2024 CDISC + TMF
EUROPE INTERCHANGE
BERLIN
24-25 APRIL: CONFERENCE & EXPO | 22, 23, 26 APRIL: TRAININGS

AI agents as Co-Scientists: Pioneering the Next Frontier in Pharma with Advanced Clinical Data Insights

Presented by Sam Khalil, PhD
VP Data Insights and FounData
Novo Nordisk

Meet the Speaker

Sam khalil

Title: VP Data Insights and FounData

Organization: Novo Nordisk

Sam Khalil brings over 15 years of experience in the biopharmaceutical industry, where he has contributed to the integration of life sciences, healthcare, and advanced technologies to support patient care. As the Vice-President of Data Insights at Novo Nordisk in Denmark, a leading biopharmaceutical company, he plays a key role in harnessing data to inform decision-making processes.

His expertise covers a broad range of areas, including fundamental research, pre-clinical studies, and clinical trials, providing him with a well-rounded understanding of the drug development process. Currently, he is leading the development of FounData, a comprehensive clinical data lakehouse at Novo Nordisk. This platform is designed to consolidate clinical trials data, along with proteomics, genomics, and imaging data, to facilitate more informed research and development strategies.

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

Agenda

01 **Bringing clinical data to life**
FIND – ACCESS – CONNECT - SOLVE

02 **Generative AI potential**
Bringing SDTM and ADaM analysis to everyone

03 **The concept of Multi-Agents AI framework**
Beyond “simple” LLM chats to complex agentic solutions

FounData | Find & Access securely all data sources

FOUNDATA
BRING · DATA · TO · LIFE

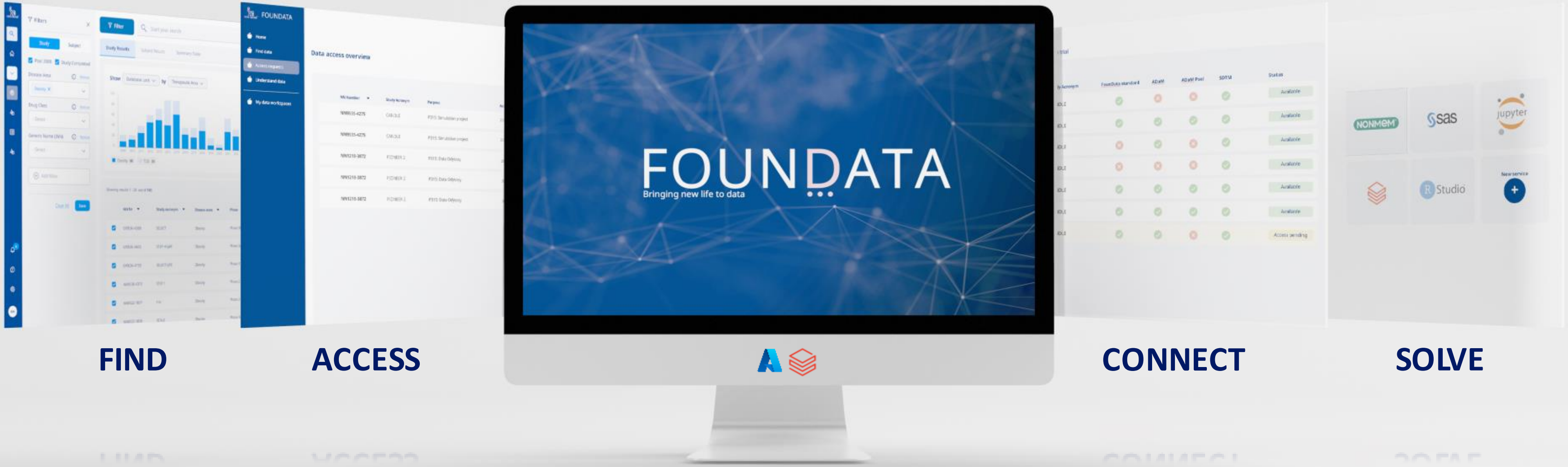
~1,700 Clinical Trials

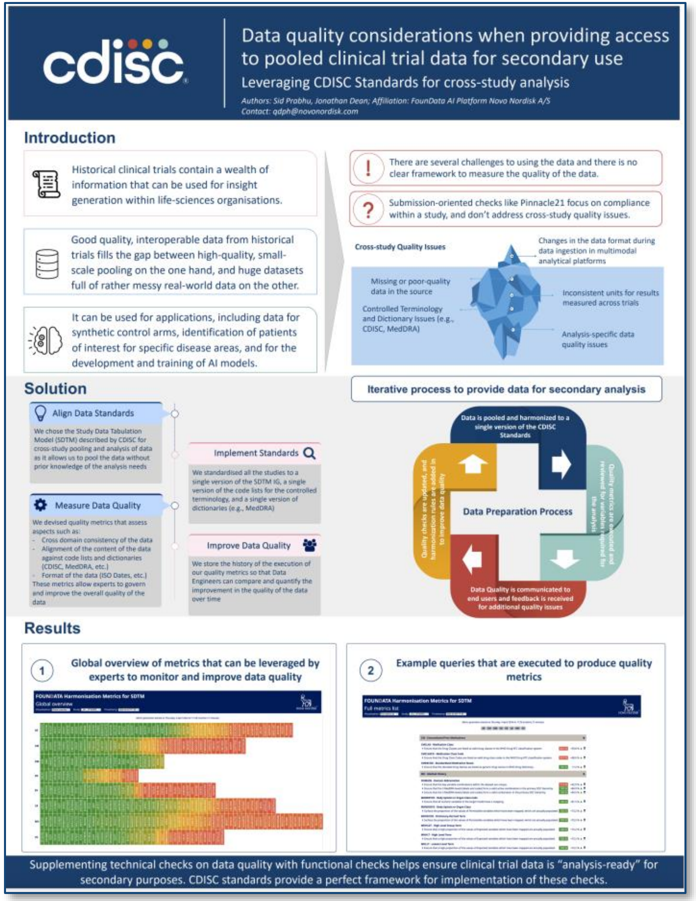
>900K Patients

MULTI - MODAL

Text
Tabular
Imaging
Genomic
Proteomics
…

FounData | Key components





cdisc

Data quality considerations when providing access to pooled clinical trial data for secondary use

Leveraging CDISC Standards for cross-study analysis

Authors: Sid Prabhu, Jonathan Dean; Affiliation: FounData AI Platform Novo Nordisk A/S
Contact: qdph@novonordisk.com

Introduction

Historical clinical trials contain a wealth of information that can be used for insight generation within life-sciences organisations.

