

DATA COLLECTION AND DATA PROCESSING

ADVANCED DATA SCIENCE TRAINING I

“People resist a census, but give them a profile page and they’ll spend all day telling you who they are.”

Max Berry, Lexicon

OUTLINE

1. What Data To Collect: Sampling Theory and Study Design
2. Modern Data Collection: APIs and Web Scraping
3. Working with your Data: Data Wrangling
4. Getting Ready for Analysis: Data Cleaning
5. Making Your Data (More) Manageable: Data Transformation
6. Ensuring Good Data: Data Quality and Data Validation

REFERENCES

DATA COLLECTION AND DATA PROCESSING

REFERENCES

Chapman, A. [2005], *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*, Report for the Global Biodiversity Information Facility, Copenhagen.

van Buuren, S. [2012], *Flexible Imputation of Missing Data*, CRC Press, Boca Raton.

Orchard, T. and Woodbury, M. [1972], *A Missing Information Principle: Theory and Applications*, Proc. Sixth Berkeley Symp. on Math. Statist. and Prob., Berkeley.

Hagiwara, S. [2012], *Nonresponse Error in Survey Sampling – Comparison of Different Imputation Methods*, Honours Thesis, Carleton University, Ottawa.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. [2001], *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models*, Survey Methodology, v.27, n.1, pp.85-95, Statistics Canada, Catalogue no. 12-001.

Survey Methods and Practices, Statistics Canada, Catalogue no.12-587-X.

REFERENCES

Rubin, D.B. [1987], *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Kutner, M., Nachtsheim, C., Neter, J. and Li, W. [2004], *Applied Linear Statistical Models*, 5th ed., McGraw-Hill/Irwin, New York.

Green, S. and Salkind, N. [2011], *Using SPSS for Windows and Macintosh – Analyzing and Understanding Data*, 6th ed., Prentice Hall, Upper Saddle River.

Wikipedia entry for [Data Cleansing](#)

Wikipedia entry for [Imputation](#)

Wikipedia entry for [Outliers](#)

Torgo, L. [2017], *Data Mining with R* (2nd edition), CRC Press.

McCallum, Q.E. [2013], *Bad Data Handbook*, O'Reilly.

REFERENCES

Kazil, J., Jarmul, K. [2016], *Data Wrangling with Python*, O'Reilly

de Jonge, E., van der Loo, M. [2013], *An Introduction to Data Cleaning with R*, Statistics Netherlands.

Pyle, D. [1999], *Data Preparation for Data Mining*, Morgan Kaufmann Publishers.

Weiss, S.M., Indurkha, I. [1999], *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann Publishers.

Buttrey, S.E. [2017], *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*, Wiley.

Aggarwal, C.C. [2013], *Outlier Analysis*, Springer.

Chandola, V., Banerjee, A., Kumar, V. [2007], *Outlier detection: a survey*, Technical Report TR 07-017, Department of Computer Science and Engineering, University of Minnesota.

Hodge, V., Austin, J. [2004], A survey of outlier detection methodologies, *Artif.Intell.Rev.*,

REFERENCES

Feng, L., Nowak, G., Welsh, A.H., O'Neill, T. [2014], *imputeR: a general imputation framework in R*.

Steiger, J.H. , [Transformations to Linearity](#), lecture notes.

Wood, F., [Remedial Measures Wrap-Up and Transformations](#), lecture notes.

Dougherty, J., Kohavi, R., Sahami, M. [1995], Supervised and unsupervised discretization of continuous features, in *Machine Learning: Proceedings of the Twelfth International Conference*, Frieditis, A., Russell, S. (eds), Morgan Kaufmann Publishers.

Orchard, T., Woodbury, M. [1972], [A Missing Information Principle: Theory and Applications](#), Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

Height Percentile Calculator, by Age and Country, <https://tall.life/height-percentile-calculator-age-country/>

Dua, D., Karra Taniskidou, E. [2017], Liver Disorders dataset, UCI Machine Learning Repository.

REFERENCES

<http://www.roymfrancis.com/scraping-instagram-choosing-hashtags/>

Munzert, S., Rubba, C., Meissner, P., Nyhuis, D. [2015], *Automated Data Collection with R, A Practical Guide to Web Scraping and Text Mining*, Wiley

Mitchell, R. [2015], *Web Scraping with Python: Collecting Data From the Modern Web*, O'Reilly.

https://www.w3schools.com/xml/xpath_intro.asp

<https://www.w3schools.com/>

<https://en.wikipedia.org/wiki/XHTML>

<https://medium.com/the-andela-way/introduction-to-web-scraping-using-selenium-7ec377a8cf72>

<https://pypi.python.org/pypi/selenium>

REFERENCES

Guyon, I., Elisseeff, A., [An Introduction to Variable and Feature Selection](#), *Journal of Machine Learning Research*, 3(Mar):1157-1182, 2003.

Cawley, G.C., Talbot, N.L.C., [Gene selection in cancer classification using sparse logistic regression with Bayesian regularization](#), *Bioinformatics*, (2006) 22 (19): 2348-2355.

Ambroise, C., McLachlan, G.J., [Selection bias in gene extraction on the basis of microarray gene-expression data](#), *PNAS*, vol.99, no.10, pp.6562–6566, 2002.

Liu, H., Motoda, H. (eds), *Computational Methods of Feature Selection*, Chapman Hall/ CRC Press.

Kononenko, I., Kukar, M. [2007], *Machine Learning and Data Mining: Introduction to Principles and Algorithms*, ch.6, Horwood Publishing.

[Lasso \(statistics\)](#) on Wikipedia

Aggarwal, C.C. [2016], *Data Mining: the Textbook*, sec. 2.4.3, Springer.

REFERENCES

Robnik-Sikonja, M., Savicky, P., [CORElearn](#) package documentation, v1.51.2, CRAN.

Ng, A., Soo, K., [Principal Component Analysis Tutorial](#), June 15, 2016.

[Principal component analysis](#), on Wikipedia

Hastie, T., Tibshirani, R., Friedman, J. [2009], [*The Elements of Statistical Learning* \(2nd ed.\)](#), ch.2, Springer.

Smith, L.I. [2002], [A Tutorial on Principal Component Analysis](#)

Shlens, J. [2014], [A Tutorial on Principal Component Analysis](#), arXiv.org

[Nonlinear dimensionality reduction](#), on Wikipedia

J. Leskovec, A. Rajaraman, J. Ullman [2015] [*Mining of Massive Datasets*](#), Cambridge University Press.

REFERENCES

Skillicorn, D. [2007], *Understanding Complex Datasets: Data Mining with Matrix Decomposition*, Chapman and Hall/CRC Press.

[CORElearn](#) documentation

[Feature selection](#), on Wikipedia

<https://simplystatistics.org/2014/10/24/an-interactive-visualization-to-teach-about-the-curse-of-dimensionality/>

Grolemund, G. [2015], *Data Wrangling with R: how to work with the structures of your data*, webinar, bit.ly/wrangling-webinar

<https://www.rstudio.com/resources/cheatsheets/>

Farrell, P., *STAT 4502 Survey Sampling Course Package*, Carleton University, Fall 2008

REFERENCES

Lessler, J. and Kalsbeek, W. [1992], *Nonsampling Errors in Surveys*, Wiley, New York

Oppenheim, N. [1992], *Questionnaire Design, Interviewing, and Attitude Measurement*, St. Martin's

Hidiroglou, M., Drew, J. and Gray, G. [1993], "A Framework for Measuring and Reducing non-response in Surveys," *Survey Methodology*, v.19, n.1, pp.81-94

Gower, A. [1994], "Questionnaire Design for Business Surveys," *Survey Methodology*, v.20, n.2
Survey Methods and Practices, Statistics Canada, Catalogue no.12-587-X

Boily, P., Schellinck, J., Hagiwara, S., *et al.* [in preparation], *Introduction to Quantitative Consulting*.

SAMPLING THEORY AND STUDY DESIGN

DATA COLLECTION AND DATA PROCESSING

“The latest survey shows that 3 out of 4 people make up 75% of the population”

D. Letterman

THE GOAL OF GOOD STUDY/SAMPLING DESIGN

We need data that can:

- provide legitimate insight into our system of interest;
- provide correct, accurate answers to relevant questions;
- support the drawing of legitimate, valid conclusions, with the ability to qualify these conclusions in terms of scope and precision.

This starts with **study design** – what data to collect and how it should be collected

“A Dartmouth graduate student used an MRI machine to study the brain activity of a salmon as it was shown photographs and asked questions. The most interesting thing about the study was not that a salmon was studied, but that the salmon was dead. Yep, a dead salmon purchased at a local market was put into the MRI machine, and some patterns were discovered. There were inevitably patterns—and they were invariably meaningless.”

NPS AND PATTERN FISHING

Two separate issues can be combined to cause **problems** with data analysis:

- drawing conclusions (inferences) from a sample about a population that are not warranted by the sample collection method (symptomatic of NPS);
- looking for any available patterns in the data and then coming up with *post hoc* explanations for these patterns.

Alone or in combination, these lead to poor (and **potentially harmful**) conclusions.

STUDIES AND SURVEYS

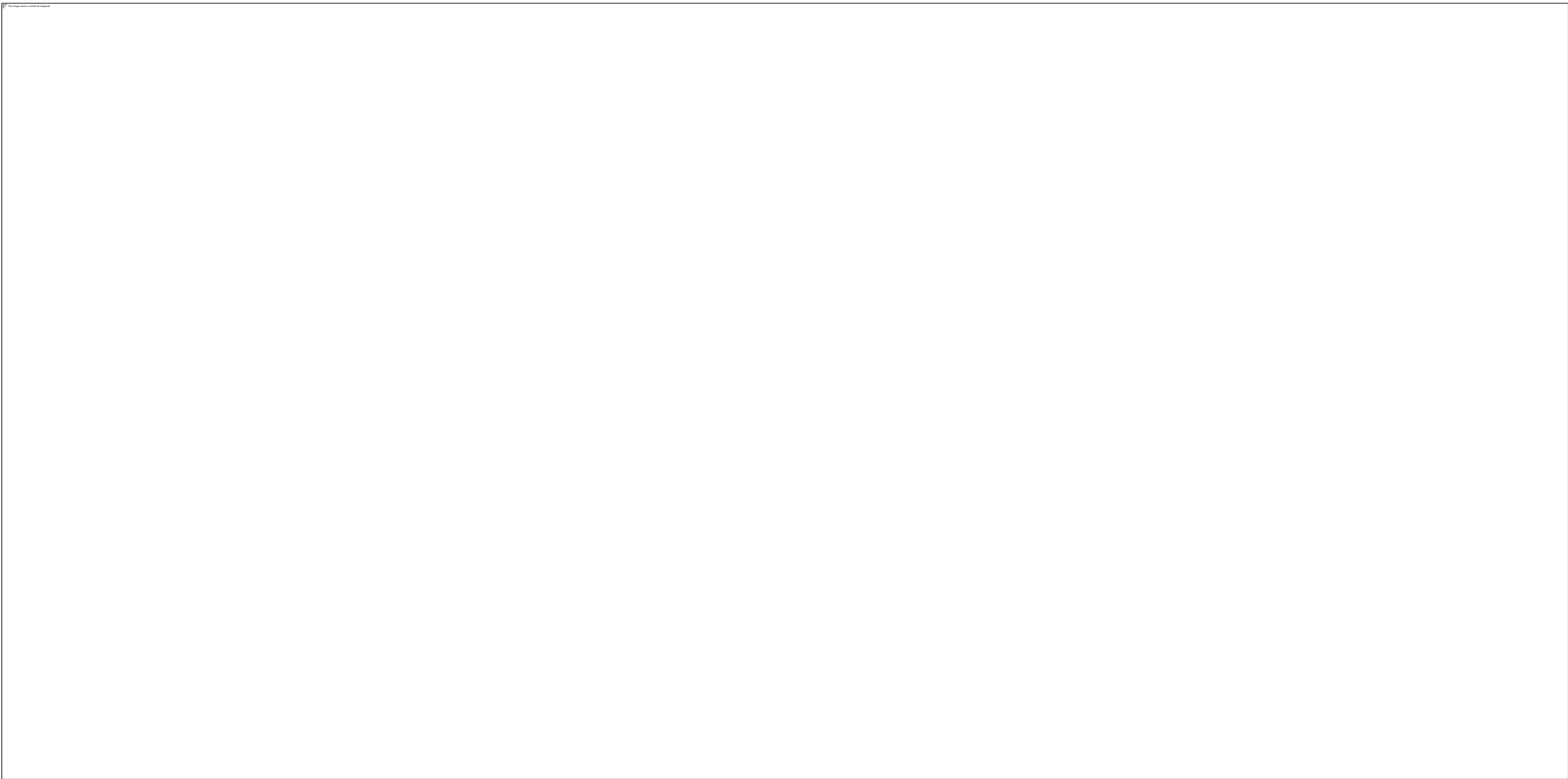
A **survey** is any activity that collects information about characteristics of interest:

- in an **organized** and **methodical** manner;
- from some or all **units** of a population;
- using **well-defined** concepts, methods, and procedures, and
- compiles such information into a **meaningful** summary form.

SAMPLING MODELS

A **census** is a survey where information is collected from all units of a population, whereas a **sample survey** uses only a fraction of the units.

When survey sampling is done properly, we may be able to use various **statistical methods** to make **inferences** about the **target population** by sampling a (comparatively) small number of units in the **study population**.



SURVEY FRAMES

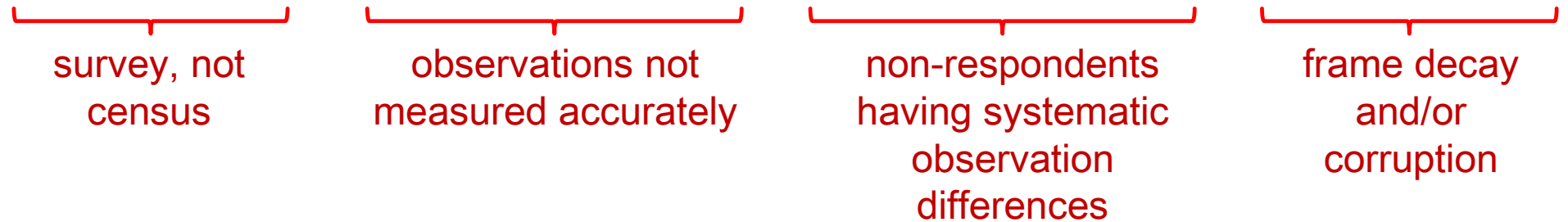
The ideal frame contains identification data, contact data, classification data, maintenance data, and linkage data, and must minimize the risk of **undercoverage** or **overcoverage**, as well as the number of duplications and misclassifications (although some issues that arise can be fixed at the data processing stage).

A statistical sampling approach is contraindicated unless the selected frame is

- **relevant** (which is to say, it corresponds, and permits accessibility to, the target population),
- **accurate** (the information it contains is valid),
- **timely** (it is up-to-date), and
- **competitively priced**.

SURVEY ERROR

Total Error = Sampling Error + Measurement Error + Non-Response Error + Coverage Error



Statistical sampling can help provide estimates, but importantly, it can also provide some control over the **total error** (TE) of the estimates.

Ideally, $TE = 0$. In practice, there are two main contributions to TE: **sampling errors** (due to the choice of sampling scheme), and **nonsampling errors** (everything else).

NONSAMPLING ERROR

Nonsampling error can be controlled, to some extent:

- **coverage error** can be minimized by selecting high quality, up-to-date survey frames;
- **non-response error** can be minimized by careful choice of the data collection mode and questionnaire design, and by using “call-backs” and “follow-ups”;
- **measurement error** can be minimized by careful questionnaire design, pre-testing of the measurement apparatus, and cross-validation of answers.

In practice, these suggestions are not that useful in modern times (landline-based survey frames are becoming irrelevant due to demographics, response rates for surveys that are not mandated by law are low, etc.). This explains, in part, the over-use of **web scraping** and **non-probabilistic sampling**.

NONPROBABILISTIC SAMPLING

Nonprobabilistic sampling (NPS) methods (designs) select sampling units from the target population using subjective, non-random approaches.

- NPS are quick, relatively inexpensive and convenient (no survey frame required).
- NPS methods are ideal for exploratory analysis and survey development.

Unfortunately, NPS are often used instead of probabilistic designs (problematic)

- the associated selection bias makes NPS methods unsound when it comes to inferences (they cannot be used to provide reliable estimates of the sampling error, the only component of TE under the analyst's direct control);
- automated data collection often fall squarely in the NPS camp – we can still analyze data collected with a NPS approach, but may not generalize the results to the target population.

PROBABILISTIC SAMPLING

Probabilistic sample designs are usually more **difficult** and **expensive** to set-up (due to the need for a quality survey frame), and take longer to complete.

They provide **reliable estimates** for the attribute of interest and the **sampling error**, paving the way for small samples being used to draw inferences about larger target populations (in theory, at least; the non-sampling error components can still affect results and generalisation).

CONFIDENCE INTERVALS

If the estimate $\hat{\beta}$ is unbiased, $E(\hat{\beta} - \beta) = 0$, then an approximate **95% confidence interval** (95% CI) for β is given approximately by

$$\hat{\beta} \pm 2\sqrt{\hat{V}(\hat{\beta})},$$

where $\hat{V}(\hat{\beta})$ is a **sampling design-specific** estimate of $V(\hat{\beta})$.

But what is a 95% CI, exactly?

SAMPLING DESIGN

Different **sampling designs** have distinct advantages and disadvantages.

They can be used to compute estimates

- for various population attributes: mean, total, proportion, ratio, difference, etc.
- for the corresponding 95% CI.

We might also want to compute sample sizes for a given **error bound** (an upper limit on the radius of the desired 95% CI), and how to determine the **sample allocation** (how many units to be sampled in various sub-population groups).

SAMPLING DESIGN – UNIVERSE OF DISCOURSE

Target population:

- N units and measurements $\mathcal{U} = \{u_1, \dots, u_N\}$

True population attributes:

- mean μ , variance σ^2 , total τ , proportion p

Sample population:

- n units and measurements $\mathcal{Y} = \{y_1, \dots, y_n\} \subseteq \mathcal{U}$

Sample population attributes:

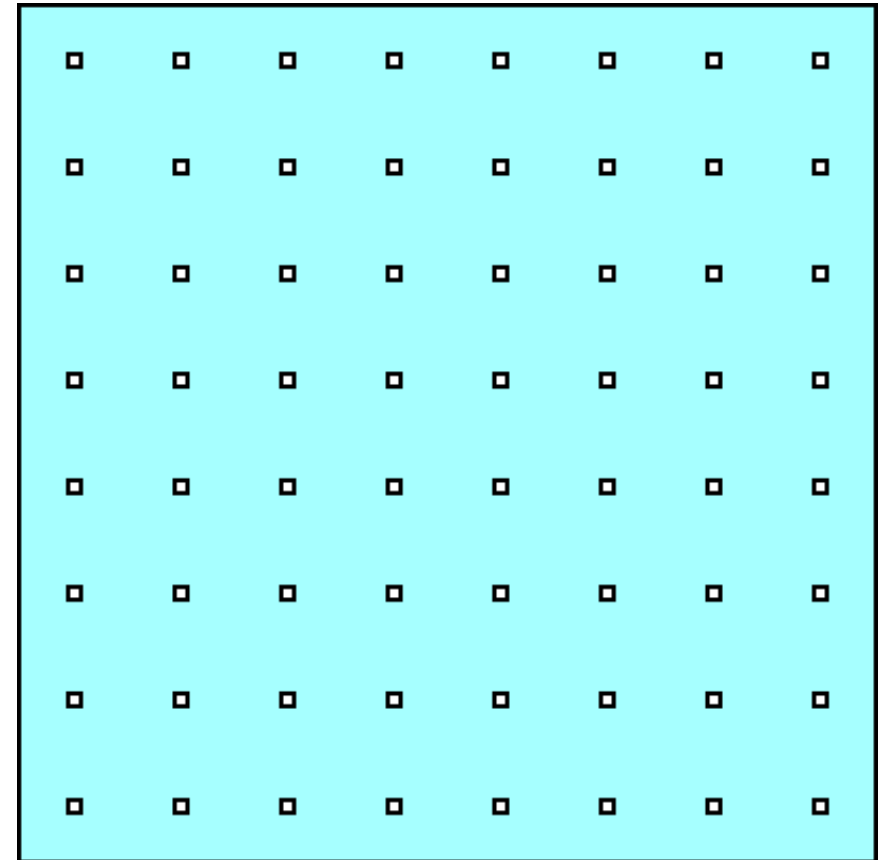
- sample mean \bar{y} , sample variance s^2 , sample total $\hat{\tau}$, sample proportion \hat{p}

SAMPLING DESIGN – UNIVERSE OF DISCOURSE

Goal: estimate the true population attributes μ , σ^2 , τ , p *via* the sample population attributes \bar{y} , s^2 , $\hat{\tau}$, \hat{p} , n , and the size N of the target population.

For a given characteristic, we define δ_i as 1 or 0 depending on whether the sample unit y_i possesses the characteristic in question or not.

We use the error bound $B = 2\sqrt{\hat{V}}$.



SIMPLE RANDOM SAMPLING (SRS)

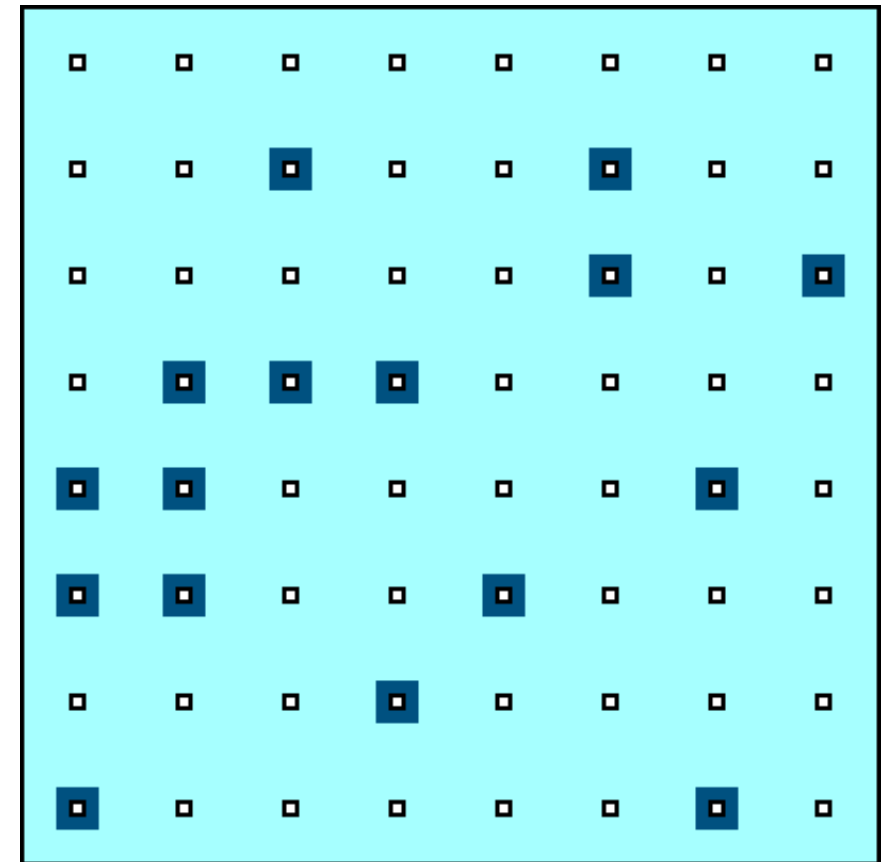
In SRS, n units are selected randomly from the frame.

