

A faint, light-gray network graph serves as the background for the entire slide. It consists of numerous small, semi-transparent circular nodes of varying sizes and colors (light gray, medium gray, and dark gray) connected by thin white lines. Some nodes are highlighted with a blue outline, and a few are filled with blue, indicating specific data points or nodes of interest.

9 Φεβρουαρίου 2019

# Data Science Workshop

# Golden Gate Pro

Λίγα λόγια

# Hello!

## Τάσος Βεντούρης

Data Scientist and  
Game Designer @  
Hattrick Ltd

You can find me at:



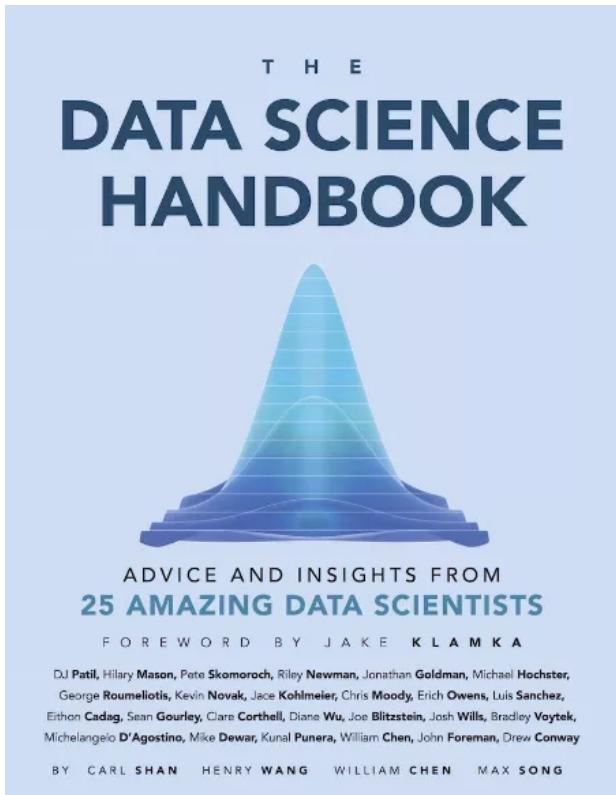
@tasosventouris



Tasos Ventouris



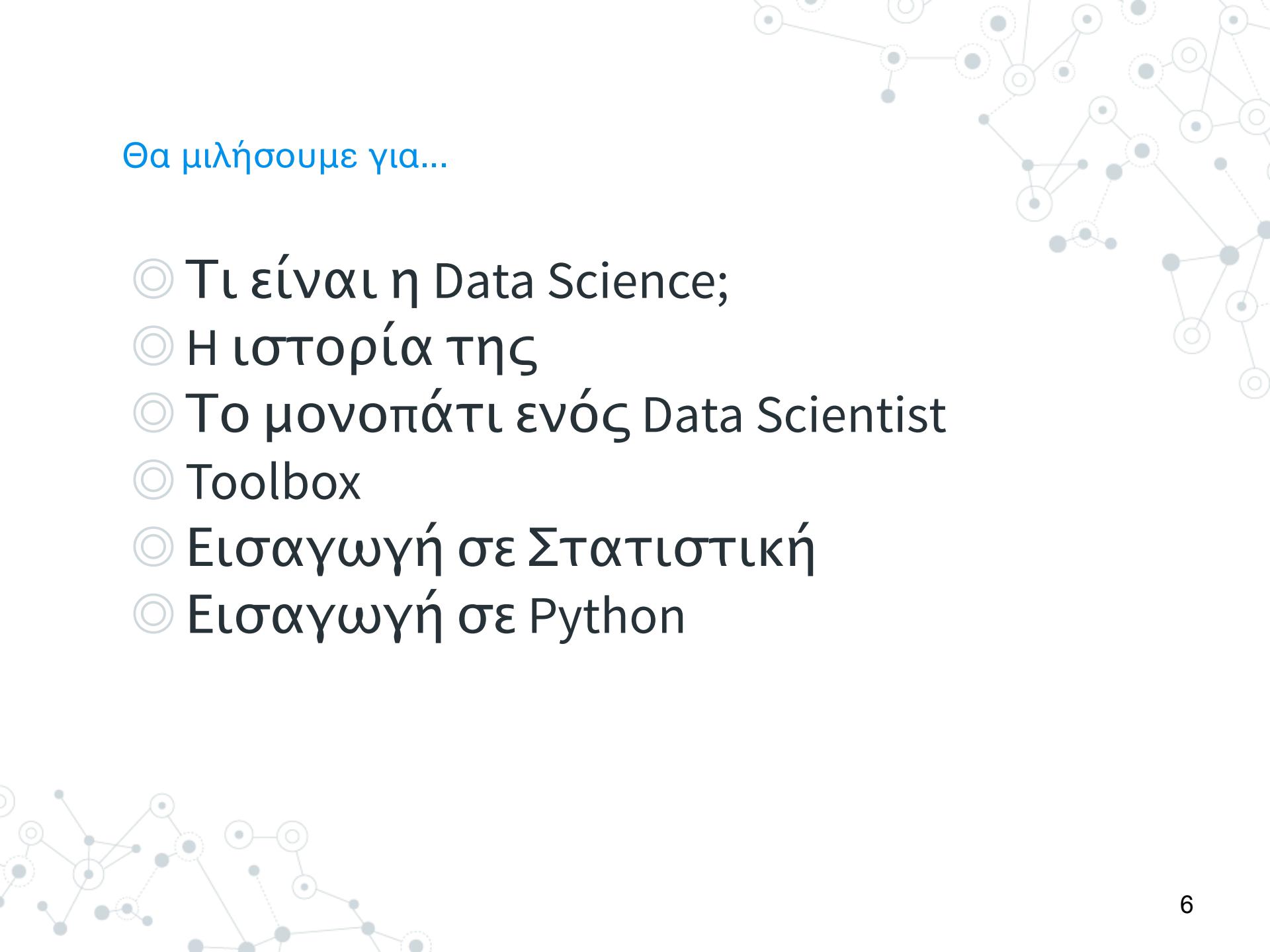
# The Data Science Handbook



<http://www.thedatasciencehandbook.com/get-the-book>

# Day 1

Εισαγωγή 😊



Θα μιλήσουμε για...

