，这也使得选取的范围更针对性，选取操作更加快速。

k-NN 算法的核心思想是给定一个训练数据集 $(\mathbf{x}, \mathbf{y})$，其包含数据的特征描述 $\mathbf{x}$ 以及待预测的目标特征 $\mathbf{y}$，对于测试样例 $(\mathbf{X}, \mathbf{Y})$ 中的每个实例，通过空间点集距离计算其邻近关系，在训练数据集中找到与该实例最邻近的 k 个实例，并通过多数投票法定义这 k 个最近邻实例的中占据多数比例的类别，并将测试样例归类为该类别。具体计算步骤如下：

---

k is the number of instances

D is the training set

for each $z \in (\mathbf{X}, \mathbf{Y})$:

    Calculate the distance d between z and each $(x, y) \in D$

    Select k elements from D having the smallest d to z

    $Y = \underset{v}{\operatorname{argmax}} \left(\sum_{(x_i, y_i) \in D_z} I(v = y_i) \right)$

这里，$|d|$ 的计算最常用的方法便是使用欧几里得距离 (Euclidean distance)，即：

$$\begin{aligned} \mathrm{d}(\mathbf{x}, \mathbf{X}) = \mathrm{d}(\mathbf{x}, \mathbf{X}) \quad &= \sqrt{(x_1 - X_1)^2 + (x_2 - X_2)^2 + \cdots + (x_n - X_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \end{aligned}$$

其中 $\mathrm{d}(\mathbf{x}, \mathbf{X})$ 表示在笛卡尔坐标系下的点集 $\mathbf{x}$ 和 $\mathbf{X}$ 的欧几里得距离，也就是特征空间中两点直线距离，这也是对两个点集相似程度的反映。

代码实现中，我们使用了 Python 机器学习软件库 scikit-learn[55] 提供的 k-NN 算法程序，实现了对用户选取的特征空间的训练样本的归类。该算法输入由特征空间中最接近的 k 个训练样本组成，而 k 值的选取对模型的预测有着直接的影响。如果 k 值过小，那么预测结果对邻近的实例点非常敏感，整个模型容易发生过拟合。相反，如果 k 值越大，学习的近似误差会增大，会使得距离实例点较远的点也起作用，致使预测发生错误。因此，与 Floodfill 引导算法相似，在我们的设计中训练样本数目 k 也是一个随用户鼠标拖动距离改变的变量，呈线性增长的对整个三维数据集进行遍历，而测试样例固定为种子点，每一次的用户鼠标移动都会对求得新的 k 个距离种子点最近的点群集合作为选取子集。我们也将算法调用的核心代码列举如下：

```
from sklearn.neighbors import NearestNeighbors
def knn_selection(x, y):
    # 初始化 mask
    mask = np.zeros(x.shape, dtype=bool)
```

```
    # 根据鼠标移动定义训练样本数目 k
    neigh = NearestNeighbors(n_neighbors=n_neighbors)
    neigh.fit(np.vstack([x, y]).transpose())
    # 获取距离种子点的 k 个最近邻点群的数据索引
    select_index = neigh.kneighbors([self.selection_origin])[1]
    # 标记选取区域
    mask[select_index] = 1
    return mask
```

基于算法引导的三维选取的工作流程图如图 2.9所示，该图给出了用户在选取过程中的操作及其对应的系统的响应。最初，用户将启动并点击按钮启动三维选取模式，此时系统将由普通浏览模式切换至三维选取模式，进行鼠标响应及选取算法的初始化；接着，用户通过鼠标输入设备在屏幕点选起始点，系统将自动选取视线方向上的数据密度最大的一个点作为种子点并保存；随后，用户拖动鼠标来定义选取范围，系统将根据鼠标移动的方向和距离实时计算阈值，该阈值与上一步保存的种子点信息将被运用到已经初始化的选取算法中，用于计算选取区域，计算出的选取区域也将由系统高亮表示，并存在数据子集 Subset 中；最后，用户退出选取模式，系统将自动切换回普通三维浏览模式，此时用户可通过系统的可视化显示设定版块对选取区域的展示效果进行调整。

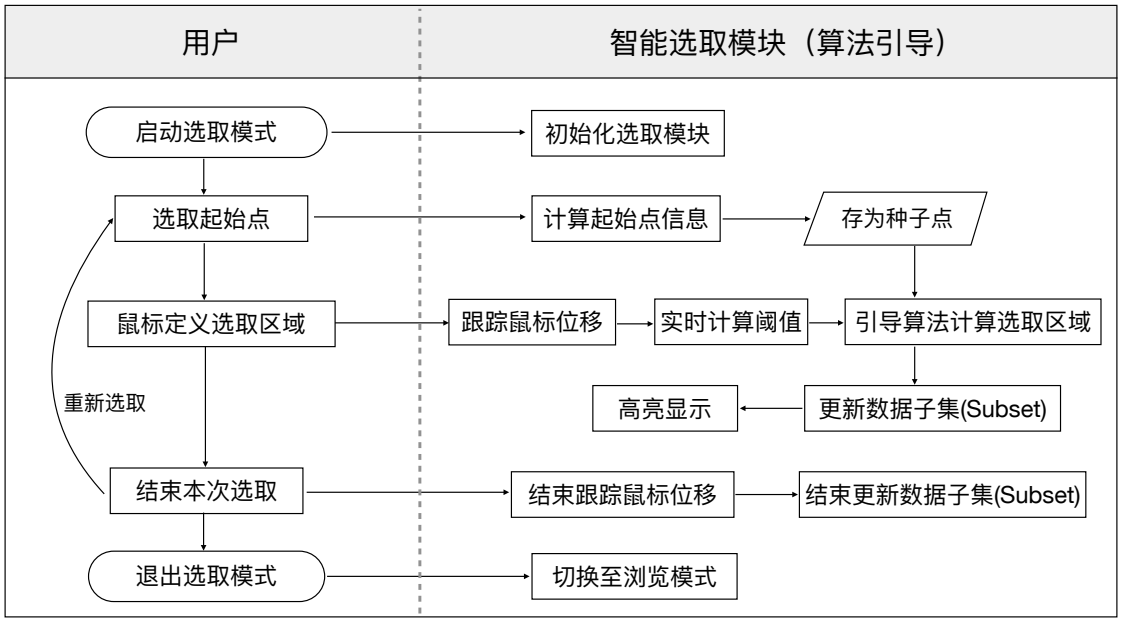

图 2.9: Glue 三维智能选取模块算法引导方案的工作流程图。

2.1.3.3 项目开发与实现

由于三维视图互联功能是基于 Glue 程序原有框架基础上的开发，因此我们需要对程序原有架构和接口类型进行理解，并对相应功能模块进行针对性开发与扩展。Glue 核心代码功能块可以分为四大块：数据模块 (Data Framework)、选取模块 (Selection/Subset Framework)、连接模块 (Linking Framework) 以及交流模块 (Communication Framework)，这四个模块协同合作，保证了 Glue 的可视化选取功能的实现，并完成了数据间的交流与连接。

- *数据模块*：该模块是 Glue 的动态数据结构和相关数据管理功能函数的集合。Glue 是支持多数据集探索的可视化软件，因此，用户在使用 Glue 的过程中往往涉及到多个数据集，这些数据集或来源于用户加载，或通过使用过程中交互创建的新子集，这些数据集都在 Glue 的数据模块下统一管理。Glue 采用树层次结构来直观的展示并管理这些数据集，由图 2.10所示。加载到 Glue 中的每个数据集以及通过操纵创建的子集都存储在称为数据对象 `Data` 的数据容器中，每个对象包含着任意数量的 n 维数组，这些 n 维数组中存储着实际数据，类型统一设置为 `Numpy Array`，数据尺寸因数据内容而异，通常从 2D 图像阵列到 1D 表格阵列不等。这些数据对象 `Data` 都收集在名为 Data Collection 的 “筐子” 中，也就是图中所示的树型存储结构的根。

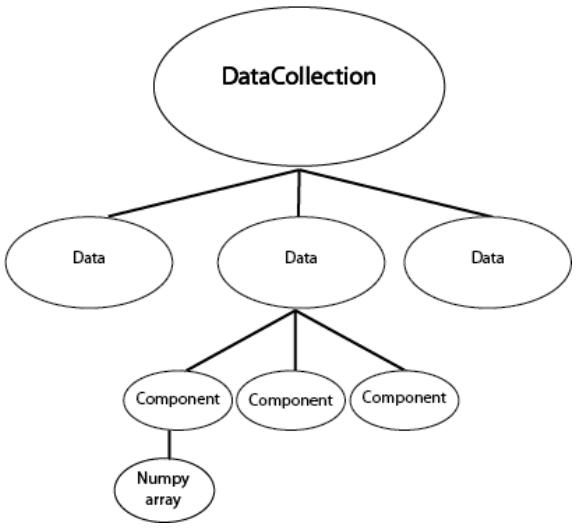

图 2.10: Glue 数据管理的树形结构图。图片来源:glueviz.org

- *数据连接模块*: 该模块实现了 Glue 数据集之间逻辑运算。Glue 的一大优势便是能够将多个不同的数据集连接起来，这个连接的建立是通过数据连接模块对数据集之间的连接完成的。不同的数据 `Component` 之间是由 `ComponentID` 标识的，通过这些唯一标示，我们便能够提取目标数据集，并在需要连接的

数据集间插入连接方程 (Link Function)，用来定义连接的方式，具体实现及逻辑图见第 2.1.2节和图 2.3。

- 视图交流模块：该模块实现了 Glue 的不同视图之间的通信功能。由于数据的复杂性和用户需要的多样性，Glue 需要支持多种不同的视图，如一维的直方图、二维散点图、三维体绘制等。这些视图所需的渲染计算量和实现方式各异，因此 Glue 将每种视图抽象成为一个显示模块，以屏蔽它们之间的差异以及内部的复杂性。为了实现视图互联的功能，不同的视图之间需要相互通信。Glue 使用了 Publish/Subscribe 的软件范式（见第 2.1.2节以及图 2.3）来实现这一功能。当一个事件被触发时，相关的视图就将事件以信息的方式发送到了 Hub。Hub 则负责将事件广播到所有订阅了该事件类事件的其它视图，这样不同视图之间就能实现基于事件触发的数据通信和同步，而无需相互了解对方的内部实现集中。

- 选取模块：该模块实现了 Glue 在用户操作过程中选取兴趣区域的功能。其工作流可被描述为：用户在视图中做出的选取操作将首先被解析为兴趣区域 (Region of Interests, ROI)，也就是对一个几何选取区域的抽象，紧接着，兴趣区域 (ROI) 由程序解析并转换为规范化描述，也就是子集状态集 (Subset State)，这些子集状态集将被映射到 Data Collection 中的各个数据集中，对应每一个数据集都将产生一个对应此次选取操作的子集集合，如图 **??**所示。

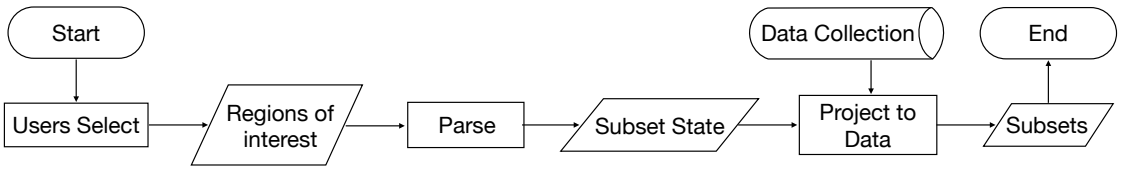

图 2.11: Glue 的选取模块流程图。

Glue 高度模块化的框架，也使得新功能的开发十分直接。由图 2.12可以看出，对新的 `Subset` 映射关系的定义并不需要对全局的代码框架进行修改，而只需对 `SubsetState` 下的子类型进行开发并嵌入到代码框架下。同样的，对 Glue 三维视图的功能开发也是针对 Glue 框架下的部分模块进行了增加，以插件 (plug-in) 的方式嵌入到软件主框架下。

整个的三维视图模块架构示意图如图 2.13所示，该图给出了 Glue 三维视图互联功能的代码实现模块及其与图形化界面之间的结构关系。图中中央（白底框图）部分是 Glue 的图形界面简化图（下面简称简化图），我们通过四周的两种底色框图

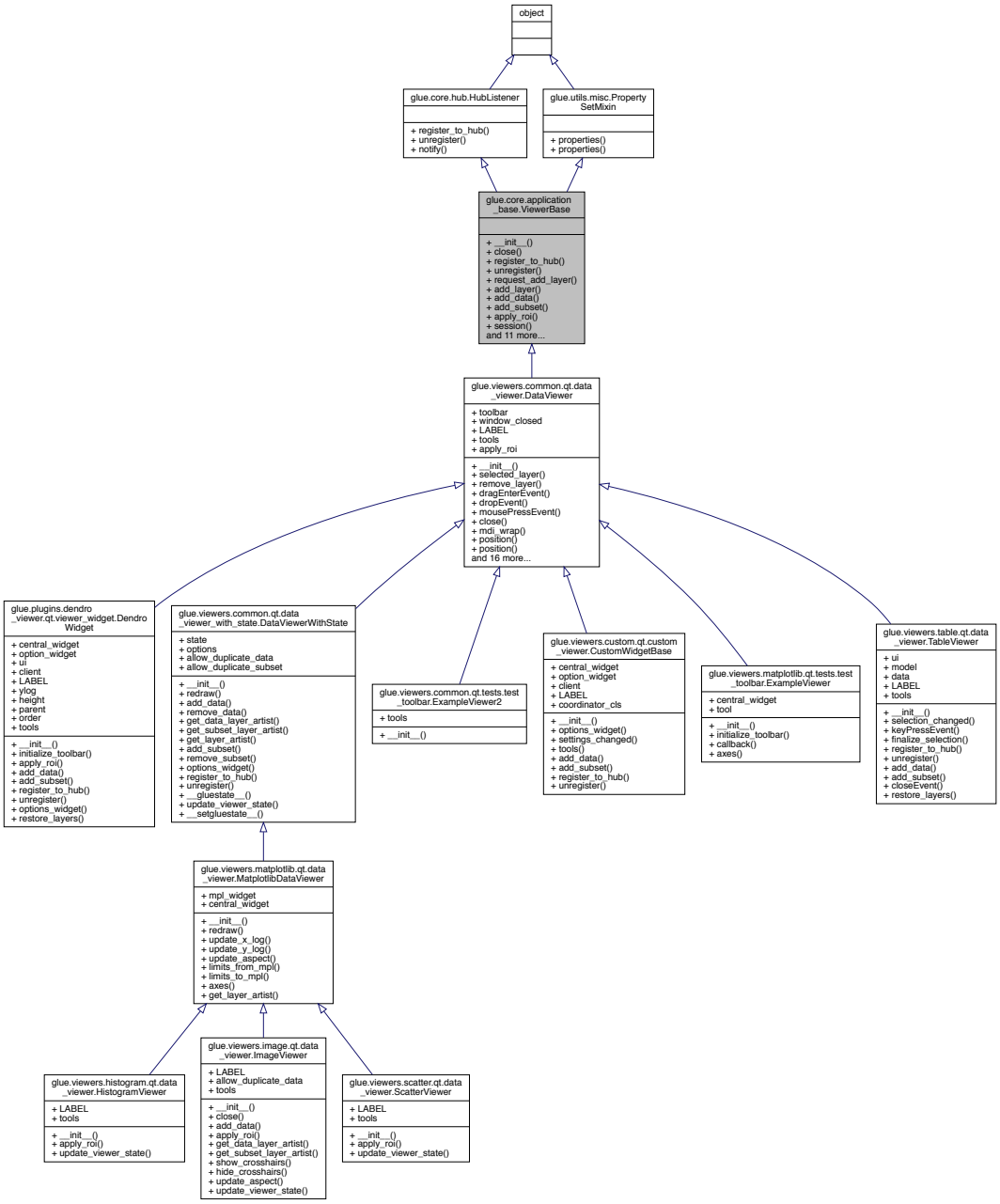

图 2.12: Glue 的 Subset 模块的 UML 图。所有的 Viewers 都建立在一个抽象的 `DataViewer` 基础上，而该抽象 DataViewer 继承了 `ViewerBase` 抽象类。由于 `ViewerBase` 继承了 `HubListener`，因此所有的 Viewers 都具备了 `HubListener` 的属性，可以实现注册 (register) 和通知 (notify) 的功能。

（蓝橙）对对应功能模块进行了分类介绍，实线箭头表明功能模块与响应的图形界面的对应关系，虚线箭头表示功能模块之间的从属或调用情况。简化图左侧的两个视图控制面板，分别对显示（例如颜色、透明度）和属性（例如拉伸比例）进行交互式控制，我们通过 PyQt 库[10]，也就是 Python 对 QT 图形界面编程语言的接口封装库，完成三维可视化的显示参数设定的用户界面及响应功能实现。简化图右侧是核心显示视图区，视图插件（Viewer Widget）、画布（Canvas）及渲染结果对象（Rendering Result）实现了三维渲染功能，通过 `layer_artist` 和 `vispy_widget` 两个类协同合作，分别负责实现场景的准备（例如坐标轴、背景颜色等）和三维渲染计算，该三维渲染计算主要通过 Vispy 绘图库对 OpenGL 进行直接调用，例如对单个对象的体绘制，以及部分再开发，例如对多个对象在同一视图的体绘制。视图区中部的工具栏模块（Toolbar），则负责视图区的附属功能实现，例如自动旋转操作即渲染结果匀速绕某坐标轴匀速旋转、编辑操作例如返回上一步操作和存储操作等，而三维选取功能也正是由该模块接入。具体的代码文档及技术实现步骤可参考 Glue-Viz-Viewers 页面[11]。

根据此框图，我们不仅实现了对 Glue 原有代码框架的功能定义与模块分类，建立了代码模块与图形界面功能块的联系，而且也为其它后续基于 Glue 的开发工作，例如 FAST 谱线可视化分析模块的开发提供了直接参考。通过三维展示与三维选取的功能实现，我们将三维视图互联由理论延伸到实际，真正提供了能够为天文学高维度数据进行可视化分析的工具，通过该工具，我们还能够在未来对三维视图互联分析方法的功能与性能进行进一步的量化研究，并可扩展到对天文以外的领域的应用性进行探讨等。

2.1.4 三维视图互联的验证与总结

本节通过天文学研究不同领域的两个示例，分别使用观测数据及数值模拟数据，来对 Glue 的三维视图互联功能进行验证。

应用案例一：银河纤维状结构查找

在银河系的星际介质中，纤维结构存在于各种规模尺度上，但是近年来才有天文学家开始发现和分类银河系最长、最高对比度的长丝，并发现它们中的一些与银河系悬臂特征有潜在的相关性[46,47]。大尺度纤维的星际介质研究可以从 Glue 的三维可视化特征中受益，因为它们的分子气体可以在 Position-Position-Velocity 的三维空间展示，其中两个位置表示气体在平面上的位置天空，速度是其通过频谱特征的多普勒频移获得的相对于太阳的速度。了解作为位置的函数的速度结构是至关重要的两个原因。首先，速度结构的连续性表明，纤维结构是单个连续特征，

[10]http://pyqt.sourceforge.net/Docs/PyQt5/

[11]https://github.com/glue-viz/glue-vispy-viewers/

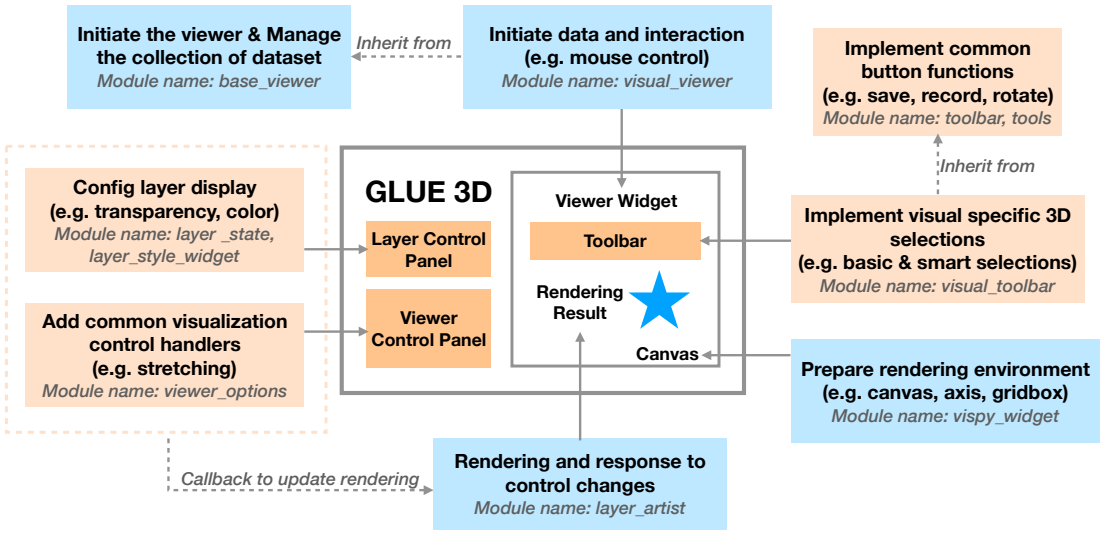

图 2.13: Glue 三维功能实现的代码架构示意图。

而不是非相关特征的叠加，因为气体的速度与其与银河中心的距离之间存在直接关系。第二，可以将纤维结构上的速度梯度与已知旋臂的速度梯度进行比较，以确定纤维结构是否确实跟追踪较大的旋臂结构。这种运动学分析可以与更传统的 2D 数据产品（例如尘埃发射 Dust Emission 和尘埃消光 Dust Extintion）相结合，以得到纤维结构的密度结构和形态的更完整的轮廓。

通过 Glue 的强大的三维视图互联功能，我们能够识别和表征银河系最密集的大规模纤维状结构。我们首先通过使用多波长方法识别二维图像中的纤维结构，通过将银河面上的二维 GLIMPSE[56] 8 μm（中红外）图像和 Hi-GAL[57] 500μm（远红外）图像相关联。因为纤维结构主要由致密的气体和灰尘组成，它们在中红外波长处于消光状态，因此在银河平面中显示为丝状“阴影”（如图 2.14顶部面板所示）。然而，纤维结构中的灰尘也吸收和再辐射较长波长的光，因此它们在远红外波长处显示为明亮的丝状特征（图 2.14上图）。一旦我们识别出两种波长的丝状形态并通过 Glue 的选取工具在二维图像上进行选取，Glue 变会自动将二维区域（绿色显示）传播到三维速度立方体的立体渲染（图 2.14下图）。然后，我们通过 Floodfill 选取工具提取属于纤维结构的分子发射（图 2.15上图），以确认我们选择了一条连续的丝状结构（图中以蓝色显示 2.15下图）而不是两个不相关的部分。一旦确定了邻接性，我们通过切割通过绿色二维选取的路径来提取自定义的位置-速度切片（PV Diagram）（图 2.15下图顶部面板中的红线）；这类似于在红色突出显示的像素上沿着速度轴塌缩，这产生了一个显示速度作为沿着红色曲线的位置的函数的图（图 2.15下图）。然后可以将跨过纤维产生的梯度输出并与已知旋臂模型的位置 - 速度迹线进行比较，以确定与旋臂特征的潜在相关性。该过程在所有银盘附近数

据上重复，以确定大尺度的银河系纤维结构的可行性样品。

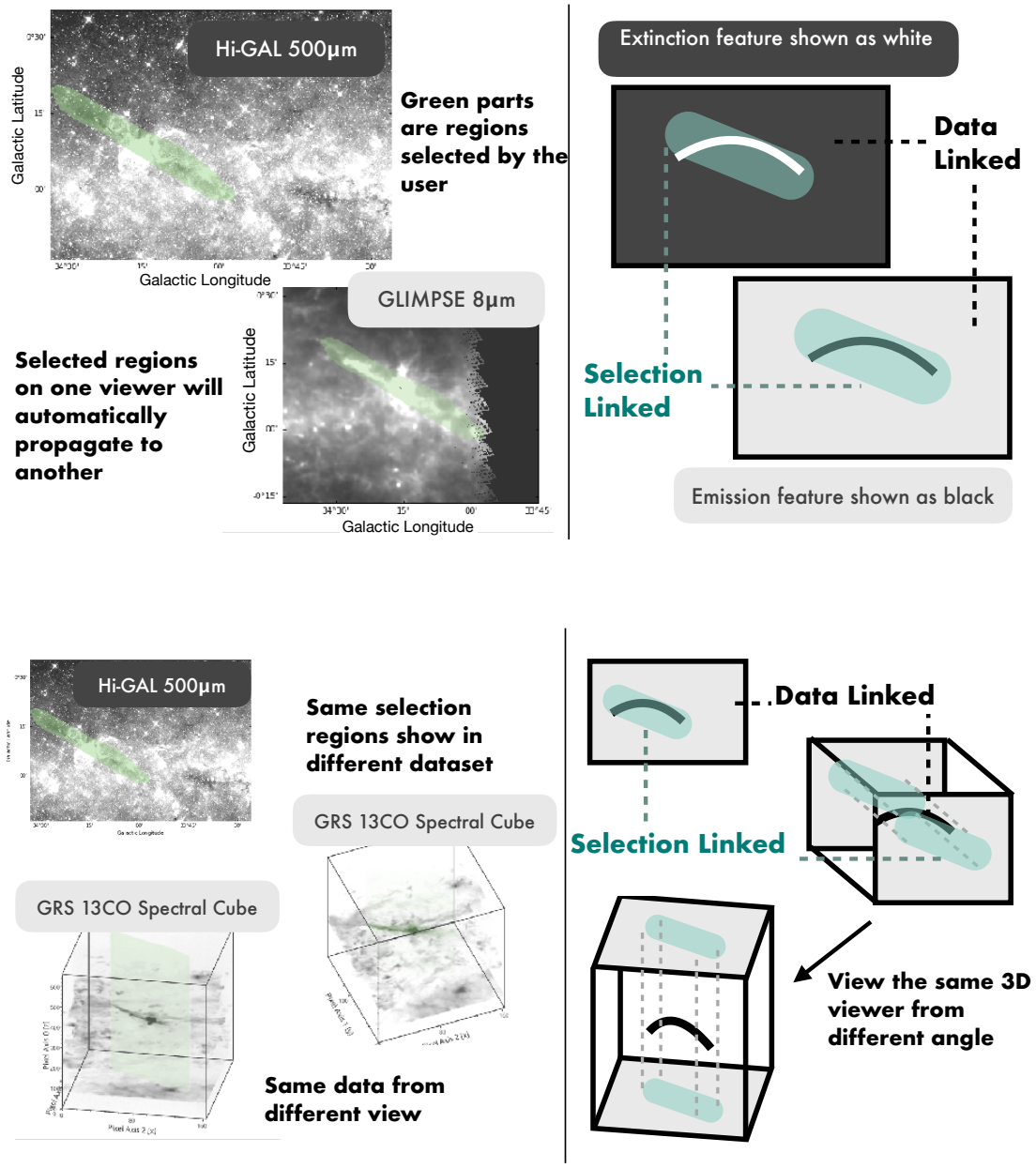

图 2.14: 通过一系列组图说明使用 Glue 软件找寻银河系内纤维结构的具体过程，图左给出的是在 Glue 中展示的观测数据和操作截图，右侧通过生动的动画简图进一步分解操作。上图：识别和选择二维红外图像上的纤维结构特征；下图：将 2D 选取区域传播到 3D 数据立方体的体绘制显示中。

应用案例二：研究恒星质量黑洞在星团中的演化和动力学关系

Glue 支持时间序列分析，这对于分析天体物理数值模拟数据非常。如图 2.16所示，Glue 被用于对一个球状星团数值模拟的可视化分析中。该星团的形态通过一

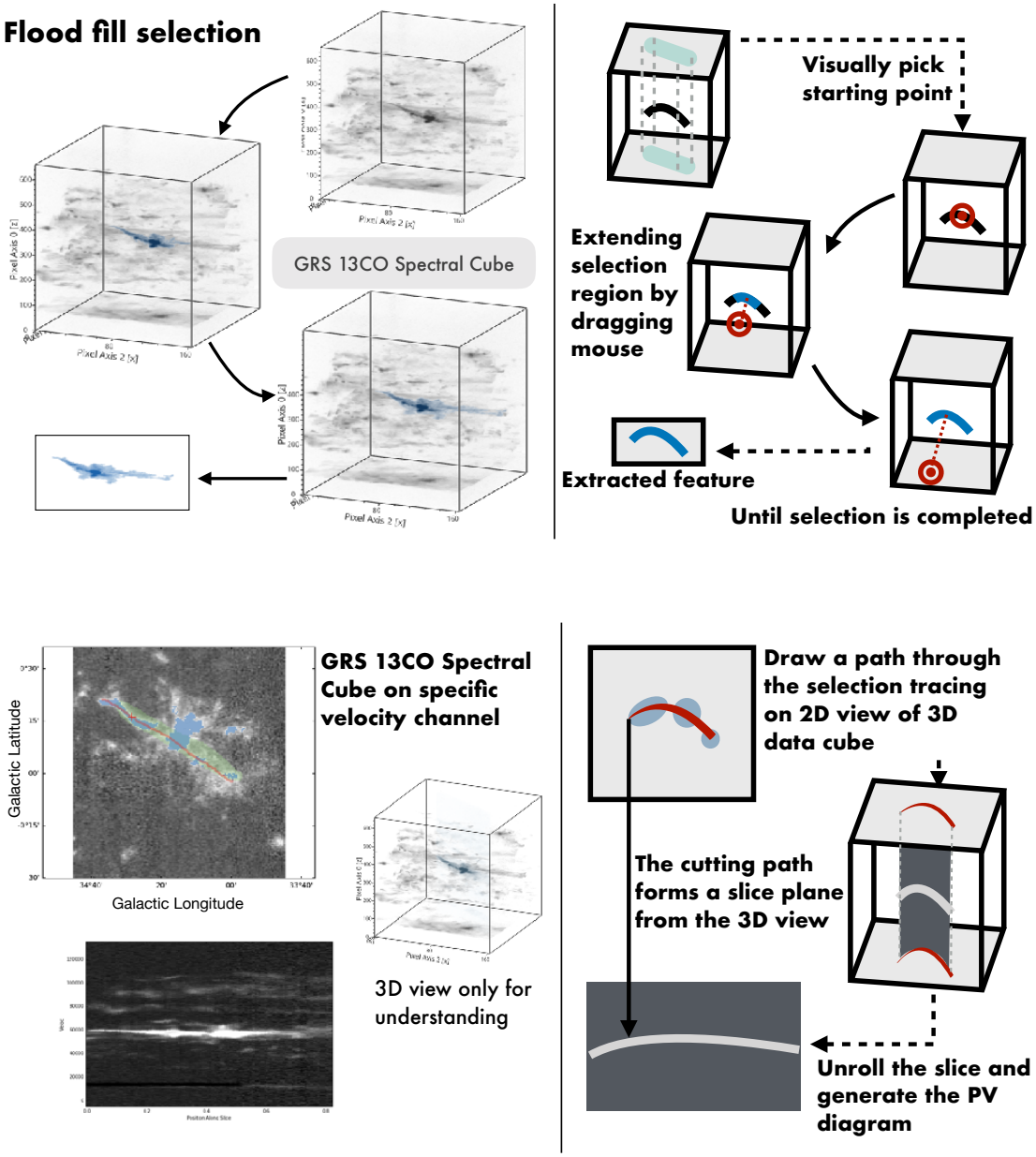

图 2.15: 上图：使用三维选取提取长纤维结构内的连续分子发射（Molecular Emission）；下图：提取位置-速度图（Position-Velocity Diagram）。

