Lancaster University

SCC.460 Data Science Fundamentals

Testing Hypotheses, Analysis and Visualisation

Slides originally developed by Dr Yehia El Khatib and edited by Dr Ioannis Chatzigeorgiou.
Lecture delivered by Dr Ignatius Ezeani.

Today's learning outcomes

- ❑ Testing Hypotheses and Variability
- ❑ Analysis
- ❑ Visualisation

Hypothesis testing

Lancaster University

- To test a **hypothesis**, form an assertion that can be tested with data.
- Make the assertion as concrete and measureable (i.e. **verifiable**) as possible.
  - 'Unfalsifiable' hypotheses are unscientific; see Karl Popper's demarcation.
- A model (to be assessed) is a set of assumptions about the data. In Data Science, models based on assumptions that involve randomness are called **chance models**.
- Hypothesis testing = Model assessment
- If we simulate data according to the model, we can compare **model predictions** to the **observed data**.
- If the data and the model predictions **are not consistent**, there is evidence against the model.

Hypothesis testing: General framework

1. Formulate the **null** hypothesis: H_0
   - Well-defined *chance model*.
2. Formulate the **alternative** hypothesis: H_1
3. Determine what the **test statistic** will be and compute the statistic of the sample
4. Obtain the null distribution of the statistic
5. Derive its **significance probability** (*p*-value)
6. Decide if there is evidence for rejecting the null hypothesis? (or the alternative hypothesis)

H_0: Engagement on social media is **unimportant**

H_1: Engagement on social media is **beneficial**

H_0: Posting on Twitter using trending hashtags **does not change** the number of new customer registrations

H_1: Posting on Twitter using trending hashtags **increases** the number of new customer registrations

p-value (or *P*-value)

- The *p*-value captures the probability of observing the data we observed, and obtaining the test statistic that we obtained under the null hypothesis.
- A low *p*-value indicates that it is highly unlikely to observe such a test statistic if the null hypothesis held. Thus, accept the alternative hypothesis.
- If the **cut-off value is set to 5%** and the *p*-value is smaller than the cut-off value, tests reject the null hypothesis. The result is called **statistically significant** (note there is a 5% chance that the null hypothesis is true).
- If the **cut-off value is set to 1%** and the *p*-value is smaller than the cut-off value, tests reject the null hypothesis. The result is called **highly statistically significant** (often used in medical studies).
- Cut-off values are not universal conventions, **use your own judgement too**.

Example E1: Jury selection (1 of 6)

Lancaster University

- Assume that we have obtained a list of all jury panels (a given *sample*) over a period of time (e.g. 1,500 entries). We also have a list of all eligible jurors (the *population*) and their ethnicity. **You have access to both categorical distributions**.

Ethnicity:	Asian	Black	Latino	White	Other
Eligible jurors:	0.15	0.18	0.12	0.54	0.01
Selected jurors:	**0.26**	**0.08**	0.08	0.54	0.04

- It looks as if people of Asian ethnicity are overrepresented on panels, while people of Black ethnicity are underrepresented. **Could this be due to chance?**

Example E1: Jury selection (2 of 6)

- **Null hypothesis (H_0):** The people on the jury panels were selected at random from the eligible population (fair / random selection).
- **Alternative hypothesis (H_1):** The people on the jury panels were not selected at random from the eligible population (biased selection).
- **Test statistic:** Sum of *absolute* differences divided by two (because we compare two distributions) known as *total variation distance* (TVD), i.e.

$$\text{TVD} = \frac{|0.15 - 0.26| + \cdots + |0.01 - 0.04|}{2} = 0.14$$

- Could this value of TVD have arisen due to chance?

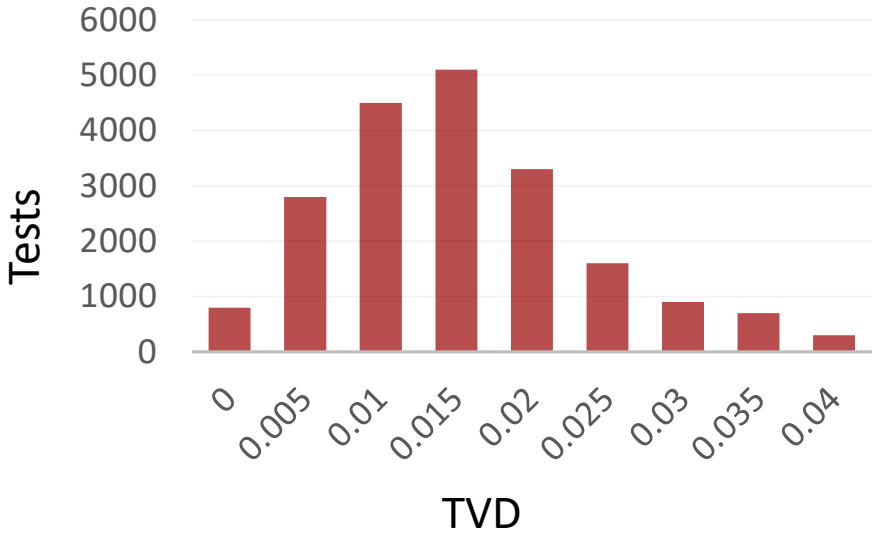
Example E1: Jury selection (3 of 6)

- **Simulation of tests:**
  1. Recall that the given sample had size 1,500.
  2. Take a new random sample (of jurors) from the population (of eligible jurors) of the same size, i.e. 1,500.
  3. Compute the TVD between the distribution of the population and the distribution of sample. Record the TVD.
  4. Repeat steps 2 and 3 many times, e.g. 20,000.
  5. Plot the TVD distribution of the 20,000 values.

Example E1: Jury selection (4 of 6)

- Let us assume that the outcome is (**case A**):

  Recall that the TVD obtained from data was **0.14**

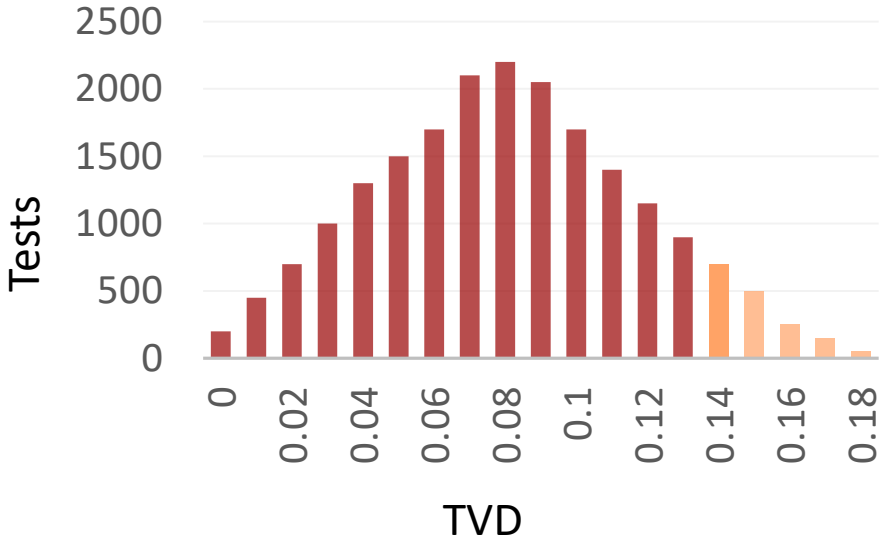


- Random selection produces TVD values between 0 and 0.04 (model predictions). The data gave a TVD value of 0.14, thus they are *inconsistent* with the model. **Tests reject the null hypothesis**.

Example E1: Jury selection (5 of 6)

- Let us assume that the outcome is (**case B**):

  Recall that the TVD obtained from data was **0.14**

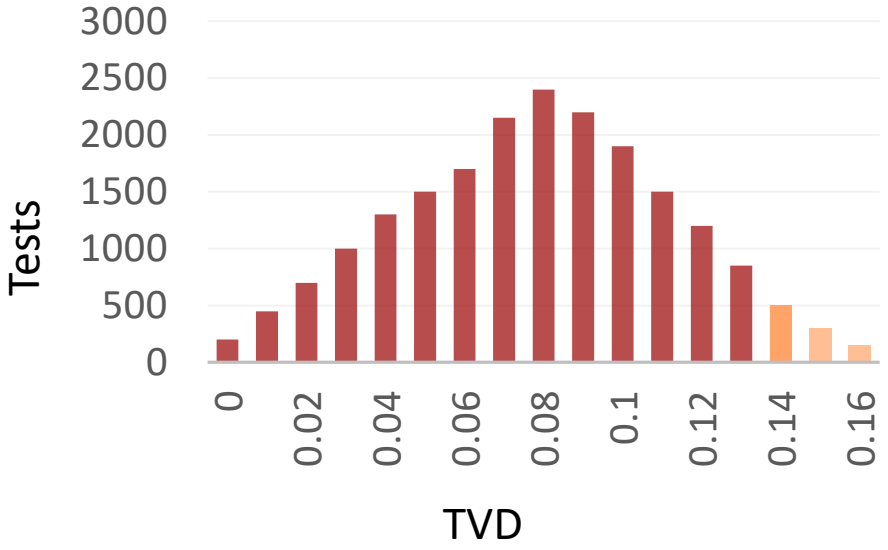


- Random selection produces TVD values between 0 and 0.18. The probability of a TVD value of 0.14 or greater is 8.25% > 5%. The data are *consistent* with the model. **Tests <u>do not reject</u> the null hypothesis**.

Example E1: Jury selection (6 of 6)

- Let us assume that the outcome is (**case C**):

  Recall that the TVD obtained from data was **0.14**



- Random selection produces TVD values between 0 and 0.16. The probability of a TVD value of 0.14 or greater is 4.75% < 5% (by a hair!). It looks as if **tests reject the null hypothesis**.

Example E2: Medical trial (1 of 5)

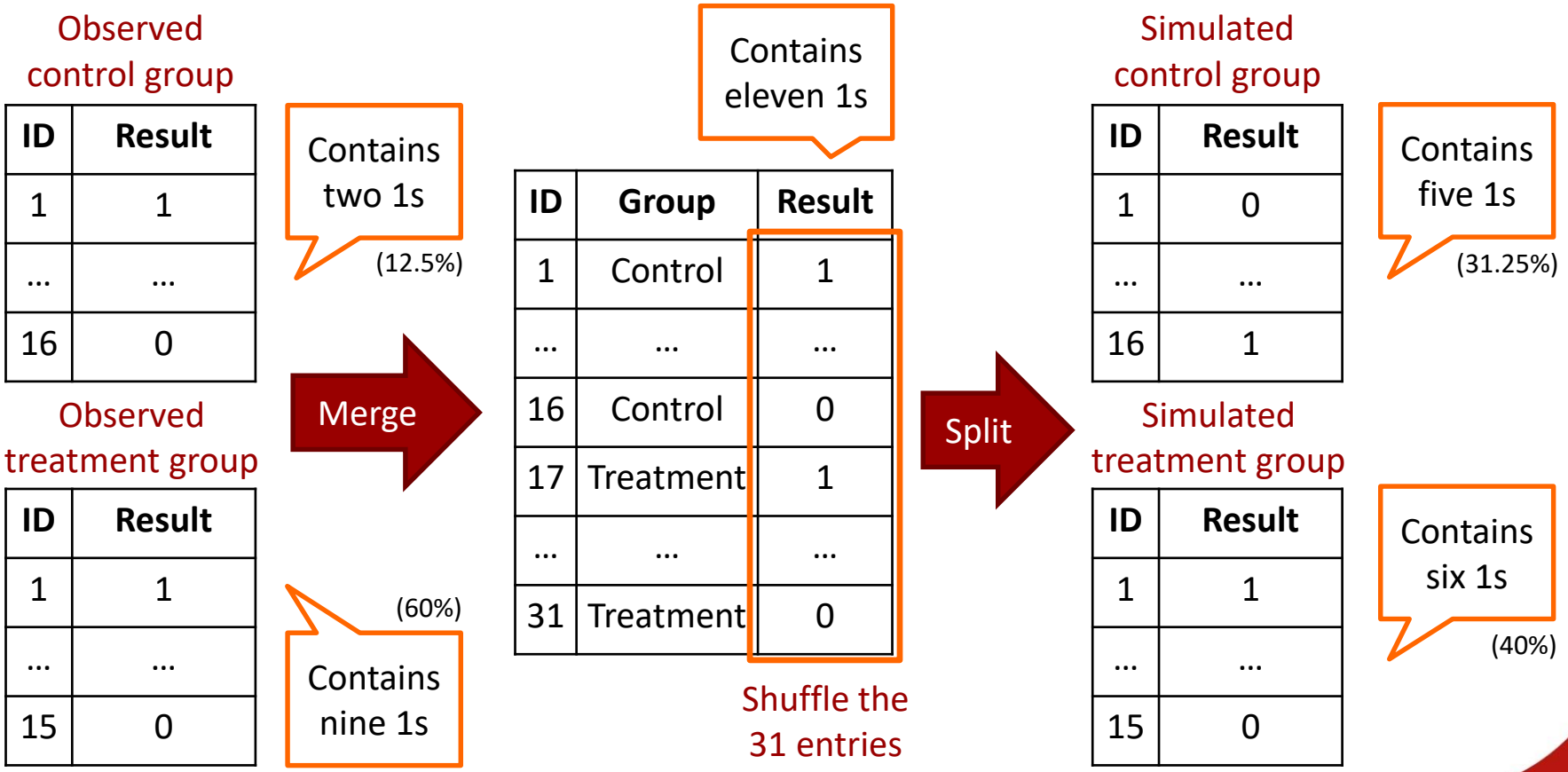
- **Randomized Controlled Trial (RCT) – A/B Testing**
  - 31 patients
  - 16 patients are in the Control Group (sample A) and are given a placebo.
  - 15 patients are in the Treatment Group (sample B) and are given medication for pain relief.
- At the end of the medical trial:
  - 2 of the 16 patients (12.5%) in the control group felt pain relief (*observed statistic of sample A*).
  - 9 of the 15 patients (60%) in the treatment group felt pain relief (*observed statistic of sample B*)
- Could this be **due to chance**? In other words, did people - who were more naturally inclined to recover - enter the treatment group?