- ◎ Τι είναι η Data Science;
- ◎ Η ιστορία της
- ◎ Το μονοπάτι ενός Data Scientist
- ◎ Toolbox
- ◎ Εισαγωγή σε Στατιστική
- ◎ Εισαγωγή σε Python

# 1.

# Data Science

Ή αλλιώς, η Επιστήμη των Δεδομένων. Τι είναι και αν αξίζει να επενδύσω σε αυτήν;



<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

— MENU  
**Harvard  
Business  
Review**

DATA

# Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

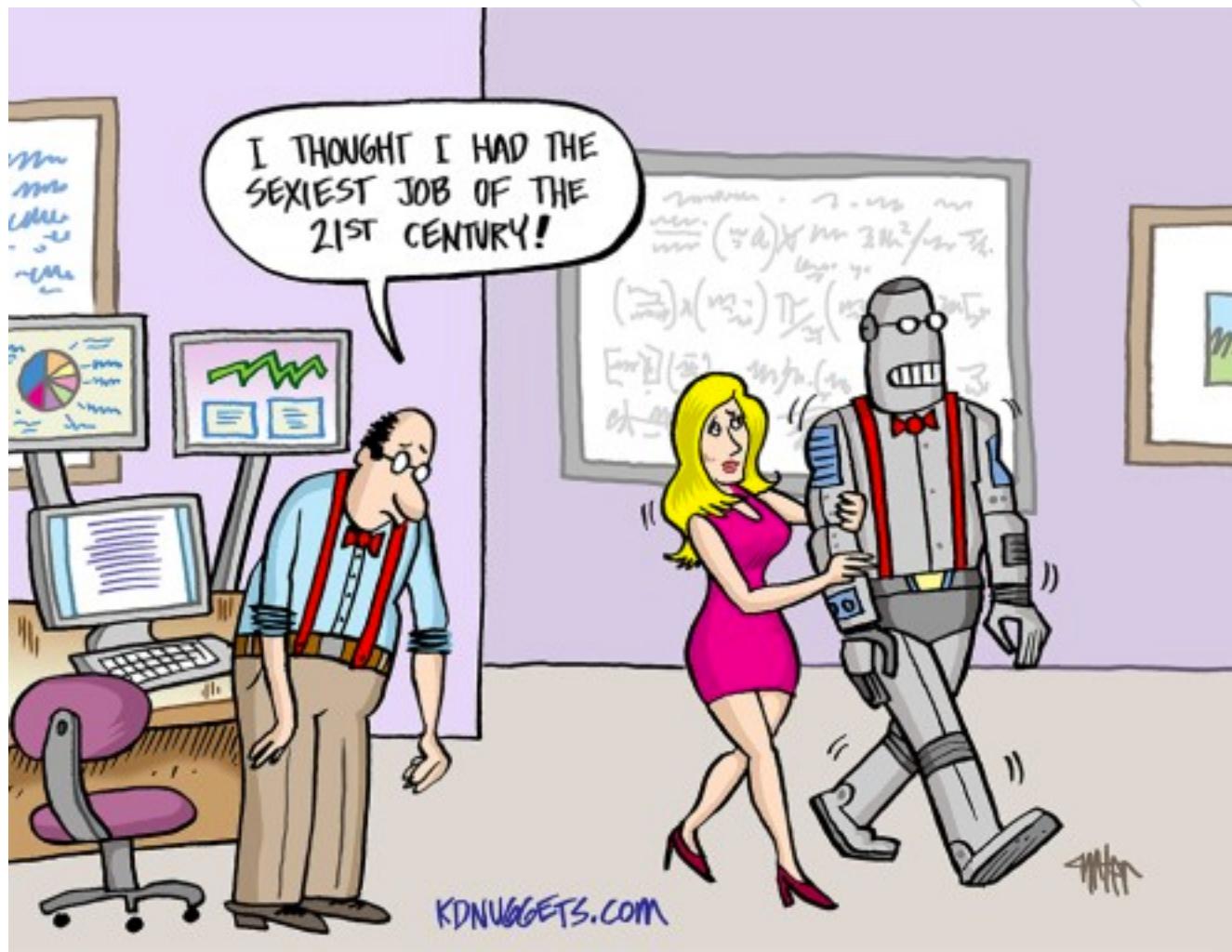
FROM THE OCTOBER 2012 ISSUE

---

 SUMMARY    SAVE    SHARE    COMMENT (4)    TEXT SIZE    PRINT    \$8.95 BUY COPIES

---

**W**hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join.





“

Η επιστήμη των Δεδομένων είναι ένα **διεπιστημονικό** πεδίο του οποίου αντικείμενο είναι η εξαγωγή της γνώσης από αδόμητα ή δομημένα δεδομένα.

--Wikipedia



“

*A field of Big Data which seeks to provide meaningful information from large amounts of complex data. Data Science combines **different fields of work** in statistics and computation in order to interpret data for the purpose of decision making.*

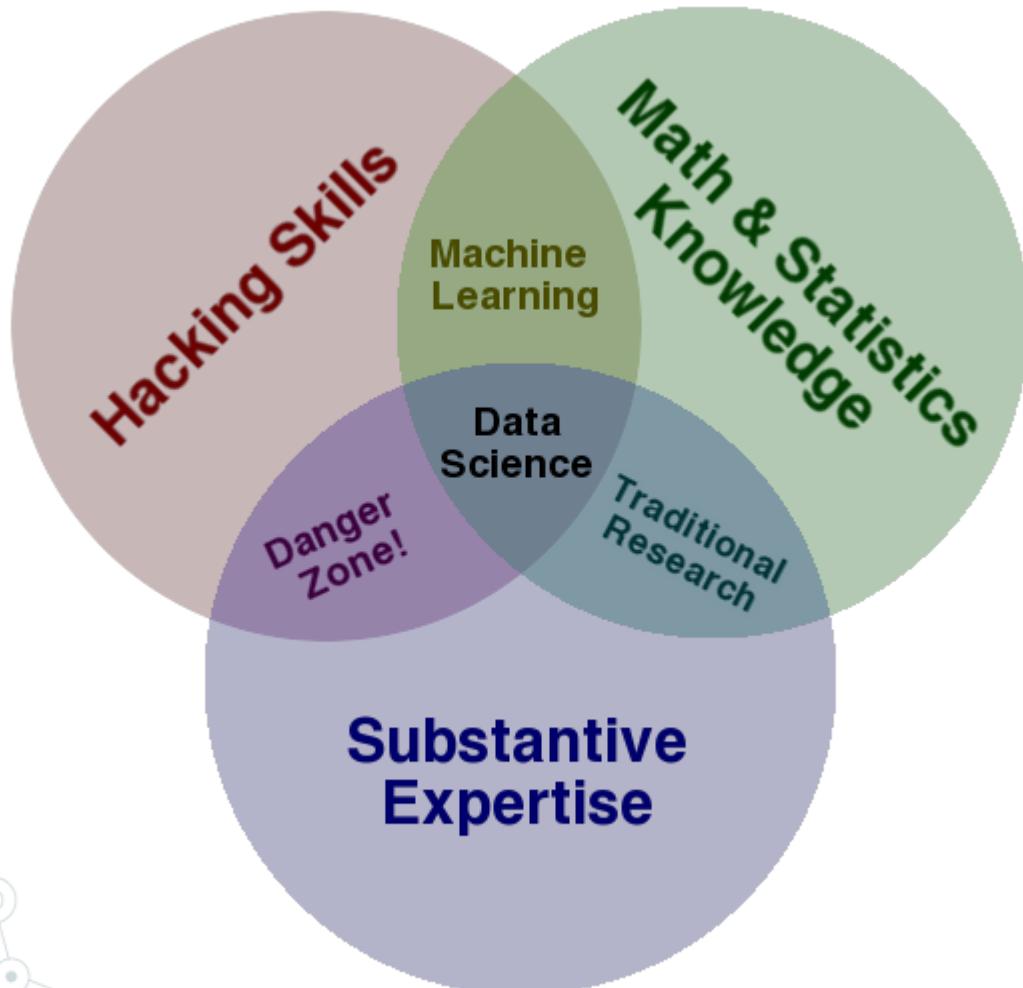
--*investopedia.com*

“

## *The creation of data products*

\*Data product = Ένα εργαλείο που δημιουργήθηκε με τη χρήση δεδομένων και βιοηθάει στη λήψη αποφάσεων.

