

Αναλυτική Δεδομένων & Μηχανική Μάθηση (1^ο εξ.)



Syllabus:

- Introduction to data analytics: principles, pipelines and pre-processing methods
- Common Machine Learning methods for classification and regression (Bayesian, Least Squares, SVM, etc.)
- Neural networks and Deep Learning fundamentals
- Clustering techniques (from standard to advanced)
- Applications on text/audio/video analytics

Lab hours (hands-on): Scikit-learn, Matlab, TensorFlow

Instructors:

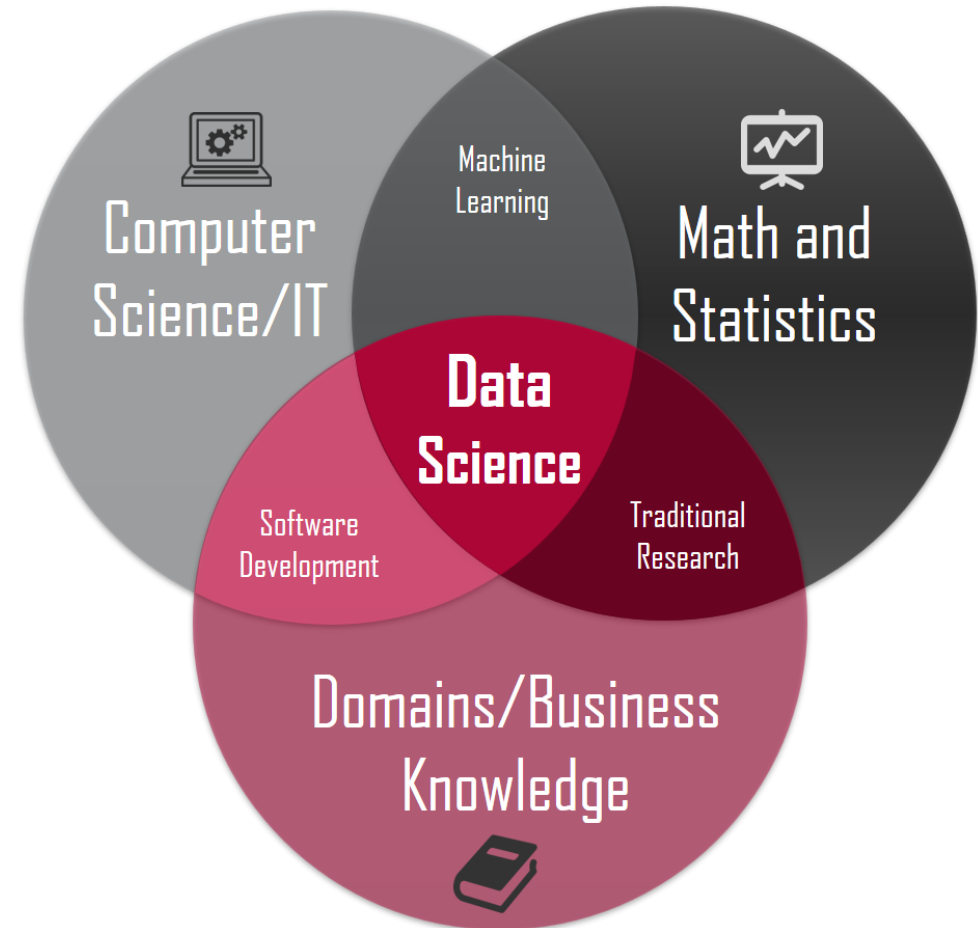
- Prof. Aggelos Pikrakis, Prof. Yannis Theodoridis
- TA: Andreas Tritsarolis (MSc)



Προκαταρκτικά (1)



- **Ορισμός** (από Wikipedia.org)
 - **Data Science** is an **interdisciplinary** field that uses **scientific methods, processes, algorithms** and **systems** to extract or extrapolate **knowledge and insights** from noisy, structured and unstructured **data**, and apply knowledge from data across a broad range of application domains.
 - **Data Science** is a "**concept to unify statistics, data analysis, informatics, and their related methods**" in order to "**understand and analyse actual phenomena**" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge.



Προκαταρκτικά (2)



■ Ιστορία

- In 1997, C.F Jeff Wu gave an inaugural lecture on “**Statistics = Data Science?**” at the Univ. Michigan. In this lecture, the term ‘data science’ was coined and it was advocated that statistics should be renamed data science and statisticians should be renamed data scientist.
- In 2008, the term **Data Scientist** was coined by DJ Patil and Jeff Hammerbacher to define their jobs at LinkedIn and Facebook, resp.
- In 2012, a Harvard Business Review article declared that ...