Example E2: Medical trial (2 of 5)

Lancaster University

- Assume that *each patient* received a score:
  - **0**: did not have pain relief; **1**: had pain relief.
- **Null hypothesis (H_0)**: The distribution of the potential control outcomes is the same as the distribution of the potential treatment outcomes.
- **Alternative hypothesis (H_1)**: The distribution of the potential control scores is different from the distribution of the potential treatment scores.
- **Overall test statistic**: Use the distance between the two average values. The observed distance is $|0.125 - 0.6| = 0.475$.
- If the null hypothesis is true (assumption of the model), both groups will have the same distribution, so *it does not matter which group is labelled "Control" and "Treatment" group*.
- The rearrangement (shuffling) of values is called a **permutation test**.

Example E2: Medical trial (3 of 5)



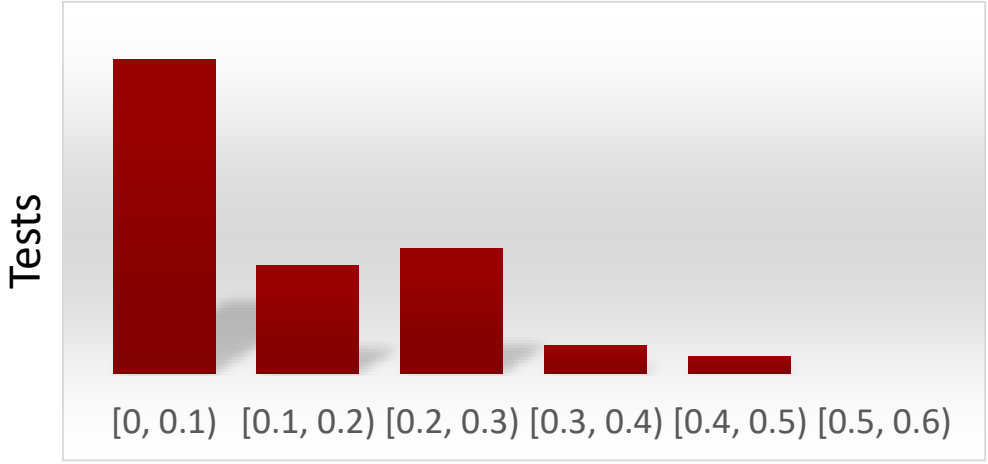
Example E2: Medical trial (4 of 5)

Lancaster University

- **Simulation of tests:**
  1. Recall that sample A had size 16 and sample B had size 15.
  2. Merge the tables of the respective samples.
  3. Shuffle the results in the merged table and use the labels to split it to two samples.
  4. Compute the statistics of the two samples.
  5. Compute the distance between the statistics and record it.
  6. Repeat steps 3, 4 and 5 many times.
  7. Plot the empirical distribution of the distance.

Example E2: Medical trial (5 of 5)

- The empirical distribution of the test statistic **under the null hypothesis** is depicted in the histogram.
- Remember that the observed distance is **0.475**.
- Use the simulated distances to compute P(distance >= 0.475).



- We find that P(distance >= 0.475) = 0.925% < 1%. Therefore, the **alternative hypothesis holds** and the treatment must have been the **cause** of the observed difference (assuming that this was a randomized controlled trial).

Example E3: Median salary (1 of 4)

Lancaster University

- Our objective is to measure a parameter of a population, for example the **median salary of a metropolitan area** (i.e. the salary that is equal to or greater than 50% of the salaries, also known as the 50th percentile).
- The median of the population (parameter) can be computed only if we know the salaries of all residents in the metropolitan area.
- If we have access to the population (via surveys), we can:
  - draw multiple random samples of a pre-determined size
  - compute the median of each sample (statistic)
  - plot the histogram of the median to visualize its **variability**
- What happens if we have **only one sample** but no access to the population (hence, we cannot draw more samples)?

Objective: Infer the parameter from the statistic

Example E3: Median salary (2 of 4)

Lancaster University

- The **bootstrap method** uses re-sampling (with replacement) of the observed sample to estimate the variability of the statistic by assuming that the sample itself has a *very similar* distribution to the population.
- A re-sample should have *the same size* as the sample.
- The idea behind bootstrap is that we compute the "statistic of inference" (here: the median) for each re-sample. The different medians will give us a way of describing the variability of the *sample median* for the observed sample drawn from the population.
- A **confidence interval** is used to describe the variability, as opposed to the range between the lowest-valued and highest-valued medians.
- Bootstrap **should not be used** when estimating very high/low percentiles (e.g. max/min) or parameters that are affected by rare elements in the population (which might not be present in the observed sample).

Example E3: Median salary (3 of 4)

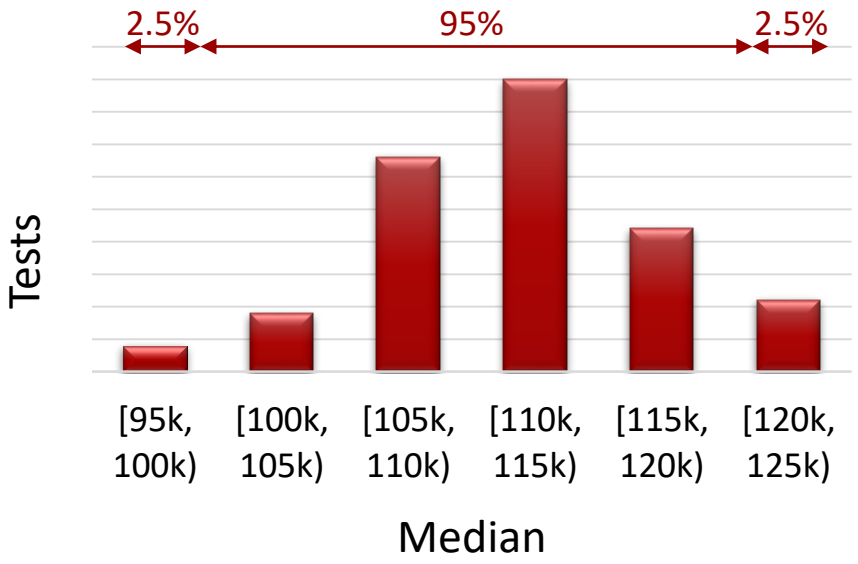
Lancaster University

- **Simulation of tests:**
  1. Assume that the size of the observed sample is 200.
  2. Randomly re-sample (with replacement) the sample; the re-sample should have the same size as the sample, i.e. 200.
  3. Compute and record the statistic of the re-sample (here: the median)
  4. Repeat steps 2 and 3 many times, e.g. 1,000.
  5. Plot the distribution of the 1,000 values of the median.
  6. Compute the confidence interval (95% or 99%).

  Note: If we repeat the simulation, the confidence interval should still contain the population parameter with the same likelihood, even if the sample statistic is different from the population parameter.

Example E3: Median salary (4 of 4)

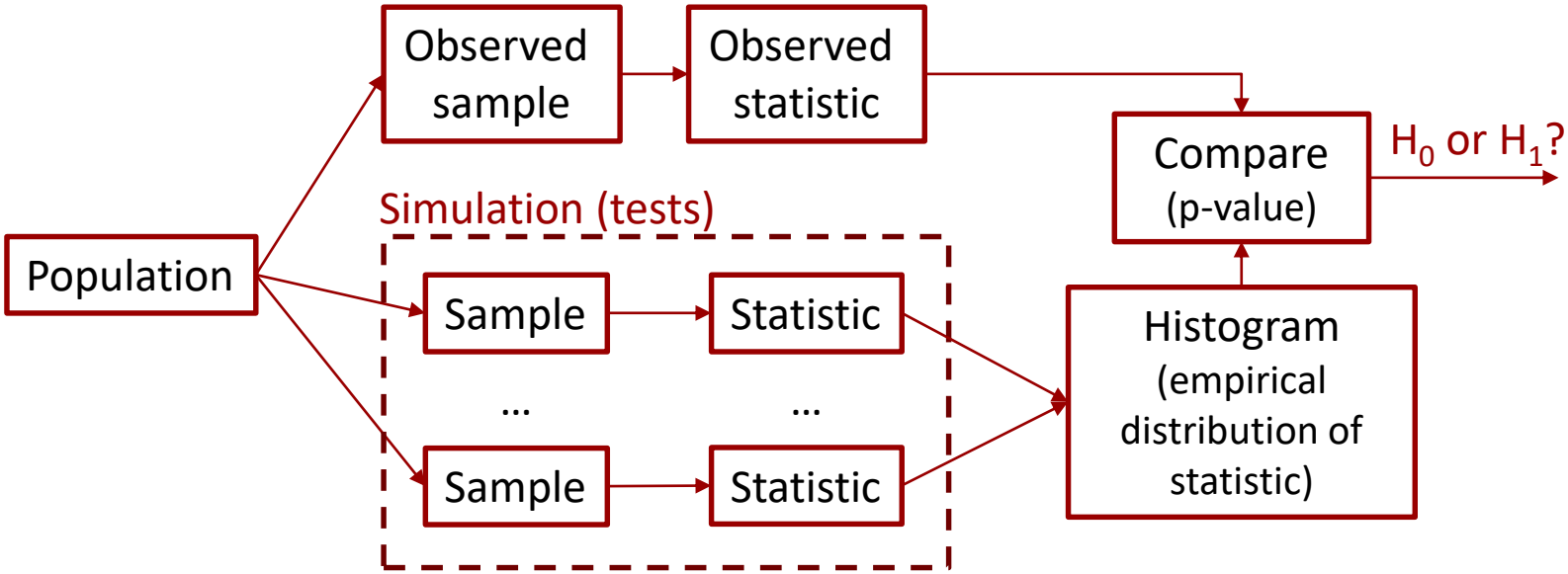
- Let us assume that the distribution of the median is shown in the figure:
- To compute the **95% confidence interval**, we need to subtract the 2.5% percentile from the 97.5% percentile (quantile in Python).
- In this example, the **95% confidence** interval is from 99.99k to 120.55k.



- Higher confidence (e.g. 99%) produces a wider interval for the estimated statistic and introduces uncertainty (trade-off).

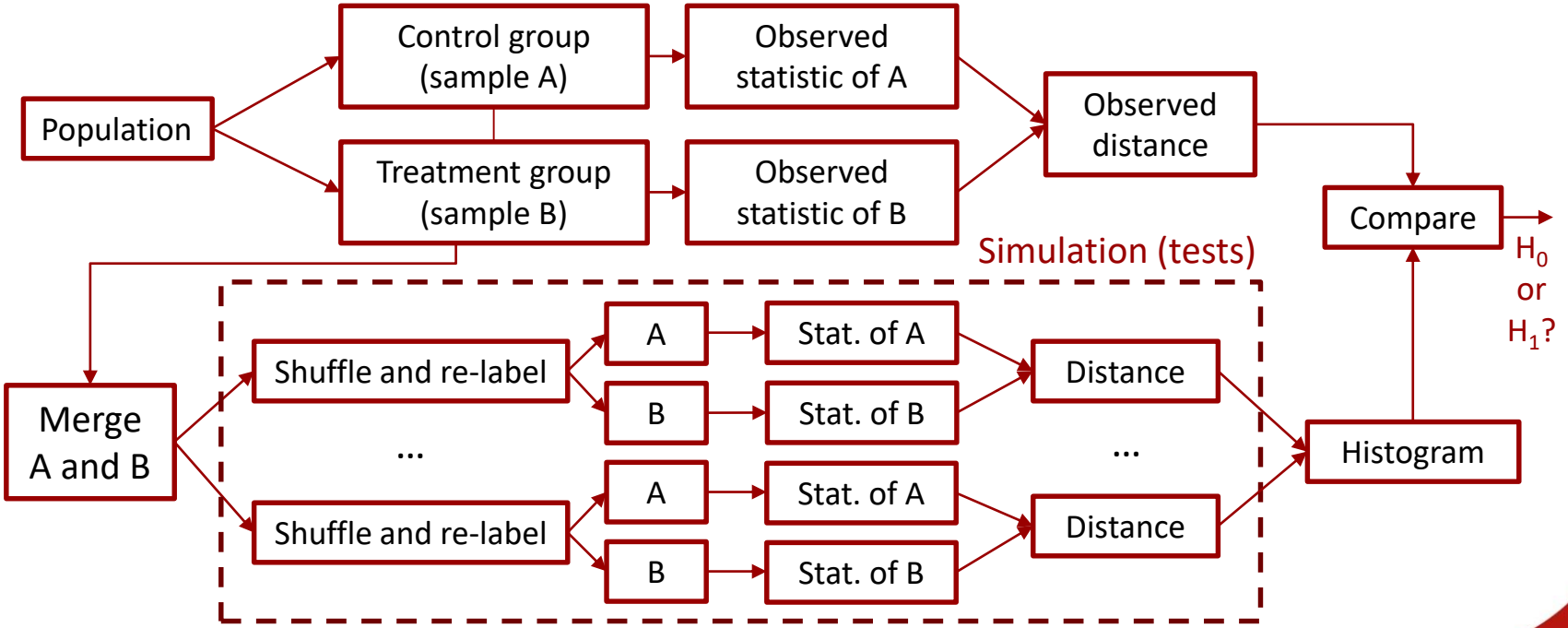
Hypothesis Testing
(can take samples from population, example E1)

- If we have access to the population and we can draw multiple samples (without replacement), we can conduct hypothesis testing.