个交互式的 3D 散点图显示处理。用户可以旋转，缩放该视图，并可以利用鼠标选择其中的一颗或多颗星。恒星的颜色代表其温度，而大小代表其质量。通过拖拽屏幕顶端滑动条，用户可以观看该星团动力学演化和恒星演化的全过程。这就相当于用户利用 Glue 作时间旅行。当用户拖动时间轴滑块时，所有的视图都相应的得到更新。

在这个案例中，Glue 被用来寻找该星团中的恒星级黑洞。恒星级双黑洞系统是

极为有意思的系统，因为它们的并合可能产生引力波（例如 Abbott, et al.(2016)[58]）。最初，该星团中所有的恒星均为主序星。但由于该星团的恒星质量由一个初始质量函数定义，星团中大质量恒星演化非常快，以至于在数值模拟即将结束的时候变成了黑洞。在该数值模拟中，天体的分类用变量 KSTAR 定义：KSTAR = 0 表示为主序星，而 KSTAR = 14 表示是黑洞。从直方图可以看出，当前系统已经产生了两个黑洞。为了研究这两个黑洞的空间分布情况，用户在直方图上选择了 KZ = 14 这个直方。于是，3D 散点图上所有 KZ = 14 的物体均被高亮显示，而数据表上相应的记录也显示为高亮，方便用户获得实际数据。

该数值模拟是利用之间 N 体代码 NBODY6++GPU[59–61] 完成的。数据的输出标准定义在 Cai, et al. (2015)[62] 一文中。

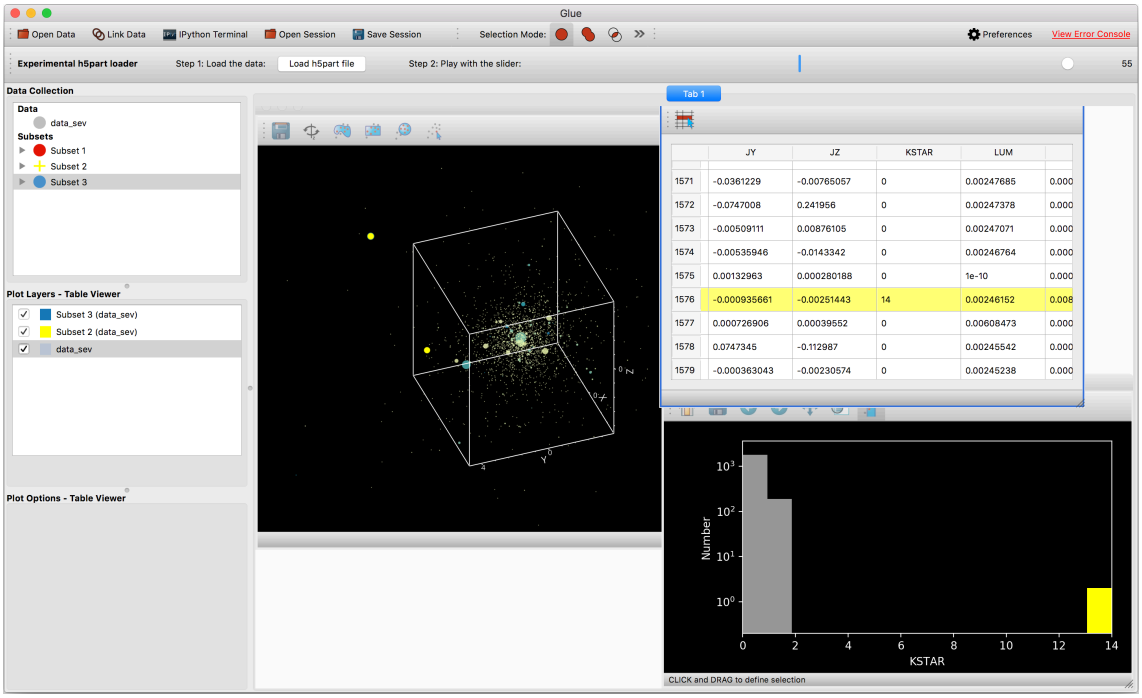

图 2.16: 通过 Glue 进行数值模拟对星团演化结果的演示。

据我们所知，天文领域中没有一个工具能够支持多数据集且多维度下 (1D/2D/3D) 的视图互联探索，且同时支持丰富的天文数据类型。Glue 是目前所知的第一个，也是唯一一个能够支持多数据同时三维可视化探索的工具。三维视图互联功能的实现，使得天文学家有了一个新的探索高维度天文数据的手段，通过以上两个应用案例，在银河纤维结构案例中的发现数据中新的结构特征，以及在星团演化案例中跟踪目标天体的演化规律，我们也可以看出，天文学家通过使用 Glue 的多样化三维展示以及灵活的三维选取功能，在短时间内便能够对多个数据集进行关联比对，并发现新的科学价值。这种从数据中挖掘科学价值，即科学数据探索，也正是 FAST 数据处理的重要目标之一，而基于 Glue 的 FAST 可视化分析模块的设计与实现，我们将在下一节详细的展开。

2.2 针对 FAST 的可视化分析模块设计

FAST 可视化分析模块不仅是对 FAST 工程需求的一个直接响应，也是天文大数据跨领域合作的一次创新。本章节首先对 FAST 可视化分析模块进行了一个具体的需求分析，重点考虑到中性氢巡天这类对大数据和可视化要求较高的研究需求；其次通过对比天文领域其它可视化工具给出了选取 Glue 作为模块软件开发框架的原因；其后结合具体需求和 Glue 的特点设计了 FAST 可视化分析模块的框架结构，并对其中的几大创新点进行分析和可行性评估；并给出了部分实现的示例及具体可行的解决方案；最后对开发过程中的代码文档管理做了总结。

2.2.1 大数据背景下的需求分析

Hassan & Fluke (2011)[7] 给出了天文数据本质与其展现手法的一个关联图（如图 2.17所示）。根据此图，天文数据可以分为五大类：二维图像数据；目录列表（主要是从处理图像数据确定的次要参数如坐标，通量，尺寸等）；光谱数据和相关产品（包括一维光谱和三维数据数据立方体，从红移获得的距离数据，源的化学成分等）；时域研究数据（包括移动物体的观测，需要在不同时期进行多次观测的可变和瞬态源）；理论研究的数值模拟，包括诸如空间位置，纬度，质量，密度，温度和粒子等属性类型，这些属性可以通过使用 “快照” 输出以明确的时间依赖性呈现。在 FAST 的科学目标中，这五类天文数据类型都很有可能被使用。

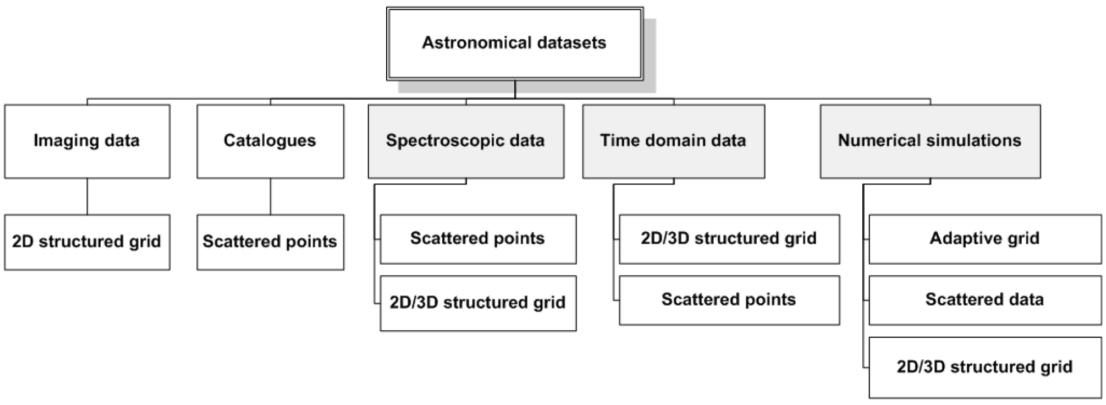

图 2.17: 天文数据的本质，展示了数据来源之间的映射[5] 和数据表示[6]，每个来源对应了最常见的数据表示。图片来源：Hassan & Fluke (2011)[7]

对应这五种类型的天文数据，可视化表现手法可以归类为：

- 散点：数据由一组点位置（$x\ y\ z$）和相关联的数据属性（例如密度，压力和温度）组成。
- 结构化网格：数据值定义在规则三维网格上，网格单元与笛卡尔轴对齐。

- 非结构化网格：数据值定义在非规则三维网格上，通过 2D/3D 形状元素的角上指定数据值。

- 自适应网格：数据值在多分辨率结构化网格上指定。粗网格用于覆盖与叠加子网格结合的整个计算域，为感兴趣区域（例如，粒子密度最高）提供更高的分辨率。

对 FAST 可视化分析系统的可视化功能而言，能够支持上述五类天文数据类型以及能够实现上述四种可视化表现手法，是从实用性考虑的基本要求。我们考虑到 FAST 工程仍处在调试观测阶段，最终数据格式和存储方式等都存在变化的可能，因此，该分析系统还需要具备高度灵活性和可扩展性，能够实现功能的增减，包括对特定数据格式的读取、增加数据分析的具体功能等。

我们还考虑到，FAST 预计每小时产生 5TB 的原始观测数据。假定每天观测时间为 8 小时，则每天将产生 40TB 的数据。FAST 不仅数据量大（每小时产生 5TB 的原始观测数据），而且数据复杂，因此只依靠流水线是不够的，往往人为介入手动分析是必须的[63,64]。Punzo et al. (2015)[65] 指出，数据可视化在射电天文中主要用在以下三个方面：找出自动数据处理流水线产生的错误或瑕疵；寻找目标源并定性地检视它们；对目标源进行量化和比较分析。但是实际中，让天文学家直接人工检视每小时 5TB 的原始数据是不现实的。这需要一个高效实时的流水线，以将数据量减少到可以人工处理的程度。关于自动流水线调度算法的开发，参加本论文第四章。这里假设流水线能自动将数据量减少到可以视觉探索的程度，那么为了实现有效的视觉探索，系统还必须满足以下需求：

- 数据检索与加载：支持对本地及远程数据的快速检索，通过指定感兴趣的目标源或天区，相应的数据（而不是全部的数据）就会自动加载并现实出来。从远程检索数据，也是对 FAST 数据存储与处理的地域限制问题的响应。FAST 计划在贵州成立大数据中心，也可能利用云服务进行部分数据存储，这些数据将被北京科学总部或各地科学家使用，因此在数据收集过程中对远程数据的快速获取将是整个数据分析的基础。

- 数据互联：同一数据集的不同维度以及不同数据集能够直接相互关联或交叉证认，对于 FAST 而言，多波段观测将对同一个目标源产生不同频段的数据，而这些数据进行分析时也可能与同源的其它观测数据进行比较；

- 数据展示与交互：加载的数据能够以不同的可视化方式展现给用户，且能够满足常用天文数据类型及用户自定义的展示需求。当天文学家对某个目标源感兴趣时，系统应当支持相应数据的选取，使得用户能够在不同的视图区，针对不同的选取任务进行灵活性选取，包括对三维展示的选取；

- 数据分析：能够满足常用数据分析需求如求平均值、对某一维度的数据取对数或者作曲线拟合等，也能够自定义分析算法；数据分析往往需要不同的软件用于完成不同的数据处理任务，为了完成一个完整的数据处理流程，往往需要结合多个软件的优势。这就要求这些软件之间遵循一定的协议，从而使得数据可以在它们之间传递。或者使用云服务的优势，通过云端高性能计算资源及完备的软件环境对某些计算复杂性任务进行处理。

2.2.2 可视化软件对比及选取

天文学领域存在各类可视化软件，在 Hassan & Fluke, (2011)[7] 就曾对天文领域所使用的可视化软件工具包的功能进行汇总，我们对可视化交互功能进行进一步细分，并结合近年来逐渐被天文领域应用的软件工具做了一个系统化调研，参考天文可视化软件比对表[12]，其简要版可见图 2.18。图中每一行是对可视化分析软件所具备的功能的统计，主要分为几大部分：互联功能、展示功能（二维及三维）、选取功能（二维及三维）、远程链接、性能（如 GPU 加速）、展示终端、与其它软件的交互（如 Hub 整合）、自定义性和主要编程语言等。绿色背景箭头表示具有该功能，红色叉号表示不具备，黄色波浪号表示有限条件下具备，图中统计了包括 Glue（目前及未来）以及其它一些天文领域常用的分析工具例如 DS9、TOPCAT 和 CARTA[13]，以及其它领域的但是可以被天文所用的软件例如地理学 ArcGIS[14]和工程科学 Igor[15]，图中也包含了一些数值分析常用的商业化软件如 DataDesk[16]、Tableau[17]、Filtergraph[18]、Plot.ly[19]和 Spotfire[20]，以及部分通用编程语言包例如 IDL 和 Mathematica[21]。

通过这些调研及总结我们可以看出，除了使用特定编程语言进行绘制（如 IDL），仅有高度可扩展的 Glue 软件可能实现所有天文常用的数据展示形式，甚至包括用户自定义的例如数图展示。此外，大部分可视化软件欠缺灵活的交互能力，三维选取功能更是欠缺，仅有 DataDesk 和 Spotfire 实现了基于形状的三维选取，智能化三维选取功能例如能够通过上下文联系判断选取区域没有一个软件具

[12]https://docs.google.com/spreadsheets/d/1NvvS471KSMir26P3ANyQnvn8n-j6IJLdlAAijMA2dD0/edit#gid=709215775

[13]https://github.com/CARTAvis/carta

[14]https://www.arcgis.com/features/index.html

[15]https://www.wavemetrics.com/index.html

[16]https://datadescription.com

[17]https://www.tableau.com/

[18]https://filtergraph.com/

[19]https://plot.ly

[20]https://spotfire.tibco.com/overview

[21]https://www.wolfram.com/mathematica/

备，相比之下，Glue 不仅实现了三维的基于选取形状的选取方式，更是开发了利用结构信息或者算法辅助等“智能”三维选取，这不仅在天文领域中有着直接应用，也是对三维选取的一个创新。此外，程序的可扩展化也是 FAST 可视化分析系统开发的一个重要因素，图中仅有 IDL、Mathematica 和 Igor 具备用户自定义功能，而 IDL 与 Mathematica 并没有图形界面，对编程能力要求较高，对于天文学家而言使用并不直接，而 Igor 支持的数据格式有限，且其主要用途是为论文发表制作二维渲染结果图片，而不是实时的探索。

此外，视图互联是可视化分析的重要元素，我们对图 2.18中给出的具备视图互联的软件做了汇总，并发现它们大多也存在着局限性：

- **DataDesk:** 1986 年面世，针对表格数据，不支持二维图像和三维体数据单元，支持视图互联但是不支持多数据集间的视图互联。价格：799 美元一个 license。
- **Spotfire:** 1996 年面世，能够对同一个数据集采用直方图、散点图、线图和三维散点图等不同的展示方式，但是并不支持二维图像及三维体数据单元，也不支持数据互联。价格：电脑端 650 美元一年，云端 2000 美元一年。
- **Tableau:** 2003 年面世，支持表格数据的视图互联，对大数据计算友好，但是输入数据格式较局限，无三维可视化功能。价格：999 美元一个 license。
- **TOPCAT:** 2003 年面世，针对天文领域的表格数据，能够通过直方图和散点图进行可视化，支持单数据集的视图互联，但是不支持非表格数据。价格：免费。
- **ArcGIS:**1999 年面世，针对地理学科的地形图和地图数据，能够采用直方图和二维图的展示方式，也能够进行网页端的展示，但是不支持三维可视化及交互功能。价格：平均一个用户一年 175 美元至 500 美元（取决于套餐）。

相比之下，Glue 不仅开源免费，还具备一系列独特且新颖的特点：数据互联：多数据集间的不同属性通过设定或转换能够在同一视图中进行展示；视图互联：在一个视图中进行的选取能够映射到其它的显示视图中；高度自定义化：用户不仅可以通过内置的 iPython 终端对数据进行定义与显示，还能够通过 Python 编程自定义用户界面及功能模块; 多样化三维交互及展示。这些特点也正是对 FAST 可视化分析模块需求的直接相应，并且 Glue 是一款基于天文需求所开发的软件，支持常用的天文数据格式，例如 `HDF5`[22]、`FITS`[21][23]和 `VOTables`[24]，而这些基本元素在大多数通用非天文可视化软件中都没有实现。

[22]https://www.hdfgroup.org/

[23]https://fits.gsfc.nasa.gov/

[24]http://www.ivoa.net/documents/VOTable/

Glue 对大数据有一定的支持，基于 Matplotlib 绘图库进行二维可视化，Glue 能够支持 10^6 行表格数据，或 10^9 个像素或体素[66]，通过测试发现，Glue 能够较流畅的可视化 16MB 的双列数据。而在三维可视化方面，由于绘图库基于 OpenGL，大部分渲染计算的工作都由图像处理器（GPU）完成，这使得实时渲染数百万体素的数据成为可能，其实际性能取决于用户的硬件配置，如 CPU 性能及 CPU 与 GPU 间的传输速度。

因此，介于 FAST 可视化分析模块的需求分析，并对比天文领域的可视化软件功能，考虑到 Glue 的高度可扩展性和对天文数据的高度兼容性，并能够针对 FAST 数据量大和数据类型复杂两大问题的特点，初步选定 Glue 作为 FAST 可视化分析模块的基本框架软件，在其基础上进行进一步开发。

View this table online at tinyurl.com/viz-features

	Glue (now)	Glue (planned)	Tableau	DataDesk	Spotfire	Plot.ly	Igor	Filtergraph	IDL	ds9	TOPCAT	Mathematica	ArcGIS	CARTA
linked views	✓	✓	✓	✓	✓	✗	~	✗	✗	✗	✓	~	✓	✗
linking data files	✓	✓	~	✗	✗	✗	✗	✗	✗	✗	~		~	✗
standard 2D plots (e.g. x-y, histograms)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗
custom 2D plots (e.g. trees)	✓	✓	~	✗	✓	~	✓	✗	✓	✗	✗			✗
image display (formats)	✓ (astro, med, GIS)	✓	✗	✗	✗	✗		✗	✓	✓	✓ (tabular data)	✓ (astro, ...?)	✓ (GIS)	✓
2D selection	✓	✓	✓	✓	✓	✗		✓	✓	✗	✓			✗
2D selection (smart)	~	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
3D point cloud plots	✓	✓	✗	✓	✓ (Desktop)	✓			✓	✗	✓	✓	✗	✗
3D surface renderings	✓	✓	✗	✗	✗	✓	✓		✓	~	✗	✓	✗	✗
3D volumetric views	✓	✓	✗	✗	✗	✗			✓	✓	✓			✗
3D selection (basic)	✓	✓	✗	✓	✓				✗		✗			✗
3D selection (smart)	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
(access to) statistics & fitting	~	✓	✗	✓	✓	~	✓	~	✓	✗		✓	~	~
remote data access	✗	✓	✓	✗	✓				✗	✓	✓			✓
GPU acceleration	~	✓		✗	Not found				~		Not found			✗
web dashboard	✗	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	~	✓	✓
Hub integration (e.g. SAMP)	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
custom widgets	✓	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗
web output	✓	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	✓	✓	✓
scripting options	Jupyter	Jupyter			R						No			Python
primary development language(s)	python	python, javascript	C++		S+	d3, javascript					Java			C++ & Java!

图 2.18: 该图给出了 Glue 与其它数据可视化软件进行功能比对的结果。

2.2.3 可视化模块框架设计

结合上述需求分析和 Glue 的特点，我们进一步细化了 FAST 可视化探索需求的方式，并在图 2.19给出了 FAST 可视化分析模块的工作流程图。基于天文学家在可视化分析中的操作模式，整个工作流可以分为四块。

- 首先是数据收集，即通过检索等操作从远端或者本地数据库中快速提取目标区域的数据，并加载到系统中。用户在打开可视化分析系统后，系统将首先进行初始化，调用所用到的软件包及插件，并提供初始的可视化图形界面和展示区域，为后续的操作做准备。其次，用户将定义数据需求，包括数据类型、数据范围和数据来源等，这些信息都会被系统解析并进行相应的格式转换，并最终加载到系统中以数据列表的方式存储并呈现给用户；

- 接下来是数据互联，即用户对加载的多个数据集之间的关联进行定义。我们的可视化分析系统的一大特点便是能够同时对多个数据集之间的关联进行可视分析，例如对同一天区的不同波段的成分信息进行关联分析。因此，我们的系统首先提供一个图形化窗口供用户直观的对不同数据集的属性进行关联，由用户定义逻辑关系，其次由系统通过逻辑关系在数据集间建立关联，并将这些关联以范式的方式列表显示出来。当然这一步骤并不是必须的，当用户仅加载单个数据集或者系统利用系统处理批量任务时，便可以省略互联而直接进入数据展示或分析；

- 然后是视觉交互与探索，这也是整个可视化分析模块的核心功能。通常情况下，用户加载的数据集的类型不尽相同，所希望完成的探索任务也有所不同，因此我们的系统将支持不同维度下的不同形式的数据可视化，为用户提供多种不同类型的可视化方式并利用 GPU 等高性能计算来实现高效展示数据。此外，我们的系统不仅对数据集进行科学高效的展示，还能够是对数据进行理解的基础。且数据的展示需要与科学相符，例如单位和坐标系统的转换等。

  对天文学家而言，单独的数据展示并不能满足数据探索的需要，用户在“看到”数据后仍然需要进一步通过交互例如调整视图，以及对可视化结果的量化例如选取某一块兴趣区域并统计其数据分布等，来发现并初步确定兴趣研究区域。因此我们的系统也将在展示的同时，采用一系列信息可视化领域的交互探索技术，例如二维/三维选取和视图互联，给与用户充分的自由度来在不同维度的空间中观察并探索数据，并实时响应用户交互，对兴趣区域进行高亮显示等，并妥善保存以备进一步的分析。

- 最后是数据分析，即用户对探索获得的兴趣数据通过合适的工具进行分析，从而最大程度的挖掘数据中的价值。为了最大程度的满足用户不同的需求，我们将通过开发内置分析工具，如 iPython 终端及拟合算法等，以及其它第三方工具的分析功能，如植入第三方工具作为插件、协议共享功能及运用云平台提供的服务，实现较完备的分析功能。输出的分析结果将被妥善保存，用户检视分析结果后便可以直接进入下一个分析任务。

上述的四个功能模块中的功能已经由 Glue 软件包含实现，其中包括对多类型

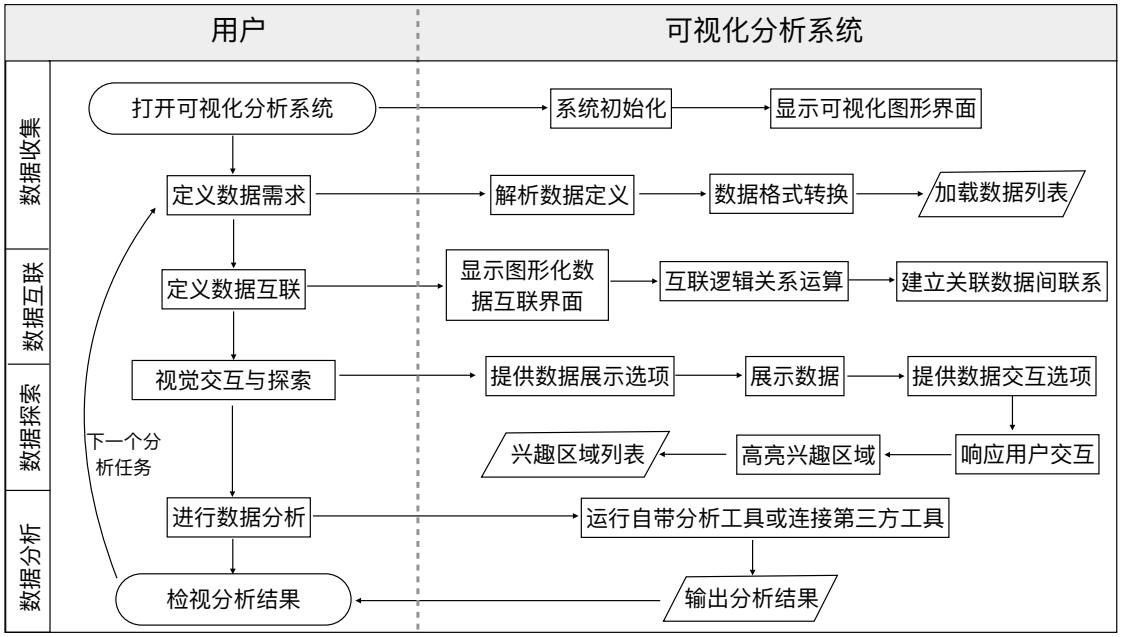

图 2.19: FAST 可视化分析模块的工作流程图。

天文数据的加载，通过范式定义对多数据进行互联（见第 2.1.2节），丰富的数据展示功能以及多样化视觉交互与探索功能，以及内置的 iPython 终端和分析工具等。但是，仍有一定数量的针对 FAST 需求的部分任务需要完善的设计与定义，这部分功能主要集中在数据收集和数据分析两个模块，我们将通过以下两小节具体阐述其中的技术创新与难点，以及各自设计解决方案。

2.2.3.1 数据收集-地图册检索

FAST 中性氢 21 厘米谱线宇宙巡天可以覆盖超过 1 万平方度的天区，这些数据至少有三个维度。传统的数据收集方式是通过用户定义某个范围的天区的坐标信息，来提取对应的数据分块并加载进行分析。这个方法的一个弊端便是对于中性氢这类分布呈弥漫无规则的成分，随机的对数据进行分块将很有可能破坏整体结构特征，因此，一种更有效的数据展示和查询方式是类似地理信息系统的图形化查询界面。也就是将整个天区的全景展示给用户，用户在全景的观察下交互式选取所希望收集的数据，并由系统自动连接数据库进行切割和加载。在系统实现中我们同时保留传统的通过输入区域范围进行检索以外，以及创新性的使用可视化地图册搜索功能提供了一个“总览-细节”的数据搜索模式，也就是为用户展示所关注天区所在的大范围天区的二维地图册。并且该部件还具有灵活的交互功能如缩放、平移和选择，因此用户可以在选择兴趣区域时保留完整的上下文。在“天空图集”中选择的区域可以准确地转换为查询，尽管选择形状会导致检索数据有不同程度的困难。然后，该查询将与 FAST 数据库和其他归档相关联，以获取该天

区中的所有已知天文物体的信息，并根据用户请求返回数据。

以上解决方案的优点十分明显：首先，它非常适合通过位置进行索引的 FAST 数据存档结构。FAST 和其他数据库将使用坐标系和位置对 FAST 的中性氢巡天数据进行索引，因此可以轻松地翻译和理解 Sky Atlas 界面的搜索请求，并且这些数据也可以轻松地与其他资源的数据集“粘合”；其次，地图集搜索窗口部件本身就是一个强大的查看器，能够在不同观测波段之间切换并提供全天视图，这些能力使其成为链接视图系统的关键部分。这种方案与“鸟类”的视图方法（Birds-view）有着异曲同工之妙，在 Birds-view 中，观看者能够对同一个数据使用不一样的视野进行观察（最初内置于谷歌地图中）。

我们设计了地图册搜索及远程检索流程图如图 2.20所示。用户打开地图册检索窗口后，系统显示初始地图册，并进行地图册列表、坐标输入框和交互控件等的初始化。紧接着用户将根据需要定义地图册显示，例如对地图册的成分进行选择或者定义显示位置，此时系统将响应用户输入，并实时更新地图册显示结果。接着，用户将通过输入坐标或者交互选取等方式定义兴趣区域，系统将解析用户的输入，转换成搜索规范语句，并在用户选取目标数据库后，与目标数据库进行连接并发送数据搜索请求，获取请求数据集并加载入系统中存为目标数据集。目标数据集便可以被用户用来进一步的探索与分析。

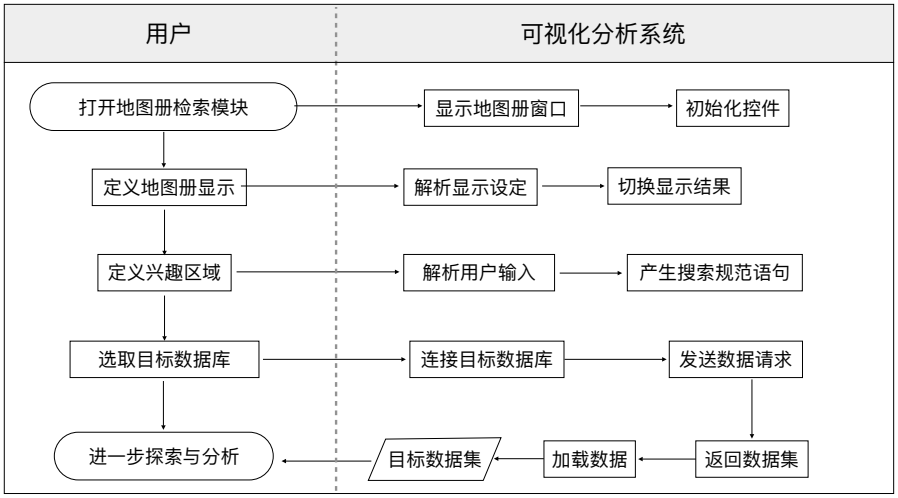

图 2.20: FAST 地图册检索功能块的工作流程图。

2.2.3.2 数据分析-第三方工具

天文数据往往需要专业的数据分析工具，而确保这些第三方工具与我们的系统之间的沟通和互操作性是整个可视化分析系统的非常重要的一部分。根据连接方式的不同，我们在系统中设计了三种可行性方案，如图 2.21所示，来实现与多样的第三方工具的互操作：

