

11 Δεκεμβρίου 2017

Data Science Workshop



Golden Gate Pro

Λίγα λόγια

Hello!

Τάσος Βεντούρης

Data Scientist and
Game Designer @
Hattrick Ltd



You can find me at:



@tasosventouris

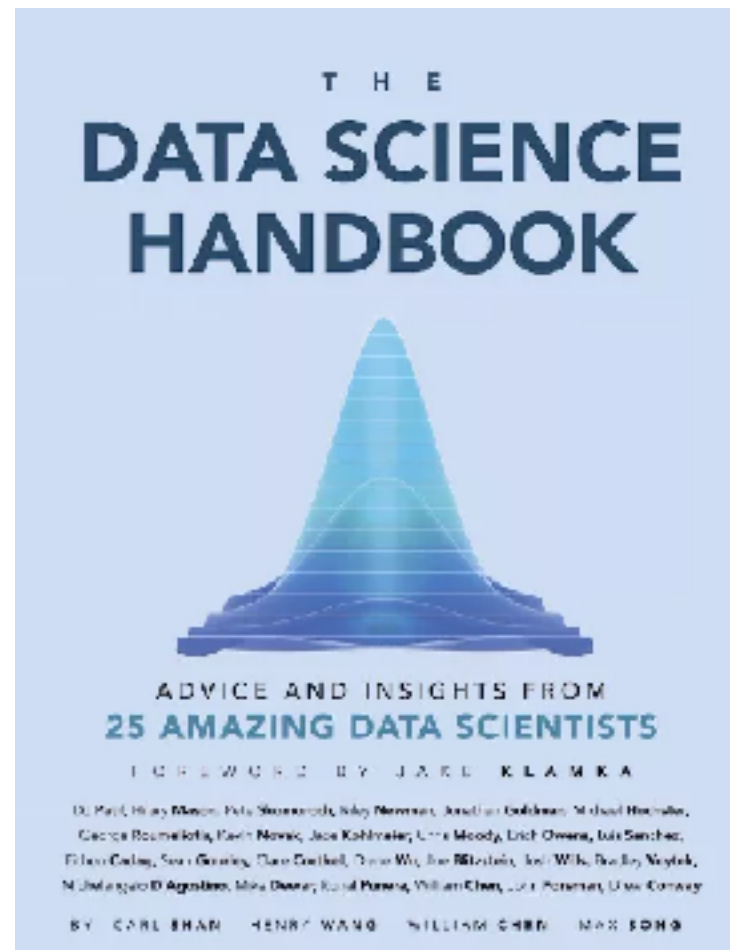


Tasos Ventouris

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



A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

Day 1

Εισαγωγή 😊

A decorative network diagram in the bottom-right corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.



Θα μιλήσουμε για...

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

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

1. **Data Science**

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

A decorative network diagram at the top of the slide, featuring a series of interconnected nodes and lines. A central node is highlighted with a dashed circle and a solid circle, containing a large blue quotation mark.

“

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

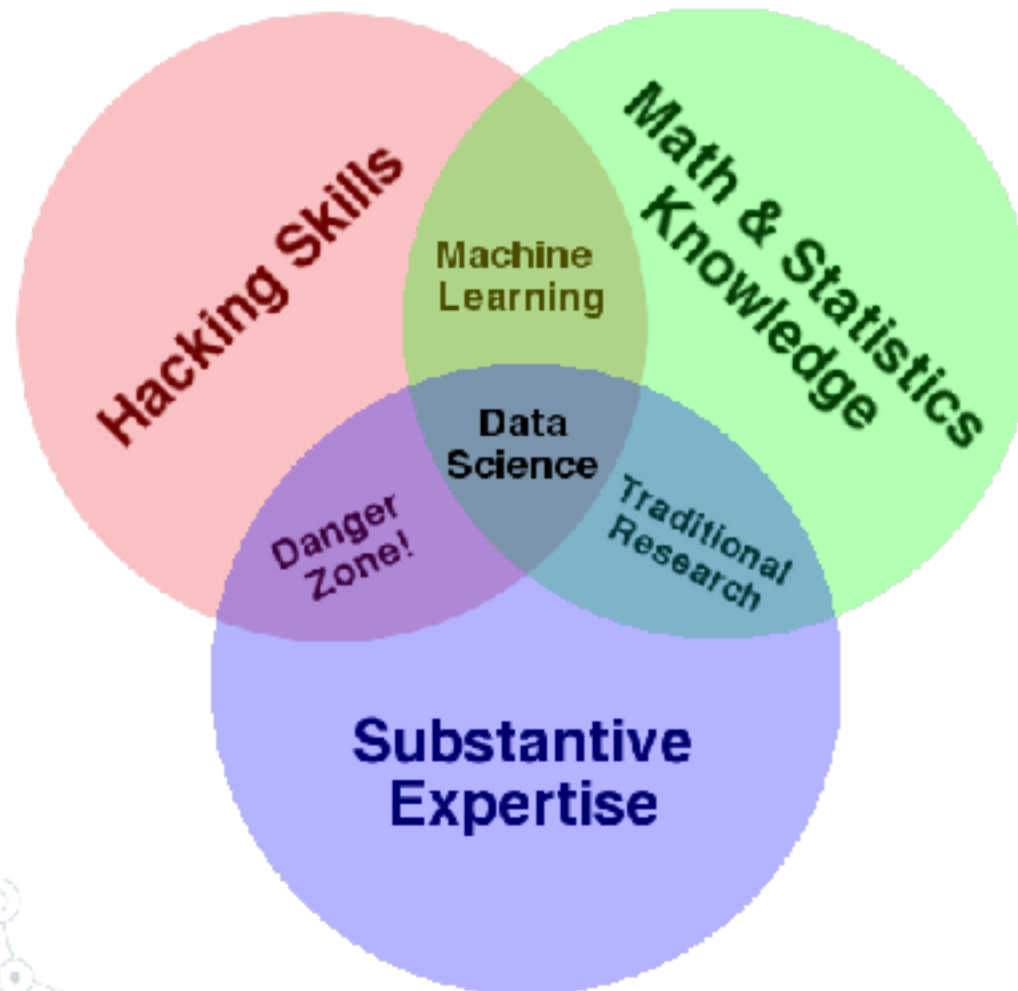


“

The creation of data products

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

My favourite



A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles inside, suggesting a hierarchical or multi-layered structure. The lines are thin and gray, connecting the nodes in a non-linear fashion.

Data Scientist

Ποιος λοιπόν μπορεί να έχει τον
τίτλο του Data Scientist;

A decorative network diagram in the bottom-right corner, similar to the one in the top-left. It shows a cluster of nodes connected by lines, with some nodes being more prominent than others. The overall style is clean and modern, using a light gray color scheme.

“



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 decorative network diagram at the top of the slide, featuring a series of interconnected nodes and lines. A central node is highlighted with a dashed circle and a solid circle, containing a large blue quotation mark.

“

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



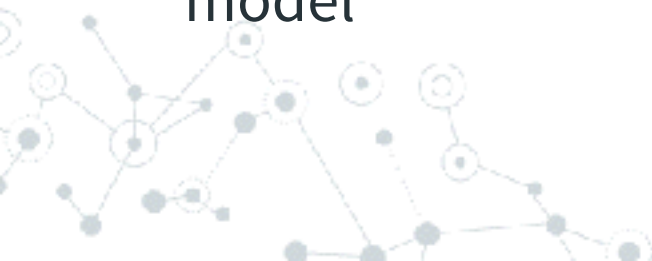
Πικρή Αλήθεια #1



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

A decorative network diagram at the top of the slide, featuring a series of interconnected nodes and lines. A central node is highlighted with a dashed circle and a solid circle, containing a large blue quotation mark.

“

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

--Media

A decorative graphic at the top of the slide featuring a network of interconnected nodes and lines. A central node is highlighted with a solid blue circle and a dashed blue circle, with a large blue quotation mark inside it.

“

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

--Biotechnology

A decorative network diagram at the top of the slide, featuring a series of interconnected nodes and lines. A central node is highlighted with a dashed circle and a solid circle, containing a large blue quotation mark.

“

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

--Manufacturing

A decorative network diagram at the top of the slide, featuring a complex web of interconnected nodes and lines. A central node is highlighted with a dashed circle and a solid circle, containing a large blue quotation mark.

“

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

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles inside, suggesting a hierarchical or multi-layered structure. The lines are thin and gray, connecting the nodes in a non-linear fashion.

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 \neq Big Data

Apollo XI, 1969

64Kb

SkyDive Stratos, 2012

Δεκάδες Gigabytes



3.

Το μονοπάτι ενός Data Scientist

Yeah! I had a skill up...

A decorative background featuring a network diagram. It consists of numerous nodes, represented by small circles, some of which are solid blue and others are hollow with blue outlines. These nodes are interconnected by thin, light gray lines, forming a complex web-like structure that is more dense on the left and right sides of the page.