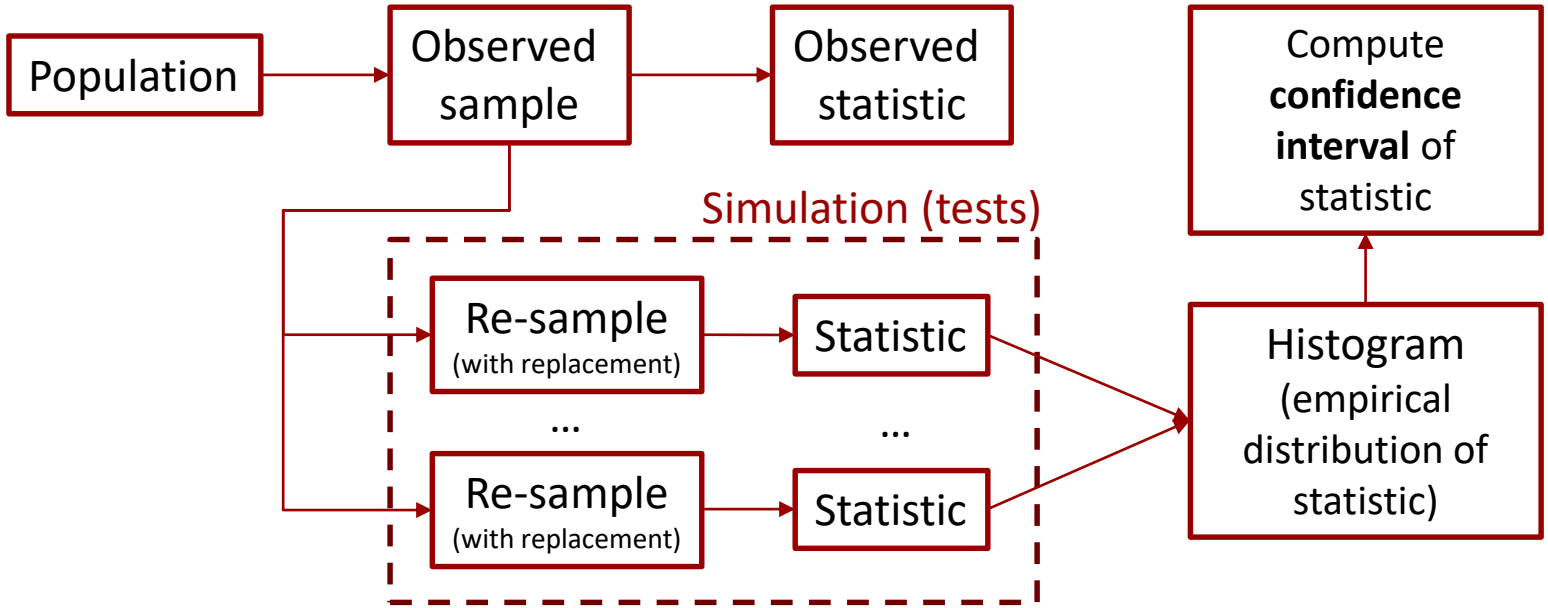
Hypothesis Testing (A/B testing, example E2)

- If we do not have access to the population and have two samples, A and B, we can conduct A/B testing to compare the distributions of the samples.



The Bootstrap Method
(can re-sample the observed sample, example E3)

- If we do not have access to the population but only to a sample, the bootstrap method can be used to measure the variability of the statistic.



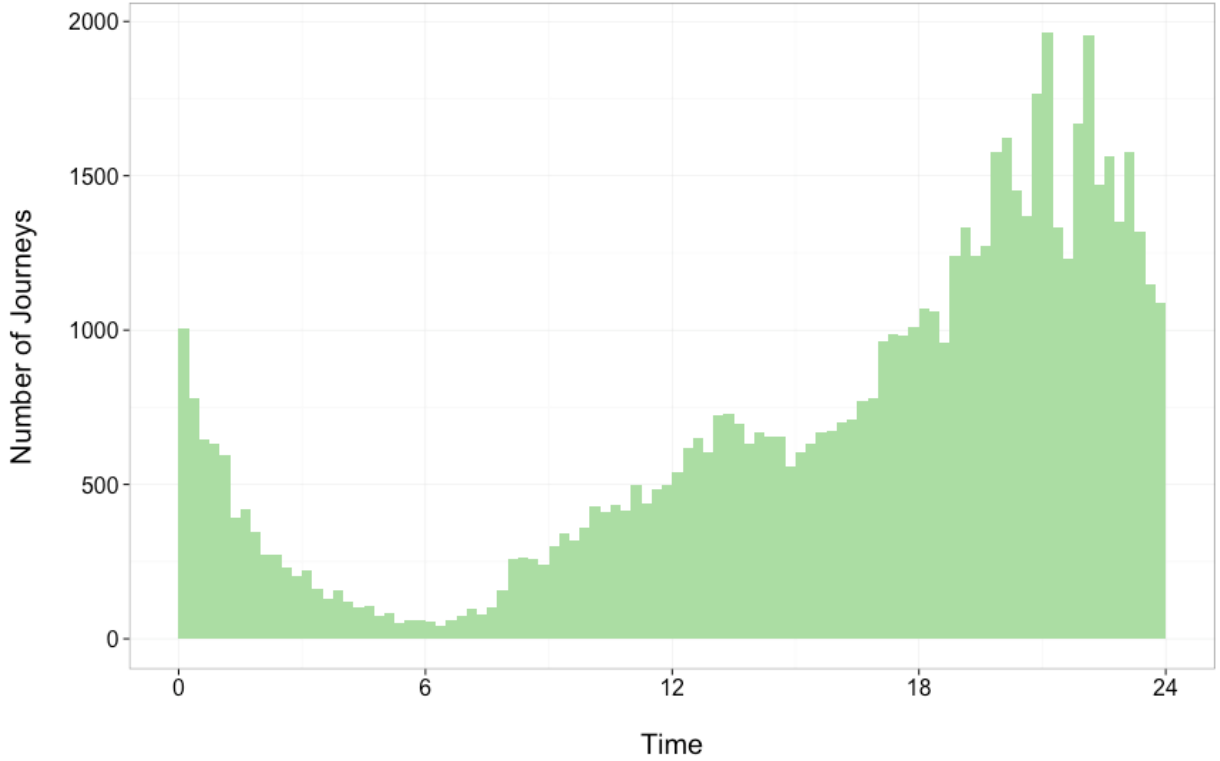
Analysis: Exploratory or Confirmatory?

Lancaster University

- Confirmatory data analysis:
  - Modeling and testing hypotheses (*as described in the first part*)
- Exploratory data analysis (EDA):
  - No model. No hypothesis.
  - To gain fundamental understanding of the data.
  - Basic tools: plots and summary statistics.
    - Generating summary statistics
    - Plotting feature distributions
    - Transforming variables
    - Looking at all pairwise relationships between features

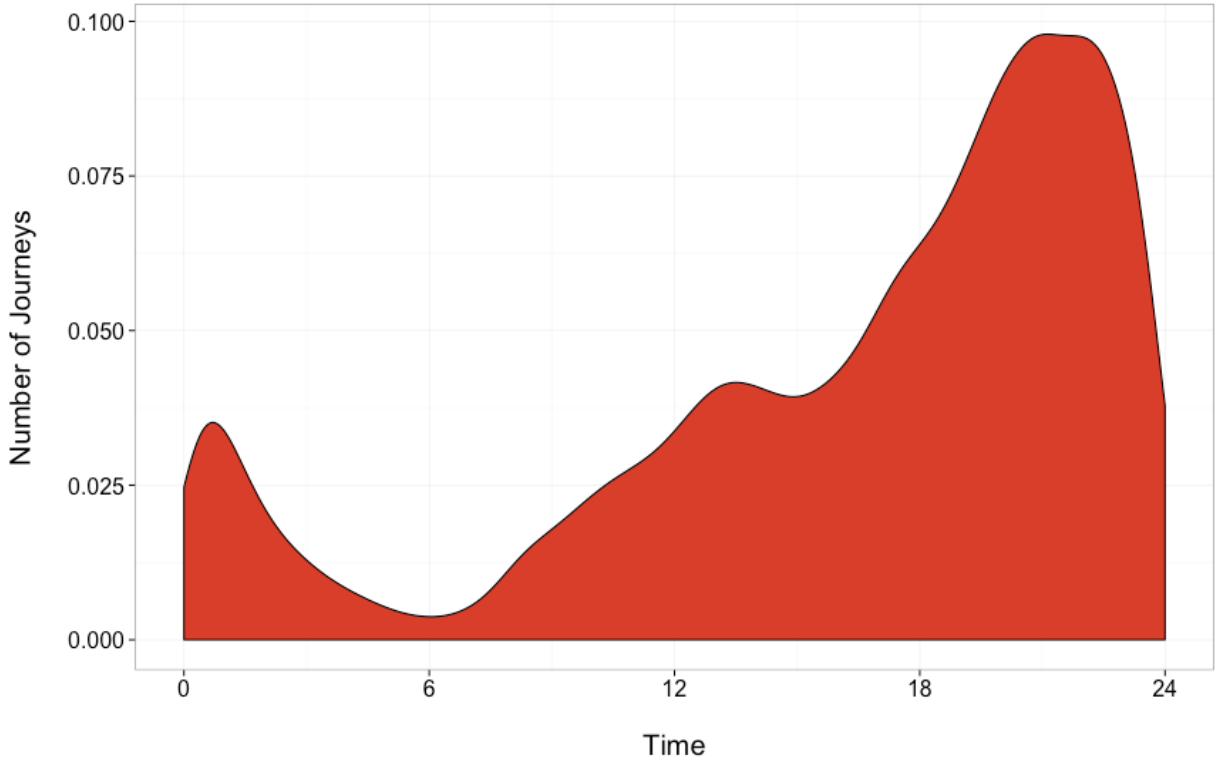
Basic plots: Distribution functions

- Histogram: exhibits the distribution of point frequencies in a dataset.



Basic plots: Distribution functions

- Probability density (PDF): presents probability of values (area under curve).



Basic plots: Distribution functions

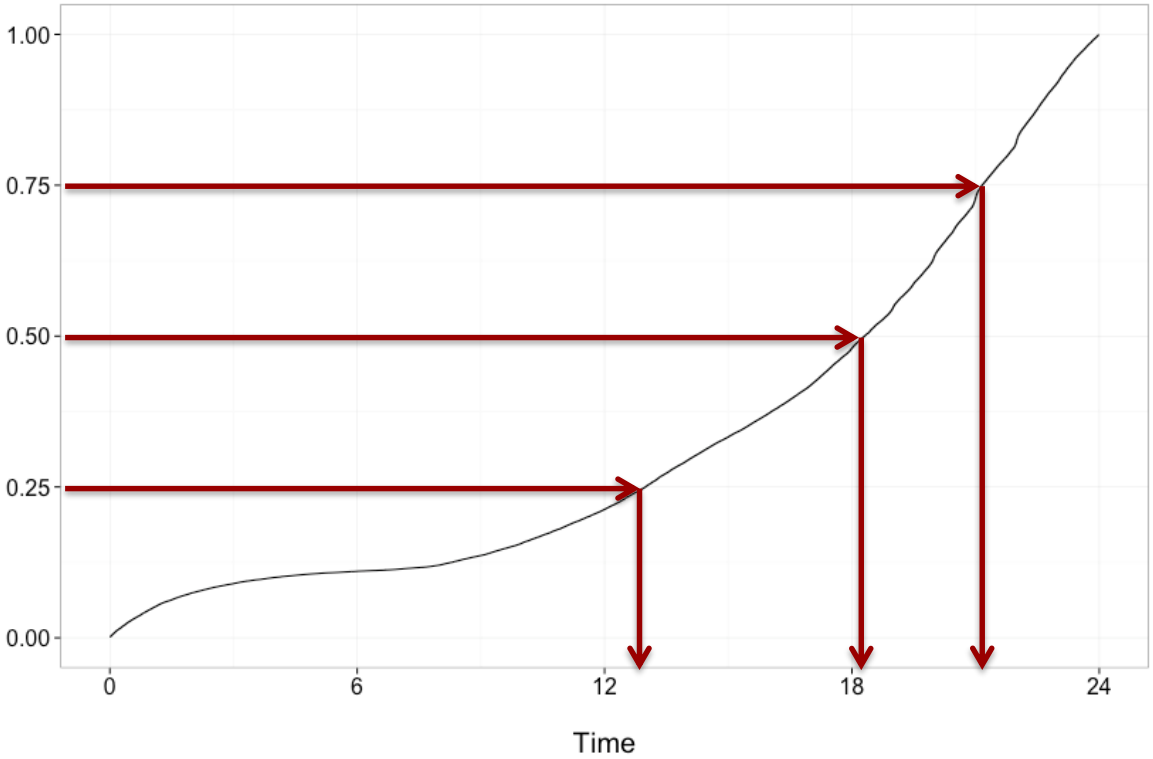
- Cumulative distribution (CDF): illustrates the percentage (probability) of values in a certain range.
- Slope signifies frequency.
- Quickly highlights:
  - Mean
  - Percentiles (including quartiles)

discrete *X*

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

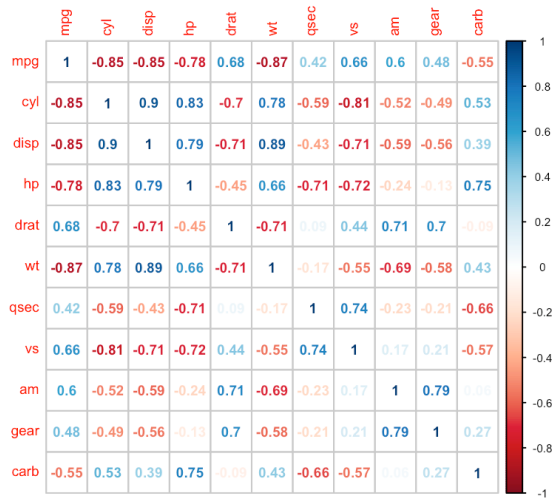
cumulative *X*

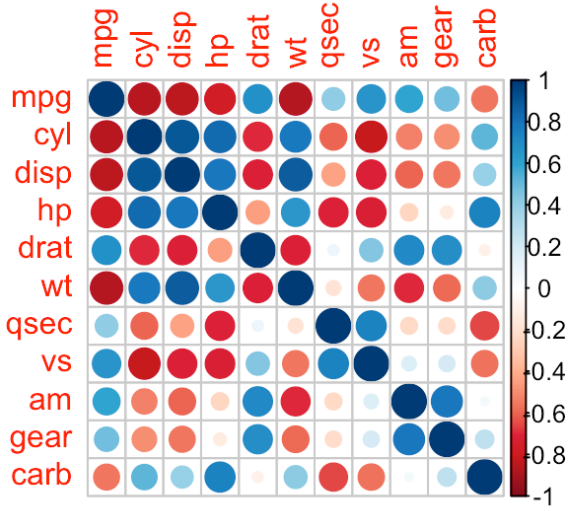
$$F(x) = \int_{-\infty}^{x} f(t)dt$$

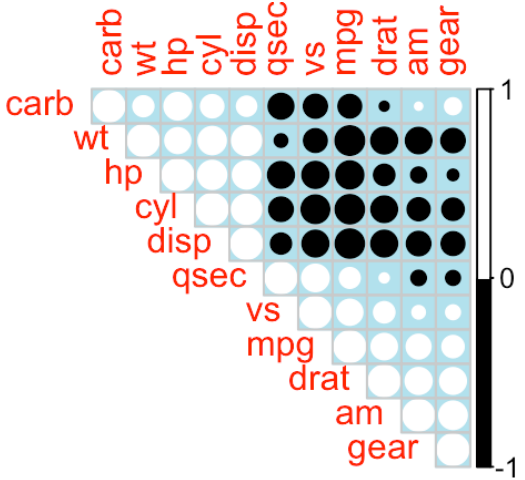


Basic plots: Correlograms

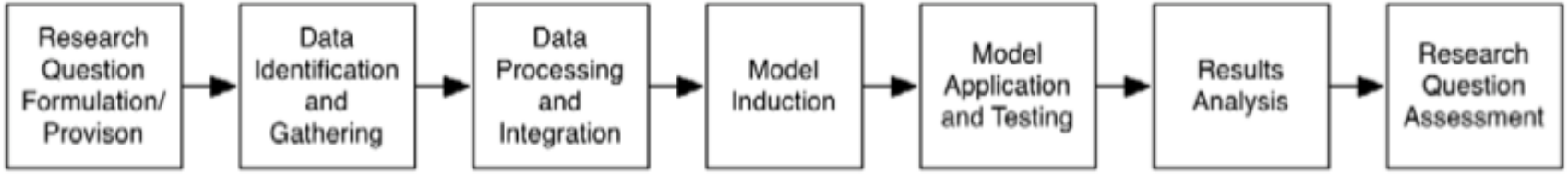
- Useful for seeing how engineered features correlate.
- Only applicable when dataset dimensionality is relatively small.
- Correlation ≠ causation!