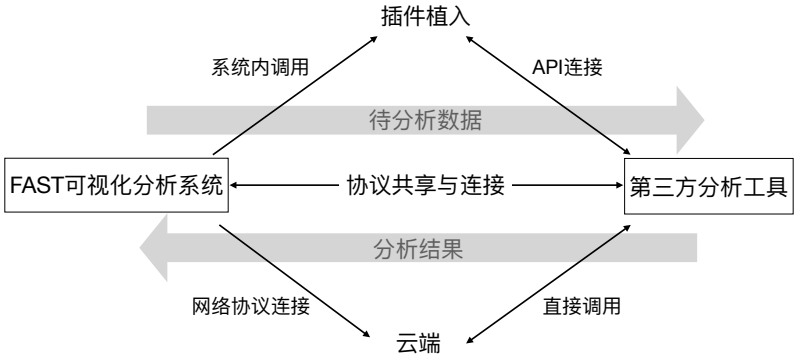

图 2.21: FAST 可视化分析系统与第三方分析工具互操作示意图。

将第三方工具作为插件植入系统中：插件植入是通过我们的系统使用第三方工具的最直接的方式。很多天文常用的数学分析方法，例如线性拟合等，都可以通过调用 Python 函数完成，因此，对于这些轻量级的功能块，我们在系统中设计了专门的通用接口，通过增加图形化界面功能，便能够直接在我们的系统内调用这些工具，同样的方法也适用于第三方天文软件工具包、可视化视图及特殊数据格式的加载等。

这个方案的优势在于通过利用这些开源的第三方工具，我们将直接受益于它们所拥有的现有全部功能，同时最小化开发人员的工作，并且能够增进与其它天文软件开发团队的合作。一个例子便是正在进行的将 WorldWide Telescope 图形窗口（WWT，微软发起的开放源码项目）内置在 Glue 的框架下，使其作为一个特定的数据查看器，实现了天文数据集在球面及地球上的 GIS 风格数据进行高级可视化。目前我们已成功地与 WWT 观众进行了初步测试，并且我们计划使用这种方法将更多的工具纳入到系统大框架下。

与第三方工具通过协议共享与连接：除了插件植入的方式，我们还可以考虑实现我们的系统与第三方工具的 “交流”，以实现功能共享。插件植入方式的一个局限性是，当第三方工具是一个功能较复杂的完备的软件时，将其通过接口开发的方式完全植入我们的系统中将会使得我们的系统变得冗余，且开发难度也将大幅增大。而传统的多个天文软件间的 “合作” 也因为不同软件的数据集间存在直接转换的障碍，而往往需要多个中间步骤的操作来完成数据共享。因此，我们需要实现我们的系统与第三方工具间的 “交流”，能够直接快速的将需要处理的数据共享，并且共享的数据能够保证在进行分析操作后能够同步更新。借此，我们便能够进一步拓展我们系统的分析能力，同时也保证了我们系统的轻量化。

与云端的第三方工具连接： 云计算提供了新颖且可行的解决方案来解决天文学中的大数据挑战。几个成功案例包括望远镜数据存档（例如 Gemini 项目-两个

8.1m 光学/红外望远镜[25]），数据密集型科学计算例如加拿大天文研究高级网络（CANFAR)[67] 和高级志愿桌面计算机组（SkyNet）[26]。

FAST 中性氢巡天数据量庞大，利用云端服务进行数据分析与计算将是趋势。此外，FAST 还存在数据存储与处理的地域限制问题：FAST 望远镜的数据接收与存储中心将落在贵州，而 FAST 的数据将由北京科学总部或各地的科学家进行调用与研究，局域网等局部传输手段不适合 FAST 这类需要跨地域合作的项目。因此，我们将在本章的设计与实现中探索利用云服务来存储并分析 FAST 数据的可能性和技术方法。

我们设计在可视化分析系统中直接添加对云计算平台的接口。这个方案的优势在于，首先，传统上天文学家必须从服务器手动下载数据集以进行分析，然而，FAST 研究的数据将很可能远远超过天文学家普通使用的计算机的带宽和存储容量，因而数据处理在云端上进行，并且从云中下载处理结果的设计更符合实际情况。其次，云架构也非常适合能够通过并行化完成的问题。一些数据探索的挑战，例如选取不同的数据集中的兴趣区域，在已预先决定种类的情况下将变得容易很多，包括直方图中的不同 bin 和数据立方体中的不同切片，这种分类也使得并行化成为可能。因此，我们也将设计并探讨如何使用云平台加速对数据集的选取，这些数据集包含了可以并行分配计算任务的 bin、切片或其他类别等。

预期完成的基于云计算平台的无缝天文系统架构图如 2.22所示。该系统大致可分为三层：用户在最高层，他们可以有不同的专业背景、用不同的客户端软件、具有不同的用户习惯以及分布在不同的地理位置等多样性。中间层是客户端软件，细分为四层：在用户交互层，用户可以通过图形用户界面或 Python 终端控制台操作和显示数据，并通过图像渲染层，由高性能的图形库如 OpenGL 等将这些数据渲染出来。在数据持久层中，这些数据被存储在高维的数据容器中，并能够自动运算出用户数据间的内在逻辑关系（例如互联数据时数据间的映射方程）。这些数据通过一个抽象的输入输出层存储，如果用户决定存储在本地文件系统，则可以存为通用的天文数据格式，如 FITS 或 HDF5 等；如果用户的数据存储在云端，则本地数据和数据处理指令将通过一个抽象网络层与云平台沟通。云平台位于最底层，大数据的存储和相关算法的实现均在这一层实现。最终，这些数据和计算任务被分布在多个计算/存储节点中。所有的存储及运算细节对用户是透明的，也就是说用户无需关心数据如何存储，在哪个节点进行计算，用什么协议与其他用户产生的数据沟通等，云平台对用户自动屏蔽了这些复杂性，使得用户能够专注在研究任务上。

[25]https://astrocompute.wordpress.com/2015/11/20/a-new-data-archive-for-gemini-fast-cheap-and-in-the- cloud/

[26]https://skynet.unc.edu/

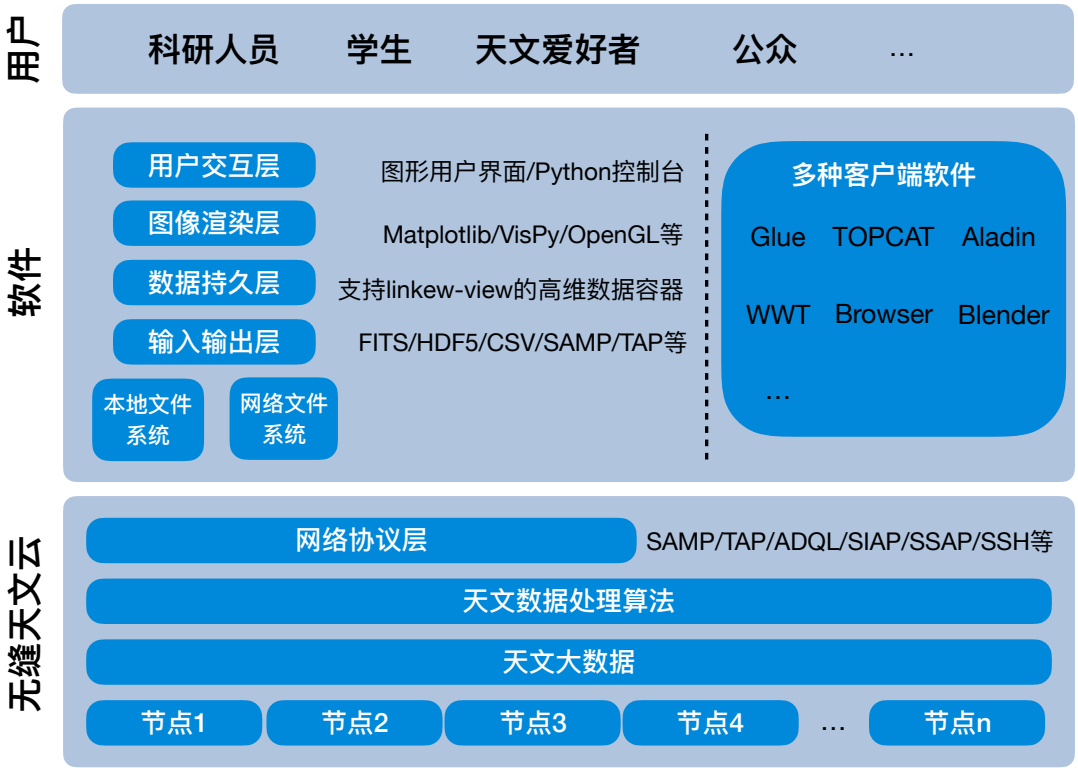

图 2.22: 基于云计算平台的无缝天文系统架构。

2.2.4 实施和方法

本小节将针对上节讨论的设计方案，对其技术方法和实施方案进行具体阐述。

2.2.4.1 嵌入和优化 Glue 中的地图集搜索窗口部件

地图册显示是地图集搜索窗口部件的基础。对于地图册显示的实现，出于减少开发工作的考虑，我们将在现有的实验性地图册插件的基础上进行完善。我们挑选天文领域常用的两个地图册类可视化交互工具，Aladin Sky Atlas[68,69] 和 WorldWide Telescope (WWT)，作为地图册显示的实验性平台。这两个工具均使用 HTML5 标记语言在网页端实现了一系列显示和交互功能。我们以 WWT 为例介绍具体实现的技术细节：

- 创建 HTTP 请求：根据用户选定的数据范围，创建一个 HTTP 请求
- 发送 HTTP 请求：将请求发送到 WWT 的 API，通过 WWT 定义的具象状态传输（RESTful）方式传递信息
- 返回 XML 数据：WWT 将接收 HTTP 请求，渲染用户的数据并返回 XML 数据
- 解析 XML 数据：从返回的 XML 数据中解析并获取视图所需的 layer 显示信息
- 初始化视图界面：通过 QT 的 QWebEnginePage 类进行 WWT 的网页界面初始化
- 显示视图信息：通过 Glue 的 API 将初始化的 WWT 网页界面内置在 Glue 软件框架中，显示上述从 XML 数据获取的视图 layer

根据上述方法，我们完成了对 WWT 和 Aladin 的实验性嵌入的开发与测试，实现了在 Glue 中将 WWT 和 Aladin 以插件的方式，根据用户选择展示不同波段的全天区的地图册，并支持鼠标输入来改变视图角度等简单视图交互功能，完整的测试代码和技术文档可以参考 Glue-WWT 页面[27]和 Glue-Aladin 页面[28]。

一旦嵌入了地图册工具并具备了基本的显示功能，我们将着手扩展 FAST 可视化分析模块的交互功能，以支持在屏幕上进行不同形状的选择，并能够从数据库获取对应形状的数据集合的功能需求。交互功能的开发需要解决以下两大挑战：

[27]https://github.com/glue-viz/glue-wwt

[28]https://github.com/glue-viz/glue-aladin

复杂形状的选取：用户选取的天区形状可能是不规则的（例如通过 free lasso 得到的形状），而一般而言，数据库查询语言（如 SQL）能描述的形状仅为简单的矩形。因此，如何将用户定义的复杂形状转换成为数据库查询语言是构建 FAST 可视化分析模块需要解决的第一个挑战。支持较复杂形状的选取模式的困难在于，一般数据库只接受规则的区域查询，例如给定的赤经和赤纬组成的框形区域，而对于圆形或 Free Lasso 这类较复杂的形状组成的查询，数据库往往不能够直接解析。因此，为了与数据库兼容，查询语句以及返回的数据集需要进行加工，例如将不规则查询区域转换成规则的查询，以及将数据库返回的数据集“剪裁”成与选取形状吻合的形状等。

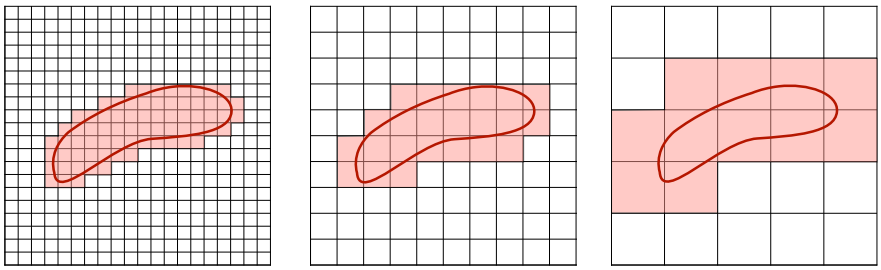

图 2.23: 不规则区域的栅格化剪裁。栅格的大小由参数 R_{ras} 定义。当 R_{ras} 选择较小时，栅格的拼接能较大程度地还原用户选择的原始区域，因此具有较小的数据冗余度，有利于提高数据传输效率。但其代价是需要构建更多的子区域查询和拼接运算。相反的，如果 R_{ras} 较大，则会出现较大区域外的空间，增加了数据传输时间，但简化了栅格化和拼接子区域所需的时间。

为了从数据库中查询如图2.23所示的不规则天区，我们需要将该天区栅格化(Rasterize)，即是把该天区分割成多块的矩形子区域。由于子区域是规则的，因此能很容易地构建数据库查询语句。从数据库返回的一系列子区域在天文学家的本地计算机中进行拼接，就得到了用户所选择天区的完整数据。如有需要，可以对多余的数据进行剪裁，从而精确地得到了用户选择的区域。这一过程需要定义一个栅格化分辨率 R_{ras}。当 R_{ras} 的选择过大的时候，数据库将返回大量用户选择区域以外的区域，这就增加了所需传输的数据量。相反的，如果 R_{ras} 过小，则将产生大量的子区域，从而意味着不仅要构建大量的子区域查询（增加数据库负担），而且也拼接这些子区域也需要更多的计算量。因此，一个优化的 R_{ras} 值是子区域数量和网络传输速度的折中（trade-off），需要根据实际情况而定。

像素空间向数据空间的映射：用户通常在二维的屏幕上进行选取，因此选取的区域是二维的。为了得到正确的数据，我们需要将屏幕的像素坐标转换成天文数据集中用到的天文坐标。二维选取向数据集的映射，主要的困难在于像素坐标系(Pixel Coordinates)与物理坐标系(World Coordinate System, WCS)之间的转换。针对天文数据类型文件的坐标转换，例如针对天文领域使用最广泛的 FITS(Flexible

Image Transport System) 文件格式，Greisen & Calabretta, 2002[70] 提出一个规范化方法来定义 FITS 格式下数据样本的坐标，并且，Calabretta & Greisen, 2002[71] 在其研究中描述了将该坐标映射到二维平面的坐标转换方法：将距离参考点 (Reference point) 的像素坐标向量乘以线性转换矩阵并标度到物理单位，其主要数学表达式为

$$x_i = s_i q_i = s_i \sum_{j=1}^{N} m_{ij}(p_j - r_j)$$

其中 p_j 是像素坐标，r_j 是由 FITS 头文件中的关键字对（下同）`CRPIX`j 给出的参考点像素坐标，m_{ij} 是由 `PC`i`_`j 或 `CD`i`_`j 给出的线性转换矩阵，N 是由 `NAXIS` 或 `WCSAXES` 给出的 WCS 表示下的维度，s_i 是由 `CDELT`i 或 1.0 给出的标度。而在 Greisen, et al. 2005[72] 中进一步扩展了该坐标转换，增加了对谱线坐标轴（例如频率、波长和速度轴）的坐标转换。这一系列的转换在天文数据处理软件包 AstroPy 中，以子软件包 (astropy.wcs) 的形式被包装[20]，因此在开发实现中我们可以直接使用。

2.2.4.2 从远程检索原始数据

我们还将通过建立与远程数据库的连接，实现从天文数据库下载原始数据并加载到 FAST 可视化分析系统中的功能。连接常用的天文数据库，不仅需要开发规范化的天文数据访问协议接口，而且还要保证与所连接的数据库端的协议相兼容。很多常用的天文数据库，包括 NASA/IPAC Extragalactic Database (NED)[29]，SIMBAD[30]和 VizieR[31]，使用的是由国际虚拟天文联盟（International Virtual Observatory Alliance, IVOA）的所定义的一系列针对天文数据远程访问的协议，目的是保证客户端与服务器之间在网络上的信息传输的兼容性。因此，我们能够在系统中开发 IVOA 协议的接口，与 FAST 数据分析相关的协议有 SIAP（简单图像访问协议）[32]，SSAP（简单光谱访问协议）[33]和 TAP（表格访问协议）[34]），或从定制数据库（例如 FAST 数据库）使用 ADQL（天文数据查询语言）[35]。

鉴于 Python 是天文领域常用的编程语言，并且 FAST 可视化分析系统的主要功能实现编程语言是 Python，使用现有的 Python 工具包进行数据请求与传输

[29]https://ned.ipac.caltech.edu

[30]http://simbad.u-strasbg.fr/simbad/

[31]http://vizier.u-strasbg.fr/viz-bin/VizieR

[32]http://www.ivoa.net/documents/SIA/

[33]http://www.ivoa.net/documents/SSA/

[34]http://www.ivoa.net/documents/TAP/

[35]http://www.ivoa.net/documents/latest/ADQL.html

的开发也将是实现与远程数据库连接的一个方案。目前天文领域类似的 Python 工具包有 Astroquery[36]，astropy.vo 和 pyvo[37]。其中，astropy.vo 与 pyvo 更加侧重对 VO 框架下的数据的请求和获取，可以运用到对 VO 相关的数据库的数据请求，而 Astroquery 则是天文工具包 astropy 的一个功能模块，具备一般网页服务的通用接口，能够使得 FAST 的远程数据库的连接更加自定义化与灵活。

除了开发针对各个天文数据库的协议 API，一个直接的方案便是利用以后的对天文数据库的接口集合工具，内置到 FAST 可视化分析系统中。我们已经实验性的将 VizieR 的基本检索功能内嵌到我们的 FAST 可视化分析系统中，VizieR[73] 是目前最全面的一个提供天文常用数据接口的在线数据库，其查询工具也允许用户对数据类型及关联数据进行定义，因此能够为 FAST 从事交叉认证研究提供帮助。与 WWT 地图册窗口内嵌的方法相似，我们通过 Glue 的高度模块化架构，将 VizieR 模块内嵌入 FAST 可视化分析系统中，并将用户在 FAST 可视化分析系统中定义的数据查询信息进行包装，通过 HTTP 请求发送到 VizieR 并获取显示信息。目前已初步实现简单的数据库搜索显示，所使用的代码及文档可见 Github 代码仓库页面[38]。

2.2.4.3 与第三方工具协议共享与连接

与第三方工具的协议共享与连接对 FAST 数据分析具有重大。国际虚拟天文台联盟（IVOA）提供的服务中，有一个 SAMP(Simple Application Messaging Protocol) 标准[74]，借助于它，可以实现各个服务功能之间的消息通信。目前已有相当数量的天文工具已经实施了使用 SAMP 的通信，例如 World Wide Telescope[75,76]、用于图像和目录可视化的 DS9 软件[77,78]，目录探索工具 TOPCAT[79,80] 等，开发一个 SAMP 连接器也使得我们的系统能与这些软件进行数据交换与功能共享，能够不通过开发便丰富我们系统的分析功能。

为了使我们的系统成为 SAMP 协议的客户端（Client），也就是作为发送方（Sender）、接收方（Recipient）或两者，能够通过 SAMP 协议与 Hub 进行交流，我们首先需要将我们的数据对象转换成 SAMP 兼容的数据类型，比如字符串（由一系列字符组成的标量值; 每个字符是一个十六进制代码为 09,0a，0d 或 20-7f 的 ASCII 字符），列表（数据项的有序数组）和映射（键值对的无序关联数组，其中每个键是字符串，每个值是数据项）等。

此外，SAMP 使用一个基于中枢的架构使得我们的系统与其他客户端进行通信，我们还需要通过填写包括客户端 ID、应用程序元数据、MType Subscription

[36]https://astroquery.readthedocs.io/en/latest/

[37]http://dev.usvao.org/pyvo

[38]https://github.com/glue-viz/glue-exp/tree/master/glue_exp/importers/vizier

以及消息和响应的映射表单注册表来注册到中心 Hub。所有这些信息使我们能够与中心建立沟通，向中心和其他客户宣传系统的存在并获取注册信息。在我们初步调研与开发下，Glue 已初步具备使用 SAMP 服务的能力，如 Youtube 视频[39]所示。一旦 SAMP 选项打开，SAMP 将创建一个用于 Glue 的客户端 (Client)，并将其注册到集线器 (Hub)，集线器是集中所有应用程序的所有消息的服务器。通过集线器，Glue 可以发送和接收特定消息，包括将选取的兴趣区域共享到其他应用程序。这样一来，所有连接的软件所具有的功能都被 “粘合” 起来，并进一步的，甚至组成一个科学网关 (Science Gateway)[81]，基于网络接口使得研究人员可以连接科学仪器，传感器流数据，以及使用先进高效的计算工具。

而 SAMP 这种通过接收和发送数据和子集实现数据与功能共享的例子也可以扩展到任何具有 API 的软件包，包括天文领域以外的软件系统，例如我们可以通过增加与 Blender 三维可视化软件包[82,83]，或 CODAP（通用在线数据分析平台）工具[40]等的协议共享和连接，来获得天文领域较为稀缺的三维可视化功能和在线分析功能，这也使得 FAST 可视化分析系统的功能更加完善也更加独特。

2.2.4.4 云架构的设计与实现

搭建 FAST 可视化分析系统的云架构的一个前提便是将数据访问计算与剩余展示分析功能进行分离，即实现一个抽象数据对象层。用户在使用过程中不需要下载数据集，系统可以根据需要返回查询数据集或者对应用户请求的绘图/计算阵列。返回数据将是一个系统的数据对象类，它是一种轻量级的方式，只存储选定/重要信息的数据。因此，GB 级原始数据可以压缩成 MB 级，但仍然包含元数据和数据体。该数据抽象层将定义新的数据对象包含运行中计算的显示数组，这些数据虽然类型不同但是在用户仍然看起来相同。

在对 Glue 软件的开发中，我们也已经完成了在亚马逊网络服务的远程数据主机上的初步测试。我们在 Amazon 服务上绘制了一个直方图，并通过 HTTP 请求将结果返回到 Glue 进行显示。因此，我们能够将这个方案扩展到支持 FAST 数据中心，在商业化云资源上存储来自 ALFALFA 巡天的一些测试数据，以构建测试主机服务器。或者利用由中国虚拟天文台（中国虚拟天文台 ChinaVO）开发的新的在线平台 AstroCloud[84]，其集成了基于云的存储空间 MyVOSpace、已完善配置在服务器上的即用软件环境以及阿里巴巴云虚拟机服务，并使用同样的数据请求来测试将一些简单的分析任务分布给远程服务器完成。对大数据的支持可以通过 blaze[41]包对新的数据对象进行备份，并非常快速地访问大数据源（如数据库）和高效计算。

[39]https://youtu.be/YB1rMCxzA3E

[40]https://codap.concord.org/

[41]http://blaze.pydata.org/

2.2.5 模块概念实现图

我们模拟了 FAST 可视化分析模块的概念实现图，如图 2.24所示。基于 Glue 的可视化分析模块的图形界面与 Glue 相似，包含六个主要功能区如图橙色框图所示，白色框图给出了功能实现的插件，用户可以在插件内进行交互式操作（红色字表示）。同时，该系统还能与其它软件工具进行数据与功能分享，例如左下角与 Topcat 软件进行互通的示例。

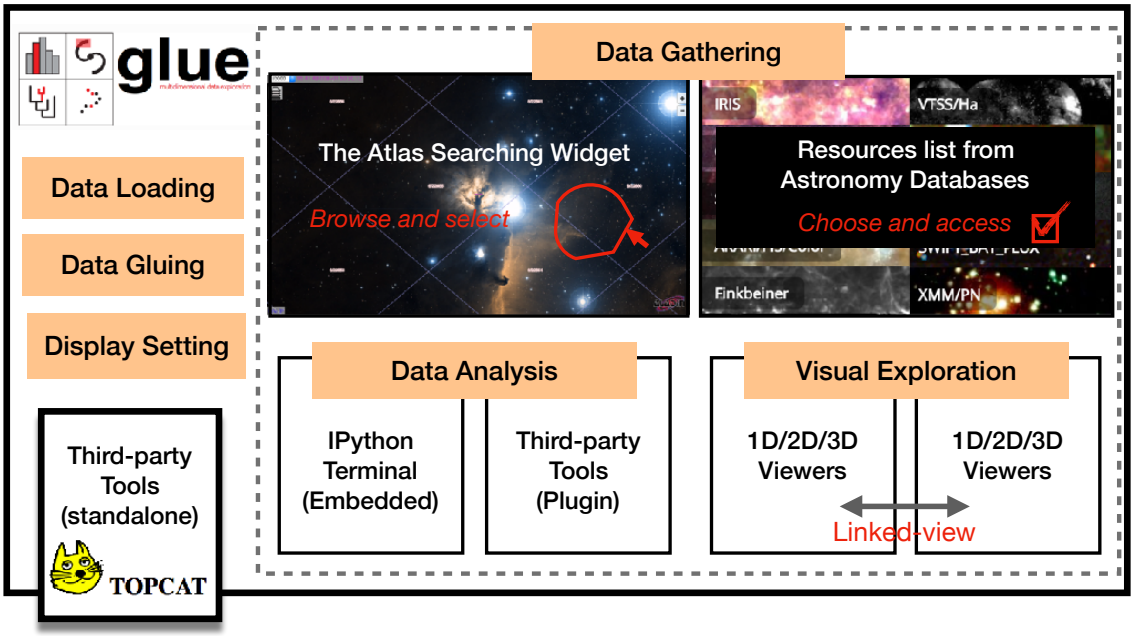

图 2.24: 基于 Glue 软件的 FAST 可视化分析系统的初步框架。

我们将保证整个系统的灵活性以顺应 FAST 数据格式的变化，因为对 FAST 的数据格式和存储方式仍在讨论中，预期最有可能数据格式将是 HDF5 或 FITS 或混合。目前的 Glue 软件可以有效地加载 FITS 和 HDF5 格式，但对于大型 HDF5 格式的支持尚未得到优化。在目前版本的 Glue 中，整个 HDF5 文件将被一次性加载到系统中，如果文件大小超过系统内存，应用程序将崩溃。因此，将开发一种加载 HDF5 数据的自适应方案十分重要。此外，如果这样的实现被证明对促进研究工作流程有帮助，还需要开发或实现专用的 HI 光谱线分析功能（例如 SpecViz[42]）用于 HI 谱线研究，例如 HI 窄线吸收[85]。

[42]http://specviz.readthedocs.io/en/latest/

2.2.6 代码及文档版本控制

版本控制是一种记录一个或若干文件内容变化，以便将来能够查阅并修改特定版本的系统。它不仅是维护工程蓝图的标准做法，也是一种软件工程的技巧，借此能在软件开发的过程中确保由不同人所编辑的代码与档案都得到同步[43]，对于FAST 可视化模块甚至整个数据处理系统的开发与维护，团队合作是不可缺少的一个环节，如何确保一个项目在多人的同时开发下仍然有条理并无冲突，是保证整个项目能够顺利进行的关键。

我们通过调研发现，常用的简单的版本控制方法是通过不断复制整个项目目录，改名加上备份时间等信息以示区分。但是这种方法的容错率低，比如不小心写错文件或者覆盖了意料之外的文件，且当文件量大时也容易产生混淆。不少本地版本控制系统，采用了基于数据库来记录文件的历次更新差异的手段，通过数据库来管理大量文件的版本信息。同时，为了使在不同系统上的开发者协同工作，而采用集中化的版本控制系统（Centralized Version Control Systems，简称 CVCS），诸如 CVS[44]、Subversion[86] 以及 Perforce[45]等，通过一个单一的集中管理服务器，保存所有文件的修订版本，而协同工作的人通过客户端连到这台服务器，取出最新的文件或者提交更新。这种集中化的版本控制系统使得项目合作更加透明化，项目成员可以清楚的看到其他成员的工作进程，并且管理员也可以轻松掌握每个开发者的权限。然而，一个显著的问题是存储方式过于单一，一旦存储项目资料的本地机器或服务器出现问题，就有丢失所有历史更新记录的风险，这对于 FAST 可视化分析系统这类复杂的程序开发，丢失记录所造成的损失将直接导致项目开发的中止并直接影响后续开发。

而分布式版本控制系统（Distributed Version Control System，简称 DVCS）中的 Git 则巧妙的解决了单一存储的问题，也因此成为 FAST 可视化分析系统乃至数据处理系统开发的最佳版本控制手段。Git 的设计理念中，服务器端仍然存有代码仓库及历史更新记录，但不同于集中管理服务器的是，客户端并不只提取最新版本的文件快照，而是把代码仓库完整地镜像下来。这么一来，任何一处协同工作用的服务器发生故障，事后都可以用一个镜像出来的本地仓库恢复。因为每一次的克隆操作，实际上都是一次对代码仓库的完整备份。对于备份，Git 并不是单纯的拷贝代码仓库的所有文件并存储，而是像一个 “快照流” 的方式，对当时的全部文件制作一个快照并保存这个快照的索引，并且为了高效，如果文件没有被修改，Git 便不再重新存储该文件，而是只保留一个链接指向之前存储的文件[87]，如图 2.25。这个设计也使得 Git 相比于 SVN 速度更快更流畅。

[43]https://en.wikipedia.org/wiki/Version_control

[44]http://www.nongnu.org/cvs/

[45]https://www.perforce.com/products

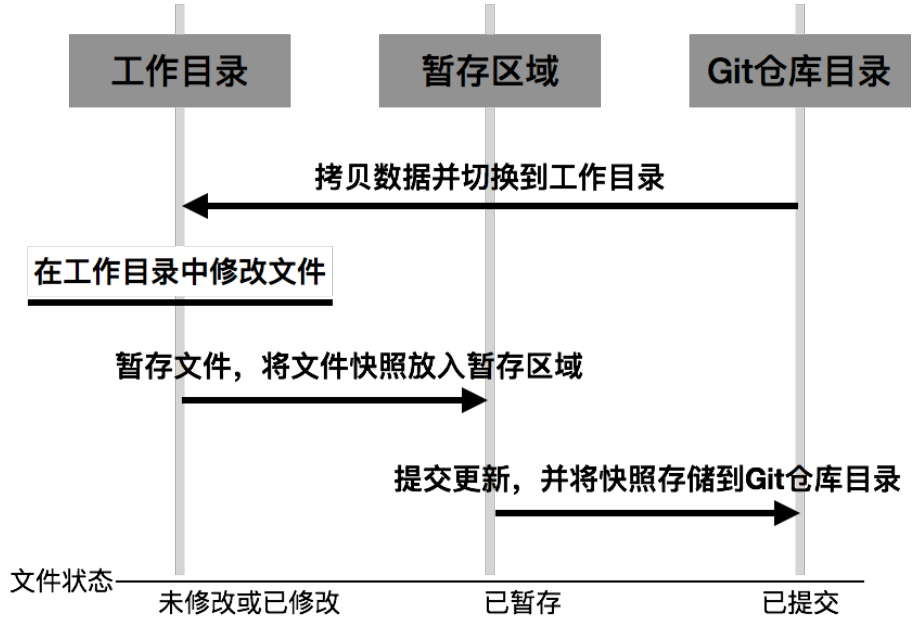

图 2.25: 该图展示了基本的 Git 版本控制的基本工作流程，主要分为三步：首先，在工作目录中修改文件；其次，暂存文件，将文件的快照放入暂存区域；最后，提交更新，找到暂存区域的文件，将快照永久性存储到 Git 仓库目录。

