# Data Science Survival Skills

Data visualization and statistics

# Importance of context

First part

**Exploratory**

Showing all your data

**Explanatory**

Second part

Showing only the relevant data

**WHO?**

**WHAT?**

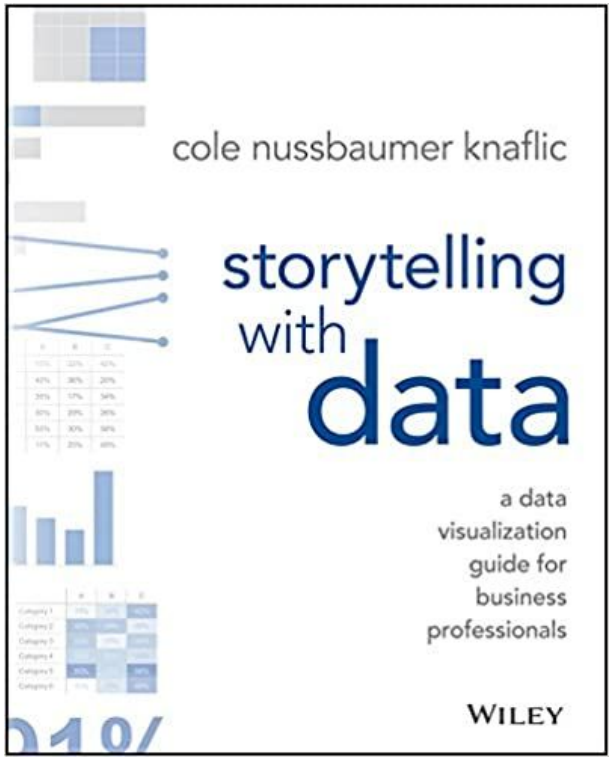
**HOW?**

# What do papers/theses do?

## Telling a story.

⇒ People love stories, and storytelling is a key skill to acquire!

# What people like or need


https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833
**Ten Simple Rules for Better Figures**

# Data needs condensation

Table I

| *Iris setosa* | | | | *Iris versicolor* | | | | *Iris virginica* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sepal length | Sepal width | Petal length | Petal width | Sepal length | Sepal width | Petal length | Petal width | Sepal length | Sepal width | Petal length | Petal width |
| 5·1 | 3·5 | 1·4 | 0·2 | 7·0 | 3·2 | 4·7 | 1·4 | 6·3 | 3·3 | 6·0 | 2·5 |
| 4·9 | 3·0 | 1·4 | 0·2 | 6·4 | 3·2 | 4·5 | 1·5 | 5·8 | 2·7 | 5·1 | 1·9 |
| 4·7 | 3·2 | 1·3 | 0·2 | 6·9 | 3·1 | 4·9 | 1·5 | 7·1 | 3·0 | 5·9 | 2·1 |
| 4·6 | 3·1 | 1·5 | 0·2 | 5·5 | 2·3 | 4·0 | 1·3 | 6·3 | 2·9 | 5·6 | 1·8 |
| 5·0 | 3·6 | 1·4 | 0·2 | 6·5 | 2·8 | 4·6 | 1·5 | 6·5 | 3·0 | 5·8 | 2·2 |
| 5·4 | 3·9 | 1·7 | 0·4 | 5·7 | 2·8 | 4·5 | 1·3 | 7·6 | 3·0 | 6·6 | 2·1 |
| 4·6 | 3·4 | 1·4 | 0·3 | 6·3 | 3·3 | 4·7 | 1·6 | 4·9 | 2·5 | 4·5 | 1·7 |
| 5·0 | 3·4 | 1·5 | 0·2 | 4·9 | 2·4 | 3·3 | 1·0 | 7·3 | 2·9 | 6·3 | 1·8 |
| 4·4 | 2·9 | 1·4 | 0·2 | 6·6 | 2·9 | 4·6 | 1·3 | 6·7 | 2·5 | 5·8 | 1·8 |
| 4·9 | 3·1 | 1·5 | 0·1 | 5·2 | 2·7 | 3·9 | 1·4 | 7·2 | 3·6 | 6·1 | 2·5 |
| 5·4 | 3·7 | 1·5 | 0·2 | 5·0 | 2·0 | 3·5 | 1·0 | 6·5 | 3·2 | 5·1 | 2·0 |
| 4·8 | 3·4 | 1·6 | 0·2 | 5·9 | 3·0 | 4·2 | 1·5 | 6·4 | 2·7 | 5·3 | 1·9 |
| 4·8 | 3·0 | 1·4 | 0·1 | 6·0 | 2·2 | 4·0 | 1·0 | 6·8 | 3·0 | 5·5 | 2·1 |
| 4·3 | 3·0 | 1·1 | 0·1 | 6·1 | 2·9 | 4·7 | 1·4 | 5·7 | 2·5 | 5·0 | 2·0 |
| 5·8 | 4·0 | 1·2 | 0·2 | 5·6 | 2·9 | 3·6 | 1·3 | 5·8 | 2·8 | 5·1 | 2·4 |
| 5·7 | 4·4 | 1·5 | 0·4 | 6·7 | 3·1 | 4·4 | 1·4 | 6·4 | 3·2 | 5·3 | 2·3 |
| 5·4 | 3·9 | 1·3 | 0·4 | 5·6 | 3·0 | 4·5 | 1·5 | 6·5 | 3·0 | 5·5 | 1·8 |
| 5·1 | 3·5 | 1·4 | 0·3 | 5·8 | 2·7 | 4·1 | 1·0 | 7·7 | 3·8 | 6·7 | 2·2 |
| 5·7 | 3·8 | 1·7 | 0·3 | 6·2 | 2·2 | 4·5 | 1·5 | 7·7 | 2·6 | 6·9 | 2·3 |
| 5·1 | 3·8 | 1·5 | 0·3 | 5·6 | 2·5 | 3·9 | 1·1 | 6·0 | 2·2 | 5·0 | 1·5 |
| 5·4 | 3·4 | 1·7 | 0·2 | 5·9 | 3·2 | 4·8 | 1·8 | 6·9 | 3·2 | 5·7 | 2·3 |
| 5·1 | 3·7 | 1·5 | 0·4 | 6·1 | 2·8 | 4·0 | 1·3 | 5·6 | 2·8 | 4·9 | 2·0 |
| 4·6 | 3·6 | 1·0 | 0·2 | 6·3 | 2·5 | 4·9 | 1·5 | 7·7 | 2·8 | 6·7 | 2·0 |
| 5·1 | 3·3 | 1·7 | 0·5 | 6·1 | 2·8 | 4·7 | 1·2 | 6·3 | 2·7 | 4·9 | 1·8 |
| 4·8 | 3·4 | 1·9 | 0·2 | 6·4 | 2·9 | 4·3 | 1·3 | 6·7 | 3·3 | 5·7 | 2·1 |
| 5·0 | 3·0 | 1·6 | 0·2 | 6·6 | 3·0 | 4·4 | 1·4 | 7·2 | 3·2 | 6·0 | 1·8 |
| 5·0 | 3·4 | 1·6 | 0·4 | 6·8 | 2·8 | 4·8 | 1·4 | 6·2 | 2·8 | 4·8 | 1·8 |
| 5·2 | 3·5 | 1·5 | 0·2 | 6·7 | 3·0 | 5·0 | 1·7 | 6·1 | 3·0 | 4·9 | 1·8 |
| 5·2 | 3·4 | 1·4 | 0·2 | 6·0 | 2·9 | 4·5 | 1·5 | 6·4 | 2·8 | 5·6 | 2·1 |
| 4·7 | 3·2 | 1·6 | 0·2 | 5·7 | 2·6 | 3·5 | 1·0 | 7·2 | 3·0 | 5·8 | 1·6 |
| 4·8 | 3·1 | 1·6 | 0·2 | 5·5 | 2·4 | 3·8 | 1·1 | 7·4 | 2·8 | 6·1 | 1·9 |
| 5·4 | 3·4 | 1·5 | 0·4 | 5·5 | 2·4 | 3·7 | 1·0 | 7·9 | 3·8 | 6·4 | 2·0 |
| 5·2 | 4·1 | 1·5 | 0·1 | 5·8 | 2·7 | 3·9 | 1·2 | 6·4 | 2·8 | 5·6 | 2·2 |
| 5·5 | 4·2 | 1·4 | 0·2 | 6·0 | 2·7 | 5·1 | 1·6 | 6·3 | 2·8 | 5·1 | 1·5 |
| 4·9 | 3·1 | 1·5 | 0·2 | 5·4 | 3·0 | 4·5 | 1·5 | 6·1 | 2·6 | 5·6 | 1·4 |
| 5·0 | 3·2 | 1·2 | 0·2 | 6·0 | 3·4 | 4·5 | 1·6 | 7·7 | 3·0 | 6·1 | 2·3 |
| 5·5 | 3·5 | 1·3 | 0·2 | 6·7 | 3·1 | 4·7 | 1·5 | 6·3 | 3·4 | 5·6 | 2·4 |
| 4·9 | 3·6 | 1·4 | 0·1 | 6·3 | 2·3 | 4·4 | 1·3 | 6·4 | 3·1 | 5·5 | 1·8 |
| 4·4 | 3·0 | 1·3 | 0·2 | 5·6 | 3·0 | 4·1 | 1·3 | 6·0 | 3·0 | 4·8 | 1·8 |
| 5·1 | 3·4 | 1·5 | 0·2 | 5·5 | 2·5 | 4·0 | 1·3 | 6·9 | 3·1 | 5·4 | 2·1 |
| 5·0 | 3·5 | 1·3 | 0·3 | 5·5 | 2·6 | 4·4 | 1·2 | 6·7 | 3·1 | 5·6 | 2·4 |
| 4·5 | 2·3 | 1·3 | 0·3 | 6·1 | 3·0 | 4·6 | 1·4 | 6·9 | 3·1 | 5·1 | 2·3 |
| 4·4 | 3·2 | 1·3 | 0·2 | 5·8 | 2·6 | 4·0 | 1·2 | 5·8 | 2·7 | 5·1 | 1·9 |
| 5·0 | 3·5 | 1·6 | 0·6 | 5·0 | 2·3 | 3·3 | 1·0 | 6·8 | 3·2 | 5·9 | 2·3 |
| 5·1 | 3·8 | 1·9 | 0·4 | 5·6 | 2·7 | 4·2 | 1·3 | 6·7 | 3·3 | 5·7 | 2·5 |
| 4·8 | 3·0 | 1·4 | 0·3 | 5·7 | 3·0 | 4·2 | 1·2 | 6·7 | 3·0 | 5·2 | 2·3 |
| 5·1 | 3·8 | 1·6 | 0·2 | 5·7 | 2·9 | 4·2 | 1·3 | 6·3 | 2·5 | 5·0 | 1·9 |
| 4·6 | 3·2 | 1·4 | 0·2 | 6·2 | 2·9 | 4·3 | 1·3 | 6·5 | 3·0 | 5·2 | 2·0 |
| 5·3 | 3·7 | 1·5 | 0·2 | 5·1 | 2·5 | 3·0 | 1·1 | 6·2 | 3·4 | 5·4 | 2·3 |
| 5·0 | 3·3 | 1·4 | 0·2 | 5·7 | 2·8 | 4·1 | 1·3 | 5·9 | 3·0 | 5·1 | 1·8 |

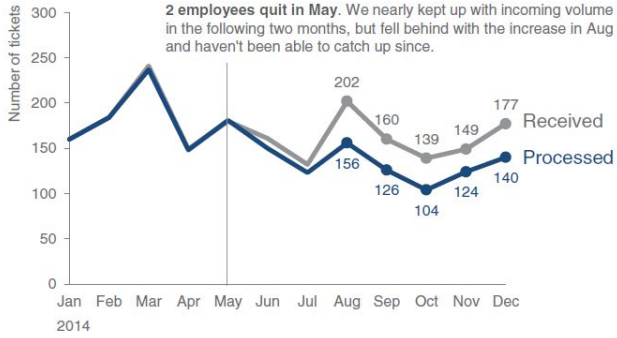
R.A. Fisher, 1936: The Iris Dataset


FIGURE 0.2 Example 1 (before): storytelling with data

## Please approve the hire of 2 FTEs

to backfill those who quit in the past year

Ticket volume over time



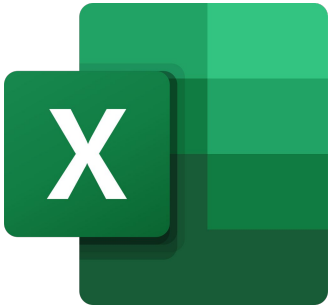

Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

FIGURE 0.3 Example 1 (after): storytelling with data

# The golden rules for effective communication

1. Understand the **context**
2. Choose an **appropriate** visual display
3. **Eliminate clutter**
4. Focus **attention** where you want it
5. Think like a **designer**
6. **Tell a story**

# Nothing is tool specific

matplotlib

bokeh

seaborn

plotly

IBM SPSS

OriginPro 2022
The Ultimate Software for Graphing & Analysis

GraphPad Prism

Prism

# Who, what and how.

**Who** is your audience? And **who** are you in respect to the audience?

**What** do you need your audience to know or do?
**What** action is required?

**Prompting action**

Here are some action words to help act as thought starters as you determine what you are asking of your audience:

accept | agree | begin | believe | change | collaborate | commence | create | defend | desire | differentiate | do | empathize | empower | encourage | engage | establish | examine | facilitate | familiarize | form | implement | include | influence | invest | invigorate | know | learn | like | persuade | plan | promote | pursue | recommend | receive | remember | report | respond | secure | support | simplify | start | try | understand | validate

# The how.

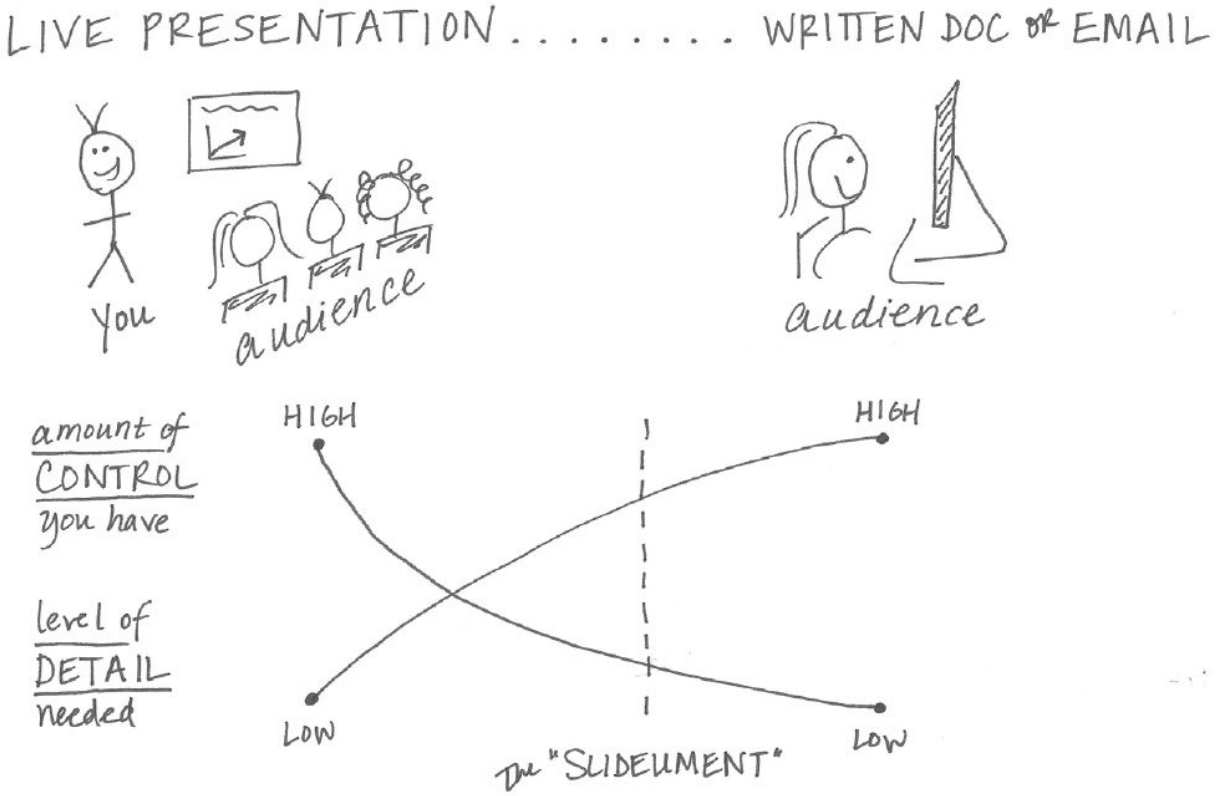



FIGURE 1.1 Communication mechanism continuum

# 3-Minute story, big idea, elevator pitch

“If you know exactly what it is you want to communicate, you can make it fit the time slot you’re given, even if it isn’t the one for which you are prepared.”