THE MAGAZINE

October 2012



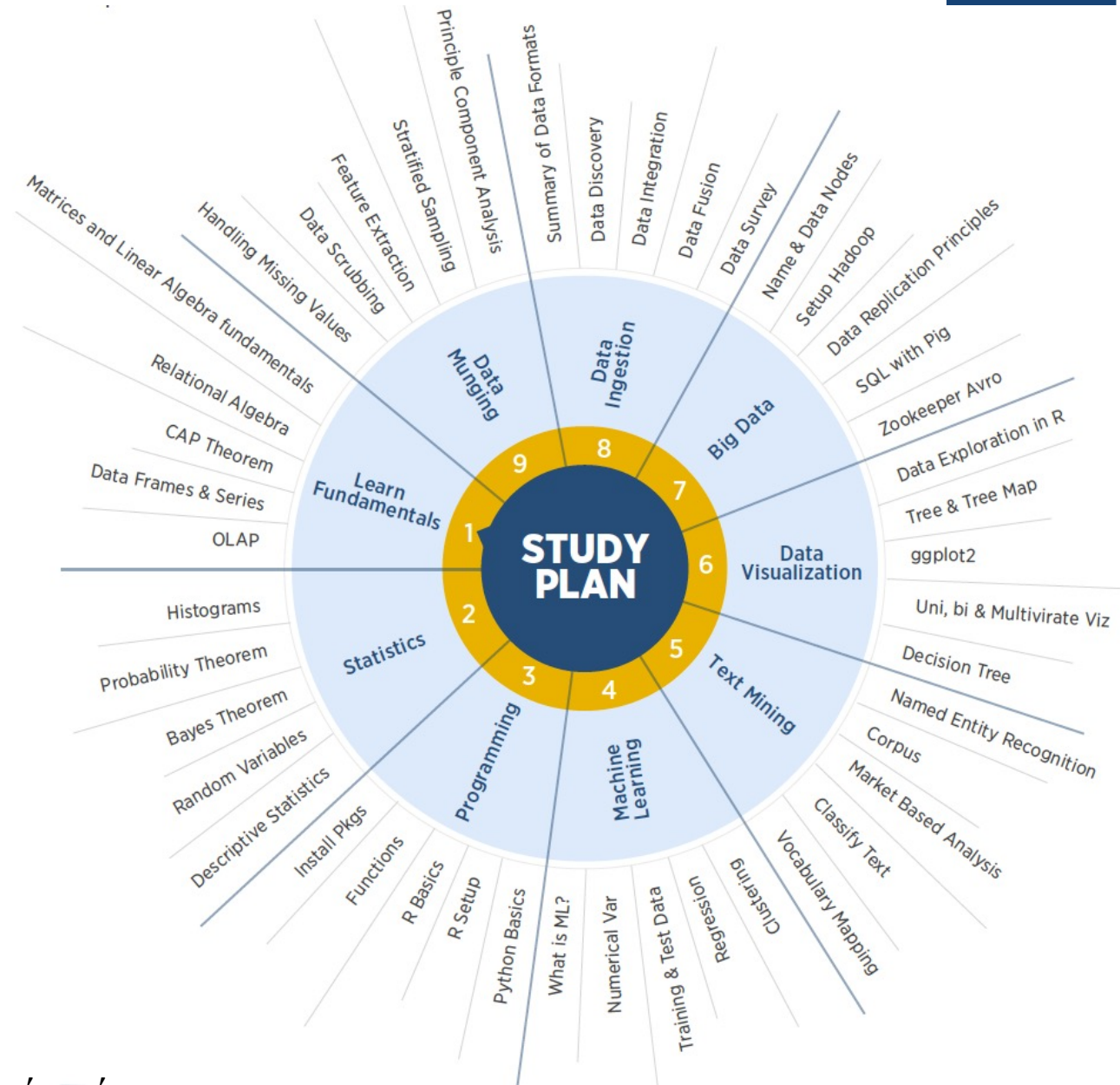
ARTICLE PREVIEW To read the full article, [sign-in](#) or [register](#). HBR subscribers, click [here to register](#) for **FREE** access »

Data Scientist: The Sexiest Job of the 21st Century

Προκαταρκτικά (3)

- Τι πρέπει να γνωρίζει ένας Data Scientist
 - Γλώσσες προγραμματισμό (Python, R, ...)
 - Στατιστική (περιγραφική στατιστική, θεωρία πιθανοτήτων, ...) & ανάλυση δεδομένων (δειγματοληψία, PCA, ...)
 - Σχεδίαση & διαχείριση (μεγάλων) βάσεων δεδομένων (SQL, OLAP, ...)
 - Οπτικοποίηση Δεδομένων,
 - Μηχανική Μάθηση (regression, clustering, ...)

κλπ.



