Statistical uncertainty

Pärt Prommik, PhD

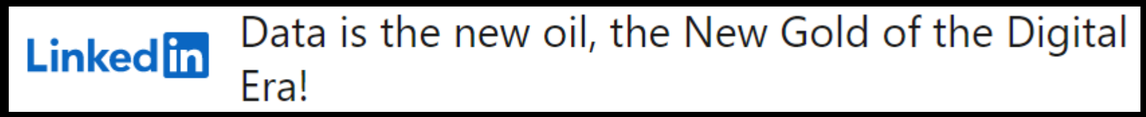
Ülo Maiväli, PhD

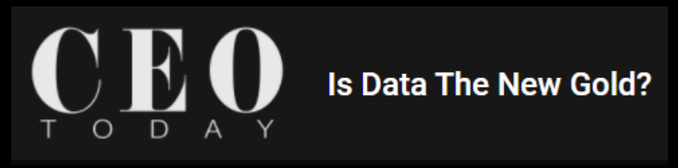
Plan for today

- Why do we need statistical uncertainty?
- Why we need R (and tidyverse)?
- Practice

id	y	x
a	0	2
b	1	1
c	2	0

Data is the new gold/oil





…when it is well-analysed

Harvard Business Review

Data Is Useless Without the Skills to Analyze It

by Jeanne Harris

tds Published in Towards Data Science

Is Data Really the New Oil in the 21st Century?

Exploring the strengths and limitations of this metaphor in the information age.

GW PRIME YOUR WHERE IS HERE

Data was, Analytics is, the New Oil

Technology innovation and the decreasing cost of data processing are fuelling the demand for actionable intelligence across sectors.

E

Without Good Analysis, Big Data Is Just a Big Trash Dump

Forbes

Data, Like Oil, Is Pretty Useless In Its Raw State

LinkedIn Big Data is Useless without Visual Analytics

Cisco UK & Ireland Blog > Innovation

Innovation

Data is neither the new oil nor the new gold…

Stu Higgins
October 9, 2018

…and is only valuable when gathered correctly and interpreted accurately

towards data science

Data is not the new oil

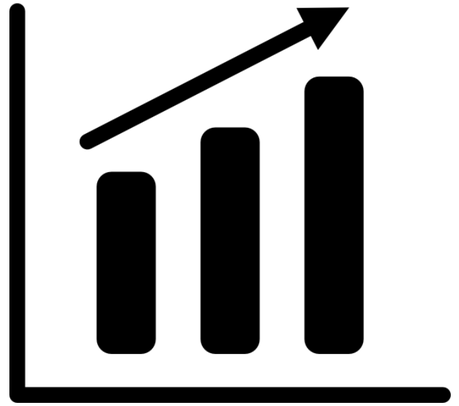
About the reality of working with data

Summary: data is not the new oil

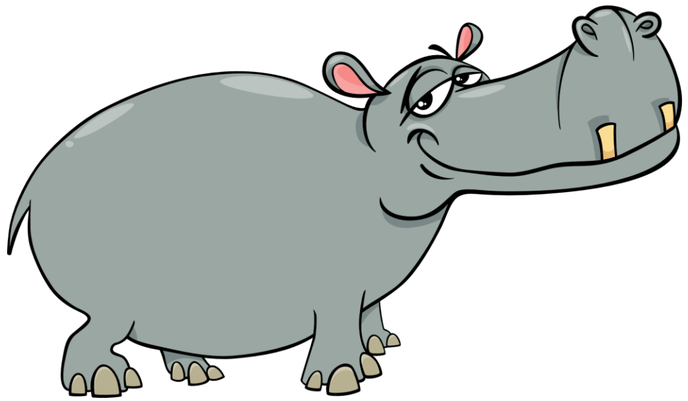
The process of working with data is messy, requires careful planning, engineering, and research, and contains a lot of unknowns and pitfalls. Most importantly, it is not always clear how to leverage the data, since data by itself is too noisy to provide value. Equating data to oil is neglecting this messy and complicated reality.

Data is meaningless without analysis

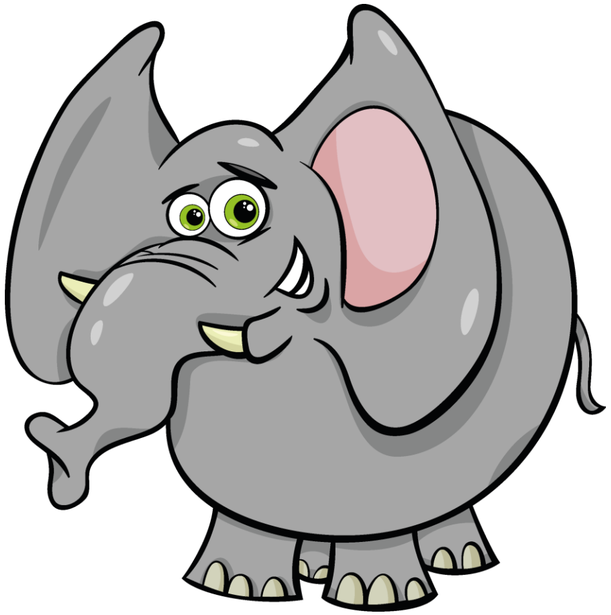
id	y	x
a	0	2
b	1	1
c	2	0



Who run faster, elephants or hippos?



VS



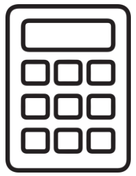
130,000

440,000

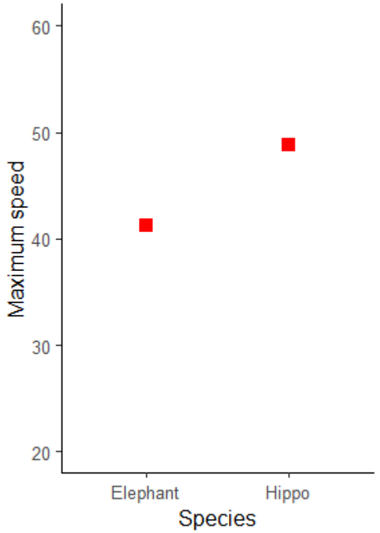
Calculating point estimates

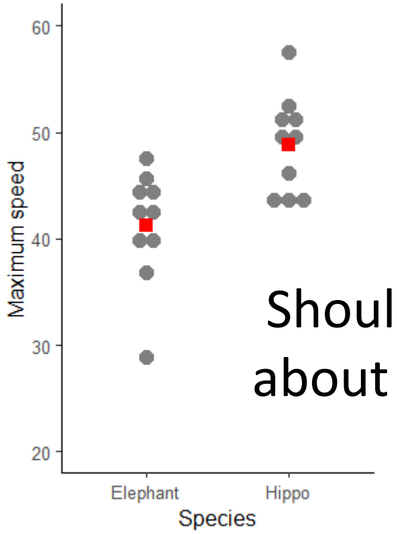
id	y	x
a	0	2
b	1	1
c	2	0

Species	Max Speed
Hippo	47.8
Hippo	48.1
Hippo	33.5
Hippo	55.0
Hippo	45.0
Hippo	43.9
Hippo	43.3
Hippo	50.5
Hippo	53.8
Hippo	56.2
Elephant	42.8
Elephant	42.3
Elephant	37.3
Elephant	43.7
Elephant	40.3
Elephant	50.8
Elephant	39.4
Elephant	35.4
Elephant	37.5
Elephant	35.4



Species	Mean Max Speed
Hippo	48.8
Elephant	37.7





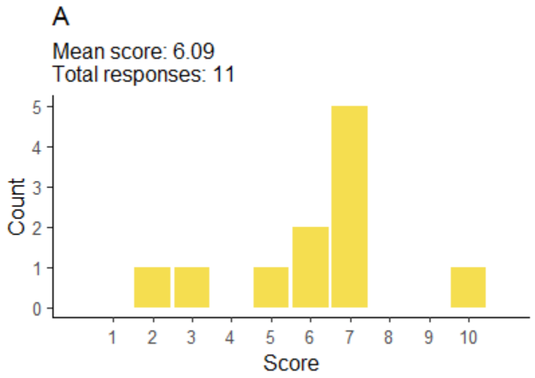
Should we care about variation?

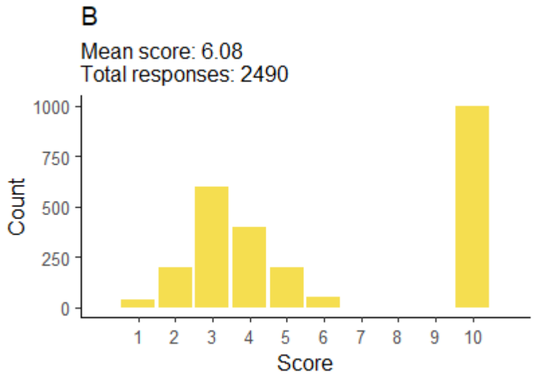
Choose the best movie for your movienight

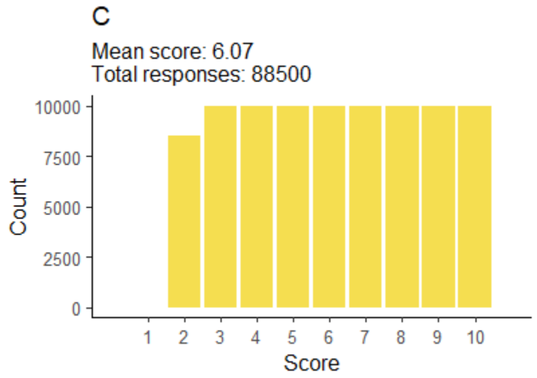
Movie title	Mean IMDB score
A	6.09
B	6.08
C	6.07
D	6.06
E	6.05
F	6.04
G	6.03
H	6.02

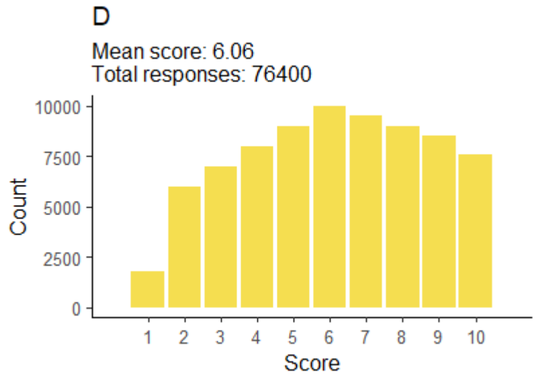
https://www.countryliving.com/life/entertainment/g21240020/fall-movies/

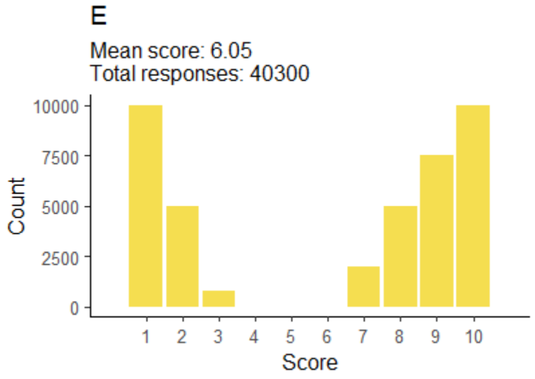
IMDB score distributions

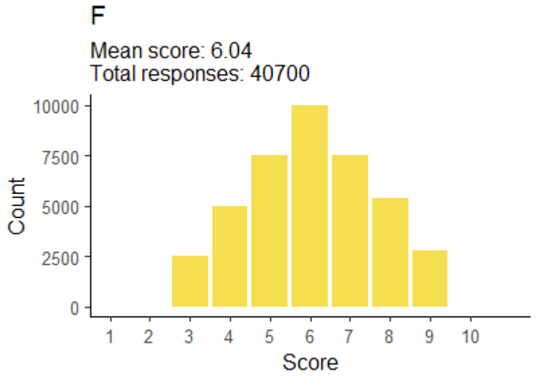


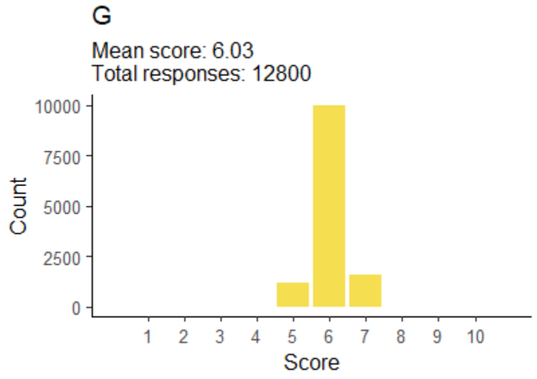


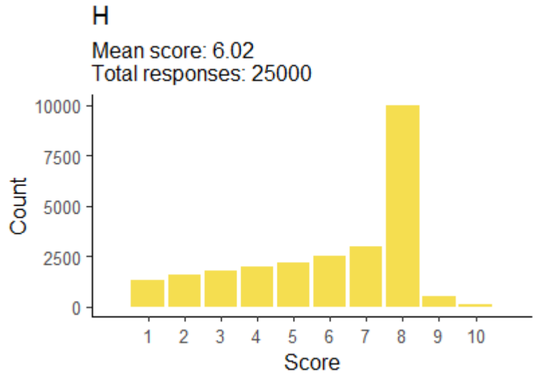






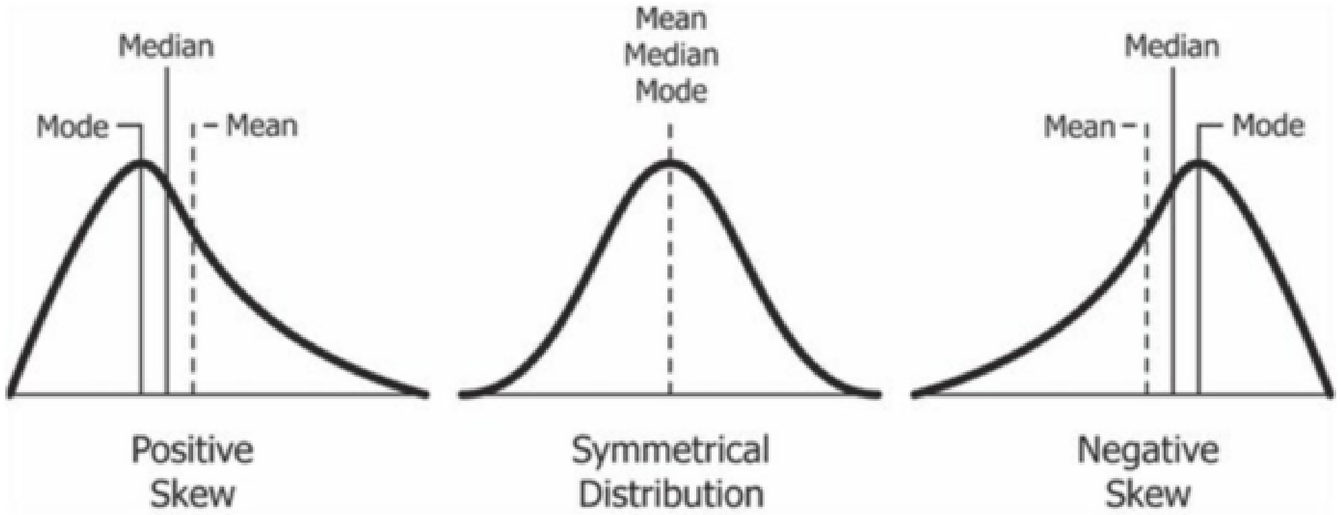






Choosing correct measures for desrcibing a distribution

Central tendency



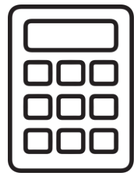
Variation

- Mean
  - Standard deviation
- Median/Mode/Geometric mean
  - Intequartile range
  - Range of percentile
  - Median absolute deviation
  - Geometric standard deviation
  - Minimum and maximum

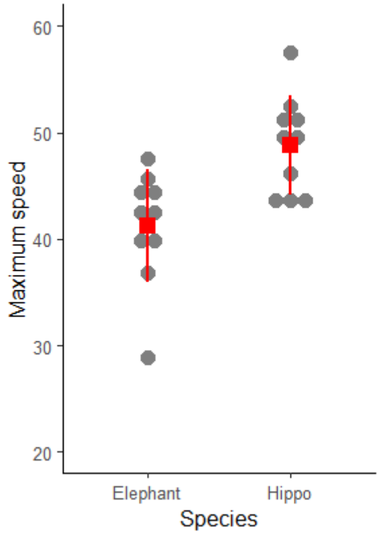
Adding variation is informative

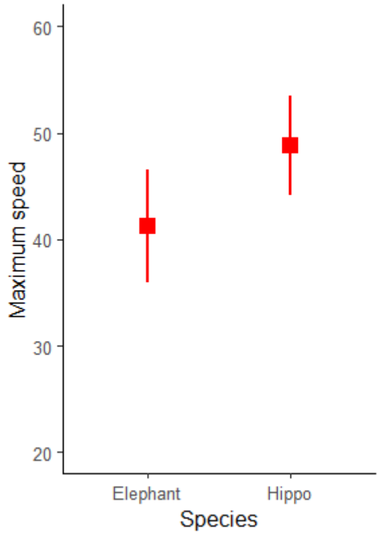
id	y	x
a	0	2
b	1	1
c	2	0

Species	Max Speed
Hippo	47.8
Hippo	48.1
Hippo	33.5
Hippo	55.0
Hippo	45.0
Hippo	43.9
Hippo	43.3
Hippo	50.5
Hippo	53.8
Hippo	56.2
Elephant	42.8
Elephant	42.3
Elephant	37.3
Elephant	43.7
Elephant	40.3
Elephant	50.8
Elephant	39.4
Elephant	35.4
Elephant	37.5
Elephant	35.4



Species	Mean Max Speed (SD)
Hippo	48.8 (5.3)
Elephant	37.7 (4.7)





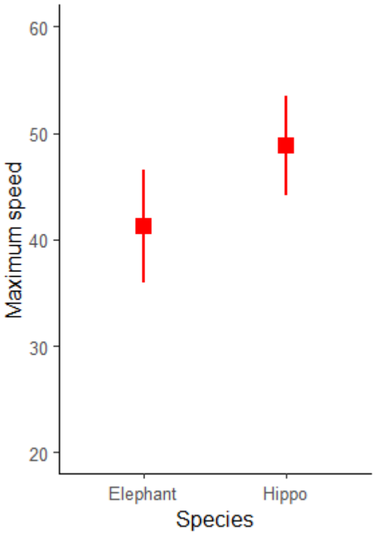
But what about the difference?

