Data
Divers

Data Science – Yellow Belt

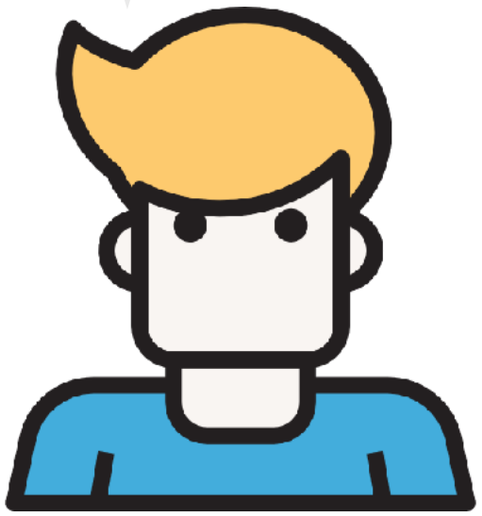
Lecture 1

by Sebastian Sauer

Introducing

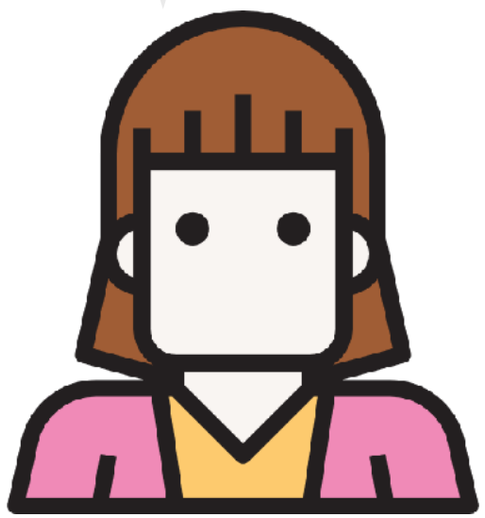
Introducing

I'm a great business man!



Don

I like data. Crunched data.



Angi

My job is to play the nerd.

Wolfi

Truly yours

Sebastian Sauer

Teacher

Scientist

Learner

https://de.wikipedia.org/wiki/Der_B%C3%BCcherwurm#/media/Datei:Carl_Spitzweg_021.jpg

Photo by Salmen Bejaoui on Unsplash

Overview

Learning goals

- **Understanding** what „**big data**“ is, with a focus on „big“.
- Being able to **differentiate typical buzzwords** such as data science, data mining …
- Knowing the **basic analytical approaches**
- Being familiar with **market research case studies** on big data.
- Having a grip on the **limits** of big data analytics.
- Knowing how to pursue the learning journey autonomously, ie., **self-lean paths**.

Agenda

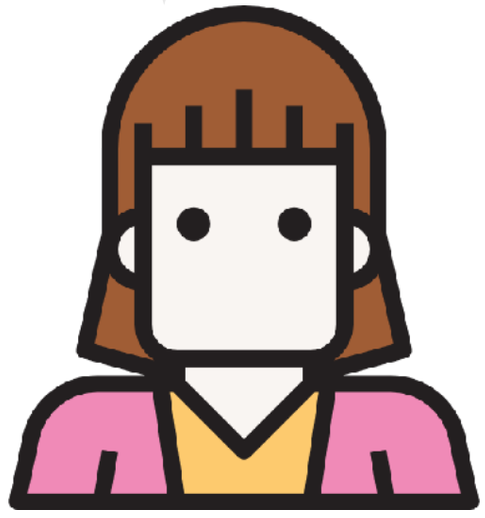
- What is Big Data?
- What made data big?
- Big Data – Buzzwords
- The two cultures of statistical modelling
- Big data market research case studies
- Why most analytical insights are unfit for business decisions
- From Mediocristan to Extremistan
- Self-learning paths

I need you

Guys. This can be a tough ride. Are you prepared to be fully concentrated?

Wolfi

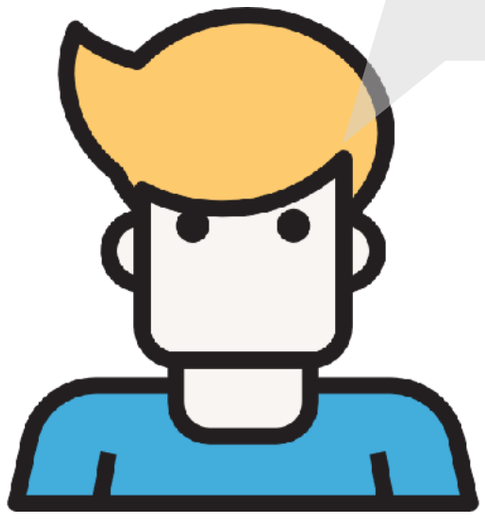
Wolfi! Don’t again scare people away.



Angi

What is „Big Data“?

Big Data: How big is „big“?

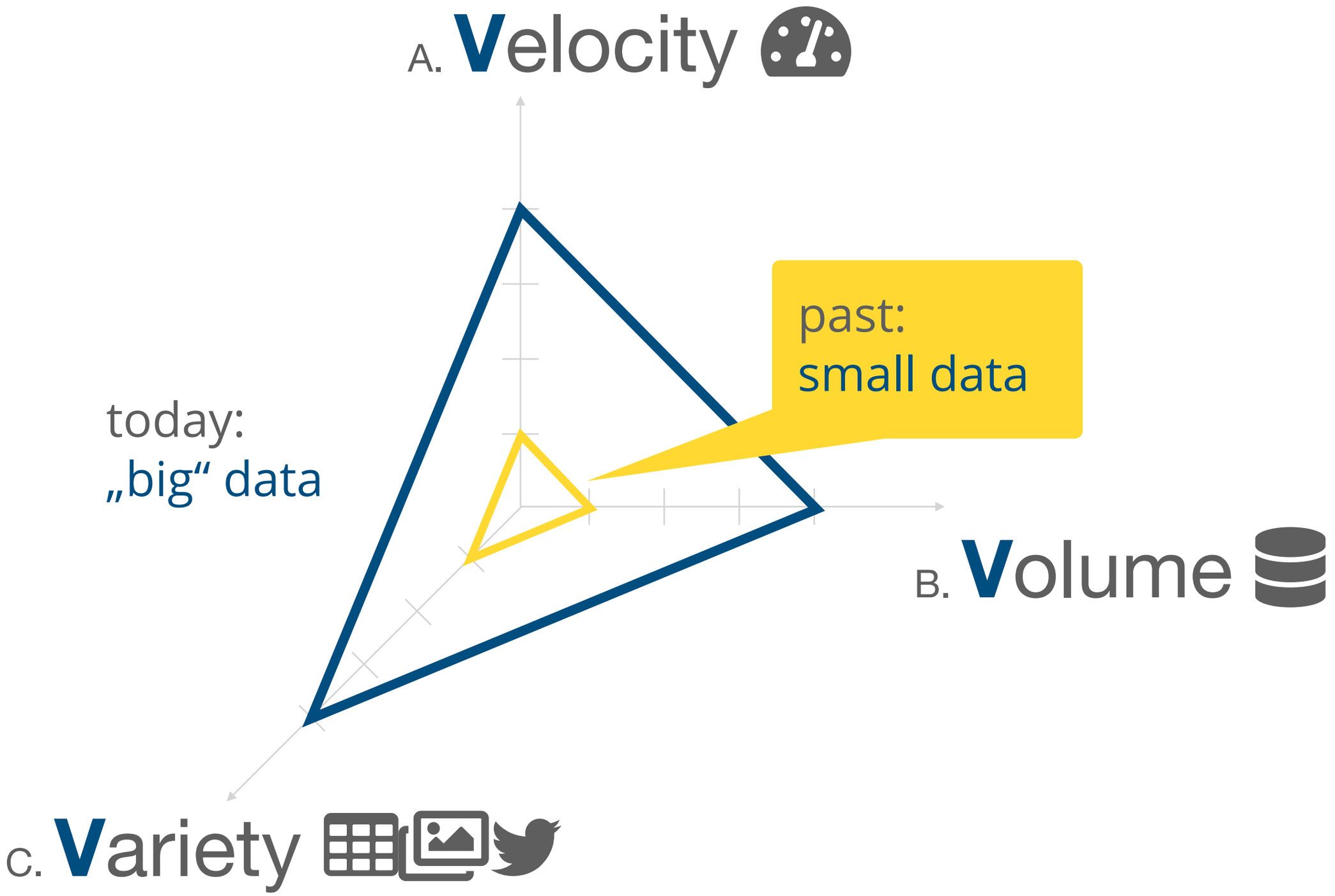


Don

Wolfi

McKinsey & Company. (2011). *Big data: The next frontier for innovation, competition, and productivity*. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

The 3 Vs of Big Data



Dorschel, J. (Ed.). (2015). *Praxishandbuch Big Data.* Springer. https://doi.org/10.1007/978-3-658-07289-6

Data analysis now and then

Hey, it was me who came up with all that classical statistics stuff – back in the 1920, you youngsters!

http://hdl.handle.net/2440/81668

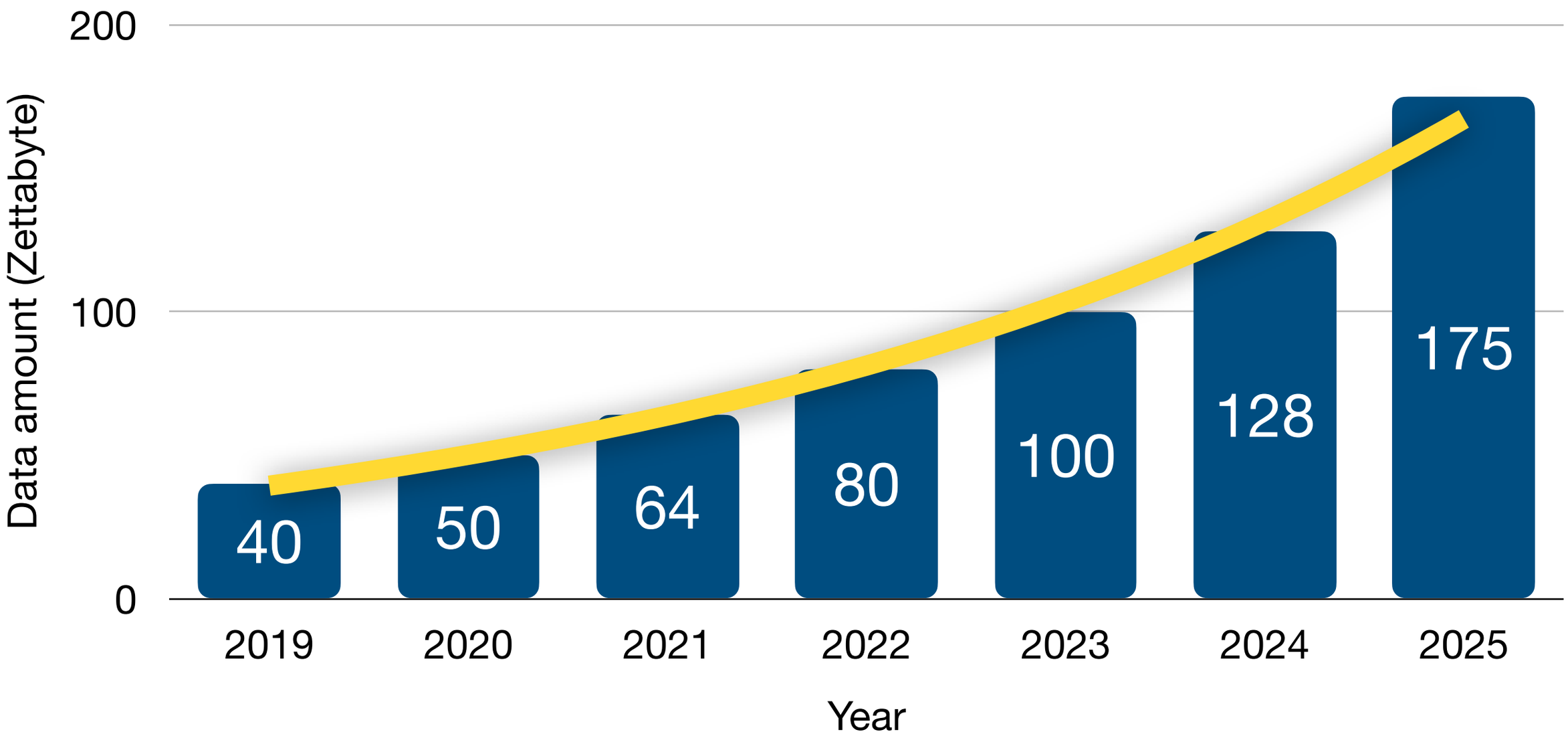
Sir Ronald Fisher

	back then	today
amount of data	small	yuuge
dimensionality	low	high
research design	mostly experiment	mostly observation
domains	agriculture	diverse
velocity	slow	fast

Fisher, R. A. (1992). Statistical Methods for Research Workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 66–70). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_6

Today's big is tomorrow's small

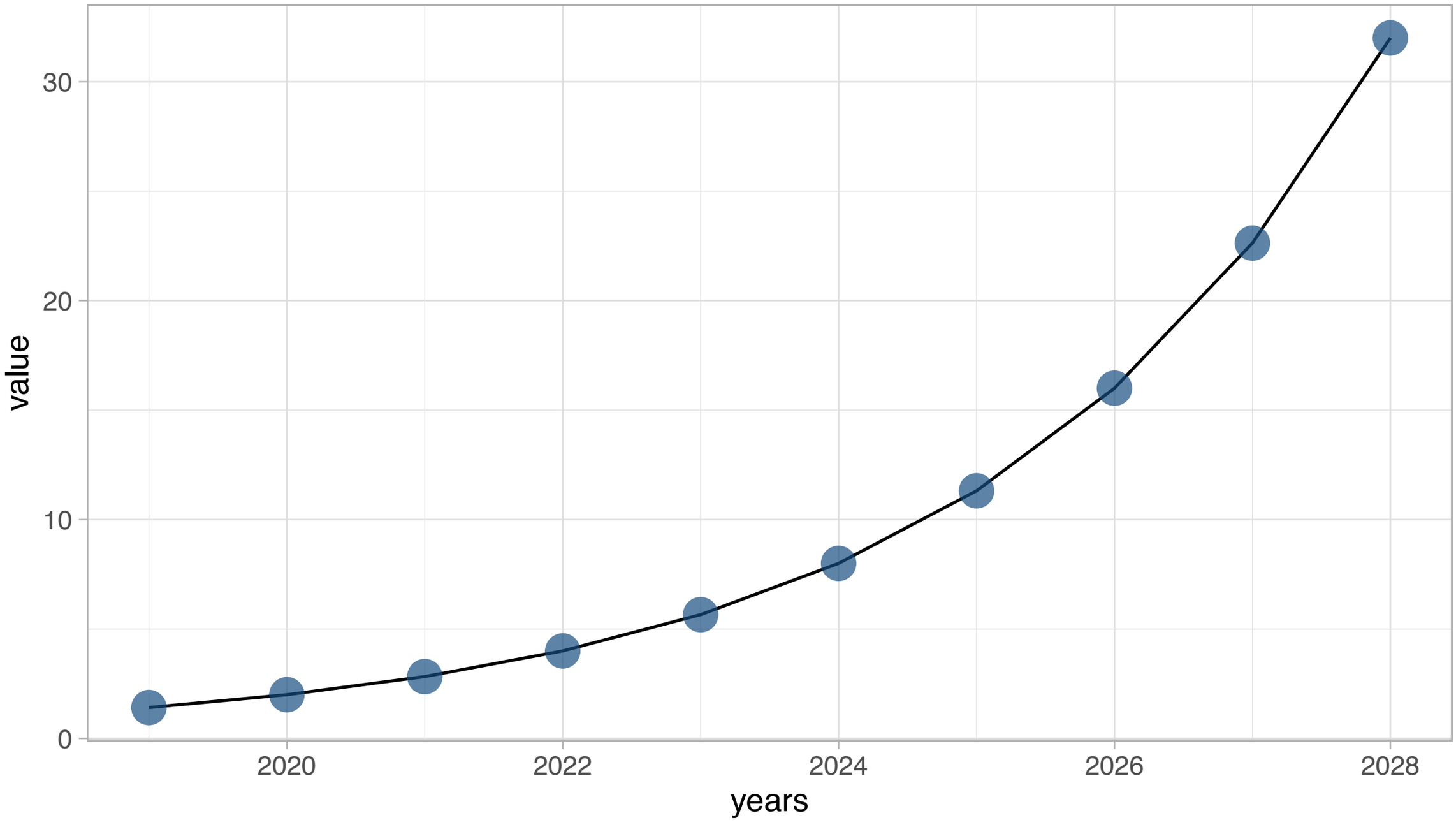
A. **V**elocity



Reinsel, D., Gantz, J., & Rydning, J. (2018). *Data Age 2025—The Digitization of the World. From Edge to Core* (#US44413318; IDC White Paper). https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Datasphere doubles every second year

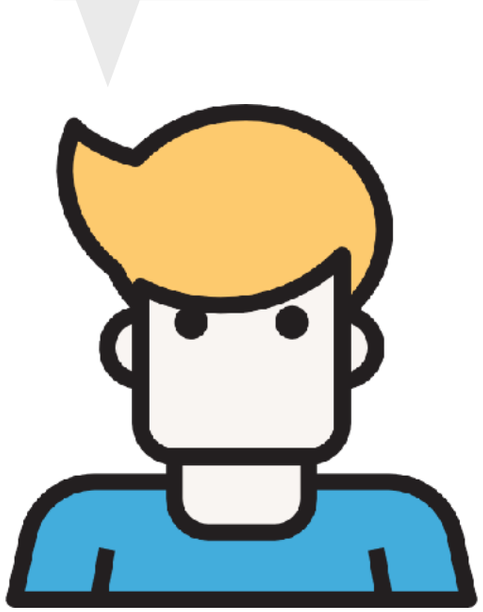
A. **V**elocity



https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/

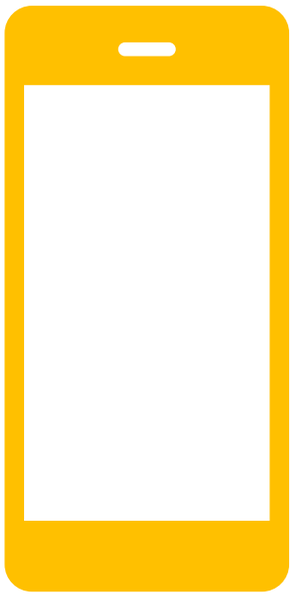
Quiz on growth processes

Stop right here. I need to check if we are on the same page.



Don

Get your phone.



Go to menti.com

Enter the PIN given.

or hit this link.

Growth process: Growing by a constant factor

- … is called **exponential growth**

Let me explain:

Wolfi

- Each year means „half“ (½) doubling:

$$v = 2^{\frac{1}{2}y}$$

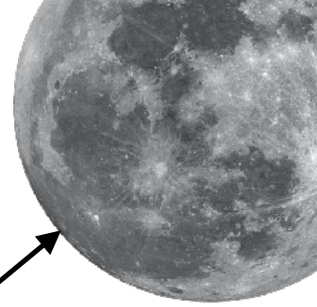
- Each exponential growth can be expressed via the ***e*** constant:

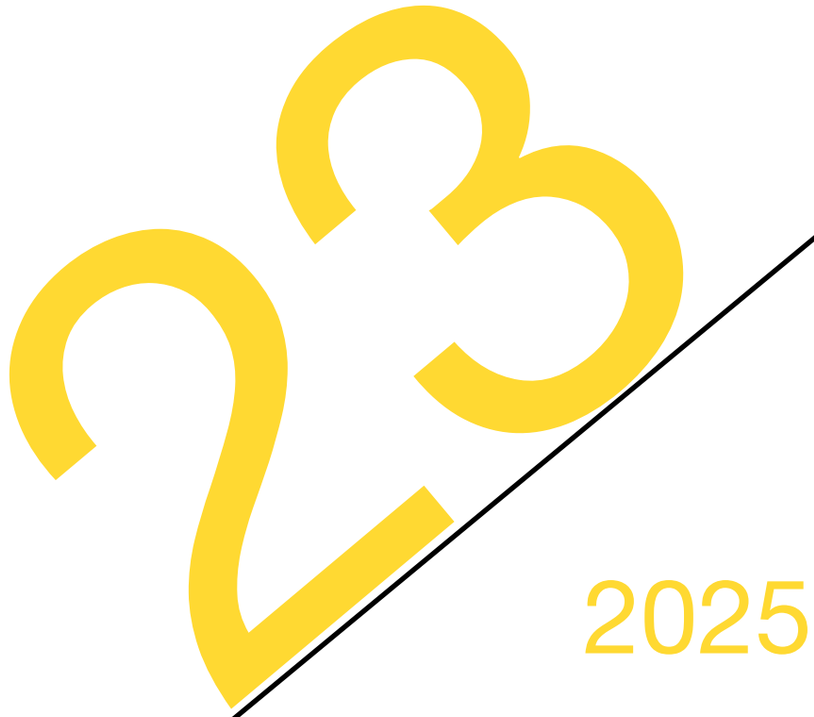
$$v = 2^{\frac{1}{2}y} = e^{(ln2)\cdot\frac{1}{2}y}$$

https://betterexplained.com/articles/understanding-discrete-vs-continuous-growth/

175 Zettabyte: yuuuge pile of DVD

B. **V**olume





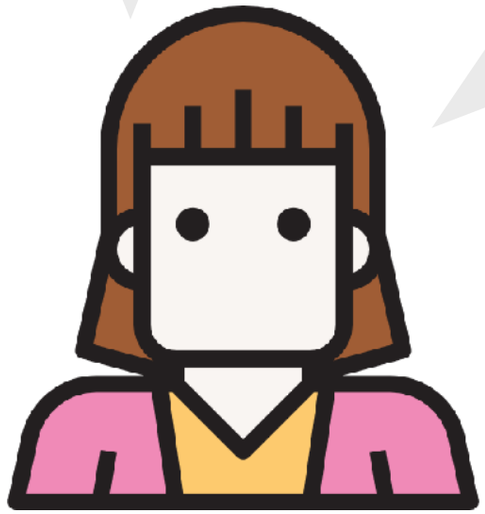
Bildquellen:
Mond: Mike Petrucci on Unsplash
Erde: NASA/Apollo 17 crew

Reinsel, D., Gantz, J., & Rydning, J. (2018). *Data Age 2025—The Digitization of the World. From Edge to Core* (#US44413318; IDC White Paper). https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

How fast does the datasphere grow?

Doubling every two years, that is exponential growth.

Let me crunch the data …

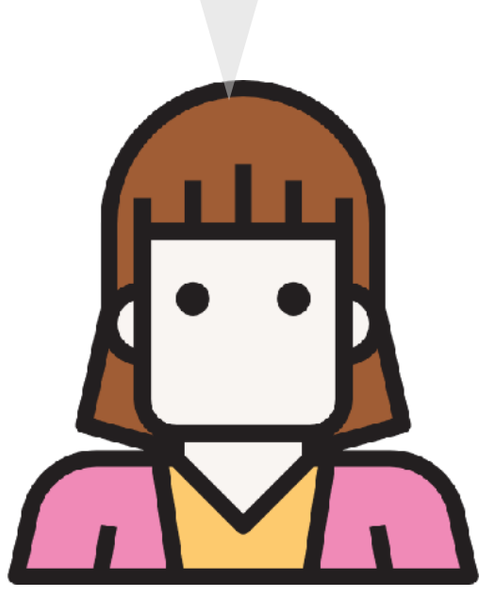


Angi

https://www.internetlivestats.com/

Exercise: Doubling data volumes

We begin with 1 unit of data. After two years we have 2, after 4 yours 4 and so on.



Angi

step	amount
0	**1**
1	2
2	4
3	...
4	...
5	?
6	
7	
8	
9	
10	
32	
64	
128	
256	

Exercise: Doubling data volumes

Being familiar with the doubling of two is kinda basis of data science.

Wolfi

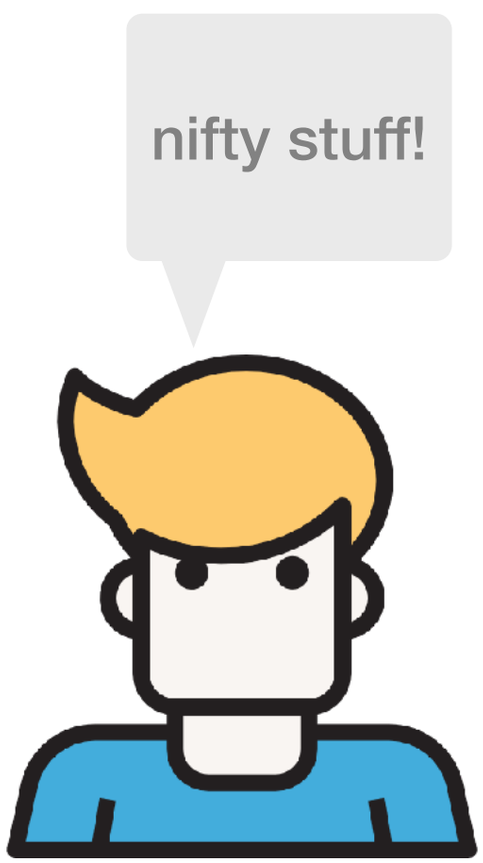
I know the first few, afterwards I take my calculator.

step	amount
0	**1**
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024
32	≈4 Mrd
64	$\approx 10^{18}$
128	$\approx 10^{36}$🥴
256	🤯🤯🤯

Big

$$2^{256} =$$

115792089237316195423570985008687907853269984665640564039457584007913129639936



Building intuition on „big“

> There was once a king in India who was a big chess enthusiast and had the habit of challenging wise visitors to a game of chess. One day a traveling sage was challenged by the king. The sage having played this game all his life all the time with people all over the world gladly accepted the Kings challenge. To motivate his opponent the king offered any reward that the sage could name. The sage modestly asked just for a few grains of rice in the following manner: the king was to put a single grain of rice on the first chess square and double it on every consequent one. The king accepted the sage’s request.
>
> The king lost.

What’s the amount of rice the king owes to the sage?

- Klick here to submit your answer.

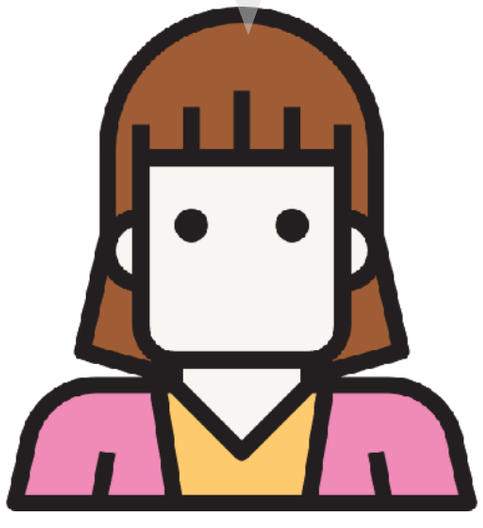
http://www-hm.ma.tum.de/ws1213/lba1/erg/erg07.pdf, https://medium.com/@purposefocuscommitment/the-rice-and-the-chess-board-story-the-power-of-exponential-growth-b1f7bd70aaca

What means „10 to the 3rd power“?

You should know that and that is simply …

Wolfi

Wolfi, wait! Let the participants explain!



Angi

„10 to the 3rd power“ equals 1,000

$$10 = 10^1$$

$$100 = 10^2$$

$$1000 = 10^3$$

6 is the exponent

$$1{,}000{,}000 = 10^6$$

6 zeros/places after the leading 1

Ask your 7th grad math teacher.

Estimating 2^{64} as decimal number

$$2^{64} = ?$$

$2^{10} \approx 10^3$

$2^{20} = 2^{10} \cdot 2^{10} \approx 10^3 \cdot 10^3 = 10^6$

$2^{40} = 2^{20} \cdot 2^{20} \approx 10^6 \cdot 10^6 = 10^{12}$

$2^{60} = 2^{40} \cdot 2^{20} \approx 10^{12} \cdot 10^6 = 10^{18}$

1,000,000,000,000,000,000

18 zeros/places after the leading 1

https://web2.0calc.com/

2^{256} – rough estimate: 1 follows by 72 zeros

$$2^{256} \approx 10^{72}$$

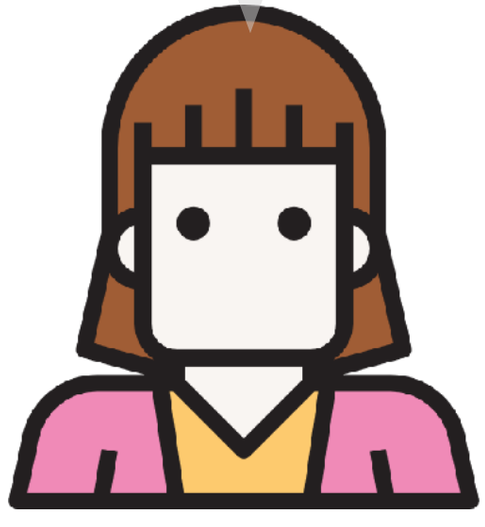
Wolfi running wild

Now you need to to the *e*-function. That's the most important number in the universe.

Wolfi

e is the growth number. Let me explain, it's fascinating …

Wolfi, stop it!



Angi

https://betterexplained.com/articles/an-intuitive-guide-to-exponential-functions-e/

Wolfi dreaming …

https://www.youtube.com/watch?v=0fKBhvDjuy0&t=306s https://en.wikipedia.org/wiki/Power_of_10

Some context to „big“

fact	figure
Covid-19 death toll in Germany (as of 2020-09-17)	10^4
number of people in Germany	10^8
number of people in the world	10^{10}
number of neutrons in the brain (except Don’s)	10^{11}
numbers of stars in the universe	10^{23}
number of atoms in the universe	10^{80}

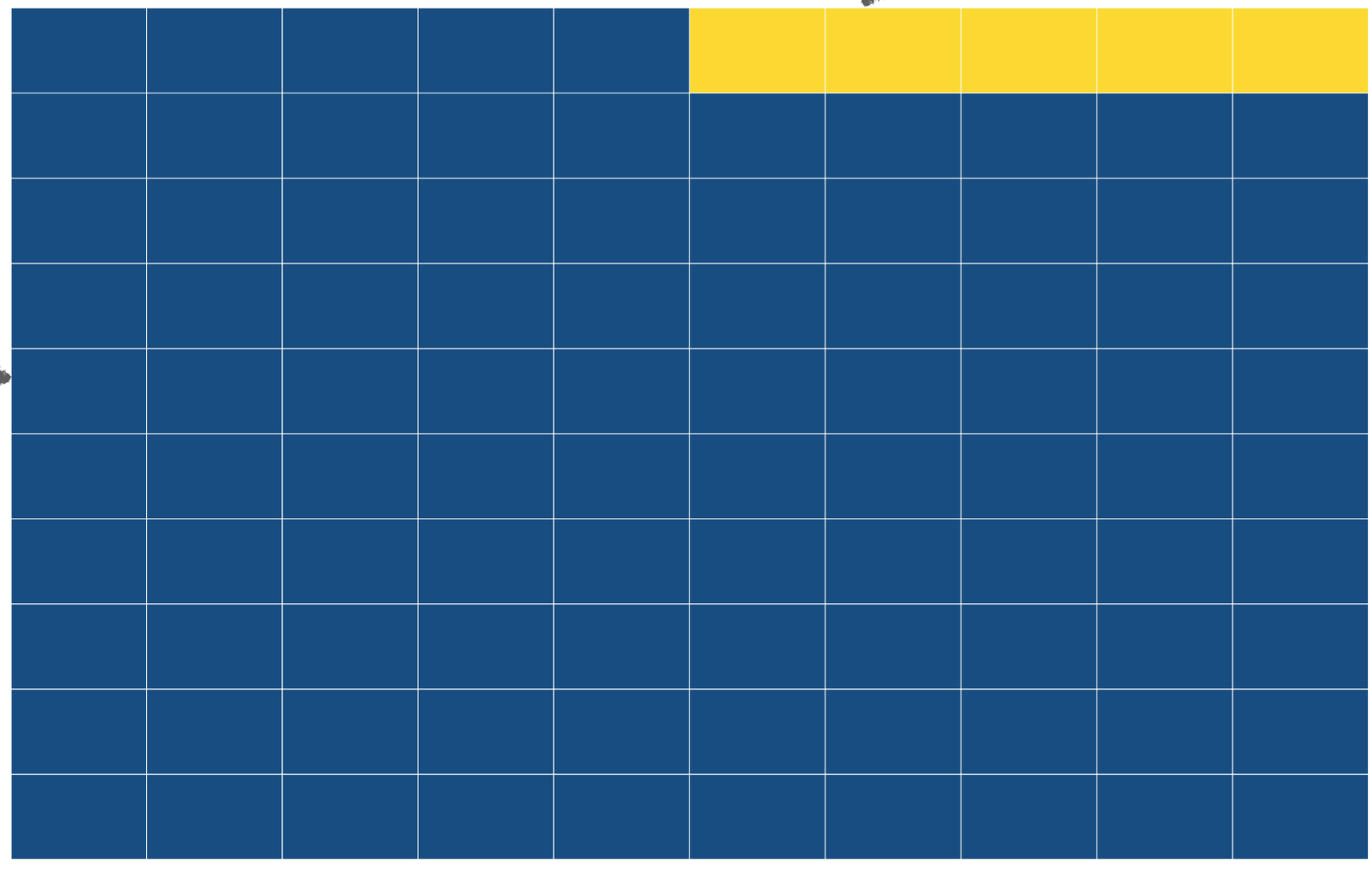
http://imgs.xkcd.com/comics/1000_times.png

Data tend to be unstructured

C. **V**ariety

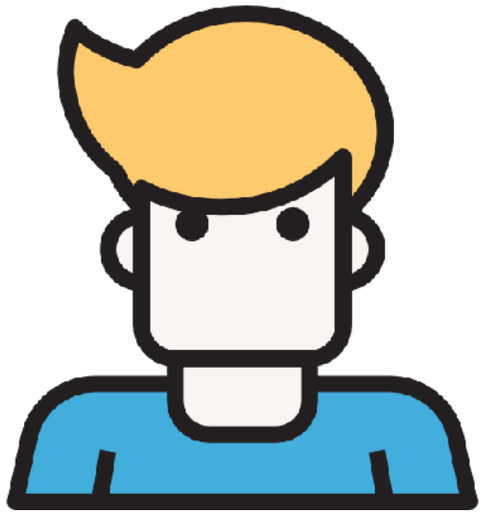
unstructured Data

rectangular data



Big data and Microsoft Excel

My dataset is 500 Megs of size. Excel won't load it!!