The BIG IDEA components (by Nancy Duarte):

- It must articulate your unique Point of View
- It must convey what’s at stake
- It must be a complete sentence

Example: “The pilot summer learning program was successful at improving students’ perceptions of science and, because of this success, we recommend continuing to offer it going forward; please approve our budget for this program.”

# Storyboarding

- Try to avoid PPTX, try to use low tech first (post its, whiteboard, paper, ...)

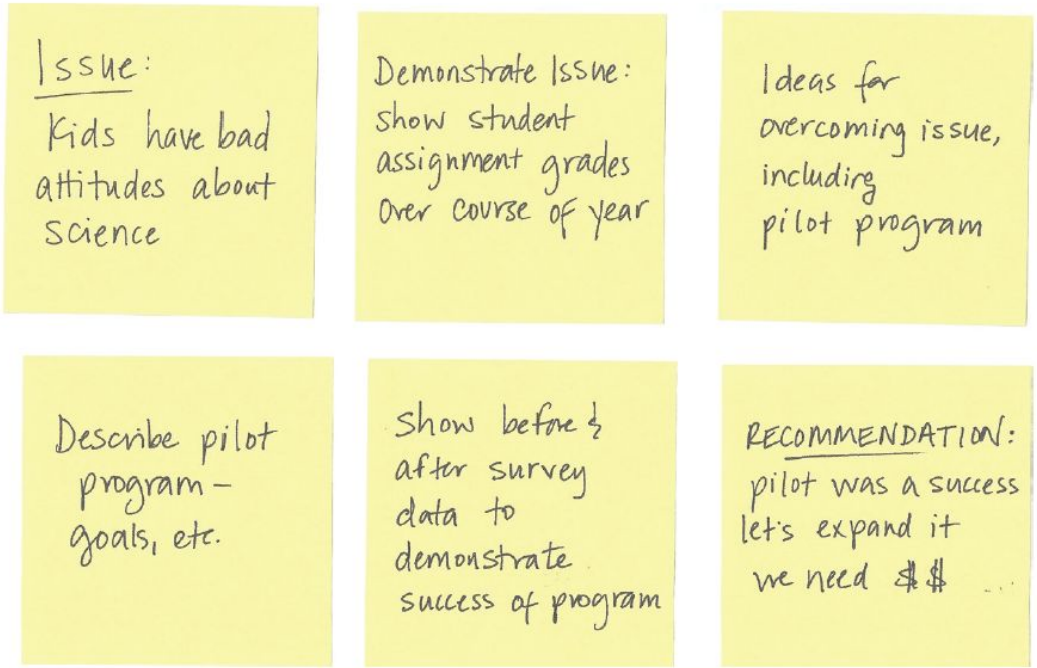
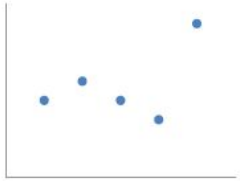



FIGURE 1.2 Example storyboard

# Choosing an effective visual


Simple text



Scatterplot

| | A | B | C |
|---|---|---|---|
| Category 1 | 15% | 22% | 42% |
| Category 2 | 40% | 36% | 20% |
| Category 3 | 35% | 17% | 34% |
| Category 4 | 30% | 29% | 26% |
| Category 5 | 55% | 30% | 58% |
| Category 6 | 11% | 25% | 49% |

Table

Line

| | A | B | C |
|---|---|---|---|
| Category 1 | 15% | 22% | 42% |
| Category 2 | 40% | 36% | 20% |
| Category 3 | 35% | 17% | 34% |
| Category 4 | 30% | 29% | 26% |
| Category 5 | 55% | 30% | 58% |
| Category 6 | 11% | 25% | 49% |

Heatmap

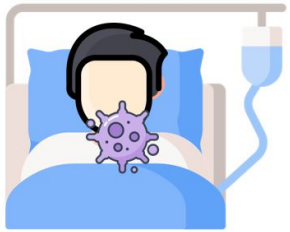
Slopegraph

FIGURE 2.1 The visuals I use most

FAU

Friedrich-Alexander-Universität
Technische Fakultät

## Head&Neck cancer is a common and serious threat



Nick

Head&Neck cancer

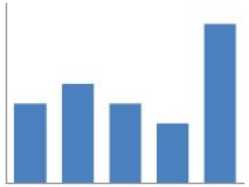
**#8** for males, #13 for females

More contrast For gray

**43%** 5 yr survival rate for males, 55% females

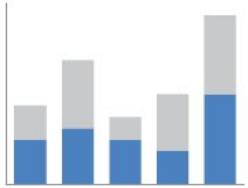
**29%** 10 yr survival rate for males, 40% females

# Choosing an effective visual

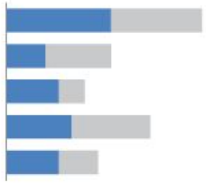


Vertical bar

Horizontal bar

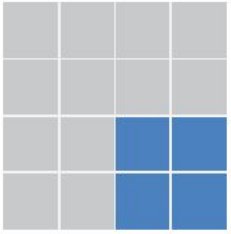


Stacked vertical bar



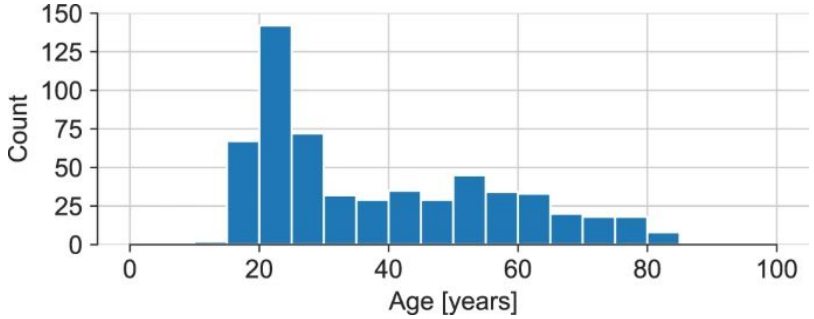
Stacked horizontal bar

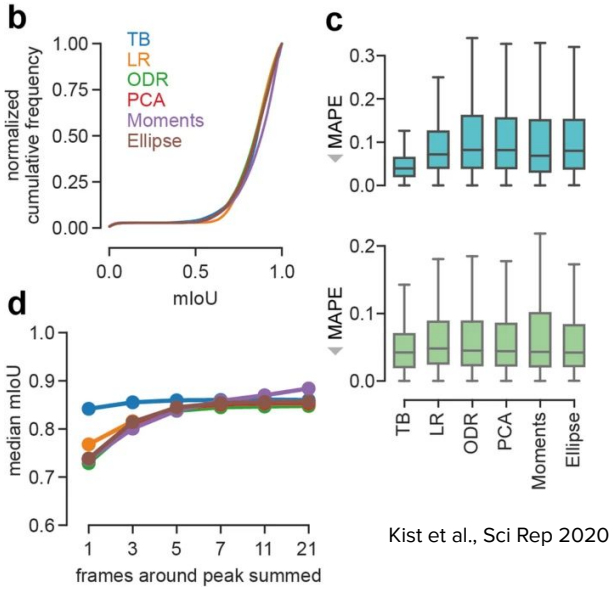
Waterfall



Square area

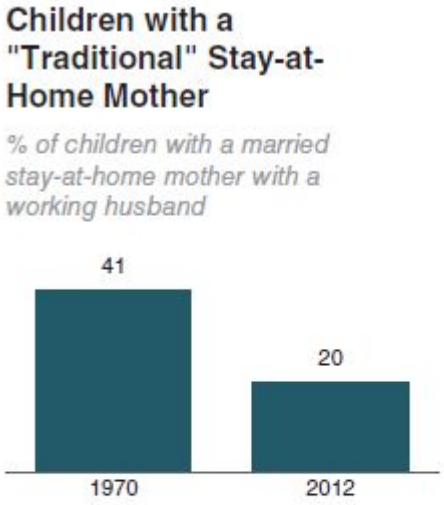
Gómez*, Kist* et al., Sci Data 2020




Kist et al., Sci Rep 2020

# Plain text (esp. for talks)




Note: Based on children younger than 18. Their mothers are categorized based on employment status in 1970 and 2012.

Source: Pew Research Center analysis of March Current Population Surveys Integrated Public Use Microdata Series (IPUMS-CPS), 1971 and 2013

Adapted from PEW RESEARCH CENTER

**FIGURE 2.2** Stay-at-home moms original graph

**20%**

of children had a
**traditional stay-at-home mom**
in 2012, compared to 41% in 1970

**FIGURE 2.3** Stay-at-home moms simple text makeover

# Tables

Heavy borders

| Group | Metric A | Metric B | Metric C |
|---|---|---|---|
| Group 1 | $X.X | Y% | Z,ZZZ |
| Group 2 | $X.X | Y% | Z,ZZZ |
| Group 3 | $X.X | Y% | Z,ZZZ |
| Group 4 | $X.X | Y% | Z,ZZZ |
| Group 5 | $X.X | Y% | Z,ZZZ |

Light borders

| Group | Metric A | Metric B | Metric C |
|---|---|---|---|
| Group 1 | $X.X | Y% | Z,ZZZ |
| Group 2 | $X.X | Y% | Z,ZZZ |
| Group 3 | $X.X | Y% | Z,ZZZ |
| Group 4 | $X.X | Y% | Z,ZZZ |
| Group 5 | $X.X | Y% | Z,ZZZ |

Minimal borders

| Group | Metric A | Metric B | Metric C |
|---|---|---|---|
| Group 1 | $X.X | Y% | Z,ZZZ |
| Group 2 | $X.X | Y% | Z,ZZZ |
| Group 3 | $X.X | Y% | Z,ZZZ |
| Group 4 | $X.X | Y% | Z,ZZZ |
| Group 5 | $X.X | Y% | Z,ZZZ |

FIGURE 2.4 Table borders

➜ Use Gestalt principles

Table

| | A | B | C |
|---|---|---|---|
| Category 1 | 15% | 22% | 42% |
| Category 2 | 40% | 36% | 20% |
| Category 3 | 35% | 17% | 34% |
| Category 4 | 30% | 29% | 26% |
| Category 5 | 55% | 30% | 58% |
| Category 6 | 11% | 25% | 49% |

Heatmap

LOW-HIGH

| | A | B | C |
|---|---|---|---|
| Category 1 | 15% | 22% | 42% |
| Category 2 | 40% | 36% | 20% |
| Category 3 | 35% | 17% | 34% |
| Category 4 | 30% | 29% | 26% |
| Category 5 | 55% | 30% | 58% |
| Category 6 | 11% | 25% | 49% |

FIGURE 2.5 Two views of the same data

# Scatter plots

plt.scatter

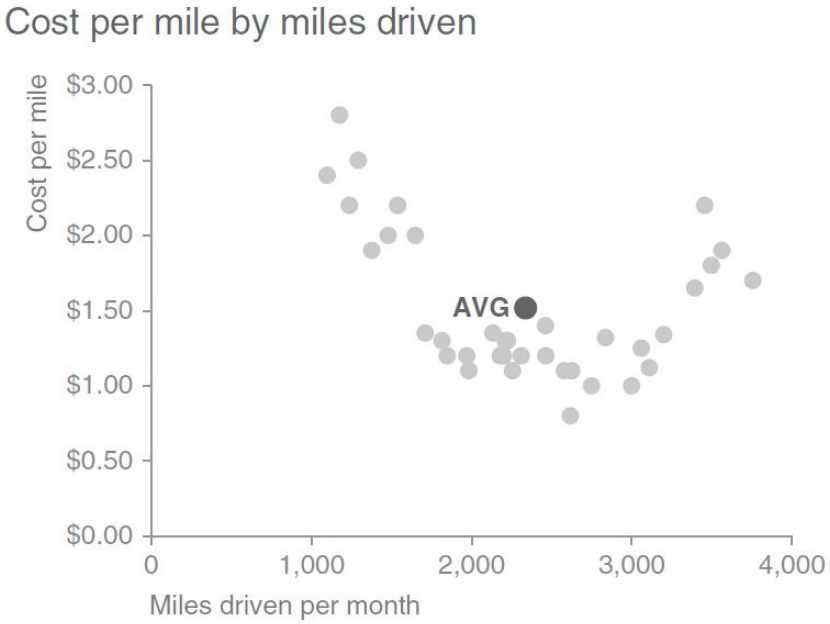



FIGURE 2.6 Scatterplot

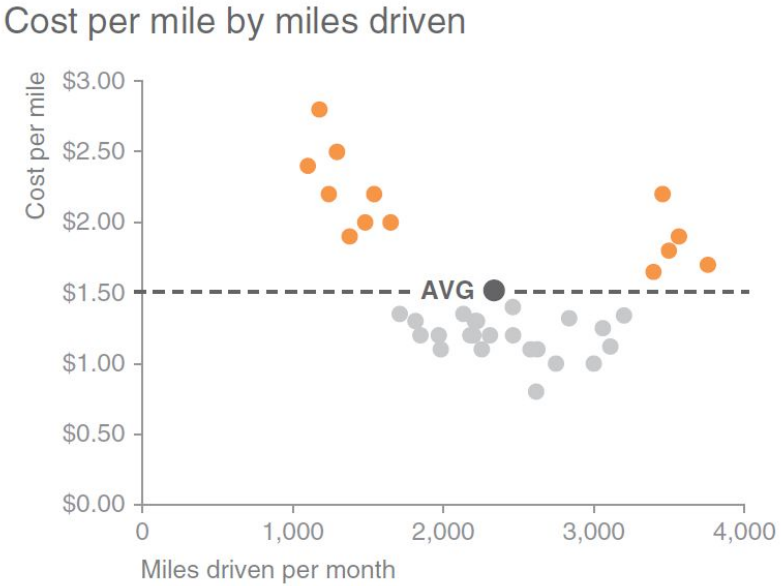



FIGURE 2.7 Modified scatterplot

# Line graphs

plt.plot

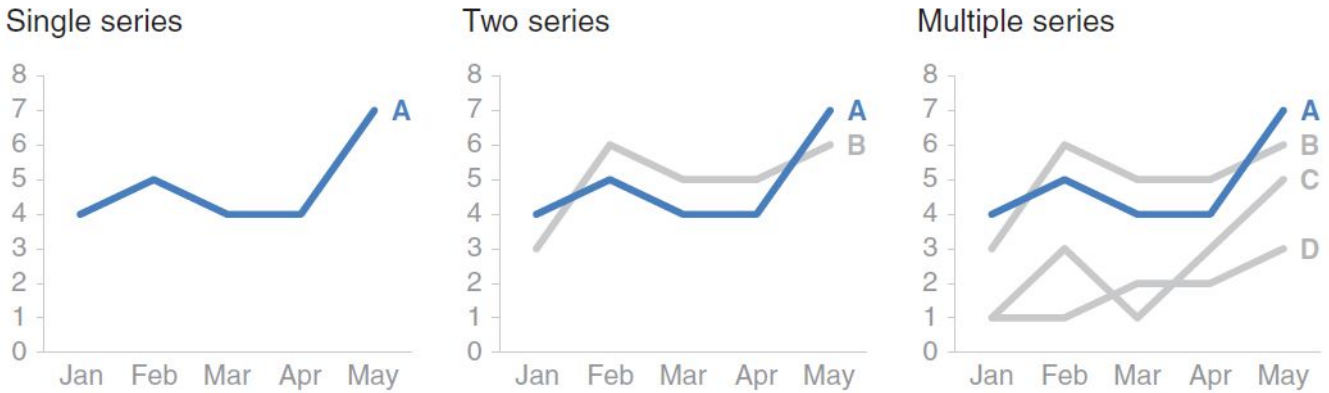



FIGURE 2.8 Line graphs


FIGURE 2.9 Showing average within a range in a line graph

plt.fill_between

# Slope graph

plt.plot


FIGURE 2.11 Modified slopegraph

- Good to show connected trends
- Well suited for homogeneity in data (all go up, all go down, or mix of everything)