id	y	x
a	0	2
b	1	1
c	2	0

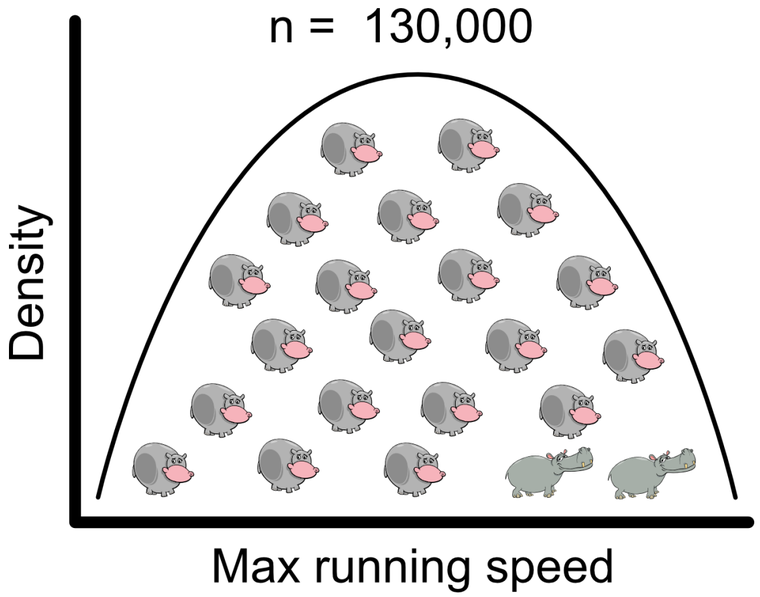
Species	Max Speed
Hippo	47.8
Hippo	48.1
Hippo	33.5
Hippo	55.0
Hippo	45.0
Hippo	43.9
Hippo	43.3
Hippo	50.5
Hippo	53.8
Hippo	56.2
Elephant	42.8
Elephant	42.3
Elephant	37.3
Elephant	43.7
Elephant	40.3
Elephant	50.8
Elephant	39.4
Elephant	35.4
Elephant	37.5
Elephant	35.4

Species	Mean Max Speed (SD)
Hippo	48.8 (5.3)
Elephant	37.7 (4.7)

Hippos run 10.7 k/h faster
(in this sample)



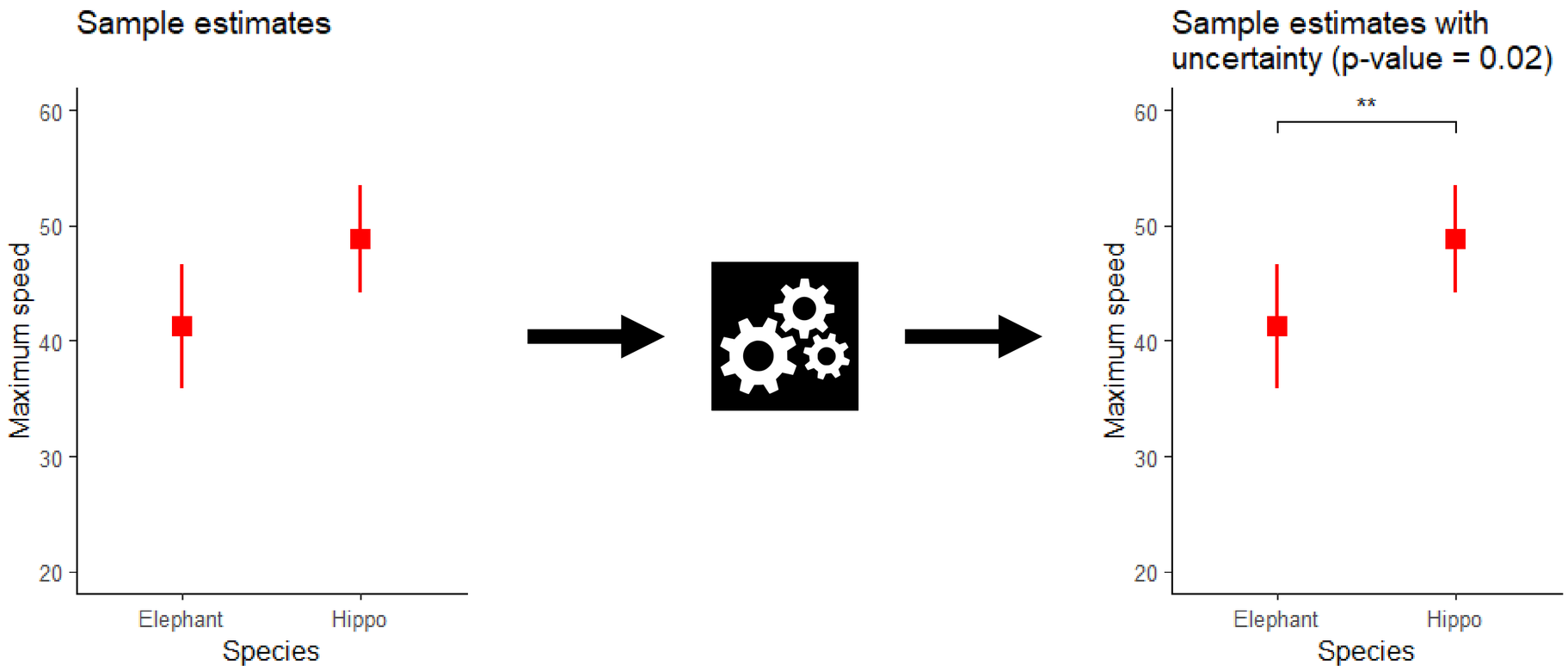
But how <u>certain</u> we are that hippos run faster in these populations?



This is why we need statistics

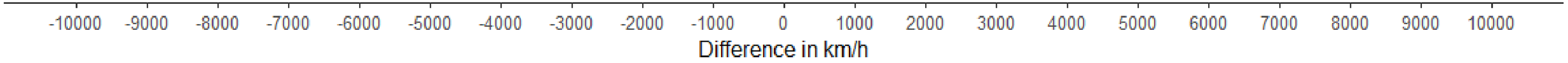
Inferential statistics is about using **sample data**
to make generalitations about a **population**.

Let’s add uncertainty (use statistics)



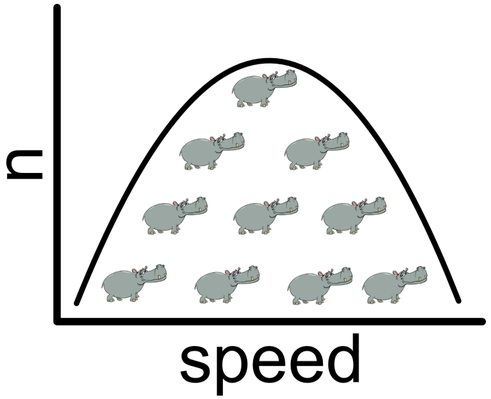
This is hypothesis testing: we have 2% probability that their maximum running speed difference could have just occured by random chance
(i.e. that the null hypothesis (Hippos = Elephants) is true).

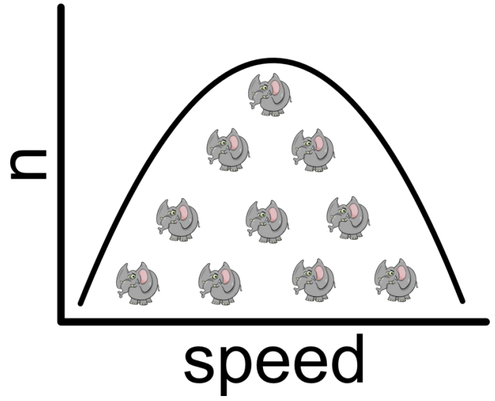
Not very intuitive and informative!

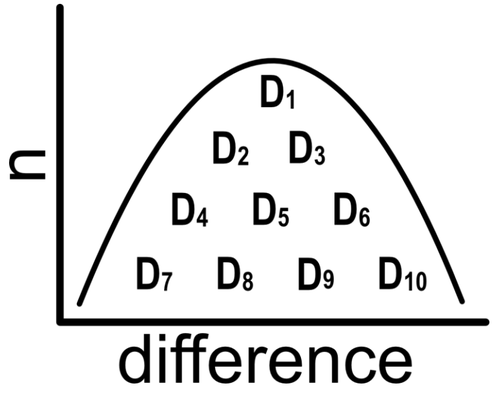


The maximum running speed of group A and B is different (p = 0.02).

Sample and population level difference

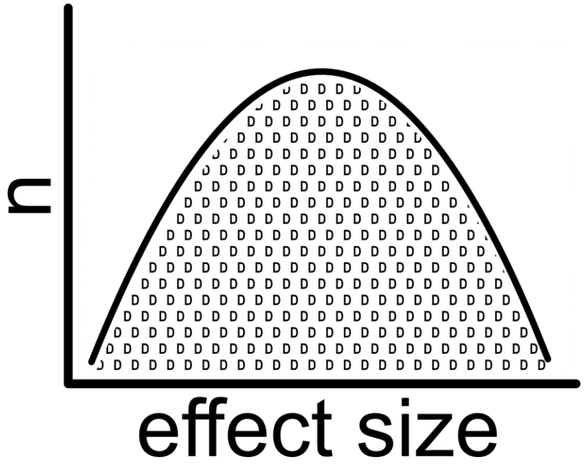






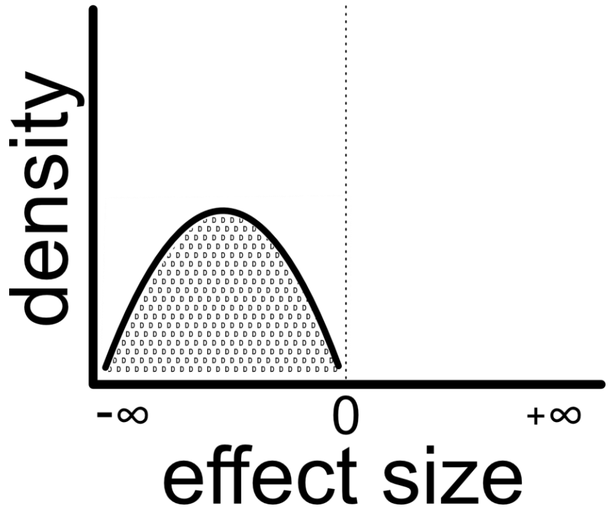
Hippo	Elephant	Difference (D)
44.2	47.6	-3.3
46.2	43.0	3.2
43.1	39.8	3.3
50.0	45.6	4.4
43.0	36.9	6.1
49.1	41.9	7.2
51.5	44.1	7.3
52.4	44.7	7.7
50.9	39.9	11.0
57.6	28.9	28.6