Advantages:

- easiest sampling design to implement
- sampling errors are well-known and easy to estimate
- does not require auxiliary information

Disadvantages:

- makes no use of auxiliary information
- no guarantee that the sample is representative
- costly if sample is widely spread out, geographically



SRS ESTIMATORS

Estimators:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \hat{t} = N\bar{y}, \quad \hat{p} = \frac{1}{n} \sum_{i=1}^n \delta_i$$

Sample Design-Specific Variance Estimates:

$$\widehat{V}(\bar{y}) = \frac{s^2}{n} \left(1 - \frac{n}{N}\right), \quad \widehat{V}(\hat{t}) = N^2 \widehat{V}(\bar{y}), \quad \widehat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} \left(1 - \frac{n}{N}\right)$$

Sample Allocation:

$$n_{\bar{y}} = \frac{4N\tilde{\sigma}^2}{(N-1)B^2 + 4\tilde{\sigma}^2}, \quad n_{\hat{t}} = \frac{4N^3\tilde{\sigma}^2}{(N-1)B^2 + 4N^2\tilde{\sigma}^2}, \quad n_{\hat{p}} = \frac{4\tilde{p}(1-\tilde{p})}{(N-1)B^2 + 4\tilde{p}(1-\tilde{p})}$$

STRATIFIED RANDOM SAMPLING (STS)

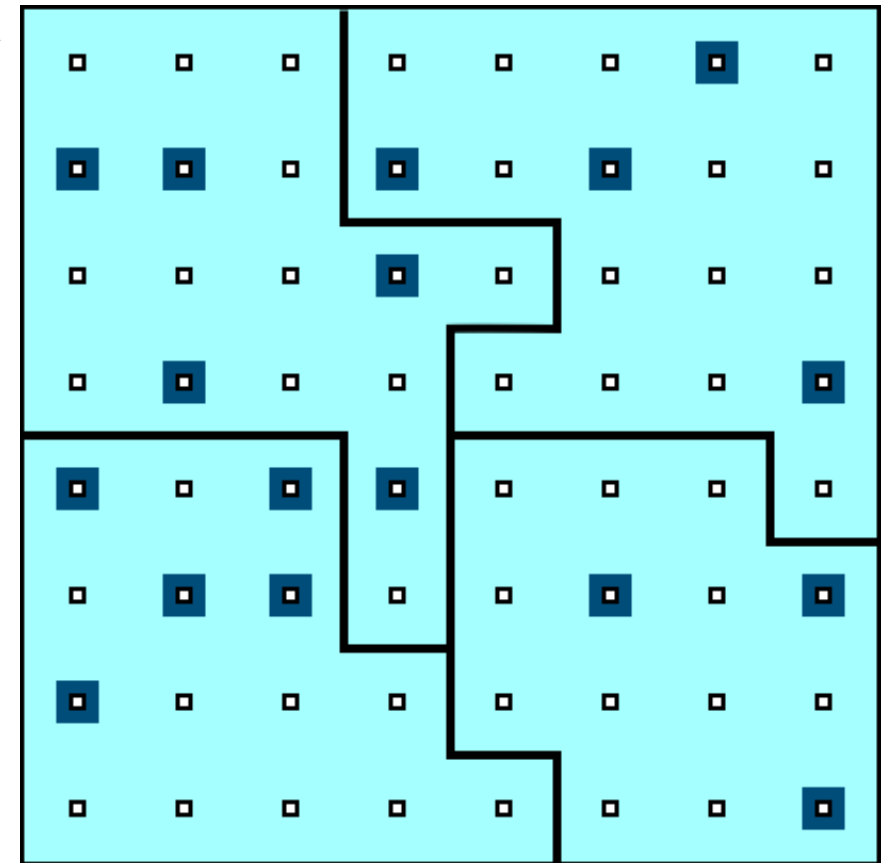
In StS, $n = n_1 + \dots + n_k$ units are selected randomly from k frame **strata**.

Advantages:

- may produce smaller error bound on estimation than SRS
- may be less expensive if elements are conveniently strat.
- may provide estimates for sub-populations

Disadvantages:

- no major disadvantage
- if there are no natural ways to stratify the frame into homo-geneous groupings, StS is roughly equivalent to SRS



STS ESTIMATORS

Estimators:

$$\bar{y}_{st} = \sum_{j=1}^k \frac{N_j}{N} \bar{y}_j, \quad \hat{\tau}_{st} = N \bar{y}_{st}, \quad \hat{p}_{st} = \sum_{j=1}^k \frac{N_j}{N} \hat{p}_j$$

Sample Design-Specific Variance Estimates:

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^k N_j^2 \hat{V}(\bar{y}_j), \quad \hat{V}(\hat{\tau}_{st}) = N^2 \hat{V}(\bar{y}_{st}), \quad \hat{V}(\hat{p}_{st}) = \frac{1}{N^2} \sum_{j=1}^k N_j^2 \hat{V}(\hat{p}_j)$$

EXERCISES

You are charged with estimating the yearly salary of data scientists in Canada.

Identify potential:

- populations (target, study, respondent, sampling frames)
- samples (intended, achieved)
- unit information (unit, response variate, population attribute)
- sources of bias (coverage, nonresponse, sampling, measurement) and variability (sampling, measurement).

EXERCISES

The file `cities.txt` contains population information for a country's cities. A city is classified as “small” if its population is below 75K, as “medium” if it falls between 75K and 1M, and as “large” otherwise.

1. Locate and load the file into the workspace of your choice. How many cities are there? How many in each group?
2. Display summary population statistics for the cities, both overall and by group.
3. Compute a 95% C.I. for the 1999 population mean using a SRS of size $n = 10$.
4. Compute a 95% C.I. for the 1999 population mean using a StS of size $(n_s, n_m, n_l) = (5, 3, 2)$.
5. Compare the estimates with the true value. Are the results surprising? If not, could they have been?

Supplemental Material

DECIDING FACTORS

In some instances, information about the **entire** population is required in order to answer questions, whereas in others it is not necessary. The **survey type** depends on multiple factors:

- the type of question that needs to be answered;
- the required precision;
- the cost of surveying a unit;
- the time required to survey a unit;
- size of the population under investigation, and
- the prevalence of the attributes of interest.

STUDY/SURVEY STEPS

Studies or surveys follow the same general steps:

1. statement of objective
2. selection of survey frame
3. sampling design
4. questionnaire design
5. data collection
6. data capture and coding
7. data processing and imputation
8. estimation
9. data analysis
10. dissemination
11. documentation

The process is not always linear, but there is a definite movement from objective to dissemination.

SURVEY FRAMES

The **frame** provides the means of **identifying** and **contacting** the units of the study population. It is generally costly to create and to maintain (in fact, there are organisations and companies that specialise in building and/or selling such frames).

Useful frames contain:

- identification data,
- contact data,
- classification data,
- maintenance data, and
- linkage data.

MODES OF DATA COLLECTION

Paper-based vs. computer-assisted

- **self-administered questionnaires** are used when the survey requires detailed information to allow the units to consult personal records; associated with high non-response rate.
- **interviewer-assisted questionnaires** use well-trained interviewers to increase the response rate and overall quality of the data; face-to-face vs. telephone.
- **computer-assisted interviews** combine data collection and data capture, which saves time.
- unobtrusive direct observation
- diaries to be filled (paper or electronic)
- omnibus surveys
- email, Internet, and social media

NPS METHODS

Haphazard

- man on the street, depends on availability of units and interviewer bias

Volunteer

- self-selection bias

Judgement

- biased by inaccurate preconceptions about the target population

Quota

- exit polling, ignores non-response bias

NPS METHODS

Modified

- starts probabilistic, switches to quota as a reaction to high non-response rates

Snowball

- “pyramid” scheme

There are contexts where NPS methods might fit a client’s or an organization’s need (and that remains their decision to make, ultimately), but they must be informed of the drawbacks, and presented with some probabilistic alternatives.

BASIC MATHEMATICAL CONCEPTS

Consider a finite population \mathcal{U} , with N units and measurements $\{u_1, \dots, u_N\}$.

The **mean** and **variance** of the population for the variable of interest are given by

$$\mu = \frac{1}{N} \sum_{j=1}^N u_j, \quad \sigma^2 = \frac{1}{N} \sum_{j=1}^N (u_j - \mu)^2.$$

If $\mathcal{Y} \subseteq \mathcal{U}$ is a **sample** of the population with n units and measurements $\{y_1, \dots, y_n\}$, then the **sample mean** and **sample variance** are given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

BASIC MATHEMATICAL CONCEPTS

Let X_1, \dots, X_n be random variables, $b_1, \dots, b_n \in \mathbb{R}$, and E , V , Cov be the **expectation**, **variance**, and **covariance** operators, respectively, i.e.:

- $E(X_i) = \mu_i$
- $\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$
- $V(X_i) = \text{Cov}(X_i, X_i) = E(X_i^2) - E^2(X_i) = E(X_i^2) - \mu_i^2 = \sigma_i^2$ and

$$E\left(\sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n b_i E(X_i) = \sum_{i=1}^n b_i \mu_i$$
$$V\left(\sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n b_i^2 V(X_i) + \sum_{i \neq j} b_i b_j \text{Cov}(X_i, X_j)$$

BASIC MATHEMATICAL CONCEPTS

The **bias** in an error component is the average of that error component if the survey is repeated many times independently under the same conditions. An **unbiased** estimate is one for which the bias is nil.

The **variability** in an error component is the extent to which that component would vary about its average value in the ideal scenario described above.

The **mean square error** of an error component is a measure of its size:

$$\text{MSE}(\hat{\beta}) = V(\hat{\beta}) + \text{Bias}^2(\hat{\beta}),$$

Where $\hat{\beta}$ is an estimator of β .

PROBABILISTIC SAMPLING DESIGNS

Simple random sampling (SRS)

Stratified random sampling (StS)

Systematic sampling (SyS)

Cluster sampling (CIS)

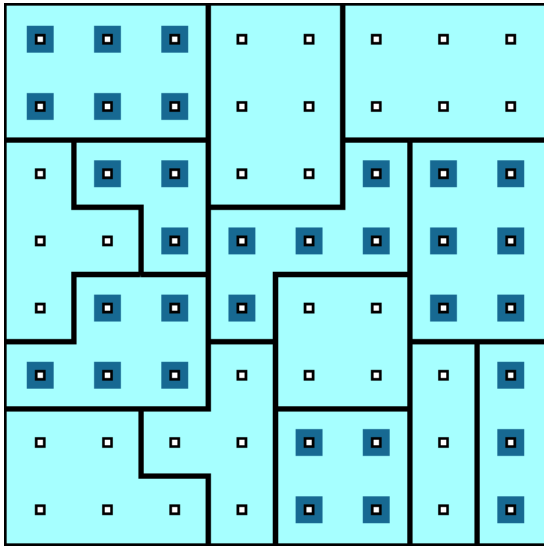
Probability proportional-to-size sampling (PPS)

Replicated sampling (ReS)

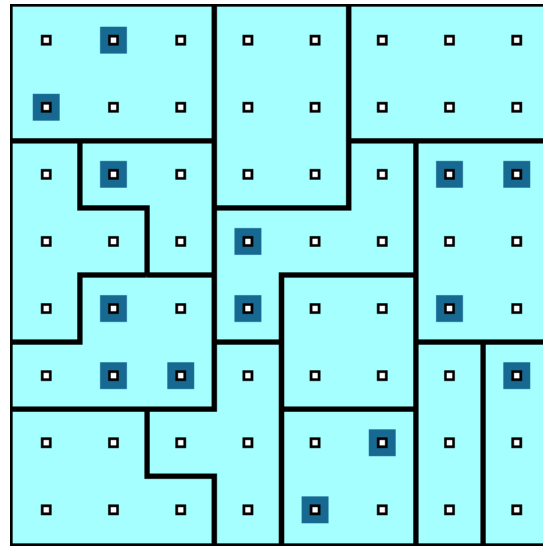
Multi-stage sampling (MSS)

Multi-phase sampling (MPS)

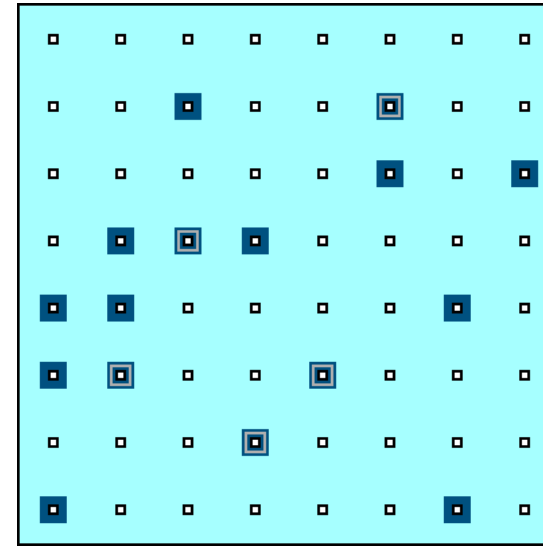
OTHER EXAMPLES OF SAMPLING DESIGNS



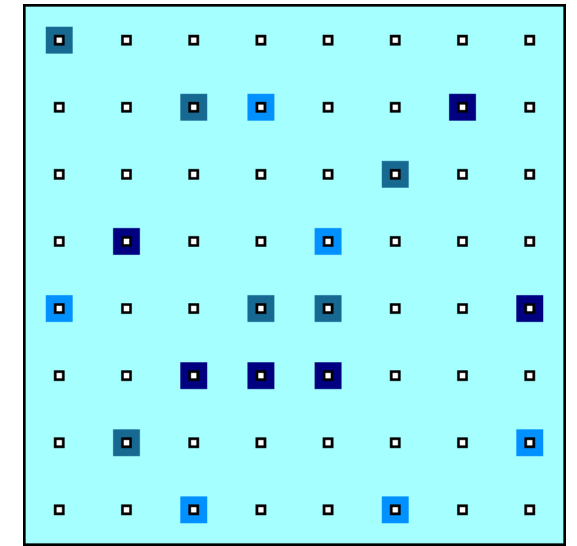
Cluster Sampling (CIS)



Multi-Stage Sampling
(MSS)



Multi-Phase Sampling
(MPS)



Replicated Sampling
(ReS)

API AND WEB SCRAPING

DATA COLLECTION AND DATA PROCESSING

“The streets of the Web are paved with data that can’t wait to be collected.”

Munzart, Rubba, Meissner, Nyhuis, Automated Data Collection with R

WORLD WIDE WEB

There was a time in the recent past where both scarcity and inaccessibility of data was a problem for researchers and decision-makers. That is **emphatically** not the case anymore.

Data abundance carries its own set of problems:

- tangled masses of data
- traditional data collection methods and classical (small) data analysis techniques may not be sufficient anymore

WEB DATA SCRAPPING EXAMPLE – NEW PHONE

Let's say you want to know what people think of a new phone. Standard approach: market research (e.g. telephone survey, reward system, etc.)

Pitfalls:

- unrepresentative sample: the selected sample might not represent the intended population
- systematic non-response: people who don't like phone surveys might be less (or more) likely to dislike the new phone
- coverage error: people without a landline can't be reached, say
- measurement error: are the survey questions providing suitable info for the problem at hand?

WEB DATA QUALITY – NEW PHONE

These solutions can be **costly, time-consuming, ineffective**.

Proxies – indicators that are strongly related to the product's popularity, without measuring it directly.

If **popularity** is defined as large groups of people preferring one product over a competitor, then sales statistics on a commercial website may provide a proxy for popularity.

Rankings on Amazon could provide a more **comprehensive** view of the phone market vs. traditional survey.

POTENTIAL ISSUES – NEW PHONE

Representativeness of the listed products

- Are all phones listed?
- If not, is it because that website doesn't sell them?
- Is there some other reason?

Representativeness of the customers

- Are there specific groups buying/not-buying online products?
- Are there specific groups buying from specific sites?
- Are there specific groups leaving/not-leaving reviews?

Truthfulness of customers and **reliability** of reviews.

IS WEB SCRAPING LEGAL?

What is a spider?

- Programs that graze or crawl the web for information rapidly
- Jumps from one page to another, grabbing the entire page content

Scraping is taking specific information from specific websites (which is the goal):
how are these **different**?

“Scraping inherently involves **copying**, and therefore one of the most obvious claims against scrapers is copyright infringement.”

LEGAL CASES – WEB SCRAPING

eBay vs. Bidder's Edge (BE)

- BE used automated programs to crawl information from different auction sites.
- Users could search listings on the BE webpage instead of going to individual auction sites.
- BE accessed eBay's sites ~100 000 times / day (1.53% of # of requests, 1.1% of total data transferred by eBay) in 1999.
- eBay alleged damages of up to \$45k- \$62K in a 10 month period.
- BE didn't steal information that wasn't public, but excessive traffic was demanding on eBay's servers.
- Your verdict?

FRIENDLY COOPERATION WITH APIS

What is an API? API stands for application program interface which is a set of routines, protocols and tools for building software applications.

Many APIs restrict the user to a certain amount of API calls per day (or some other limits).

These limits should be obeyed.

Supplemental Material

WHY AUTOMATED DATA COLLECTION?

With regards to social scientific data:

- sparse financial resources
- little time or desire to collect data by hand
- want to work with up to date, high quality data rich sources
- document process from beginning (data collection) to end (publication) so it can be reproduced

Issues with manual collection:

- non-reproducible process
- prone to errors and cumbersome
- subject to heightened risks of “death by boredom”

WHY AUTOMATED DATA COLLECTION?

Advantages of program-based solutions:

- reliability
- reproducibility
- time-efficient
- assembly of higher quality datasets

AUTOMATED COLLECTION CHECKLIST

Is **web scraping** or **statistical text processing** (automated or semi-automated data collection) really necessary?

Criteria:

- Do you plan to repeat the task from time to time e.g. to update your database?
- Do you want others to be able to replicate your data collection process?
- Do you deal with online sources of data frequently?
- Is the task non-trivial in terms of scope and complexity?

AUTOMATED COLLECTION CHECKLIST

Criteria: (continued)

- If the task can be done manually, do you lack the resources to let others do the work?
- Are you willing to automate the process by means of programming?

If most of the answers are positive then an automated approach may be the right choice.

WORLD WIDE WEB

The way we **share**, **collect**, and **publish** data has changed over the past few years due to the ubiquity of the *World Wide Web* (WWW).

Private businesses, **government**, and **individual users** are posting and sharing all kinds of data and information.

At every moment, new channels generate vast amounts of data on human behaviour.

OPEN SOURCE SOFTWARE

Another trend:

- growth and increasing popularity and power of **open source software** (source code can be inspected, modified, and enhanced by anyone).

Community aspect → ever-changing and improving

R and **Python** are open source software that can be used for data analysis in the social sciences and other domains

They incorporate **interfaces** to other programming languages and software solutions.

DATA CLEANING AND DATA PROCESSING

Data collection proper is only the tip of the iceberg.

Data cleaning and data processing is **essential** (as well as time-consuming).

Tasks:

- Selecting the columns (variables) of interest
- Re-labeling these columns
- Modifying the data type of the columns so that the data can be used the way we want

DATA CLEANING AND DATA PROCESSING

Tasks: (continued)

- Editing and/or extracting data in a column
- Deciding how to deal with missing data (which can be a challenge)
- Multiple other tasks, depending on the data and its uses

Certain tasks can be automated, others cannot.

QUESTIONS ABOUT DATA QUALITY

1. What type of data is most suited to answer your questions?
2. Is the quality of the data sufficiently high to answer your question?
3. Is the information systematically flawed?

Can you avoid the dreaded: “well, it’s the best data we have...”?

DATA QUALITY

First-hand information: for example, a tweet, or a news article

Second-hand data: data that has been copied from an offline source or scraped from elsewhere.

Sometimes you can't remember or retrace the source of data when it is second-hand.

Does it still make sense to use it? It depends.

Cross-validation is standard for use of any secondary data.

DATA QUALITY AND USER'S PURPOSE

Data quality depends on the **application**.

For example,

- Sample of tweets collected on a random day could be used to analyze the use of a hashtags or the gender-specific use of words
- Not as useful if collected on Election Day to predict the election outcomes (**collection bias**)

DATA SOURCES (TRADE-OFFS)

Automated vs. Traditional

Accuracy vs. Completeness

Coverage vs. Validity

Speed vs. Cost

etc.

DATA COLLECTION PROCESS (5 STEPS)

1. Know exactly what kind of information you need

- Specific: GDP of all OECD countries for last 10 years; sales of top 10 shoe brands in 2017
- Vague: people's opinion on shoe brand X

2. Find out if there are any web data sources that could provide direct or indirect information on your problem

- Easier for specific facts: shoe store's webpage will provide information about shoes that are currently in demand i.e. sandals, boots, etc.
- Tweets may contain opinion trends on *anything*
- Commercial platforms can provide information on product satisfaction

DATA COLLECTION PROCESS (5 STEPS)

3. Develop a theory of the data generation process when looking into potential sources

- When was the data generated?
- When was it uploaded to the Web?
- Who uploaded the data?
- Are there any potential areas that are not covered? consistent? accurate?
- How often is the data updated?

DATA COLLECTION PROCESS (5 STEPS)

4. Balance advantages and disadvantages of potential data sources

- Validate the quality of data used
- Are there other independent sources that provide similar information to crosscheck against
- Can you identify original source of secondary data

5. Make a decision

- Choose data source that seems most suitable
- Document reasons for this decision
- Collect data from several sources to validate data sources

IS WEB SCRAPING LEGAL?

Ethical Guidelines:

- Work as transparently as possible
- Document data sources at all time
- Give credit to those who originally collected and published the data
- If you did not collect the information, you probably need permission to reproduce it
- Don't do anything illegal.

Crawling another company's information to process and resell it is a common complaint.

LEGAL CASES – WEB SCRAPING

Associated Press (AP) vs. Meltwater

- Meltwater offers software that scrapes news information based on specific keywords.
- Clients order summaries on topics containing excerpts of news articles.
- AP said their content was stolen and that Meltwater needed a license before distributing the information that was scraped.
- The judge found in favour of AP arguing that Meltwater is a competitor.
- **Your verdict?**

LEGAL CASES – WEB SCRAPING

Facebook vs. Pete Warden

- Pete Warden scraped basic information from Facebook users' profiles, to offer services to manage communication and networks.
- His process, according to him, was in line with robots.txt.
- After his first blog post using the data he scraped from Facebook, he was asked to delete the data.
- Facebook contends that robots.txt has no legal force and they could sue anyone for accessing their site even if they complied with the scraping instructions, that the only legal way to access any website with a crawler was to obtain prior written permission.
- Your verdict?

LEGAL CASES – WEB SCRAPING

United States vs Aaron Swartz

- Swartz co-created RSS, markdown and Infogami.
- He was arrested in 2011 for having illegally downloaded millions of articles from the archives of JSTOR.
- The case was dismissed after his suicide in January 2013.
- Your verdict?

LESSONS LEARNED

Not clear which scraping actions are illegal and which are legal.

Re-publishing content for commercial purposes is considered more problematic than downloading pages for research/analysis.

Robots.txt: *Robots Exclusion Protocol* is a file that tells scrapers what information on the site may be harvested.

Be friendly! Not everything that can be scraped requires to be scraped. Scraping programs should behave “nicely”, provide the data you seek, and be efficient, in this order.

robots.txt cqads.carleton.ca/robots.txt

```
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html
```

```
User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /INSTALL.sqlite.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
Disallow: /UPGRADE.txt
Disallow: /xmlrpc.php
```

```
User-agent: Twitterbot
Allow: /
```

```
User-agent: *
Disallow: /esi/
Disallow: /webview
Disallow: /vweb
Disallow: /news/sponsored
Disallow: /search
Disallow: /19849159/
```

theweathernetwork.com/robots.txt

```
User-agent: *
Disallow:
Crawl-delay: 10
```

cfl.ca/robots.txt

CONTACT DATA PROVIDERS

Any data accessed by HTTP forms is stored in some sort of database.

Ask proprietors of the data first if they will grant access to the database or files.

The larger the amount of data you want, **the better it is for both parties to communicate before starting to harvest data.**

For small amounts of data, that's less important.

SCRAPING DO'S AND DON'T'S

1. Stay identifiable

2. Reduce traffic

- Accept compressed files
- If scraping the same resources multiple times, check first if it has changed before accessing again
- Retrieve only parts of a file

SCRAPING DO'S AND DON'T'S

3. Do not bother server with multiple requests

- Many requests per second can bring smaller servers down
- Webmasters may block you if your scraper behaves this way
- One or two request per second is fine

4. Write modest scraper (efficient and polite)

- No reason to scrape pages daily or repeat same task over and over; make your scraper as efficient as possible
- Do not over-scrape pages
- Select resources you want to use and leave the rest untouched

DEVELOPER TOOLS

Developer tools allow us (among other things) to see the correspondence between the HTML code for a page, and the rendered version we see in the browser.

