DATA COLLECTION AND DATA PROCESSING

SETTING THE STAGE

"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

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

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.

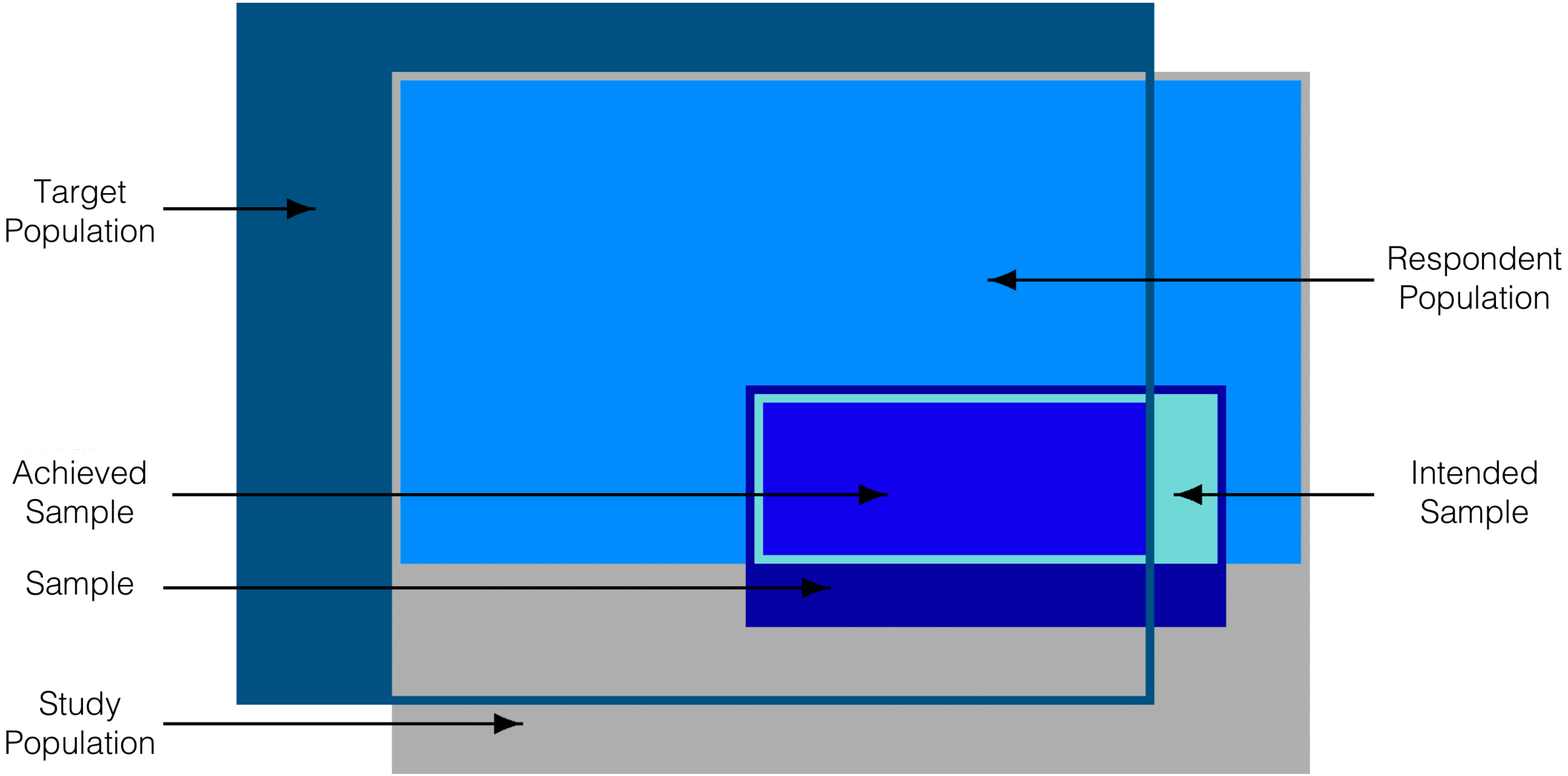
Alone or in combination, these lead to poor (and **potentially harmful**) conclusions.

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.