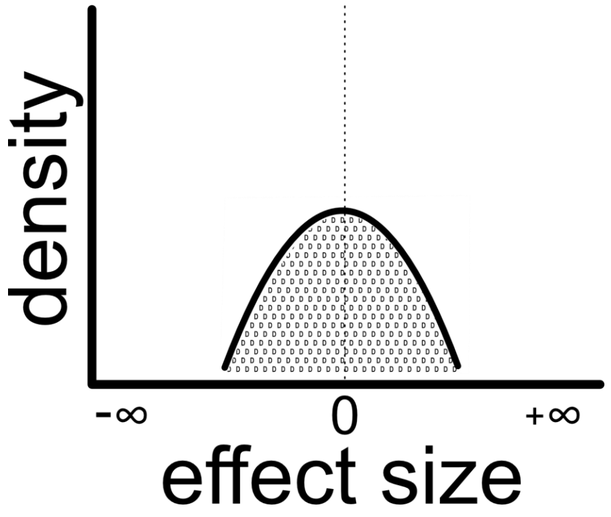
id	y	x
a	0	2
b	1	1
c	2	0

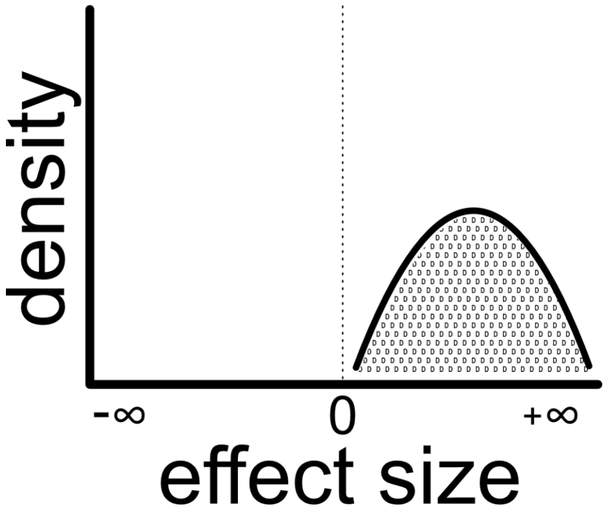


Effect direction

effect size = Hippos - Elephants

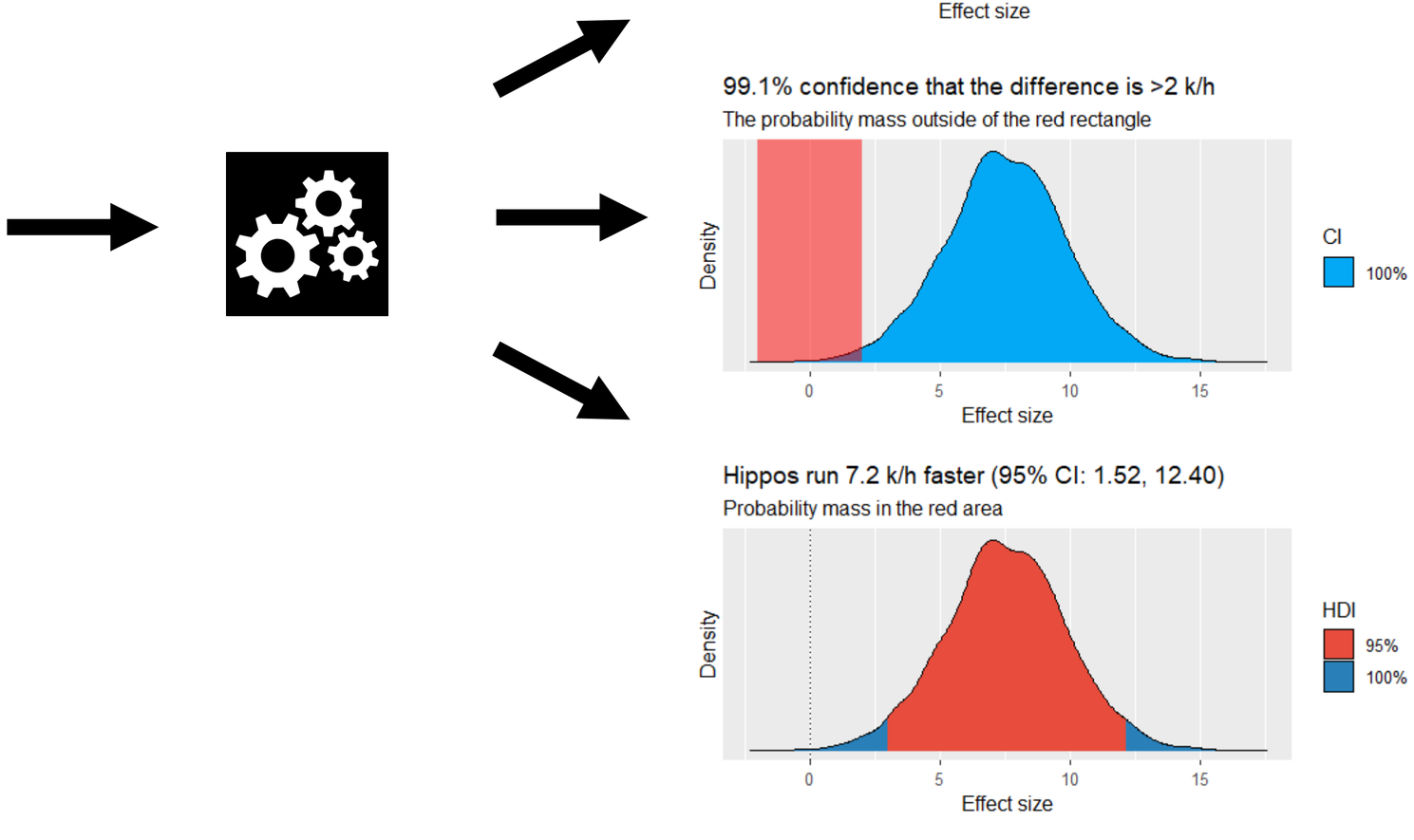




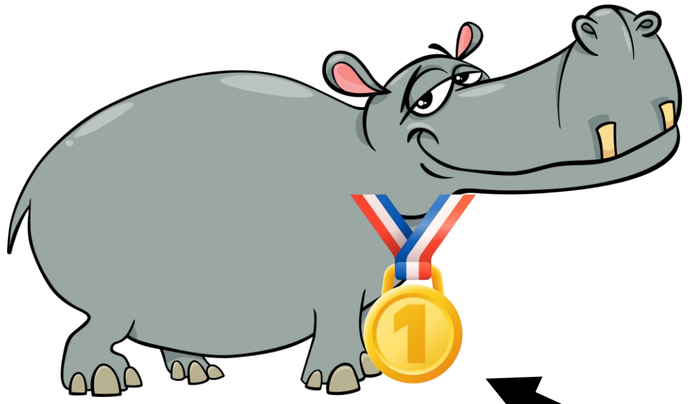


Using effect size for visualising uncertainty

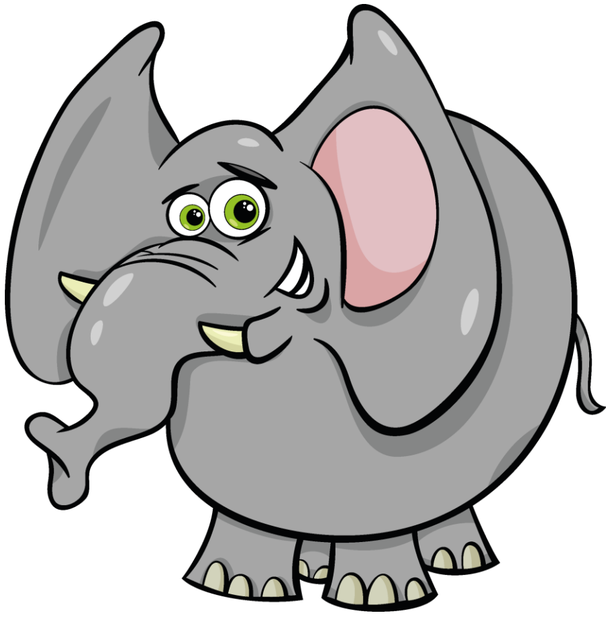
id	y	x
a	0	2
b	1	1
c	2	0



Question > Sample data + Statistics > Population level answer



VS



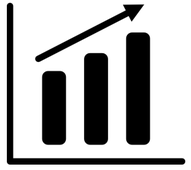
10

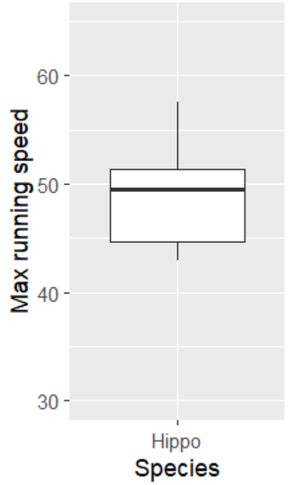
10

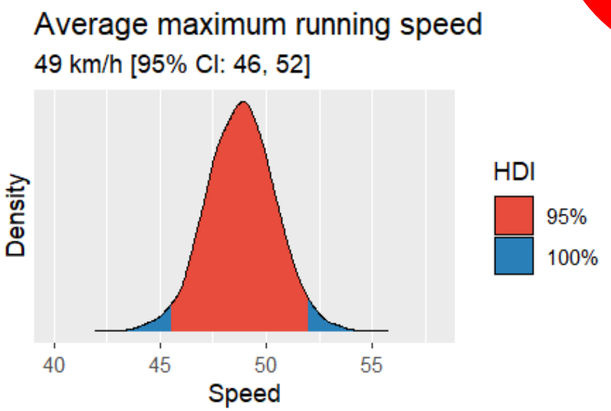
id	y	x
a	0	2
b	1	1
c	2	0

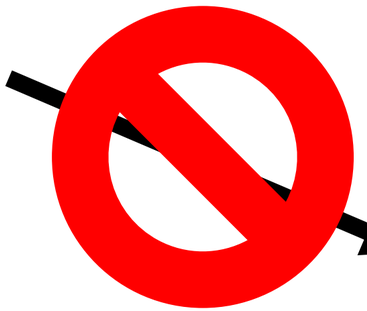
Statistical uncertainty is the uncertainty associated with the use of sample data to make statesments about the wider population

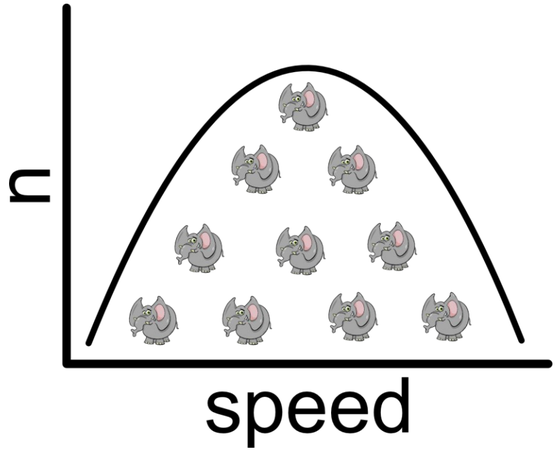
id	y	x
a	0	2
b	1	1
c	2	0



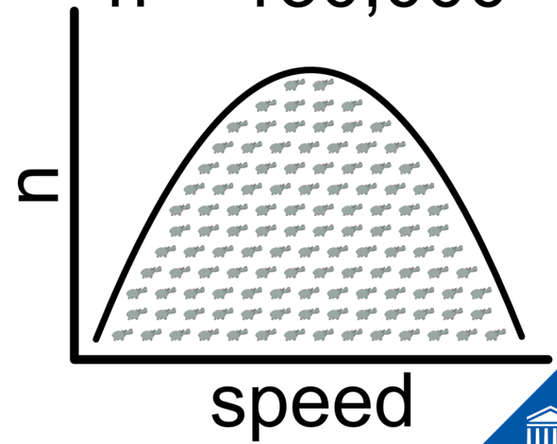




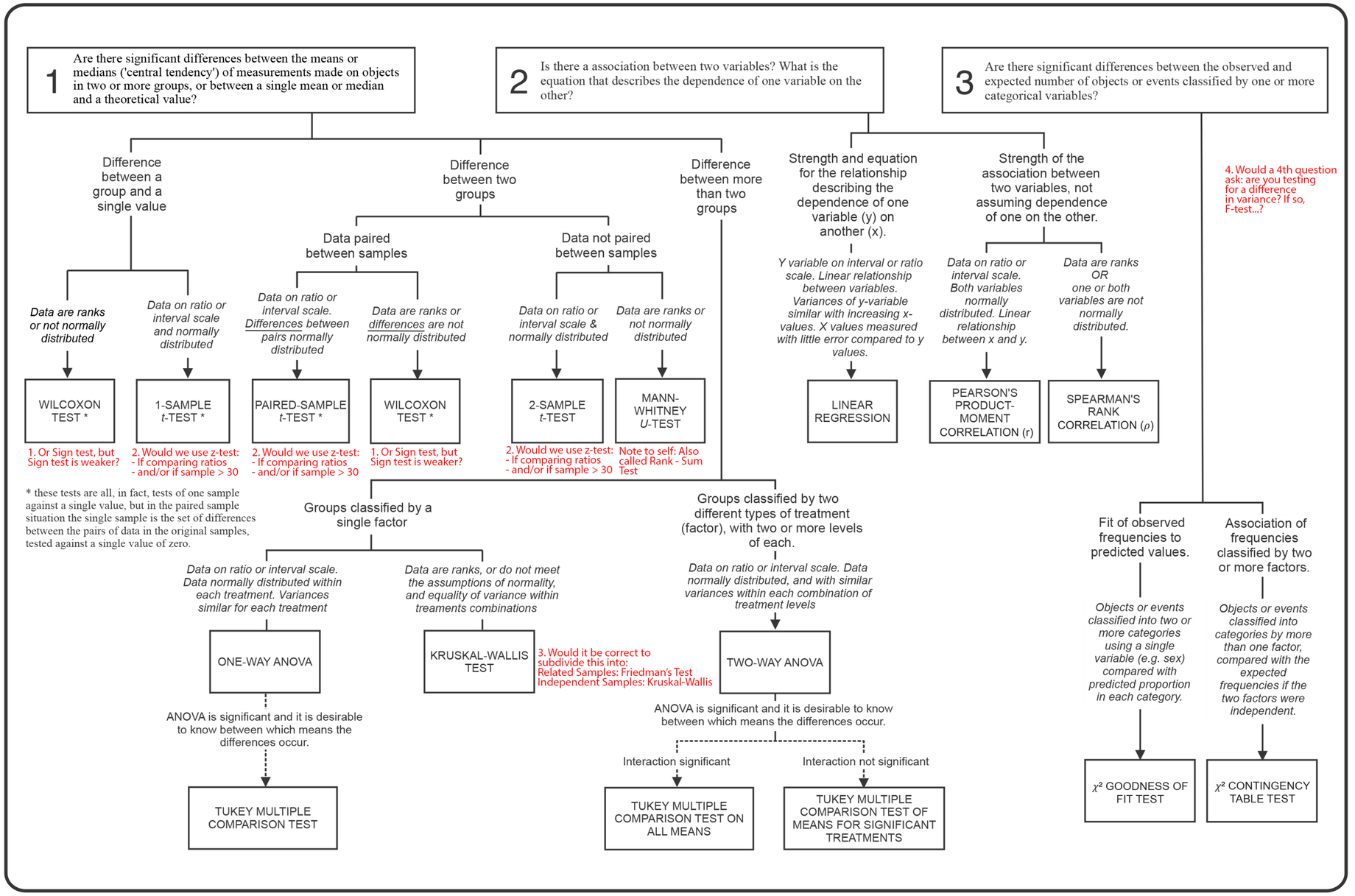




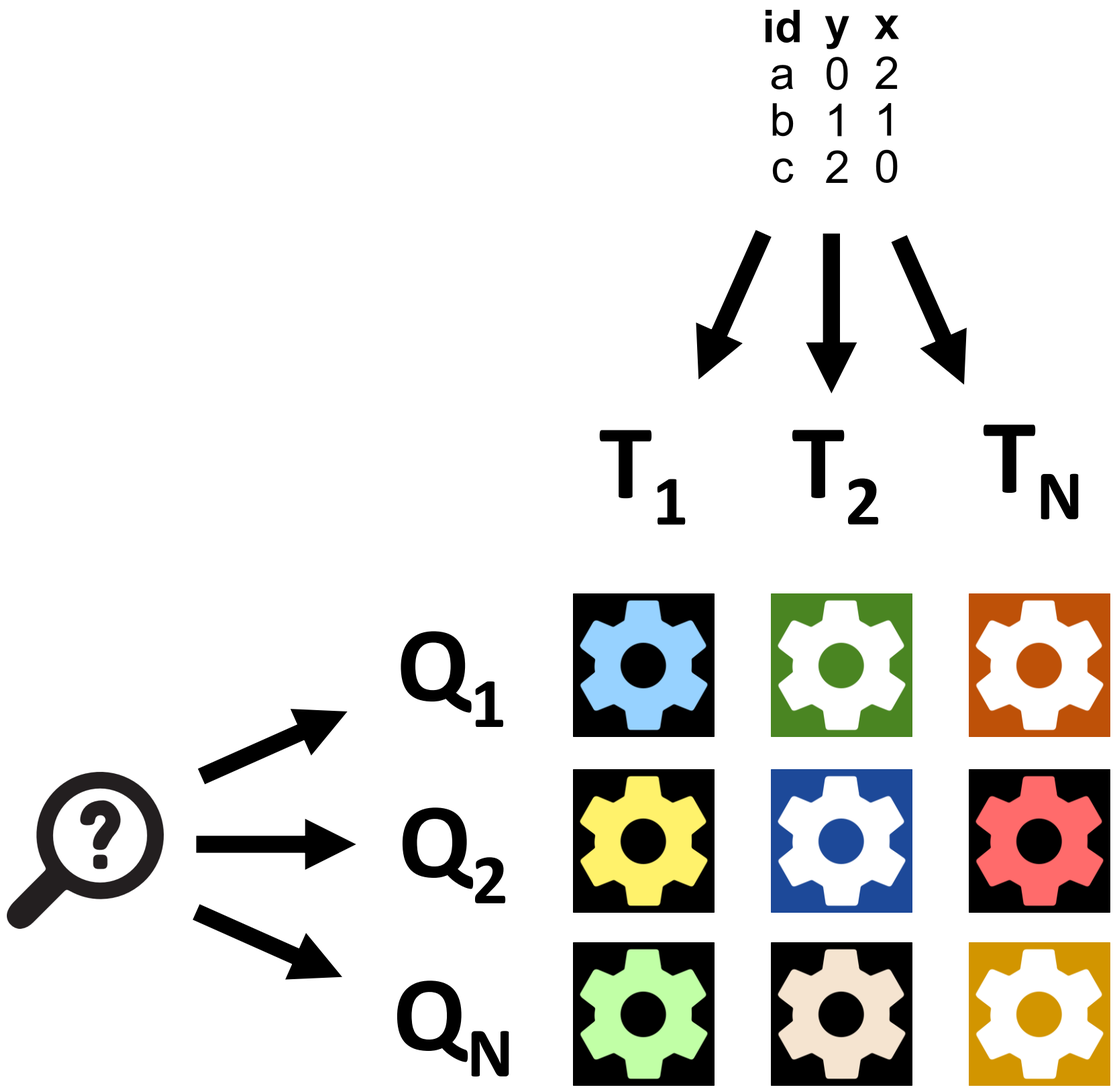
n = 130,000



How is the magic (statistics) applied?

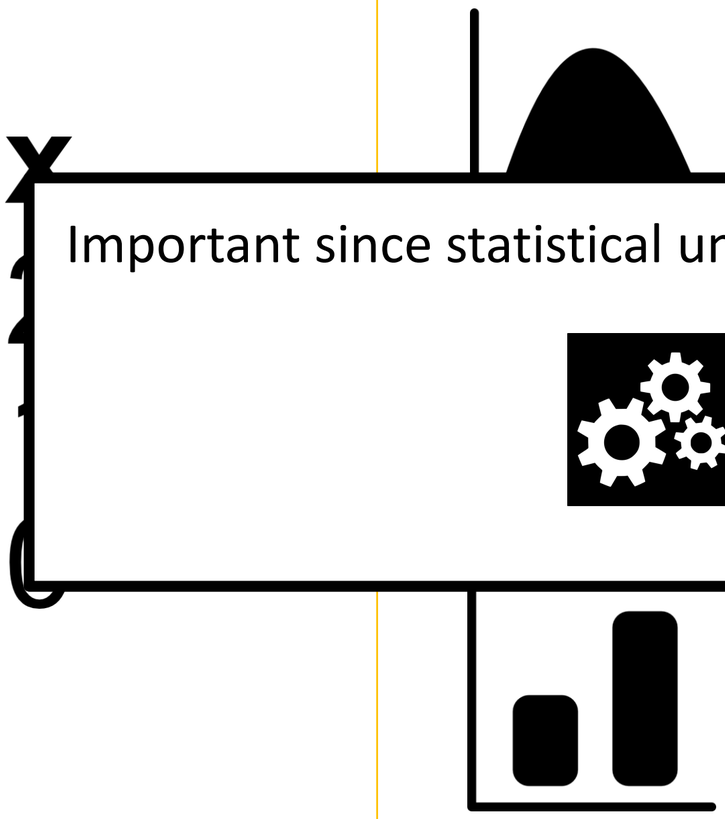


How statistical tests are chosen



Must know your data (variable types)

id	y	x
a	0	2
b	1	1
c	2	0

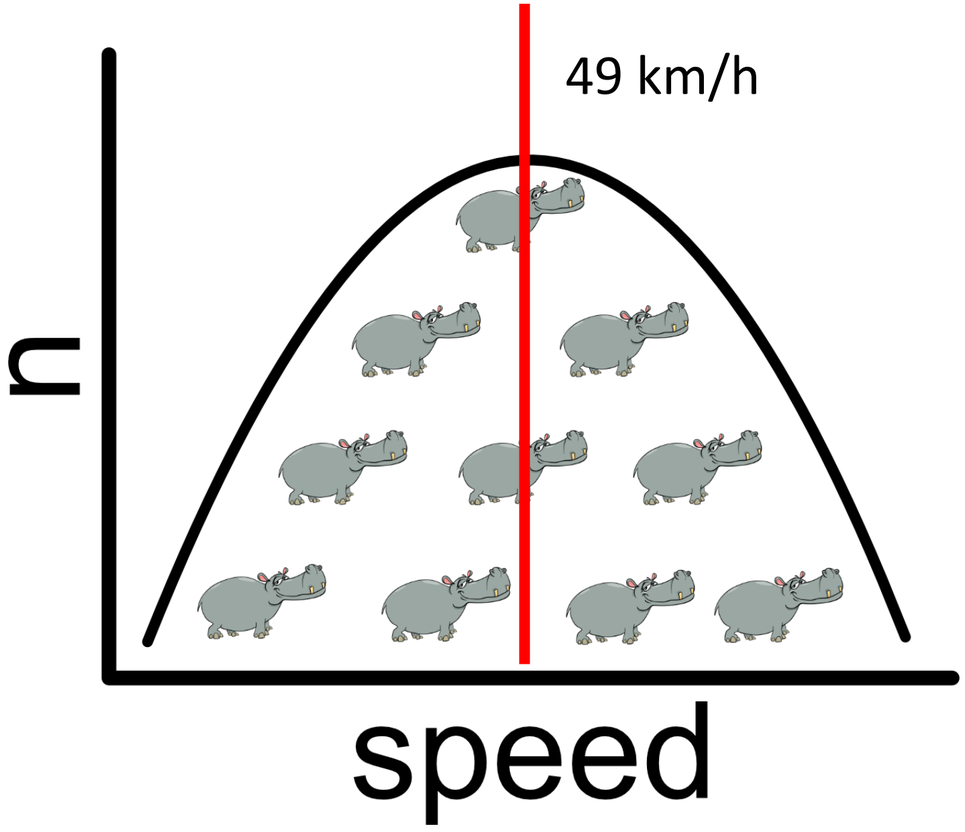


Important since statistical uncertainty is calculated!

	Variable type	Value examples
	Integer	1 5 200
	Numeric	0.05 1
		ia“ nma“ y text“
		UE FALSE TRUE
	Factor	Female Male Unknown

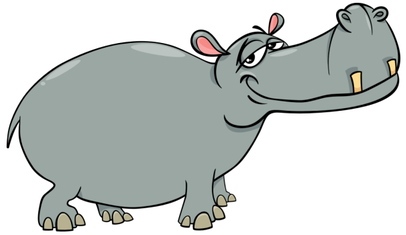
Different study questions

- **Descriptive**
  - What is the maximum running speed of hippos?



Different study questions

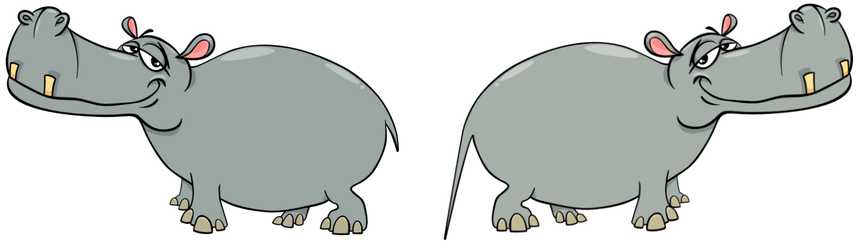
- Descriptive
  - What is the average running speed of hippos?
- **Comparative**
  - How different is the running speed of hippos from elephants?