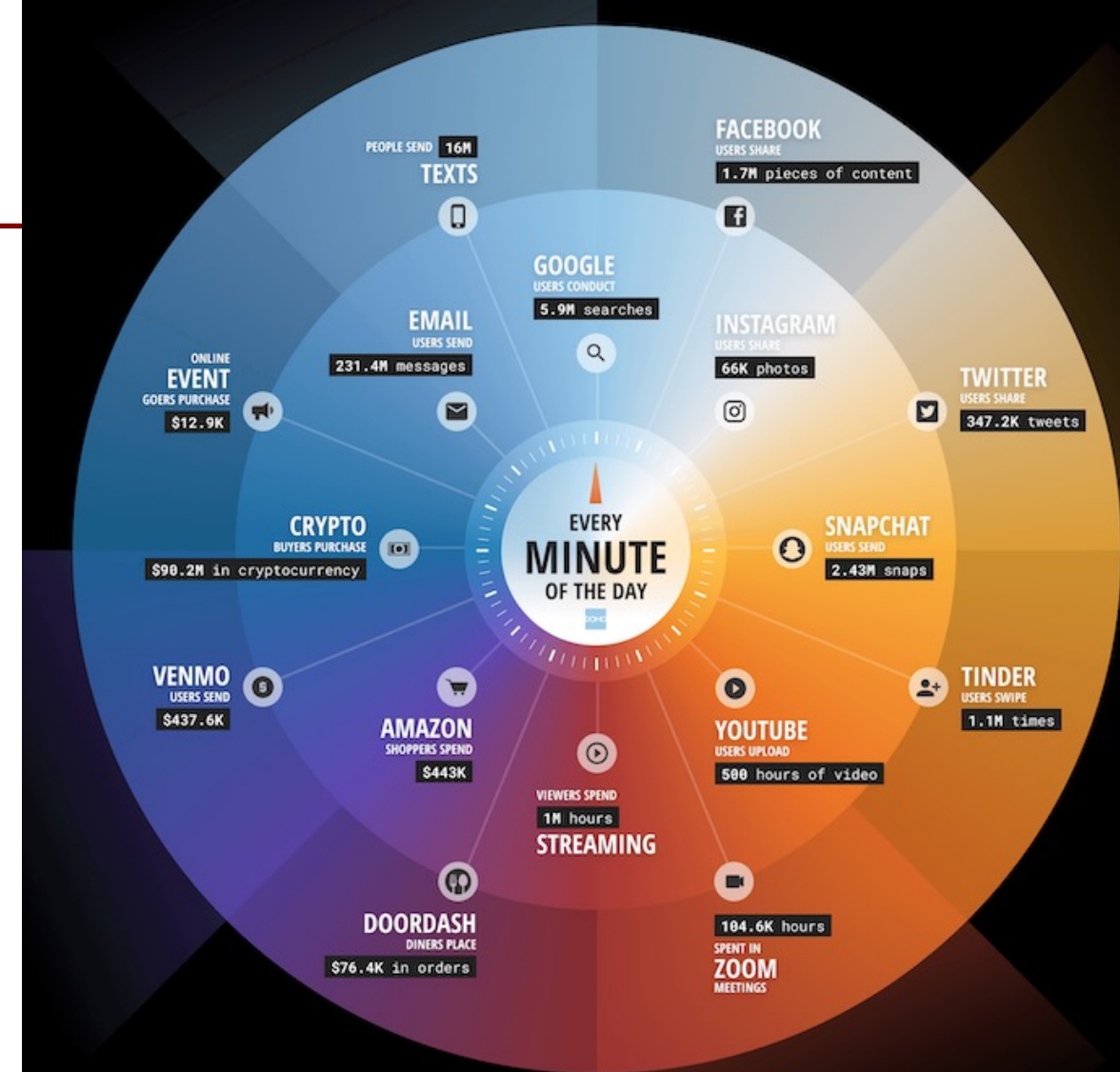
πηγή εικόνας:

<https://www.simplilearn.com/the-numbers-game-deciphered-guide-pdf>

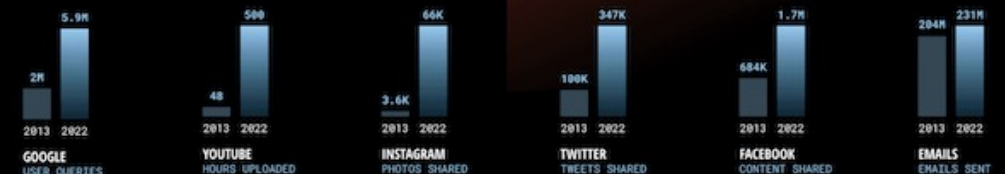
Η εποχή των Big Data

Every minute of the day (2022) ...

- **Google** users conduct 5.9M searches
- **Facebook** users share 1.7M pieces of content
- **Twitter** users share 347.2K tweets
- **Instagram** users share 66K photos
- **Amazon** shoppers spend \$443K
- etc.