My favourite



# Data Scientist

Ποιος λοιπόν μπορεί να έχει τον  
τίτλο του Data Scientist;



“



**Josh Wills**  
@josh\_wills

Follow



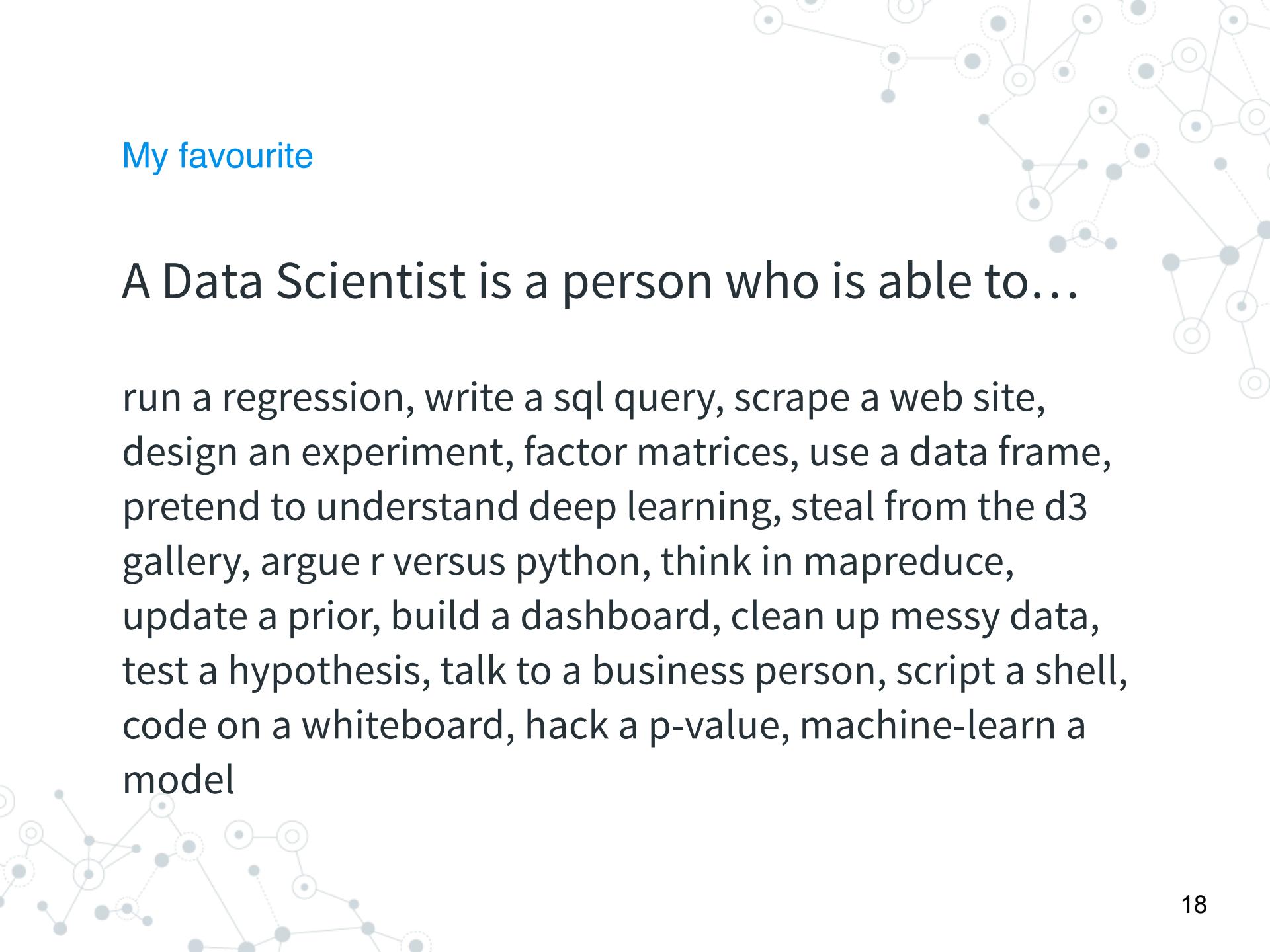
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.



*A Data Scientist is a **statistician**  
who lives in San Francisco 😊*

# Πικρή Αλήθεια #1





My favourite

A Data Scientist is a person who is able to...

run a regression, write a sql query, scrape a web site,  
design an experiment, factor matrices, use a data frame,  
pretend to understand deep learning, steal from the d3  
gallery, argue r versus python, think in mapreduce,  
update a prior, build a dashboard, clean up messy data,  
test a hypothesis, talk to a business person, script a shell,  
code on a whiteboard, hack a p-value, machine-learn a  
model



# Πικρή Αλήθεια #2

For the rest...

you are just THE “data-guy”

(or THE “math-guy”)





# **Ποια είναι η αξία της Data Science;**

Τι απάντησαν 5 Data Scientist



“

*When you ask me what the value of data science is, it's almost, like explaining the value of water to a fish.*

--Media



“

*We're going to help the business go to new places that it  
hasn't yet even thought of going.*

--Biotechnology



“

*If we didn't have a data science capability we would lose  
money.*

--Manufacturing



“

*We have an asset: it's the data. And what you do with that data dictates whether you'll be differentiated in the future.*

--Retail



“

*The business realized that, nowadays, we cannot be competitive if we are not data-savvy enough.*

--Banking



2.

## Η ιστορία

Πως ξεκίνησαν όλα;

## Χρονοδιάγραμμα

- ◎ 1960 - Computer Science = Data Science από Peter Naur
- ◎ 1974 - Πρώτη φορά σε δημοσίευση από Peter Naur
- ◎ 1996 - Συνέδριο με τίτλο “Data Science, classification, and related methods”
- ◎ 1997 - Ομιλία του Jeff Wu με τίτλο “Statistics = Data Science?”
- ◎ 2001 - William S. Cleveland χρησιμοποίησε τη Data Science ως ανεξάρτητο όρο σε άρθρο της “International Statistical Review”
- ◎ 2002 - Committee on Data for Science & Technology. Νέο περιοδικό με τίτλο Data Science Journal
- ◎ 2003 - The Journal of Data Science από Columbia University
- ◎ 2008 - DJ Patil & Jeff Hammerbacher χρησιμοποίησαν τον τίτλο Data Scientist
- ◎ 2012 - Άρθρο από Harvard Business Review με τίτλο “Data Scientist: The Sexiest Job of the 21st Century”

Data Science ≠ Big Data

Apollo XI, 1969

64Kb

SkyDive Stratos, 2012

Δεκάδες Gigabytes

3.