Completing the pipeline



- We have:
  - Formulated our research question and objectives
    - Identified our methodology
    - Inspected ethical considerations and conflicts
  - Gathered our datasets and cleaned them
  - Integrated and transformed datasets
  - Processed our data using big data tools and technologies
  - Engineered our features and built models
  - Inspected our data and tested hypotheses

What is visualisation?

- “The action or fact of visualizing; the power or process of forming a mental picture or vision of something not actually present to the sight; a picture thus formed.” [oed.com]
- “The use of computer-generated, interactive, visual representations of data to amplify cognition.” [Card, Mackinlay, & Shneiderman 1999]
- “Transformation of the symbolic into the geometric” [McCormick et al. 1987]

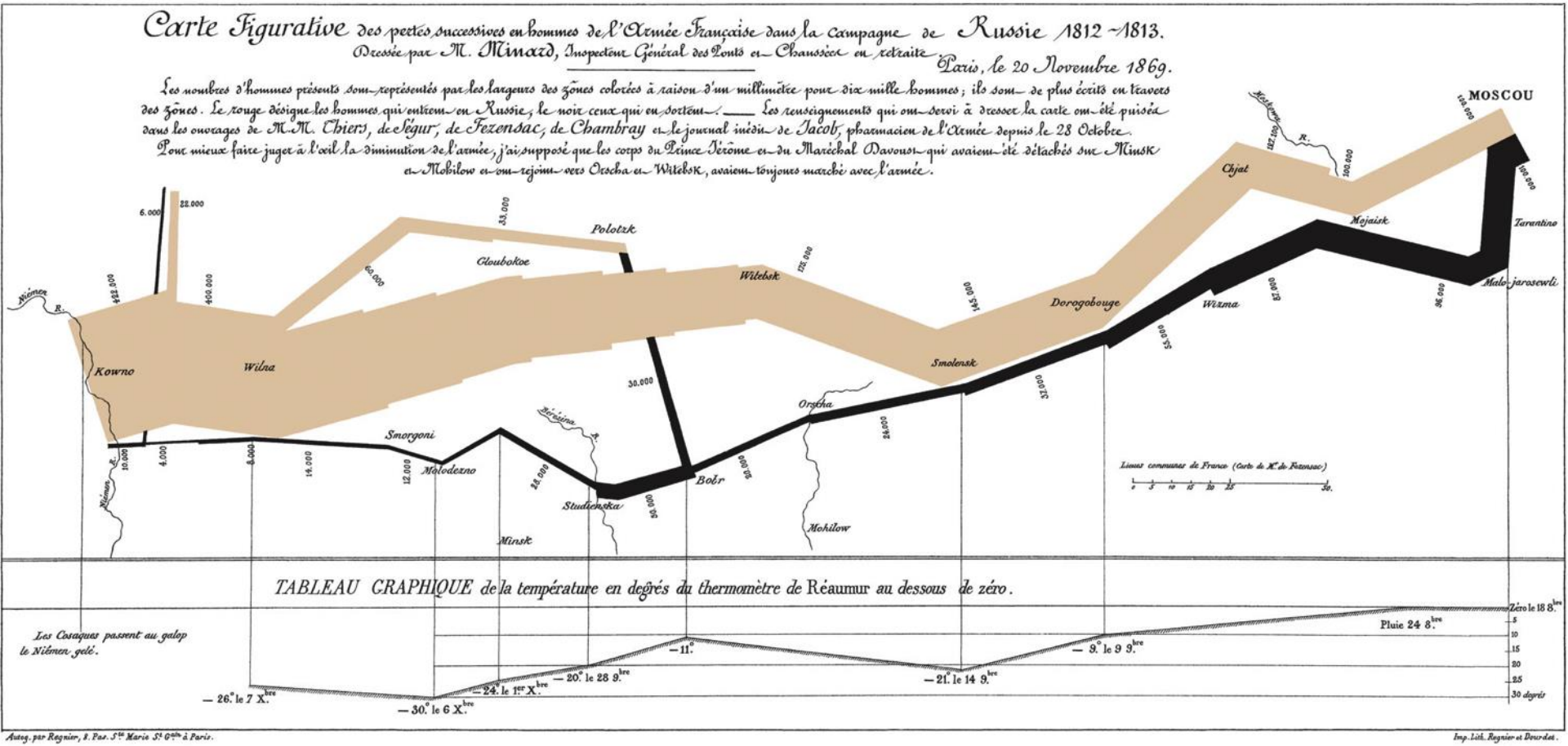
What to answer before visualising

- What is the message I want to convey?
  - Is any part of the message more important?
- Who is my audience?
  - What do I assume they know and care about?
  - What do I want them to do with what I give them? Is that clear to them?
- What is the communication channel?
  - Is the visual element on its own?
  - Is there room for multi-modal communication?
- Is the channel one- or two-way? What will I do with feedback/participation?
- What is the outcome that I want?

The outcome?

- Could be far beyond a visual language to communicate results/data/etc.
  - Education, showing a different side, summarizing or dissecting rich data
  - Raising awareness
  - Combining disparate pieces of information
  - Asking difficult questions
  - Starting a dialogue
  - Getting feedback
  - Collective brain storming
  - Crowd-sourced / collaborative exploration
  - Having fun!

Example: Summarising a wealth of data



Example: Dissecting a rich data set

Lancaster University

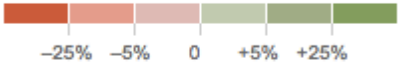
All Spending | Types of Spending | Changes | Department Totals

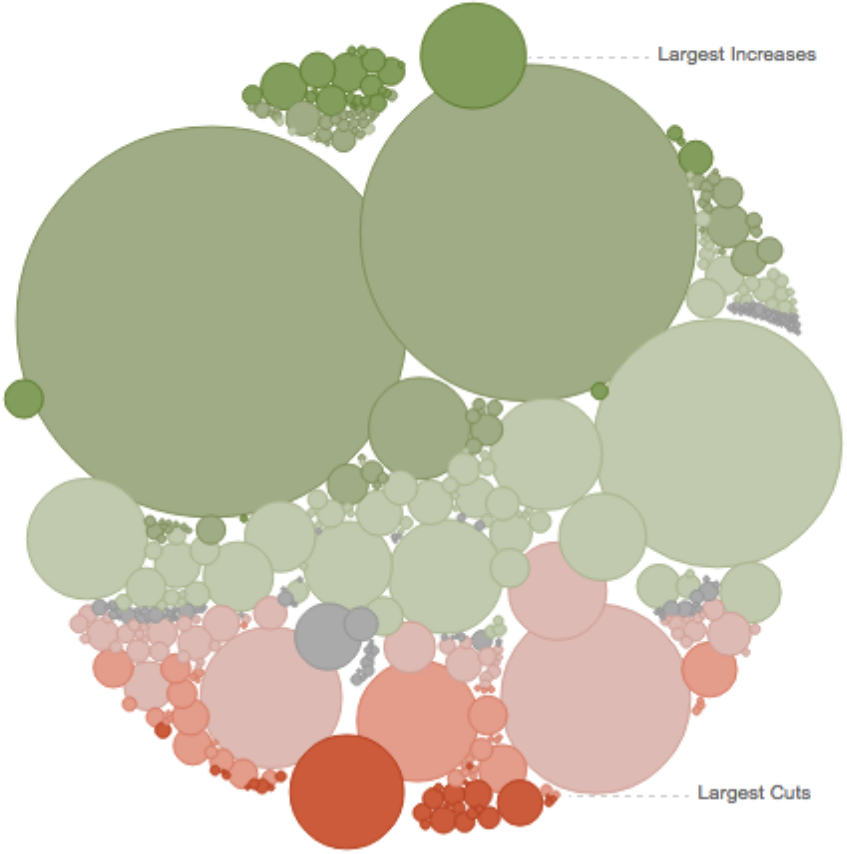
How $3.7 Trillion Is Spent

Mr. Obama's budget proposal includes $3.7 trillion in spending in 2013, and forecasts a $901 billion deficit.

Circles are sized according to the proposed spending.

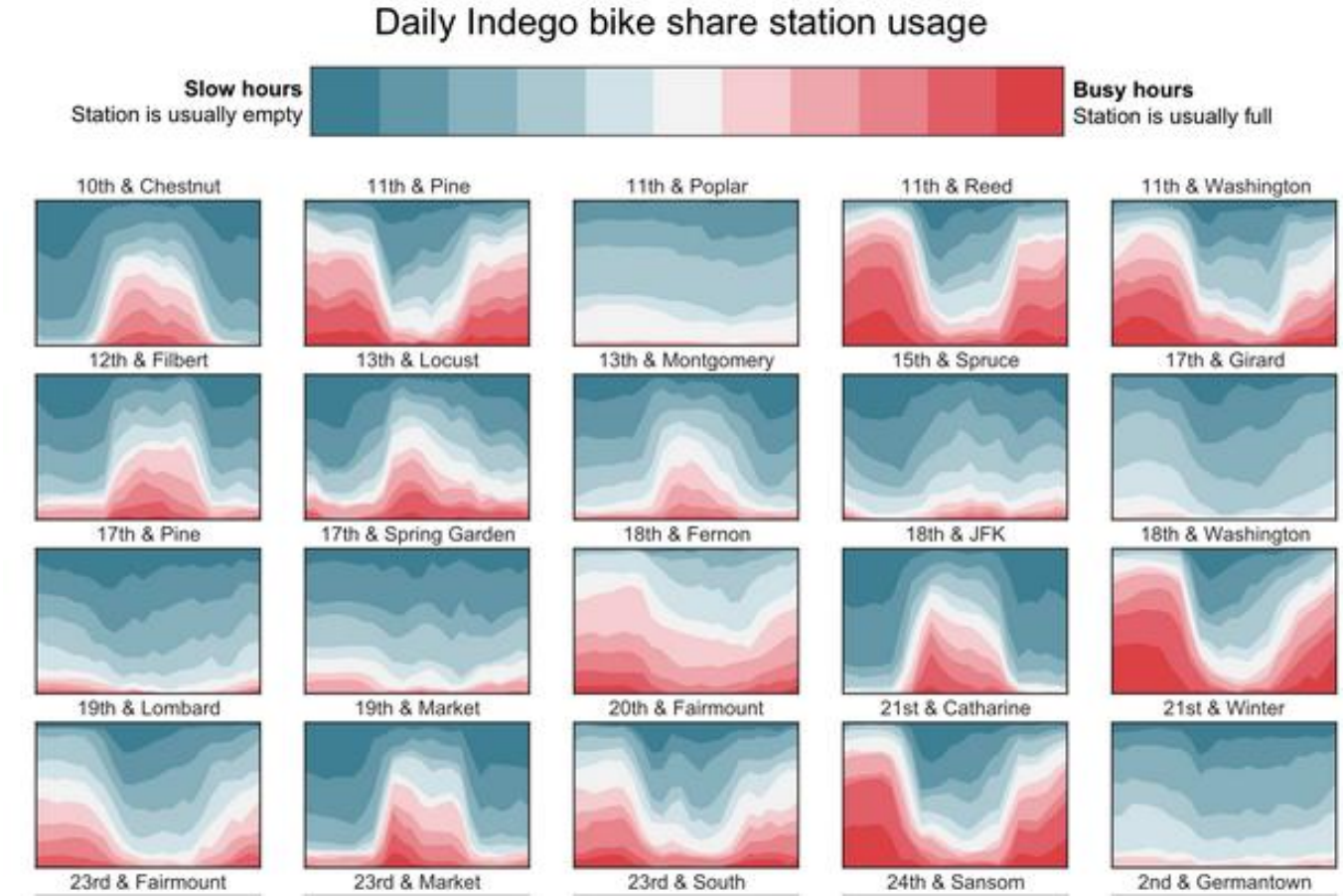
Color shows amount of cut or increase from 2012.





The proposal forecasts a $901 billion deficit.