# Bars and how to NOT use them

plt.bar(h)

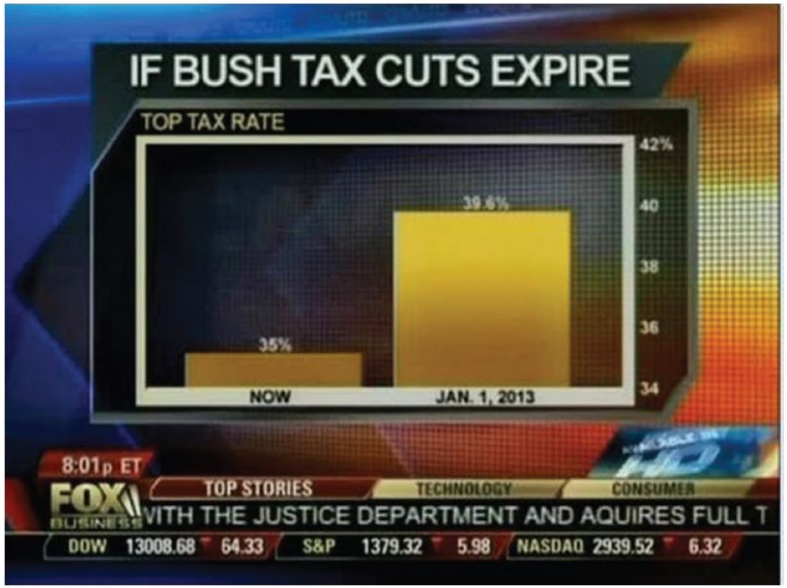
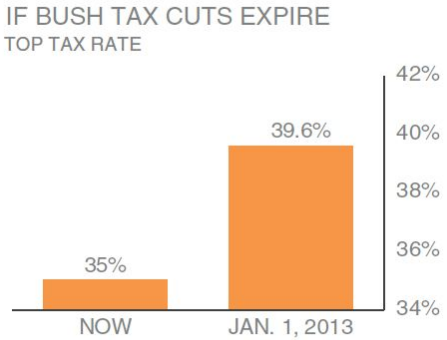



FIGURE 2.12 Fox News bar chart

**Non-zero baseline**: as originally graphed




**Zero baseline**: as it should be graphed

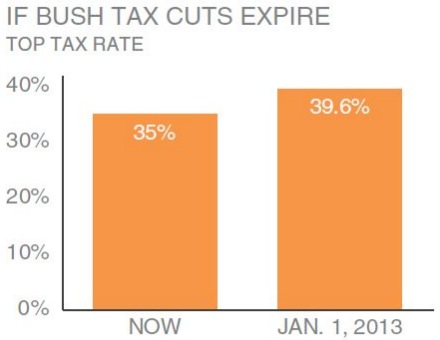



FIGURE 2.13 Bar charts must have a zero baseline


FIGURE 2.14 Bar width

# Bar charts

plt.bar(h)

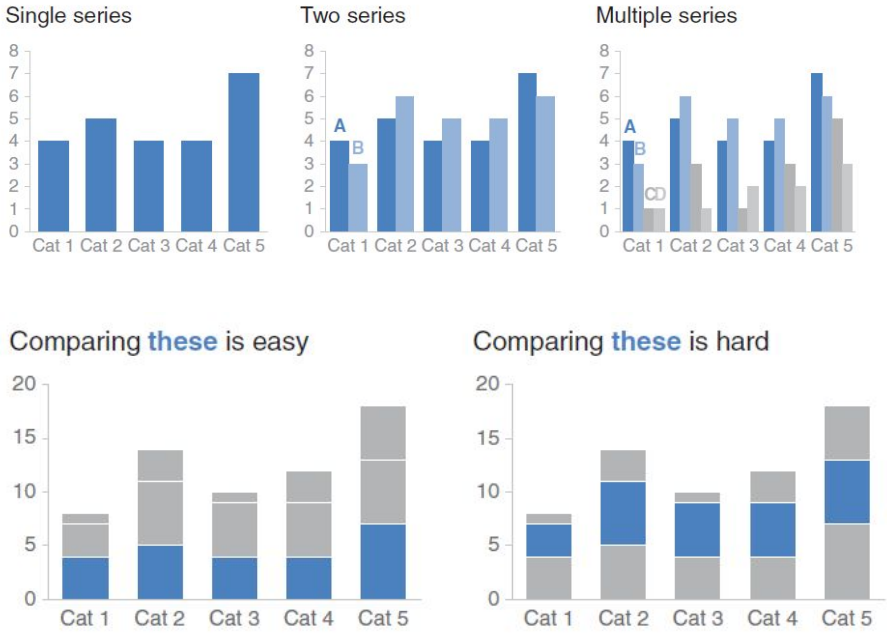



FIGURE 2.16 Comparing series with stacked bar charts

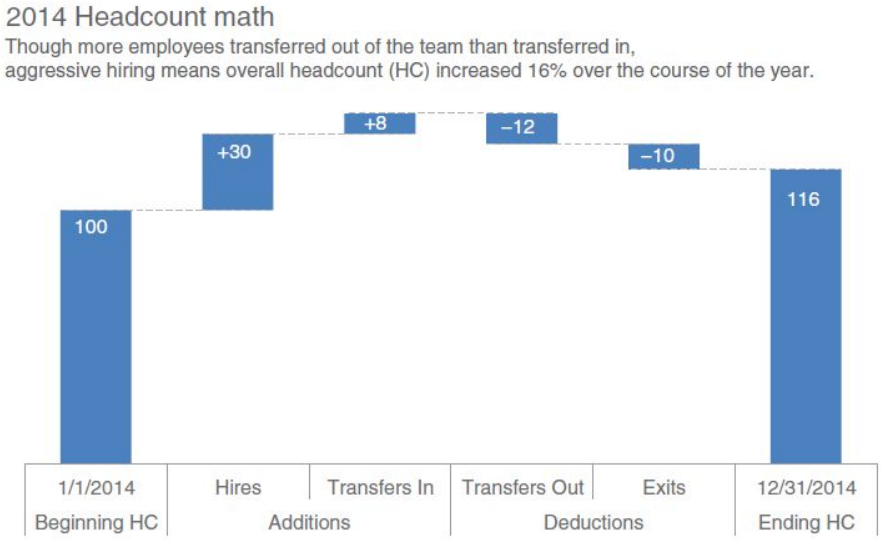



FIGURE 2.17 Waterfall chart

# Horizontal bar charts

Especially with long category names!