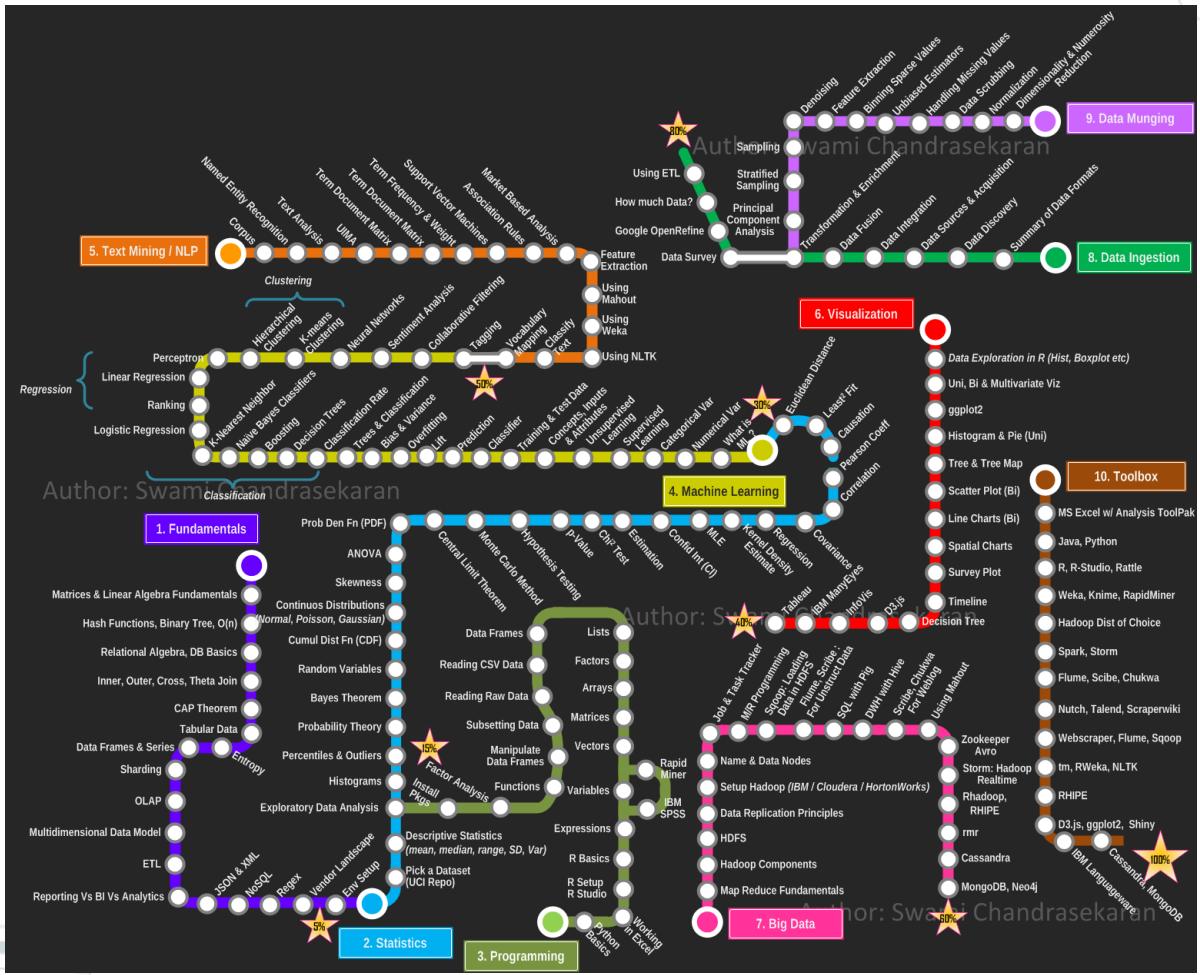
## To μονοπάτι ενός Data Scientist

Yeah! I had a skill up...

# Πικρή Αλήθεια #3



# Το μυοπάτι του Data Scientist



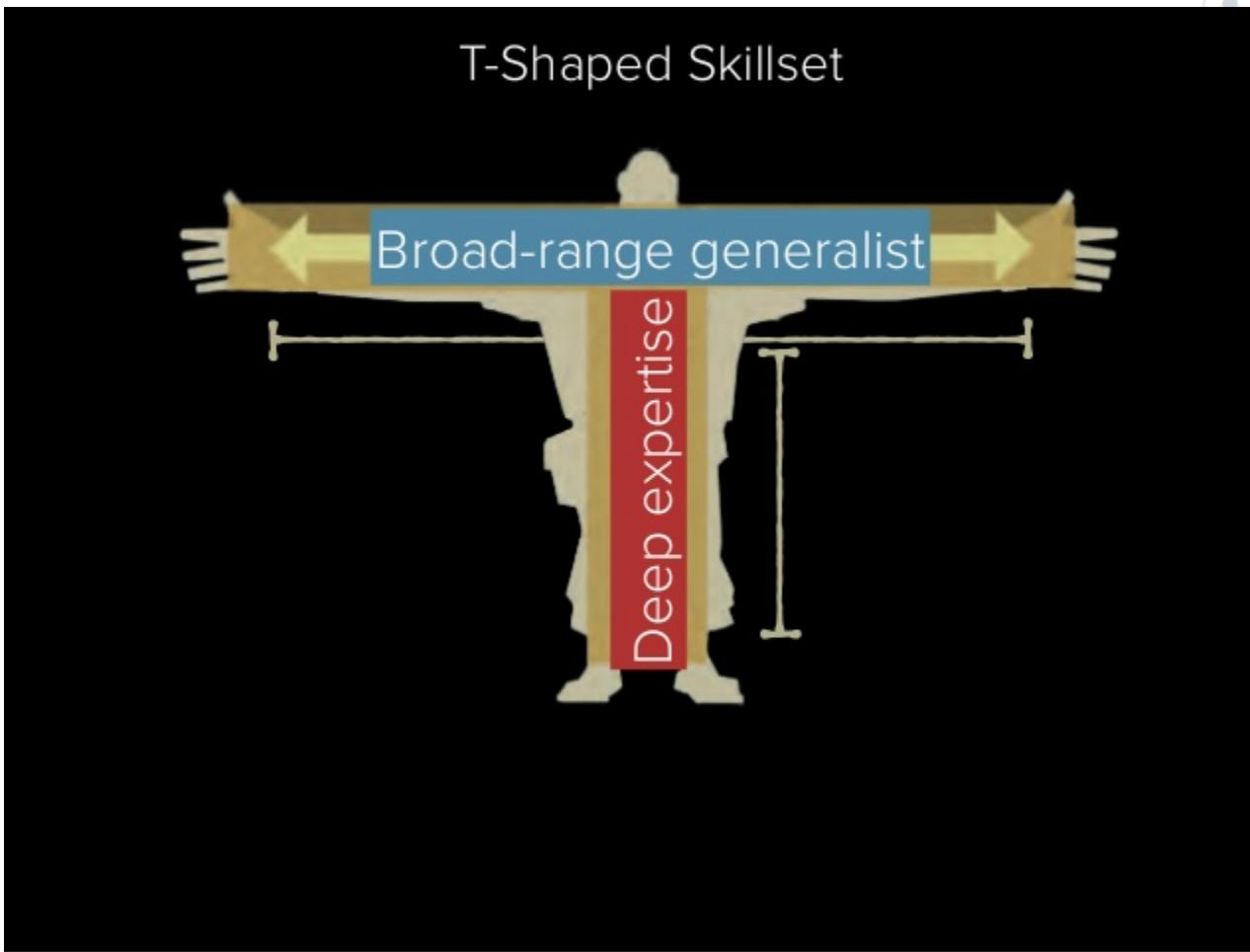
Focus on....