Don

That's not big data. Get some decent analytical software.

Wolfi

https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

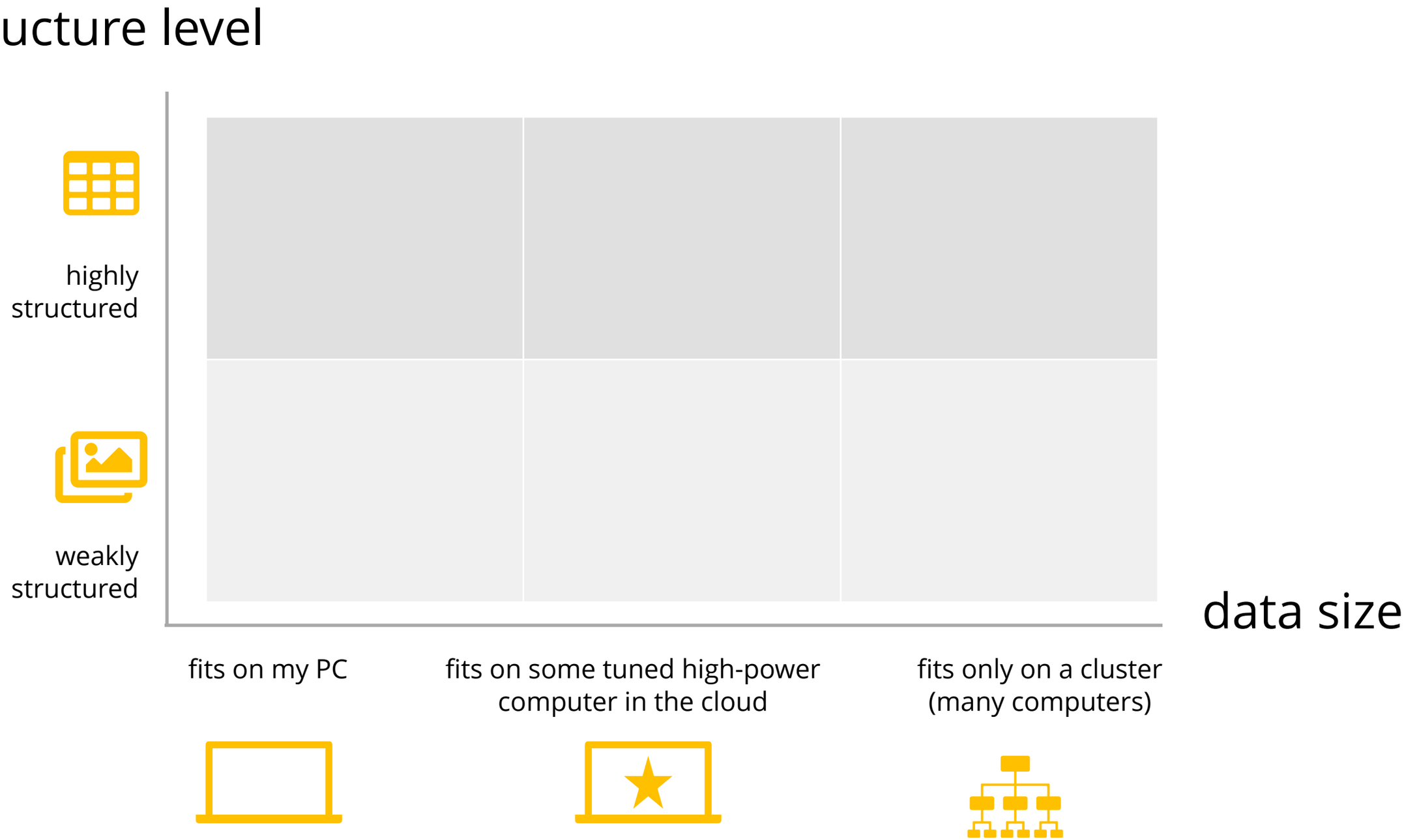
Analytical technologies as a function of data size

Size	adequate technology
some kilobytes	Get a pencil and a pocket calculator.
some megabytes	Each analytical software will swallow that amount.
some gigabytes	Get more RAM to your machine.
some hundreds gigabytes	Install simple SQL server on your machine.
smaller than 1 terabyte	Rent some supercomputer in the cloud.
some terabytes	Use Hadoop.

https://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

Pragmatic classification of your data formats

Place dots in this conceptboard. Discuss.

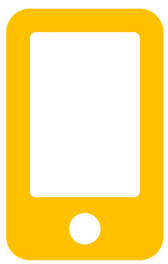


Recap – Big Data

- Big can be defined as „bigger than what fits on your PC“.
- The 3V’s are commonly used for defining big data.
- We need to change data analysis methods as the data have changed too.
- Data is growing veeery rapidly.
- It’s difficult to image bigness.

What made data big?

Some drivers of data growth



real-time data

machines, sensors

image and videos

https://www.bigdataframework.org/business-drivers-for-big-data/, https://www.absatzwirtschaft.de/big-data-analytics-im-marketing-13941/

Data formats used by your customers

- What data formats do your customers use?
- Example data formats include survey multiple choice, images, free text, click rates, ...
- Collect such data formats using this conceptboard

Big data = evil data?

Read up these stories for some spicy view on the dark side of big data.

Wolfi

Fhom, H. S. (2015). Big Data: Opportunities and Privacy Challenges. *arXiv:1502.00823 [cs]*. http://arxiv.org/abs/1502.00823

Case study: Predicting your personality

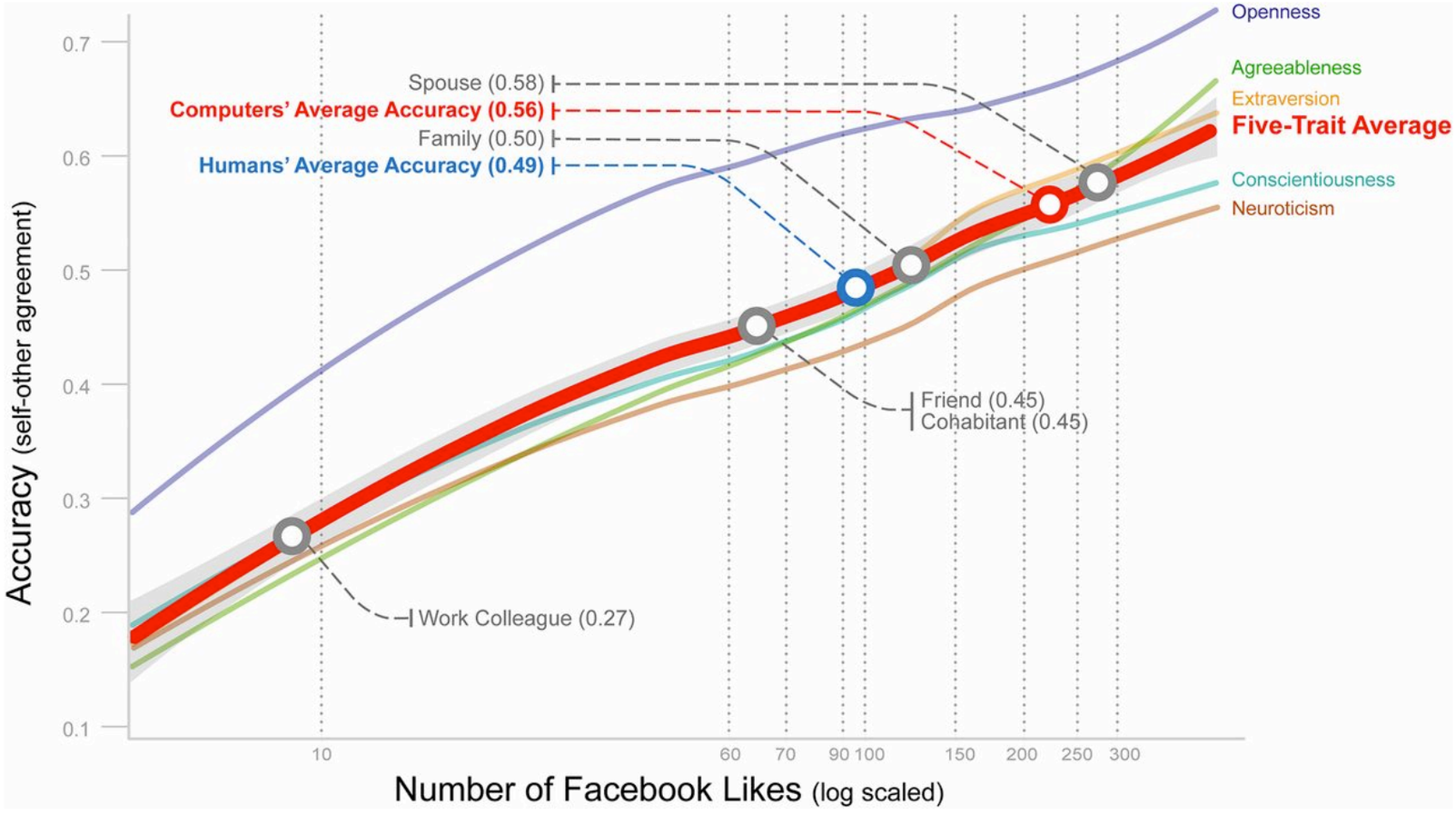
- The computer outperformed even close friends when it comes to predicting a personality, study finds.

amount of data

70000 individuals

analytic method

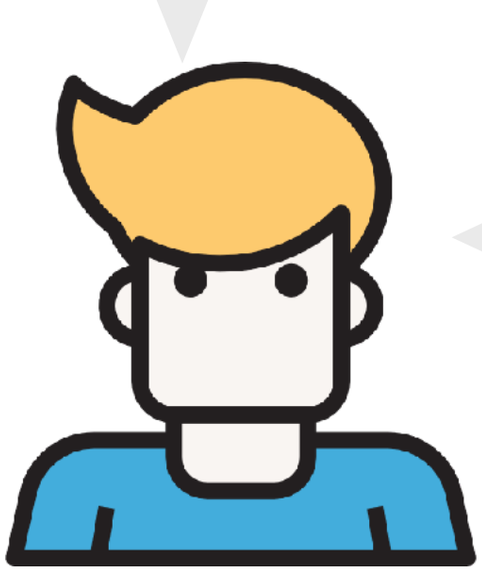
LASSO-Regression, cross validation



Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Reflecting on the power of the algorithms

Yeah! I got social media data of every human, so I'll send ads that match your personality.



Don

Is the machine capable of predicting sexual preferences? That could be nicely exploited.

It's too early to assess the quality of this method.

Wolfi

Welcome to the dark side

RESEARCH ARTICLE

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski, David Stillwell, and Thore Graepel

PNAS April 9, 2013 110 (15) 5802-5805; https://doi.org/10.1073/pnas.1218772110

> We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Recap – Drivers of Big Data

- Drivers of data growth include real-time, mobile data, sensors, images, videos and more.
- Big data technologies raise data privacy issues.
- Big data technologies can possibly exploited by bad intentions.
- It’s too early to evaluate whether the powers and promises are real or overhyped.

Big Data – Buzzwords

Buzzwords

Statistics

Data Science

Data Mining

Machine Learning

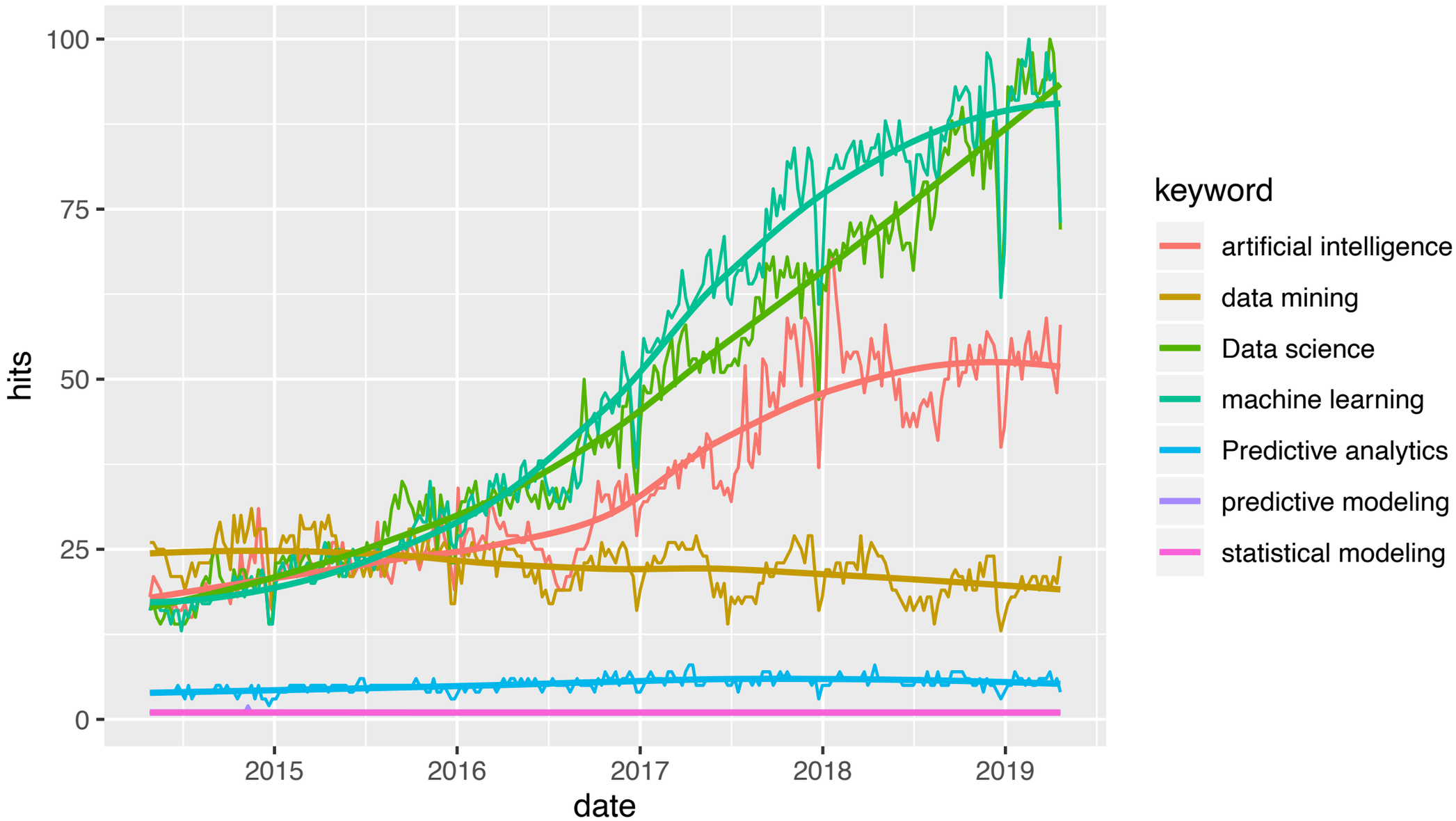
AI*

Deep Learning

*Artificial Intelligence

Data Science emerged as the new sexy term

Google Search hits to big data buzzwords



Google Search term analysis, 2019-05

Statistics

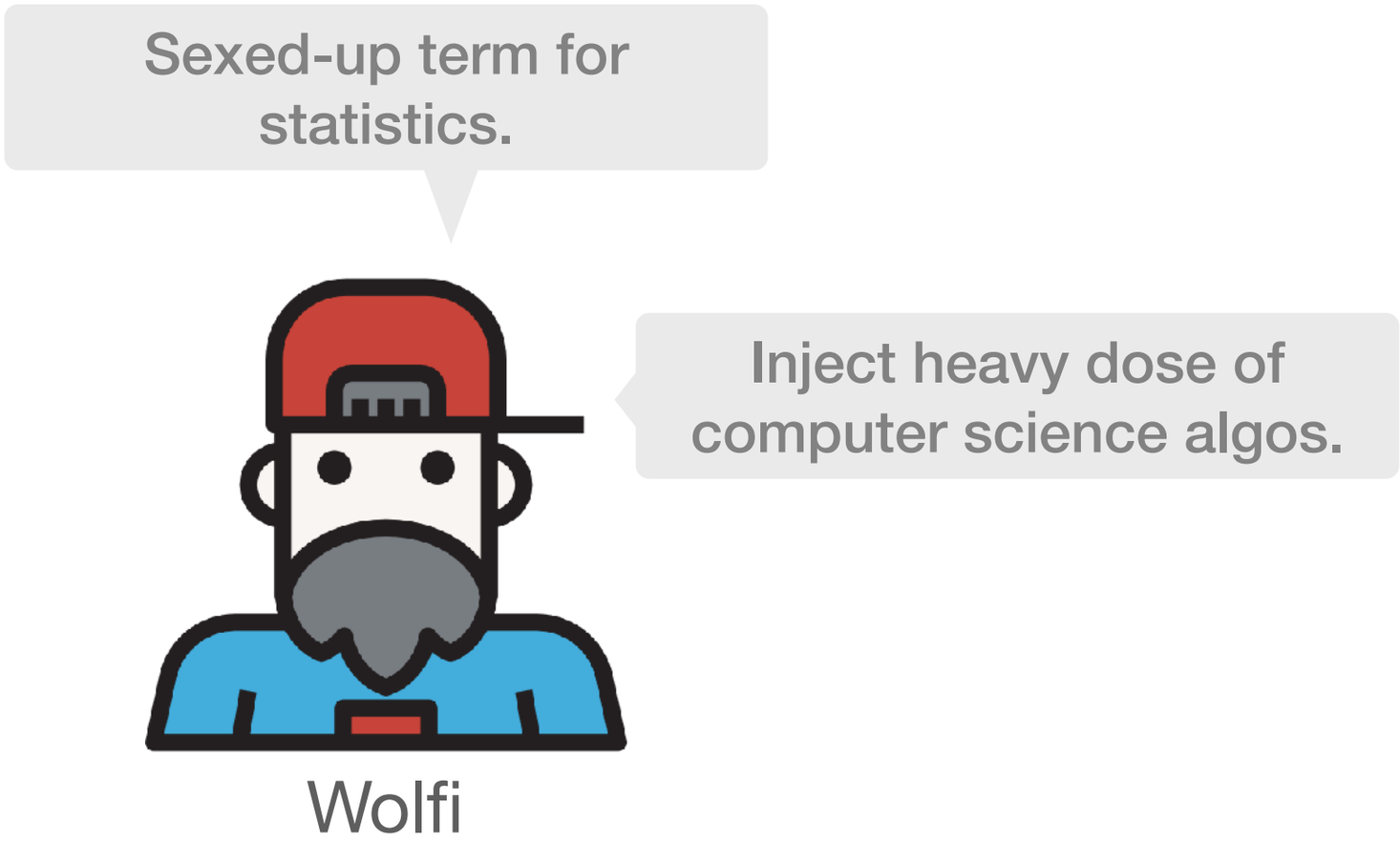
Science of uncertainty.

Think probability calculus.

Wolfi

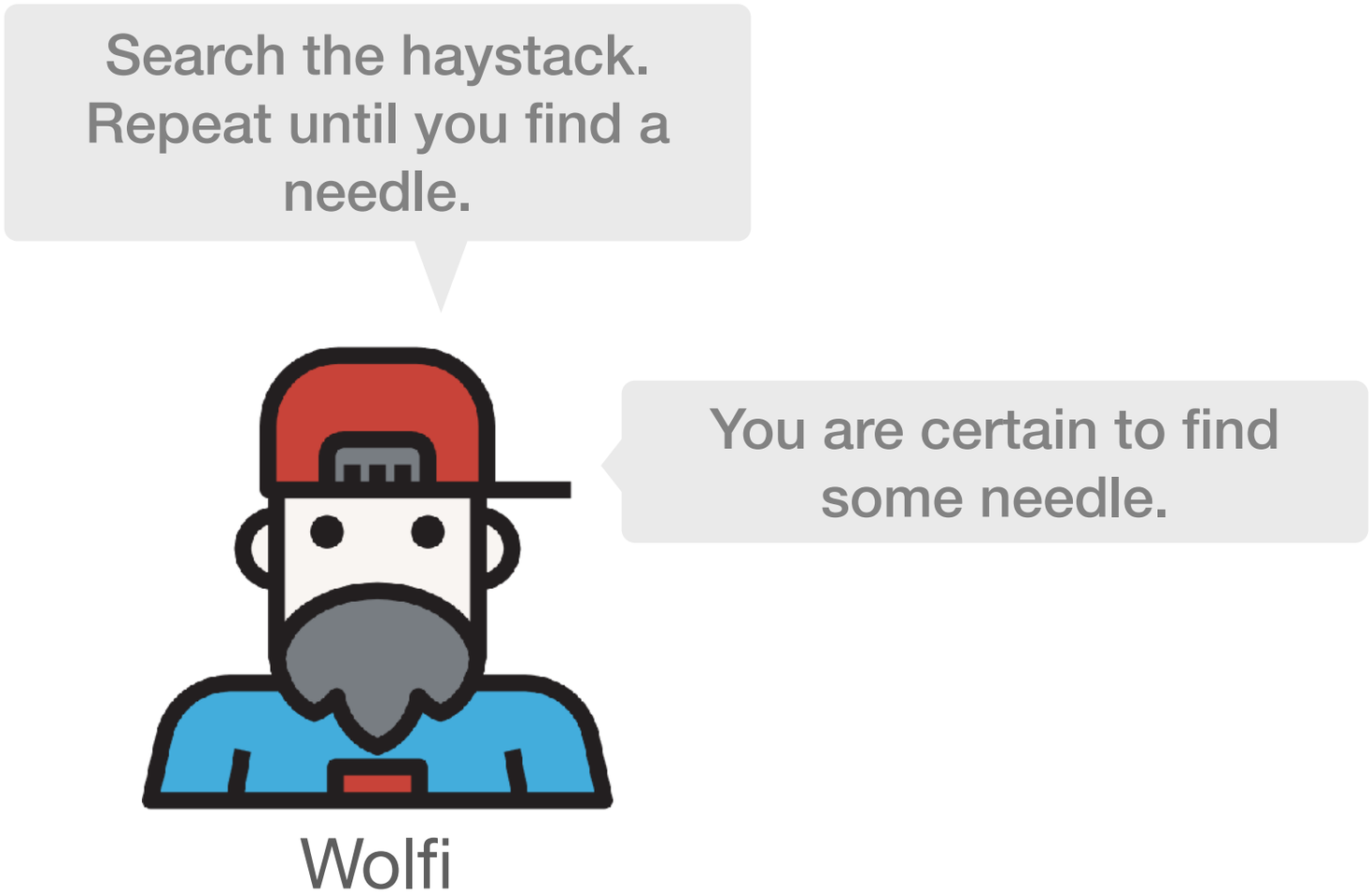
Spiegelhalter, D. (2019). *The art of statistics: Learning from data*. Penguin UK. https://www.penguin.com.au/books/the-art-of-statistics-9780241258767

Data Science



Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2017). Modern Data Science with R (Chapman & Hall/CRC Texts in Statistical Science). Chapman and Hall/CRC., Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren* (1. Auflage 2019). Springer. https://www.springer.com/de/book/9783658215866

Data Mining



Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (Bd. 54). https://dx.doi.org/10.1007/978-3-642-19721-5

Machine Learning

Machine Learning

The cool bit is that the machine is not explicitly programmed but learns from data.

Wolfi

Artificial Intelligence (AI)

AI

Catch-all term for the latest smart technology such as speed recognition, autonomous driving etc.

Wolfi

Deep Learning (artificial neural networks)

Deep Learning

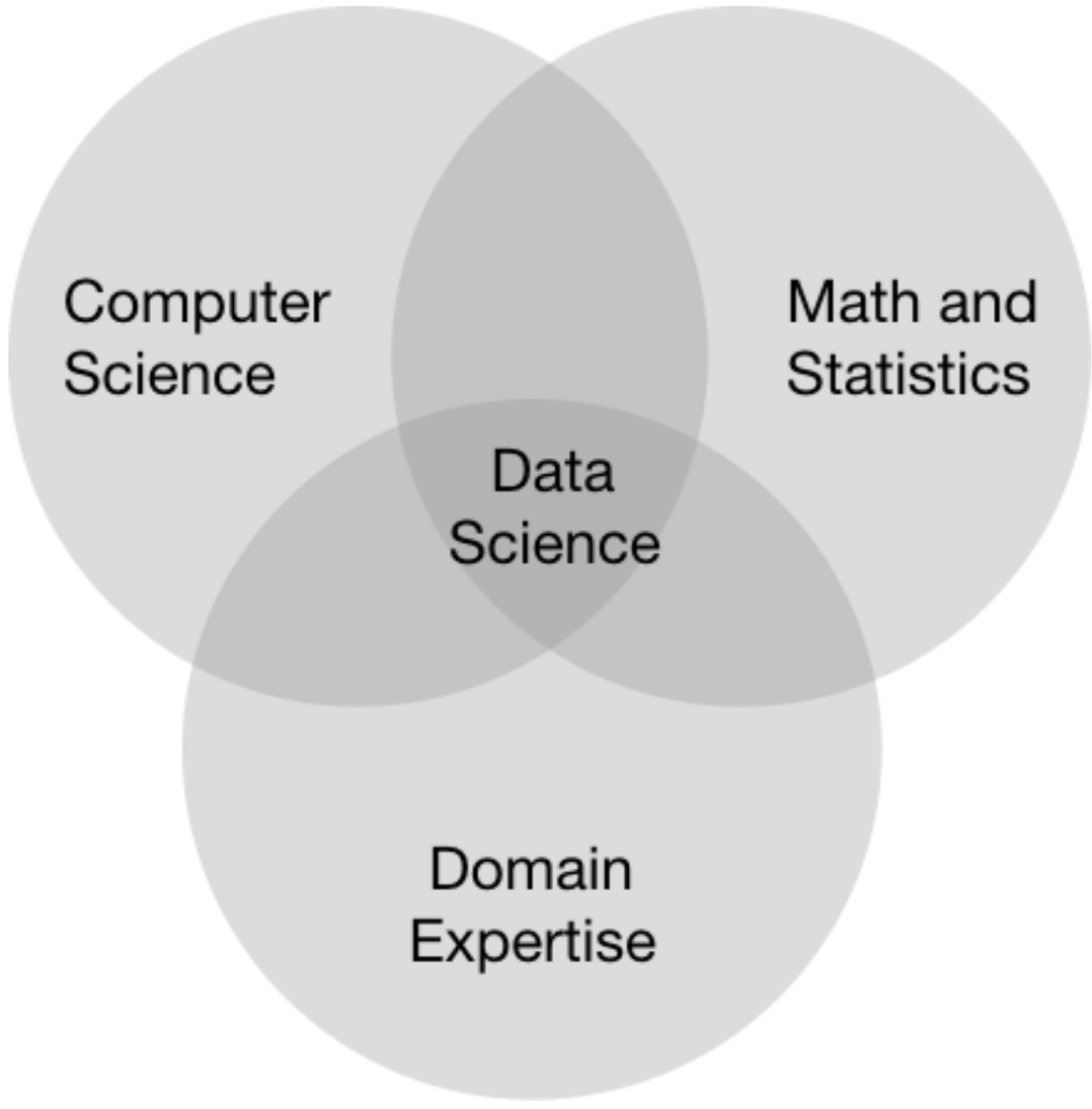
A certain algorithm of machine learning. Powerful in some domains.

Wolfi

(Buzz)Words – overview

- Statistics: Science of measuring uncertainty. Probability theory is pivotal.
- Data Science: Applied statistics drawing heavily on computers. Less probability, more algorithms.
- Data Mining. Searching large amounts of data for some useful patterns.
- Machine Learning: Teaching machines (computers) to learn patterns from data instead of programming hard-wired rules to the machine.
- Artificial intelligence: If a machine routinely completes a task which would need intelligence if accomplished by a human, some kind of AI is said to be present.
- Deep Learning: Class of algorithms that uses multiple layers to progressively extract higher level features from raw input.