Target Population
Respondent Population
Achieved Sample
Intended Sample
Sample
Study Population

SURVEY ERROR

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

- Sampling Error: survey, not census
- Measurement Error: observations not measured accurately
- Non-Response Error: non-respondents having systematic observation differences
- Coverage Error: frame decay and/or corruption

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).

SAMPLING DESIGN

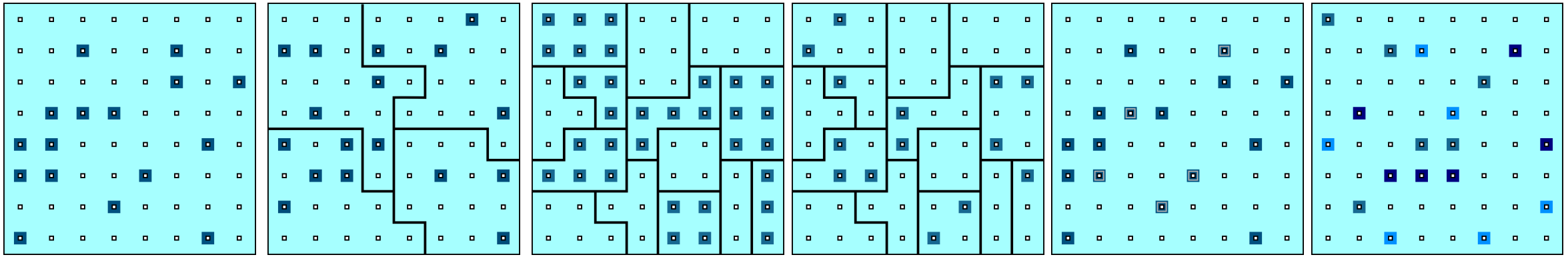
Different **probabilistic sampling designs** have distinct advantages and disadvantages.

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).

PROBABILISTIC SAMPLING DESIGNS



Simple Random Sampling (SRS)

Stratified Sampling (StS)

Cluster Sampling (CIS)

Multi-Stage Sampling (MSS)

Multi-Phase Sampling (MPS)

Replicated Sampling (ReS)

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

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."

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.

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

VS.

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

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.

APPROACHES TO DATA CLEANING

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**
- **Data Entry Issue**
- **Invalid Entry**
- **Expected Blank**

Not all analytical methods can easily accommodate missing observations:

- **Discard** the missing observation (not usually recommended)
- Come up with a **replacement value** (imputation)