Unlike “View Source”, developer tools shows the *dynamic* version of the HTML (i.e. the HTML is shown with any changes made by JavaScript since the page was first received).

Inspecting a page’s various elements and discovering where they reside in the HTML file is crucial to efficient web scraping.

DEVELOPER TOOLS

Firefox

- right click page → Inspect Element

Safari

- Safari → Preferences → Advanced → Show Develop Menu in Menu Bar
- Develop → Show Web Inspector

Chrome

- right click page → Inspect



HOME LIVE SHOWS **ERB** MUSIC VIDEOS GALLERY SHOP PRESS ARCHIVE CONTACT

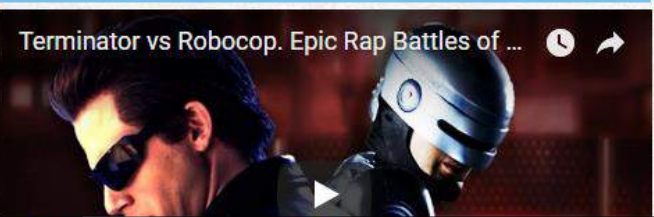
Shaka Zulu vs Julius Caesar



Eastern Philosophers vs Western Philosophers



Terminator vs Robocop



David Copperfield vs Harry Houdini





CONTACT

`div.zoogole-feature.block.layout_full` | 491.52 x 275.22

Shaka Zulu vs Julius Caesar. Epic Rap Battle...



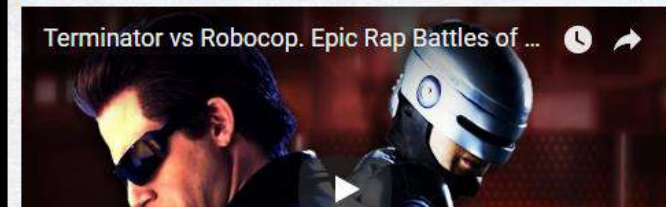
Eastern Philosophers vs Western Philosophers

Eastern Philosophers vs Western Philosophers...  





Terminator vs Robocop

Terminator vs Robocop. Epic Rap Battles of ... ⌚ ➦



David Copperfield vs Harry Houdini

David Copperfield vs Harry Houdini. Epic Rap...  



The screenshot shows the 'Elements' panel in Chrome DevTools. The DOM tree is expanded to show the 'page-content-wrap' div. Inside this div, there is a 'zoogole-content block' with a 'content-width' attribute. The attribute value is a long list of numbers from 100 to 1360 in increments of 20, wrapped in angle brackets. The list is: <100 >120 >140 >160 >180 >200 >220 >240 >260 >280 >300 >320 >340 >360 >380 >400 >420 >440 >460 >480 >500 >520 >540 >560 >580 >600 >620 >640 >660 >680 >700 >720 >740 >760 >780 >800 >820 >840 >860 >880 >900 >920 >940 >960 >980 >1000 >1020 >1040 >1060 >1080 >1100 >1120 >1140 >1160 >1180 >1200 >1220 >1240 >1260 >1280 >1300 >1320 >1340 >1360 >1380 >1400>.

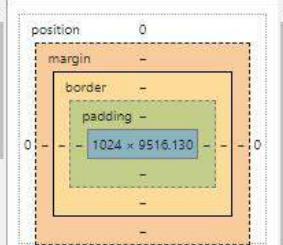
```

***
    <div class="zoogole-columns-inner site-wrap"> == $0
      <div class="zoogole-column zoogole-column-1-of-2
col_1_of_2 block layout_half" data-column-id=
"3098090">
        <div class="zoogole-feature block layout_full
block-title-feature" data-block-id="3105122">...
      </div>
      <div class="zoogole-feature block layout_full"
data-block-id="3105121">...</div>
      <div class="zoogole-feature block layout_full
block-title-feature" data-block-id="3007634">...
    </div>
    <div class="zoogole-feature block layout_full"
data-block-id="3007639">...</div>
    <div class="zoogole-feature block layout_full
block-title-feature" data-block-id="2948276">...
  </div>

```

[Styles](#)
[Event Listeners](#)
[DOM Breakpoints](#)
[Properties](#)
[Accessibility](#)

```
Filter                                :hov .cls +
element.style {
}
#usersite-                            application-6fbwf72b7b2e.css:1
container .zoogole-columns-inner {
  display: webkit-box;
  display: ms-flexbox;
  position: relative;
  -webkit-box-pack: justify;
  -ms-flex-pack: justify;
  -ms-flex-wrap: wrap;
  flex-wrap: wrap;
  display: flex;
  justify-content: space-between;
}
div {                                user agent stylesheet
  display: block;
}
```



Filter ☐ Show all

- ▶ color rgb(0...
- ▶ display flex
- ▶ flex-wrap wrap
- ▶ font-family Mul, s...
- ▶ font-size 14px

XPATH

XPath is a query (domain-specific) language

- Used to select specific pieces of information from marked-up documents such as HTML, XML or variants such as SVG, RSS
- Information stored in marked-up documents needs to be converted into formats suitable for processing and statistical analysis
- Implemented in R package `XML`
- Process steps:
 1. Specifying the data of interest
 2. Locating it in a specific document
 3. Tailoring a query to the document to extract the desired info.



Robert Gentleman

'What we have is nice, but we need something very different'

Source: Statistical Computing 2003, Reisenburg

Rolf Turner

'R is wonderful, but it cannot work magic'

answering a request for automatic generation of data from a known mean and 95% CI

Source: [R-help](#)

[The book homepage](#)

Notebook: XPath Basics

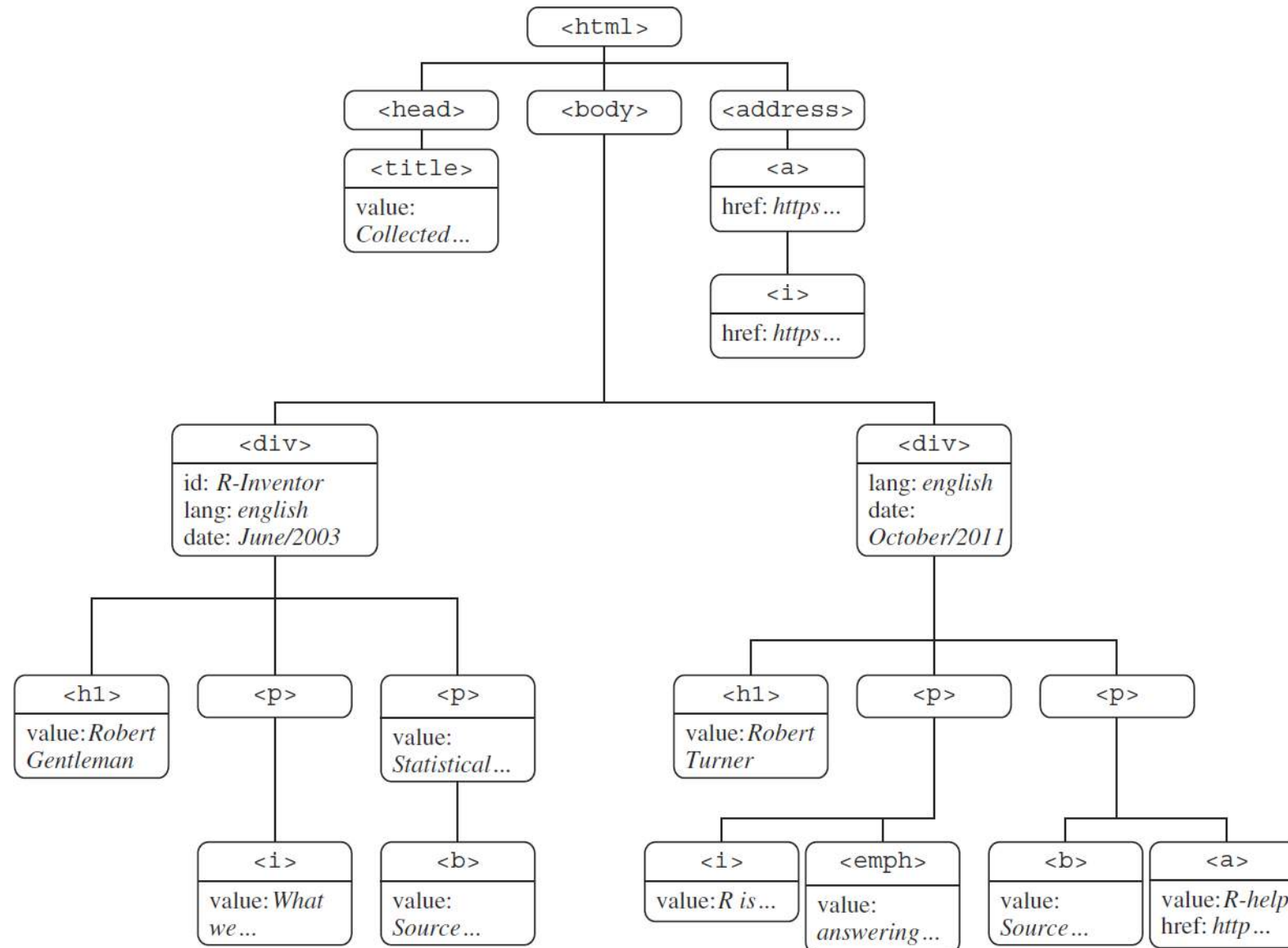


```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
<html>
<head><title>Collected R wisdoms</title></head>
<body>
<div id="R Inventor" lang="english" date="June/2003">
  <h1>Robert Gentleman</h1>
  <p><i>'What we have is nice, but we need something very different'</i></p>
  <p><b>Source: </b>Statistical Computing 2003, Reicensburg</p>
</div>

<div lang="english" date="October/2011">
  <h1>Rolf Turner</h1>
  <p><i>'R is wonderful, but it cannot work magic'</i> <br><emph>answering a request
for automatic generation of 'data from a known mean and 95% CI'</emph></p>
  <p><b>Source: </b><a href="https://stat.ethz.ch/mailman/listinfo/r-help">R-help</a>
</p>
</div>

<address>
<a href="http://www.r-datacollectionbook.com"><i>The book homepage</i></a><a></a>
</address>

</body>
</html>
```



XPATH – BASIC STRUCTURE

HTML/XML tags have **attributes** and **values**.

HTML files must be parsed before they can be queried by XPath.

XPath queries require a path and a document to search.

- paths consist of hierarchical addressing mechanism (succession of nodes, separated by forward slashes ["/"])
- a query takes the form `xpathSApply(doc, path):`
 - `xpathSApply(parsed_doc, "/html/body/div/p/i")`

would find all `<i>` tags inside a `<p>` tag inside a `<div>` tag in the `body` of the `html` file.

XPATH – NODE RELATIONS

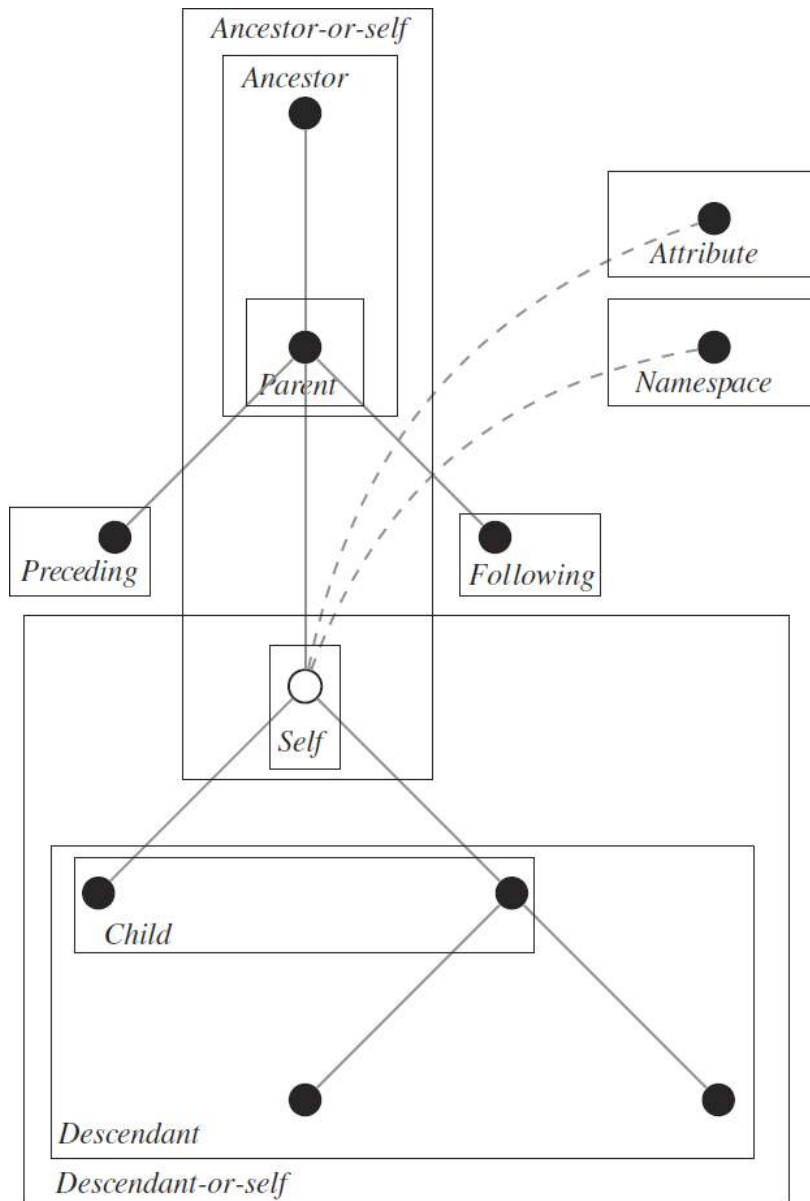
Absolute (or even relative) paths cannot always succinctly select nodes in large or complicated files.

Family tree analogy: nodes' placement in the parsed tree often mimic the relations in extended families.

Relations are denoted according to `node1/relation::node2`.

Examples:

- `"//a/ancestor::div"` returns all `<div>` nodes that are ancestor to an `<a>` node.
- `"//a/ancestor::div//i"` returns all `<i>` nodes contained in a `<div>` node that is an ancestor to an `<a>` node, etc.



Axis name	Result
ancestor	Selects all ancestors (parent, grandparent, etc.) of the current node
ancestor-or-self	Selects all ancestors (parent, grandparent, etc.) of the current node and the current node itself
attribute	Selects all attributes of the current node
child	Selects all children of the current node
descendant	Selects all descendants (children, grandchildren, etc.) of the current node
descendant-or-self	Selects all descendants (children, grandchildren, etc.) of the current node and the current node itself
following	Selects everything in the document after the closing tag of the current node
following-sibling	Selects all siblings after the current node
namespace	Selects all namespace nodes of the current node
parent	Selects the parent of the current node
preceding	Selects all nodes that appear before the current node in the document except ancestors, attribute nodes, and namespace nodes
preceding-sibling	Selects all siblings before the current node
self	Selects the current node

XPATH – PREDICATES

A predicate is a function that applies to a node's name, value, or attributes and that returns a logical *TRUE* or *FALSE*.

Predicates modify the path input of an XPath query. Nodes for which the relation is true are selected by the query.

Predicates are denoted by square brackets, placed after a node.

Examples:

- `"//p[position()=1]"` returns the first `<p>` node relative to its parent node
- `"//p[last()]"` returns the last `<p>` node relative to its parent node
- `"//div[count(./@*)>2]"` returns all `<div>` nodes with 2+ attributes, etc.

Function	Description	Example
<code>name(<node>)</code>	Returns the name of <node> or the first node in a node set	<code>//*[name()='title'];</code> Returns: <title>
<code>text(<node>)</code>	Returns the value of <node> or the first node in a node set	<code>//*[text()='The book homepage'];</code> Returns: <i> with value <i>The book homepage</i>
<code>@attribute</code>	Returns the value of a node's <i>attribute</i> or of the first node in a node set	<code>//div[@id='R Inventor'];</code> Returns: <div> with attribute <i>id</i> value <i>R Inventor</i>
<code>string-length(str1)</code>	Returns the length of <code>str1</code> . If there is no string argument, it returns the length of the string value of the current node	<code>//h1[string-length()>11];</code> Returns: <h1> with value <i>Robert Gentleman</i>
<code>translate(str1, str2, str3)</code>	Converts <code>str1</code> by replacing the characters in <code>str2</code> with the characters in <code>str3</code>	<code>//div[translate(./@date, '2003', '2005')='June/2005'];</code> Returns: first <div> node with date attribute value <i>June/2003</i>
<code>contains(str1, str2)</code>	Returns TRUE if <code>str1</code> contains <code>str2</code> , otherwise FALSE	<code>//div[contains(@id, 'Inventor')];</code> Returns: first <div> node with id attribute value <i>R Inventor</i>
<code>starts-with(str1, str2)</code>	Returns TRUE if <code>str1</code> starts with <code>str2</code> , otherwise FALSE	<code>//i[starts-with(text(), 'The')];</code> Returns: <i> with value <i>The book homepage</i>
<code>substring-before(str1, str2)</code>	Returns the start of <code>str1</code> before <code>str2</code> occurs in it	<code>//div[substring-before(@date, '/')='June'];</code> Returns: <div> with date attribute value <i>June/2003</i>
<code>substring-after(str1, str2)</code>	Returns the remainder of <code>str1</code> after <code>str2</code> occurs in it	<code>//div[substring-after(@date, '/')=2003];</code> Returns: <div> with date attribute value <i>June/2003</i>
<code>not(arg)</code>	Returns TRUE if the boolean value is FALSE, and FALSE if the boolean value is TRUE	<code>//div[not(contains(@id, 'Inventor'))];</code> Returns: the <div> node that does not contain the string <i>Inventor</i> in its id attribute value
<code>local-name(<node>)</code>	Returns the name of the current <node> or the first node in a node set—without the namespace prefix	<code>//*[local-name()='address'];</code> Returns: <address>
<code>count(<node>)</code>	Returns the count of a nodeset <node>	<code>//div[count(./a)=0];</code> Result: The second <div> with one <a> child
<code>position(<node>)</code>	Returns the index position of <node> that is currently being processed	<code>//div/p[position()=1];</code> Result: The first <p> node in each <div> node
<code>last()</code>	Returns the number of items in the processed node list <node>	<code>//div/p[last()];</code> Result: The last <p> node in each <div> node

UK GOV PRESS RELEASES – BACKGROUND

The United Kingdom Government publishes all of its press releases online, at gov.uk/government/announcements.

As of 29 March 2018, there were 65K+ press releases available on the site.

Questions:

- Can we predict which agency released an announcement based on its textual content alone?
- Are there topics that seem to return to the forefront over and over?



Announcements

You can use the filters to show only results that match your interests

Contains

Announcement type

All announcement types

Policy area

All policy areas

Department

All departments

Person

All people

World locations

All locations

65,716

 announcements

Get updates to this list  [email](#) [feed](#)

Welsh innovation is key to Britain's future export success

24 March 2018 WO Speech

Preventing Hunger as a Weapon of War

23 March 2018 FCO Speech

"Our vote today against this resolution is a vote against the politicization of the Commission on the Status of Women."

23 March 2018 FCO Speech

Lord Ahmad welcomes conclusions of the 37th Session of the UN Human Rights Council

23 March 2018 FCO Speech

Rt Hon Mark Field MP speech at Global FinTech Investor Forum

23 March 2018 FCO Speech

Foreign Secretary statement on Iran

The Foreign Secretary has made the following statement on the protests in Iran.

Published 1 January 2018

From: [Foreign & Commonwealth Office](#) and [The Rt Hon Boris Johnson MP](#)



The Foreign Secretary Boris Johnson said:

“ The UK is watching events in Iran closely. We believe that there should be meaningful debate about the legitimate and important issues the protesters are raising and we look to the Iranian authorities to permit this.”

“ We also believe that, particularly as we enter the 70th anniversary year of the Universal Declaration on Human Rights, people should be able to have freedom of expression and to demonstrate peacefully within the law.”

“ We regret the loss of life that has occurred in the protests in Iran, and call on all concerned to refrain from violence and for international obligations on human rights to be observed.”

Individual press releases contain:

- title
- date of publication
- publishing organisations/individuals
- text of the release

Focus on 2017, and releases from

- Wales Office
- Foreign Office
- Department of Science and Technology
- Department for Environment, Food & Rural Affairs

Notebook: UK Gov Press Releases

REGULAR EXPRESSIONS

Main task in web scraping is to collect **relevant** information for the research problem from lots of textual data.

We care about systematic elements of textual data, especially if quantitative methods are eventually going to be applied.

Systematic structures can be

- numbers
- names (countries, etc.)
- addresses (mailing, e-mailing, URLs, etc.)
- specific character strings, etc.

REGULAR EXPRESSIONS

Regular expressions (regexps) allow for the systematic extraction of the information components.

Regexps are abstract sequences of strings that match concrete recurring patterns in text.

Can be used to extract from plain text, HTML, and XML.

Useful when information is hidden within *atomic* values.

Notebook: Python Regular Expressions and More

BEAUTIFUL SOUP

Simple web requests require some networking code to fetch a page and return the HTML contents.

Browsers do a lot of work to intelligently parse totally improper HTML syntax, something like:

```
<a href="crummy.com"> <b>link text<a> </b>
```

Beautiful Soup is a Python library that helps extract data out of HTML and XML files. It parses HTML files, even if they're broken.

BEAUTIFUL SOUP

BS does not simply convert bad HTML to good XHTML, to be parsed with an XML parser.

BS allows a user to fully inspect the (proper) HTML structure it produces, programmatically.

When BS is done its work on an HTML file, the result is an API for traversing, searching, and reading the document's elements.

BEAUTIFUL SOUP

Typical HTML elements to be extracted/read come in various formats:

- text
- tables
- form field values
- images
- videos
- etc.