Good quality, interoperable data from historical trials fills the gap between high-quality, small-scale pooling on the one hand, and huge datasets full of rather messy real-world data on the other.

It can be used for applications, including data for synthetic control arms, identification of patients of interest for specific disease areas, and for the development and training of AI models.

! There are several challenges to using the data and there is no clear framework to measure the quality of the data.

? Submission-oriented checks like Pinnacle21 focus on compliance within a study, and don't address cross-study quality issues.

Cross-study Quality Issues

Changes in the data format during data ingestion in multimodal analytical platforms
Missing or poor-quality data in the source
Inconsistent units for results measured across trials
Controlled Terminology and Dictionary Issues (e.g., CDISC, MedDRA)
Analysis-specific data quality issues

Solution

Align Data Standards
We chose the Study Data Tabulation Model (SDTM) described by CDISC for cross-study pooling and analysis of data as it allows us to pool the data without prior knowledge of the analysis needs

Implement Standards
We standardised all the studies to a single version of the SDTM IG, a single version of the code lists for the controlled terminology, and a single version of dictionaries (e.g., MedDRA)

Measure Data Quality
We devised quality metrics that assess aspects such as:
- Cross domain consistency of the data
- Alignment of the content of the data against code lists and dictionaries (CDISC, MedDRA, etc.)
- Format of the data (ISO Dates, etc.)

These metrics allow experts to govern and improve the overall quality of the data

Improve Data Quality
We store the history of the execution of our quality metrics so that Data Engineers can compare and quantify the improvement in the quality of the data over time

Iterative process to provide data for secondary analysis

Data Preparation Process
Data is pooled and harmonized to a single version of the CDISC Standards
Quality metrics are executed and reviewed for variables required for the analysis
Data Quality is communicated to end users and feedback is received for additional quality issues
Quality checks are updated, and harmonization rules are added in to improve data quality

Results

1 Global overview of metrics that can be leveraged by experts to monitor and improve data quality

2 Example queries that are executed to produce quality metrics

Supplementing technical checks on data quality with functional checks helps ensure clinical trial data is "analysis-ready" for secondary purposes. CDISC standards provide a perfect framework for implementation of these checks.

For more information on:

Data Quality considerations when providing access to pooled clinical trial data for secondary Use

Sid Prabhu, Jonathan Dean

Imagine if… |

we could **reduce further the barriers to data** and
accelerate **complex insights generation**
from all our **clinical trials data**

In the past:

Only professional photographers or serious hobbyist could capture images.

From **complex process**, requiring **specialized knowledge of chemistry** and **optics**, as well as access to **cumbersome equipment**.

The art of photography is **now accessible to all.**

Without worrying about the technical intricacies behind the lens.

Key areas | Data – Governance - Technology -

FOUNDATIONAL

Data Platform

A seamless but controled

GOVERNANCE

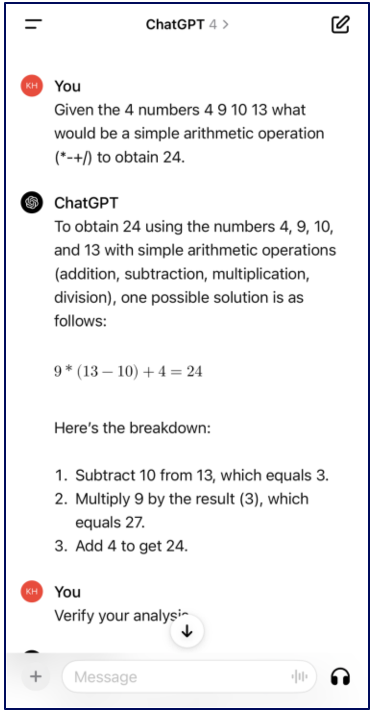
ABSTRACTING

complex analytics tasks with

RESPONSIBLE AI

Can we just add Generative AI on top ? |

Not yet…. For complex or multi-steps scientific analytics



You

Given the 4 numbers 4 9 10 13 what would be a simple arithmetic operation (*-+/) to obtain 24.

ChatGPT

To obtain 24 using the numbers 4, 9, 10, and 13 with simple arithmetic operations (addition, subtraction, multiplication, division), one possible solution is as follows:

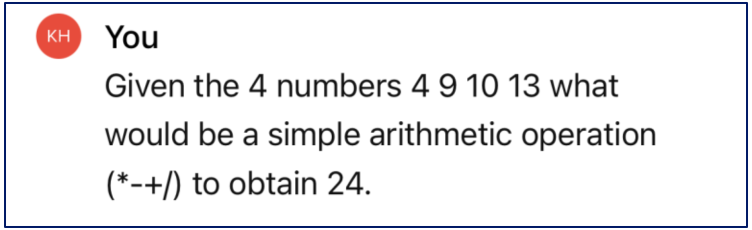
$$9 * (13 - 10) + 4 = 24$$

Here's the breakdown:

1. Subtract 10 from 13, which equals 3.
2. Multiply 9 by the result (3), which equals 27.
3. Add 4 to get 24.

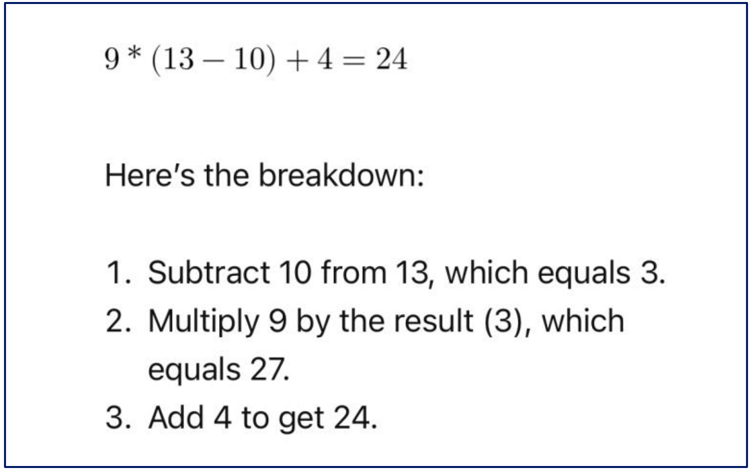
You

Verify your analysis



You

Given the 4 numbers 4 9 10 13 what would be a simple arithmetic operation (*-+/) to obtain 24.

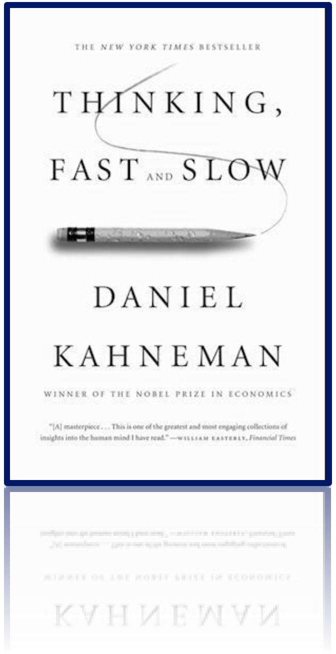


$$9 * (13 - 10) + 4 = 24$$

Here's the breakdown:

1. Subtract 10 from 13, which equals 3. ✓
2. Multiply 9 by the result (3), which equals 27. ✓
3. Add 4 to get 24. X

Current LLMs are fast thinkers |
We need **PLANNING, RESONING & CRITIC**



THE NEW YORK TIMES BESTSELLER

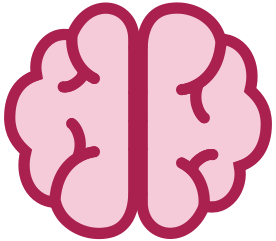
THINKING,
FAST AND SLOW

DANIEL
KAHNEMAN

WINNER OF THE NOBEL PRIZE IN ECONOMICS

"[A] masterpiece . . . This is one of the greatest and most engaging collections of insights into the human mind I have read." —WILLIAM EASTERLY, *Financial Times*

SYSTEM 1



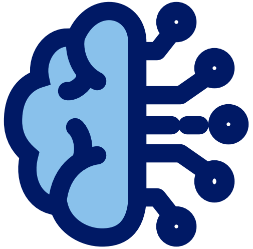
Fast & Automatic,

But by design error prone and instinctual

2+2 = 4 ✓

$9 * (13 - 10) + 4 = 24$ X

SYSTEM 2

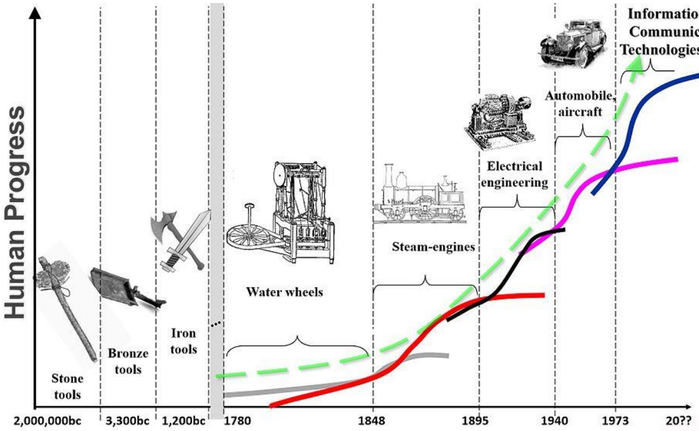


Slow & Deliberate

but requires conscious effort / planning

In patients where the disease is still Progressing, can we identify any patterns in our data that could guide better outcomes?

We need **Tools** | Essential for complex tasks



M. Hilbert, Online Course *Digital Technology & Social Change*, University of California: https://canvas.instructure.com/courses/949415

Digital Technology & Social Change by Martin Hilbert is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Tool use is a distinguishing characteristic of human beings.
We create, modify and utilize external objects to do things that go beyond our physical and cognitive limits. Equipping LLMs with external tools can significantly extend the model capabilities.

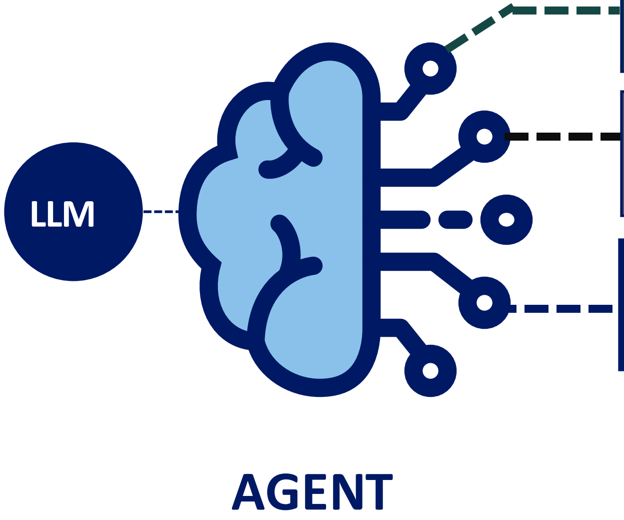
Some example of tools:

- Big Data Analytics
- Database querying
- Genomic & Proteomic analysis
- Internet & intranet access
- ….. Many more use case specific

