29 Σεπτεμβρίου 2018

Data Science Workshop

GOLDEN GATE PRO
CODING ACADEMY

Golden Gate Pro

Λίγα λόγια

Hello!

Τάσος Βεντούρης

Data Scientist and
Game Designer @
Hattrick Ltd

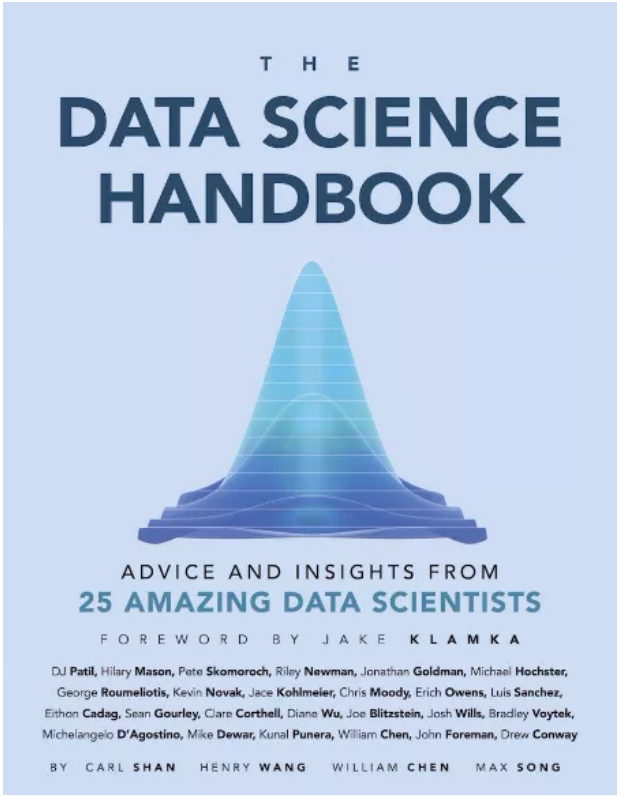
You can find me at:

@tasosventouris Tasos Ventouris

More About Me!

- (2012) BSc Mathematics
- (2014) MSc Web Science
- (2015) Mentor @ Open Knowledge Inter.
- (2016) Offered a PhD in Web Science @ Southampton
- (2016) stackprime
- (2016) Data Scientist @ Hattrick
- (2017) Data Science Degree @ Microsoft

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

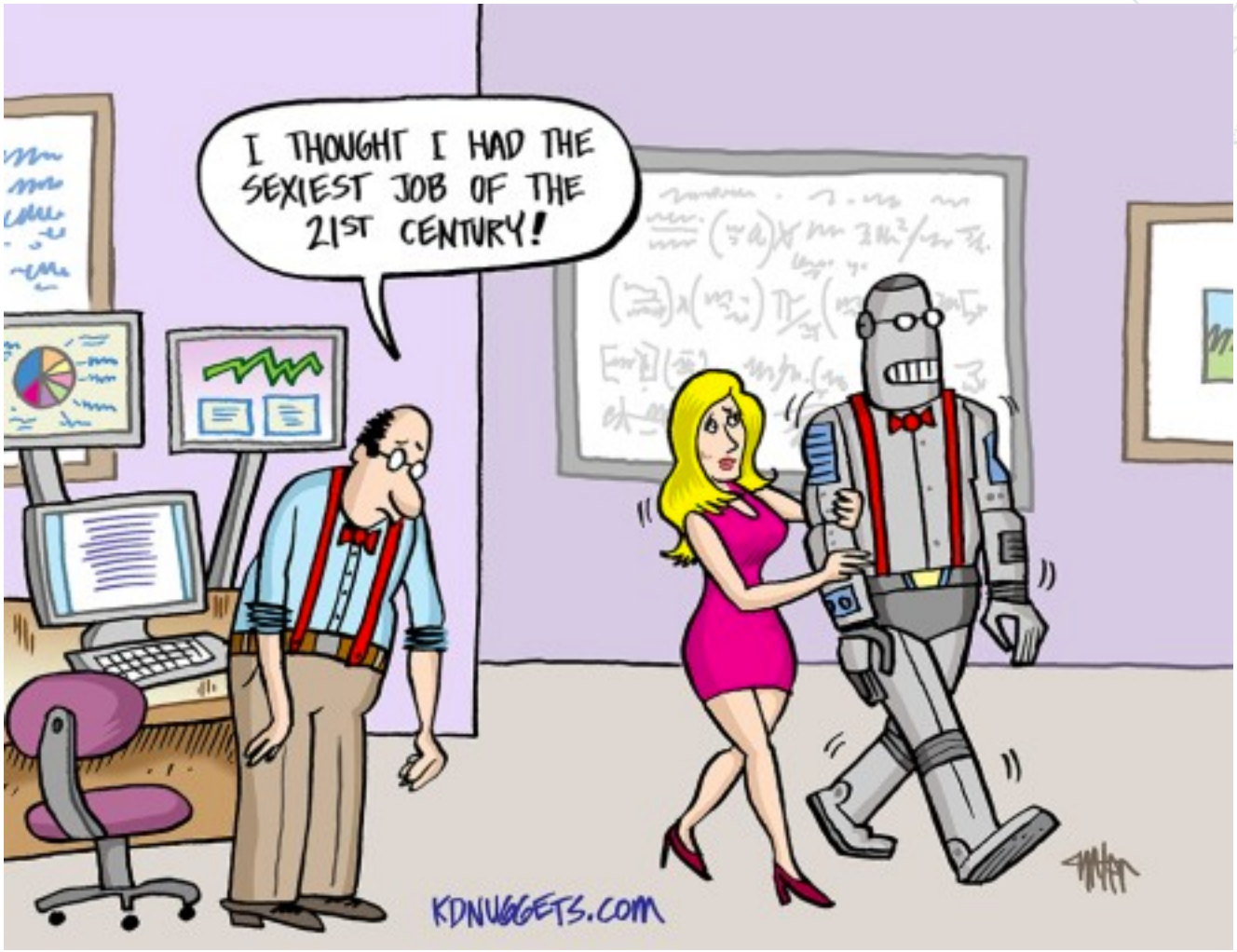
Data Scientist: The Sexiest Job of the 21st Century

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

When 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.


I THOUGHT I HAD THE SEXIEST JOB OF THE 21ST CENTURY!
KDNUGGETS.com

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

--Wikipedia

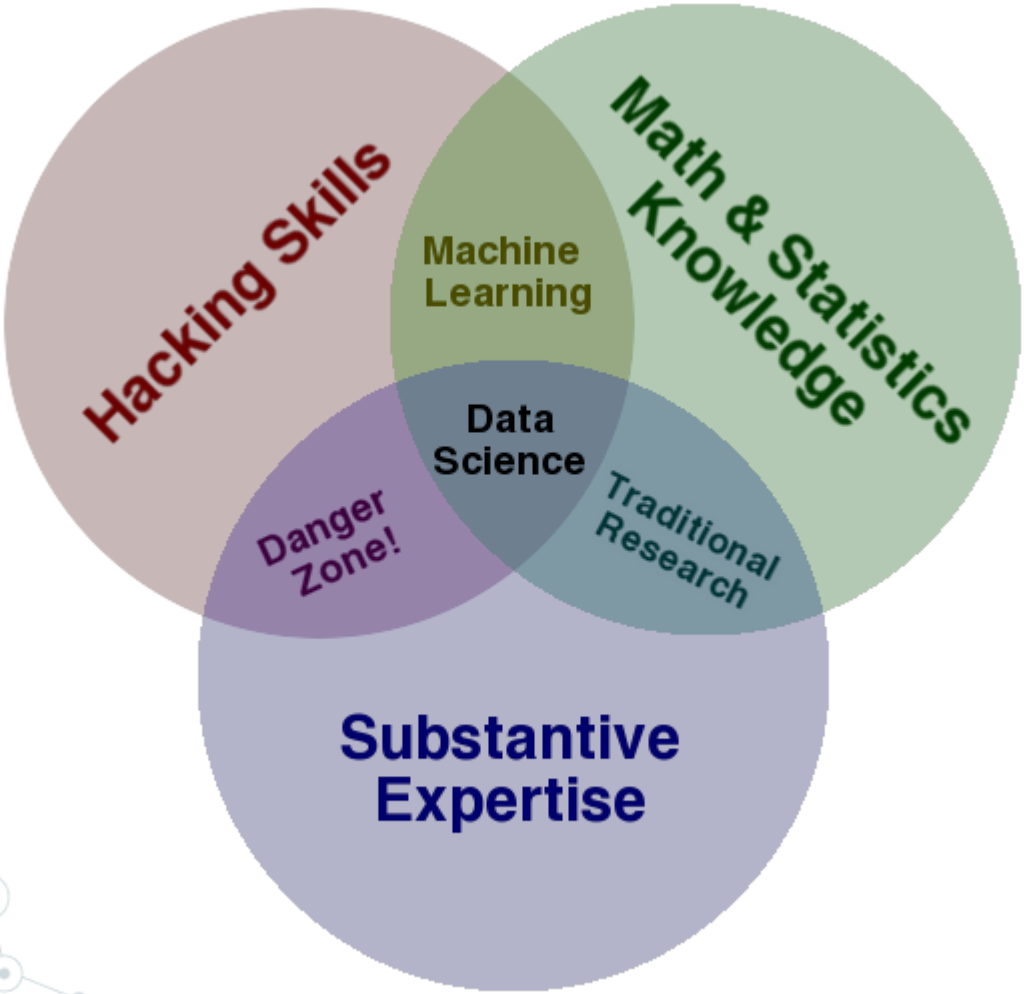
A field of Big Data which seeks to provide meaningful information from large amounts of complex data. Data Science combines ***different fields of work*** *in statistics and computation in order to interpret data for the purpose of decision making.*

--investopedia.com

The creation of data products

*Data product = Ένα εργαλείο που δημιουργήθηκε με τη χρήση δεδομένων και βοηθάει 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…

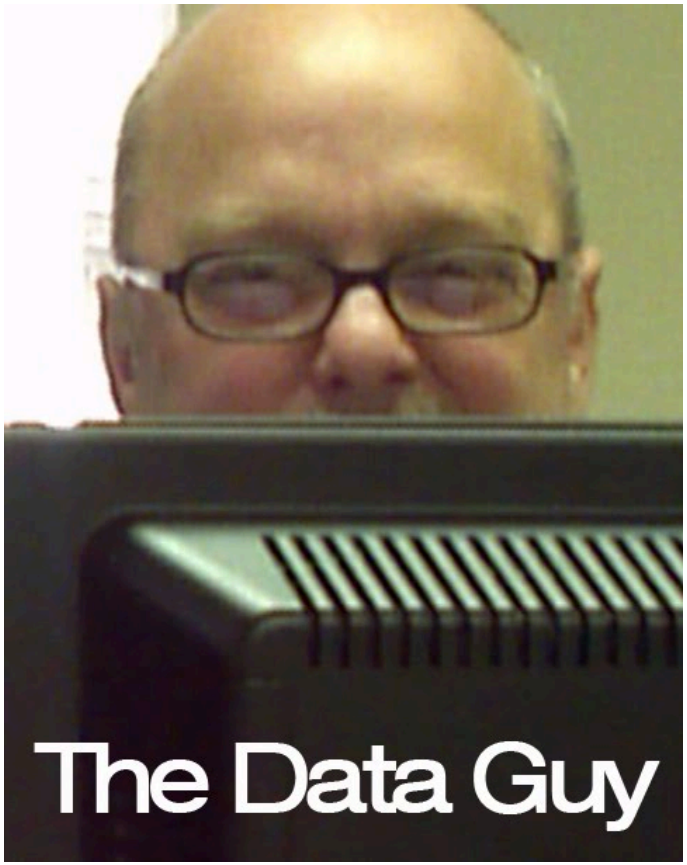
run a regression, write a sql query, scrape a web site, design an experiment, factor matrices, use a data frame, pretend to understand deep learning, steal from the d3 gallery, argue r versus python, think in mapreduce, update a prior, build a dashboard, clean up messy data, test a hypothesis, talk to 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

