

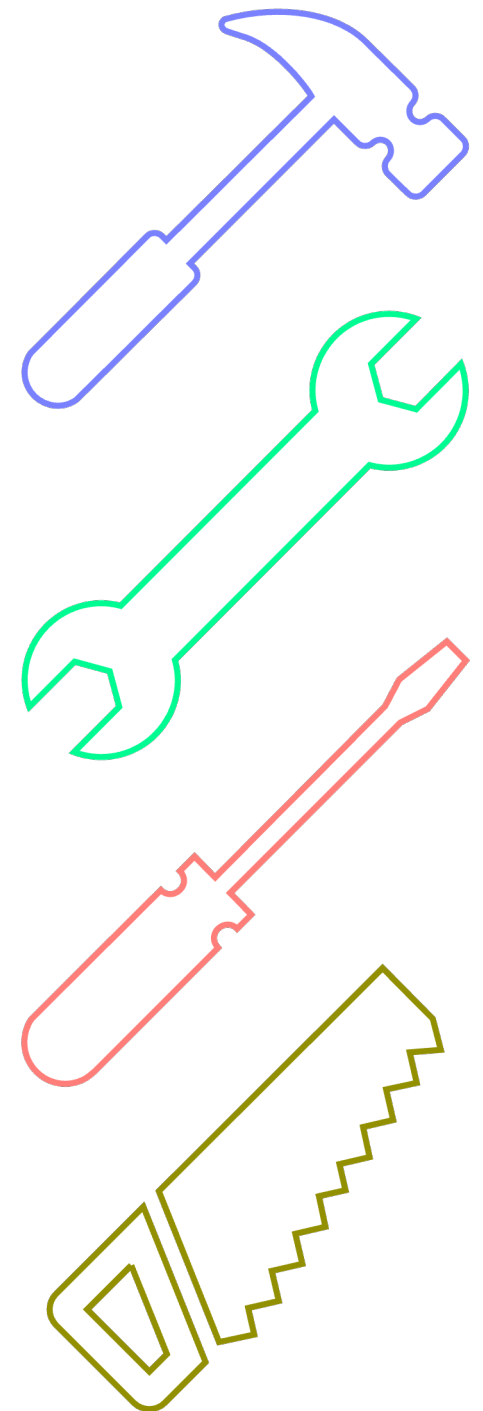
Python as a Research Tool in Preventive Medicine

March 21, 2023

Colby Witherup Wood

Lead Data Scientist

Northwestern IT Research Computing Services



Topics for today

Python at the university

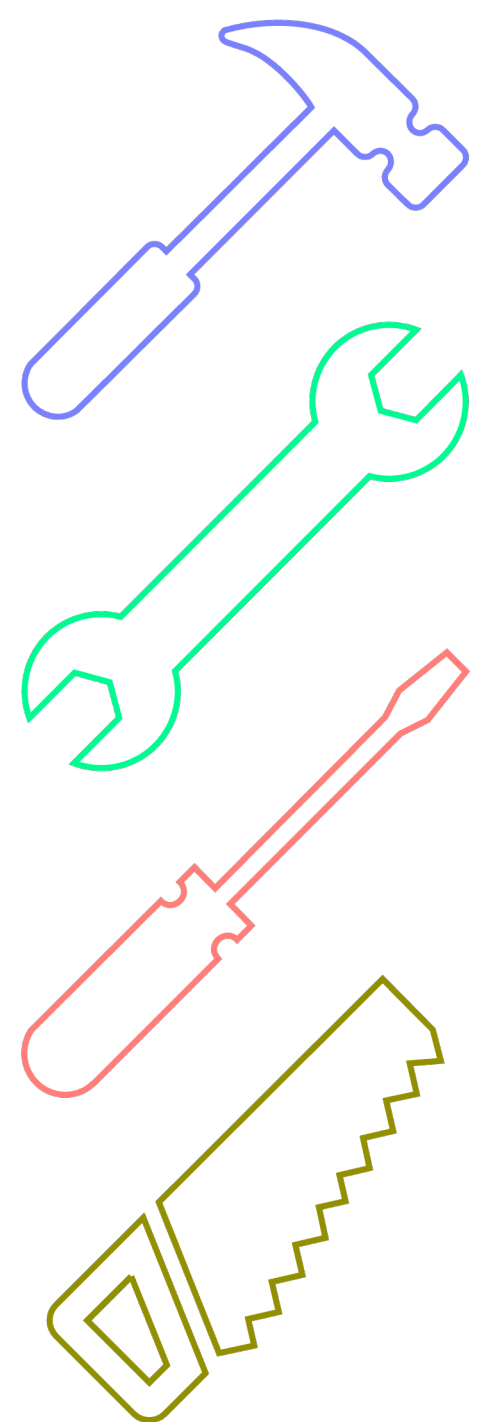
R vs. Python

Python as a research tool:

Where it excels

Real projects

How to learn Python



Research Data Science

1-on-1 Consults
Faculty Projects
Trainings

Our team

2 Data Scientists

.5 Data Visualization Specialist

12 Student Consultants

bit.ly/rcsconsult

Research Data Science

Our team

2 Data Scientists

.5 Data Visualization Specialist

12 Student Consultants

bit.ly/rcsconsult

1-on-1 Consults

- all researchers (undergrad through faculty, all schools and departments)
- about 300 free consults per year
- troubleshooting code, planning out data science projects, giving advice, etc.

Faculty Projects

Trainings

Research Data Science

Our team

2 Data Scientists

.5 Data Visualization Specialist

12 Student Consultants

bit.ly/rcsconsult

1-on-1 Consults

- I see a lot of code
- I get to learn about the interesting research happening at NU

Faculty Projects

Trainings

Research Data Science

Our team

2 Data Scientists

.5 Data Visualization Specialist

12 Student Consultants

bit.ly/rcsconsult

1-on-1 Consults

- ~50% in R, ~25% in Python, and ~25% in other languages or tools
- Python: 34% Weinberg, 24% McCormick, 16% Feinberg
- R: 40% Weinberg, 22% Feinberg, 14% School of Communication

Faculty Projects

Trainings

Research Data Science

Our team

2 Data Scientists

.5 Data Visualization Specialist

12 Student Consultants

bit.ly/rcsconsult

1-on-1 Consults

Faculty Projects

- Longer projects writing code to create research products
- Primary to the research (data collection, algorithms, analyses, machine learning models, etc.)
- Or secondary (data pipelines, automation, databases, visualizations or apps for the public, etc.)