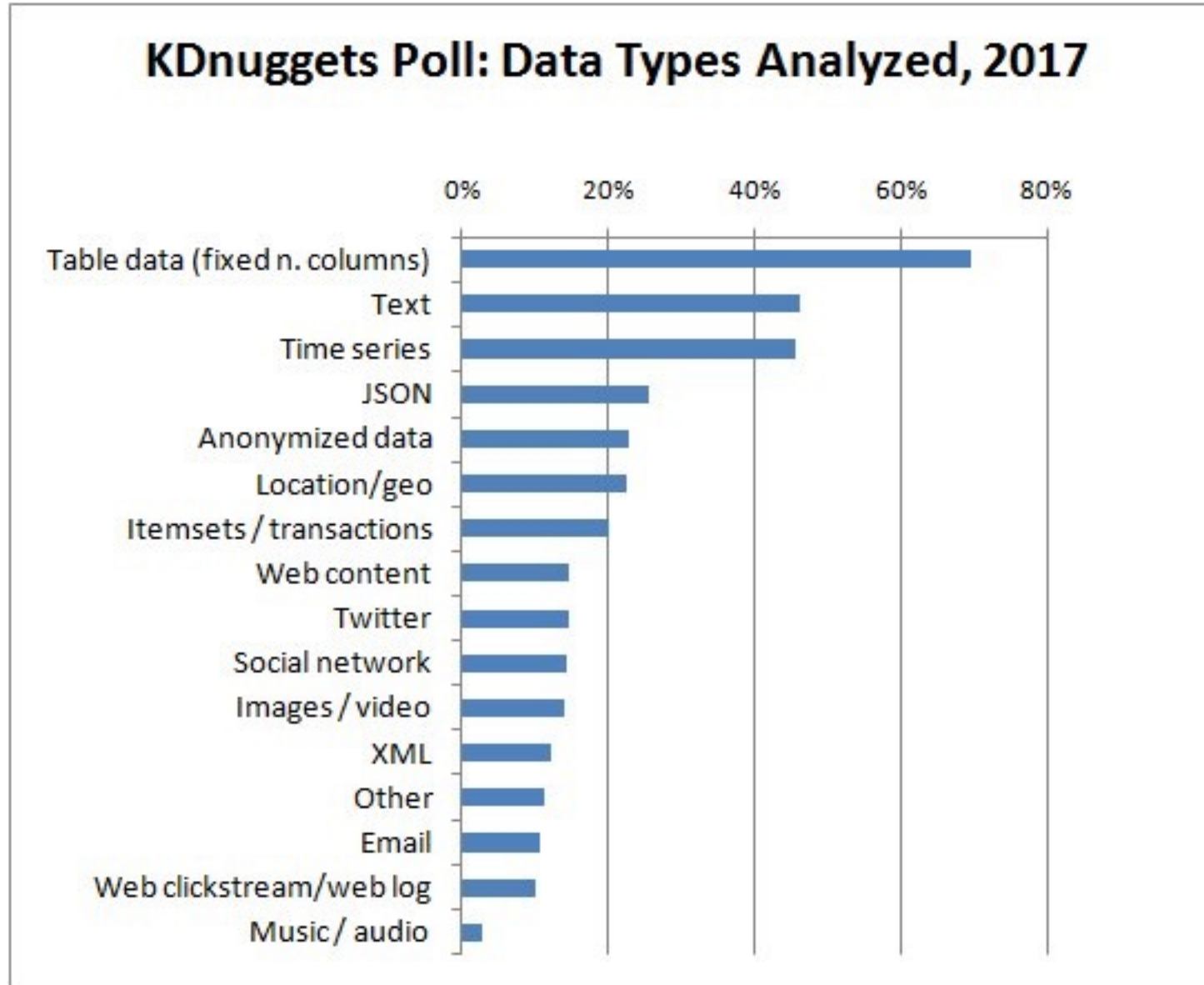


DATA NEVER SLEEPS 1.0 VS. 10.0



πηγή εικόνας:
<https://www.domo.com/data-never-sleeps>

Τι δεδομένα αναλύουμε συνήθως ...