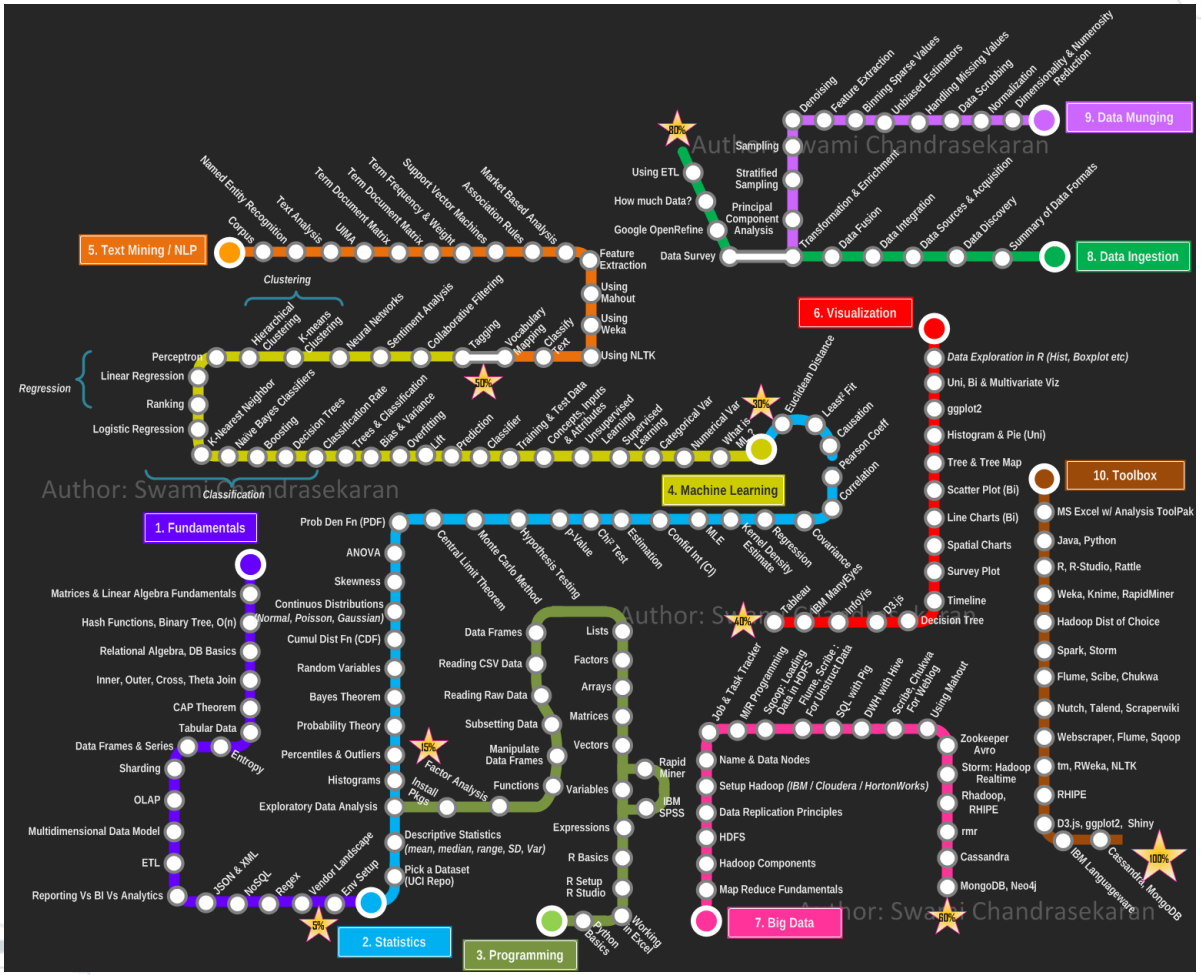
Δεκάδες Gigabytes

3.
Το μονοπάτι ενός Data Scientist

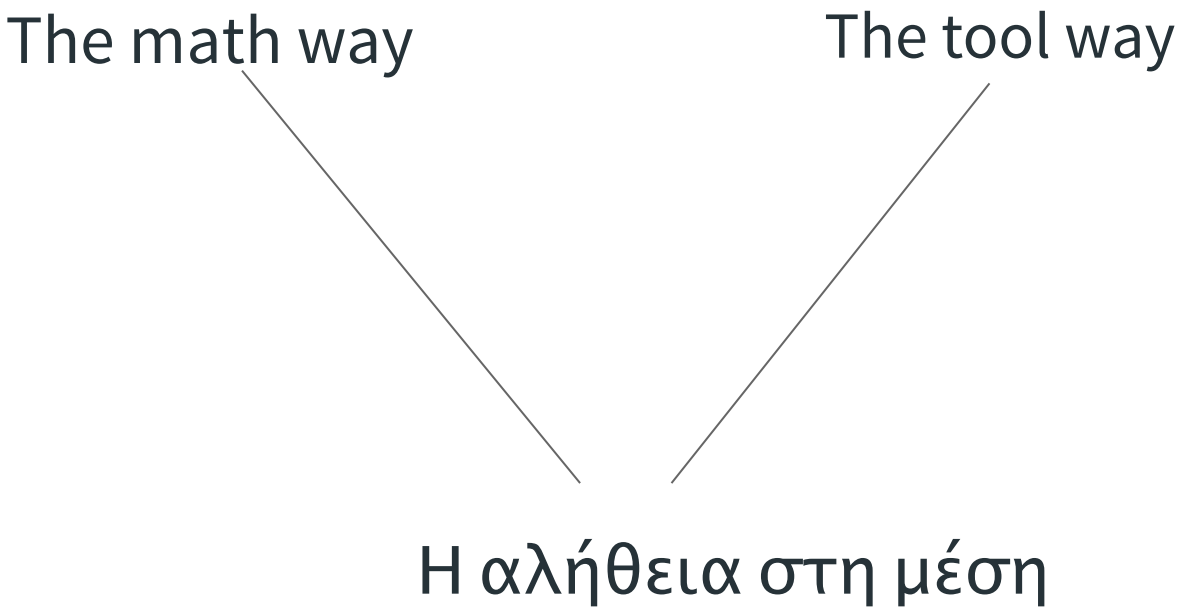
Yeah! I had a skill up...

Πικρή Αλήθεια #3

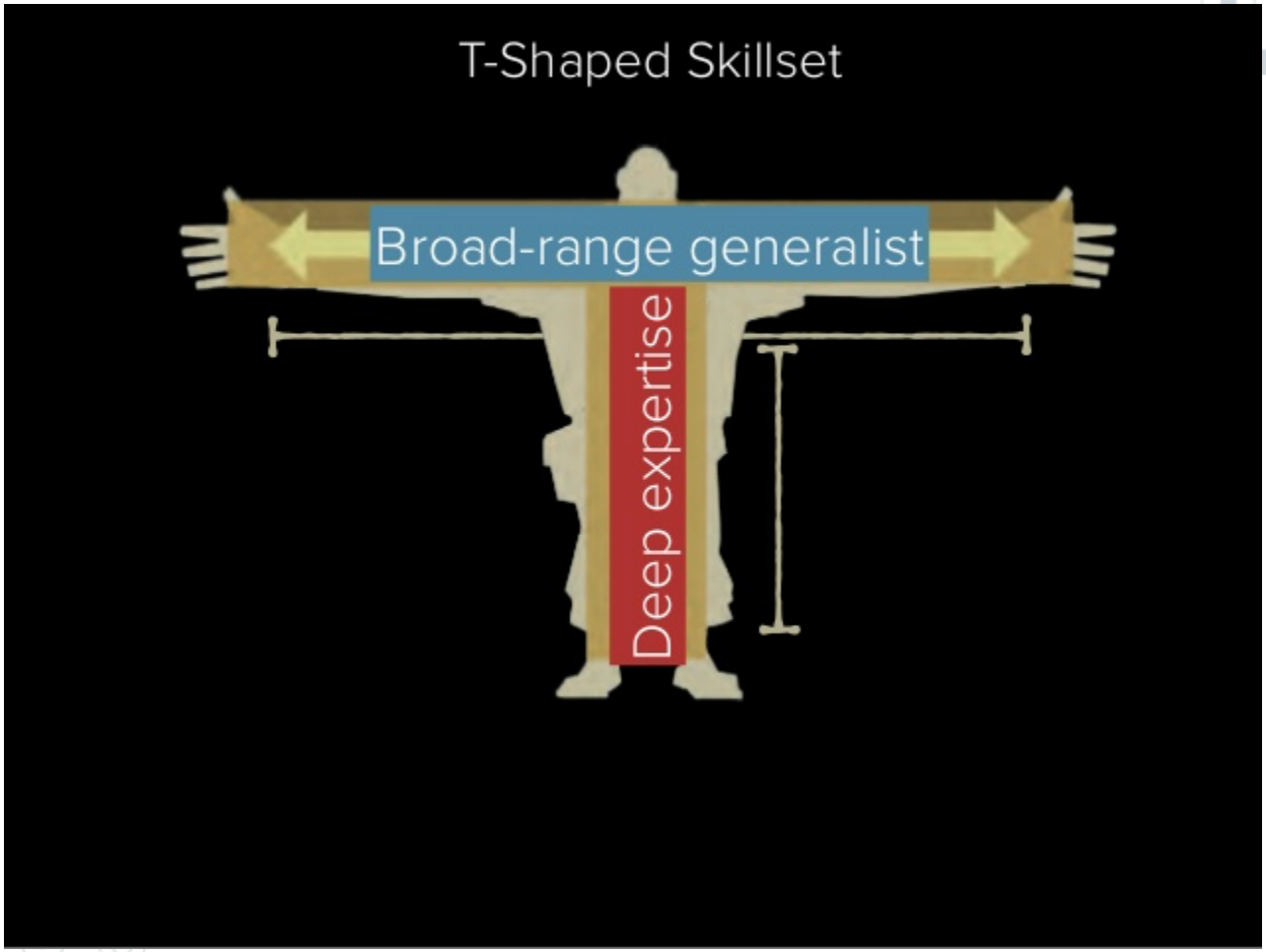
Το μονοπάτι του Data Scientist



Focus on….



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

Δεκαδικός αριθμός

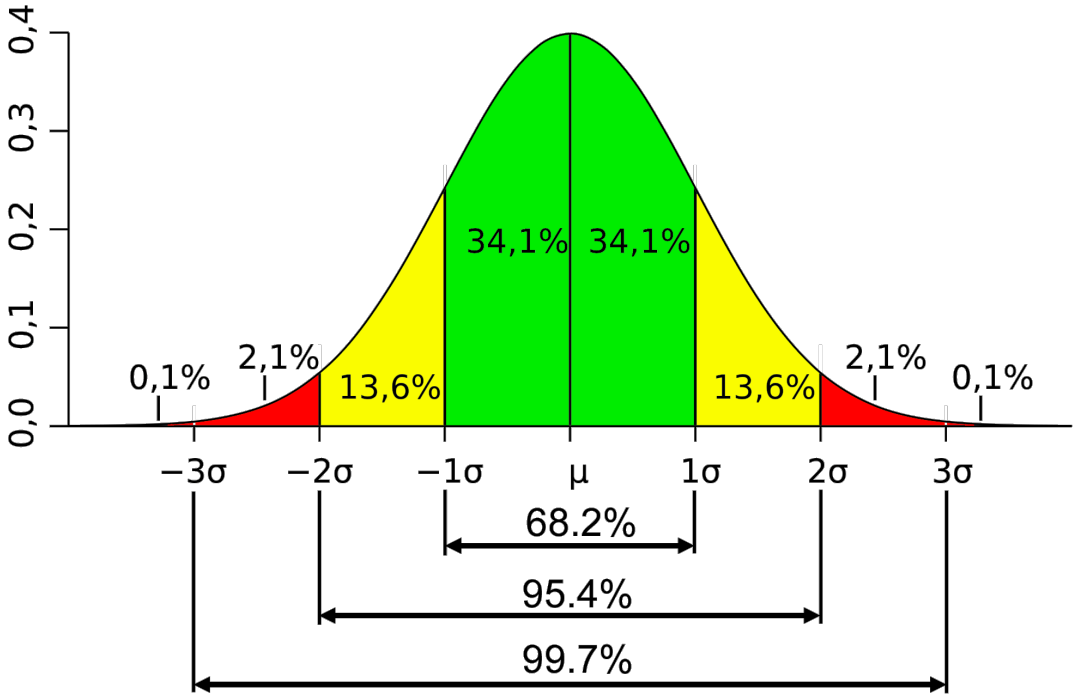
≠

Κατηγορική μεταβλητή

Μέτρα Θέσης

- Μέση τιμή (x)
- Διάμεσος (δ)
- Εύρος (R)
- Διακύμανση (s^2)
- Τυπική Απόκλιση (s)
- Συντελεστής Μεταβολής (s/x)

Normal Distribution

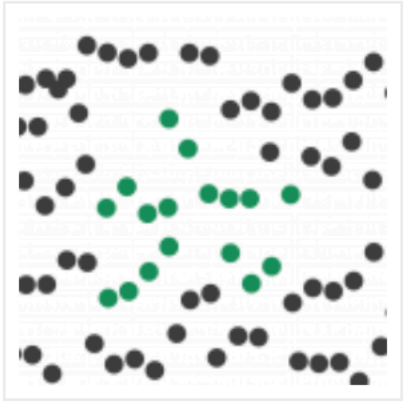


Statistical Bias

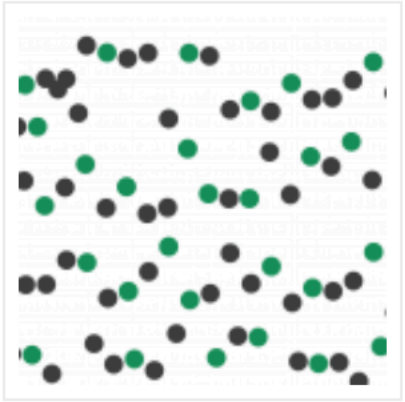
Τα 8 θανάσιμα αμαρτήματα

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

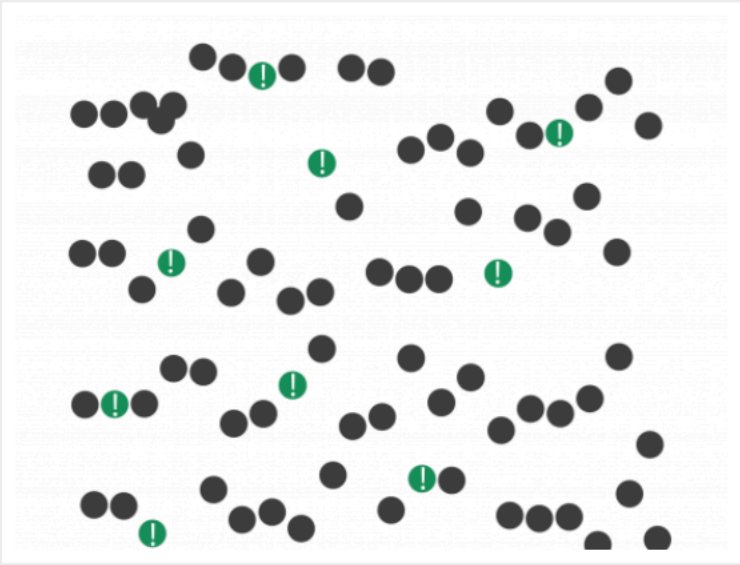


proper random sampling

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

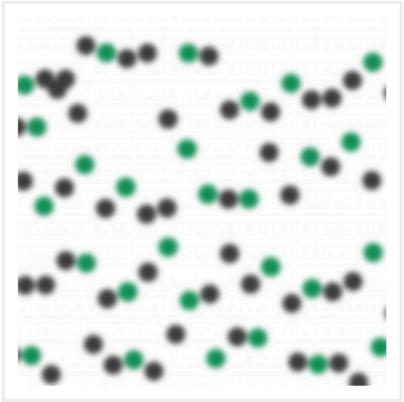
Statistical bias #2: Self-Selection bias

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.