IMPUTATION METHODS

List-wise deletion

Mean or most frequent imputation

Regression or correlation imputation

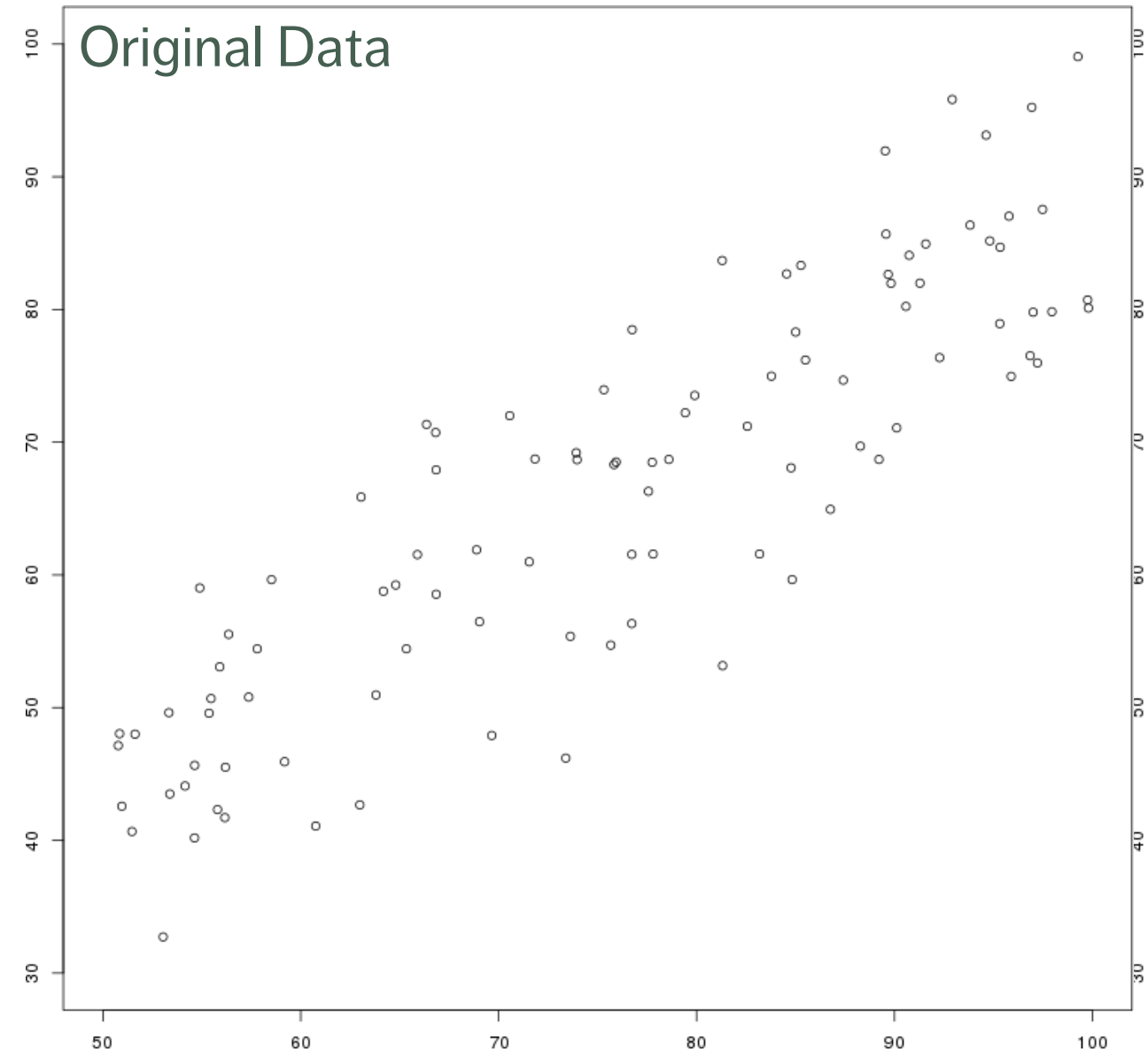