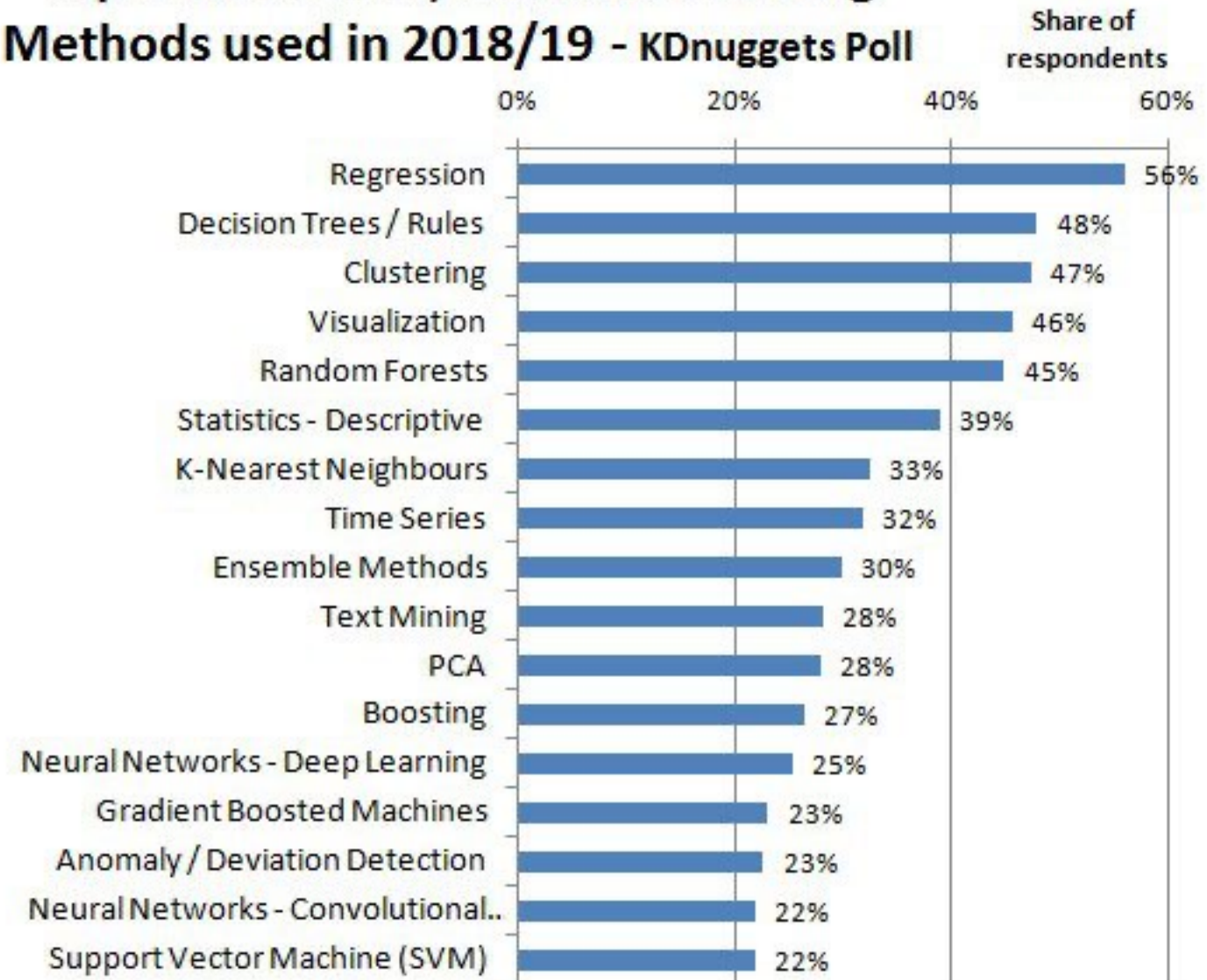


πηγή εικόνας:
kdnuggets.com

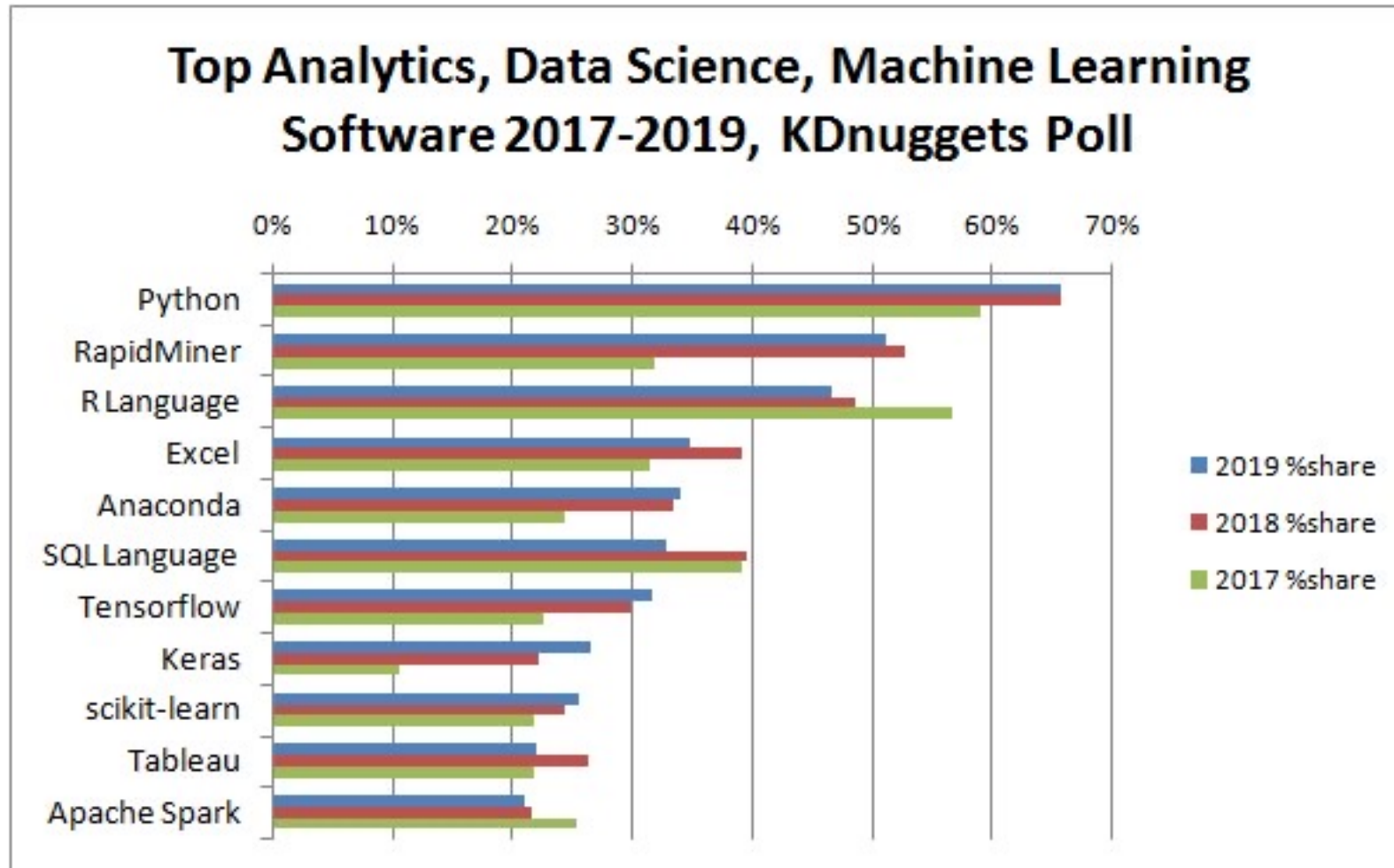
Με ποιες τεχνικές ...



