

BIG DATA & DATA ANALYTICS



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS



Γεράσιμος Ραζής (razis@uth.gr)

Θέματα συζήτησης

- Big Data
- Analytics
 - Descriptive
 - Predictive
 - Prescriptive
- MapReduce
- Hadoop
- NoSQL
- Data Mining -
Machine Learning



Big Data (Μεγάλα Δεδομένα)

Σύνολα δεδομένων τόσο μεγάλα ή σύνθετα που ξεφεύγουν από τις δυνατότητες καταγραφής, αποθήκευσης και ανάλυσης των παραδοσιακών τεχνικών επεξεργασίας δεδομένων

Bytes

- 1KB = 2^{10} bytes
- 1MB = 2^{20} bytes
- 1GB = 2^{30} bytes
- 1TB = 2^{40} bytes
- 1PB = 2^{50} bytes
- 1EB = 2^{60} bytes
- 1ZB = 2^{70} bytes
- 1YB = 2^{80} bytes



<http://en.wikipedia.org/wiki/Yottabyte>

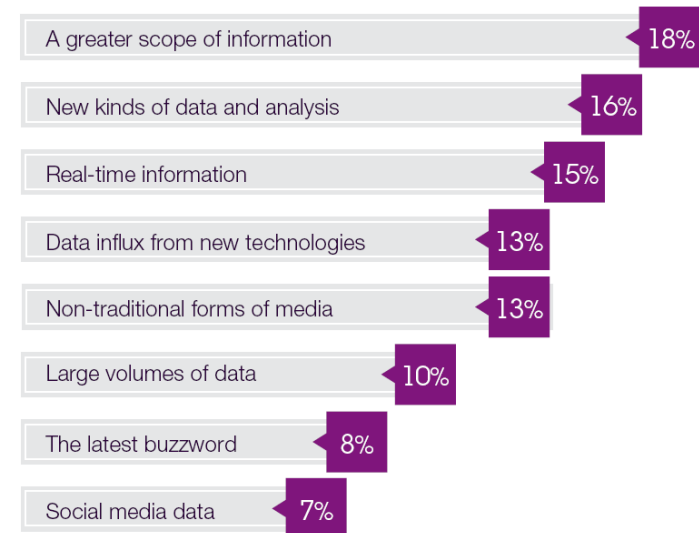


1Yottabyte

Πώς ορίζονται τα Big Data;

- Τεράστιες ποσότητες δομημένων, ημιδομημένων και αδόμητων δεδομένων
- Τα Big Data είναι ο συνδυασμός εξελίξεων στην τεχνολογία που συνέβησαν τα τελευταία 50 έτη

Defining big data



Respondents were asked to choose up to two descriptions about how their organizations view big data from the choices above. Choices have been abbreviated, and selections have been normalized to equal 100%. Total respondents=1144.

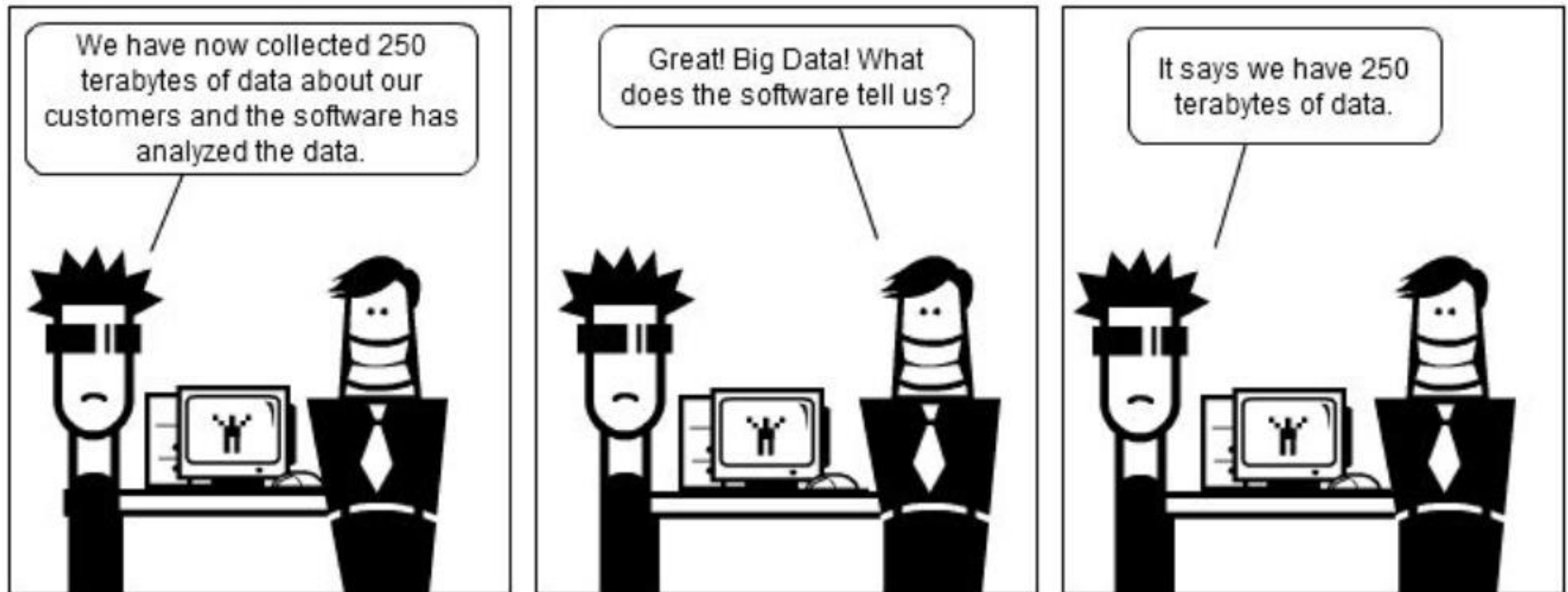
Figure 1: Respondents were split in their views of big data.

The Big Data Challenge

View more social media cartoons at

www.socmedsean.com

The Big Data Challenge



Μορφές δεδομένων

Δομημένα

- Σχεσιακές βάσεις δεδομένων
- XML
- JSON

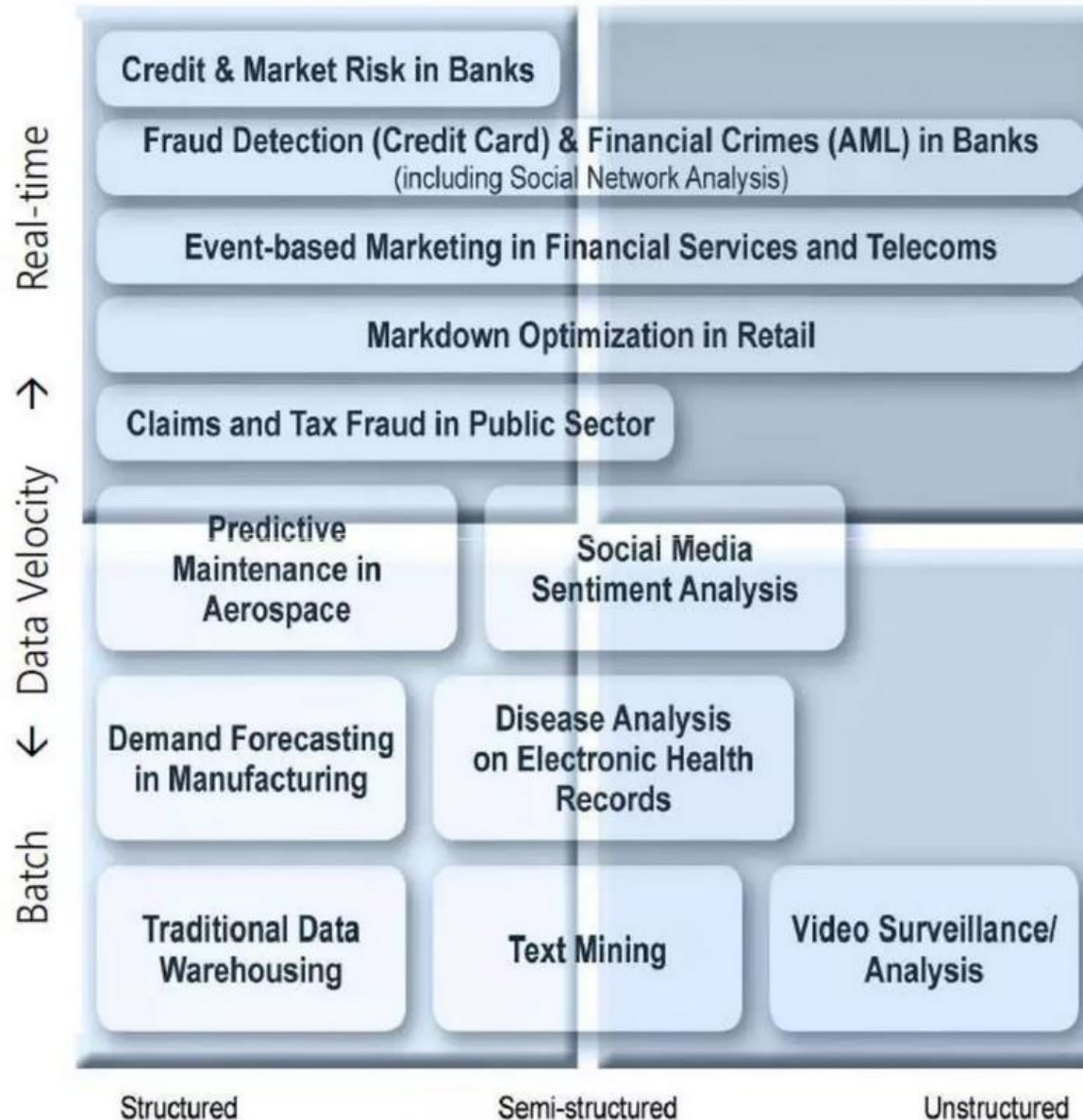
Ημιδομημένα

- CSV
- Emails
- Tweets
- Facebook status
- Σχόλια σε Blogs

Αδόμητα

- Κείμενο
- Εικόνα
- Ήχος
- Βίντεο

Μορφές δεδομένων



Τεχνολογίες που προηγήθηκαν των Big Data

- Σχεσιακές Βάσεις Δεδομένων
- Data warehouses και data marts (ημερήσια ή εβδομαδιαία ενημέρωση)
- Object Oriented Βάσεις Δεδομένων

BLOB = Binary Large Objects

Αλλαγές τελευταίων ετών

- Μείωση τιμών αισθητήρων
- Μείωση κόστους για αποθήκευση – επεξεργασία
- Αλλαγή συμπεριφοράς χρηστών – αποδοχή διάθεσης προσωπικών πληροφοριών
- Σημαντική πρόοδος σε αλγορίθμους Μηχανικής Μάθησης



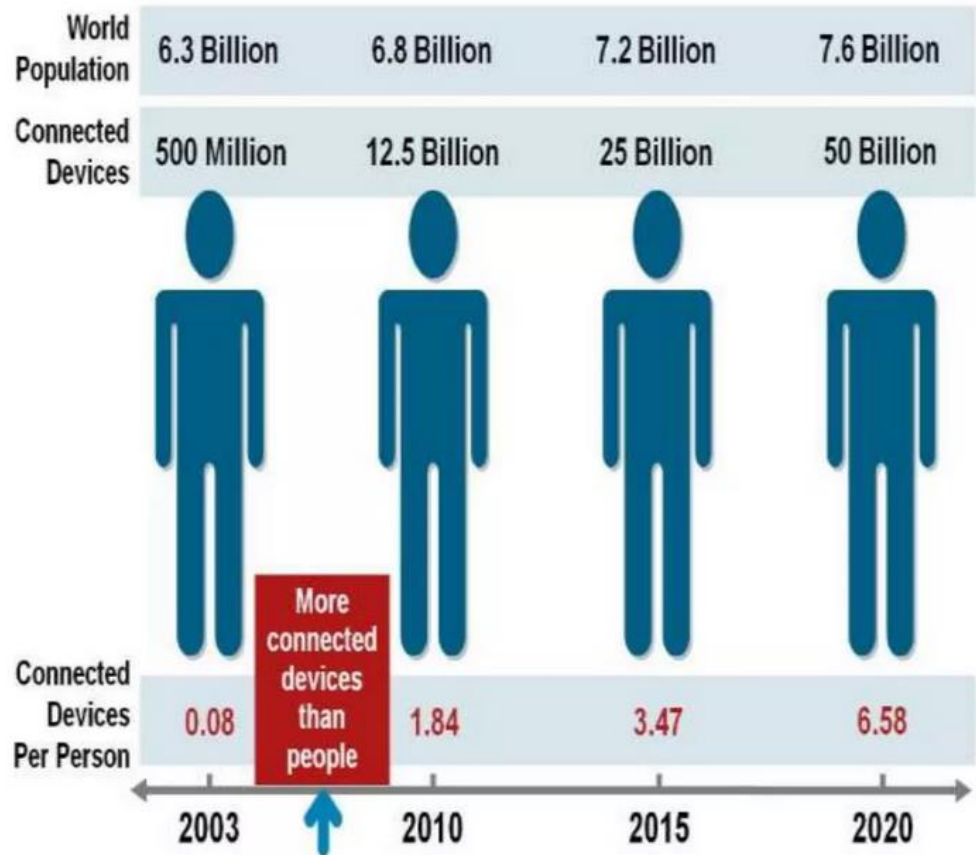
Σημείο καμπής

- Το κόστος της υπολογιστικής επεξεργασίας και αποθήκευσης έφτασε σε κομβικό σημείο ανάμεσα στο 2008 και το 2010
- Περισσότερες επιχειρήσεις έχουν πλέον την δυνατότητα να διαχειρίζονται Big Data

Μορφές δεδομένων

Figure 1. The Internet of Things Was "Born" Between 2008 and 2009

- Το Facebook «παράγει» 10TB δεδομένων ημερησίως
- Το Twitter (X) «παράγει» 7TB δεδομένων ημερησίως
- Η IBM υποστηρίζει πως το 90% των αποθηκευμένων δεδομένων σήμερα παράχθηκε τα τελευταία 2 χρόνια

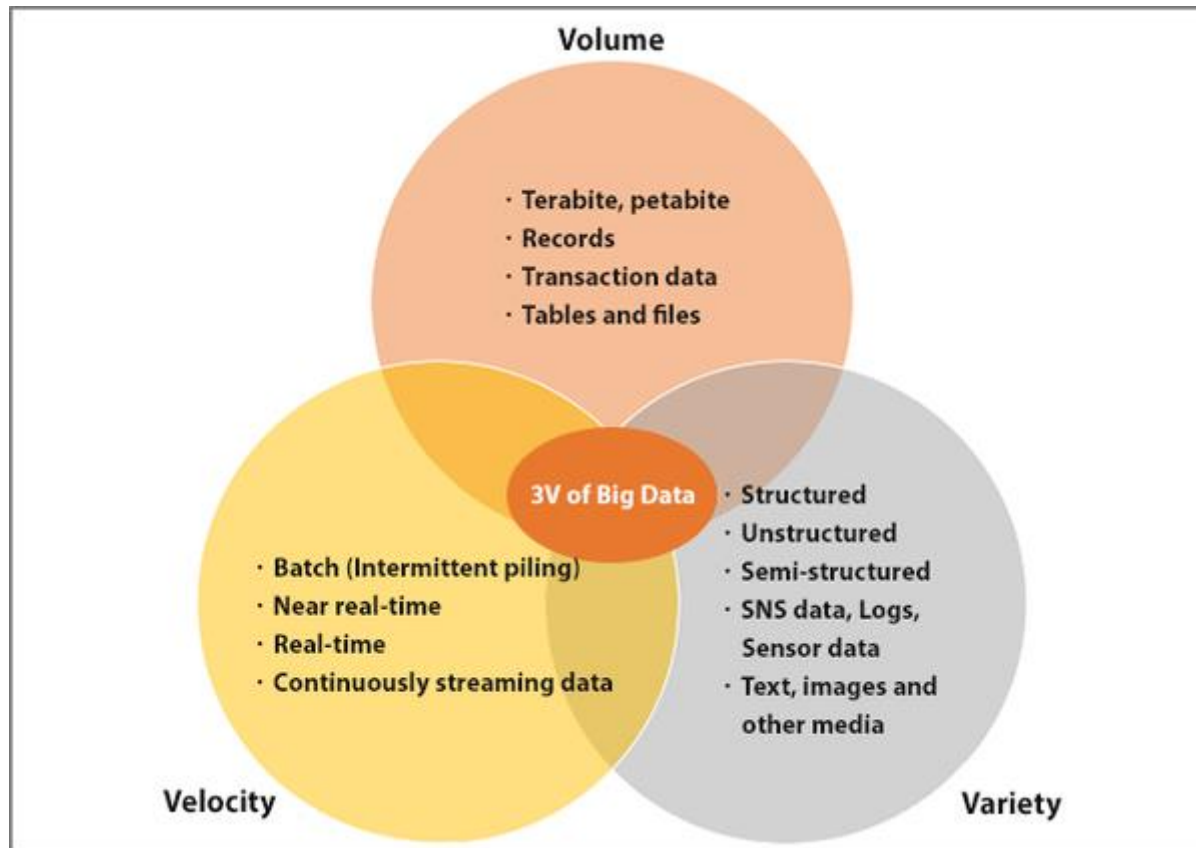


Source: Cisco IBSG, April 2011

Big Data από τη “ματιά” της GenAI



The 3 Vs of Big Data



Volume (Όγκος – Ποσότητα δεδομένων)

- Terabytes έως Petabytes δεδομένων
- Η ποσότητα των δεδομένων που συλλέγονται αυξάνεται συνεχώς
- Ότι θεωρείται σήμερα ως μεγάλα δεδομένα στο μέλλον θα είναι ακόμα μεγαλύτερο

Variety

(Ποικιλομορφία)

- Συγκέντρωση δεδομένων από διάφορες πηγές εντός και εκτός επιχείρησης
- Αισθητήρες
- Έξυπνες συσκευές
- Μορφές δεδομένων
 - Κείμενο
 - Δεδομένα πλοήγησης στο διαδίκτυο
 - Tweets
 - Δεδομένα αισθητήρων
 - Ήχος
 - Βίντεο
 - Αρχεία καταγραφής (logs)
 - ...