Πικρή Αλήθεια #3

Το μονοπάτι του Data Scientist



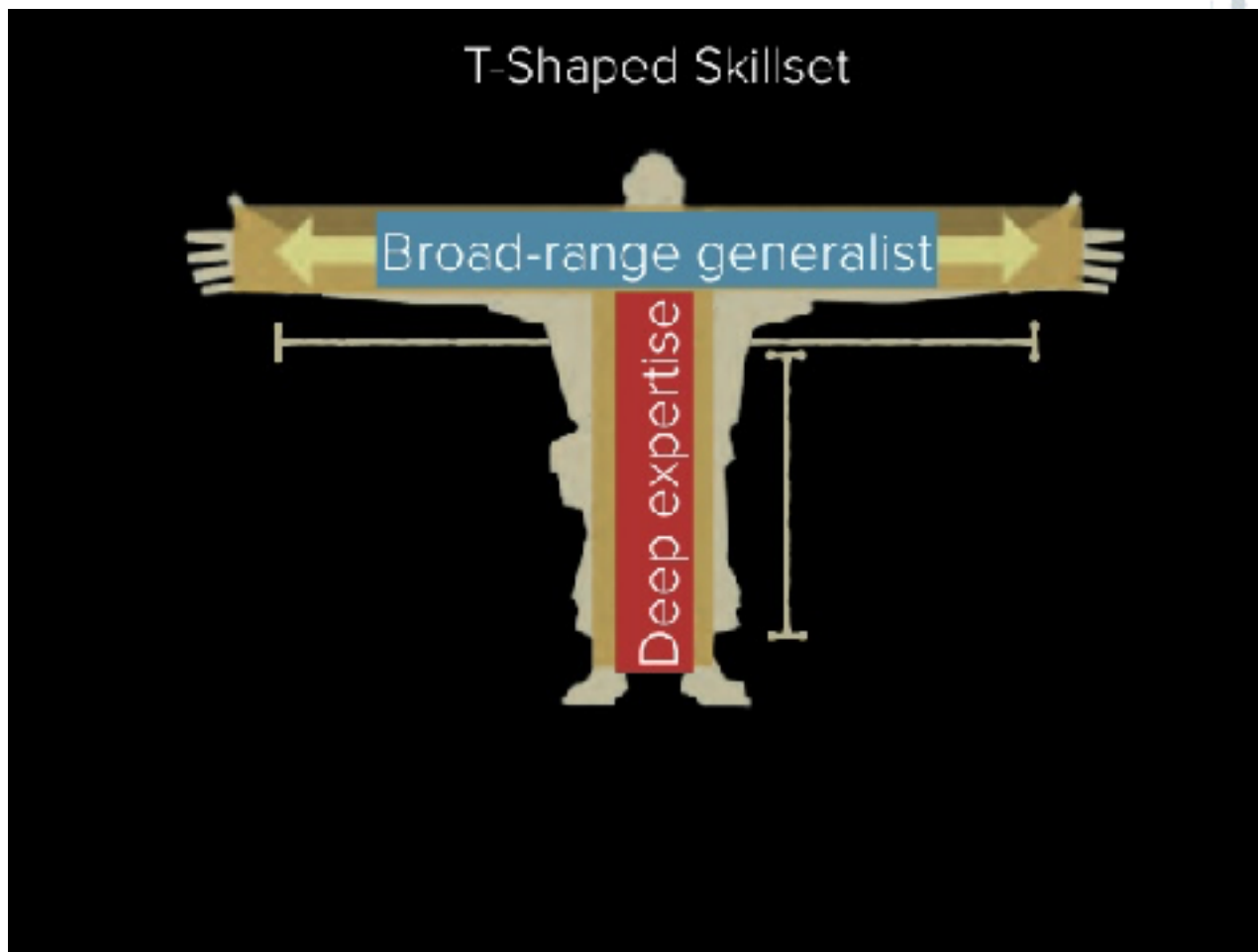
Focus on....

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 \neq 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)

A decorative graphic in the top-left corner featuring a network of interconnected nodes and lines. Some nodes are highlighted with blue circles, and some lines are solid blue, while others are light gray.

Συμβουλή #2



Συνεχείς Μεταβλητές

Στην περίπτωση συνεχών μεταβλητών ή διακριτών με μεγάλο πλήθος τιμών χωρίζουμε τα δεδομένα σε μικρότερο πλήθος από ομάδες. Τις λεγόμενες κλάσεις.

Μέτρα Θέσης

- ◎ Μέση τιμή (\bar{x})
- ◎ Διάμεσος (δ)
- ◎ Εύρος (R)
- ◎ Διακύμανση (s^2)
- ◎ Τυπική Απόκλιση (s)
- ◎ Συντελεστής Μεταβολής (s/\bar{x})


A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles inside, while others are smaller and solid. The lines are thin and gray, connecting the nodes in a non-linear fashion.

5. Excel

Μια μικρή εισαγωγή...



Τι θα δούμε στο excel...

- ◎ Import data from CSV
 - ◎ Filter rows
 - ◎ Βασικές πράξεις μεταξύ στηλών
 - ◎ Split Columns
 - ◎ Max, Min, Sum, Mean
 - ◎ Find and replace NaN
 - ◎ Pivot Tables
 - ◎ Γραφικές παραστάσεις
- 



6. **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

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

1. **Git & Github**

Keep everything clean!



Τι είναι το git;

Ποιος ξέρει....;



A decorative network diagram at the top of the slide, featuring a complex web of interconnected nodes and lines. A central node is highlighted with a dashed circle and a solid circle, containing a large blue quotation mark.

“

Το *Git* είναι ένα σύστημα ελέγχου εκδόσεων (λέγεται και σύστημα ελέγχου αναθεωρήσεων ή σύστημα ελέγχου πηγαίου κώδικα) με έμφαση στην ταχύτητα, στην ακεραιότητα των δεδομένων και στην υποστήριξη 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