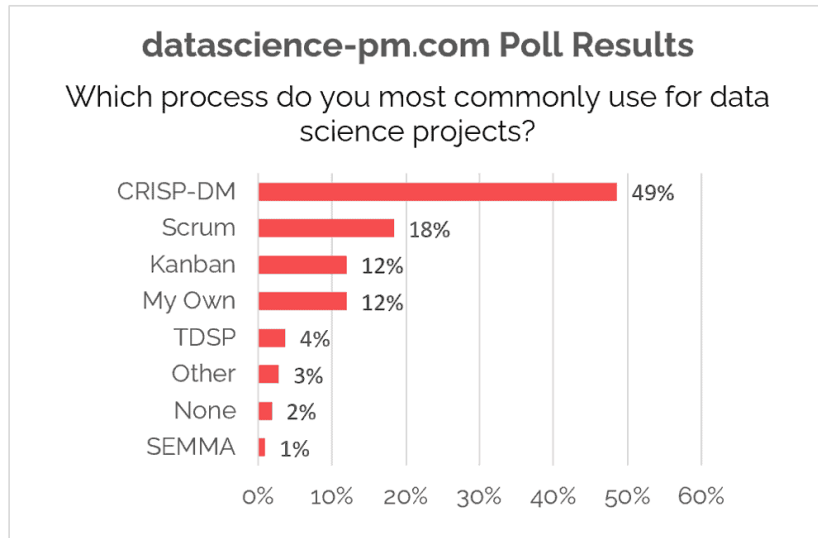
Top Data Science, Machine Learning Methods used in 2018/19 - KDnuggets Poll



Με ποιο λογισμικό ...

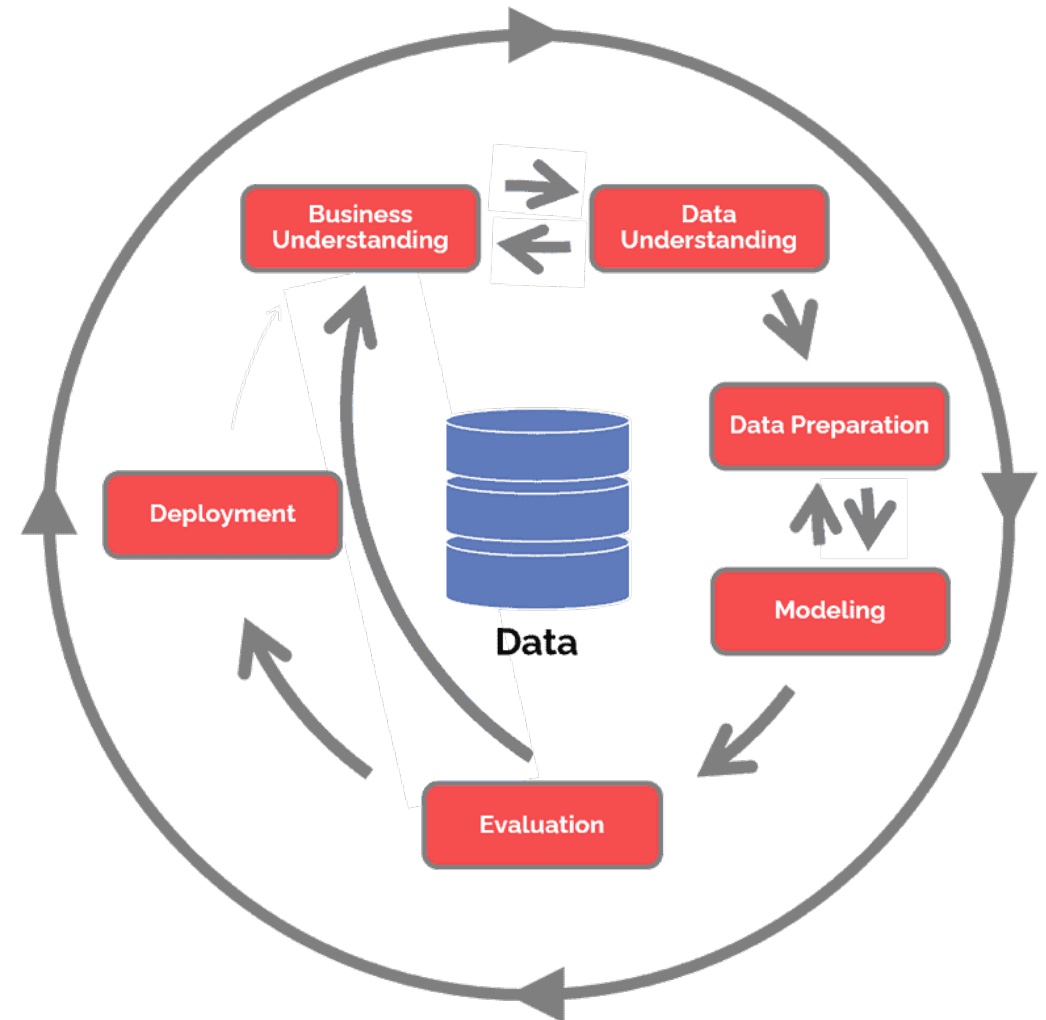


Με ποια διαδικασία ...



CRISP-DM

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?