Velocity (Ταχύτητα)

- Η ταχύτητα με την οποία δημιουργούνται τα δεδομένα συνεχώς αυξάνεται
- Ορισμένες εφαρμογές απαιτούν λήψη αποφάσεων σε πραγματικό χρόνο (real time)
 - ανίχνευση απάτης (credit card fraud detection)
 - συστήματα συστάσεων (recommendation systems)

2012 Big Data @ work study

IBM Global Business Services
Business Analytics and Optimization
Executive Report

IBM Institute for Business Value

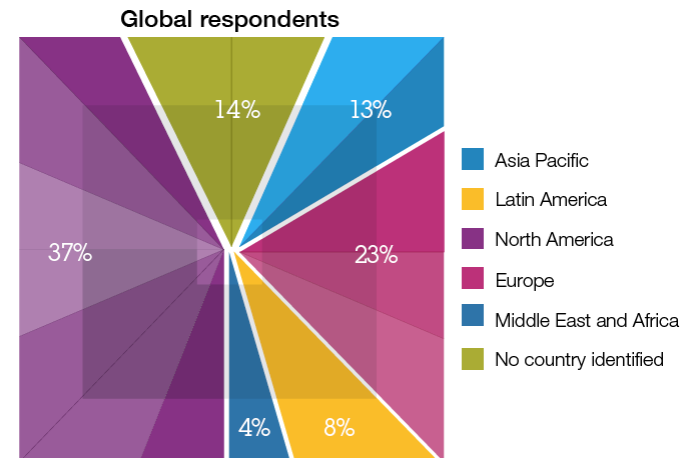
In collaboration with Said Business School at the University of Oxford



- IBM Institute of Business Value + University of Oxford Said Business School
- 1144 επιχειρήσεις σε 95 χώρες

Analytics: The real-world use of big data

How innovative enterprises extract value from uncertain data

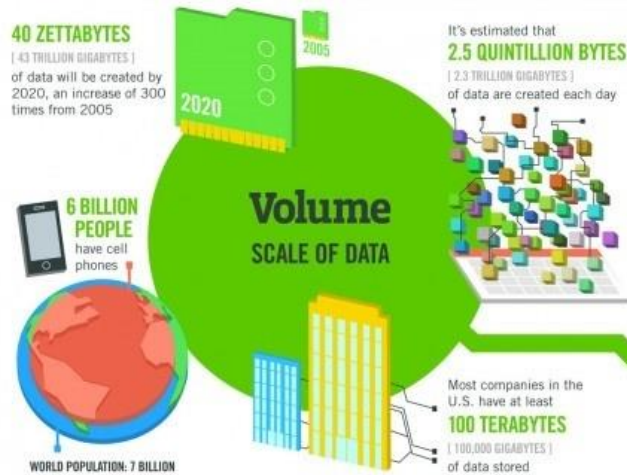


Συμπεράσματα έρευνας 2012 Big Data @ work study

- 63% πιστεύει ότι τα Big Data δημιουργούν συγκριτικό πλεονέκτημα (37% σε αντίστοιχη έρευνα* του 2010)
- Οι επιχειρήσεις προσεγγίζουν πραγματιστικά τα Big Data
- Μόλις 7% ορίζει τα Big Data σε σχέση με δεδομένα Κοινωνικών Δικτύων

*IBM's 2010 New Intelligence
Enterprise Global Executive Study
and Research Collaboration

The 4 Vs of Big Data



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT are shared on Facebook every month



By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month



400 MILLION TWEETS are sent per day by about 200 million monthly active users

Variety
DIFFERENT FORMS OF DATA



The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Velocity
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS**

— almost 2.5 connections per person on earth



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions



Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR

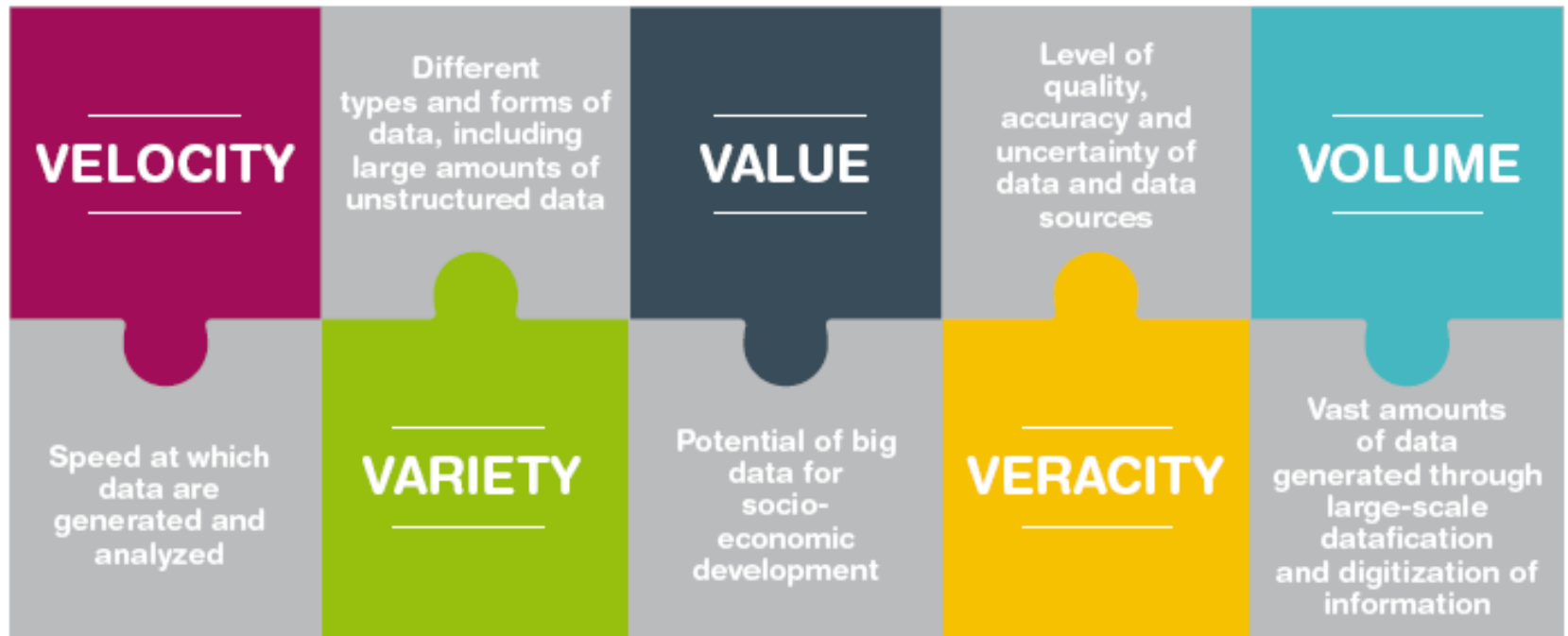


27% OF RESPONDENTS

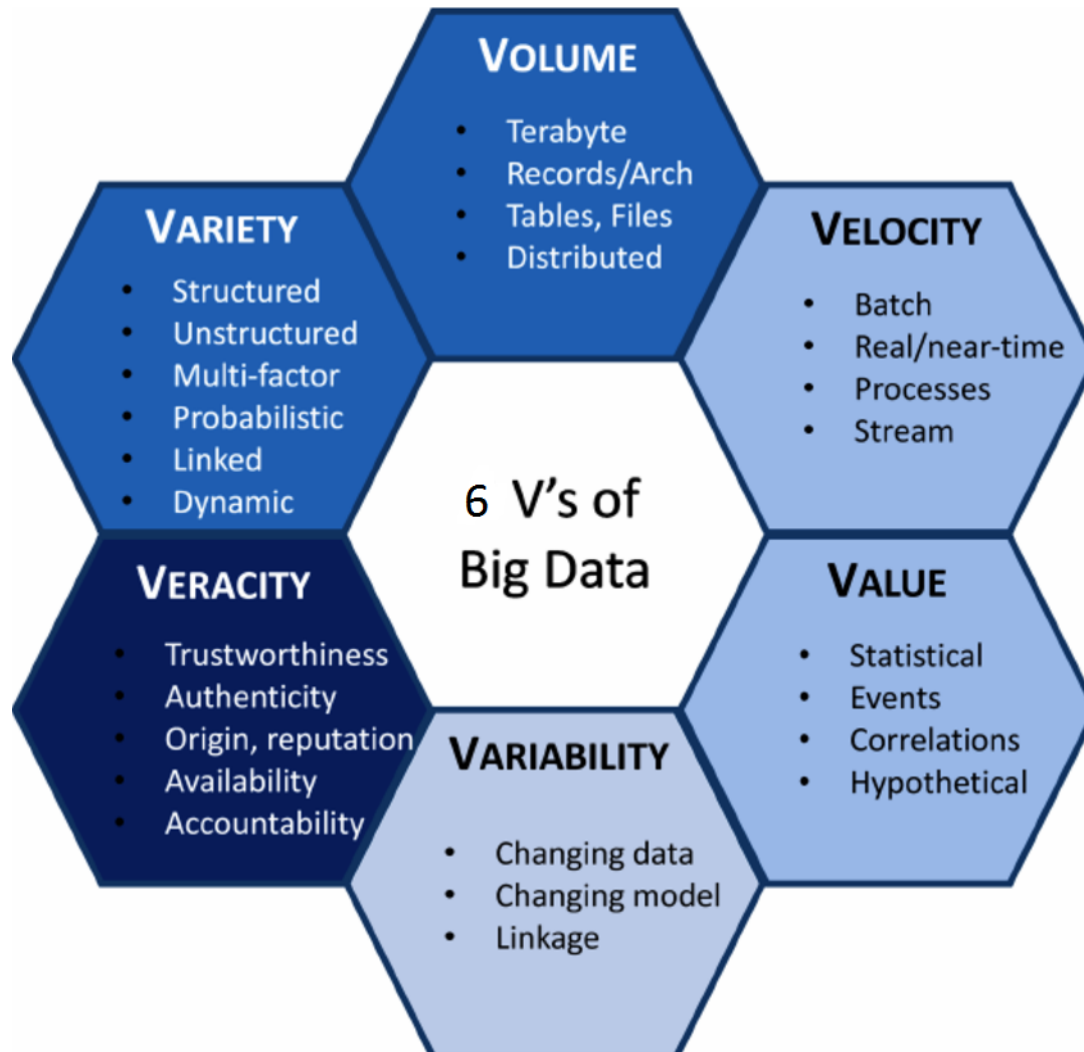
in one survey were unsure of how much of their data was inaccurate

Veracity
UNCERTAINTY OF DATA

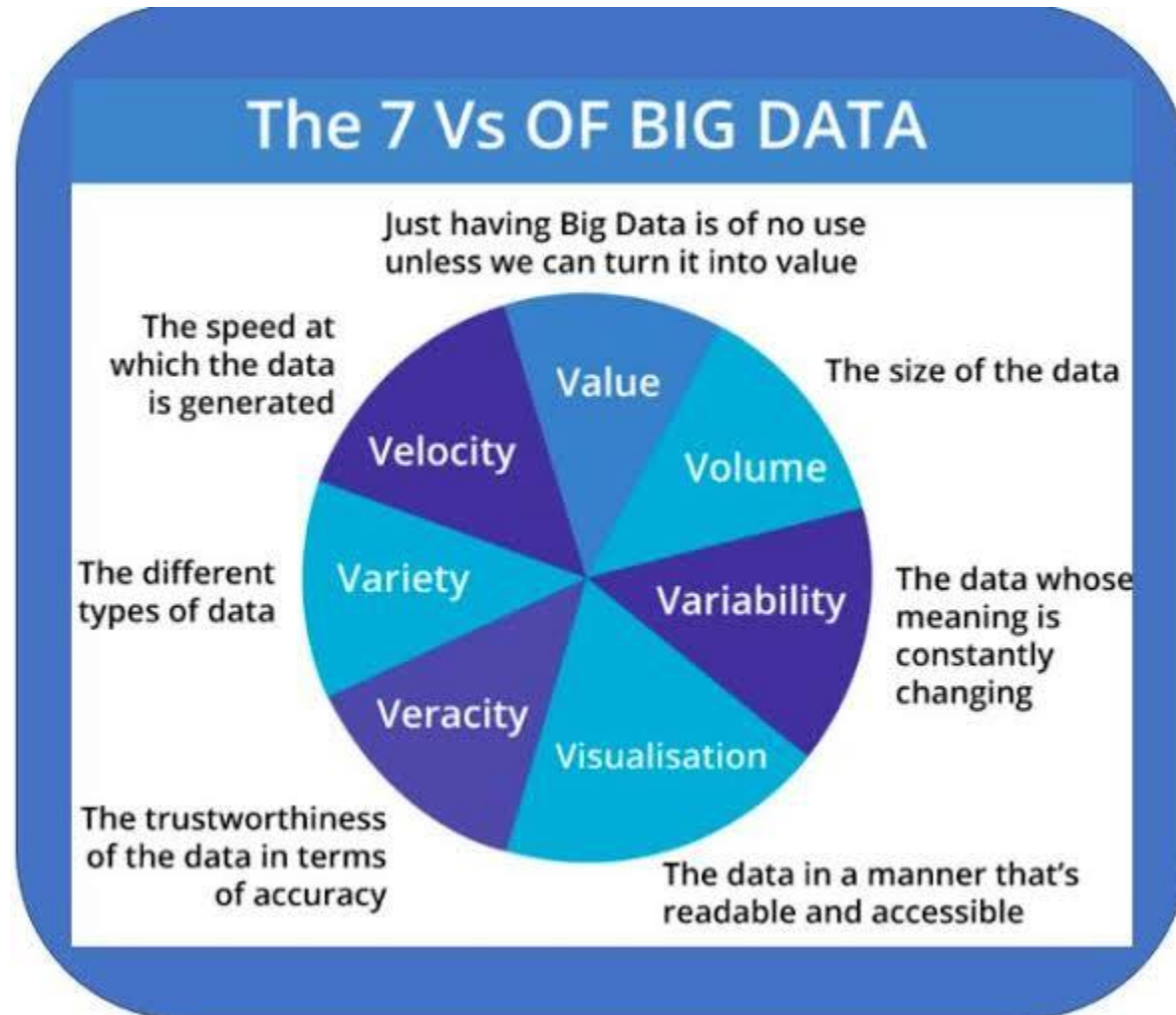
The 5 Vs of Big Data



The 6 Vs of Big Data



The 7 Vs of Big Data



The 8 Vs of Big Data



The 10 Vs of Big Data



Διαχείριση των Big Data

Αποθήκευση

- Καταγραφή και αποθήκευση των δεδομένων

Επεξεργασία

- Καθάρισμα και ανάλυση των δεδομένων (analytics)