More on Data Science

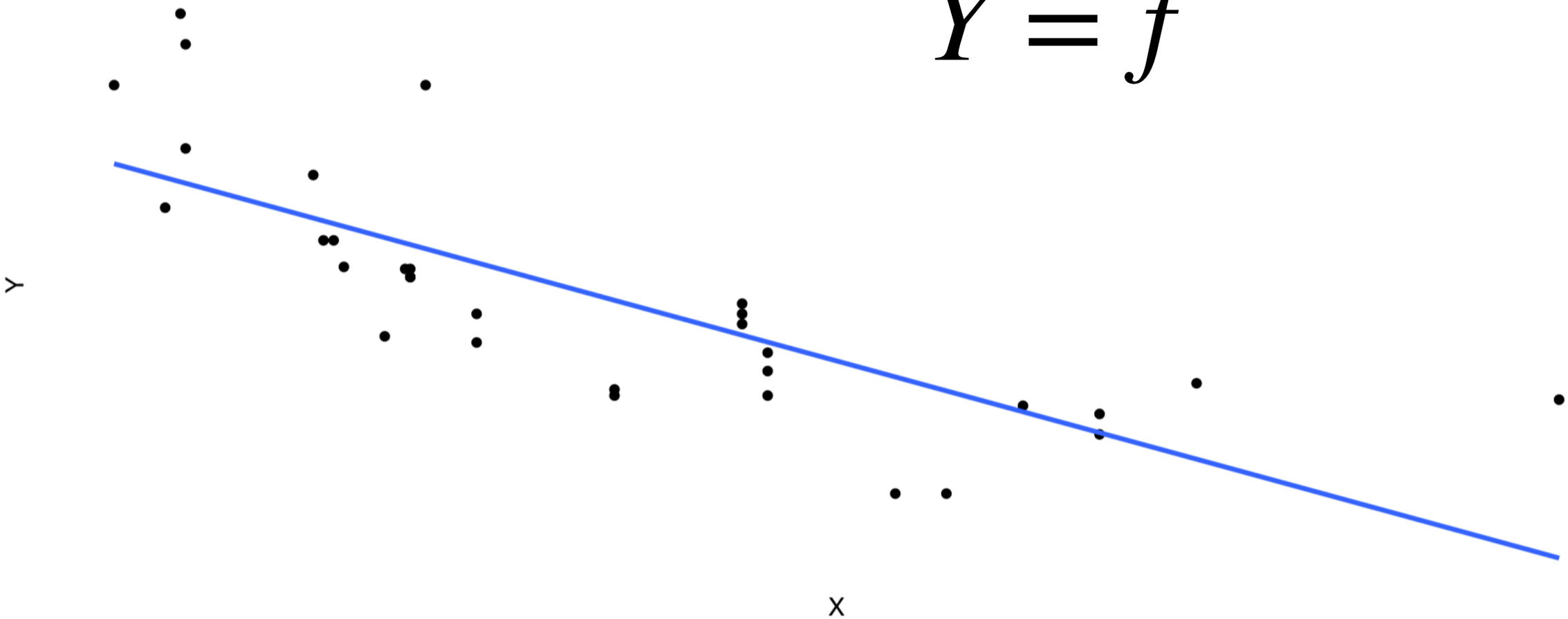


http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram but see https://towardsdatascience.com/why-the-data-science-venn-diagram-is-misleading-16751f852063

Common ground: Quantifying patterns in data

$$Y = f(x) + \epsilon$$

$$\hat{Y} = \hat{f}$$

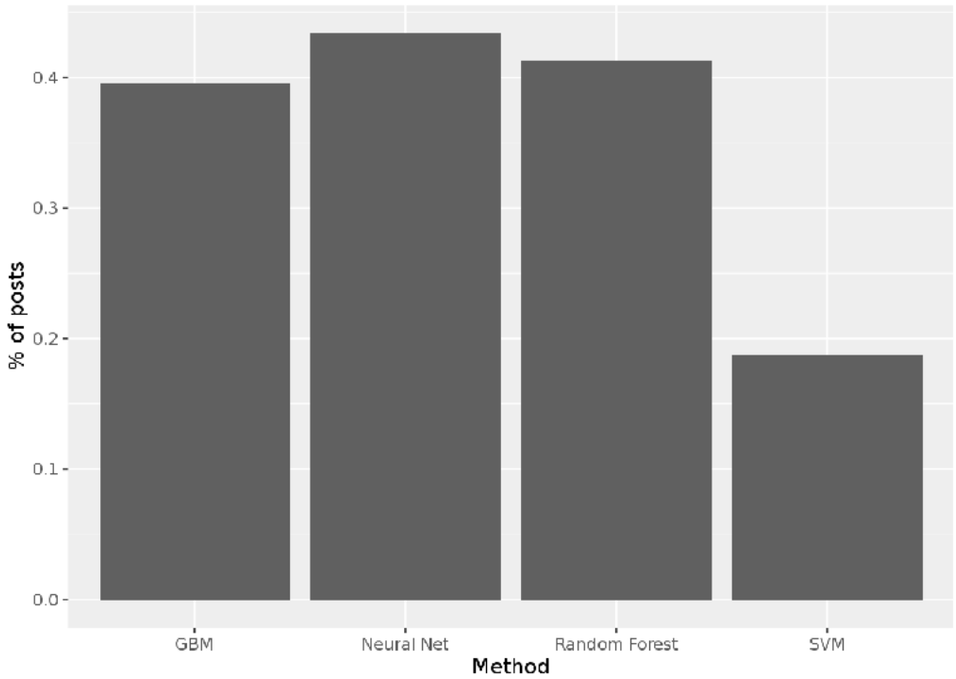


James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

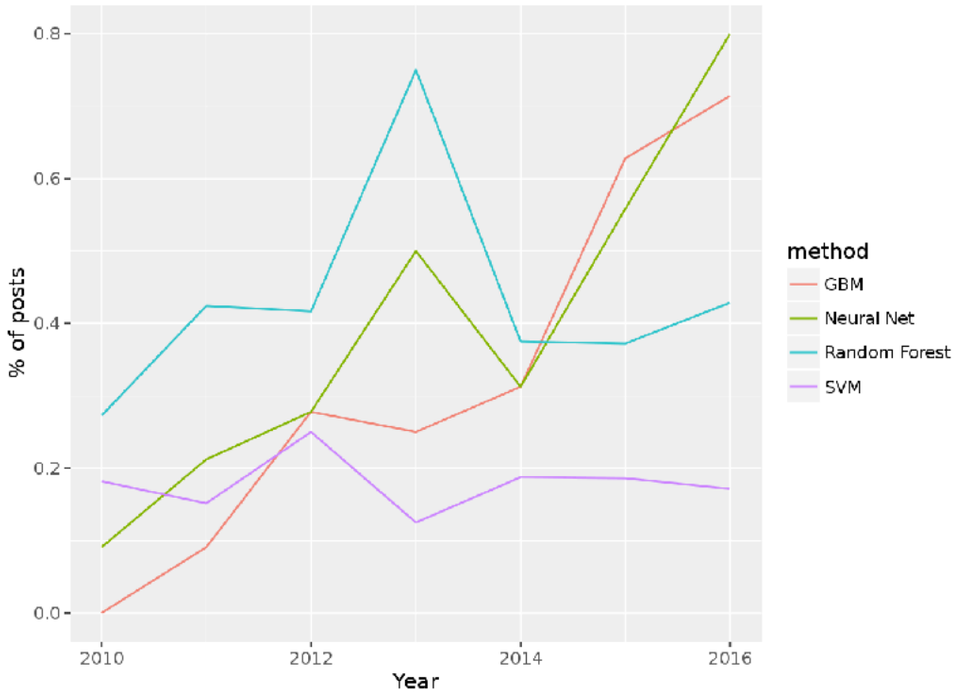
The best model/method/approach ...

What algos are most successful on kaggle.com?

- Neural nets are among the top scoring algos.



- And they appear to stay on the rise.

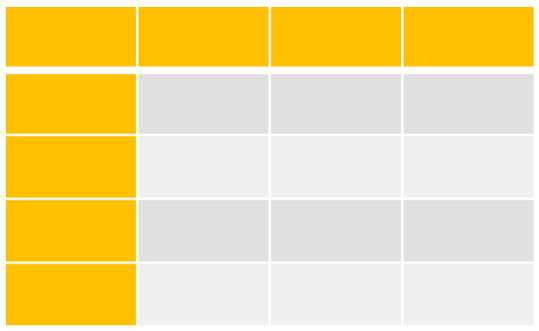


https://en.wikipedia.org/wiki/Kaggle, https://www.kaggle.com/msjgriffiths/r-what-algorithms-are-most-successful-on-kaggle/report

Are neural nets always the best choice?

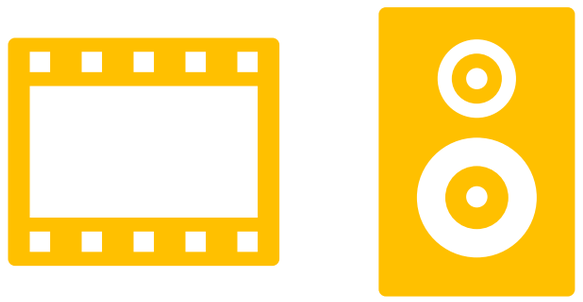
- According to Kaggle's CEO and founder, Anthony Goldbloom, there are two winning approaches:

rectangular data



decision trees ensembles such as random forest or boosting

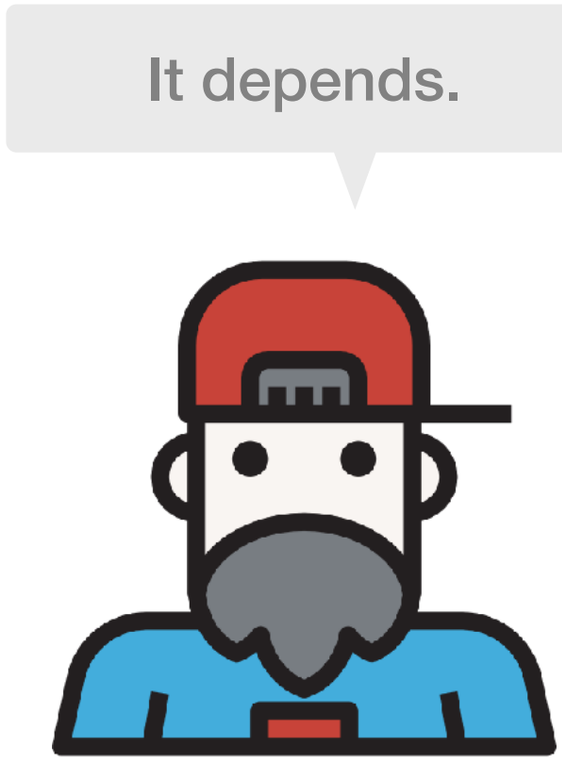
image or speech data



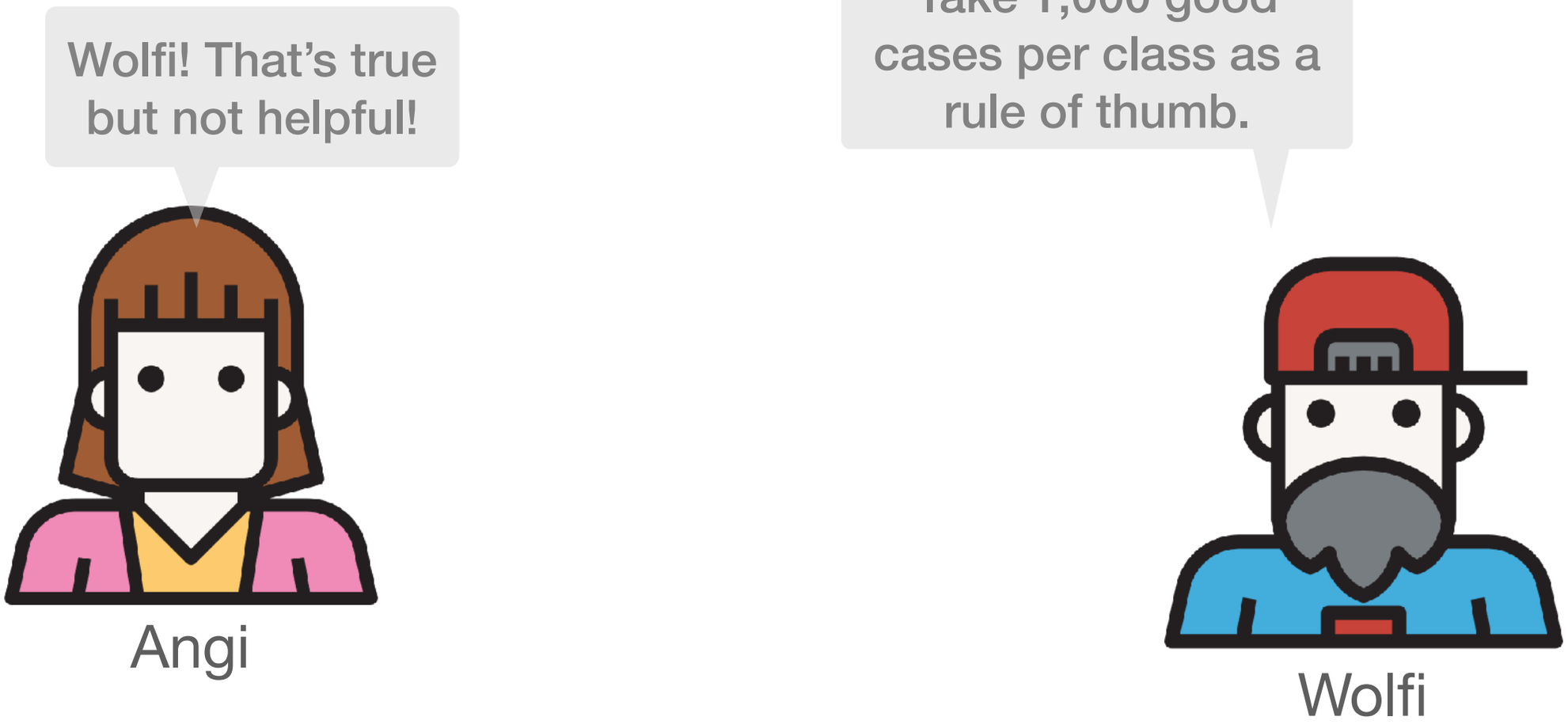
deep learning (aka, neural nets)

https://www.import.io/post/how-to-win-a-kaggle-competition/

How large should my sample be?

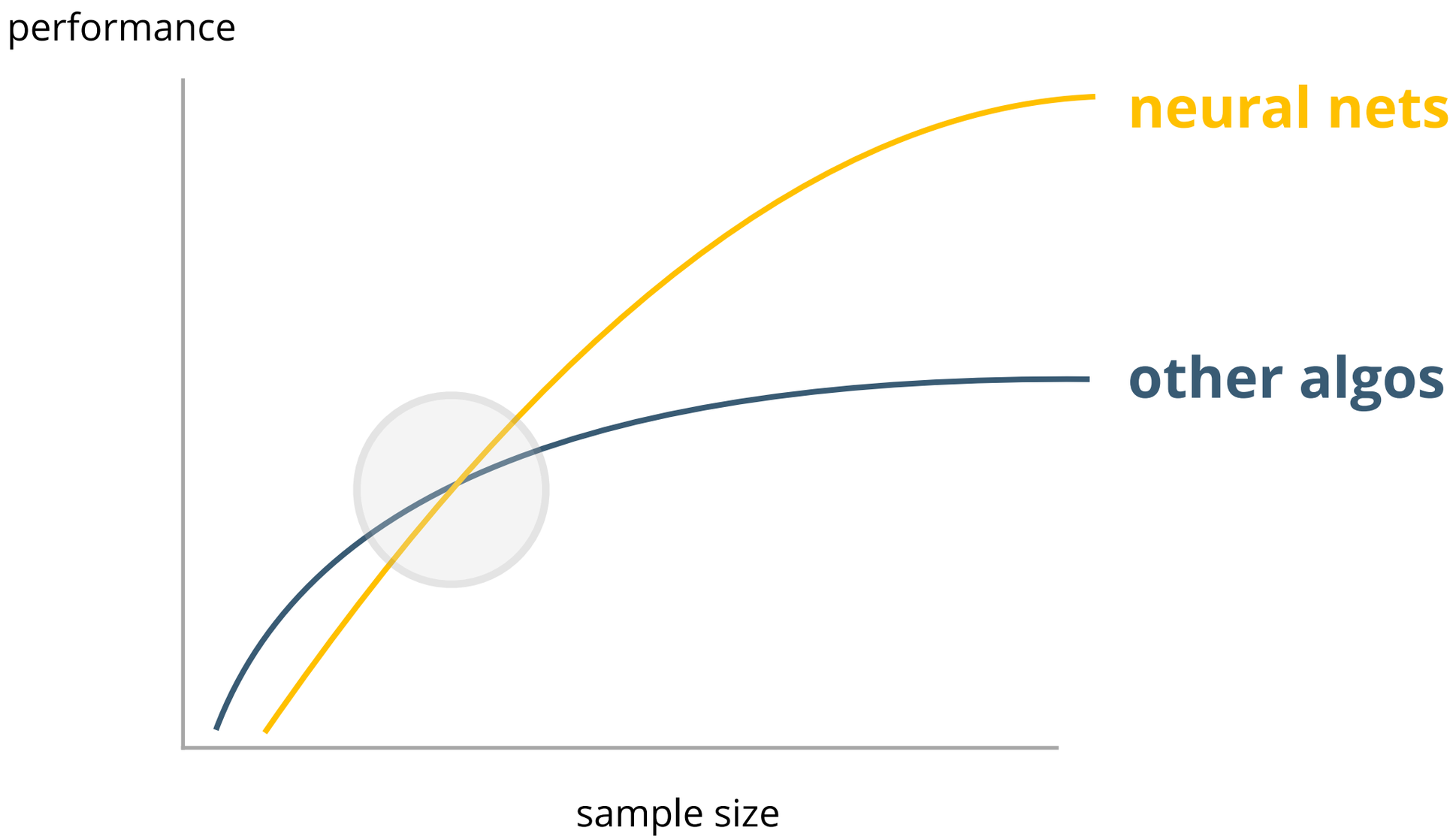


How large should my sample be? 2/2



https://petewarden.com/2017/12/14/how-many-images-do-you-need-to-train-a-neural-network/

What about neural nets and sample size?



https://indatalabs.com/blog/trends-computer-vision-software-market?cli_action=1555888112.716, https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063

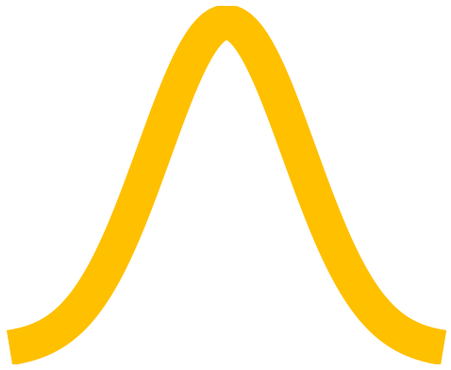
Recap – Buzzwords

- All those shiny terms are in effect just different perspectives on modern data analysis.
- The core principle is to find patterns in mass data.
- There’s no single best approach/method.
- Tree-based models work well for rectangular data.
- Neural nets work well for less structured data such as images or speech.
- Neural nets can better exploit large data corpora.

The (two) cultures of data analysis

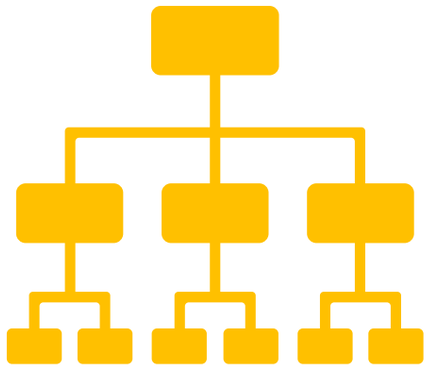
Breiman's two cultures

stochastic modelling



building on probability theory

algorithmic modelling



building on computer science models

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–215. https://projecteuclid.org/euclid.ss/1009213726

Data Science vs. Statistician

Created by H Alberto Gongora
from Noun Project

Created by Alfonso López-Sanz
from Noun Project

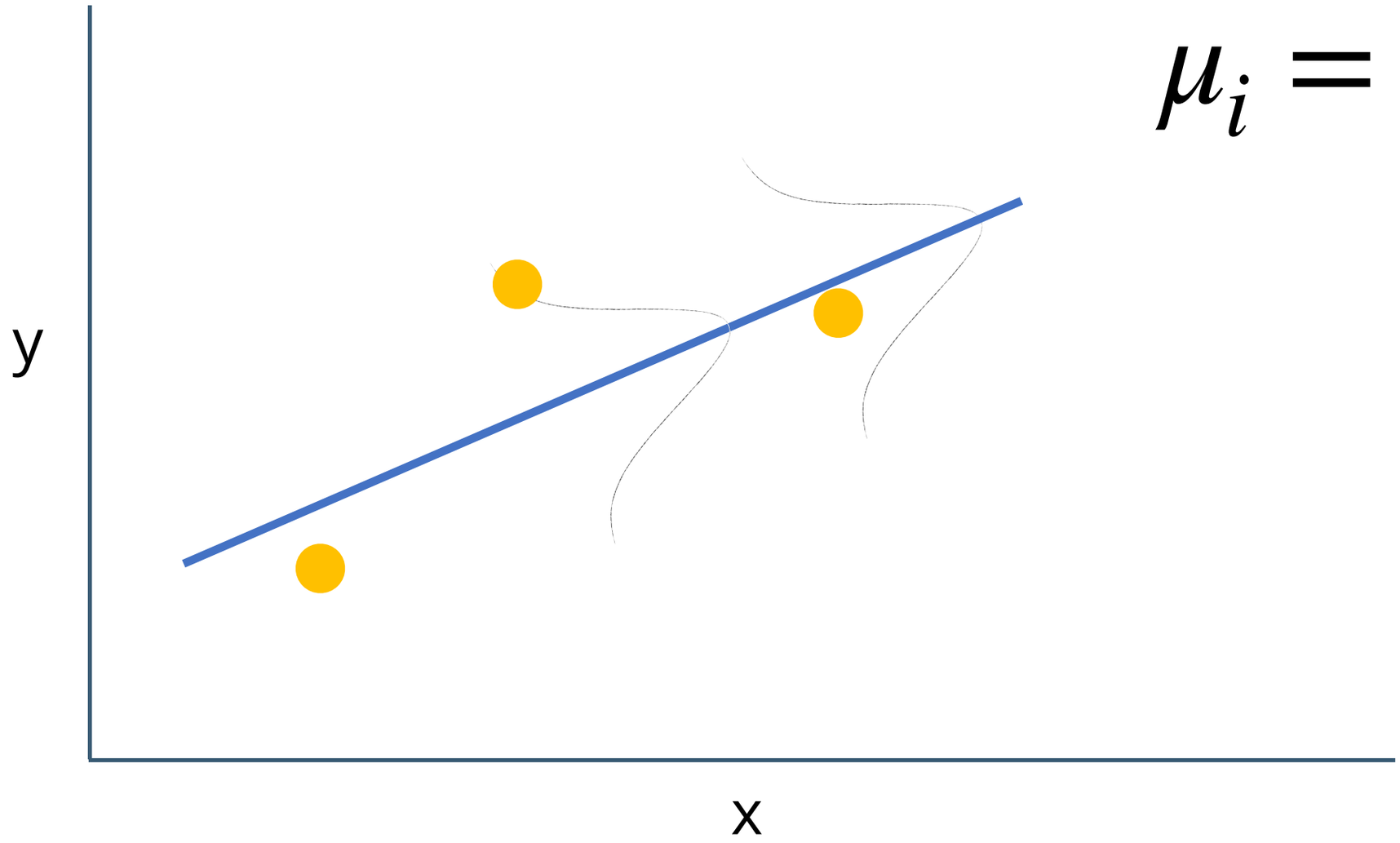
- Watch the video
- Report back the (alleged) characteristics of a data scientist vs. a statistician

https://www.youtube.com/watch?v=uHGlCi9jOWY

Culture 1: Classical statistics

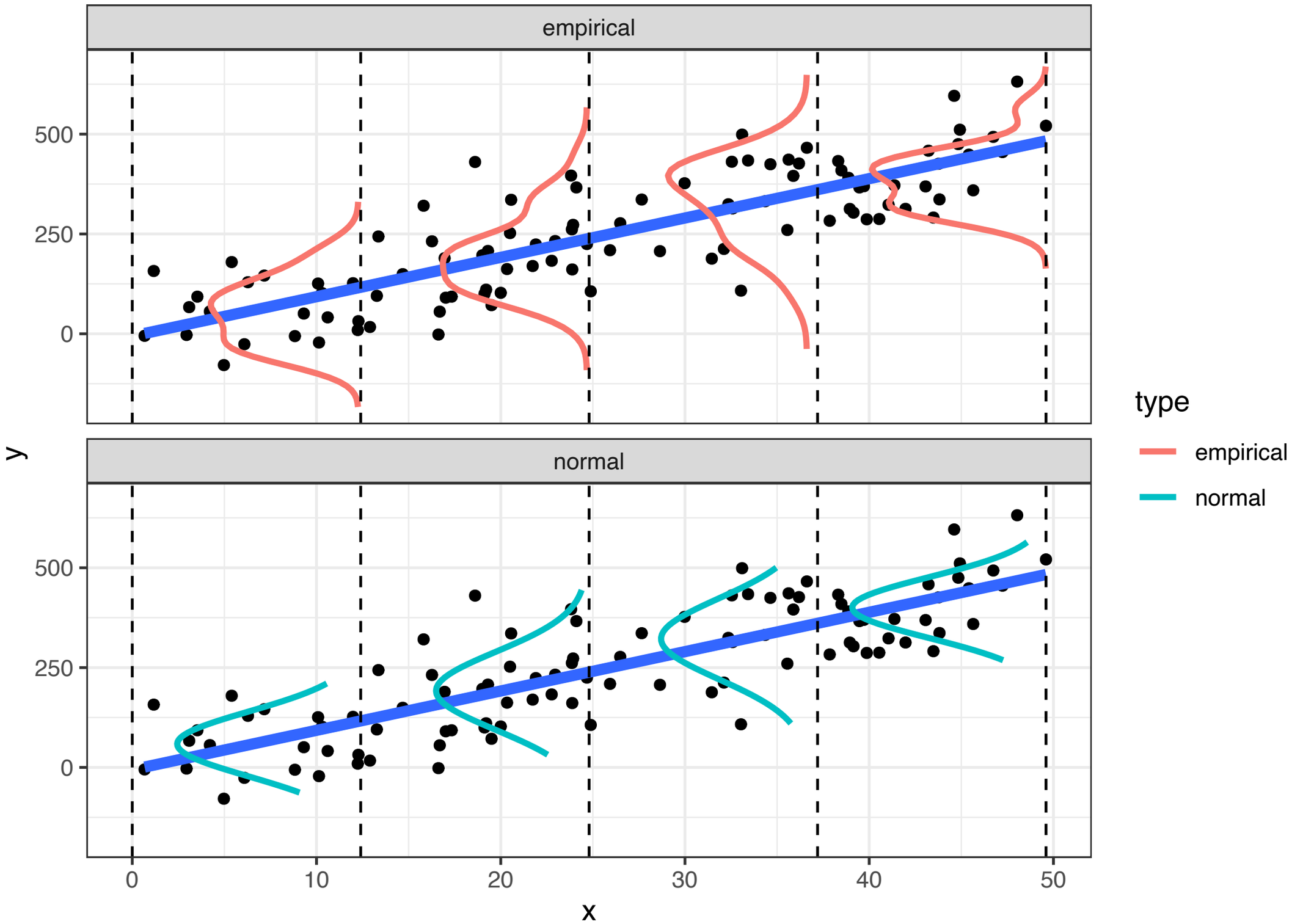
$$y_i = N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$



https://medium.com/quick-code/maximum-likelihood-estimation-for-regression-65f9c99f815d

Stochastic model



https://stackoverflow.com/questions/31794876/ggplot2-how-to-curve-small-gaussian-densities-on-a-regression-line

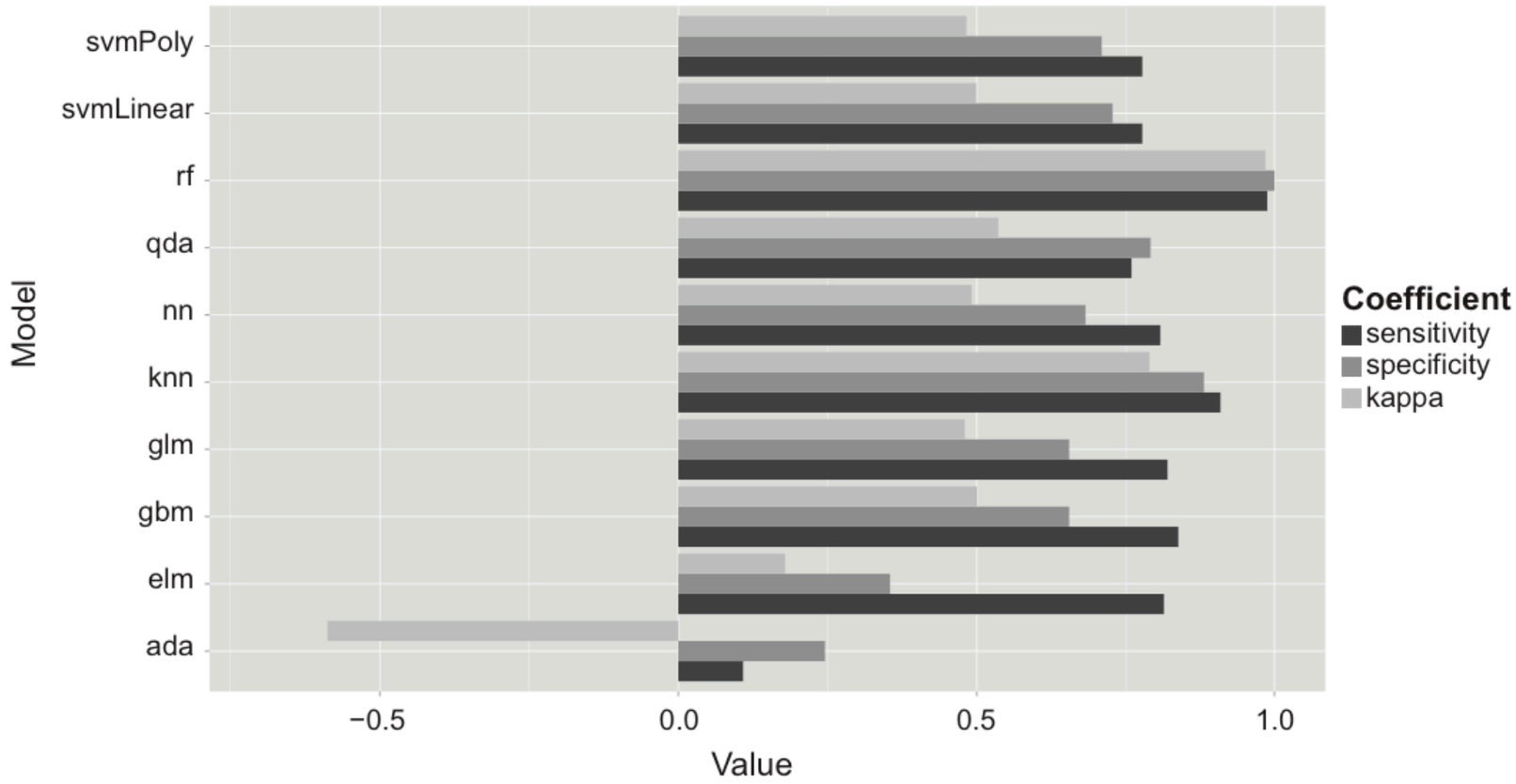
Case study: Regression in real live

amount of data

276 individuals,
14 survey items

analytic method

10 machine learning algos

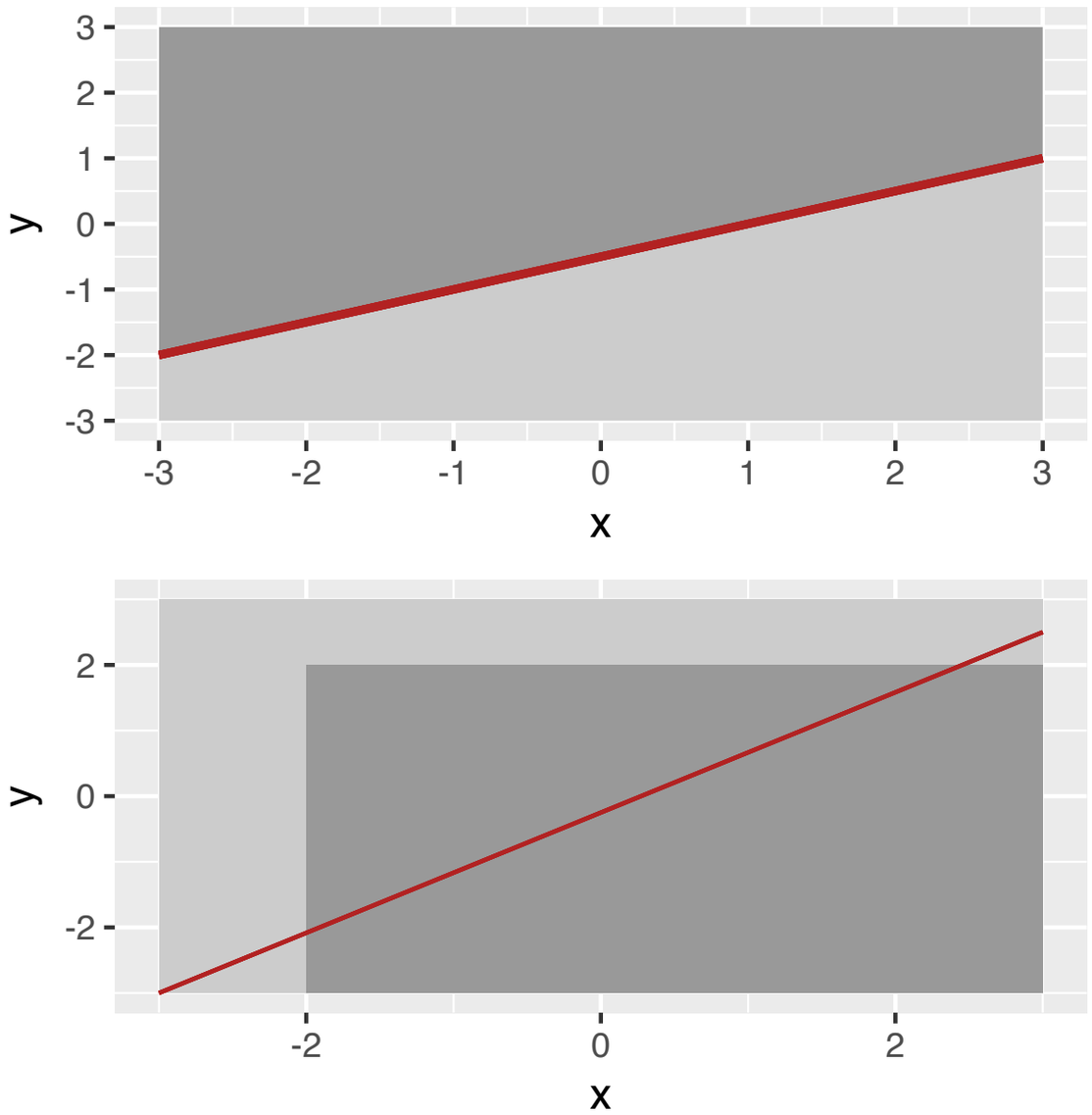


Machine learning provided higher predictive accuracy …

… compared to classical regression

Sauer, S., Buettner, R., Heidenreich, T., Lemke, J., Berg, C., & Kurz, C. (2018). Mindful Machine Learning. *European Journal of Psychological Assessment*, *34*, 6–13. http://dx.doi.org/10.1027/1015-5759/a000312

Again, there's no free lunch



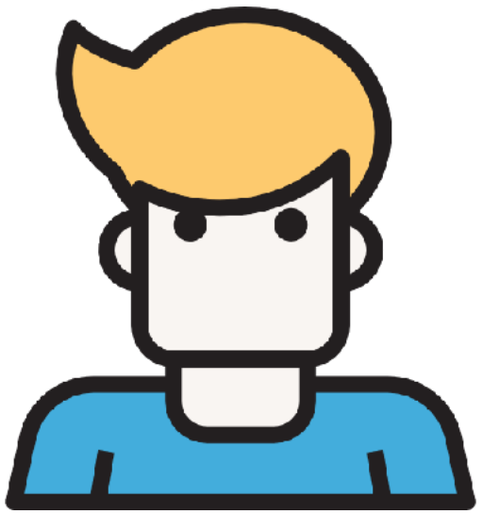
Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren* (1. Auflage 2019). Springer. https://www.springer.com/de/book/9783658215866

One model to rule them all

Most classical off-the-shelf statistical devices such as t-test, ANOVA etc. are just special cases of the linear model.