通过 Github[46]，即 Git 的网页端源代码托管服务网站，我们对 FAST 中性氢数据处理流水线的开发与文档管理，如图 2.26所示。通过此开源平台，不仅能够直观易懂的进行合作开发及记录管理，而且还能够通过此开源社区的活跃度，吸收编程爱好者对项目进行开发，从而加快工程进度。储存在代码仓库上的 FAST 中性氢数据处理脚本也将是对本领域的一份贡献。

[46]https://github.com

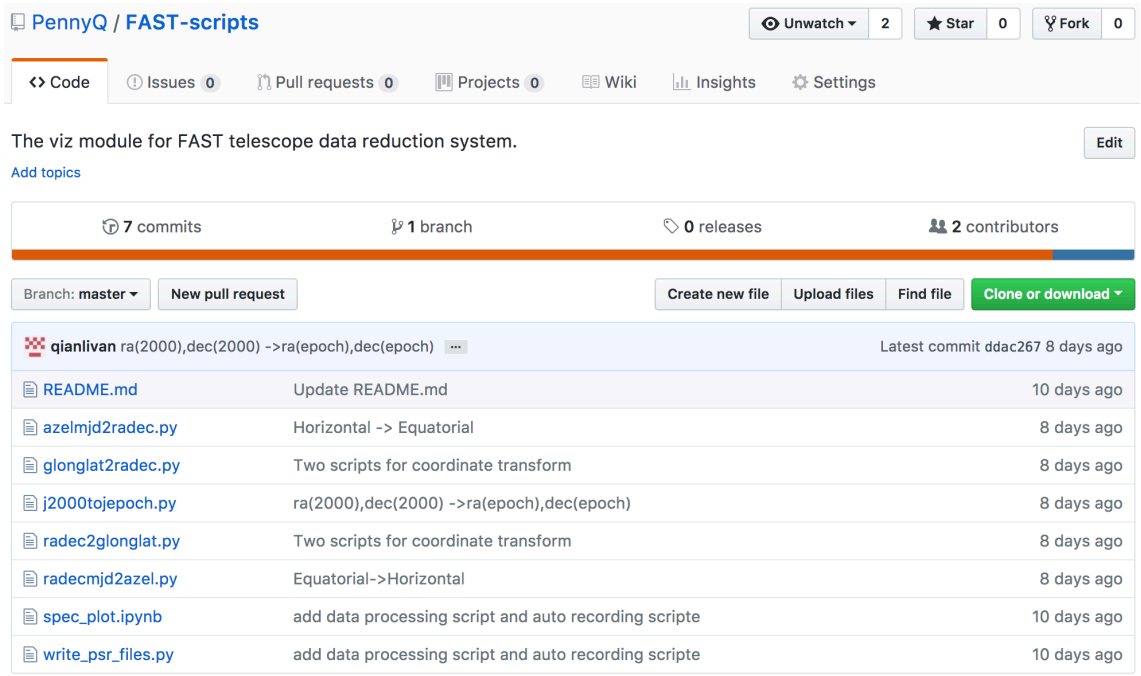

图 2.26: 该图给出了我们在 Github 上搭建的用于管理 FAST 中性氢数据处理系统开发的代码仓库界面。

2.3 总结

数据可视化是数据处理的一种重要手段。FAST 项目不可避免将产生复杂的、高维度的数据集。天文学家往往需要探索这些数据集之间的联系（如交叉证认）以及数据集中不同子集的逻辑联系。普通的数据可视化软件支持 2D 和 3D 的散点图可视化，但无法满足高维度数据处理的需求。而且，这些软件一般并非针对天文数据处理而开发，无法满足天文学家所需的特定工作流程。

Glue 软件正是为了填补这方面软件缺失而开发的。该软件是一个基于 Python 的开源数据化软件。针对高维度数据探索的需求，Glue 开创性地使用了“视图互联”的方法。Glue 允许用户利用多个视图（如散点图、直方图等）显示同一个数据的不同维度。视图与视图间的数据之间存在着逻辑联系，因此当用户在一个视图中选取特点的部分时，相应的数据也会在其它视图中被高亮显示出来。

对三维渲染和选取的实现，满足了 FAST 对三维交互展示的需要。我们和哈佛史密松天体物理中心的团队合作，利用 VisPy 库，实现了用 OpenGL 硬件加速来渲染大数据集的目的。在两维的屏幕上通过鼠标选择三维空间中的物体一直是计算机图形学中的一个挑战，但这个功能对于交互式的视觉分享至关重要。因此，我们开发了一套三维选取的算法。虽然 Glue 的视图互联功能早在 2014 年就已经实现，但三维选取的实现使得 Glue 第一次实现三维视图互联的渲染和人际交互能力。

基于视图互联系统开发的具有高度可扩展性的 FAST 数据可视化分析模块, 具备可交互性、三维展示和选取、具备视图互联等强大的可视化分析功能，能够深入探索单个数据集及其内部属性之间的联系, 方便得出不同数据集关于同一个数据源的比较关系, 还能够可视化展示这些数据和数据间的联系, 并给与用户极大的自由使得他们能够自由选取兴趣区域并进行进一步分析。

第三章　FAST 数据产品的三维展示

FAST 数据产品展示旨在为 FAST 巡天数据进行科学化呈现，特别是中性氢巡天产生的高维度数据，使得科学成果能够以更加生动形象的方式被用于数据理解和传播。ALFALFA 巡天就通过 Google Sky 作为平台来传播天文数据给教育和科普[88]，并且 ALFALFA 还开发了一个专门的 IDL 软件 GRIDVIEW 来实现对三维数据立方体的可视化[89]。对于中性氢数据的三维可视化，Taylor 利用 FRELLED 工具实现了三维实时渲染[90]，而 Punzo 基于医学软件 Slicer[1]的开发了对中性氢数据可视化与三维选取的扩展模块 AstroSlicer[91]。而除了三维渲染，虚拟现实等新技术也逐渐被天文学领域发掘并应用。

本章节侧重于介绍对高维度天文数据展示的新技术的设计与开发，并通过实例给出具体实现方法及其展示结果。这不仅是对 FAST 科学数据产品展示需求的一个回应，也是天文与可视化跨学科相互补充的一次实践，是对天文可视化领域的一次创新。本章节旨在通过实践队 FAST 数据处理系统的展示模块给出一定的建议和方向，而不是也不能够对所有可能的技术和软件进行穷尽。

3.1　基于 Blender 开发三维数据单元的可视化工具

近年来，随着科技的进步，无论是地面的还是天空的观测设备都在不断增加、改进和升级，各波段拥有各自优秀的望远镜建成并投入使用，如阿塔卡马大型毫米波/亚毫米波阵列 (ALMA)[92]、澳大利亚平方千米阵列 (ASKAP)[93] 以及位于智利的光学观测设备 Very Large Telescope(VLT) [2] 等，随着观测能力越来越强，接收系统获得的数据量也越来越大，数据中包含的多维度信息也越来越丰富。

如何从“高维度”数据中完整展现所有的信息，对于天文这样一个以数据为基础的学科来说至关重要。利用数据可视化方法，可以发现数据背后的隐藏信息，还原和重构信息之间的相互关系，正如美国计算机科学家布鲁斯 · 麦考梅克在 1987 年关于科学可视化的定义[3]中阐述的，数据可视化能够 “利用计算机图形学来创建视觉图像，帮助人们理解科学技术概念或结果的那些错综复杂而又往往规模庞大的数字表现形式”。对天文数据进行可视化有利于展示复杂天体如分子云、星系的结构，建立恒星形成、星系演化等模型，并且有助于对三维宇宙大尺度结构的研究。可以说，有效地展示多维度信息是天文研究的重要手段。

[1] www.slicer.org

[2] http://www.eso.org/public/teles-instr/vlt/

[3] https://zh.wikipedia.org/zh-cn/%E7%A7%91%E5%AD%A6%E5%8F%AF%E8%A7%86%E5%8C%96

而目前的天文数据处理软件大多数没有完善的三维展示功能，对云、烟等无定形态物质和大规模粒子的三维模拟方法仍处在探索和发展中，部分天文软件采用利用二维图像加上一维谱线的方式体现三维特征，例如 Starlink[4][94] 软件包，或是按一定序列动态展示二维图像的方式显示三维特征，例如 GIPSY[95][5]等，但这些传统的二维图像显示模式已不能满足科学家们的需求，未来的天文研究需要更完善的数据展示方法和工具。

随着越来越多的科学家和团组意识到三维可视化对天文领域研究的重要性，并致力于开发满足宇宙及星系结构研究需求的三维可视化工具，可视化工具包的开发成为天文领域一个热门的研究方向，目前已经有一批软件被开发并且使用，例如美国国家射电天文台 (National Radio Astronomy Observatory) 开发的图像展示与分析工具 CASA Viewer[6]，小型科学家团队开发的用于光滑粒子流体动力学的光线跟踪算法 Splotch[7]，以及斯威本科技大学 (Swinburne University of Technology) 开发的三维可视化软件 S2PLOT[96] 等等，这些软件为相关领域的天文学家带来了便利。

与此同时，天文数据结构复杂、无定形态等特征，与其他领域的数据具有高度相似性，例如医学上展现组织结构的三维影像，地理上的等高线图等。由于有广泛的社会需求，这些领域已拥有比较成熟的三维展示软件，例如医学图像软件 3D Slicer[97]。如果天文数据能转换成相应的格式，也可以使用这些软件研究天文领域的复杂图像如恒星形成区等[2] 。本节为天文学家介绍一款动画媒体领域使用的专业三维建模软件—Blender，及其在天文领域的优势和应用，并详细说明对三维数据单元和 N 体模拟数据的三维可视化工具的开发过程。

3.1.1 Blender 简介

本文使用的 Blender[98] 工具是一款支持高质量的建模、动画和渲染的开源三维建模软件，在 GNU GLP 的官方下载下支持 Linux、Mac OS X、Windows 操作系统平台，且在不同平台下界面保持一致。Blender 内核由 C 语言写成，界面部分由 C++ 写成，Python 作为其应用程序编程接口 (Application Programming Interface，API) 及脚本编辑语言。为了方便用户进行更灵活的开发，Blender 将用户界面 (User Interface，UI) 的开发代码封装到 BPY 库中，用户可以通过编写 Python 脚本自定义用户界面，并将自定义用户界面置于工具栏、属性栏等不同位置，还可以添加各种控件。

[4]http://starlink.jach.hawaii.edu/starlink

[5]http://www.astro.rug.nl/∼gipsy/

[6]https://safe.nrao.edu/ wiki/bin/view/Software/CasaViewer

[7]http:// wwwmpa.mpa-garching.mpg.de/∼kdolag/Splotch/

其灵活的图形界面，使得科学家在导入脚本后便可像使用一般软件那样利用 Blender 实现天文数据的可视化与简单分析，同时，Blender 支持很多数据格式，包括静态图片常用的 PNG、GIF、JPEG、TIF 格式，动画的 AVI、H.264、Quicktime、MPEG 格式等，以及常用的三维模型文件格式，同时，Blender 软件包含 Python 文本编辑器和命令行窗口，因此它能够方便地处理各种类型的天文数据例如 FITS、HDF 等，通过导入相应的 Python 科学扩展包也可以对数据进行快速读取和计算[99]。

与普通三维建模软件相比，Blender 具有跨平台性、用户界面可扩展性以及能够直接展示和处理天文数据的特点; 与天文软件包相比，Blender 具有更好的可视化效果，并能够实时地观察数据对象的三维结构，如图 3.1，分别展示了不同工具包对同一数据的三维可视化效果，数据源为范围 z<=3300km/s 邻近星系星表 Cosmicflow-1 Distance[100]。

从图 3.1可以看到，Blender 可以实时地从不同角度对一个数据源的形态结构进行观察，自由灵活地调整视角，并且数据展示的清晰度更高，结构特征表现得更为明显，用户还可以通过材质纹理等个性化设置将数据模型渲染成高分辨率彩色图像，将研究成果更好地展现给公众。

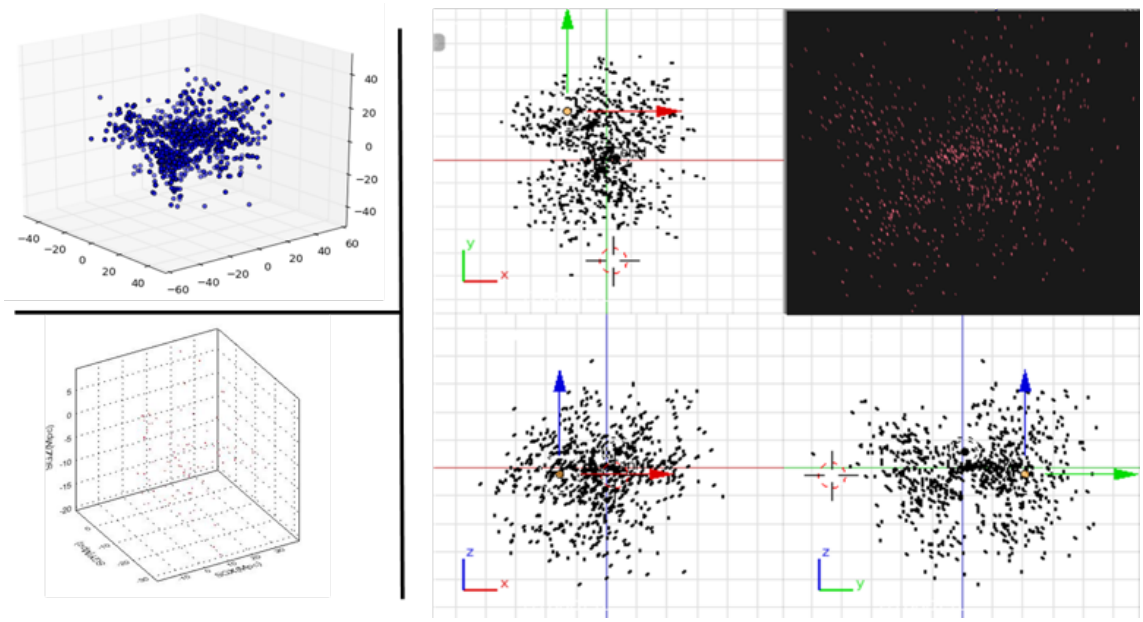

图 3.1: 比较 Blender 与普通软件的可视化效果。左上图为利用 Python 自带的绘图库 Matplotlib 对数据的三维可视化，左下为用 IDL 绘制的结果，右侧为 Blender 的四视图效果和渲染结果。

由于对不同的数据结构处理需要使用不同的 Python 扩展包，重新编写程序，因此为了节约时间，方便进一步分析，在 Blender 环境下开发了对特定结构数据三维数据单元和 N 体模拟结果进行三维可视化的工具包。

3.1.1.1 基于 Blender 开发三维数据单元的可视化工具

谱线三维数据单元在射电天文观测中有着广泛的运用，大多数中性氢 (HI)、CO 等数据结构均为三维数据单元，它由两个空间维度 (通常是 RA 和 Dec) 、一个频率或波长维度以及通量值构成，通常的天文处理软件只是根据频率或波长谱线展现相应的二维图像及其每一点的通量谱线，对于三维数据单元的整体结构并没有一个完整展示。而在研究中，不仅应该考虑单独分子云的特征，更应该把握其形态结构和动力学结构，估计周围物质对它的影响，才可以更好地研究其形成和演化规律。因此，对三维数据单元的三维可视化能够更好地了解它的动态结构，为研究恒星形成和星系演化做准备，甚至可以在宇宙大尺度结构的研究上做一些科学推测。

下面以 FCRAO 14m 望远镜对 G25.4-0.14 分子云[101] 的 CO 谱线观测数据为展示基于 Blender 开发的三维可视化插件工具。示例数据是一个三维数据单元，其文件为标准 FITS 格式，包含头文件和图像数据两部分，在头文件中用键值对的方式存储了文件基本信息以及与观测、定标有关的一些信息，图像数据部分为三维数组结构，每个数组元素存储了当前位置的源的通量值，对应银道面坐标银经范围是 25.20° ∼ 26. 00°、银纬范围是 -0.504° ∼ 0.117°，径向速度范围是 90.476 ∼ 105.354 km/s，根据以下公式计算出分子云中心位置距离太阳为约 5.7 kpc。

$$v_r = v\cos\alpha - v_{sun}\sin l$$
$$R^2 = R_{sun}^2 + d^2 - 2R_{sun}d\cos l$$

其中 $\cos\alpha = \frac{R_{sun}}{R}\sin l$; v_r 和 v 分别为分子云的径向速度和旋转速度 (Rotational Velocity); v_{sun} 为太阳围绕星系中心的速度; l 为以太阳为中心的银道坐标系下的银经; R 为分子云相对星系中心的距离; R_{sun} 为太阳相对星系中心的距离; d 为待求的分子云距离，具体计算参见文[101]。

工具包的设计思路参考了开源工具包 FRELLED[8]，使用体绘制技术展示三维数据单元，并将整个可视化工具分为三大模块，分别为数据预处理模块、三维模型重建模块和数据分析模块，每一个模块有一个用户界面面板和相应的一系列控件实现用户交互，这三大功能模块的具体实现方法将分别在 3.1.1.2、3.1.1.3和 3.1.1.4节进行阐述， 3.1.1.5节对当前工作进行总结。

3.1.1.2 数据预处理

对标准 FITS 格式的谱线数据进行坐标转换、图像提取等操作，为后续的三维建模和数据分析模块做准备，预处理过程主要分为以下三个步骤:

[8]http:// www.rhysy.net/frelled.html

1. 通过 Python 的 PYFITS 扩展库[102,103] 读取 FITS 文件的头信息中的关键字值和数据部分内容，对应关键字含义可参考 FITS 格式说明文档[70]。

2. 根据变量 naxes 的值，对使用了齐次坐标的数据集进行格式转换，将数据集转换为图像坐标下的的三维坐标格式。

3. 格式转换完成后，通过 Matplotlib 中的 Pyplot 库，将三维 Data Cube 数据按等间距不同方向投影分割成成一系列切片图（Slicers），并通过设置图片的 Alpha 值[9]来改变图片的透明度。为了获得更好的可视化效果，我们需要对每个切片图片做“剪裁”操作，也就是设定每个切片所要展现的通量数值范围，最后将约 900 张图片集存放在指定位置备用。处理前后对比图如图 **??**所示，通过对比我们可以发现 adaptive 方法可以有效的减少每个切片图片的噪声，这样在重建 cube 后使得数据结构更清晰。

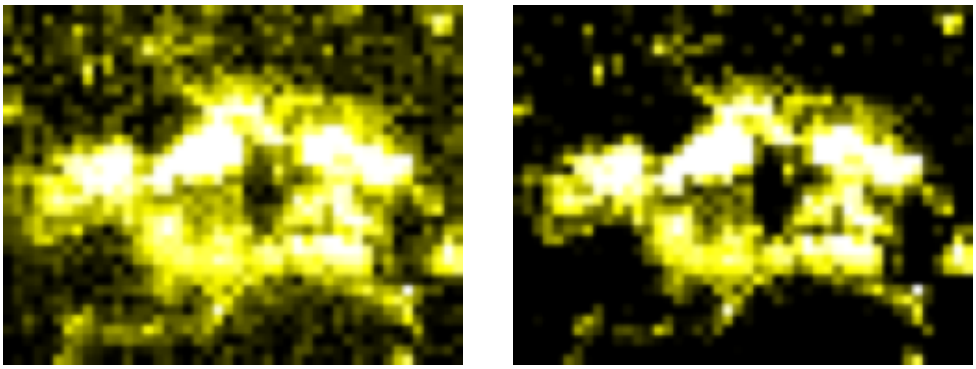

图 3.2: ”Adaptive” 处理前后效果对比，左为处理前，右为处理后。

3.1.1.3 三维模型重建

将已处理的图片作为纹理贴图到新建的 Object 对象上，并完成图像坐标像世界坐标的转换，设置 Object 对象纹理时使用的是 UV 贴图，可以在 Blender 的 3D Viewport 中可以不通过渲染，使用 Texture Mode 实时、多角度观察重建的三维模型的结构特征，具体流程如 3.3。

3.1.1.4 数据分析

因为 Blender 内置 Python API 的特点，我们可以基于现有的 Python 科学数据扩展包进行数据分析功能的开发和扩展，这里仅对已开发功能做一个简单介绍。

[9]Alpha 通道是一个 8 位的灰度通道，该通道用 256 级灰度来记录图像中的透明度信息，定义透明、不透明和半透明区域，其中黑表示全透明，白表示不透明，灰表示半透明

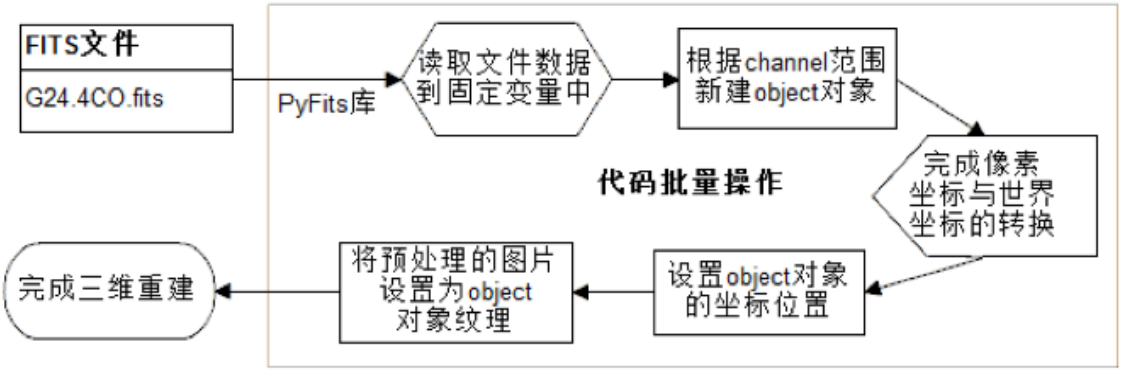

图 3.3: 三维模型重建流程示意图。

1. 在三维模型重建完成后，选定特定区域进行数据分析，在 Blender 的四视图下，可以准确定位兴趣区域。

2. 通过 Python 的 Matplotlib 绘图库，绘制选取区域的等高线图（Contours）和三维等高线序列图。

3. 自动将相关参数传到 SDSS 或 NED 数据库获得相应的光学波段成图数据，并可以自动保存数据图像以进一步探究该区域是否有恒星或星系。

3.1.1.5 小结

本工具具有良好的可移植性，不需要配置使用环境，在 Blender 中导入脚本即可，且图形化界面易于使用和操作。功能上实现了对 FITS 格式的三维数据单元数据的实时可视化及简单分析，并具有可扩展性，可以通过 Python 编程补充和完善现有的数据处理功能。

传统的二维处理方式不能解释隐藏规律的特点，而利用三维可视化能够从大量的数据中提取有用的信息，或者得到其他方式不容易观察的数据特征。对于三维数据单元数据，二维图像展示方法丢失的在频率维度的信息可以通过这个工具被展示，根据对应的旋转曲线和速度分布推导出其空间分布关系，并还原到三维模型中，从而得到整体的结构信息。从图 3.4对 G25.4-0.14 分子云的二维和三维展示的效果对比中可以发现，在二维图中的结构看上去很像一个环状结构的分子云，但在三维图中环状结构并不明显，很可能是由于投影效应而形成二维的环。根据 3.1.1.1式，还可以把径向速度和分子云的距离相对应。第三维度的径向速度实际上可以看做是分子云的距离，因此图 3.4展示的是分子云的近似三维空间结构，直接展示了巨分子云复杂的三维空间分布，其中有很多子结构是二维图像无法展示的。

由此可见三维可视化对研究分子云、星系等复杂结构的天体特征及其演化，建立物理模型等有较大的帮助。

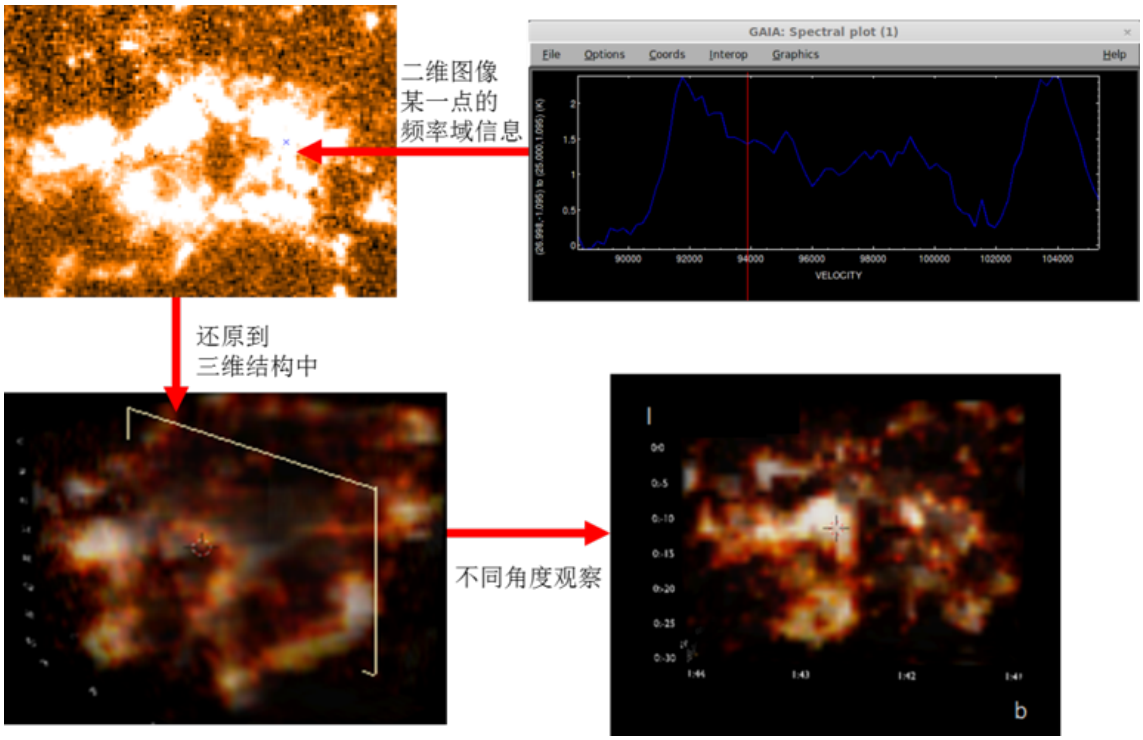

图 3.4: 展示了不同维度的演示效果。左上为利用 Starlink 软件展示的频率在 Frequency=466 通道 (115.23 GHz) 的二维图像，右上为二维图像中某一点的 CO(1-0) 谱线图，下面两幅是 Blender 的三维重建模型效果图，以不同的角度展示。

3.1.2 利用 Blender 展示动态 N 体模拟结果

Blender 不仅可以对实测的三维数据单元数据进行三维展示，对于大规模数值模拟 (例如 N-body 模拟) 生成的数据也有比较好的可视化效果。在现代天体物理学中，N 体模拟是研究各种动力学系统演化的一个重要途径。利用现代超级计算机的优势，大规模的 N 体模拟已达到数亿级的粒子数规模。N 体模拟使得全尺度的天体物理模拟得以达到空前的准确率和所需的动态范围，当然同时也对数据存储和可视化提出了巨大的挑战。下面的例子展示了利用 Blender 为 N 体模拟数据进行可视化处理，制作三维动画的过程。

随着 SverreAarseth 在 20 世纪 60 年代开发出第一个 NBODY 代码-NBODY1，时至今日，新版本的 N 体模拟的代码 NBODY6++[59,104] 正在被不断完善和广泛使用中，它是一个在大规模并行计算机组上运行的、通过图形处理器加速的、用来模拟星团的直接 N-body 代码[10]。NBODY6++ 运用四阶 Hermite 积分，并有一系

[10]直接 N-body 代码 (Direct N-body code) 指的是能够直接解运动方程而不需要做任何的假设或

列先进的处理方法例如独立时间帧 (Individual Time Step，ITS) 、Ahmad-Cohen Neighbor Scheme、KS 正则化[60] 等等来加速计算。

本例中的数据通过 NBODY6++ 代码，将以 Plummer Model[11]为密度原型的两个球状星团及其碰撞过程模拟出来。为了重构模拟的动态演化过程，输出时按照固定的时间间隔，创建一系列按时间顺序排列的帧 (Step#0, Step#1, ⋯, Step#n) 。通过 Blender 工具包开发，导入所有帧中的数据并以动态图的形式展现出来，实现 N 体模拟数据的动态三维可视化，具体的数据预处理过程和动画实现方法分别见 3.1.2.1、 3.1.2.2节。

3.1.2.1 数据预处理

1. 本例中将 NBODY6++ 模拟的数据存储在 HDF5 中，HDF5 是一个为大规模数值数据集进行优化的高性能 IO 库。多维度 (包括位置、速度、加速度、密度等) 的粒子数据以分开的数组存储。通过调用 Python 处理 HDF5 格式的扩展库 h5py 包，可以方便地将模拟数据导入 Blender 中。

2. 因为在 NBODY6++ 中应用了独立时间帧策略，在给定的时间点仅有活跃粒子 (active particles) 信息会被整合。星团中心的星经常会交会 (即将碰撞但是没有发生碰撞) ，因此他们的动力学信息必须常更新。相应地，星团边缘的星大多数在不受扰动的轨迹上运动，这也就意味着它们需要少得多的计算。因此，在 HDF5 输出中，仅存储活跃粒子信息，并通过 Hermite scheme 为非活跃粒子插入粒子数据，以节省存储空间:

$$r_{p,i}(t) = r_0 + v_0(t-t_0) + a_{0,i}\frac{(t-t_0)^2}{2} + \dot{a}_{0,i}\frac{(t-t_0)^3}{6} + O(\Delta_r),$$
$$v_{p,i}(t) = v_0 + a_{0,i}(t-t_0) + \dot{a}_{0,i}\frac{(t-t_0)^2}{2} + O(\Delta_v),$$

3. 如果使用回溯查找的方式查询每个缺省粒子的最近更新信息，则每次查找的时间复杂度平均为 $\mathcal{O}(n^2)$ ，效率很低。这里通过“以空间换取时间”, 定义一个定长数组 latest_particle，数组长度为粒子总数，初始化信息为 #Step0 中每个粒子的属性，按对应 ID 存放所有粒子的初始状态信息。在每个时间点读取粒子信息时对 latest_particle 数组进行更新，保证数组中存储的是每个粒子的“最新”信息，这样一来查找的时间复杂度降为 $\mathcal{O}(n)$，减少了时间耗费。