Πρόσβαση

- Ανάκληση και οπτικοποίηση των δεδομένων



Analytics (Αναλυτική)

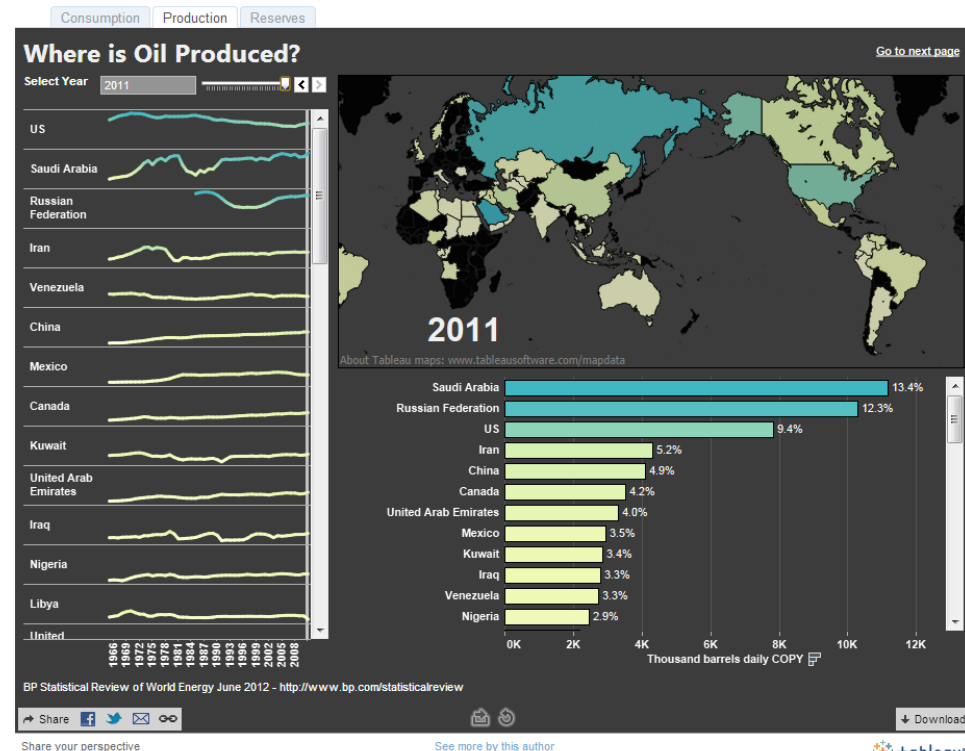
Συστηματική υπολογιστική ανάλυση δεδομένων ή στατιστικών για την ανακάλυψη, ερμηνεία και επικοινωνία σημαντικών προτύπων

Analytics (Αναλυτική)



Descriptive Analytics (Περιγραφική Αναλυτική)

- Περιγραφή των δεδομένων με στόχο την **κατανόησή τους**
- Δημιουργία **διαίσθησης** για τα δεδομένα
- Δημιουργία αναφορών (reports)
- Εντοπισμός καταστάσεων που χρήζουν προσοχής
- Ομαδοποίηση αντικειμένων με παρόμοια χαρακτηριστικά (clustering)



<http://www.tableausoftware.com/learn/gallery>

Predictive Analytics (Προγνωστική Αναλυτική)

- Χρήση των δεδομένων που διαθέτουμε για ορισμένα αντικείμενα προκειμένου να προβλέψουμε τις τιμές άλλων αντικειμένων
- Υπάρχουν πολλά μοντέλα πρόβλεψης με καλύτερες και χειρότερες επιδόσεις ανά πρόβλημα

«It is difficult to make predictions, especially about the future» Niels Bohr



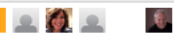
Kashmir Hill, Forbes Staff

Welcome to The Not-So Private Parts where technology & privacy collide

+ Follow (1,551)

TECH | 2/16/2012 @ 11:02AM | 2,211,757 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



325 comments, 170 called-out

+ Comment Now

+ Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. [Target](#), for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the [New York Times](#) how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole —

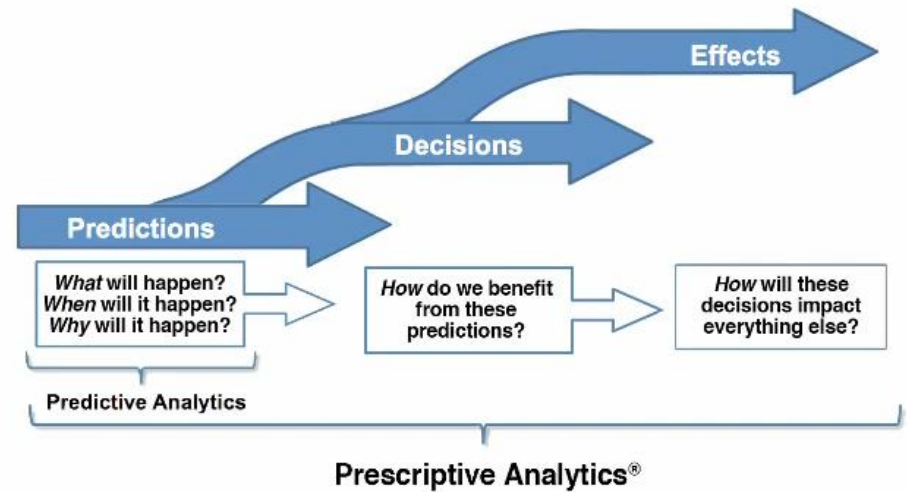
<http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>



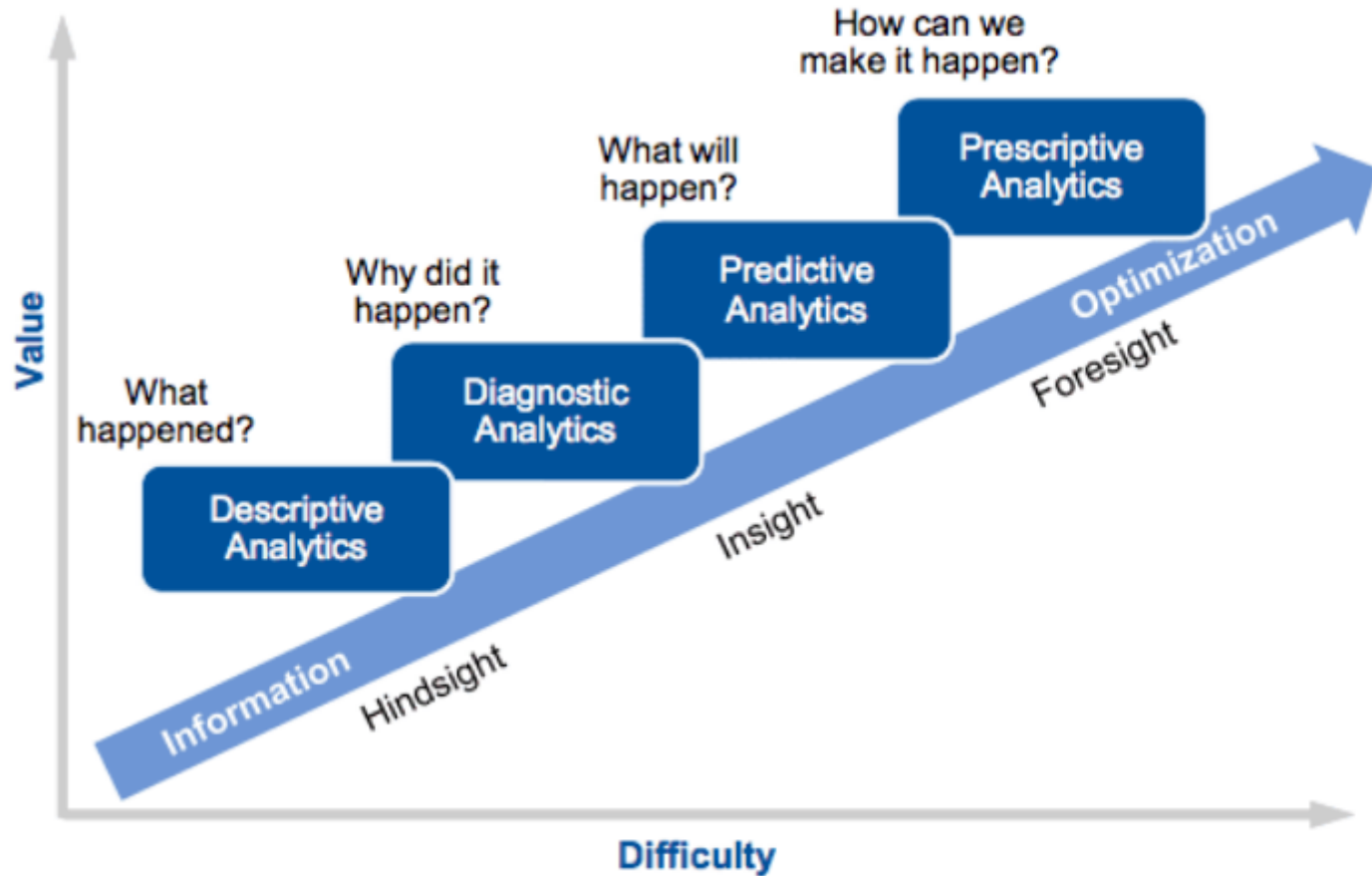
Target has got you in its aim

Prescriptive Analytics (Καθοδηγητική αναλυτική)

- Πως οι προβλέψεις για το μέλλον μπορούν να αλλάξουν τις αποφάσεις που παίρνουμε έτσι ώστε να αποκτήσουμε πλεονέκτημα στο μέλλον;
- Τι παρενέργειες θα έχουν οι αποφάσεις που θα λάβουμε για το ευρύτερο σύστημα;



Gartner Analytic Ascendancy model



Source: Gartner (March 2012)

NETFLIX

The screenshot shows the Netflix Prize announcement page. At the top, the Netflix logo is on the left, and a large yellow banner with the word "COMPLETED" in red letters is on the right. Below the banner, there are navigation links: Home, Rules, Leaderboard, and Update. The main content area features a "Congratulations!" message in a white box with a blue border. The message reads: "The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences. On September 21, 2009 we awarded the \$1M Grand Prize to team 'BellKor's Pragmatic Chaos'. Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum. We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love." The background of the page shows a blurred view of the Netflix website interface with silhouettes of people looking at a screen.

NETFLIX

Netflix Prize

COMPLETED

Home Rules Leaderboard Update

Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their [algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

FAQ | Forum | Netflix Home

© 1997-2009 Netflix, Inc. All rights reserved.

Netflix's New 'My List' Feature Knows You Better Than You Know Yourself (Because Algorithms)

The Huffington Post | By Dino Grandoni
Posted: 08/21/2013 1:44 pm EDT | Updated: 08/22/2013 8:31 am EDT



30 14 7 108

Share Tweet Email Comment

GET TECHNOLOGY NEWSLETTERS:
Enter email

FOLLOW: [Netflix](#), [Netflix My List](#), [Netflix Recommendation](#), [Netflix Recommendations](#), [Technology News](#)

Netflix is throwing off one of the remaining vestiges of its DVD-by-mail business, Instant Queue, and replacing it with a solution that gives even more control to the algorithms that are increasingly driving its subscribers' experience.

http://www.huffingtonpost.com/2013/08/21/netflix-my-list_n_3790472.html



Hadoop και MapReduce

Πλαίσιο λογισμικού για δημιουργία εφαρμογών που επεξεργάζονται τεράστιες ποσότητες δεδομένων παράλληλα σε μεγάλα συμπλέγματα με τρόπο αξιόπιστο και ανεκτικό σε σφάλματα

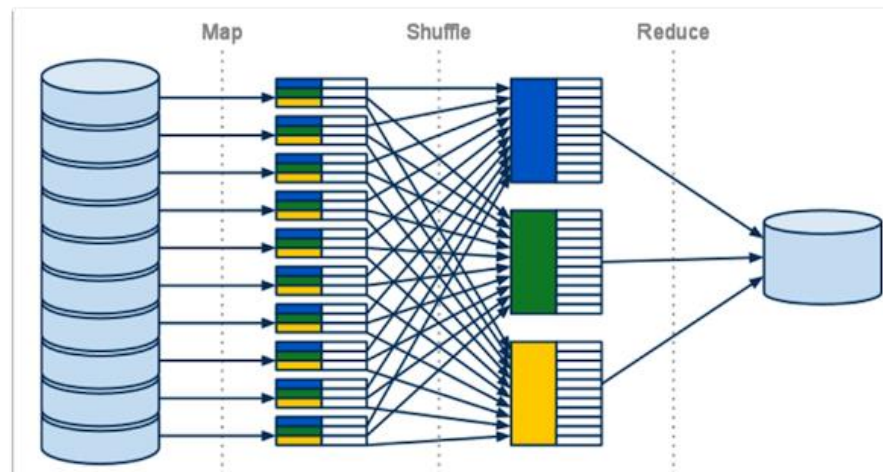
MapReduce

- Το MapReduce είναι ένα υπολογιστικό μοντέλο που χρησιμοποιείται ευρύτατα για αποδοτική κατανεμημένη επεξεργασία πάνω σε μεγάλα σύνολα δεδομένων
- Εκτελείται σε clusters υπολογιστών
- Ο προγραμματιστής χρειάζεται να γράψει 2 συναρτήσεις την συνάρτηση map και την συνάρτηση reduce
- Περιγράφηκε αρχικά σε άρθρο του 2003 (Google)
- Το Hadoop είναι open source υλοποίηση του MapReduce



MapReduce functions

- Τα προβλήματα «σπάνε» σε 2 φάσεις
 - **Map:** τα δεδομένα του προβλήματος διαχωρίζονται σε μη επικαλυπτόμενα τμήματα της μορφής <key, value> και ανατίθενται σε διεργασίες που παράγουν αποτελέσματα επίσης της μορφής <key, value>
 - **Reduce:** τα αποτελέσματα της Map φάσης τροφοδοτούνται σε διεργασίες που τα συνοψίζουν σε μικρότερο αριθμό εγγραφών
- Η συνάρτηση reduce εκτελείται μετά την συνάρτηση map



