“**Speedrun: Data Science Fundamentals (update)**”是发布在 learn.palantir.com 上的一门实战导向课程，由 Ontologize 团队（前 Palantir 工程师组成）提供 1, 2。
该课程旨在通过一个完整的**临床研究（Clinical Study）**案例，指导用户如何在 Palantir Foundry 中利用数据科学工具链处理从原始数据到交互式分析报告的全流程 1, 2。以下是基于来源对该课程内容的详细介绍：

1. 核心目标与业务场景

课程的核心目标是让用户熟悉 Foundry 中的数据科学工作流，包括数据预处理、调和（**Reconciliation**）、模型分析和结果发布 1。

- 业务场景：分析临床研究数据，通过患者的人口统计学信息、不良事件（Adverse Events）数据以及患者反馈，来理解与药物消耗相关的风险 2, 3。
- 多样化数据源：课程涵盖了结构化表格（CSV）和非结构化数据（PDF 文件）的处理 1, 2。

2. 关键技术流程与工具

A. 数据集成与预处理 (Pipeline Builder)

用户首先在 **Pipeline Builder** 中构建数据管道，对数据进行清洗和统一 1, 4。

- 多源合一：将人口统计数据（DM）、不良事件数据（AE）以及患者反馈媒体集（Media Set）导入 4, 5。
- **AIP** 与文本提取：利用 **AIP (Artificial Intelligence Platform)** 从 PDF 格式的患者反馈中提取原始文本 1, 6。
- **LLM** 情感分析：使用大语言模型（如 GPT-4 或 Gemini 2.0 Flash）对提取的文本进行处理，将其转换为结构化数据，包括患者 **ID**、情感评分和摘要 7, 8。
- 数据连接 **(Joins)**：通过左连接（Left Join）将上述结构化后的反馈数据与人口统计和不良事件表格合并，形成一个全方位的患者信息资产 9。

B. 交互式数据分析 (Jupyter Lab)

在数据准备就绪后，课程引导用户进入 **Jupyter Lab** 环境进行深入探索 10。

- 环境优势：在 Foundry 内部使用 Jupyter，既能保留数据科学家熟悉的 Python 开发体验，又能直接访问受治理的“组织单一事实来源（Source of Truth）” 10, 11。
- 可视化：通过安装 matplotlib 等库，编写 Python 代码生成直观的直方图，分析不同治疗组（安慰剂组 vs. 治疗组）的年龄分布以及不良事件的发生频率 12, 13。

C. 构建数据应用 (Streamlit)

为了将分析结果分享给非技术决策者，课程教授如何创建 **Streamlit** 应用程序 14。

- 交互式看板：用户可以构建一个“不良事件年龄分布探索器”，允许其他用户通过下拉菜单选择特定的不良事件类型（如焦虑或食欲下降），动态查看相关的年龄分布直方图和患者反馈摘要 15, 16。
- 版本控制与共享：Streamlit 应用依托于代码仓库，支持分支管理，并可以通过链接直接分享给拥有权限的同事 17。

3. 管理与治理：数据血缘 (Data Lineage)

课程强调了在 Foundry 中进行数据科学工作的独特优势——数据出处（**Data Providence**）18。

- 端到端可见性：通过 Data Lineage 应用，用户可以清晰地看到从原始 PDF 和 CSV 到 Pipeline Builder 转换，再到最终 Jupyter 分析和 Streamlit 应用的完整路径 18, 19。
- 可重复性：这种透明度确保了分析结果是可追溯且可重复的，避免了传统本地数据科学流程中常见的“数据丢失”或“过程黑箱”问题 18。

4. 学习建议与后续

- 权限要求：参加课程需要具备 Pipeline Builder、Jupyter Workspace、Streamlit 和 AIP 的访问权限 2, 20。
- 后续扩展：完成基础流程后，课程鼓励用户尝试在 Jupyter 环境中进一步训练机器学习模型，并参考 Foundry 提供的 AIP 示例进行更高级的应用开发 21。

通过这门课程，用户不仅能学会使用单一工具，更能掌握如何将 **AIP**、**Pipeline Builder** 和 **Code Workspaces** 协同工作，构建端到端的数据科学解决方案 1, 18。