者简化，它的优点是计算结果的高度准确性，缺点则是运算复杂度太高。

[11]The Plummer model 或 Plummer sphere 是一个密度定律，由 H. C．Plummer 第一次使用来拟合球状星团的观测结果。它经常被 N 体模拟用作玩具模型 (toy model) 来模拟恒星系统。

3.1.2.2 可视化实现与结论

本例中通过 Blender 中每一个实例对象对应模拟数据一帧的方式实现模型建立，即将数据预处理得到的每一帧的完整粒子信息，给相应的网格数据对象的顶点赋值，然后根据这些网格数据对象创建对应的实例模型。由于一个模拟数据往往包含多个帧，因此需要建立多个模型，但是 Python 作为脚本语言效率较低，为了提高速度，这一过程可以通过 C 语言扩展实现。

动态演示部分通过逐帧动画[12]方法实现。为每一个实例模型的隐藏属性的设置关键帧及其生存时间 (Life Time) ，保证每个实例模型只在固定的时间出现一帧，由于视觉暂留效应，人眼所看到的则是连续的一段动画，并且因为设置关键帧的函数封装在 BPY 库中，函数调用耗费的时间很低，动画生成的效率较高，并能获得较好的可视化效果，如图 3.5，由四幅按时间序列的图展示了用 NBODY6++ 代码模拟两个星团碰撞的过程，用户可以选择不同的背景和纹理获得更好的渲染效果，生成形象逼真的动态视频，达到良好的宣传和出版效果，在上传的视频[13]中展示了利用 Blender 制作的三维可视化动画。

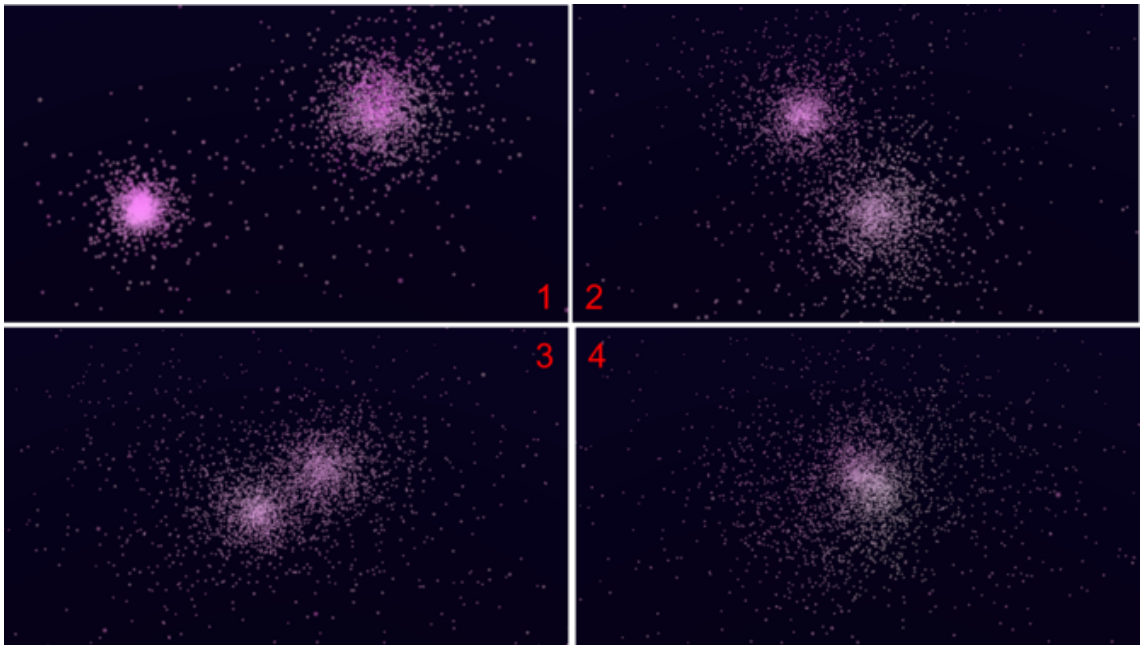

图 3.5: N 体模拟效果图。

由于 Blender 并不是专业的天文软件，对于数据的分析等功能扩展必须依赖于 Python 脚本编程完成，因此相对于 Paraview[14]等专门作数据处理的软件来说，

[12]逐帧动画是一种常见的动画形式 (Frame By Frame) ，其原理是在“连续的关键帧”中分解动画动作，也就是在时间轴的每帧上逐帧绘制不同的内容，使其连续播放而成动画

[13]视频链接 http://i.youku.com/u/UMTQ2MzYxNzY5Ng==/playlists

[14]http://www.paraview.org

性能方面有一定的局限性。但相比之下，专业处理软件往往更专注于分析和计算功能，在可视化效果方面较差，因此可以通过开发展示 N-body simulation 数据的工具包作为辅助工具，将科学成果用更直观、更美化的方式展现给读者。

3.1.3 Blender 总结与 FAST 应用

Blender 对天文研究的优势在于其跨平台（Linux、Mac 和 Windows）运行及图形界面与 Python 脚本编程结合的操作方式[99]。FAST 数据分析与展示很大可能会涉及不同的系统平台，而 Blender 的图形界面则能够在不同系统下都保持显示与功能的统一性，且并不需要额外配置渲染引擎，这将大大简化环境配置的复杂性并提高软件的实用性。此外，Naiman, (2016)[83] 开发了一个开源的 Python 库 AstroBlender，其能够通过封装的程序接口调用 Blender 作为渲染引擎，从而使得天文数据读取、三维渲染和结果输出都可以通过脚本语言自定义完成，这也使得 Blender 能够真正融入 FAST 可视化分析模块作为其一部分。

在展示功能方面，通过上文对三维数据立方体及数值模拟数据进行渲染的实例，我们可以看出作为专业动画软件，Blender 具备高质量的输出结果且高效的渲染能力，通过用户设置，渲染结果往往能够达到出版刊物要求的质量，而 Blender 自带的视频编辑功能也使得天文学家能够通过图形界面操作来获得高质量的视频展示。因此，Blender 也被用在天文外领域例如生物中对蛋白质结构[105] 和地球科学中对火星、地球及月球的表面结构研究[106]。FAST 的中性氢巡天预计将产生大天区范围的高维度数据，我们希望能够通过 Blender 实时渲染高分辨率的巡天天区的二维积分图，这不仅能够直接运用到第二章介绍的可视化分析模块的地图册搜索板块中，作为地图册的背景参考图，还能够为 FAST 巡天进度提供一个直观性的展示结果。在三维显示功能上，我们希望进一步拓展 Blender 对数据立方体的实时展示的可能性，理论上 Blender 并不适合作为实时展示较大的数据集（如百 MB），不仅软件运行速度明显变慢，且产生的中间产品（例如切片图片）也会占用一定的存储空间，但是如果我们能够自动化 Blender 的三维渲染过程，将最终的输出作为实时展示的显示结果，那么凭借其优越的动画特效效果，将会为 FAST 产品的最终展示增添色彩。

3.2 立体图和虚拟现实的三维展示

3.2.1 背景介绍

为了满足对更真实影像的需求，显示技术已从二维发展至三维，除了一般的影像与色彩外，还提供了立体空间的视觉感受，而这种对深度信息的视觉显示对高维度天文数据的理解至关重要。St. John et al. (2001)[107] 和 H. Piringer, et al.

(2004)[108] 的研究都证明三维显示能够消除重叠效应从而更容易直观的发现第三维度展现的结构信息。

天文应用中三维展示通常指三维渲染（3D Rendering）[15]，也就是通过计算机将三维数据渲染为二维图像使其能够在屏幕上显示。一种常用的“非实时”的渲染就是制作视频或图像，来阐述所蕴含的科学问题。这些渲染结果能方便的放入论文发表和网页中用来展示和交流研究成果。通常的渲染结果是静态的，用户通过直接观察便可以获得信息。“Paper of future”[16]中指出，未来的数据展示不单单是“看”，而是用户可以能够自由操作与选取，这样的技术已经存在并且并不难获得，例如 Alyssa Goodman 在其 2008 的 Nature 文章中[109] 使用的 3D PDF 图片，展示 L1448 恒星形成区的 CO 成分自重力三维图，在 PDF 阅读中，读者能够点击图片并通过鼠标旋转观察不同角度，还能够自由选取要隐藏或展示的“零件”部分[17]，而获得这样的 3D 对象可以通过常用的天文软件来生成，如 Paraview 就有保存三维渲染结果为.obj 对象的功能。

当然，非实时的渲染限于数据量较小或者显示种类比较单一的情况。当数据达到 GB 或者 TB 时（FAST 数据便是这个级别），三维渲染的结果往往耗时很大，而且另一个限制是对于不定形态的对象，例如云、尘埃类数据，往往并不能直接用面对象（Mesh）来表示，而常用的三维输出格式.obj 要求其渲染对象是面对象形式。因此，“实时”的渲染在天文数据展示和分析中更直接。通过计算机实时计算出对应交互操作的显示效果，使得用户能够有更大的自由度从二维的屏幕上操作视图角度并观察三维对象，从而从二维显示中获得更多维度的信息。这种三维实时渲染在大型游戏中已经广泛使用，随着高性能计算和 GPU 在天文领域的使用，天文学家已经能够对百 GB 级别的高维度数据进行高质量的实时三维渲染，例如 A. Hassan 等人对 204GB 数据立方体进行每秒 30 帧的渲染[110]，且渲染结果也越来越多的采用色彩、透明度等可视化因素。技术的发展使得交互式三维实时可视化不仅可以用来展示成果，还可以用来进行数据分析。

相比于实时三维渲染，立体显示 (Stereo Display 或 3D Display) 在二维屏幕上显示三维信息的基础上，还能够结合用户的头部或眼部的活动改变或增加显示的三维信息，其基本原理是在二维图片上叠加或者增强关于“深度”的信息，使观察者产生立体视觉[18]。这种立体视觉的原理是由两眼时差所造成的，由于人眼的左眼与右眼相距约 6.5 厘米，在观看物体时的角度略有不同，接收到的影像便有些微的差异。接收到影像后，大脑把略有差异的两影像结合，便成了带有深度资讯的视

[15]https://en.wikipedia.org/wiki/3D_rendering

[16]https://dx.doi.org/10.22541/au.148769949.92783646

[17]https://helpx.adobe.com/cn/acrobat/using/displaying-3d-models-pdfs.html

[18]https://en.wikipedia.org/wiki/Stereo_display

觉影像[19]。

而近年来兴起的虚拟现实技术在天文领域也有一定的影响。随着计算机硬件水平的提升和 GPU 的应用，真实的天文数据能够被快速处理为虚拟环境展示所能容纳的格式并导入虚拟现实设备中进行展示甚至互动，一个开创型例子是 Manitoba 大学对关于星系的谱线射电数据立方体进行展示和交互[111]，该实例是通过使用 CAVE 环境[112] 实现对星系的三维数据立方体进行展示，用户通过两个手柄还可以对展示结果进行一定的交互。并且，随着天文观测数据开始呈现“大天区”的特点[113,114]，同时灵敏度、分辨率、数据尺寸和细节都会增加，创建用户位于数据中的沉浸式球面全景是非常有用的，而这种 360 度天区全景图实现手段也已经被提出[115]。此外，虚拟现实也已被用于天文教育及科普[116] 中。这些都预示着虚拟现实技术或将对 FAST 巡天数据展示提供一个创新点。

本节我们将介绍利用不同的立体显示和虚拟现实技术，使用基于 Pan-STARRS 1(PS1) 巡天[117] 和 The Two Micron All-Sky Survey(2MASS) 巡天的星际尘埃红化的三维天图数据[118]，来展示新型高维度可视化的结果。该天图包含了 PS1 巡天 8 亿颗恒星的高质量光谱，其中约 2 亿条与 2MASS 巡天光谱数据吻合。天区覆盖范围约占总天区的四分之三，直到北天区 $\delta \approx -30^\circ$ ，距离上可延伸到几千秒差距。该天图数据包含了从纤维状到大型云团等丰富的结构，对其进行三维立体展示有利于增强对银河系三维结构的理解与研究。

3.2.2 基于 Jupyter Notebook 和 WebGL 的网页三维显示

Jupyter Notebook[119] 是一个交互式笔记本，支持运行包括 Python 在内的 40 多种编程语言。iPython 作为其 Python 内核，被广泛用于提供交互式 Python 编程，其运行结果可以单独保存为.ipynb 格式的文件，且可以导入 nbviewer（Jupyter Notebook Viewer）进行渲染和 HTML 及 PDF 的输出。也因此，Jupyter Notebook 被认为能够为未来的论文发表提供可交互的附图，读者在网页阅读时能够对附图进行放大、缩小、旋转等一系列操作，同时还可以获得生成附图的代码并对其进行进一步修改利用[20]。而结合 WebGL 技术，我们便可兼容的网页浏览器中渲染 3D 图形的 JavaScript API，无需加装插件，只需要编写网页代码即可实现 3D 图像的展示。

通过开源工具包 iPyvolume[21]，对尘埃三维天图数据进行网页交互展示仅需数行代码便可实现。iPyvolume 是由 Maarten A. Breddels 开发的用来在 Jupyter notebook 对高维度天文数据进行三维可视化的工具包，可以完成包括体绘制、Scatter

[19] 黄怡菁, 黄乙白,& 谢汉萍.(2010). 3D 立体显示技术.

[20] https://www.authorea.com/users/23/articles/8762-the-paper-of-the-future/_show_article

[21] https://github.com/maartenbreddels/ipyvolume

Plots 和矢量图等可视化并提供基本的显示设定。该数据库可以在散点图中渲染达到一百万个点源，而且还可以绘制带箭头的向量场。渲染不仅可以在 Jupyter Notebook 里完成，还可以创建一个单独的 HTML 页面。除了静态渲染，iPyvolume 还支持动画渲染，用于播放带有时间序列的数据。简单的 GUI 控制例如 Slider Bar、按钮等也被整合在数据包中。该数据库是经过高度包装的，因此从数据读取到渲染显示所需的代码编写非常优化，如图 3.6所示，该图给出了在 Jupyter Notebook 环境下，使用 iPyvolume 包对星系尘埃天图数据中距离小于 $1Kpc$ 的 Orion 环区域进行渲染的结果，相关代码和渲染结果可参考 Github 页面[22]。在全屏模式下，用户可以虚拟现实头戴式显示器（例如 Google Cardboard）结合智能手机，透过眼镜观看立体显示的结果，具体细节我们将在下一节展开。

图 3.6: 该示意图显示了在 Jupyter Notebook 平台，通过 iPyvolume 工具包对尘埃三维数据中特定区域进行网页交互式展示的结果。

3.2.3 三维立体电影展示三维尘埃天图全景

网页端的三维显示对于某块天区非常实用，但是相对于巡天数据范围广、数据量大、坐标系统复杂的特点，自定义程序编写并设置渲染条件更切合天文需求，而程序编写则要求我们对三维图片的生成过程有深入的了解。通过上一节的讨论我们理解到，想要在二维图片上展示三维效果，我们需要叠加或增强图片深度信息（获取视差，生成有差异的图像）。一种较直接的获取视差的方式是在渲染过程

[22]http://nbviewer.jupyter.org/github/PennyQ/ipython_scripts/blob/master/ipyvolume_3d_dust.ipynb?flush_cache=true

中增加相机数量，通过两个或多个相机之间位置差异来模拟人眼视差。而另一种比较复杂且较多用于电影行业的技术是后期制作中将二维转换成三维[23]，把画面中需要前后分开的各个物体 roto 出来，然后人工地赋予每个层特定的深度值，然后再通过算法转换将深度值转换成视差值并结合原始图像生成最终立体图像。

三维立体图片最经典的就是 Anaglyph 影像，其包含两个透过不同颜色的滤镜的图像，而且会重叠成一个图像。这个解决方案的原理是，一个影像由红色滤镜显示，而另一个影像则由青蓝色或蓝色滤镜显示。使用者需要一对特别的红蓝眼镜来观看看影像的立体 3D 效果，其中红蓝色镜片分别置于左右两边眼睛。在这种情况下，使用者的每一只眼睛只能看到独有的影像，不能看到另一只眼睛所看到的影像。图 3.7给出了一个 Anaglyph 影像的示例，借助 3D 红蓝眼镜的辅助，用户可以从二维屏幕上感受到三维深度信息。

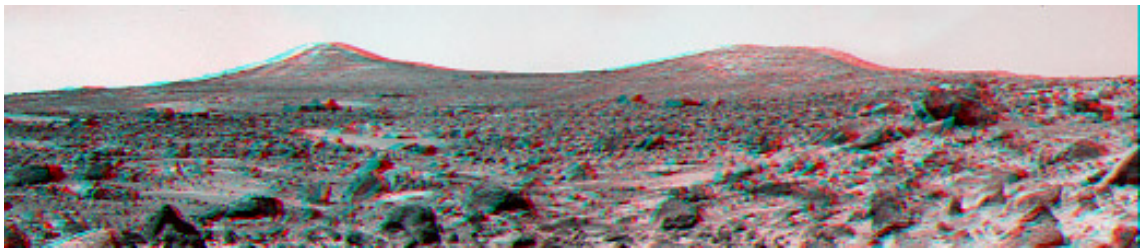

图 3.7: 该图给出了一个从 NASA 的火星探路者 (Mars Pathfinder) 任务收集到的图像，此图片进行前后景相位加工，给出了生动的由近及远的火星地表特征。图片来源:https://en.wikipedia.org/wiki/Anaglyph_3D

对尘埃三维天图进行三维立体化也是同样的原理。原有显示尘埃三维天图的视频在尘埃三维天图网页被展示[24]，其通过设置相机不同位置和移动轨迹，实现了对 Local Dust（即相机正对银河系中心相反点），绕转太阳（相机以 50-pc 为半径轴围绕太阳）以及 Galaxy Tour（相机以几 kpc 为半径轴巡回银盘）等不同角度的渲染。我们通过以下向量运算，三维立体显示在保持原相机运动轨迹的基础上，通过双相机同时渲染从而获得深度信息并合成三维立体图:

$$\delta = \vec{N} \times \vec{AF} \div (|\vec{N}||\vec{AF}| \sin\theta) \cdot |\vec{A1}|$$

其中 $\vec{N}$ 是朝向坐标系北极的单位向量，A 为原中心相机点，A1 和 A2 为新添加的左相机点和右相机点，F 是相机焦点位置，而 θ 是左或右相机偏离中线的角度。求得的 δ 为左右相机相对于中心相机的偏移向量，由于中心相机位置已知，因此由 $A1 = A + \delta$ 和 $A2 = A - \delta$ 求得左右相机的位置。

[23]https://www.fxguide.com/featured/art-of-stereo-conversion-2d-to-3d-2012/

[24]http://argonaut.skymaps.info/

合成效果的优劣关键在于双相机间距和相机距离投影平面的距离的选择，这里就需要对汇合角度和正负视差的定义先做一个解释。如图3.8所示，双眼汇合角度是双眼与被观测物体产生的夹角，角度愈大，物件愈感靠近，相反的，角度愈小则物件愈觉得远离。而该角度过大意味着物体太近，以致于双眼对焦造成不适，而太小则表示物体太远立体感会丧失。根据双眼汇合点与屏幕的前后位置关系，我们可以将视差效果分为正视差和负视差。当目标物在左眼图像中向右偏，在右眼图像中向左偏，双眼焦距（汇合点）会被引导落到显示器的后面如图左半部所示。而当目标物在左眼图像中向左偏，在右眼图像中向右偏，双眼焦距（汇合点）会被引导落到显示器的前面，如图右半部所示。

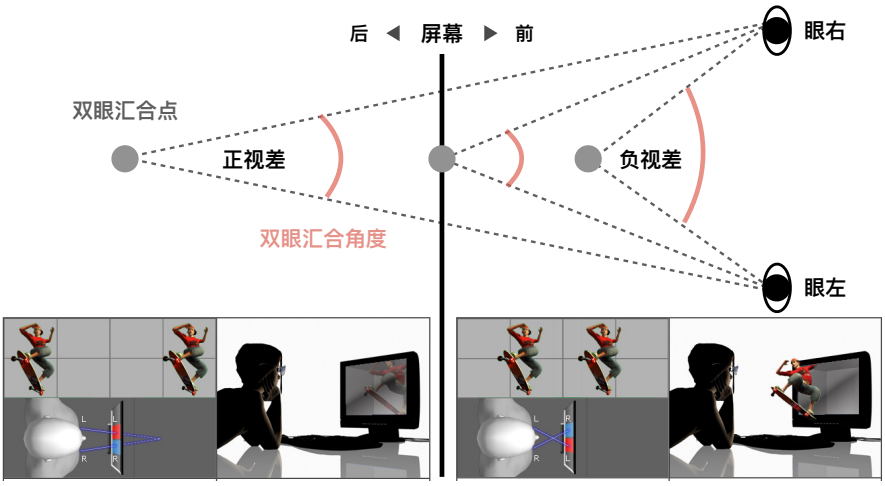

图 3.8: 该图显示了双眼汇合角度和正负视差的原理，对应汇合点在屏幕位置的不同，观察者所看到的立体图像会呈现出陷在屏幕中或凸出屏幕外的特点。图下部两个示意图来源：reallusion 网页

通过调研我们发现[25]，几个决定立体感的判断标准在于：相机间距应小于或等于相机距离投影平面距离的 1/20，这也是舒适观看的最大可接受间距；另一个限制是保证正负视差在合理范围内，且负视差不超过自然双眼间距。一个常用的判断手法是利用定义:

$$\theta = 2\arctan(DX/2D)$$

其中 DX 是投影点在屏幕上的左右位移，D 是相机距离投影平面的距离。对于屏幕上所有的投影点，其 θ 的绝对值不应超过 $1.5deg$，这是绝大多数观众所容易定影的角度。而对于负视差，θ 值为负数，鉴于当负视差情况下部分物体切出了投影平面从而加大定影的困难，限制 θ 值更接近于 0 也是一个可以增加立体感的因素。

[25]http://paulbourke.net/stereographics/stereorender/

基于以上规则和现有的 Python 代码基础上，我们增加了立体显影的渲染模式，对两种运动轨迹：绕行太阳[26]及 Grand Tour[27]进行了渲染、视频编辑及画面增强，所使用代码和文档保存在 Github 代码仓库[28]。

3.2.4 虚拟现实和 360 度全景图

虚拟现实系统与传统的普通显示设备相比，具有下面三个基本特征：即三个“I” 即 immersion - interaction - imagination（沉浸—交互—构想），人在虚拟系统中的主导作用被大大增强[120]。传统设备下，人只能从计算机系统的外部去观测处理的结果，通过键盘、鼠标与计算环境中的单维数字信息发生作用，只能以定量计算为主的结果中启发从而加深对事物的认识。而在虚拟现实环境下，人能够沉浸到计算机系统所创建的环境中，利用多种传感器与多维信息的环境发生交互作用，并有可能从定性和定量综合集成的环境中得到感知和理性的认识从而深化概念和萌发新意。也正如 Bryson 在其 “Virtual reality in scientific visualization”[121] 中对虚拟现实的定义：“利用计算机和人机交互接口，来产生一个包含可交互对象的三维空间，且使用者有强烈的三维立体感。”

事实上，虚拟现实与天文结合并不难实现。利用 Cardboard 和智能手机就可以组成一个虚拟现实设备。一种通过 OpenGL 三维渲染真实数据并网页显示的方式已在上一节提到，而另一种非常适合 “沉浸式” 观察的方法则是 360 度天区全景图。广义上的全景图是视角超过人的正常视角的图像，而 360 度全景图指水平视角 360 度，垂直视角 180 度的图像。而此图像最大的三个特点是：全方位，能够全面展示 360 度球型范围内的所有景致；在例子中可通过输入设备（如鼠标）进行交互，观看场景的各个方向；实景，三维全景大多是在照片基础之上拼合得到的图像，最大限度的保留了场景的真实性；三维立体效果，虽然照片是平面的，但是通过软件处理之后得到的三维全景，却能给人以三维立体的空间感觉，使观者犹如身在其中[29]。

普通天文全景图像，除却长宽比例为至少为 2：1，还需要对其元数据，也就是关于文件的一组标准化信息如作者姓名、分辨率、色彩空间、版权以及应用于文件的关键字等，进行编辑与修改，从而全景图能够被正确解读和显示。一个来源于 Google Street View 的全景图元数据模板[30]如下所示：

```
<rdf:Description rdf:about="" xmlns:GPano="http://ns.google.com/
```

[26]https://www.youtube.com/watch?v=y1ce6z0EY0A

[27]https://www.youtube.com/watch?v=LER5cIwhppo

[28]https://github.com/PennyQ/stereo_3D_dust_map

[29]https://baike.baidu.com/item/360%E5%BA%A6%E5%85%A8%E6%99%AF%E5%9B%BE?fr=aladdin

[30]https://developers.google.com/streetview/spherical-metadata

```
photos/1.0/panorama/">
    <GPano:UsePanoramaViewer>True</GPano:UsePanoramaViewer>
    <GPano:CaptureSoftware>Photo Sphere</GPano:CaptureSoftware>
    <GPano:StitchingSoftware>Photo Sphere</GPano:StitchingSoftware>
    <GPano:ProjectionType>equirectangular</GPano:ProjectionType>
    <GPano:PoseHeadingDegrees>350.0</GPano:PoseHeadingDegrees>
    <GPano:InitialViewHeadingDegrees>90.0</GPano:InitialViewHeadingDegrees>
    <GPano:InitialViewPitchDegrees>0.0</GPano:InitialViewPitchDegrees>
    <GPano:InitialViewRollDegrees>0.0</GPano:InitialViewRollDegrees>
    <GPano:InitialHorizontalFOVDegrees>75.0</GPano:InitialHorizontalFOVDegrees>
    <GPano:CroppedAreaLeftPixels>0</GPano:CroppedAreaLeftPixels>
    <GPano:CroppedAreaTopPixels>0</GPano:CroppedAreaTopPixels>
    <GPano:CroppedAreaImageWidthPixels>4000</GPano:CroppedAreaImageWidthPixels>
    <GPano:CroppedAreaImageHeightPixels>2000</GPano:CroppedAreaImageHeightPixels>
    <GPano:FullPanoWidthPixels>4000</GPano:FullPanoWidthPixels>
    <GPano:FullPanoHeightPixels>2000</GPano:FullPanoHeightPixels>
    <GPano:FirstPhotoDate>2012-11-07T21:03:13.465Z</GPano:FirstPhotoDate>
    <GPano:LastPhotoDate>2012-11-07T21:04:10.897Z</GPano:LastPhotoDate>
    <GPano:SourcePhotosCount>50</GPano:SourcePhotosCount>
    <GPano:ExposureLockUsed>False</GPano:ExposureLockUsed>
</rdf:Description>
```

我们通过元数据编辑工具如 ExifTool[31]，及 Python 编程如 Python XMP Toolkit 包[32]，将以上信息添加到图像数据的元数据中，再将处理后的图像传到全景图显示软件例如 vrEmbed 全景图网页显示工具[33]，以实现对天文图像进行虚拟现实的转换。具体代码和步骤可参考我们归档记录的 Github 代码仓库页面[34]。具体效果如图 3.9所示，该图将尘埃三维天图数据的 SDF 图像[122] 通过 vrEmbed 网页服务进行在线显示。

3.2.5 混合现实和微软 HoloLens

混合现实（Mixed Reality，简称 MR）是结合真实和虚拟世界创造了新的环境和可视化，混合现实环境下的物理实体和数字对象共存并能实时相互作用，来模拟

[31] https://www.sno.phy.queensu.ca/ phil/exiftool/

[32] https://github.com/python-xmp-toolkit/python-xmp-toolkit

[33] http://vrembed.org

[34] https://github.com/PennyQ/stereo_3D_dust_map/tree/master/360-sphere-photo

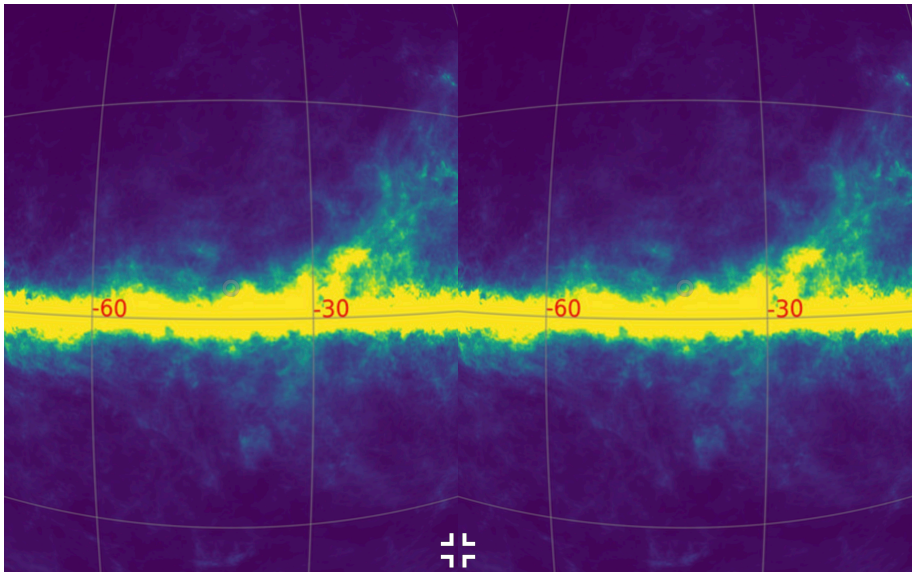

图 3.9: 该图显示了我们结合立体显像原理及 Python 编程实现了网页端对尘埃三维天图 SFD 图像的立体显影，该显示结果具有高度自定义化，包括对颜色、分辨率和网格坐标表示，结合谷歌的 Cardboard 和智能手机就能够进行 360 度全景体验，可交互网页页面可参考 tiny.cc/SFD-stereo

真实物体[123]。与虚拟现实（VR）不同的是，虚拟现实更多的为用户展现的是虚拟环境，而混合现实技术不仅限于真实或虚拟的世界，而是虚拟与现实混合的世界，通过沉浸式技术实现增强现实和增强虚拟，如 Paul Milgram[8] 提出的 “现实-虚拟” 区间 (reality-virtuality continuum)（如图 3.10）中所示，混合现实既包含了增强现实也包含增强虚拟。而增强现实领域最突出的设备便是微软 HoloLens[35]。