Παράδειγμα MapReduce: Μέτρηση λέξεων

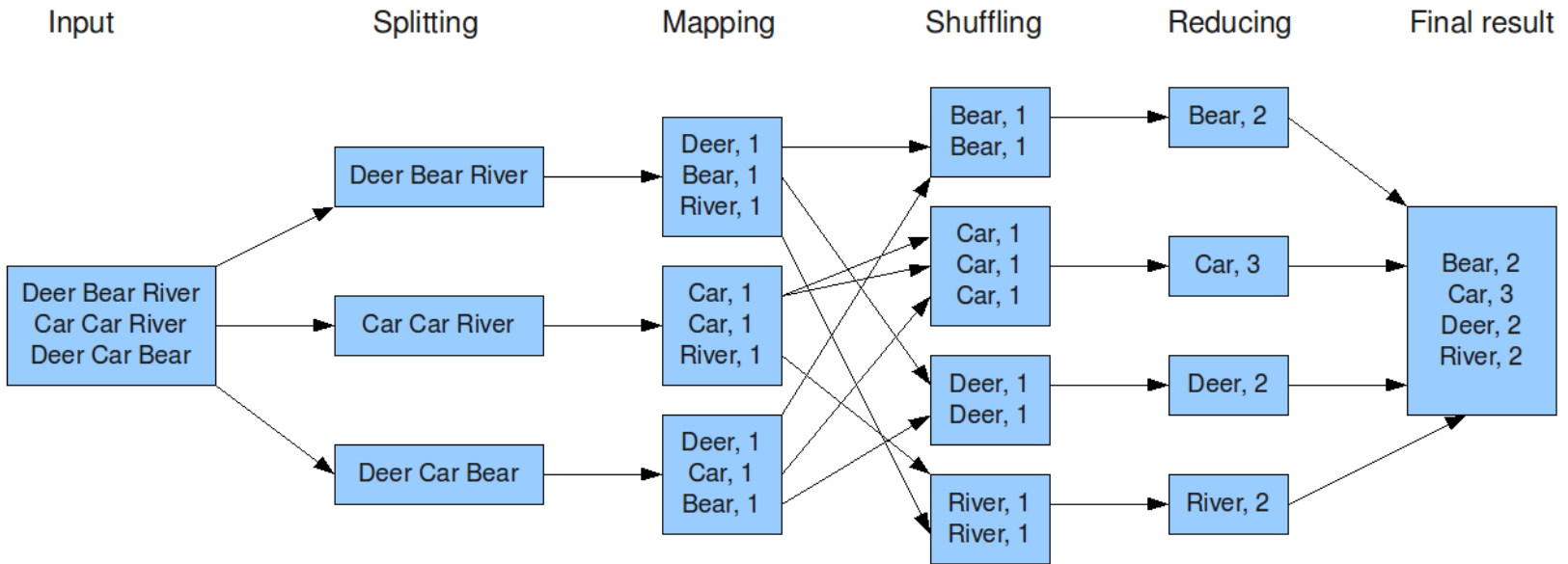
- Πρόβλημα: Υπολογισμός της συχνότητας εμφάνισης λέξεων σε ένα σύνολο πολλών κειμένων
- Θα πρέπει να γραφεί μια map συνάρτηση και μια reduce συνάρτηση

```
//key: όνομα του εγγράφου  
//value: περιεχόμενο του εγγράφου  
map(String key, String value)  
for each w in value  
    emitIntermediate(w,1)
```

```
//key: μια λέξη  
//value: μια λίστα από μετρήσεις  
reduce(key, values)  
c = 0  
for each v in values  
    c += v  
emit(c)
```

Παράδειγμα μέτρησης λέξεων

The overall MapReduce word count process

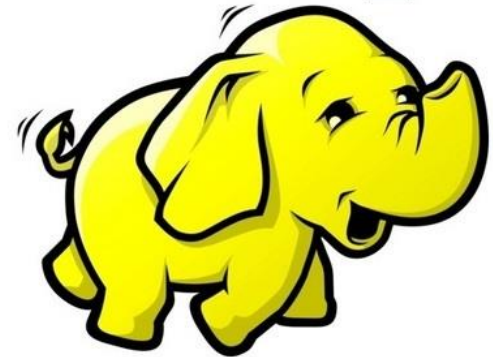


<http://xiaochongzhang.me/blog/?p=338>

Hadoop

- Έχει γραφεί σε Java
- Υποστηρίζει προγραμματισμό σε java αλλά και άλλες γλώσσες προγραμματισμού
- Αποτελείται από 2 υποσυστήματα
 - HDFS
 - MapReduce
- Έχει μεγάλη αποδοχή (Yahoo!, Twitter, Amazon, Facebook κ.α.)

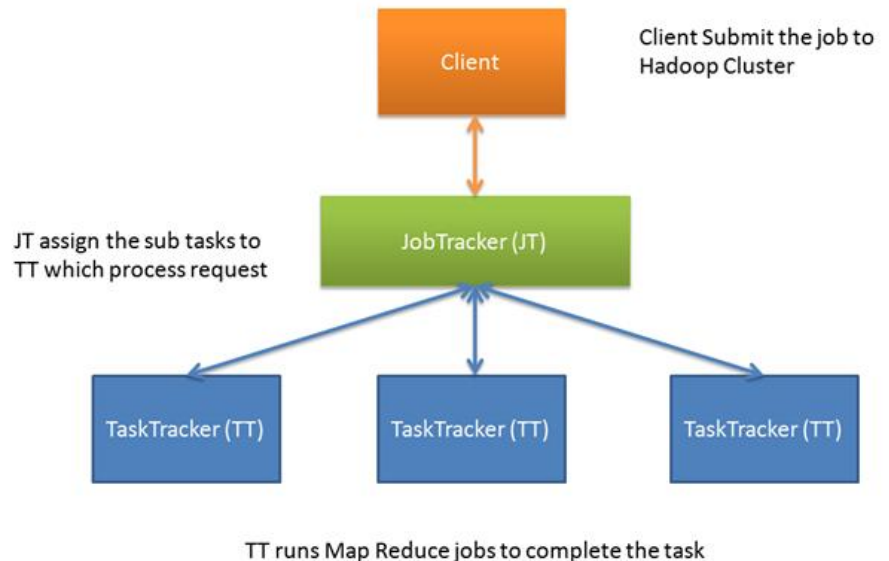
hadoop



<http://hadoop.apache.org/>

Ρόλοι

- **JobTracker:** αναθέτει εργασίες (map ή reduce) στα υπόλοιπα μηχανήματα
- **TaskTracker:** εκτελεί ένα αντίγραφο του προγράμματος MapReduce σε ένα τμήμα των δεδομένων
- Ένας κεντρικός JobTracker διαχειρίζεται πολλούς TaskTrackers



Apache Mahout



Scalable machine learning
and data mining

Apache Mahout has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining



NoSQL

Μηχανισμός για την αποθήκευση και ανάκτηση δεδομένων που διαμορφώνεται με άλλα μέσα από τις σχέσεις πίνακα που χρησιμοποιούνται σε σχεσιακές ΒΔ

Σχεσιακές Βάσεις Δεδομένων

- Τα δεδομένα είναι οργανωμένα σε **πίνακες** με καλά ορισμένες σχέσεις μεταξύ τους
- **ACID**
 - **Atomicity:** Μια συναλλαγή γίνεται είτε στο σύνολό της είτε καθόλου
 - **Consistency:** Κάθε συναλλαγή μεταφέρει την ΒΔ από μια συνεπή κατάσταση σε άλλη συνεπή επίσης κατάσταση
 - **Isolation:** Οι άλλες λειτουργίες δεν μπορούν να προσπελάσουν αλλαγές στα δεδομένα συναλλαγών που δεν έχουν ολοκληρωθεί
 - **Durability:** Η βάση δεδομένων μπορεί να ανακάμψει σε περίπτωση αστοχιών υλικού
- Πλεονεκτήματα
 - Ώριμη τεχνολογία που υποστηρίζει χιλιάδες εφαρμογές σε λειτουργία σήμερα
 - Υψηλές αποδόσεις
 - Πληθώρα εργαλείων
 - Εκπαιδευμένο προσωπικό
- Μειονεκτήματα
 - Κόστος
 - Κλιμάκωση
 - Δύσκολη τροποποίηση της βάσης

Διαδομένες σχεσιακές Βάσεις Δεδομένων

Εμπορικές

- Oracle Database
- IBM DB2
- Microsoft SQL Server
- SAP Sybase SQL Anywhere

Ανοικτού κώδικα

- MySQL
- PostgreSQL
- Apache Derby
- SQLite
- H2
- HSQLDB

NoSQL Βάσεις Δεδομένων

- Οι NoSQL ΒΔ χρησιμοποιούνται συχνά για την αποθήκευση Big Data



- Πλεονεκτήματα
 - Ευκολότερη κλιμάκωση (high scalability)
 - Υψηλές επιδόσεις
 - Αποθήκευση μη δομημένων δεδομένων
- Μειονεκτήματα (features)
 - Weak (eventual) consistency
 - No schema
 - No transactions
 - No SQL

NoSQL landscape

- Key–value stores
 - Redis, Riak
- Column Family Stores
 - Cassandra, HBase
- Document databases
 - MongoDB, CouchDB
- Graph databases
 - Neo4J, Infogrid, HyperGraphDB





**AN SQL QUERY GOES INTO A BAR,
WALKS UP TO TWO TABLES AND ASKS...**



'CAN I JOIN YOU?'





Data Analytics (Αναλυτική Δεδομένων)

Ανάλυση ακατέργαστων δεδομένων για την εξαγωγή συμπερασμάτων,
συχνά με χρήση αυτόματων διαδικασιών και αλγορίθμων

Moneyball

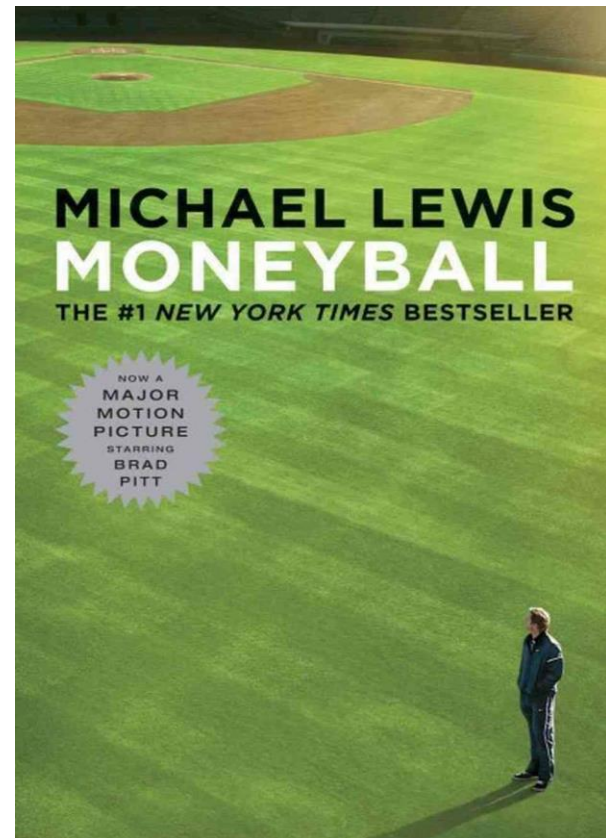
Billy Beane, general manager of MLB's Oakland A's and protagonist of Michael Lewis's *Moneyball* had a problem: **how to win in the Major Leagues with a budget that's smaller than that of nearly every other team.**

Conventional wisdom long held that big name, highly athletic hitters and young pitchers with rocket arms were the ticket to success.

But Beane and his staff, buoyed by massive amounts of carefully **interpreted statistical data**, believed that wins could be had by more affordable methods such as hitters with high on-base percentage and pitchers who get lots of ground outs.

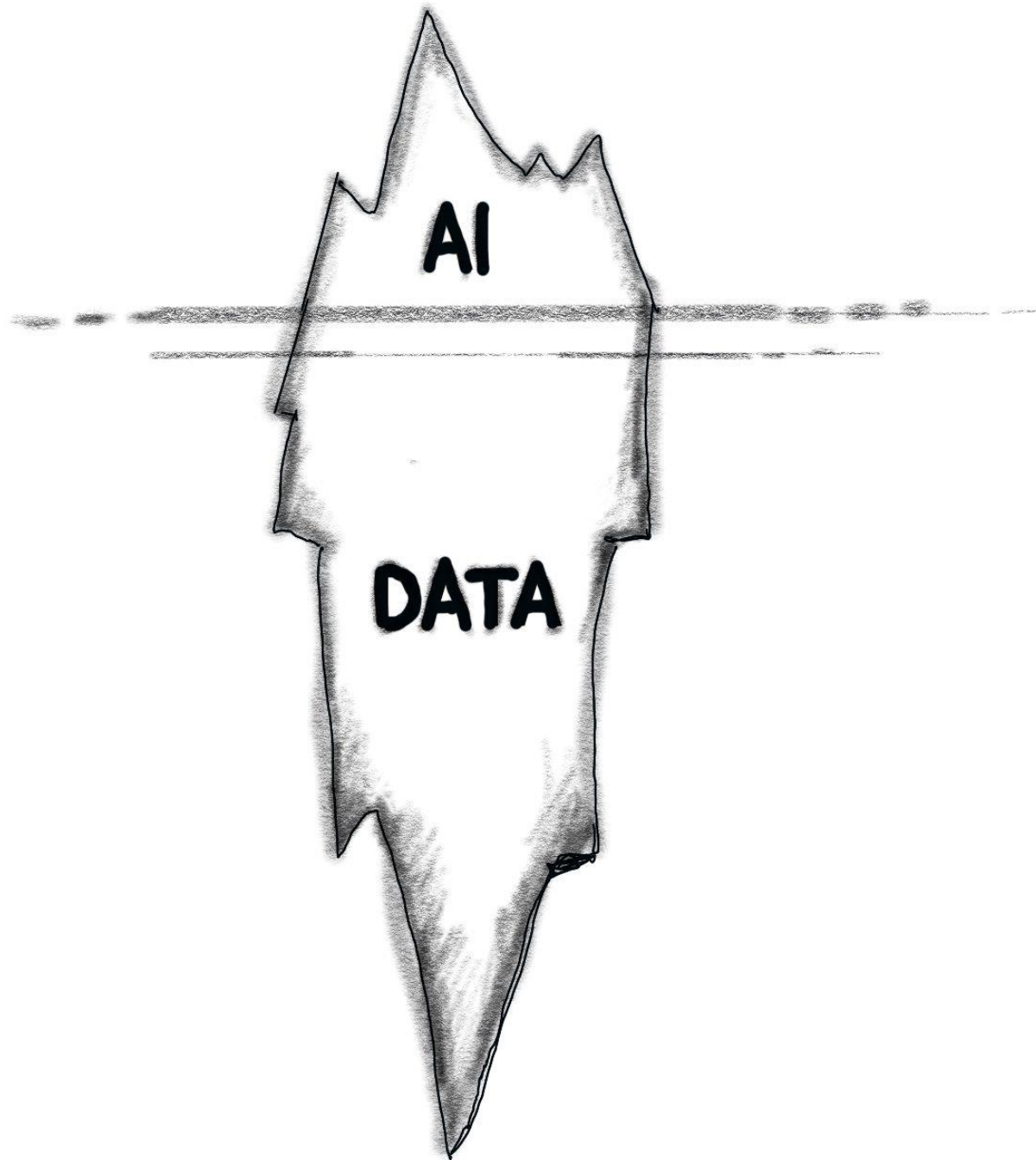
Given this information and a tight budget, Beane defied tradition and his own scouting department to build winning teams of young affordable players and inexpensive cast-off veterans.

