



Biostat 202: Data Management, Storage, & Cleaning

or

*The Many Lives of
Research Data*

Aaron Wolfe Scheffler

Assistant Professor, Biostatistics &
Epidemiology

aaron.scheffler@ucsf.edu

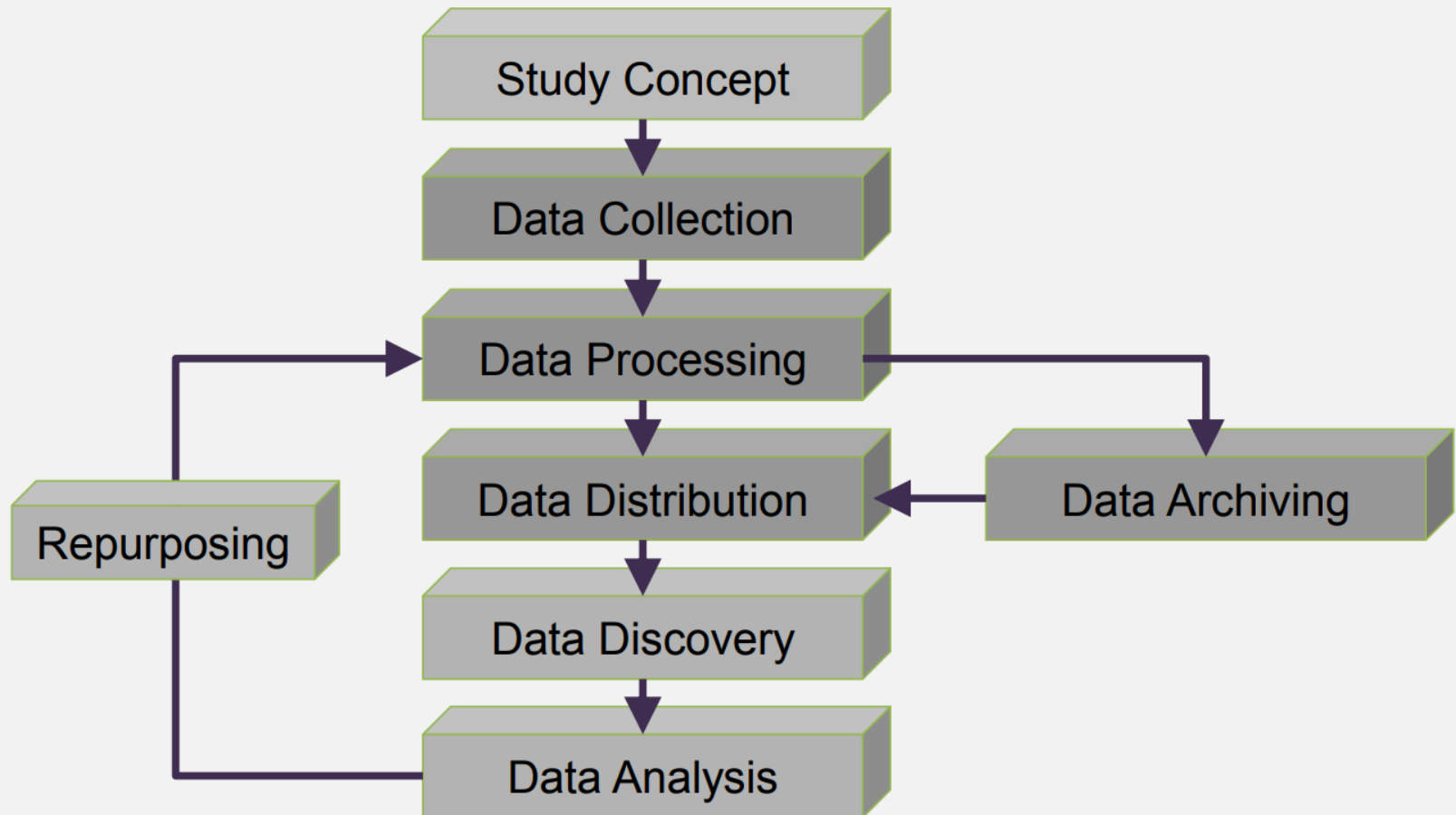
Last time – getting data

1. It helps to know the format and layout of a dataset you want to analyze.
2. Paying attention to data formatting issues can save LOTS of time later.
3. Defining variable roles is important in Orange Data Mining.
4. Access to rich data is easy.
5. Finding your way around can be hard.
6. Look for tools (tutorials, browsers),
7. Or get professional help (Library, ARS and CTSI).