It provides **idiomatic** ways of navigating, searching, modifying the parse tree of the HTML file (it's a huge time-saver).

```
html_doc = """
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>

<p class="story">Once upon a time there were three little sisters; and their names were
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.</p>

<p class="story">...</p>
"""
```

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')

print(soup.prettify())
# <html>
# <head>
# <title>
#   The Dormouse's story
# </title>
# </head>
# <body>
# <p class="title">
#   <b>
#     The Dormouse's story
#   </b>
# </p>
# <p class="story">
#   Once upon a time there were three little sisters; and their names were
#   <a class="sister" href="http://example.com/elsie" id="link1">
#     Elsie
#   </a>
#   ,
#   <a class="sister" href="http://example.com/lacie" id="link2">
#     Lacie
#   </a>
#   and
#   <a class="sister" href="http://example.com/tillie" id="link2">
#     Tillie
#   </a>
#   ; and they lived at the bottom of a well.
# </p>
# <p class="story">
#   ...
# </p>
# </body>
# </html>
```

```
soup.title
# <title>The Dormouse's story</title>

soup.title.name
# u'title'

soup.title.string
# u'The Dormouse's story'

soup.title.parent.name
# u'head'

soup.p
# <p class="title"><b>The Dormouse's story</b></p>

soup.p['class']
# u'title'

soup.a
# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>

soup.find_all('a')
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

soup.find(id="link3")
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>
```



```
for link in soup.find_all('a'):
    print(link.get('href'))
# http://example.com/elsie
# http://example.com/lacie
# http://example.com/tillie
```