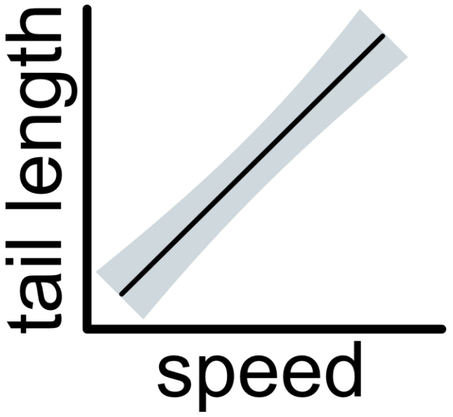


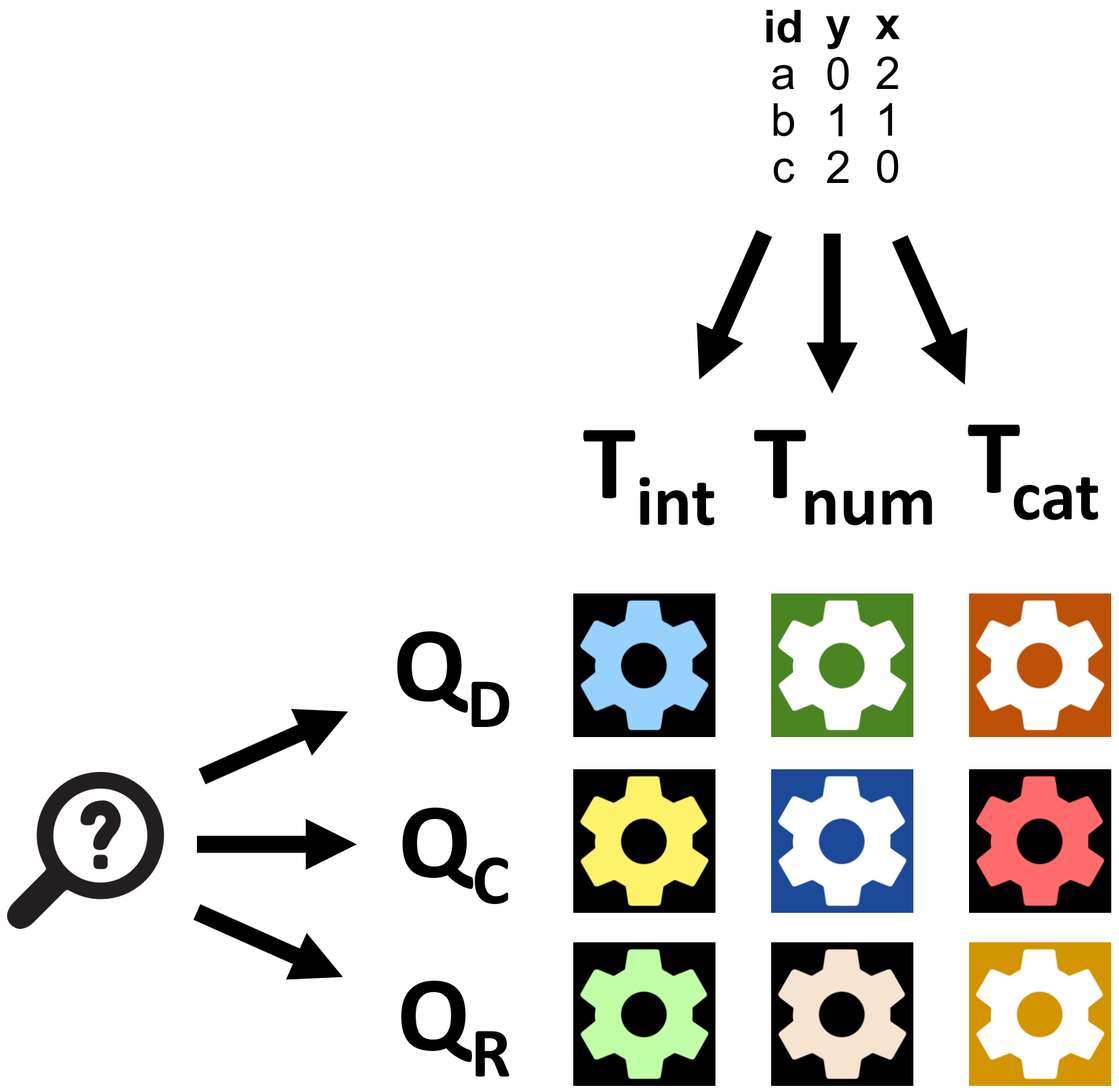
VS

Different study questions

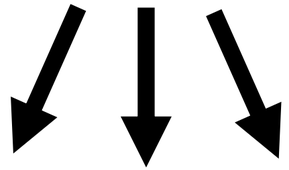
- Descriptive
  - What is the average running speed of hippos?
- Comparative
  - How different is the running speed of hippos from elephants?
- **Relationship**
  - How does tail length affects running speed of hippos?



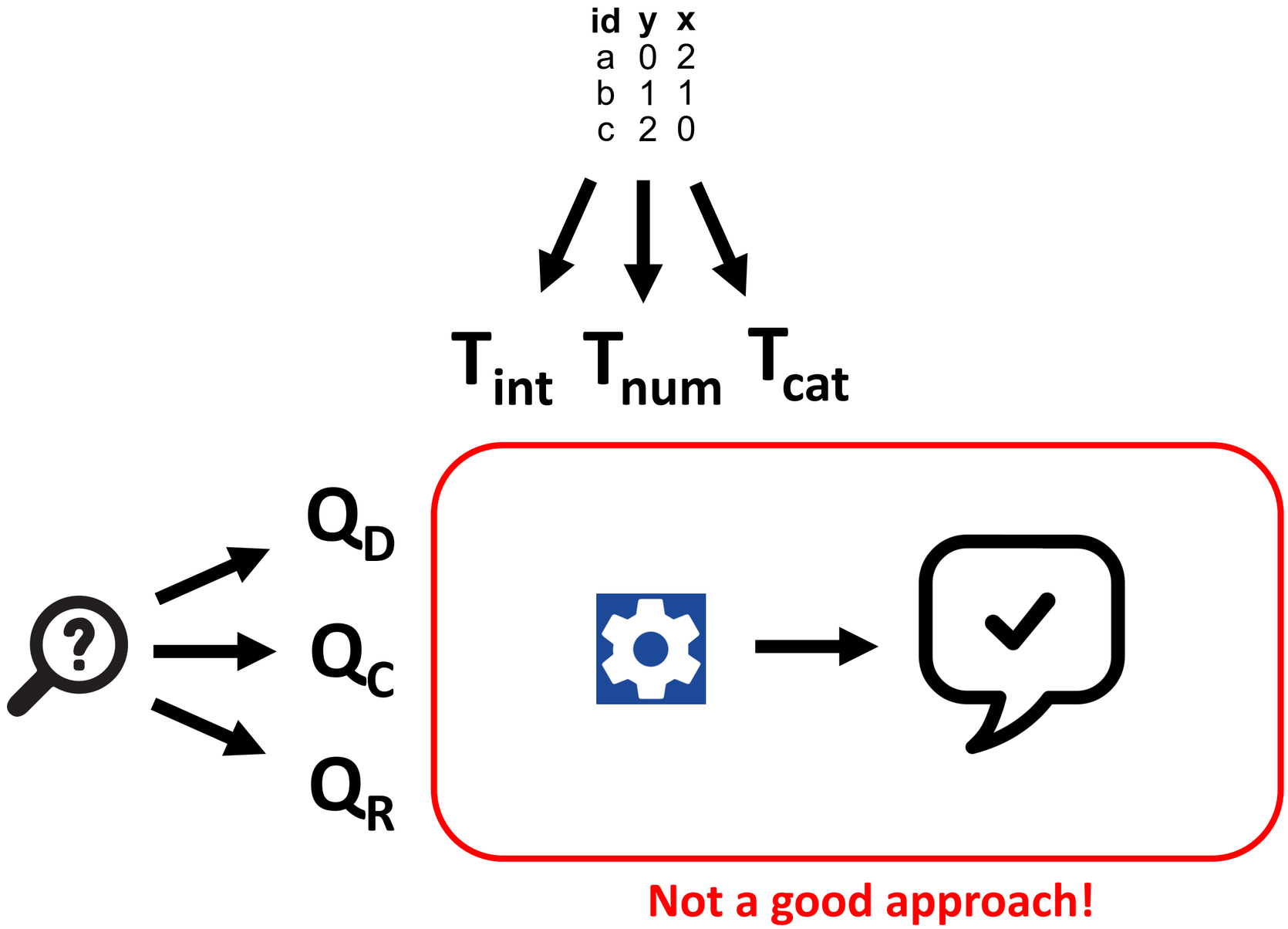



id y x
a 0 2
b 1 1
c 2 0
T int
T num
T cat
Q D
Q C
Q R
?

id	y	x
a	0	2
b	1	1
c	2	0



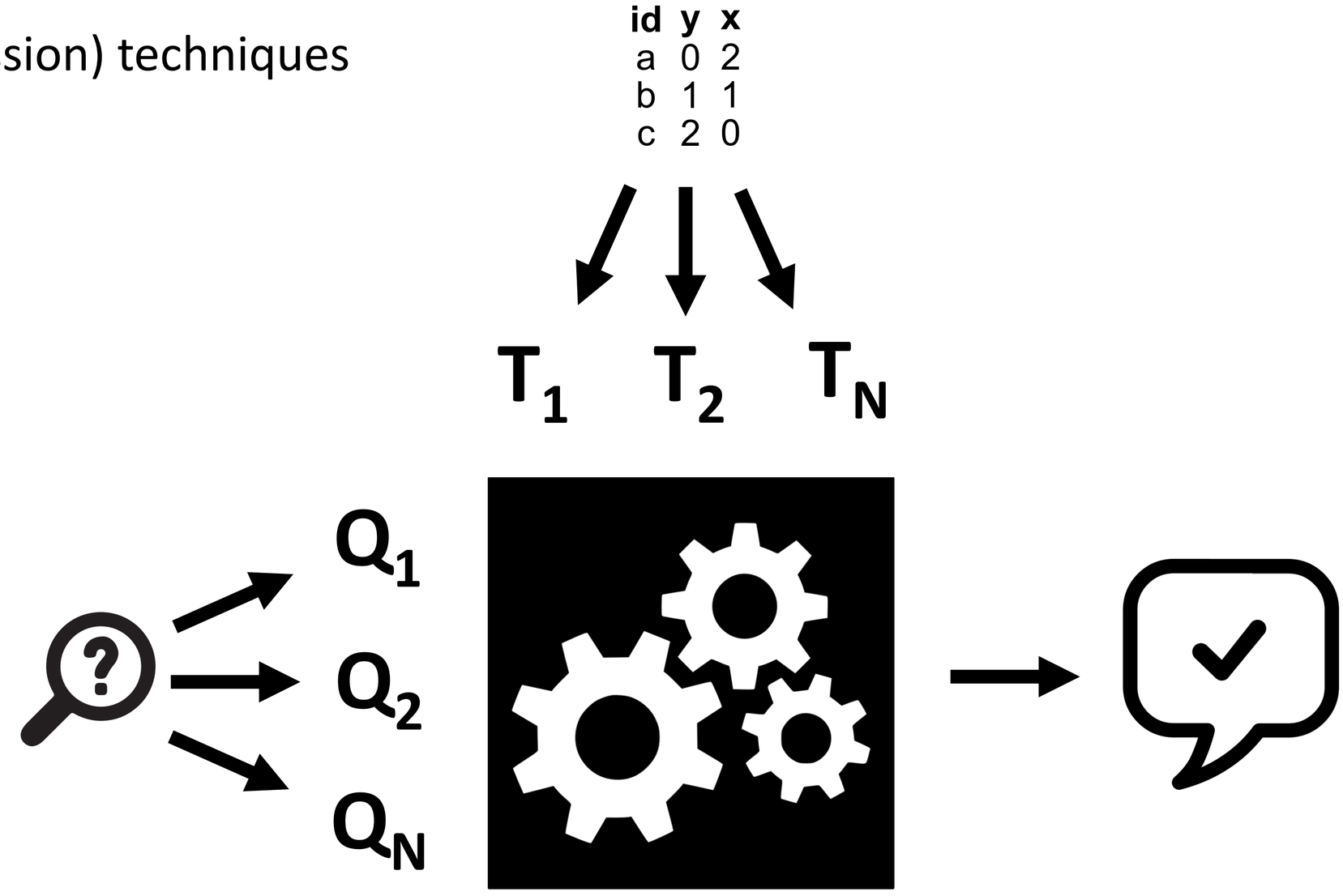
T_{int} T_{num} T_{cat}



Not a good approach!

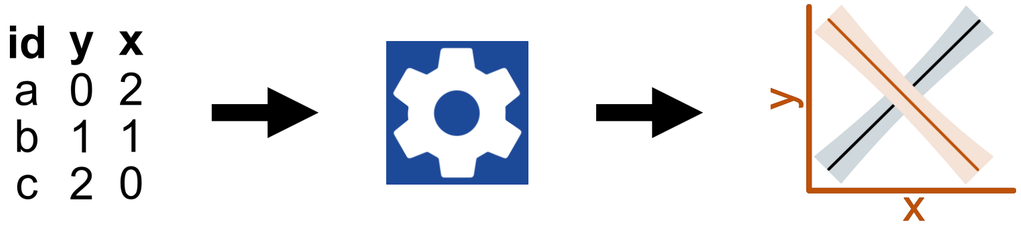
A better approach

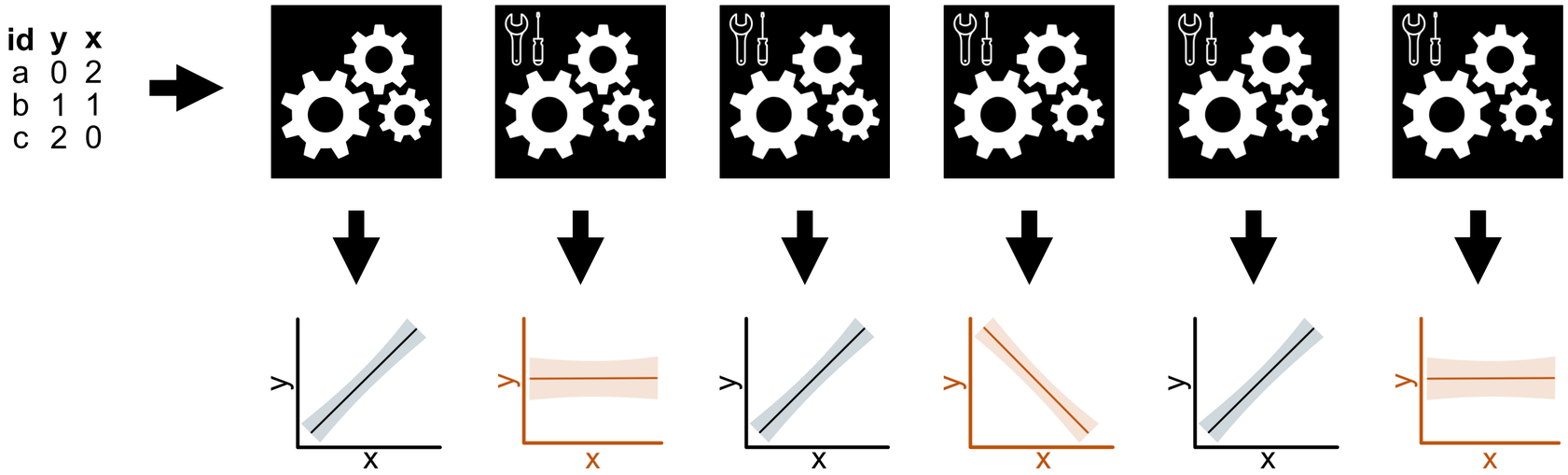
1. Learn modelling (regression) techniques



A better approach

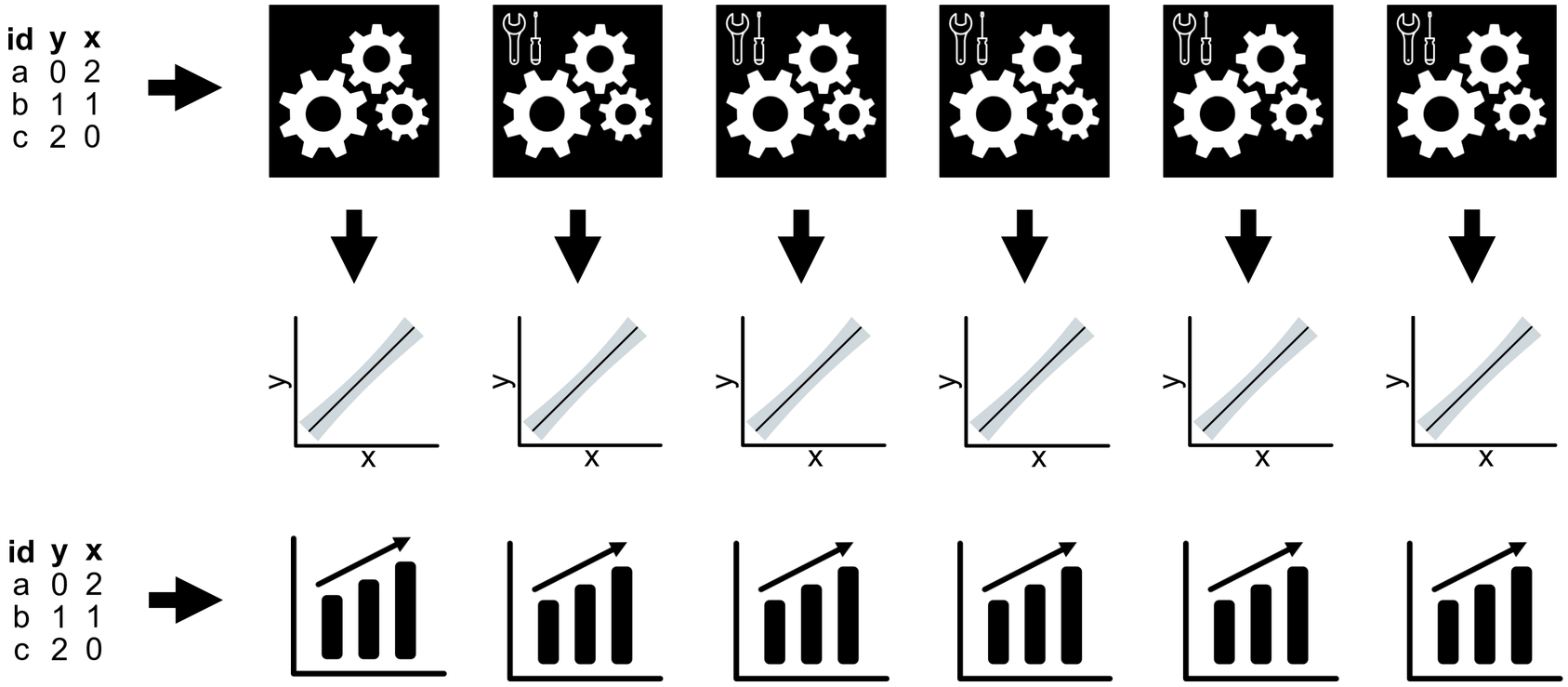
1. Learn modelling (regression) techniques
2. Try to answer your question with multiple models