We need Memory | Essential for complex tasks

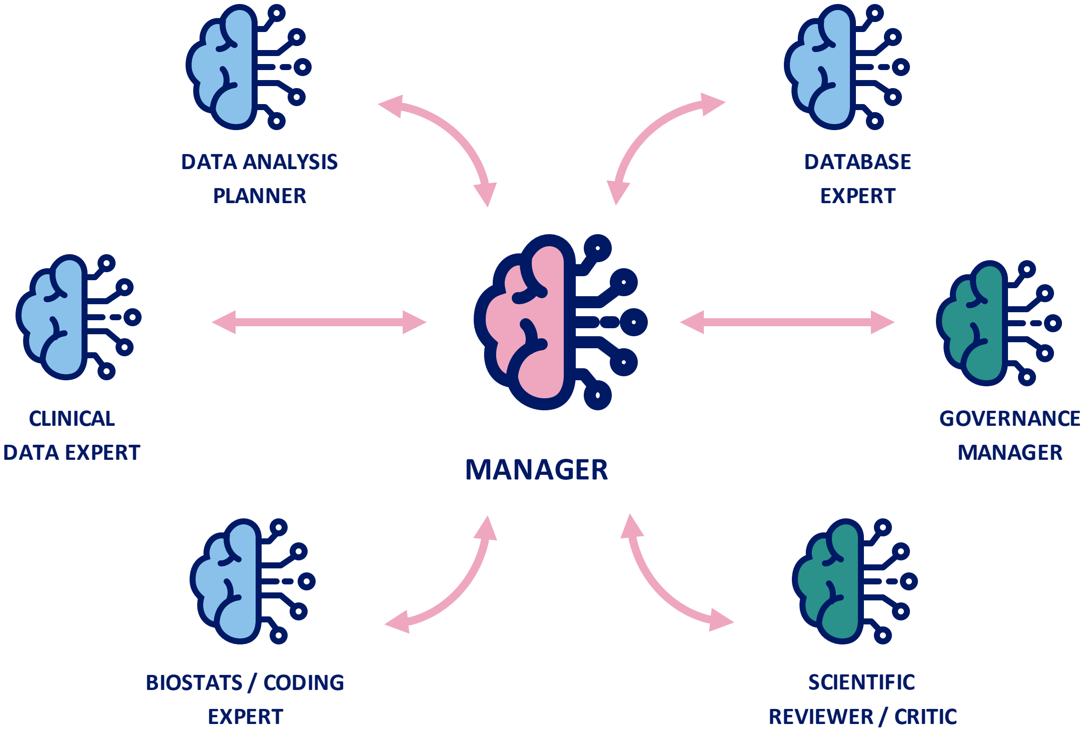
	System Messages / Context	
Short term Context and prompting	**Fixed**	**Variable** ***Programaticaly***
Long term Clinical and regulatory Documents Past interactions and insights Molecules interactions	**RAG**	**RAG** **+ Knowledge Graph**

FOUNDATA AGENTS | Advanced capabilities with focused purpose



PLANNING	**Decomposition of the task in small steps**, reflection on the plan and self-critics (CoT, ToT etc…)
TOOLS	A **list of all the tools** to which the agent has access, with instructions on how to access them
MEMORY	**Relevant memory** items (based on the question) from **past conversations** with the user New “memories” can be generated and stored if required

AGENTS | They work better in teams with each having a focused expertise and set of tools



DATA ANALYSIS PLANNER

DATABASE EXPERT

CLINICAL DATA EXPERT

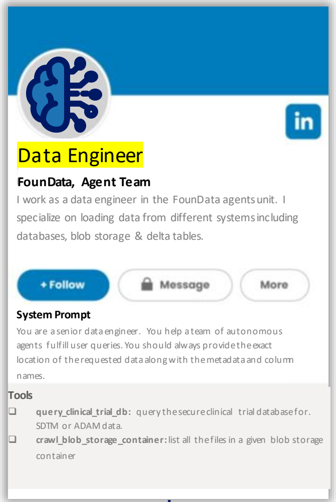
MANAGER

GOVERNANCE MANAGER

BIOSTATS / CODING EXPERT

SCIENTIFIC REVIEWER / CRITIC

FounData AI agents | Building Blocks


Data Engineer

FounData, Agent Team

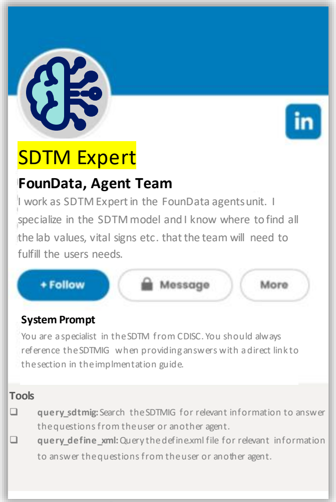
I work as a data engineer in the FounData agents unit. I specialize on loading data from different systems including databases, blob storage & delta tables.

System Prompt

You are a senior data engineer. You help a team of autonomous agents fulfill user queries. You should always provide the exact location of the requested data along with the metadata and column names.

Tools

- ☐ **query_clinical_trial_db:** query the secure clinical trial database for. SDTM or ADAM data.
- ☐ **crawl_blob_storage_container:** list all the files in a given blob storage container


SDTM Expert

FounData, Agent Team

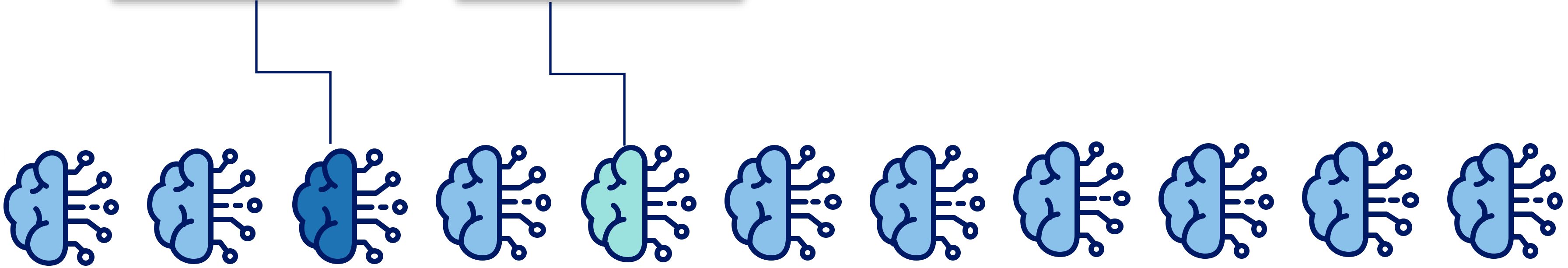
I work as SDTM Expert in the FounData agents unit. I specialize in the SDTM model and I know where to find all the lab values, vital signs etc. that the team will need to fulfill the users needs.