```
print(soup.get_text())
# The Dormouse's story
#
# The Dormouse's story
#
# Once upon a time there were three little sisters; and their names were
# Elsie,
# Lacie and
# Tillie;
# and they lived at the bottom of a well.
#
# ...
```


SELENIUM

Selenium is a tool to automate web browser interactions (in Python). It is used primarily for automating web applications for testing purposes, but it has other applications.

Mainly, it allows the user to open a browser and to perform tasks as a human being would, such as:

- clicking buttons
- entering information in forms
- searching for specific information on the web pages
- etc.

SELENIUM

Selenium requires a driver to interface with the chosen browser. Firefox, for example, requires *geckodriver*.

Other supported browsers will have their own drivers available:

Chrome: <https://sites.google.com/a/chromium.org/chromedriver/downloads>

Edge: <https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/>

Firefox: <https://github.com/mozilla/geckodriver/releases>

Safari: <https://webkit.org/blog/6900/webdriver-support-in-safari-10/>

SIMULATING A WEB BROWSER

Selenium automatically controls a complete browser, including rendering the web documents and running JavaScript.

This is useful for pages with a lot of dynamic content that isn't in the base HTML.

Selenium can program actions like "click on this button", or "type this text", and at any point you have access to the dynamic HTML of the current state of the page, like what you see in Developer Tools.

USING APIS

An API is a website's way of giving programs access to their data, without the need for scraping.

That is, an API provides **structured access** to **structured data**.

For example, a finance site might offer an API with financial data, or the *New York Times* might offer an API for news articles.

In either case, the data will be in a pre-defined, structured format (often JSON).

USING APIS

The APIs we'll consider have R/Python libraries that encapsulate all required networking and encoding.

This means that it suffices to read the library documentation to know what to do.

Exercise: Use Zomato to find which Canadian city has the best sushi restaurants (<https://github.com/fatihsucu/pyzomato>).

YOUTUBE API – KHAN ACADEMY

Millions of videos are available through YouTube.

It's not obvious how one would scrape video content off the web in general (other than the URLs); some videos have associated text content (**transcripts**).

We use the YouTube API to scrape that content.

Notebook: YouTube Transcripts



Home



Trending



History

BEST OF YOUTUBE



Music



Sports



Gaming



Movies



TV Shows



News



Live



360° Video



Browse channels

Sign in now to see your channels and recommendations!

SIGN IN



Statistics

68 videos • 3,290,303 views • Last updated on Jul 2, 2014



Khan Academy

SUBSCRIBE

Introduction to statistics. Will eventually cover all of the major topics in a first-year statistics course (not there yet!)

1



Statistics: The average | Descriptive statistics | Probability and Statistics | Khan Academy

Khan Academy

2



Statistics: Sample vs. Population Mean

Khan Academy

3



Statistics: Variance of a population | Probability and Statistics | Khan Academy

Khan Academy

4



Statistics: Sample variance | Descriptive statistics | Probability and Statistics | Khan Academy

Khan Academy

5



Statistics: Standard deviation | Descriptive statistics | Probability and Statistics | Khan Academy

Khan Academy

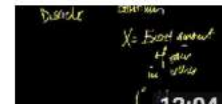
6



Statistics: Alternate variance formulas | Probability and Statistics | Khan Academy

Khan Academy

7



Introduction to Random Variables

Khan Academy

DATA WRANGLING

DATA COLLECTION AND DATA PROCESSING

LEARNING OBJECTIVES

Become familiar with the tidy data format

Increase familiarity of data wrangling functions

Identify R packages that facilitate data processing

DATA WRANGLING

A fair amount of time (up to 80%, perhaps) must be spent on data processing (both cleaning and manipulation).

The main goals of **data wrangling** are to:

- make the data useable by a specific piece of software
- reveal pre-analysis insights in the data

TIDY DATA

Tidy data has a specific structure:

- each variable is a column
- each observation is a row
- each type of observational unit is a table

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

FUNCTIONALITY

Data wrangling functions should allow the analyst to:

- extract a subset of variables from the data frame
- extract a subset of observations from the data frame
- sort the data frame along any combination of variables in increasing or decreasing order
- to create new variables from existing variables
- to create (so-called) pivot tables, by observation groups
- database functionality (joins, etc.)
- etc.

FUNCTIONALITY

In R, this can be achieved in various ways. Current favoured packages include:

- `tidyr`
- `dplyr` (data transformation)
- `lubridate` (dates and times)
- `stringr` (string manipulation)
- `purrr` (functions)
- `readr` (data import)

For equivalent Python modules, consult Kazil & Jarmul's *Data Wrangling with Python*.

EXERCISES

What would the following dataset look like in a tidy format?

storm	stat	value
Alex	wind	68
Alex	pressure	130
Allison	wind	55
Allison	pressure	121
Bobbie	wind	72
Bobbie	pressure	118

EXERCISES

How would you go from the table on the left to the table on the right?

storm	stat	value
Alex	wind	68
Alex	pressure	130
Allison	wind	55
Allison	pressure	121
Bobbie	wind	72
Bobbie	pressure	118

stat	mean	std dev
wind	65	8.9
pressure	123	6.2

EXERCISES

Run section 9 of the notebook `CSPS_04_R_Basics.ipynb` to explore how the packages `tidyr` and `dplyr` help the process of data wrangling in R.

EXERCISES

Turn the data found in `cities.txt` into a tidy dataset.

DATA CLEANING

DATA COLLECTION AND DATA PROCESSING

“Obviously, the best way to treat missing data is not to have any.”

T. Orchard, M. Woodbury

“The most exciting phrase to hear, the one that heralds the most discoveries, is not “Eureka!” but “That's funny...”.”

LEARNING OBJECTIVES

Recognize the strengths and weaknesses of both major data cleaning approaches

Identify methods to handle missing observations

Increase familiarity with various anomaly detection or outlier tests

FOUR VERY IMPORTANT REMARKS

NEVER work on the original dataset. Make copies along the way.

Document **ALL** your cleaning steps and procedures.

If you find yourself cleaning too much of your data, **STOP**. Something might be off with the data collection procedure.

Think **TWICE** before discarding an entire record.

APPROACHES TO DATA CLEANING

There are two **philosophical** approaches to data cleaning and validation:

- methodical
- narrative

The **methodical** approach consists of running through a **check list** of potential issues and flagging those that apply to the data.

The **narrative** approach consists of **exploring** the dataset and trying to spot unlikely and irregular patterns.

TAKE-AWAYS

The narrative approach is similar to working out a crossword puzzle with a pen and putting down potentially wrong answers every once in a while to see where that takes you.

The mechanical approach is similar to working it out with a pencil, a dictionary, and never jotting down an answer unless you are certain it is correct.

You'll solve more puzzles (and it will be flashier) the first way, but you'll rarely be wrong the second way.

Be comfortable with both approaches.

TYPES OF MISSING OBSERVATIONS

Blank fields come in 4 flavours:

- **Nonresponse**
an observation was expected but none had been entered
- **Data Entry Issue**
an observation was recorded but was not entered in the dataset
- **Invalid Entry**
an observation was recorded but was considered invalid and has been removed
- **Expected Blank**
a field has been left blank, but expectedly so

TYPES OF MISSING OBSERVATIONS

Too many missing values (of the first three type) can be indicative of **issues with the data collection process** (more on this later).

Too many missing values (of the fourth type) can be indicative of **poor questionnaire design**.

THE CASE FOR IMPUTATION

Not all analytical methods can easily accommodate missing observations.

There are two options:

- **Discard** the missing observation
 - not recommended, unless the data is missing completely randomly in the dataset as a whole
 - acceptable in certain situations (such as a small number of missing values in a large dataset)
- Come up with a **replacement value**
 - main drawback: we never know for a fact what the true value would have been
 - often the best available option

MISSING MECHANISMS

Missing Completely at Random (MCAR)

- item absence is independent of its value or of auxiliary variables

Missing at Random (MAR)

- item absence is not completely random; can be accounted by auxiliary variables with complete info

Not Missing at Random (NMAR)

- reason for nonresponse is related to item value (also called **non-ignorable non-response**)

IMPUTATION METHODS

List-wise deletion

Mean or most frequent imputation

Regression or correlation imputation

Stochastic regression imputation

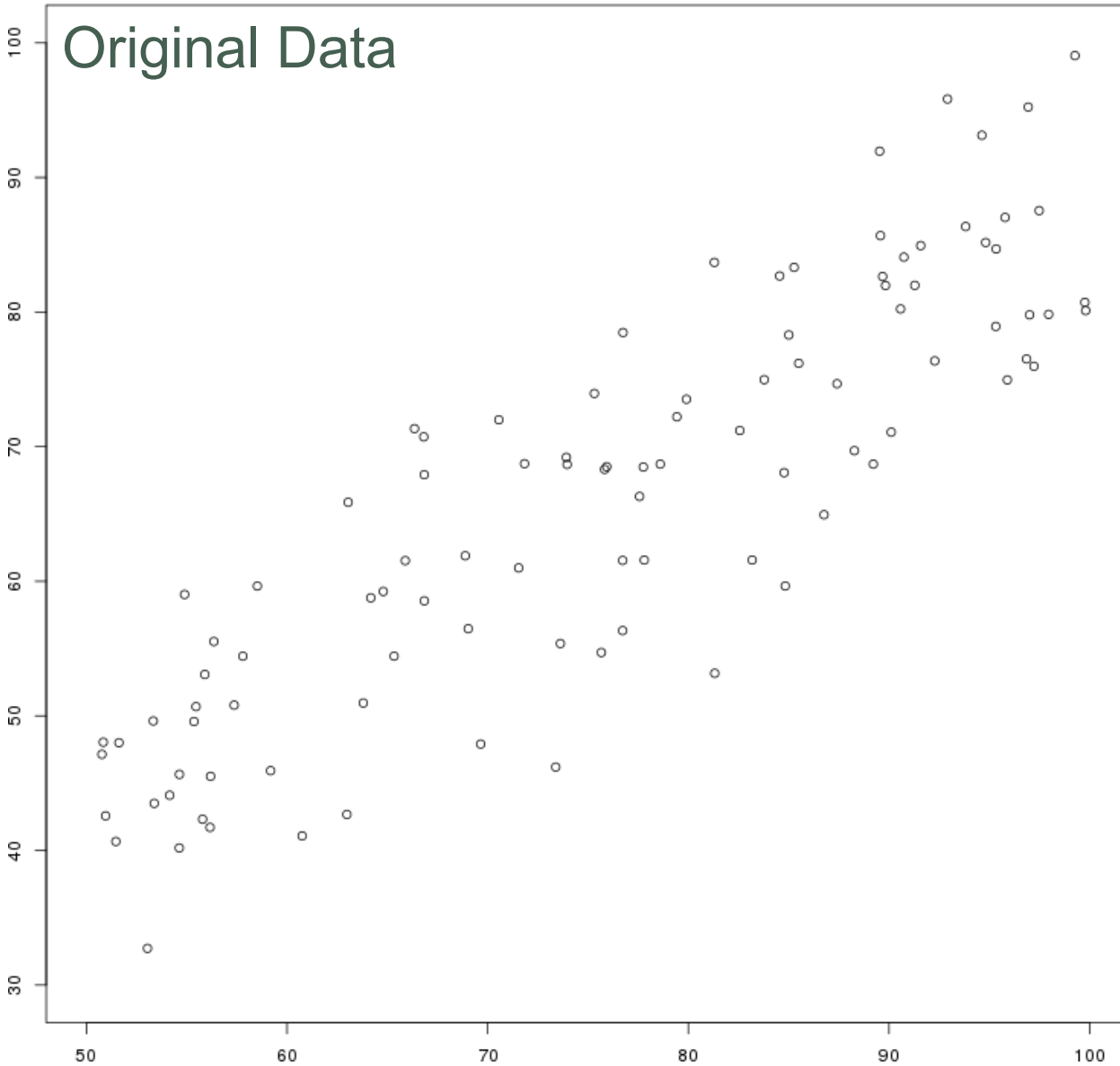
Last observation carried forward

k -nearest neighbours imputation

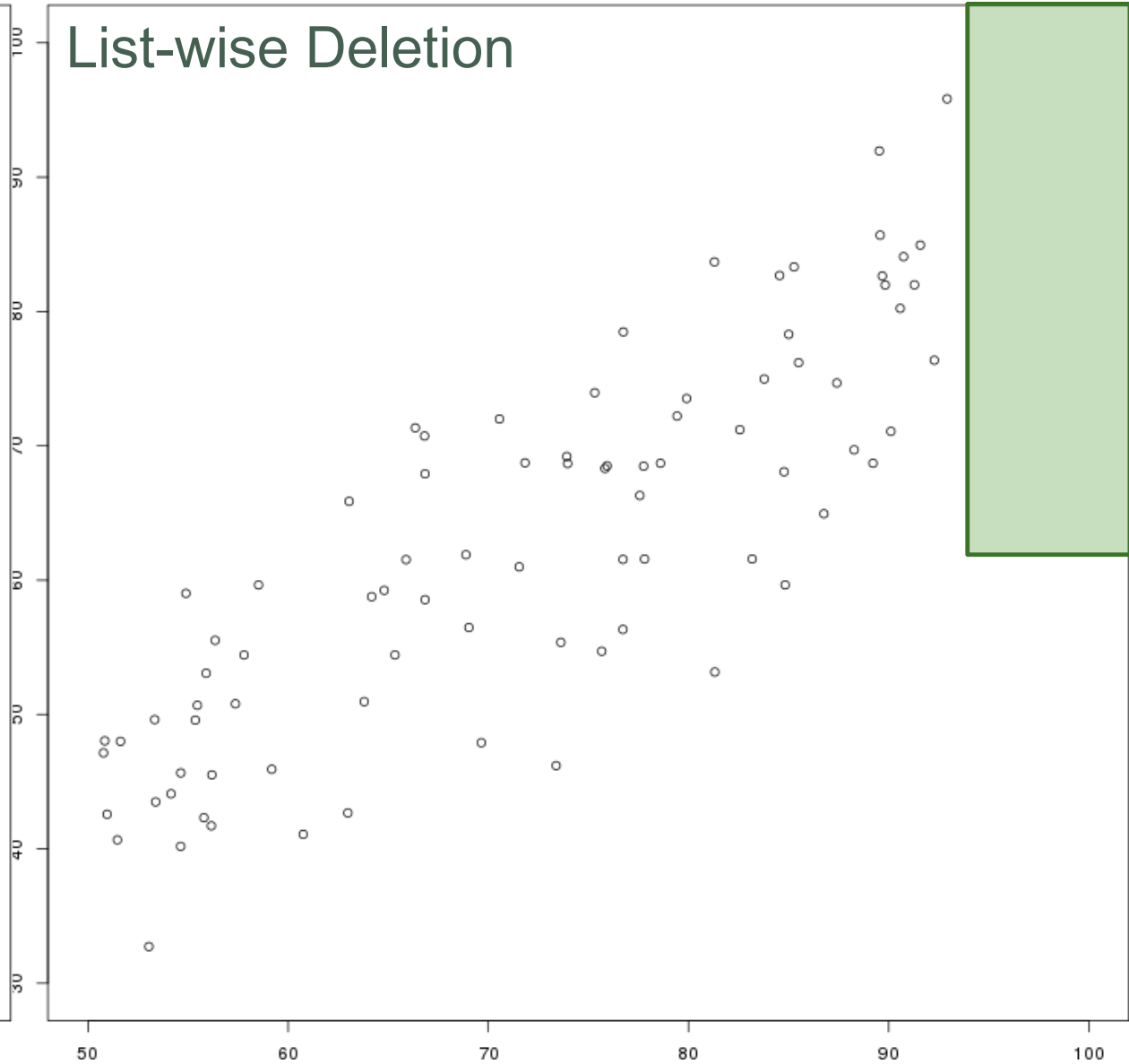
Multiple imputation

Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.

Original Data

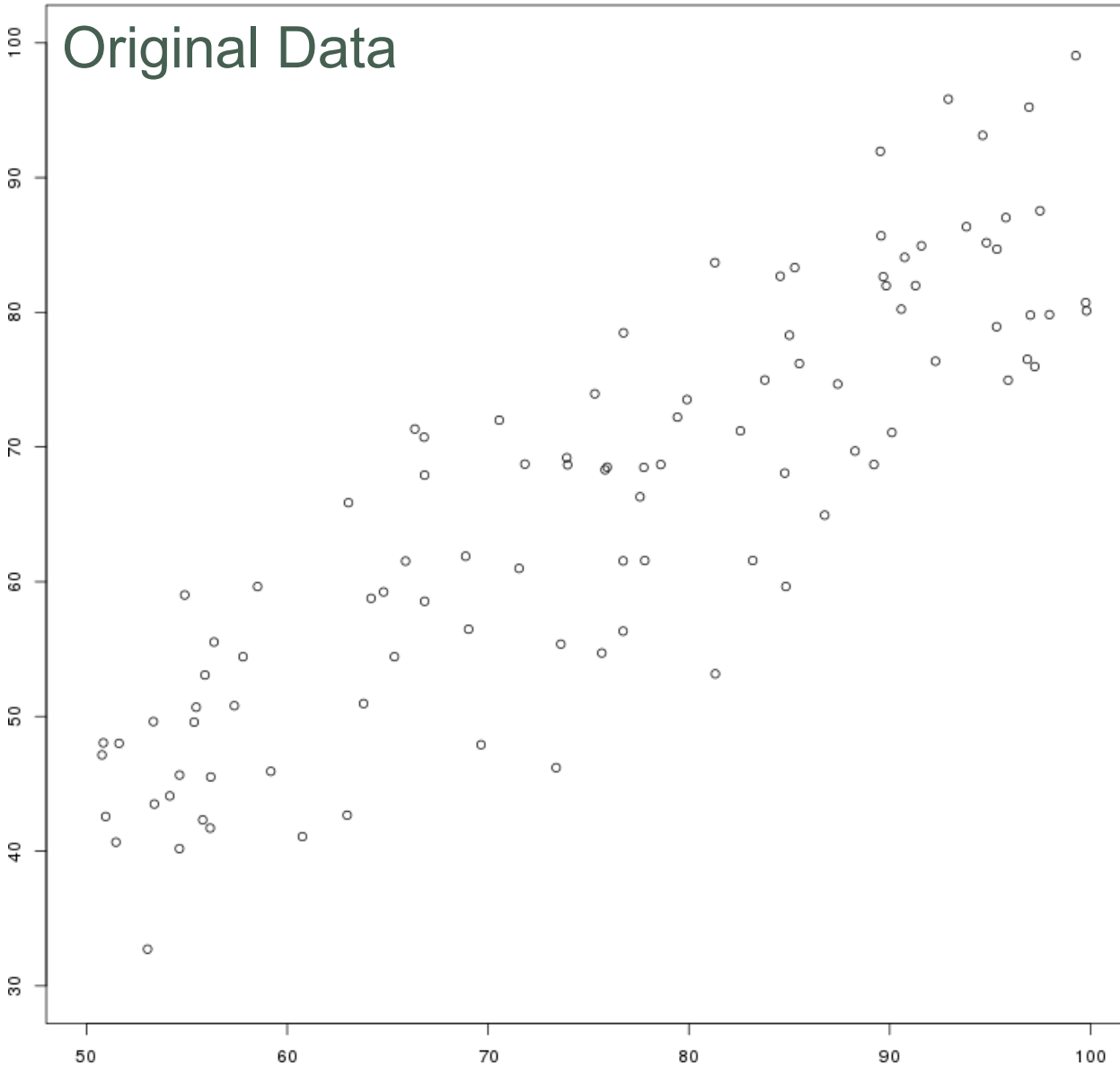


List-wise Deletion

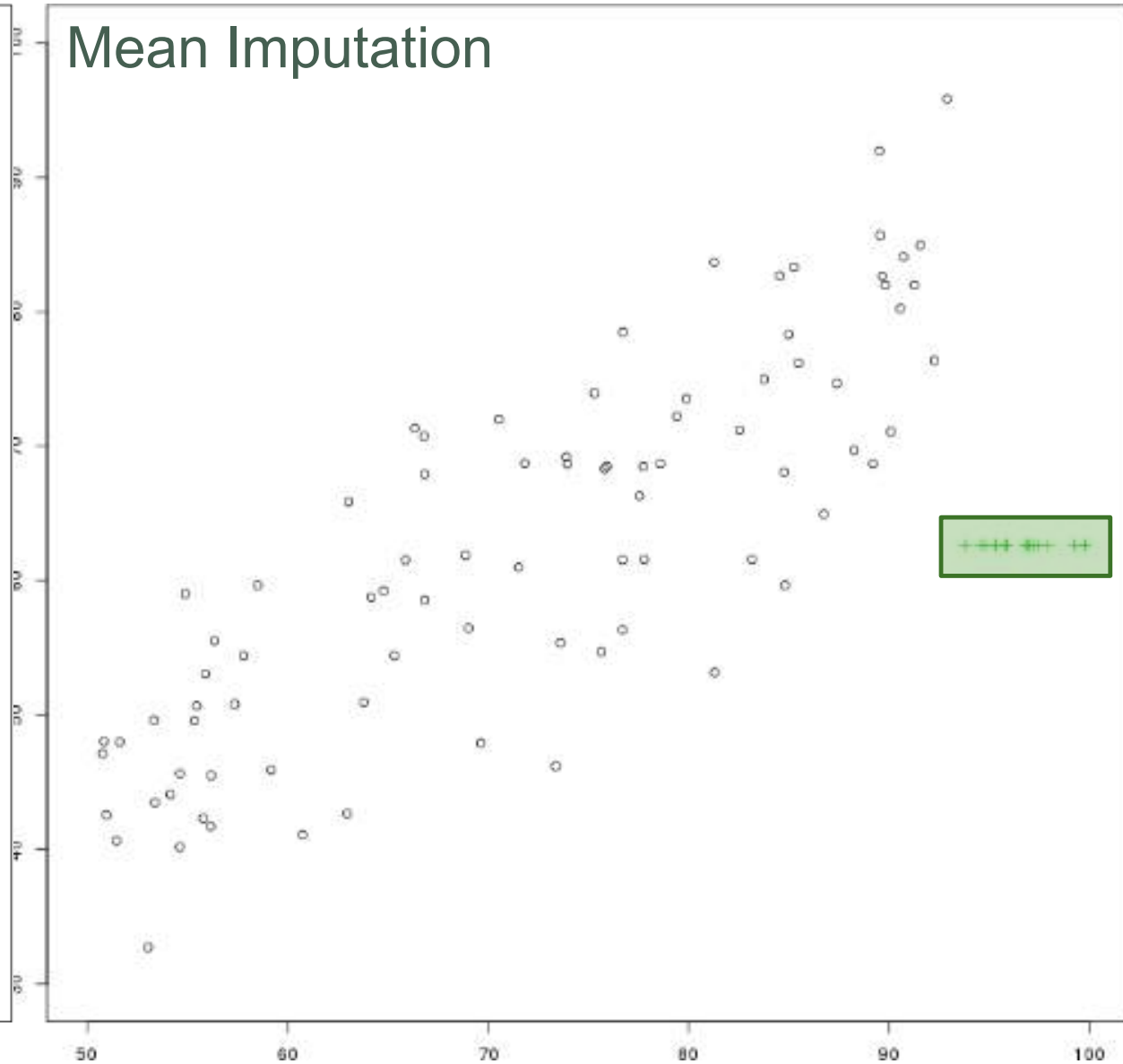


Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.

Original Data

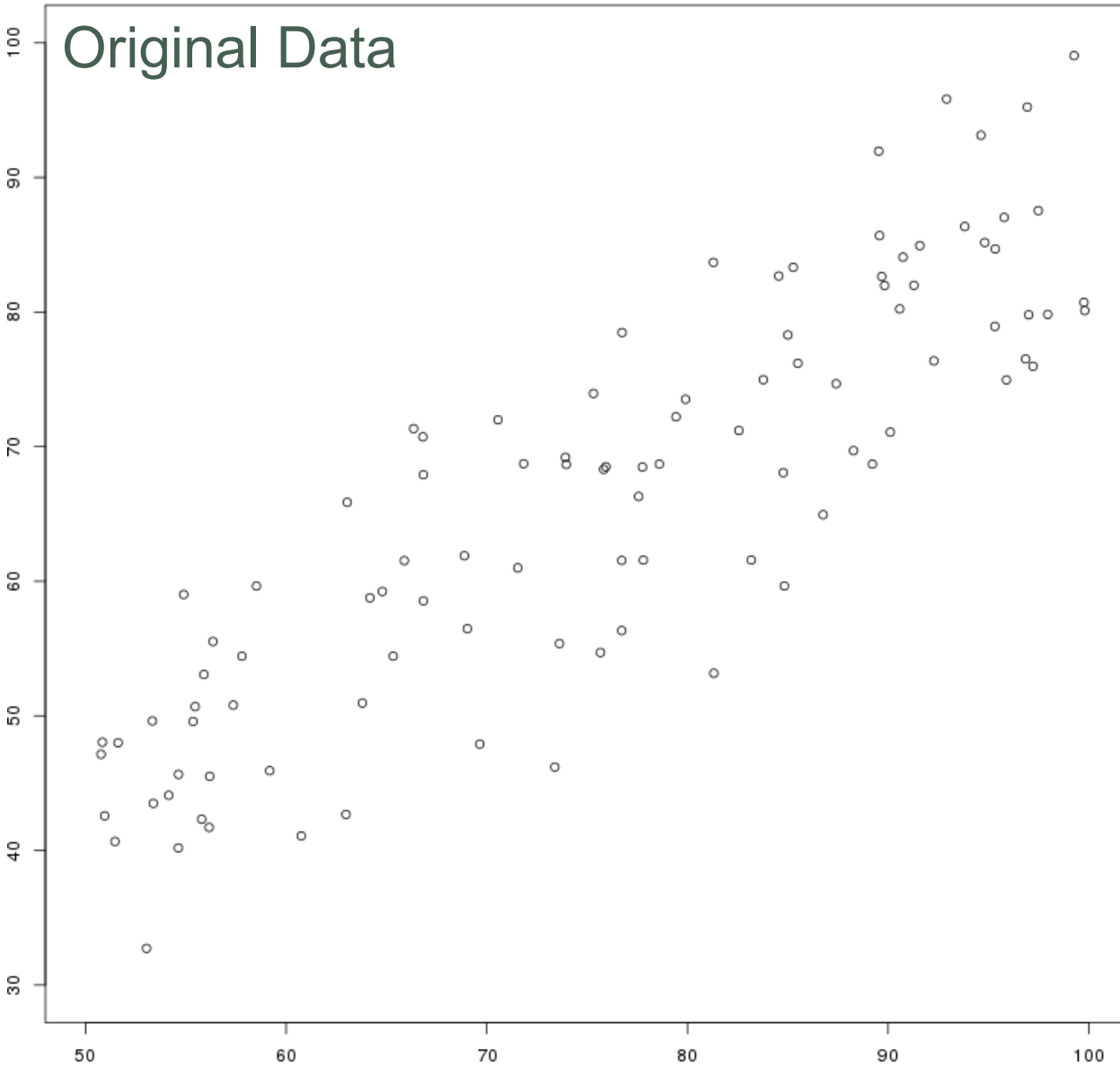


Mean Imputation

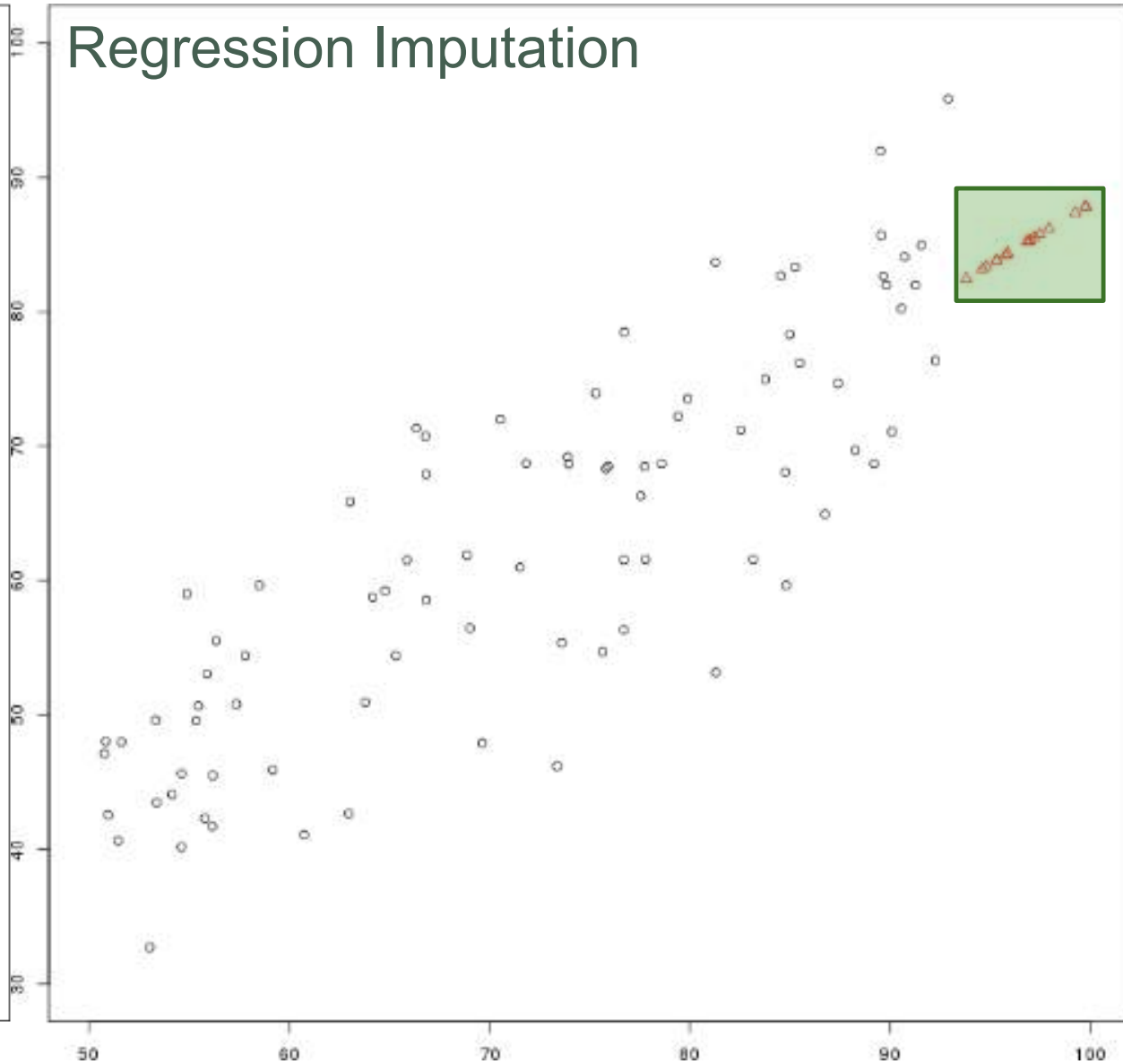


Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.

Original Data

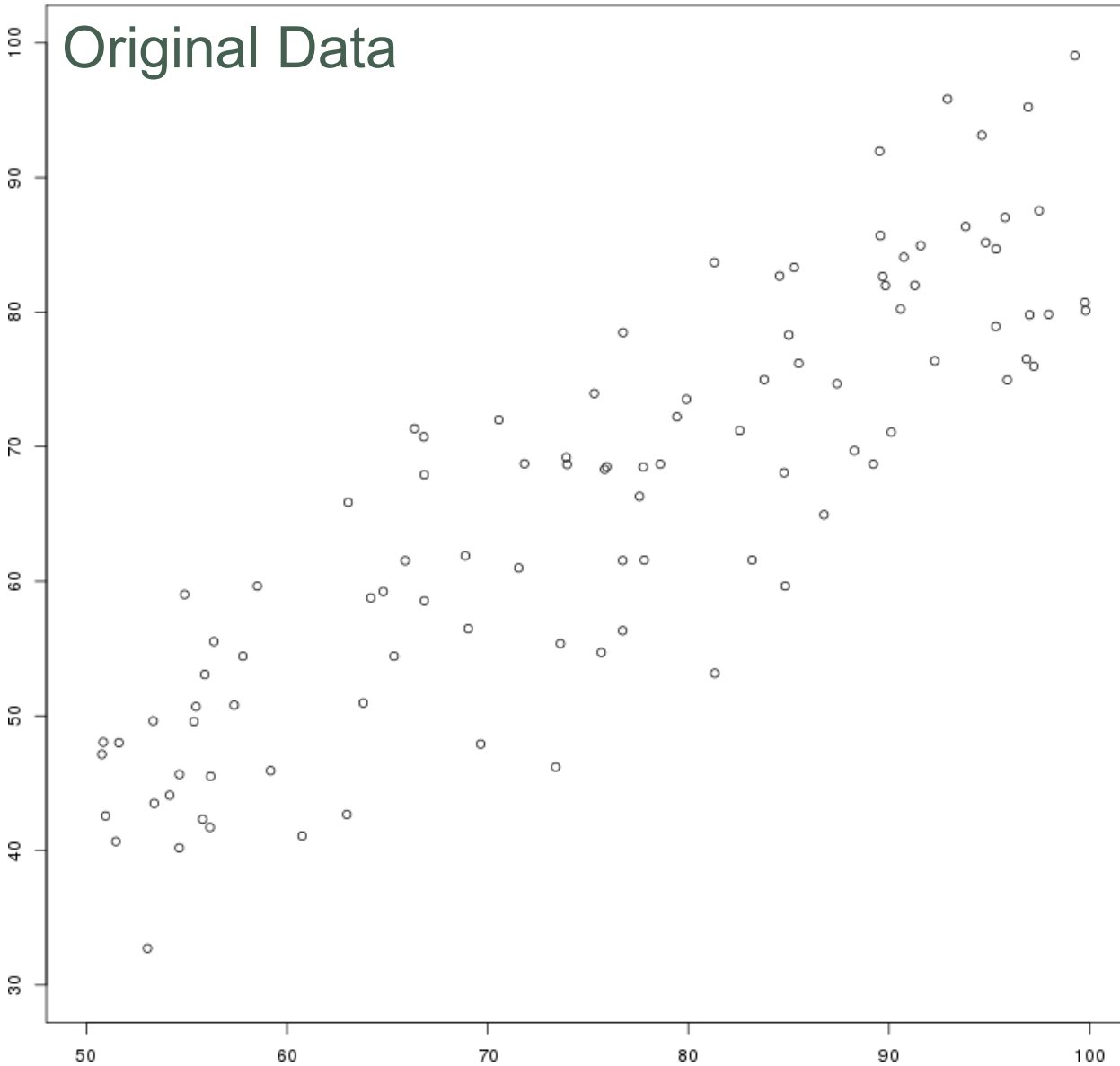


Regression Imputation

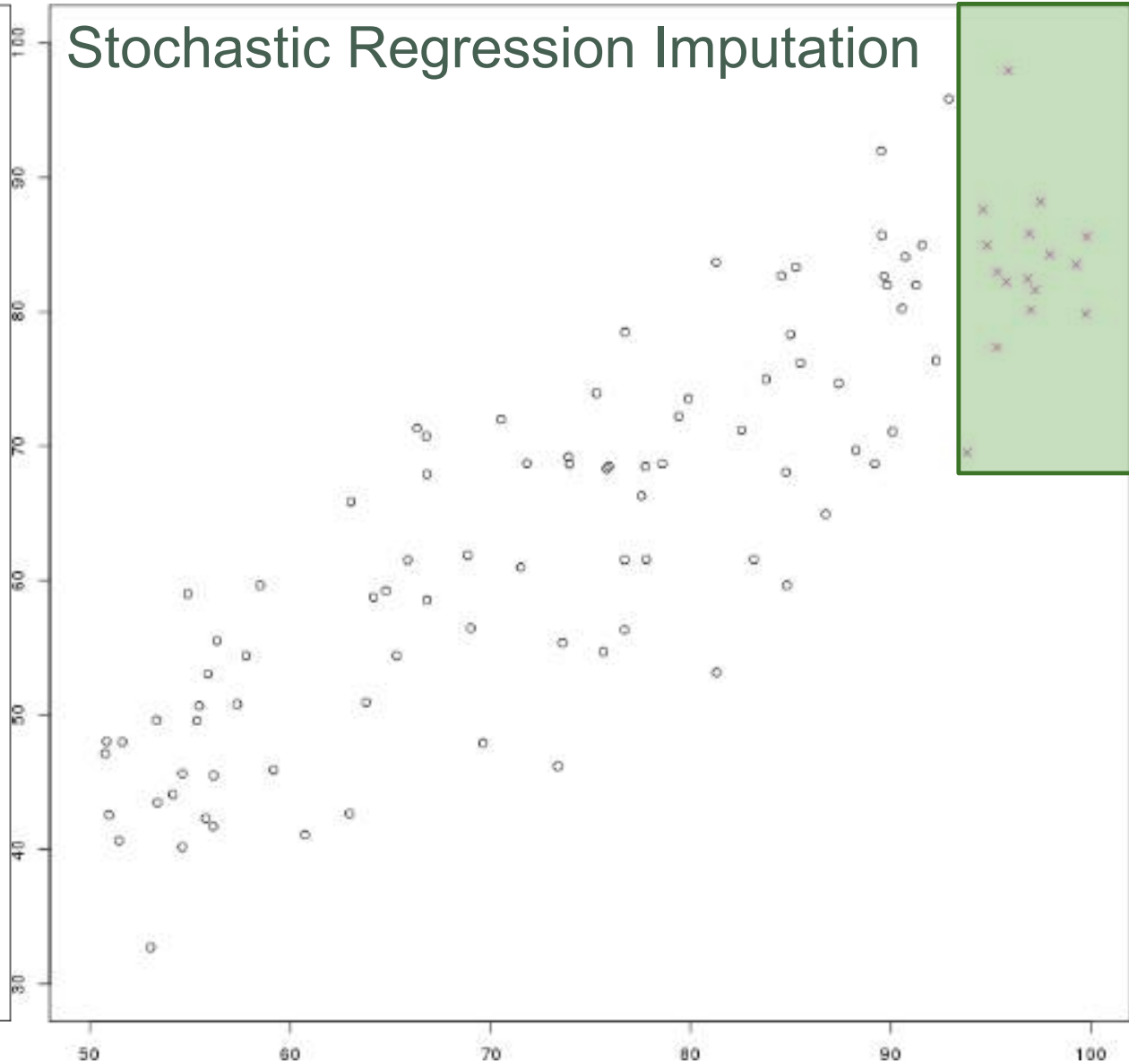


Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.

Original Data



Stochastic Regression Imputation



MULTIPLE IMPUTATION

Imputations increase the noise in the data.

In **multiple imputation**, the effect of that noise can be measured by consolidating the analysis outcome from multiple imputed datasets.

Steps:

1. Repeated imputation creates m versions of the dataset.
2. Each of these datasets is analyzed, yielding m outcomes.
3. The m outcomes are pooled into a single result for which the mean, variance, and confidence intervals are known.

MULTIPLE IMPUTATION

Advantages

- **flexible**; can be used in a various situations (MCAR, MAR, even NMAR in certain cases).
- accounts for **uncertainty** in imputed values
- fairly easy to implement

Disadvantages