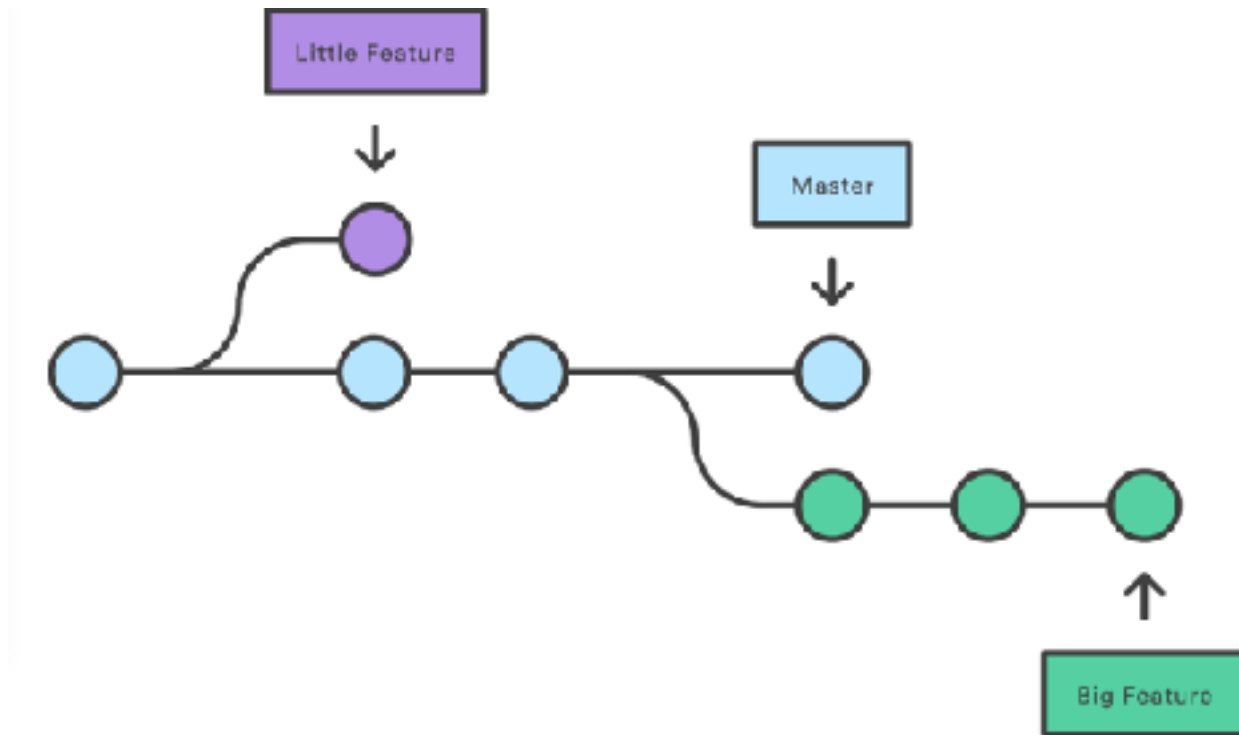
pull

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

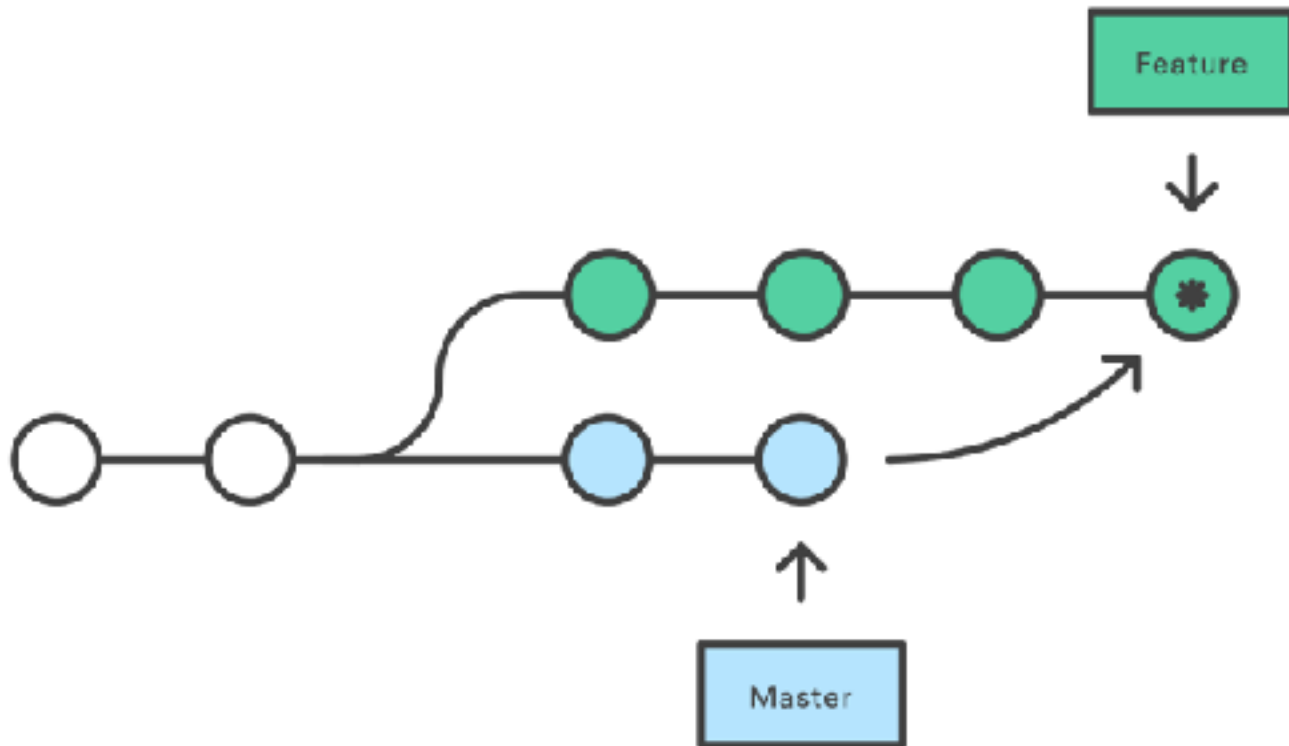
git pull



Branch



Merge





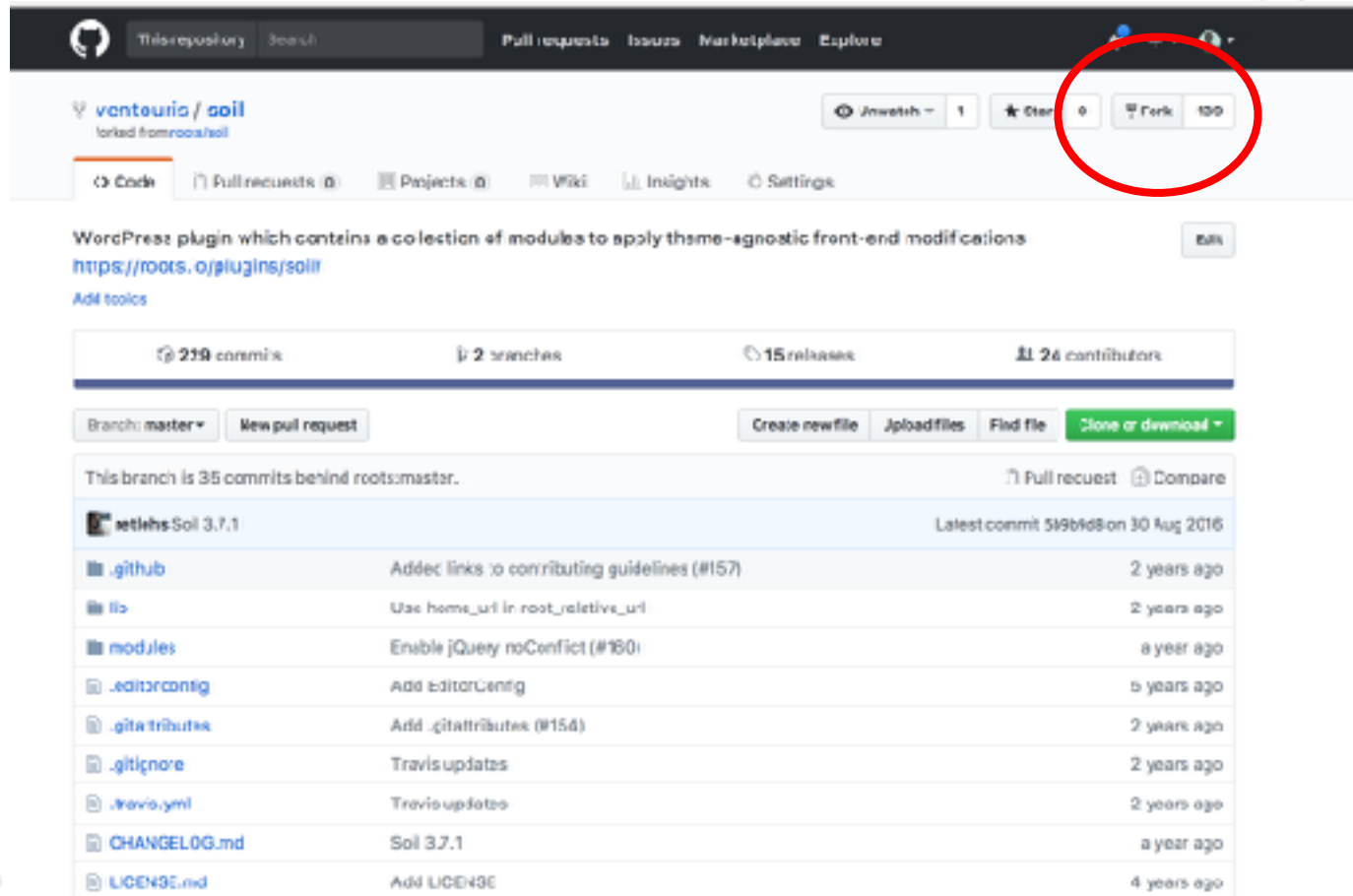
...continue

git branch <name> (new)

git checkout <name> (switch)

git merge <name> (merge with current)

Fork on Github



WordPress plugin which contains a collection of modules to apply theme-agnostic front-end modifications
<https://roots.org/plugins/soil/>

Add tools

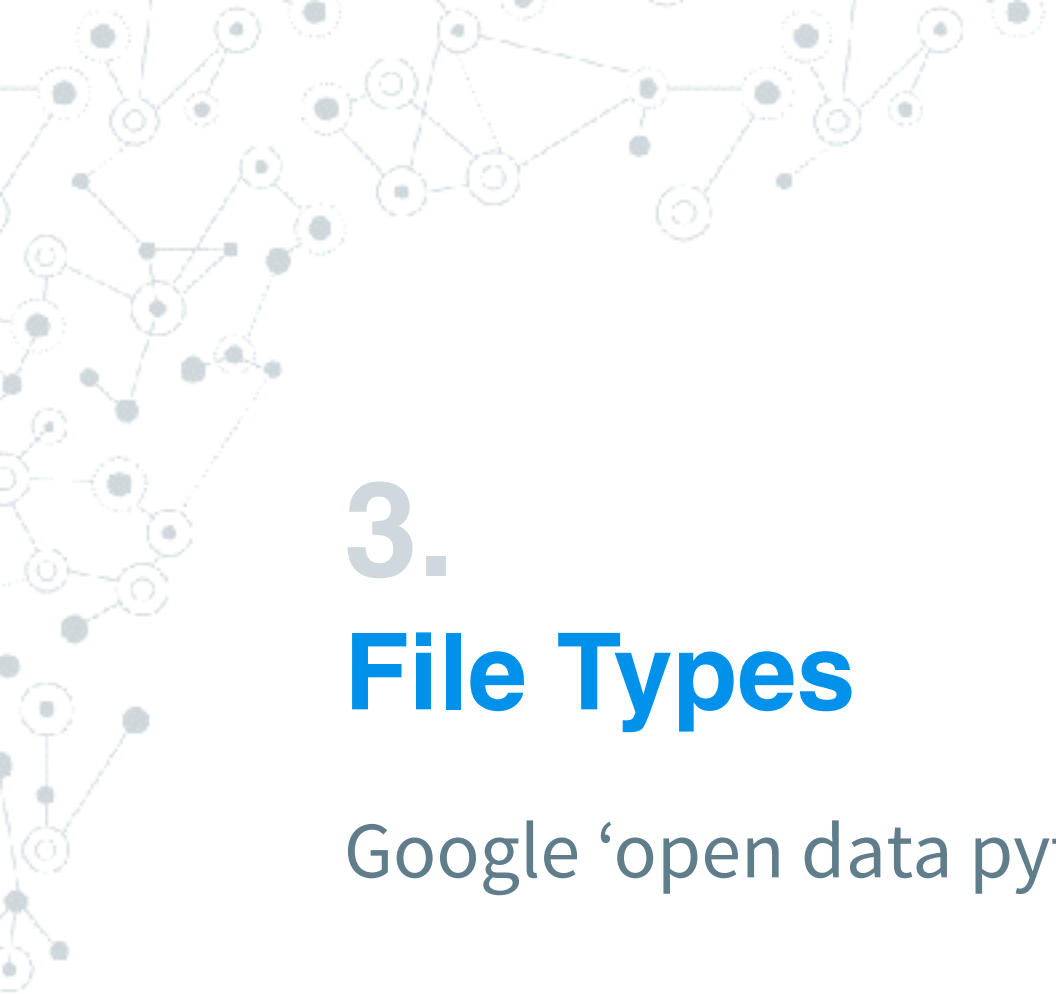
219 commits 2 branches 15 releases 24 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