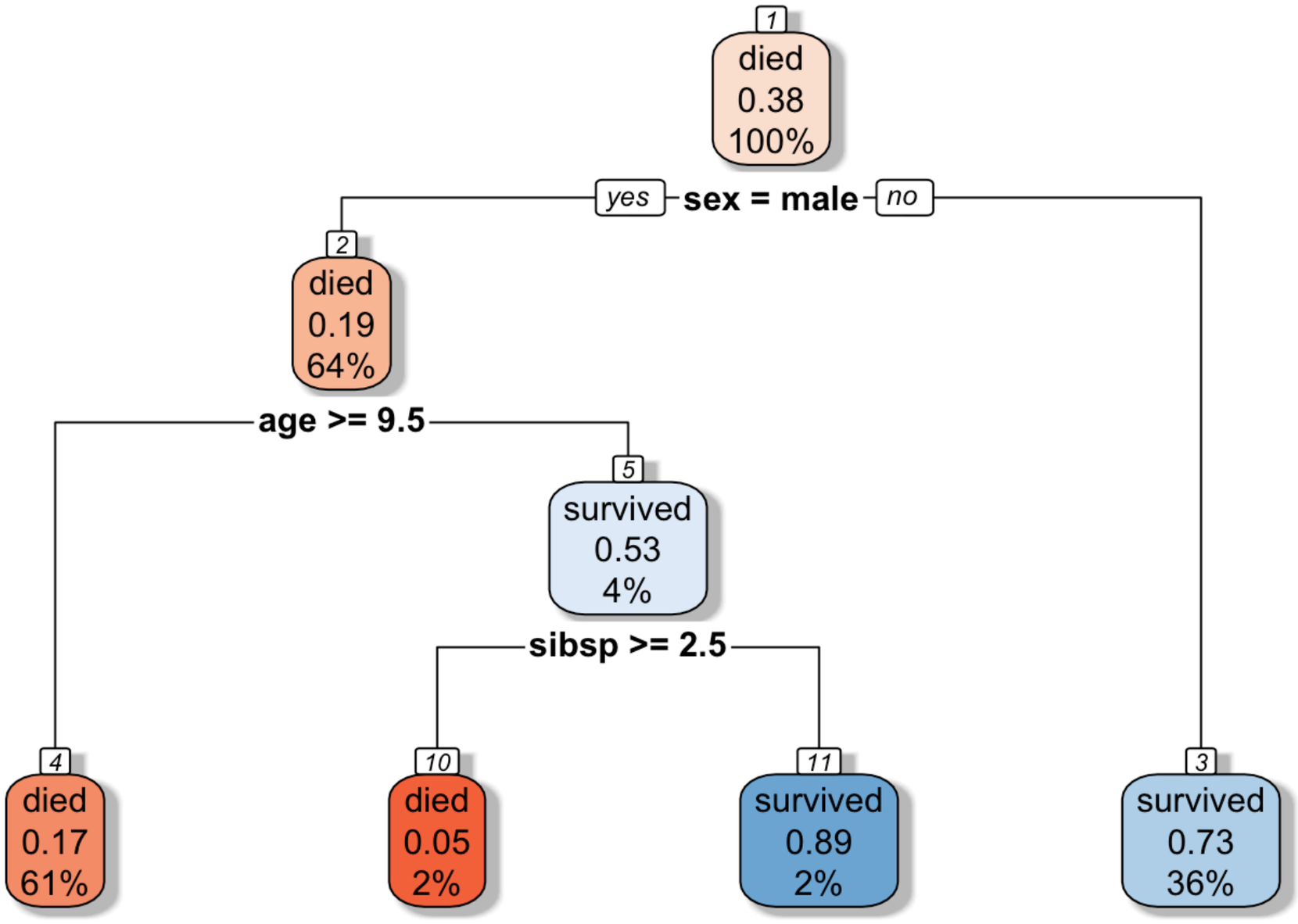
Wolfi

That's what I've kept on saying since ages! Trust me.



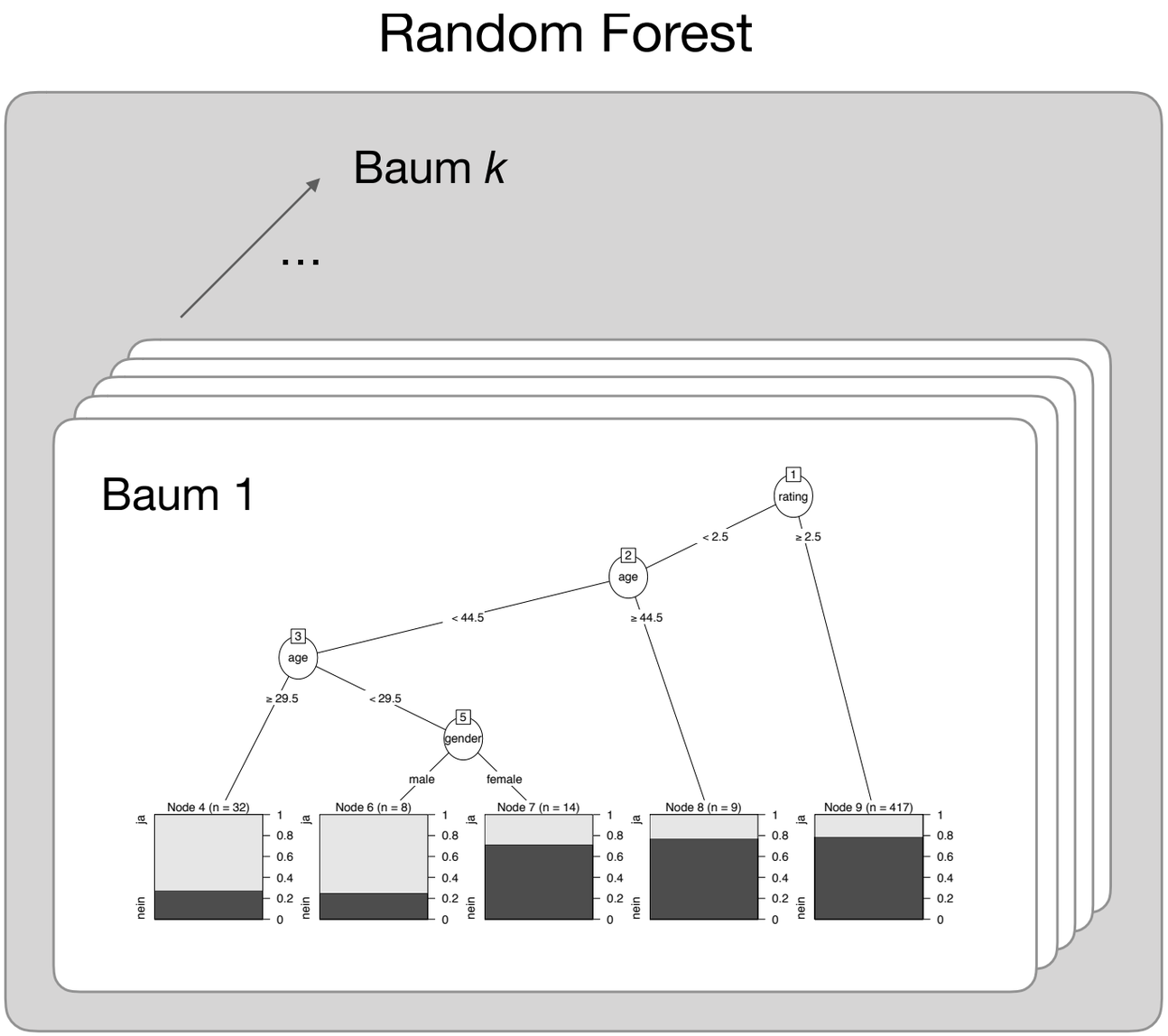
Don

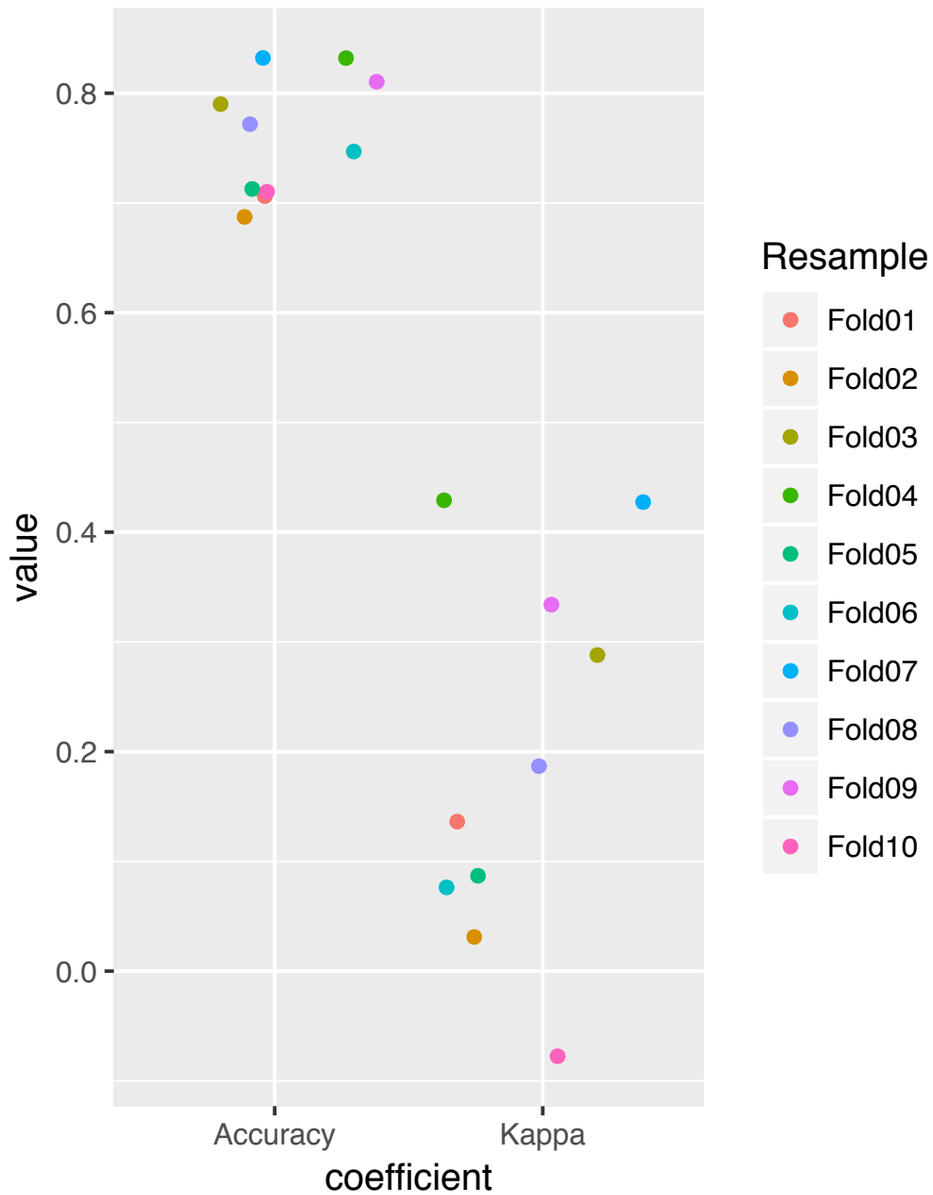
Culture 2: algorithmic modelling: Decision Tree



https://blog.exploratory.io/visualizing-a-decision-tree-using-r-packages-in-explortory-b26d4cb5e71f

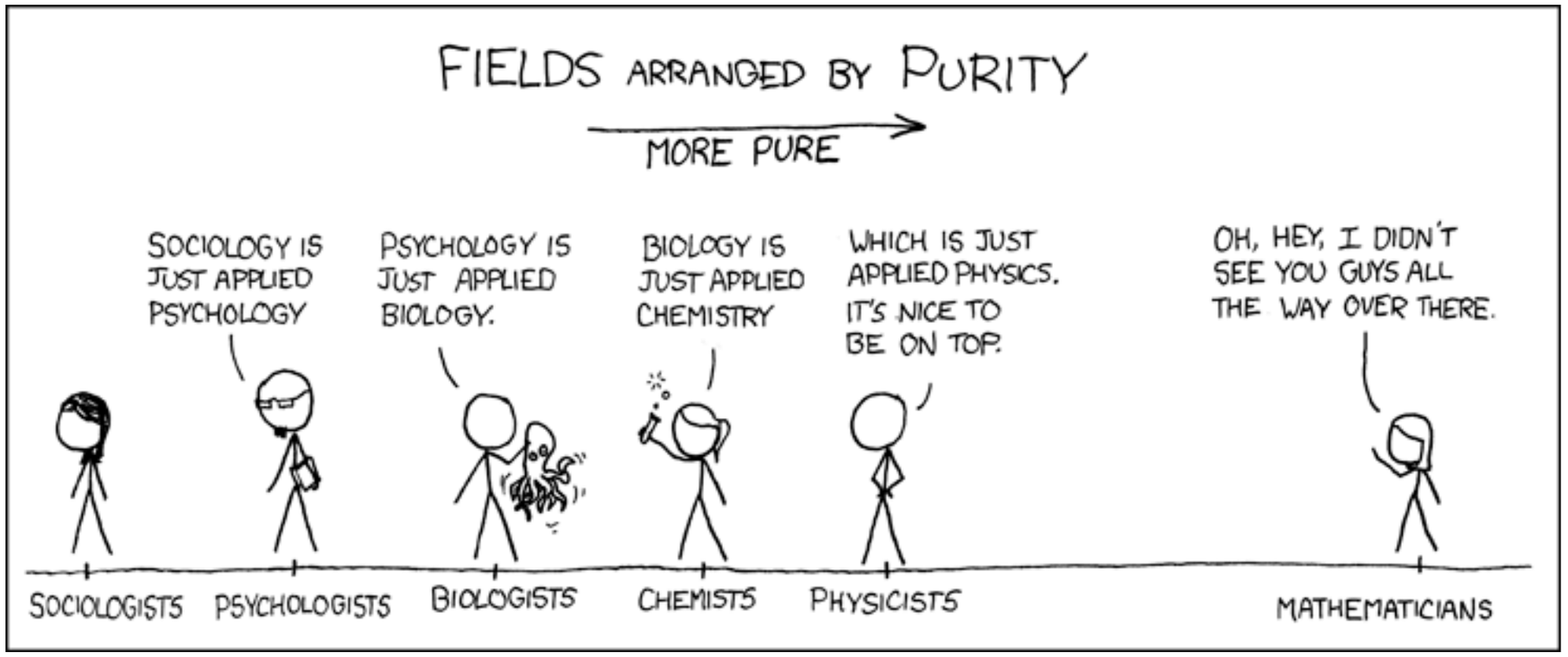
Many trees make a forest: Ensemble models





Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren* (1. Auflage 2019). Springer. https://www.springer.com/de/book/9783658215866

Fields/science branches overlap (?)



https://xkcd.com/435/

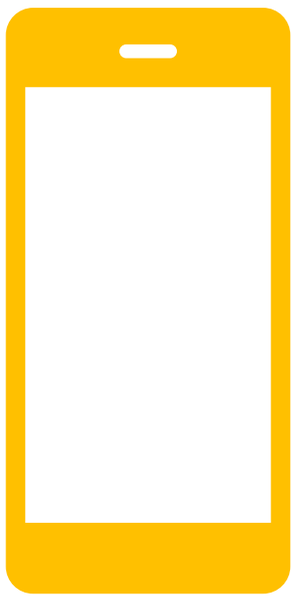
Describe the analysis

- Gather in the break-out rooms
- Work through this demonstration of the machine learning technique
- Prepare to report back the main insights

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

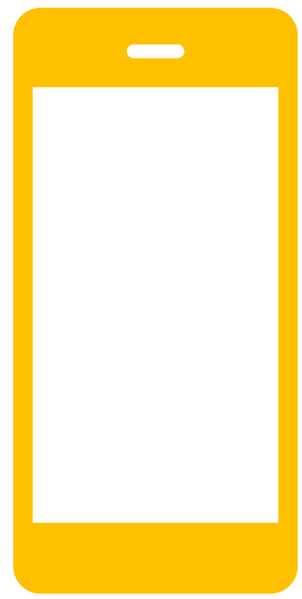
Social competition quiz

Get your phone.



Go to kahoot.it

Enter the PIN given.



or click here.

Recap – Cultures of data science

- There are two tribes of data analysis:
  - a) based on stochastic models and
  - b) on computer-science models.
- Did I mention there's no free lunch?
- Tree-based models are a good starter for a data-science analysis.
- In practice, one should compare performance of multiple algos.

Big data market research case studies

There are three types of market research studies

Epistemological goals of research studies

descriptive

„What consumer types exist?“

associative

„Do Facebook likes predict personality?“

causal

„Does *consumer time on the Webshop* increase sales?“?

„effect of X on Y“

„X impacts Y“

„X influences Y“

„X leads to Y“

Hernán, M. A., Hsu, J., & Healy, B. (2019). A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *Chance*, *32*(1), 42–49. https://doi.org/10.1080/09332480.2019.1579578

Analytical taxonomy

Prediction

Classification

abc

Text Mining

Data Reduction

Image/speech recognition

Clustering

...

Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques (Bd. 54). https://dx.doi.org/10.1007/978-3-642-19721-5

Sentiment mining

abc

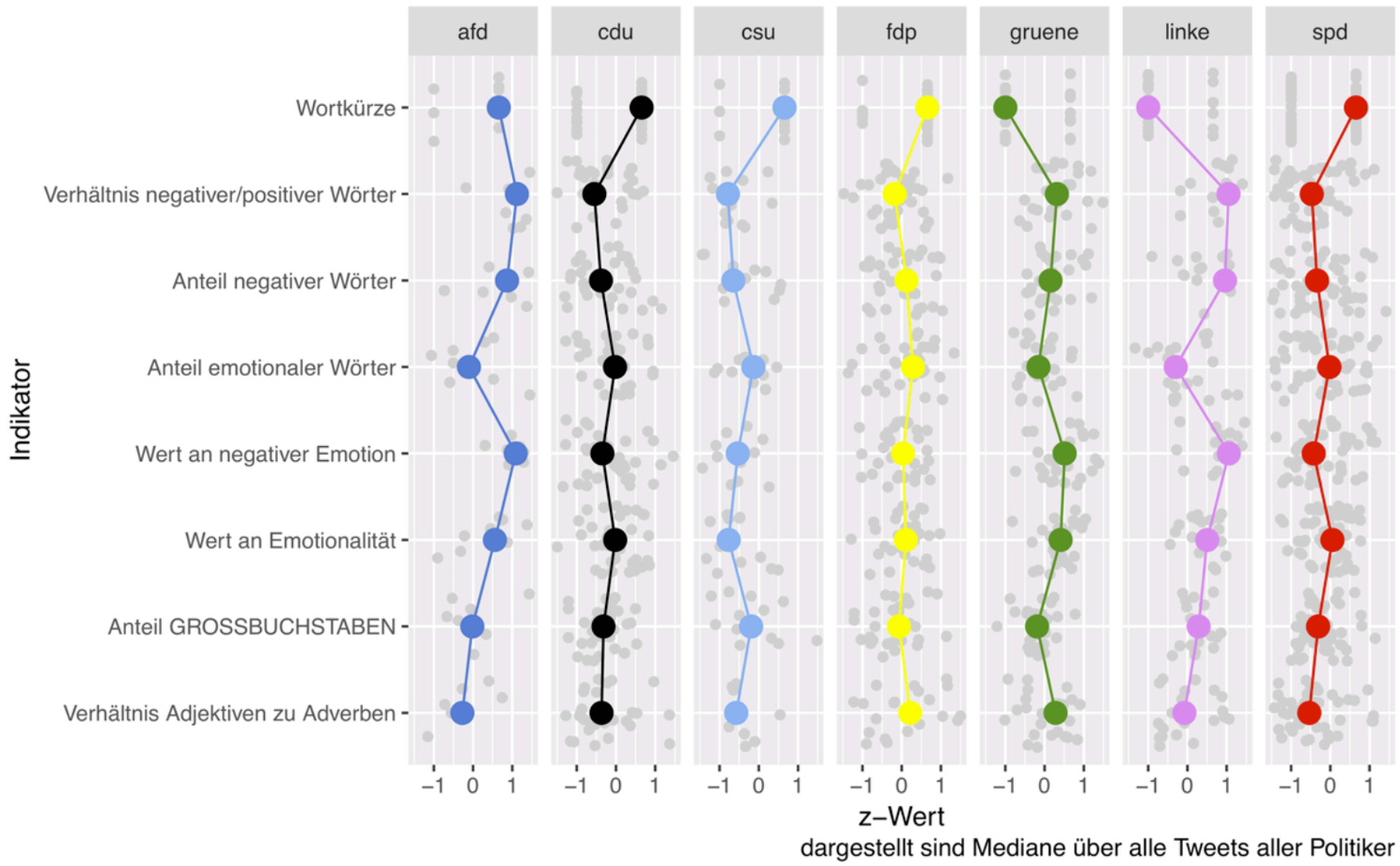
Text Mining

amount of data

6 Million words

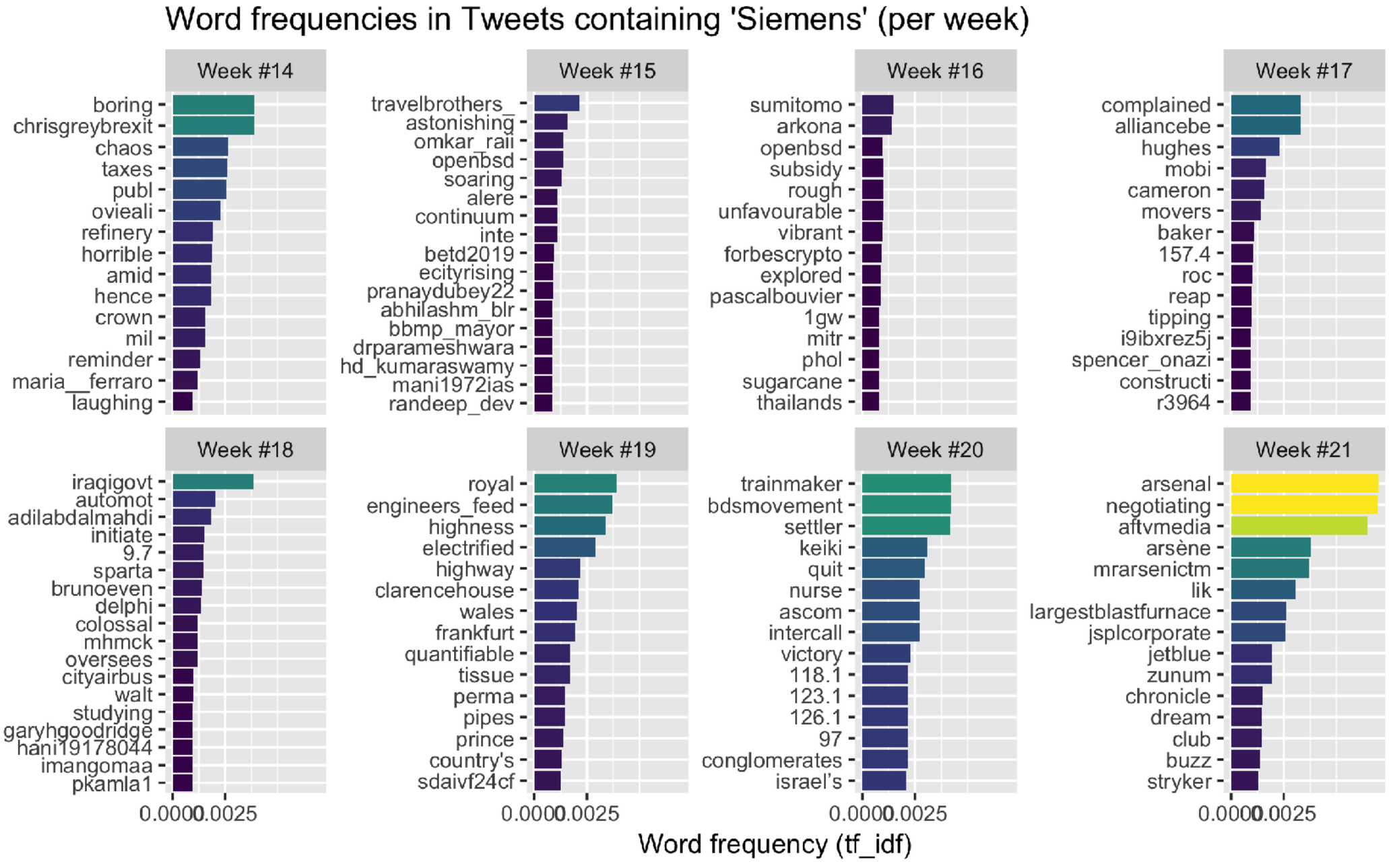
analytic method

Twitter-Scraping, Textmining



https://github.com/sebastiansauer/polits_tweet_mining

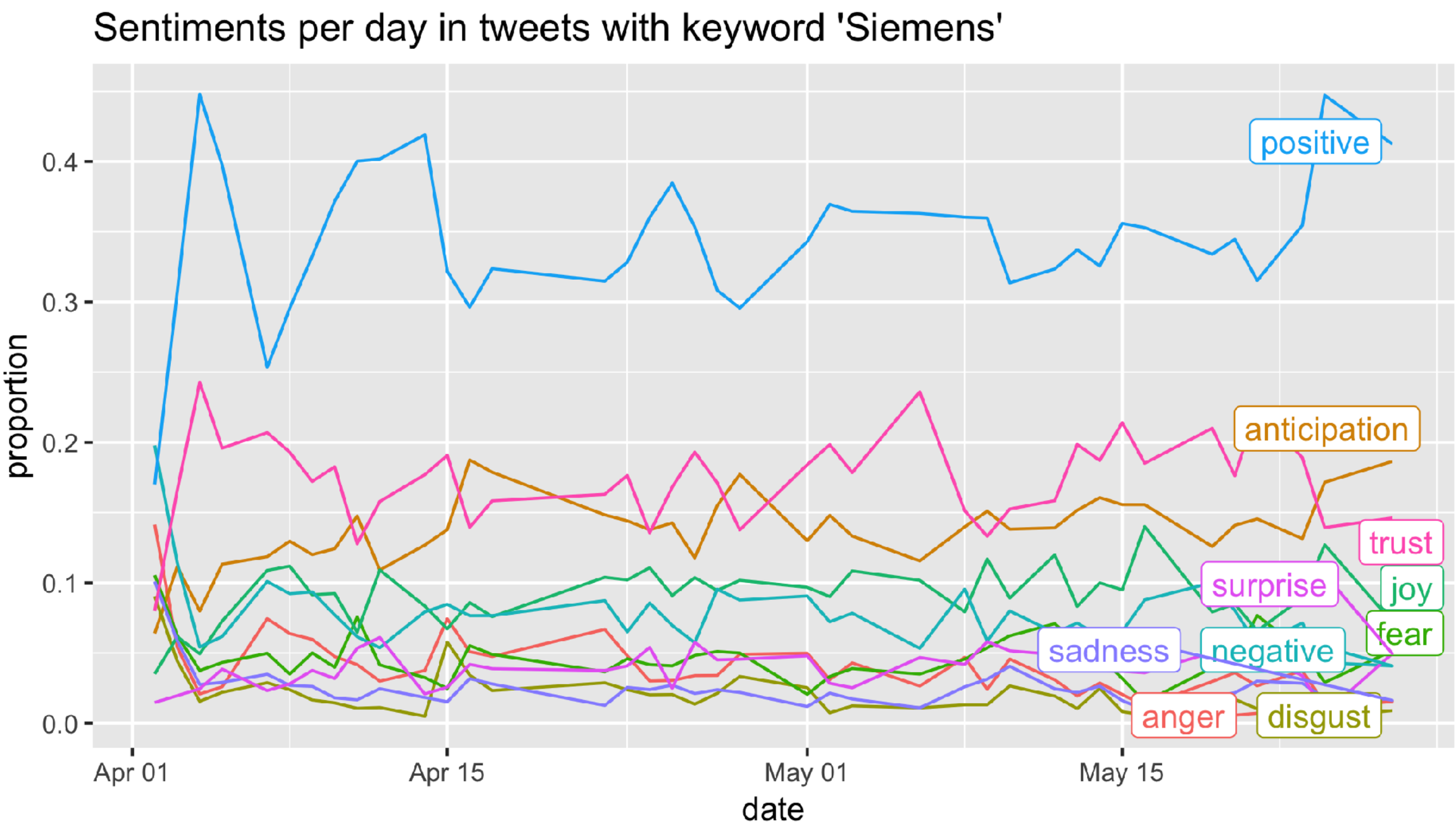
What do people say about Siemens?



Own analysis, unpublished

Different emotions attached to „Siemens“

abc
Text Mining



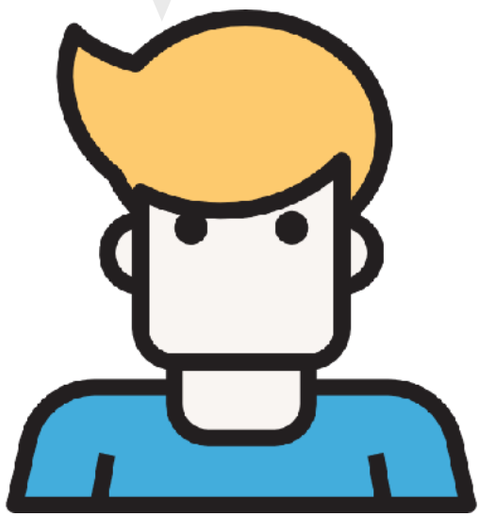
Own analysis, unpublished

IBM predicts which employees will quit

Prediction

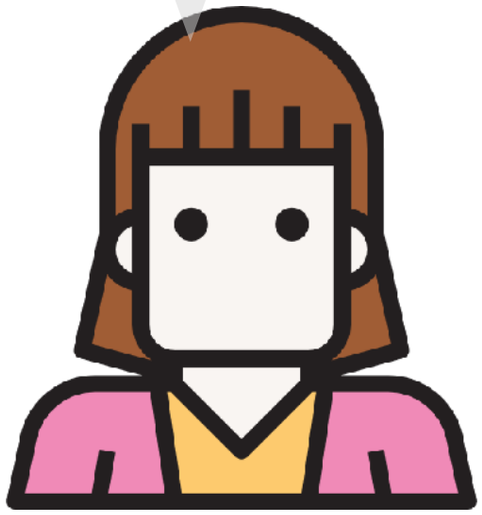
according to a study

If I knew you wants to quit, I'll fire ´em anyway!



Don

Predictive quality depends on data quality.



Angi

Privacy concerns ahead.