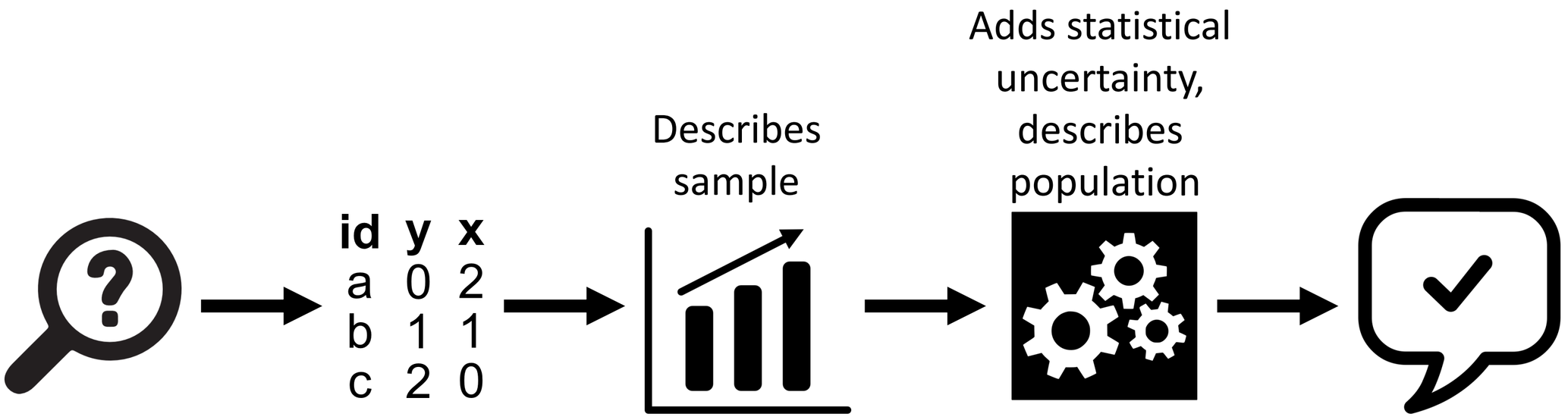


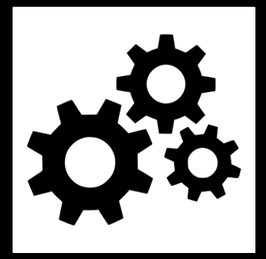
A better approach

1. Learn modelling (regression) techniques
2. Try to answer your question with multiple models
3. Compare their findings with EDA



Analyses give consistent findings - a more confident answer to your question!

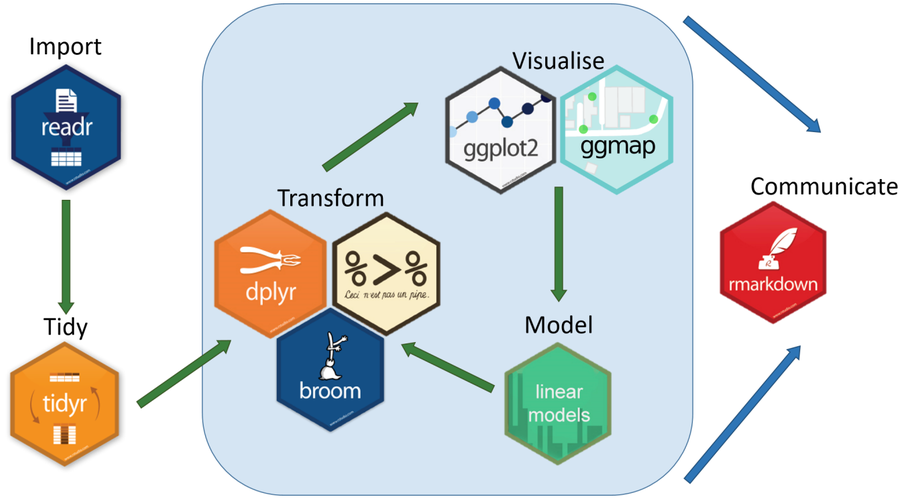

Describes sample
Adds statistical uncertainty, describes population
id y x
a 0 2
b 1 1
c 2 0




tidyverse
b~r+(m|s)

What is R?

- R is a language and environment for statistical computing and graphics.
- Its functionality is divided into modular packages (>18,000).
- Likely, R is the best software for statistical analysis
  - Free
  - Good graphics capabilities
  - Active community (Stack Overflow)



Set-up

Rtools

Windows

RStudio

Import

readr

Tidy

tidyr

Transform

dplyr

%>%

broom

Visualise

ggplot2

ggmap

Model

linear models

Communicate

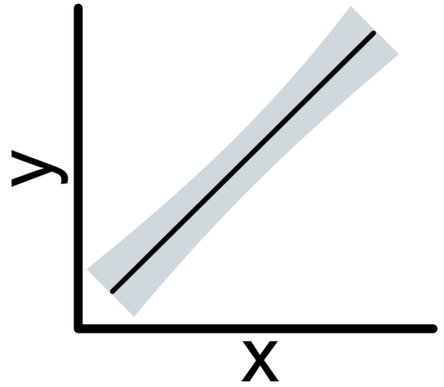
rmarkdown

Code

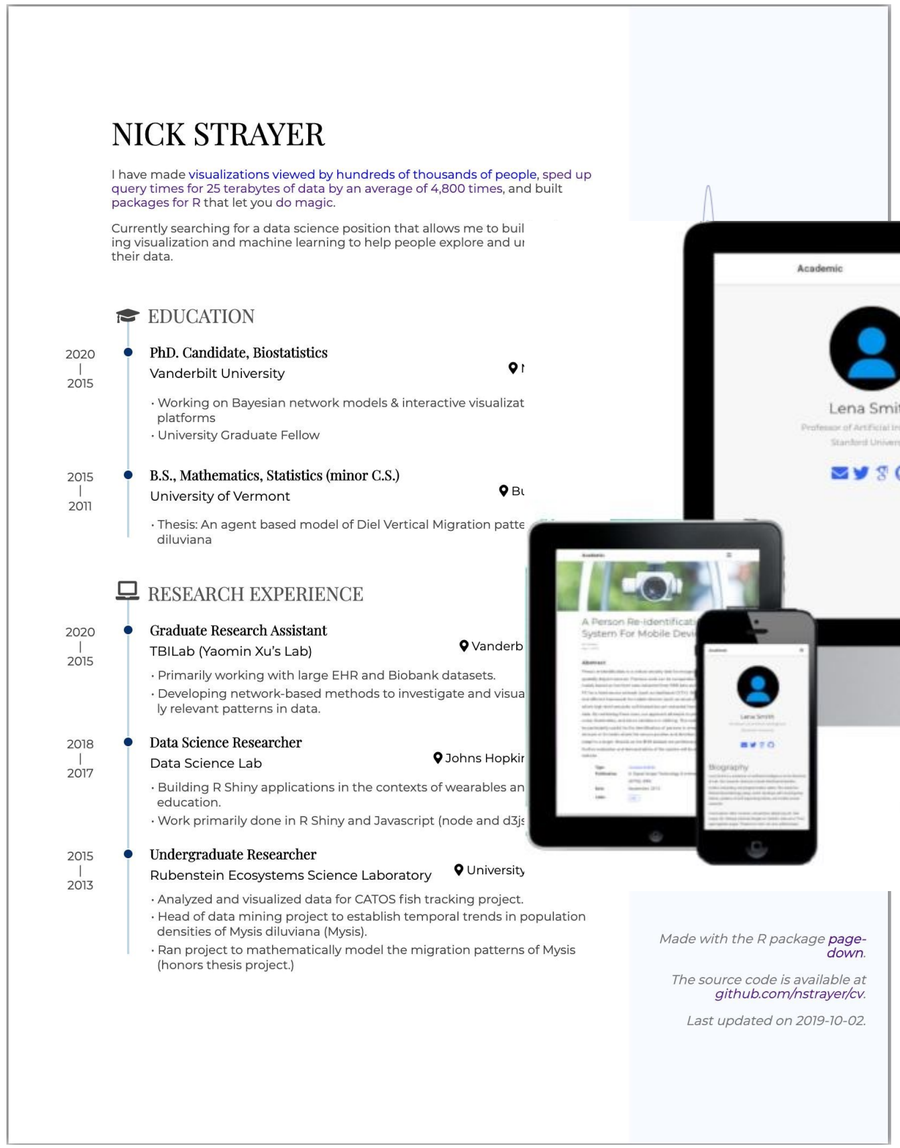
```
library(tidyverse)

print(figure)
```

Output



Not only for statistical analysis



NICK STRAYER

I have made visualizations viewed by hundreds of thousands of people, sped up query times for 25 terabytes of data by an average of 4,800 times, and built packages for R that let you do magic.

Currently searching for a data science position that allows me to buil ing visualization and machine learning to help people explore and u their data.

EDUCATION

2020 | 2015 — **PhD. Candidate, Biostatistics**
Vanderbilt University
- Working on Bayesian network models & interactive visualizat platforms
- University Graduate Fellow

2015 | 2011 — **B.S., Mathematics, Statistics (minor C.S.)**
University of Vermont
- Thesis: An agent based model of Diel Vertical Migration patte diluviana

RESEARCH EXPERIENCE

2020 | 2015 — **Graduate Research Assistant**
TBILab (Yaomin Xu's Lab) — Vanderb
- Primarily working with large EHR and Biobank datasets.
- Developing network-based methods to investigate and visua ly relevant patterns in data.

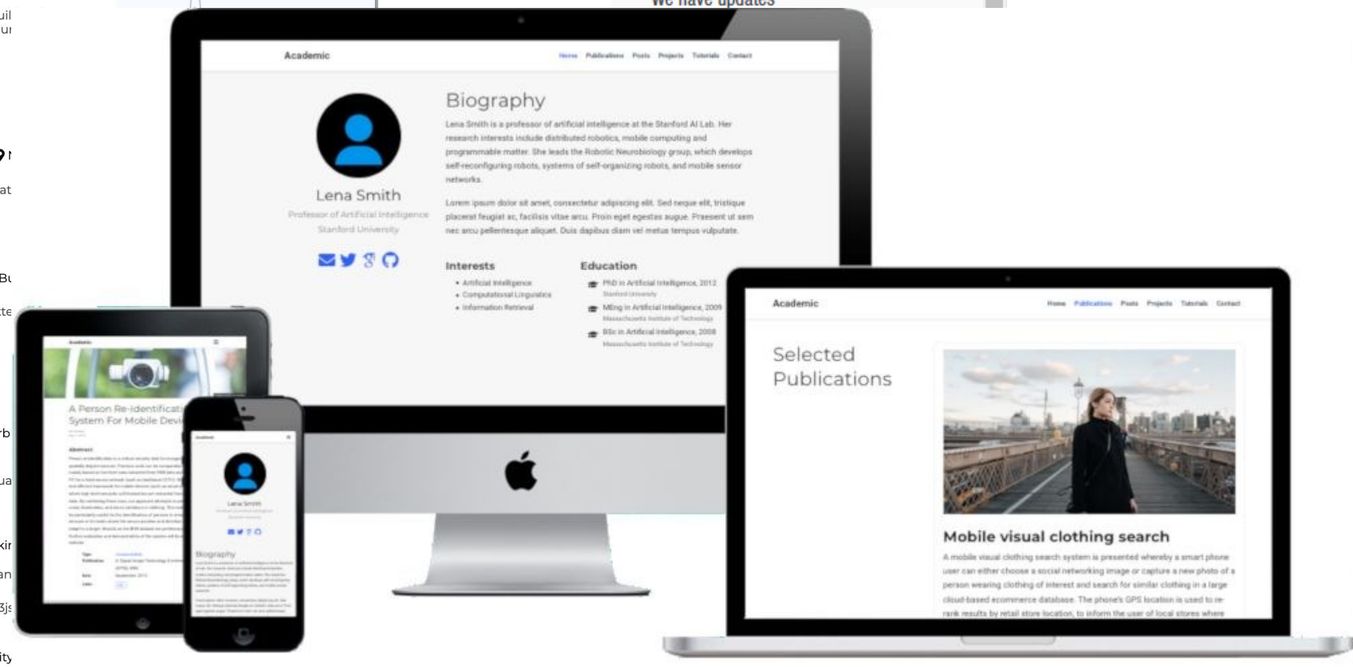
2018 | 2017 — **Data Science Researcher**
Data Science Lab — Johns Hopkir
- Building R Shiny applications in the contexts of wearables an education.
- Work primarily done in R Shiny and Javascript (node and d3j

2015 | 2013 — **Undergraduate Researcher**
Rubenstein Ecosystems Science Laboratory — University
- Analyzed and visualized data for CATOS fish tracking project.
- Head of data mining project to establish temporal trends in population densities of Mysis diluviana (Mysis).
- Ran project to mathematically model the migration patterns of Mysis (honors thesis project.)

Made with the R package pagedown.

The source code is available at github.com/nstrayer/cv.

Last updated on 2019-10-02.



bookdown.org/content/4857/

Apps | Evaluation and the... | Common statistical... | Everything is a Regr... | Reading li

What and why

My assumptions about you

How to use and understand this project

R setup

We have updates

3.2 Sampling to summarize

3.3 Sampling to simulate prediction

3.4 ~~Summary~~ Let's practice with **brms**

Session info

Statistical rethinking **with brms,**

e tidyverse: Second

on of Richard McElreath's (2020a) text,

ian course with examples in R and

the models he covered with Paul Bürkner's

brms package (Bürkner, 2017, 2018, 2020), which makes it easy to fit Bayesian regression models in **R** (R Core Team, 2020) using Hamiltonian Monte Carlo. I also prefer plotting and data wrangling with the packages from the **tidyverse** (Wickham, 2019; Wickham et al., 2019). So we'll be using those methods, too.


tidyverse

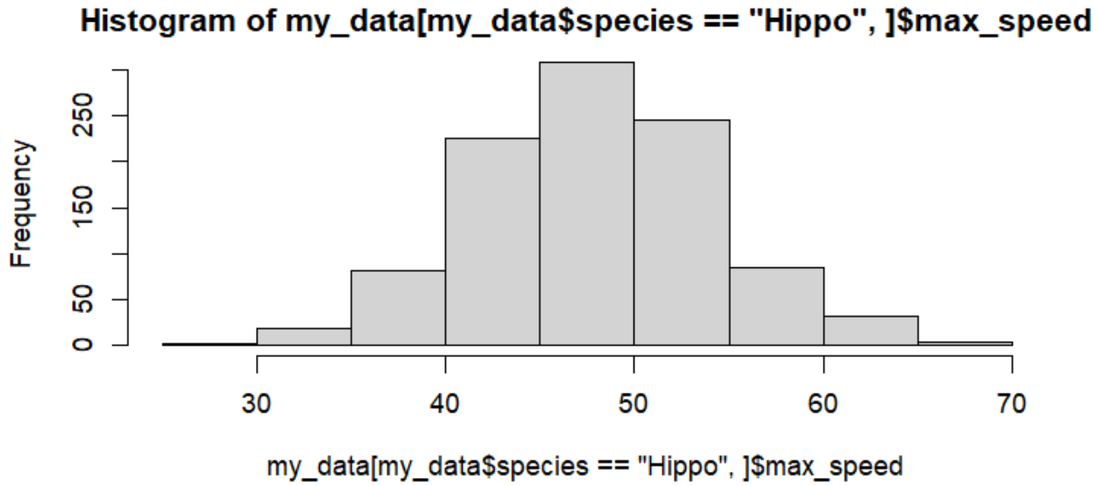
English vs Base R vs Tidyverse

English

Use my hippos-and-elephants data, filter out hippos and plot their maximum running speed values on a histogram.