Stochastic regression imputation

Last observation carried forward

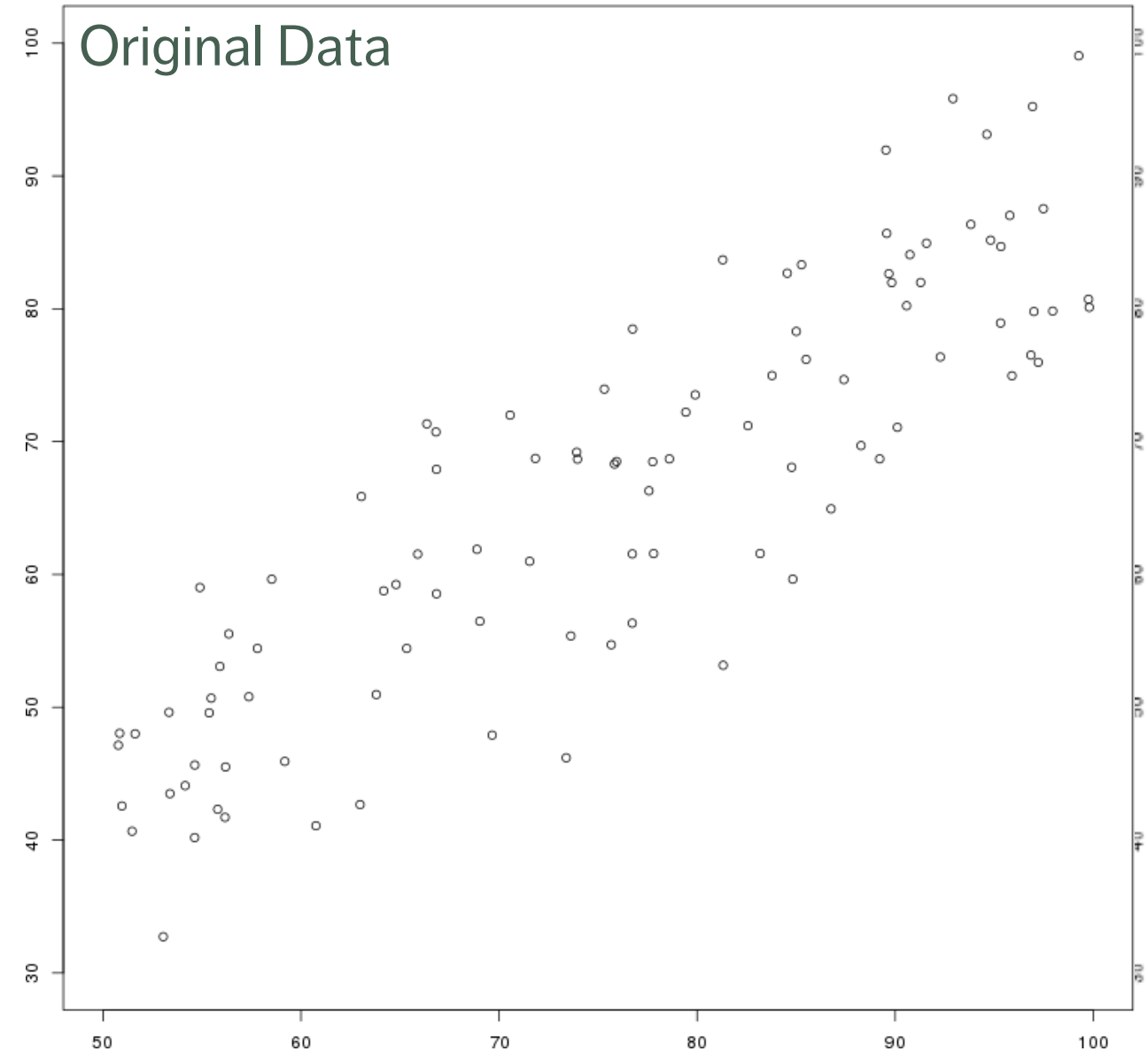
