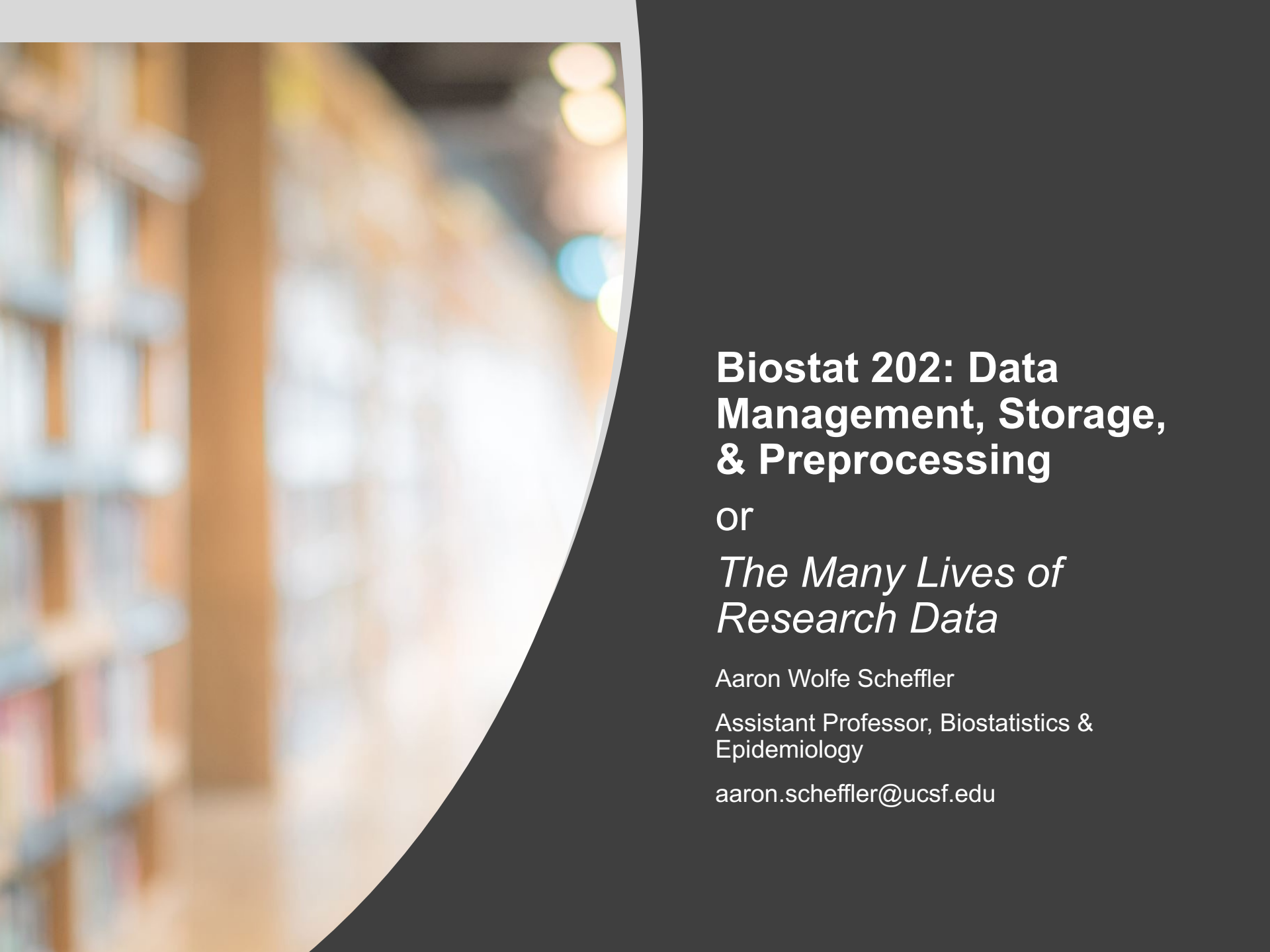




Biostat 202: Data Management, Storage, & Preprocessing

or

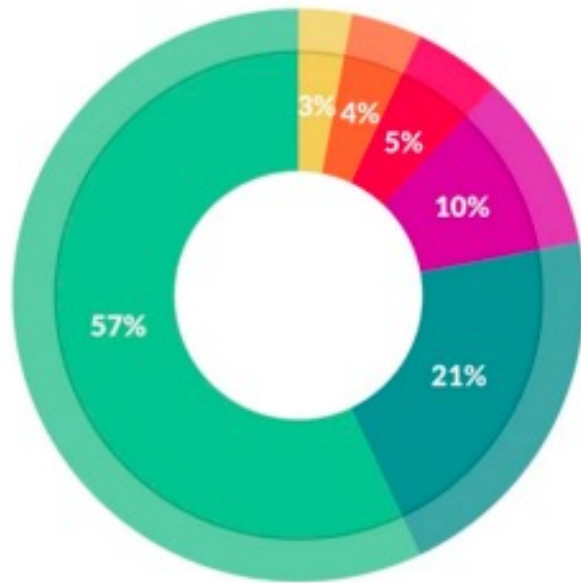
*The Many Lives of
Research Data*

Aaron Wolfe Scheffler

Assistant Professor, Biostatistics &