Trainings

Research Data Science

Our team

2 Data Scientists

.5 Data Visualization Specialist

12 Student Consultants

bit.ly/rcsconsult

1-on-1 Consults

Faculty Projects

Trainings

- R and Python bootcamps
- More advanced R and Python
- Git, command line, data viz tools, SQL, etc.
- BYOD groups for researchers

Research Data Science

Our team

2 Data Scientists

.5 Data Visualization Specialist

12 Student Consultants

bit.ly/rcsconsult

1-on-1 Consults

Faculty Projects

Trainings

- Python Fundamentals

Registrants: 1,676

- R Fundamentals Registrants:

1,328

Research Data Science

Our team

2 Data Scientists

.5 Data Visualization Specialist

12 Student Consultants

bit.ly/rcsconsult

Primary Coding Languages of Student Applicants for 2023-24

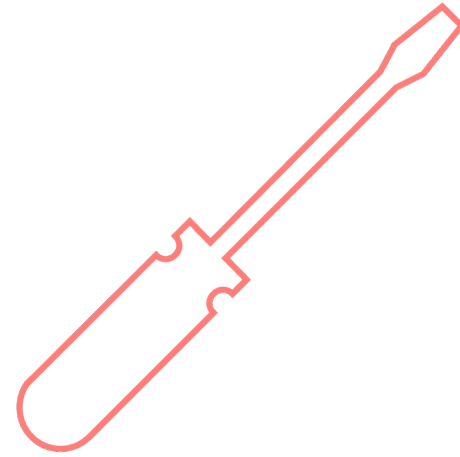
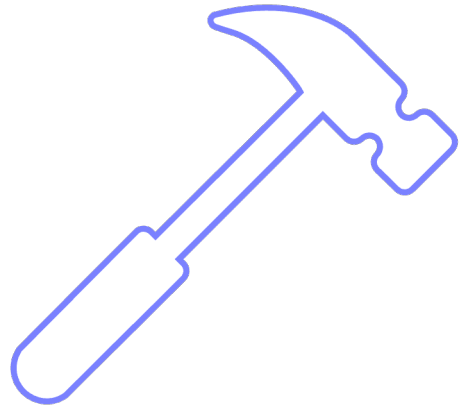
60 Grad Students

- 35 Python
- 29 R

39 Undergrads

- 31 Python
- 9 R

Should I learn R or Python?



Should I learn R or Python?

You should think of R and Python as two different tools

**People who code in only one can do
*almost everything***

People who code in only one can do *almost everything*

Both R and Python are:

- supported
- evolving
- open access
- available on the cloud
- available on Quest analytics nodes

People who code in both use each language for different things

Christina Maimone uses R for:

- Visualization
- Exploratory data analysis, especially with groups
- Simple exploratory text analysis (word counts, bag of words dictionaries)
- Statistical analysis
- Easy, quick web applications with Shiny
- Producing analysis reports with R Markdown

Christina Maimone uses Python for:

- Large scale file processing
- Extracting text data from various formats (pdf, html)
- Part of speech text tagging, text analysis models, topic modeling
- Machine learning models, esp. if need to search/tune parameters
- Robust or multi-page web applications
- Applications with extensive interaction with databases
- Building command line utilities
- Non-statistical simulations
- Hierarchical data or data not in data frames
- Web scraping
- Collecting large amounts of data from APIs

Christina Maimone uses R for:

- Visualization
- Exploratory data analysis, especially with groups
- Simple exploratory text analysis (word counts, bag of words dictionaries)
- Statistical analysis
- Easy, quick web applications with Shiny
- Producing analysis reports with R Markdown

Christina Maimone uses Python for:

- Large scale file processing
- Extracting text data from various formats (pdf, html)
- Part of speech text tagging, text analysis models, topic modeling
- Machine learning models, esp. if need to search/tune parameters
- Robust or multi-page web applications
- Applications with extensive interaction with databases
- Building command line utilities
- Non-statistical simulations
- Hierarchical data or data not in data frames
- Web scraping
- Collecting large amounts of data from APIs

Why researchers who know R want to learn Python:

1. They want to use Python tutorials or scripts written by others
2. They are leaving academia, hoping to get a high-paid job as a data scientist
3. They use R, but they don't love it or feel like they just don't get it
4. They have a task that can't be completed in R

Why researchers who know R want to learn Python:

1. They want to use Python tutorials or scripts written by others
 2. They are leaving academia, hoping to get a high-paid job as a data scientist
 3. They use R, but they don't love it or feel like they just don't get it
 4. They have a task that can't be completed in R
1. Learn minimal Python

Why researchers who know R want to learn Python:

1. They want to use Python tutorials or scripts written by others
 2. They are leaving academia, hoping to get a high-paid job as a data scientist
 3. They use R, but they don't love it or feel like they just don't get it
 4. They have a task that can't be completed in R
1. Learn minimal Python
 2. I have a lot of thoughts...

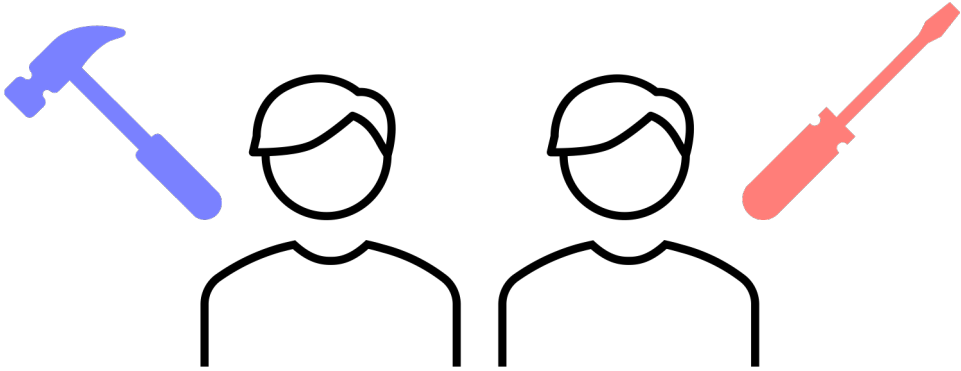
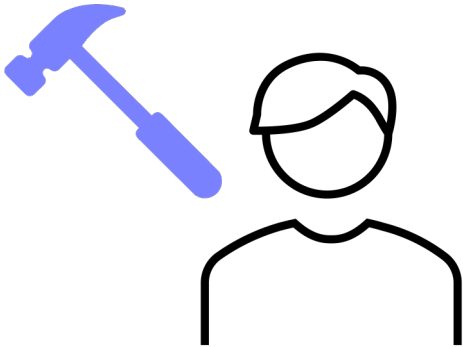
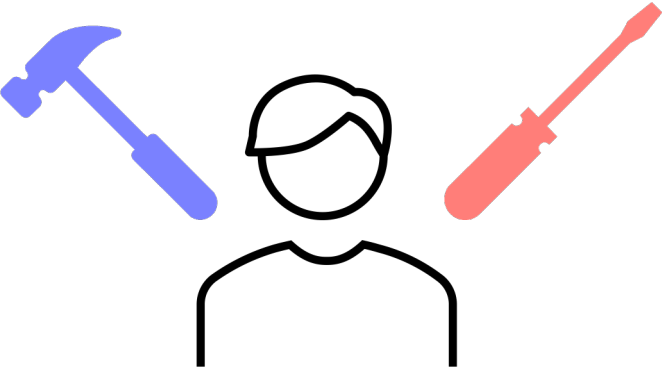
Why researchers who know R want to learn Python:

1. They want to use Python tutorials or scripts written by others
 2. They are leaving academia, hoping to get a high-paid job as a data scientist
 3. They use R, but they don't love it or feel like they just don't get it
 4. They have a task that can't be completed in R
1. Learn minimal Python
 2. I have a lot of thoughts...
 3. Try learning Python

Why researchers who know R want to learn Python:

1. They want to use Python tutorials or scripts written by others
 2. They are leaving academia, hoping to get a high-paid job as a data scientist
 3. They use R, but they don't love it or feel like they just don't get it
 4. They have a task that can't be completed in R
1. Learn minimal Python
 2. I have a lot of thoughts...
 3. Try learning Python
 4. **Learn Python or hire someone**

There are multiple ways to get both the R (or SAS) tool AND the Python tool on your team



Python

Python

Python and R are fundamentally very different languages, though both are easy to read.