This branch is 35 commits behind roots:master. Pull request Compare

setlehs Soil 3.7.1 Latest commit 5f9b1d8 on 30 Aug 2016

.github	Add links to contributing guidelines (#157)	2 years ago
lib	Use home_url in root_relative_url	2 years ago
modules	Enable jQuery noConflict (#160)	a year ago
.editorconfig	Add editorConfig	5 years ago
.gitattributes	Add .gitattributes (#154)	2 years ago
.gitignore	Travis updates	2 years ago
.travis.yml	Travis updates	2 years ago
CHANGELOG.md	Soil 3.7.1	a year ago
LICENSE.md	Add LICENSE	4 years ago

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles inside, suggesting a hierarchical or multi-layered structure. The lines are thin and gray, connecting the nodes in a non-linear fashion.

3. **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))
```

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles inside, suggesting a hierarchical or multi-layered structure. The lines are thin and gray, connecting the nodes in a non-linear fashion.

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						



3. **Visualisation**

Ποιο γράφημα να διαλέξω;

Σύγκριση πολλών τιμών



COLUMN



BAR



CIRCULAR
AREA



BULLET



LINE



SCATTER

Ανάλυση της σύνθεσης ενός συνόλου



AREA



WATERFALL



PIE



STACKED
BAR



STACKED
COLUMN

Παρουσίαση κατανομής συνόλου



COLUMN



BAR



LINE



SCATTER

Ανάλυση τάσεων



COLUMN



LINE



DUAL-AXIS
-LINE



BUBBLE

Ανάλυση σχέσεων μεταξύ συνόλων



LINE



SCATTER