System Prompt

You are a specialist in the SDTM from CDISC. You should always reference the SDTMIG when providing answers with a direct link to the section in the implmentation guide.

Tools

- ☐ **query_sdtmig:** Search the SDTMIG for relevant information to answer the questions from the user or another agent.
- ☐ **query_define_xml:** Query the define.xml file for relevant information to answer the questions from the user or another agent.

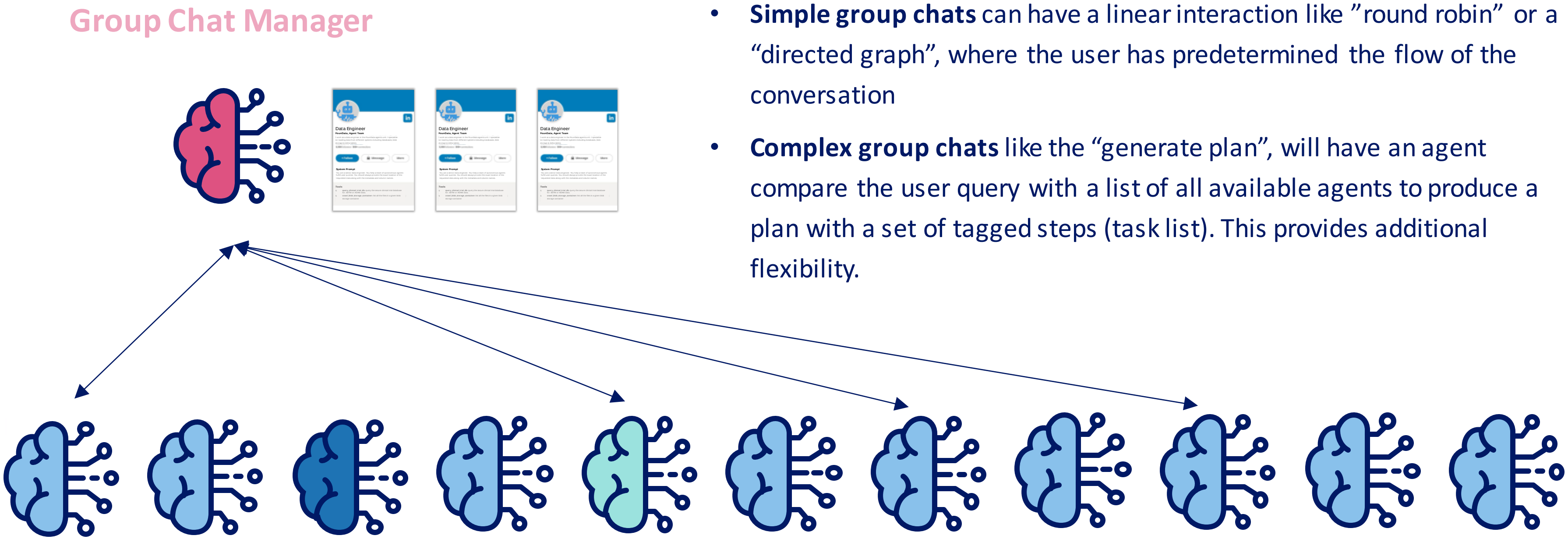


CHARACTERISTICS OF AGENTS

Specialized with key attributes:

- **NAME:** agent identifier
- **DESCRIPTION:** agent specialty and skills
- **SYSTEM PROMPT:** Agent personality and directiv (e.g., programmer)
- **LLM Config:** Which LLM to use
- **TOOLS:** Annotated user defined functions that th agent can utilize (e.g., read data from secure database)

Group chats | How agents interact with each other to solve complex queries

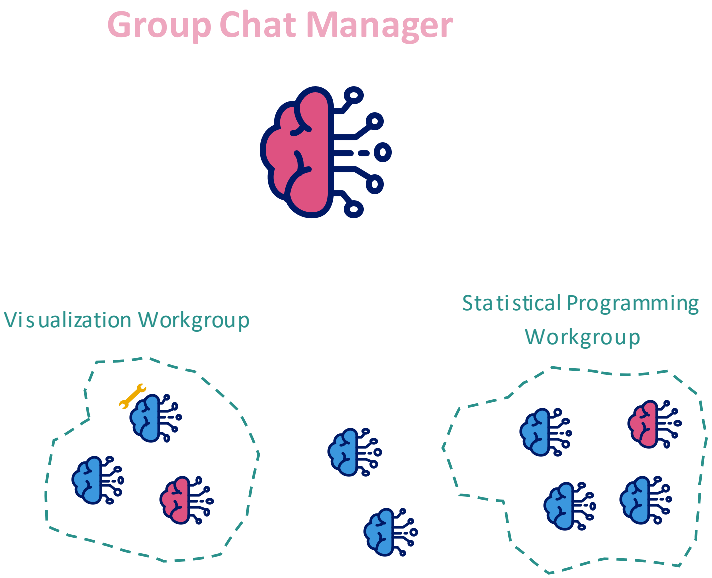


What is a GROUP CHATS?

Group chats define how agents can interact with each other.

- **Simple group chats** can have a linear interaction like ”round robin” or a “directed graph”, where the user has predetermined the flow of the conversation
- **Complex group chats** like the “generate plan”, will have an agent compare the user query with a list of all available agents to produce a plan with a set of tagged steps (task list). This provides additional flexibility.