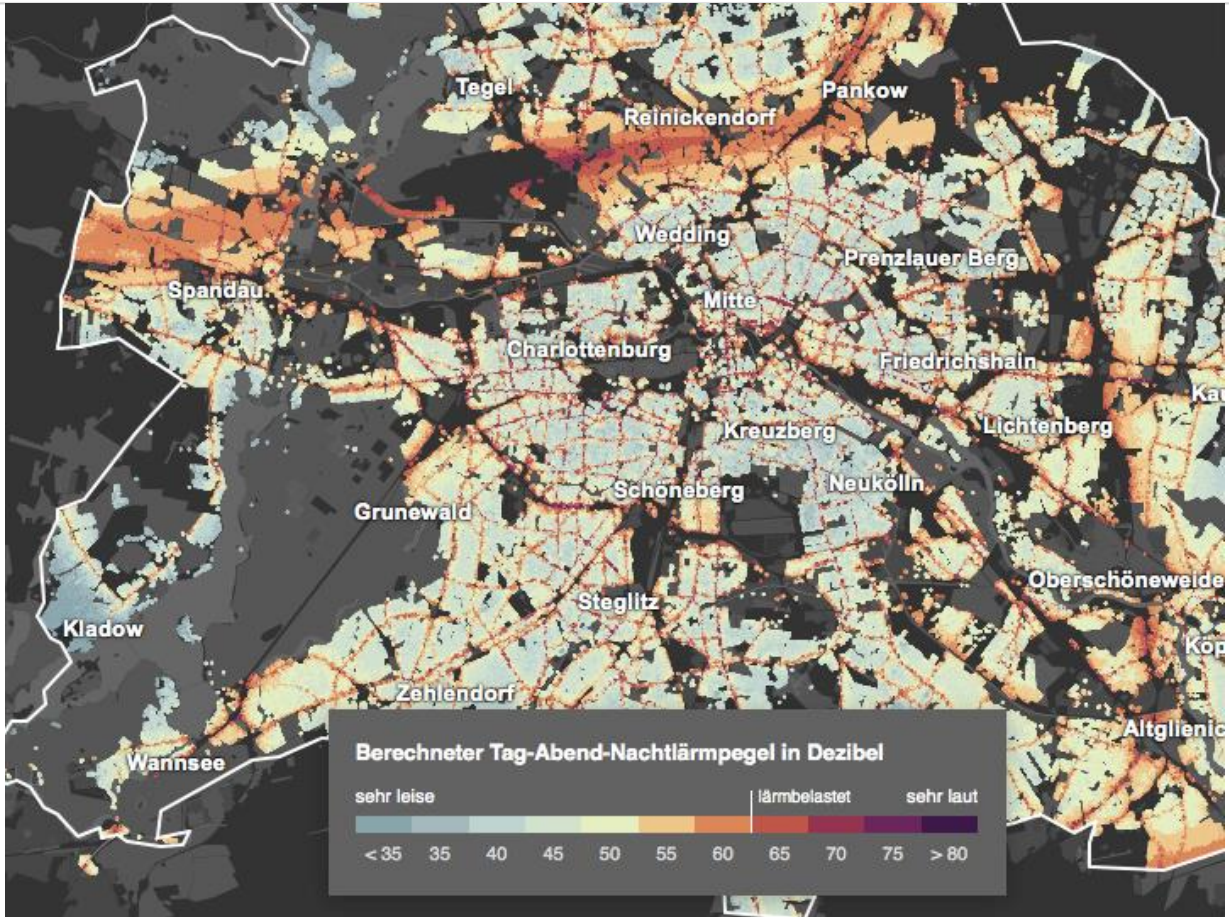
Example: Dissecting a rich data set



Example: Showing a different side

To watch: https://www.youtube.com/watch?v=RY2TF6QEOGs
Colour signifies time of day (e.g., white = morning).
Length of queue represents amount of traffic.

Example: Crowd-sourced / collaborative exploration



Do

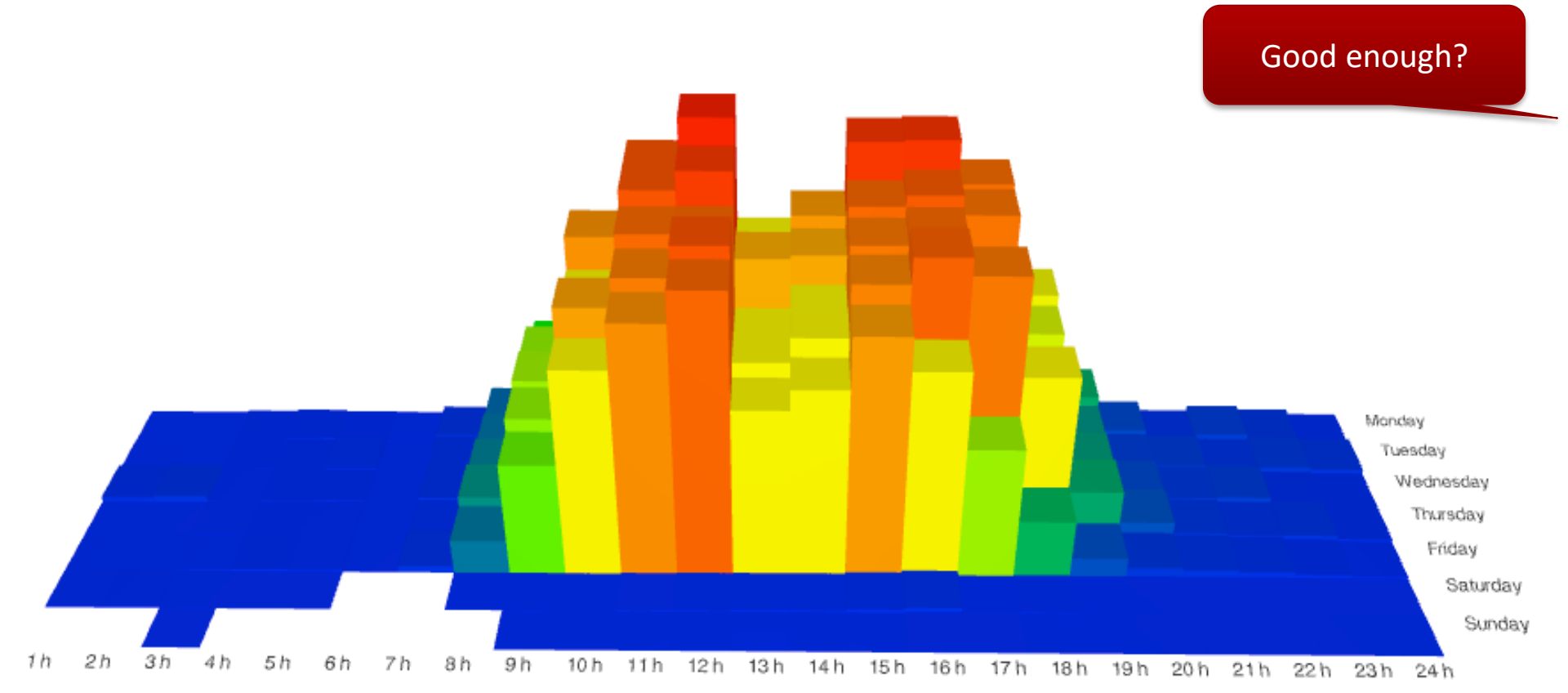
- Identify dimensions
- Use basic plots (for EDA and beyond)
- Embrace uncertainty
- Accentuate the take away points
- Utilise graphing libraries

Identify dimensions

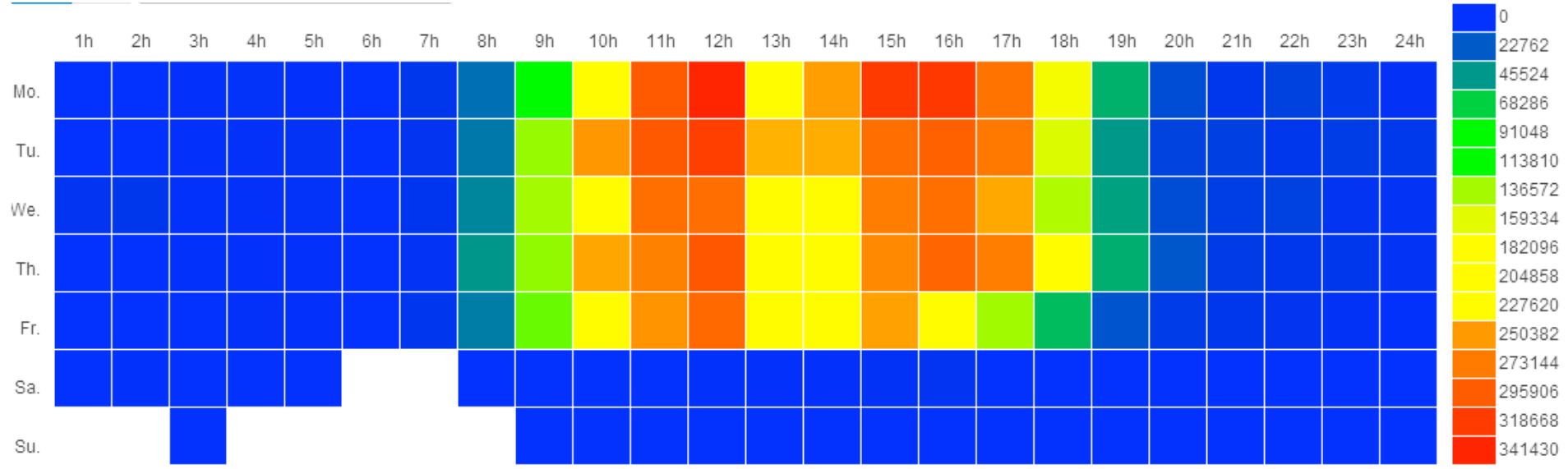
- What dimensions of data do you want to display?
  - Most media communicate between 2 and 5 dimensions without confusion
- What do you want to show?
  - Display distributions, Untangle dependencies, etc.
- This will dictate what you choose for visualising different dimensionalities.

Identify dimensions

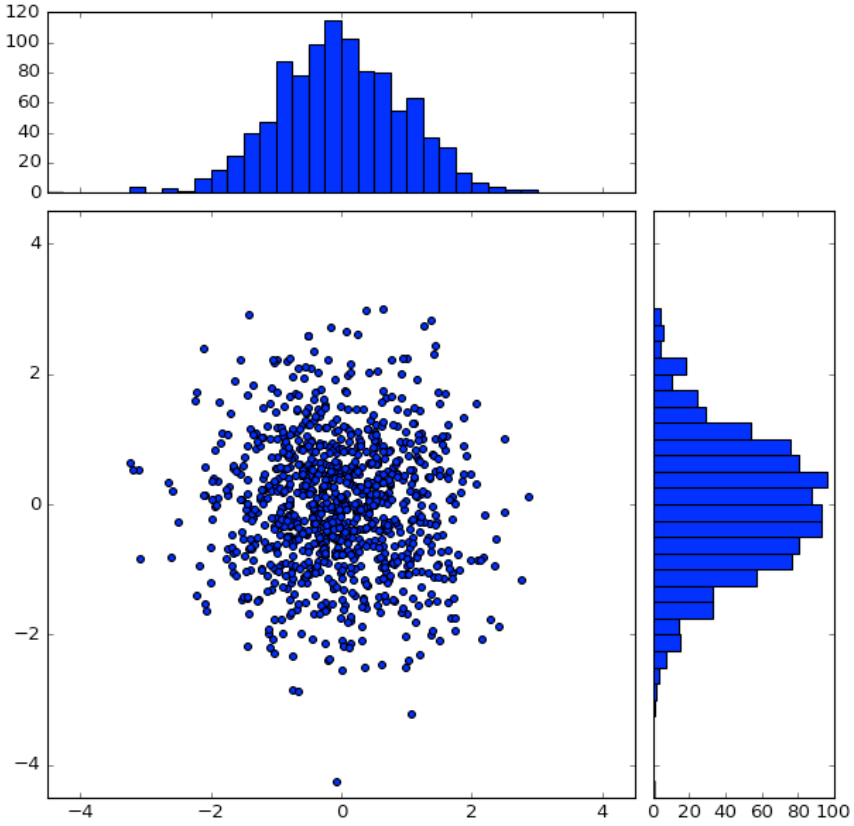
Lancaster University



Identify dimensions



Which dimensions are important?