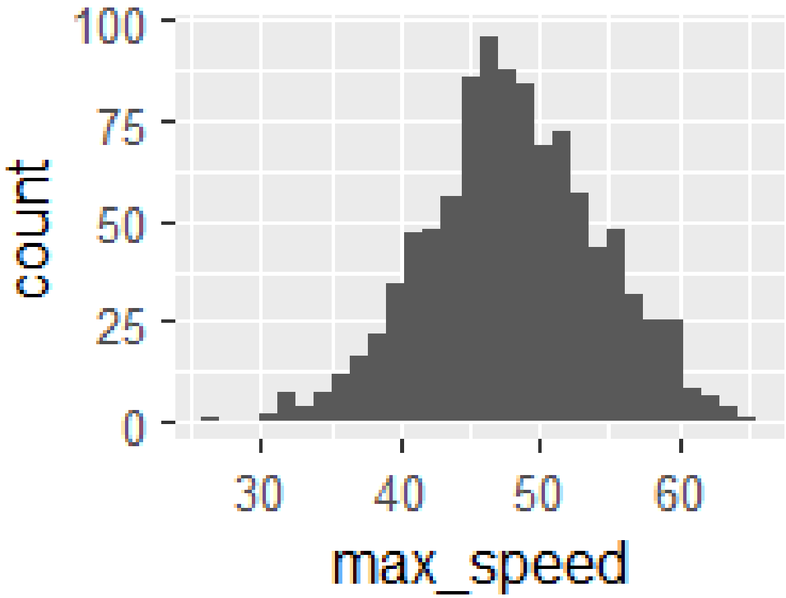
Base R

```
hist(x = my_data[my_data$species ==
"Hippo",]$max_speed)
```

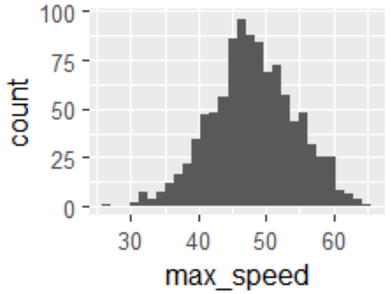


Tidyverse

```
my_data %>%
  filter(species == "Hippo") %>%
    ggplot() +
      geom_histogram(aes(x = max_speed))
```



It is a programming language



A

```
my_data %>%
  filter(species == "Hippo") %>%
    ggplot() +
      geom_histogram(aes(x = max_speed))
```

B

```
my_data %>%
  filter(species != "Elephant") %>%
    ggplot(aes(x = max_speed))+
      geom_histogram()
```

C

```
my_data %>%
  filter(species %>% str_detect("pp")) %>%
    ggplot(aes(x = max_speed))+
      geom_histogram()
```

D

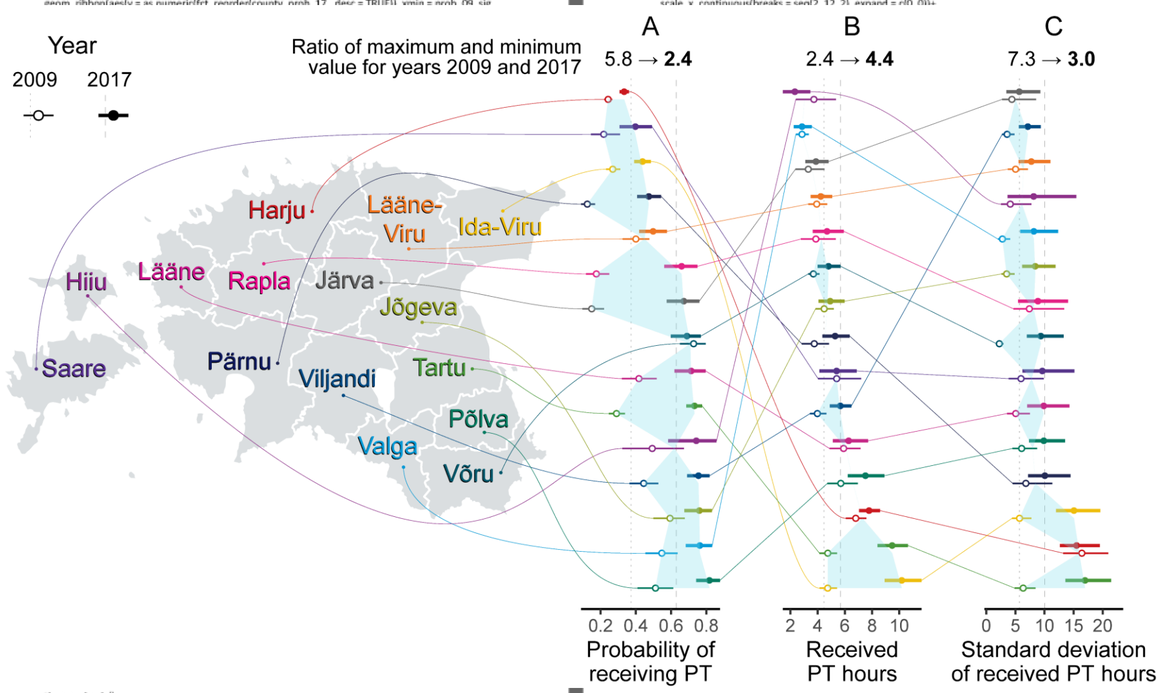
```
filtered_data = my_data %>%
      filter(species == "Hippo")


filtered_data %>%
  ggplot(aes(x = max_speed))+
    geom_histogram()
```

Why to choose the hard way - R?

```{r}
ratio_prob_09 = round_tidy(max(data_for_plot$prob_09)/min(data_for_plot$prob_09), 1) %>% toString()
ratio_PT_09 = round_tidy(max(data_for_plot$PT_09)/min(data_for_plot$PT_09), digits = 1)
ratio_sigma_09 = round_tidy(max(data_for_plot$sigma_09)/min(data_for_plot$sigma_09), digits = 1)
ratio_prob_17 = round_tidy(max(data_for_plot$prob_17)/min(data_for_plot$prob_17), digits = 1) %>% toString()
ratio_PT_17 = round_tidy(max(data_for_plot$PT_17)/min(data_for_plot$PT_17), digits = 1)
ratio_sigma_17 = round_tidy(max(data_for_plot$sigma_17)/min(data_for_plot$sigma_17), digits = 1)
values_prob = paste(ratio_prob_09, sprintf('\u2192'), ratio_prob_17)
values_PT = paste(ratio_PT_09, sprintf('\u2192'), ratio_PT_17)
values_sigma = paste(ratio_sigma_09, sprintf('\u2192'), ratio_sigma_17)

A = data_for_plot2 %>%
ggplot(aes())+
  geom_vline(xintercept=mean(data_for_plot$prob_09), linetype=3, color = "#aaaaaa", size = 0.15)+
  geom_vline(xintercept=mean(data_for_plot$prob_17), linetype=2, color = "#aaaaaa", size = 0.15)+
  scale_color_manual(values = c("Harju" = "#cb191f", "Ida-Viru" = "#eebd0c", "Saare" = "#55338a", "Pärnu" = "#232a55", "Tartu" = "#489739", "Lääne" = "#c52786", "Viljandi" = "#004a83", "Jõgeva" = "#95a329", "Põlva" = "#067b61", "Võru" = "#03586c", "Valga" = "#0699d6", "Hiiu" = "#8d3089", "Rapla" = "#e32285", "Lääne-Viru" = "#ee7725", "Järva" = "#676767")+
  geom_errorbarh(aes(y = fct_reorder(county, prob_17, .desc = TRUE), color = county, xmin = prob.lower_09, xmax = prob.upper_09), height = 0, size = 0.4, alpha = 1)+
  geom_errorbarh(aes(y = fct_reorder(county, prob_17, .desc = TRUE), color = county, xmin = prob.lower_09, xmax = prob.upper_09), height = 0, size = 0.1)+
  geom_point(aes(y = fct_reorder(county, prob_17, .desc = TRUE), color = county, x = prob_09), size = 1, shape = 21, fill = "white", alpha = 1)+
  geom_errorbarh(aes(y = fct_reorder(county, prob_17, .desc = TRUE), color = county, xmin = prob.lower_17, xmax = prob.upper_17), height = 0, size = 0.8, position = position_nudge(y = 0.22))+
  geom_errorbarh(aes(y = fct_reorder(county, prob_17, .desc = TRUE), color = county, xmin = prob.lower_17, xmax = prob.upper_17), height = 0, size = 0.1, position = position_nudge(y = 0.22))+
  geom_point(aes(y = fct_reorder(county, prob_17, .desc = TRUE), color = county, x = prob_17), size = 1.5, stroke = 0, position = position_nudge(y = 0.22))+
#shaded area between
```




```
theme_classic()+
```

```
  scale_x_continuous(expand = c(0.1, 0.1))+
  ylab("County")+
  xlab("Standard deviation\nof received PT\nhours (95% CI)")+
  guides(color = F)+
   ggtitle(label = "C", subtitle = values_sigma)+
  theme(axis.line.y = element_blank(), axis.ticks.y = element_blank(), plot.margin = unit(c(0, 1, 0, 0),"cm"), axis.text.y = element_blank(), axis.title.y = element_blank(), plot.title = element_text(hjust = 0.5, size = 13), plot.subtitle = element_text(hjust = 0.5, size = 10))

A+B+C+

  ggsave("plotid/trends in PT_presentation.eps", device=cairo_pdf, width = 130, height = 125, units="mm")
```

The Popularity of Data Science Software

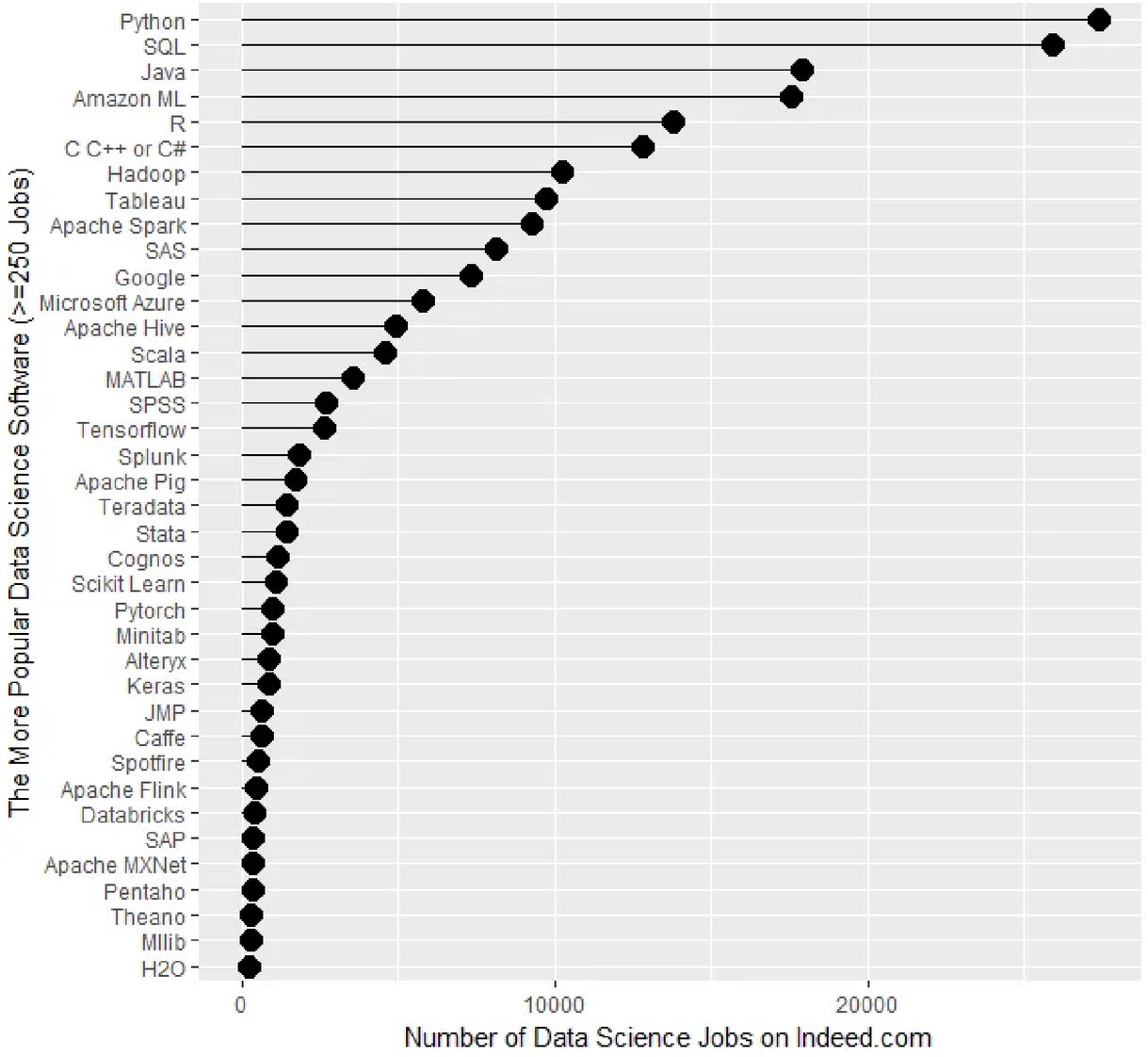
by Robert A. Muenchen

Actuate, Alpine, Alteryx, Angoss, Apache Flink, Apache Hive, Apache Mahout, Apache MXNet, Apache Pig, Apache Spark, BMDP, C, C++ or C#, Caffe, Cognos, DataRobot, Domino Data Labs, Enterprise Miner, FICO, FORTRAN, H2O, Hadoop, Info Centricy or Xeno, Java, JMP, Julia, KNIME, Lavastorm, MATLAB, Megaputer or PolyAnalyst, Microsoft, Minitab, NCSS, Oracle Data Miner, Prognoz, Python, R, RapidMiner, Salford SPM, SAP, SAS, Scala, Spotfire, SPSS, SPSS Modeler, SQL, Stata, Statgraphics, Statistica, Systat, Tableau, Tensorflow, Teradata, Vowpal Wabbit, WEKA/Pentaho, and XGboost.

- Job Advertisements
- Scholarly Articles
- Surveys of Use
- Books
- Blogs
- Discussion Forum Activity
- Growth in Capability

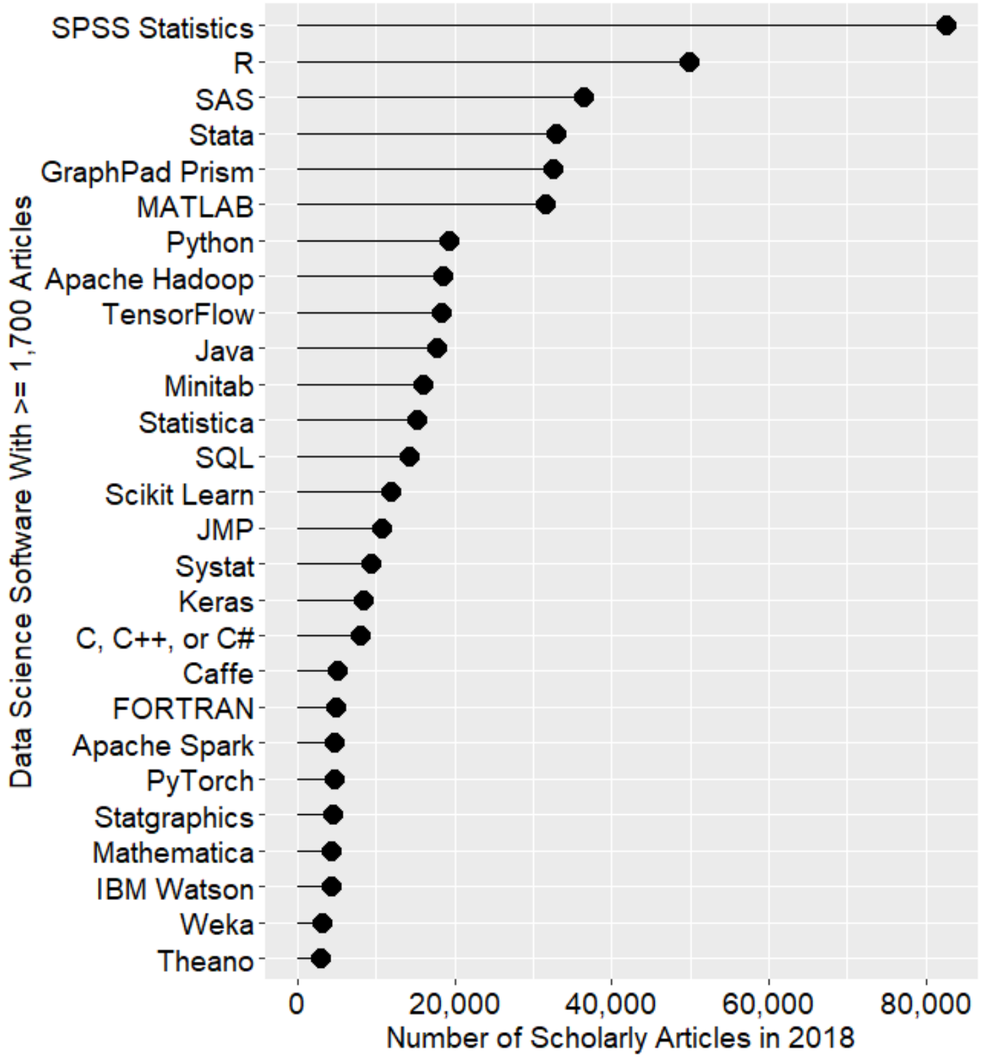
https://r4stats.com/articles/popularity/

Job Advertisements



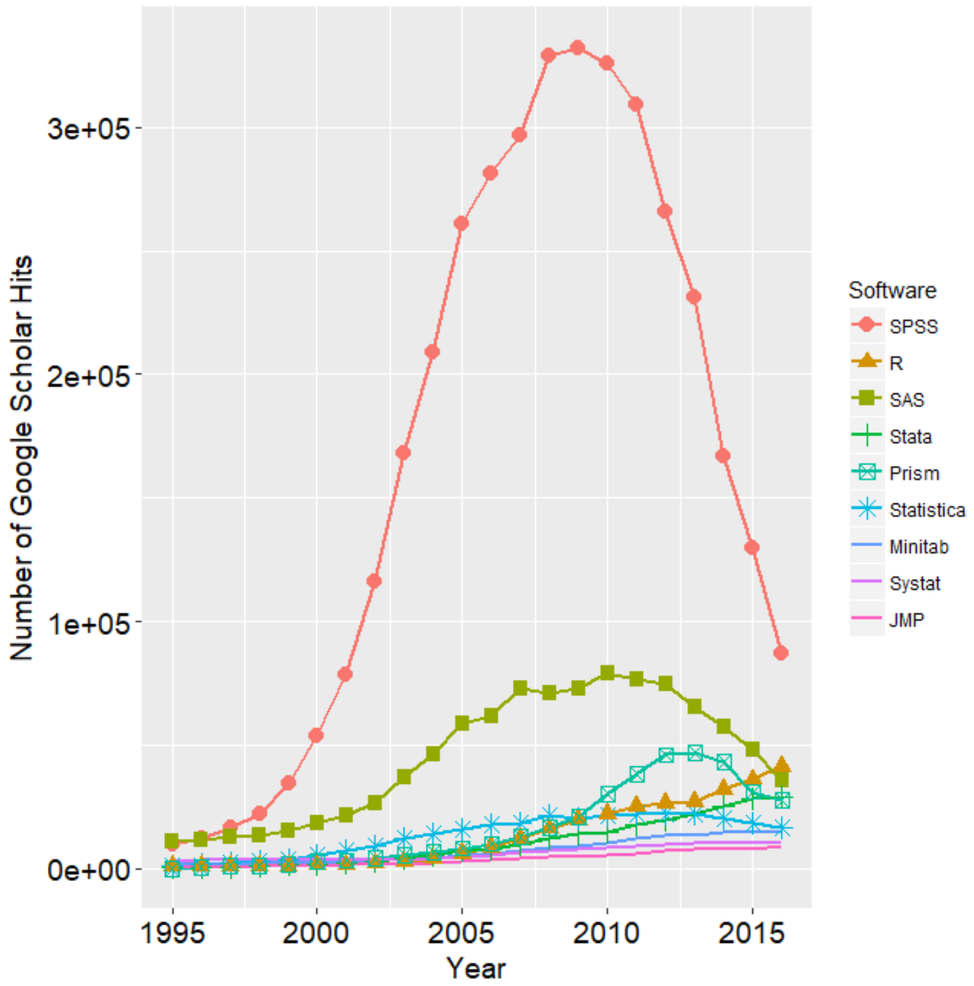
https://r4stats.com/articles/popularity/