R and SAS were written for statistical analysis.

Python was written as a general-purpose language for writing software.

Python

R is usually used as a functional language – tidy up your data, then apply functions to it; store data in rows and columns

Python is usually used as a structured (procedural) – loop through your data and do different things depending on conditions met or not met – or object-oriented – it sometimes makes more sense to store data in a hierarchical shape instead of rows and columns – or functional language

Python excels at:


- Customization
- Automation and pipelines (works well with others, especially command line programs)
- Custom algorithms
- Reproducibility (conda package manager)
- Scale (runs faster, easier to run in background, easier to loop)
- Coding alongside visualizations and comments (Jupyter Notebooks)
- Anything approaching software development

Python project examples

P-RIFTEHR: Family Trees from EHRs

JOURNAL ARTICLE

Using electronic health record data to link families: an illustrative example using intergenerational patterns of obesity

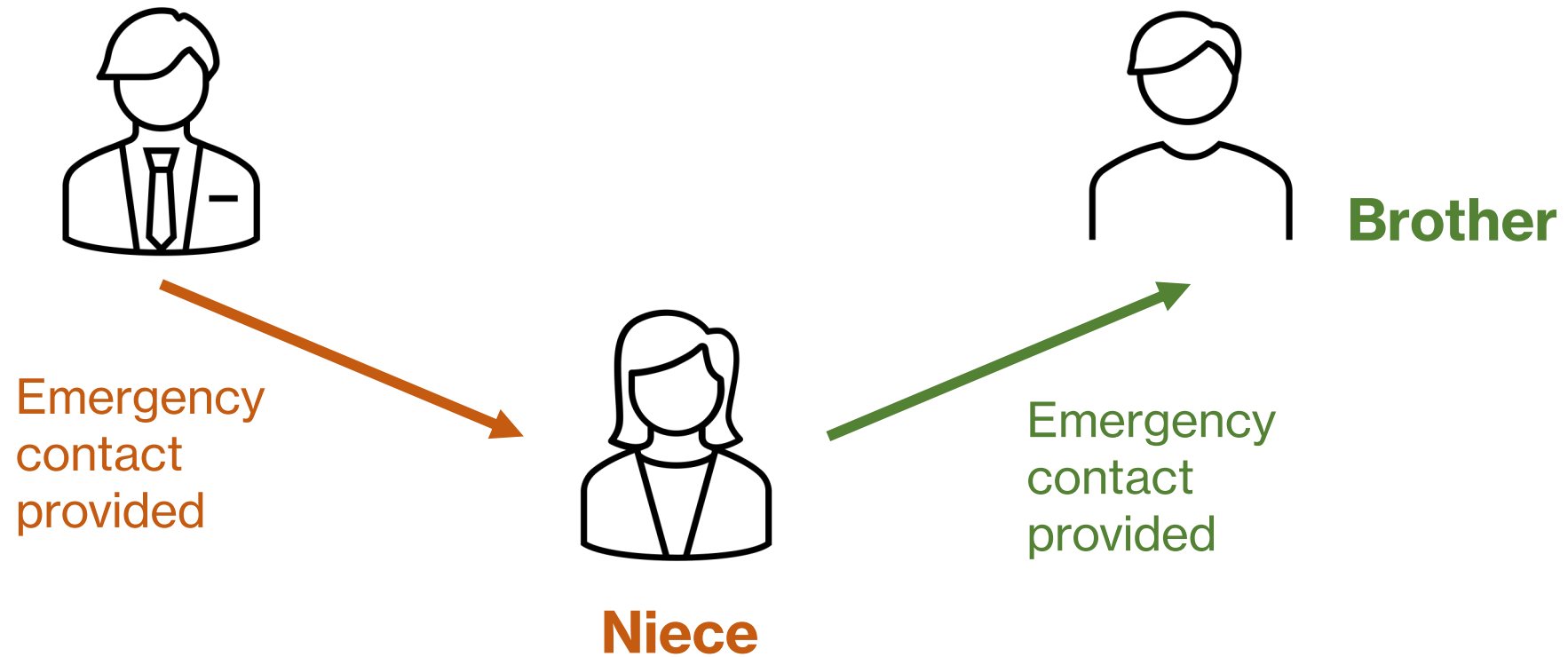
Amy E Krefman , Farhad Ghamsari, Daniel R Turner, Alice Lu, Martin Borsje, Colby Witherup Wood, Lucia C Petito, Fernanda C G Polubriaginof, Daniel Schneider, Faraz Ahmad, Norrina B Allen