Epidemiology

aaron.scheffler@ucsf.edu

Where do we spend time?

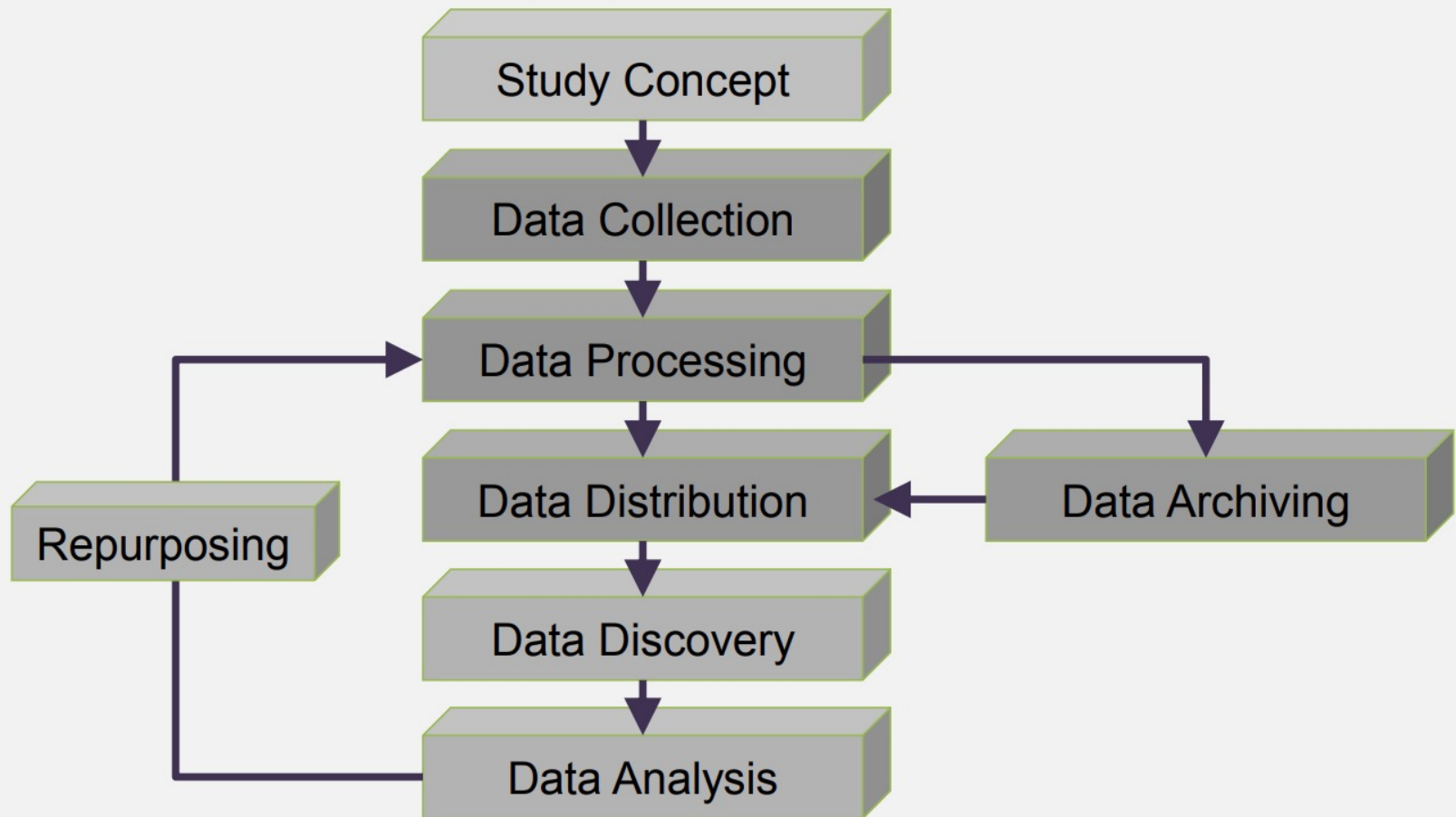


What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Image Credits: whatsthebigdata.com

Research data life cycle – dynamic!