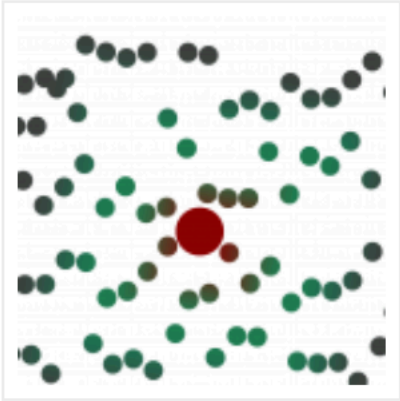
Statistical bias #3: Recall bias

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.



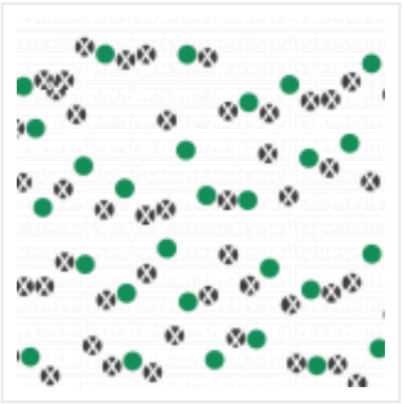
Statistical bias #4: Observer bias

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).

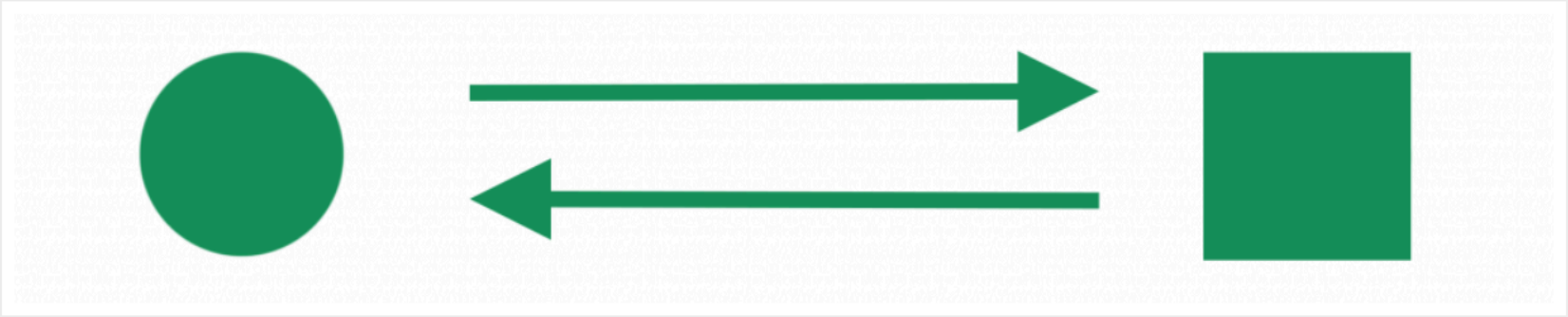


Statistical Bias #6: Omitted Variable Bias

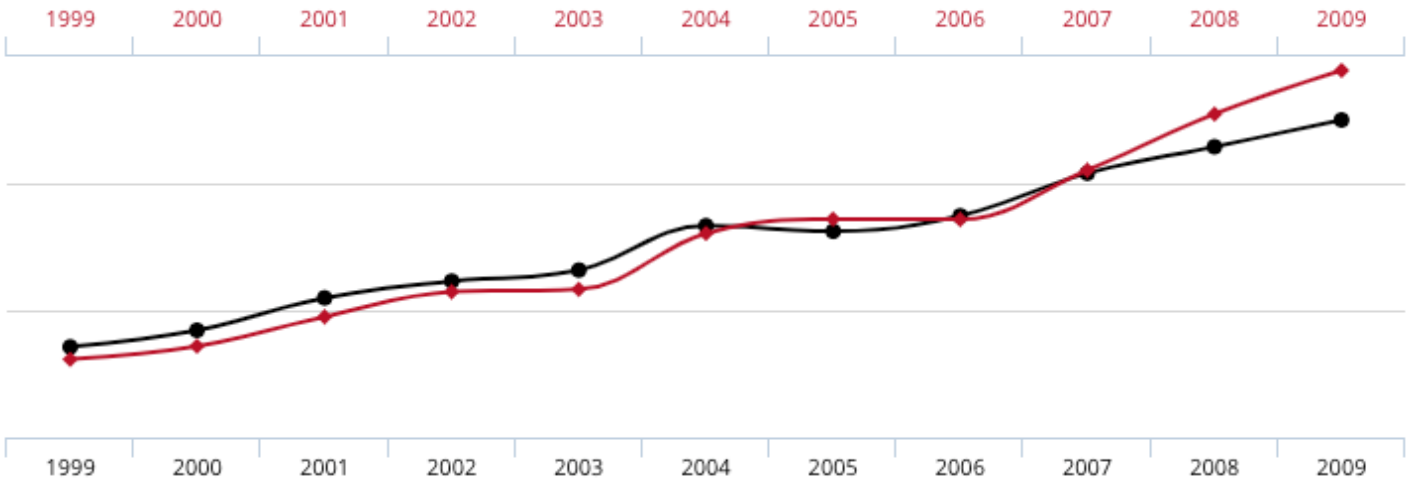
Omitted Variable Bias occurs, when you are leaving out one or more important variables from your model. This issue comes up especially often regarding Predictive Analytics.

Statistical Bias #7: Cause-effect Bias

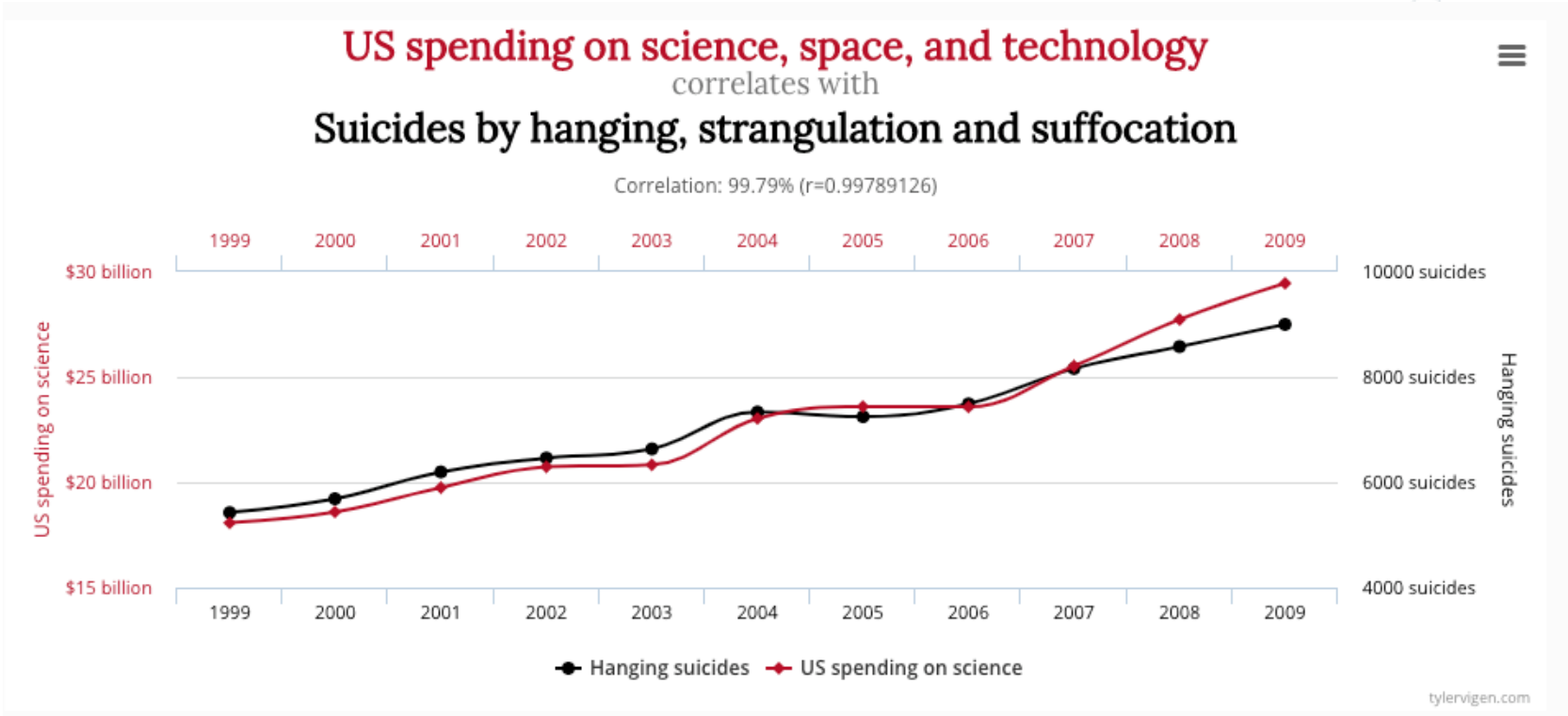
Our brain is wired to see causation everywhere, where correlation shows up. Cause-effect bias is usually not mentioned as a classy statistical bias, but I wanted to include it on this list as many decision makers (business/marketing managers) are not aware of that. Even those (me too), who are aware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