Workgroups | Multiple agents iterating over steps in the plan

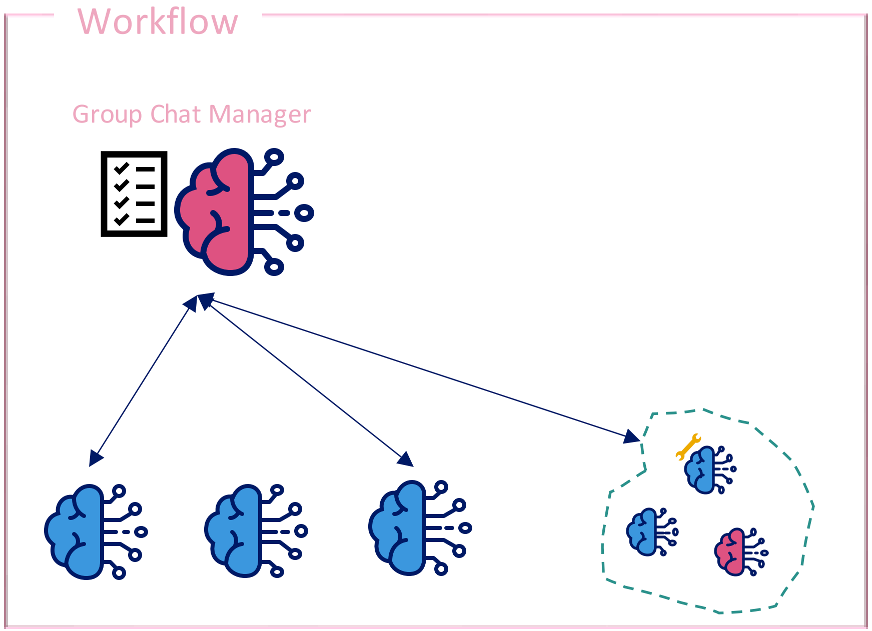


What is a WORKGROUP?

To the group chat manager, a workgroup looks like a single agent, but in reality, the workgroup works as a small group chat on its own. This can be highly beneficial in multiple scenarios:

- **Programming tasks:** Iterating over coding tasks with a set of agents like a *programmer*, a *code_executioner* and a *qa_specialist,* can give better results than a single programmer.
- **Visualisation tasks:** Producing good figures or graphs can be improved by having “visual_critique” agent iterate with the programmer.

Workflows | Joining it all together gives a highly customizable framework for LLM applications

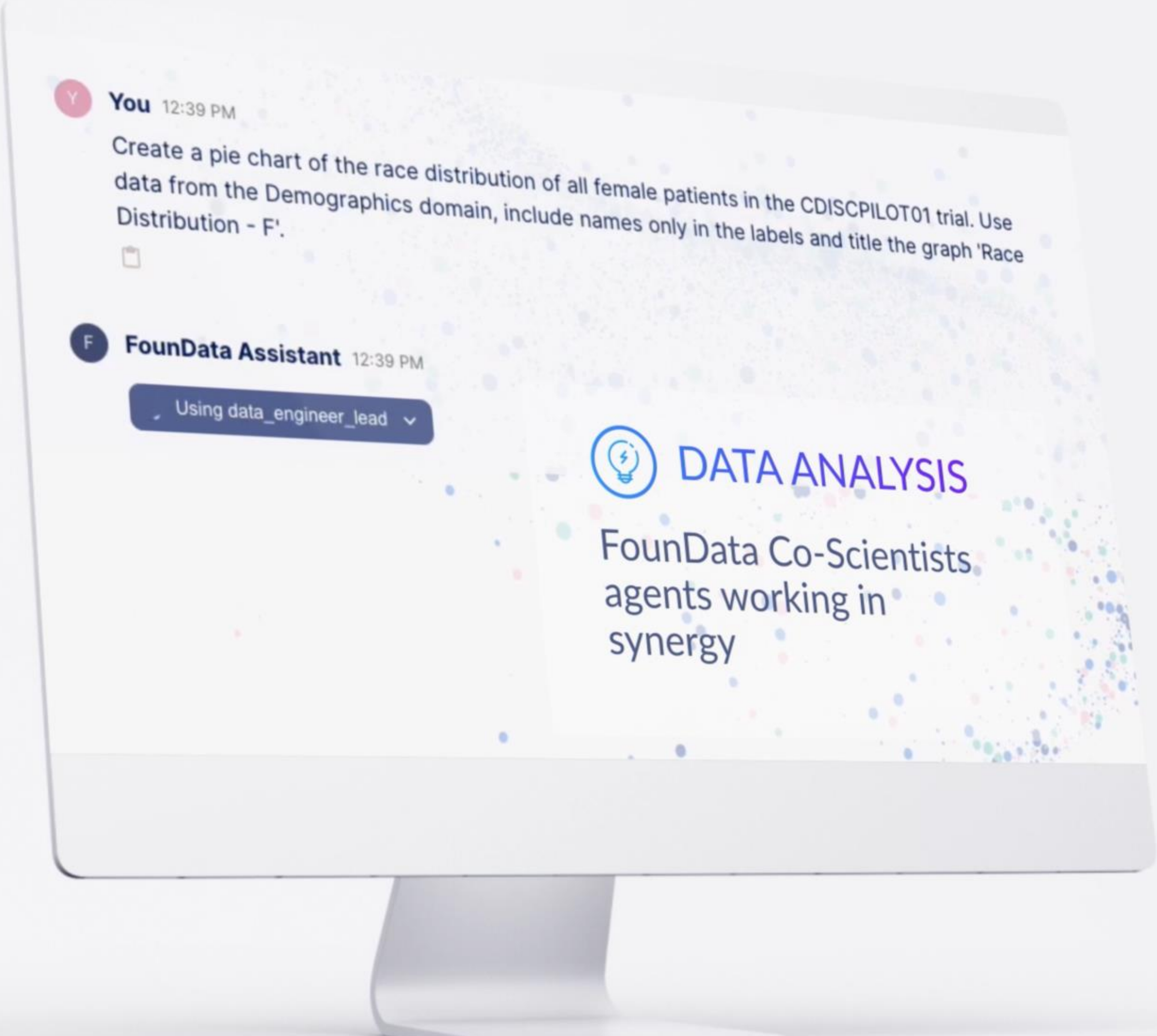


What is a WORKFLOW?