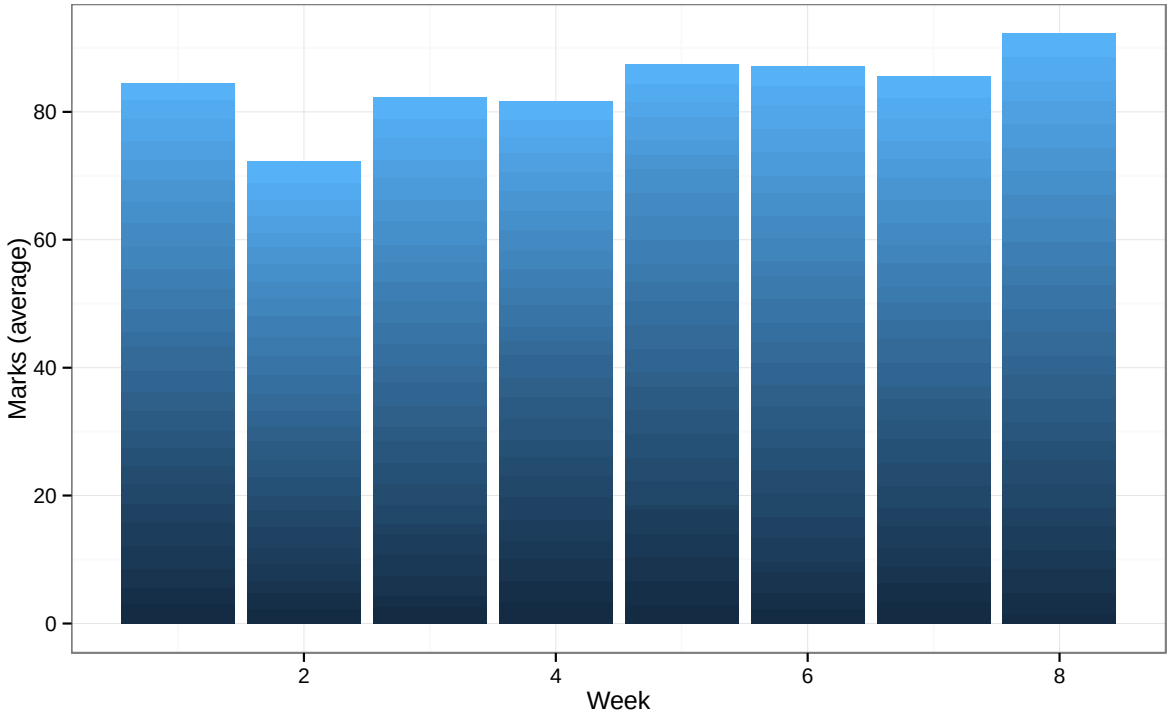


Basic plots

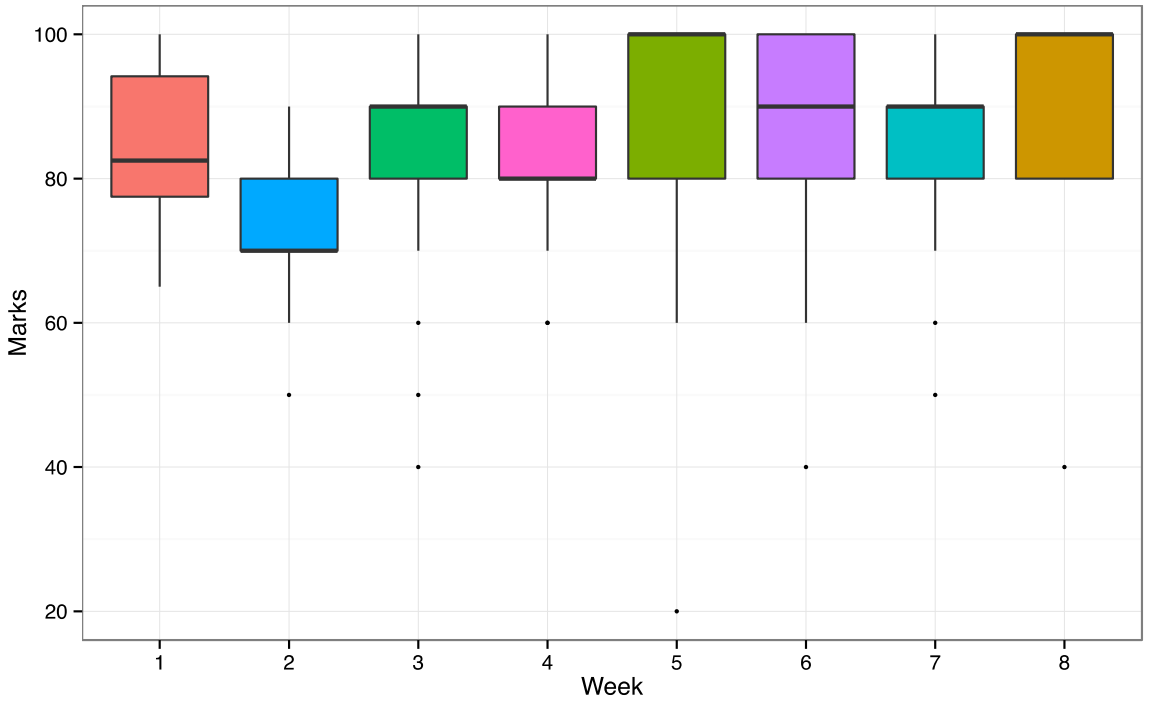
Lancaster University

- Basic plots are very useful
  - Exploratory analysis (see previous slides)
  - Internal communication
- Simple visual representation can illuminate your path
  - Examine how data is distributed
  - Rapidly investigate how features are varied for separate classes
  - Identify anomalies and misinterpretations

Acknowledge the nuances of data

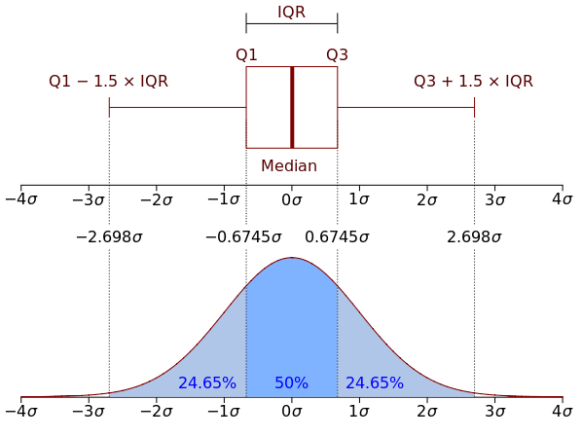


Acknowledge the nuances of data

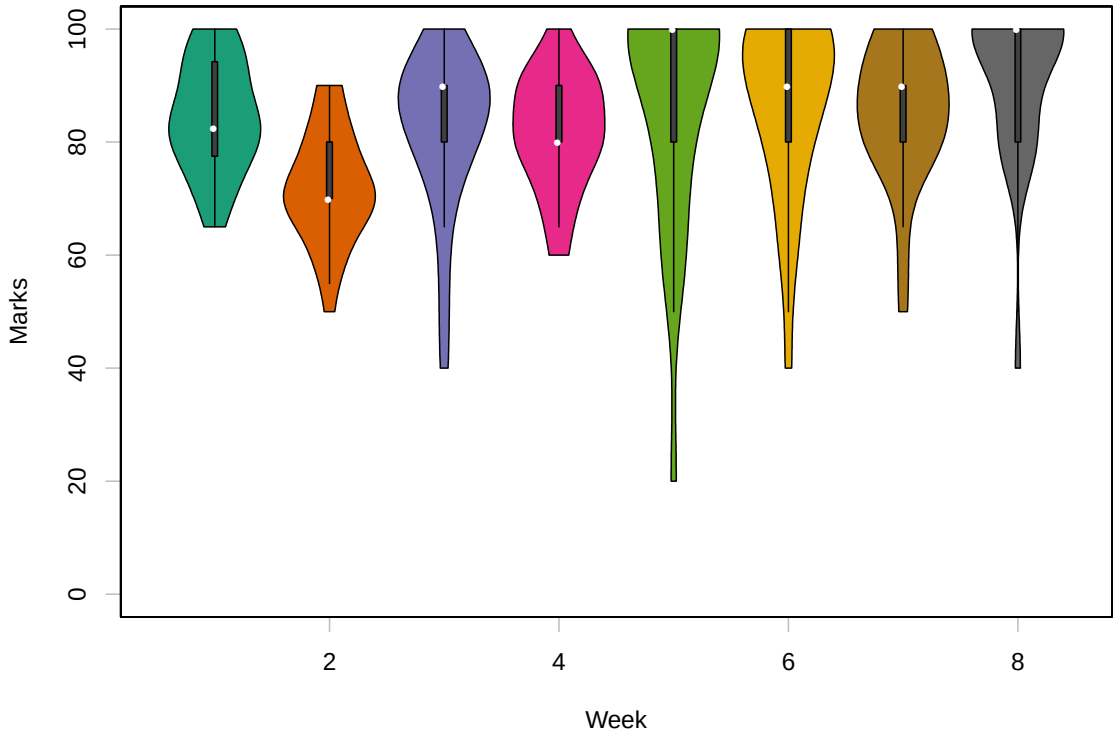


Box plots (aka box-and-whisker diagrams)

- Median, InterQuartile Range
- (potential) Outliers
- Easy assessment of the data distribution



Acknowledge the nuances of data



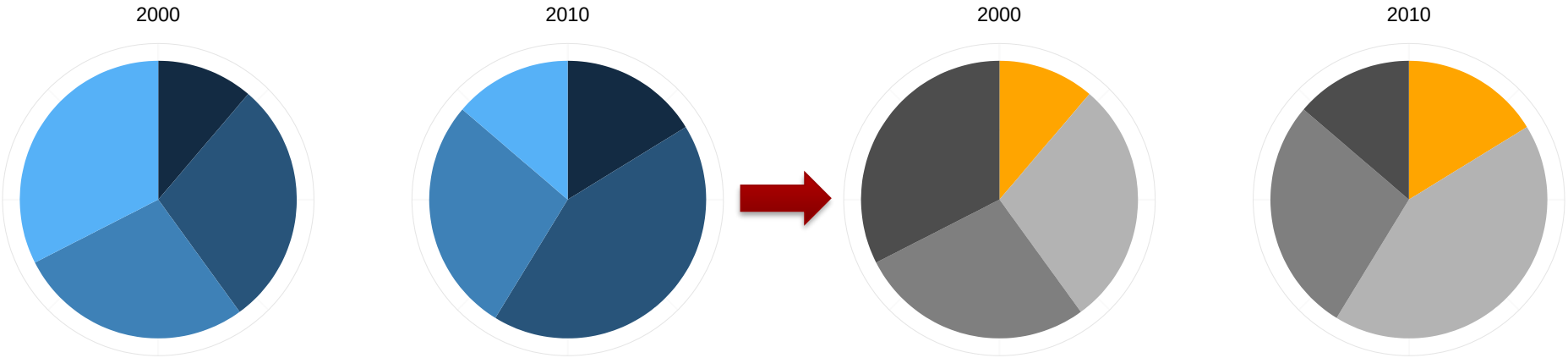
Violin plots

- The **outer shape** represents frequency of data values, i.e., the probability density (the thickest section represents the *mode,* i.e., the value that is repeated more often than any other).
- The **thick line** signifies the InterQuartile Range (IQR).
- The **central** (white) **dot** represents the median.

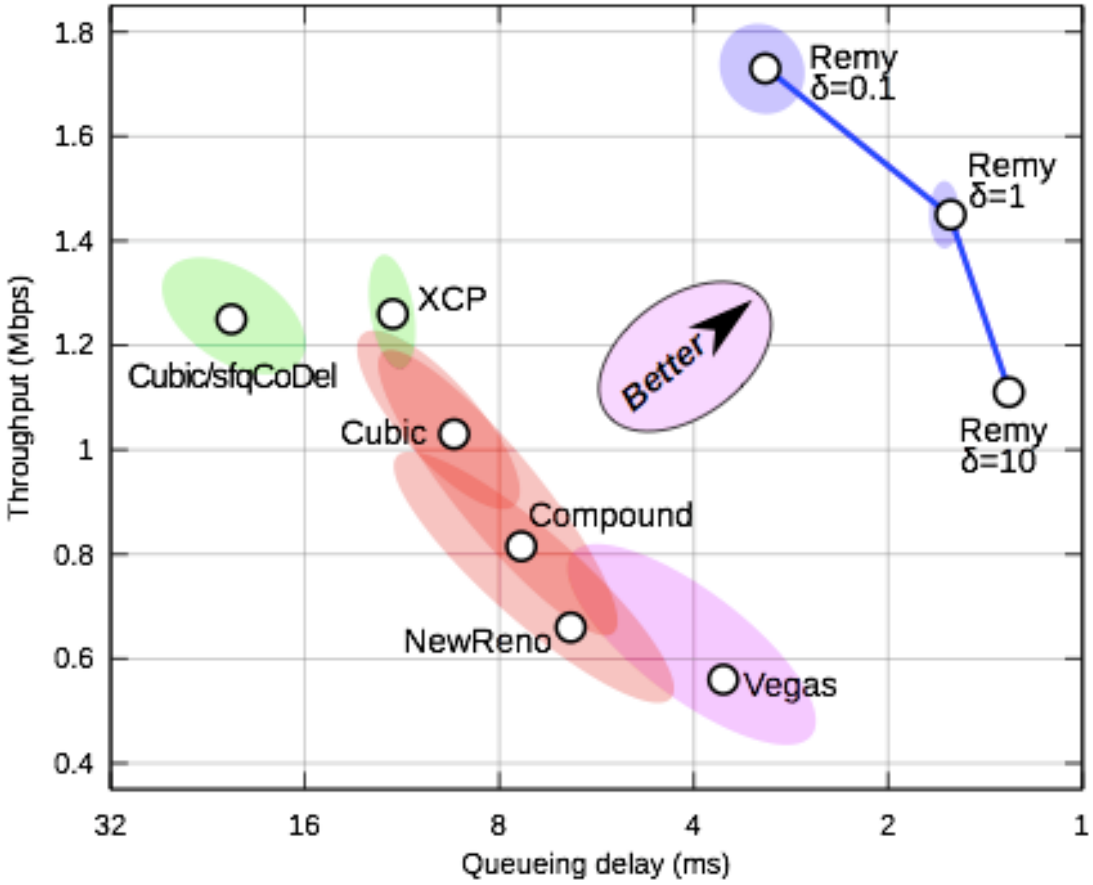
Accentuate the take away points

- We are all bombarded with information all day long.
- Don’t assume people will be interested as you are.
  - Help them along; make sure they “take the medicine”
- Know your audience: what they care about, their level of knowledge.