Journal of the American Medical Informatics Association, ocad028,

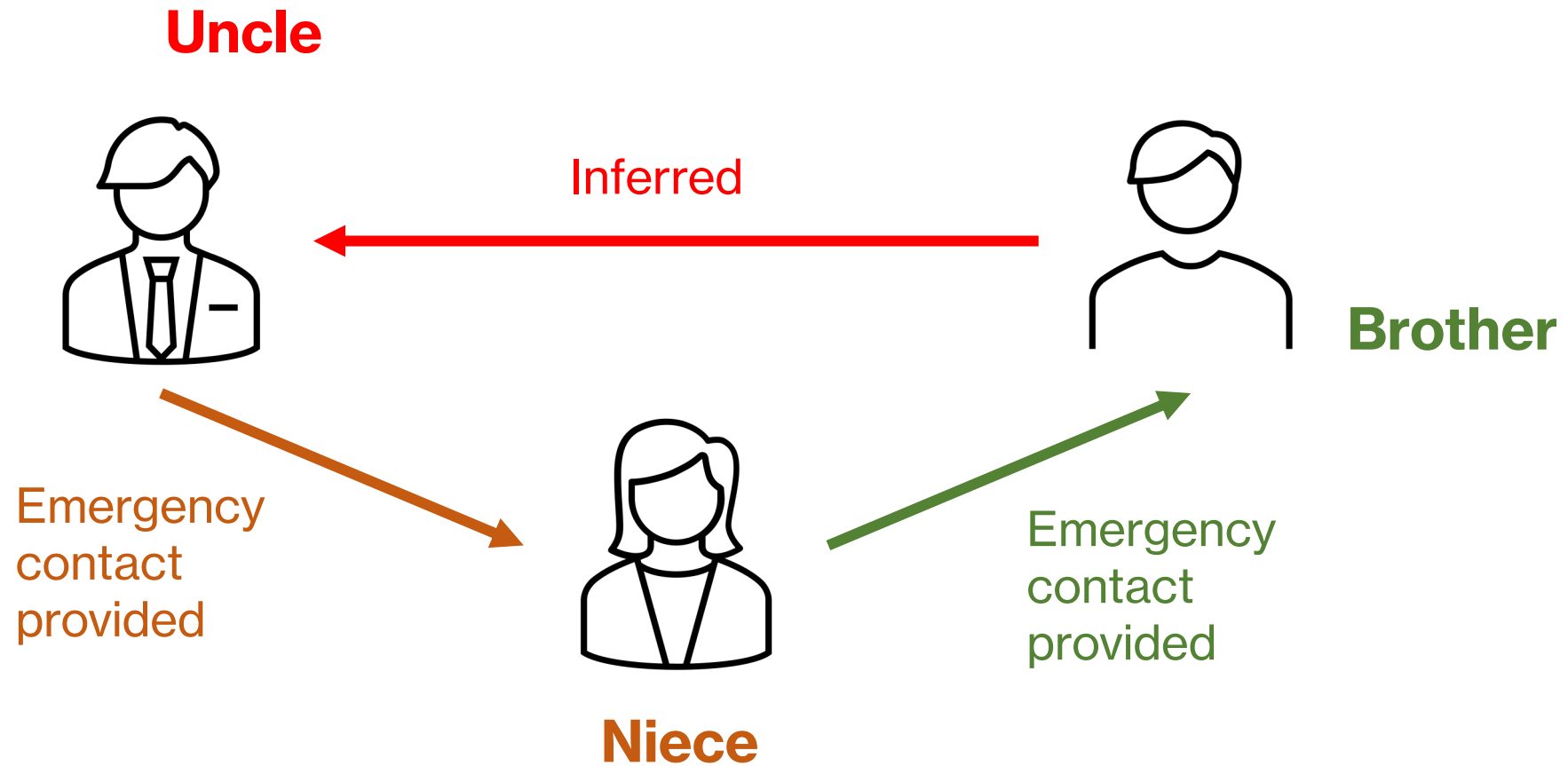
<https://doi.org/10.1093/jamia/ocad028>

Published: 28 February 2023 **Article history** ▼

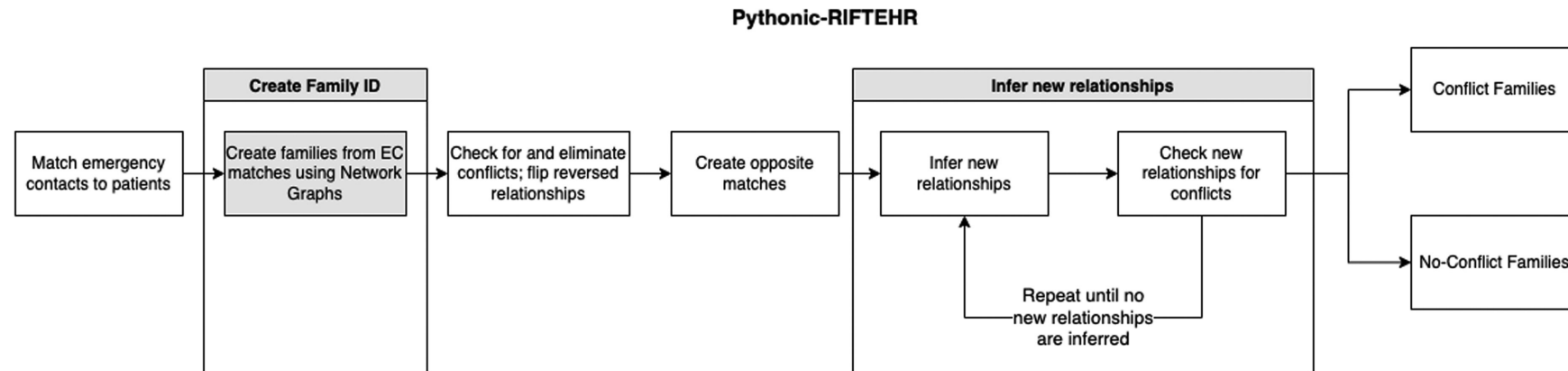
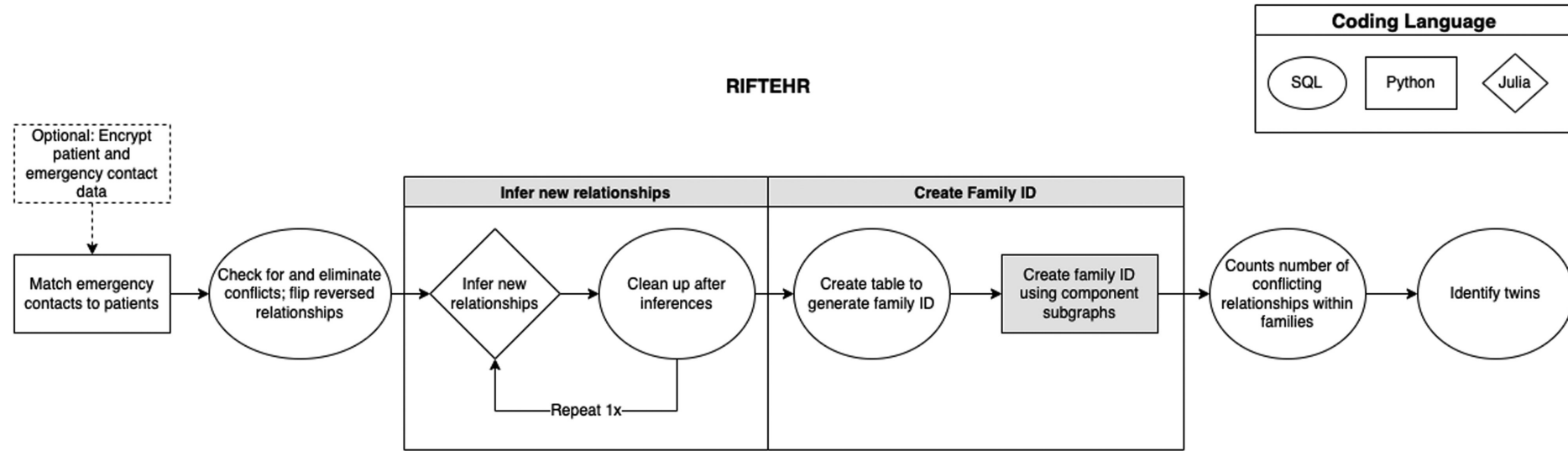
P-RIFTEHR: Family Trees from EHRs