Outline for today

- Why discuss data management and storage?
- Data management
- Data storage
- Data cleaning

Research data life cycle – dynamic!



What might your data look like?

- Survey responses
- Demographics
- Focus group interviews
- Queried EHR data
- Images
- Laboratory measurements
- Computer modeling
- Simulations
- Specimen samples

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



Data management

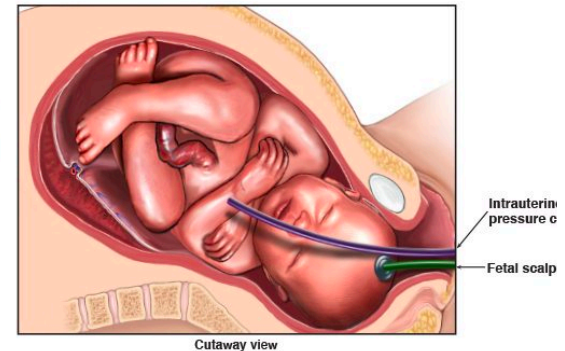
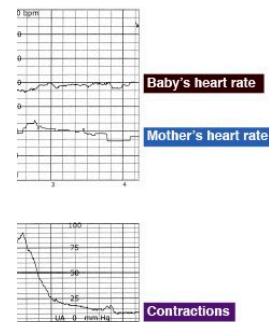
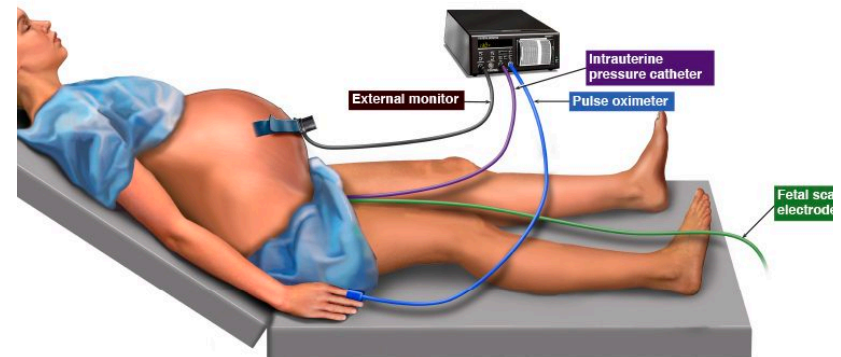
Save yourself time in the future

- Earlier you consider data management issues, the better.
- Storage needs?
- Access needs?
- Software needs?
- Security needs?
 - E.g. protected health information (PHI)

KEEP YOUR DATA NEAR YOUR SOFTWARE!

Case study from my work: Electronic fetal monitoring to predict hypoxic-ischemic encephalopathy

- Storage needs? 100 Gb
for fetal heart rate signals
alone
- Access needs? Multisite
team
- Software needs?
Statistical and machine
learning software required
- Security needs? Secure
server located physically at
Kaiser