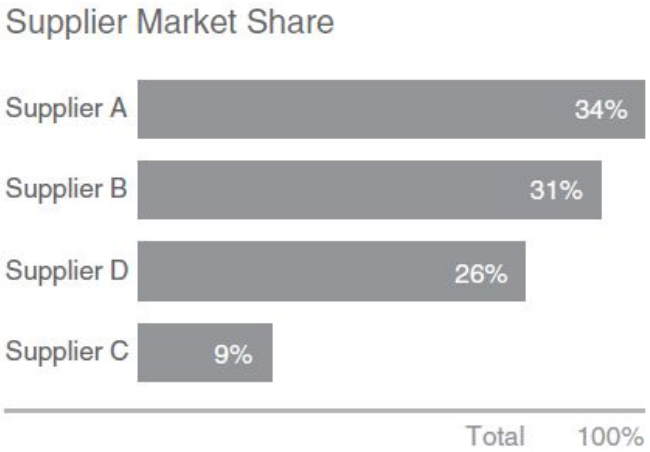



FIGURE 2.23 An alternative to the pie chart

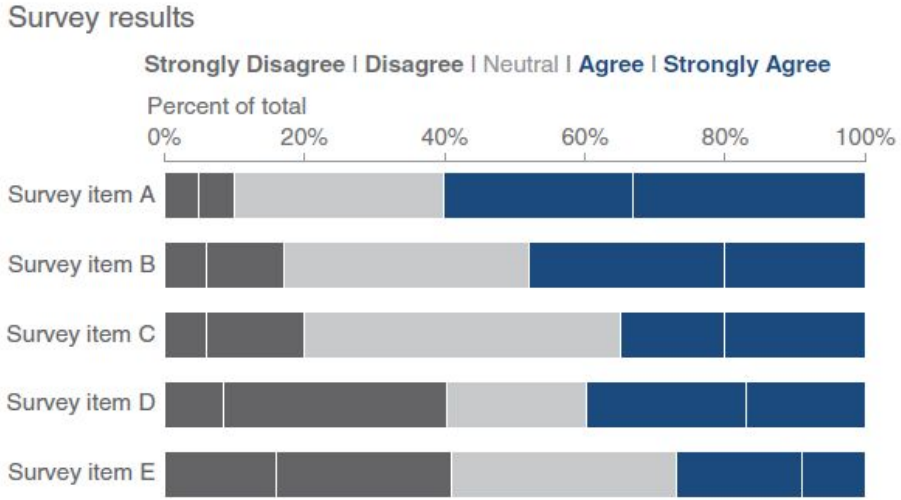



FIGURE 2.19 100% stacked horizontal bar chart

# What to avoid in general?

- Pie charts
- 3D effects
- Random color
- Secondary y-axis

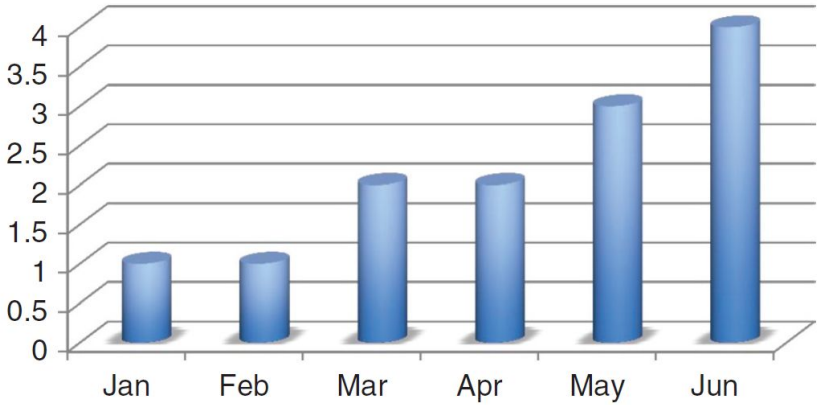



FIGURE 2.25 3D column chart

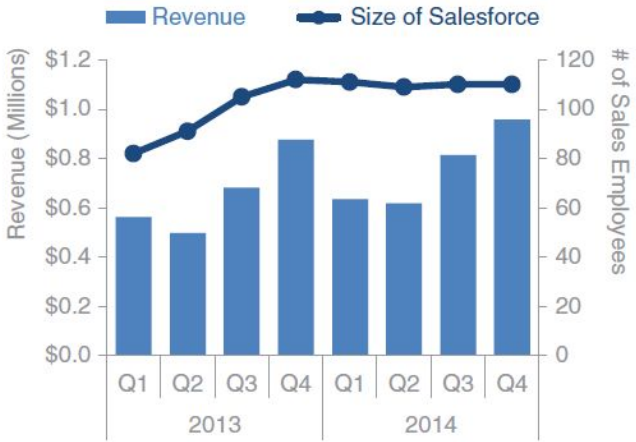



FIGURE 2.26 Secondary y-axis

Alternative 1: **label directly**


Alternative 2: **pull apart vertically**


FIGURE 2.27 Strategies for avoiding a secondary *y*-axis

# Remove clutter

**proximity, similarity, enclosure, closure, continuity, and connection**

Reprise: **Gestalt principles**

FIGURE 3.2 You see columns and rows, simply due to dot spacing

FIGURE 3.7 Gestalt principle of closure

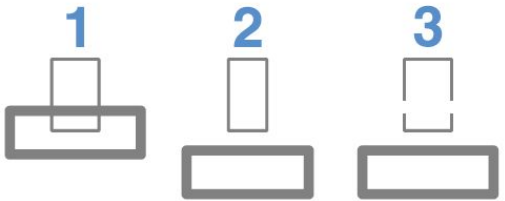



FIGURE 3.9 Gestalt principle of continuity

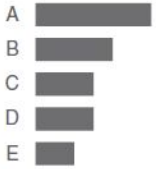



FIGURE 3.10 Graph with *y*-axis line removed

FIGURE 3.4 You see rows due to similarity of color

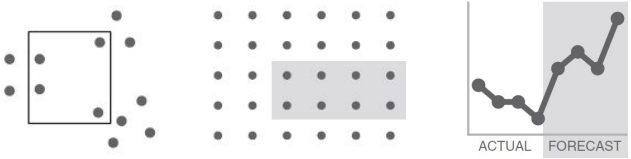



FIGURE 3.5 Gestalt principle of enclosure

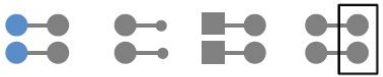


FIGURE 3.11 Gestalt principle of connection

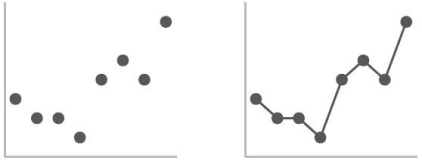


FIGURE 3.12 Lines connect the dots

# Lack of visual order

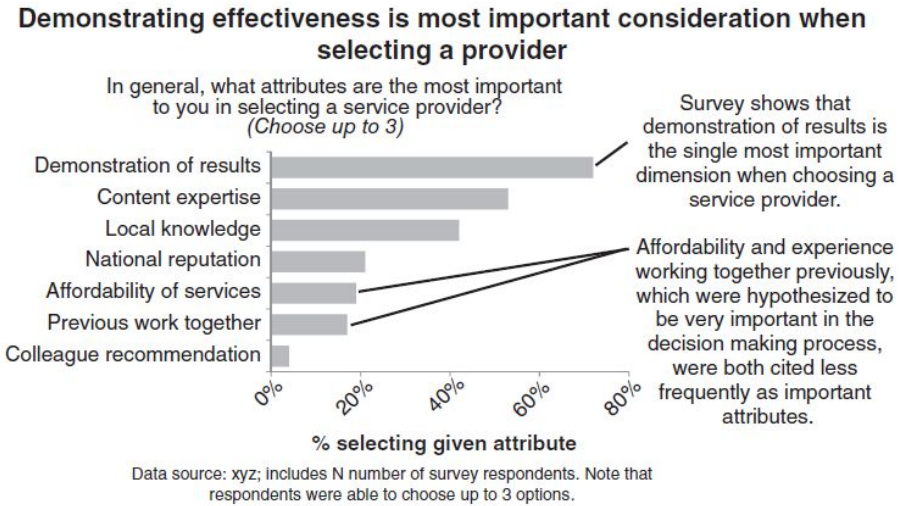



FIGURE 3.13 Summary of survey feedback

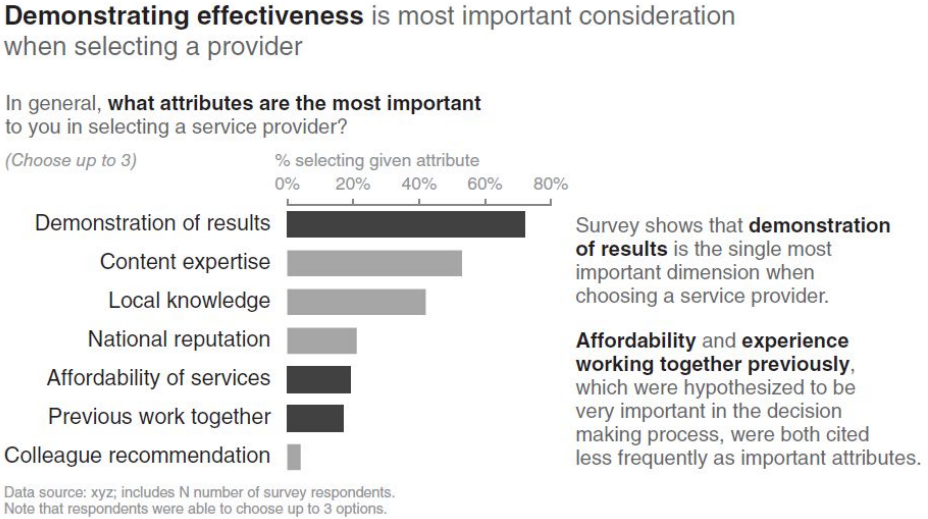
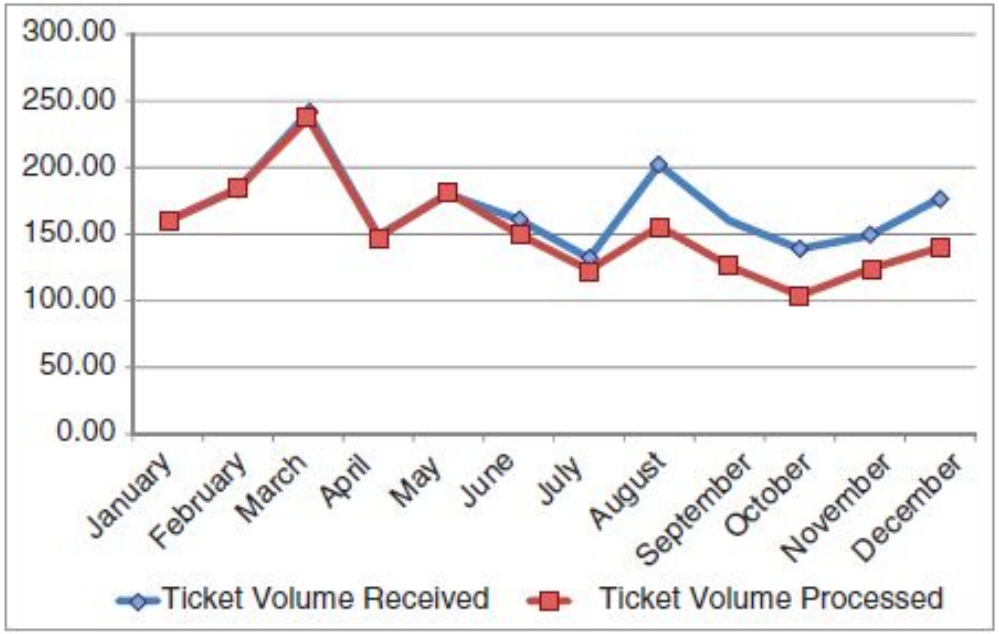


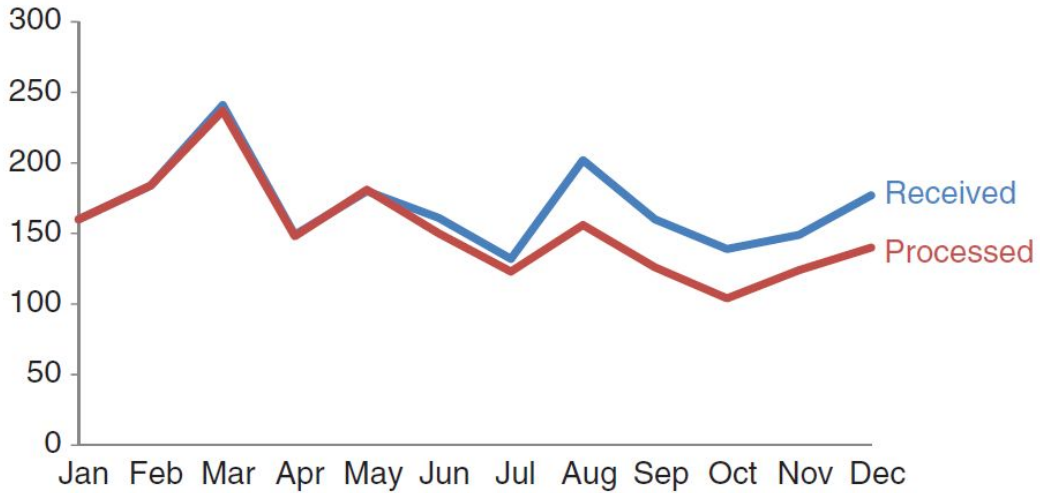

FIGURE 3.14 Revamped summary of survey feedback

# How to remove clutter

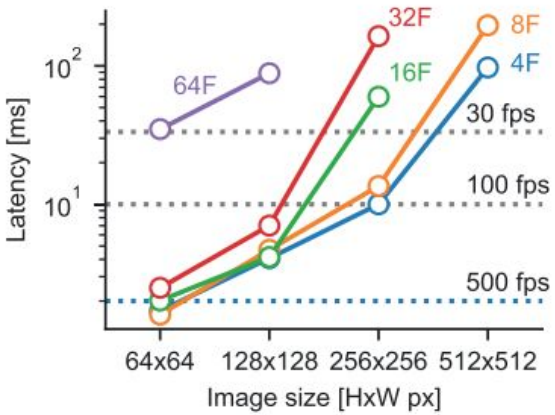


# What did we do?

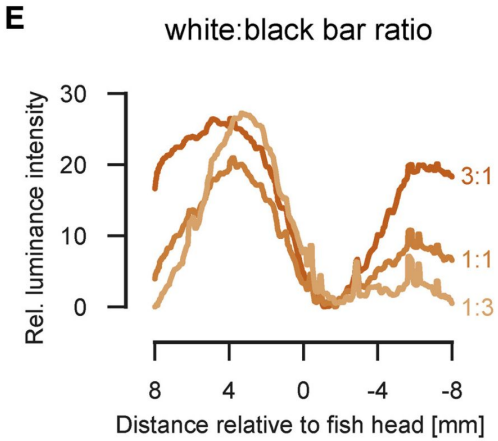
- Remove chart border (Gestalt: closure)
- Remove grid lines (if they don't help)
- Remove data markers
  (personal opinion: add them on purpose!)
- Clean up x-axis
- Label data directly
- Leverage consistent color



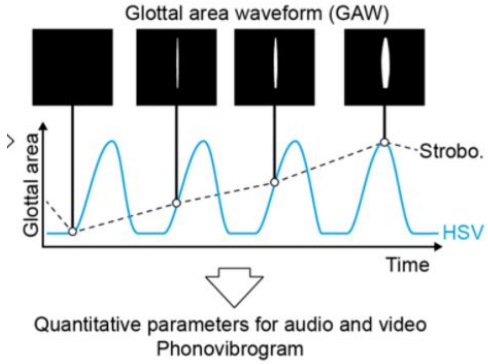
# Real life examples




Kist and Döllinger, IEEE Access 2020




Kist and Portugues, Cell Reports 2019




Kist et al., Sci Reports 2021

# Drawing attention


FIGURE 4.2 Count the 3s example

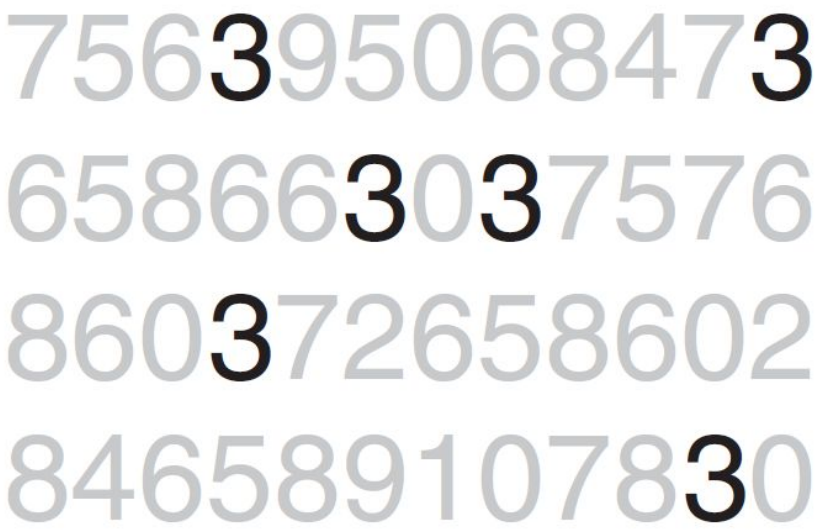



FIGURE 4.3 Count the 3s example with preattentive attribute

# Attention/Saliency tricks

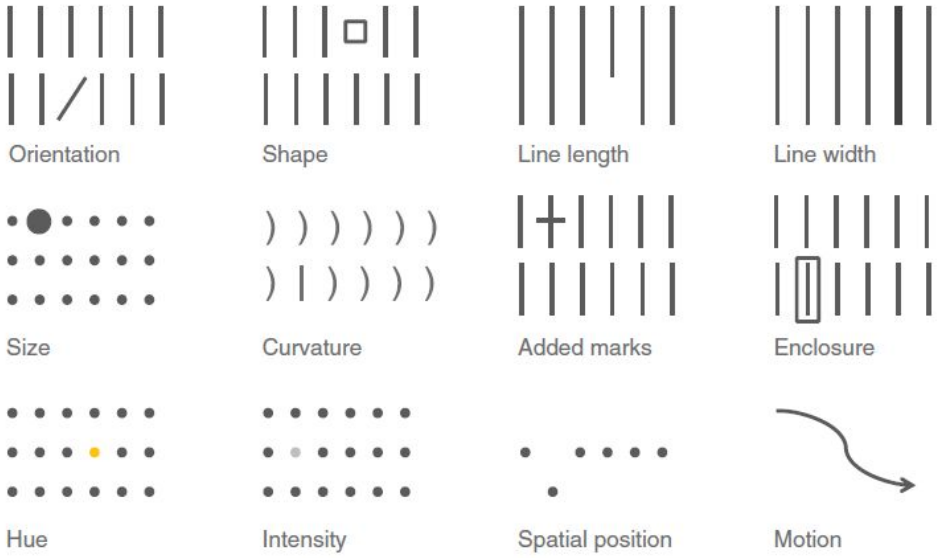



FIGURE 4.4 Preattentive attributes

In text:

- **Bold**
- *Italics*
- Underline
- Color
- Size

- Separate spatially

- Outline (enclosure)

# More than 10 per 1,000, 3 are noise-related

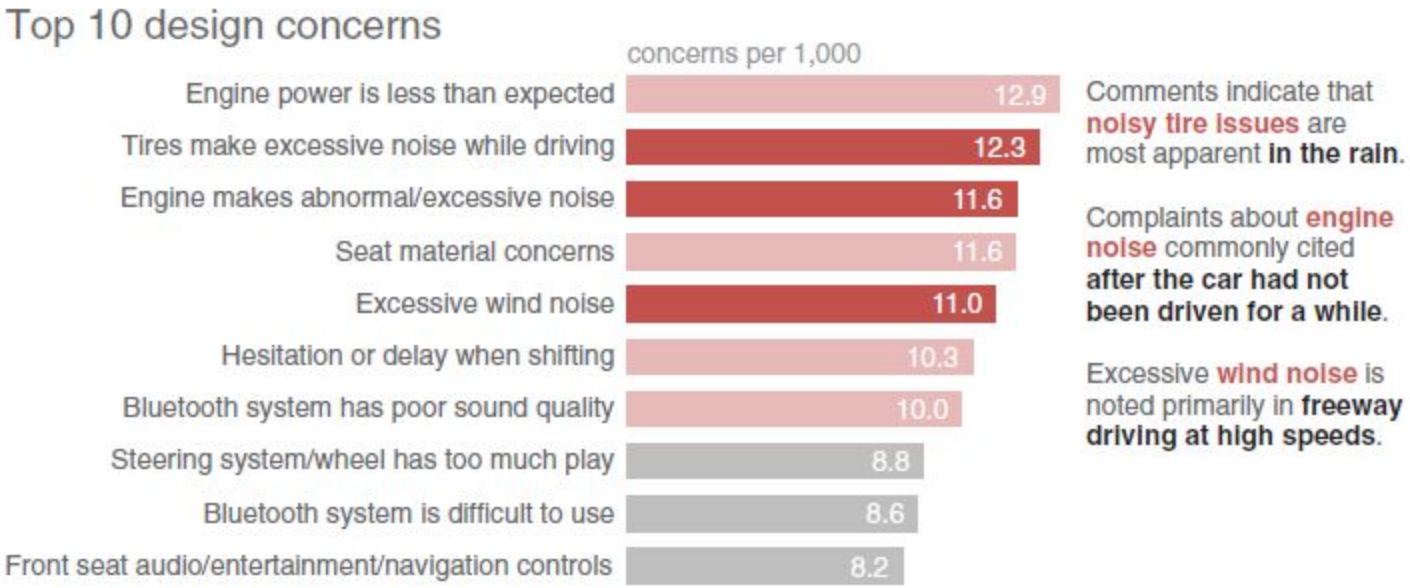



FIGURE 4.9 Create a visual hierarchy of information

# Add data directly to labels


FIGURE 4.13 Too many data labels feels cluttered

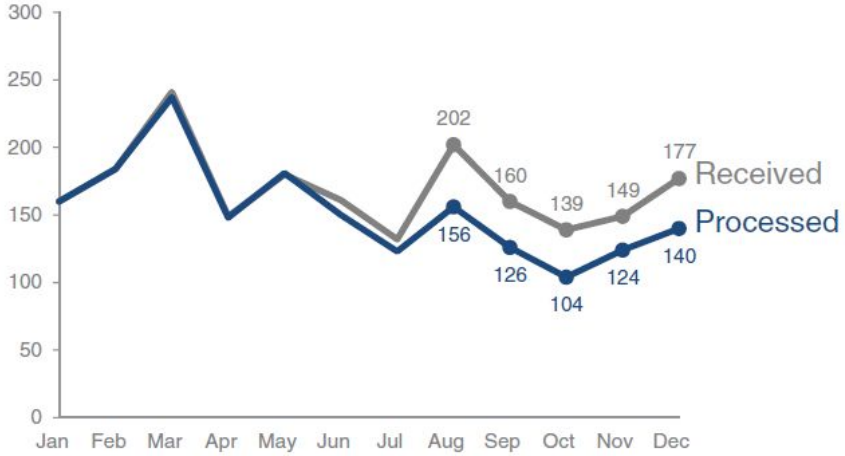



FIGURE 4.14 Data labels used sparingly help draw attention

# Use color sparingly

## Country Level Sales Rank Top 5 Drugs

Rainbow distribution in color indicates sales rank in given country from #1 (red) to #10 or higher (dark purple)

| Country | A | B | C | D | E |
|---|---|---|---|---|---|
| AUS | 1 | 2 | 3 | 6 | 7 |
| BRA | 1 | 3 | 4 | 5 | 6 |
| CAN | 2 | 3 | 6 | 12 | 8 |
| CHI | 1 | 2 | 8 | 4 | 7 |
| FRA | 3 | 2 | 4 | 8 | 10 |
| GER | 3 | 1 | 6 | 5 | 4 |
| IND | 4 | 1 | 8 | 10 | 5 |
| ITA | 2 | 4 | 10 | 9 | 8 |
| MEX | 1 | 5 | 4 | 6 | 3 |
| RUS | 4 | 3 | 7 | 9 | 12 |
| SPA | 2 | 3 | 4 | 5 | 11 |
| TUR | 7 | 2 | 3 | 4 | 8 |
| UK | 1 | 2 | 3 | 6 | 7 |
| US | 1 | 2 | 4 | 3 | 5 |