P-RIFTEHR: Family Trees from EHRs



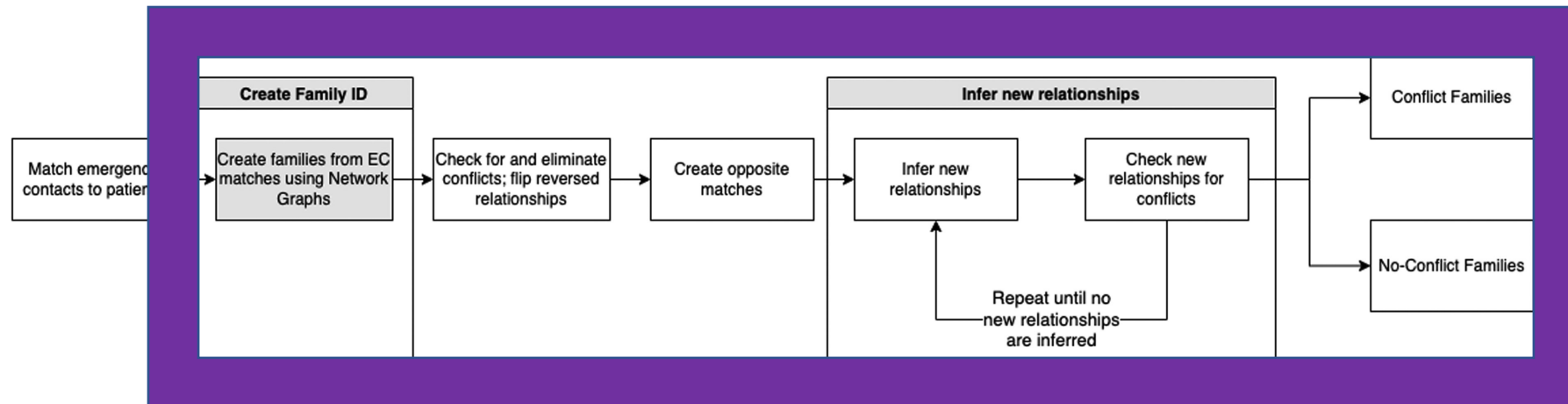
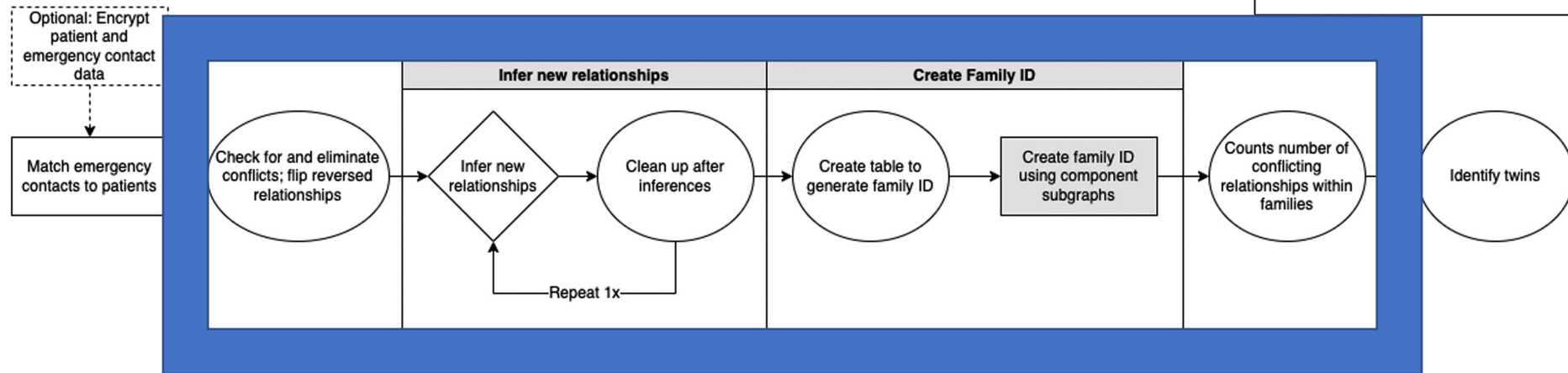
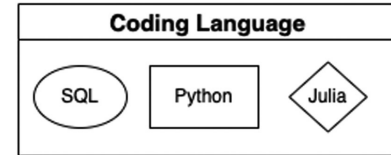
P-RIFTEHR: Family Trees from EHRs



P-RIFTEHR: Family Trees from EHRs

Inference and checks

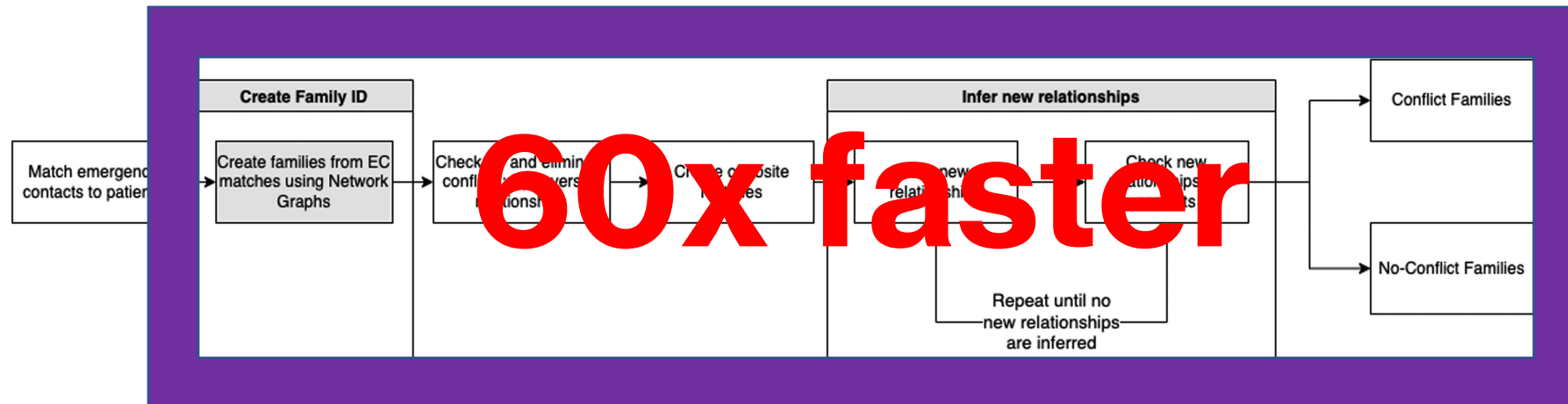
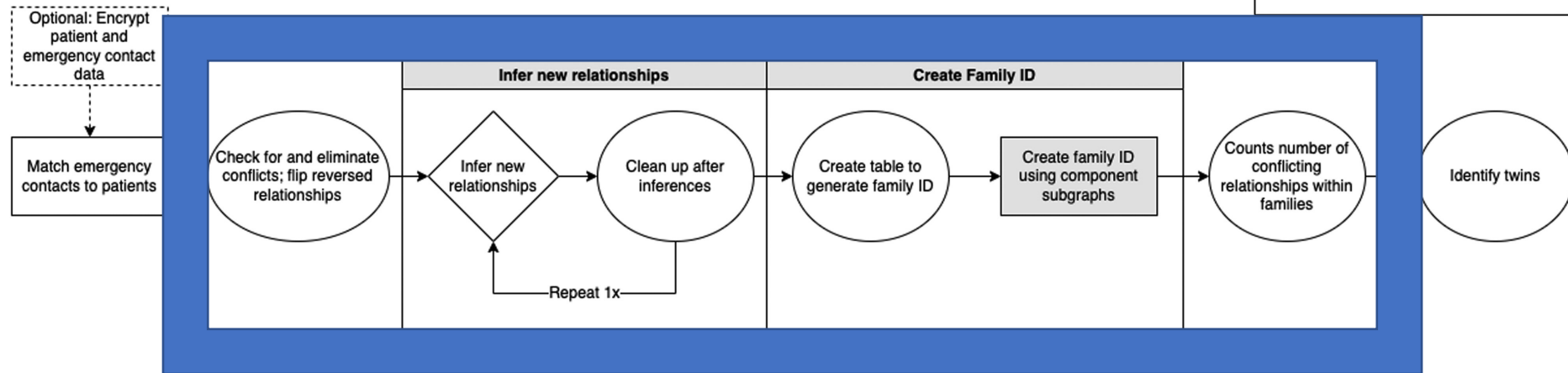
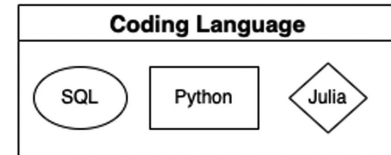
RIFTEHR