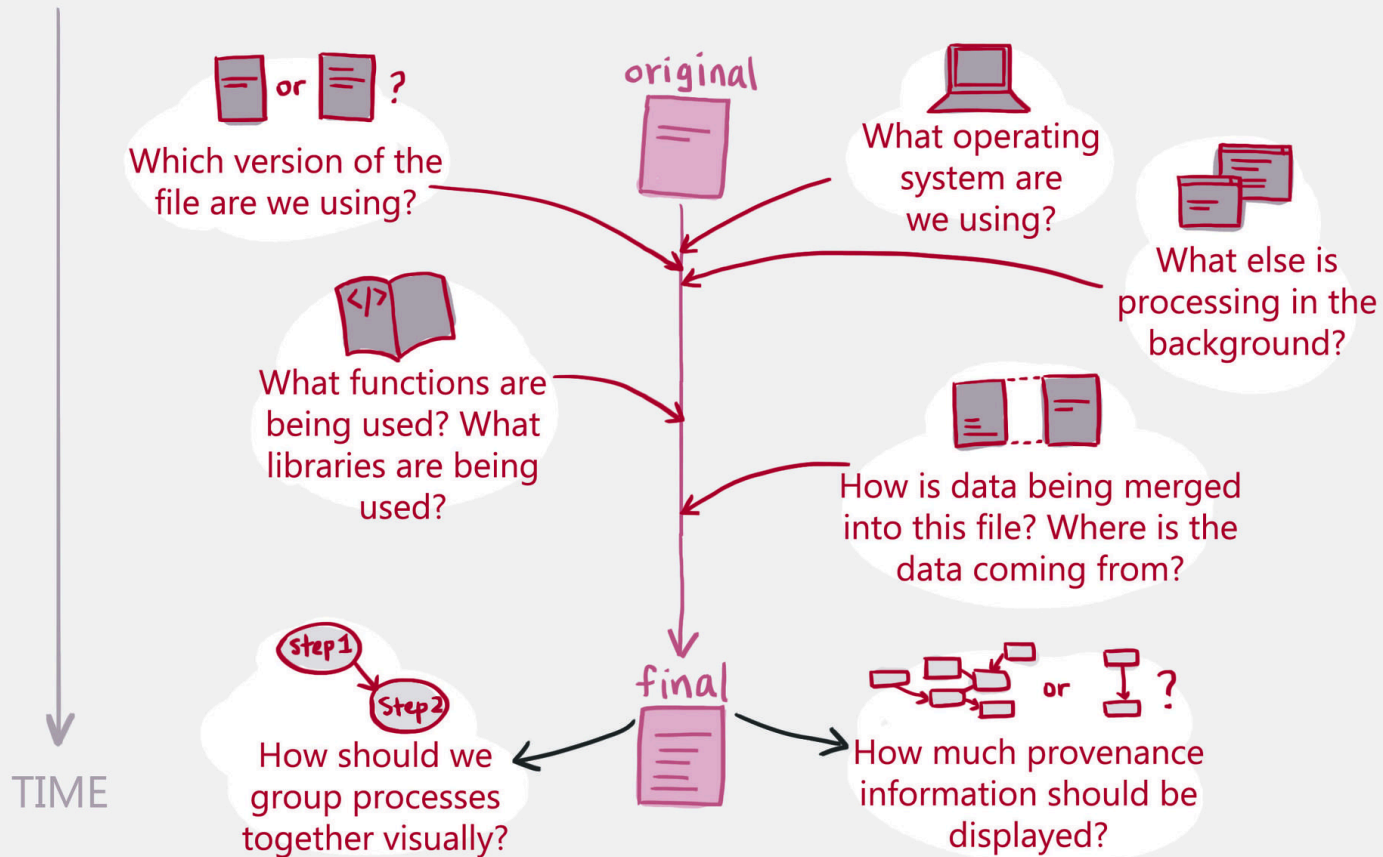


Ensure data integrity

- How would someone else use your data?
- Develop comprehensive data documentation, i.e. metadata
- Metadata should include
 - Provenance/Lineage
 - Methodology
 - Codebooks
- Cornell guide on data documentation
<https://data.research.cornell.edu/content/readme>

Data provenance and processing

Provenance is about capturing the **process** that documents or data were put through.



Example metadata and codebook: NIS

Metadata

Codebook

NIS Database Documentation

The National (Nationwide) Inpatient Sample (NIS) is a large p States, containing data on more than seven million hospital st regional estimates and enables analyses of rare conditions, ur

The following links provide detailed documentation for the NIS restrictions, and resources necessary for researchers working

Description of NIS Database

- [NIS Overview](#)
 - [HCUP Partners in the NIS](#)
- Introduction to the NIS
 - 2017 ([PDF](#) file, 684 KB; [HTML](#))
 - [Prior Years](#)
- [NIS Related Reports](#)
 - [Prior Years](#)
- [Checklist for Working With the NIS](#)