FIGURE 4.15 Use color sparingly

## **Top 5 drugs**: country-level sales rank

RANK 1 2 3 4 5+

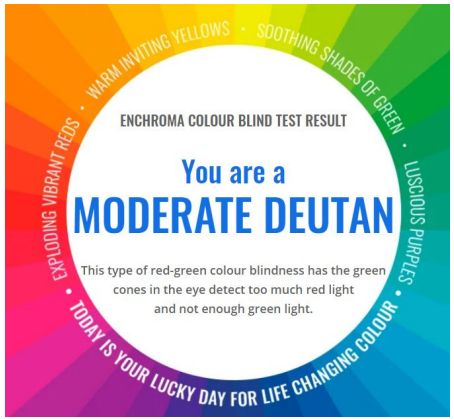
COUNTRY | DRUG

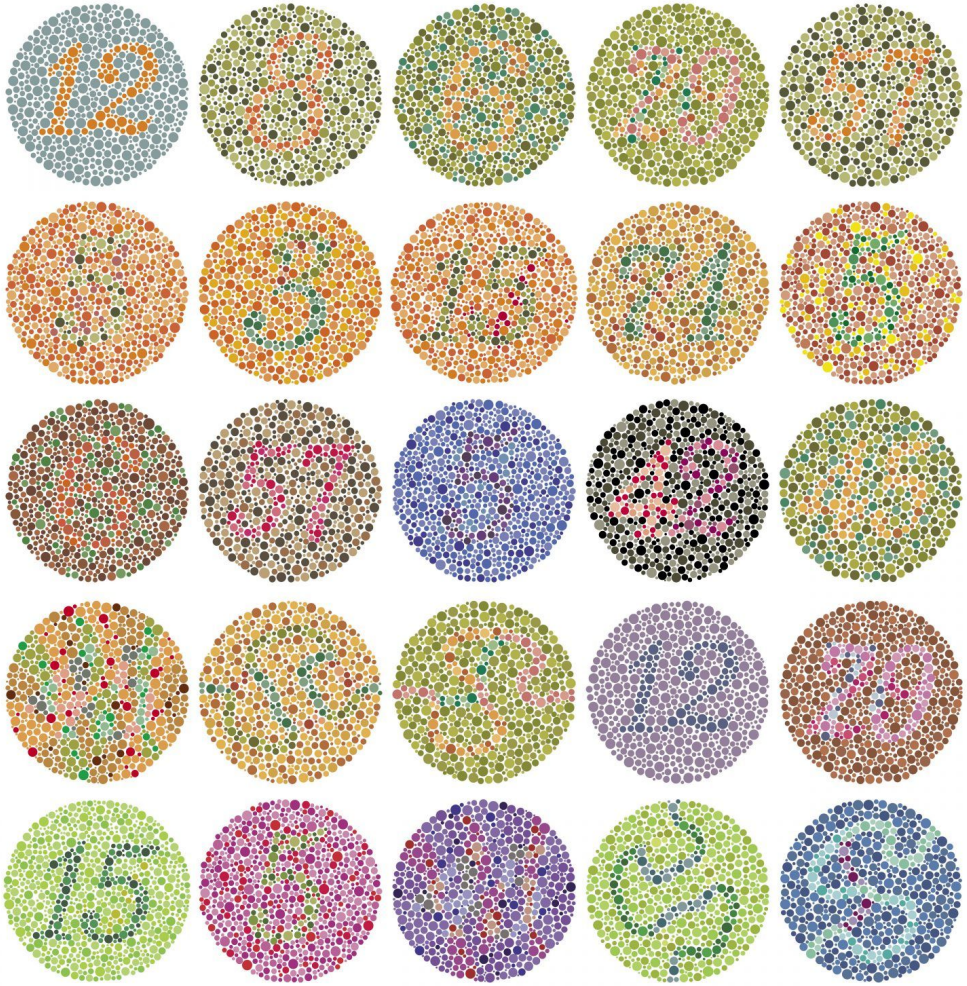
| | A | B | C | D | E |
|---|---|---|---|---|---|
| Australia | 1 | 2 | 3 | 6 | 7 |
| Brazil | 1 | 3 | 4 | 5 | 6 |
| Canada | 2 | 3 | 6 | 12 | 8 |
| China | 1 | 2 | 8 | 4 | 7 |
| France | 3 | 2 | 4 | 8 | 10 |
| Germany | 3 | 1 | 6 | 5 | 4 |
| India | 4 | 1 | 8 | 10 | 5 |
| Italy | 2 | 4 | 10 | 9 | 8 |
| Mexico | 1 | 5 | 4 | 6 | 3 |
| Russia | 4 | 3 | 7 | 9 | 12 |
| Spain | 2 | 3 | 4 | 5 | 11 |
| Turkey | 7 | 2 | 3 | 4 | 8 |
| United Kingdom | 1 | 2 | 3 | 6 | 7 |
| United States | 1 | 2 | 4 | 3 | 5 |

# Think about colors

## Design with colorblind in mind

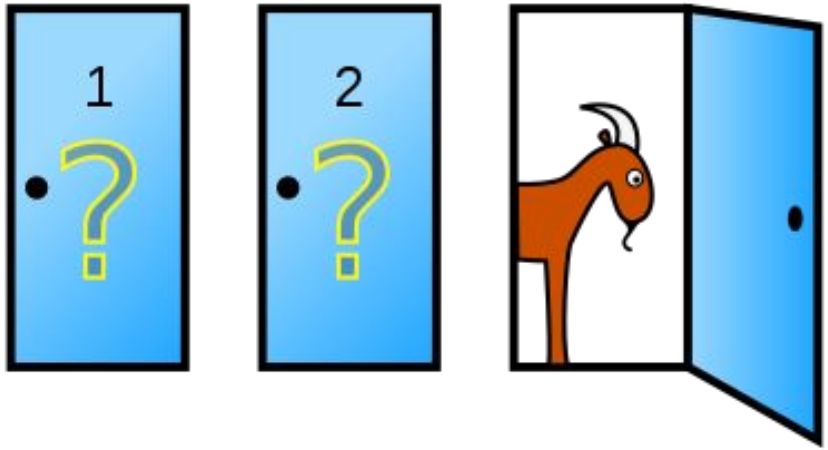
Roughly 8% of men (including my husband and a former boss) and half a percent of women are colorblind. This most frequently manifests itself as difficulty in distinguishing between shades of red and shades of green. In general, you should avoid using shades of red and shades of green together. Sometimes, though, there is useful connotation that comes with using red and green: red to denote the double-digit loss you want to draw attention to or green to highlight significant growth. You can still leverage this, but make sure to have some additional visual cue to set the important numbers apart so you aren't inadvertently disenfranchising part of your audience. Consider also using bold, varying saturation or brightness, or adding a simple plus or minus sign in front of the numbers to ensure they stand out.



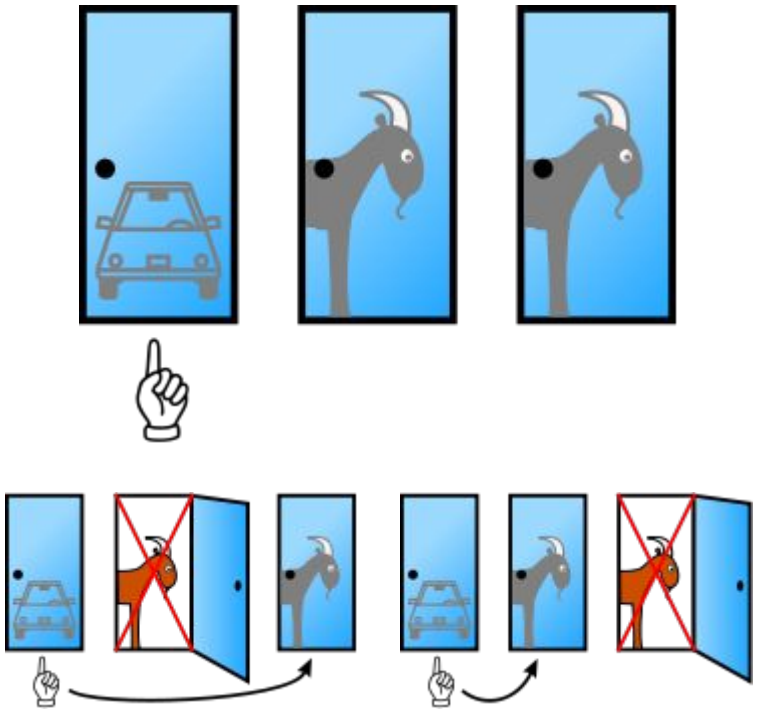
# Statistics

# The Monty Hall problem

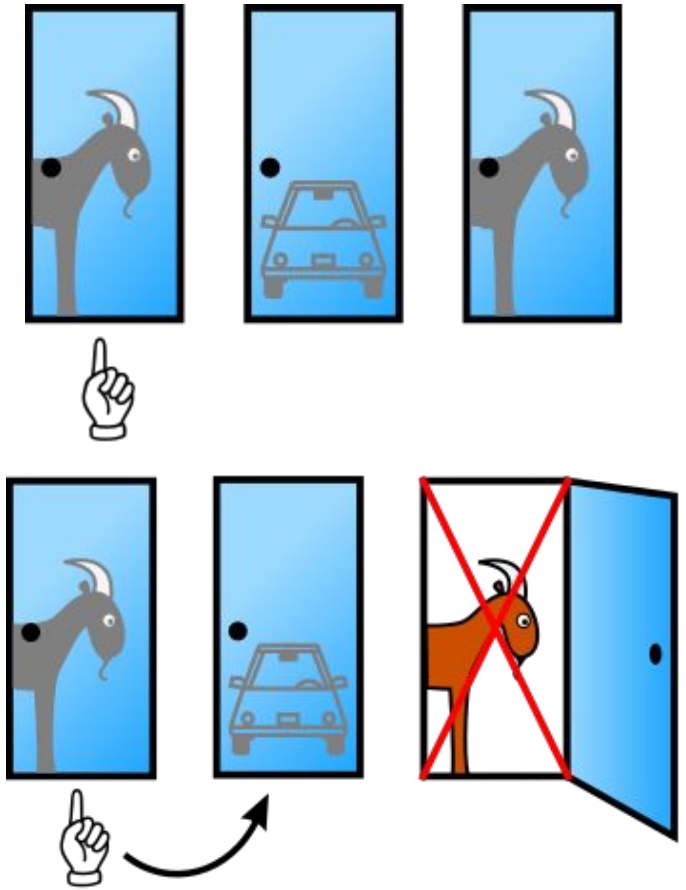



After opening the first door (e.g. #3), did the probabilities change?
Should you change your selection?!

YES!

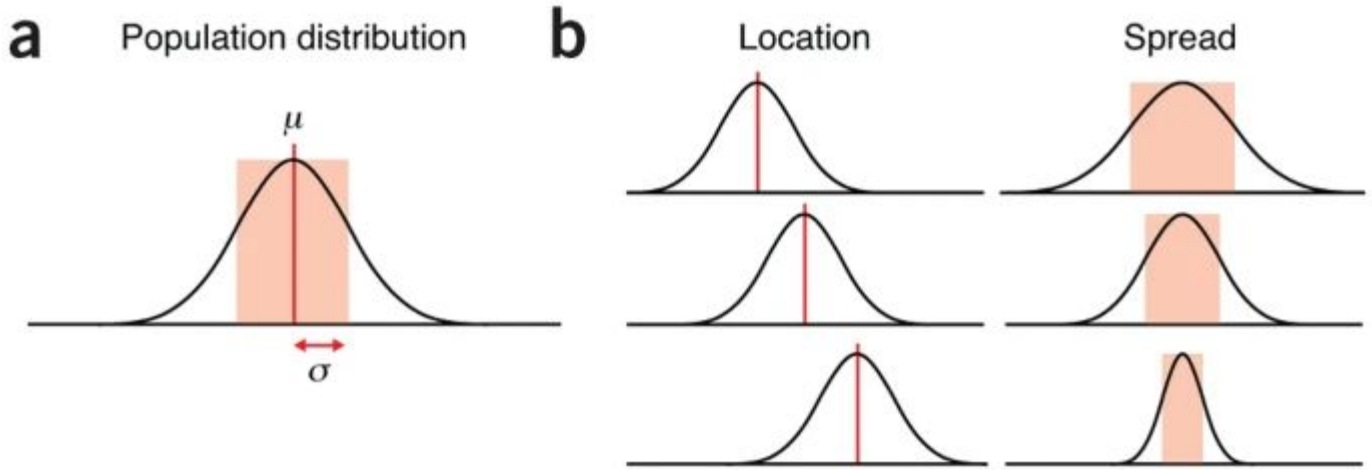


⅓ would be bad choice…



⅔ would be good choice…
Moderator has to choose a goat

# Localisation and spread of the data




μ and σ of the population

Krzywinski and Altmann, Nat Methods 2013

# Variance and s.d.

Krzywinski and Altmann, Nat Methods 2013

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Population variance
(square root of standard deviation)

if pop. mean is NOT KNOWN

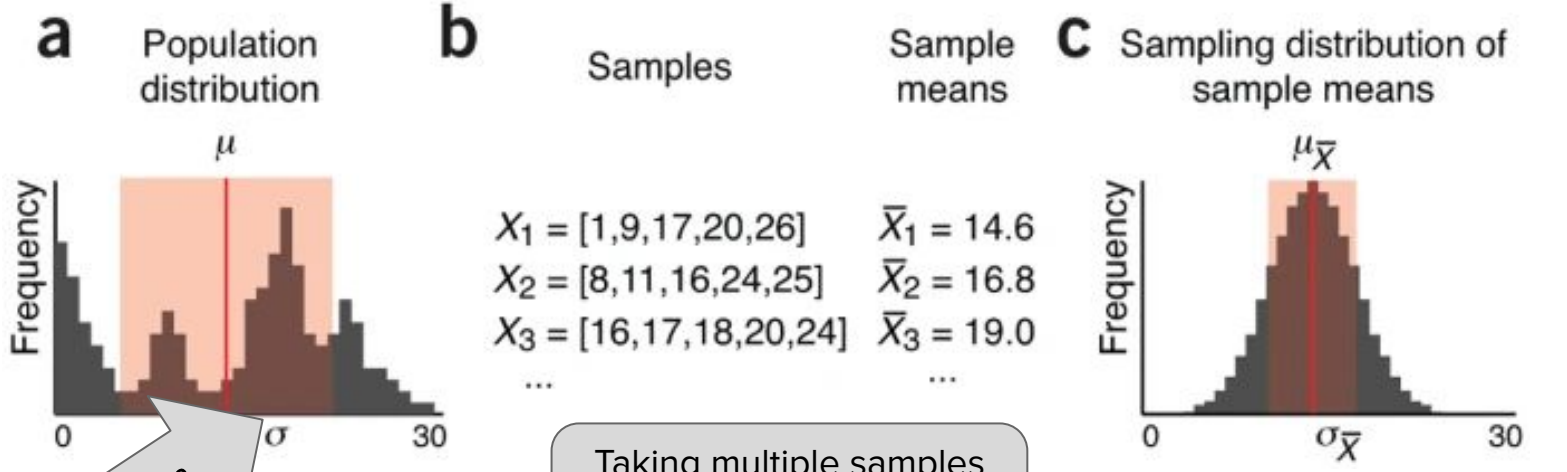
$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