<https://www.goodreads.com/book/show/1301.Moneyball>



Data analytics (Data mining)

- Μετασχηματισμός δεδομένων σε κανόνες ή σε πρότυπα (patterns) με νόημα
- Κατηγορίες τεχνικών data analytics
 - *Κατευθυνόμενες τεχνικές:* πρόβλεψη χαρακτηριστικών για ένα συγκεκριμένο στοιχείο
 - *Μη κατευθυνόμενες τεχνικές:* εντοπισμός patterns σε δεδομένα ή ομαδοποίηση των δεδομένων
- Παλινδρόμηση
- Κατηγοριοποίηση
- Συσταδοποίηση



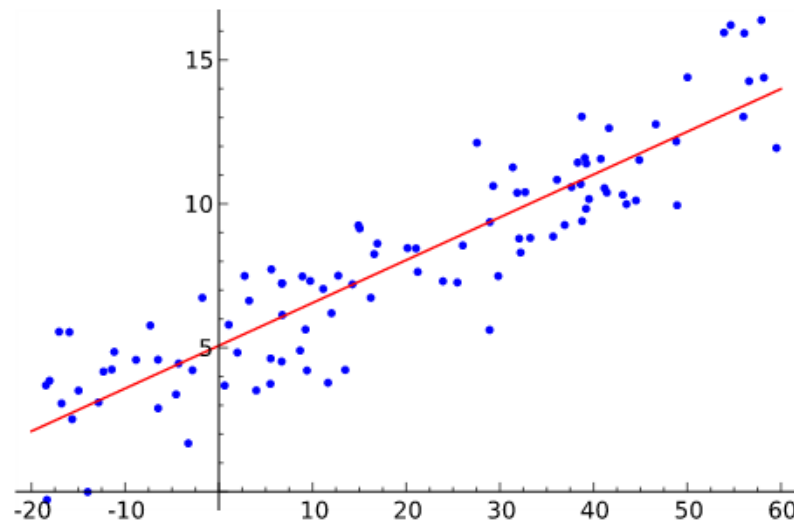
WEKA

- Δυνατότητες προεπεξεργασίας δεδομένων
- 100+ αλγόριθμοι για κατηγοριοποίηση
- 20+ αλγόριθμοι για συσταδοποίηση
- Δυνατότητες οπτικοποίησης δεδομένων και αποτελεσμάτων
- Open source
- Έχει γραφεί σε Java
- Μπορεί να χρησιμοποιηθεί ως callable library



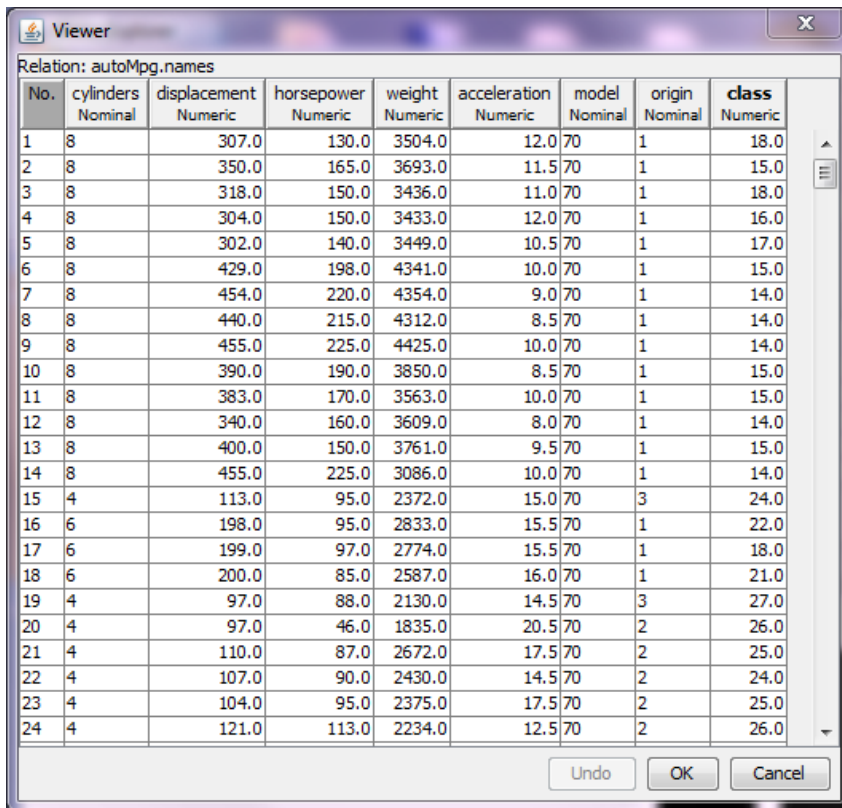
Παλινδρόμηση (Regression)

- Κλασσική στατιστική μέθοδος
- Ένα σύνολο από ανεξάρτητες μεταβλητές παράγουν ένα αποτέλεσμα (εξαρτημένη μεταβλητή)
- Προβλέπει την τιμή της εξαρτημένης μεταβλητής βάσει των τιμών των ανεξάρτητων μεταβλητών



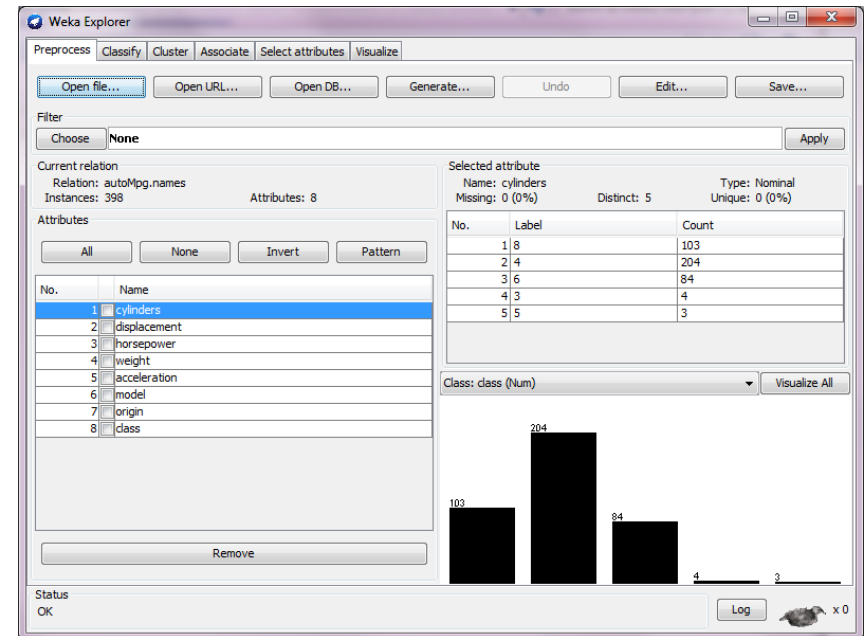
Παράδειγμα με Παλινδρόμηση

class = απόσταση σε μίλια ανά γαλόνι καυσίμου



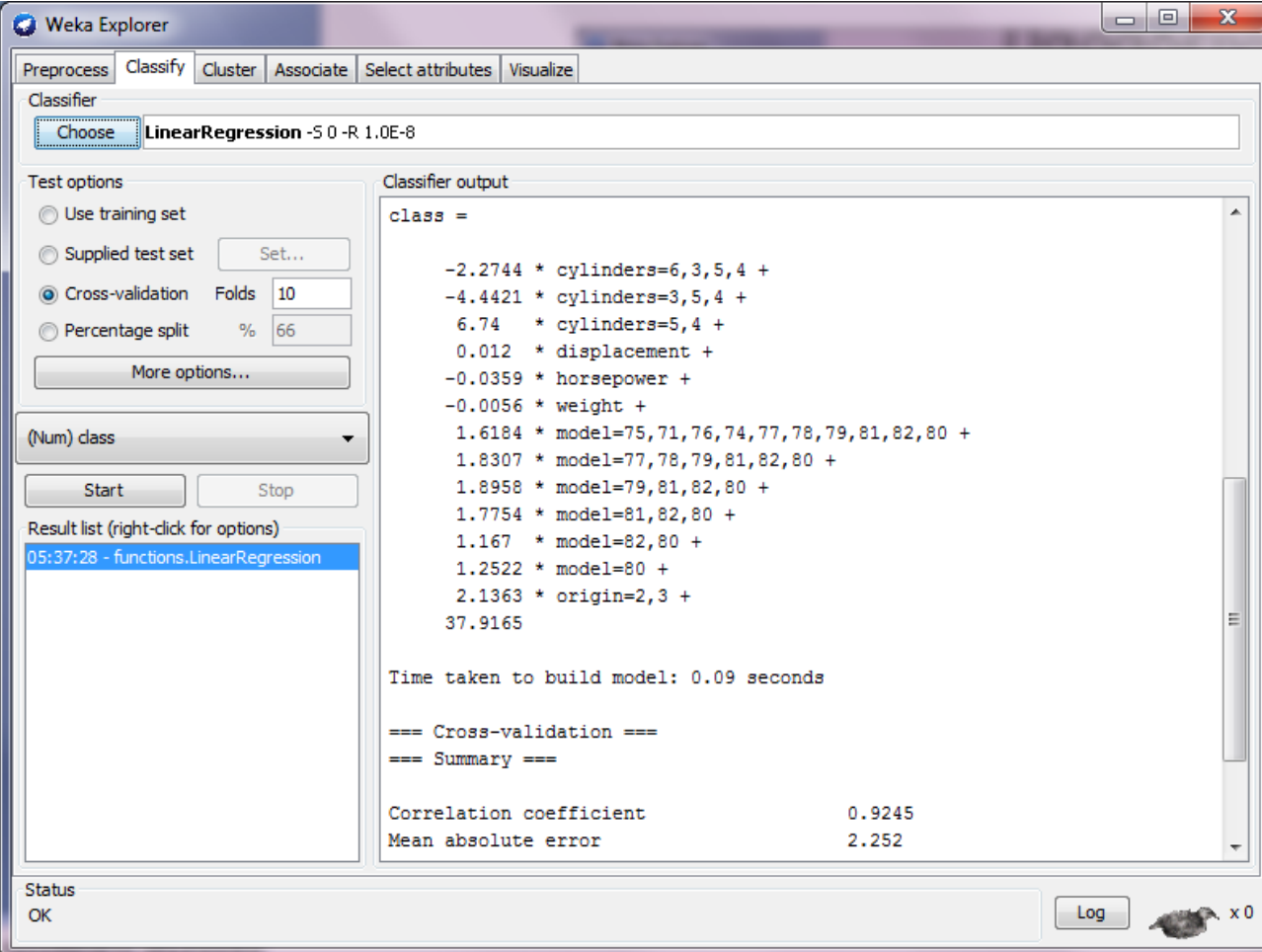
Relation: autoMpg.names

No.	cylinders Nominal	displacement Numeric	horsepower Numeric	weight Numeric	acceleration Numeric	model Nominal	origin Nominal	class Numeric
1	8	307.0	130.0	3504.0	12.0	70	1	18.0
2	8	350.0	165.0	3693.0	11.5	70	1	15.0
3	8	318.0	150.0	3436.0	11.0	70	1	18.0
4	8	304.0	150.0	3433.0	12.0	70	1	16.0
5	8	302.0	140.0	3449.0	10.5	70	1	17.0
6	8	429.0	198.0	4341.0	10.0	70	1	15.0
7	8	454.0	220.0	4354.0	9.0	70	1	14.0
8	8	440.0	215.0	4312.0	8.5	70	1	14.0
9	8	455.0	225.0	4425.0	10.0	70	1	14.0
10	8	390.0	190.0	3850.0	8.5	70	1	15.0
11	8	383.0	170.0	3563.0	10.0	70	1	15.0
12	8	340.0	160.0	3609.0	8.0	70	1	14.0
13	8	400.0	150.0	3761.0	9.5	70	1	15.0
14	8	455.0	225.0	3086.0	10.0	70	1	14.0
15	4	113.0	95.0	2372.0	15.0	70	3	24.0
16	6	198.0	95.0	2833.0	15.5	70	1	22.0
17	6	199.0	97.0	2774.0	15.5	70	1	18.0
18	6	200.0	85.0	2587.0	16.0	70	1	21.0
19	4	97.0	88.0	2130.0	14.5	70	3	27.0
20	4	97.0	46.0	1835.0	20.5	70	2	26.0
21	4	110.0	87.0	2672.0	17.5	70	2	25.0
22	4	107.0	90.0	2430.0	14.5	70	2	24.0
23	4	104.0	95.0	2375.0	17.5	70	2	25.0
24	4	121.0	113.0	2234.0	12.5	70	2	26.0



<http://www.ibm.com/developerworks/library/os-weka1/>

Παράδειγμα με Παλινδρόμηση (αποτελέσματα)



The screenshot displays the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'LinearRegression -S 0 -R 1.0E-8'. The 'Test options' section shows 'Cross-validation' selected with 10 folds and 66% split. The 'Classifier output' pane shows the following regression equation:

```
class =  
-2.2744 * cylinders=6,3,5,4 +  
-4.4421 * cylinders=3,5,4 +  
6.74 * cylinders=5,4 +  
0.012 * displacement +  
-0.0359 * horsepower +  
-0.0056 * weight +  
1.6184 * model=75,71,76,74,77,78,79,81,82,80 +  
1.8307 * model=77,78,79,81,82,80 +  
1.8958 * model=79,81,82,80 +  
1.7754 * model=81,82,80 +  
1.167 * model=82,80 +  
1.2522 * model=80 +  
2.1363 * origin=2,3 +  
37.9165
```

Time taken to build model: 0.09 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.9245
Mean absolute error	2.252

The 'Result list' shows a single entry: '05:37:28 - functions.LinearRegression'. The status bar at the bottom indicates 'OK' and a 'Log' button.

Αλγόριθμοι κατηγοριοποίησης

Pages in category "Classification algorithms"

The following 85 pages are in this category, out of 85 total. [This list may not reflect recent changes.](#)

(

- [\(1+ε\)-approximate nearest neighbor search](#)

*

- [Calibration \(statistics\)](#)
- [Soft independent modelling of class analogies](#)
- [Statistical classification](#)
- [Variable kernel density estimation](#)

A

- [AdaBoost](#)
- [ALOPEX](#)
- [Alternating decision tree](#)
- [Analogical modeling](#)
- [Averaged one-dependence estimators](#)
- [Artificial neural network](#)
- [Types of artificial neural networks](#)
- [Automated Pain Recognition](#)

B

- [Boosting \(machine learning\)](#)
- [BrownBoost](#)

C

- [C4.5 algorithm](#)
- [Cascading classifiers](#)
- [Case-based reasoning](#)
- [Chi-square automatic interaction detection](#)
- [Classifier chains](#)
- [Co-training](#)
- [CoBoosting](#)
- [Compositional pattern-producing network](#)
- [Conceptual clustering](#)

D

- [Decision boundary](#)
- [Decision tree learning](#)
- [Deductive classifier](#)

E

- [Elastic matching](#)
- [Evolving classification function](#)

F

- [Feature Selection Toolbox](#)

G

- [Generalization error](#)
- [Gesture Description Language](#)
- [Gradient boosting](#)
- [Group method of data handling](#)

H

- [Hierarchical classification](#)
- [Hyper basis function network](#)

I

- [ID3 algorithm](#)
- [IDistance](#)
- [Information gain \(decision tree\)](#)
- [Information gain ratio](#)

J

- [Jackknife variance estimates for random forest](#)

K

- [K-nearest neighbors algorithm](#)
- [Kernel method](#)

L

- [Large margin nearest neighbor](#)
- [Latent class model](#)
- [Learning vector quantization](#)
- [Least-squares support vector machine](#)
- [Linear classifier](#)
- [Linear discriminant analysis](#)
- [Locality-sensitive hashing](#)
- [Logic learning machine](#)
- [LogitBoost](#)

M

- [Margin classifier](#)
- [Margin-infused relaxed algorithm](#)
- [Mathematics of artificial neural networks](#)
- [Multi-label classification](#)
- [Multiclass classification](#)

- [Multifactor dimensionality reduction](#)
- [Multilayer perceptron](#)
- [Multinomial logistic regression](#)
- [Multiple discriminant analysis](#)
- [Multispectral pattern recognition](#)

N

- [Naive Bayes classifier](#)
- [Nearest centroid classifier](#)
- [Nearest neighbor search](#)
- [Normal discriminant analysis](#)

O

- [One-class classification](#)
- [Operational taxonomic unit](#)
- [Optimal discriminant analysis and classification tree analysis](#)
- [Ordinal regression](#)

P

- [Perceptron](#)
- [Probabilistic latent semantic analysis](#)
- [Probit model](#)

Q

- [Quadratic classifier](#)

R

- [Radial basis function network](#)
- [Random forest](#)
- [Random subspace method](#)
- [Relevance vector machine](#)
- [Rules extraction system family](#)

S

- [Support vector machine](#)
- [Syntactic pattern recognition](#)

T

- [Textual case-based reasoning](#)
- [Tsetlin machine](#)