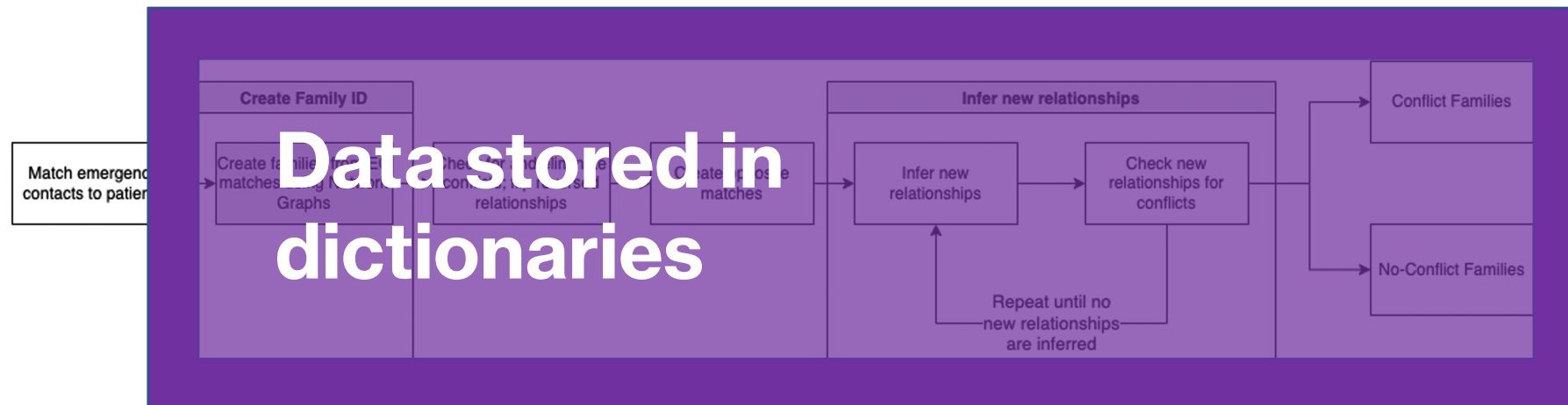
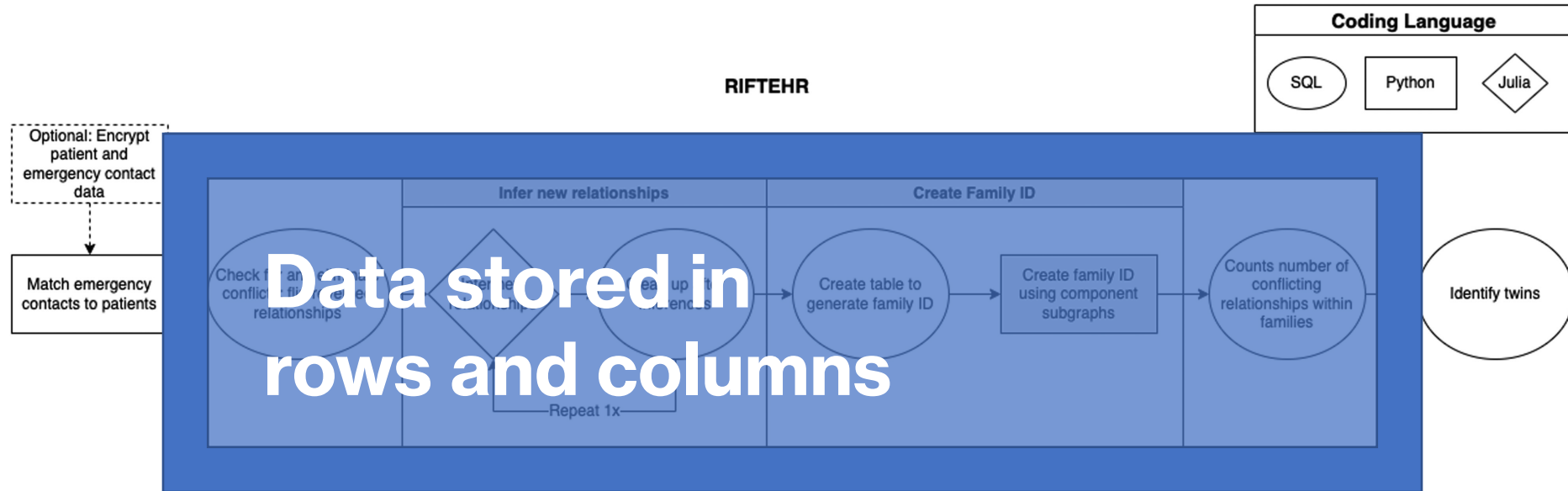
P-RIFTEHR: Family Trees from EHRs