Biased sample std
without Bessel correction

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

Unbiased sample variance
with Bessel correction

# Looking at the distribution of means



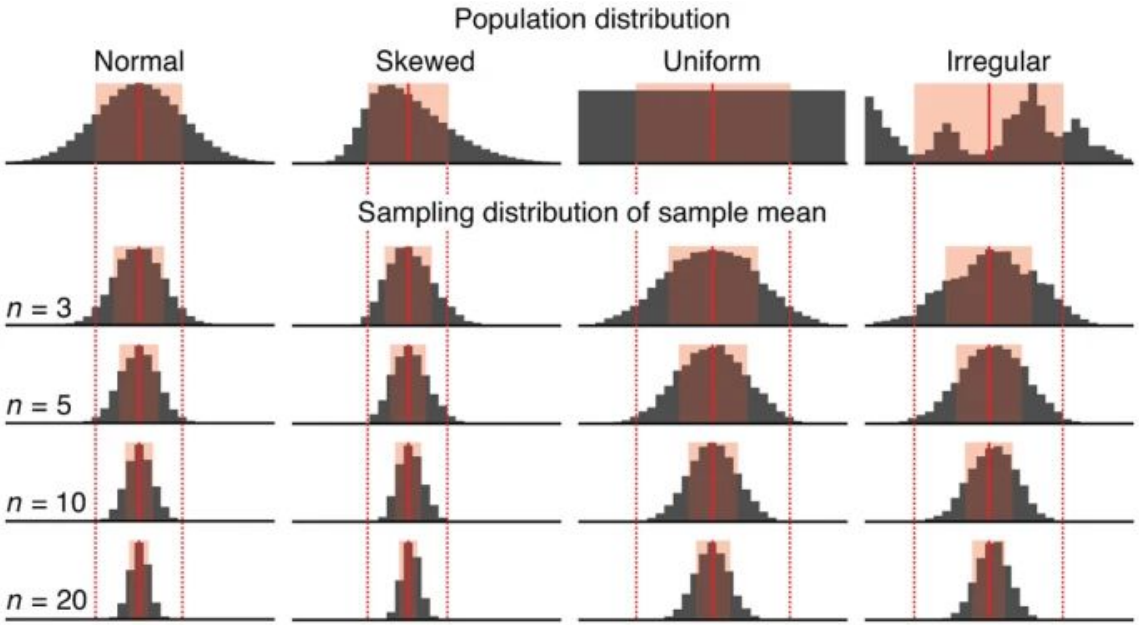

Pop. distribution not known

Taking multiple samples with n=5, compute the mean of X

Central limit theorem: Mean is normally distributed

Krzywinski and Altmann, Nat Methods 2013

# Central Limit Theorem across pop. distribution

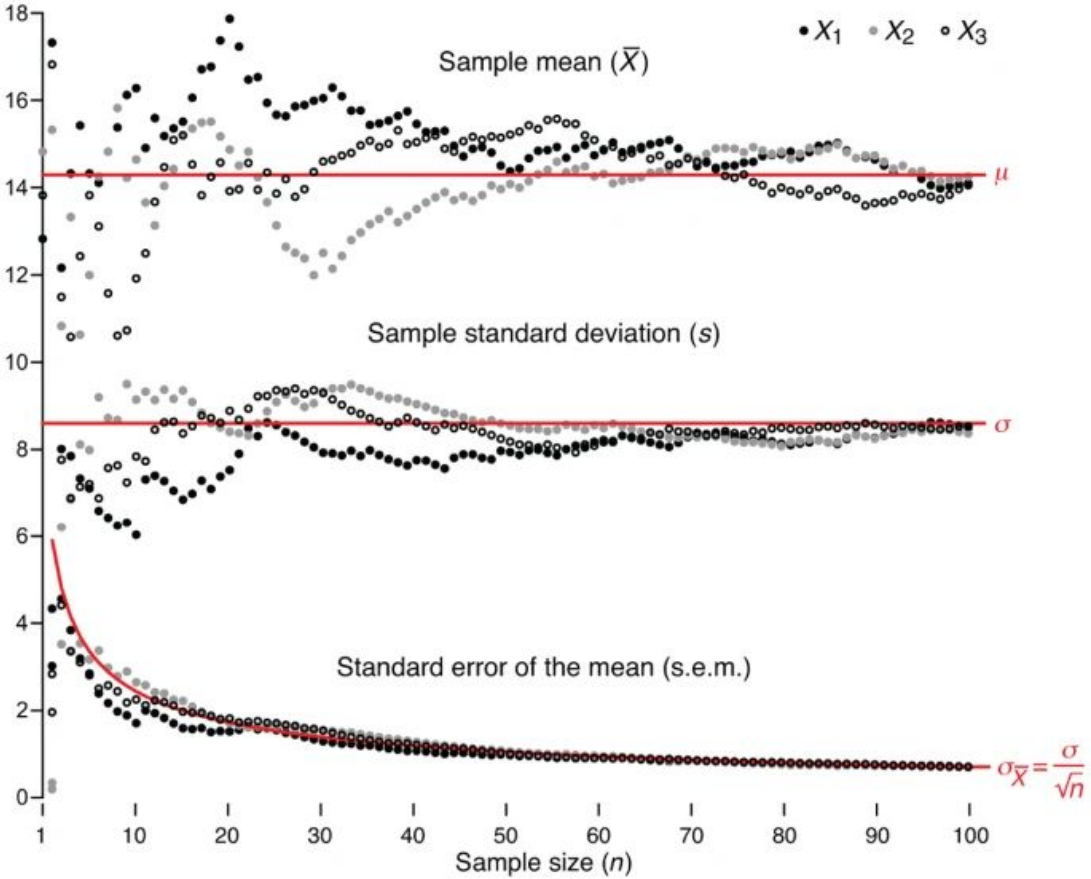



Regardless the underlying distribution, the sample means will follow a normal distribution, allowing standard statistics to be applied.

The standard error of the mean (SEM) is the deviation around the estimated mean of the sample means

Krzywinski and Altmann, Nat Methods 2013

# How sample size influences statistics



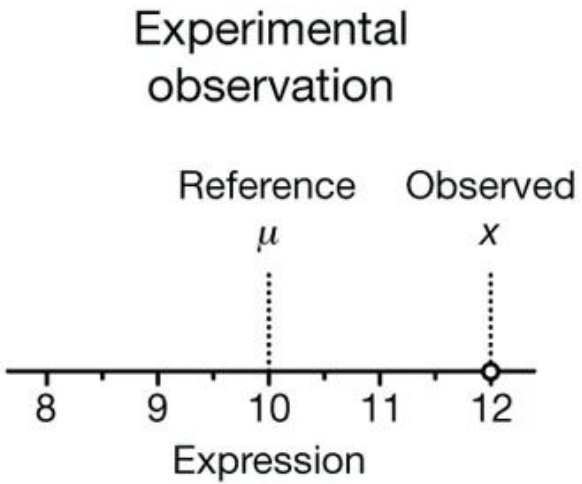

Krzywinski and Altmann, Nat Methods 2013

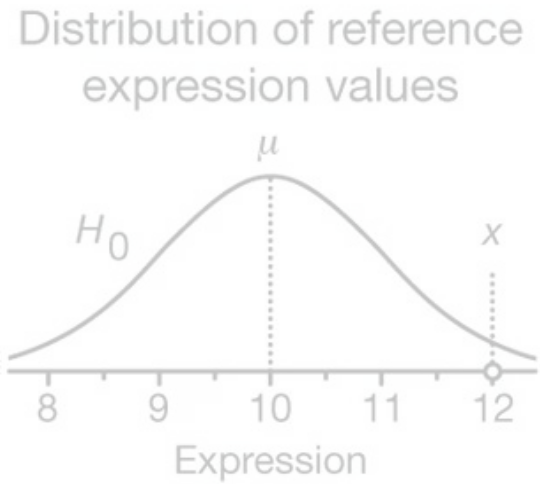
# Statistical testing

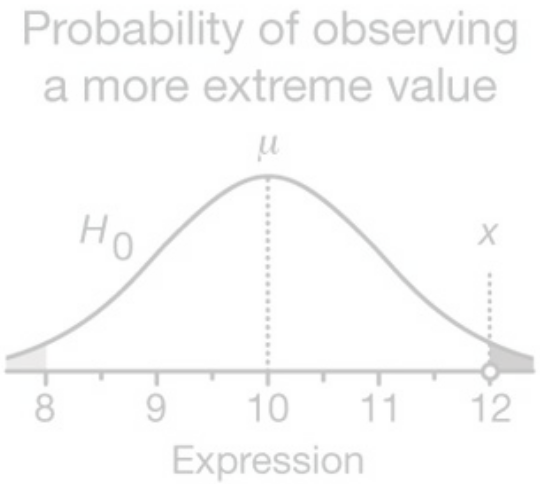
Example: protein expression measurement with Western Blot, reference: 10, observed (x): 12

Is x=12 a common measurement? => H0 probability

Now we can find the probability P to observe >= x

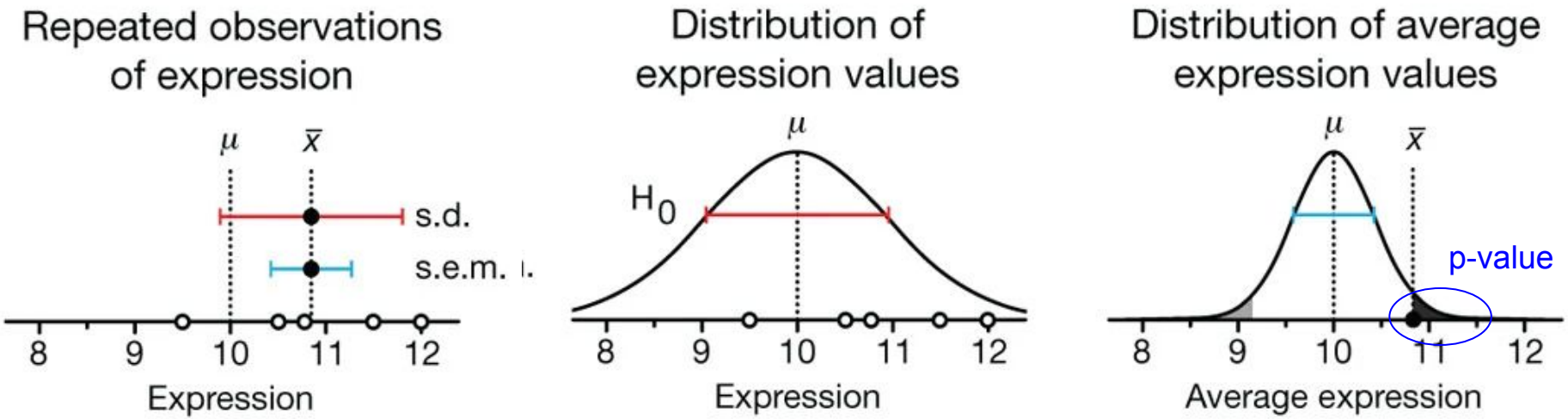


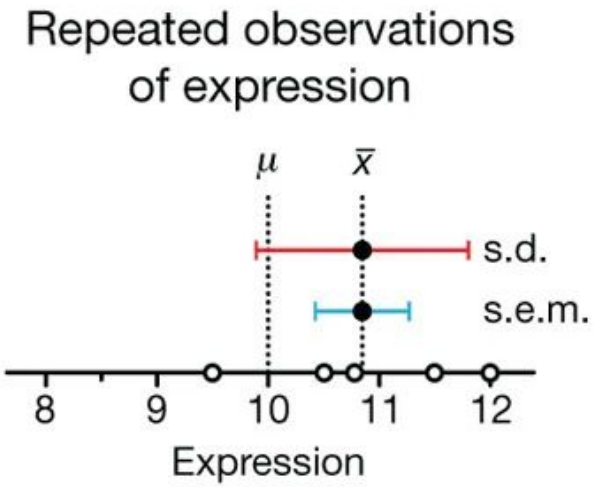

Krzywinski and Altmann, Nat Methods 2013

# Wait, how do we know the H0 spread?!




Krzywinski and Altmann, Nat Methods 2013

# What do we need?

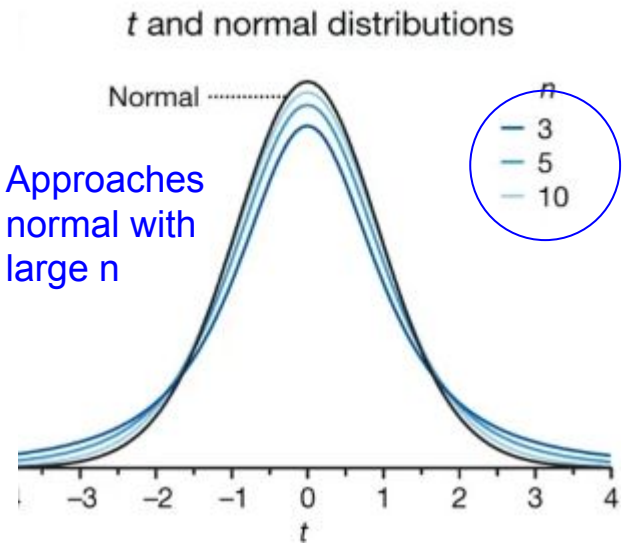



Krzywinski and Altmann, Nat Methods 2013

# t-Statistic

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Probability density function (PDF)




Approaches normal with large n

Depends on n, i.e. DOF

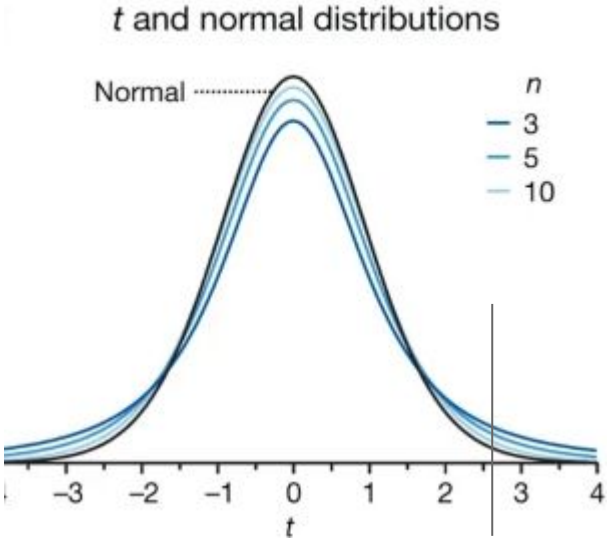
Student's ***t*-distribution** has the probability density function given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where $\nu$ is the number of *degrees of freedom* and $\Gamma$ is the gamma function.