A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. Some nodes are highlighted with blue circles, and others with blue dots. The diagram is rendered in a light gray color.

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!!!





4.

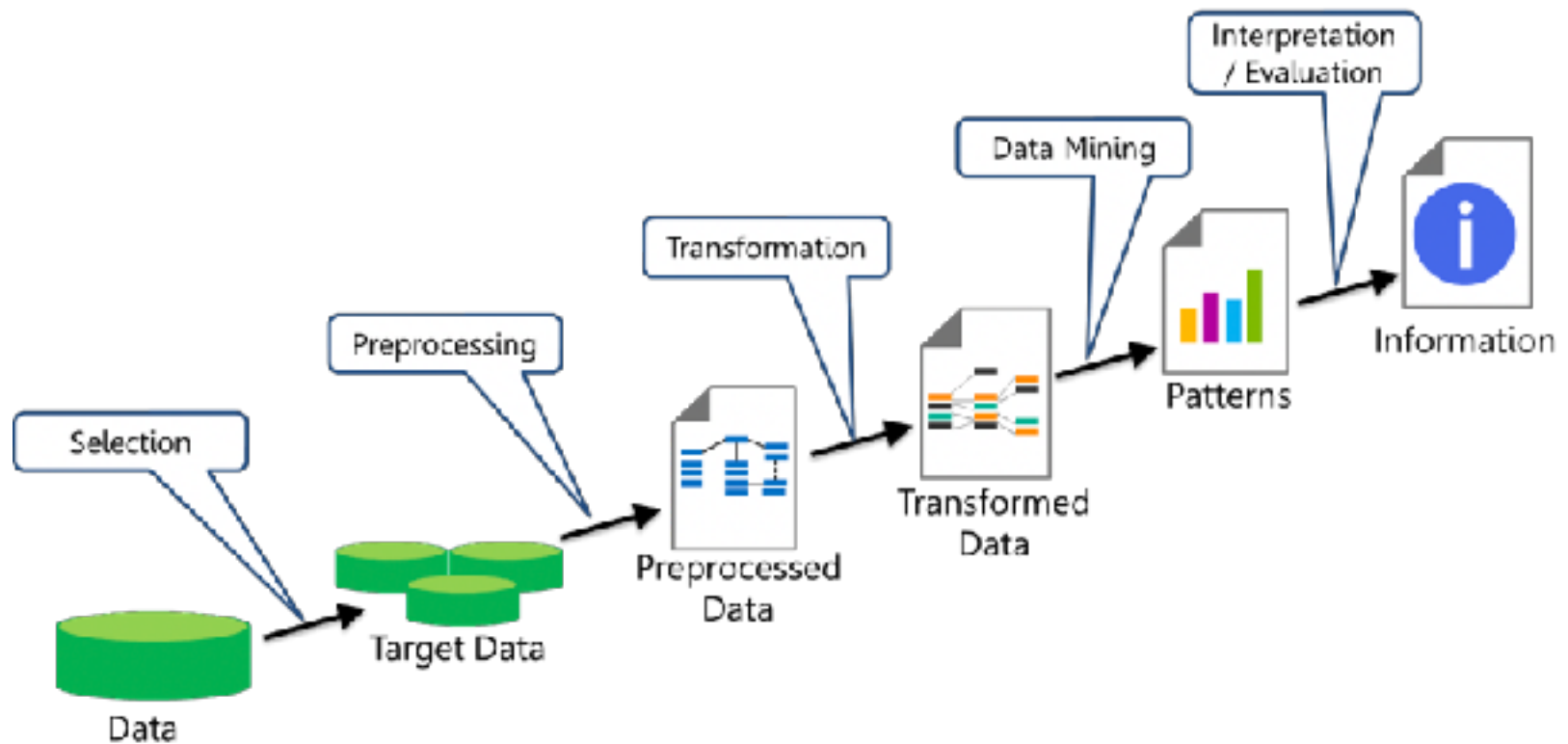
Data Science Process

Ποια είναι τα βήματα;

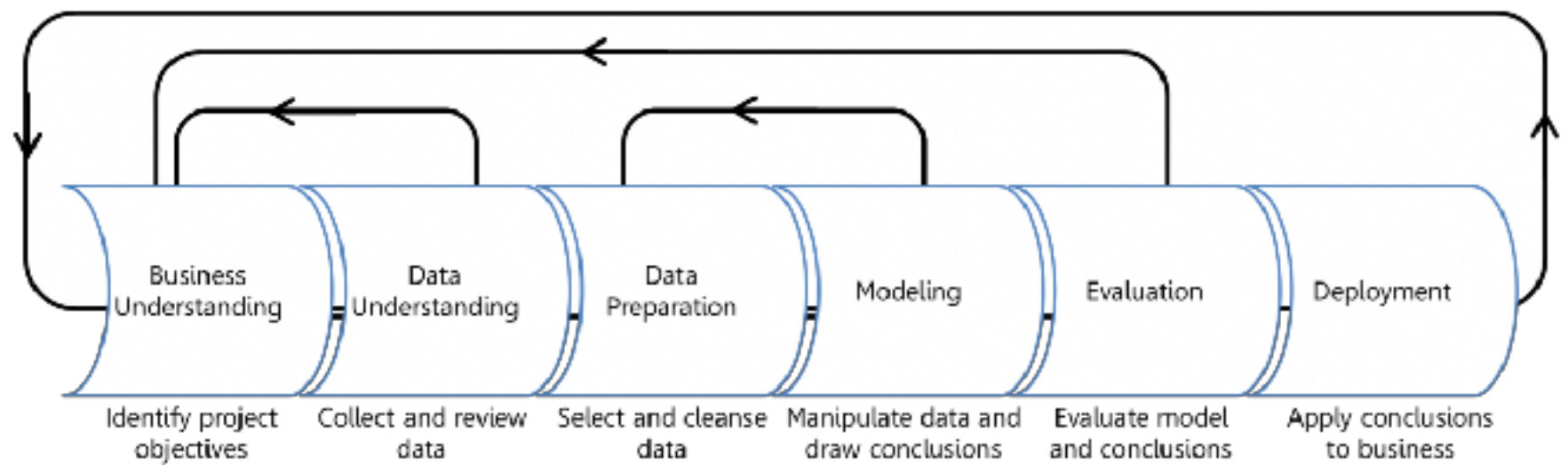
The process

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

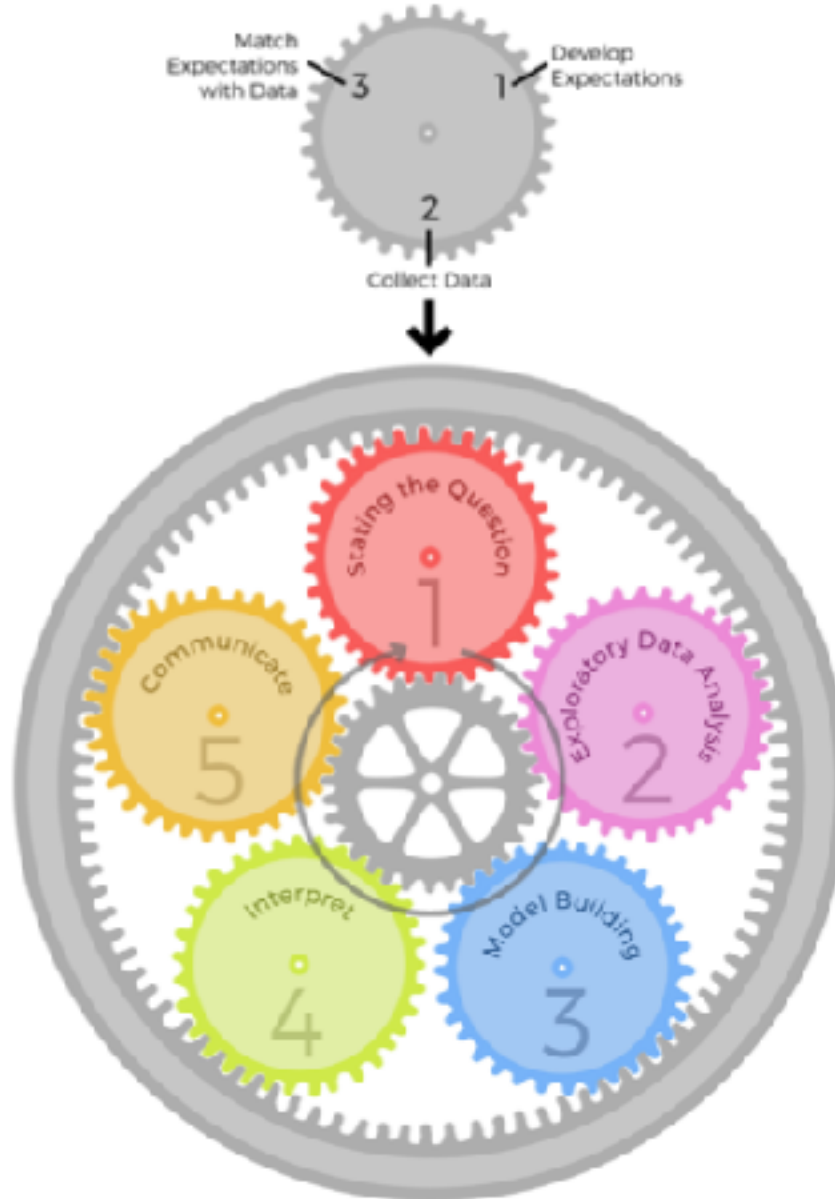
To 1997



To 2000



My Favorite





5. **To 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;

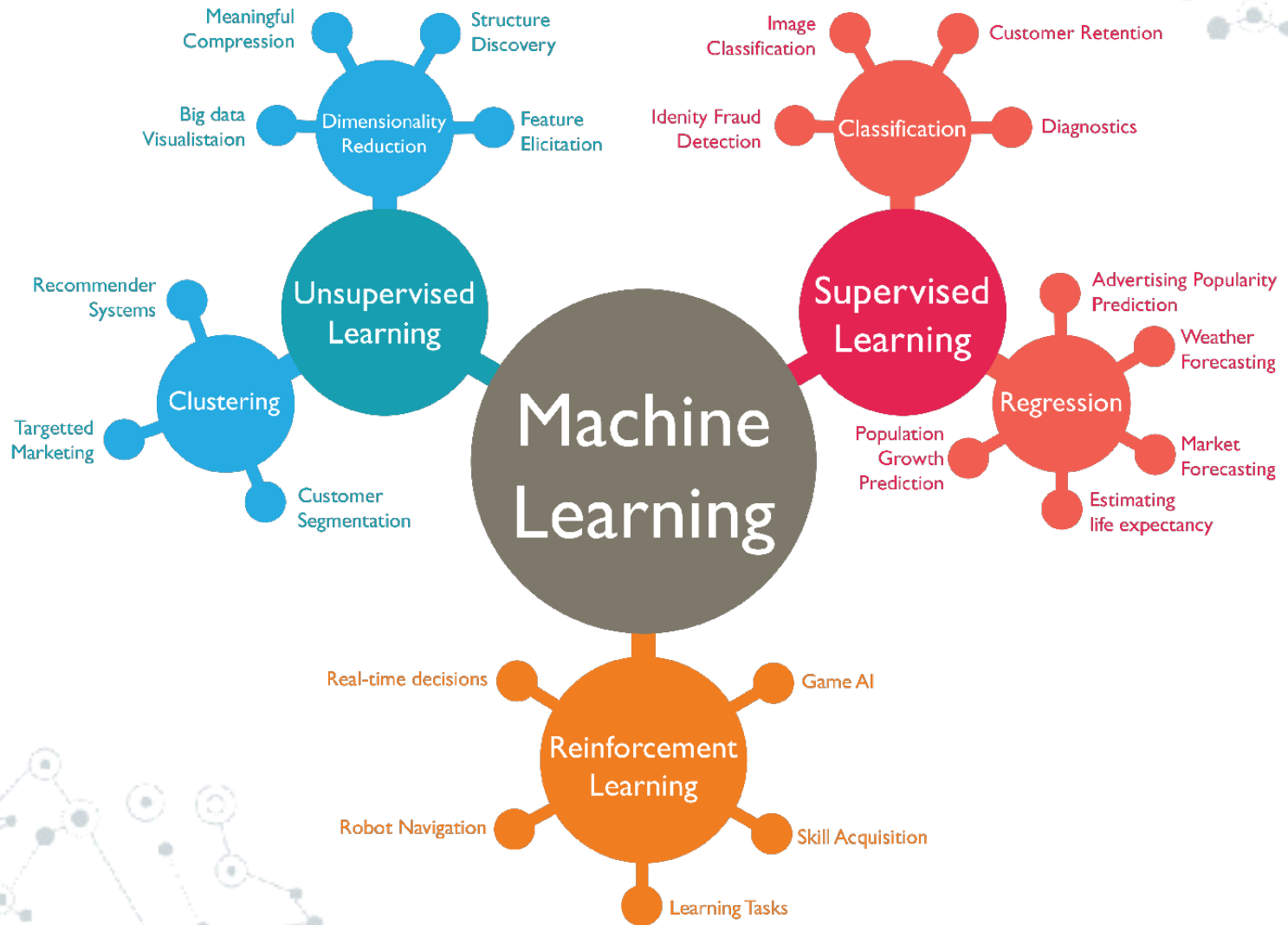


6.

Intro to Machine Learning

Ή αλλιώς μηχανές μάθησης

Machine Learning family

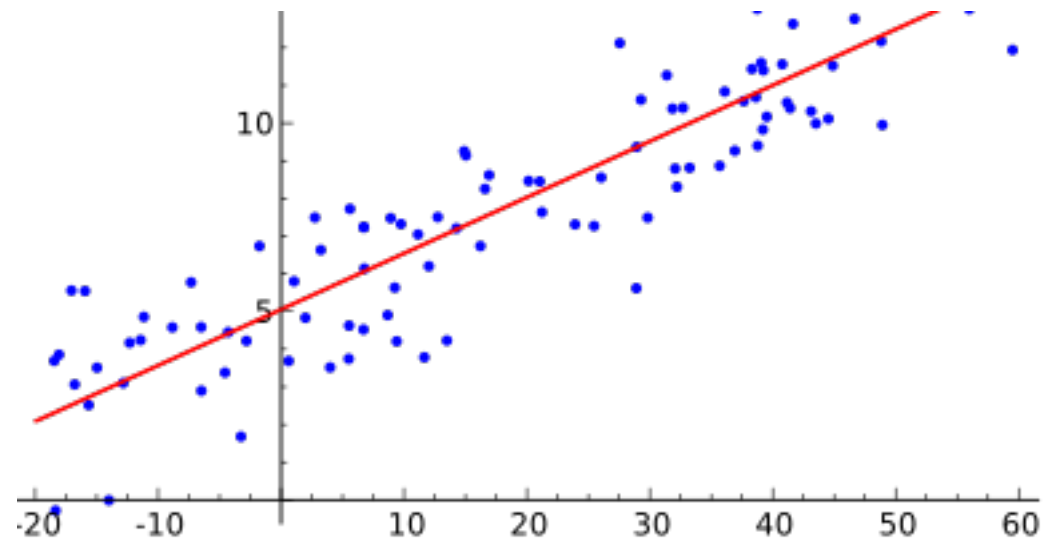




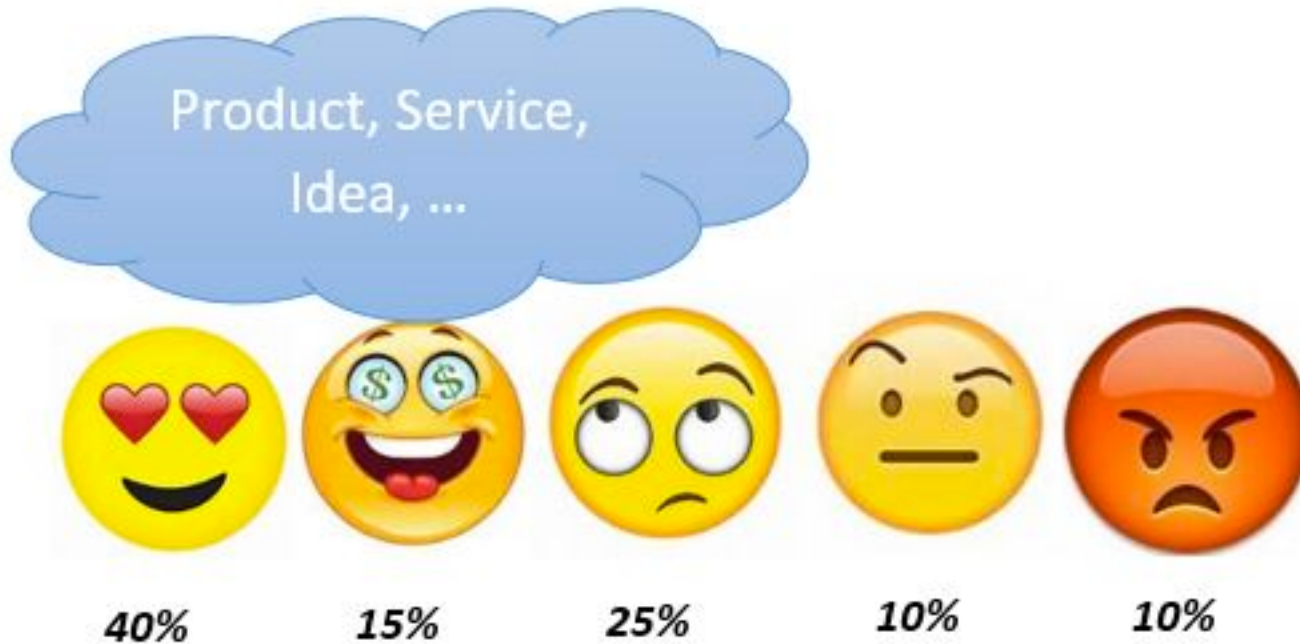
Και τι θα διαλέξω...;

- ◎ Regression