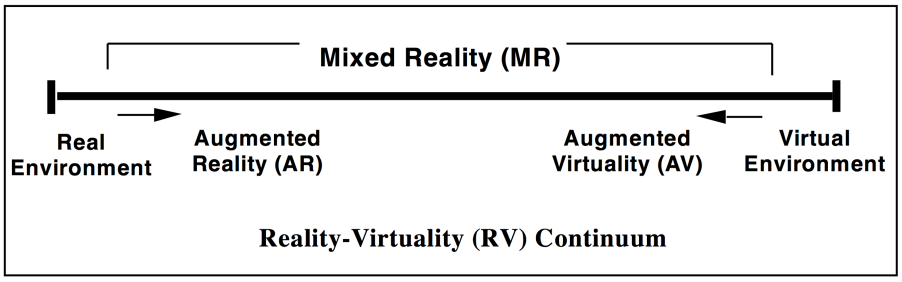

图 3.10: 该图引用 Paul Milgram[8] 提出的“现实-虚拟”区间 (Reality-Virtuality Continuum) 图。区间向左至右依次表示现实环境、增强现实 (AR, Augmented Reality)、增强虚拟 (AV，Augmented Vituality)，直到向右至无穷表示虚拟环境。而混合现实区间包括了 AR 和 AV 两个部分。

微软 HoloLens 是一个基于 Windows 10 系统的智能眼镜产品。它采用先进的传感器、高清晰度 3D 光学头置式全角度透镜显示器以及环绕音效。它允许在增

[35]http://hololens.com

强现实中用户界面可以与用户透过眼神、语音和手势互相交流。HoloLens 使用的传感器是一种高效节能的深度摄像头，具有 120°×120° 的视野。传感器提供的其他功能包括头部跟踪，视频拍摄，以及声音捕捉。除了高性能的 CPU 和 GPU，HoloLens 带有全息处理器（HPU）这一协处理器用于从所述的各种传感器集成数据，并处理诸如空间映射，手势识别和语音识别的任务。

HoloLens 开发大多是通过 Unity 平台[36]，Unity 是类似于 Blender 的利用交互的图型化开发环境为首要方式的多平台的综合型游戏开发工具，它能够轻松创建三维视频游戏、实现建筑可视化和实时三维动画等，用户也可以通过编程进行高阶开发，其主要编程语言为 C#。通过调研我们发现，使用 Unity 对天文数据三维建模的最重要的两个步骤是[124]：

加载天文数据到 *Unity* 环境下： Unity 并不是专业的天文学建模软件，所以常用的天文数据格式例如 FITS 并不能直接加载到 Unity 中。考虑到 Unity 的主要编程语言是 C#，因此我们可以将天文数据转存为编程语言通用的二进制文件 (Binary File)，这个步骤可以通过已存在的天文工具包例如 AstroPy 来完成。被加载到内存的数据在 Unity 中显示为三维纹理（3D Texture），这个三维纹理可以被看做是一个对照表，用来给出三维坐标的对应方程。这个对应方程可以被 Unity 存为 Unity asset 而且可以被重复使用。

三维体渲染 - *Ray Casting* 算法： 在有了坐标的对应方程后，我们只需要对对应坐标位置进行三维体渲染就可以生成，最直接的方法是通过 Ray Casting 算法[29]：对于要在屏幕上渲染的每个像素，沿着当前视线投射射线，并且沿着该射线，以规则间隔的间隔检索数据。使用这种方法，沿着视线方向积累这些值：每一点（体素）上的值是由色彩（RGB 通道）和不透明度（Alpha 通道）来表示。相关程序代码可参见Github 代码仓库。

在导入真实的天文数据到 HoloLens 环境中后，一个实验性案例是通过 HoloLens 进行虚拟天文课堂演示。其设计思路为，展示者将一副二维图像展示到课堂任意位置，并通过声音指令实时渲染对应数据的三维展示，对于数据中的任一兴趣区域，通过 AirTap（选取手势）或预定的声音指令进行选取后，更加详细的三维模型将被渲染并可以由手势置于教室任意位置供观察者进行任意角度的观察，由图 3.11所示，具体视频可参考YouTube 视频链接。

[36]https://unity3d.com

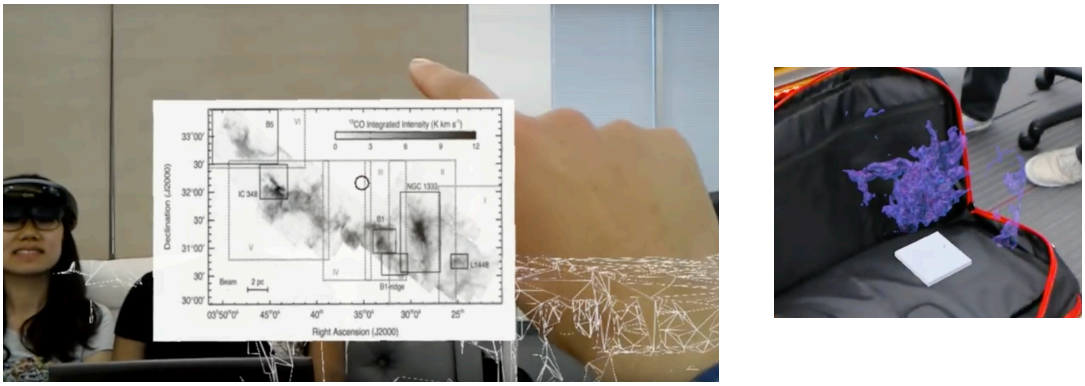

图 3.11: 该图展示了我们设计与开发的 HoloLens 虚拟天文课堂应用的场景与混合现实渲染效果。左图显示了老师将图像共享显示给学生并进行交互式讲解的场景，图中二维图像展示了 Perseus 分子云团的积分强度图，来源于 [H. Arce, et al. 2010][9] 的 Fig. 1；右图是对 Perseus 分子云团大天区中某一块兴趣区域进行选取并置于现实场景某一特定位置供学生多角度观察的场景，图中显示的是 L1448 恒星形成区的 CO 成分的三维等值面渲染结果。

3.3 总结

天文从传统上来说是一门观测科学。现代天文学的蓬勃发展离不开先进的观测设备。这些分布在全球各地以及大气层外的望远镜，一方面提供了高质量的数据，另一方面也产生了海量的、高维度的、具有复杂逻辑关系的数据。现代计算机图行学和高性能图像加速硬件的蓬勃发展，开创了天文数据可视化的新局面，也更加丰富了的数据产品的展示。

作为传统数据处理模式的补充和完善，三维可视化不仅可以补充二维显示对信息丢失的缺陷，还可以通过分析整体结构获得更多的信息。对于天文学这样一门以数据为研究基础的学科来说，三维可视化研究正逐渐成为数据分析中不可缺少的一部分。

本章节调研并应用新的三维可视化技术与软件，为 FAST 的数据产品展示提供了新的角度和方法。首先，通过对 Blender 软件的天文可视化模块的开发，完成了对天文数据立方体和数值模拟数据的实时三维展示，证实了随着观测数据量的增长及对三维渲染引擎的需求，Blender 作为可高速并行计算的可扩展三维建模软件，能够通过一定的开发对数据进行展示，然而，这种展示并不是真正意义上的三维体渲染，其通过渲染一系列二维切片并叠加的方法来模拟三维结构，产生了不必要的中间输出并增加存储负担，且在某些角度下切片间隙会很大程度上影响最终可视化效果，鉴于此，基于 Blender 作为 FAST 高维度数据的展示工具，在数据量和数据分辨率上较有局限。但是，将 Blender 单独作为渲染引擎（如 AstroBlender[83]），则是对天文领域专业渲染软件不足的现状的一个补充，并能够直接导出三维体渲

染结果，为 FAST 实现高质量高效率的三维渲染结果。

其次，通过使用 Jupyter Notebook 和 WebGL 技术，实现对 FAST 数据的网页端实时三维渲染，在静态渲染情况下，通过一定的立体显像技术来叠加三维信息并配合虚拟现实设备例如 3D 眼镜，可以实现对 FAST 数据和成果的较易实现但又十分新颖的“三维”展示，在可交互情况下例如使用可跟踪用户动作的设备如 Google Cardboard，便可以进一步拓展数据展示范围，进行对全天图数据的展示。这可应用于 FAST 中性氢巡天即将产生的大天区数据展示。

最后，虚拟现实技术作为超三维显示技术，在显示的同时还能够提供交互与选取等功能，这不仅能够大大增强观察者的可视化体验，探索更自然的三维选取模式，同时还能够提升 FAST 数据成果展示的吸引力，加深科研影响力与天文科普的吸引力，并且当科学家置身于天文虚拟环境中并能够自由的观察交互，这很大可能能够启发科学灵感。

第四章　批量任务管理工具的开发

4.1　背景介绍

批量任务管理在数值模拟中十分重要。数值模拟被广泛用于研究各种尺度的动力学系统，在天体物理学中，这些尺度范围从行星系统、开放的团簇和球状星团，到星系甚至是大宇宙尺度[125]。理论上数值模拟程序从提交到计算机队列的那一刻开始，就能够保持运行状态直到达到终止标准。然而实际上，由于软件和硬件的问题，数值模拟程序往往会由于各种原因而中断，包括常规计算机维护、作业调度限制、停电或非计划中断。在发生中断的情况下，需要进行人员监督来纠正错误，例如通过调整输入参数并随后重新提交模拟。重新启动模拟程序容易出现人为错误。随着研究中不断增长的规模和精度要求，多次长时间模拟的手动记账变得越来越困难。也因此，开发自动化记录（bookkeeping）功能是代替人工繁杂操作的最直接有效的方法。

同样的，FAST 的关键科学项目之一，即脉冲星搜寻也对批量任务管理有极大的需求。FAST 的脉冲星搜寻将使用 19 波束馈源阵列接收机，其覆盖频率范围是 1.05-1.45 GHz[126]，预计数千个新的脉冲星会被发现。然而，搜索脉冲星需要极高的计算能力，而且在测量中会产生大量的数据。FAST 的采样率为每秒 20,000 次，数据记录在 16K 通道数字后端。因此，每个波束接收器每秒将产生 80 MB 数据。目前，每 30 秒将数据记录到文件中，每个文件的大小为 2.4 GB。19 光束接收器将每小时生成 2280 个文件，每 10 小时夜生成 22800 个文件。每个文件将使用 PRESTO 软件套件[127] 进行处理，以搜索周期信号并识别脉冲星。一组计算机将用于在并行计算环境中处理数据。因此，每次运行的过程将会超过几百个，每个进程可以生成多达几十个脉冲星候选。这些候选体中有一些来自射频干扰（RFI），需要通过人或智能人员进行检查。因此，密集记账需要跟踪所有的数据文件、流程、RFI 和脉冲星候选，以及相关的脉冲星参数。当数据传输速率较低时，这种记录在传统的脉冲星搜索中通过人工手动处理，但在 FAST 脉冲星数据处理中变得不可能。

同时，对脉冲星观测的密集记账也同样需要重启的操作。在脉冲星观测中，当存储设备空间接近饱和时，硬件的不稳定会造成观测数据的记录不定时出现中断并丢包，而这类问题的解决方式往往是购置新的存储设备进行扩容，通常需要一定的时间进行购置和安装，但在此期间观测任务仍然安装计划安排进行，记录的中断监测与重启则需要人为的干预，而人为操作不仅会耽误第一时间对观测记录的续写而造成数据丢包，还容易因此多次操作产生失误从而使得记录结果产生混乱。因此，在开发过程中的 FAST 数据处理流水线中，自动化跟踪所有脉冲星搜索过

程和数据文件以及记录脉冲星候选和参数十分关键。

鉴于这些需求，我们开发了 SiMon（Simulation Monitor），也就是用于检测数值模拟程序的开源 Python 工具，作为对批量任务管理快速增长的需求的响应。通过后台进程模式（Daemon Mode）定期检查模拟代码的运行过程，并从输出文件中得到有效信息并记录。更加自动化的是，当 SiMon 检测到中断时会自动重新启动程序，并实时的备份运行状态与输出结果。该工具的主要目的和优点是应用自动化工作流程，以便从生成初始条件到监控和控制任务进行，直到完成所有任务并且正确处理所得到的数据为止。因此，天文学家只需要指定初始参数空间，并指派工作流模式，那么 SiMon 就可以自动监控大批量任务的进行直到达到终止目标。这对于涉及大规模参数空间和大量数值研究以及从观测数据处理大量数据集来说尤其有用。

需要指出的是，SiMon 不仅实现了作业排队和优先级管理，而且还解决了作业监视和参数空间调度，而传统的作业调度程序，如 Slurm[1] 和 OpenLava[2]，则不具备这些功能。在天文研究中，如运行数值模拟或进行天文观测数据处理时，一个作业经常会由于代码错误或硬件问题而中断。传统的作业调度软件不具备监视功能，不能重启中断的进程。SiMon 不仅可以重启被中断的进程，而且能自动分析日志文件，找出中断的原因，并自动调节被中断作业的参数。更近一步的，天文学家只需要指定待运行的作业（即参数空间），SiMon 将根据当前系统的软硬件资源自动并行地调度。与传统的作业调度软件要求用户手动地逐一提交作业相比，工作效率大大提高。

4.2 设计与实现

4.2.1 基于农场的设计思路

现代农场的自动化运行机制与批量任务管理有很多相似的地方：大片作物在田间同时生长，使产量最大化; 同样的，数值模拟程序在计算机上并行运行，并期待以最小化等待时间获得运行结果。这样一个类比也延伸到运行一个数值模拟集合的生命周期，如下所示：

准备土壤: 对农场而言，适宜的土壤需要在种植作物前准备好。同样，在进行调度观察数据处理流水线计算之前，也需要对计算底层环境进行配置。该准备包括检测硬件环境（例如，确定可用计算，存储器和存储资源）并配置软件环境（例如，编译数字代码和库依赖性，在计算群集上部署作业提交脚本等）。

播种: 每种作物需要种子才能生长。同样，每个模拟都需要一套初始条件来启

[1]https://slurm.schedmd.com

[2]http://www.openlava.org

动。对于流水线调度而言，种子播种是在参数空间中产生相应的初始条件。为了保持清楚的数据结构，每个数值模拟任务的初始条件和模拟输出数据都包含在一个单独的目录中。因此，该播种环节也包括为每个模拟创建适当的目录结构。

培育: 培育是照顾或种植作物的行为。农场中作物生长于调度和启动数值模拟相似，培育作物于监控模拟程序相似。此外，数值模拟程序可能会遇到中断等不确定因素干扰，所得到的数据应该被正确备份，并且崩溃的模拟能够重新启动。当模拟任务完成时，应该使用释放的计算资源来安排下一个模拟任务。

收获: 模拟程序结束后，通过处理数据可以 “收获” 模拟结果，例如生成图表以及转换为研究人员后续数据分析需要的格式。

另外，正如农作物易受害虫影响，模拟任务也容易受到软件错误和硬件问题的影响。模拟在从新程序（NEW）到程序结束（DONE）的过程中，可能会经过运行（RUN），暂停（STALL）和终止（STOP）等状态。如果模拟任务经历了重复的中断，它可能转入错误（ERROR）状态，这就需要人为的监督。本质上，每个单独模拟的生命周期可以被抽象为有限状态机，如图 4.1所示。因此，模拟农业管理模式的主要目的是为了便于将状态机集合自动地从 “NEW” 的状态转变为 “DONE” 的状态。因为我们希望整个过程中能最低限度人力监督，因此整个管理过程应该更加自动化，同时也允许用户在必要时进行手动控制。

4.2.2 软件设计流程

SiMon 工具是一个基于开源的和 Python 编程语言的自动数值模拟任务管理工具的轻量级实现。本节详细介绍了对管理过程的四个阶段（准备土壤，播种，培育和收割）的技术实现，以及对用户界面和可扩展性进行讨论。

***1.* 准备土壤：设定环境变量**

在开始运行数值模拟程序之前，SiMon 工具需要一些全局环境变量，在设定文件 `SiMon.conf` 中能够自由设定。随着用户启动 SiMon 管理工具，设定文件也会被加载。我们列出了关键性参数的设定如下：

- `Root_dir`: 用于模拟数据存储的根目录。SiMon 假设在根数据目录下，每个模拟都有自己的目录用于存储初始条件和生成的数据。
- `Daemon_sleep_time`: SiMon 会检查所有数值模拟任务的间隔时间段。
- `Max_concurrent_jobs`: 同一时间需要运行的数值模拟任务的数量。
- `Max_restarts`: 一个数值模拟任务可被重启的最大次数，任务会被标为 ERROR 如果运行次数超过此值。

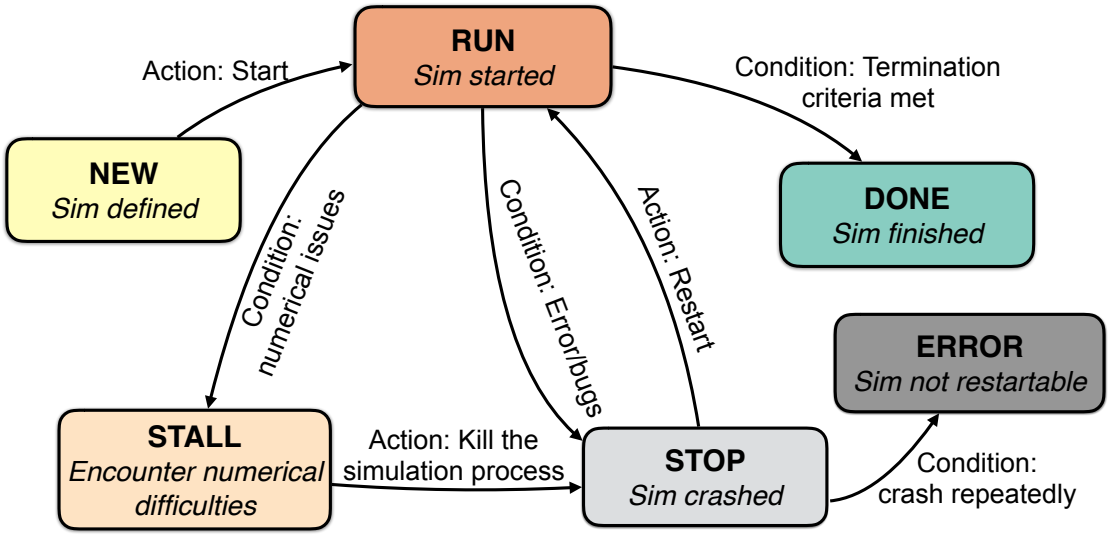

图 4.1: 模型的生命周期的状态机模型。模拟的生命周期可以被建模为有限状态机，其经历几个状态的转变：当代码被初始化并且初始条件被加载时，它进入 “NEW” 的状态。随后，代码进入模型演化阶段，也就是 “RUN” 状态。由于各种问题，运行中的模拟可能会进入 “STALL” 或 “STOP” 状态。如果模拟经历重复的中断，这表明在代码或初始条件中存在错误，需要进行人力监督。在这种情况下，模拟从 “STOP” 转换为 “ERROR”，需要进行人工监督。最终，代码完成了模式的演进，因此进入 “DONE” 状态。除非完成所有的数值模拟任务，否则将触发新的模拟循环。

- `Log_level`: 记录详细程度。越靠后详细程度越低: `CRITICAL`, `ERROR`, `WARNING` 和 `INFO`。默认的记录程度是 `INFO`，此程度上所有的信息都会被记录。
- `Stall_time`: 如果一个数值模拟任务最后一次更新的时间距离当前时间小于该值，那么该任务将被标记为 “STALL”。

2. 播种：部署与设定初始环境

为了在给定的参数空间中初始化一个模拟集合，我们提供了生成初始条件和对每个数值模拟任务配置文件的功能，并在文件系统上部署了适当的结构。为了保持清楚的数据结构，初始条件、配置文件和模拟输出都包含在单独的目录中。

所有的子目录都被收集在父目录中，如配置文件所示。当在集合中迭代运行数值模拟任务时，工作流会解析每个数值模拟配置文件，以获得有关如何控制程序运行和模拟优先级的信息。每个单独的数值模拟配置文件由初始条件生成器根据参数空间规范和全局设置自动生成。如果需要，用户可以重新编辑相应的配置文件来覆盖任何模拟程序的默认设置。

3. 培育：自动数值模拟监测和调度

监控和调度是 SiMon 的核心功能，它们是自动的且需要最少的人力监督。在守护进程模式下，工作流作为服务在后台运行。SiMon 还提供交互式仪表板，以控制交互模式下的数值模拟任务。

在图 4.2中，我们介绍了一般工作流程，可以分为三个步骤：

步骤 1 - 准备：数值模拟的工作目录由配置文件确定。在模拟数据根目录上执行深度优先搜索（BFS）[128] 以构建分层模拟集合。每个模拟任务都有自己的配置文件，该文件被解析以确定应该使用哪个代码并加载相应的模块。

步骤 2 - 监控：准备任务启动运行的输入文件、输出和诊断文件。在后台运行模式下，确定管理模拟的实时状态，并根据状态机模型启动管理动作，如图 4.1所示。在交互模式下，提供信息显示板 (Dashboard)，并允许用户手动控制运行。

步骤 3 - 输出：在交互模式下，数值模拟程序运行的一个概况将被展示，用户可以手动监控和控制任务。在后台运行模式下，其调度算法采取的每个自动操作都记录在一个单独的文件中。

4. 收获：自动化数值模拟数据处理

当一个数值模拟任务完成时，后台处理流水线就会自动启动，例如更新并显示用户运行状态的结果等。

4.2.3 后台运行模式的实现

为了实现 SiMon 自动化后台运行，我们需要训练 SiMon 根据数值模拟任务的状态从而进行对应的操作，即设计程序的决策树模型。机器学习中，决策树模型

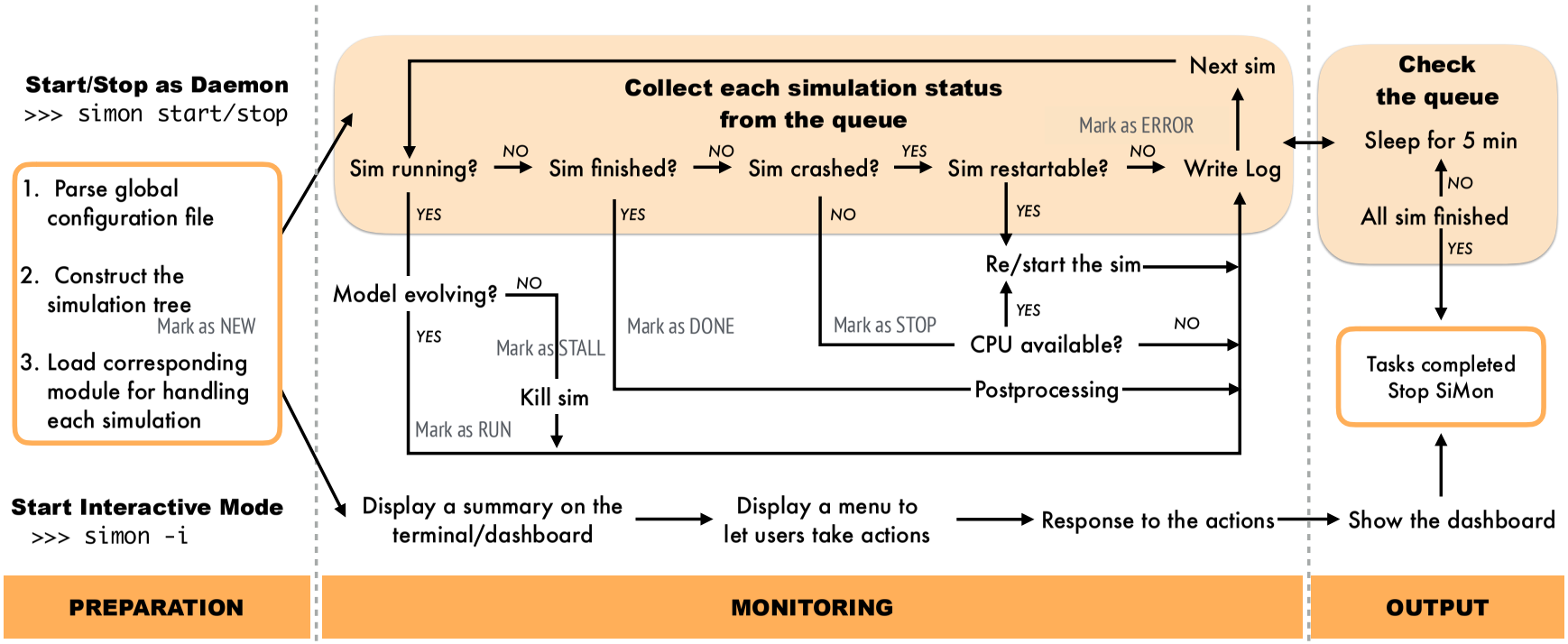

图 4.2: SiMon 的逻辑流程图。SiMon 工具支持两种运行模式：交互模式和后台运行模式。经过准备阶段、模拟阶段和输出阶段，SiMon 将维护一个数值模拟任务的队列的进行情况，队列的更新由后台运行程序定期完成，或当用户调用交互模式时完成。SiMon 收集所有托管的任务的实时状态，并在其交互式仪表板中显示信息。

是一个预测模型，它表示对象属性和对象值之间的一种映射，树中的每一个节点表示对象属性的判断条件，分支表示符合节点条件的对象，树的叶子节点表示对象所属的预测结果，例如对应的操作等。这里，我们借鉴决策树模型中的 CART（Classification and Regression tree）分类回归树，来实现 SiMon 后台逻辑流的设计。主要原因是，对于每个模拟任务的状态判断通常是逻辑上的是与否两种，而 CART[129] 正是一棵二叉树，被用于对连续性关联变量（回归）或类别式变量（分类）的预测。CART 采用二元切分法，通过递归的方式建立树结构，并保证每个节点在分裂的时候都是通过最好的方式将剩余的样本划分成两类。其主要算法总结如下：

输入：训练数据集 D 以及停止计算的条件

输出：*CART* 决策树

根据训练数据集，从根节点开始，递归的对每个节点做以下步骤：

1. 设结点的训练数据集为 D，计算现有特征对该数据集的基尼指数，此时，对每一个特征 A，对其可能取的每个值 a，根据样本点对 $A{=}a$ 的测试为“是”或“否”将 D 分割成 D_1 和 D_2 两部分，并利用公式 $\mathrm{Gini}(p) = 1 - \sum_{k=1}^{K} p_k^2$ 计算 $A{=}a$ 时的基尼指数，其中 K 是类别的数目，p_k 表示样本属于第 k 类的概率值

2. 在所有可能的特征 A 以及它们所有可能的切分点 a 中，选择基尼指数最小的特征及其对应的切分点作为最优特征与最优切分点，依最优特征与最优切分点，从现结点生成两个子节点，将训练数据集依特征分配到两个子结点中

3. 对两个子结点递归调用步骤 *1* 和 *2*，直至满足停止条件

4. 生成 *CART* 决策树

在 SiMon 的实现中，我们对 CART 决策树模型进行了一定的修改定义。首先，特征 A 是由模拟任务状态基和记录文件更新情况共同定义的，类别 K 固定为 2，分别是任务处于该状态与否；其次，我们通过迭代每次将一个任务带入决策树模型，而非一次性输入队列中所有的模拟任务，这种设计能够避免多并发读写可能出现的冲突。程序实现上，我们通过 Python 的 `if-then` 逻辑实现了基于当前任务列表的根据任务状态的操作分类，这里我们检查每个模拟任务运行情况并收集状态信息，随后的动作是根据状态机模型进行的，SiMon 会在特定行动期间进入睡眠状态。重复此过程，直到每个模拟任务完成。我们将主要逻辑部分代码简化为以下伪代码所示：

```
queue = BFS(parameter space)
while(queue is not empty):
```

```
    for sim in queue:
        if sim is running:
            if not sim is evolving:
                Kill(sim)
                Mark(sim, STALL)
        else if sim is finished:
            Mark(sim, DONE)
            Dequeue(sim, queue)
            Finalize(sim)
        else if sim is crashed:
            if sim is restartable:
                if CPU is available:
                    Restart(sim)
                else:
                    Mark(sim, STOP)
            else:
                Mark(sim, ERROR)
                Dequeue(sim, queue)
                GenerateWarning(sim)
        else if sim not crashed:
            if CPU is available:
                Start(sim)
            else:
                Mark(sim, NEW)
        Write(Log files)
    Sleep(a period of time)
Quit(SiMon)
```

而后台运行模式的设计，我们考虑到能够对以下功能的优化：

- *数值模拟数据的自动备份*：模拟代码必须支持重新启动，允许用户继续以前中断的模拟任务，这可以通过周期性存储模拟数据来实现，例如记录重启文件。SiMon 会始终保留两个后续的重新启动文件，以保证在重新启动文件写入期间即使不希望的中断发生，也不会阻止稍后重新启动该运行。
- *自动重启任务*：SiMon 通过一个回滚机制来自动重新启动。当最近的一个重新启动文件被证明是有问题的，例如文件已损坏，那么第二个最近的重新启

动文件将被使用（回滚）。每一个数值模拟程序有一个祖先节点（也就是所有的数值模拟任务从那里开始）和几个子节点（数值模拟任务从那里重启）。在拓扑学上，回滚重启方案形成一个树结构，其中当前的模拟任务是一个分支。自动仿真重新启动的示意图如图 4.3所示，其中回滚方案被示为模拟任务 4，图中的模拟任务 1-3 显示了真实的退化树结构。

模拟任务的状态信息通过树节点传播。当模拟中断时，SiMon 从初始（根）模拟开始搜索重启树。如果重启树存在，则 SiMon 会从所有的重新启动（叶节点）开始搜索，直到达到最大模型时间的节点。这个最旧的节点是重启候选节点。重新启动时，其状态设置为 “RUN”，表示原始仿真正在其重新启动的仿真之一中运行。如果模拟任务经历重新启动，并以状态 “ERROR” 结束，则将该信息传播到其所有子节点，重启树中的该分支将随后终止。

- *自动调度*: 为了最大限度地利用可用的硬件资源，我们设计 SiMon 在计算机资源一旦可用的情况下，就自动安排新任务。任务调度的顺序取决于队列优先级，如图 4.4所示。后台进程定期检查每个模拟的状态，并根据图 4.1中详细描述的逻辑执行操作。相比之下，如果手动管理模拟程序，则当用户在模拟完成或中断时很难立即响应，当机器空转时就会浪费时间。另外，如果用户试图始终保持所有处理器出于繁忙状态（手动并行化），则会更加困难。通过这种基于优先级的调度方案，不仅使空闲时间最小化，而且还能够自动并行化来启动模拟任务。

- *自动化记录*: 由于 SiMon 中的操作都是自动化进行的，因此记录这些操作以备将来参考很重要。我们通过日志文件记录的方式，将后台程序执行的每个操作都记录下来，一个日志文件的例子如下:

```
5/10 0:28AM: sim_1 [INFO] Started.
5/10 1:28AM: sim_2 [INFO] Restarted.
5/10 2:28AM: sim_3 [WARNING] Crashed.
5/10 3:28AM: sim_4 [ERROR] Not restartable.
```