Krzywinski and Altmann, Nat Methods 2013

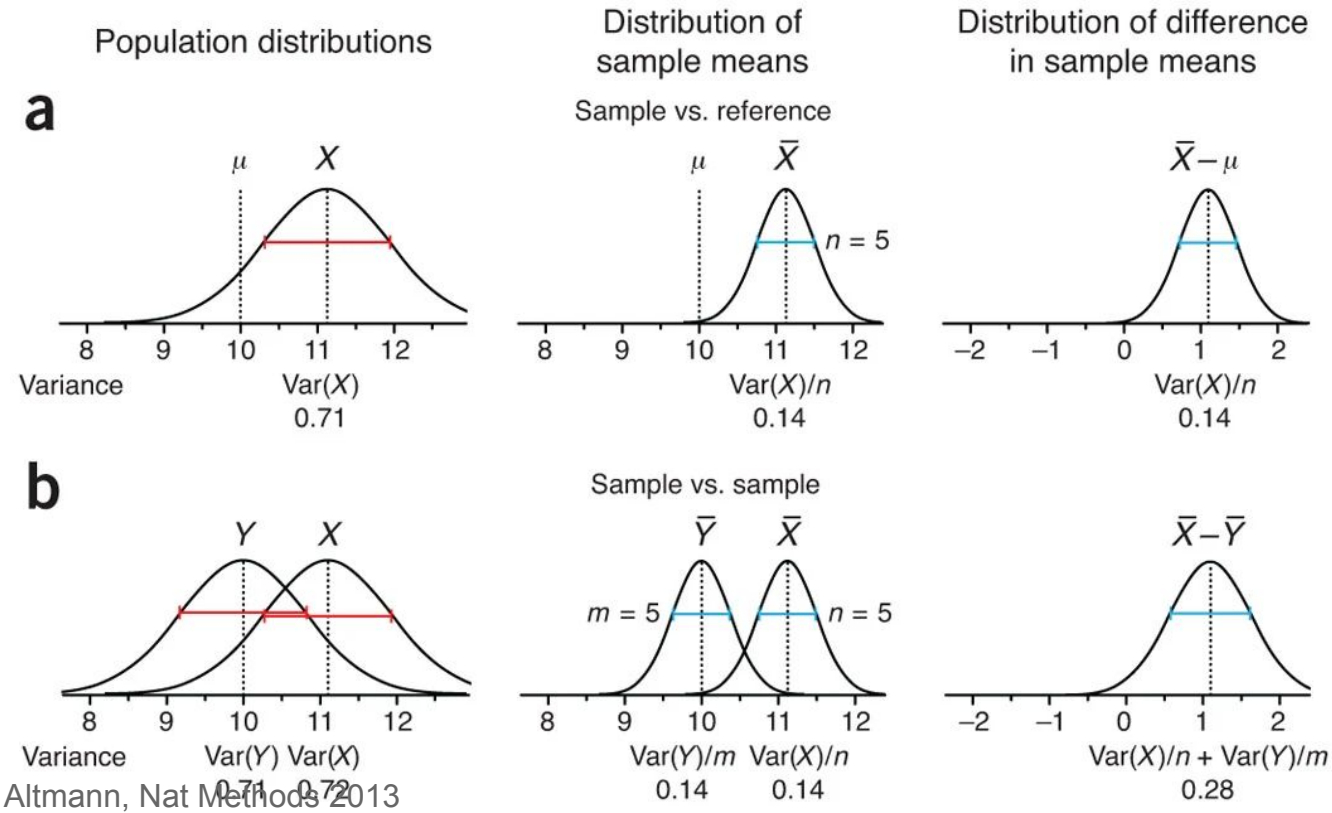
# Finding p-value




p-value at given t

Krzywinski and Altmann, Nat Methods 2013

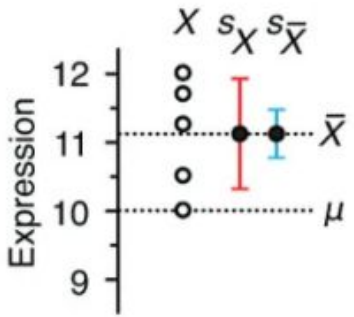
# Comparing more than 2 samples




Krzywinski and Altmann, Nat Methods 2013

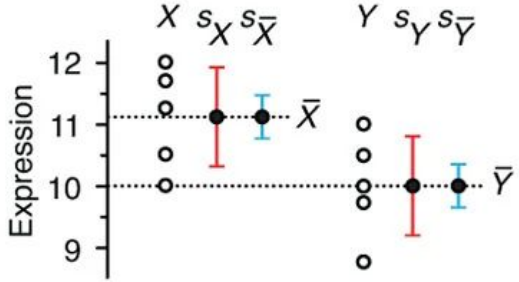
# One- and two-sample test

One-sample *t*-test




$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

Two-sample *t*-test




$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_{\bar{X} - \bar{Y}}}$$

$$s^2_{\bar{X} - \bar{Y}} = s^2_{\bar{X}} + s^2_{\bar{Y}} \approx s_p^2/n + s_p^2/m$$

Pooled sd, n=m

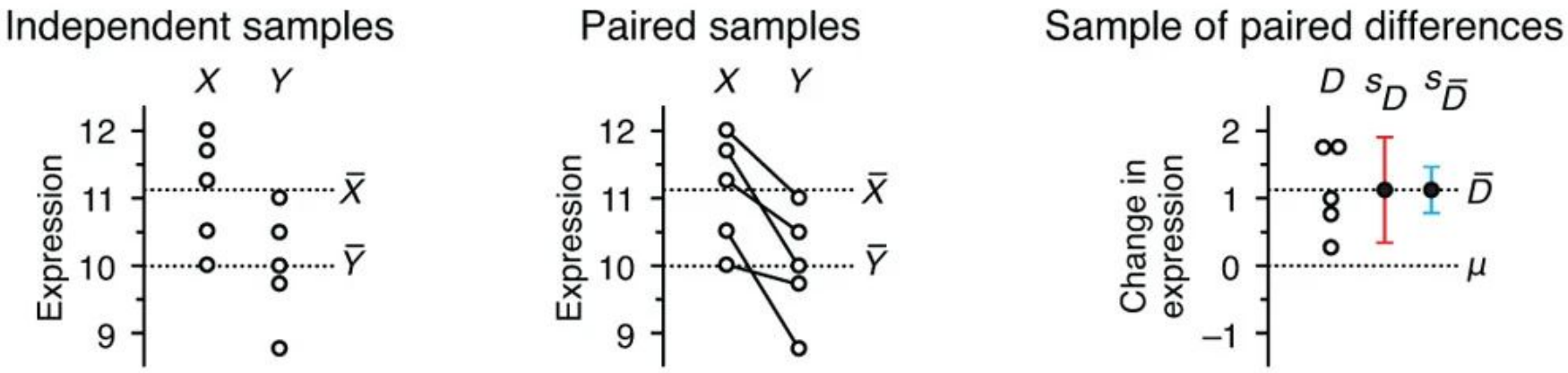
$$s_p = \sqrt{\frac{s^2_{X_1} + s^2_{X_2}}{2}}.$$

Pooled sd, n!=m

$$s_p = \sqrt{\frac{(n_1 - 1)\, s^2_{X_1} + (n_2 - 1)\, s^2_{X_2}}{n_1 + n_2 - 2}}$$

Krzywinski and Altmann, Nat Methods 2013

# Paired t-test




Krzywinski and Altmann, Nat Methods 2013

# Assumptions for the t-test

- Population is normally distributed (robust if it's not)
- Variances are equal (robust if they are not)
- Samples are NOT CORRELATED! (very important)

Krzywinski and Altmann, Nat Methods 2013

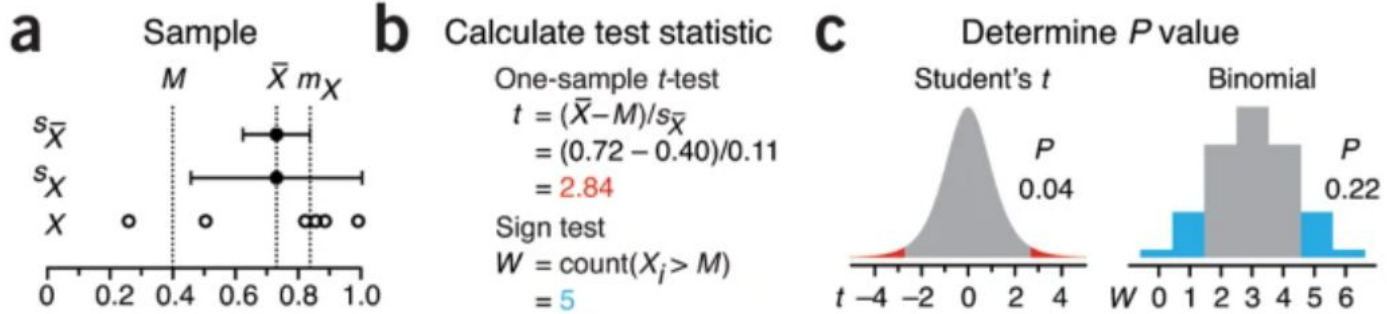
# Non-parametric tests

- nonparametric tests are more suitable for data that come from skewed distributions or have a discrete or ordinal scale.
- Nonparametric tests such as the **sign and Wilcoxon rank-sum tests** relax distribution assumptions BUT lower sensitivity owing to less information inherent in their assumptions.

Krzywinski and Altmann, Nat Methods 2013

# Sign rank test

**Figure 1: A sample can be easily tested against a reference value using the sign test without any assumptions about the population distribution.**

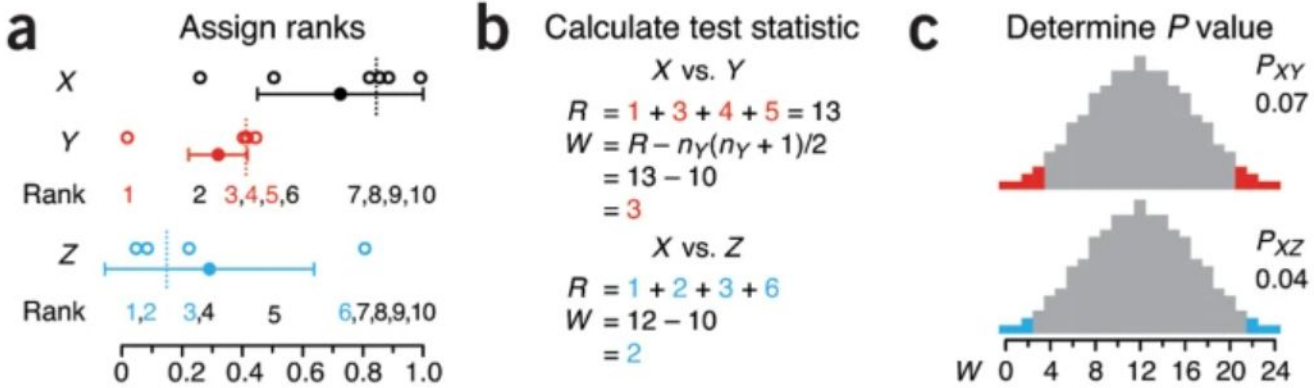



(**a**) Sample $X$ ($n = 6$) is tested against a reference $M = 0.4$. Sample mean $\bar{X}$ is shown with s.d. ($s_X$) and s.e.m. error bars ($s_{\bar{X}}$). $m_X$ is sample median. (**b**) The $t$-statistic compares $\bar{X}$ to $M$ in units of s.e.m. The sign test's $W$ is the number of sample values larger than $M$. (**c**) Under the null, $t$ follows Student's $t$-distribution with five degrees of freedom, whereas $W$ is described by the binomial with 6 trials and $P = 0.5$. Two-tailed $P$ values are shown.

Krzywinski and Altmann, Nat Methods 2013

# Wilcoxon sum ranked test

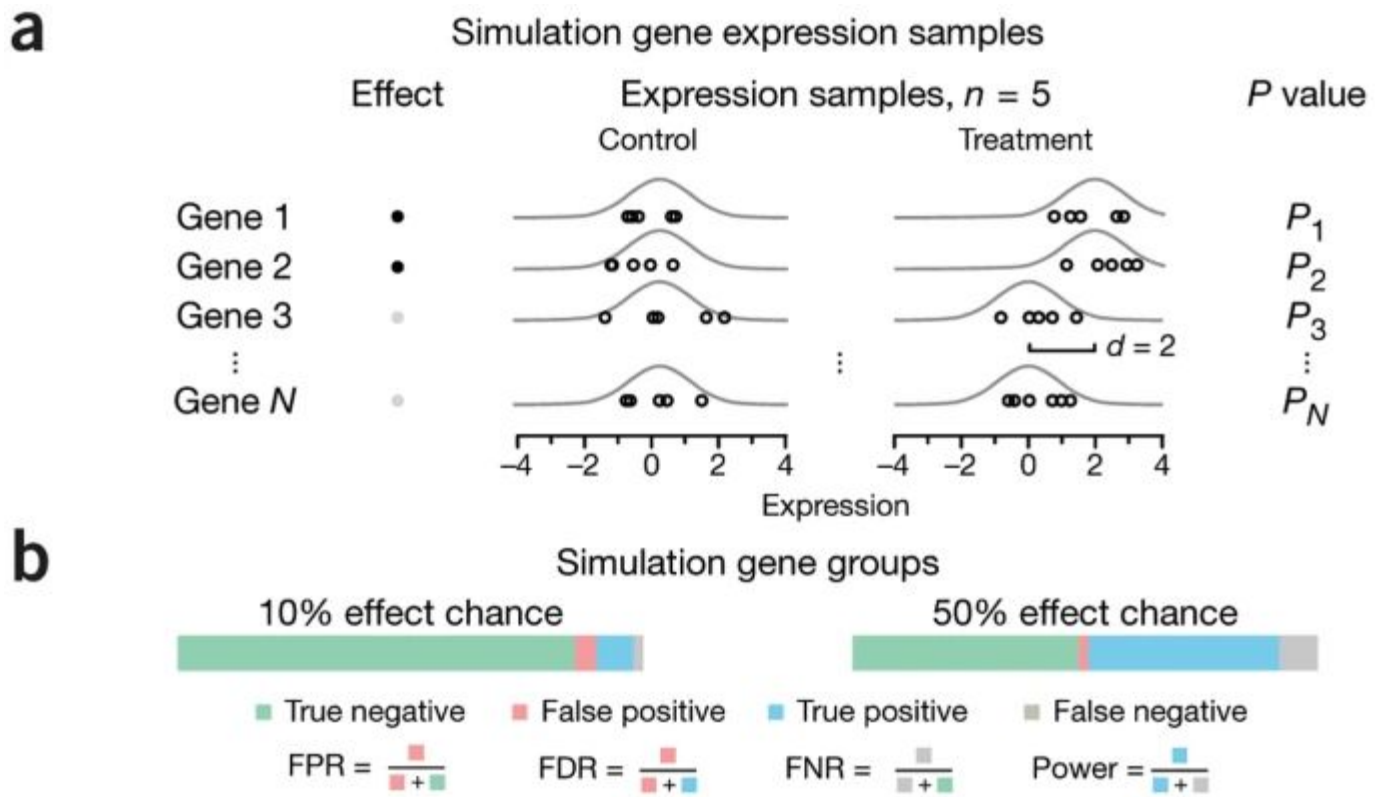
**Figure 2: Many nonparametric tests are based on ranks.**




(**a**) Sample comparisons of *X* vs. *Y* and *X* vs. *Z* start with ranking pooled values and identifying the ranks in the smaller-sized sample (e.g., 1, 3, 4, 5 for *Y*; 1, 2, 3, 6 for *Z*). Error bars show sample mean and s.d., and sample medians are shown by vertical dotted lines. (**b**) The Wilcoxon rank-sum test statistic *W* is the difference between the sum of ranks and the smallest possible observed sum. (**c**) For small sample sizes the exact distribution of *W* can be calculated. For samples of size (6, 4), there are only 210 different rank combinations corresponding to 25 distinct values of *W*.