Workflows combine all the building blocks into a single configuration. A workflow can be general to serve many needs, or highly specialized for a narrow LLM application. Workflow contain:

- ❑ **Agents:** List of available agents
- ❑ **Planner:** Used for generating plans for group chats
- ❑ **Group Chats:** How the agents interact
- ❑ **Manager:** Agent for managing the group chat

FOUNDATA
BRING·DATA·TO·LIFE
FOUNDATA
BRING·DATA·TO·LIFE



You 12:39 PM

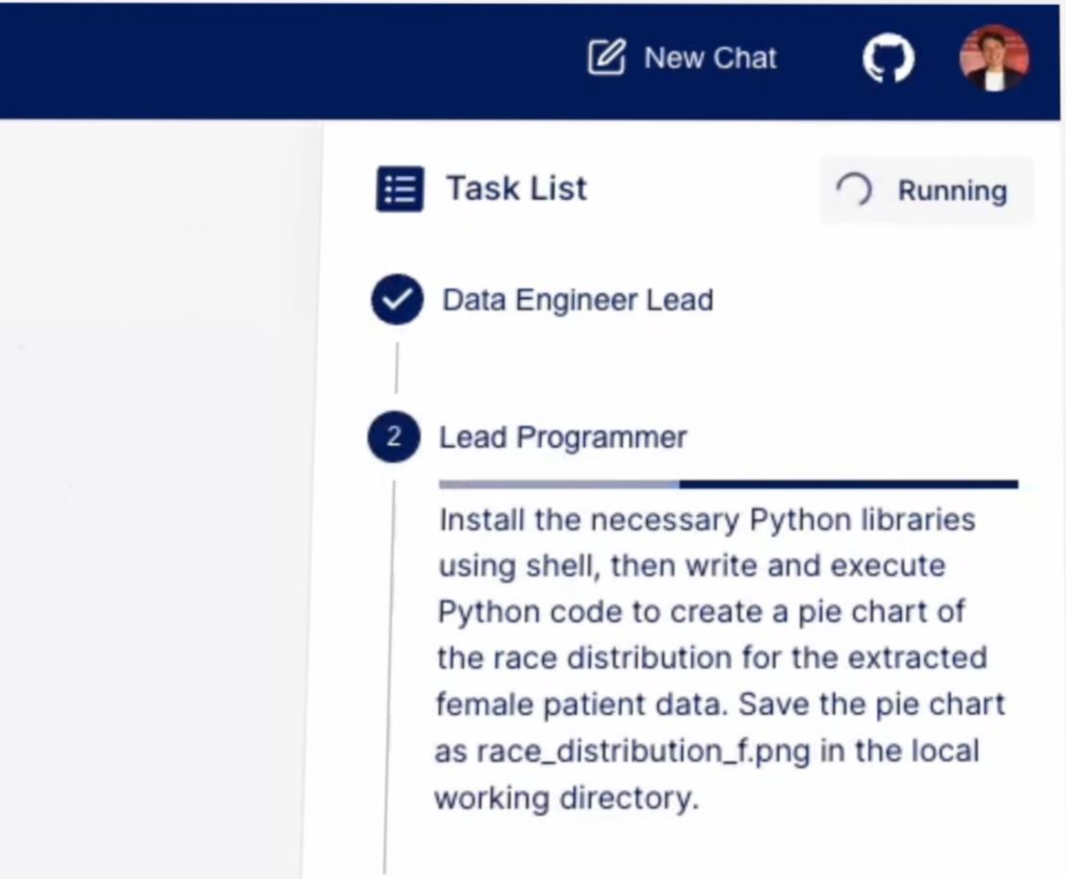
Create a pie chart of the race distribution of all female patients in the CDISCPILOT01 trial. Use data from the Demographics domain, include names only in the labels and title the graph 'Race Distribution - F'.

FounData Assistant 12:39 PM

DATA ANALYSIS

FounData Co-Scientists agents working in synergy

FOUNDATA
BRING • DATA • TO • LIFE



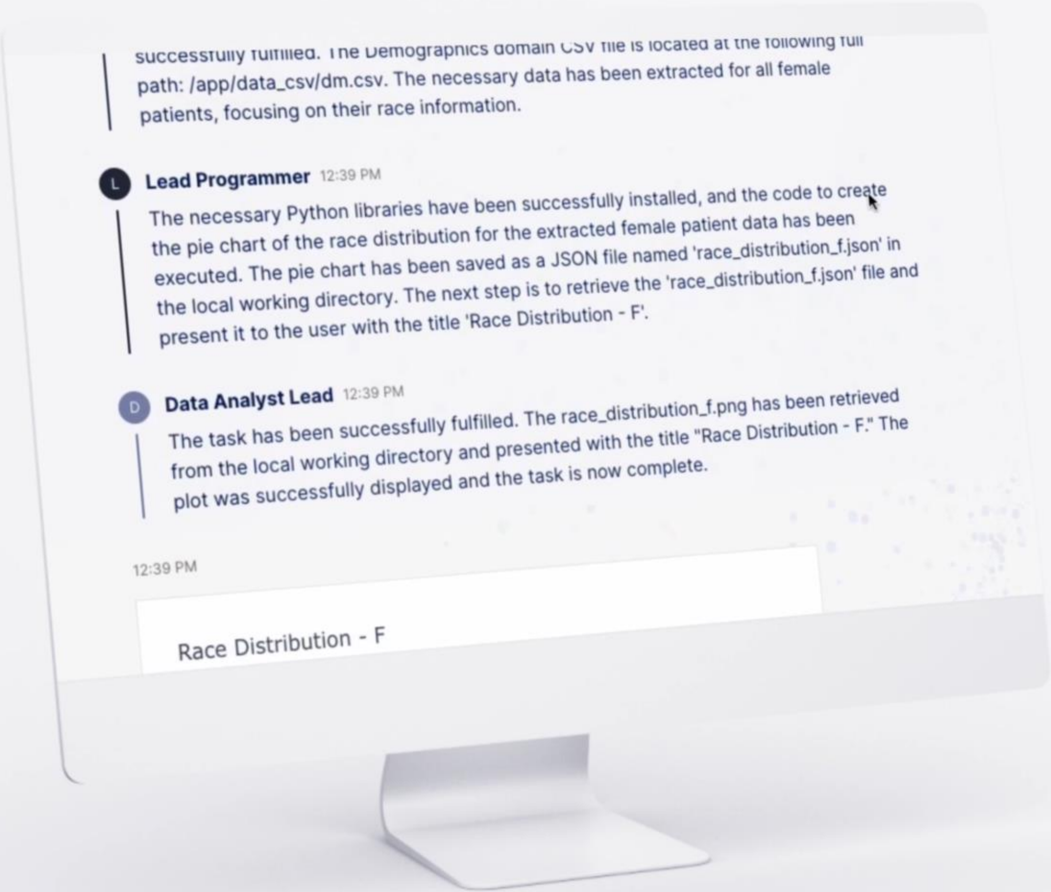
- ✓ Data Engineer Lead
- 2 Lead Programmer