Wolfi

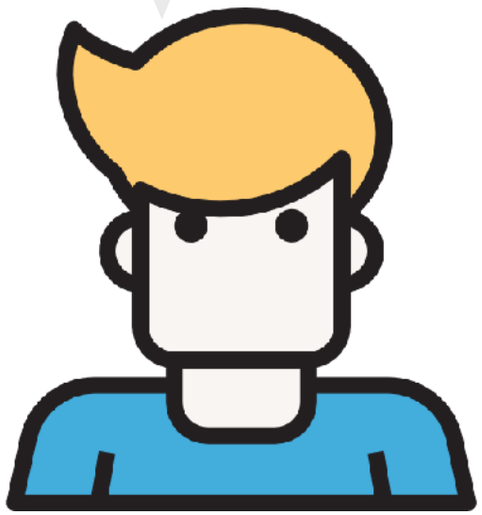
https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html, https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d

Market research in the advent of Big Data

- Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics, 24(1), 1–7. https://doi.org/10.1016/j.iedeen.2017.06.002
- Bosch, V. (2016). Big data in market research: Why more data does not automatically mean better information. Marketing Intelligence Review, 8(2), 56–63. https://content.sciendo.com/downloadpdf/journals/gfkmir/8/2/article-p56.xml
- Culotta, A. (2014). Estimating county health statistics with twitter. Proceedings of the 32Nd annual ACM conference on human factors in computing systems, 1335–1344. https://doi.org/10.1145/2556288.2557139 [Volltext]
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
- Liu, X., Shin, H., & Burns, A. C. (2019). Examining the impact of luxury brand's social media marketing on customer engagement : Using big data analytics and natural language processing. Journal of Business Research, S0148296319302954. https://doi.org/10.1016/j.jbusres.2019.04.042 [Volltext]
- Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. J. (2013). Mining facebook data for predictive personality modeling. Seventh international AAAI conference on weblogs and social media. [Volltext]

Check out big data business models

Let's just do what the consultants say!



Don

crystal balls

Boston Consulting Group

PWC

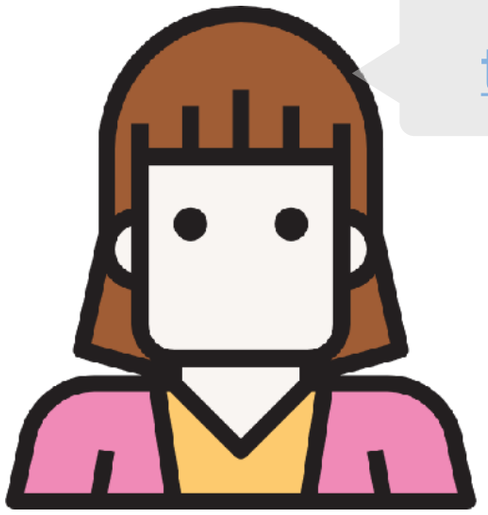
McKinsey

idc

https://image-src.bcg.com/Images/Seven_Ways_Big_Data_Apr_2014_tcm9-143370.pdf, https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.pdf, https://www.pwc.de/de/prozessoptimierung/assets/pwc-big-data-bedeutung-nutzen-mehrwert.pdf, https://www.import.io/wp-content/uploads/2017/04/Seagate-WP-DataAge2025-March-2017.pdf

What are your data analysis needs?

Try to map your ideas to the taxonomies presented above.



epistemological goals

- **descriptive**
- **associative**
- **causal**

analytical goals

- Prediction
- Classification
- abc Text Mining
- Data Reduction
- Image/speech recognition
- Clustering

https://app.conceptboard.com/board/nynx-1f65-gzan-k5hs-2hpb

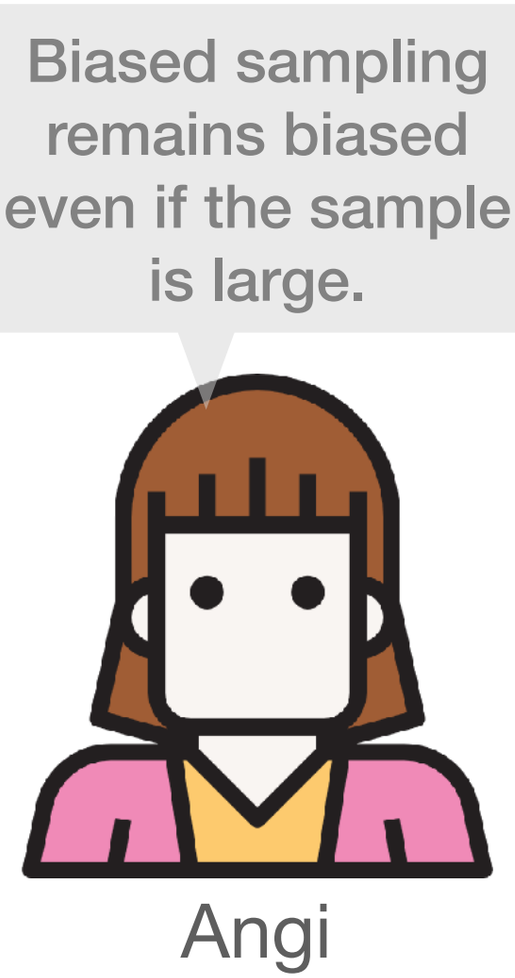
Recap – Market research case studies

- Epistemological goals can be classified in
  - decriptive
  - associative
  - causal
- Analytical goals can be classified as
  - prediction, classification, clustering, text mining, image recognition, data reduction, …
- Consultants propose some business models. However, it’s more important to tailer your ideas to your customers’ needs.

Why most analytical insights are unfit for business decisions

Examples of biased samples

Challenge 1: Biased samples

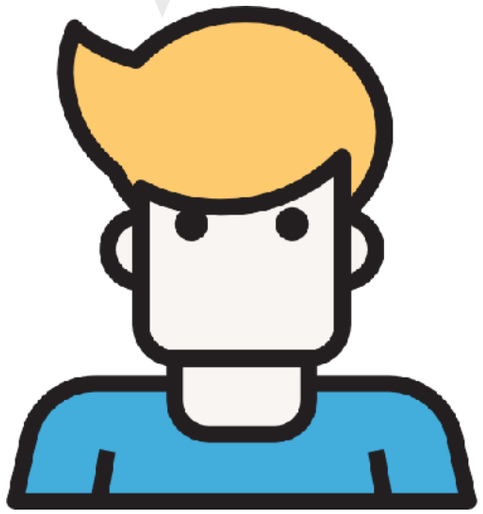


Bosch, V. (2016). Big data in market research: Why more data does not automatically mean better information. Marketing Intelligence Review, 8(2), 56–63, https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html

Explain „Correlation ≠ Causation“

Challenge 2: Missing causal structure

Something with storks and babies?

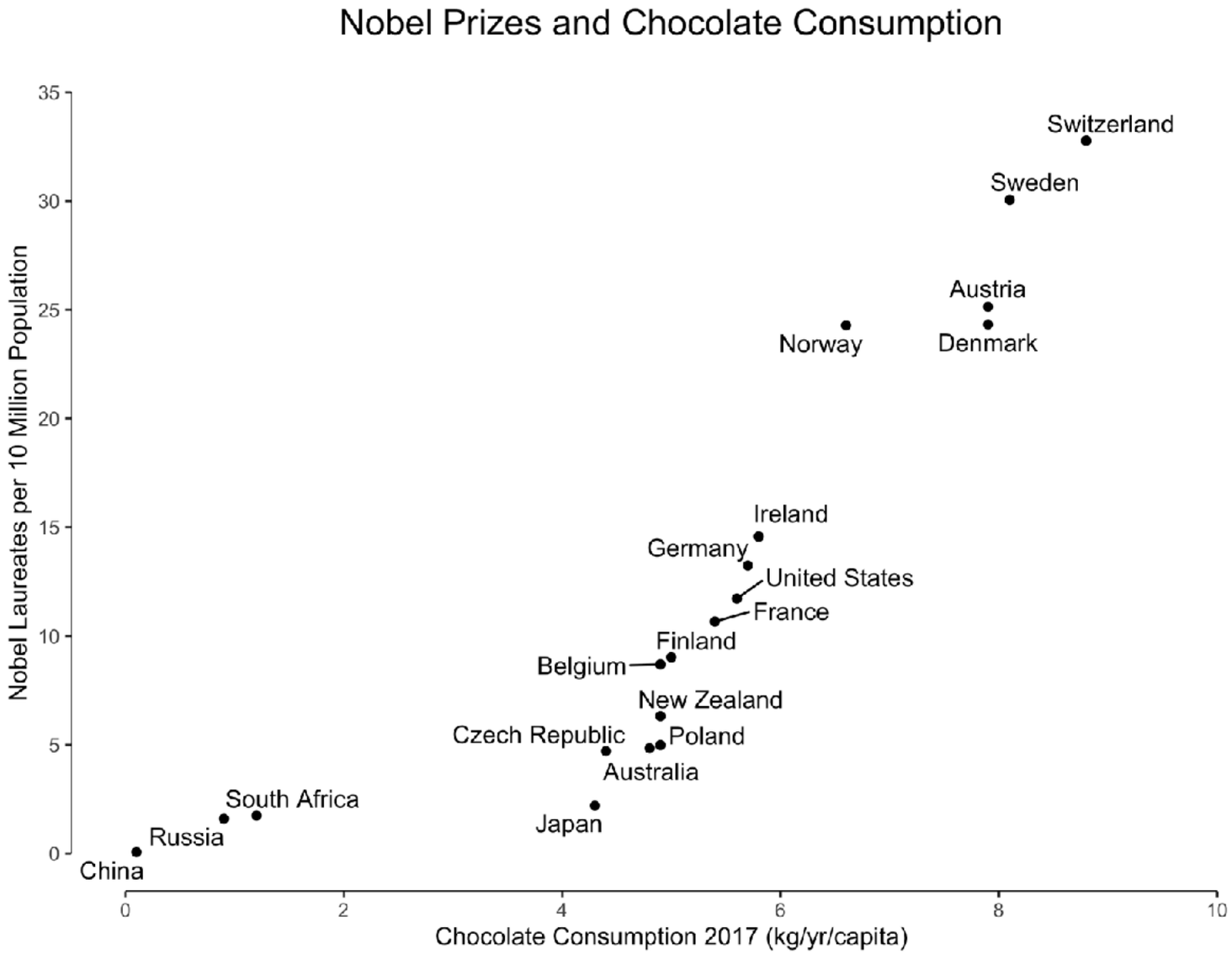


Don

Yeah, but there’s more to it.

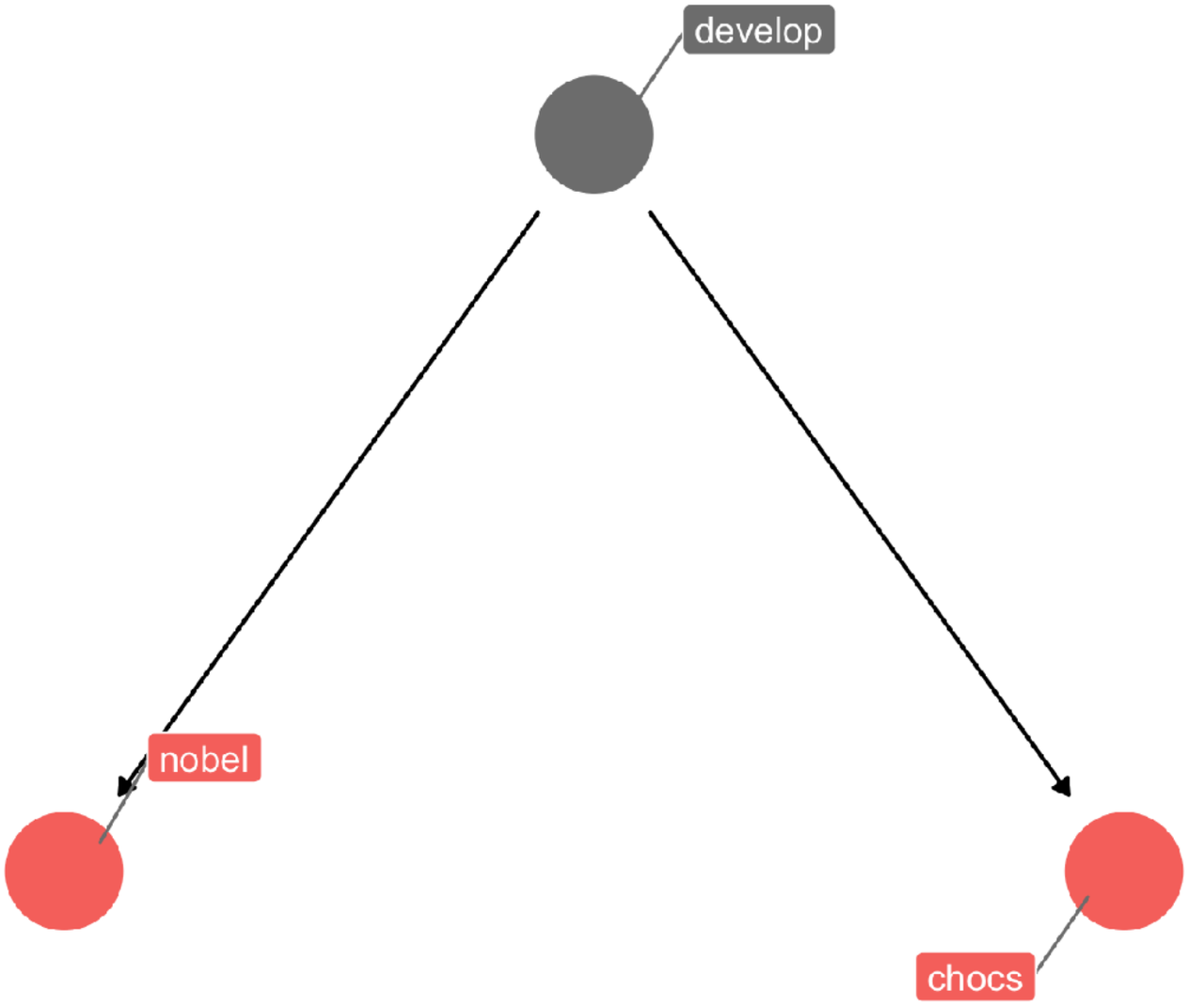
Wolfi

Spurious correlation: Example

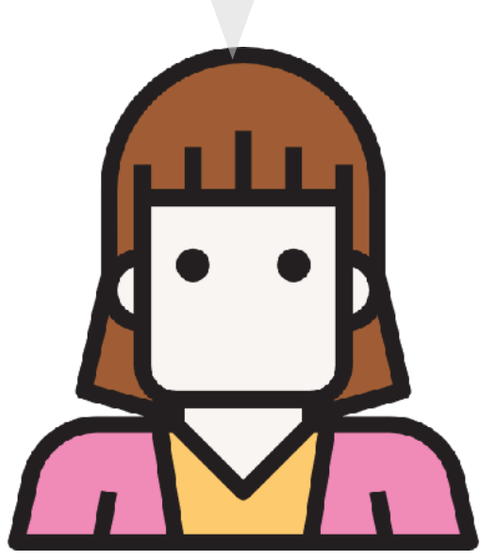


Dablander, F. (2020). *An Introduction to Causal Inference* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/b3fkw

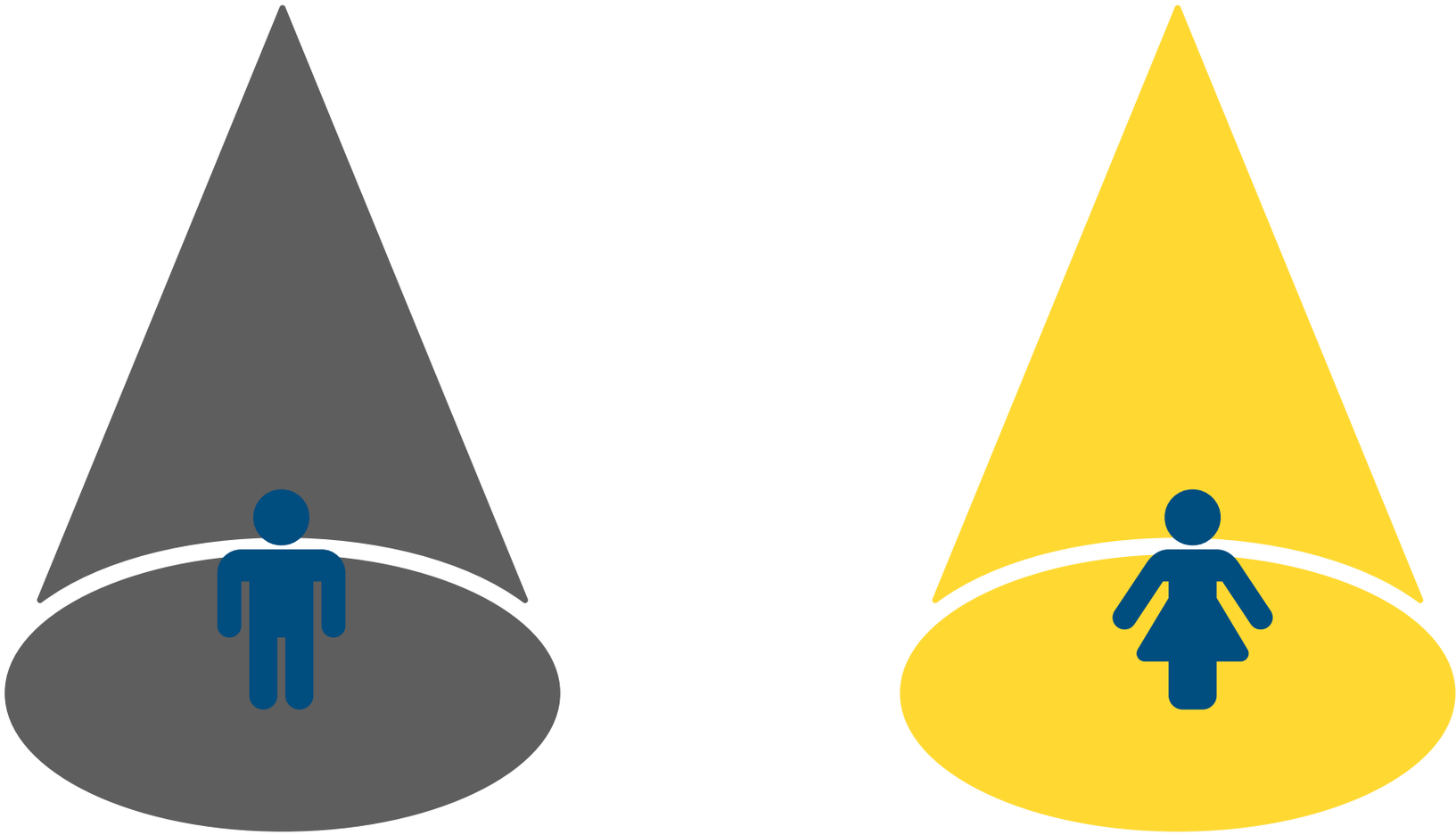
A (simplistic) causal model for the chocolate example



There'll be a fake correlation between *nobel* and *chocolate* consumption – unless you control for *develop*!



Statistical associations can be real … or spurious

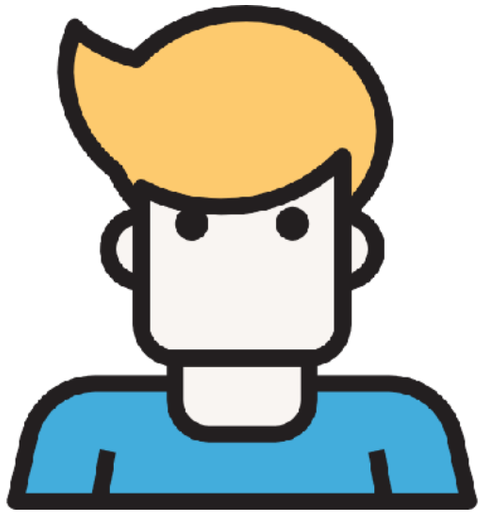


Fake (spurious)

real (causal)

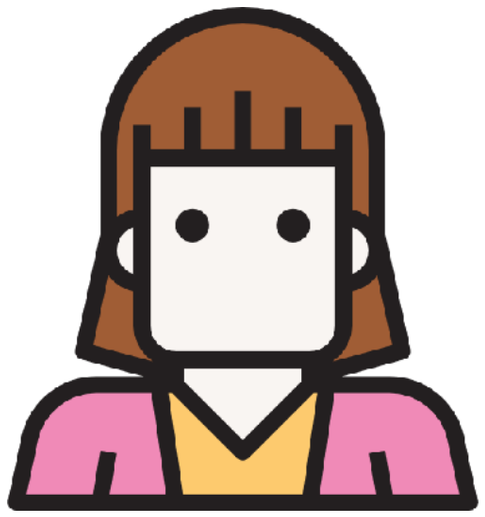
Angi's got a job from Don

What's the worth of my house? Two bedrooms! Great bedrooms!



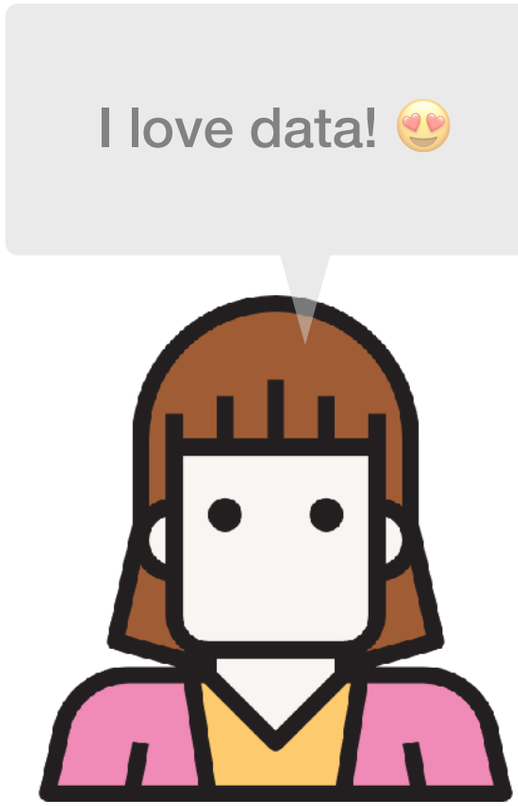
Don

I'll find out. The science way.



Angi

Here's a glimpse on her data



Angi

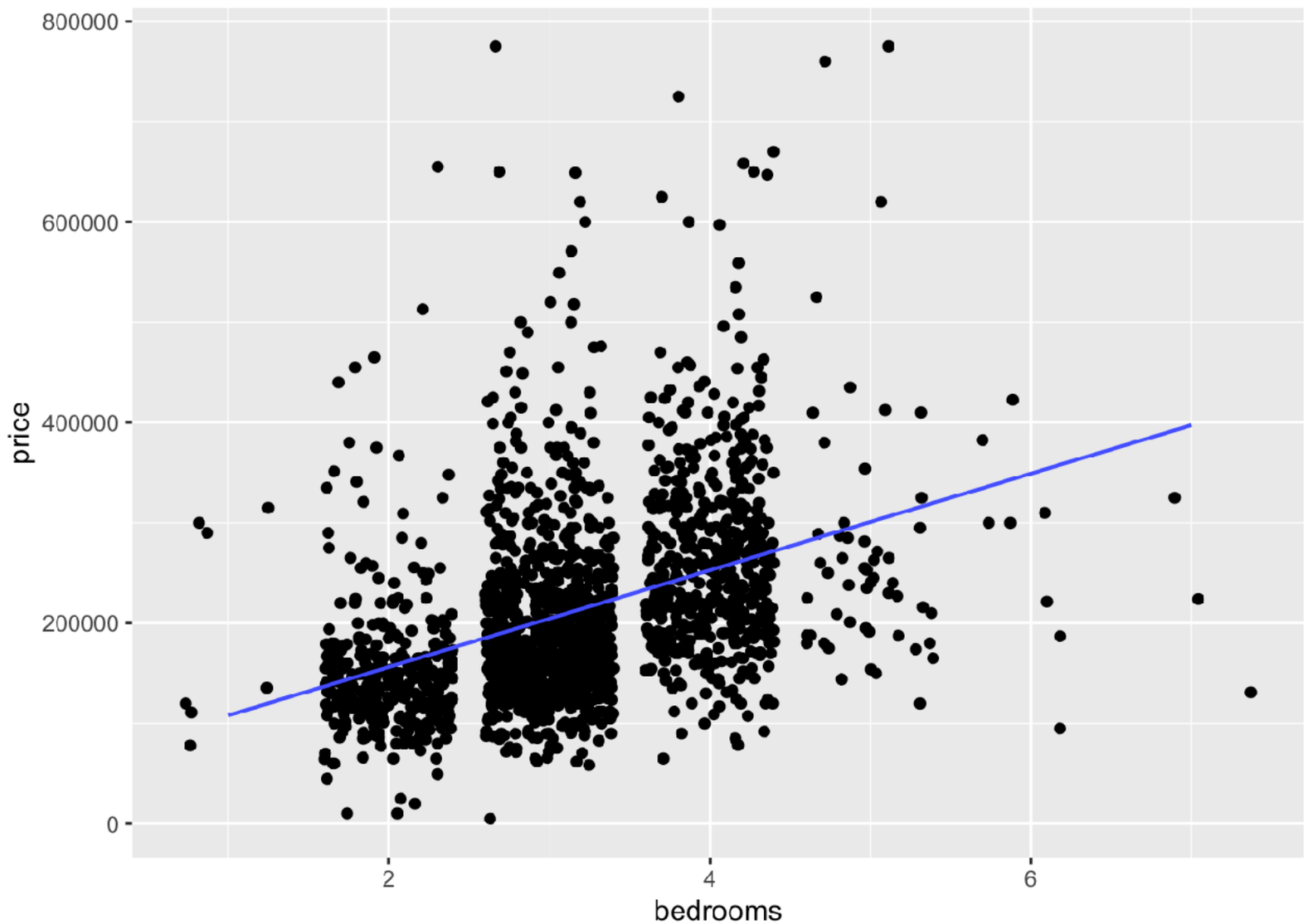
price	square-feet	age	a/c	fireplace	heating	…
132.500	84,17	42	No	Yes	Electricity	
181.115	181,44	0	No	No	Gas	
109.000	180,60	133	No	Yes	Gas	
155.000	180,60	13	No	Yes	Gas	
86.060	78,04	0	Yes	No	Gas	
120.000	107,02	31	No	Yes	Gas	
153.000	255,67	33	No	Yes	Oil	
170.000	154,40	23	No	Yes	Oil	
90.000	151,62	36	No	No	Electricity	
122.900	131,55	4	No	No	Gas	
…	…	…	…	…	…	

Corvetti, C. (2006). *Saratoga Houses*. https://rdrr.io/cran/mosaicData/

Model 1: Price as a function of number of bedrooms

Hey Don! More bedrooms, more bucks!

Angi



Angi reporting back to Don. Don isn't happy.

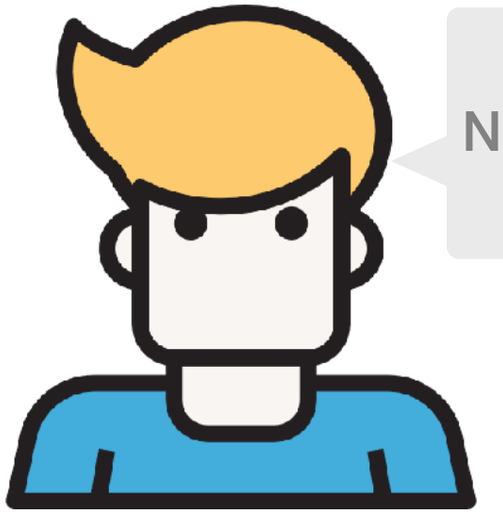
```
model1 <- lm(price ~ bedrooms,
data = SaratogaHouses)
coef(model1)
```

```
## (Intercept)    bedrooms
##    59862.96    48217.81
```

```
dons_house <- data.frame(bedrooms = 2)
predict(model1, dons_house)
```

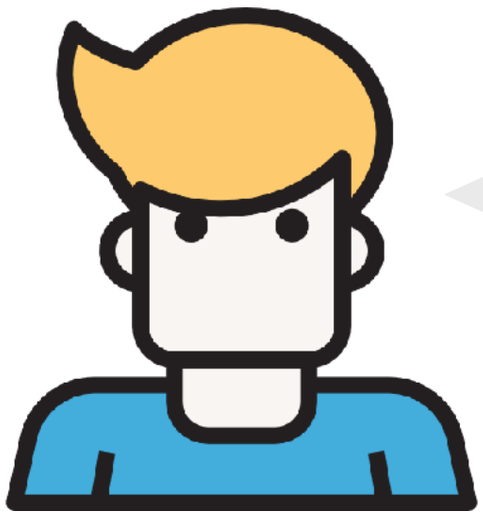
```
##        1
## 156298.6
```

Angi



Don

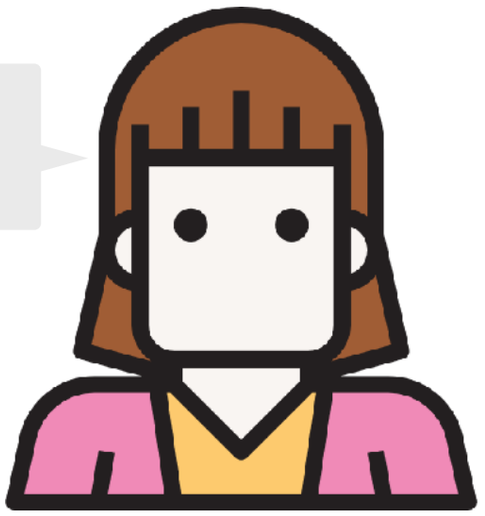
Don's got an idea: Split each bedroom into two



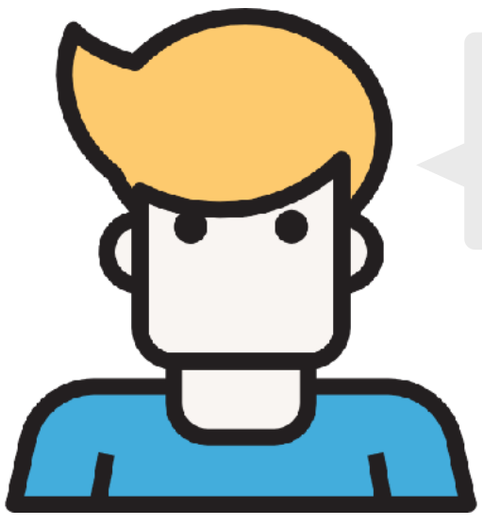
I'll split each bedrooms into two!

Don

Wait …



Angi



Crunch the data – now!