- m may need to be fairly **large** when there are many missing values in numerous features, which slows down the analyses
- what happens if the analysis output is not a single value but some more complicated mathematical object?

TAKE-AWAYS

Missing values cannot simply be ignored.

The missing mechanism cannot typically be determined with any certainty.

Imputation methods work best when values are missing completely at random or missing at random, but imputation methods tend to produce biased estimates.

In single imputation, imputed data is treated as the actual data; multiple imputation can help reduce the noise.

Is stochastic imputation best? In our example, yes – but beware the *No-Free Lunch* theorem!

SPECIAL DATA POINTS

Outlying observations are data points which are **atypical** in comparison to

- the unit's remaining features (*within-unit*),
- the field measurements for other units (*between-units*),

or as part of a collective subset of observations.

Outliers are observations which are **dissimilar to other cases** or which **contradict known dependencies** or rules.

Careful study is needed to determine whether outliers should be retained or removed from the dataset.

SPECIAL DATA POINTS

Influential data points are observations whose absence leads to **markedly different** analysis results.

When influential observations are identified, remedial measures (such as data transformations) may be required to minimize their undue effects.

Outliers may be influential data points, yet influential data points need not be outliers (weighted data).

DETECTING ANOMALIES

Outliers may be anomalous along any of the unit's variables, or in combination.

Anomalies are by definition **infrequent**, and typically shrouded in **uncertainty** due to small sample sizes.

Differentiating anomalies from noise or data entry errors is **hard**.

Boundaries between normal and deviating units may be **fuzzy**.

When anomalies are associated with malicious activities, they are typically **disguised**.

DETECTING ANOMALIES

Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used.

Graphical methods are easy to implement and interpret.

- **Outlying Observations**

- box-plots, scatterplots, scatterplot matrices, 2D tour, Cooke's distance, normal qq plots

- **Influential Data**

- some level of analysis must be performed (leverage)

Once anomalous observations have been removed from the dataset, previously “regular” units may become anomalous.

OUTLIER TESTS

Supervised methods use a historical record of labeled anomalous observations:

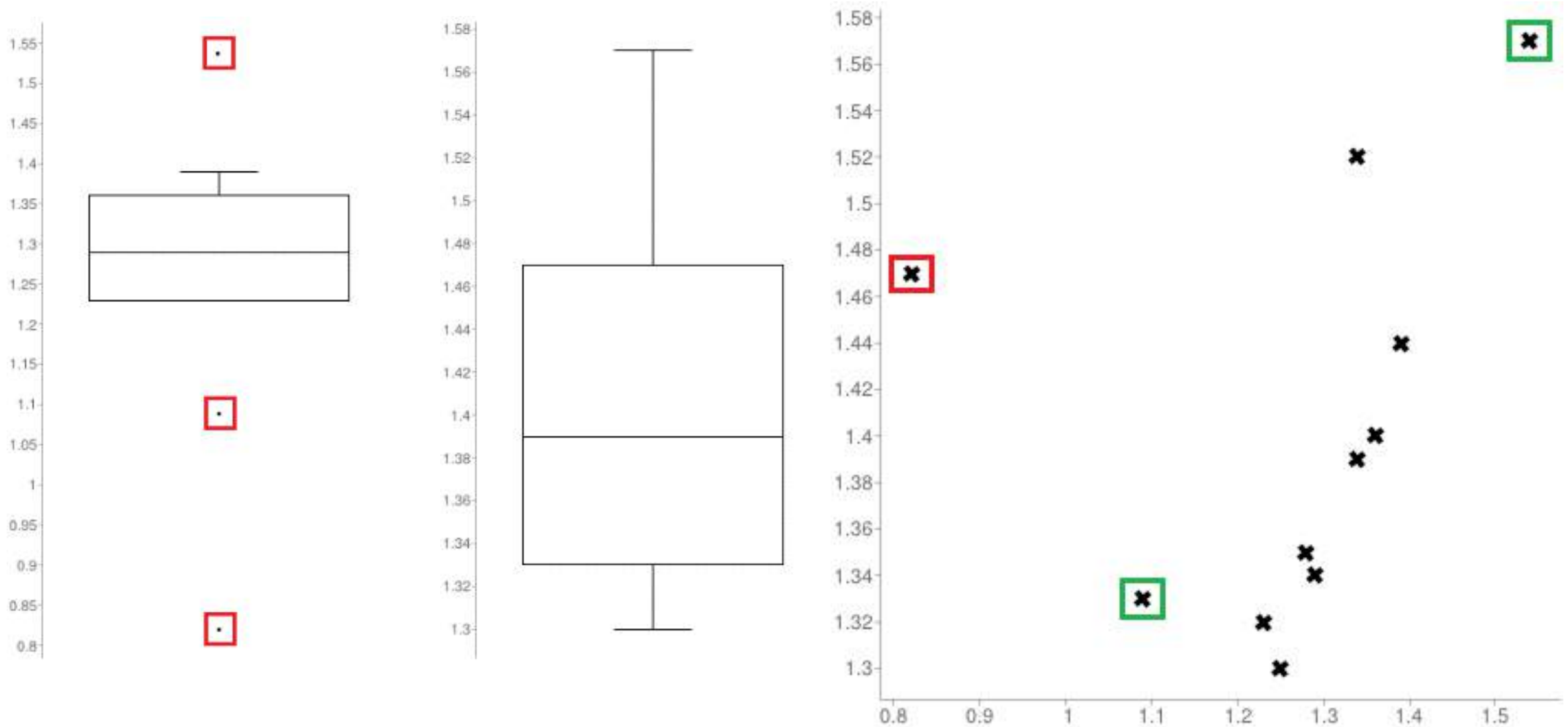
- domain expertise required to tag the data
- classification or regression task (probabilities and inspection rankings)
- rare occurrence problem (more on this later)

Unsupervised methods don't use external information:

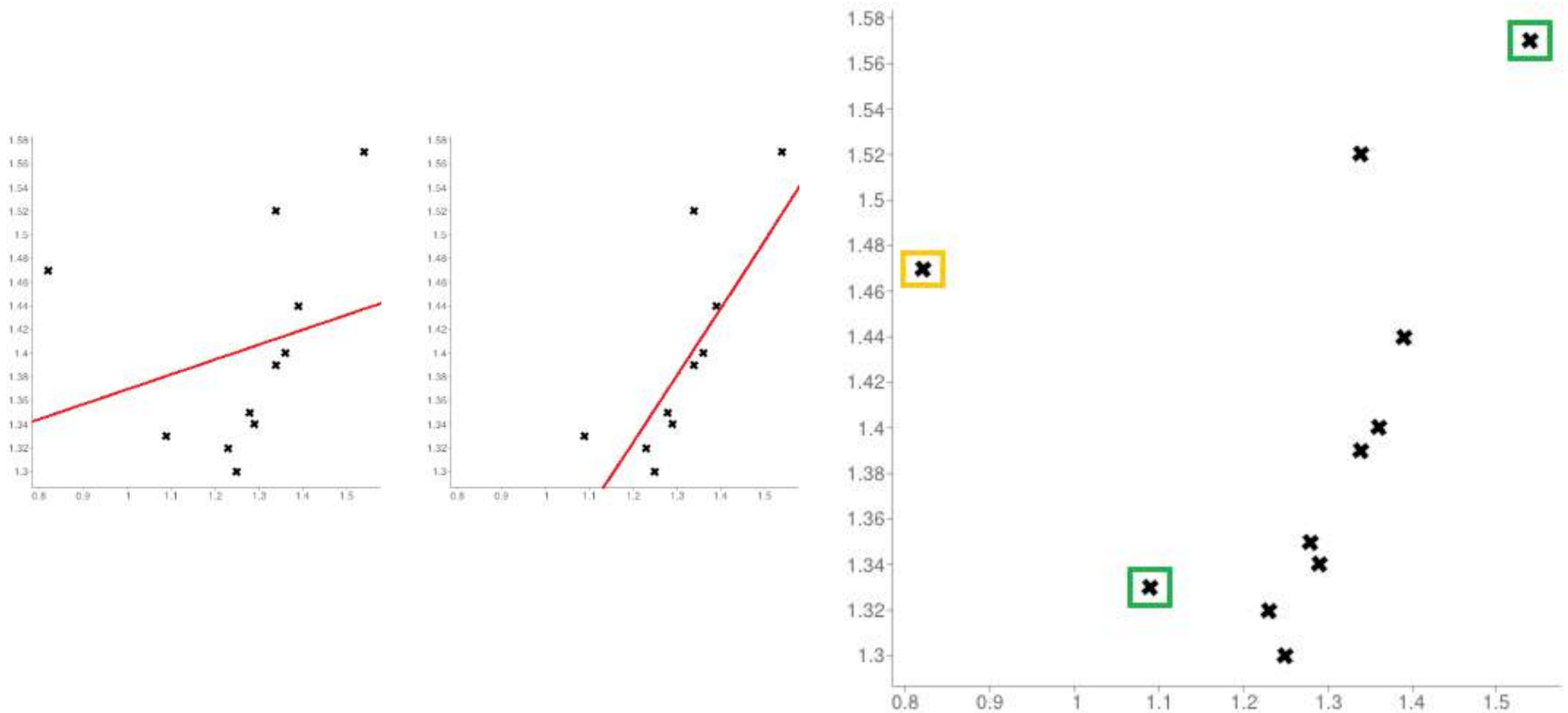
- traditional methods and tests
- can also be seen as a clustering or association rules problem

Semi-supervised methods also exist.

Queuing dataset: processing rate vs. arrival rate



Queuing dataset: processing rate vs. arrival rate



TAKE-AWAYS

Identifying influential points is an iterative process as the various analyses have to be run numerous times.

Fully automated identification and removal of anomalous observations is NOT recommended.

Use transformations if the data is NOT normally distributed.

Whether an observation is an outlier or not depends on various factors; what observations end up being influential data points depends on the specific analysis to be performed.

EXERCISES

The ability to monitor and perform early forecasts of various river algae blooms is crucial to control any ecological harm they can cause.

The `algae_bloom.csv` dataset is used to train a learning model consists of:

- **chemical properties** of various water samples of European rivers
- the **quantity of seven algae** in each of the samples, and
- the **characteristics of the collection process** for each sample.

What is the data science motivation for such a model, given that we **can** actually analyze water samples to determine if various harmful algae are present or absent?

EXERCISES

The answer is simple: chemical monitoring is **cheap** and **easy to automate**, whereas biological analysis of samples is **expensive** and **slow**.

Another answer is that analyzing the samples for harmful content does not provide a better understanding of algae bloom **drivers**: it just tells us which samples contain the harmful algae.

Can our model provide a more thorough understanding of the algae situation?

EXERCISES

Locate and determine the structure of the algae bloom dataset, and provide a summary of its features.

Compute the number of missing values for each record.

Identify some potential anomalous observations in the same dataset.

What strategies could you use to deal with such observations / records?

Supplemental Material

PROS AND CONS

Methodical (syntax)

- Pros: checklist is **context-independent**; pipelines **easy to implement**; common errors and invalid observations **easily identified**
- Cons: may prove **time-consuming**; cannot identify new types of errors

Narrative (semantics)

- Pros: process may simultaneously yield **data understanding**; false starts are (at most) as costly as switching to mechanical approach
- Cons: may miss important sources of errors and invalid observations for datasets with **high number of features**; domain knowledge may bias the process by neglecting uninteresting areas of the dataset

TOOLS AND METHODS

Methodical

- list of potential problems (Data Cleaning Bingo)
- code which can be re-used in different contexts

Narrative

- visualization
- data summary
- distribution tables
- small multiples
- data analysis

Data Cleaning Bingo

random'missing' values	outliers	values'outside'of' expected'range'4 numeric	factors' incorrectly/inconsiste ntly'coded	date/time'values'in' multiple'formats
impossible'numeric' values	leading'or'trailing' white'space	badly'formatted' date/time'values	non4random'missing' values	logical' inconsistencies' across'fields
characters'in' numeric'field	values'outside'of' expected'range'4 date/time	DCB!	inconsistent'or'no' distinction'between' null,'0,'not'available,' not' applicable,missing	possible'factors' missing
multiple'symbols' used'for'missing' values	???	fields'incorrectly' separated'in'row	blank'fields	logical'iconsistencies' within'field
entire'blank'rows	character'encoding' issues	duplicate'value'in' unique'field	non4factor'values'in' factor	numeric'values'in' character'field

OUTLIER TESTS

Normality is an assumption for most tests.

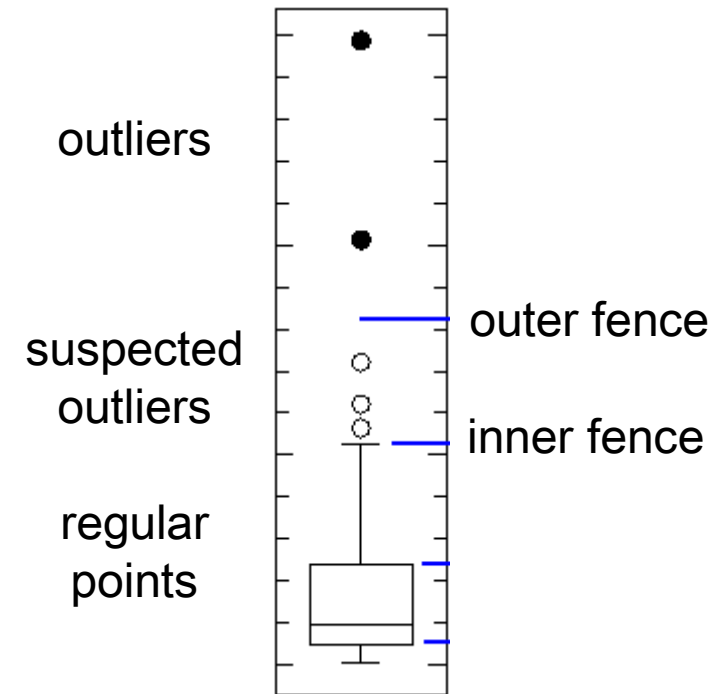
Tukey's Boxplot test: for normally distributed data, regular observations typically lie between the **inner fences**

$$Q_1 - 1.5 \times (Q_3 - Q_1) \text{ and } Q_3 + 1.5 \times (Q_3 - Q_1).$$

Suspected outliers lie between the inner fences and the outer fences

$$Q_1 - 3 \times (Q_3 - Q_1) \text{ and } Q_3 + 3 \times (Q_3 - Q_1).$$

Outliers lie beyond the outer fences.



OUTLIER TESTS

The **Grubbs Test** is a univariate test. Consider

- x_i : value of feature X for i^{th} unit, $1 \leq i \leq N$
- \bar{x} : mean value of feature X
- s_x : standard deviation of feature X
- α : significance level
- $T(\alpha, N)$: value of the t -distribution at significance $\alpha/2N$

The i^{th} unit is an **outlier along feature X** if

$$|x_i - \bar{x}| \geq \frac{s_x(N-1)}{\sqrt{N}} \times \sqrt{\frac{T^2(\alpha, N)}{N-2+T^2(\alpha, N)}}$$

OUTLIER TESTS

The **Dixon Q Test** is used in experimental sciences to find outliers in (extremely) small datasets (dubious validity).

The **Mahalanobis Distance** (linked to the leverage) can be used to find multi-dimensional outliers (when relationships are linear).

Other tests:

- **Tietjen-Moore** (for a specific # of outliers)
- **generalized extreme studentized deviate** (for unknown # of outliers)
- **chi-square** (outliers affecting goodness-of-fit)
- **DBSCAN, OR_h and LOF** (unsupervised outlier detection)

IMPUTATION METHODS

List-wise deletion: remove units with at least one missing values.

- Assumption: MCAR
- Cons: can introduce bias (if not MCAR), reduction in sample size, increase in standard error

Mean or Most Frequent Imputation: substitute missing values by average value or most frequent value

- Assumption: MCAR
- Cons: distortions of distribution (spike at mean) and relationships among variables

IMPUTATION METHODS

Regression or Correlation Imputation: substitute missing values by using regression based on other variables (with complete information)

- Assumption: MAR
- Cons: artificial reduction in variability, over-estimation of correlation

Stochastic Regression Imputation: regression imputation with random error terms added

- Assumption: MAR
- Cons: increased risk of type I error (false positives) due to small std error

IMPUTATION METHODS

Last Observation Carried Forward (LOCF): substitute the missing values with previous values (in a longitudinal study)

- Assumption: MCAR, values do not vary greatly over time
- Cons: may be too “generous”, depending on the nature of study

k -Nearest-Neighbour Imputation (k NN): substitute the missing entry with the average from the group of the k most **similar** complete respondents

- Assumption: MAR
- Cons: difficult to choose appropriate value for k . Possible distortion in data structure ($k > 1$)

DATA REDUCTION AND TRANSFORMATIONS

DATA COLLECTION AND DATA PROCESSING

LEARNING OBJECTIVES

Familiarity with the following concepts:

- Dimensionality of data
- Curse of Dimensionality
- Feature selection
- Principal Component Analysis (PCA)
- Data transformation
- Scaling
- Discretization

DIMENSIONALITY OF DATA

In data analysis, the **dimension** of the data is the number of variables (or attributes) that are collected in a dataset, represented by the number of columns.

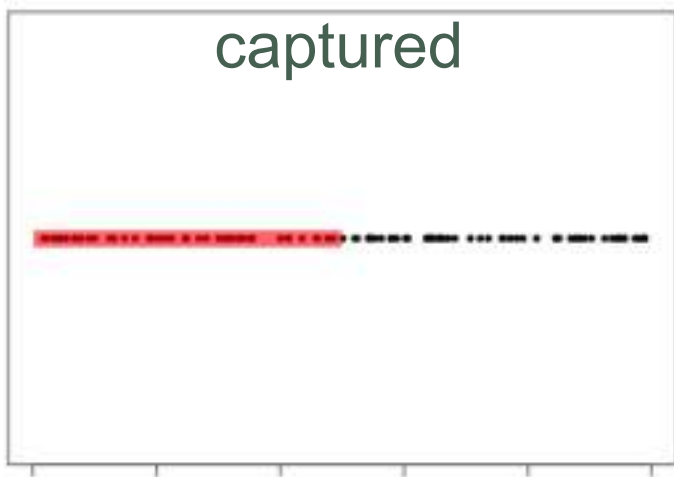
Here the term dimension is an extension of the use of the term to refer to the size of a vector.

We can think of the number of variables used to describe each object (row) as a vector describing that object.

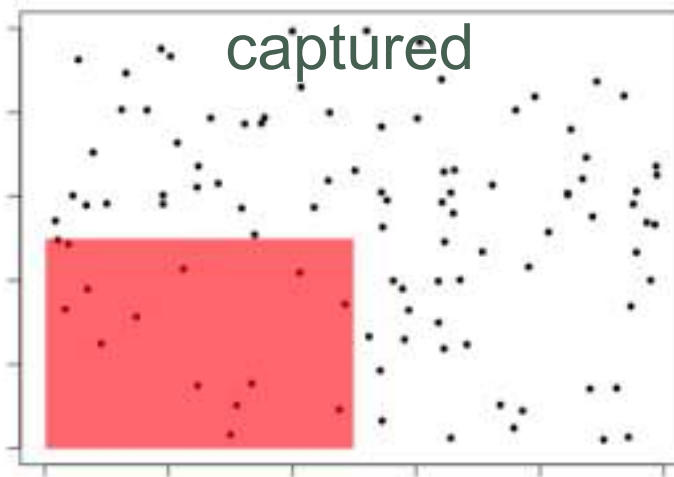
(Note – the term dimension is used differently in business intelligence contexts)

CURSE OF DIMENSIONALITY

42% of data is
captured



14% of data is
captured



7% of data is
captured



$N = 100$ observations, uniformly distributed on $[0,1]^d$, $d = 1, 2, 3$.
 % of observations captured by $[0,1/2]^d$, $d = 1, 2, 3$.

SAMPLING OBSERVATIONS

Question: does every row of the dataset need to be used?

If rows are selected randomly (with or without replacement), the resulting sample might be **representative** of the entire dataset.