k-nearest neighbours imputation

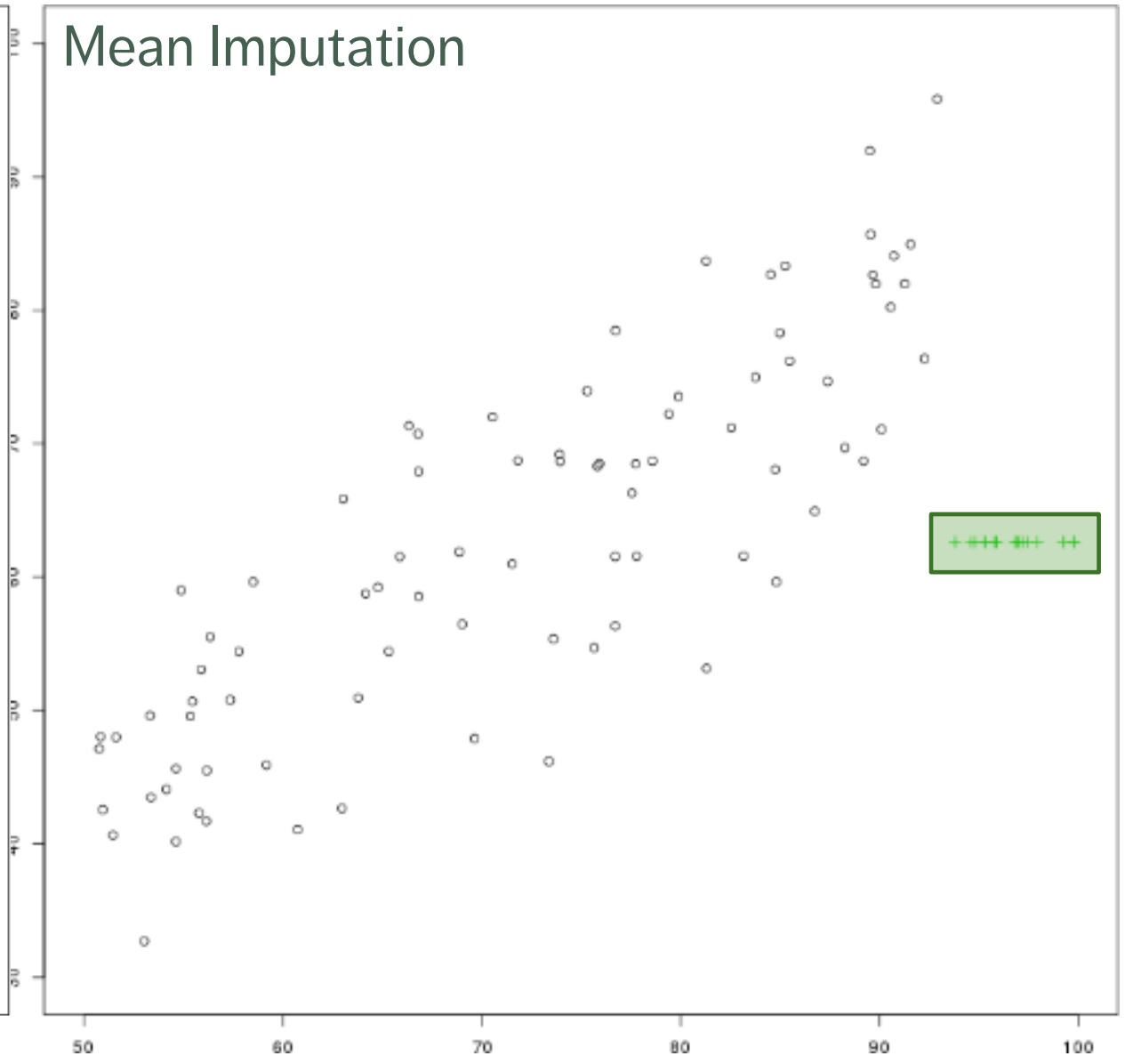
Multiple imputation

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