Restrictions on Use

- [HCUP Data Use Agreement Training](#)
- Data Use Agreement for the National (Nationwide) Databases ([PDF](#) file, 85 KB; [HTML](#))
- [Requirements for Publishing with HCUP Data](#)

File Specifications and Load Programs

- [NIS File Specifications](#)
- [Nationwide SAS Load Programs](#)
- [Nationwide SPSS Load Programs](#)
- [Nationwide Stata Load Programs](#)

Data Elements

- [NIS Description of Data Elements](#)
 - [Prior Years](#)
- [NIS Summary Statistics](#)
- Prior to Data Year 2012
 - [Availability of AHA Hospital Identifiers](#)
 - [Why the NIS should not be used to make State-level estimates](#)

Abbreviations: DX, diagnosis; PR, procedure; Q1, di quarter 3; Q4, discharge quarter 4.

HCUP Data Element	Descriptive Title	Years
☐ Select All	Data elements change freq, date generated.	
☐ ADAYWK	Admission day of week	1988-1997
☐ AGE	Age in years at admission	1988-2017
☐ AGE_NEONATE	Neonatal age (first 28 days after birth) indicator	2012-2017
☐ AGEDAY	Age in days (when AGE is less than 1 year)	1988-2011

Variable	Description	Value	Value Description
ADAYWK	Admission day of week	1	Sunday
		2	Monday
		3	Tuesday
		4	Wednesday
		5	Thursday
		6	Friday
		7	Saturday
		.	Missing
		.A	Invalid
		.B	Unavailable from source (coded 1988-1997 data only)

Retention requirements

- Your data will have multiple “demands”: funding source, publisher, and data application.
- Many funding sources (NSF, NIH) require formal data management and sharing plans.
- In addition, you might be required to archive the data for some period of time.

Retention Best Practices: It Depends

- IRB OHRP Requirements: 45 CFR 46 = *3 years* from completion
- NIH: = minimum of *3 years* final report submission.
- HIPAA Requirements: = minimum of *6 years* from date of signed authorization
- FDA Requirements 21 CFR 312.62.c = *2 years* following the date a marketing application is approved for the drug or the investigation is discontinued and FDA is notified.
- VA Requirements: At present records for any research that involves the VA must *be retained indefinitely* per VA federal regulatory requirements.
- Intellectual Property Requirements - Any research data used to support a patent through must be *retained for the life of the patent* in accordance with Intellectual Property Policy.
- Questions of data validity: *you must retain all of the original research data until such questions or allegations have been completely resolved.*

Promote and protect yourself!

- Well documented data allows others to make use of your data collection - increases your research profile.
- Proper knowledge of data management and retention protects you against liability.
- Data integrity also protects you against questions of data validity.

Show me the data!

SHARE



66K



3



Two papers relying on hospital records of COVID-19 patients have been retracted because the company that purportedly analyzed the raw data won't allow their validity to be independently validated. AP PHOTO/MANU FERNANDEZ

Two elite medical journals retract coronavirus papers over data integrity questions

By [Charles Pillar](#), [Kelly Servick](#) | Jun. 4, 2020 , 5:30 PM

Data management at UCSF - personal tools

- DMPTool (free)
 - Helps researchers generate data management plan
 - <https://dmptool.org/>
- UCSF Library Data Science Initiative (free)
 - Consultations, workshops and resources
 - <https://www.library.ucsf.edu/ask-an-expert/data-science/>