4.2.3.1 交互模式的实现

在交互模式下，SiMon 给用户提供了所有数值模拟程序的当前状态，并且允许用户干预。交互模式是通终端输入过命令行 `simon` 启动的。该交互模式还提供了一个状态显示板，其中概述了所有模拟的实时状态（参见图 4.5）。通过该状态板，用户可以选择和操纵模拟的运行方式（参见图 4.2）。

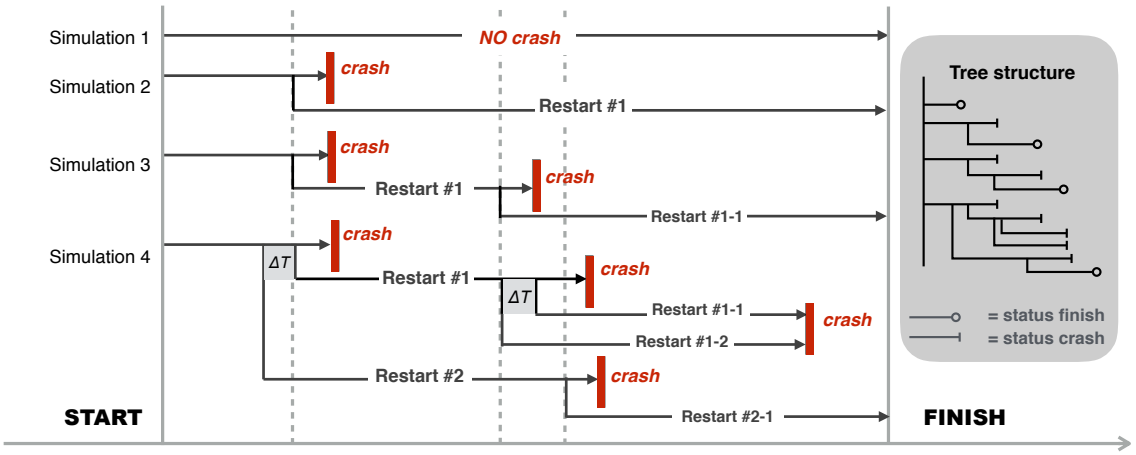

图 4.3: SiMon 使用层次拓扑来索引一个模拟集合。起始状态下，每个模拟程序都是根节点下的叶节点。随着模拟程序的运行，其中一些可能会崩溃并随后重新启动。重新启动的程序被认为是原始模拟的子节点。因此，原始模拟的状态是通过传播重新启动的信息来获得的。控制原始模拟本质上是控制子节点中的重启模拟。以这种方式，模拟树将动态生长，直到所有模拟完成。在这个例子中，我们管理了四个模拟程序的集合。模拟程序 1 不中断完成：不需要重新启动；模拟程序 2 在某一时刻暂停，但是经过一次重启后 `Restart #1` 得以完成；模拟程序 3 在第一次重启后 (`Restart #1`) 继续崩溃并重启了一次（图中的 `Restart #1-1`），才得以完成；模拟程序 4 是最复杂的一种情况，*回滚重启*在这种情况下被多次使用使得程序能够完成。在第一次中断后，程序第一次重启 `Restart #1` 并很快再次崩溃，其后的重启 `Restart #1-1` 并没有回滚到合适的重启点来解决数值计算问题，因此随后另一个重启被 SiMon 工具启动。这个随后的重启回滚到更前一次“快照”来避免将要遇到的数值计算问题，这个新的重启 `Restart #1-2` 遇到了另一个数值问题并再次崩溃。自此所有从 `Restart #1` 存储的“快照”都已被用来重启程序，但是它们都没有成功的让程序完成。相关 `Restart #1` 的文件因此被定义为不可重启，SiMon 从 `Restart #1` 回滚 ΔT 段时间来回到更上一级程序崩溃的时间点，再经历了又一次崩溃 `Restart #2-1` 后，程序得以完成到最终 DONE 状态。图右侧的树结构给出了整个重启过程的简略版。

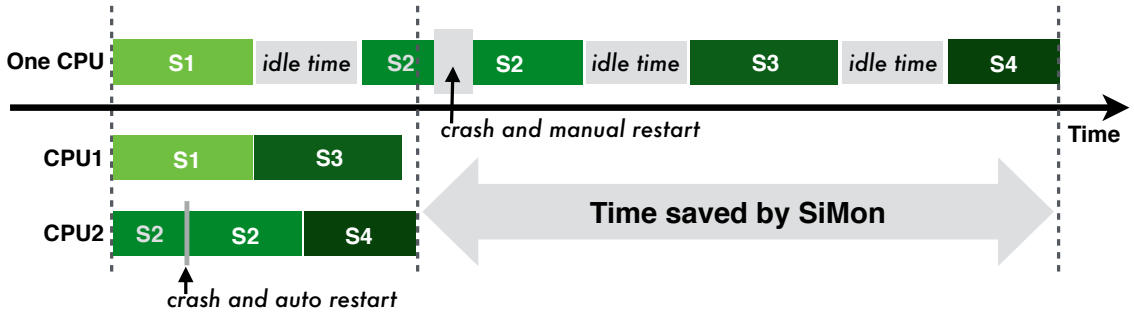

图 4.4: Simon 的优先级作业调度方案与人工手动顺序模拟管理的比较。S1，S2，S3 和 S4 是四个模拟。需要完成它们的时间由它们的长度表示; 工作优先级用颜色表示（较浅的颜色具有较高的优先级）。这四个模拟是在双 CPU 机器上启动的。一开始，CPU1 和 CPU2 都是空闲的，所以 S1 和 S2 分别由于它们的高优先级而被安排在它们上。其后 S2 经历中断，但已经被 SiMon 自动重新启动。当 S1 完成后，CPU1 变为空闲，所以 SiMon 立即启动 S3。S3 启动后不久，S2 完成，CPU2 变为空闲，随后 S4 启动。该图显示出 SiMon 的自动化调度方案大大减少了在多处理器机器上运行多次模拟程序的总时间。

4.2.3.2 其它

用户界面: SiMon 的后台进程模式没有用户界面，因为它不需要人为的监督。SiMon 的交互模式具有一个文本用户界面（TUI）。这样可以确保天文学家可以使用它，而不需要额外的渲染支持（例如，图形用户界面，GUI），这也与天文学家常用的通过安全 Shell（SSH）登录到计算节点的情况相似。

工具扩展能力: SiMon 允许用户管理任意大量的数字代码，每个数字代码可以有明显的不同，包括输入或后处理要求。为了概况所有可能的操作，我们在表 4.1和表 4.3列出了最常见的操作。我们还实现了一个常用的模块去支持任意普遍的数值模拟环境，这通过 `module_common.py` 模块实现。这里我们假设模拟程序可以通过 UNIX 命令行或者设置文件来控制。

代码管理: 我们通过 GitHub 软件平台[3]对代码和文档进行管理，同时，我们将工具发布到 PyPI 平台，供用户使用命令行 `pip install astrosimon` 来快速安装稳定版本。

[3] https://github.com/maxwelltsai/SiMon

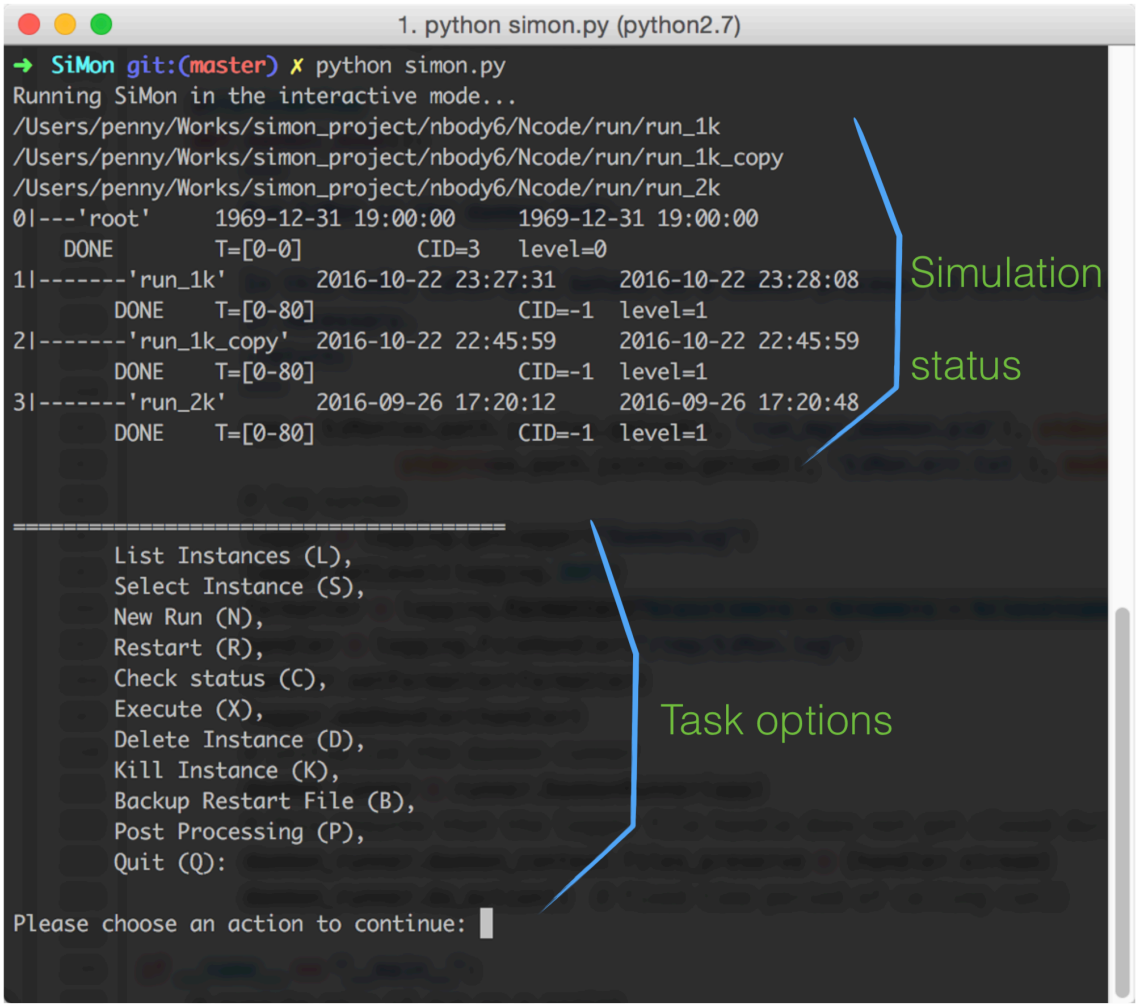

图 4.5: SiMon 的交互式状态板，为用户提供了所有模拟的当前状态的概述，并提供手动控制这些模拟的操作符。每个模拟，包括重新启动的模拟，都被分配一个唯一的 ID，允许用户选择一个或多个模拟，并对它们应用管理操作。有关可能的管理操作符的清单，请参阅表 4.2。

表 4.1: 数值模拟任务的属性列表。属性列表的顺序可能与实际的数字代码不同，可以通过 Python 的 `dict` 数据结构进行扩展。

Category	**Properties**	**Example**
Paths	Configuration file, input files, output directory	`/path/to/data/dir`
Type	The numerical code used to carry out the simulation	`NBODY6`, and all codes supported by `AMUSE`[4]
Model	Start/Termination criteria, current model time	$t = 5, t_{\rm start} = 0, t_{\rm end} = 10$
Process	Process ID, process launch timestamp	`PID=12345`
Commands	Commands to start/restart/stop a simulation	`./simulation_code`
Relation	IDs of the parental simulations and sub-simulations	`sim_id=2, parent=1, children=[5,6,7]`
Status	Current status of the simulation	`RUN/STALL/STOP/DONE/ERROR`

表 4.2: 交互模式下支持的人工操作

Task name	Description
List Simulations	Generate a status overview of all managed simulations
Select Simulations	Select multiple simulation and execute command in batch
New Run	Start new simulations from beginning point
Restart	Restart the simulation from crashing point
Check status	Check the recent or current calculation results and print it
UNIX shell	Execute an UNIX shell command in the simulation directory
Stop Simulations	Send a stop request to the simulation code
Delete Simulations	Delete the simulation instance and all its substance
Kill Simulations	Kill the UNIX process associate with a simulation task
Backup Restart File	Backup simulation checkpoint files (for future restarting purpose)
Post Processing	Perform (post)-processing (usually) after the simulation is done
Quit	Quit the SiMon interactive mode

表 4.3: 用于控制任意模拟程序的通用方法列表。SiMon 工具包实现了所有这些方法，这些方法的实际功能可由配置文件定义（使用 shell 命令且并不需要 Python 编程）或自定义代码特定的模块（需要 Python 编程）。

Method Name	**Description**
`sim_init()`	Perform necessary initialization procedures to start the simulation.
`sim_start()`	Start the simulation
`sim_restart()`	Restart the simulation
`sim_get_status()`	Get the current status of the simulation
`sim_stop()`	Stop the simulation using the mechanism provided by the code
`sim_kill()`	Kill the simulation process forcibly (when the code stalls)
`sim_backup_checkpoint()`	Backup the restart files
`sim_delete()`	Delete the simulation data
`sim_clean()`	Clean up the simulation data except for the input files and restart files
`sim_reset()`	Reset the simulation, leaving only the input files
`sim_shell_exec()`	Execute a UNIX shell command on the simulation data directory
`sim_finalize()`	Finalize the simulation (e.g. perform data processing) after finished.

4.3 SiMon 总结与展望

在天文研究中，经常需要运行多个计算作业，以实现对一批观测数据的数据加工或者对一个物理参数空间的数值模拟。这些计算作业不仅计算量大、耗时长，而且很可能会由于物理参数不正确、代码出错或硬件问题而中断。为了完成这些计算作业，天文学家需要反复登陆计算集群，逐一监视这些作业的运行情况。对于出错的作业，还必须通过日志文件分析的原因，并做出相应重启的操作。这些工作不仅耗时繁琐，而且也容易出现人为错误。FAST 参数的数据量巨大，而这些原始数据必须实时处理归档，通过人工来完成这些工作是非常困难的。

目前常见的作业调度软件种类繁多，并且有不少免费开源的软件可供使用，例如 Slurm 和 OpenLava 等。然而，这些软件实质上只是实现了作业按优先级的排队功能，并不具备作业监视的功能。当一个作业中断时，天文学家依然需要手动重启。换而言之，这些作业没有具备自动化天文工作流的作用。

为了满足 FAST 天文数据流水线调度以及天体物理数值模拟的需求，我们开发了 Simulation Monitor（SiMon）工具。该工具不仅具备了普通作业调度软件所实现的作业按优先级排队的功能，而且还能监视多个并发作业的运行情况。对于被中断的作业，SiMon 将自动分析其日志文件，试图找出问题所在，然后自动调整参数并重启作业。如果多次的自动重启均无法完成某一特点的作业，SiMon 将该作业标记为 “ERROR”，从而提醒天文学家人工处理。另外，天文学家不再需要手动去运行这些作业。通过 SiMon，天文学家只需指定一个参数空间，SiMon 就可以根据当前系统的计算资源智能调度这些作业，使得总耗时最少。毫无疑问，这就大大减轻了天文学家的工作量，从而允许天文学家将时间用在对天文数据的探索和解读上。

第五章　总结与展望

5.1　总结

500 米口径球面射电望远镜（FAST）是国家重大科学工程。该工程于 2016 年竣工，现在已经进入了全面的调试和优化阶段。FAST 的科学目标包括了大规模中性氢巡天、脉冲星搜寻、星际分子探测、观测脉冲星计时阵列（Pulsar Timing Array)、搜寻地外智能生命等。作为射电天文的战略性巡天设备，FAST 预计产生 60PB 的数据，作为对比，当前世界上最快的超级计算机 “天河二号” 的存储容量为 2PB。因此，如何存储和处理这些数据，将其高效准确的转化为科学产品，无疑对于其科学目标的成败具有决定作用。

本论文提出了当代天文时代下 FAST 数据处理系统的构想和实现方案。FAST 的数据处理系统并不仅限于传统意义上的 Data Reduction Pipeline(DRP)，其应该包括数据收集、归档、挖掘和理解的所有现代科学研究工作模块。数据收集和归档部分主要由望远镜控制和 DRP 完成，在这方面，FAST 工程可以很大程度借鉴前人的工作并加以开发利用，例如 ALFALFA 中性氢 DRP。而天文数据挖掘和理解则是目前需要发展和更多投入的两个部分，随着天文数据量的喷井式增长，新的分析手段、展示方法和管理技术需要被用于从海量数据中获得有用信息并用其回答科学问题。本论文结合天文学研究与其它领域研究，围绕新的技术方法进行了一系列研究与开发，包括可视化展示分析、立体显影和虚拟现实以及自动化批量任务管理等，具体如下：

通过 Glue 平台实现了三维视图互联，为天文高维度数据的可视化分析提供了一种新的有效的手段。不仅实现了多样化高效率的三维展示功能，而且创新性的实现了基于不同三维展示的多样化三维选取，这在天文领域以及可视化分析领域都是最新的。

通过大量调研与实践设计了可靠且新颖的 FAST 可视化分析模块并实现了其部分功能。该可视化模块不仅具备强大的从一维到三维可视化功能，还能够顺应大数据时代的需求，通过地图册搜索等方式快速获取目标远程数据，并通过云端计算资源完成计算复杂性任务。通过该模块的高度可扩展化，第三方工具将通过内嵌或外接两种方式与本模块进行数据与功能的共享，从而大大增强了该模块的数据分析功能。该可视化分析模块不仅是 FAST 工程的直接需求，切实可行的服务于 FAST 数据处理，更是天文大数据跨领域合作的一大创新，将可视化、云计算、软件开发与天文研究紧密的结合在一起。

对三维可视化相关领域的研究进行了总结，提出并实现了新颖的 FAST 高维度数据可视化方案，例如使用三维建模软件 Blender 进行天文数据三维可视化的

尝试，包括对数据立方体和数值模拟数据的实时展示与渲染；例如使用立体显像技术，通过叠加深度信息的方法来高维度天文数据，使观察者通过显像设备就能简单的体验到三维；例如使用最新的混合现实技术将虚拟场景中渲染的物体真实展现在三维场景中供用户交互式观察。这些新的技术实现都是对 FAST 产品展示及天文数据可视化的一大补充。

针对 FAST 的数据量巨大、原始数据必须得到实时处理的挑战，我们开发了 Simulation Monitor（SiMon）工具，可以用于自动并发调度 FAST 的数据处理流水线。该工具不仅具备了普通作业调度软件所实现的作业按优先级排队的功能，而且还能监视多个并发作业的运行情况。对于被中断的作业，SiMon 将自动分析其日志文件，试图找出问题所在，然后自动调整参数并重启作业。如果多次的自动重启均无法完成某一特点的作业，SiMon 将该作业标记为 “ERROR”，从而提醒天文学家人工处理。另外，天文学家不再需要手动去运行这些作业。通过 SiMon，天文学家只需指定一个参数空间，SiMon 就可以根据当前系统的计算资源智能调度这些作业，使得总耗时最少。这就大大减轻了 FAST 天文学家的工作量，从而使得他们可以集中精力在天文数据的探索和研究中。

5.2 展望

FAST 将不可避免地产生巨大规模的数据。当天文学家对这些数据进行分析的时候，传统的做法是将它们逐一地下载到他们的本地计算机上。显然，这个做法对于 FAST 数据分析是不现实的。即便天文学家有足够的带宽和时间下载数据，也几乎不可能有足够的存储空间，更何况下载所需要的时间将极大地拖延了天文研究的进展。事实上，对大数据进行迁移或下载是不现实、不理智的。唯一有效的解决方案是对数据进行在线处理。这需要依赖当前先进且快速发展的云计算／云存储技术。进一步的，考虑到天文学家需要要不同的科学软件完成不同的数据处理工作，我们需要实现这些软件之间的互操作性。这需要借助虚拟天文台技术和无缝天文的理念来完成。

云计算是当前大数据时代的产物。其应用不仅解决了大数据存储和检索的诸多挑战，而且也极大地方便了我们的日常生活。我们几乎每个人都在不知不觉地使用云计算技术。例如，当我们用手机拍照时，所拍的照片可能自动同步到我们的电脑上。当我们开车使用地图导航时，许多的地图应用可以实时地显示当前的路况信息，我们从而可以合理的规划路线，避免交通拥堵。设想这些技术用在 FAST 项目中：位于贵州的 FAST 团队发现一个有价值的天体目标，并对数据进行初步处理，然后通知位于北京的 FAST 团队。北京团队收到了信息之后打开了他们的数据处理软件，相关的数据就自动下载并显示处理，于是他们可以立即对数据进行进一步分析。这种工作流无疑对天文学家产生了极大的便利：没有地域和时间的

限制，随时随地处理天文数据。许多的天文学研究具有实时的要求，例如当 LIGO 或 VIRGO 探测到双中子星并合产生的引力波时，可能会需要通知位于智利的望远镜在一分钟之内做出反应，及时观测该双中子星系统产生的电磁辐射[130]。这种情况下，云计算将对于科学成果的产出产生了决定性作用。

云计算和云存储技术屏蔽了背后的技术复杂性。例如，当我们在使用搜索引擎的时候，我们从来都不需要考虑数据存储在哪个服务器上，需要用哪个 CPU 来对数据进行处理。云平台会自动选择最优化的处理方式并实时返回相关结构。同理，云计算在 FAST 项目的使用，将使得天文学家无限考虑如何获得数据，获得什么数据。只要给定一个天区范围，相关的数据将按需下载并显示到天文学家的计算机上。这样，天文学家就能专注于天文数据处理的的工作上。

云计算在天文学上已经得到了初步的应用。国际虚拟天文台联盟 (IVOA) 开发了一套完整的虚拟天文台网络协议，如 Table Access Protocol (TAP), Simple Image Access Protocol (SIAP), Simple Spectra Access Protocol (SSAP) 等。这些协议已经在天文软件如 Aladin、TOPCAT 中实现。如图 5.1所示，用户可以在 TOPCAT 软件中选择使用 Gaia 数据，然后指定目标天体名称或天区范围，然后数据将自动从云端按需下载到用户的计算机中。

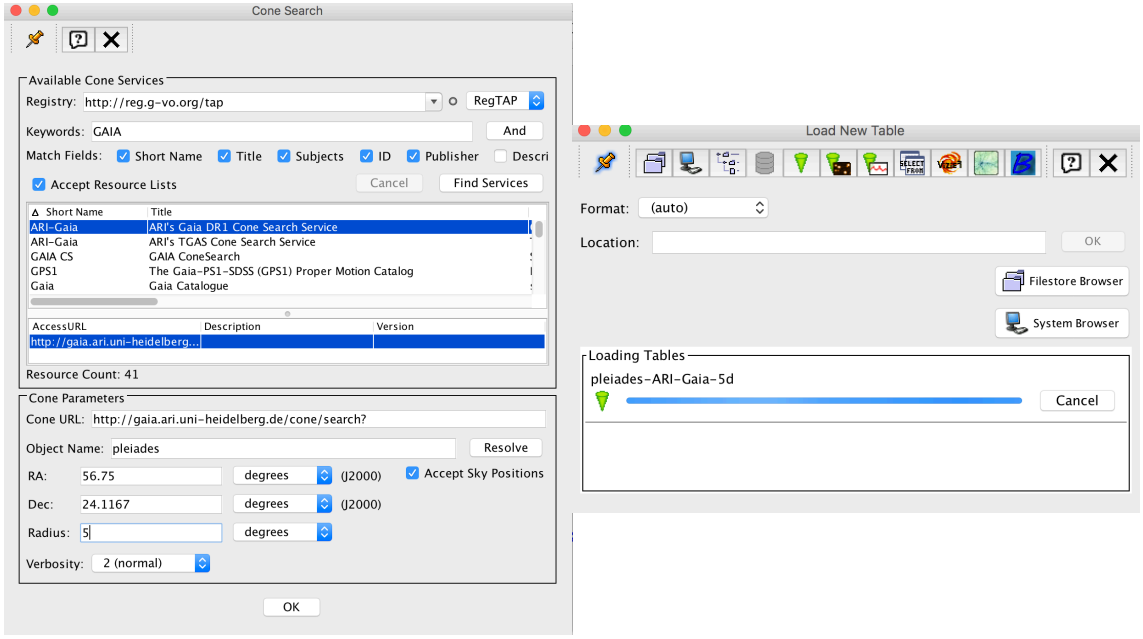

图 5.1: 该图显示了使用 Top 软件自动下载 Gaia 数据的图形界面。左图展示了用户指定 Gaia 数据源并注册其为兴趣区域，在点击 OK 后，关联数据会被自动下载如右图所示。

我们希望搭建 FAST 的云平台，结合云计算能力和数据处理分析能力的 FAST 数据处理系统将为 FAST 科学研究提供更广阔的空间，并帮助天文学家透过海量复杂的数据洞察宇宙的奥秘。将云技术用到了 FAST 项目中，天文学家就不再需

要关心数据的具体存储位置和处理数据具体用到哪个服务器等。所有的这些技术细节将被云平台自动优化处理。而为了方便天文学家使用不同的天文数据处理软件，本文提出了将虚拟天文台技术和云计算结合的观点：不同的天文软件之间的数据互通是基于虚拟天文台的协议实现的，而云平台则提供了数据源。这样，当天文学家 A 需要研究某个天区时，只需在软件 A 中指定相应天区的坐标，点击确定，相关的数据就会自动下载到其计算机中并显示处理。当其完成数据处理后，处理过的数据将自动上传到云平台。此时天文学家 B 可在平台上看到数据已经得到更新。当其需要用另一天文软件 B 进行处理时，因为天文软件 B 也遵循相应的虚拟天文台协议，因此可用无缝接受到来自 A 的数据，这一架构将实现的无缝天文的理念。

参考文献

[1] NAN R, LI D, JIN C, et al. The Five-Hundred-Meter Aperture Spherical Radio Telescope (FAST) Project[J]. arXiv.org, 2011(06): 989–1024.

[2] GOODMAN A A. Principles of High-Dimensional Data Visualization in Astronomy[J]. arXiv.org, 2012(5): 505–.

[3] Ellsworth-Bowers T P, Rosolowsky E, Glenn J, et al. The Bolocam Galactic Plane Survey. XII. Distance Catalog Expansion Using Kinematic Isolation of Dense Molecular Cloud Structures with ^{13}CO(1-0)[J]. ApJ, 2015, 799: 29. DOI: 10.1088/0004-637X/799/1/29.

[4] Furlan E, Fischer W J, Ali B, et al. The Herschel Orion Protostar Survey: Spectral Energy Distributions and Fits Using a Grid of Protostellar Models [J]. ApJS, 2016, 224: 5. DOI: 10.3847/0067-0049/224/1/5.

[5] BRUNNER R J, DJORGOVSKI S G, PRINCE T A, et al. Massive datasets in astronomy[J]. Handbook of massive data sets, 2002.

[6] GALLAGHER R S. Computer visualization: Graphics techniques for scientific and engineering analysis[M]. 1st ed. Boca Raton, FL, USA: CRC Press, Inc., 1994.

[7] HASSAN A, FLUKE C J. Scientific Visualization in Astronomy: Towards the Petascale Astronomy Era[J]. arXiv.org, 2011(02): 150–170.

[8] PAUL MILGRAM A U F K, Haruo Takemura. Augmented reality: a class of displays on the reality-virtuality continuum[J/OL]. Proc.SPIE, 1995, 2351: 2351 – 2351 – 11. http://dx.doi.org/10.1117/12.197321. DOI: 10.1117/12.197321.

[9] ARCE H G, BORKIN M A, GOODMAN A A, et al. THE COMPLETE SURVEY OF OUTFLOWS IN PERSEUS[J]. The Astrophysical Journal, 2010, 715 (2): 1170–1190.

[10] JIN C J, NAN R D, GAN H Q. The fast telescope and its possible contribution to high precision astrometry[J]. Proceedings of the International Astronomical Union, 2007, 3(S248): 178–181. DOI: 10.1017/S1743921308018978.

[11] LI D, NAN R, PAN Z. The Five-hundred-meter Aperture Spherical Radio Telescope Project and its Early Science Opportunities[J]. arXiv.org, 2012 (S291): 325–330.

[12] Li D, Goldsmith P F. H I Narrow Self-Absorption in Dark Clouds[J]. ApJ, 2003, 585: 823–839. DOI: 10.1086/346227.

[13] Chang T C, Pen U L, Bandura K, et al. An intensity map of hydrogen 21-cm emission at redshift z˜0.8[J]. Nature, 2010, 466: 463–465. DOI: 10.1038/nature09187.

[14] Smits R, Lorimer D R, Kramer M, et al. Pulsar science with the Five hundred metre Aperture Spherical Telescope[J]. A&A, 2009, 505: 919–926. DOI: 10.1051/0004-6361/200911939.

[15] Kalberla P M W, Kerp J. The Hi Distribution of the Milky Way[J]. ARA&A, 2009, 47: 27–61. DOI: 10.1146/annurev-astro-082708-101823.

[16] Djorgovski S G, Williams R. Virtual Observatory: From Concept to Implementation[C]//Kassim N, Perez M, Junor W, et al. Astronomical Society of the Pacific Conference Series: volume 345 From Clark Lake to the Long Wavelength Array: Bill Erickson's Radio Science. [S.l.: s.n.], 2005: 517.

[17] Quinn P J, Barnes D G, Csabai I, et al. The International Virtual Observatory Alliance: recent technical developments and the road ahead[C]//Quinn P J, Bridger A. Proc. SPIE: volume 5493 Optimizing Scientific Return for Astronomy through Information Technologies. 2004: 137–145. DOI: 10.1117/12.551247.

[18] GIOVANELLI R, HAYNES M P, KENT B R. The Arecibo Legacy Fast ALFA Survey: I. Science Goals, Survey Design and Strategy[J]. arXiv.org, 2005(6): 2598–2612.

[19] Hoppmann L, Staveley-Smith L, Freudling W, et al. A blind H I mass function from the Arecibo Ultra-Deep Survey (AUDS)[J]. MNRAS, 2015, 452: 3726–3741. DOI: 10.1093/mnras/stv1084.

[20] Astropy Collaboration, Robitaille T P, Tollerud E J, et al. Astropy: A community Python package for astronomy[J]. A&A, 2013, 558: A33. DOI: 10.1051/0004-6361/201322068.

[21] Wells D C, Greisen E W, Harten R H. FITS - a Flexible Image Transport System[J]. A&AS, 1981, 44: 363.

[22] Hanisch R J, Farris A, Greisen E W, et al. Definition of the Flexible Image Transport System (FITS)[J]. A&A, 2001, 376: 359–380. DOI: 10.1051/0004-6361:20010923.