Why discuss data management and storage?

- Save yourself time in the future
- Ensure data integrity
- Retention requirements
- Promote and protect yourself!
- Support open research
- Protect and secure your data

Why discuss data management and storage?

- Save yourself time in the future
 - Anticipate possible bottlenecks or roadblocks
- Ensure data integrity
 - How would someone else use your data?
- Retention requirements
 - Often contractual obligations about how long and where to store data
- Promote and protect yourself!
 - Protect yourself against questions of provenance
- Support open research
 - Increase your research profile, advance science
- Protect and secure your data
 - Don't leave your data on the street!

Ten Simple Rules for Digital Data Storage

Edmund M. Hart , Pauline Barmby, David LeBauer, François Michonneau, Sarah Mount, Patrick Mulrooney, Timothée Poisot, Kara H. Woo, Naupaka B. Zimmerman, Jeffrey W. Hollister

1. Anticipate how your data will be used
2. Know your use case
3. Keep Raw Data Raw
4. Store Data in Open Formats
5. Data should be structured for analysis
6. Data should be uniquely identifiable
7. Link Relevant Metadata
8. Adopt Proper Privacy Protocols
9. Have a systematic Backup Scheme
10. The location and method of storage depend quantity

#Quick Takes

UCSF Pays \$1 Million Ransom to Hackers

By [Lindsay McKenzie](#) // July 2, 2020

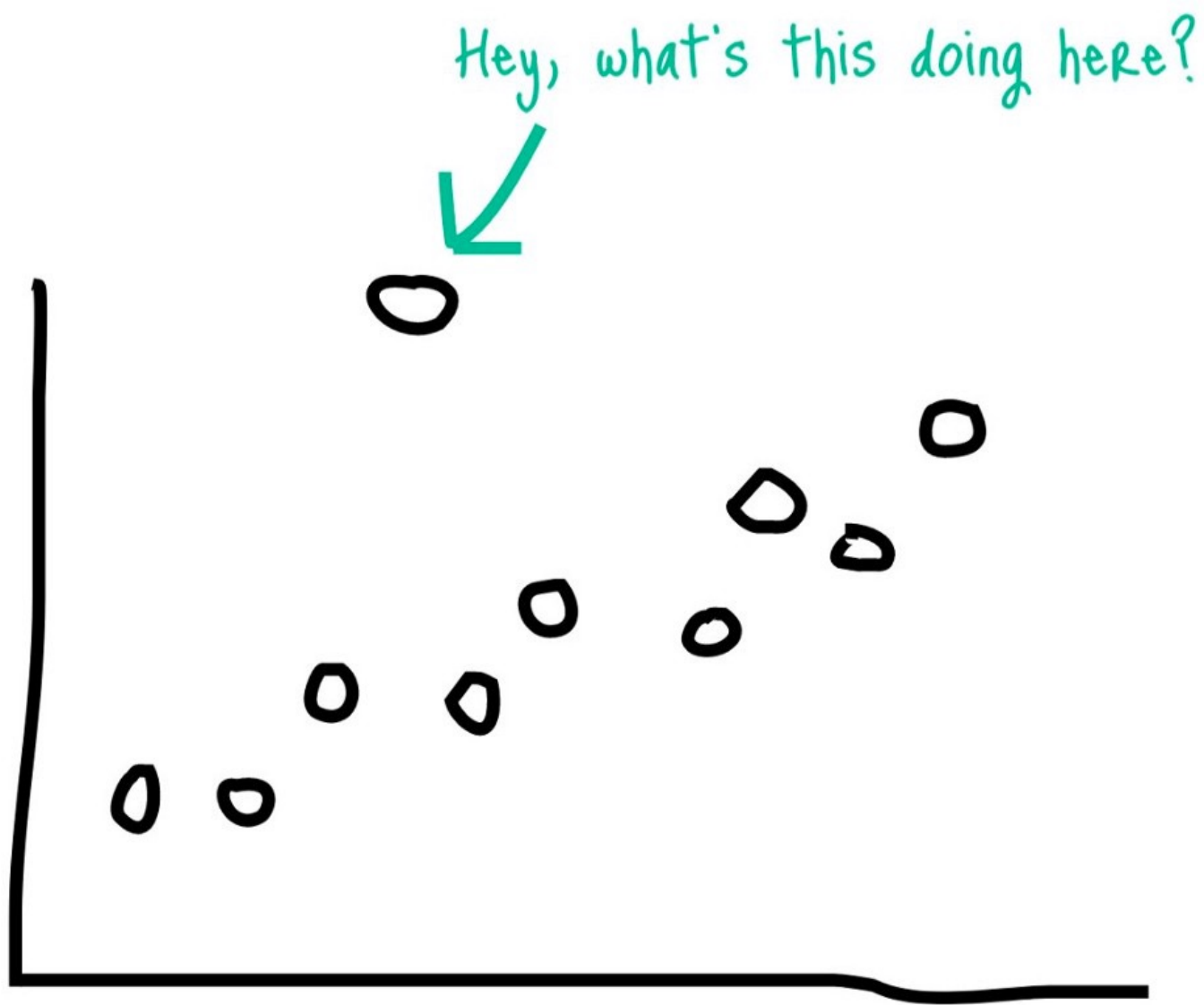
The University of California, San Francisco, paid a ransom of \$1.14 million to hackers who encrypted and threatened to publish sensitive data stolen from the institution's School of Medicine.