W

- [Whitening transformation](#)
- [Winnow \(algorithm\)](#)

http://en.wikipedia.org/wiki/Category:Classification_algorithms

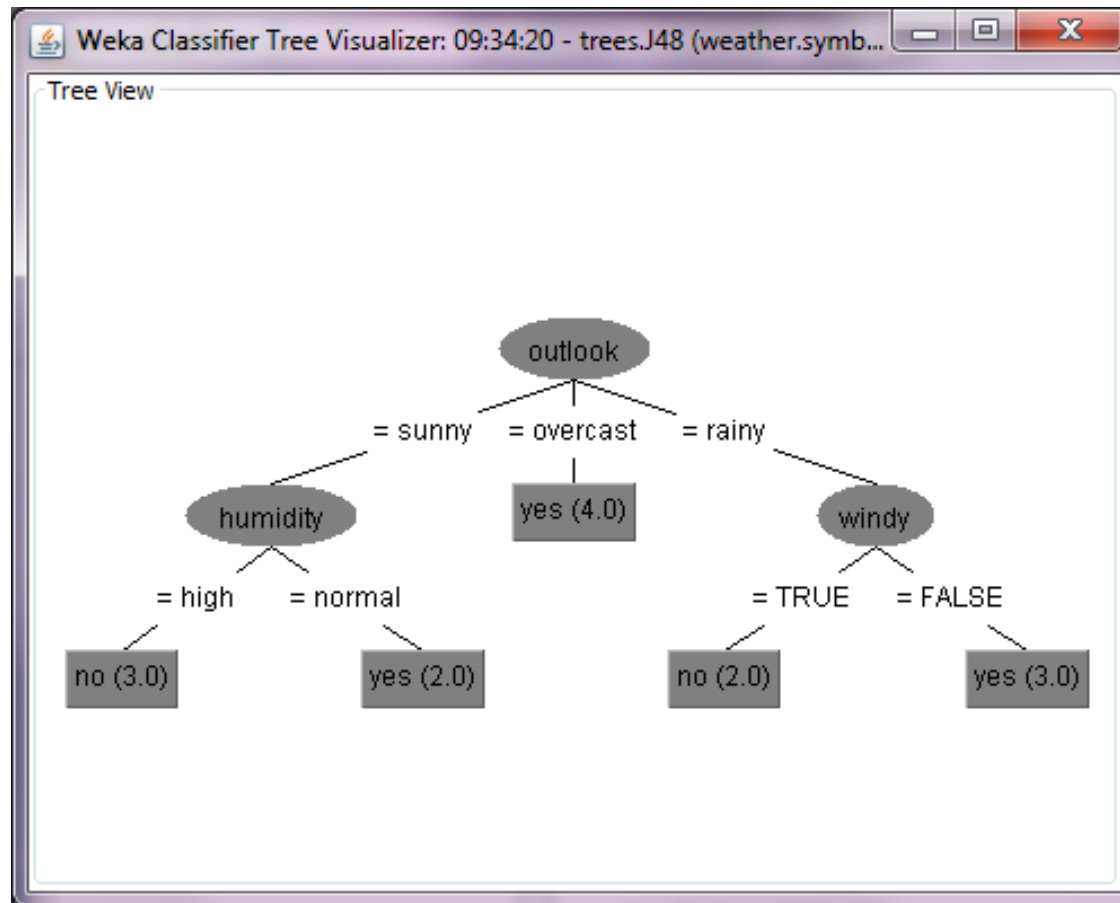
Παράδειγμα με κατηγοριοποίηση (Δένδρα Απόφασης)

- Βήματα δημιουργίας δένδρου απόφασης:
 1. Επιλογή μεταβλητής ως ρίζα
 2. Δημιουργία διακλάδωσης για κάθε πιθανή τιμή της μεταβλητής
 3. Διαχωρισμός των δεδομένων σε υποσύνολα που ανατίθενται σε κάθε κλάδο του δένδρου
 4. Επανάληψη βημάτων 2 και 3 και 4 αναδρομικά
- Κρίσιμη απόφαση: ποιες μεταβλητές θα προηγηθούν
- Αλγόριθμος C4.5 (Weka J48)

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

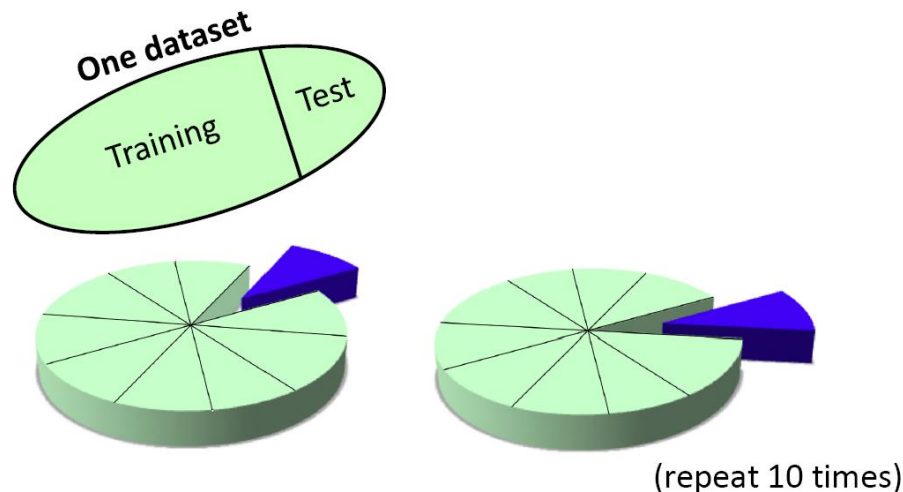
Δένδρο Απόφασης

Τα δένδρα απόφασης γίνονται εύκολα κατανοητά



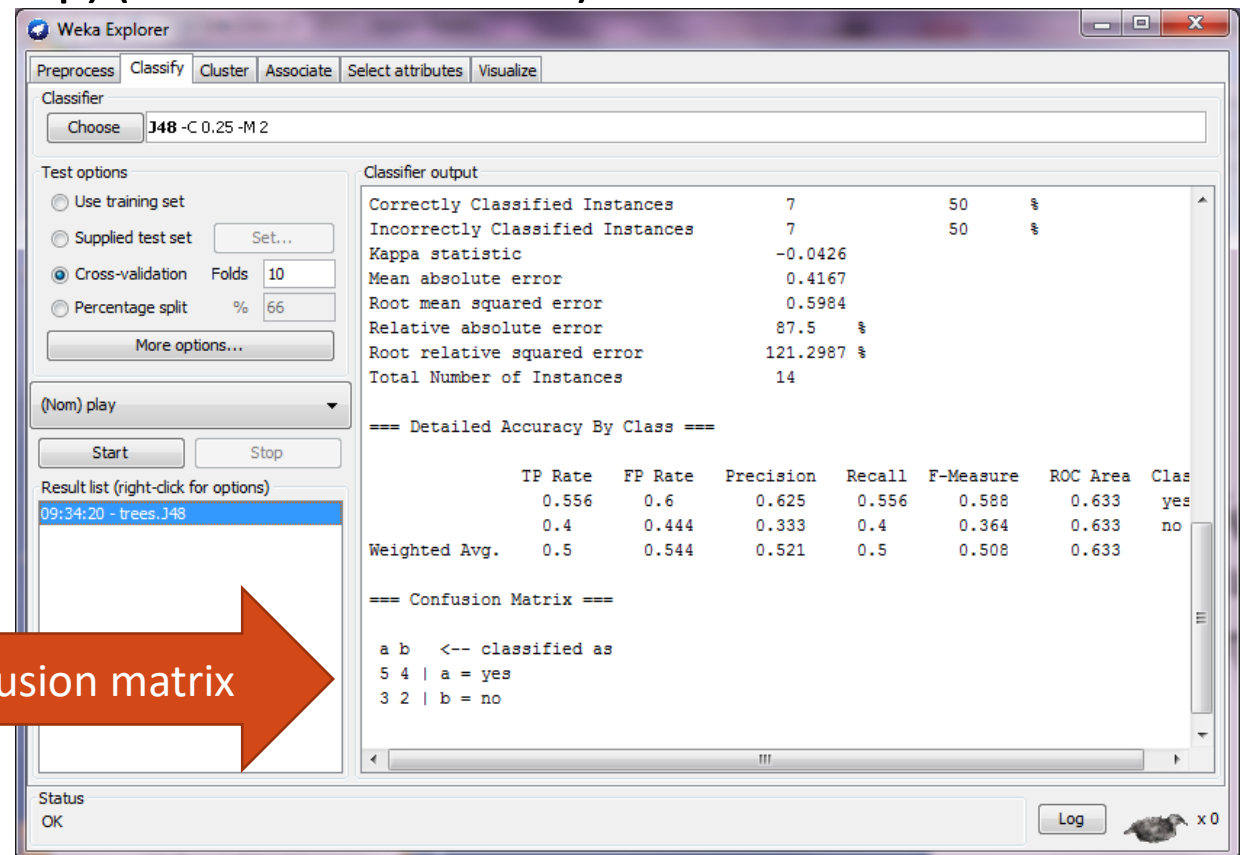
Cross-validation

- Τα δεδομένα πρέπει να χωρίζονται σε training και test έτσι ώστε να εκτιμηθεί το **overfitting**
- Το Weka υποστηρίζει n-fold cross validation
- 10-fold cross validation
 - Τα δεδομένα χωρίζονται σε 10 τμήματα (folds)
 - Με την σειρά επιλέγεται 1-fold
 - Εκτελείται ο αλγόριθμος
 - Υπολογίζεται ο μέσος όρος όλων των αποτελεσμάτων



Δένδρο Απόφασης (αποτελέσματα)

- Σωστά ταξινομημένες περιπτώσεις
- Εσφαλμένα ταξινομημένες περιπτώσεις
- Πίνακας σύγχυσης (Confusion matrix)



The screenshot shows the Weka Explorer interface with the Classifier tab selected. The classifier used is J48 -C 0.25 -M 2. The test options are set to Cross-validation with 10 folds. The classifier output is displayed in a table format, showing various performance metrics. A detailed accuracy by class table is also shown, along with a confusion matrix. An orange arrow points from the text 'Confusion matrix' to the confusion matrix section of the output.

Classifier output

Correctly Classified Instances	7	50	%
Incorrectly Classified Instances	7	50	%
Kappa statistic	-0.0426		
Mean absolute error	0.4167		
Root mean squared error	0.5984		
Relative absolute error	87.5	%	
Root relative squared error	121.2987	%	
Total Number of Instances	14		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
yes	0.556	0.6	0.625	0.556	0.588	0.633	yes
no	0.4	0.444	0.333	0.4	0.364	0.633	no
Weighted Avg.	0.5	0.544	0.521	0.5	0.508	0.633	

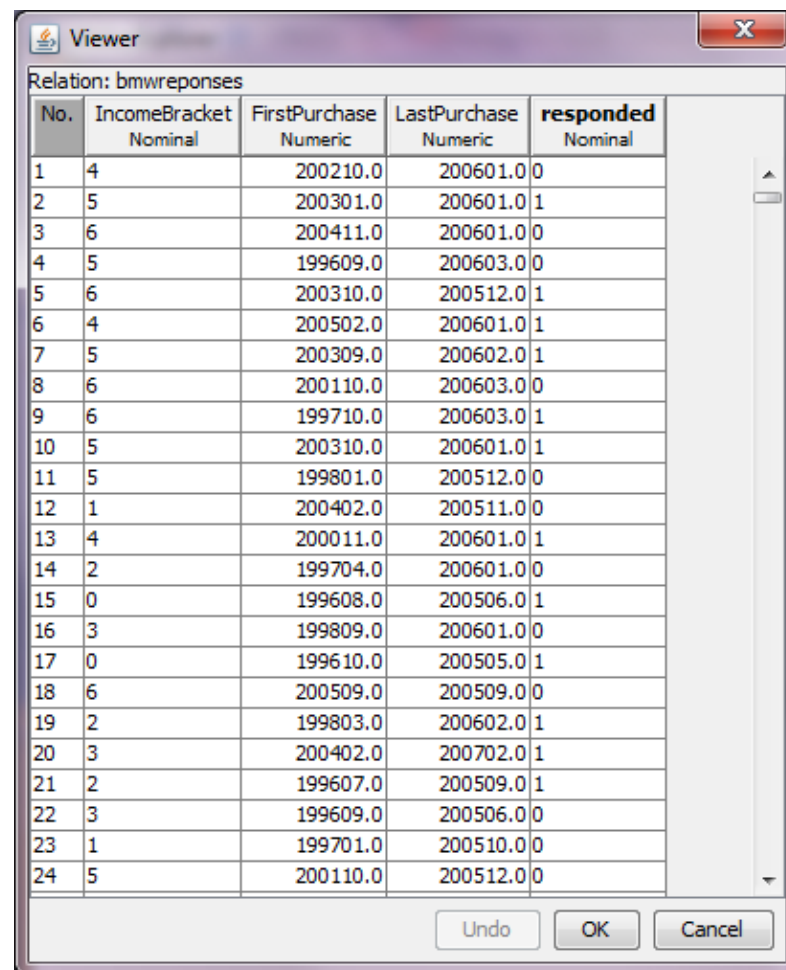
=== Confusion Matrix ===

```
a b <-- classified as
5 4 | a = yes
3 2 | b = no
```

Confusion matrix

Παράδειγμα με κατηγοριοποίηση (BMW προώθηση εγγύησης)

- Εισοδηματική κατηγορία
 - 0=\$0-\$30k
 - 1=\$31k-\$40k,
 - 2=\$41k-\$60k,
 - 3=\$61k-\$75k,
 - 4=\$76k-\$100k,
 - 5=\$101k-\$150k,
 - 6=\$151k-\$500k,
 - 7=\$501k+
- Έτος/μήνας πρώτης αγοράς BMW
- Έτος/μήνας πιο πρόσφατης αγοράς BMW
- Εάν ανταποκρίθηκαν στην προσφορά εκτεταμένης εγγύησης στο παρελθόν



Relation: bmwreponses

No.	IncomeBracket Nominal	FirstPurchase Numeric	LastPurchase Numeric	responded Nominal
1	4	200210.0	200601.0	0
2	5	200301.0	200601.0	1
3	6	200411.0	200601.0	0
4	5	199609.0	200603.0	0
5	6	200310.0	200512.0	1
6	4	200502.0	200601.0	1
7	5	200309.0	200602.0	1
8	6	200110.0	200603.0	0
9	6	199710.0	200603.0	1
10	5	200310.0	200601.0	1
11	5	199801.0	200512.0	0
12	1	200402.0	200511.0	0
13	4	200011.0	200601.0	1
14	2	199704.0	200601.0	0
15	0	199608.0	200506.0	1
16	3	199809.0	200601.0	0
17	0	199610.0	200505.0	1
18	6	200509.0	200509.0	0
19	2	199803.0	200602.0	1
20	3	200402.0	200702.0	1
21	2	199607.0	200509.0	1
22	3	199609.0	200506.0	0
23	1	199701.0	200510.0	0
24	5	200110.0	200512.0	0

Αποτελέσματα με τον C4.5 (Weka J48)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) responded

Start Stop

Result list (right-click for options)

- 10:30:50 - trees.J48
- 10:37:08 - lazy.IBk

Classifier output

Correctly Classified Instances 1642 54.7333 %

Incorrectly Classified Instances 1358 45.2667 %

Kappa statistic 0.0933

Mean absolute error 0.49

Root mean squared error 0.5038

Relative absolute error 98.0236 %

Root relative squared error 100.7747 %

Total Number of Instances 3000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.592	0.499	0.551	0.592	0.571	0.55
	0.501	0.408	0.543	0.501	0.521	0.55
Weighted Avg.	0.547	0.454	0.547	0.547	0.546	0.55