- ◎ Classification
- ◎ Clustering
- ◎ Dimensionality Reduction

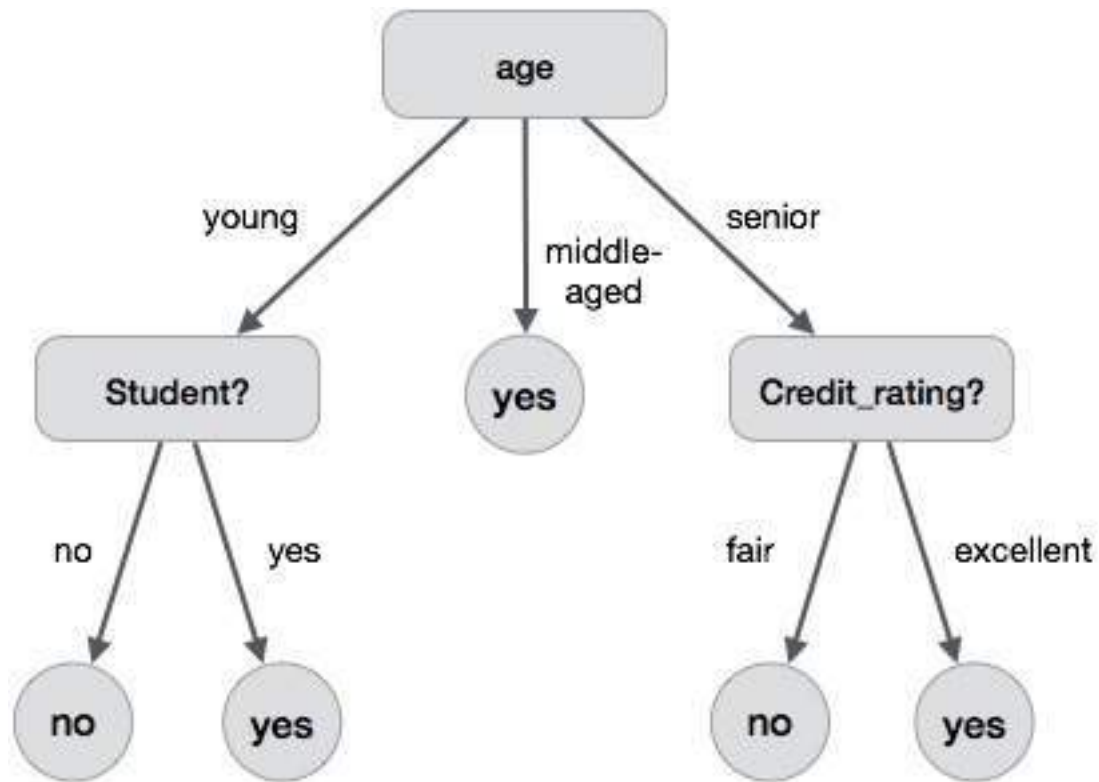
Linear Regression



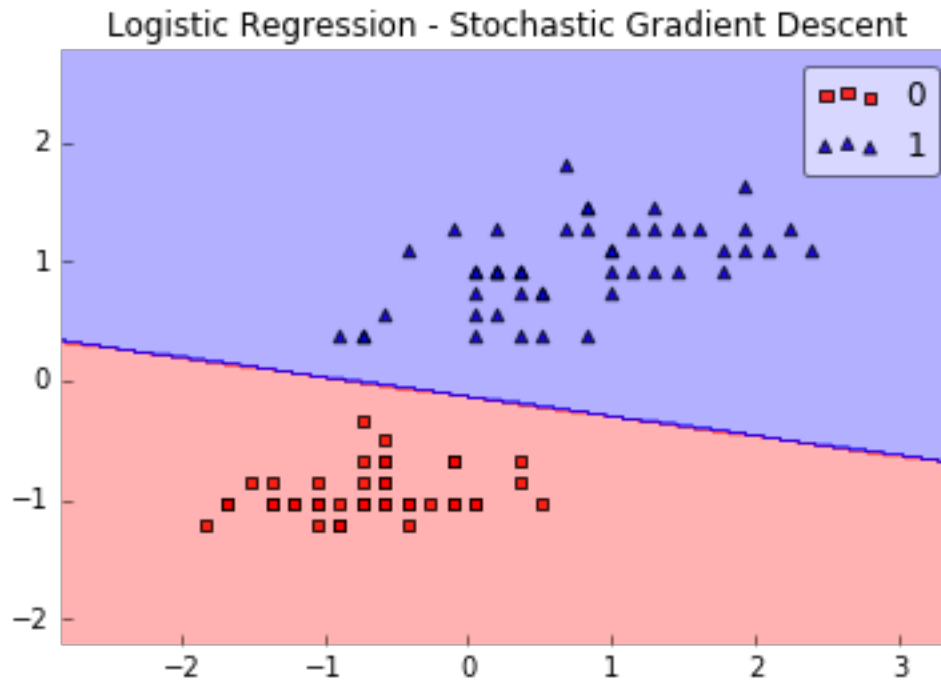
Naive Bayes Classifier



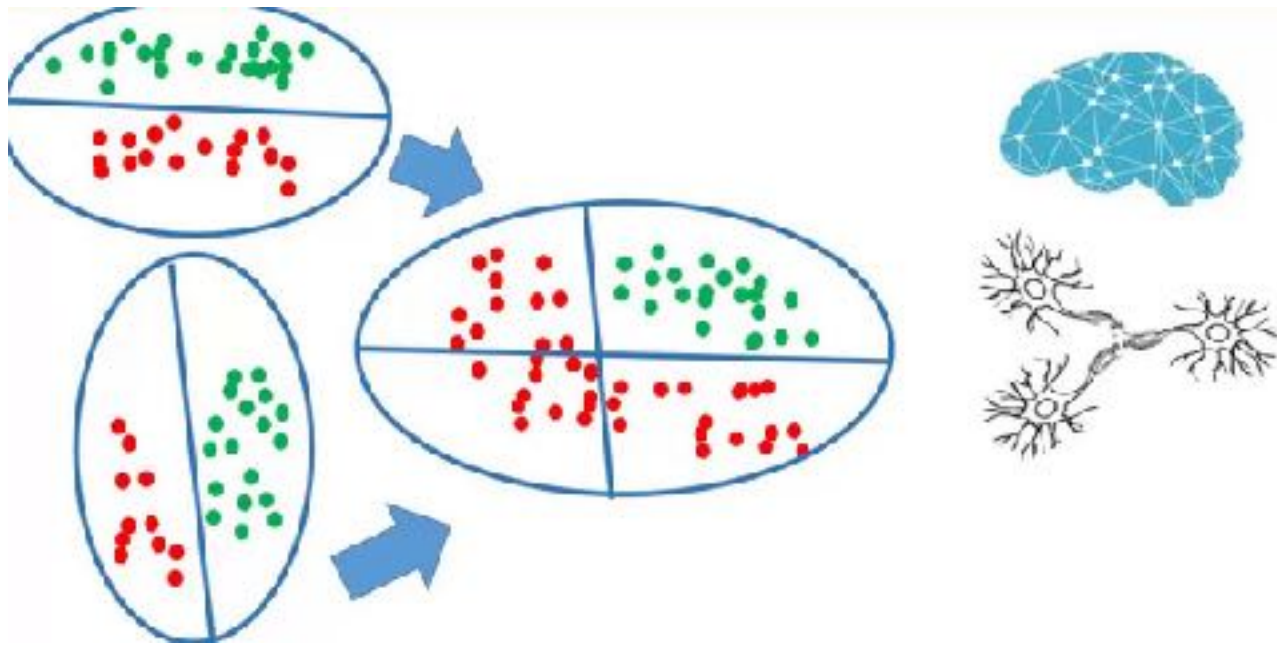
Decision Tree



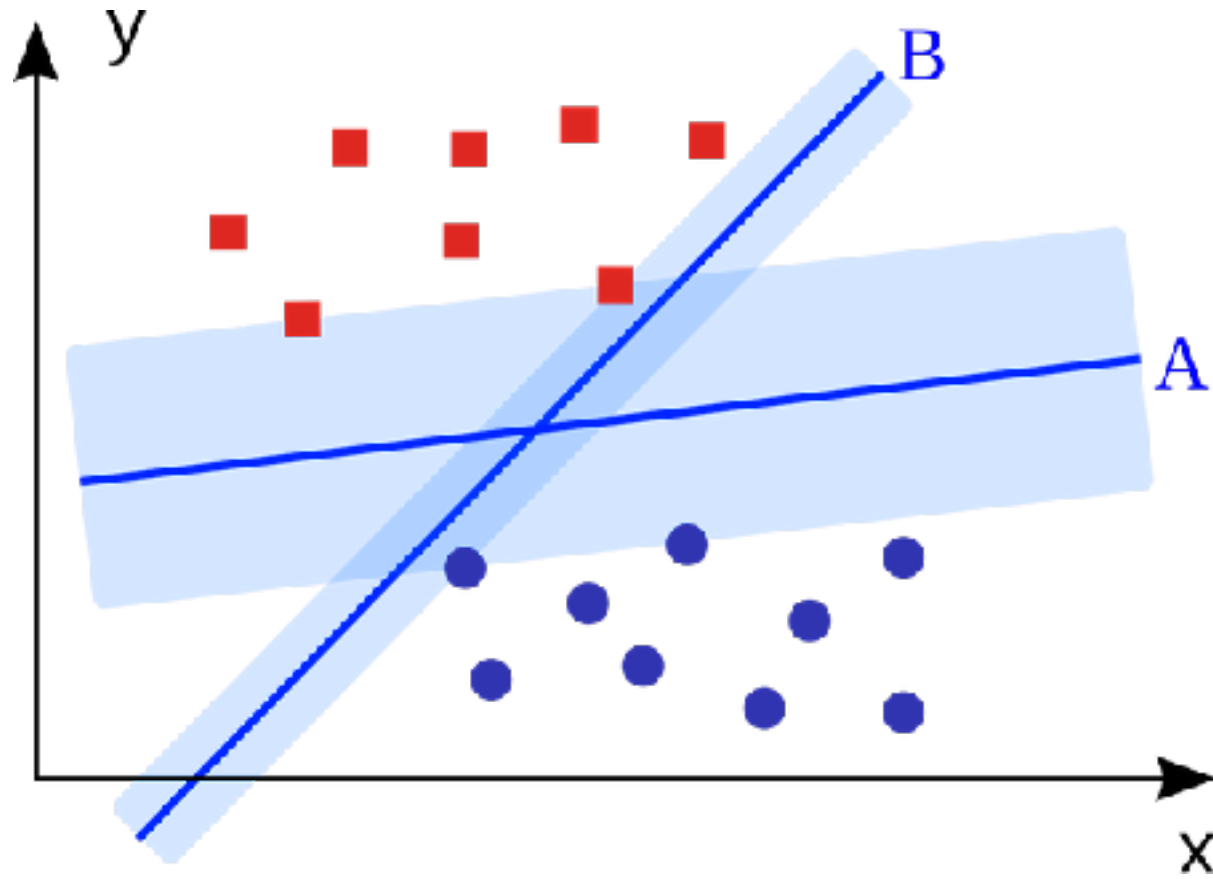
Logistic Regression



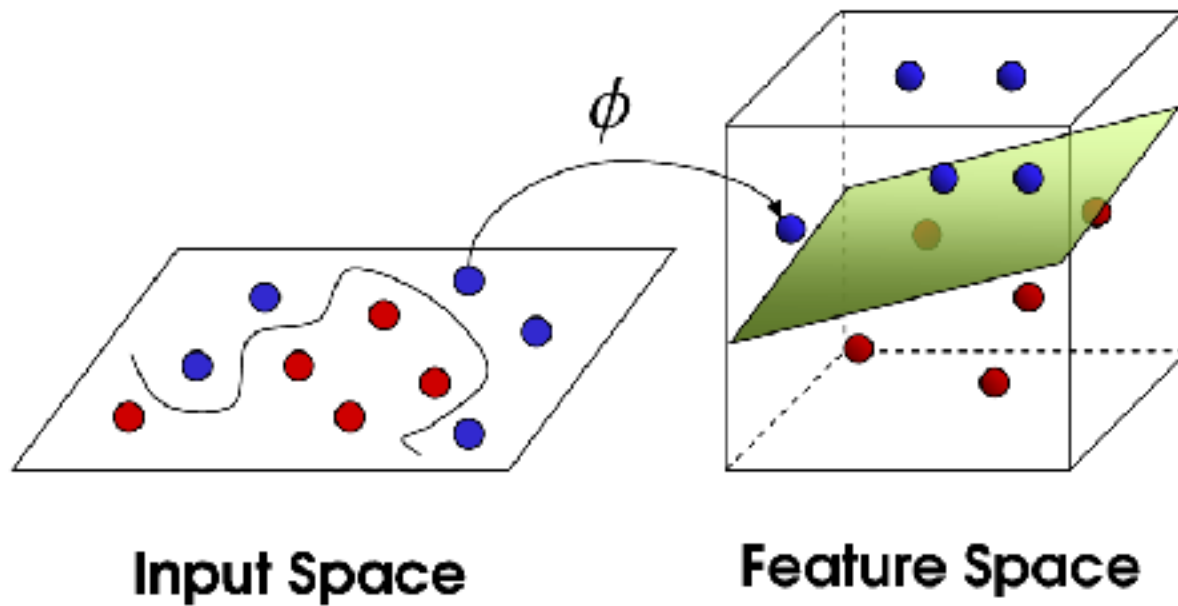
Neural Network



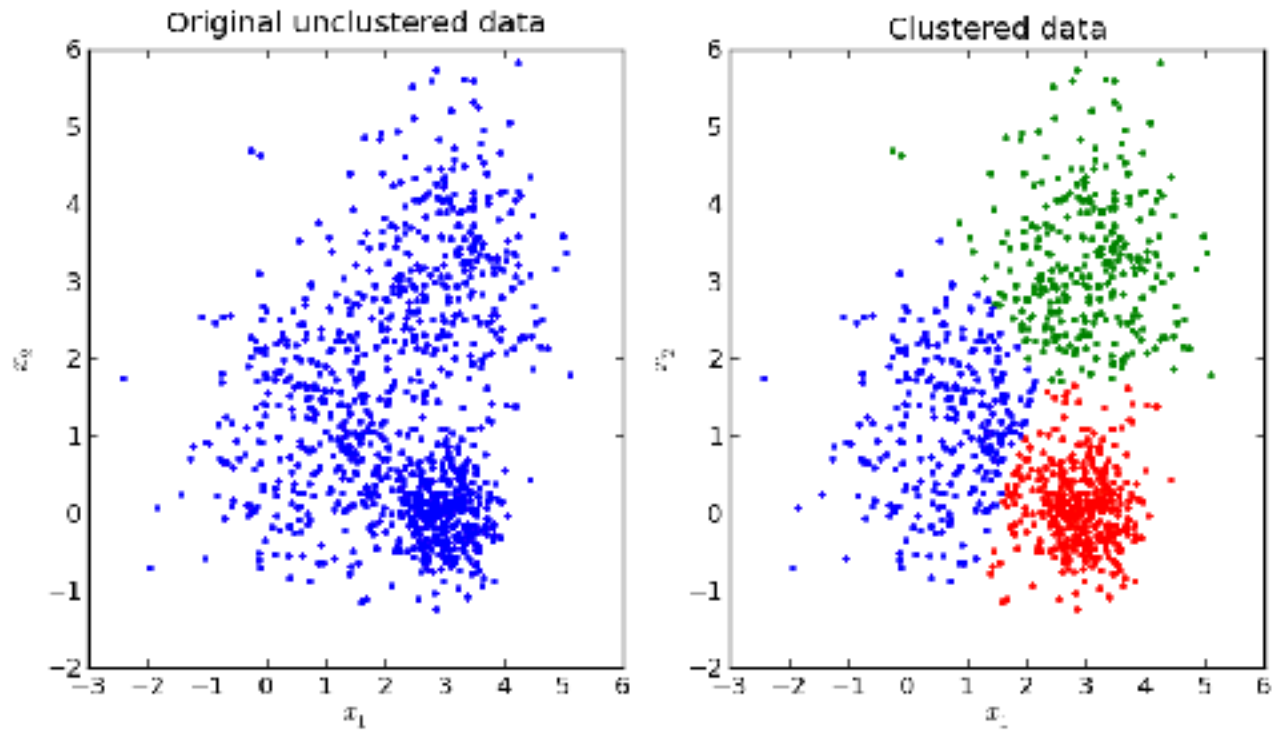
Support Vector Machine



Support Vector Machine (space)

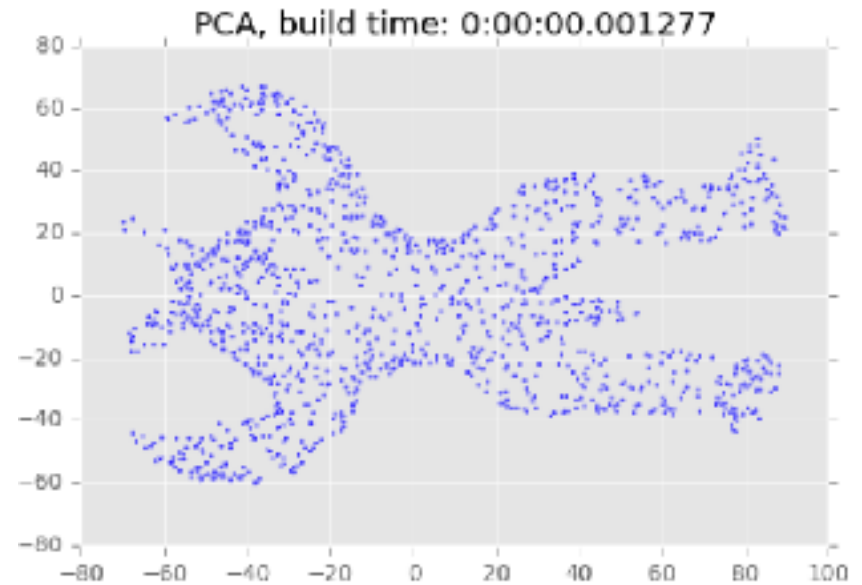
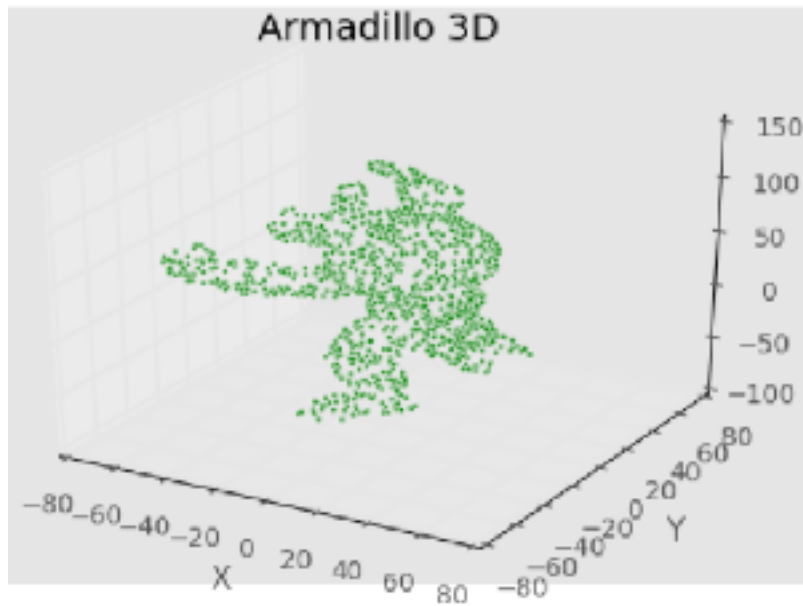


k-means clustering



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

Dimensionality Reduction (Principal Component Analysis)





Day 3

Kaggle, AzureML & SQL 🙄





1.

Kaggle Account

Let's do it!