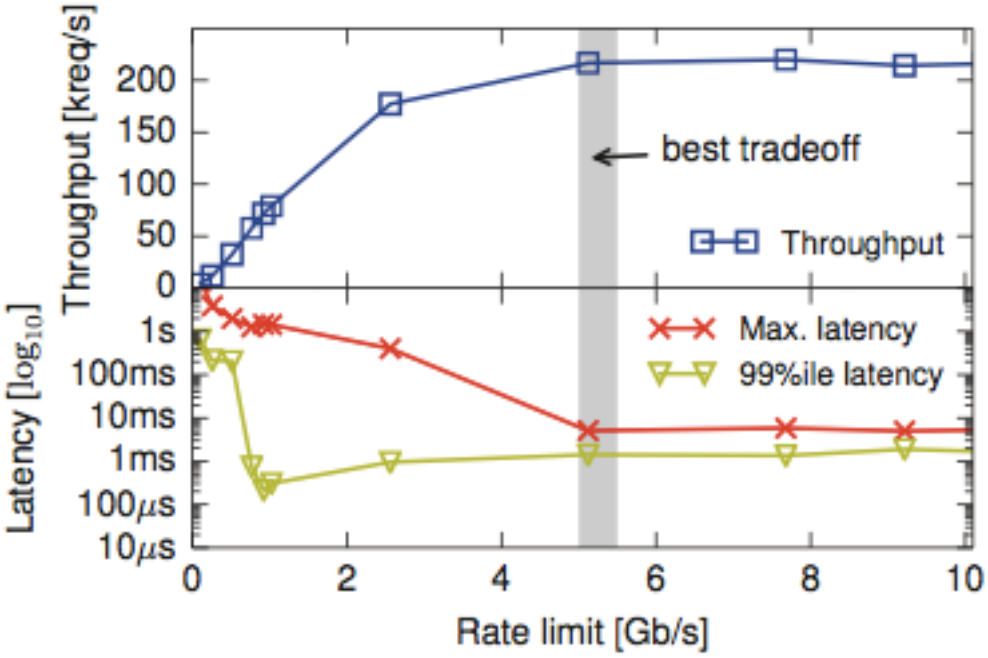
Accentuate the take away points



Accentuate the take away points



Accentuate the take away points



Familiarise yourself with graphing libraries

Lancaster University

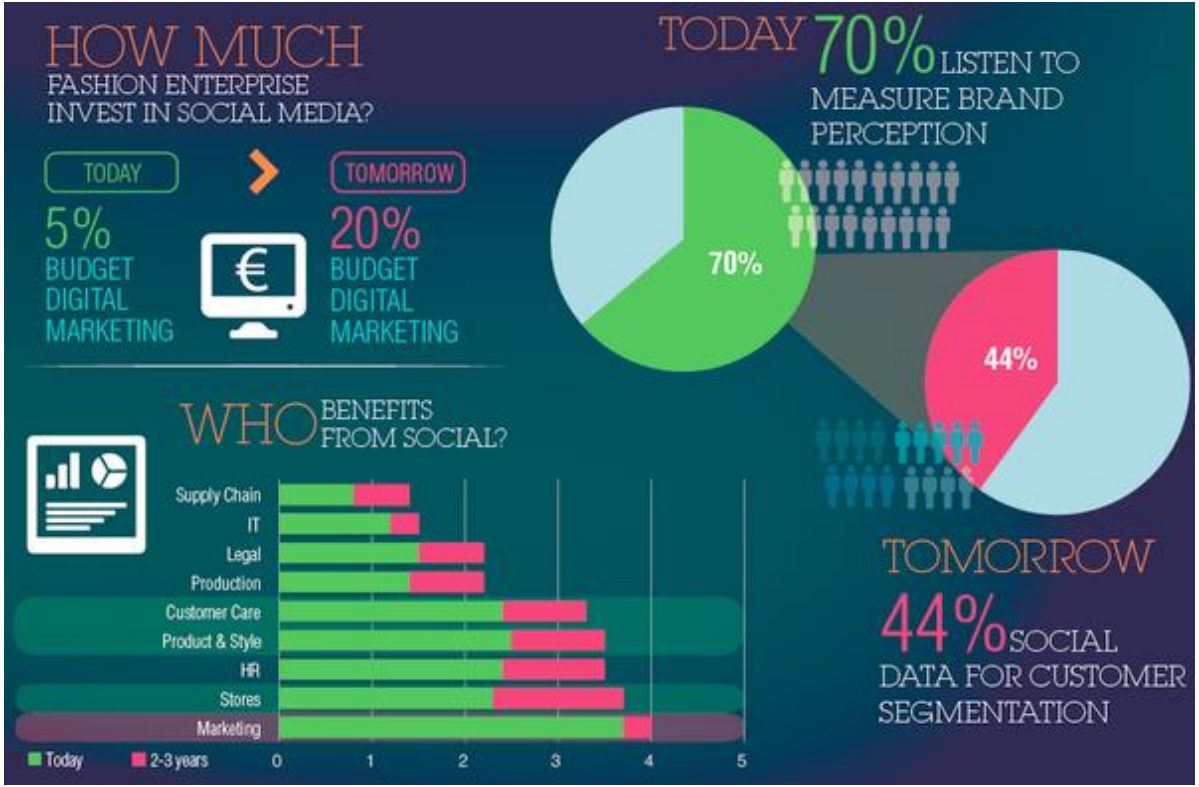
- Power tools.
  - Allow you to prototype and create low-effort exploratory analyses.

- Some examples:
  - Python: matplotlib, Seaborn, VisPy, Bokeh, Pygal
  - Javascript: Chart.js, D3, Flot, HighCharts, Protovis, amCharts, Google Charts
  - R: ggplot2, lattice, ggvis, rCharts, rgl
  - Cross-language: Plotly, Jupyter

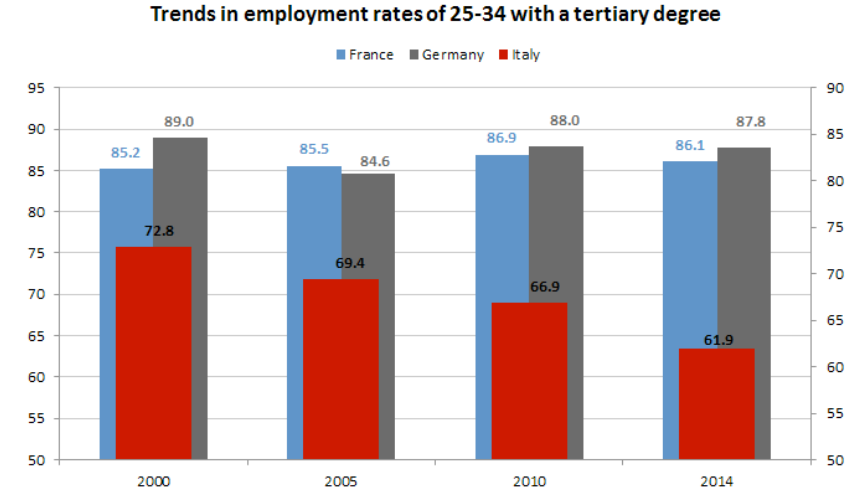
Don’t

- Forget/ignore the basics
  - Specify units
  - Use consistent scales
  - Communicate the data
  - Only communicate what is relevant
  - Know how to use colours
- Mindlessly plot anything with geo-coordinates on a map
  - Be conscious of the message you are trying to convey
- Feel the need to conform

Basics: specifying units



Basics: use consistent scales



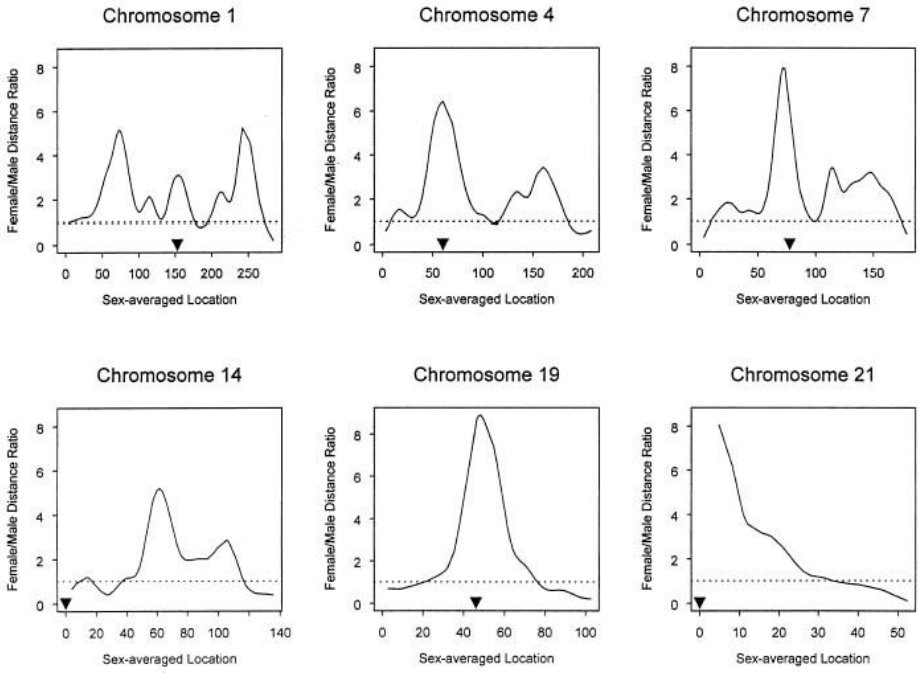
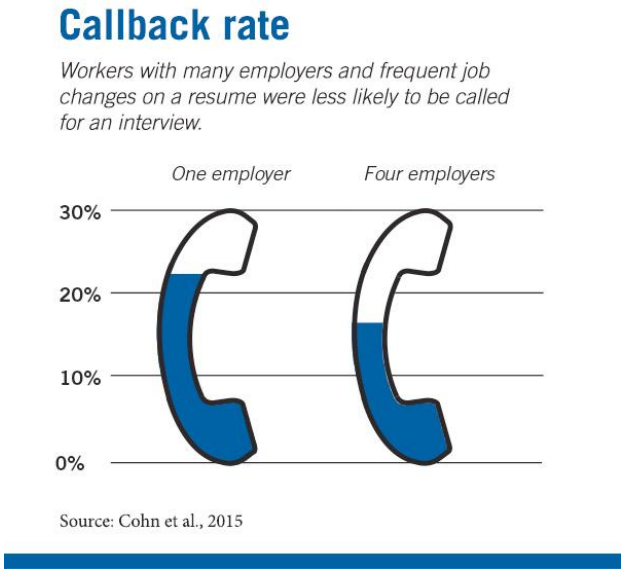
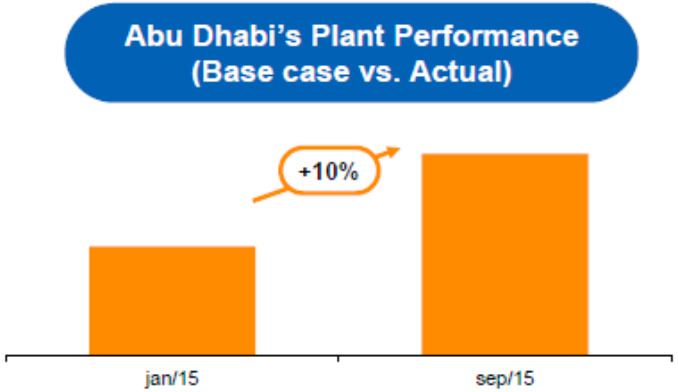


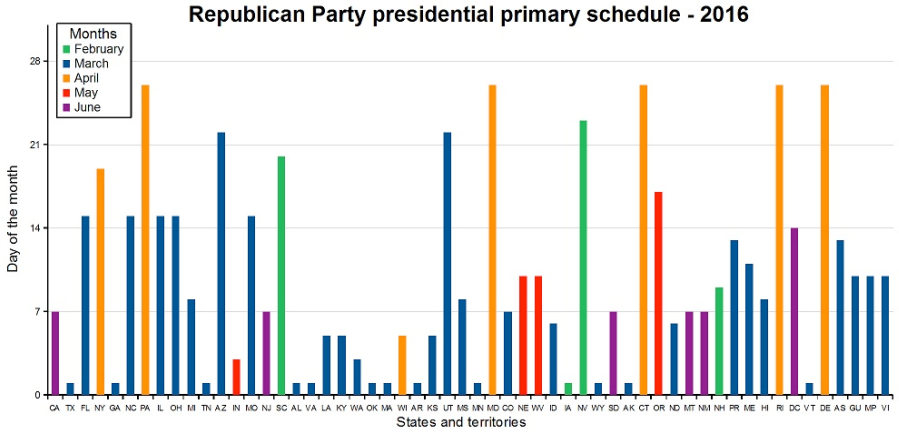
Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

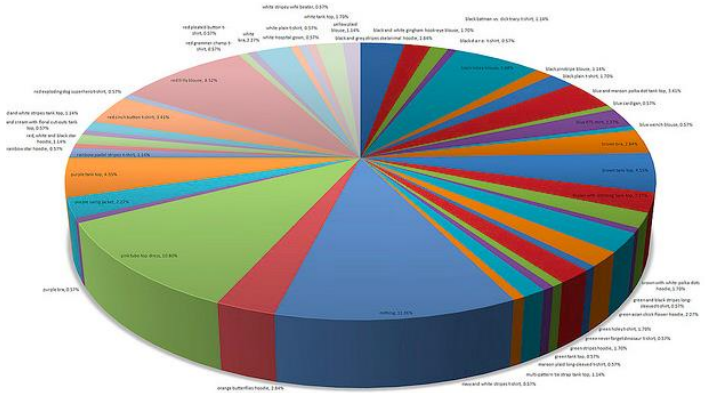
Basics: Communicate the data

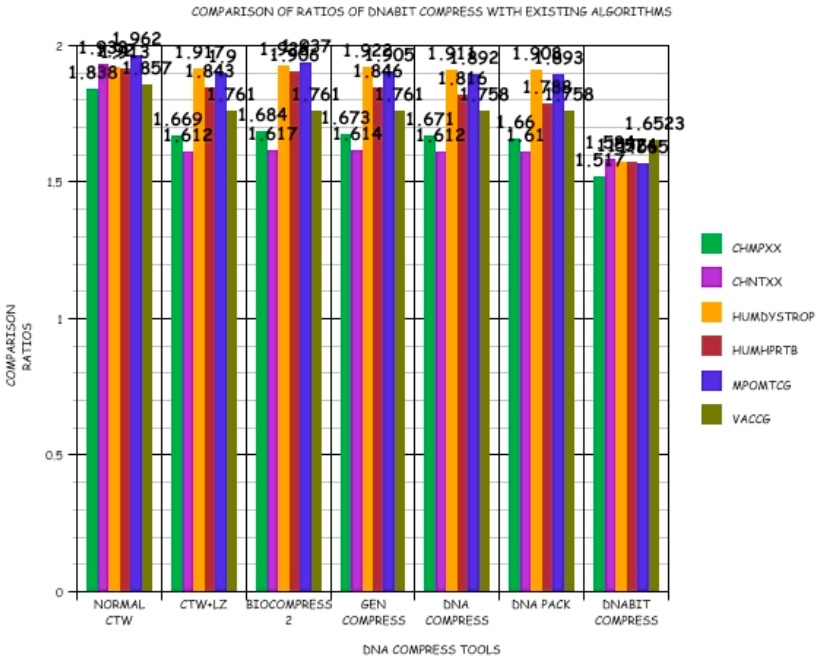




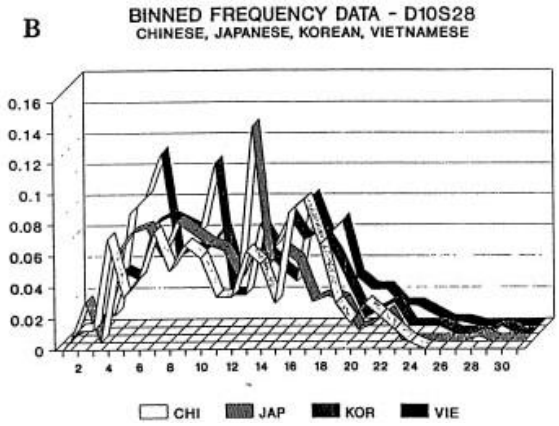
Basics: Only communicate what is relevant