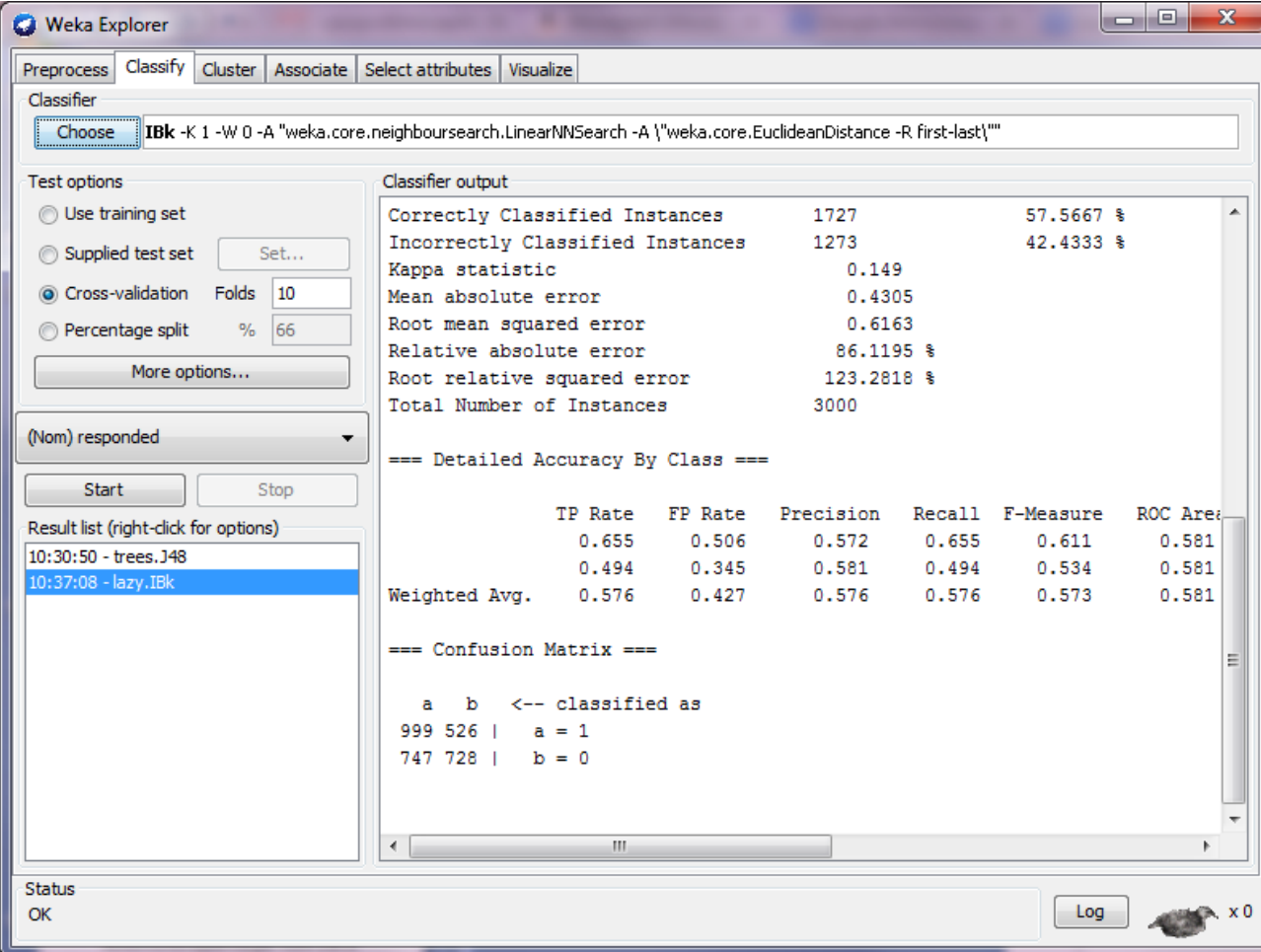
=== Confusion Matrix ===

```
a b <-- classified as
903 622 | a = 1
736 739 | b = 0
```

Status OK

Log x 0

Αποτελέσματα με τον k-nearest neighbor (K-NN)



The screenshot shows the Weka Explorer interface with the Classifier tab selected. The classifier is set to IBk with parameters: `-K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A {}weka.core.EuclideanDistance -R first-last"`. The test options are set to Cross-validation with 10 folds. The classifier output is displayed in a text area, showing overall performance metrics and a detailed accuracy table by class.

Classifier output

Correctly Classified Instances	1727	57.5667 %
Incorrectly Classified Instances	1273	42.4333 %
Kappa statistic	0.149	
Mean absolute error	0.4305	
Root mean squared error	0.6163	
Relative absolute error	86.1195 %	
Root relative squared error	123.2818 %	
Total Number of Instances	3000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.655	0.506	0.572	0.655	0.611	0.581
	0.494	0.345	0.581	0.494	0.534	0.581
Weighted Avg.	0.576	0.427	0.576	0.576	0.573	0.581

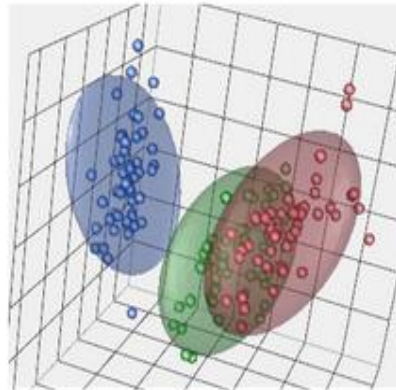
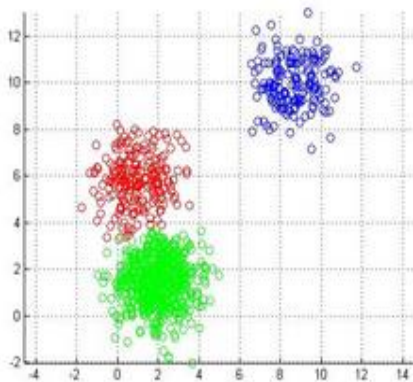
=== Confusion Matrix ===

a	b	<-- classified as	
999	526		a = 1
747	728		b = 0

The interface also shows a result list with two entries: "10:30:50 - trees.J48" and "10:37:08 - lazy.IBk" (highlighted). The status bar at the bottom indicates "OK" and has a "Log" button.

Συσταδοποίηση (Clustering)

- Ομαδοποιεί τα δεδομένα σε clusters εντοπίζοντας πρότυπα που μπορεί να είναι κρυμμένα
- Ο χρήστης θα πρέπει να ορίσει τον αριθμό των clusters εκ των προτέρων



- Αλγόριθμος k-means
 1. Τα δεδομένα κανονικοποιούνται
 2. Επιλέγονται k τυχαίες εγγραφές ως επίκεντρα των clusters
 3. Υπολογίζεται η απόσταση κάθε εγγραφής από τα επίκεντρα και ανατίθεται στο cluster που είναι πλησιέστερα
 4. Υπολογίζεται το νέο επίκεντρο ως μέσος όρος των εγγραφών του cluster
 5. Επαναλαμβάνονται τα βήματα 3 και 4 μέχρι να σταθεροποιηθούν τα επίκεντρα

Παράδειγμα με Συσταδοποίηση (BMW επισκέπτης → πελάτης)

Viewer

Relation: car-browsers

No.	Dealership Numeric	Showroom Numeric	ComputerSearch Numeric	M5 Numeric	3Series Numeric	Z4 Numeric	Financing Numeric	Purchase Numeric
1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0
3	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0
5	1.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0
6	1.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0
7	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
8	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0
9	1.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0
10	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
11	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0
12	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0
13	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0
14	1.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0
15	1.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0
16	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0
17	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
18	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
19	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0
20	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
21	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0
22	1.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0
23	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0
24	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0
25	1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0
26	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0
27	1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0
28	1.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0
29	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Cluster

Choose SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode

Use training set

Supplied test set Set...

Percentage split % 66

Classes to clusters evaluation

(Num) Purchase

Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

11:27:52 - SimpleKMeans

Cluster output

	ComputerSearch	M5	3Series	Z4	Financing	Purchase
ComputerSearch	0.43	0.6538	0	1		
M5	0.53	0.4615	0.963	1		
3Series	0.55	0.3846	0.4444	0.8		
Z4	0.45	0.5385	0	0.8		
Financing	0.61	0.4615	0.6296	0.8		
Purchase	0.39	0	0.5185	0.4		

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

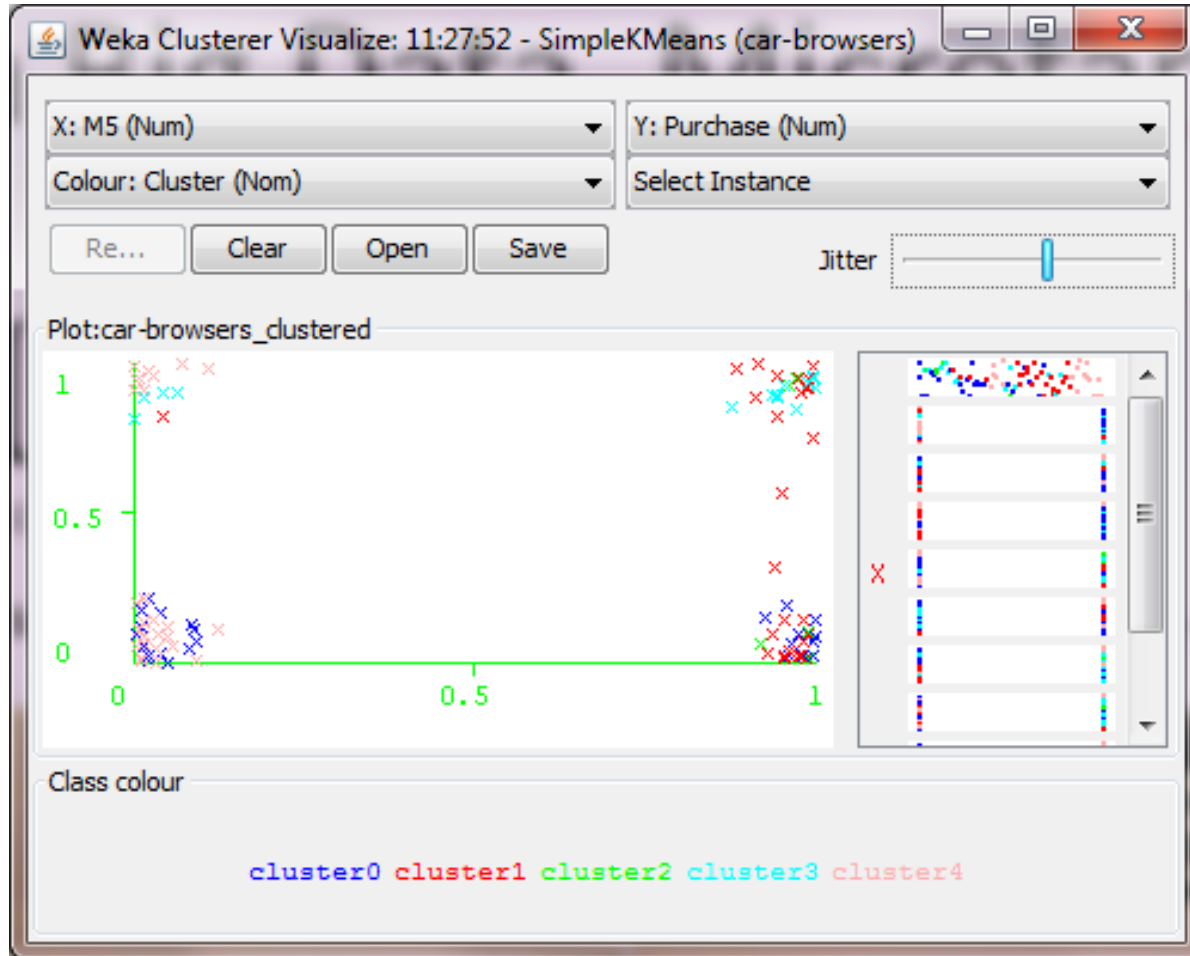
0	26 (26%)
1	27 (27%)
2	5 (5%)
3	14 (14%)
4	28 (28%)

Status OK

Log x 0

SimpleKMeans

Οπτικοποίηση αποτελεσμάτων





Λογισμικά για Data Science

Εργαλεία για Data Science

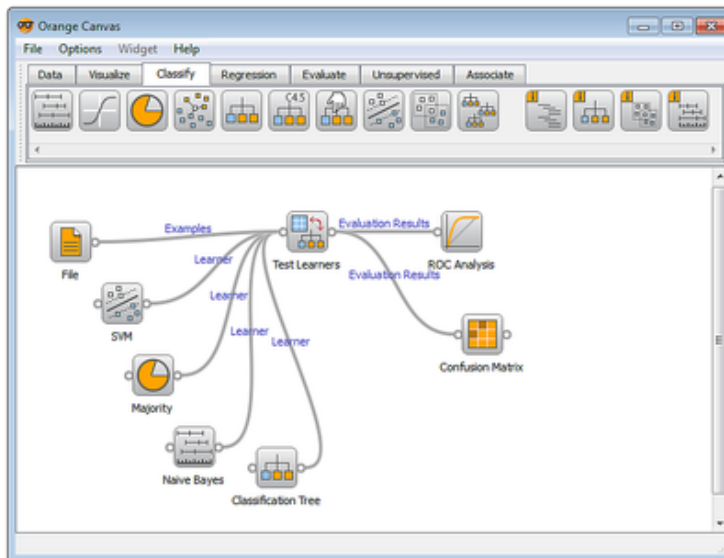
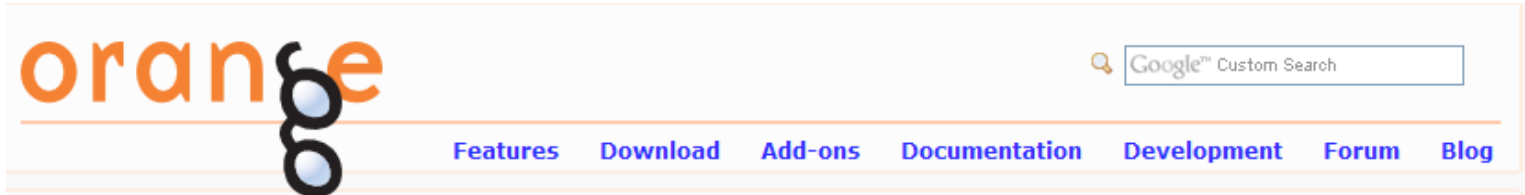
Open Source

- R
- Python SciPy/NumPy
 - Orange
 - Scikit-learn
- Java Weka
- Matlab clones
 - Scilab
 - Octave
- Processing (Visualization)
- Gephi

Commercial

- SAS
- IBM SPSS
- Matlab
- SAP HANA
- Splunk

Python Orange



Open source data visualization and analysis for novice and experts. Data mining through visual programming or Python scripting. Components for machine learning. Add-ons for bioinformatics and text mining. Packed with features for data analytics.



[\(Downloads for other systems and versions\)](#)

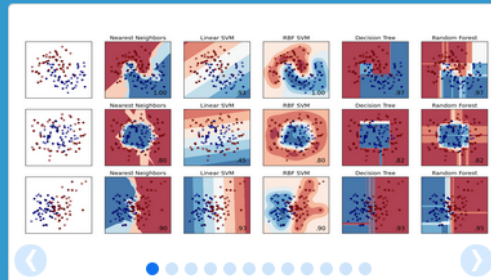
Python scikit-learn



Home Installation Documentation ▾ Examples

Google™ Custom Search

Search x



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.

Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: K-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: PCA, Isomap, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

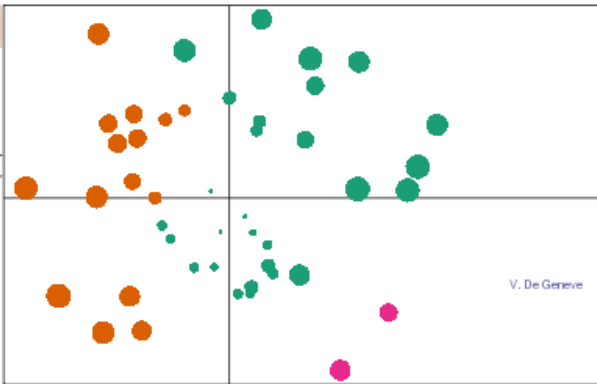
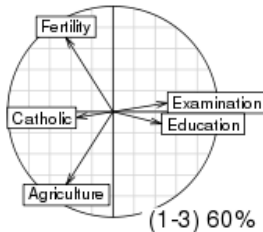
Application: Transforming input data such as text for use with machine learning algorithms.

Modules: preprocessing, feature extraction. — Examples

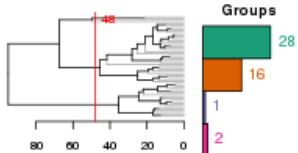


The R Project for Statistical Computing

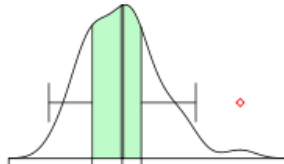
PCA 5 vars
`prcomp(x = data, cor = cor)`



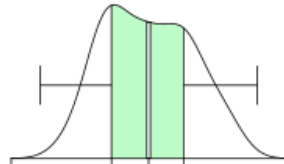
Clustering 4 groups



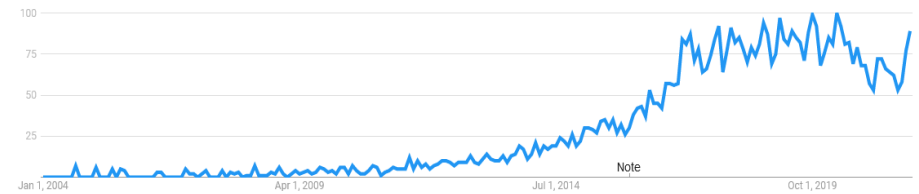
Factor 1 [41%]



Factor 3 [19%]



Interest over time ?