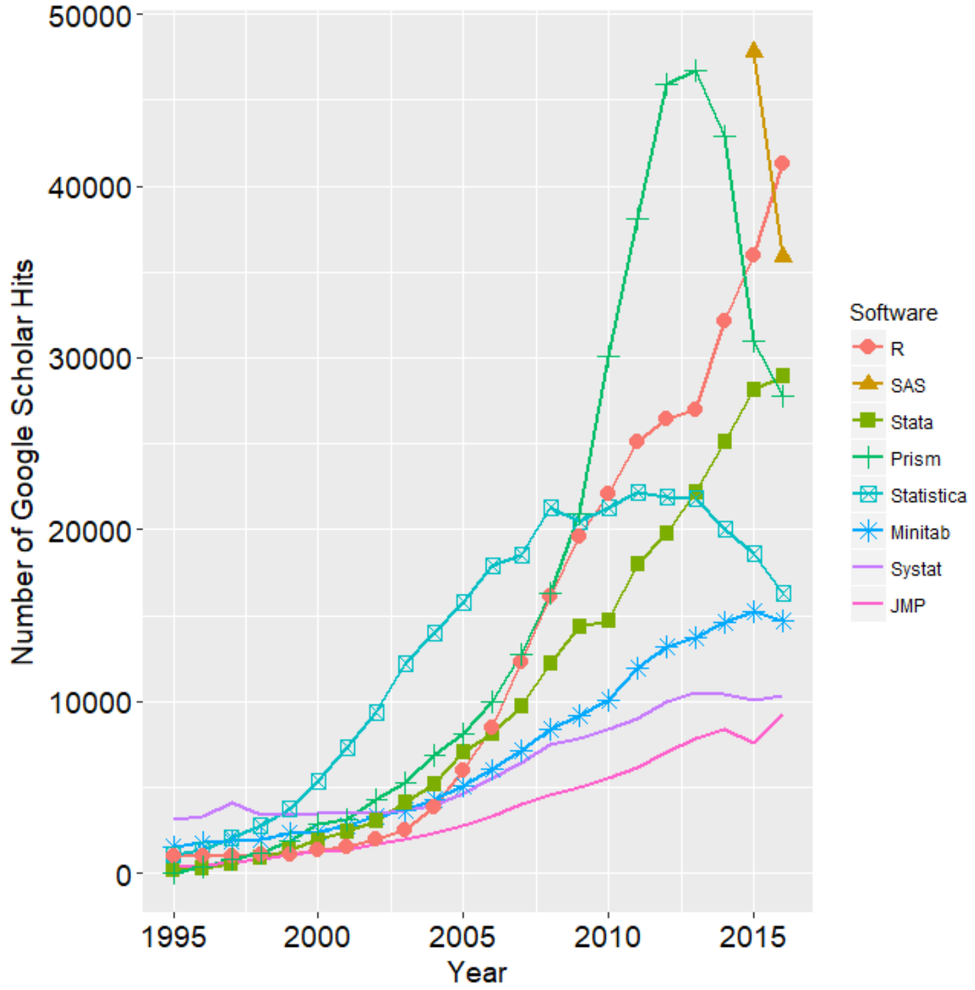
Scholarly Articles



https://r4stats.com/articles/popularity/

Scholarly Articles: trends for classic statistics packages (1995-2015)

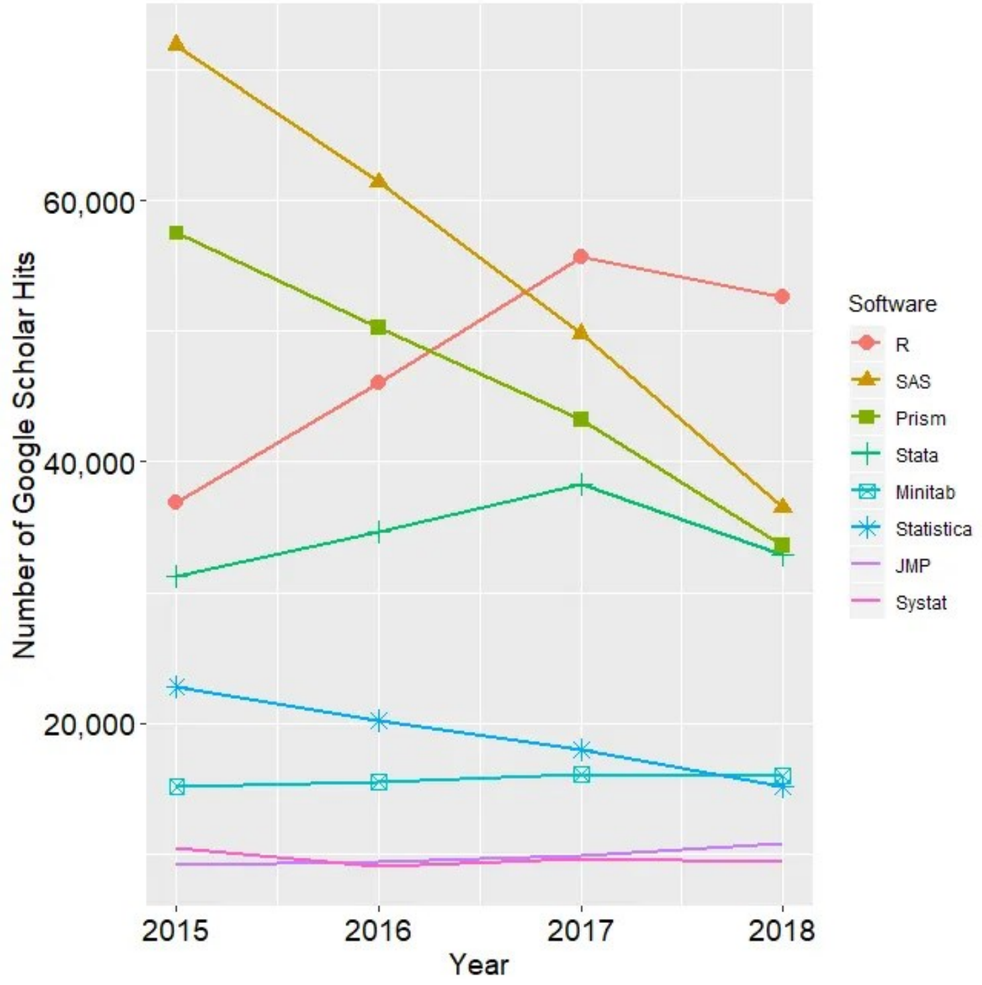




The number of Google Scholar citations for each classic statistics package from 1995 through 2016, this time with SPSS removed and SAS included only in 2014 and 2015. The removal of SPSS and SAS expanded scale makes it easier to see the rapid growth of the less popular packages.

https://r4stats.com/articles/popularity/

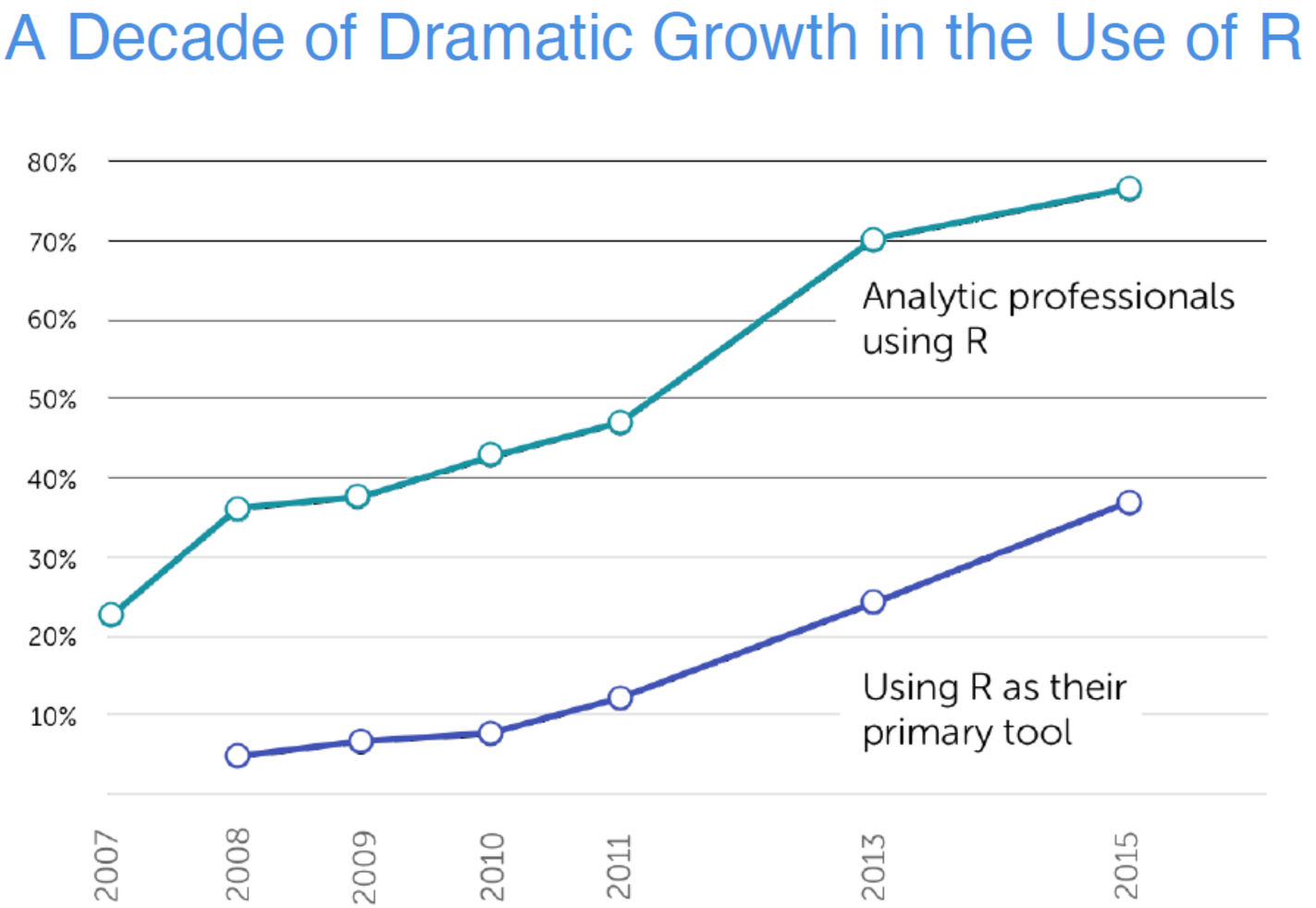
Scholarly Articles: trends for classic statistics packages (2015-2018)



https://r4stats.com/articles/popularity/

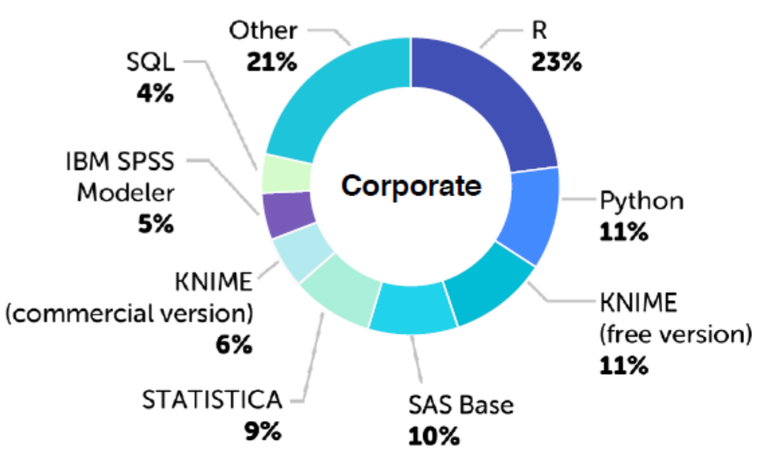
Surveys of Use: Rexer Analytics (2017)

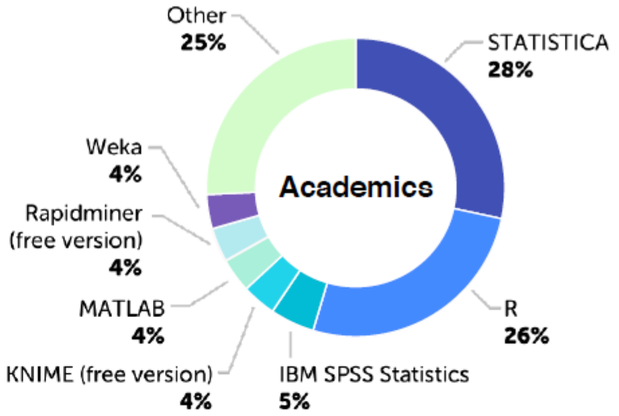
- 67 questions
- 10,000+ invitations emailed & promoted by newsgroups, vendors and bloggers
- Respondents: 1,123 analytic professionals from 91countries
- Data collected in first half of 2017

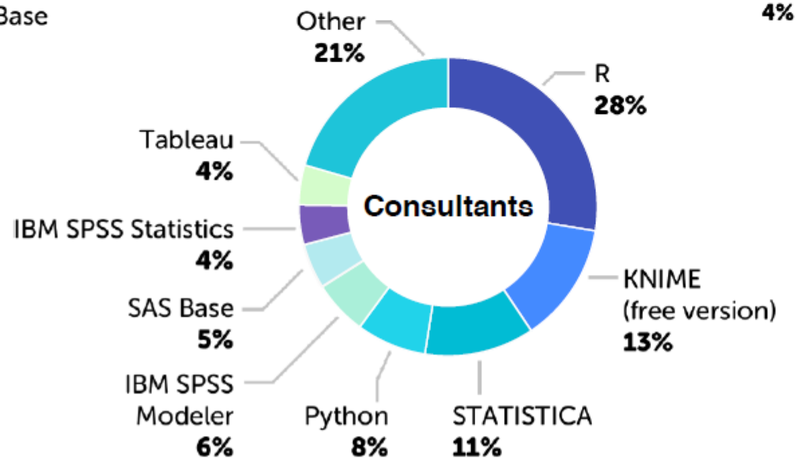


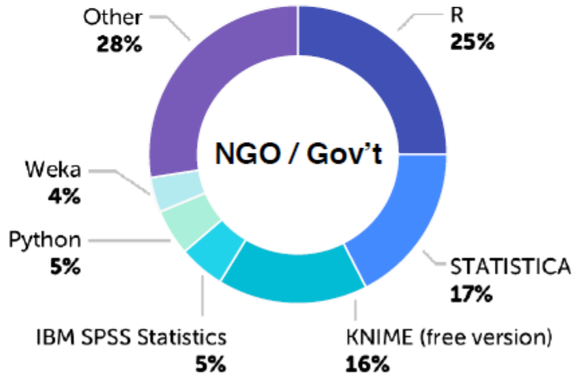
https://www.rexeranalytics.com/data-science-survey

Surveys of Use: Rexer Analytics (2017)