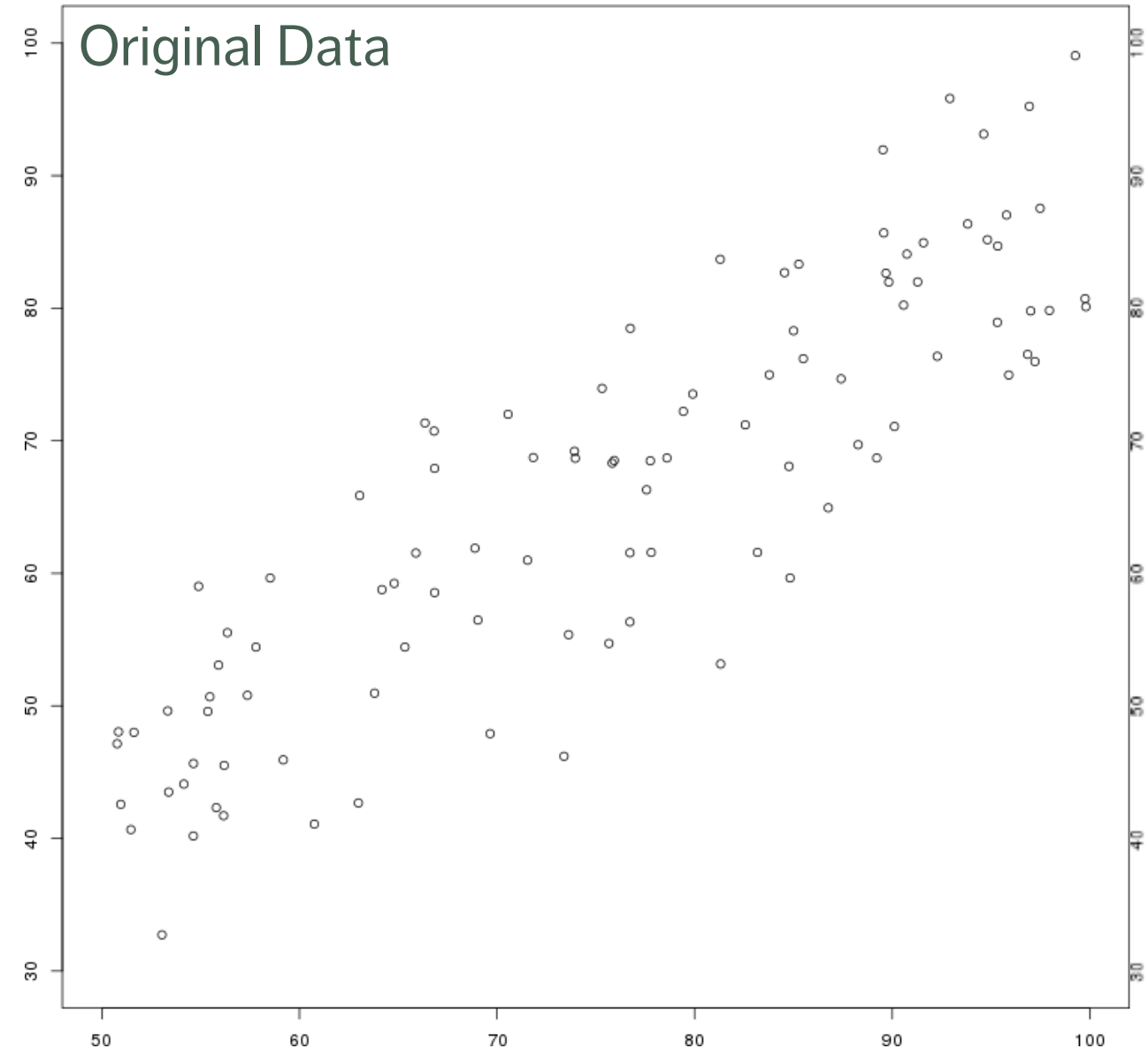


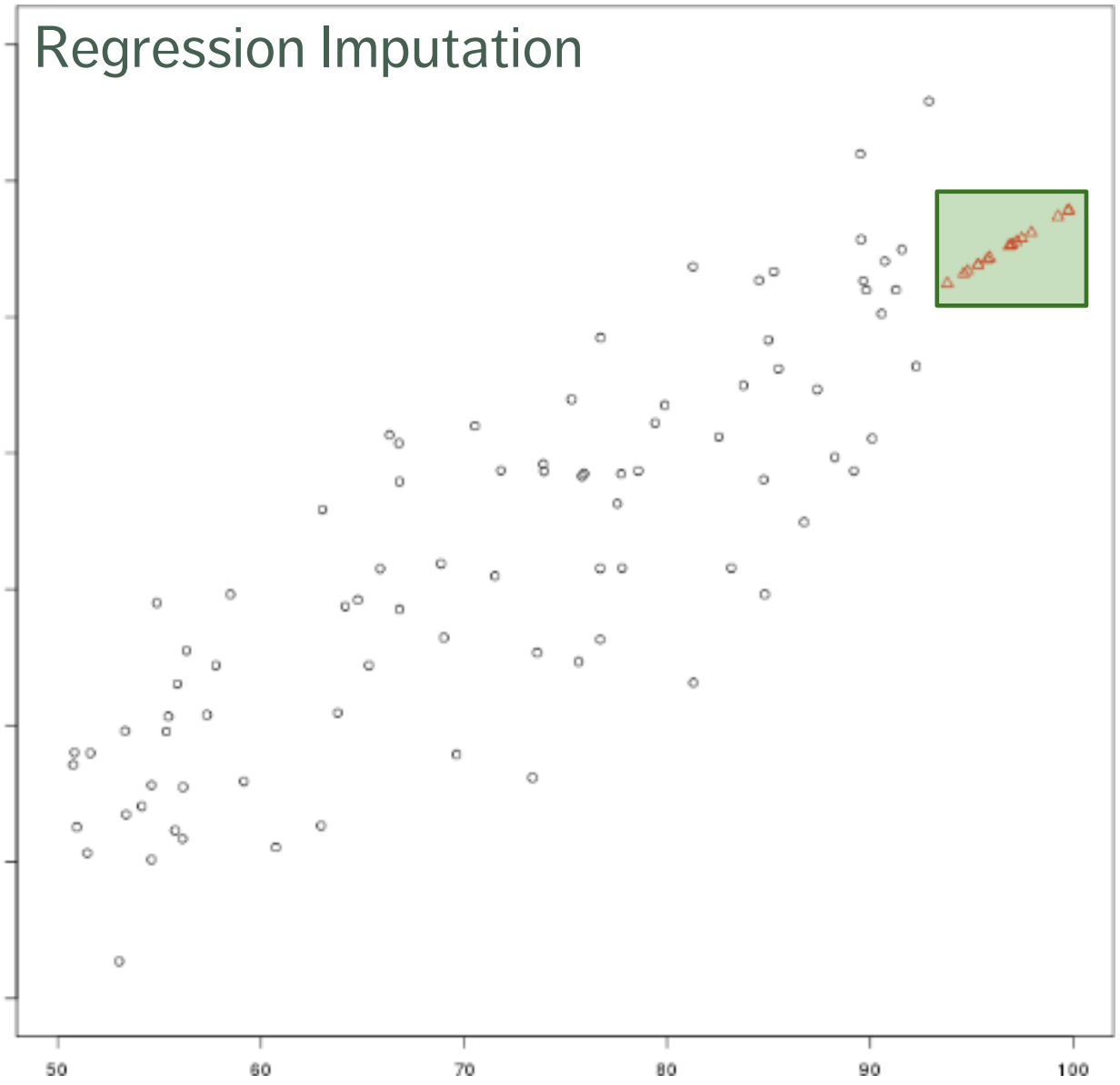
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