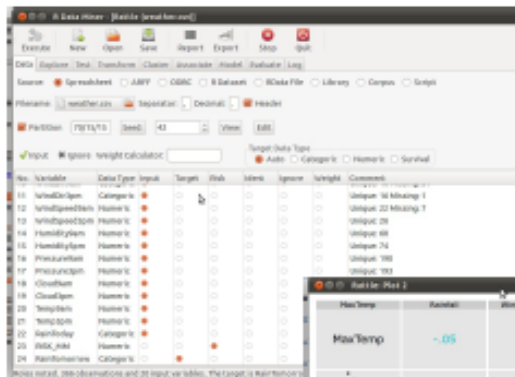
R Rattle

Rattle



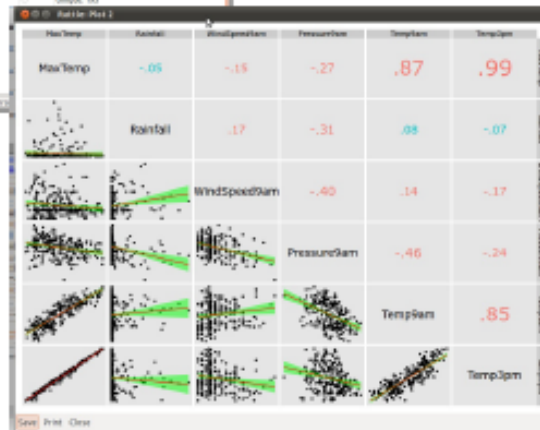
Data Mining using R

Rattle is the R Analytical Tool To Learn Easily, building on the strength of the worlds most powerful statistical software environment, R.



See your data in new and interesting ways.

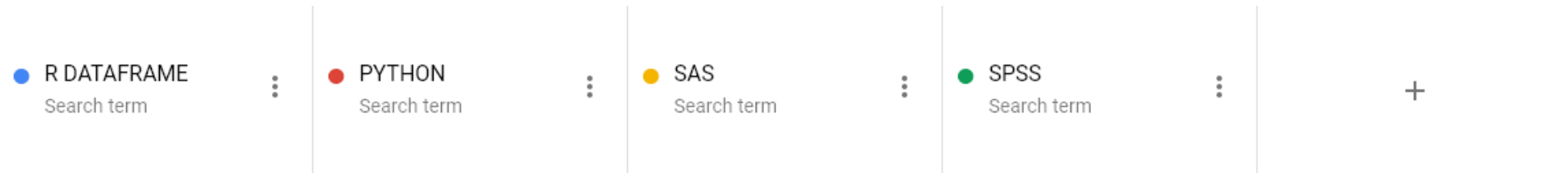
Then build a data mining model in just 4 clicks of the mouse button



Rattle presents statistical and stunning visual summaries of data, with 2-D and 3-D interactive explorations. All of this is built on top of the interationally recognised best-of-breed graphics capabilities of R.

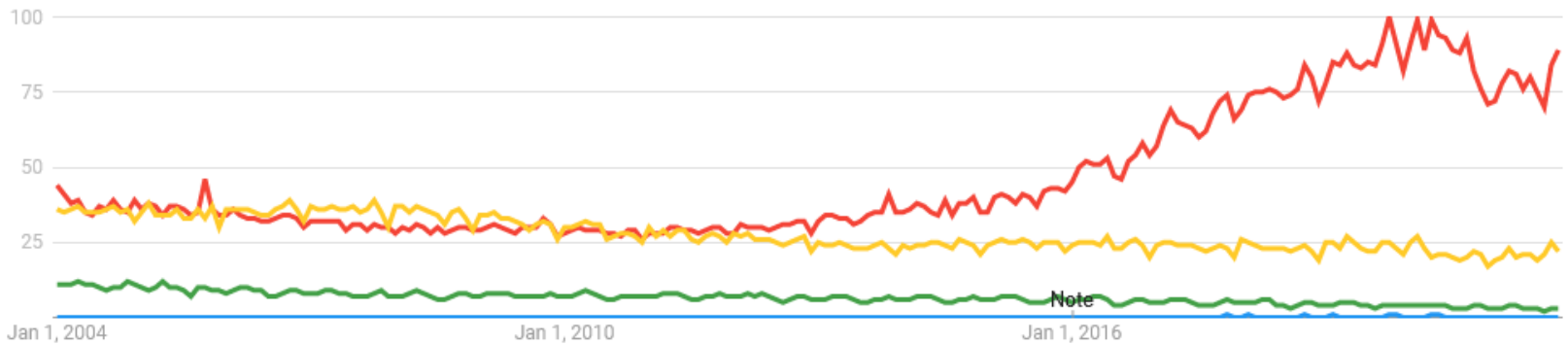
Data can be transformed into the shapes required for modelling.

R, Python, SAS, SPSS



Worldwide ▾ 2004 - present ▾ All categories ▾ Web Search ▾

Interest over time ?





Χρήσιμοι Πόροι

Πηγες Δεδομένων

- Google public data explorer
 - <https://www.google.com/publicdata/directory>
- UCI Machine Learning Repository
 - <http://archive.ics.uci.edu/ml/>
- PublicData.eu
 - <http://publicdata.eu/no/>
- Open Government Data
 - <http://www.data.gov/>
- NYC Open Data
 - <https://data.cityofnewyork.us/>
- Open Source Sports
 - <http://www.opensourcesports.com/>
- GapMinder
 - <http://www.gapminder.org/data/>
- Quandl - Intelligent Search for Numerical Data
 - <http://www.quandl.com/>
- Infochips
 - <http://www.infochimps.com/marketplace>

Kaggle

The image shows the top navigation bar of the Kaggle website. It includes the Kaggle logo on the left, followed by menu items: 'Customer Solutions', 'Competitions', and 'Community'. On the right side of the navigation bar are 'Sign Up' and 'Login' buttons. Below the navigation bar is a large orange banner with the text: 'Welcome to Kaggle, the leading platform for predictive modeling competitions. Here's how to jump into competing on Kaggle—'. To the right of this text are three columns with icons and titles: 'Enter' (with a magnifying glass icon), 'Build' (with a gear icon), and '...Win!' (with a star icon). Each column has a short paragraph of text explaining the step. At the bottom left of the banner, there is a link: 'New to Data Science? Visit our Wiki > Learn about hosting a competition > in-Class & Research competitions >'. Below the banner is a section titled 'Active Competitions' with a sub-section 'All Competitions'. It contains a table of active competitions.

Active Competitions	
All Competitions	Active Competitions
	Flight Quest 2: Flight Optimization Optimize flight routes based on current weather and traffic. 41 days 34 teams \$220,000
	Personalized Web Search Challenge Re-rank web documents using personal preferences. 2 months 38 teams \$9,000
	See Click Predict Fix Predict which 311 issues are most important to citizens. 20 days 191 teams \$4,000
	As the World Churns Predict which customers will leave an insurance company in the next 12 months. 44 days 11 teams

The image shows a screenshot of the Kaggle Wiki page for 'Tutorials'. The navigation bar is the same as in the previous image. Below the navigation bar, there is a search bar with the text 'Wiki (Beta) >' on the left and 'Search' on the right. The main content area has a heading 'Tutorials' and a sub-section 'Tutorials by Kaggle'. Below this, there are several links to tutorial articles: 'Getting Started With Python For Data Science', 'Getting in Shape for The Sport of Data Science (youtube.com)', and 'Getting Started competitions'. There are also links to 'Digit Recognizer', 'Titanic: Machine Learning from Disaster', and 'Data Analysis in R'. The page also includes a search bar and a 'This article is a stub. You can help us by expanding it.' message.

This article is a stub. You can help us by expanding it.

Tutorials by Kaggle

[Getting Started With Python For Data Science](#)
Our product wiz Chris introduces you to the use of the Python programming language for data science including environment setup and code examples. A good place to start.

[Getting in Shape for The Sport of Data Science \(youtube.com\)](#)
A tutorial by our chief scientist, Jeremy Howard, giving a brief overview of a (highly successful) data scientist's toolkit.

Getting Started competitions

[Digit Recognizer](#)
The goal in this competition is to take an image of a handwritten single digit, and determine what that digit is. This competition is designed to introduce people to Machine Learning.

[Titanic: Machine Learning from Disaster](#)
This competition, in which we ask you to predict who was likely to survive the wreck of the *Titanic*, provides an ideal starting place for people who may not have a lot of experience in data science and machine learning.

Data Analysis in R

Coursera's Data Science Moocs



Data Analysis

Learn about the most effective data analysis methods to solve problems and achieve insight.



Oct 28th 2013 (8 weeks long)

Go to class

About the Course

You have probably heard that this is the era of "Big Data". Stories about companies or scientists using data to recommend movies, discover who is pregnant based on credit card receipts, or confirm the existence of the Higgs Boson regularly appear in Forbes, the Economist, the Wall Street Journal, and The New York Times. But how does one turn data into this type of insight? The answer is data analysis and applied statistics. Data analysis is the process of finding the right data to answer your question, understanding the processes underlying the data, discovering the important patterns in the data, and then communicating your results to have the biggest possible impact. There is a critical shortage of people with these skills in the workforce, which is why Hal Varian (Chief Economist at Google) says that being a statistician will be the sexy job for the next 10 years.

This course is an applied statistics course focusing on data analysis. The course will begin with an overview of how to organize, perform, and write-up data analyses. Then we will cover some of the most popular and widely used statistical methods like linear

About the Instructor



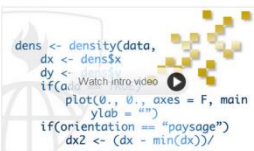
Course Details

Workload: 3-5 hours/week
Taught In: English
Subtitles Available In: English



Computing for Data Analysis

This course is about learning the fundamental computing skills necessary for effective data analysis. You will learn to program in R and to use R for reading data, writing functions, making informative graphs, and applying modern statistical methods.



Jan 6th 2014 (4 weeks long)

Enroll for Free

About the Course

In this course you will learn how to program in R and how to use R for effective data analysis. You will learn how to install and configure software necessary for a statistical programming environment, discuss generic programming language concepts as they are implemented in a high-level statistical language. The course covers practical issues in statistical computing which includes programming in R, reading data into R, creating informative data graphics, accessing R packages, creating R packages with documentation, writing R functions, debugging, and organizing and commenting R code. Topics in statistical data analysis and optimization will provide working examples.

Recommended Background

Some familiarity with programming concepts will be useful as well basic knowledge of statistical reasoning. At Johns Hopkins, this course is taken by first-year graduate students in Biostatistics.

About the Instructor



Course Details

Workload: 3-5 hours/week
Taught In: English
Subtitles Available In: English



Introduction to Data Science

Join the data revolution. Companies are searching for data scientists. This specialized field demands multiple skills not easy to obtain through conventional curricula. Introduce yourself to the basics of data science and leave armed with practical experience extracting value from big data.



May 1st 2013 (8 weeks long)

View class archive

About the Course

Commerce and research is being transformed by data-driven discovery and prediction. Skills required for data analytics at massive levels – scalable data management on and off the cloud, parallel algorithms, statistical modeling, and proficiency with a complex ecosystem of tools and platforms – span a variety of disciplines and are not easy to obtain through conventional curricula. Tour the basic techniques of data science, including both SQL and NoSQL solutions for massive data management (e.g., MapReduce and contemporaries), algorithms for data mining (e.g., clustering and association rule mining), and basic statistical modeling (e.g., linear and non-linear regression).

Course Syllabus

- Part 0: Introduction**
- Examples, data science articulated, history and context, technology landscape
- Part 1: Data Manipulation, at Scale**
- Databases and the relational algebra

About the Instructor



Course Details

Workload: 8-10 hours/week
Taught In: English
Subtitles Available In: English
[Preview Lectures](#)



Courses Partners About | Christos Gogos

Stanford Machine Learning

Learn about the most effective machine learning techniques, and gain practice implementing them and getting them to work for yourself.



Oct 14th 2013 (10 weeks long)

Enroll for Free

About the Course

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI. In this class, you will learn about the most effective machine learning techniques, and gain practice implementing them and getting them to work for yourself. More importantly, you'll learn about not only the theoretical underpinnings of learning, but also gain the practical know-how needed to quickly and powerfully apply these techniques to new problems. Finally, you'll learn about some of Silicon Valley's best practices in innovation as it pertains to machine learning and AI.

This course provides a broad introduction to machine learning, datamining, and statistical pattern recognition. Topics include: (i) Supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks). (ii) Unsupervised learning (clustering, dimensionality reduction, recommender systems,

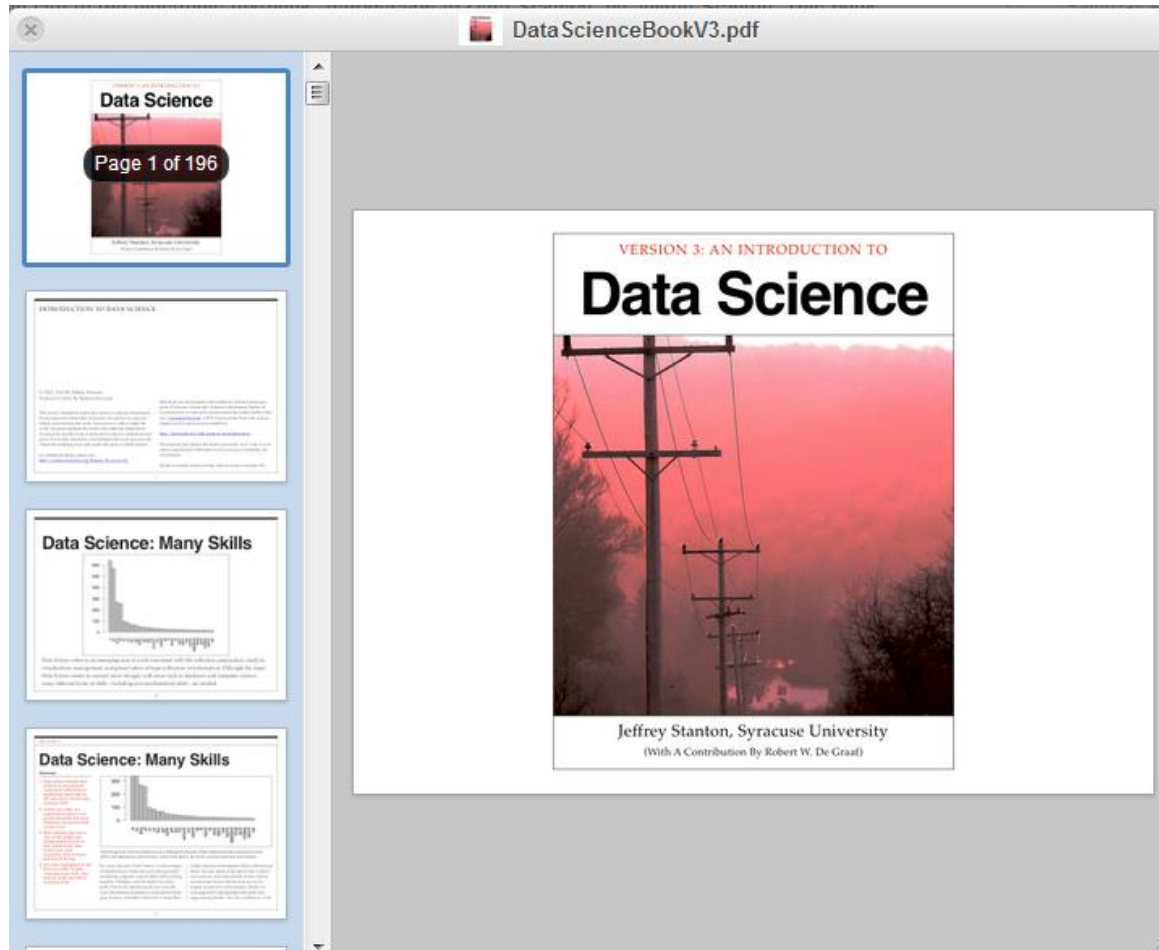
About the Instructor



Course Details

Workload: 5-7 hours/week
Taught In: English
Subtitles Available In: English
[Preview Lectures](#)

DataScience



http://jsresearch.net/wiki/projects/teachdatascience/Teach_Data_Science.html

Types of Data Professionals

Data Engineer



ML Engineer



Data Scientist



Data Analyst