**Data
storage**

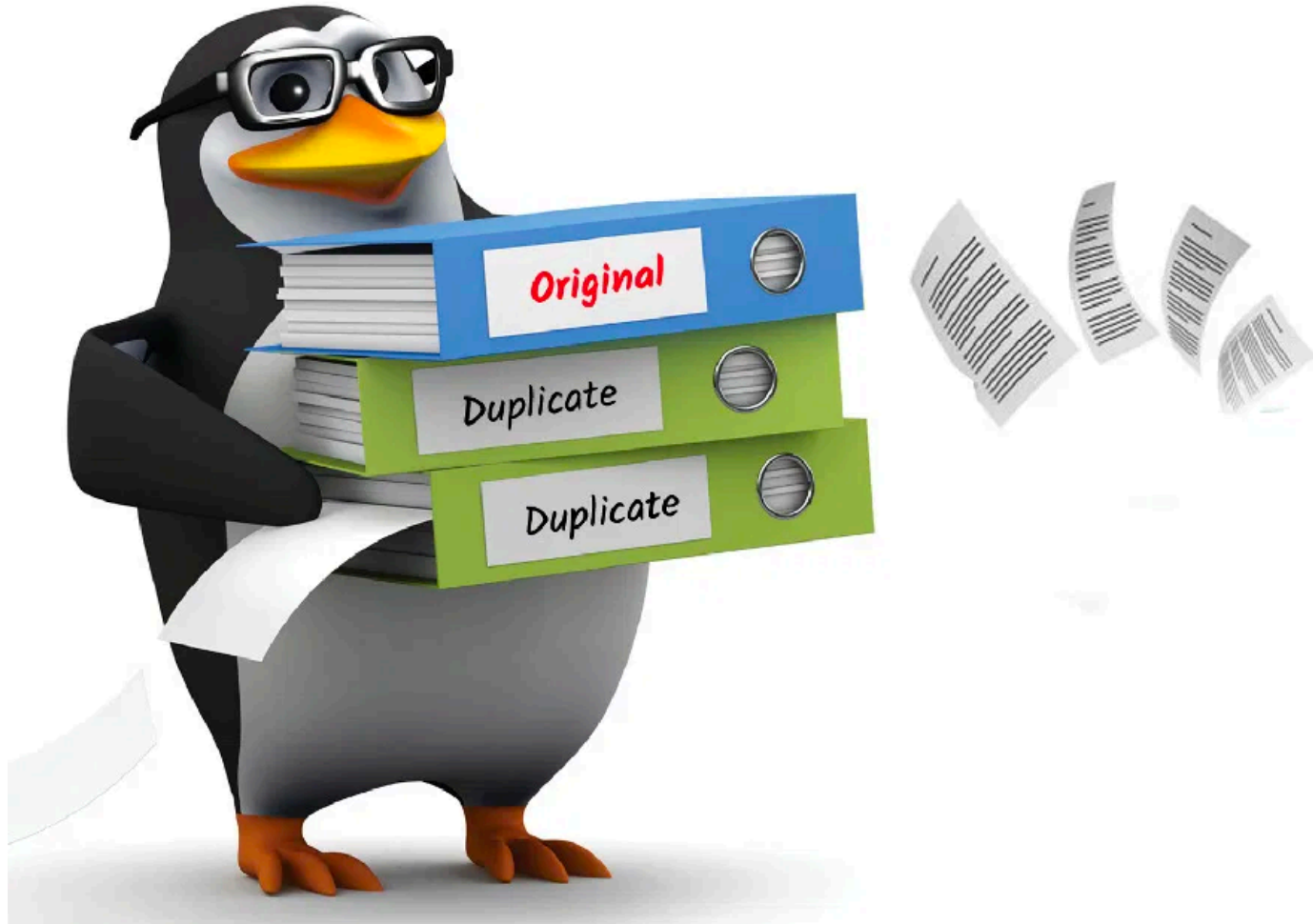
Protect and secure your data

- Data is like currency. Would you leave your money on the street?
- Proper data storage ensures that you will not lose your data.
- Data security is also needed to be compliant with IRB, funding agencies, etc...

Your laptop is vulnerable: data loss and intrusions



UCSF storage: redundant and secure



Does UCSF storage differ From data on your laptop?

- Size?
- Goals?
- Back-up?
- Version control?
- Privacy?
 - Anonymizing
 - HIPPA
 - Stored only on secure server



Data storage: 3-2-1 rule

- Three copies of your data
- Two types of media (harddrives)
- One offsite location

Many UCSF data services will offer forms of this in their subscription plans.