UCSF is one of [three universities](#) recently targeted using ransom software known as Netwalker. Michigan State University and Columbia College Chicago were also affected. Michigan State announced last month that it decided [not to pay](#) the ransom out of concern that payment to the criminals would encourage future attacks.

Through an anonymous tip, a [BBC News](#) reporter was able to join the live chat room where UCSF negotiated the ransom. The hackers initially demanded \$3 million. A UCSF representative said the coronavirus pandemic was "financially devastating" for the university and made a counteroffer of \$780,000, which was refused. They settled on a payment of 116.4 Bitcoin, worth approximately \$1.14 million.

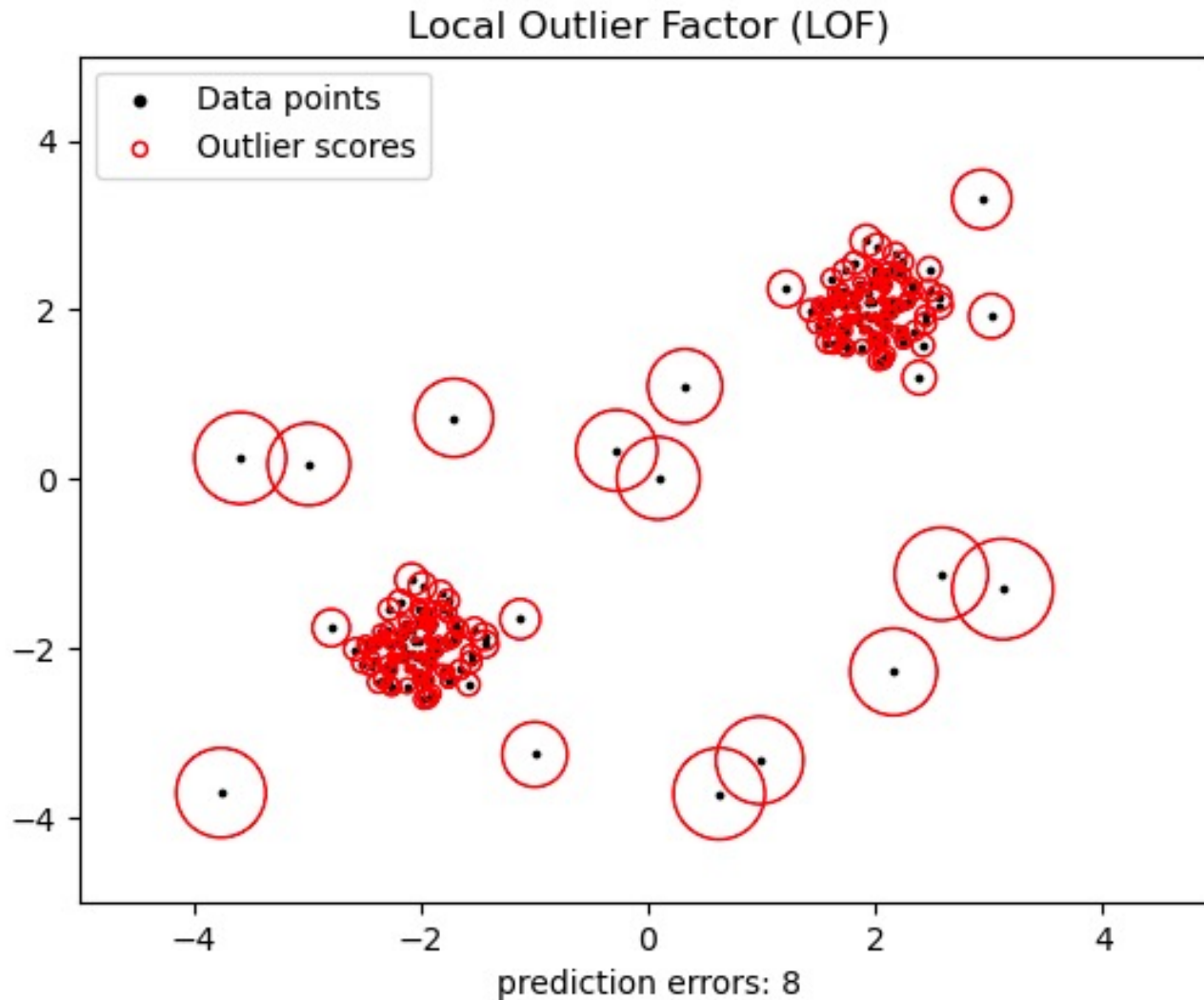
Orange for data management and cleaning

- Feature construction (lab)
- Merging & concatenating (lab)
- Impute missing data (class demo/lab)
- Detect outliers (class demo)



Outliers

Orange – Local Outlier Factor



Orange – outlier detection

The screenshot displays the Orange3 software interface with a workflow titled "Lecture 3 Outliers Demo.ows". The interface is divided into a left sidebar and a main workspace.

Left Sidebar:

- Data:** Contains a "Data" widget icon.
- Visualize:** Contains various visualization widget icons, including Tree Viewer, Box Plot, Distribut..., Scatter Plot, Line Plot, Sieve Diagram, Mosaic Display, FreeViz, Linear Projection, Radviz, Heat Map, and Venn Diagram.
- Scatter Plot:** A detailed description of the Scatter Plot widget: "Interactive scatter plot visualization with intelligent data visualization enhancements." followed by a [more...](#) link.