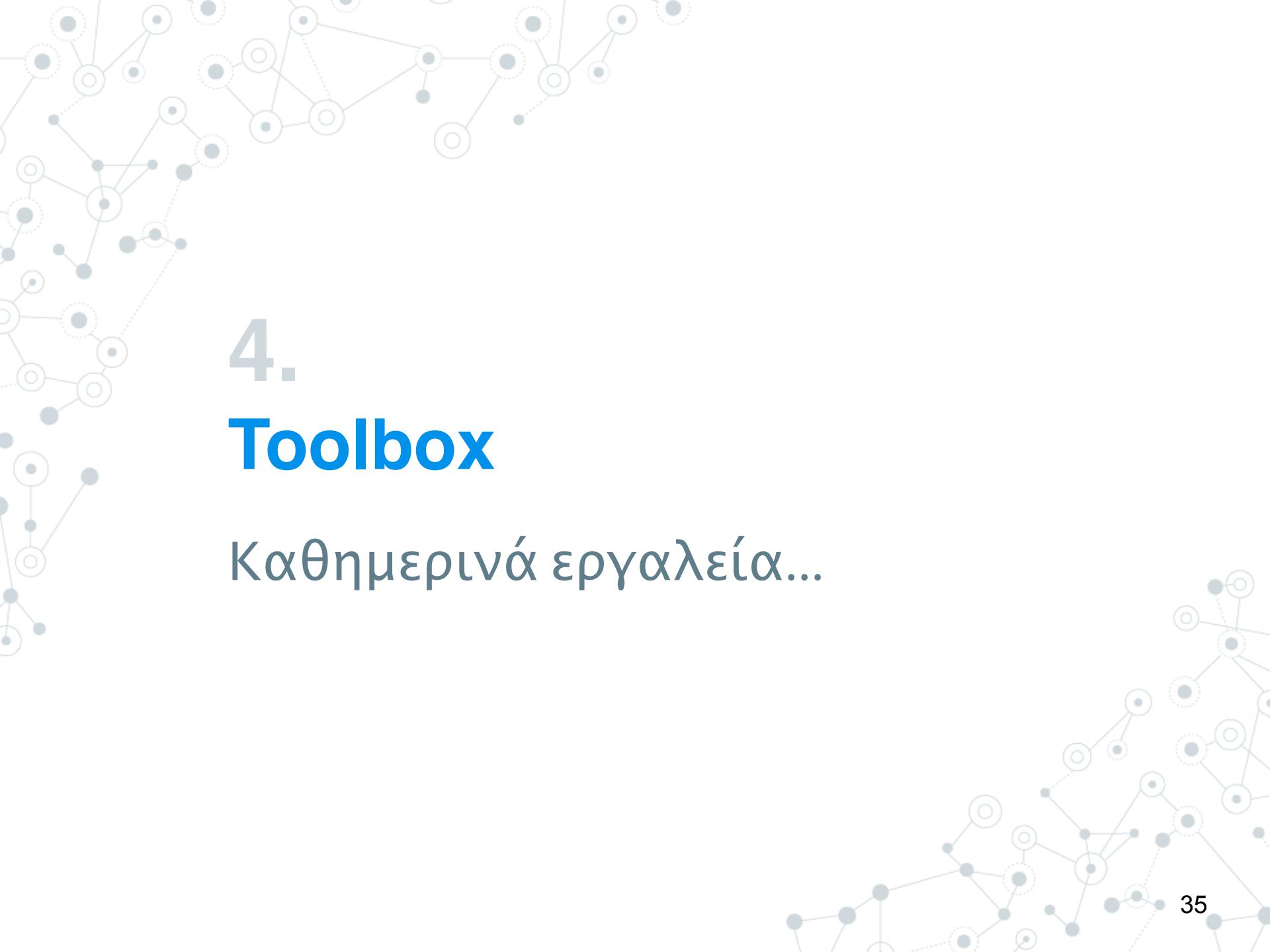
The math way

The tool way

Η αλήθεια στη μέση

Στην πράξη....





## 4. **Toolbox**

Καθημερινά εργαλεία...



Θα χρειαστείς...

- Git
- Virtual Machines
- Excel!!!!
- Python/R
- SQL



Extra:

- Dataiku
- Azure ML



# Συμβουλή

## #1

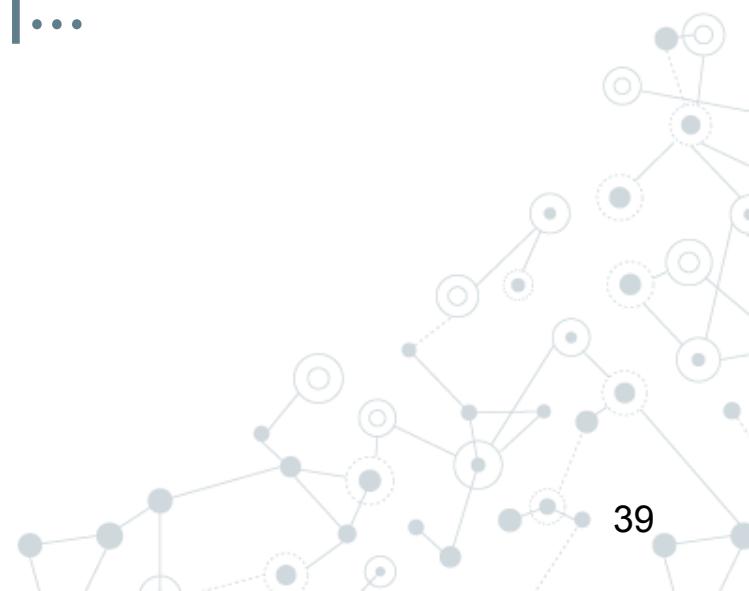


Μυαλό = Επεξεργασία & Αποφάσεις  
Μυαλό ≠ Αποθήκευση



## 5. **Στατιστική**

Μια μικρή επανάληψη...



Έννοιες και η σημασία τους...