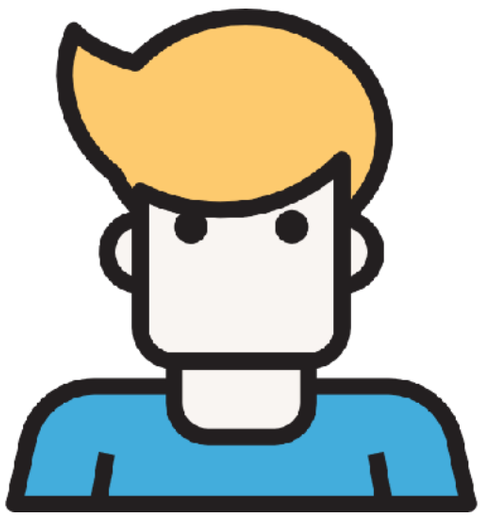
Don

With 4 bedrooms, the price rises to 250k, model 1 says

```
dons_new_house <- data.frame(bedrooms = 4)
predict(model1, dons_new_house)
```

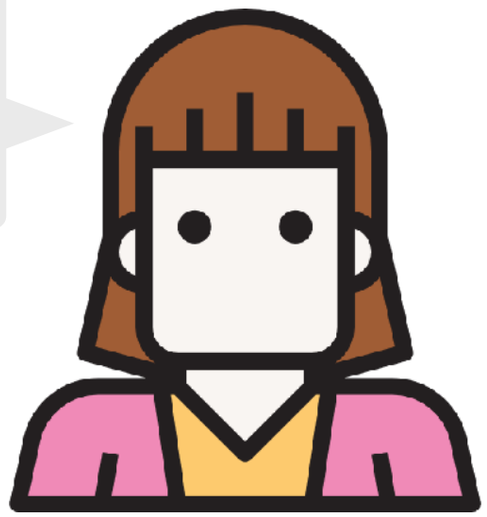
```
##        1
## 252734.2
```

I nailed it!
Now I'll earn 250k,
a full 100k plus!🤑



Don

Not so fast …



Angi

Wolfi wants you to know a thing

Wolfi

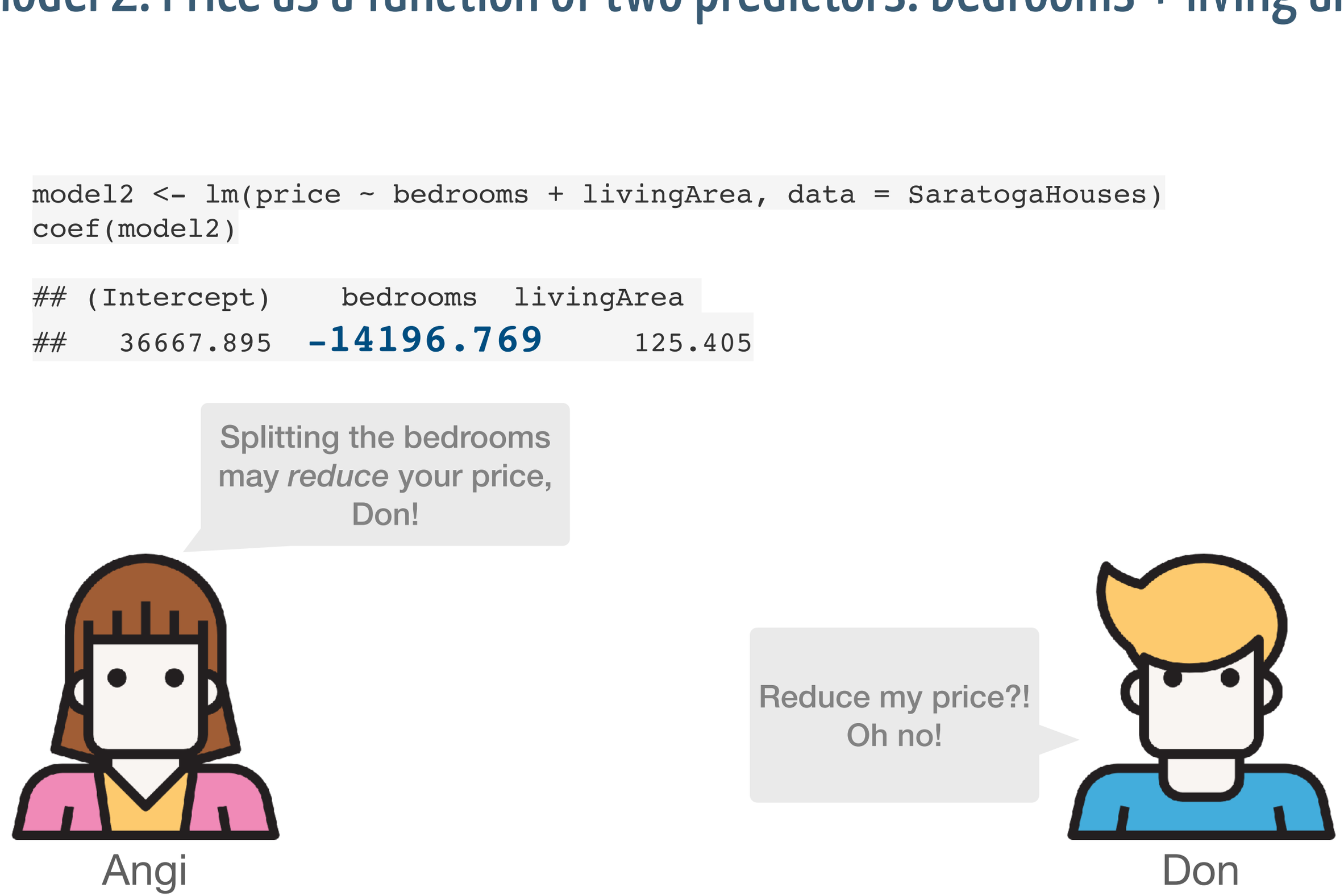
It something looks to good to be true,

it probably is.

Model 2: Price as a function of two predictors: bedrooms + living area

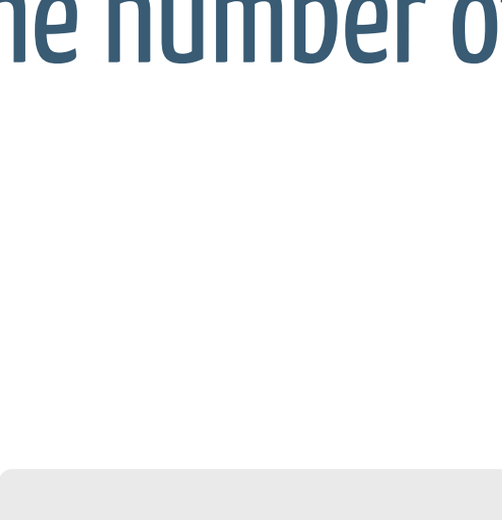
```
model2 <- lm(price ~ bedrooms + livingArea, data = SaratogaHouses)
coef(model2)
```

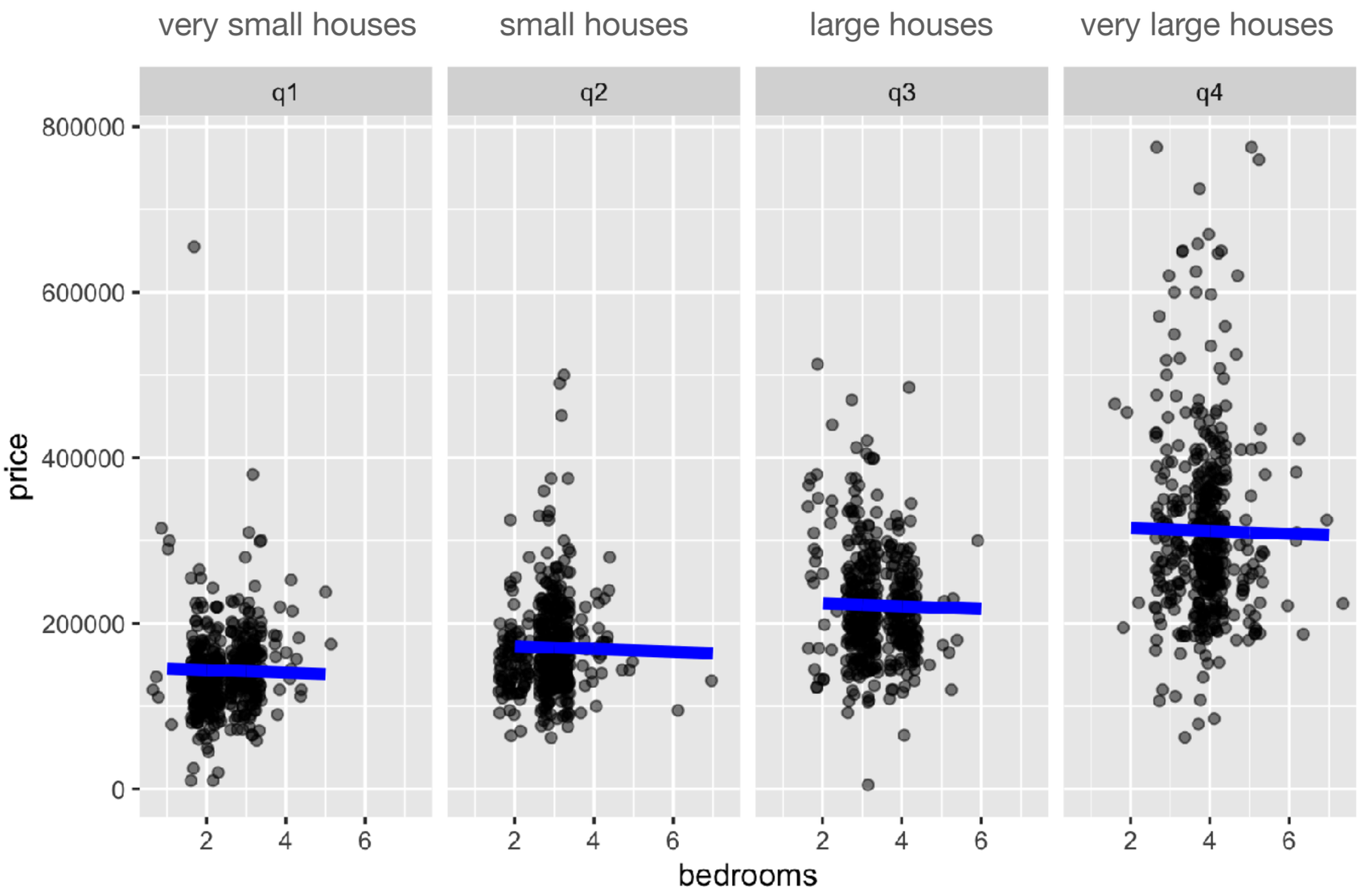
```
## (Intercept)    bedrooms  livingArea
##   36667.895  -14196.769     125.405
```



The number of bedrooms is negatively associated with price

Now a **NEGATIVE** association!
NE-GA-TIVE!





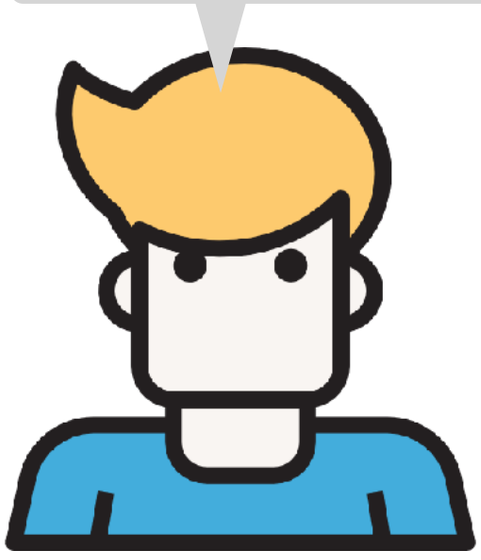
Wolfi shares his wisdom

Adding predictors will often change the association to the outcome of the other predictors.

Wolfi

The borders of statistics

But which model should I trust? Model 1 or model 2?



Don

Statistics cannot tell.

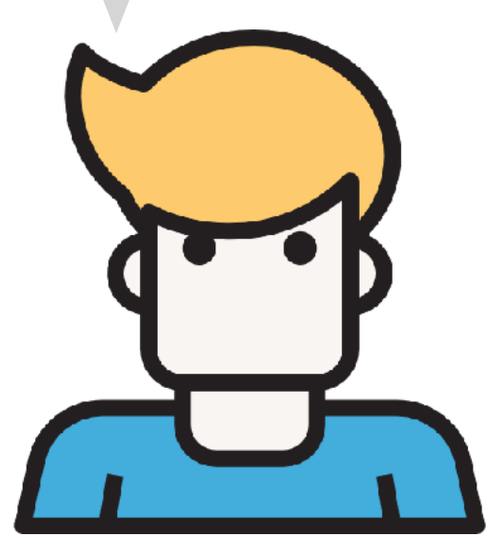
Wolfi

Business is about decision making

Many observational studies are unfit for decision making.

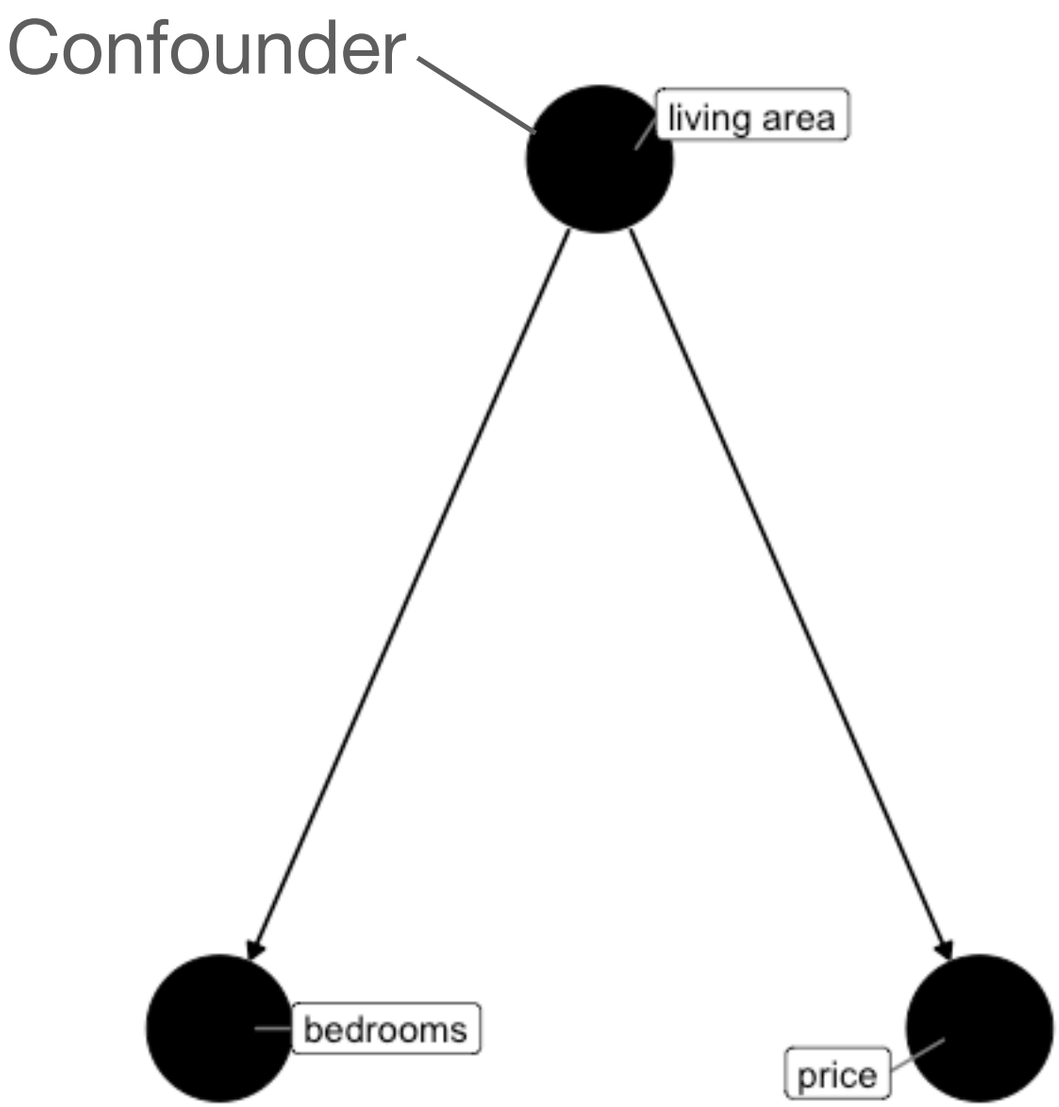
Wolfi

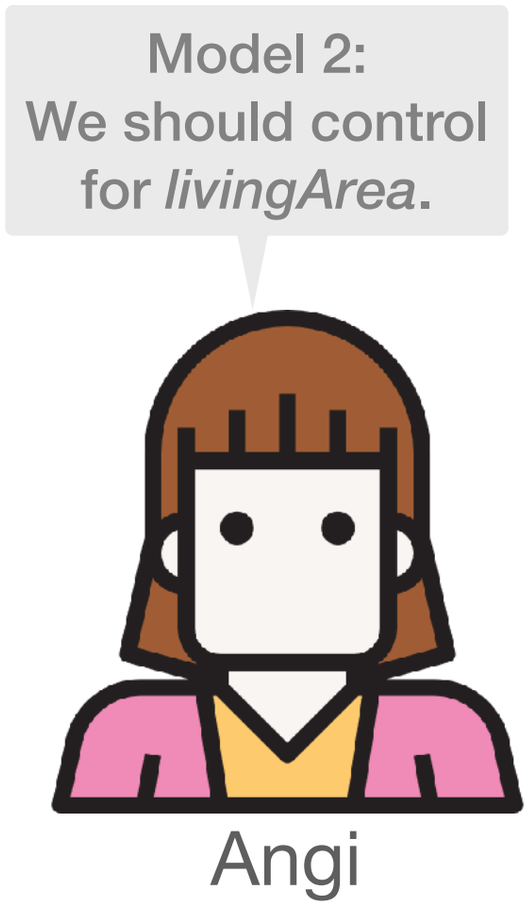
Who are you? The slayer of science or what?



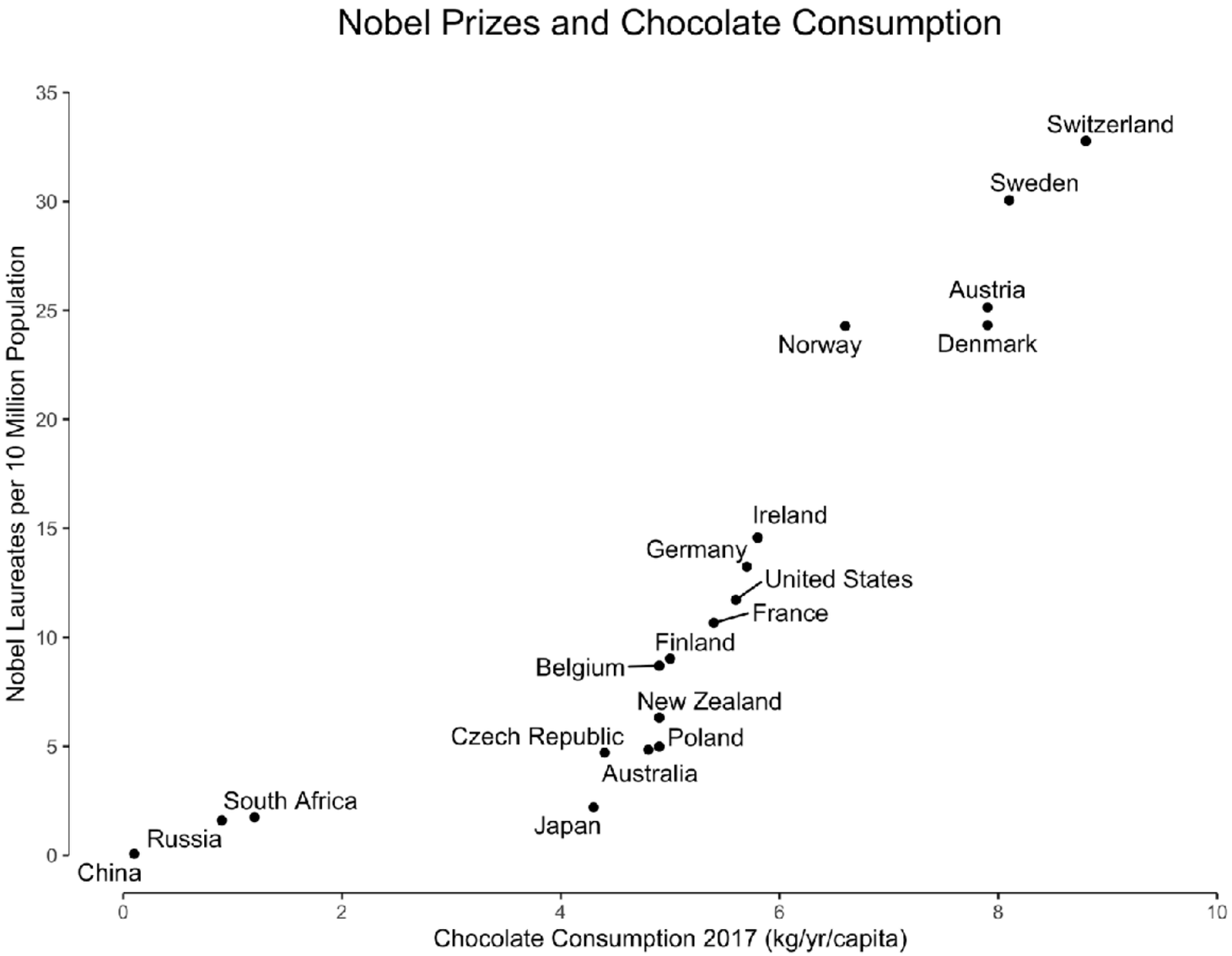
Don

Model 2: Angi's model: livingArea as a confounder

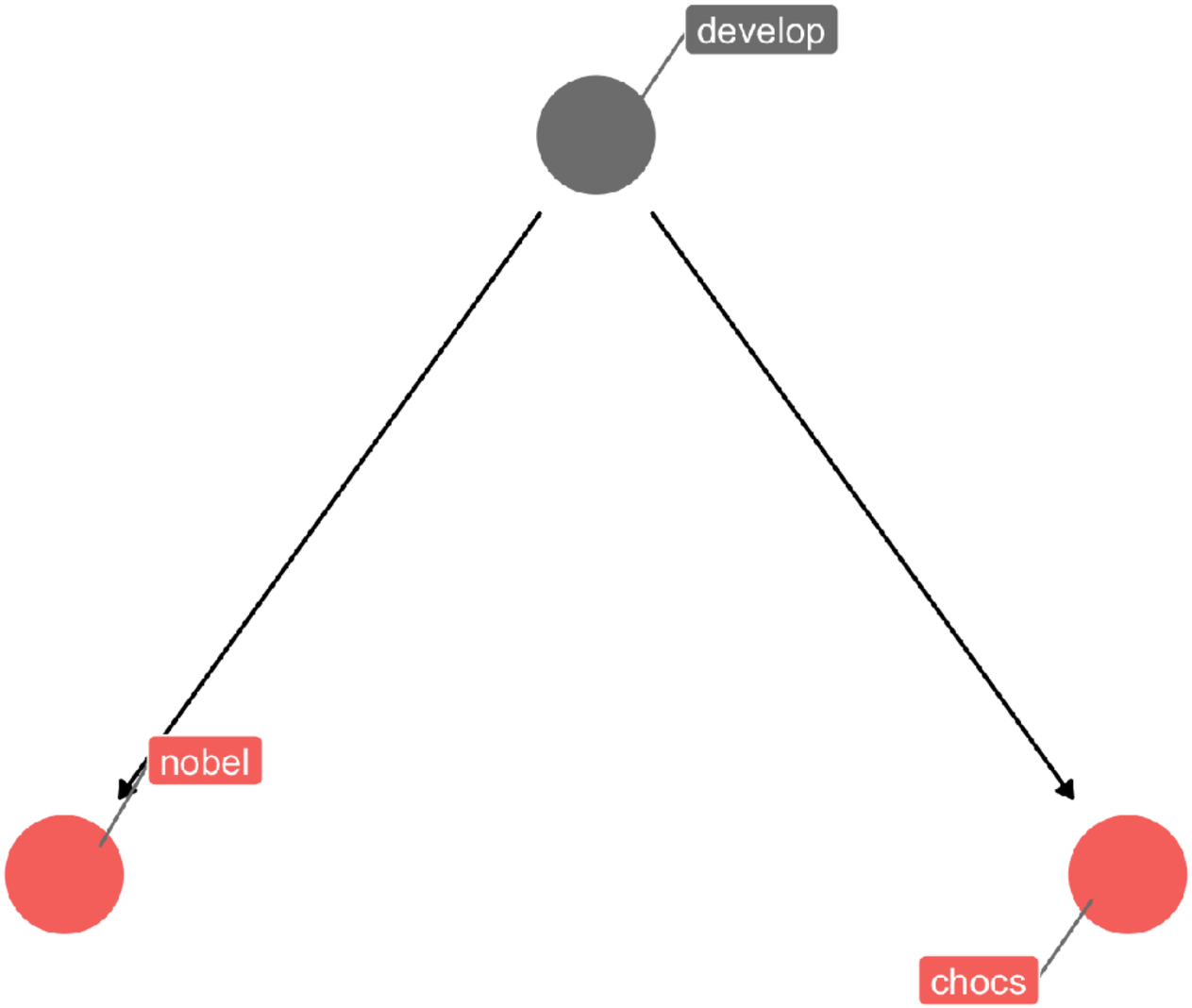




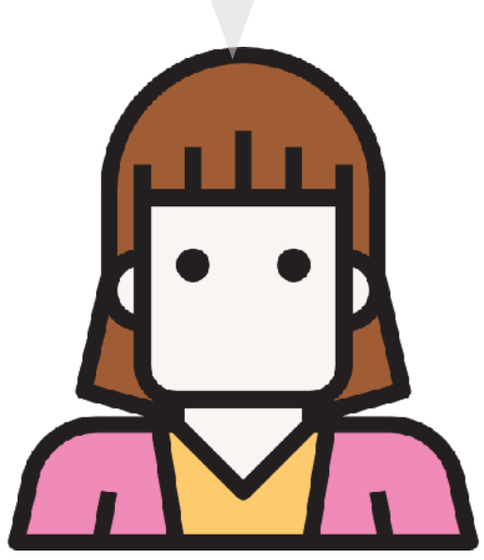
Spurious correlation: Example



A (simplistic) causal model for the chocolate example

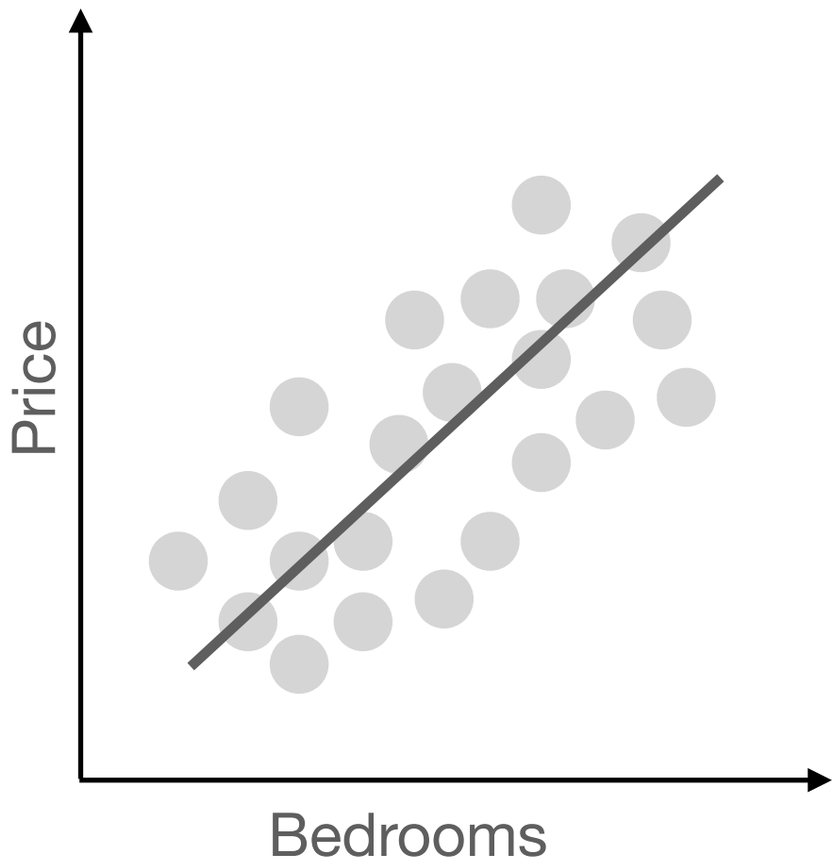


There'll be a fake correlation between *nobel* and *chocolate* consumption – unless you control for *develop*!



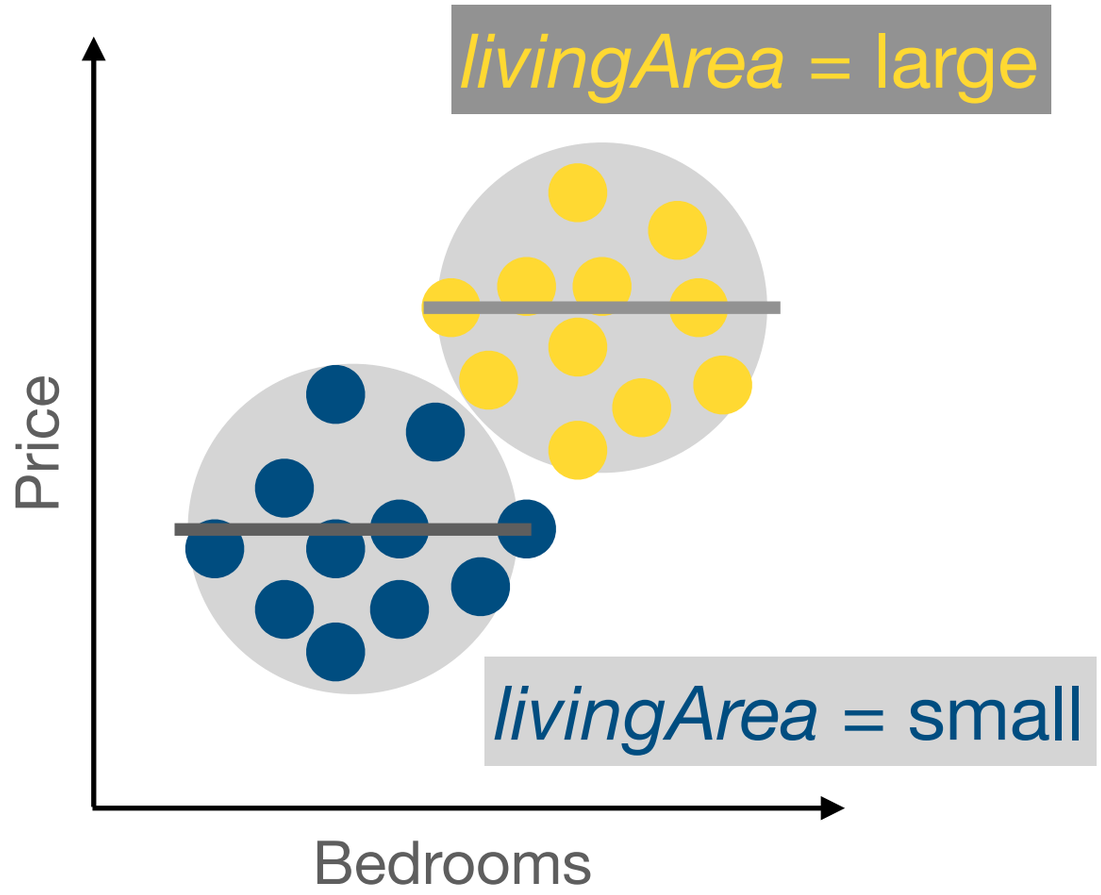
Controlling the confounder is the key

Model 1: Confounder *livingArea*
NOT controlled



Spurious correlation appears

Model 2: Confounder l*ivingArea*
controlled

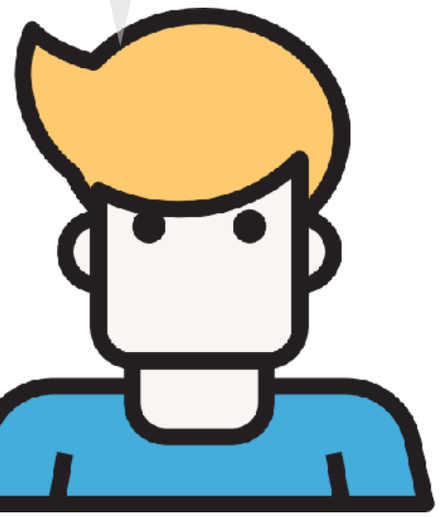


Spurious correlation disappears

Model 1 does not fit the data

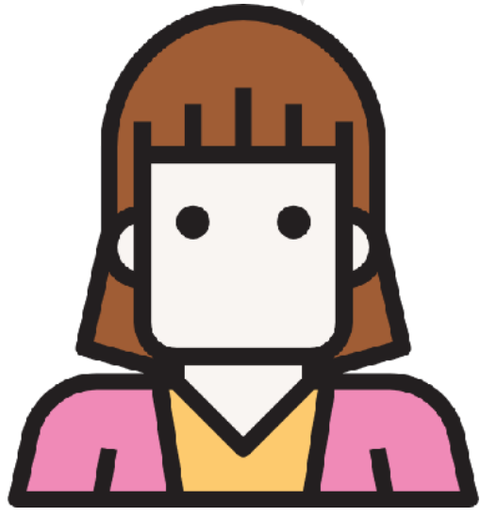
Don still likes model 1 though

Model 1 holds that *price* is influenced by *bedrooms* only. No confounders!



Don

Model 1 is falsified by the data.

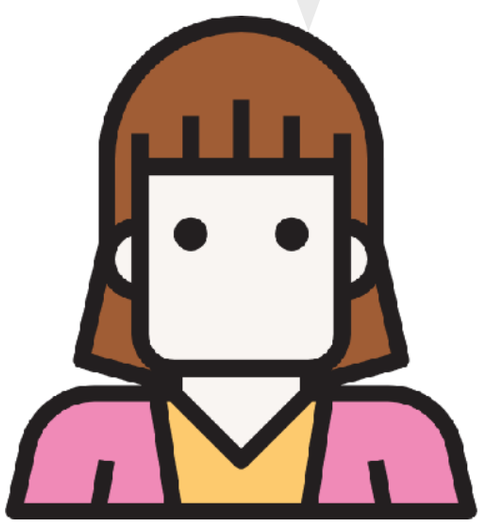


Angi

Journal of Applied Psychology reasons about causal questions

Indicators of causal hypotheses in recent issues (4/5 of 2020)

10 out of 12 studies presented their research questions using causal language.