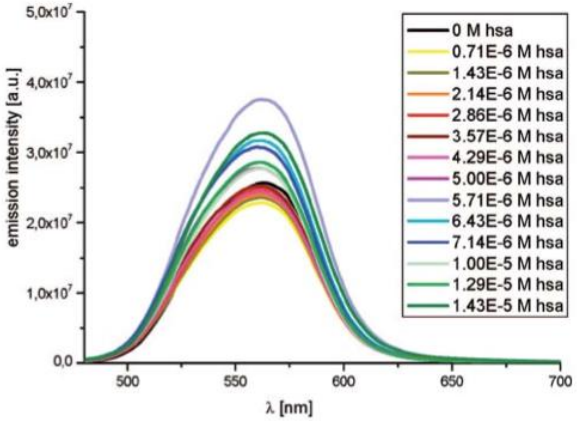




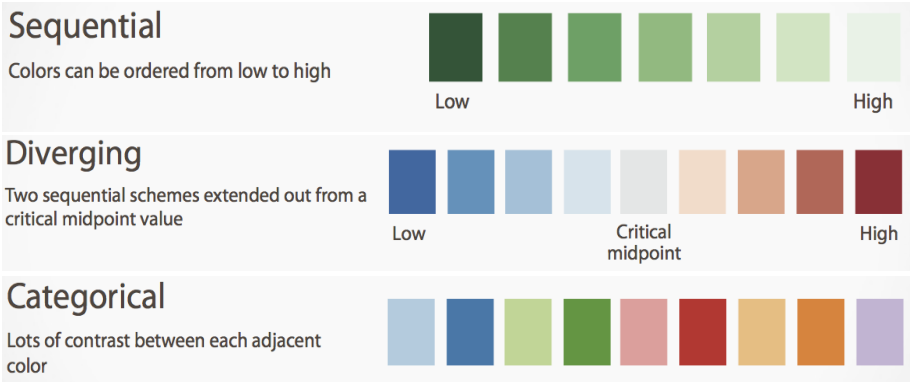


Basics: Know how to use colours





Choose colors based on the information you want to convey.

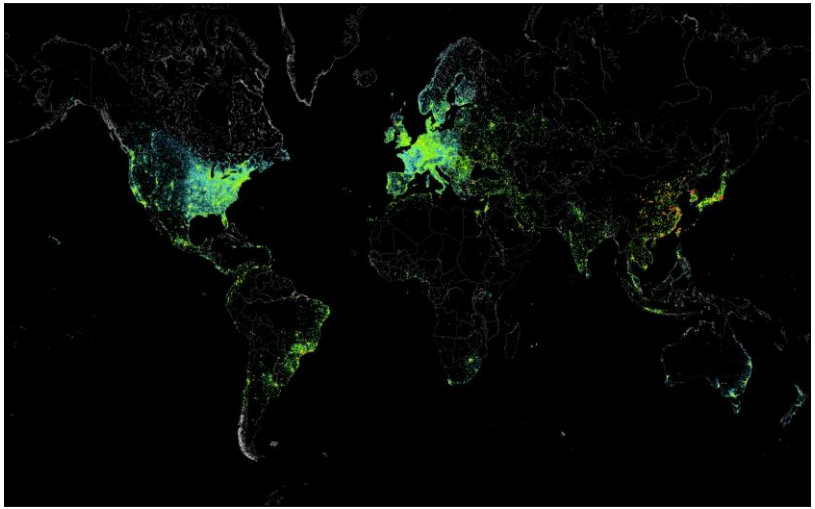


Use online resources.

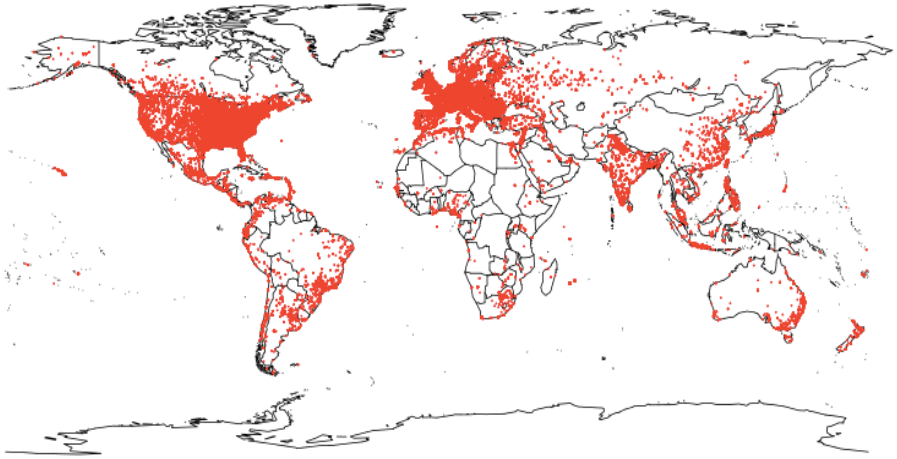
–http://colorbrewer2.org/

–https://color.adobe.com/

Geographical plotting: Spot the difference

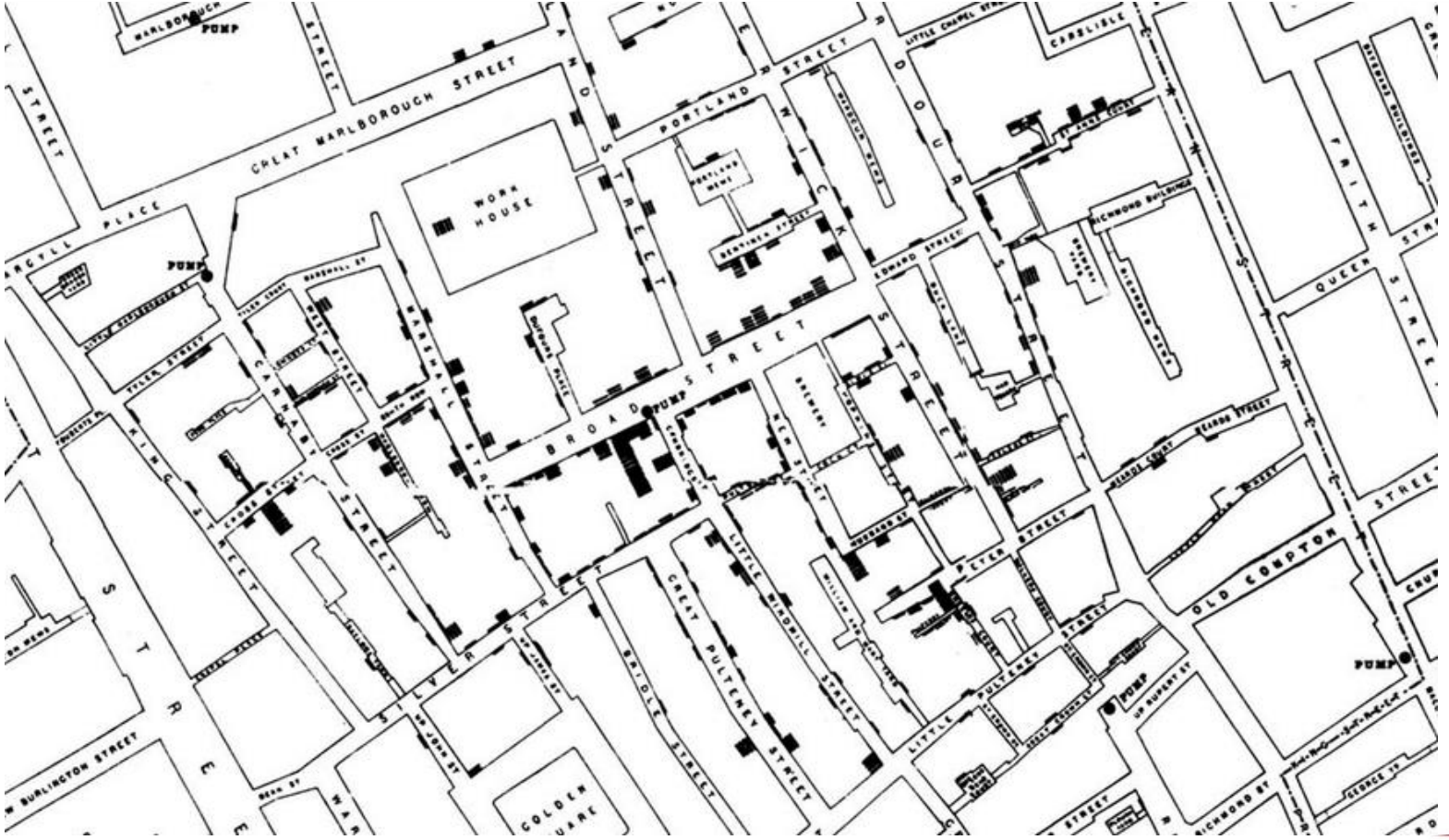


Internet users



Users of a popular website

Geographical plotting: a good example

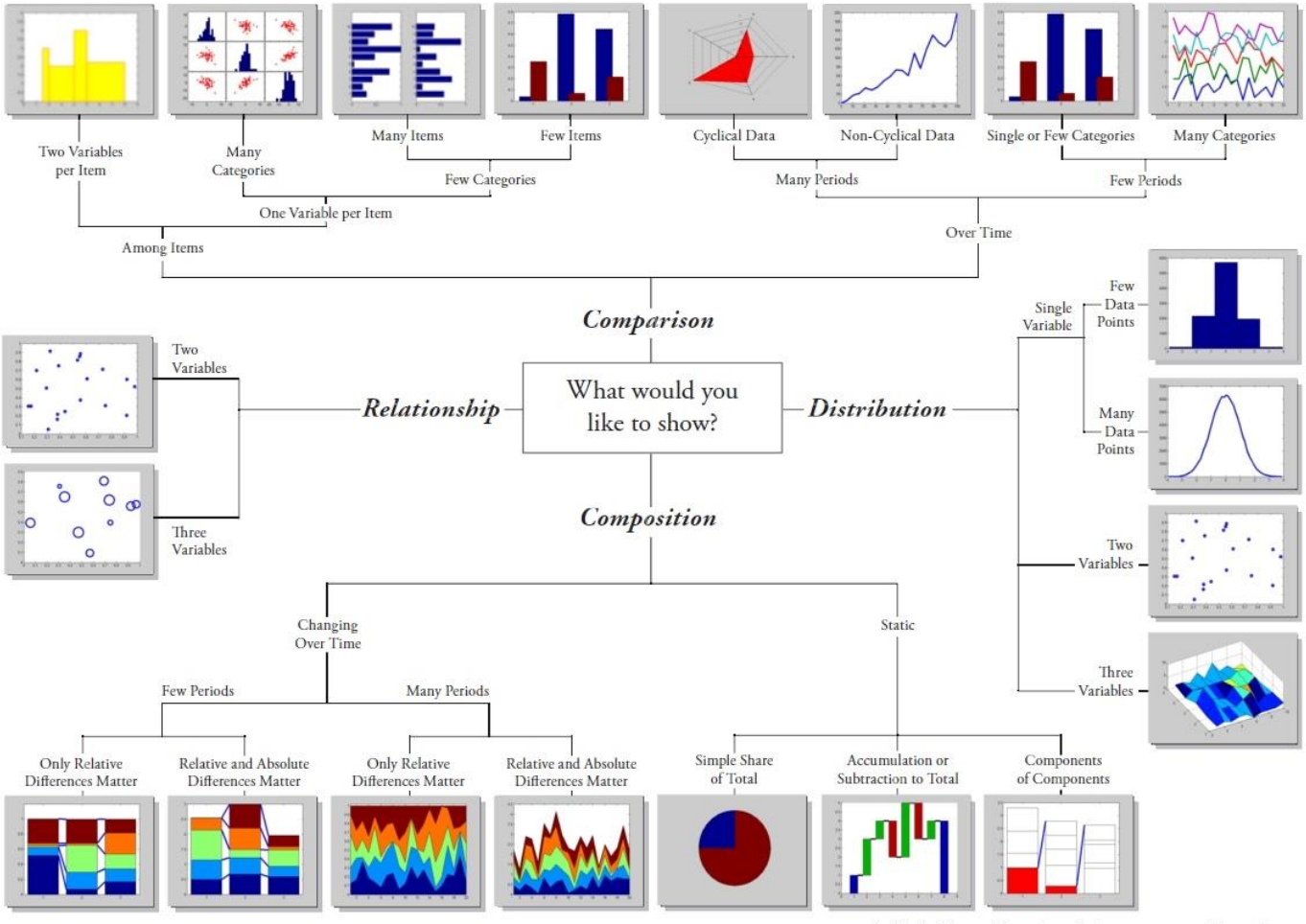


So, which plot to use?

Lancaster University

- Think about what message you want to convey.
- Why?
  - Explore
  - Convince
  - Communicate
- Experiment with different plots.
- Try (with caution) others' tips.
  - e.g. Andrew Abela's Chart Chooser

Chart Suggestions—A Thought-Starter



Modified with permision -Doug Hull
blogs.mathworks.com/videos
hull@mathworks.com 2009
www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

Further reading

Lancaster University

- E. Tufte, “The Visual Display of Quantitative Information”, 2nd edition, 2001.
- I. Meirelles, “Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations”, 2013.
- K. Börner and D.E. Polley, “Visual Insights: A Practical Guide to Making Sense of Data”, 2014.
- Catalogue of examples:
  - M. Lima, “Visual Complexity: Mapping Patterns of Information”, 2011.
  - A. Cairo, “The Functional Art: An introduction to information graphics and visualization (Voices That Matter)”, 2012.
- http://www.visualisingdata.com/
- J.W. Tukey, “Exploratory Data Analysis”, 1977.