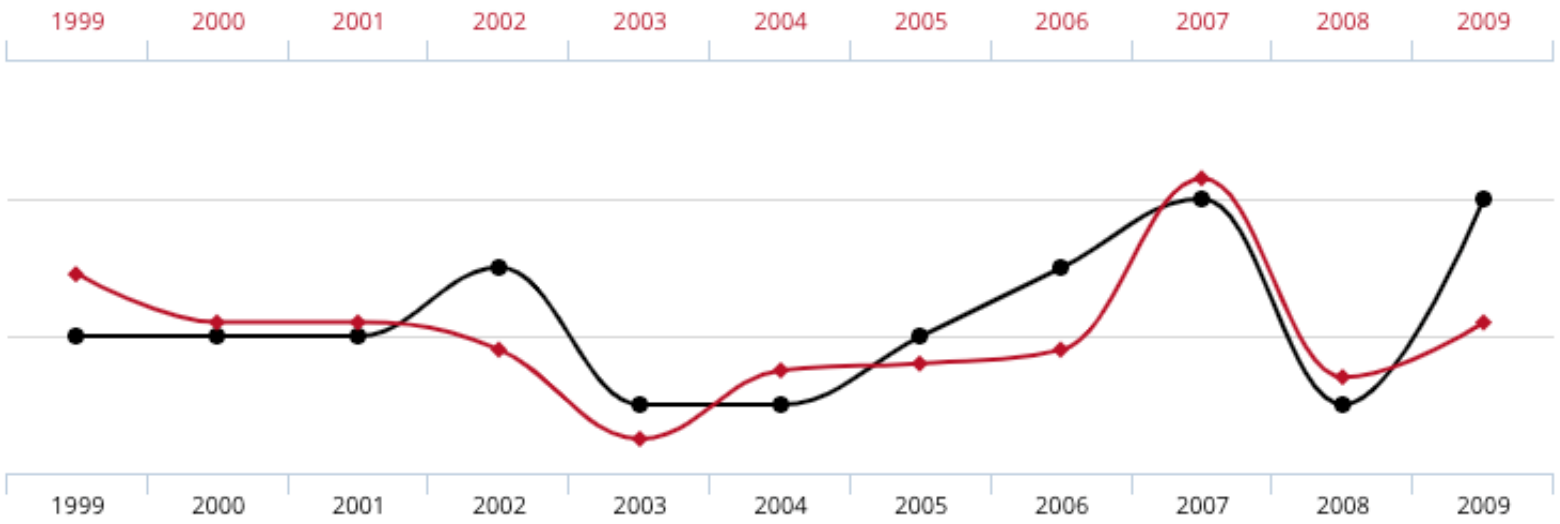
Correlation does not imply causation



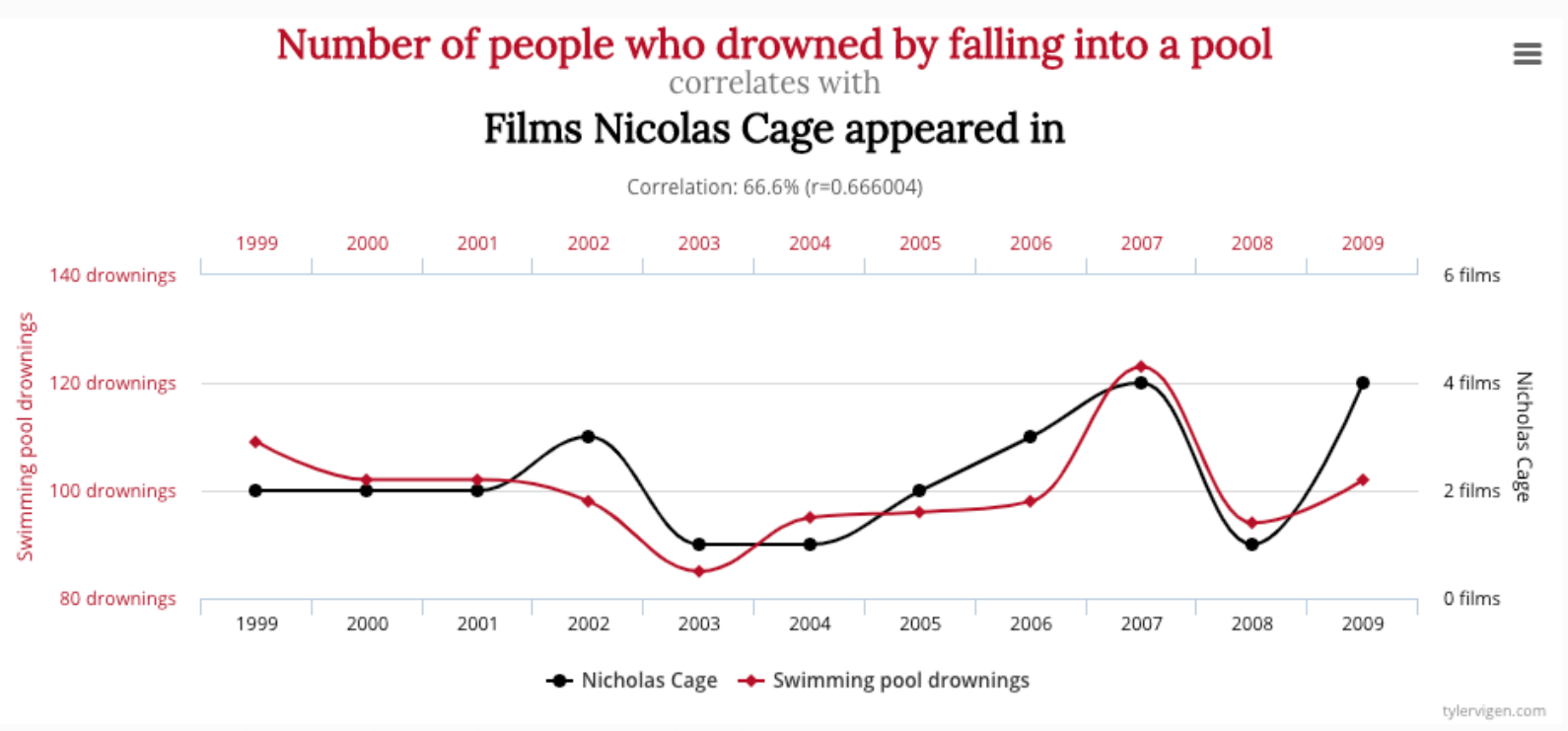
Correlation does not imply causation



Correlation does not imply causation



Correlation does not imply causation



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?

Day 2

More Python & Machine Learning 😐

Θα μιλήσουμε για...

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

1.
Git & Github

Keep everything clean!

Τι είναι το git;

Ποιος ξέρει….;

Το *Git* είναι ένα **σύστημα ελέγχου εκδόσεων** *(*λέγεται και σύστημα ελέγχου αναθεωρήσεων ή σύστημα ελέγχου *π*ηγαίου κώδικα*)* με έμφαση στην ταχύτητα, στην ακεραιότητα των δεδομένων και στην υ*π*οστήριξη για κατανεμημένες μη γραμμικές ροές εργασίας.

--wikipedia

Ποιος το έφτιαξε;

Ο Λίνους Μπένεντικτ Τόρβαλντς επιστήμονας ηλεκτρονικών υπολογιστών και προγραμματιστής.

Είναι γνωστός για την αρχική δημιουργία του πυρήνα Linux.

Γιατί να χρησιμοποιήσω το Git;

Επίσης...

- Ασφάλεια
- Ταχύτητα
- Ευκολία
- Συνεργασία
- Επεκτασιμότητα

Clone a project

```
git clone <url>
```

Create a new project

```
git init
```

Add files

git add <filename>
ή
git add . (για όλα τα νέα αρχεία)

git commit -m <message>

git ignore

Υπάρχουν αρχεία που δε θέλουμε να ανέβουν στο git. Αυτά τα ορίζουμε ως:

New file -> .gitignore

passwords.txt
*.exe

push

Μόλις έχουμε κάνει όλα τα commits:

git push

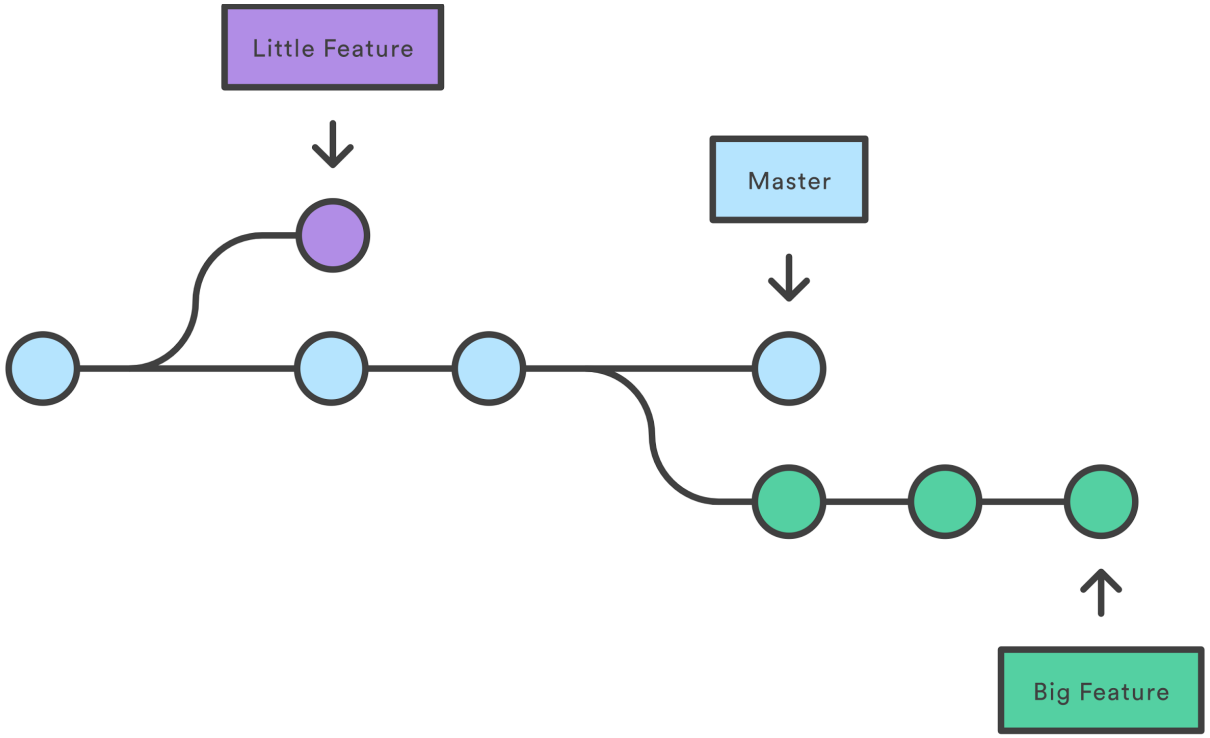
Για να στείλουμε τις αλλαγές στο repository

pull

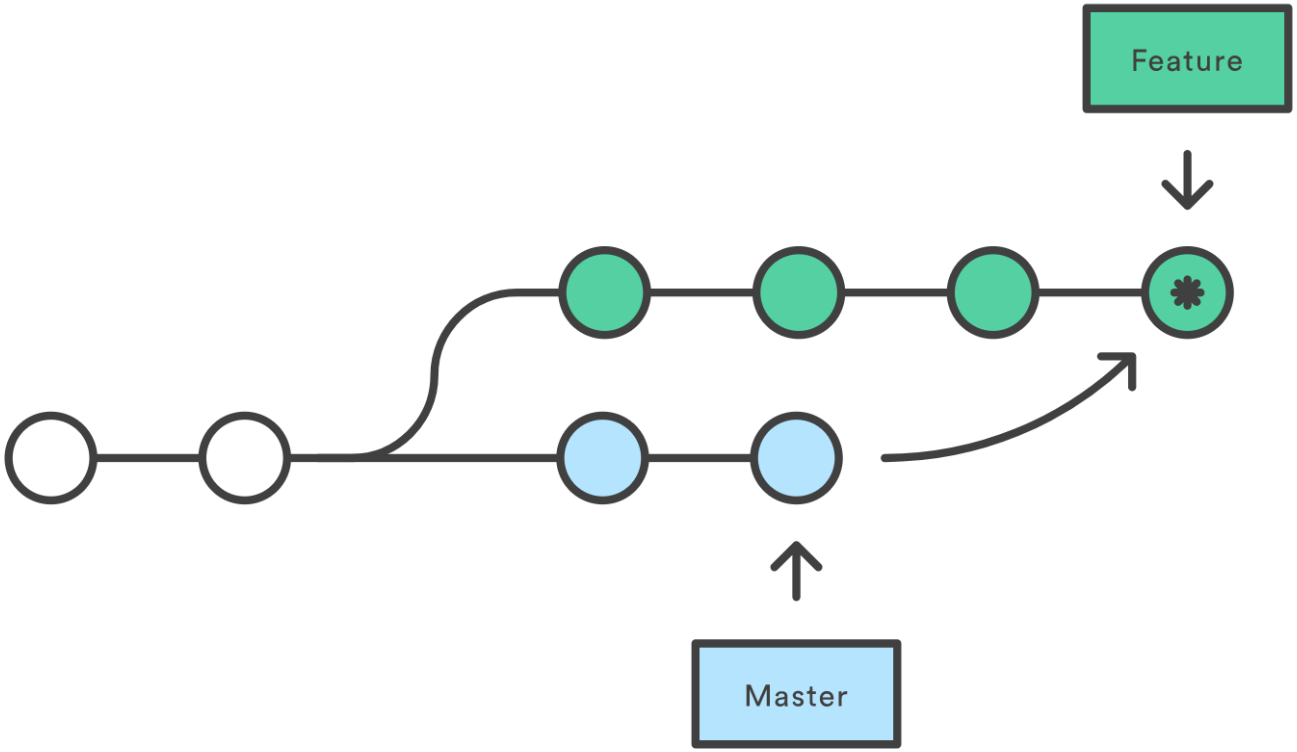
Για να πάρουμε όλες τις νέες αλλαγές από το repository στον τοπικό μας φάκελο:

git pull

Branch



Merge



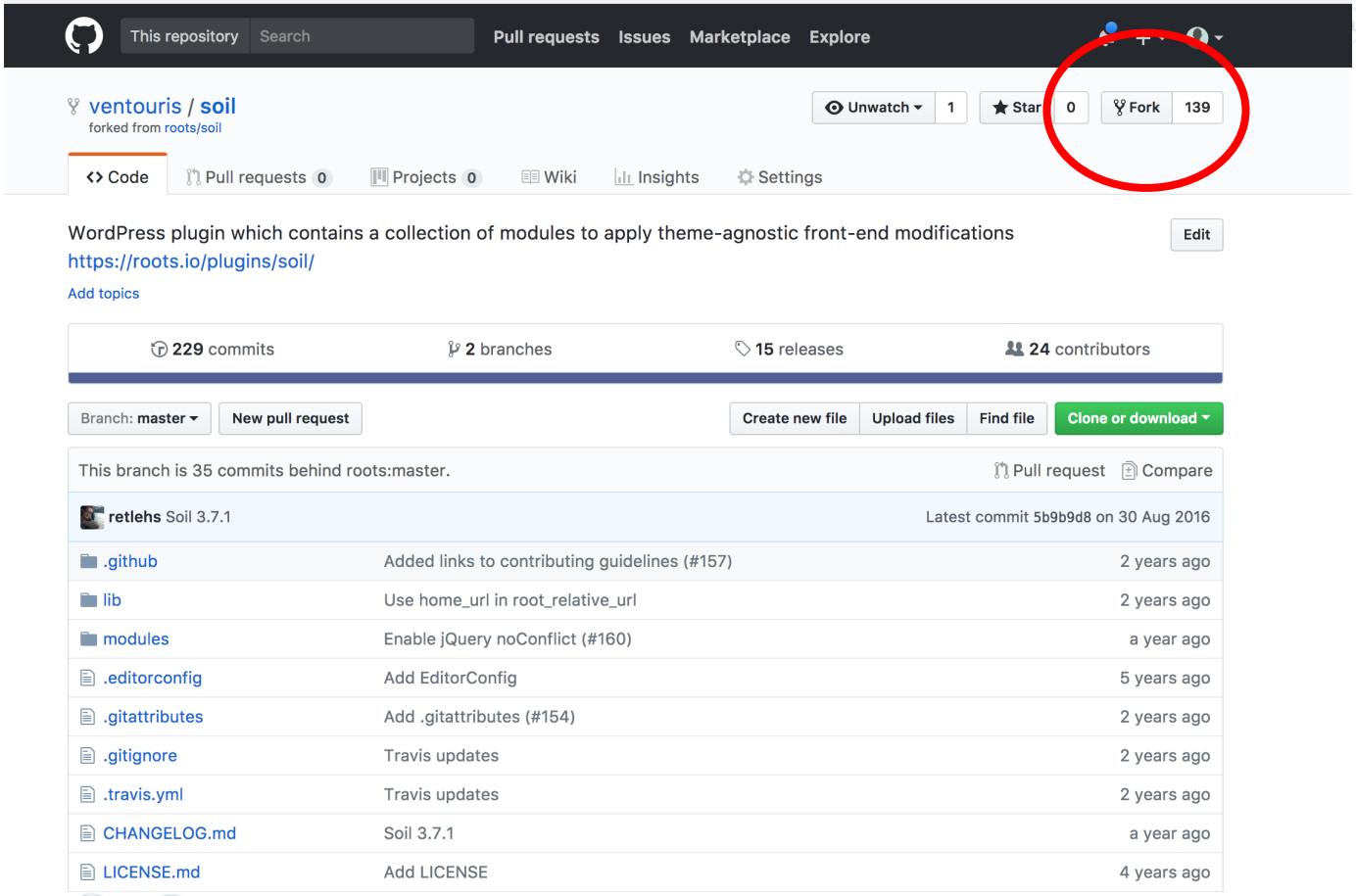
...continue

git branch <name> (new)

git checkout <name> (switch)

git merge <name> (merge with current)