Angi

Title	quotes (abstract)	causal language?
The generation and function of moral emotions in teams: An integrative review.	„influence on individual team members' moral emotions“	yes
On melting pots and salad bowls: A meta-analysis of the effects of identity-blind and identity-conscious diversity ideologies.	„improve intergroup relations“ „the effects of identity-blind ideologies“	yes
Political affiliation and employment screening decisions: The role of similarity and identification processes.	„to examine the effects of“	yes
A dynamic account of self-efficacy in entrepreneurship.	„self-efficacy energizes action because“	yes
Coworker support and its relationship to allostasis during a workday: A diary study on trajectories of heart rate variability during work.	„We examined the effect of“	yes
A theoretical assessment of dismissal rates and unit performance, with empirical evidence.	"utility analysis suggests that increasing dismissal rates can improve performance“	yes
Motivation to lead: A meta-analysis and distal-proximal model of motivation and leadership.	„the three MTL types partially explained the relationship“	**no**
Putting leaders in a bad mood: The affective costs of helping followers with personal problems.	„how such helping acts may impact leaders“ „ leaders with high (vs. low) managerial experience were less affected by“	yes
When goals are known: The effects of audience relative status on goal commitment and performance.	„investigating how the perceived relative status of a goal audience influences goal commitment“	yes
Selecting response anchors with equal intervals for summated rating scales.		**no**
It hurts me too! (or not?): Exploring the negative implications for abusive bosses.	„we propose that perpetrated abuse impacts these supervisor outcomes“	yes
How can employers benefit most from developmental job experiences? The needs–supplies fit perspective.	„developmental job experiences (DJE) lead to positive work-related outcomes“	yes

Nothing's a practical as a good theory

You need a causal model in order to disentangle the true correlations in a observational study.

Wolfi

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Digging deeper

The post is cool, but kinda heavy going.

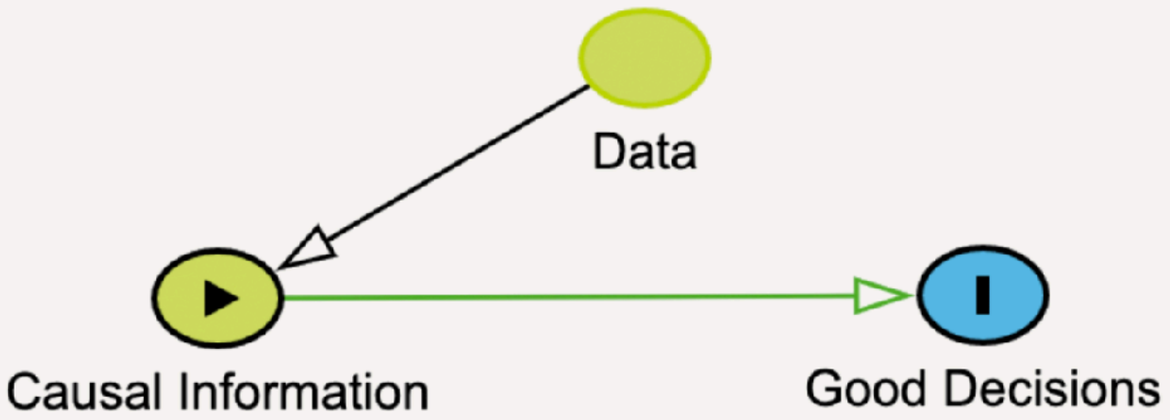
Wolfi

Want to make good business decisions? Learn causality

EDDIE LANDESBERG, MOLLY DAVIES, AND STEPHANIE YEE

December 19, 2019 · San Francisco, CA

Tweet this post! Post on LinkedIn



Made with DAGitty

https://multithreaded.stitchfix.com/blog/2019/12/19/good-marketing-decisions/

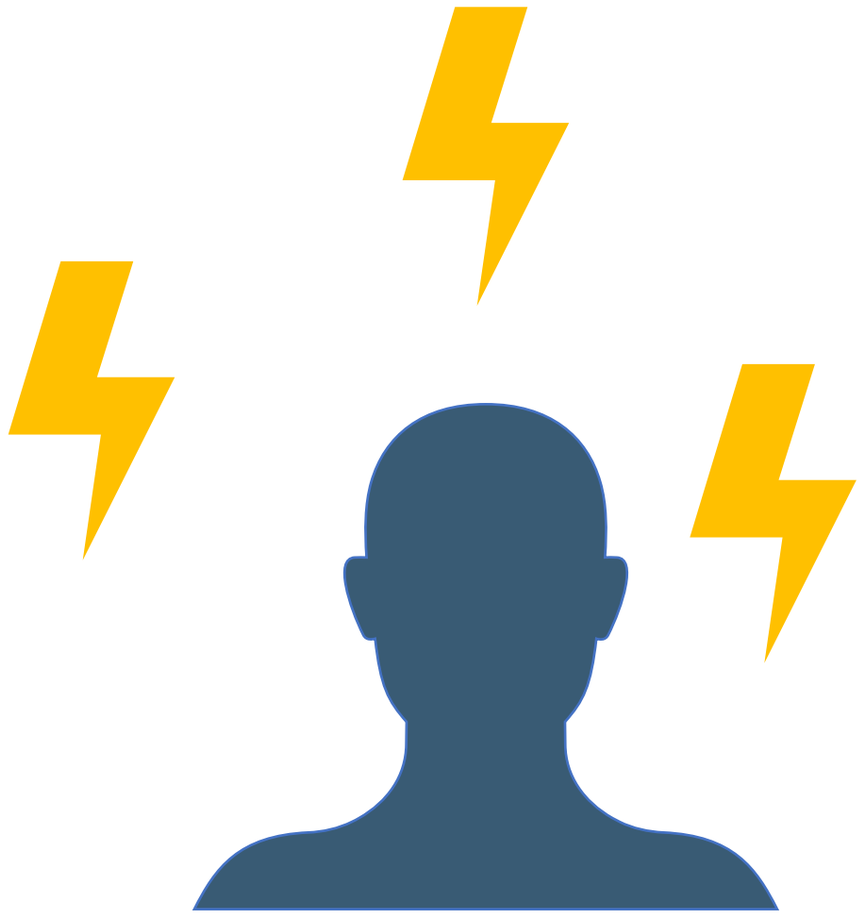
Recap – causal analysis

- In order to decide something, you need to causal knowledge
- Correlation is not sufficient for establishing causal association.
- In fact, it is even not even necessary for having causal association, but we didn't pick that up.
- Confounding is a typical problem that yields spurious correlation.

From Mediocristan to Extremistan

Some have strong feelings against prediction

Challenge 3: long-tailed variables



Nassim Taleb speaks out against (some) prediction.

Taleb, N. N. (2019). *Technical incerto, vol 1: The statistical consequences of fat tails, papers and commentaries*. https://nassimtaleb.org/2020/01/final-version-fat-tails/

„Normal“ and long-tailed variables

Mediocristan

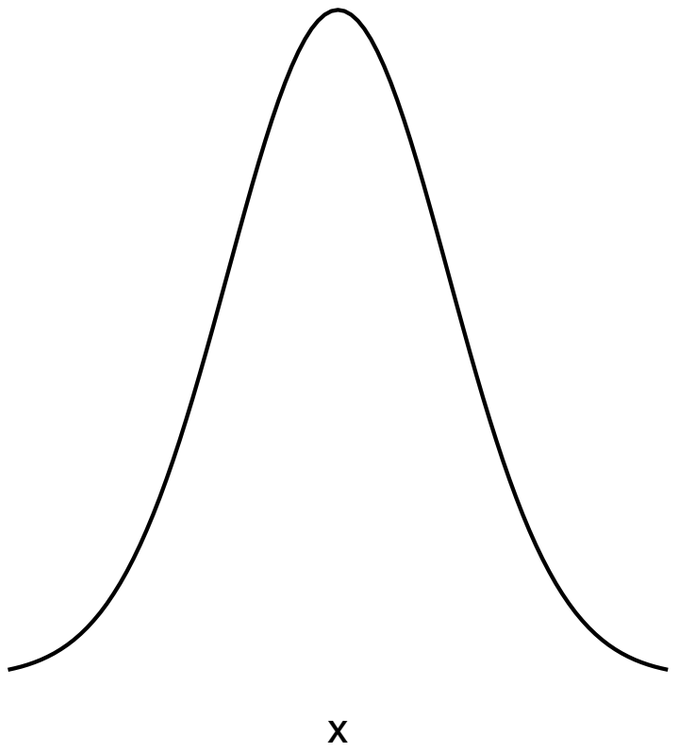
- height
- dice/coin flipping
- weight
- IQ
- blood pressure
- errors of a machine

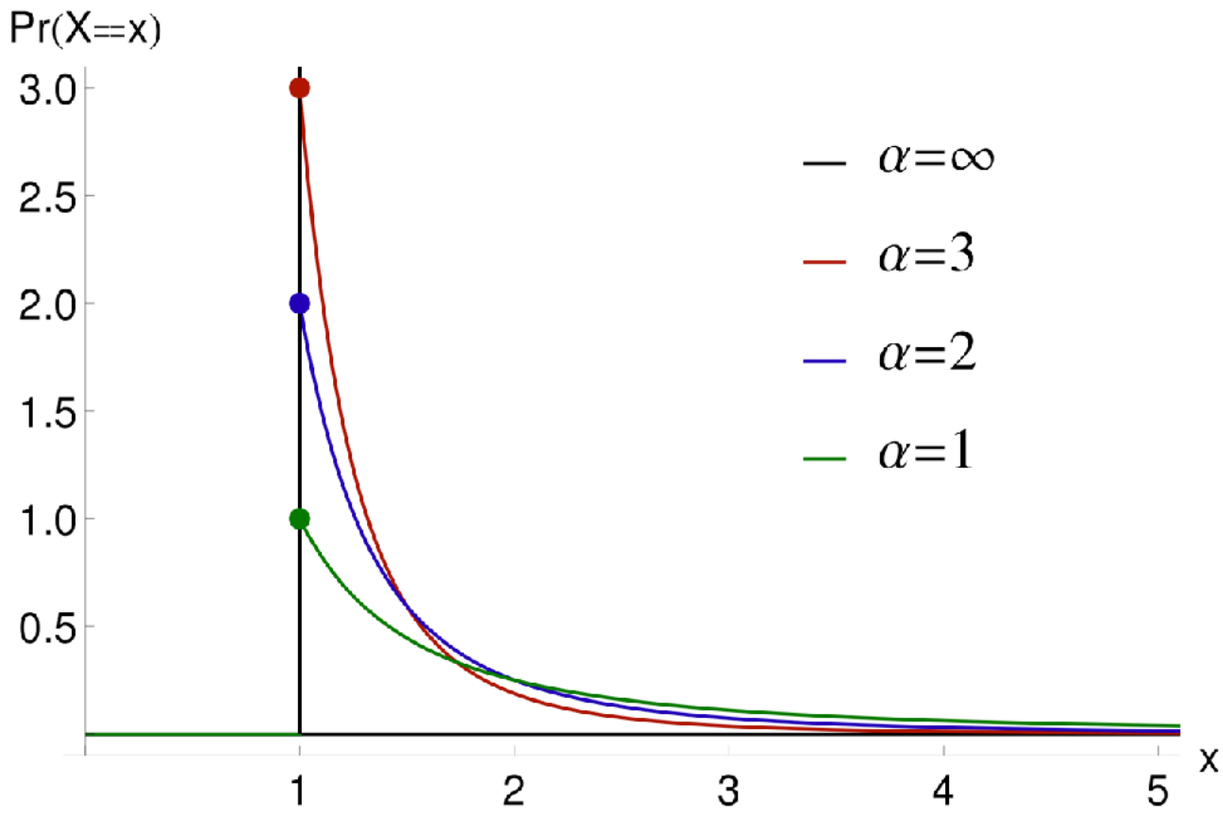
Extremistan

- wealth
- book copies sold
- fame
- stock market
- earth quakes
- pandemics
- success on tinder
- size of meteorites
- city sizes

Taleb, N. N. (2007). Black Swans and the Domains of Statistics. *The American Statistician*, *61*(3), 198–200. https://doi.org/10.1198/000313007X219996

Mediocristan and Extremisten





https://david-salazar.github.io/2020/05/26/r-squared-and-fat-tails/, https://youtu.be/_ZLrZc9NPVw

Power laws make fat tails

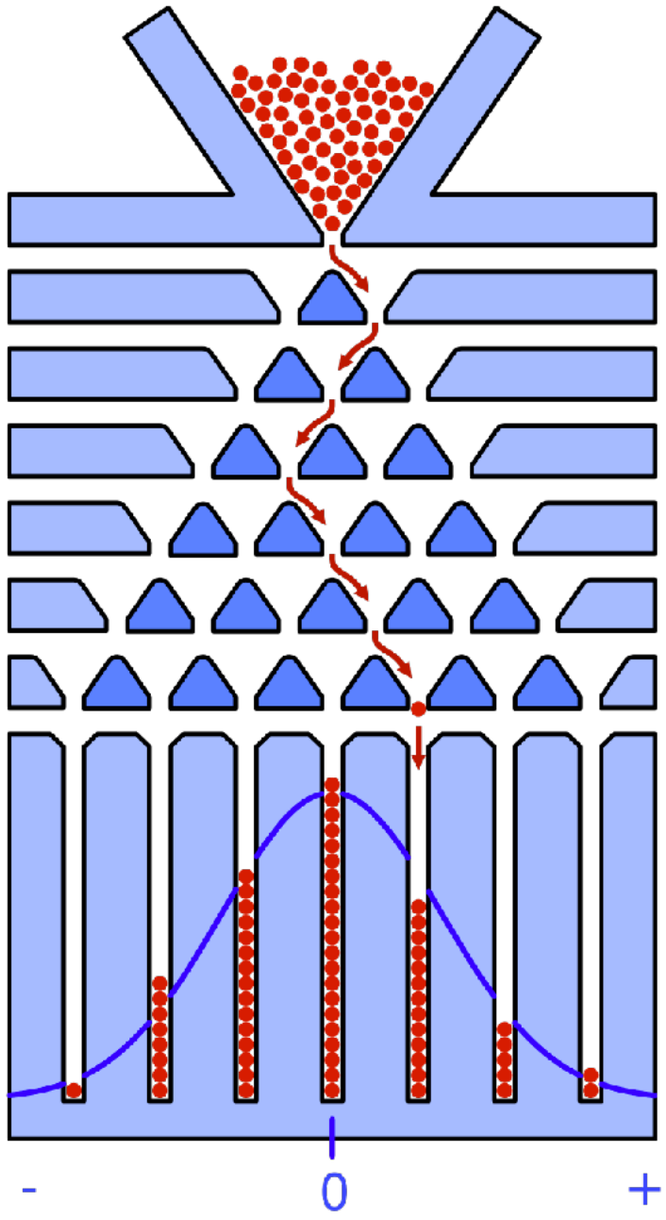
$$S(x) = 1 - P(X > x)$$

$$S(x) \approx \frac{1}{x^{\alpha}}$$

- The lower the alpha, the fatter the tail.

https://david-salazar.github.io/2020/05/19/understanding-the-tail-exponent/

What makes normal normal?



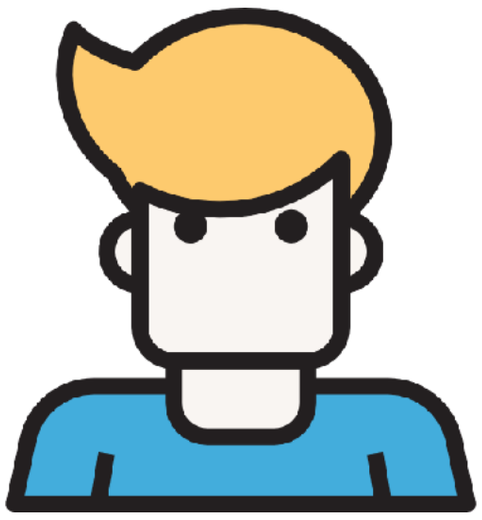
Video

https://en.wikipedia.org/wiki/Bean_machine

Intuition to Power laws

Explain the gist of such fat-tailed-stuff. In money terms. That's what I understand.

It's a Power law variable if the ratio of people with $ 16 million compared to $ 8 million is the same as the ratio of people with $ 2 million and $ 1 million



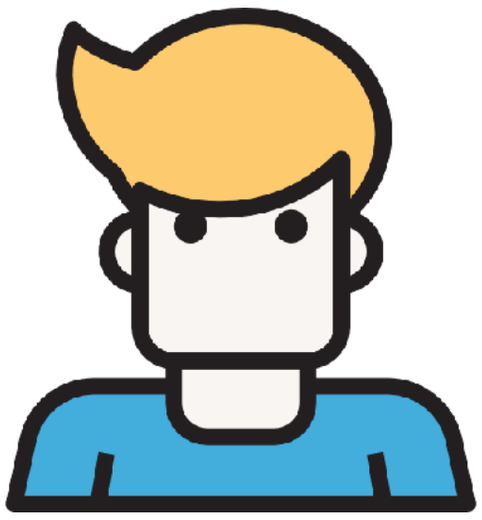
Don

Wolfi

What's Pareto?

What's Pa-Re-To?
I heard it deals with rich people.

It's a power law type, where 20% of the folks own 80% of the wealth.

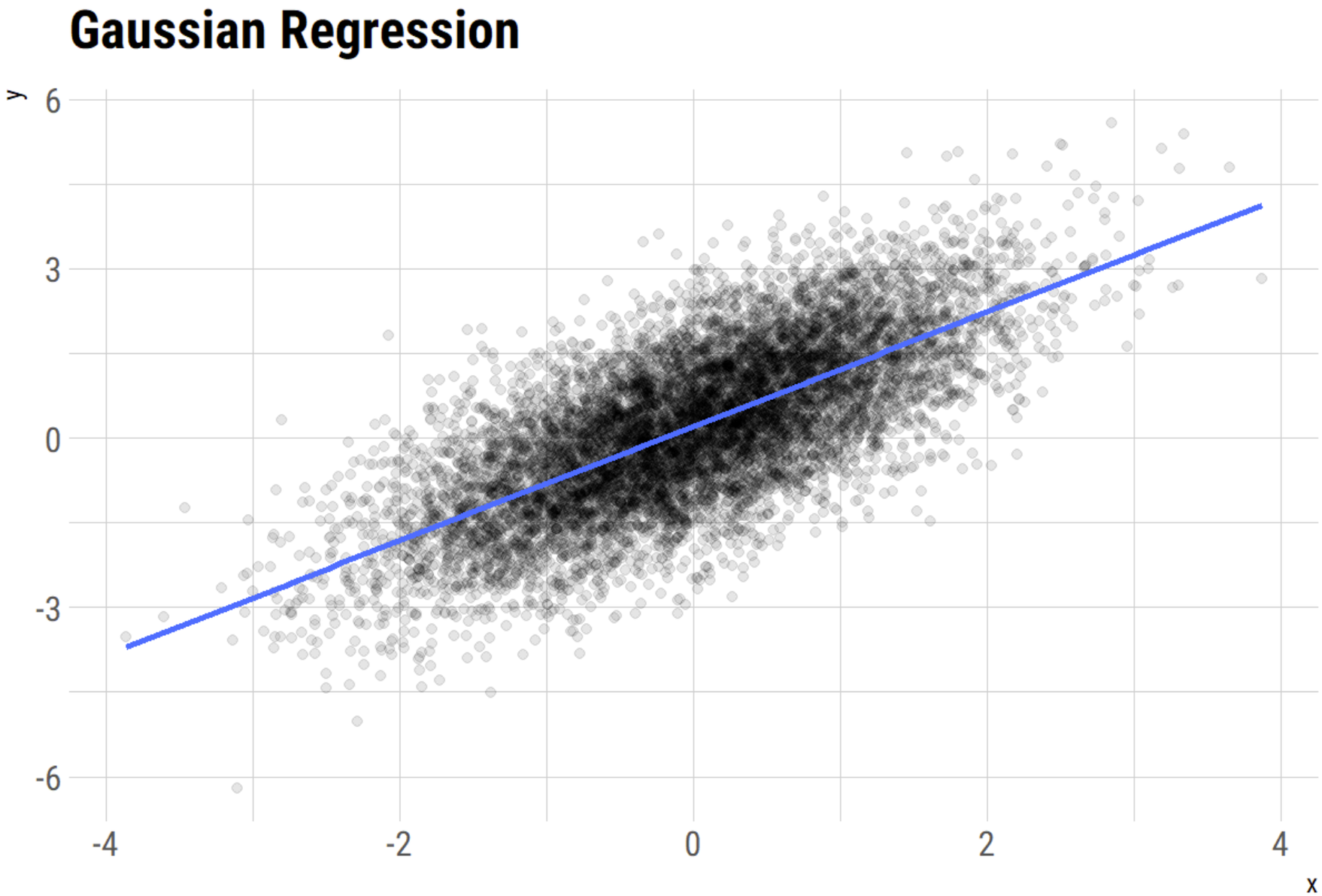


Don

Wolfi

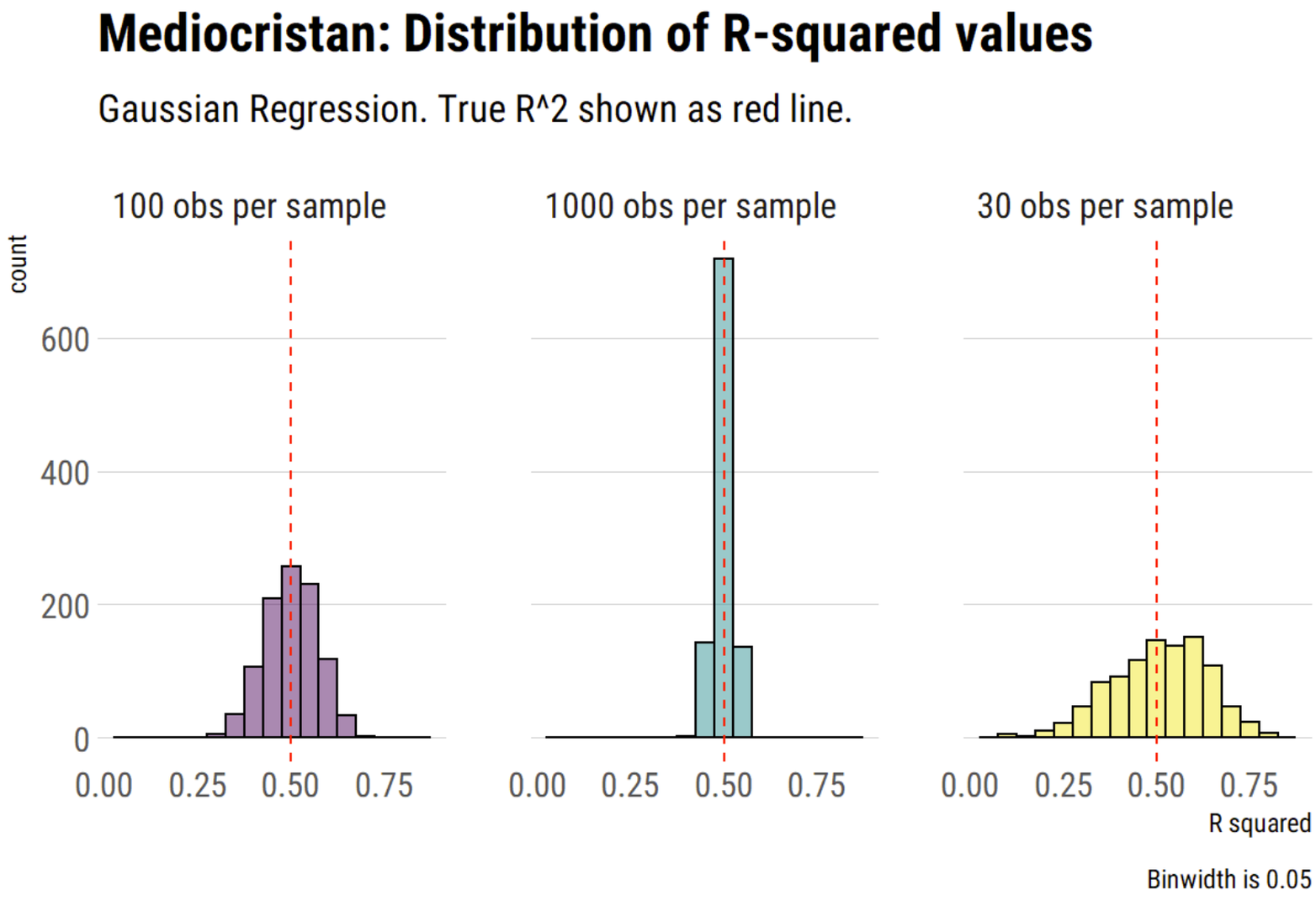
Regression with Gaussian errors (Mediocristan)

Simulation based on true (population) R^2 = 50%



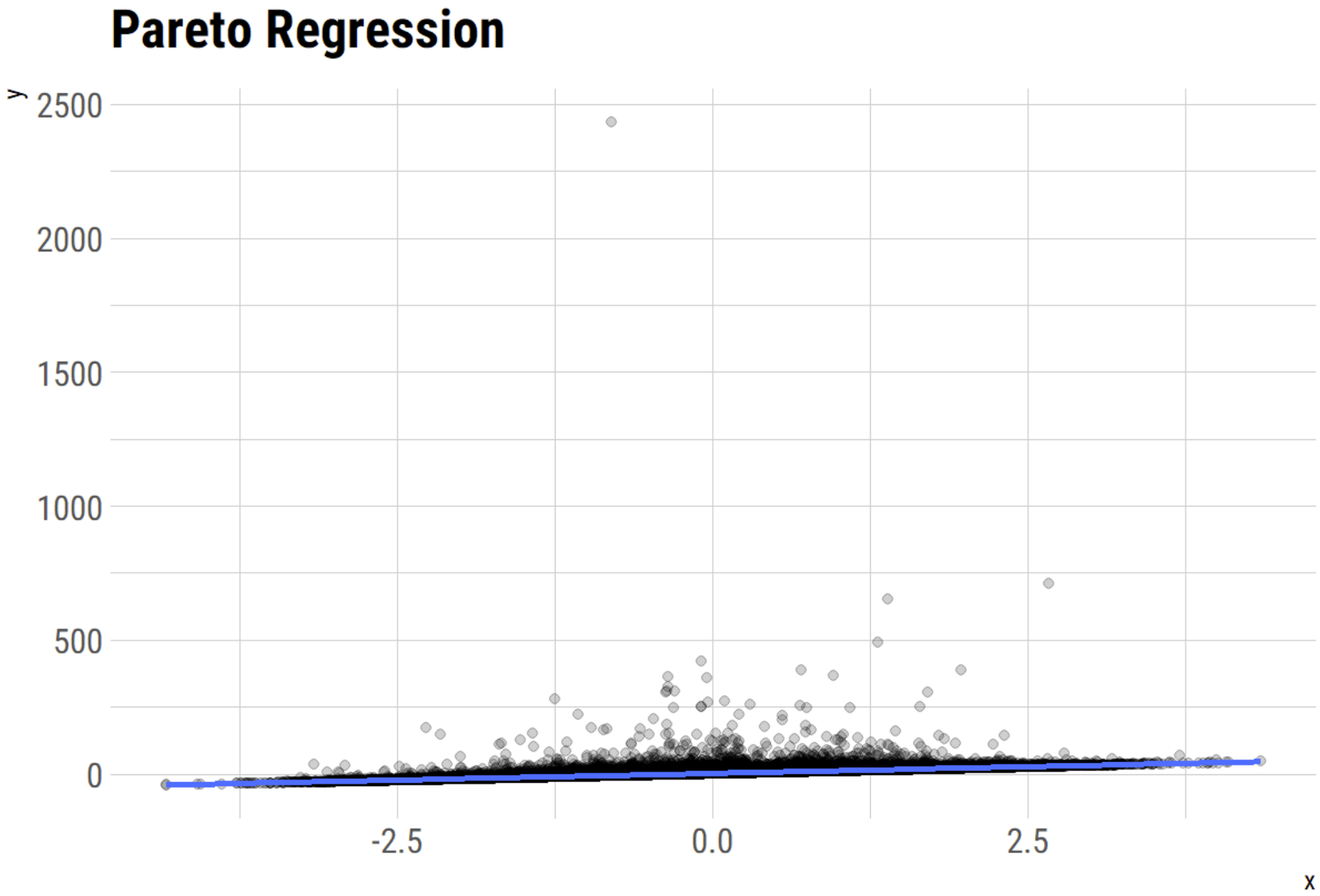
Regression with Gaussian errors: works!

Simulated samples recover true R^2.



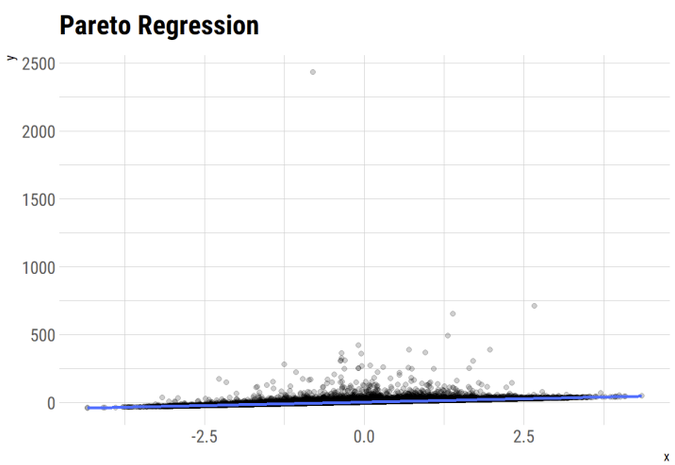
Regression with Pareto errors: Extremisten

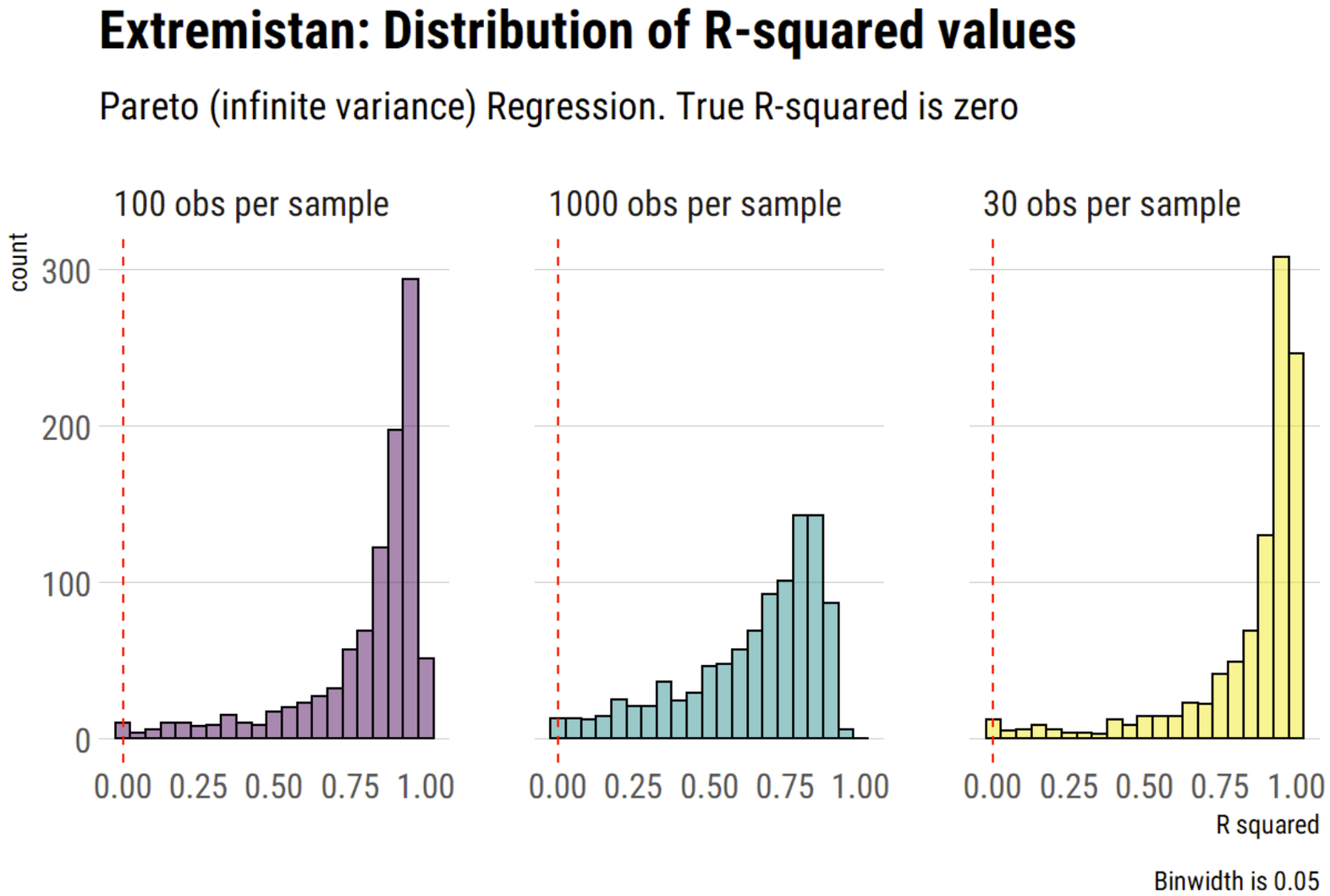
Simulated based on $R^2 = 0\%$



Regression with Pareto errors: FAILS!