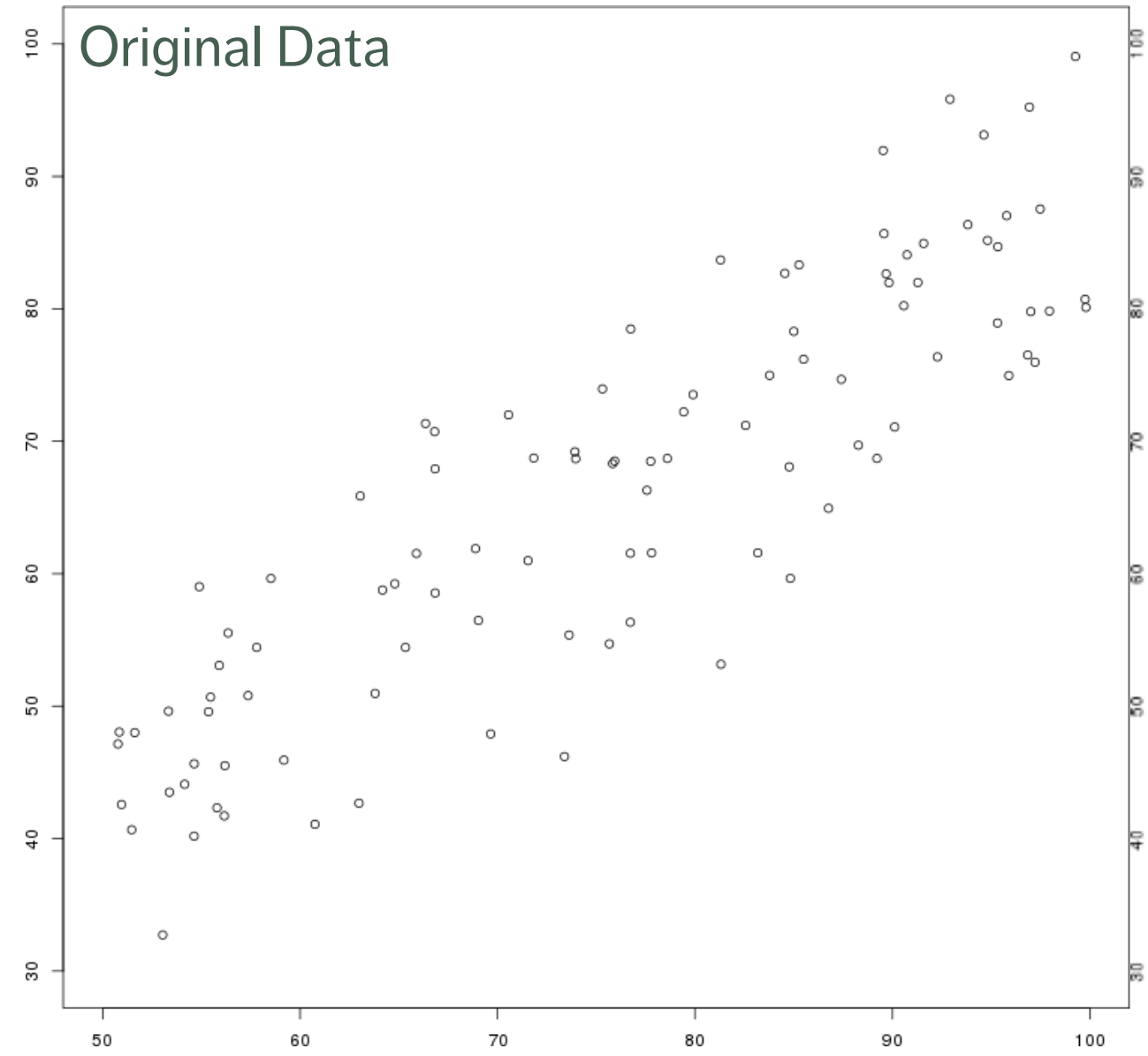


Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





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



SPECIAL DATA POINTS

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.