◎ **Τυχαίο δείγμα**

◎ **Μεταβλητές**

- **Κατηγορικές - Ποιοτικές**
  - Ονομαστική (χρώμα ματιών, τόπος γέννησης, φύλο)
  - Διάταξης (μορφωτικό επίπεδο, κλάσεις ηλικιών)
- **Ποσοτικές**
  - Διακριτή (πόσες σοκολάτες τρώω κάθε μέρα)
  - Συνεχής (το ύψος ανθρώπων)



# Συμβουλή

## #2

Κατηγορικές Μεταβλητές

Δεκαδικός αριθμός

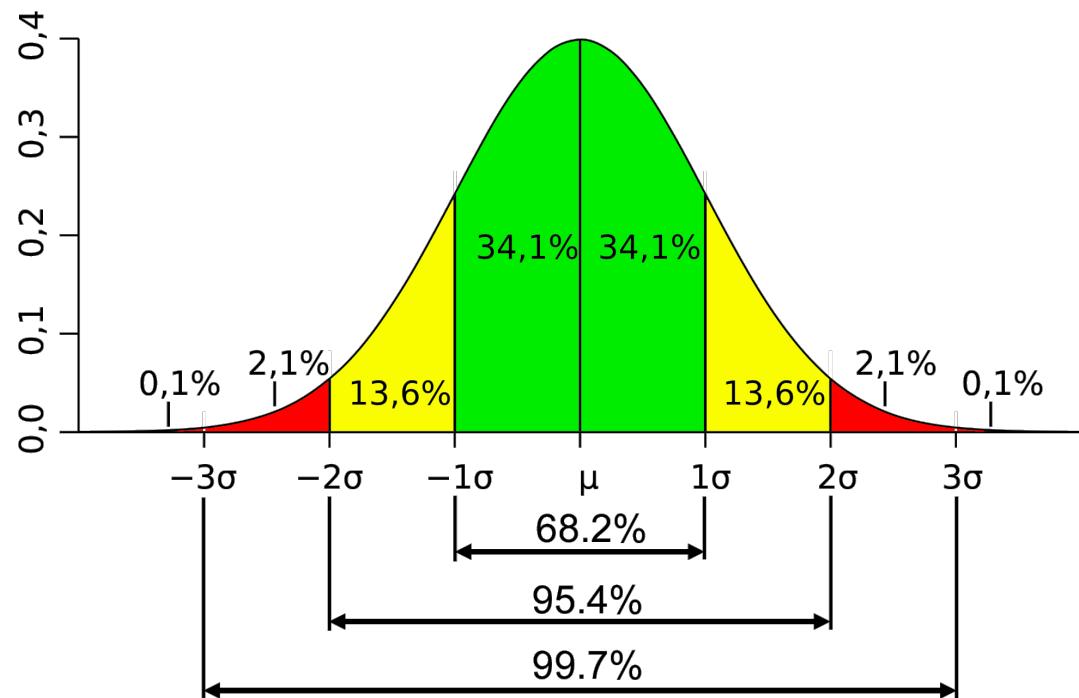
$\neq$

Κατηγορική μεταβλητή

## Μέτρα Θέσης

- ◎ Μέση τιμή ( $x$ )
- ◎ Διάμεσος ( $\delta$ )
- ◎ Εύρος ( $R$ )
- ◎ Διακύμανση ( $s^2$ )
- ◎ Τυπική Απόκλιση ( $s$ )
- ◎ Συντελεστής Μεταβολής ( $s/x$ )

## Normal Distribution

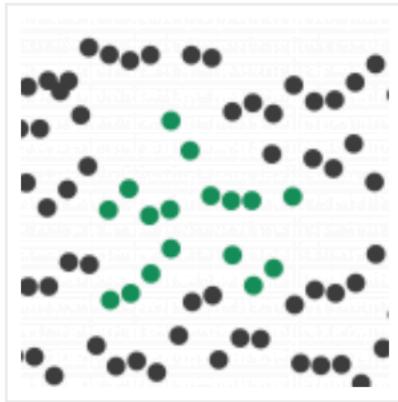


# Statistical Bias

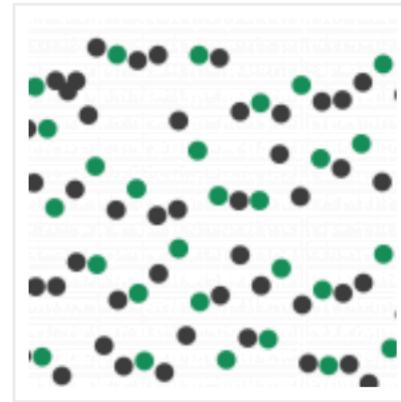
Τα 8 θανάσιμα αμαρτήματα

# Statistical bias #1: Selection bias

Selection bias occurs, when you are selecting your sample or your data wrong. Usually this means accidentally working with a specific subset of your audience instead of the whole, hence your sample is not representative of the whole population. There are many underlying reasons, but by far the most typical I see: collect and work only with data that is *easy to access*.



*selection bias*

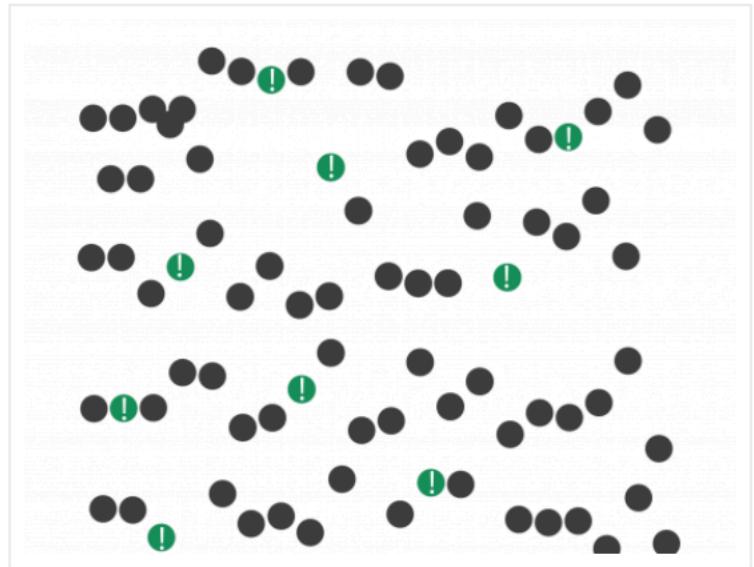


*proper random sampling*

Source: <https://data36.com/statistical-bias-types-explained/>

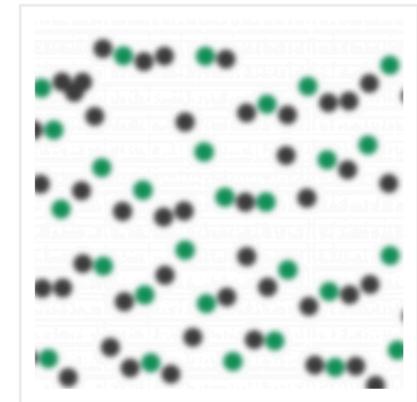
# Statistical bias #2: Self-Selection bias

Self-selection bias is a subcategory of selection bias. If you let the subjects of your analyses/researches select themselves, that means that less proactive people will be excluded. The bigger issue is that self-selection is a specific behaviour – that implies other specific behaviours – thus this sample does not represent the entire population.