Drawbacks:

- if the signal of interest is rare, sampling might drown it altogether
- if aggregation is happening down the road, sampling will necessarily affect the numbers (passengers vs. flights)
- even simple operations on a large file (finding the # of lines, say) can be taxing on the memory and in terms of computation time – **prior information on the dataset structure can help**

FEATURE SELECTION

Removing **irrelevant** or **redundant** variables is a common data processing task.

Motivations:

- modeling tools do not handle these well (variance inflation due to multicollinearity, etc.)
- dimension reduction ($\# \text{ variables} \gg \# \text{ observations}$)

Approaches:

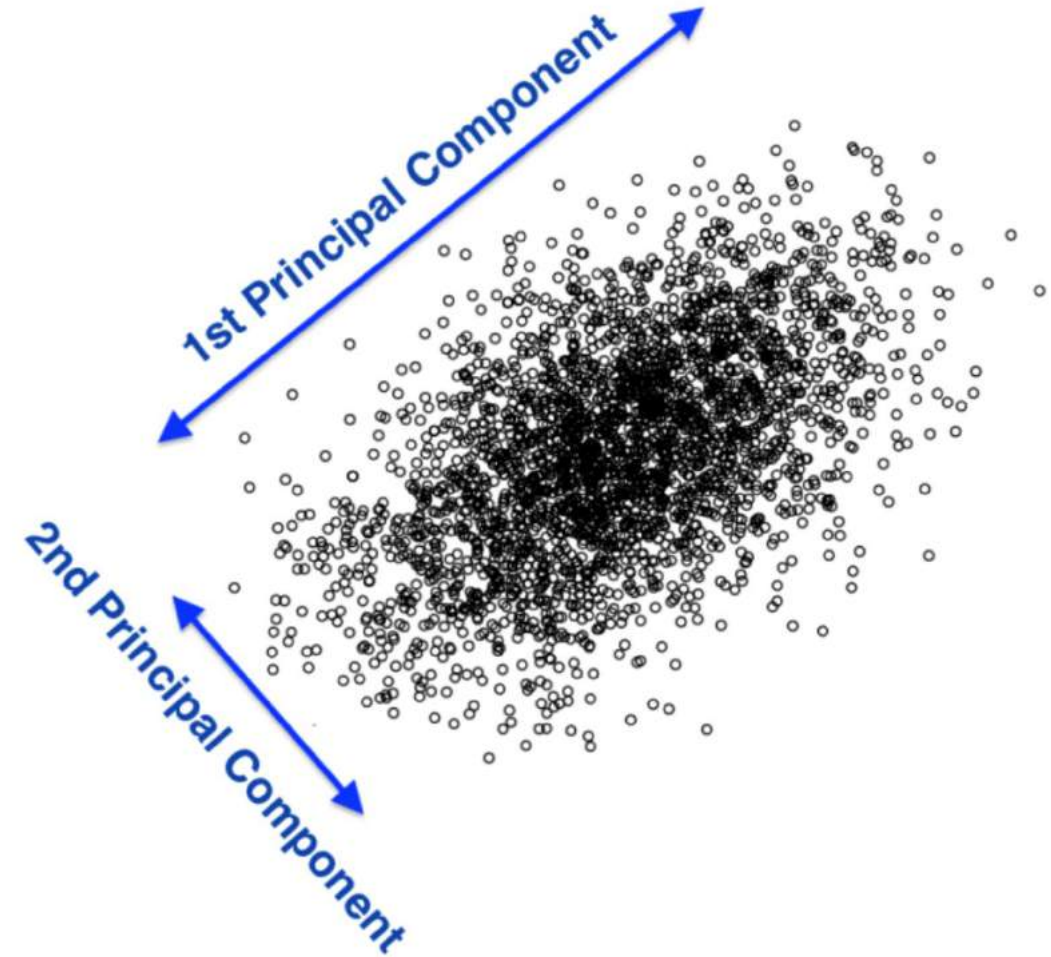
- filter vs. wrapper
- unsupervised vs. supervised

PRINCIPAL COMPONENT ANALYSIS

Motivational Example: Nutritional Content of Food

What is the best way to differentiate food items? Vitamin content, fat, or protein level? A bit of each?

Principal Component Analysis (PCA) can be used to find the combinations of variables along which the data points are **most spread out**.



DIFFERENTIATION

Vitamin C is present in various levels in fruit and vegetables, but not in meats. It **separates** vegetables from meats, and specific vegetables from one another (to some extent), but the meats are **clumped together** (left).

The situation is reversed for *Fat* levels, so the **combination** of vitamin C and fat **separates** vegetables from meats, and **spreads** vegetables and meats (right).

Vitamin C

- Parsley
- Kale
- Broccoli
- Cauliflower
- Soybeans
- Yam
- Guinea Hen

Vitamin C - Fat

- Parsley
- Kale
- Broccoli
- Cauliflower
- Cabbage
- Spinach
- Yam
- Sweet Corn
- Guinea Hen
- Bluefish
- Mackerel
- Chicken
- Beef
- Pork
- Lamb

COMMON TRANSFORMATIONS

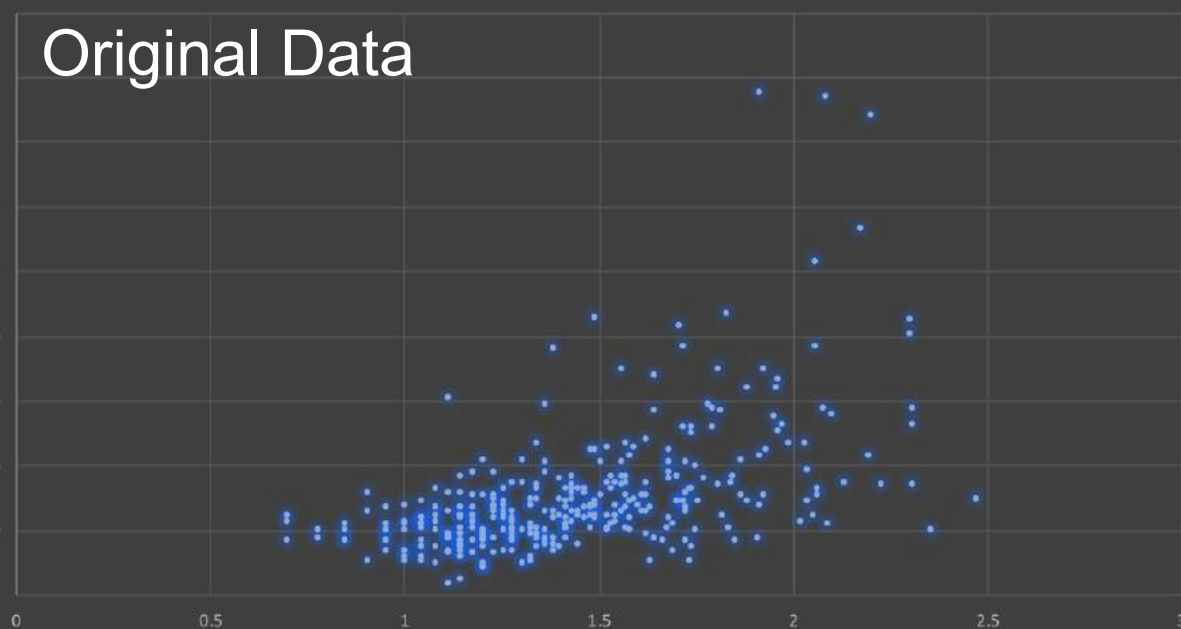
Models sometimes require that certain data assumptions be met (normality of residuals, linearity, etc.).

If the raw data does not meet the requirements, we can either

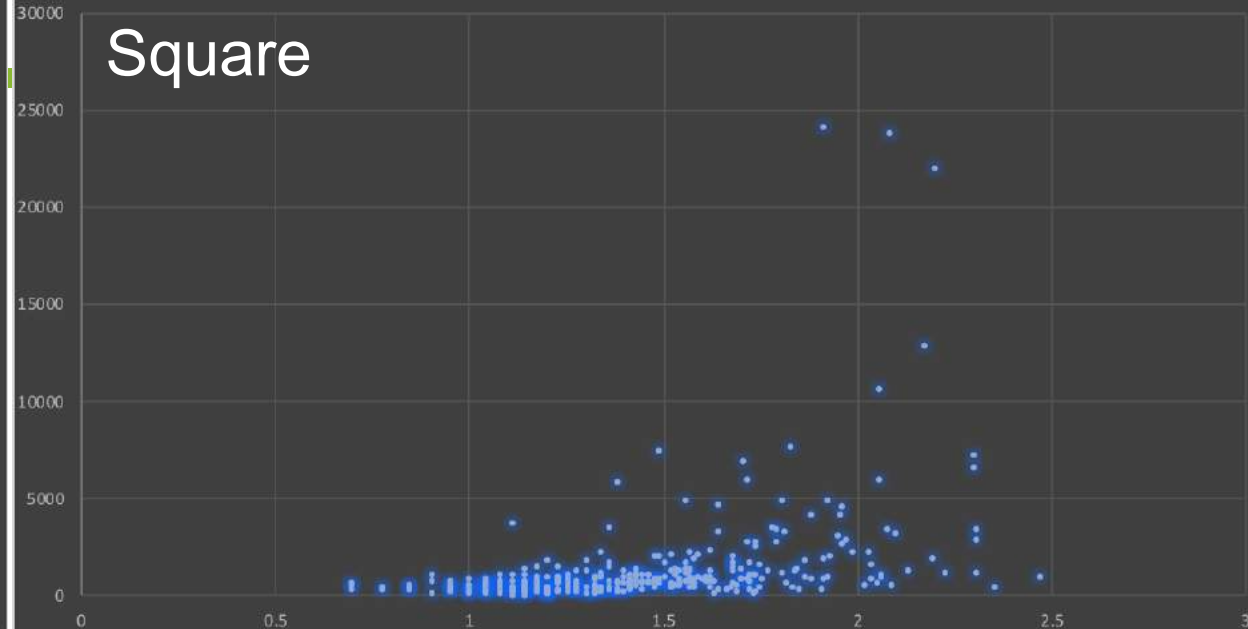
- abandon the model
- attempt to **transform** the data

The second approach requires an inverse transformation to be able to draw conclusions about the original data.

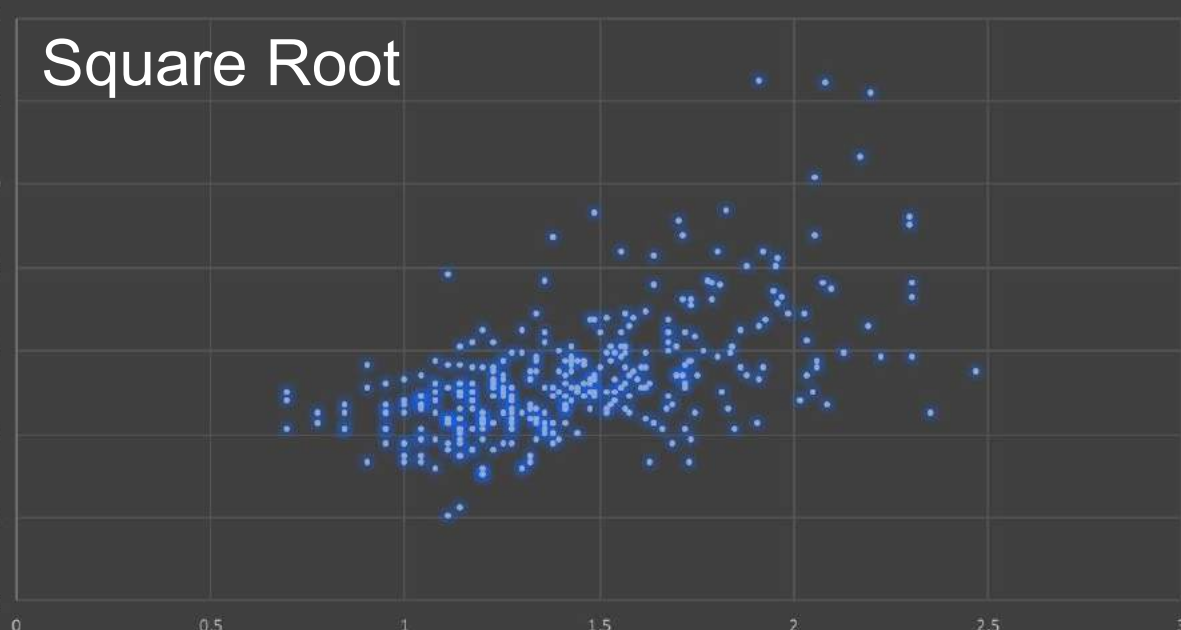
Original Data



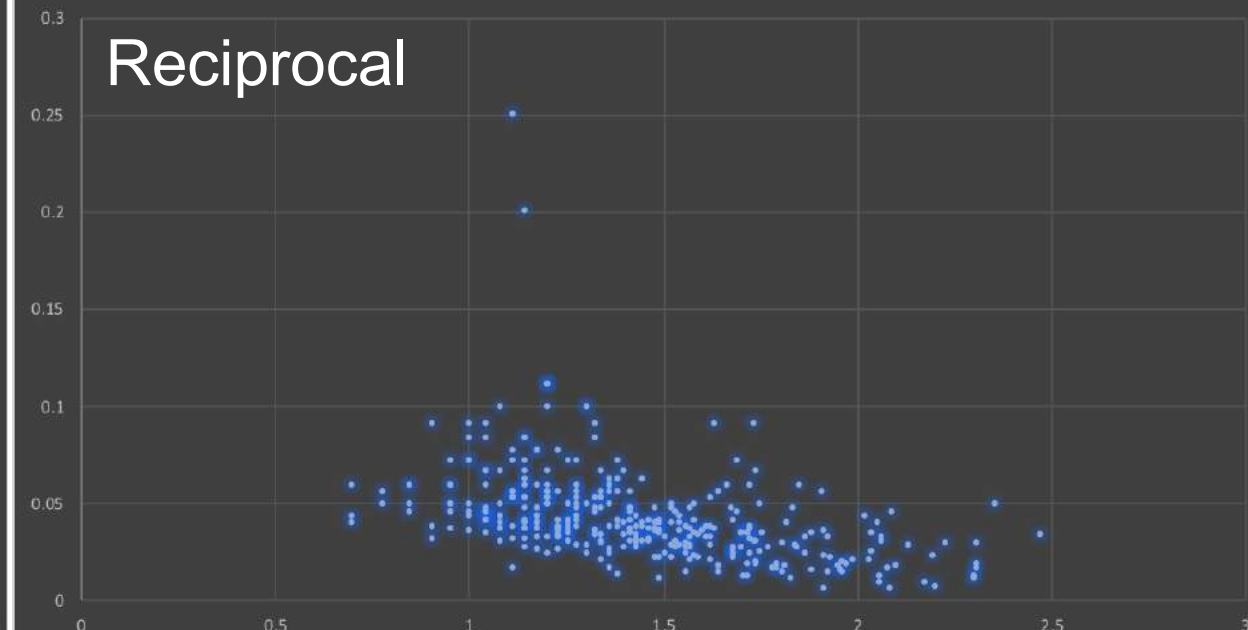
Square



Square Root



Reciprocal



SCALING

Numeric variables may have different **scales** (weights and heights, for instance).

The variance of a large-range variable is typically greater than that of a small-range variable, introducing a bias (for instance).

Standardization creates a variable with mean 0 and std. dev. 1:

$$Y_i = \frac{X_i - \bar{X}}{s_X}$$

Normalization creates a new variable in the range [0,1]: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

DISCRETIZING

To reduce computational complexity, a numeric variable may need to be replaced by an **ordinal** variable (from *height* value to “*short*”, “*average*”, “*tall*”, for instance).

Domain expertise can be used to determine the bins' limits (although that could introduce unconscious bias to the analyses)

In the absence of such expertise, limits can be set so that either

- the bins each contain the same number of observations
- the bins each have the same width
- the performance of some modeling tool is maximized

CREATING VARIABLES

New variables may need to be introduced:

- as **functional relationships** of some subset of available features
- because modeling tool may require **independence of observations**
- because modeling tool may require **independence of features**
- to simplify the analysis by looking at **aggregated summaries** (often used in text analysis)

Time dependencies → time series analysis

Spatial dependencies → spatial analysis

Supplemental Material

LOCAL METHODS IN HIGH DIMENSIONS

A model is said to be **local** if it depends solely on the data *near* the input vector (k NN is local, linear regression isn't).

With a **large training set**, we could increase k (in a k NN model, say) and get enough data points to provide a solid approximation to the theoretical boundary.

The **Curse of Dimensionality** (CoD) is the breakdown of this approach in high-dimensional spaces: when the # of features increases, the # of observations required to maintain predictive power also increases... **but at a substantially higher rate.**

MANIFESTATIONS OF COD

Let $x_i \sim U^1(0,1), i = 1, \dots, N$ be i.i.d.

For any $z \in [0,1]$ and $\varepsilon > 0$ such that

$$I_1(z; \varepsilon) = \left[z - \frac{\varepsilon}{2}, z + \frac{\varepsilon}{2} \right] \subseteq [0,1],$$

we expect $|I_1(z; \varepsilon) \cap \{x_i\}_{i=1}^N| \approx \varepsilon \cdot N$

In other words, a subset whose edge is ε percent of the original set in \mathbb{R} contains ε percent of the observations.

MANIFESTATIONS OF COD

Let $x_i \sim U^2(0,1), i = 1, \dots, N$ be i.i.d.

For any $z \in [0,1]^2$ and $\varepsilon > 0$ such that

$$I_2(z; \varepsilon) = \left[z_1 - \frac{\varepsilon}{2}, z_1 + \frac{\varepsilon}{2} \right] \times \left[z_2 - \frac{\varepsilon}{2}, z_2 + \frac{\varepsilon}{2} \right] \subseteq [0,1]^2,$$

we expect $|I_2(z; \varepsilon) \cap \{x_i\}_{i=1}^N| \approx \varepsilon^2 \cdot N$

In other words, a subset whose edge is ε percent of the original set in \mathbb{R}^2 contains ε^2 percent of the observations.

MANIFESTATIONS OF COD

Let $x_i \sim U^p(0,1), i = 1, \dots, N$ be i.i.d.

For any $z \in [0,1]^p$ and $\varepsilon > 0$ such that

$$I_p(z; \varepsilon) = \prod_{j=1}^p \left[z_j - \frac{\varepsilon}{2}, z_j + \frac{\varepsilon}{2} \right] \subseteq [0,1]^p,$$

we expect $|I_p(z; \varepsilon) \cap \{x_i\}_{i=1}^N| \approx \varepsilon^p \cdot N$

In other words, a subset whose edge is ε percent of an original set in \mathbb{R}^p contains ε^p percent of the observations.

MANIFESTATIONS OF COD

To capture r percent of the observations uniformly distributed in a unit p -hypercube, we need a hyper-subset with edge

$$\varepsilon_p(r) = r^{1/p}.$$

For instance, for $r = 33\%$, we need a subset with edge

- $\varepsilon_1(1/3) \approx 0.33$ in \mathbb{R}
- $\varepsilon_2(1/3) \approx 0.58$ in \mathbb{R}^2
- $\varepsilon_{10}(1/3) \approx 0.90$ in \mathbb{R}^{10}

Locality is lost!

SUPERVISED FILTER METHODS

Correlation between a feature X and a target variable Y :

$$\rho_{X,Y} = \frac{\sum_{i=1}^N (y_i - \bar{y}) (x_i - \bar{x})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}}$$

Features which are highly correlated with the target variable are retained, but this approach is limited if the relationship to the target variable is **non-linear**.

SUPERVISED FILTER METHODS

Mutual Information of nominal target Y from nominal feature X :

$$I(Y; X) = H(Y) - H(Y|X)$$

where the **entropy** and **conditioned class entropy** are given by

$$H(Y) = - \sum_c P(Y = c) \cdot \log P(Y = c) \quad (v, c \text{ represent levels of } X, Y).$$

$$H(Y|X) = - \sum_{v,c} P(X = v, Y = c) \cdot \log \frac{P(X=v, Y=c)}{P(X=v)}$$

$I(Y; X)$ measures the amount of information that can be obtained about Y by knowing X .

LASSO AND VARIANTS

Stepwise Selection is a form of *Occam's Razor*. at each step, a new feature is considered for **inclusion** or **removal** from the current features set based on some criterion (F -test, t -test, etc.).

Limitations:

- tests are biased, since they are based on the same data.
- adjusted R^2 only takes into account the number of features in the final fit, and not the d.f. that have been used in the entire model.
- if cross-validation is used, stepwise selection has to be repeated for each sub-model (that's not usually done).
- classic example of p -hacking (results without hypothesis).

LASSO AND VARIANTS

In what follows, we assume that we have N **centered** and **scaled** $x_i = (x_{1,i}, \dots, x_{p,i})^T$ and a target observation y_i .

Let $\hat{\beta}_{LS,j} = \left[(X^T X)^{-1} X^T y \right]_j$ be the j^{th} OLS coefficient and set a threshold $\lambda > 0$, whose value depends on the training dataset.

In general, there are **no restrictions** on the values taken by the coefficients $\hat{\beta}_{LS,j}$ – larger magnitudes imply that corresponding features **play an important role** in predicting the target.

LASSO AND VARIANTS

Ridge regression is a method to **regularize** the regression coefficients (the effect is to shrink the coefficient values)

The problem consists in solving

$$\arg \min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + N\lambda \|\boldsymbol{\beta}\|_2^2 \},$$

which yields ridge coefficients

$$\hat{\beta}_{RR,j} = \frac{\hat{\beta}_{LS,j}}{1 + N\lambda}$$

LASSO AND VARIANTS

Regression with best subset selection is a method that sets some regression coefficients to 0 (potentially).

The problem consists in solving

$$\arg \min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + N\lambda \|\boldsymbol{\beta}\|_0 \}, \text{ where } \|\boldsymbol{\beta}\|_0 = \sum_j \text{sign}(|\beta_j|),$$

which yields coefficients

$$\hat{\beta}_{BS,j} = \begin{cases} 0 & \text{if } |\hat{\beta}_{LS,j}| < \sqrt{N\lambda} \\ \hat{\beta}_{LS,j} & \text{if } |\hat{\beta}_{LS,j}| \geq \sqrt{N\lambda} \end{cases}$$

LASSO AND VARIANTS

Least Absolute Shrinkage and Selection Operator (LASSO) is a regression method for **feature selection** and **regularization**.

The problem consists in solving

$$\arg \min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + N\lambda \|\boldsymbol{\beta}\|_1 \}$$

which yields lasso coefficients

$$\hat{\beta}_{L,j} = \hat{\beta}_{LS,j} \cdot \max \left(0, 1 - \frac{N\lambda}{|\hat{\beta}_{LS,j}|} \right)$$

LASSO AND VARIANTS

LASSO combines the properties of **ridge regression** (shrinkage) and **best subset selection** (feature selection).

Ridge regression can be viewed as linear regression with prior **normal distributions** assigned to the coefficients; these are **Laplace distributions** in lasso regression.

Lasso selects **at most** $\max\{p, N\}$ features, and usually selects no more than one feature in a group of highly correlated variables.

Extensions: elastic nets; group, fused and adaptive lassos; bridge regression

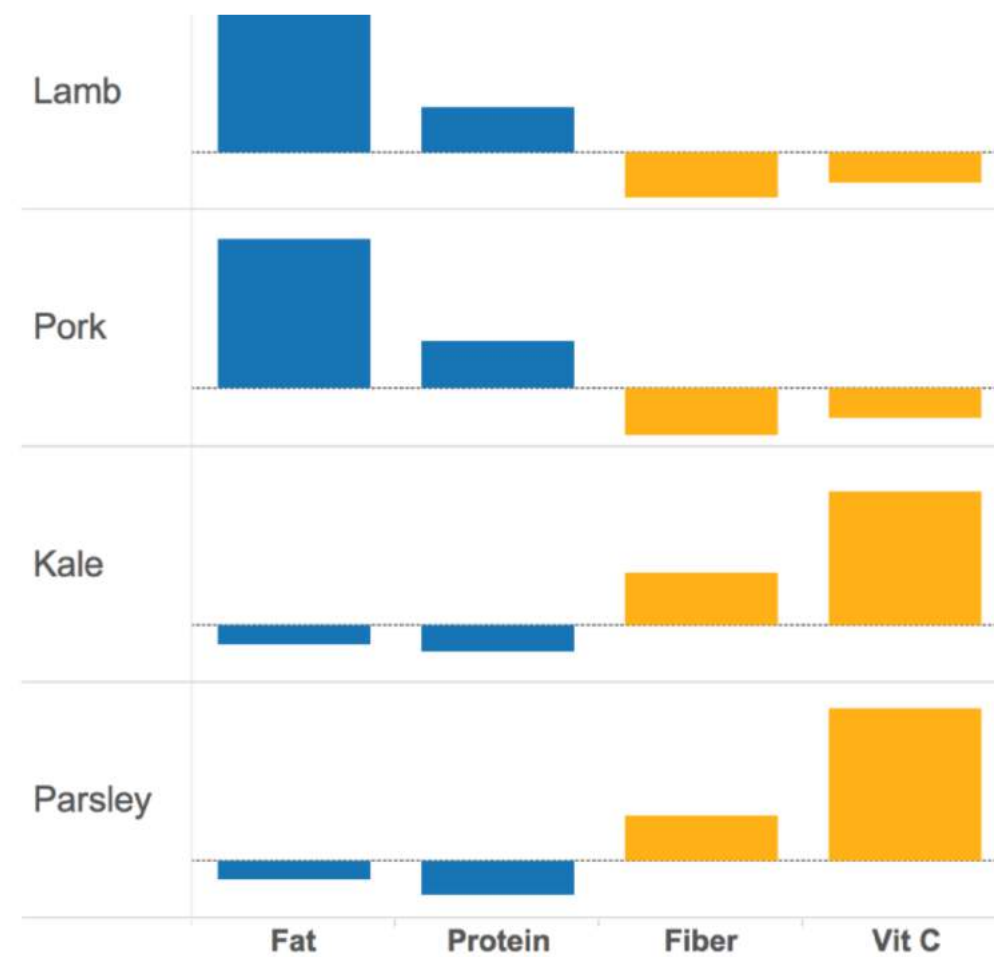
PRINCIPAL COMPONENTS

Presence of nutrients appears to be correlated among food items.

In the (small) sample consisting of Lamb, Pork, Kale, and Parsley, *Fat* and *Protein* levels seem in step, as do *Fiber* and *Vitamin C*.

In a larger dataset, the correlations are $r = 0.56$ and $r = 0.57$.

How much could 2 variables explain?



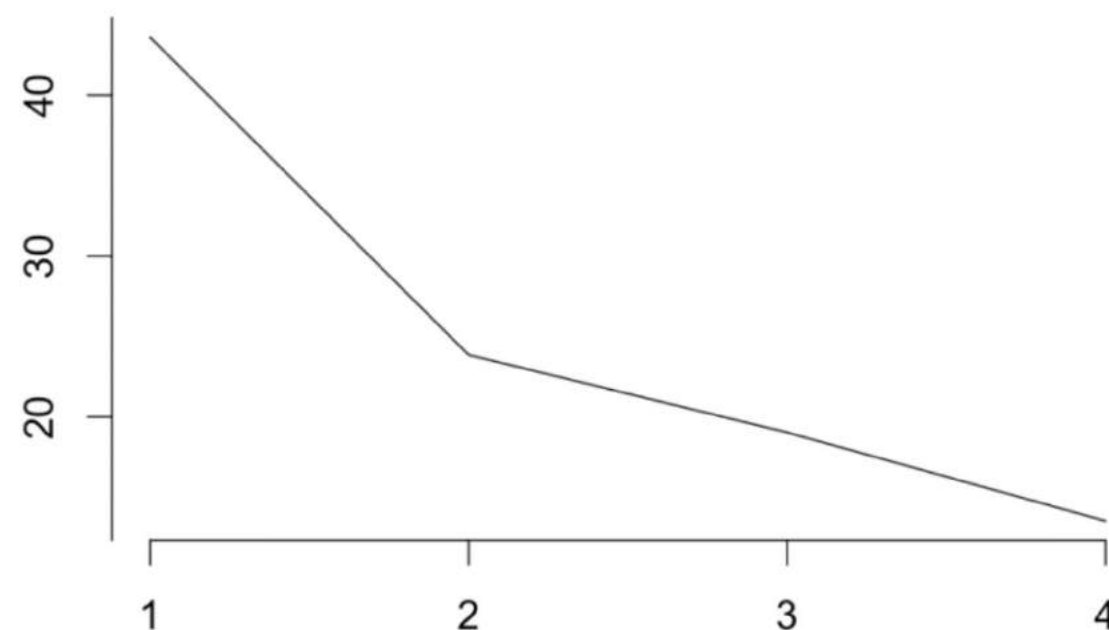
RETAINING PRINCIPAL COMPONENTS

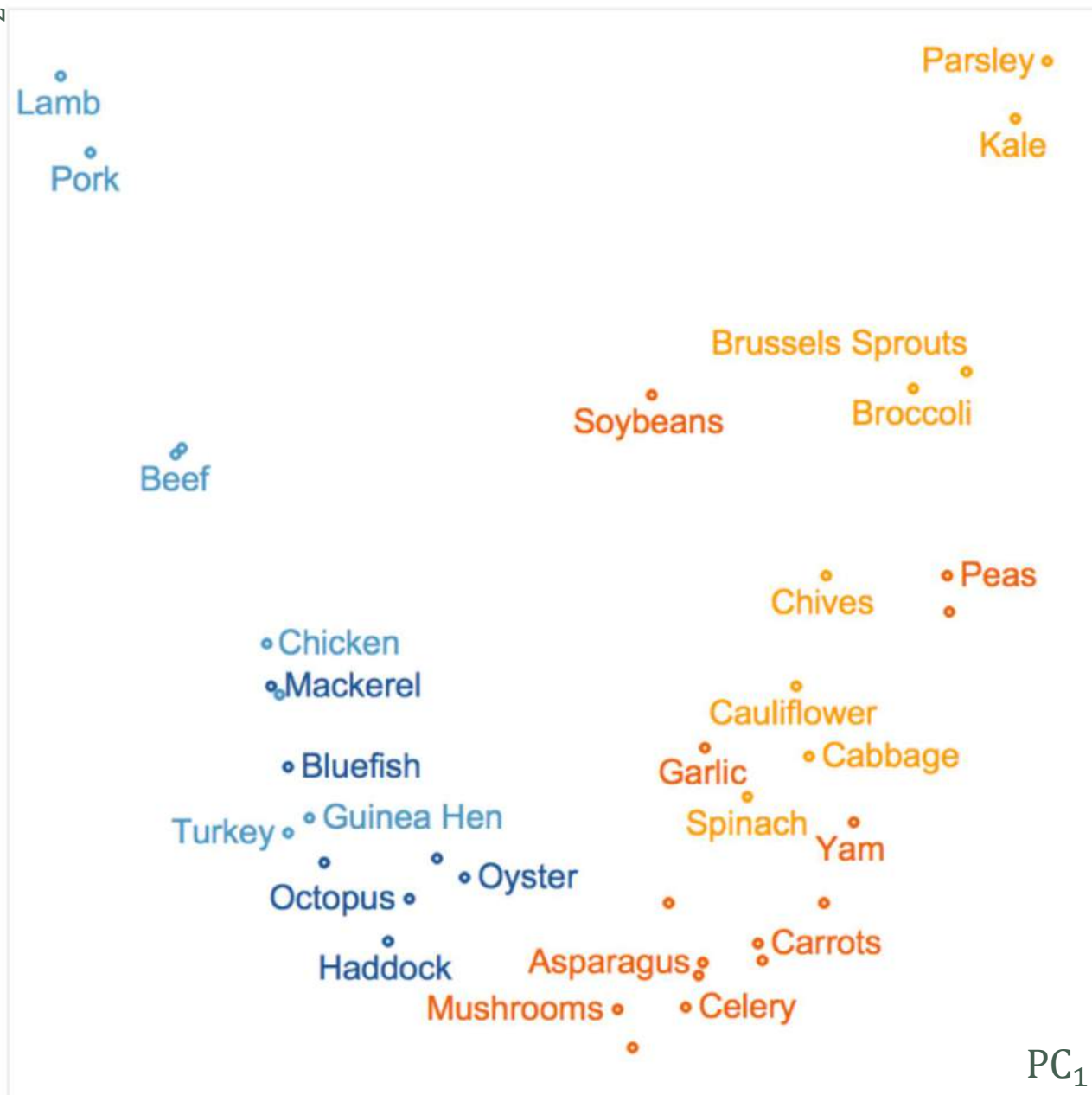