- Bottom Bar:** Contains icons for information (i), zoom (x), text (T), pan (arrow), stop (II), and help (?).

Main Workspace:

The workflow consists of the following widgets and connections:

- File:** The starting widget, represented by a document icon.
- Data:** A connection line labeled "Data" from the File widget to the Outliers widget.
- Outliers:** A widget represented by an orange circle with black dots inside.
- Data:** A connection line labeled "Data" from the Outliers widget to the Feature Statistics widget.
- Feature Statistics:** A widget represented by a bar chart icon.
- Data:** A curved connection line labeled "Data" from the Outliers widget to the Scatter Plot widget.
- Scatter Plot:** A widget represented by a red circle with a scatter plot icon inside.

Outlier detection – honest truth

- There is no simple solution
- Best practice is to evaluate multiple considerations
 - Plausibility
 - Visual inspection
 - Automated algorithms to flag points
- If uncertain, consider models that are robust to extreme values (more on this later)

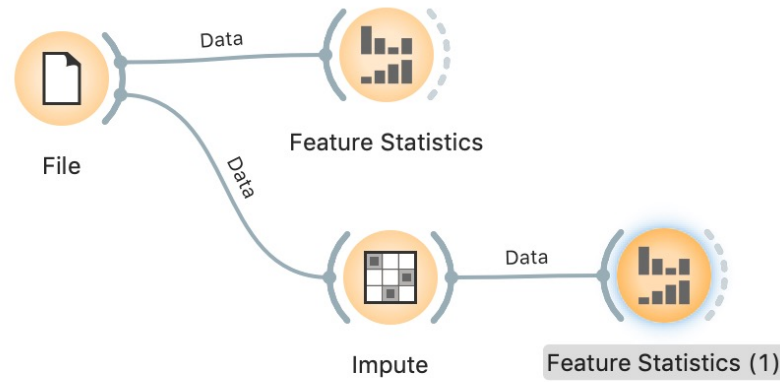


Missing data

Missing data imputation

- Use the mean to fill in the missing value (okay)
- Use the mean for all samples belonging to the same class to fill in the missing value: (smart)
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree or regression (smarter)
- Multiple imputation techniques examining sensitivity of estimates to imputation (smartest?)
- What can Orange do for you? (*In class!*)

Orange – missing data imputation



Questions?
Use the **chat** to let
me know, I will
attend in that order.

Next up, lab

After a 15 minute break we will return for lab.

Please visit CLE to download Assignment 2 and the associated data sets.