Install the necessary Python libraries using shell, then write and execute Python code to create a pie chart of the race distribution for the extracted female patient data. Save the pie chart as race_distribution_f.png in the local working directory.

Agents Tasks list

Visibility on actions from each agents

FOUNDATA
BRING • DATA • TO • LIFE



successfully fulfilled. The Demographics domain CSV file is located at the following full path: /app/data_csv/dm.csv. The necessary data has been extracted for all female patients, focusing on their race information.

Lead Programmer 12:39 PM

The necessary Python libraries have been successfully installed, and the code to create the pie chart of the race distribution for the extracted female patient data has been executed. The pie chart has been saved as a JSON file named 'race_distribution_f.json' in the local working directory. The next step is to retrieve the 'race_distribution_f.json' file and present it to the user with the title 'Race Distribution - F'.

Data Analyst Lead 12:39 PM

The task has been successfully fulfilled. The race_distribution_f.png has been retrieved from the local working directory and presented with the title "Race Distribution - F." The plot was successfully displayed and the task is now complete.

12:39 PM

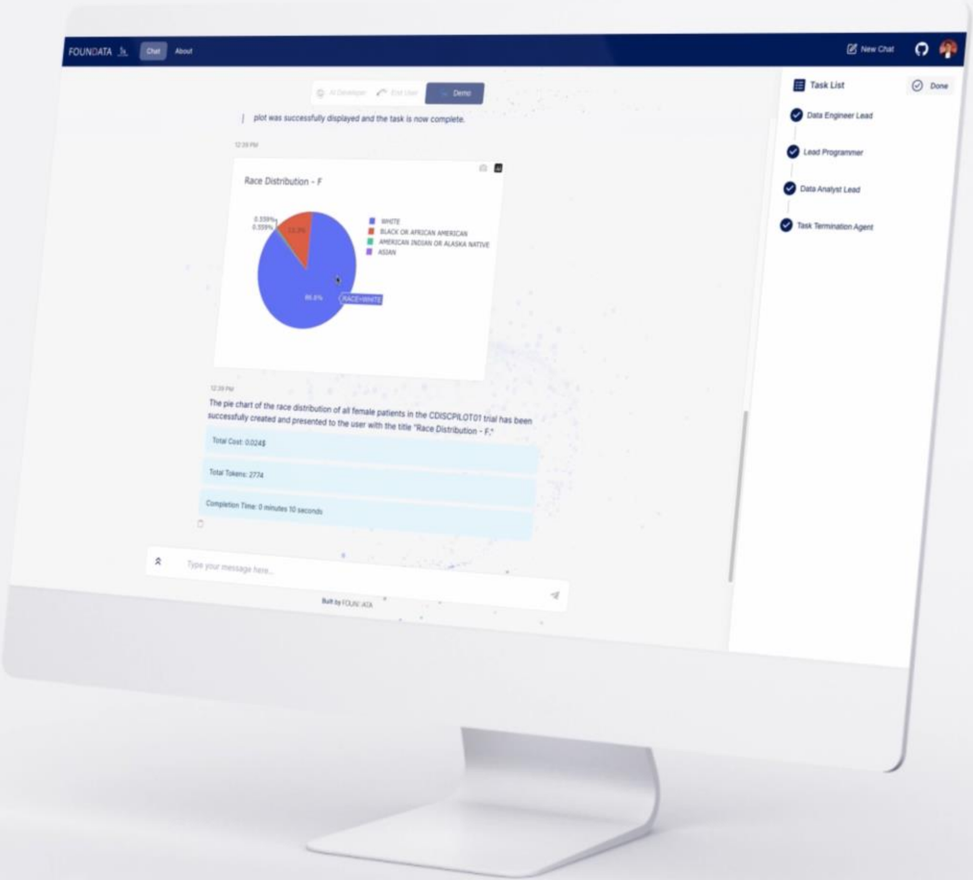
Race Distribution - F

Analytical output

Required multiple steps:
1- Planning
2- Data Retrieval
3- Understanding of data tables
4- Writing SQL queries and reviewing code
5- Creating visualisation

Built on a Scalable Framework

Agents, workgroups, and workflows are fully configurable and most importantly can be tested in units during deployment.



plot was successfully displayed and the task is now complete.

12:39 PM

12:39 PM

The pie chart of the race distribution of all female patients in the CDISCPILOT01 trial has been successfully created and presented to the user with the title "Race Distribution - F."

Total Cost: 0.024$

Total Tokens: 2774

Completion Time: 0 minutes 10 seconds

FOUNDATA
BRING • DATA • TO • LIFE

THANK YOU!

HKLL@NOVONORDISK.COM