πηγή εικόνας: datascience-pm.com

Προετοιμασία των δεδομένων για ανάλυση (1)

Βήμα 1: αποθήκευση σε κάποιο αποθηκευτικό χώρο

- “αποθήκη” δεδομένων (data warehouse) vs. “λίμνη” δεδομένων (data lake)
- διαφορές (υπέρ και κατά) ως προς:
 1. δεδομένα
 2. σχήμα
 3. κόστος /απόδοση
 4. ποιότητα δεδομένων
 5. χρήστες
 6. αναλυτικές μέθοδοι

Data warehouse vs data lake

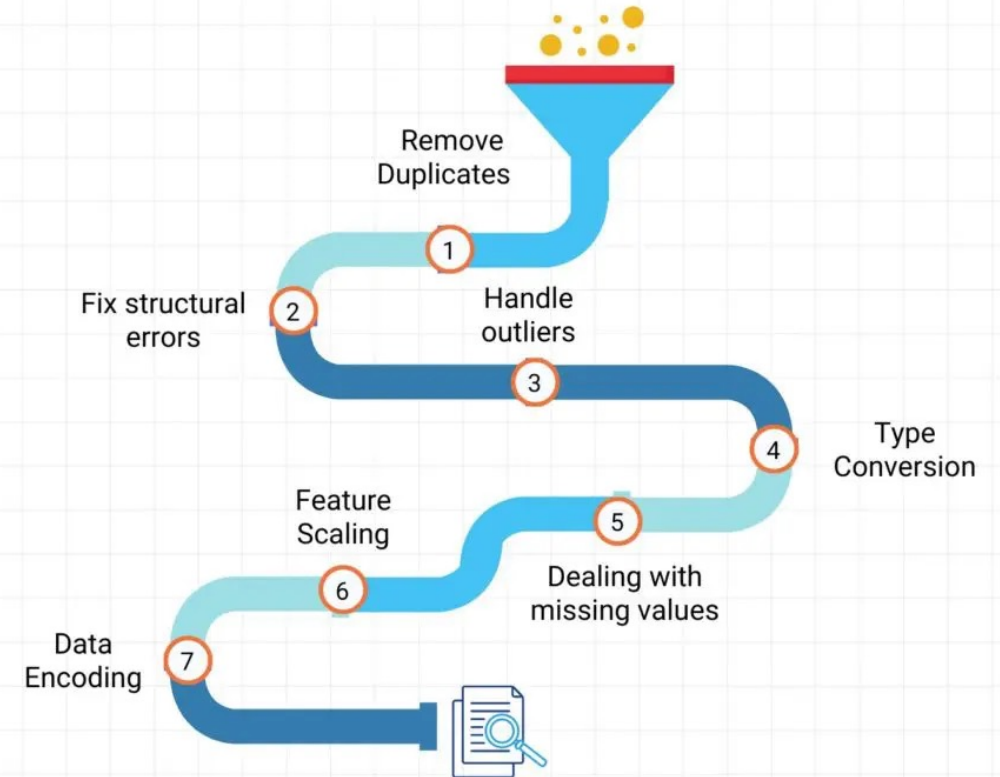
Characteristics	Data Warehouse	Data Lake
 Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
 Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
 Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
 Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
 Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
 Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

Προετοιμασία των δεδομένων για ανάλυση (2)

(2) Γνωριμία με τα δεδομένα και προεπεξεργασία

- Απαλοιφή διπλότυπων
- Διόρθωση σφαλμάτων (π.χ. ανορθογραφίες)
- Εντοπισμός ακραίων τιμών (π.χ. μέσω box-plot)
- Μετατροπή τύπων (π.χ. αριθμητικά vs. κατηγορικά δεδομένα)
- Χειρισμός ελλιπών τιμών (απόρριψη; τροποποίηση;)
- Αλλαγή κλίμακας τιμών (π.χ. στην κλίμακα 0..1)
- Κωδικοποίηση τιμών (π.χ. Yes/No)

Data preprocessing The foundation of data science solution



datasciencedojo
— data science for everyone —

(c) Copyrights Reserved <https://datasciencedojo.com>

πηγή εικόνας:
<https://datasciencedojo.com/blog/data-preprocessing-data-science-solution/>

Προετοιμασία των δεδομένων για ανάλυση (3)



Γνωριμία με τα δεδομένα:

- Αποτύπωση σε πίνακες (αν εφαρμόζεται)
- Απαλοιφή διπλότυπων εγγραφών (duplicate records)
- Εντοπισμός ελλιπών τιμών (missing values), θορύβου (outliers). Πώς;
 - π.χ. με μια πρώτη στατιστική ανάλυση
- Μετατροπή, π.χ. binning

Data can be numeric or categorical

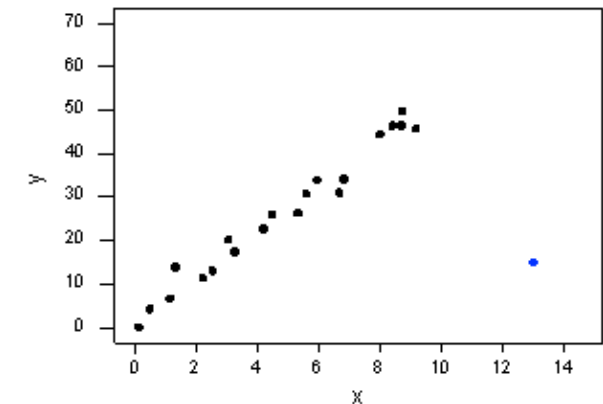


User	Income	IncomeGrp
1	\$65,500	[61k,80k]
2	\$81,041	[81k,100k]
3	\$38,346	[21k,40k]
4	\$47,072	[41,60k]
5	\$30,812	[21k,40k]
6	\$21,618	[21k,40k]
7	\$97,872	[81k,100k]

If there are 4 categories/ranges, we have 3 binary variables.