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.

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)

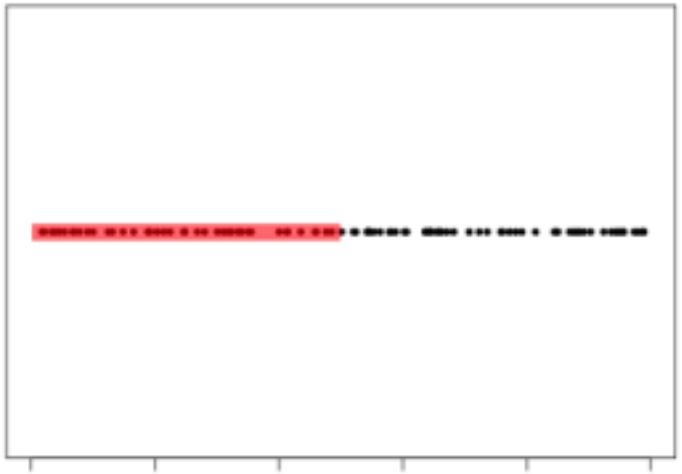
Unsupervised methods don't use external information:

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

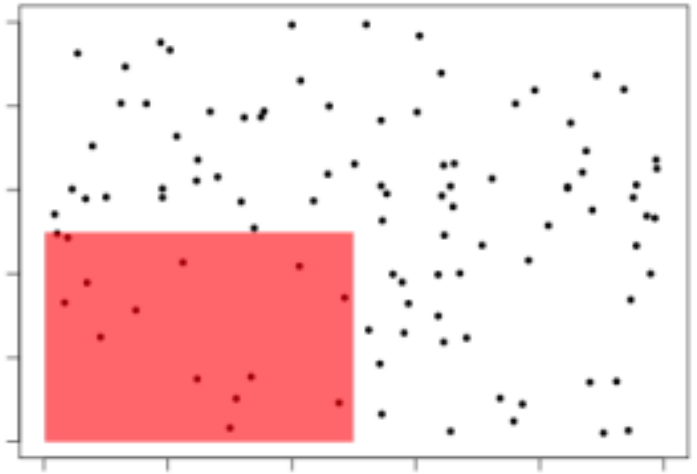
Semi-supervised methods also exist.

CURSE OF DIMENSIONALITY

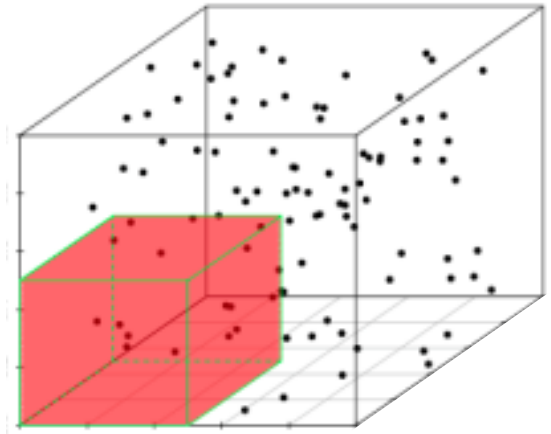
42% of data is captured



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 $[0,1/2]^d, d = 1,2,3.$

FEATURE SELECTION AND DIMENSION REDUCTION

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

Motivations:

- modeling tools do not handle these well (variance inflation due to multicolinearity, etc.)
- dimension reduction (# variables ≫ # observations)

Approaches:

- filter vs. wrapper (regularization), unsupervised vs. supervised

Dimension Reduction: PCA, UMAP, Manifold Learning, etc.

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 (power, reciprocal, log, Box-Cox, etc.)

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

SCALING AND DISCRETIZING

Numeric variables may have different **scales** (weights and heights, 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 $[0,1]$: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

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).

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.