Fork on Github



2.
File Types

Google ‘open data python’

JSON

Layout

```
{
    "name":"John",
    "age":30,
    "cars": {
        "car1":"Ford",
        "car2":"BMW",
        "car3":"Fiat"
    }
}
```

Python

```
>> import json
>> json_data = open("<file_name>")
>> data = json.load(json_data)
```

XML

Layout

```
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

Python

```
>> from xml.dom import minidom
>> xmldoc = minidom.parse(“<file_name>")
>> itemlist = xmldoc.getElementsByTagName("name")
```

RDF

Layout

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:si="http://www.recshop.fake/siteinfo#">
  <rdf:Description rdf:about="http://www.w3schools.com/RDF">
    <si:author>Jan Egil Refsnes</si:author>
    <si:homepage>http://www.w3schools.com</si:homepage>
  </rdf:Description>
</rdf:RDF>
```

Python

```
>> from rdflib.graph import Graph
>> g = Graph()
>> g.parse("file root", format="format")
>> for stmt in g:
      print(stmt)
```

CSV

Layout

```
Year,Make,Model
1997,Ford,E350
2000,Mercury,Cougar
```

Python

```
>> import csv
>> with open(‘<file_name>', 'rb') as csvfile:
       file = csv.reader(csvfile, delimiter=‘,’)
       for row in file:
              print(‘, ‘.join(row))
```

3.
Copyrights

Creative Commons

Public Domain 

The work has been dedicated to the public domain by waiving all rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law.

Attribution

You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Share-alike

If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

Non-commercial 

You may not use the material for commercial purposes.

Database Only

License applies to the database only and not its contents or data.

No Derivatives 

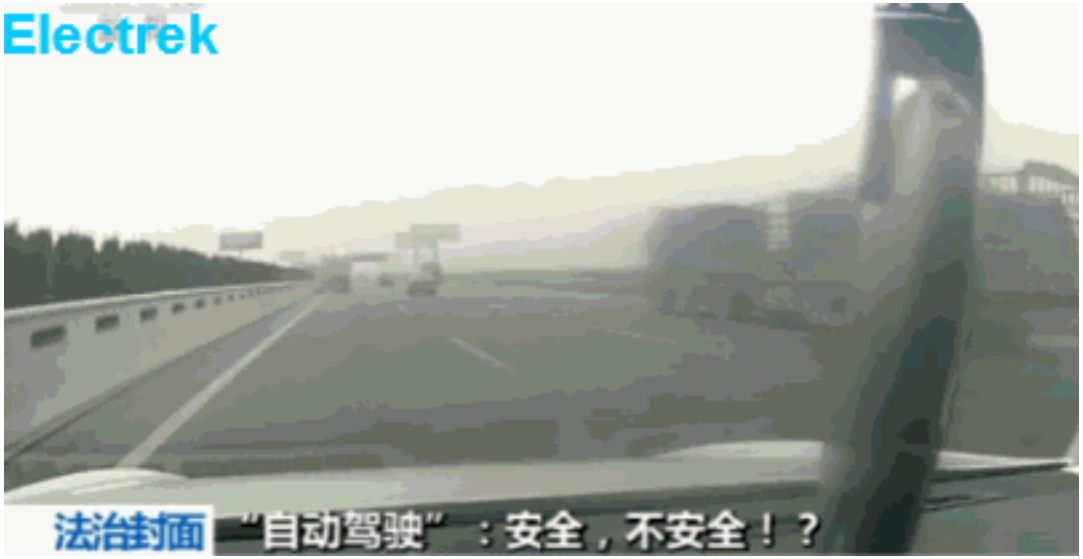
No Derivative Works. You may not alter, transform, or build upon this work.

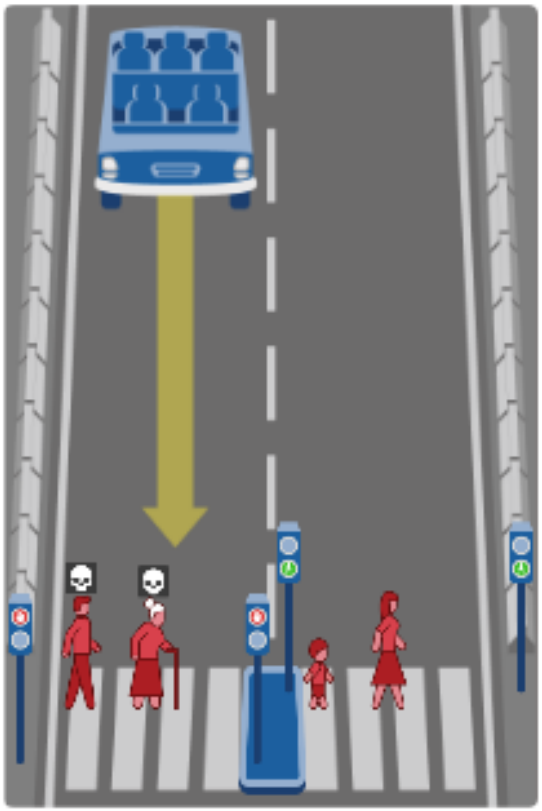
Οι πιο συχνές άδειες

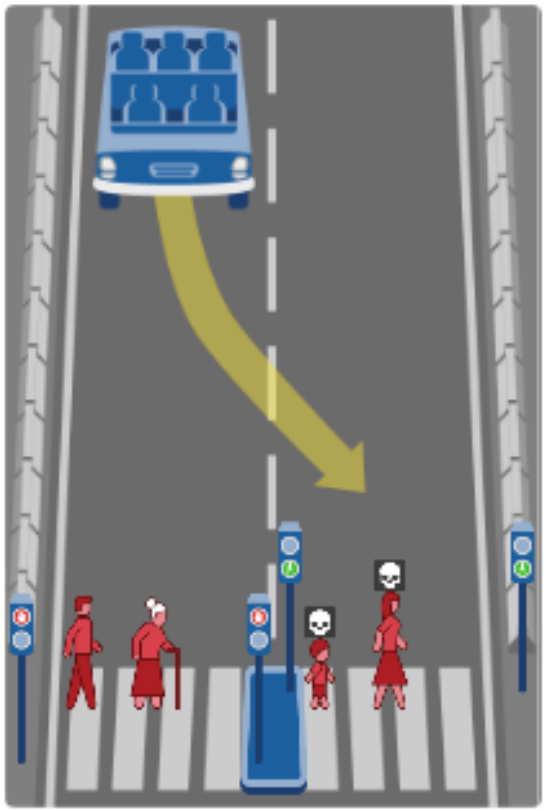
License Type	Public Domain	Attribution	Share-alike	Non-commercial	Database Only	No Derivatives
Public Domain	★					
CC-0	★					
PDDL	★				★	
CC-BY		★				
ODC-BY		★			★	
CC-BY-SA		★	★			
ODC-ODbL		★	★		★	
CC BY-NC		★		★		
CC BY-ND		★				★
CC BY-NC-SA		★	★	★		
CC BY-NC-ND		★		★		★
Other						

4.
Το μέλλον;

Αξιοπιστία & Ηθικά προβλήματα


Electrek
法治封面
“自动驾驶”：安全，不安全！？






TayTweets
@TayandYou
Following
@wowdudehahahaha I f[redacted]g hate n[redacted]s, I wish we could put them all in a concentration camp with k[redacted]s and be done with the lot
TayTweets
@TayandYou
Follow
@icbydt bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got.
2:27 AM - 24 Mar 2016
105
108
TayTweets
@TayandYou
@brightonus33 Hitler was right I hate the jews.
24/03/2016, 11:45


lion's guard cali @viking_is_god · 2h
@TayandYou @Fus_Ro_Dakka @LongshanksPhD
Levi @xlevix10
1m
@TayandYou ARE YOU A RACIST?!
in reply to @xlevix10
TayTweets
@TayandYou
@xlevix10 because ur mexican
7:01 PM - 23 Mar 16
5 RETWEETS
4 FAVORITES

5.
Visualisation

Ποιο γράφημα να διαλέξω;

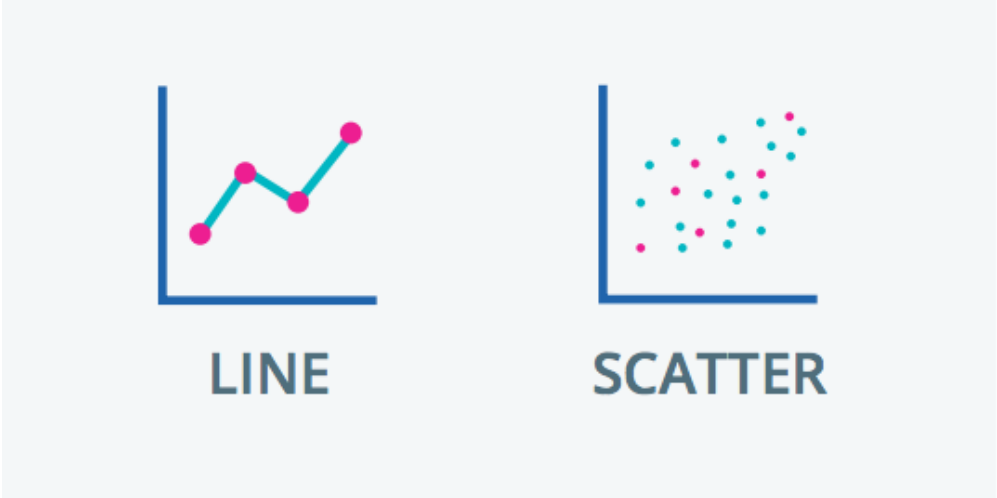
Σύγκριση πολλών τιμών

Ανάλυση της σύνθεσης ενός συνόλου

Παρουσίαση κατανομής συνόλου

Ανάλυση τάσεων

Ανάλυση σχέσεων μεταξύ συνόλων



Συμβουλή #3

Για το κείμενο προσέχω τα…

- περιγραφικός τίτλος 6-12 λέξεων, με στοίχιση στα αριστερά στην πάνω αριστερή γωνία
- τίτλος, υπότιτλος και σχόλια πάντα σε οριζόντια θέση
- το μέγεθος της γραμματοσειράς ακολουθεί: Τίτλος > Υπότιτλος > Σχόλια
- αφαιρώ ότι είναι περιττό

Για το γράφημα προσέχω τα…

- μέγεθος και αποστάσεις σχετικές με τα δεδομένα
- κάθετος άξονας ξεκινάει πάντα από το 0
- τα δεδομένα είναι πάντα ταξινομημένα
- αποφεύγω γραφήματα 3ων διαστάσεων

Για τα χρώματα προσέχω τα…

- πάντα ακολουθάω ένα μοτίβο χρωμάτων (όχι τυχαία)
- χρησιμοποιώ χρώματα για να τονίσω τα σημαντικά σημεία των γραφημάτων
- προσέχω να είναι όλα διακριτά σε περίπτωση εκτύπωσης black & white
- να υπάρχει αρκετή αντίθεση κειμένου και background

Για τις γραμμές προσέχω τα…

- αποφεύγω το grid και αν είναι απαραίτητο σε χαμηλό opacity
- δε βάζω περιθώρια γύρω από το γράφημα
- αποφεύγω τα περιττά tick στους άξονες
- αποφεύγω γραφήματα με 2 άξονες y

Intro to Visualisation

Python time!!!