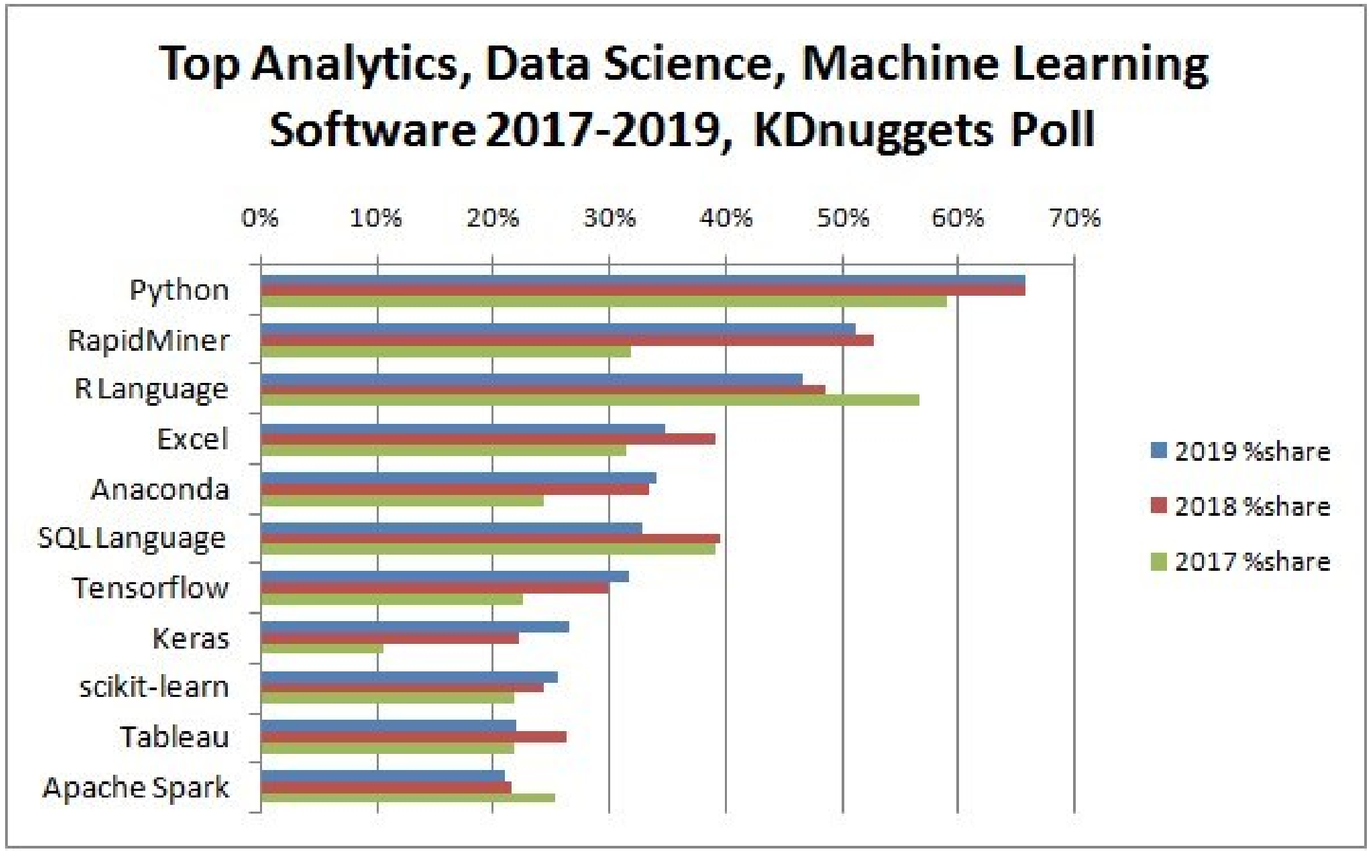






https://www.rexeranalytics.com/data-science-survey

Surveys of Use: KDnuggets



https://www.kdnuggets.com/2020/06/data-science-tools-popularity-animated.html

Books on Amazon.com

Software	Number of Books	
SAS	576	
SPSS Statistics	339	
R	240	[Corrected from blog post: 172]
JMP	97	
Hadoop	89	
Stata	62	
Minitab	33	
Enterprise Miner	32	

https://r4stats.com/articles/popularity/

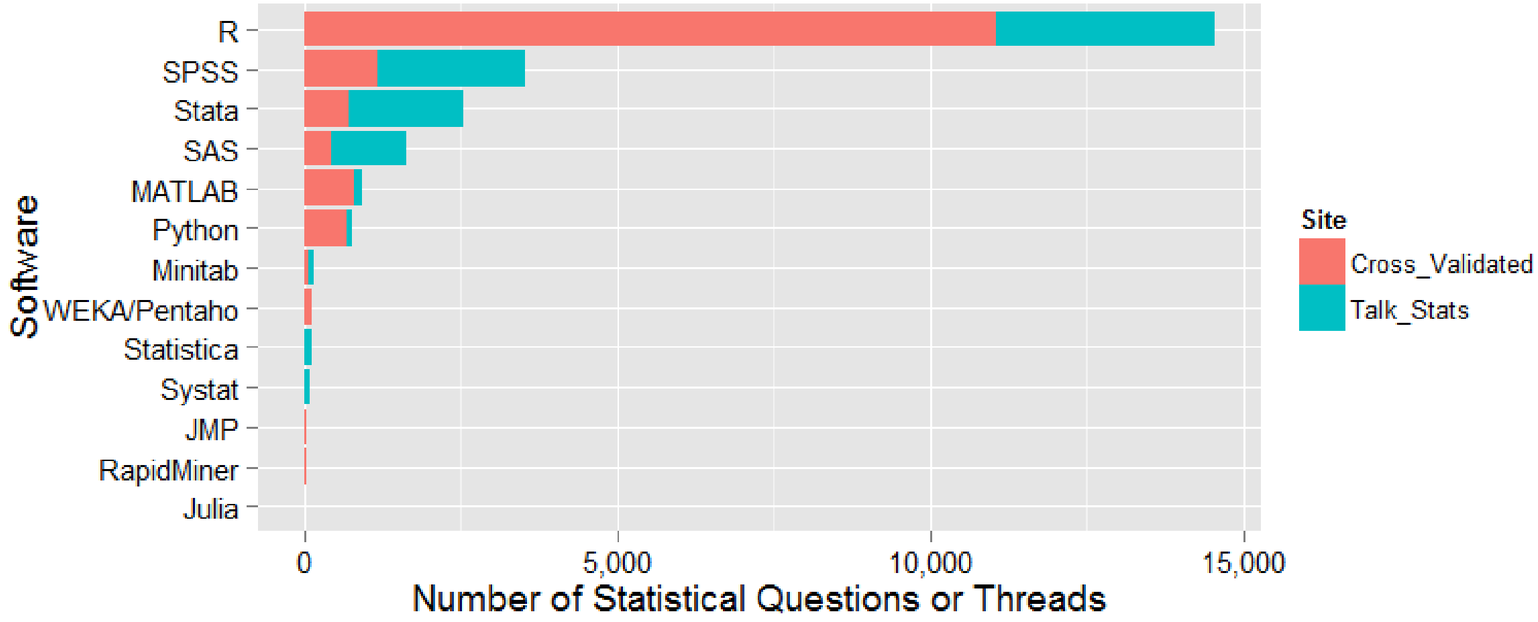
Blogs

Software	Number of Blogs	Source
R	550	R-Bloggers.com
Python	60	SciPy.org
SAS	40	PROC-X.com, sasCommunity.org Planet
Stata	11	Stata-Bloggers.com

Number of blogs devoted to each software package on April 7, 2014, and the source of the data.

https://r4stats.com/articles/popularity/

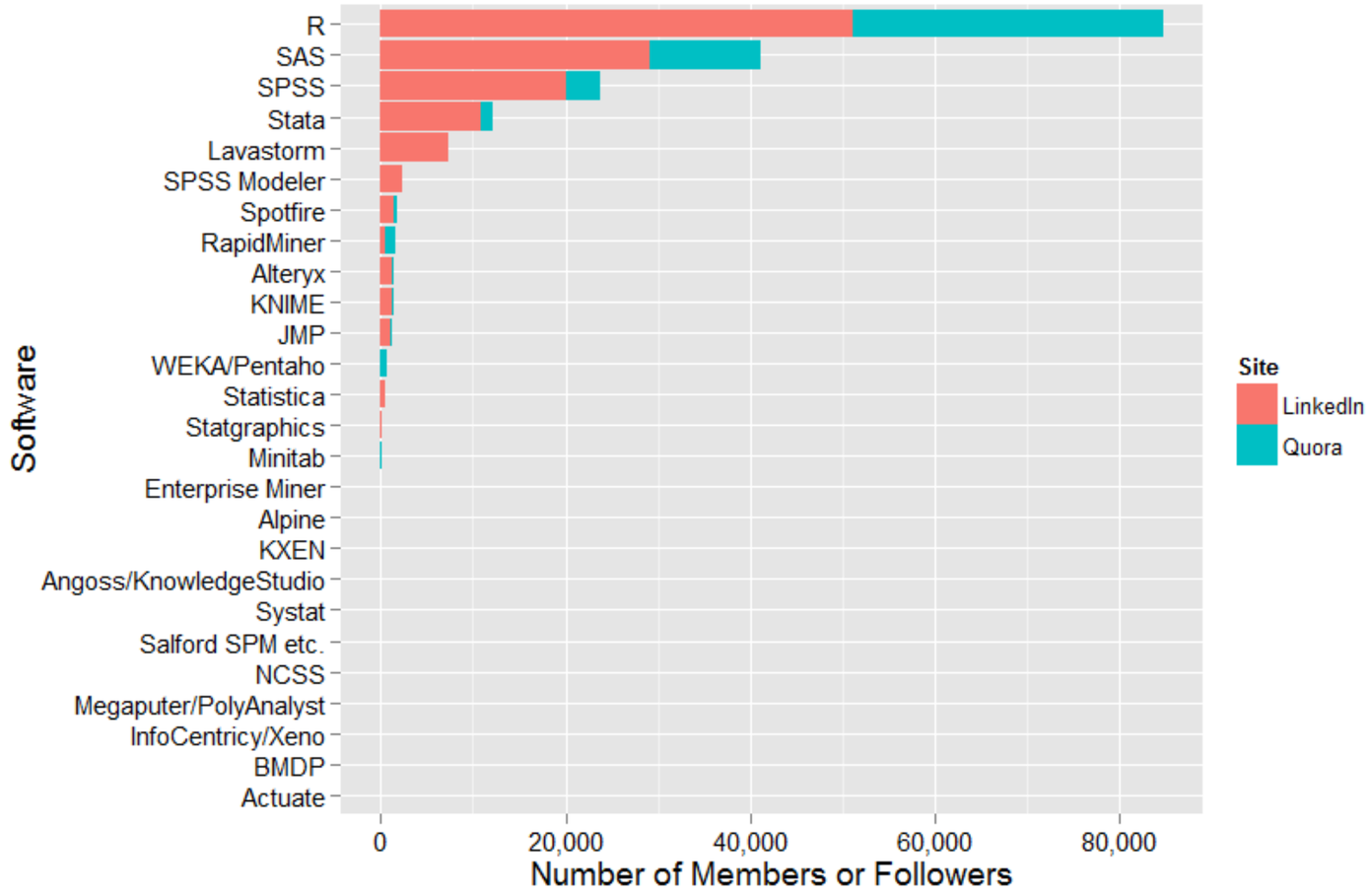
Discussion Forum Activity: Cross Validated and Talk Stats



Number of statistical questions or threads on Cross Validated and Talk Stats.

https://r4stats.com/articles/popularity/

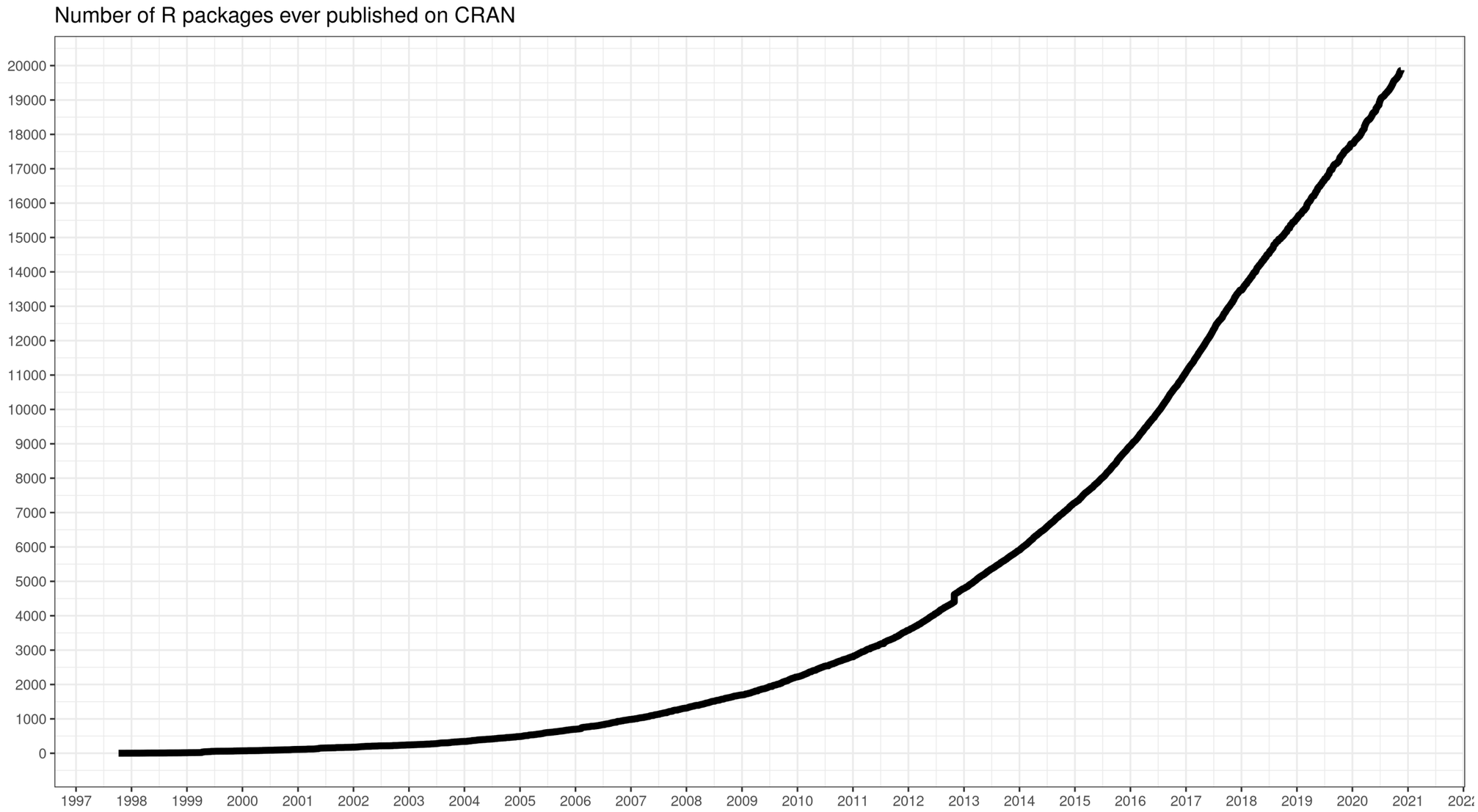
Discussion Forum Activity: LinkedIn and Quora



Number of members or followers on LinkedIn and Quora.

https://r4stats.com/articles/popularity/

Growth in Capability



https://r4stats.com/articles/popularity/

Croat Med J. 2020 Feb; 61(1): 66–68.
doi: 10.3325/cmj.2020.61.66

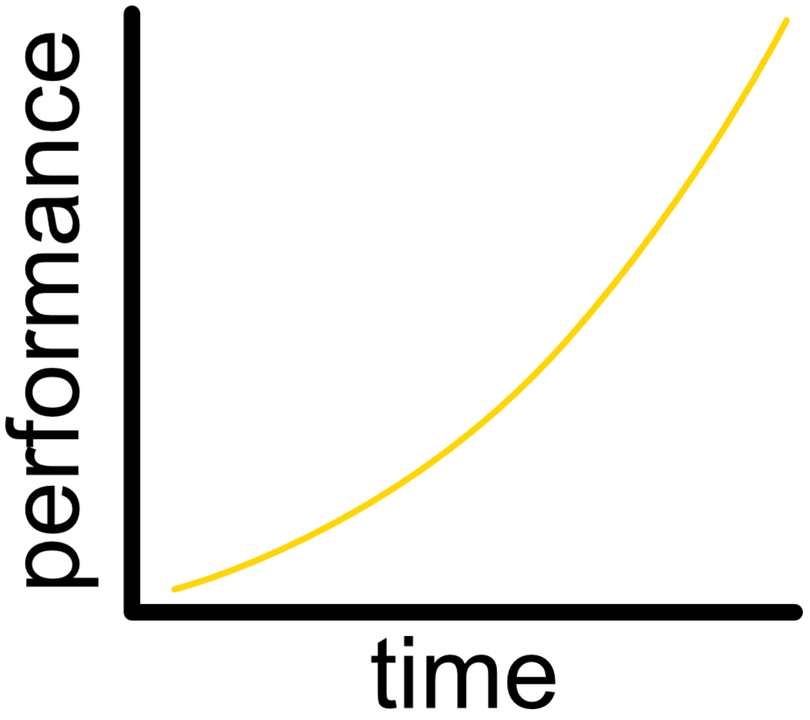
PMCID: PMC7063554
PMID: 32118381

R software: unfriendly but probably the best

Branimir K. Hackenberger

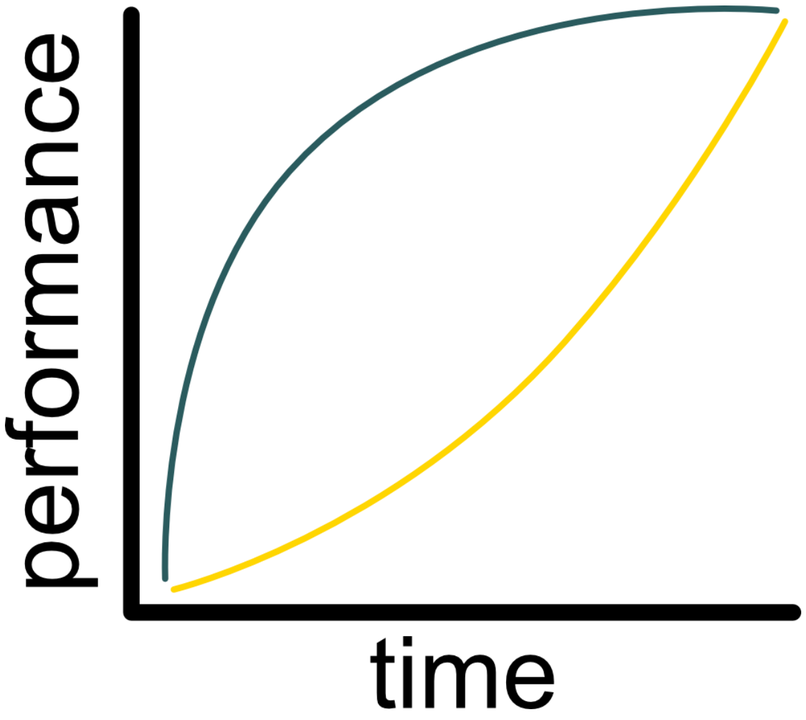
R is perhaps the today's best tool for statistical data processing, ranging from frequentist statistics, Bayesian statistics, meta-analysis, machine and deep learning to parallel computing and Big Data processing. For free!

Expect fustration



Photos: https://www.viator.com/tours/Las-Vegas/Emerald-Cave-Kayak-Tour/d684-170456P1; https://www.tribpub.com/gdpr/theactivetimes.com/

Expect fustration



Photos: https://www.viator.com/tours/Las-Vegas/Emerald-Cave-Kayak-Tour/d684-170456P1; https://www.tribpub.com/gdpr/theactivetimes.com/

Your friends

Google

(R OR tidyverse) AND "your problem"

stackoverflow