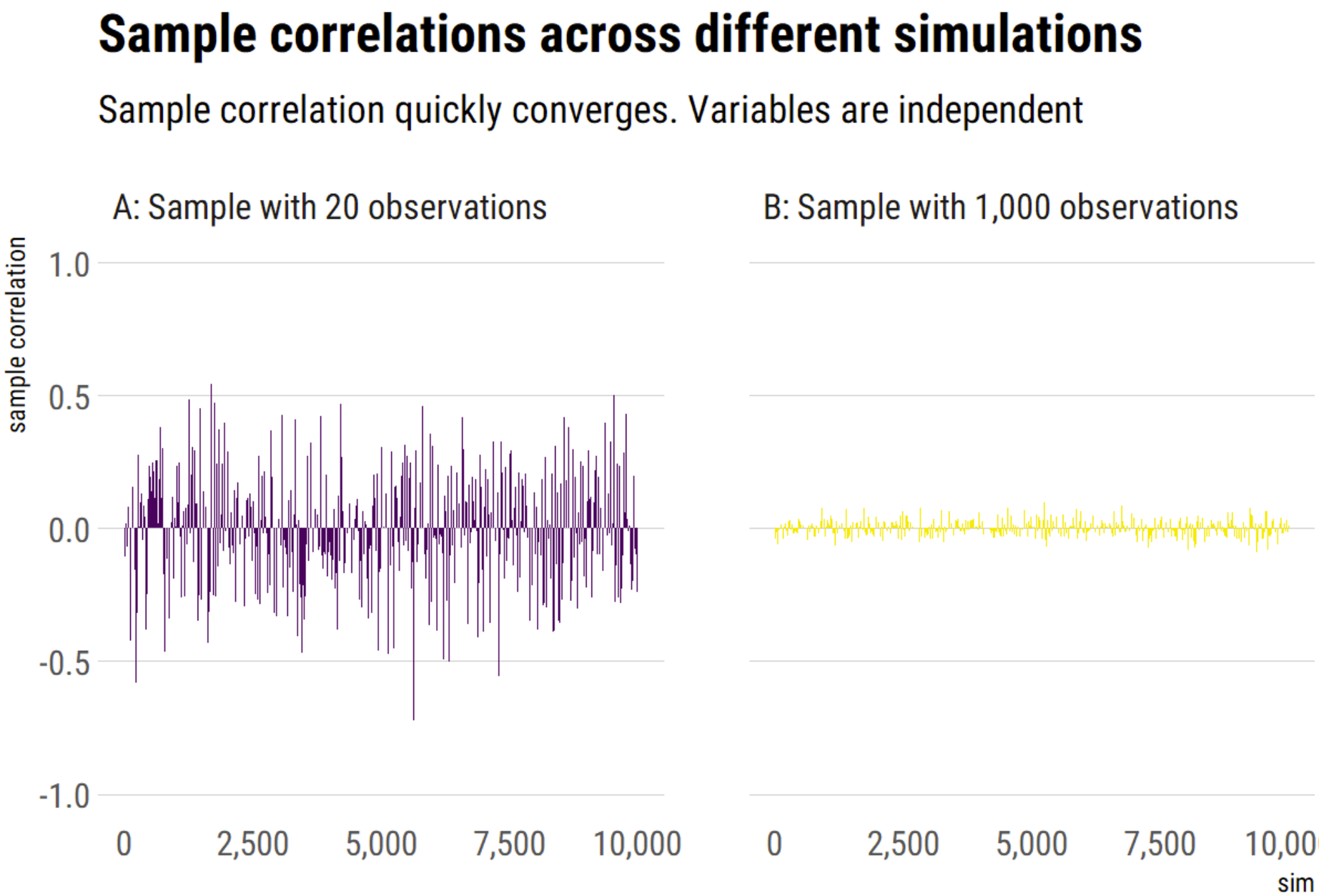
Simulated samples FAIL DRAMATICALLY to recover true R^2.





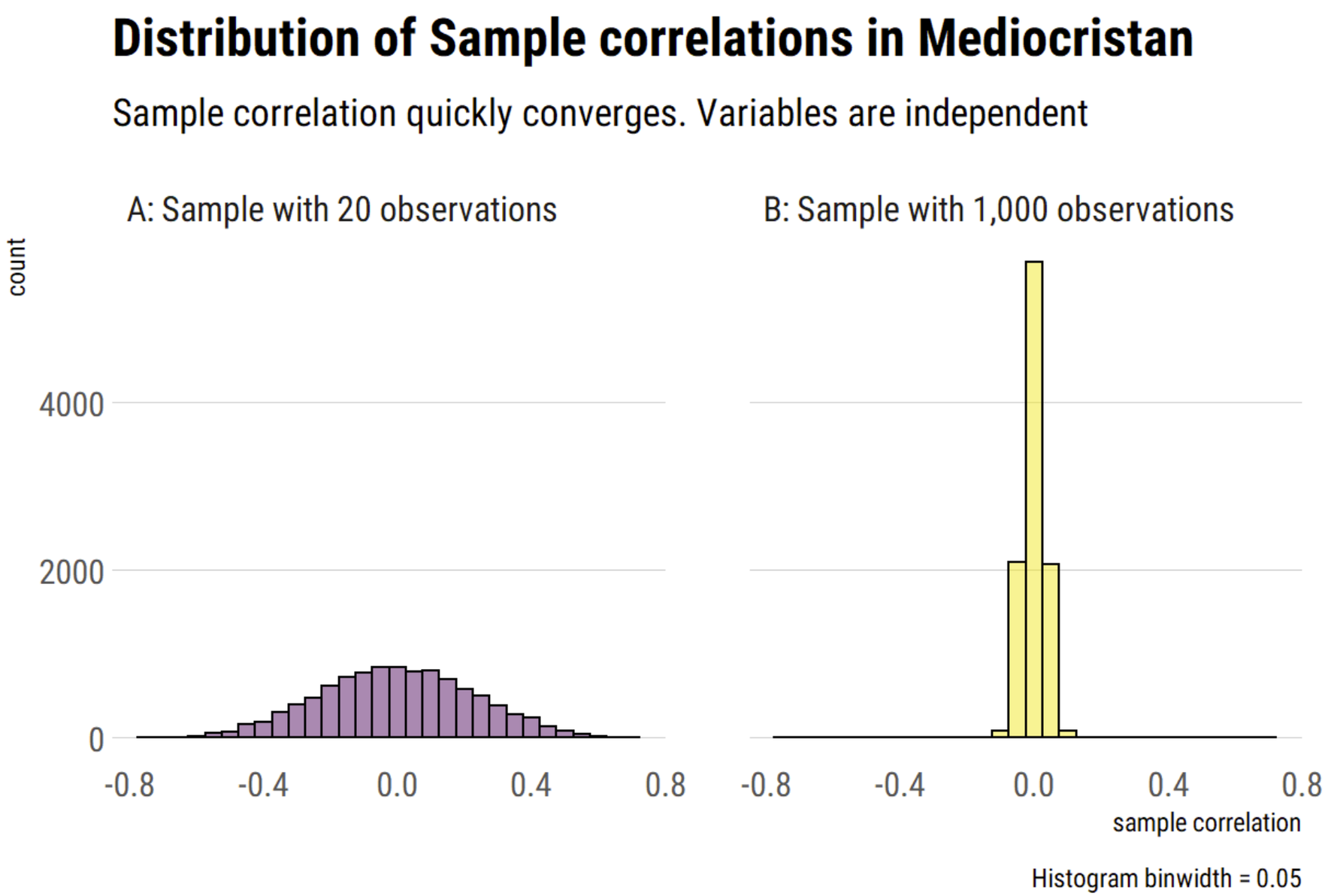
Correlation in Mediocristan

Simulation based on true (population) r = 0



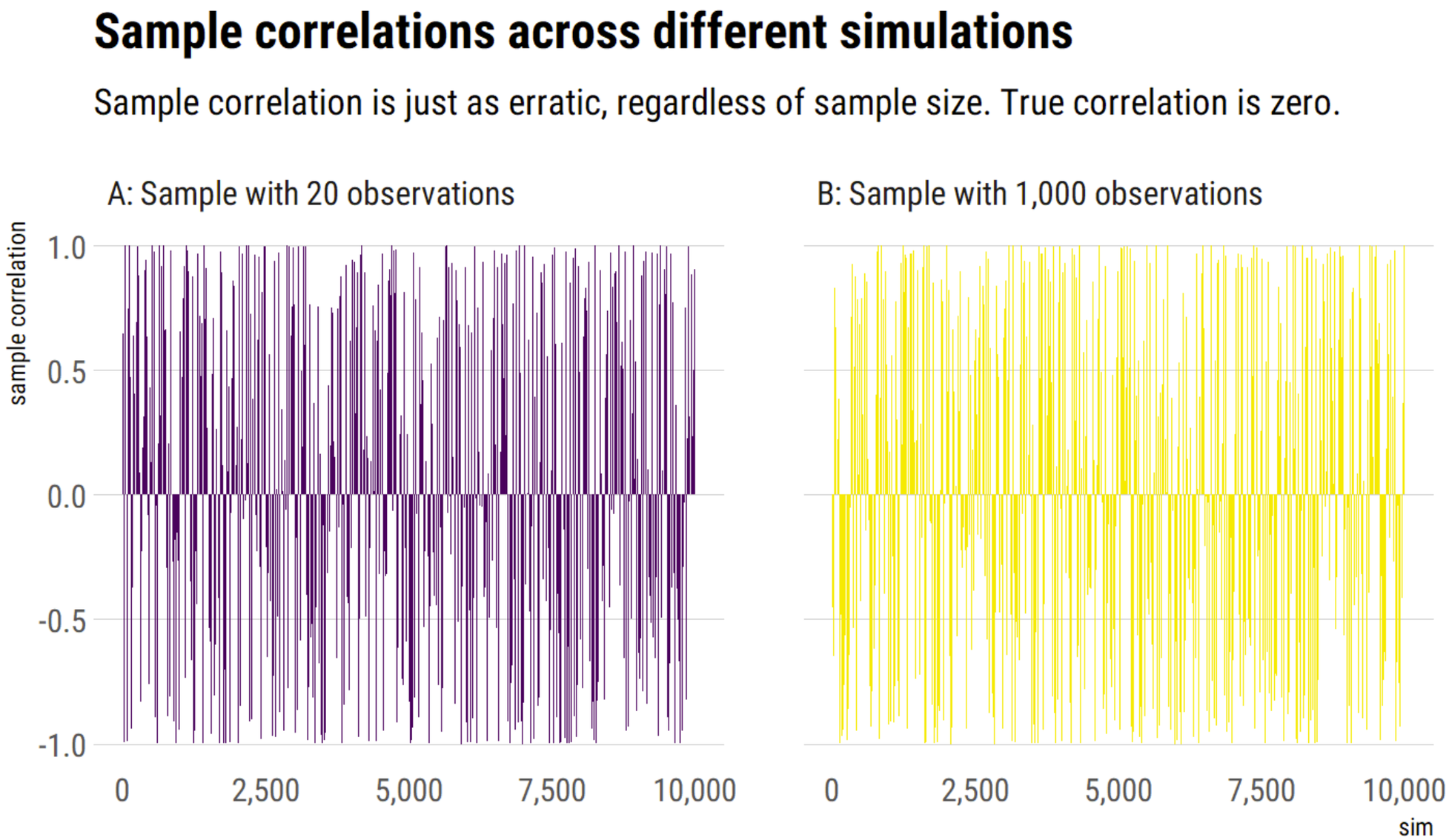
Correlation in Mediocristan

Simulation based on true (population) r = 0



Correlation in Extremistan

Simulation based on true (population) r = 0

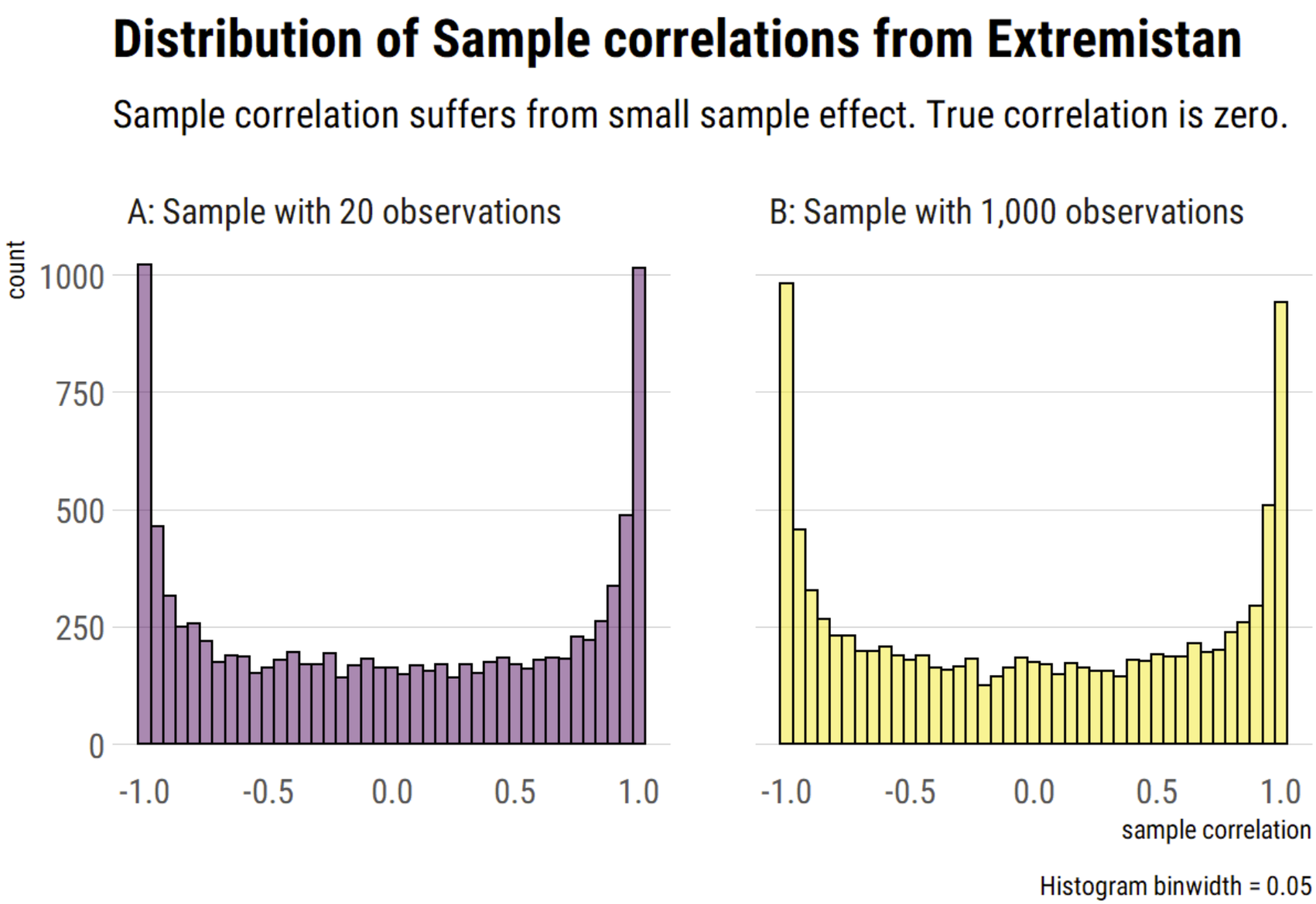


bivariate t-student distribution with exponent 2/3

https://david-salazar.github.io/2020/05/22/correlation-is-not-correlation/

Correlation in Extremistan

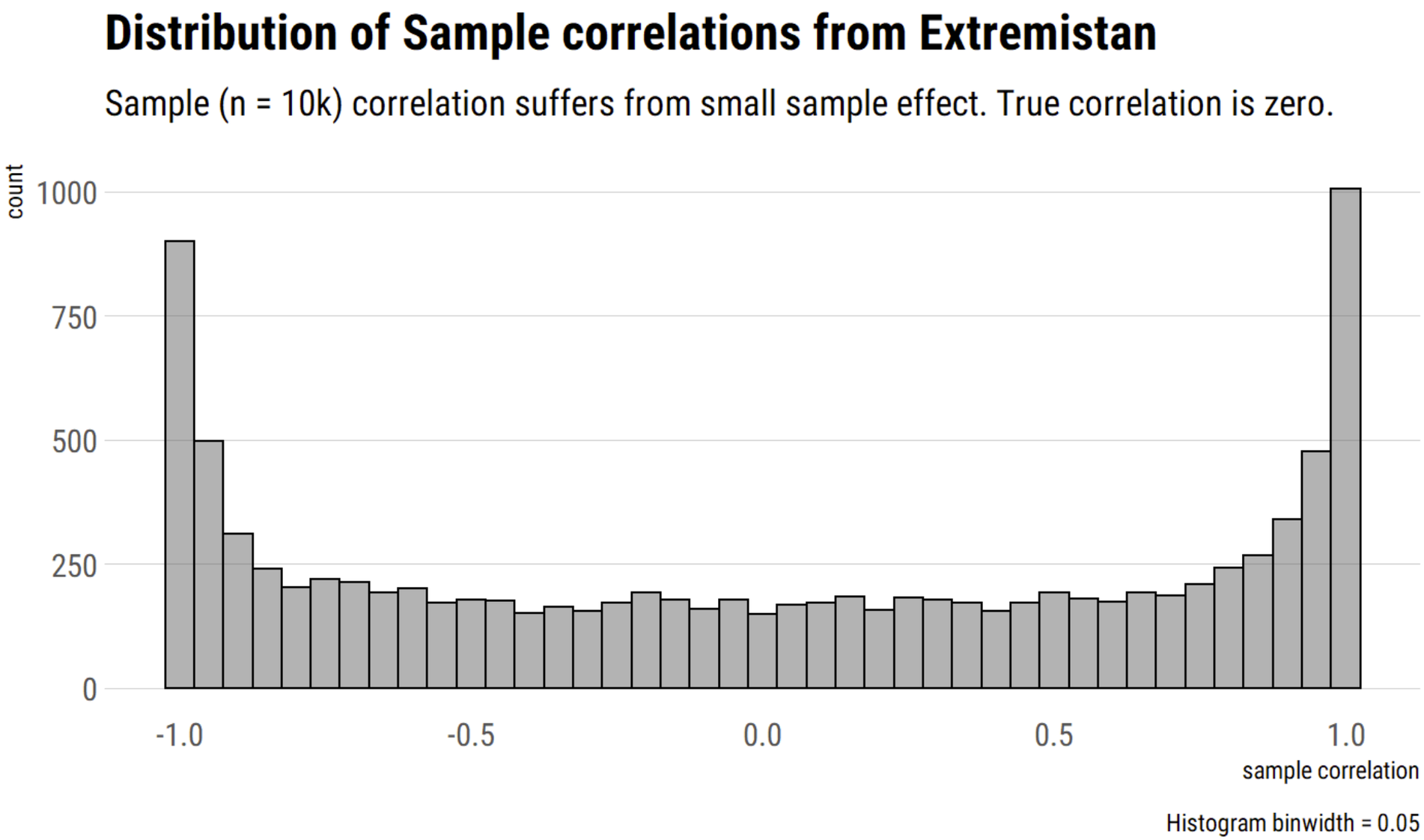
Simulation based on true (population) r = 0



https://david-salazar.github.io/2020/05/22/correlation-is-not-correlation/

Even yuuge samples do not help in Extremistan

Everybody go home.



https://david-salazar.github.io/2020/05/22/correlation-is-not-correlation/

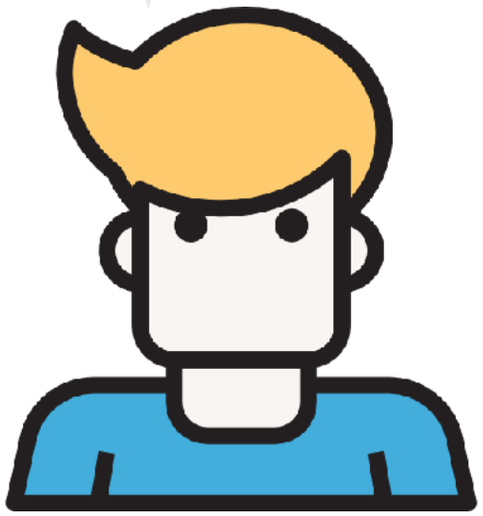
Recap – Extremistan

- Check-out whether your variables are normal (or thin-tailed) or long-tailed.
- Typical data-analysis methods work well only for thin-tailed variables (Mediocristan).
- Estimation and prediction breaks down in Extremistan. Maybe better go home and don’t speak about it.

Self-learning paths

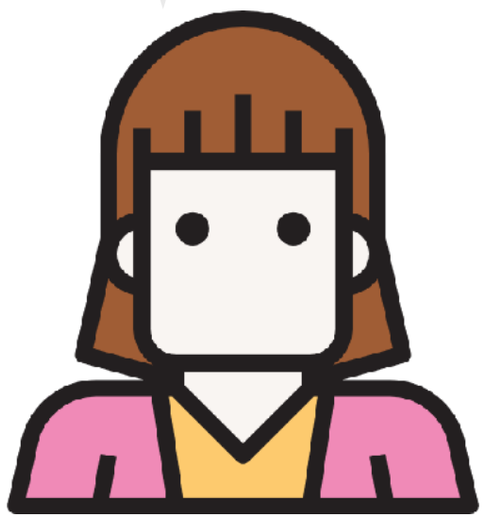
OK, everybody give some recommendation

Oh yeah!



Don

It’s a long path.

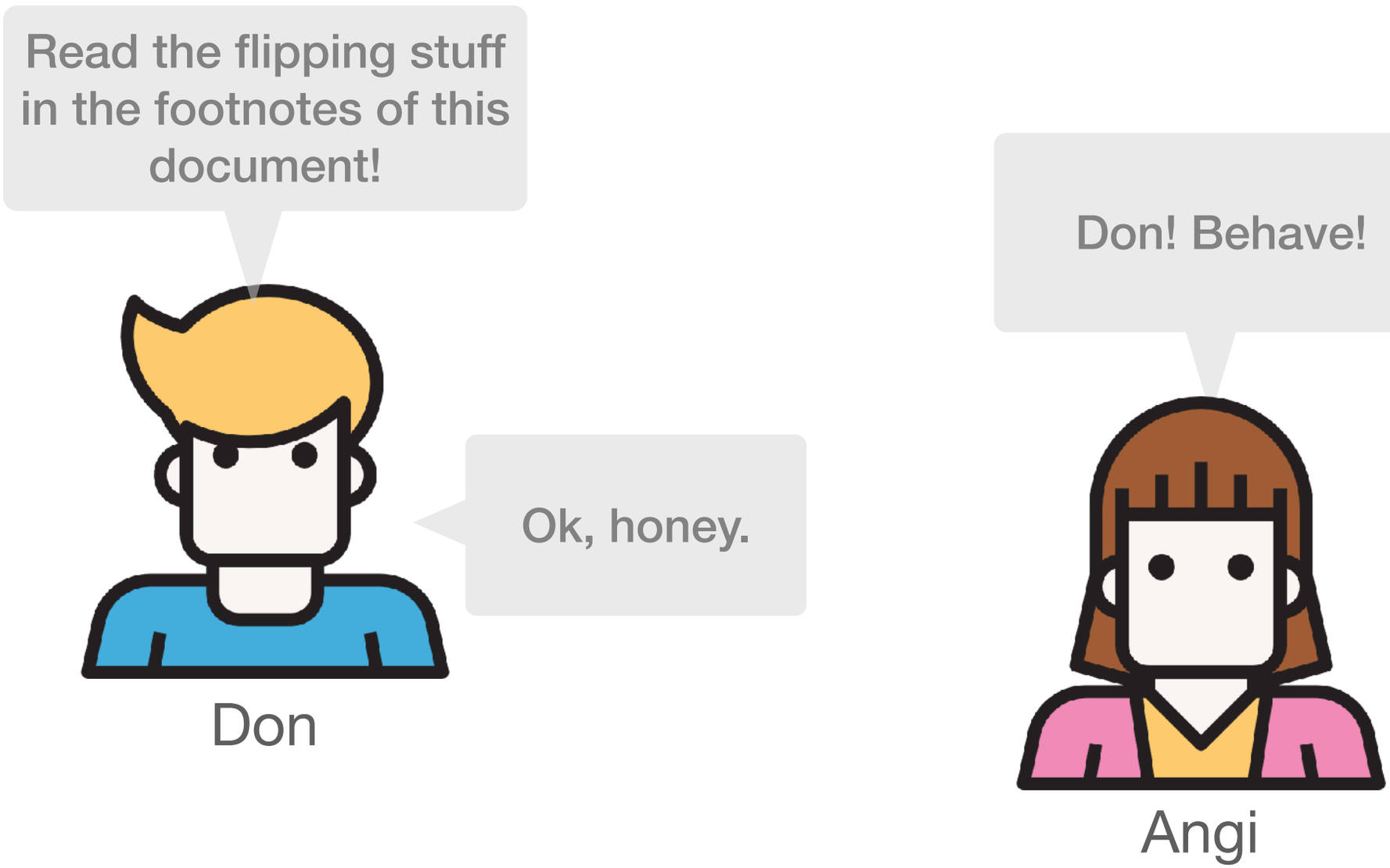


Angi

Do you really want that?

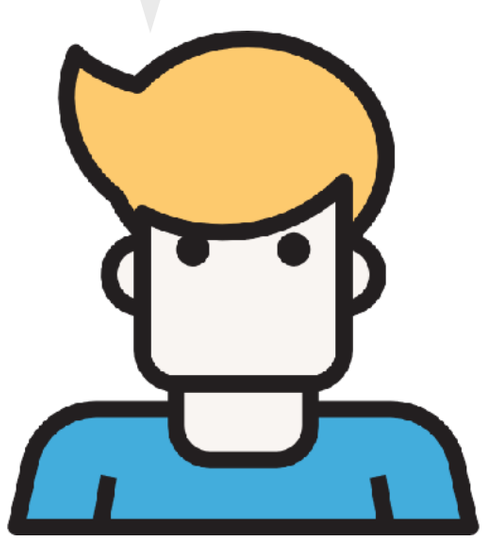
Wolfi

Don's advise

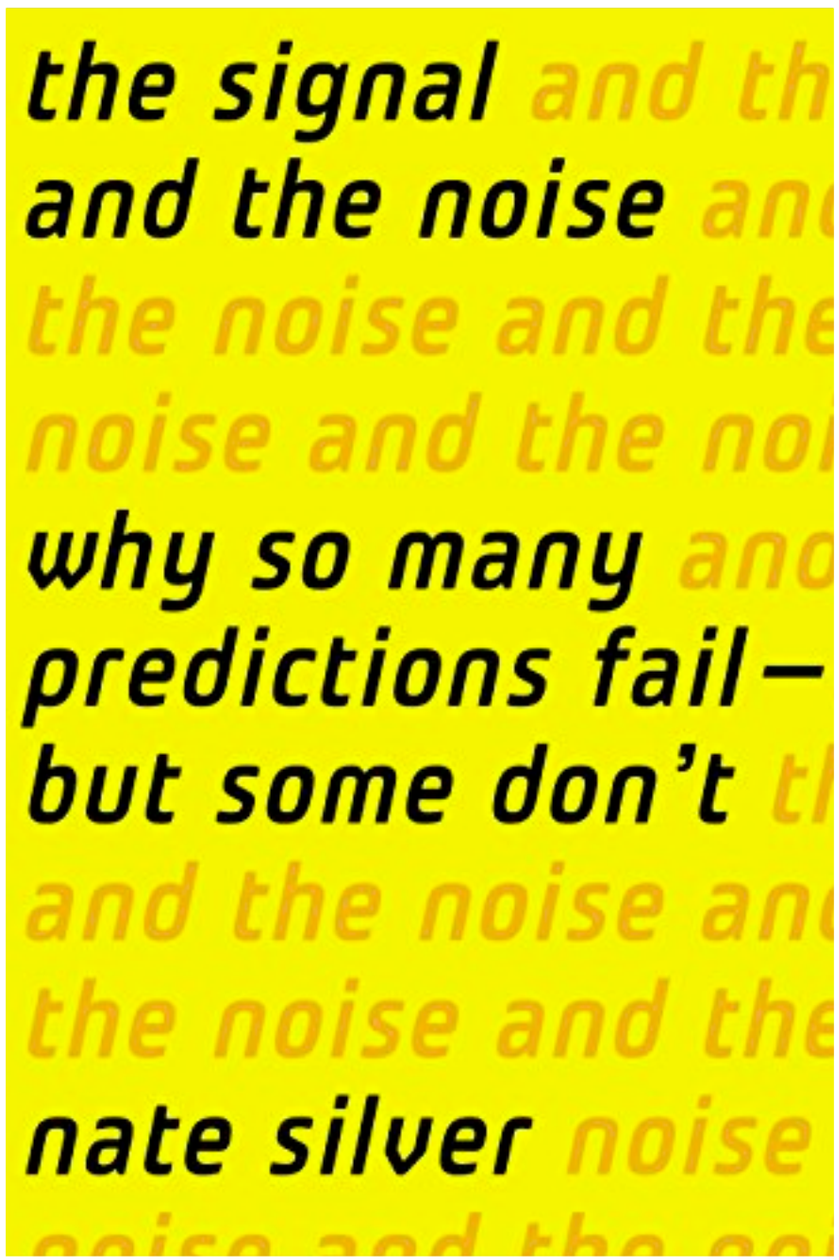

Read the flipping stuff in the footnotes of this document!
Ok, honey.
Don
Don! Behave!
Angi

Nate Silver: The Signal and the Noise

OK. Cool read if you don't like math.



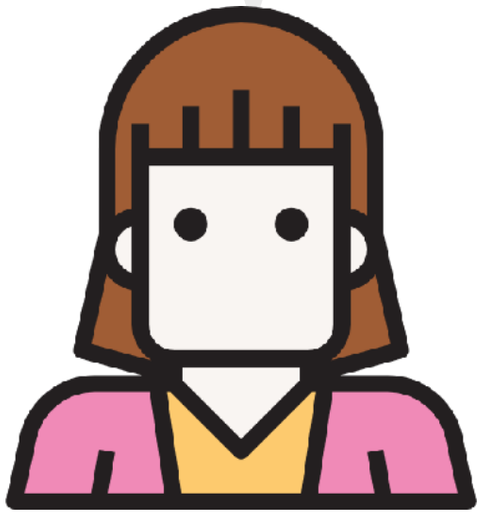
Don



Silver, N. (2012). *The signal and the noise: The art and science of prediction*. Penguin UK. https://en.wikipedia.org/wiki/The_Signal_and_the_Noise

Coursera's data science courses

Cool if you like practical competence.



Angi

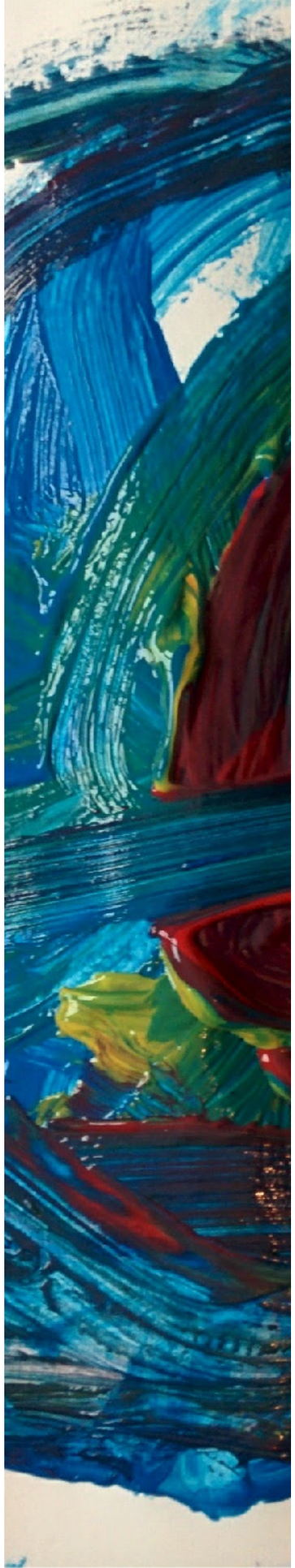
coursera

Nassim Taleb: Enfant Terrible of intellectual establishment

Listen to Nassim Taleb if you like strong stuff.

Wolfi

- Taleb on Twitter
- Podcast
- Video on fat tails



Sebastian Sauer

sebastian.sauer@data-divers.com

data-divers.com