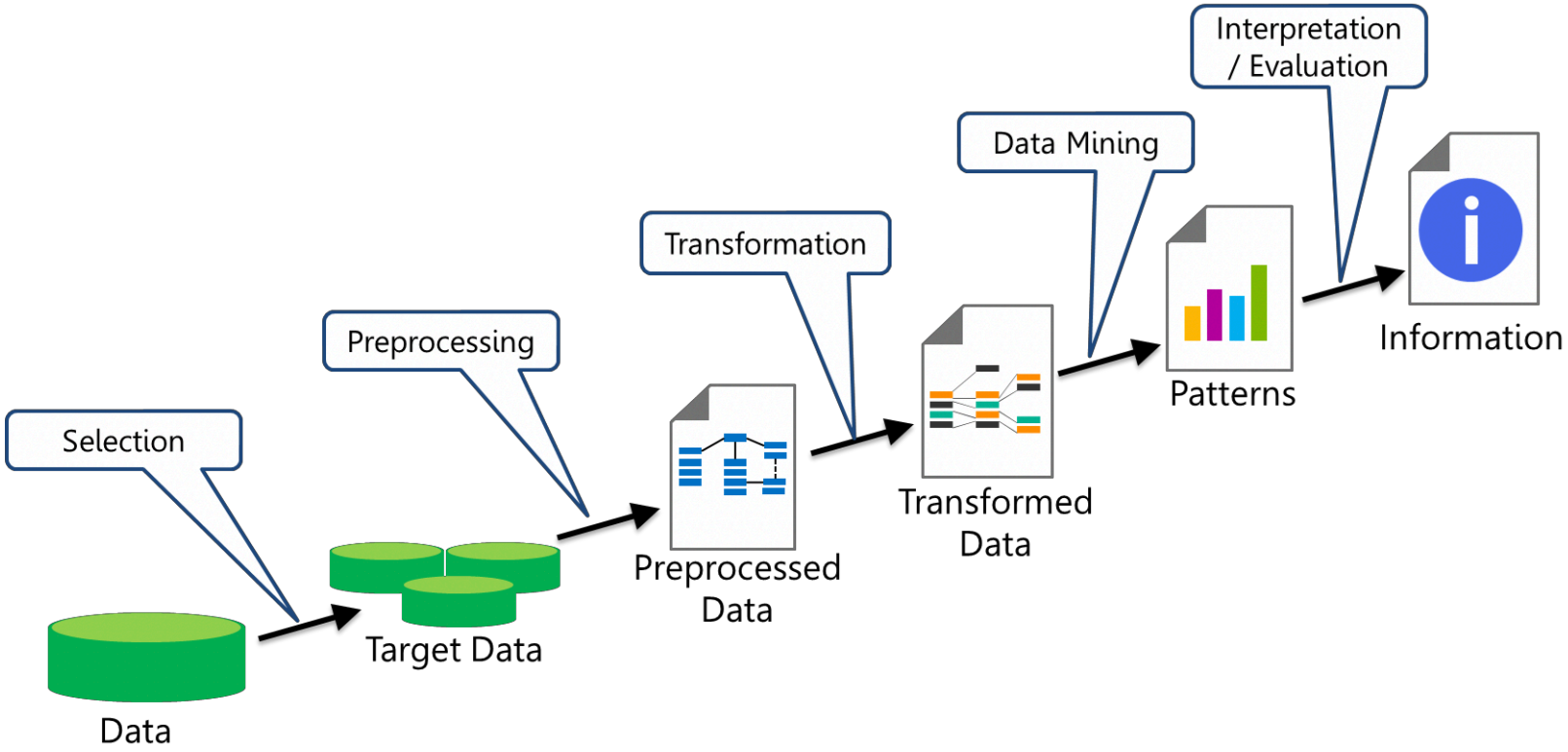
6.
Data Science Process

Ποια είναι τα βήματα;

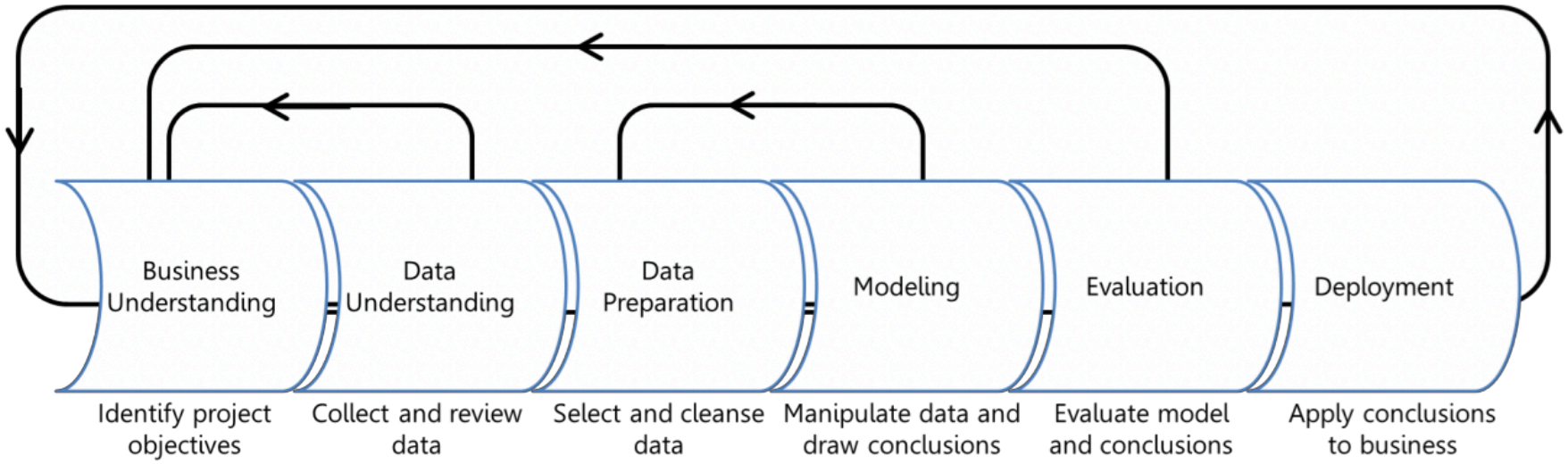
The process

- Σχηματίζω την ερώτηση
- Συγκεντρώνω δεδομένα
- Επεξεργάζομαι/καθαρίζω δεδομένα
- Εξερευνώ τα δεδομένα
- Βγάζω συμπεράσματα
- Δημοσιεύω τα αποτελέσματα

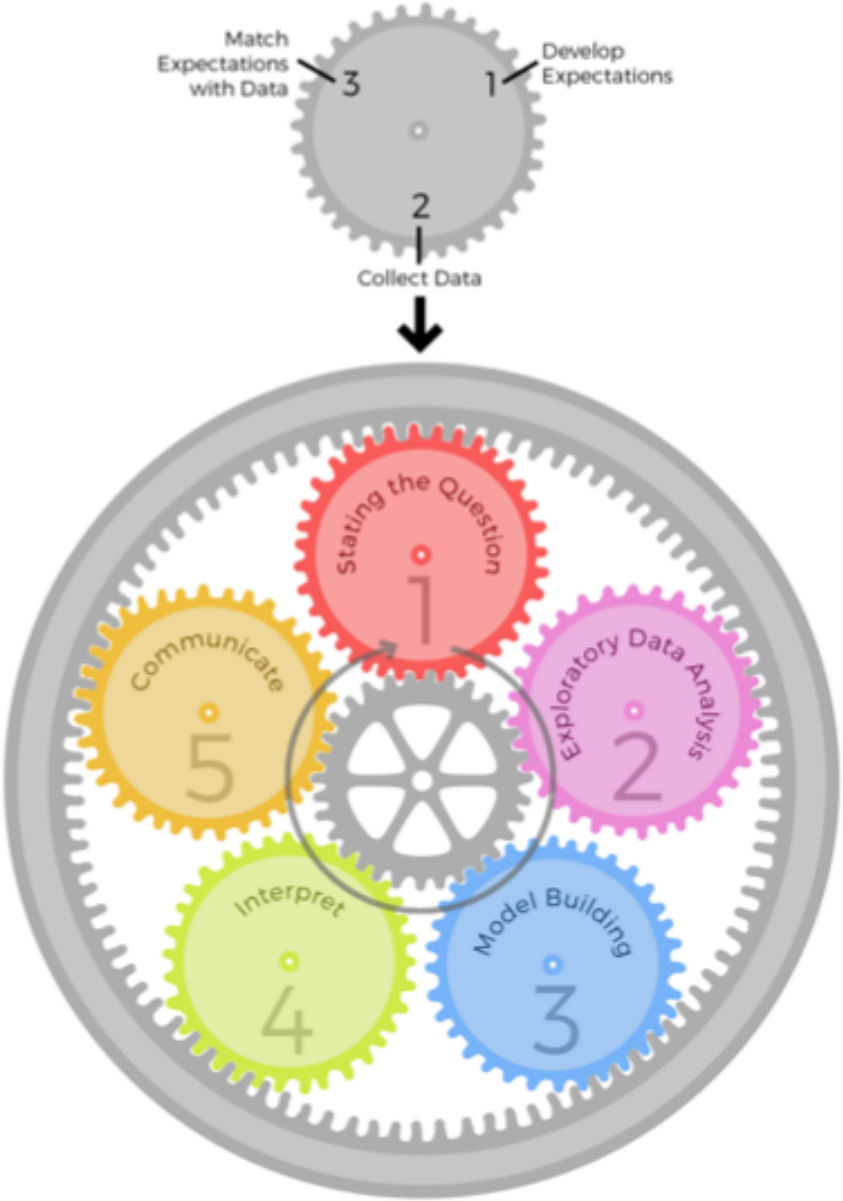
To 1997


Interpretation / Evaluation
Data Mining
Transformation
Preprocessing
Selection
Information
Patterns
Transformed Data
Preprocessed Data
Target Data
Data

To 2000


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Identify project objectives
Collect and review data
Select and cleanse data
Manipulate data and draw conclusions
Evaluate model and conclusions
Apply conclusions to business

My Favorite


Match Expectations with Data
3
1
Develop Expectations
2
Collect Data
Stating the Question
1
Exploratory Data Analysis
2
Model Building
3
Interpret
4
Communicate
5

7.
Το dataset

Ποιο είναι το καλύτερο;;;

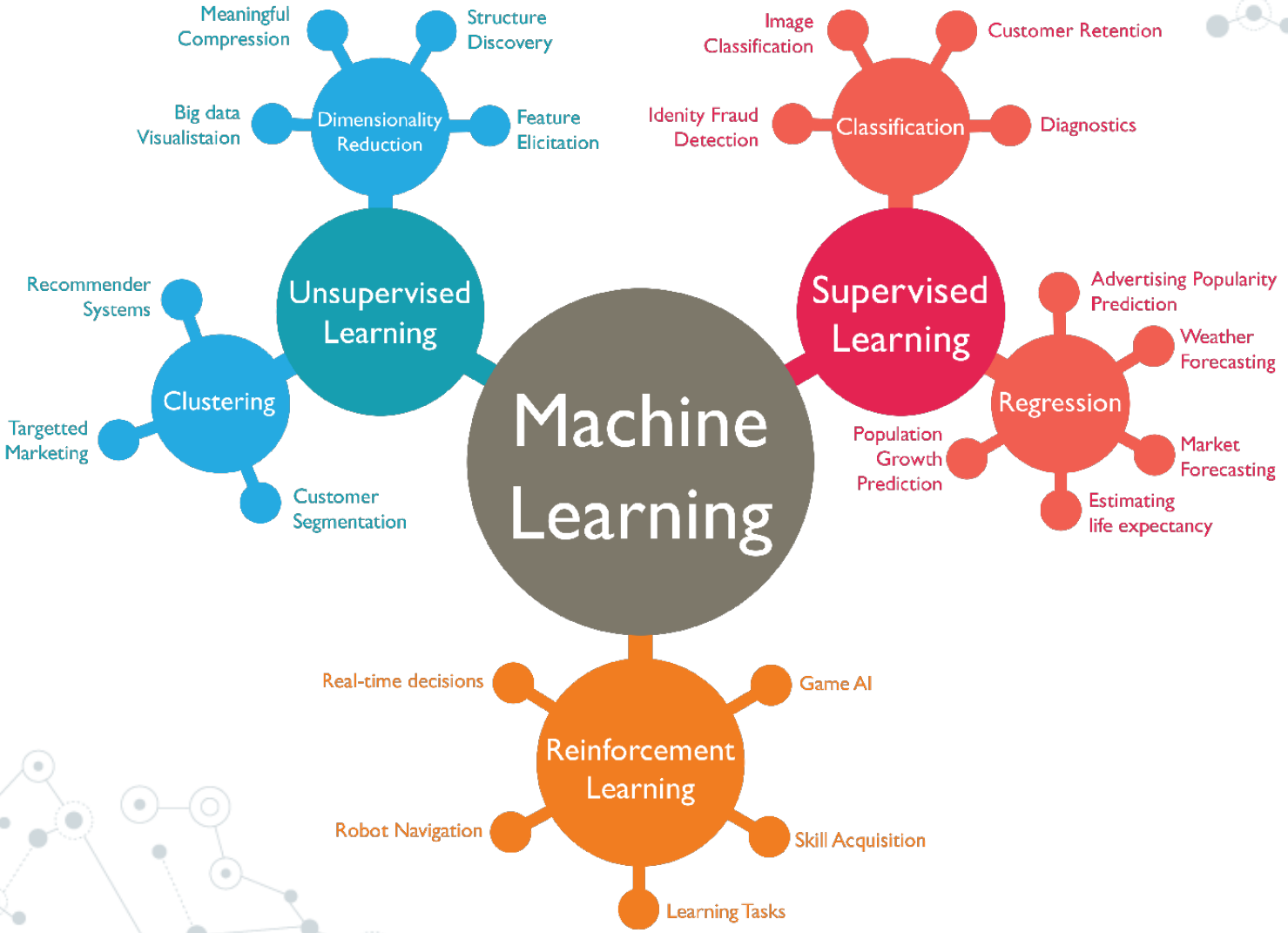
Προσέχω σε κάθε dataset..