Data management at UCSF - storage

- UCSF Box [free, not great for backup]
 - Secure, enterprise, cloud-based storage solution. Offers unlimited amounts of storage and an encrypted, HIPAA-compliant folder.
 - <https://it.ucsf.edu/service/cloud-storage-box>
- MyResearch (\$)
 - Combined secure environment and software resources
 - Backed up automatically for a fee
 - <https://myresearch.ucsf.edu/myresearch>
- UCSF Backup Services (\$)
 - <https://it.ucsf.edu/service/backup-services>

Data management at UCSF - professional database help

- Data Management Consultation (\$) - EHR
 - <https://consult.ucsf.edu/data-management>
- Data System Services, DEB (\$)
 - REDCAP programming
 - <https://epibiostat.ucsf.edu/data-systems-services>

Make UCSF work for you!

*As members of this community,
you have access and
membership to many
resources!*



Data Cleaning

Garbage in,
garbage out



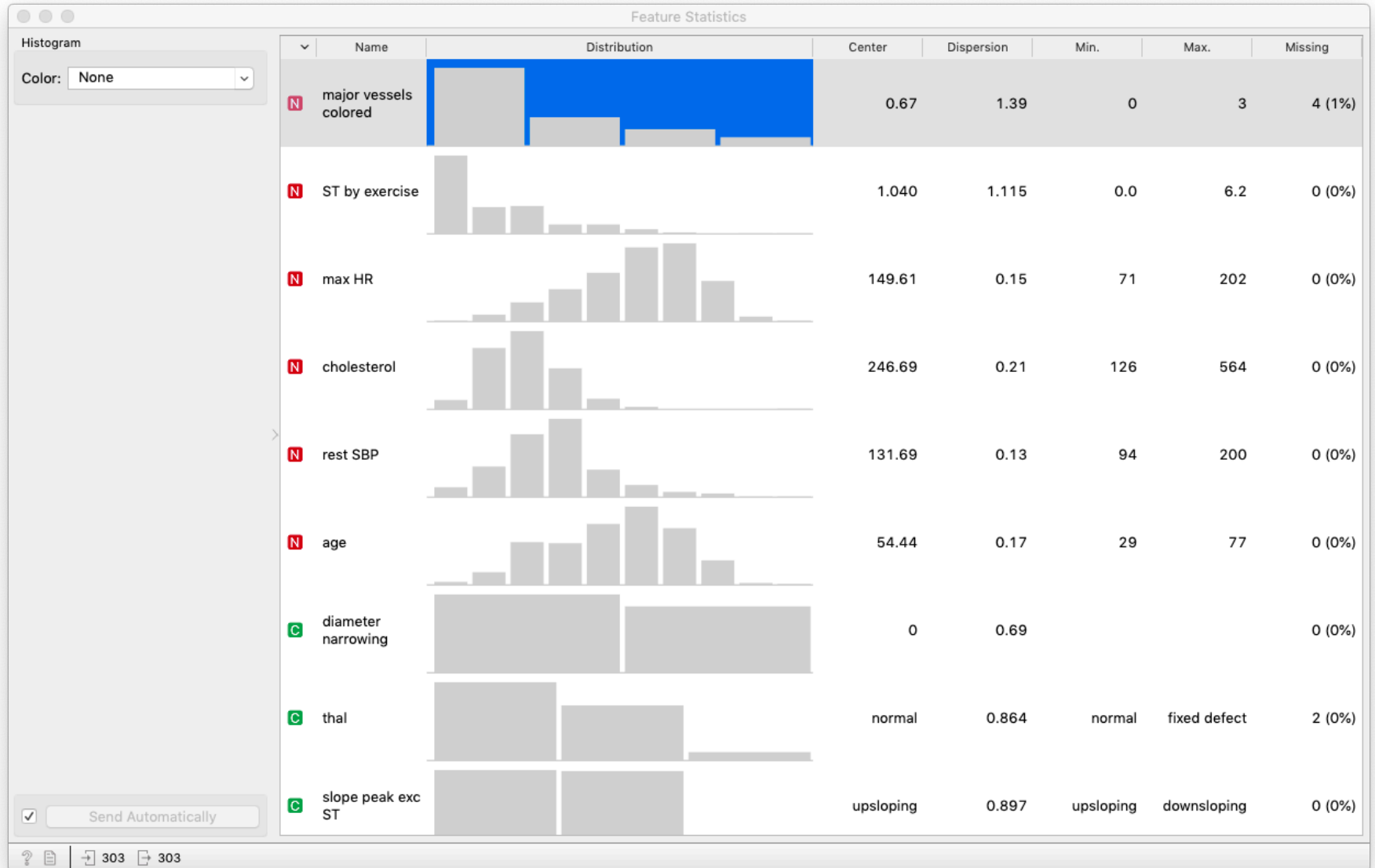
Data cleaning

- Major steps in data cleaning
- Visualize the data
- Outlier detection
- Missingness