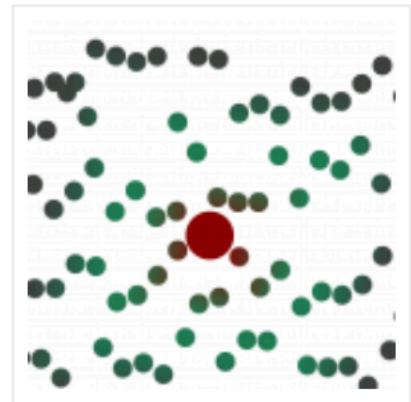
## Statistical bias #3: Recall bias

Recall bias is another common error of interview/survey situations, when the respondent doesn't remember correctly for things. It's not bad or good memory – humans have selective memory by default. After a few years certain things stay, others fade. It's normal, but it makes researches much more difficult.



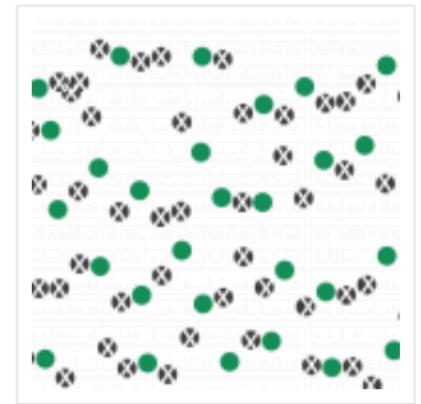
## Statistical bias #4: Observer bias

Observer bias is happening, when the researcher subconsciously projects his/her expectations to the research. It can come in many forms. Eg. (unintentionally) influencing the participants (only at interviews and surveys) or doing some serious **cherry picking** (focusing rather on the statistics that support our hypothesis, than to the statistics, that doesn't.)



## Statistical bias #5: Survivorship bias

Survivorship bias is a statistical bias type, where the researcher is focusing only to that part of the data set, that already went through some kind of pre-selection process – and missing those data-points, that fell off during this process (because they are not visible anymore).



# Statistical Bias #6: Omitted Variable Bias

Omitted Variable Bias occurs, when you are leaving out one or more important variables from your model. This issue comes up especially often regarding Predictive Analytics.

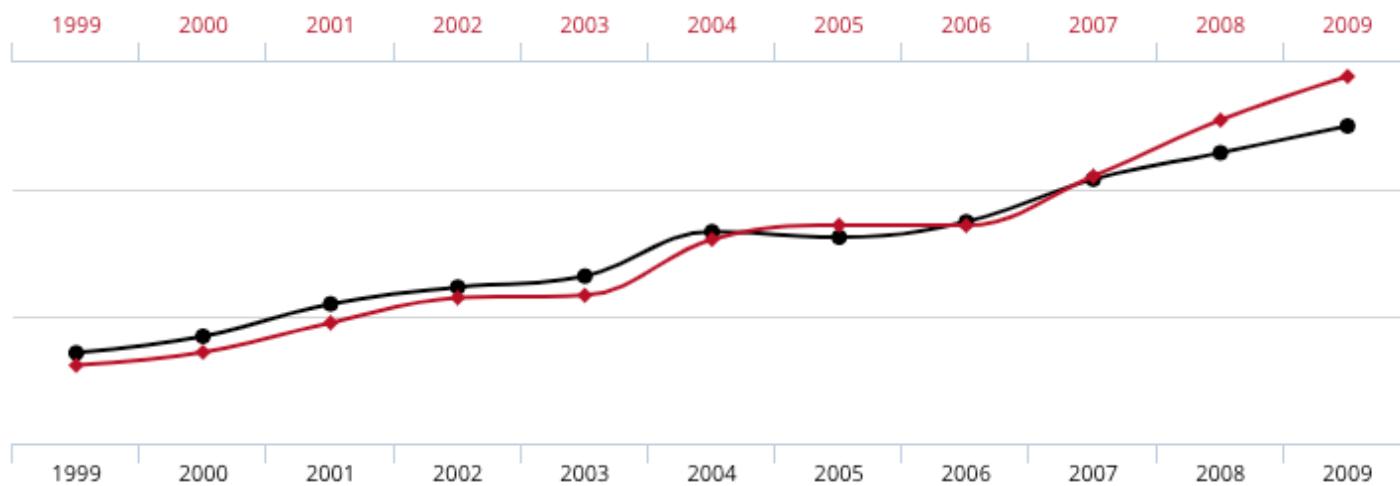
a b c d e f

# Statistical Bias #7: Cause-effect Bias

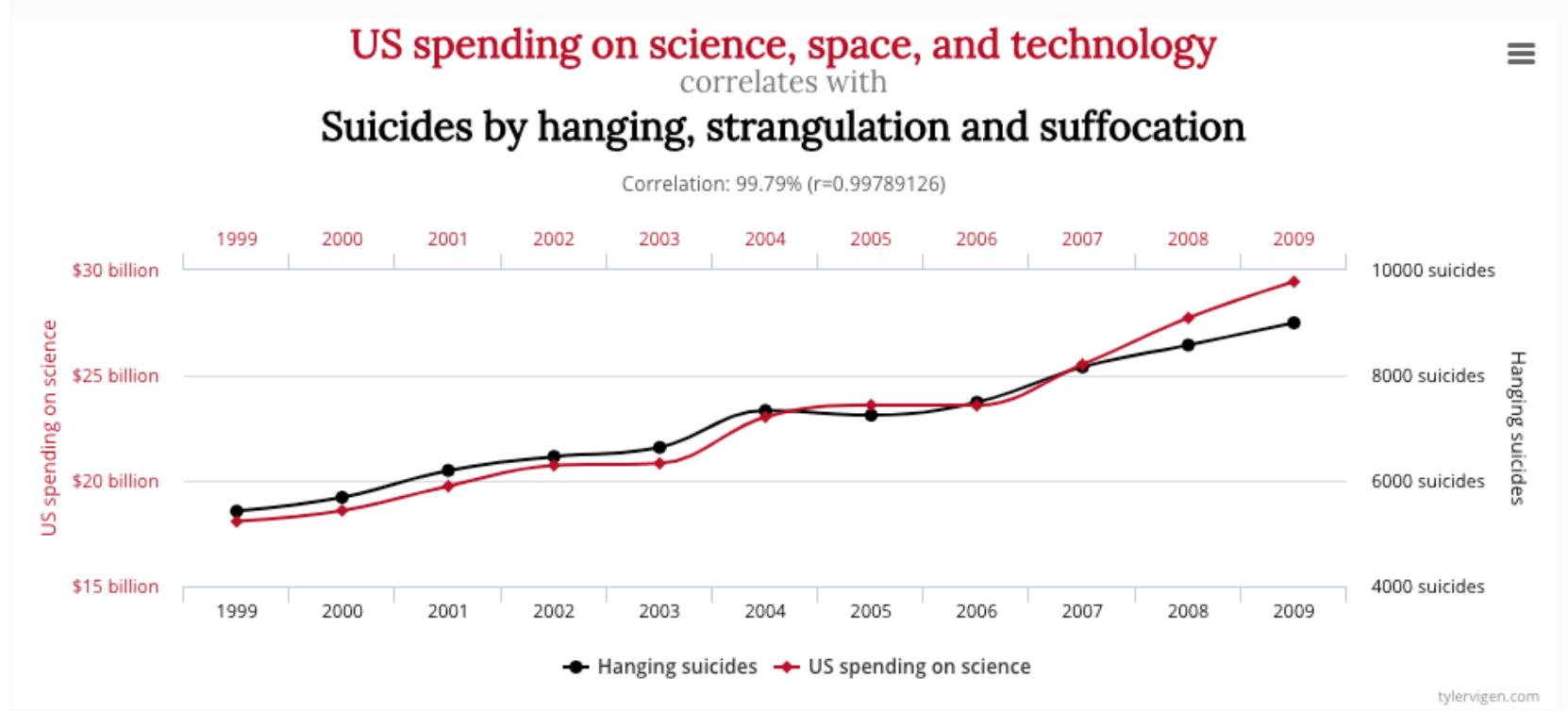
Our brain is wired to see causation everywhere, where correlation shows up. Cause-effect bias is usually not mentioned as a classy statistical bias, but I wanted to include it on this list as many decision makers (business/marketing managers) are not aware of that. Even those (me too), who are aware of it, have to remind themselves from time to time: correlation does not imply causation.