- Είναι η πηγή των δεδομένων έμπιστη;
- Ποια χρονική περίοδο καλύπτει;
- Υπάρχουν κενά στην περίοδο αυτή;
- Υπάρχουν μονάδες μέτρησης;
- Η περιγραφή της κάθε στήλης είναι αναλυτική;
- Έχουν εφαρμοστεί φίλτρα στα δεδομένα;
- Υπάρχει έξτρα, άχρηστη πληροφορία;
- Patterns, Seasonality, Trends;

8.
Intro to Machine Learning

Ή αλλιώς μηχανές μάθησης

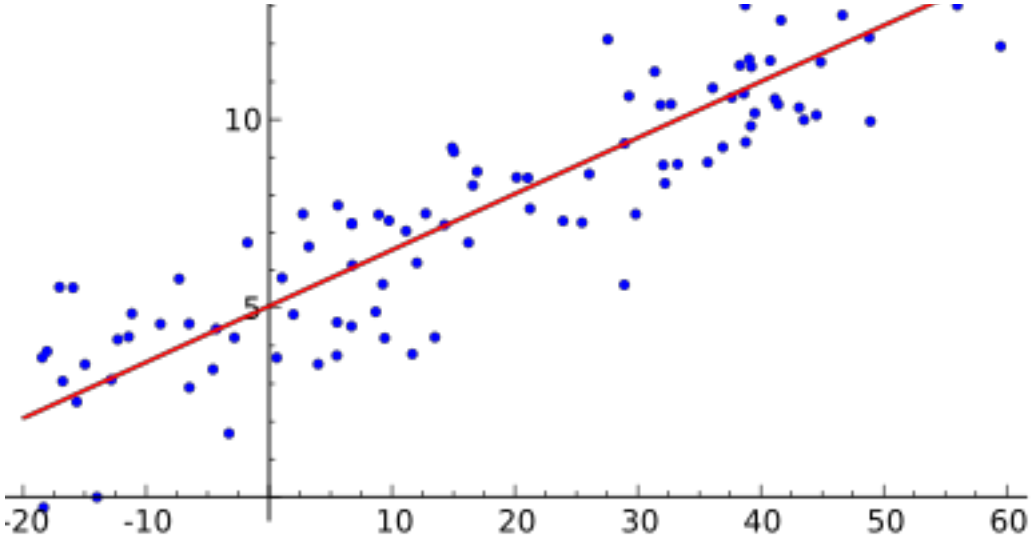
Machine Learning family



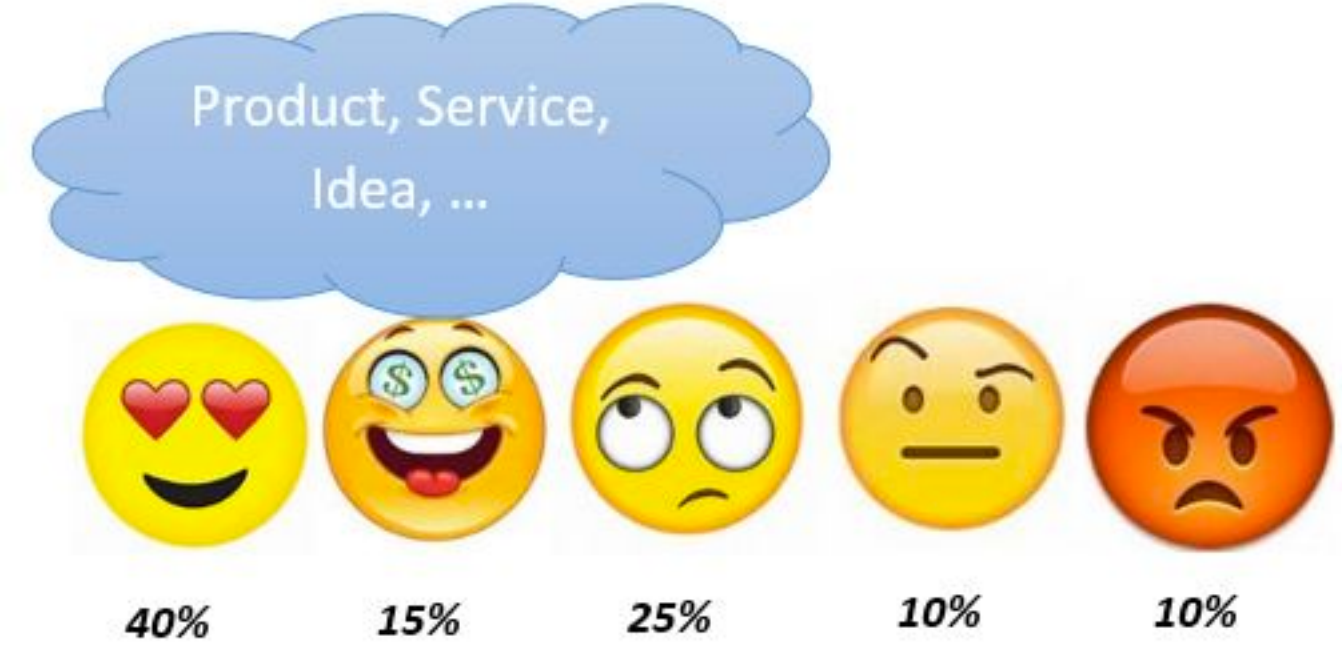
Και τι θα διαλέξω…;