Major steps in data cleaning

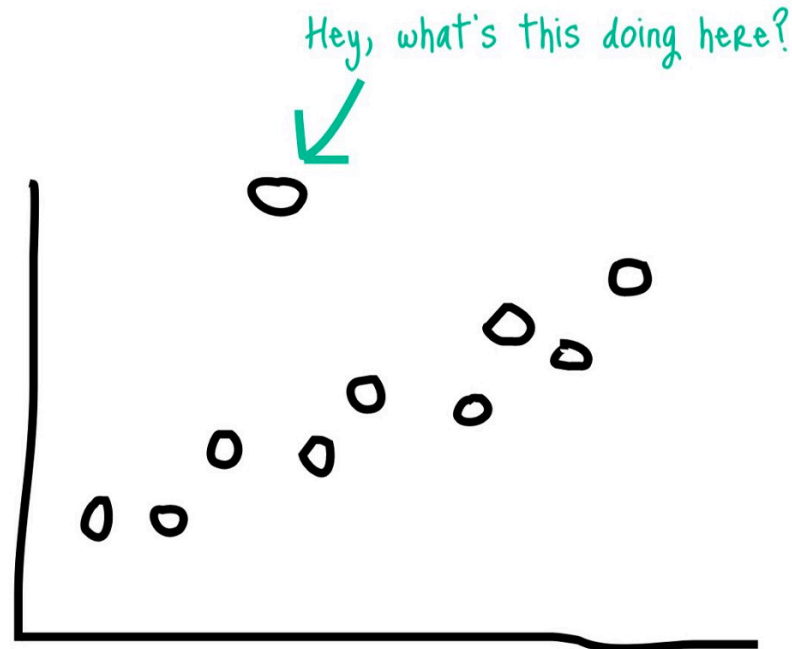
- **Visualize the data**
- Remove unwanted data (duplicate entries, irrelevant entries)
- Fix structural errors
- Resolve inconsistencies
- Plausibility check
- **Detect and filter outliers**
- **Characterize and address missing data**
- Transform data
 - Derive new variables
 - Center/scale data
- Reduce data, i.e. dimensional reduction

Visualize the data – “Feature Statistics” widget



Outlier detection (1)

- An outlier is a data point that differs significantly from other observations.
- Outliers can be detected visually, but in large datasets we need algorithmic methods to detect



Outlier detection (2)

- Outliers can be due chance or data corruption
- Outliers can influence out algorithms, cause poor prediction or confusing results
- Ultimately, you can choose to keep them, discard them, or transform them (think Fare variable in titanic).

MAKE SURE TO DOCUMENT YOUR DECISIONS!!!

Missing data

- Missing data may be due to
 - Refusal to answer
 - Equipment malfunction
 - Data not entered due to error
- Blank entries are good to indicate missing
- Missing data may need to be inferred or imputed to avoid bias

How to handle missing data?

- Ignore the data (bad)
- Fill in the missing values manually (e.g. call the patient, not viable generally)
- “Impute” or fill in the missing values (best if original source cannot be verified)

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 for data management and cleaning

- Visualize data
- Feature construction
- Merging & concatenating
- Impute missing data
- Detect outliers

Summary

1. Data management and storage are necessary not only to protect your research efforts but also for compliance.
2. Data integrity can help you and others use your data.
3. Don't store important data on your laptop! Use 3-2-1 rule when possible.
4. Garbage in, garbage out
5. Data cleaning decisions have critical impact on subsequent analysis and interpretation.



Coming
up....

Starting next week, we will
begin a 3-week journey into
machine learning and
algorithms.