Inference and checks

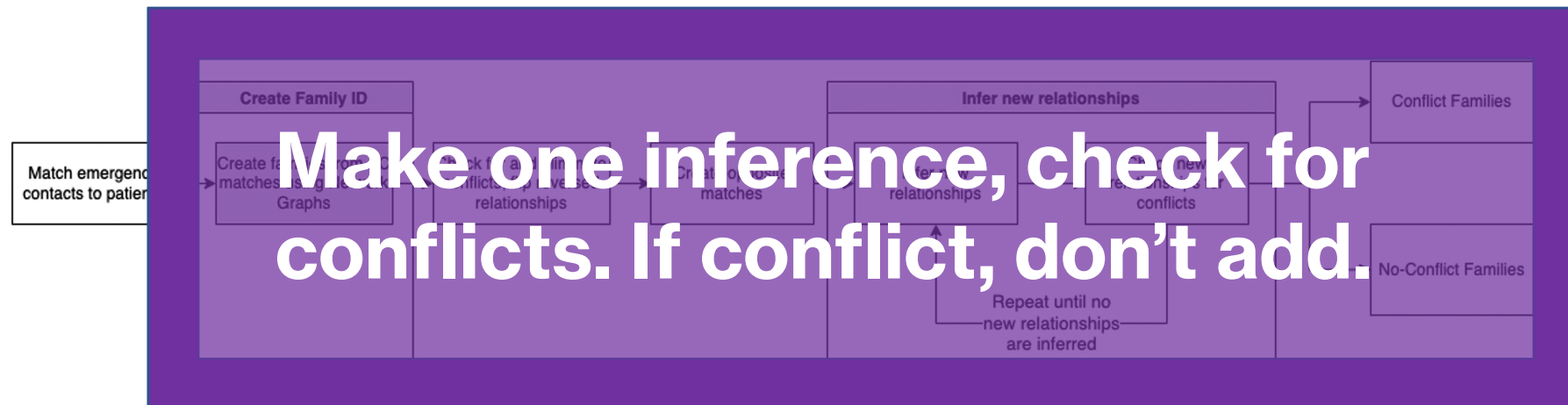
RIFTEHR



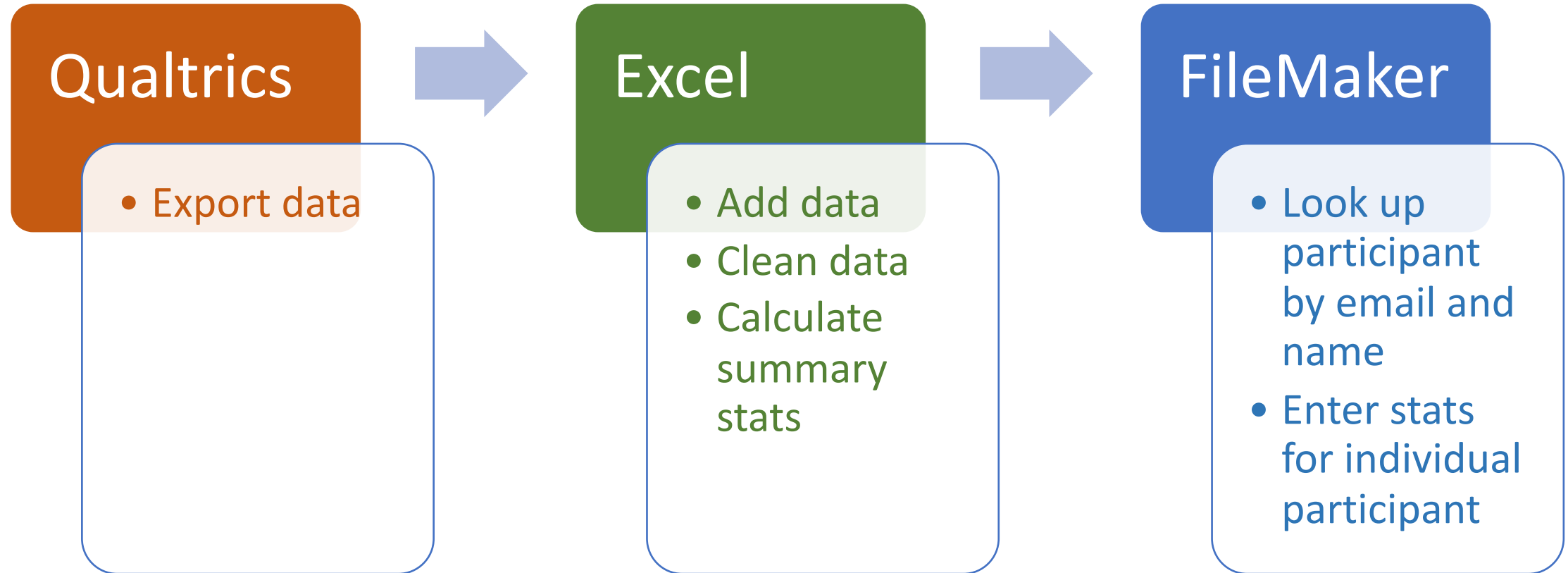
P-RIFTEHR: Family Trees from EHRs



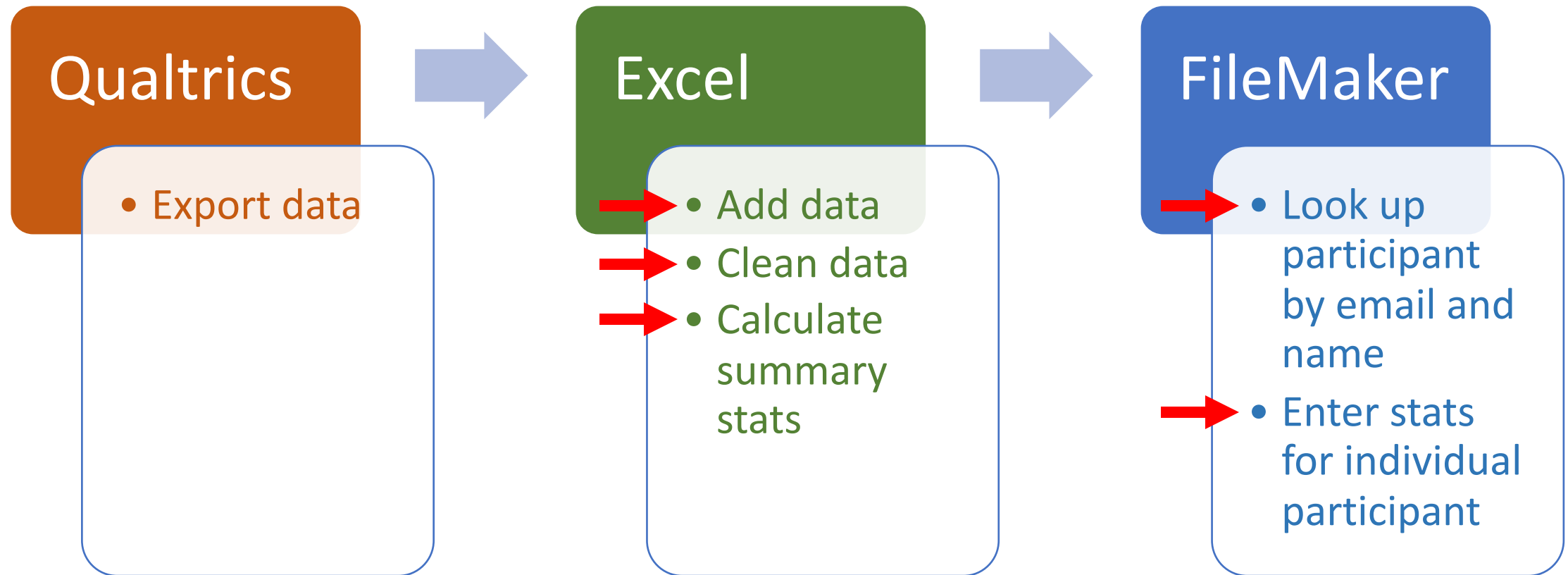
P-RIFTEHR: Family Trees from EHRs



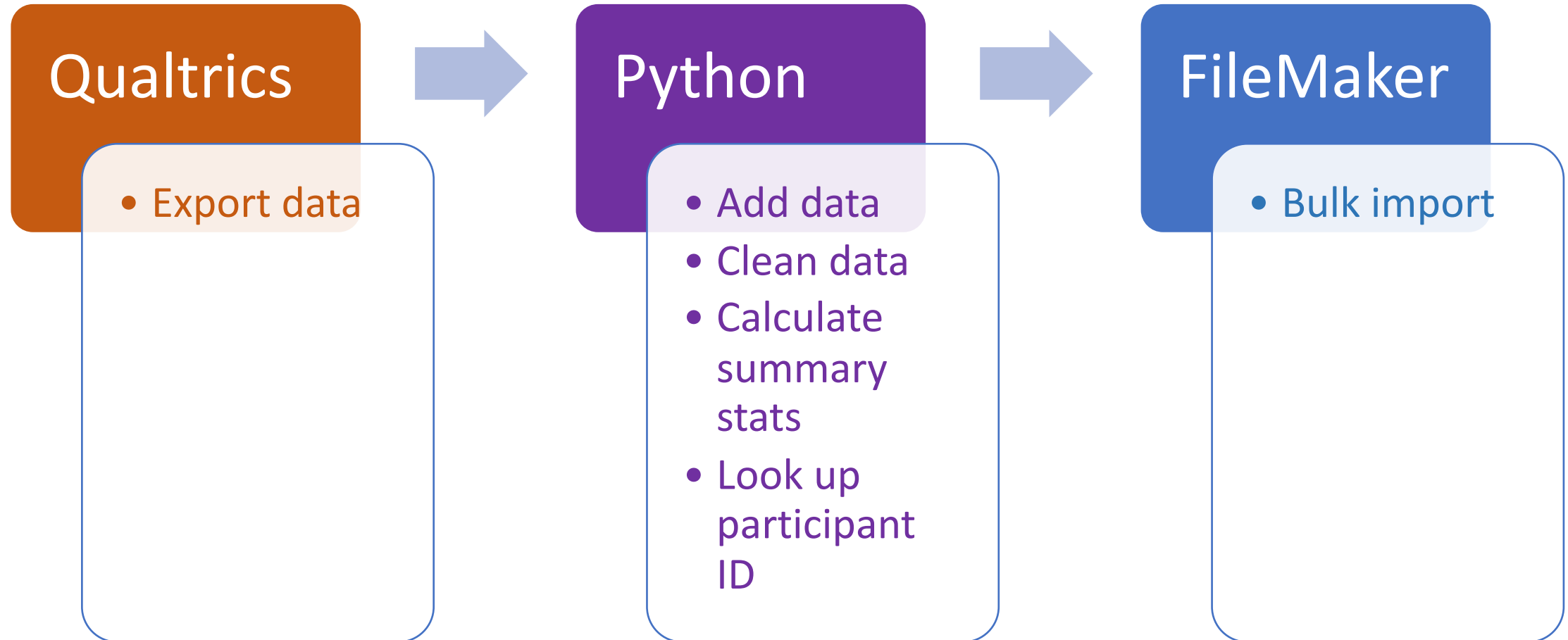
Lab workflow pipeline and automation

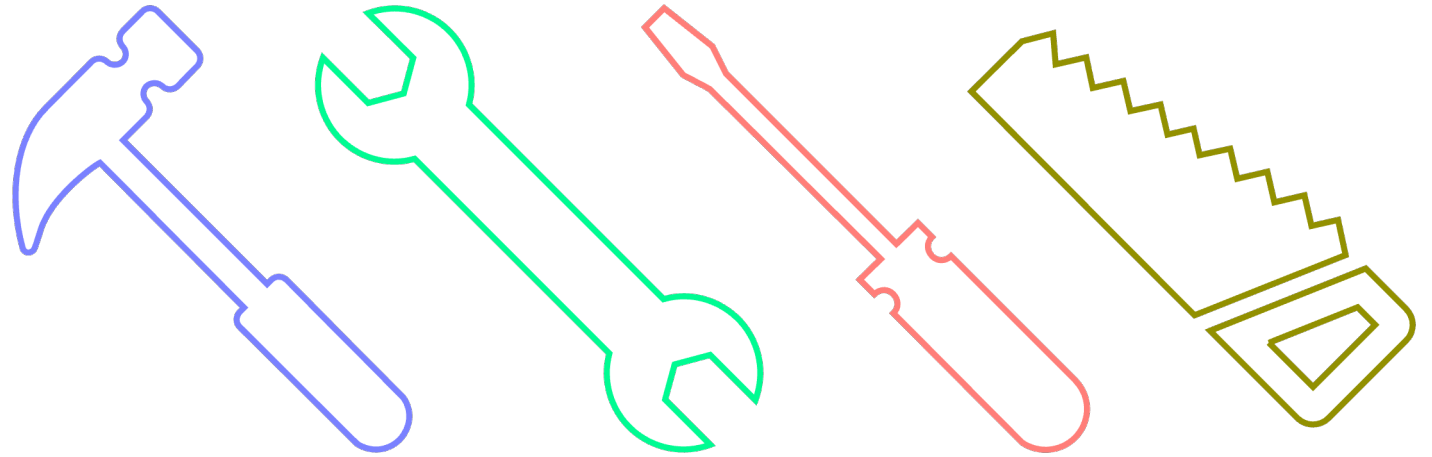


Lab workflow pipeline and automation



Lab workflow pipeline and automation





Predicting health outcomes with Machine Learning

New models for prediction, clustering, and feature selection are developed in Python first.

Learning Python

- Chicago campus in-person Python Fundamentals bootcamp: June 28, 29, 30 with me
- Evanston version in early July with me
- Evanston in-person pandas (rows and columns), data viz with matplotlib, Python for Automation, more, plus remote workshops in Machine Learning, Intermediate topics
- Work through my notebooks on your own (links to run them in the cloud without installing anything):
https://github.com/aGitHasNoName/pythonBootcamp_3Day

Learning Python

- Many YouTube videos will teach you Python in ~4 hours. These are pretty good, but they will have you download Python and a Python IDE in many complicated ways.
- We recommend you download the Anaconda distribution of Python, which comes with multiple IDEs (PyCharm, Spyder, and my favorite, Jupyter Lab)
- Once you've installed Anaconda, you can catch up on the videos after they go through installation.

Questions?

colby.witherup@northwestern.edu

[bit.ly/rcs consult](https://bit.ly/rcs_consult)