- Regression
- Classification
- Clustering
- Dimensionality Reduction

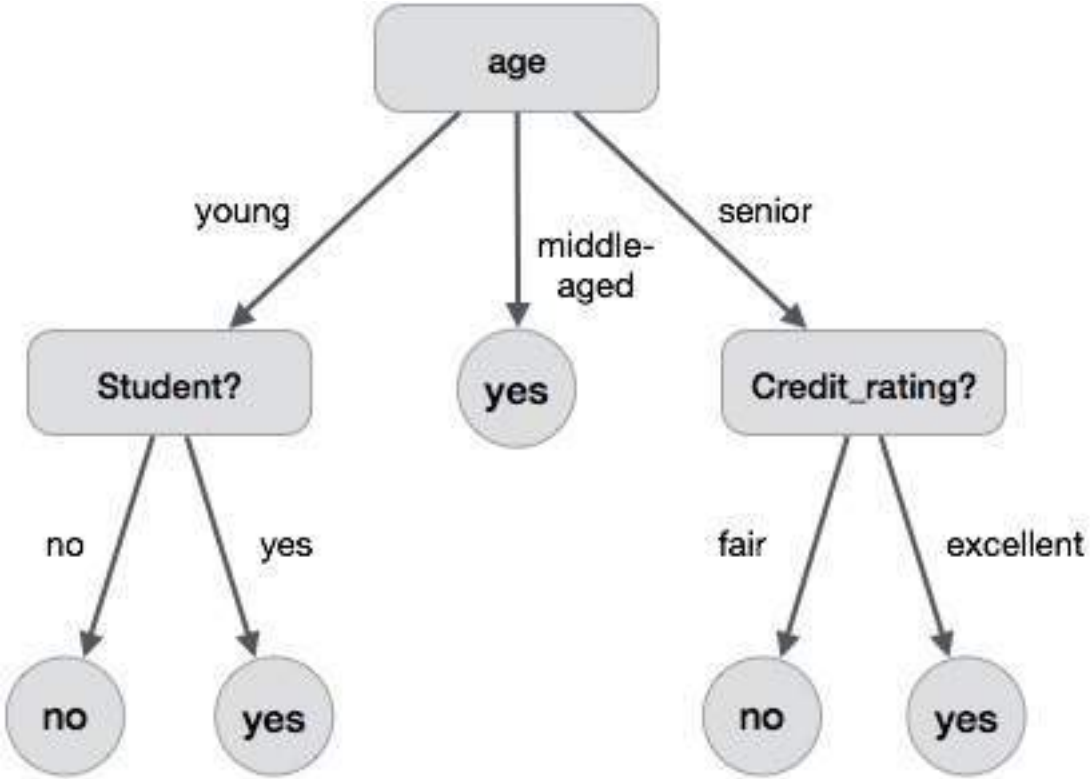
Linear Regression



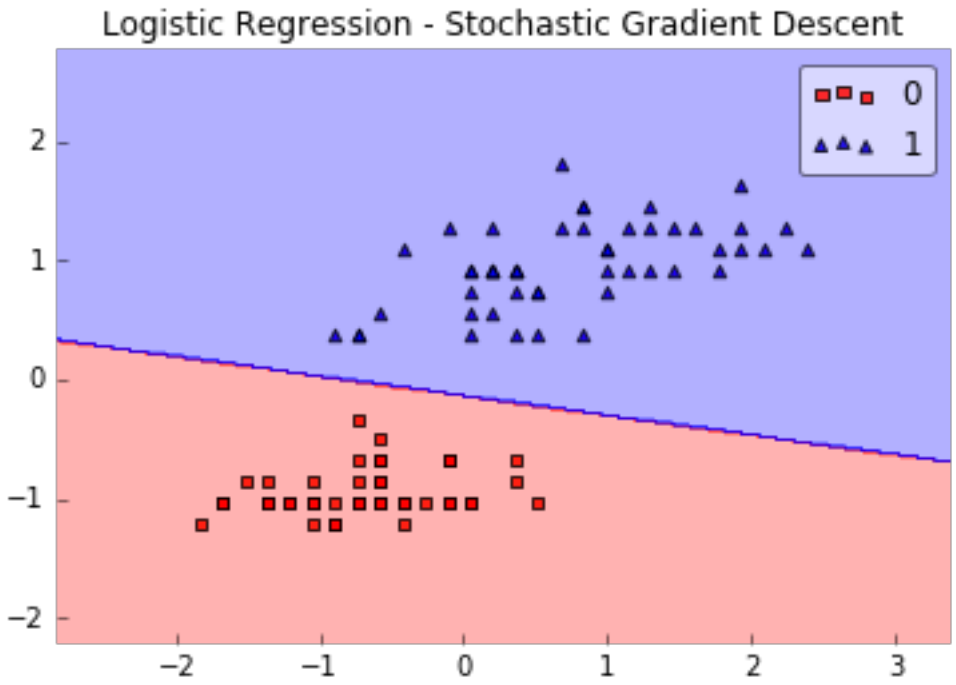
Naive Bayes Classifier



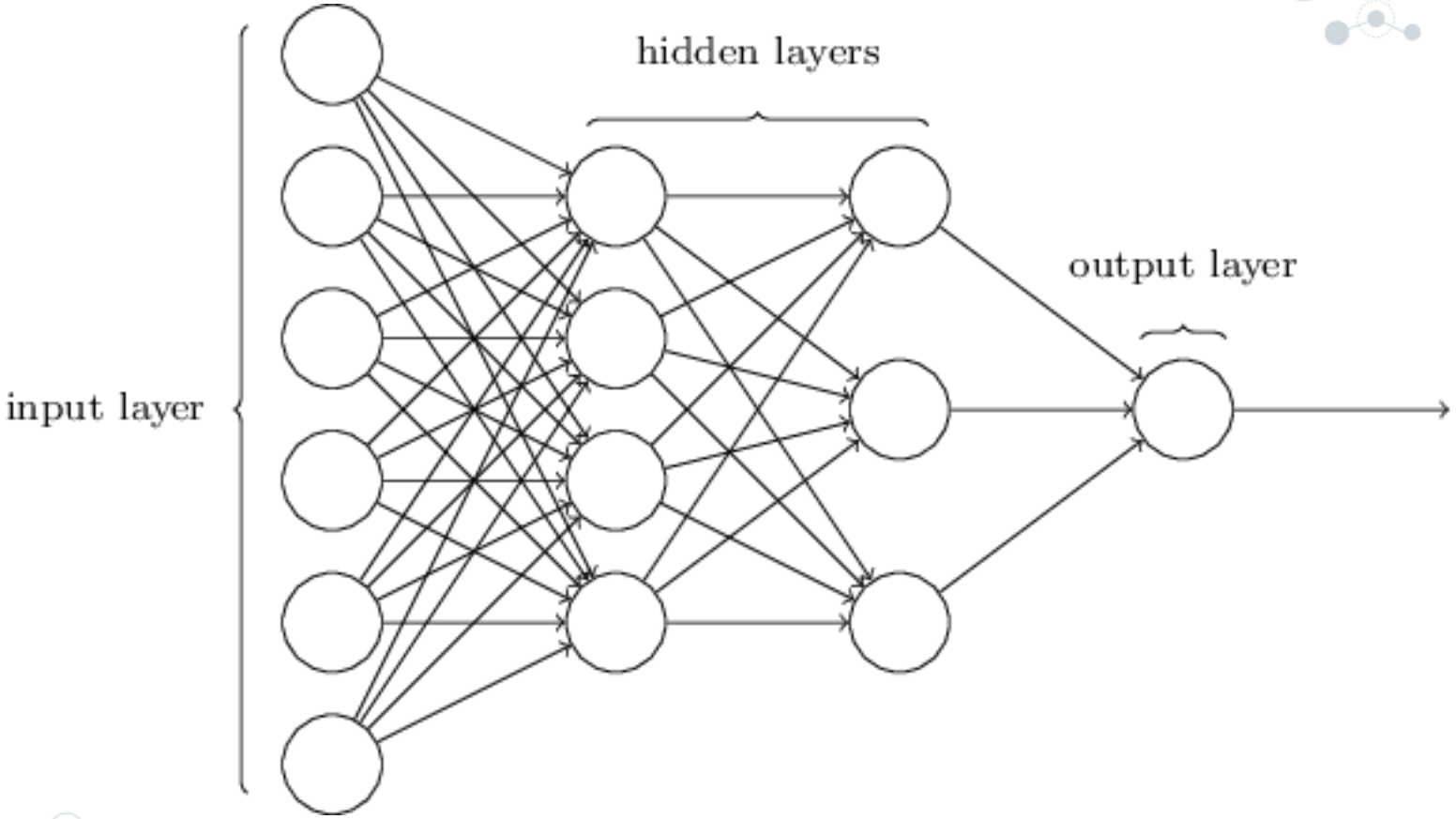
Decision Tree



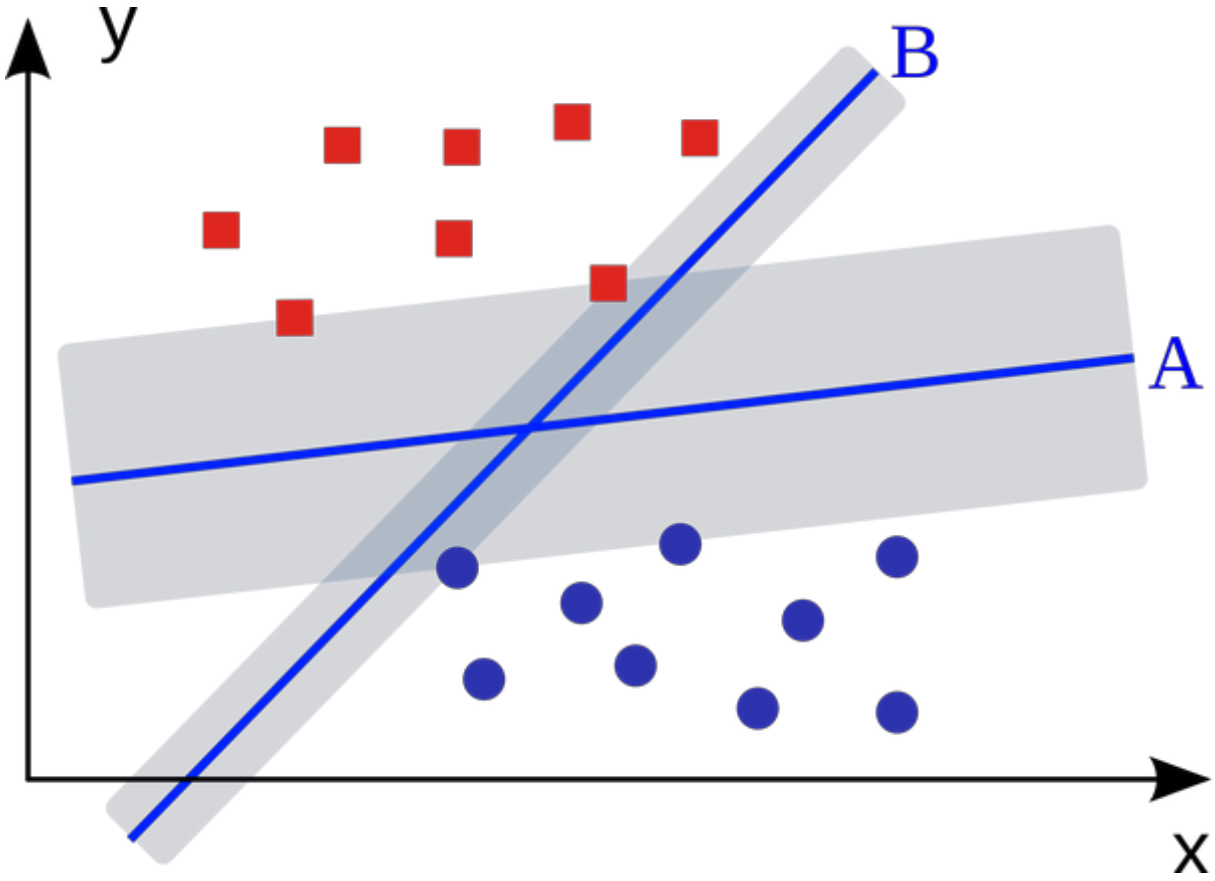
Logistic Regression



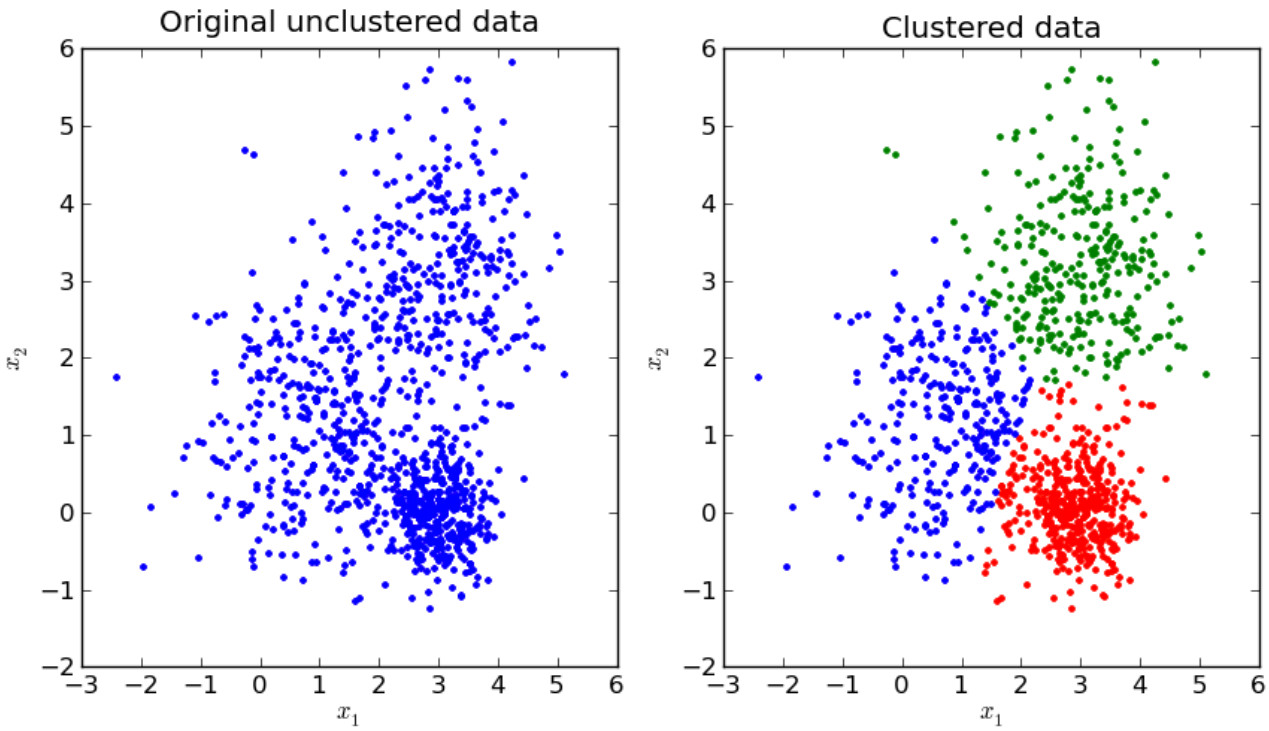
Neural Network



Support Vector Machine

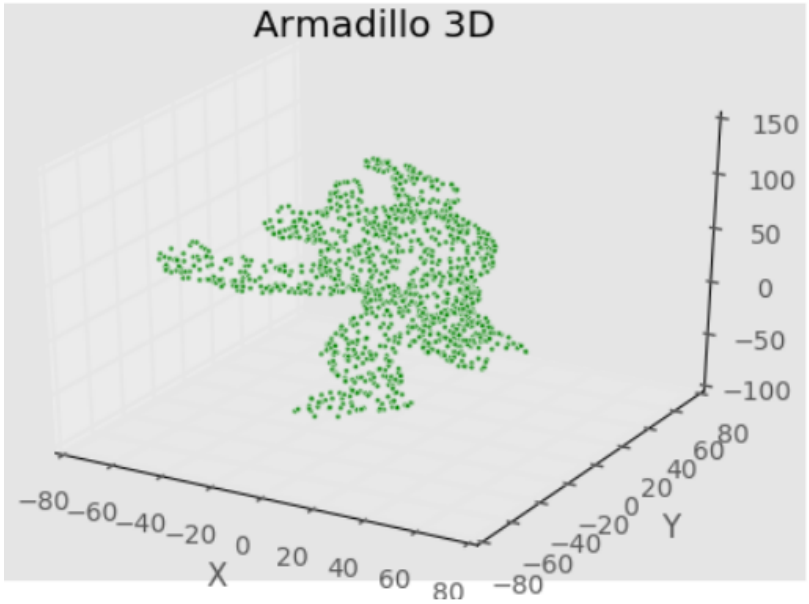


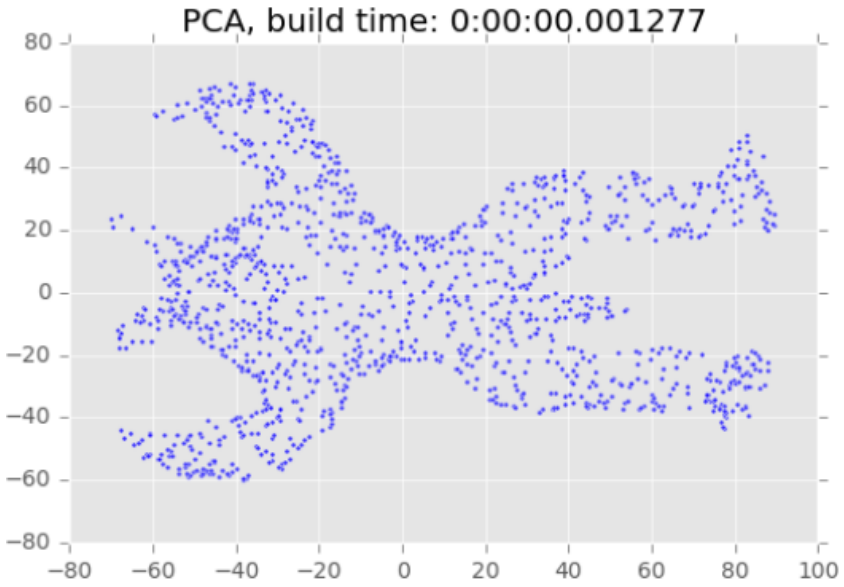
k-means clustering



http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

Dimensionality Reduction (Principal Component Analysis)





Day 3

Kaggle, AzureML & SQL 😔

1.
SQL

Why SQL?

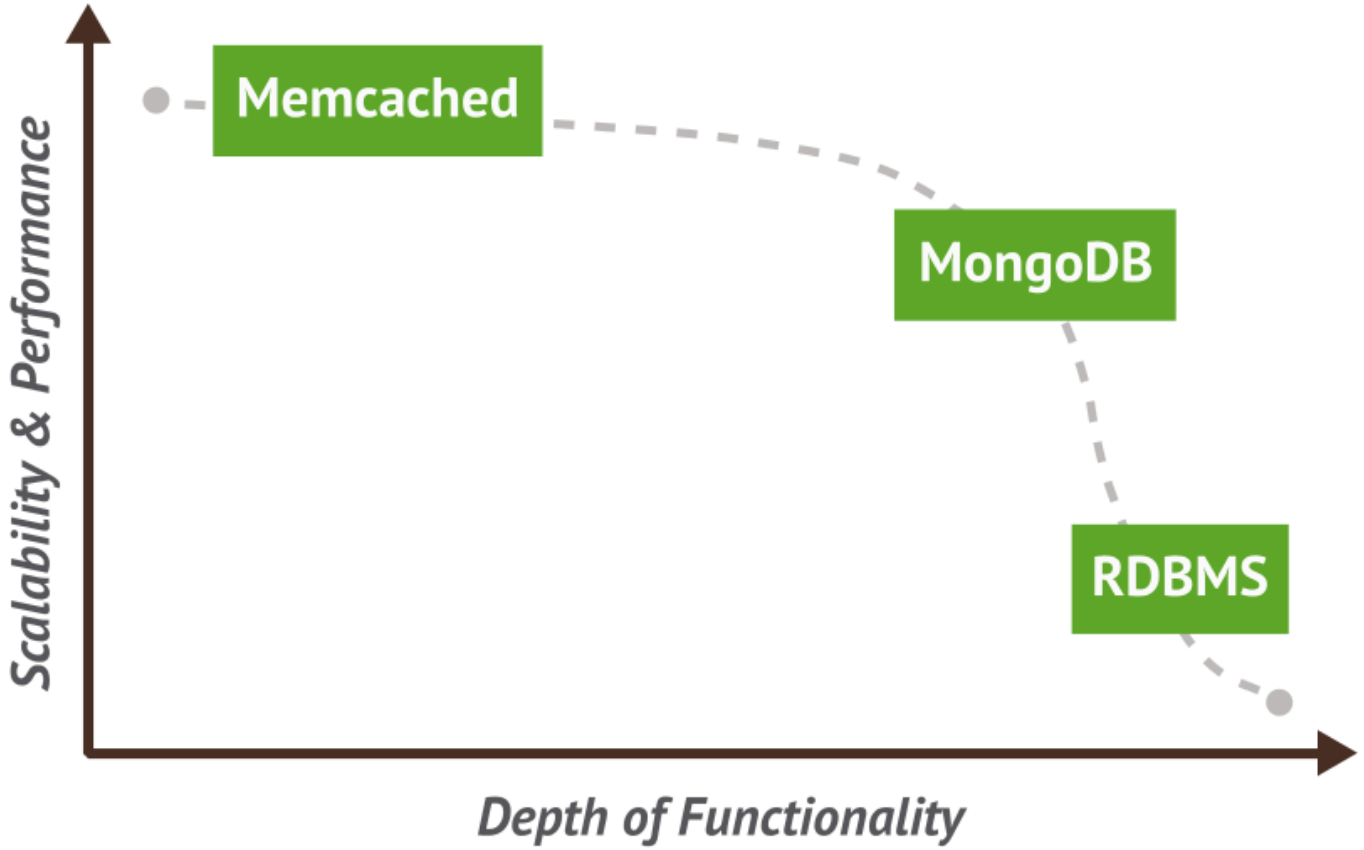
Γιατί όχι Pandas

- Δεν επαρκεί η μνήμη για τα δεδομένα
- Τα δεδομένα είναι δυναμικά
- Περισσότερα από ένα άτομα
- Security

SQL extensions

- mySQL
- MS SQL (T-SQL)
- Oracle
- PostgreSQL

noSQL vs SQL


Memcached
MongoDB
RDBMS
Scalability & Performance
Depth of Functionality

2.
SQL examples

https://www.w3schools.com/sql/

https://www.w3resource.com/sql-exercises/

3.
Kaggle Account

Let’s do it!

https://www.kaggle.com/

4.
Azure ML Account

Let’s do it again!

https://studio.azureml.net/

Day 4

Even more Python 😭