The **proportion of the spread** in the data which can be explained by each principal component is shown in the scree plot.

How many PCs are retained in the analysis?

- keep the PCs where the cumulative proportion is below some threshold
- keep the PCs leading to a kink

Here, 2 PCs \approx 68% of the spread.



PC₂

DIFFERENTIAION (REPRISE)

PC₁ differentiates vegetables from meats; PC₂ differentiates two **sub-categories** within these:

- **Meats** are concentrated on the left (low PC₁ values).
- **Vegetables** are concentrated on the right (high PC₁ values).
- **Seafood** have lower *Fat* content (low PC₂ values) and are concentrated at the bottom.
- **Non-leafy veggies** have lower *Vitamin C* content (low PC₂ values) and are also bunched at the bottom.

PCA IN THEORY

PCA attempts to fit a *p*-**ellipsoid** to centered and scaled* data. Ellipsoid axes represent the principal components of the data. Small axes are components along which the variance is “small”; removing these component leads to a “small” loss of information.

Procedure:

1. Centre and scale the data: matrix X
2. Compute the data's covariance matrix $K = X^T X$
3. Compute K 's eigenvalues Λ and orthonormal eigenvector matrix W
4. Each eigenvector w represents an axis, whose variance is given by the associated eigenvalue λ

PCA IN THEORY

The eigenvectors w are also called the **loadings**.

Typically, the eigenvalues are ordered in **decreasing** sequence, so that the first loading corresponds to the largest axis.

K positive semi-definite \Rightarrow eigenvalues $\lambda = s^2$ are positive; s is a singular value of X (i.e. a diagonal entry of Σ in the **singular value decomposition** $X = U\Sigma W^T$).

The PCA decomposition of X is $T = XW = U\Sigma$.

PCA IN THEORY

The link between the PCs and the eigenvectors can be made explicit:

- the **first** principal component is the loading which maximizes the variance of the first column of T

$$\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \{\text{Var}(\mathbf{t}^1)\}$$

- but T is centered so the variance is simply

$$\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \{t_{1,1}^2 + \dots + t_{1,N}^2\}$$

- using the PC decomposition of X , this becomes

$$\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \{\langle \mathbf{x}_1 | \mathbf{w} \rangle^2 + \dots + \langle \mathbf{x}_N | \mathbf{w} \rangle^2\} = \arg \max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\}$$

PCA IN THEORY

- which by definition of the norm is $\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}\}$
- since $\|\mathbf{w}\| = 1$, the loading also satisfies

$$\mathbf{w}^1 = \arg \max_{\|\mathbf{w}\|=1} \left\{ \frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

- using Lagrange multipliers, it can be shown that the critical points of $\frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}}$ are exactly the eigenvectors of $\mathbf{K} = \mathbf{X}^T \mathbf{X}$ (of which there are p)
- if \mathbf{w} (unit) and $\lambda^* \geq 0$ are such that $\mathbf{K} \mathbf{w} = \lambda \mathbf{w}$, then

$$\frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \frac{\mathbf{w}^T \mathbf{K} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \frac{\mathbf{w}^T \lambda \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \lambda \frac{\mathbf{w}^T \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \lambda$$

PCA NOTES

The loading that explains the most variance along a single axis is the eigenvector of the empirical covariance matrix corresponding to the largest eigenvalue, and that variance is proportional to the eigenvalue.

The process is repeated, yielding **orthonormal** principal components PC_1, \dots, PC_r , where $r = \text{rank}(\mathbf{X})$.

GENERALIZATIONS

Nonlinear PCA-like methods attempt to find **principal manifolds**.

- self-organizing maps
- auto-encoders
- curvilinear component analysis
- manifold sculpting

Rather than reducing the dimensionality, we may **expand** it with kernel PCA (this is equivalent to replacing the usual inner product by more exotic objects).

LIMITATIONS

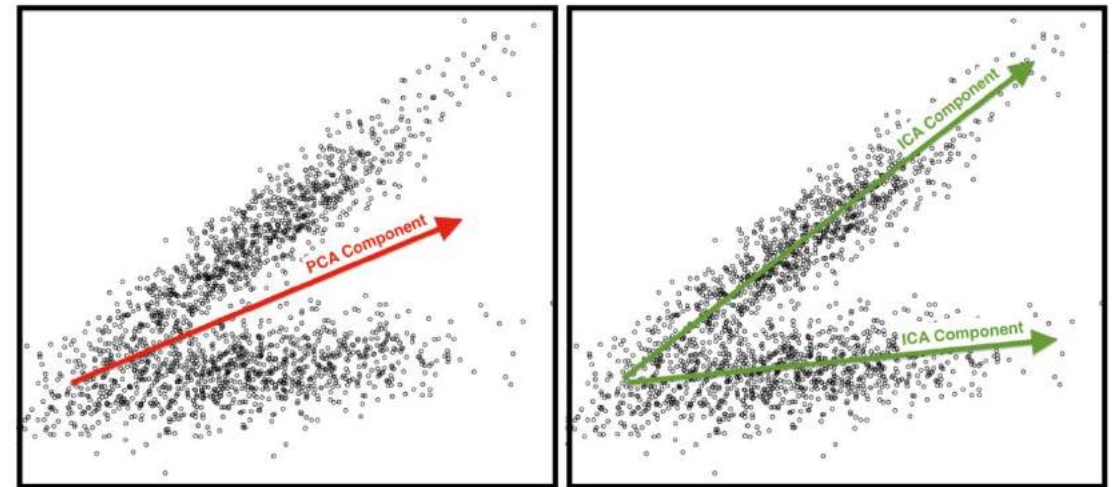
PCA is dependent on scaling (not unique).

With no prior domain expertise, interpreting PCs may be difficult.

Assumptions are **not always met**

- important structures and spread are correlated (i.e. counting pancakes)
- PCs are orthogonal (what about ICA?)
- change of basis framework (i.e. Ferris wheel tracking)

Sensitive to outliers.



HIGH DIMENSIONALITY AND 'BIG' DATA

Datasets can be “big” in a variety of ways:

- too large for the **hardware** to handle (cannot be stored or accessed properly due to # of observations, # of features, or the overall size)
- dimensions can go against specific **modeling assumptions** (# of features \gg # observations)

Examples:

- Multiple sensors recording 100+ observations per second in a large geographical area over a long time period = **very big dataset**.
- In a corpus' *Term Document Matrix* (cols = terms, rows = documents), the number of terms is usually substantially higher than the number of documents, leading to **excessively sparse data**.

FEATURE SELECTION METHODS

Filter methods inspect each variable individually and score them according to some **importance metric**.

The less relevant features (i.e. importance score below some set threshold) are then removed.

Wrapper methods seek feature subsets for which the evaluation criterion used by the eventual analytical method is “optimized”.

The process is **iterative**, and typically computationally intensive: candidate subsets are used in the analysis until one produces an acceptable evaluation metric for the analysis.

FEATURE SELECTION METHODS

Unsupervised methods determine the importance of a feature based only on its values.

Supervised methods evaluate each feature's importance by studying the relationship with a **target feature** (correlation, etc.)

Wrapper methods are usually supervised.

Unsupervised filter methods: removing constant variables, ID-like variables (different on all observations), features with low variability, etc.

SUPERVISED FILTER METHODS

Correlation between a feature X and a target variable Y :

$$\rho_{X,Y} = \frac{\sum_{i=1}^N (y_i - \bar{y}) (x_i - \bar{x})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}}$$

Features which are highly correlated with the target variable are retained, but this approach is limited if the relationship to the target variable is **non-linear**.

OTHER SUPERVISED METRICS

Classification Tasks

- Gain Ratio
- Inf Gain
- Gini
- MDL, etc.

Regression Tasks

- MSE of Mean
- MAE of Mean
- Relief (evaluates features simultaneously), etc.

COMMON TRANSFORMATIONS

In the regression context, transformations are **monotonic**:

- logarithmic
- square root, inverse, power: W^k
- exponential
- Box-Cox, etc.

Transformations on X may achieve linearity, but usually at some price (correlations are not preserved, for instance). Transformations on Y can help with non-normality and unequal variance of error terms.

BOX-COX TRANSFORMATION

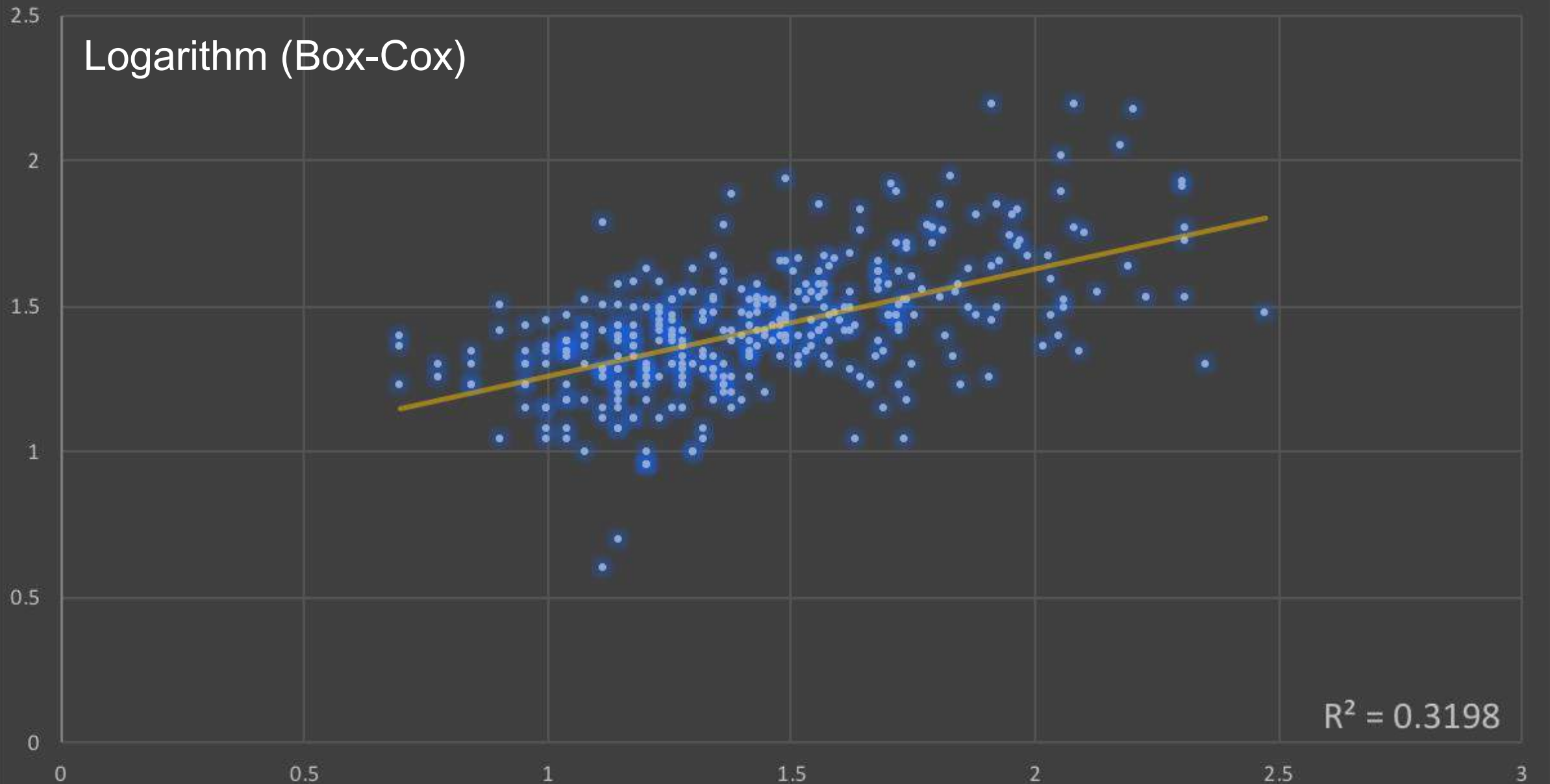
Assume the usual model $Y_j = \sum_i \beta_i X_{j,i} + \varepsilon_j$ with either

- skewed residuals
- not-constant variance
- non-linear trend

The **Box-Cox transformation** $Y_j \mapsto Y_j'(\lambda)$ suggests a choice: select λ which maximizes the corresponding log-likelihood

$$Y_j'(\lambda) = \begin{cases} \text{gm}(\mathbf{Y}) \times \ln(Y_j), & \lambda = 0 \\ \lambda^{-1} \text{gm}(\mathbf{Y})^{1-\lambda} \times (Y_j^\lambda - 1), & \lambda \neq 0 \end{cases}$$

Logarithm (Box-Cox)



BOX-COX TRANSFORMATION

The procedure provides a **guide** to select a transformation.

Theoretical rationales may exist for a particular choice of λ .

Residual analysis is still required to ensure that the choice was appropriate.

The resulting parameters have the least squares property only with respect to the transformed data points.

DATA QUALITY AND DATA VALIDATION

DATA COLLECTION AND DATA PROCESSING

Martin: Data is messy.

Allison: Even when it's been cleaned?

Martin: Especially when it's been cleaned.

P. Boily, *The Great Balancing Act*

LEARNING OBJECTIVES

Understand common sources of data error and types of potential issues

Understand difference between accuracy and precision

Understand, at a high level, some techniques for detecting data issues

Familiarity with some examples of data validity issues

SOUND DATA

The ideal dataset will have as few issues as possible with:

- **Validity:** data type, range, mandatory response, uniqueness, value, regular expressions
- **Completeness:** missing observations
- **Accuracy and Precision:** related to measurement and/or data entry errors; target diagrams (accuracy as bias, precision as standard error)
- **Consistency:** conflicting observations
- **Uniformity:** are units used uniformly throughout?

Checking for data quality issues at an early stage can save headaches later in the analysis.



accurate and
precise



precise but
not accurate



accurate but
not precise



neither accurate
nor very precise

COMMON SOURCES OF ERROR

When dealing with **legacy**, **inherited** or **combined** datasets (that is, datasets over which you have little control):

- Missing data given a code
- 'NA'/'blank' given a code
- Data entry error
- Coding error
- Measurement error
- Duplicate entries
- Heaping

DETECTING INVALID ENTRIES

Potentially invalid entries can be detected with the help of:

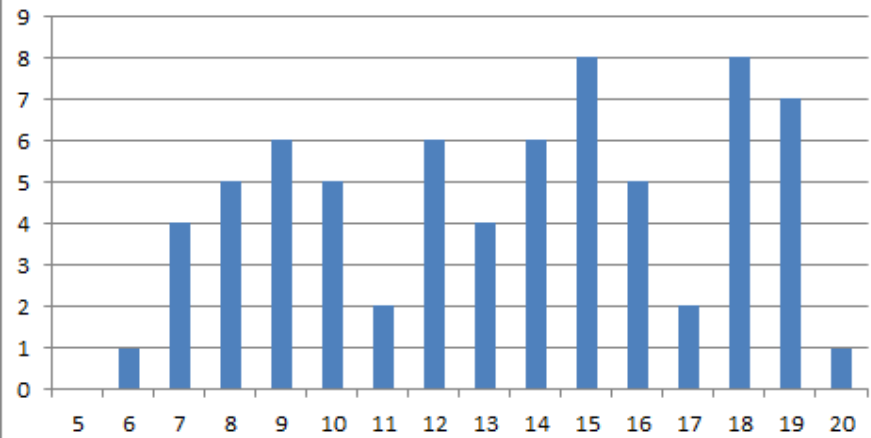
- **Univariate Descriptive Statistics**
count, range, z-score, mean, median, standard deviation, logic check
- **Multivariate Descriptive Statistics**
n-way table, logic check
- **Data Visualization**
scatterplot, scatterplot matrix, histogram, joint histogram, etc.

This step might allow for the identification of potential outliers.

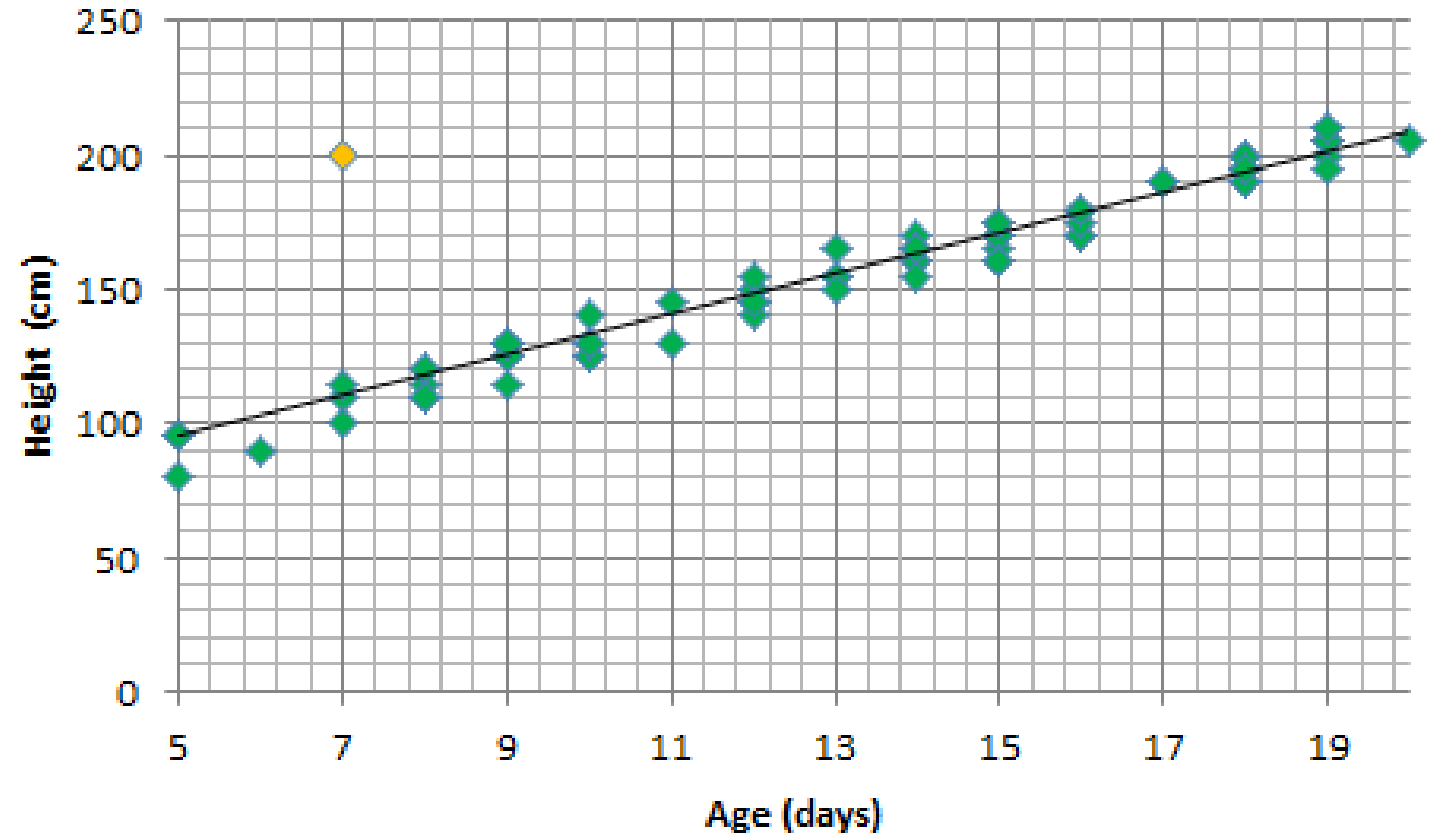
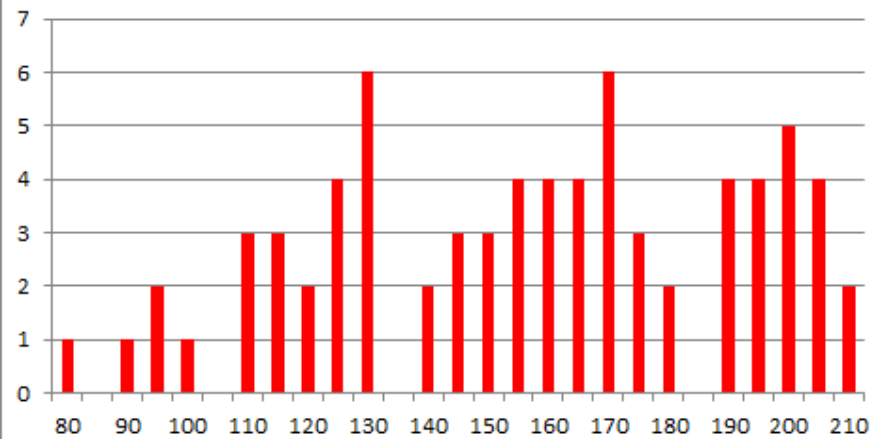
Failure to detect invalid entries \neq all entries are valid.

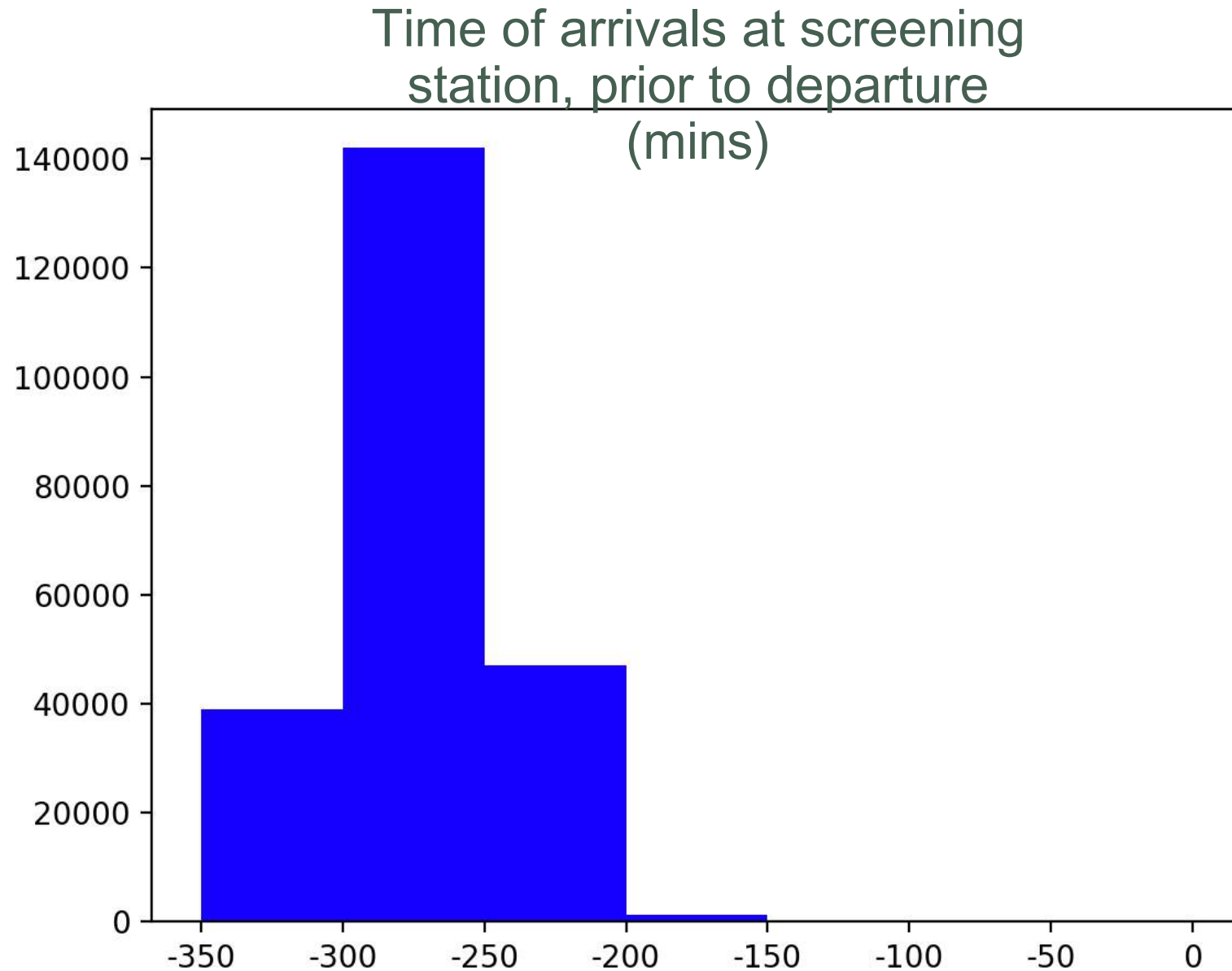
Small numbers of invalid entries recoded as “missing.”

Plant Age (weeks)

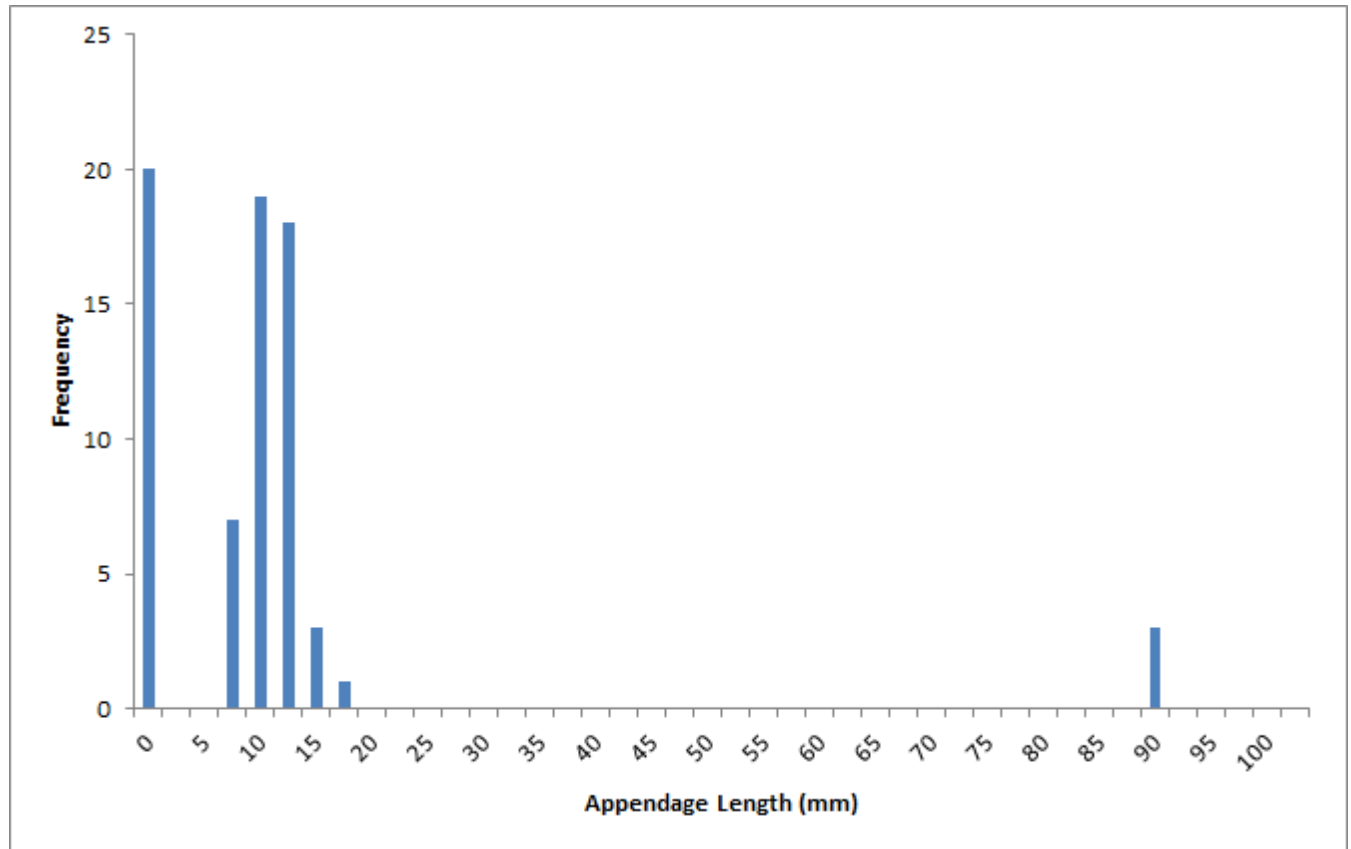


Plant Height (cm)





<i>Appendage length (mm)</i>	
Mean	10.35
Standard Deviation	16.98
Kurtosis	16.78
Skewness	4.07
Minimum	0
First Quartile	0
Median	8.77
Third Quartile	10.58
Maximum	88
Range	88
Interquartile Range	10.58
Mode	0
Count	71



TAKE-AWAYS

Don't wait until after the analysis to find out there was a problem with data quality.

Univariate tests don't always tell the whole story.

Visualizations can help.

Context is crucial – you may need more context about the data in order to make sense of what you see... but whatever the situation, you need to understand the dataset quality.