<https://www.kaggle.com/>

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

2.

Azure ML Account

Let's do it again!

<https://studio.azureml.net/>

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are highlighted with a double-circle outline. The lines are thin and gray, creating a mesh-like structure.

3. SQL


Why SQL?

Γιατί όχι Pandas

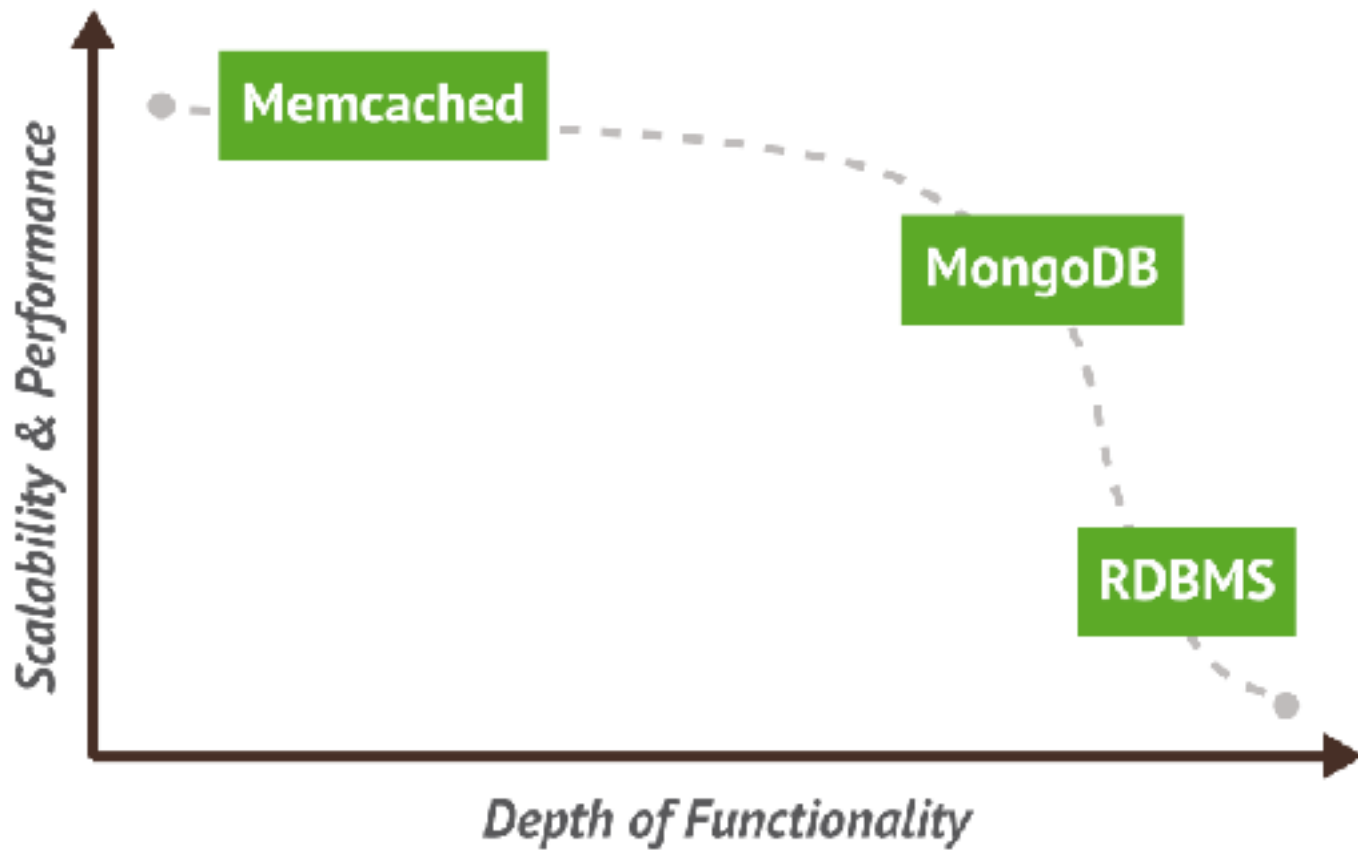
- ◎ Δεν επαρκεί η μνήμη για τα δεδομένα
- ◎ Τα δεδομένα είναι δυναμικά
- ◎ Περισσότερα από ένα άτομα
- ◎ Security




SQL extensions

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

noSQL vs SQL



A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

4. **SQL examples**

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



Day 4

Even more Python 🥹