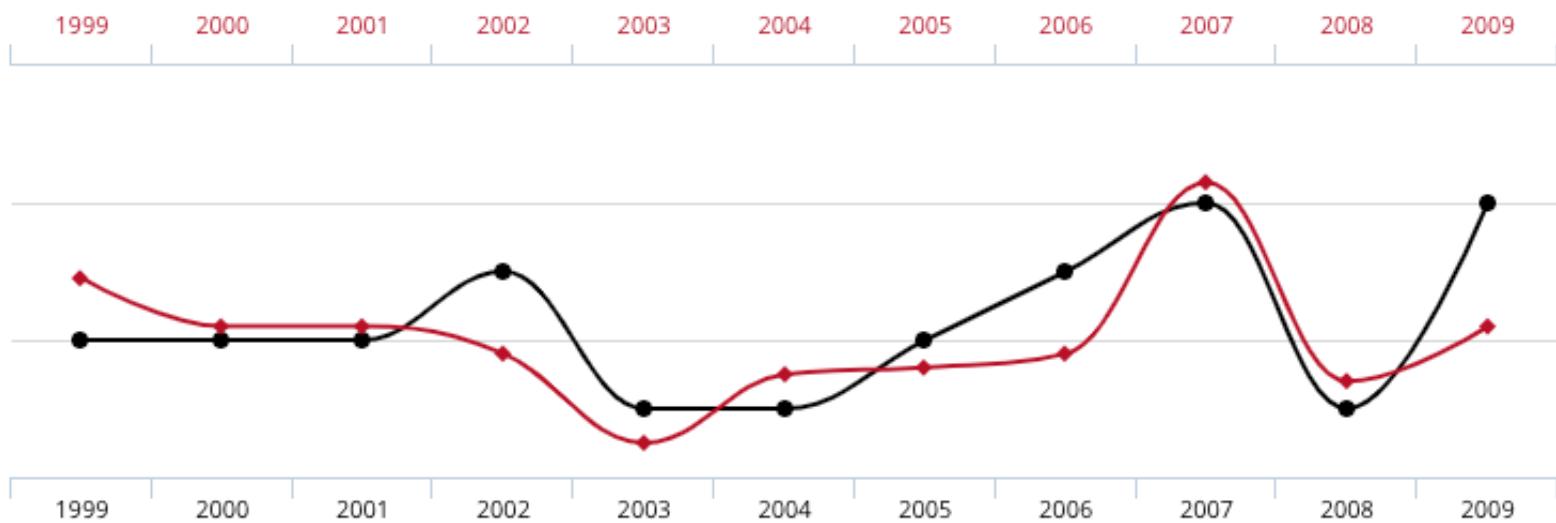
Correlation does not imply causation



## Correlation does not imply causation



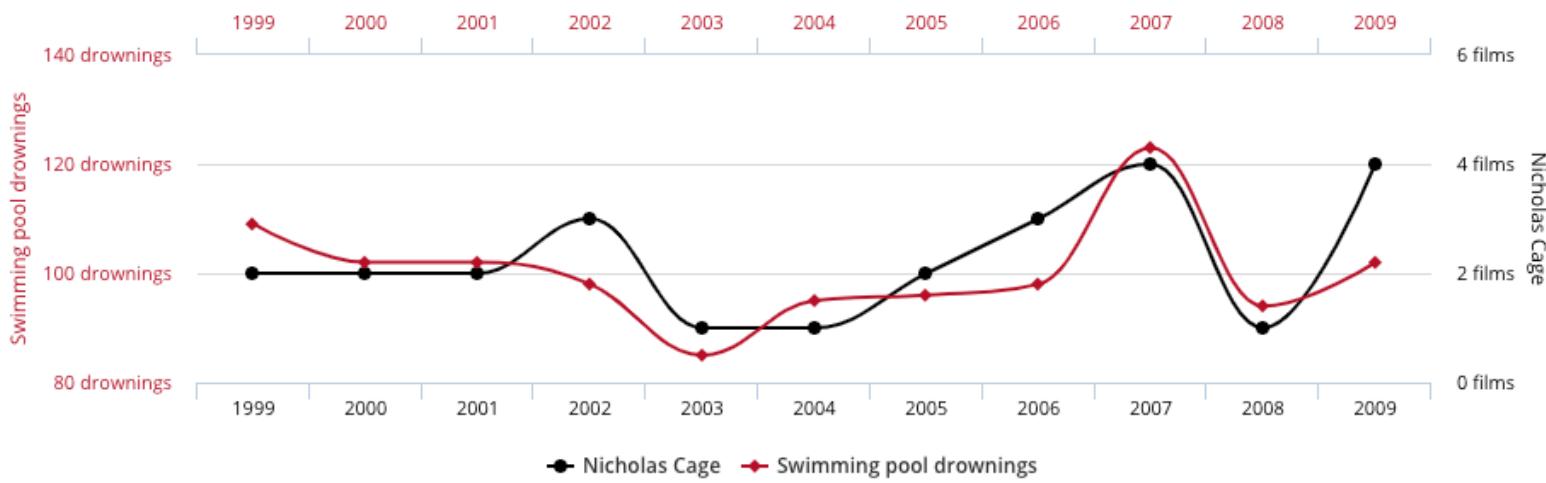
Correlation does not imply causation



## Correlation does not imply causation

Number of people who drowned by falling into a pool  
correlates with  
Films Nicolas Cage appeared in

Correlation: 66.6% ( $r=0.666004$ )





# More fun on..

<http://tylervigen.com/spurious-correlations>



# Statistical Bias #8: Funding Bias

I briefly mentioned Funding Bias (sometimes called sponsorship bias) already in [Statistical Bias Types part 1](#). We are talking about it, when the results of a scientific study is biased in a way, that it supports the financial sponsor of the research.

\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$\$



# 5. Python

Get our hands dirty!

## Η επόμενη μέρα

- Git
- Data file types & Copyrights
- More Python (😍)
- Working on a Data Science Project
- Intro to Machine Learning

# Thanks!

Any questions?