[23] MCCORMICK B H. Visualization in scientific computing[J/OL]. SIGBIO Newsl., 1988, 10(1): 15–21. http://doi.acm.org/10.1145/43965.43966. DOI: 10.1145/43965.43966.

[24] FRENKEL K A. The art and science of visualizing data[J/OL]. Commun.

ACM, 1988, 31(2): 111–121. http://doi.acm.org/10.1145/42372.42373. DOI: 10.1145/42372.42373.

[25] Snell R L, Loren R B, Plambeck R L. Observations of CO in L1551 - Evidence for stellar wind driven shocks[J]. ApJ, 1980, 239: L17–L22. DOI: 10.1086/183283.

[26] Arce H G, Borkin M A, Goodman A A, et al. A Bubbling Nearby Molecular Cloud: COMPLETE Shells in Perseus[J]. ApJ, 2011, 742: 105. DOI: 10.1088/0004-637X/742/2/105.

[27] Norris R P. The Challenge of Astronomical Visualisation[C]//Crabtree D R, Hanisch R J, Barnes J. Astronomical Society of the Pacific Conference Series: volume 61 Astronomical Data Analysis Software and Systems III. [S.l.: s.n.], 1994: 51.

[28] OOSTERLOO T. Visualisation of radio data[J]. Publications of the Astronomical Society of Australia, 1995, 12(2): 215 – 218. DOI: 10.1017/S1323358000020294.

[29] ROTH S D. Ray casting for modeling solids[J/OL]. Computer Graphics and Image Processing, 1982, 18(2): 109 – 144. http://www.sciencedirect.com/science/article/pii/0146664X82901691. DOI: https://doi.org/10.1016/0146-664X(82)90169-1.

[30] AMATI G, GIANLUCA D, DI RICO G, et al. Astromd: a 3d visualization and analysis tool for astrophysical data[M]. [S.l.: s.n.], 2011.

[31] AHRENS J, HEITMANN K, HABIB S, et al. Quantitative and comparative visualization applied to cosmological simulations[J/OL]. Journal of Physics: Conference Series, 2006, 46(1): 526. http://stacks.iop.org/1742-6596/46/i=1/a=073.

[32] LI H, FU C W, HANSON A. Visualizing multiwavelength astrophysical data [J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1555–1562. DOI: 10.1109/TVCG.2008.182.

[33] Mickaelian A M. Astronomical surveys and big data[J]. Baltic Astronomy, 2016, 25: 75–88.

[34] FEIGELSON E D, BABU G J. Big data in astronomy[J/OL]. Significance, 2012, 9(4): 22–25. http://dx.doi.org/10.1111/j.1740-9713.2012.00587.x. DOI: 10.1111/j.1740-9713.2012.00587.x.

[35] ZHANG Y, ZHAO Y. Astronomy in the big data era[J]. Data Science Journal, 2015(14): 11. DOI: http://doi.org/10.5334/dsj-2015-011.

[36] Beaumont C N, Goodman A A, Kendrew S, et al. The Milky Way Project:

Leveraging Citizen Science and Machine Learning to Detect Interstellar Bubbles[J]. ApJS, 2014, 214: 3. DOI: 10.1088/0067-0049/214/1/3.

[37] WILLS G. Linked data views[M/OL]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 217–241. https://doi.org/10.1007/978-3-540-33037-0_10. DOI: 10.1007/978-3-540-33037-0_10.

[38] GRESH T, LICHATOWICH J, SCHOONMAKER P. An estimation of historic and current levels of salmon production in the northeast pacific ecosystem: Evidence of a nutrient deficit in the freshwater systems of the pacific northwest[J/OL]. Fisheries, 2000, 25(1): 15–21. http://dx.doi.org/10.1577/1548-8446(2000)025<0015:AEOHAC>2.0.CO;2. DOI: 10.1577/1548-8446(2000)025<0015:AEOHAC>2.0.CO;2.

[39] TUKEY J. Addison-wesley series in behavioral science: Exploratory data analysis[M/OL]. Addison-Wesley Publishing Company, 1977. https://books.google.nl/books?id=UT9dAAAAIAAJ.

[40] WONG P C, BERGERON R D. Multivariate visualization using metric scaling [C/OL]//VIS '97: Proceedings of the 8th Conference on Visualization '97. Los Alamitos, CA, USA: IEEE Computer Society Press, 1997: 111–ff. http://dl.acm.org/citation.cfm?id=266989.267036.

[41] Beaumont C, Goodman A, Greenfield P. Hackable User Interfaces In Astronomy with Glue[C]//Taylor A R, Rosolowsky E. Astronomical Society of the Pacific Conference Series: volume 495 Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV). [S.l.: s.n.], 2015: 101.

[42] Beaumont C, Robitaille T, Borkin M. Glue: Linked data visualizations across multiple files[M]. [S.l.: s.n.], 2014.

[43] AMAR R, EAGAN J, STASKO J. Low-level components of analytic activity in information visualization[C/OL]//INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization. Washington, DC, USA: IEEE Computer Society, 2005: 15–. https://doi.org/10.1109/INFOVIS.2005.24. DOI: 10.1109/INFOVIS.2005.24.

[44] HASSAN A H, FLUKE C J, BARNES D G. Interactive Visualization of the Largest Radioastronomy Cubes[J]. arXiv.org, 2010(2): 100–109.

[45] CAMPAGNOLA L, KLEIN A, LARSON E, et al. VisPy: Harnessing The GPU For Fast, High-Level Visualization[C/OL]//HUFF K, BERGSTRA J. Proceedings of the 14th Python in Science Conference. Austin, Texas, United States, 2015. https://hal.inria.fr/hal-01208191.

[46] GOODMAN A A, ALVES J, BEAUMONT C N, et al. THE BONES OF THE MILKY WAY[J]. The Astrophysical Journal, 2014, 797(1): 53–13.

[47] ZUCKER C, BATTERSBY C, GOODMAN A. The Skeleton of the Milky Way[J]. arXiv.org, 2015(1): 23.

[48] MUNZNER T. Ak peters visualization series: Visualization analysis and design[M/OL]. CRC Press, 2014. https://books.google.com/books?id=dznSBQAAQBAJ.

[49] Ridge N A, Di Francesco J, Kirk H, et al. The COMPLETE Survey of Star-Forming Regions: Phase I Data[J]. AJ, 2006, 131: 2921–2933. DOI: 10.1086/503704.

[50] Rosolowsky E W, Pineda J E, Kauffmann J, et al. Structural Analysis of Molecular Clouds: Dendrograms[J]. ApJ, 2008, 679: 1338–1351. DOI: 10.1086/587685.

[51] BISHOP C M. Pattern recognition and machine learning (information science and statistics)[M]. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[52] SANGYOON LEE G J K C M P, Jinseok Seo. Evaluation of pointing techniques for ray casting selection in virtual environments[J/OL]. Proc.SPIE, 2003, 4756: 4756 – 4756 – 7. http://dx.doi.org/10.1117/12.497665. DOI: 10.1117/12.497665.

[53] ARGELAGUET F, ANDUJAR C. Efficient 3d pointing selection in cluttered virtual environments[J]. IEEE Computer Graphics and Applications, 2009, 29 (6): 34–43. DOI: 10.1109/MCG.2009.117.

[54] ALTMAN N S. An introduction to kernel and nearest-neighbor nonparametric regression[J/OL]. The American Statistician, 1992, 46(3): 175–185. http://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879. DOI: 10.1080/00031305.1992.10475879.

[55] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in python[J/OL]. J. Mach. Learn. Res., 2011, 12: 2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195.

[56] Benjamin R A, Churchwell E, Babler B L, et al. GLIMPSE. I. An SIRTF Legacy Project to Map the Inner Galaxy[J]. PASP, 2003, 115: 953–964. DOI: 10.1086/376696.

[57] Molinari S, Swinyard B, Bally J, et al. Hi-GAL: The Herschel Infrared Galactic Plane Survey[J]. PASP, 2010, 122: 314. DOI: 10.1086/651314.

[58] Abbott B P, Abbott R, Abbott T D, et al. Observation of Gravitational Waves

from a Binary Black Hole Merger[J]. Physical Review Letters, 2016, 116(6): 061102. DOI: 10.1103/PhysRevLett.116.061102.

[59] Spurzem R. Direct N-body Simulations[J]. Journal of Computational and Applied Mathematics, 1999, 109: 407–432.

[60] Aarseth S J. Gravitational N-Body Simulations[M]. [S.l.: s.n.], 2003: 430.

[61] Wang L, Spurzem R, Aarseth S, et al. NBODY6++GPU: ready for the gravitational million-body problem[J]. MNRAS, 2015, 450: 4070–4080. DOI: 10.1093/mnras/stv817.

[62] Cai M X, Meiron Y, Kouwenhoven M B N, et al. Block Time Step Storage Scheme for Astrophysical N-body Simulations[J]. ApJS, 2015, 219: 31. DOI: 10.1088/0067-0049/219/2/31.

[63] Sancisi R, Fraternali F, Oosterloo T, et al. Cold gas accretion in galaxies[J]. A&A Rev., 2008, 15: 189–223. DOI: 10.1007/s00159-008-0010-0.

[64] Boomsma R, Oosterloo T A, Fraternali F, et al. HI holes and high-velocity clouds in the spiral galaxy NGC 6946[J]. A&A, 2008, 490: 555–570. DOI: 10.1051/0004-6361:200810120.

[65] PUNZO D, VAN DER HULST J M, ROERDINK J B T M, et al. The role of 3-D interactive visualization in blind surveys of H I in galaxies[J]. Astronomy and Computing, 2015, 12: 86–99.

[66] HUNTER J D. Matplotlib: A 2d graphics environment[J]. Computing in Science Engineering, 2007, 9(3): 90–95. DOI: 10.1109/MCSE.2007.55.

[67] Gaudet S, Dowler P, Goliath S, et al. The Canadian Advanced Network For Astronomical Research[C]//Bohlender D A, Durand D, Dowler P. Astronomical Society of the Pacific Conference Series: volume 411 Astronomical Data Analysis Software and Systems XVIII. [S.l.: s.n.], 2009: 185.

[68] Bonnarel F, Fernique P, Bienaymé O, et al. The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources[J]. A&AS, 2000, 143: 33–40. DOI: 10.1051/aas:2000331.

[69] Boch T, Fernique P. Aladin Lite: Embed your Sky in the Browser[C]//Manset N, Forshay P. Astronomical Society of the Pacific Conference Series: volume 485 Astronomical Data Analysis Software and Systems XXIII. [S.l.: s.n.], 2014: 277.

[70] Greisen E W, Calabretta M R. Representations of world coordinates in FITS [J]. A&A, 2002, 395: 1061–1075. DOI: 10.1051/0004-6361:20021326.

[71] Calabretta M R, Greisen E W. Representations of celestial coordinates in FITS[J]. A&A, 2002, 395: 1077–1122. DOI: 10.1051/0004-6361:20021327.

[72] Greisen E W, Calabretta M R, Valdes F G, et al. Representations of spectral coordinates in FITS[J]. A&A, 2006, 446: 747–771. DOI: 10.1051/0004-6361: 20053818.

[73] Ochsenbein F, Bauer P, Marcout J. The VizieR database of astronomical catalogues[J]. A&AS, 2000, 143: 23–32. DOI: 10.1051/aas:2000169.

[74] Taylor M B, Boch T, Taylor J. SAMP, the Simple Application Messaging Protocol: Letting applications talk to each other[J]. Astronomy and Computing, 2015, 11: 81–90. DOI: 10.1016/j.ascom.2014.12.007.

[75] Goodman A, Fay J, Muench A, et al. WorldWide Telescope in Research and Education[C]//Ballester P, Egret D, Lorente N P F. Astronomical Society of the Pacific Conference Series: volume 461 Astronomical Data Analysis Software and Systems XXI. [S.l.: s.n.], 2012: 267.

[76] GOODMAN A, FAY J, MUENCH A, et al. WorldWide Telescope in Research and Education[J]. arXiv.org, 2012.

[77] Joye W A, Mandel E. New Features of SAOImage DS9[C]//Payne H E, Jedrzejewski R I, Hook R N. Astronomical Society of the Pacific Conference Series: volume 295 Astronomical Data Analysis Software and Systems XII. [S.l.: s.n.], 2003: 489.

[78] Joye W A. New Features of SAOImage DS9[C]//Gabriel C, Arviset C, Ponz D, et al. Astronomical Society of the Pacific Conference Series: volume 351 Astronomical Data Analysis Software and Systems XV. [S.l.: s.n.], 2006: 574.

[79] Taylor M B. TOPCAT STIL: Starlink Table/VOTable Processing Software [C]//Shopbell P, Britton M, Ebert R. Astronomical Society of the Pacific Conference Series: volume 347 Astronomical Data Analysis Software and Systems XIV. [S.l.: s.n.], 2005: 29.

[80] TAYLOR M. TOPCAT: Desktop Exploration of Tabular Data for Astronomy and Beyond[J]. Informatics, 2017, 4(3): 18–18.

[81] SCIACCA E, VITELLO F, BECCIANI U, et al. Vialactea science gateway for milky way analysis[J/OL]. Future Generation Computer Systems, 2017. http://www.sciencedirect.com/science/article/pii/S0167739X17309561. DOI: http://dx.doi.org/10.1016/j.future.2017.08.038.

[82] ROOSENDAAL T, SELLERI S. The official blender 2.3 guide: Free 3d creation suite for modeling, animation, and rendering[M]. San Francisco, CA, USA: No Starch Press, 2004.

[83] Naiman J P. AstroBlend: An astrophysical visualization package for Blender

[J]. Astronomy and Computing, 2016, 15: 50–60. DOI: 10.1016/j.ascom.2016.02.002.

[84] Cui C, He B, Yu C, et al. AstroCloud: A Distributed Cloud Computing and Application Platform for Astronomy[J]. ArXiv e-prints, 2017.

[85] Li D, Goldsmith P, Dickey J. Atomic Hydrogen in Molecular Clouds[C]//IAU Symposium: volume 221 IAU Symposium. [S.l.: s.n.], 2003.

[86] PILATO M, COLLINS-SUSSMAN B, FITZPATRICK B. O'reilly series: Version control with subversion[M/OL]. O'Reilly Media, 2008. https://books.google.com/books?id=EW-3srTM84sC.

[87] CHACON S. Pro git[M]. 1st ed. Berkely, CA, USA: Apress, 2009.

[88] GOODMAN A A, UDOMPRASERT P S, KENT B, et al. Astronomy Visualization for Education and Outreach[C]//Astronomical Data Analysis Software and Systems XX. ASP Conference Proceedings. [S.l.: s.n.], 2011: 659–.

[89] KENT B R. The GRIDView Visualization Package[C]//Astronomical Data Analysis Software and Systems XX. ASP Conference Proceedings. [S.l.: s.n.], 2011: 625–.

[90] Taylor R. FRELLED: A realtime volumetric data viewer for astronomers[J]. Astronomy and Computing, 2015, 13: 67–79. DOI: 10.1016/j.ascom.2015.10.002.

[91] PUNZO D, VAN DER HULST J M, ROERDINK J B T M, et al. SlicerAstro: A 3-D interactive visual analytics tool for HI data[J]. Astronomy and Computing, 2017, 19: 45–59.

[92] Brown R L, Wild W, Cunningham C. ALMA - the Atacama large millimeter array[J]. Advances in Space Research, 2004, 34: 555–559. DOI: 10.1016/j.asr.2003.03.028.

[93] Johnston S, Bailes M, Bartel N, et al. Science with the Australian Square Kilometre Array Pathfinder[J]. PASA, 2007, 24: 174–188. DOI: 10.1071/AS07033.

[94] Currie M J, Berry D S, Jenness T, et al. Starlink Software in 2013[C]//Manset N, Forshay P. Astronomical Society of the Pacific Conference Series: volume 485 Astronomical Data Analysis Software and Systems XXIII. [S.l.: s.n.], 2014: 391.

[95] van der Hulst J M, Terlouw J P, Begeman K G, et al. The Groningen Image Processing SYstem, GIPSY[C]//Worrall D M, Biemesderfer C, Barnes J. Astronomical Society of the Pacific Conference Series: volume 25 Astronomical Data Analysis Software and Systems I. [S.l.: s.n.], 1992: 131.

[96] DOLAG K, REINECKE M, GHELLER C, et al. Splotch: visualizing cosmological simulations[J/OL]. New Journal of Physics, 2008, 10(12): 125006. http://stacks.iop.org/1367-2630/10/i=12/a=125006.

[97] PIEPER S, HALLE M, KIKINIS R. 3d slicer[C]//2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821). 2004: 632–635 Vol. 1. DOI: 10.1109/ISBI.2004.1398617.

[98] HESS R. Blender foundations: The essential guide to learning blender 2.6[M]. [S.l.]: Focal Press, 2010.

[99] Kent B R. Visualizing Astronomical Data with Blender[J]. PASP, 2013, 125: 731. DOI: 10.1086/671412.

[100] Tully R B, Shaya E J, Karachentsev I D, et al. Our Peculiar Motion Away from the Local Void[J]. ApJ, 2008, 676: 184–205. DOI: 10.1086/527428.

[101] Ai M, Zhu M, Xiao L, et al. Properties of the UCHII region G25.4NW and its associated molecular cloud[J]. Research in Astronomy and Astrophysics, 2013, 13: 935–944. DOI: 10.1088/1674-4527/13/8/005.

[102] Barrett P E, Bridgman W T. PyFITS, a FITS Module for Python[C]// Mehringer D M, Plante R L, Roberts D A. Astronomical Society of the Pacific Conference Series: volume 172 Astronomical Data Analysis Software and Systems VIII. [S.l.: s.n.], 1999: 483.

[103] Barrett P, Hsu J C, Hanley C, et al. PyFITS: Python FITS Module[M]. [S.l.: s.n.], 2012.

[104] Spurzem R, Berentzen I, Berczik P, et al. Parallelization, Special Hardware and Post-Newtonian Dynamics in Direct N - Body Simulations[C]//Aarseth S J, Tout C A, Mardling R A. Lecture Notes in Physics, Berlin Springer Verlag: volume 760 The Cambridge N-Body Lectures. 2008: 377. DOI: 10.1007/978-1-4020-8431-7_15.

[105] Zini M F, Porozov Y, Andrei R M, et al. BioBlender: Fast and Efficient All Atom Morphing of Proteins Using Blender Game Engine[J]. ArXiv e-prints, 2010.

[106] Florinsky I V, Filippov S V. Development of virtual morphometric globes using Blender[J]. ArXiv e-prints, 2015.

[107] JOHN M S, COWEN M B, SMALLMAN H S, et al. The use of 2d and 3d displays for shape-understanding versus relative-position tasks [J/OL]. Human Factors, 2001, 43(1): 79–98. https://doi.org/10.1518/001872001775992534. DOI: 10.1518/001872001775992534.

[108] PIRINGER H, KOSARA R, HAUSER H. Interactive focus+context visual-

ization with linked 2d/3d scatterplots[C]//Proceedings. Second International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2004. 2004: 49–60. DOI: 10.1109/CMV.2004.1319526.

[109] Goodman A A, Rosolowsky E W, Borkin M A, et al. A role for self-gravity at multiple length scales in the process of star formation[J]. Nature, 2009, 457: 63–66. DOI: 10.1038/nature07609.

[110] Hassan A H, Fluke C J, Barnes D G. A Distributed GPU-Based Framework for Real-Time 3D Volume Rendering of Large Astronomical Data Cubes[J]. PASA, 2012, 29: 340–351. DOI: 10.1071/AS12025.

[111] Ferrand G, English J, Irani P. 3D visualization of astronomy data cubes using immersive displays[J]. ArXiv e-prints, 2016.

[112] CRUZ-NEIRA C, SANDIN D J, DEFANTI T A. Surround-screen projection-based virtual reality: The design and implementation of the cave[C/OL]// SIGGRAPH '93: Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques. New York, NY, USA: ACM, 1993: 135–142. http://doi.acm.org/10.1145/166117.166134. DOI: 10.1145/166117.166134.

[113] Abazajian K N, Adelman-McCarthy J K, Agüeros M A, et al. The Seventh Data Release of the Sloan Digital Sky Survey[J]. ApJS, 2009, 182: 543–558. DOI: 10.1088/0067-0049/182/2/543.

[114] Skrutskie M F, Cutri R M, Stiening R, et al. The Two Micron All Sky Survey (2MASS)[J]. AJ, 2006, 131: 1163–1183. DOI: 10.1086/498708.

[115] Kent B R. Spherical Panoramas for Astrophysical Data Visualization[J]. PASP, 2017, 129(5): 058004. DOI: 10.1088/1538-3873/aa5543.

[116] TARNG W, LIOU H. The application of virtual reality in astronomy education [J/OL]. Adv. Technol. Learn., 2007, 4(3): 160–169. http://dl.acm.org/citation.cfm?id=1722204.1722211.

[117] Chambers K C, Magnier E A, Metcalfe N, et al. The Pan-STARRS1 Surveys [J]. ArXiv e-prints, 2016.

[118] Green G M, Schlafly E F, Finkbeiner D P, et al. A Three-dimensional Map of Milky Way Dust[J]. ApJ, 2015, 810: 25. DOI: 10.1088/0004-637X/810/1/25.

[119] SHEN H. Interactive notebooks: Sharing the code[J/OL]. Nature, 2014, 515(7525): 151–152. https://doi.org/10.1038/515151a. DOI: 10.1038/515151a.

[120] SHERIDAN T B. Interaction, imagination and immersion some research needs [C/OL]//VRST '00: Proceedings of the ACM Symposium on Virtual Reality

Software and Technology. New York, NY, USA: ACM, 2000: 1–7. http://doi.acm.org/10.1145/502390.502392. DOI: 10.1145/502390.502392.

[121] BRYSON S. Virtual reality in scientific visualization[J/OL]. Commun. ACM, 1996, 39(5): 62–71. http://doi.acm.org/10.1145/229459.229467. DOI: 10.1145/229459.229467.

[122] Schlegel D J, Finkbeiner D P, Davis M. Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds[J]. ApJ, 1998, 500: 525–553. DOI: 10.1086/305772.

[123] MILGRAM P, KISHINO F. A taxonomy of mixed reality visual displays[J]. IEICE Trans. Information Systems, 1994, E77-D(12): 1321–1329.

[124] FERRAND G, ENGLISH J, IRANI P. 3D visualization of astronomy data cubes using immersive displays[J]. arXiv.org, 2016.

[125] Trenti M, Hut P. Gravitational N-body Simulations[J]. ArXiv e-prints, 2008.

[126] Nan R, Li D, Jin C, et al. The Five-Hundred Aperture Spherical Radio Telescope (fast) Project[J]. International Journal of Modern Physics D, 2011, 20: 989–1024. DOI: 10.1142/S0218271811019335.

[127] Ransom S M, Eikenberry S S, Middleditch J. Fourier Techniques for Very Long Astrophysical Time-Series Analysis[J]. AJ, 2002, 124: 1788–1809. DOI: 10.1086/342285.

[128] LEISERSON C E, SCHARDL T B. A work-efficient parallel breadth-first search algorithm (or how to cope with the nondeterminism of reducers)[M]// SPAA '10: Proceedings of the Twenty-second Annual ACM Symposium on Parallelism in Algorithms and Architectures. New York, NY, USA: ACM, 2010: 303–314. DOI: 10.1145/1810479.1810534.

[129] BREIMAN L, OTHERS. Classification and Regression Trees[M/OL]. New York: Chapman & Hall, 1984: 358. http://www.crcpress.com/catalog/C4841.htm.

[130] The LIGO Scientific Collaboration, The Virgo Collaboration. GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral[J]. ArXiv e-prints, 2017.

作者简历

作者基本情况

钱旭冉，女，安徽省安庆市人，生于 1990 年 9 月 21 日，中国科学院国家天文台博士研究生。

2015.10 – 2017.10	哈佛史密松天体物理中心 (CfA) Predoc Fellow （国家公派联合培养）
2012.9 – 今	中国科学院国家天文台博士研究生（硕博连读）
2008.9 – 2012.7	北京语言大学本科生

联系方式

E-mail: qianxuran@nao.cas.cn

通讯地址：北京市朝阳区大屯路甲 20 号中国科学院国家天文台，100012

攻读学位期间发表的学术论文及科研成果

1. **钱旭冉**，朱明，蔡栩：基于 *Blender* 的天文数据三维可视化，天文研究与技术，第 12 卷，第 2 期，2015 年 4 月
2. **Qian, P. X.**, Cai, M. X., Portegies Zwart, S., & Zhu, M., *SiMon:Simulation Monitor for Computational Astrophysics*, PASP, 2017, 129: 979

攻读博士学位期间的获奖情况

1. 2017 年度国家奖学金（博士）
2. 2016 年度 AMD-国台天文台奖学金二等奖
3. 2014-2015 年度国科大 “优秀学生干部”
4. 2014-2015 年度 “三好学生”

致　谢

在国家天文台度过的六年时光，恍惚间竟已进入尾声。回想多年前懵懂踏入天文这个充满奇幻色彩的领域，心中充满好奇，想要探索，想要亲手解开许多奥秘。从计算机科学到天文技术，由获取知识到应用知识解决问题，这些转变的过程中有着很多艰辛。但可幸的是，身边一直有老师、同学、朋友、亲人的陪伴与鼓励。

首先希望感谢我的导师朱明老师。朱老师在学术上非常严谨，而这种严谨的态度慢慢纠正了我在科研中对细节的不重视的问题，记得在写第一篇论文的时候，我总是急急的码好内容就想提交，但是朱老师告诉我论文撰写需要静下心来缓缓完成，并给我的论文提出了很多建议，最终论文也非常顺利的被接收；朱老师非常平易近人和关心学生，跟朱老师的相处模式亦师亦友，我比较缺乏天文背景，曾多次向朱老师请教基础的天文知识，但是朱老师总是非常耐心的给我讲解，一个小时或多个小时直到我能明白，朱老师对学生的邮件会及时回应；朱老师也十分关心学生的生活状态，在我对选择专业困惑时，或者在对未来发展担忧时，朱老师总是坚定的给我最适合我发展的建议；朱老师非常鼓励学生进行会议演讲或出国交流，这也使得我这个履历尚浅的学生得以见到更广阔的世界和遇到不同的人，也使得我有了两年非常难得的在美国交流的机会，毕业在即的我深深地感受到这种开放性指导方式的对科研独立性和交流能力培养的重要性。古人云：师者，传道授业解惑者也！朱老师在我心里就是一位师者，非常感激有朱老师的一路陪伴。

其次希望感谢我在美国的指导老师 Alyssa Goodman 教授。Alyssa 是我心中的女性楷模，她学识丰富，和她相处的每个细节都能有所收获，小到如何设置 MacOS 的显示比例，大到论文和 Proposal 撰写，她总是拥有非常多且独具创意的思想，而她也非常乐于分享并指导学生将这些点子实现；作为一位哈佛名教授，她却是那么的谦逊和耐心，她会在会议时我有理解困难的时候主动用更通俗易懂的方式讲解给我听，在日常社交中她也会主动为我创造话题，在与她的邮件或短信交流中，她会主动表达感谢和歉意，Alyssa 不仅教会了我如何更好的科研，更是通过她的实际行动指导我如何做一个更好的人。

我还要感谢在美国留学的另一位指导老师 Michelle Borkin 教授。Michelle 是我心目中另一位非常佩服的女性学者，她的鼓励和耐心给予我极大的信心，她对我的项目提出了非常多的意见，而且她也非常支持和鼓励我的想法，包括对未来的规划，她不仅会在大方向上给与我指导，也会在细节处理上与我一同疏通，她的言传身教也使得我在待人处事方法学到很多。

感谢国台的好友们。感谢艾美、张博和杨超对我工作和学习上的帮助，感谢周渝涛和刘晓杰与我一起分享开心和快乐，感谢赵欣、张洋、徐小钧、刑树果、赵哲等学生会大家庭成员的共同合作和成长，感谢同组的郭元旗、苑利霞、钟益、李

然帮我远程处理和递交材料，感谢肖莉师姐和陈如荣师兄为我解答问题，感谢崔辰州老师和张海燕老师为我申请出国准备推荐信，感谢李菂老师对我项目的指导，也感谢 FAST 科学部的焦倩、黄梦林师姐、岳友玲师兄、李会贤师姐、钱磊师兄、于萌师兄以及毕业的吴碧雨、郝巧丽等，感谢你们的一路支持与陪伴。

感谢英国的合作者 Tom Robitaille，Tom 是一位极具才华的天文学家和程序员，他给予我很多编程和项目管理的指导。感谢同组的哈佛研究生 Catherine Zucker 和 Hope Chen，他们都在我项目需要帮助的时候及时帮助，并给予我很多语言、文化和生活上的帮助。感谢 Simon Portegies Zwart 教授给予的论文和工作的帮助，感谢 Doug Finkbeiner 教授对我项目的指导，也感谢 Tom Dame 教授给予我项目上的意见。

感谢我在 CfA 的好友们。感谢岳楠楠、Joyce Lee、Şeyma Mercimek，感谢她们在我初来美国时的陪伴和关心，也感谢我们四个女生相伴的所有快乐时光，它们是我最美好的回忆。感谢 Riwaj Pokhrel 和张高原，感谢我们一起度过的电影之夜与美食聚会。感谢 Junko Udea，感谢你带给我的美国的第一份友谊。感谢龙曦的美食推荐和练车建议，感谢邱艳丽、朱辉、崔晓红（晓红姐）、李宗南、熊放、李尚活、杨君等，感谢你们的用心陪伴，也感谢我的办公室室友们 Marco Zennaro、Cara Battersby、Giovanni Dipierro、Fernando Rico 等，感谢你们与我分享你们国家的文化。

感谢我的美国舍友 Abie Baafi 和 Robin Neldridge，感谢她们像我的两个大姐姐般对我生活上的帮助和帮我疏解生活上的烦恼，非常感激和她们一起分享两年的喜怒哀乐和不同的风俗文化，还有她们做的美味的美国食物。感谢我的好友和邻居 Nishu Karna，感谢她经常周末开车载我购物游玩，也很感激她与我分享开心和悲伤。感谢 Chichi & Gershon Larsen 夫妇在波士顿对我的照顾，感谢 Matt Chin 像大哥哥般对我的关心和鼓励。感谢 Alex & Christa Mayfield 夫妇对我的关照。感谢波士顿的 Aliza Llovet, Mike Lee & Lisa Jeon 等为我在波士顿的生活带来了很多欢乐。

最后，感谢国家天文台研究生部的杜红荣主任、艾华老师以及马怀宇老师。感谢 CfA 的 Sarah Block 和 Christine Crowley。感谢中国科学院国家天文台（NAOC）、中国科学院大学（UCAS）和哈佛史密松天文台 (CfA) 提供的优越的学习和科研环境，也感谢国家留学基金委 (CSC) 给我提供的出国交流的机会和基金支持。

谨把本文献给我的父母以及 Maxwell Cai。