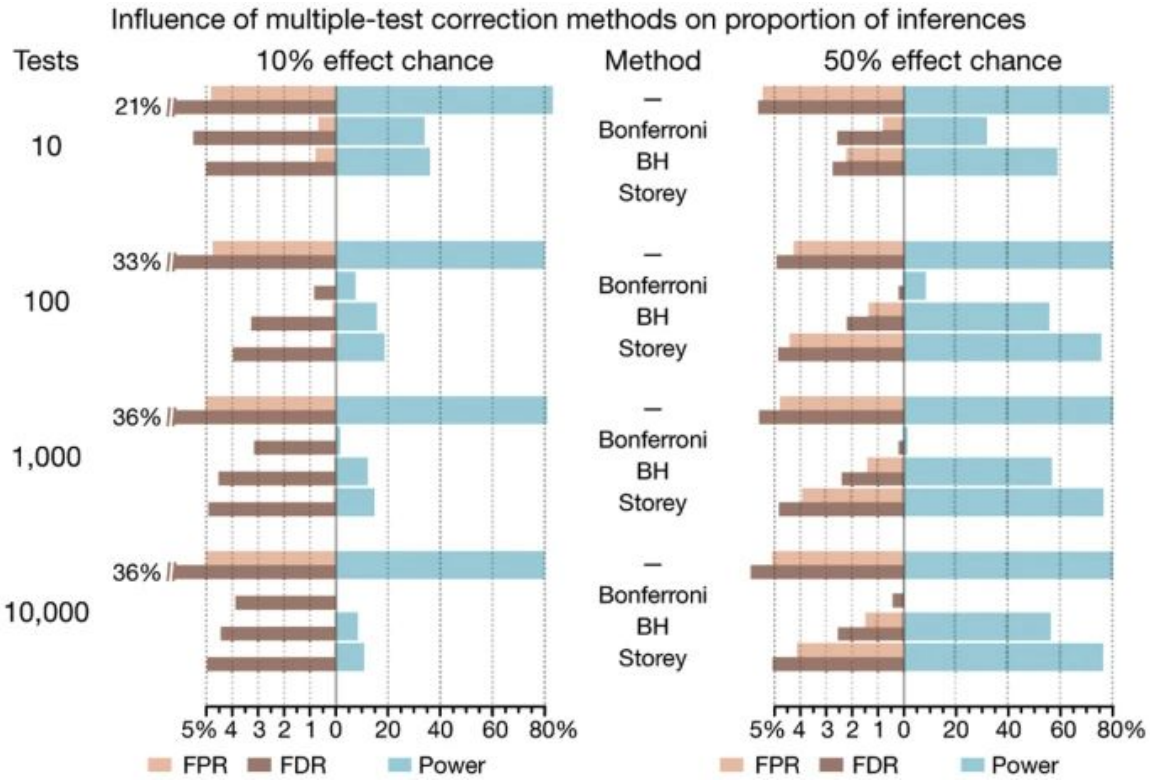
Krzywinski and Altmann, Nat Methods 2013

# Multiple test correction




Krzywinski and Altmann, Nat Methods 2013

# Multiple test correction




Krzywinski and Altmann, Nat Methods 2013

**Bonferroni**:
Kills statistical power, simple but conservative correction (P' = N*P)

**Benjamini-Hochberg (BH)** scales P values in inverse proportion to their rank when ordered

**Storey** only works for larger test numbers

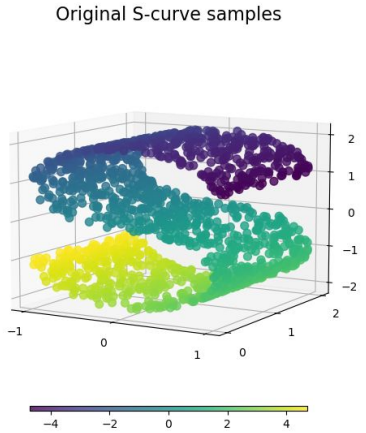
# Statistics cheat sheet

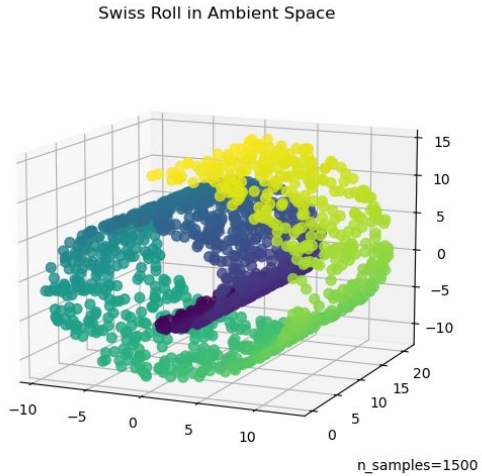
https://towardsdatascience.com/demystifying-statistical-analysis-1-a-handy-cheat-sheet-b6229bf992cf

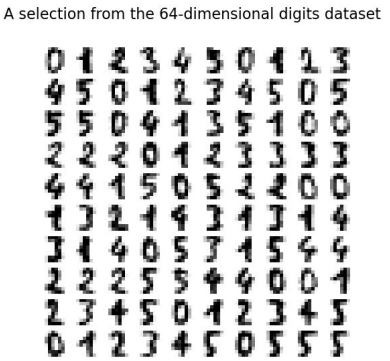
| Predictor / Covariate / Independent Variable | | Criterion / Measure / Dependent Variable: Categorical — 1 Variable, 2 Categories | Categorical — 1 Variable, >2 Categories | Continuous — 1 Variable | Continuous — >1 Variable |
|---|---|---|---|---|---|
| Categorical | 1 Variable<br>2 Categories<br>Between-subjects | $\chi^2$ Test<br>(Crosstabs → Statistics → ☑ Chi-square) | | **Independent *t* Test**<br>(Compare Means → Independent-Samples) | |
| | 1 Variable<br>2 Categories<br>Within-subjects | | | **Paired *t* Test**<br>(Compare Means → Paired Samples) | |
| | 1 Variable<br>>2 Categories<br>Between-subjects | | | **One-Way ANOVA**<br>(Compare Means → One-way ANOVA) | **One-Way MANOVA**<br>(General Linear Model → Multivariate → Add Dependent Variables) |
| | 1 Variable<br>>2 Categories<br>Within-subjects | | | **Repeated Measures ANOVA**<br>(General Linear Model → Repeated Measures → Add Within-Sbj Factors) | **Repeated Measures MANOVA**<br>(Repeated Measures ANOVA → Add Measures) |
| | >1 Variable<br>All Categorical<br>Between-subjects | **Binomial Logistic Regression with Categorical Predictors**<br>(Regression → Binary Logistic → Categorical Covariates) | **Multinomial Logistic Regression**<br>(Regression → Multinomial Logistic) | **Factorial ANOVA**<br>(General Linear Model → Univariate → Add Fixed Factors) | **Factorial MANOVA**<br>(One-Way MANOVA → Add Fixed Factors) |
| | >1 Variable<br>All Categorical<br>Mixed Within- & Between-subjects | | | **Mixed-Design ANOVA**<br>(Repeated Measures ANOVA → Add Between-Sbj Factors) | **Mixed-Design MANOVA**<br>(Mixed-Design ANOVA → Add Measures) |
| | >1 Variable<br>Mixed Categorical & Continuous | | | **One-Way ANCOVA**<br>(One-Way ANOVA → Add Covariates) | **One-Way MANCOVA**<br>(One-Way MANOVA → Add Covariates) |
| Continuous | 1 Variable | **Binomial Logistic Regression**<br>(Regression → Binary Logistic) | | **Simple Linear Regression**<br>(Regression → Linear) | **Multivariate Linear Regression**<br>(General Linear Model → Multivariate → Add Dependent Variables → No Fixed Factors → Add Covariates) |
| | >1 Variable | | | **Multiple Linear Regression**<br>(Regression → Linear) | |

# WHY IS THIS IMPORTANT?

# Manifold learning




Can we project this to a lower dimensional space?

# Manifold learning

Journal of Machine Learning Research 9 (2008) 2579-2605 Submitted 5/08; Revised 9/08; Published 11/08

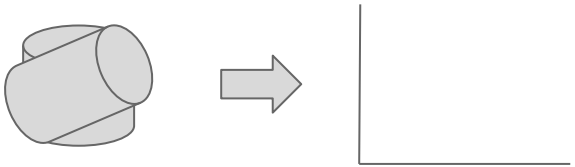
## Visualizing Data using t-SNE

**Laurens van der Maaten** LVDMAATEN@GMAIL.COM
*TiCC*
*Tilburg University*
*P.O. Box 90153, 5000 LE Tilburg, The Netherlands*

**Geoffrey Hinton** HINTON@CS.TORONTO.EDU
*Department of Computer Science*
*University of Toronto*
*6 King's College Road, M5S 3G4 Toronto, ON, Canada*

**Laurens van der Maaten**

*Research scientist in machine learning and computer vision.*



# T-distributed Stochastic Neighbour Embedding

1

For each pair of points $i$ and $j$ in the high-dimensional space:

- Compute the probability $p_{ij}$ that point $j$ would be picked as a neighbor of point $i$ based on a Gaussian distribution centered at $i$.
- $p_{ij}$ depends on the distance between $i$ and $j$ and a parameter $\sigma_i$ (perplexity-based bandwidth).

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Sigma is perplexity-based bandwidth

2

Map the high-dimensional points $x_i$ to low-dimensional points $y_i$.

Compute similarities $q_{ij}$ in the low-dimensional space, but now using a **Student's t-distribution** instead of a Gaussian:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

The t-distribution has heavier tails than a Gaussian, which helps separate clusters and avoid crowding points together.

3

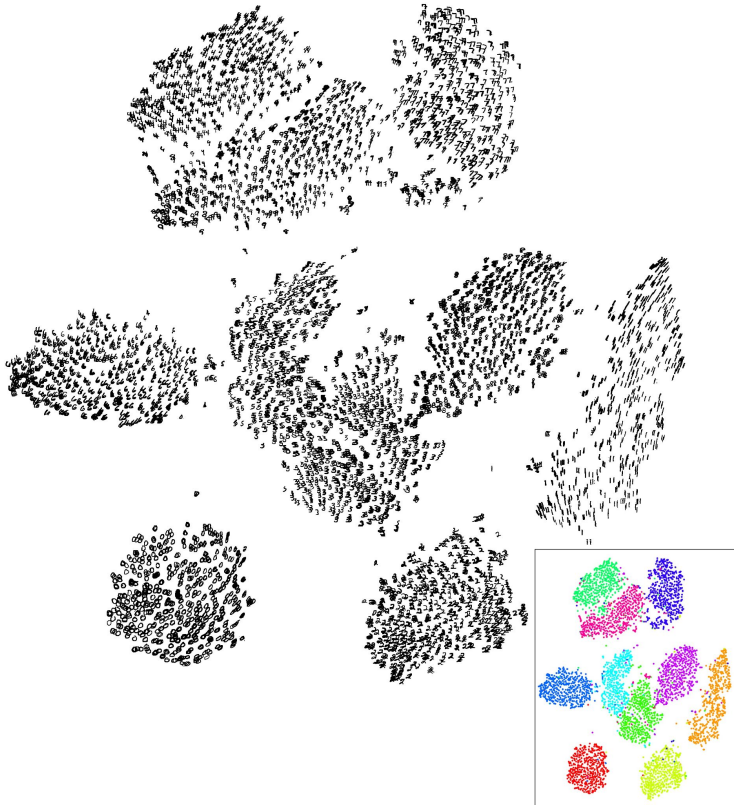
Use **Kullback-Leibler (KL) divergence** as the objective function to minimize:

$$KL(P\|Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

Intuition:

- $p_{ij}$: Represents relationships in the high-dimensional space.
- $q_{ij}$: Represents relationships in the low-dimensional space.
- Minimize the difference between these distributions so that the local structure is preserved.

# t-SNE



1. Those hyperparameters (e.g. learning rate, perplexity) really matter
2. Cluster sizes in a t-SNE plot mean nothing
3. Distances between clusters might not mean anything
4. Random noise doesn't always look random.
5. You can see some shapes, sometimes
6. For topology, you may need more than one plot

https://distill.pub/2016/misread-tsne/

## sklearn.manifold.TSNE

*class* sklearn.manifold.TSNE(*n_components=2, *, perplexity=30.0, early_exaggeration=12.0, learning_rate='warn', n_iter=1000, n_iter_without_progress=300, min_grad_norm=1e-07, metric='euclidean', init='warn', verbose=0, random_state=None, method='barnes_hut', angle=0.5, n_jobs=None, square_distances='legacy'*) [source

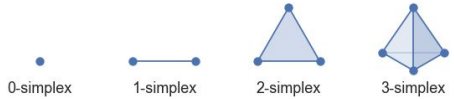
T-distributed Stochastic Neighbor Embedding.

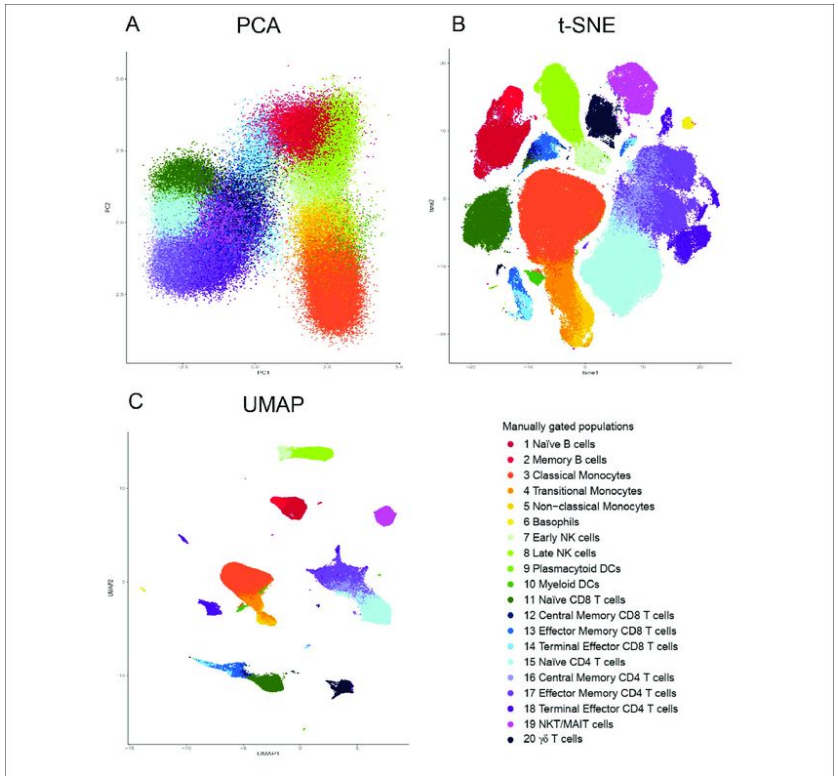
t-SNE [1] is a tool to visualize high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. t-SNE has a cost function that is not convex, i.e. with different initializations we can get different results.

It is highly recommended to use another dimensionality reduction method (e.g. PCA for dense data or TruncatedSVD for sparse data) to reduce the number of dimensions to a reasonable amount (e.g. 50) if the number of features is very high. This will suppress some noise and speed up the computation of pairwise distances between samples. For more tips see Laurens van der Maaten's FAQ [2].

# UMAP

## Čech complex → working on simplex




https://pair-code.github.io/understanding-umap/

### How to (mis)read UMAP

While UMAP offers a number of advantages over t-SNE, it's by no means a silver bullet - and reading and understanding its results requires some care. It's worth revisiting our previous work on (mis)reading t-SNE, since many of the same takeaways apply to UMAP:

**1. Hyperparameters really matter**

Choosing good values isn't easy, and depends on both the data and your goals (eg, how tightly packed the projection ought to be). This is where UMAP's speed is a big advantage - By running UMAP multiple times with a variety of hyperparameters, you can get a better sense of how the projection is affected by its parameters.

**2. Cluster sizes in a UMAP plot mean nothing**

Just as in t-SNE, the size of clusters relative to each other is essentially meaningless. This is because UMAP uses local notions of distance to construct its high-dimensional graph representation.

**3. Distances between clusters might not mean anything**

Likewise, the distances between clusters is likely to be meaningless. While it's true that the global positions of clusters are better preserved in UMAP, the distances between them are not meaningful. Again, this is due to using local distances when constructing the graph.

**4. Random noise doesn't always look random.**

Especially at low values of `n_neighbors`, spurious clustering can be observed.

**5. You may need more than one plot**

Since the UMAP algorithm is stochastic, different runs with the same hyperparameters can yield different results. Additionally, since the choice of hyperparameters is so important, it can be very useful to run the projection multiple times with various hyperparameters.

# Done for today

# Homework

In this homework you will work with the wine quality dataset. In the first step, you need to apply a dimensionality reduction to the features to visualize them in 2D. After that, you will have to come up with a hypothesis and apply statistical tests to analyse it.