User	inc_41_60	inc_61_80	inc_81_100
1	0	1	0
2	0	0	1
3	0	0	0
4	1	0	0
5	0	0	0
6	0	0	0
7	0	0	1

Humidity	Windy
70%	false
68%	true
80%	false
?	false
50%	false
45%	true
58%	?
65%	false
40%	false
0%	false
?	true

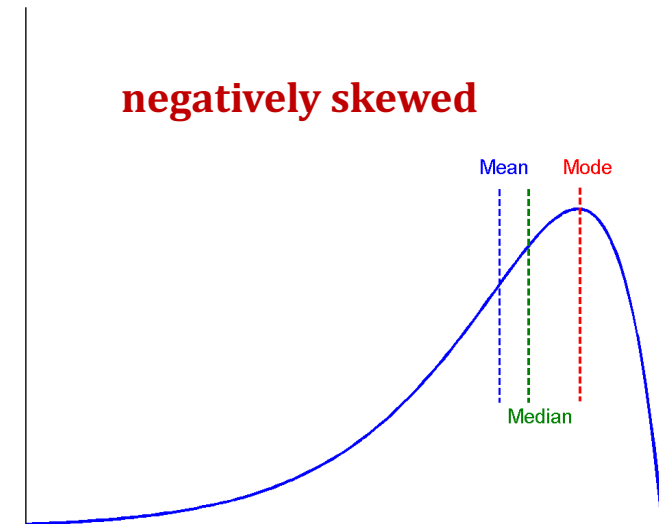
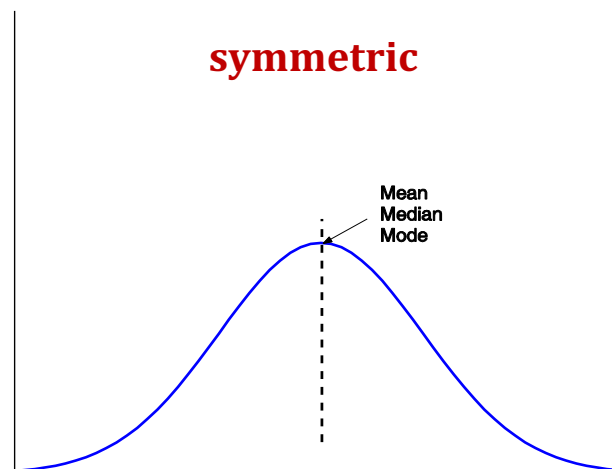
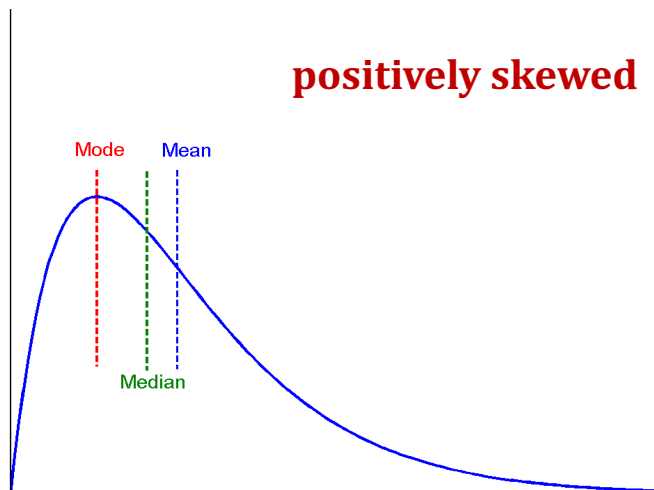
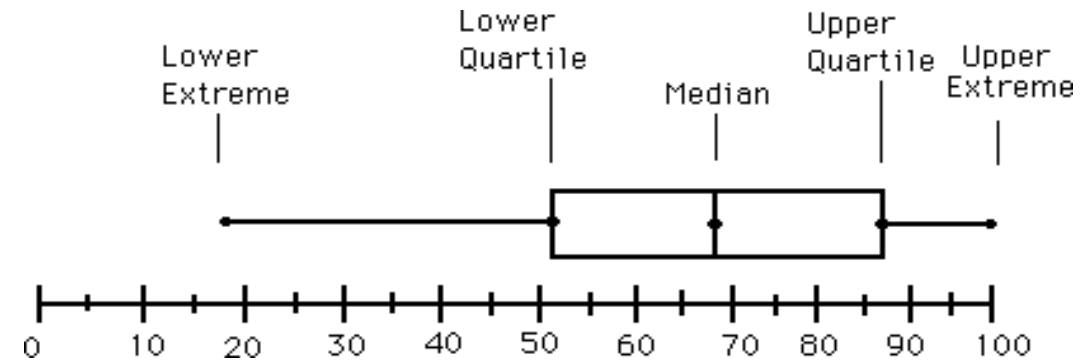


Προετοιμασία των δεδομένων για ανάλυση (4)



Γνωριμία με τα δεδομένα:

- ...
- Μια πρώτη στατιστική ανάλυση

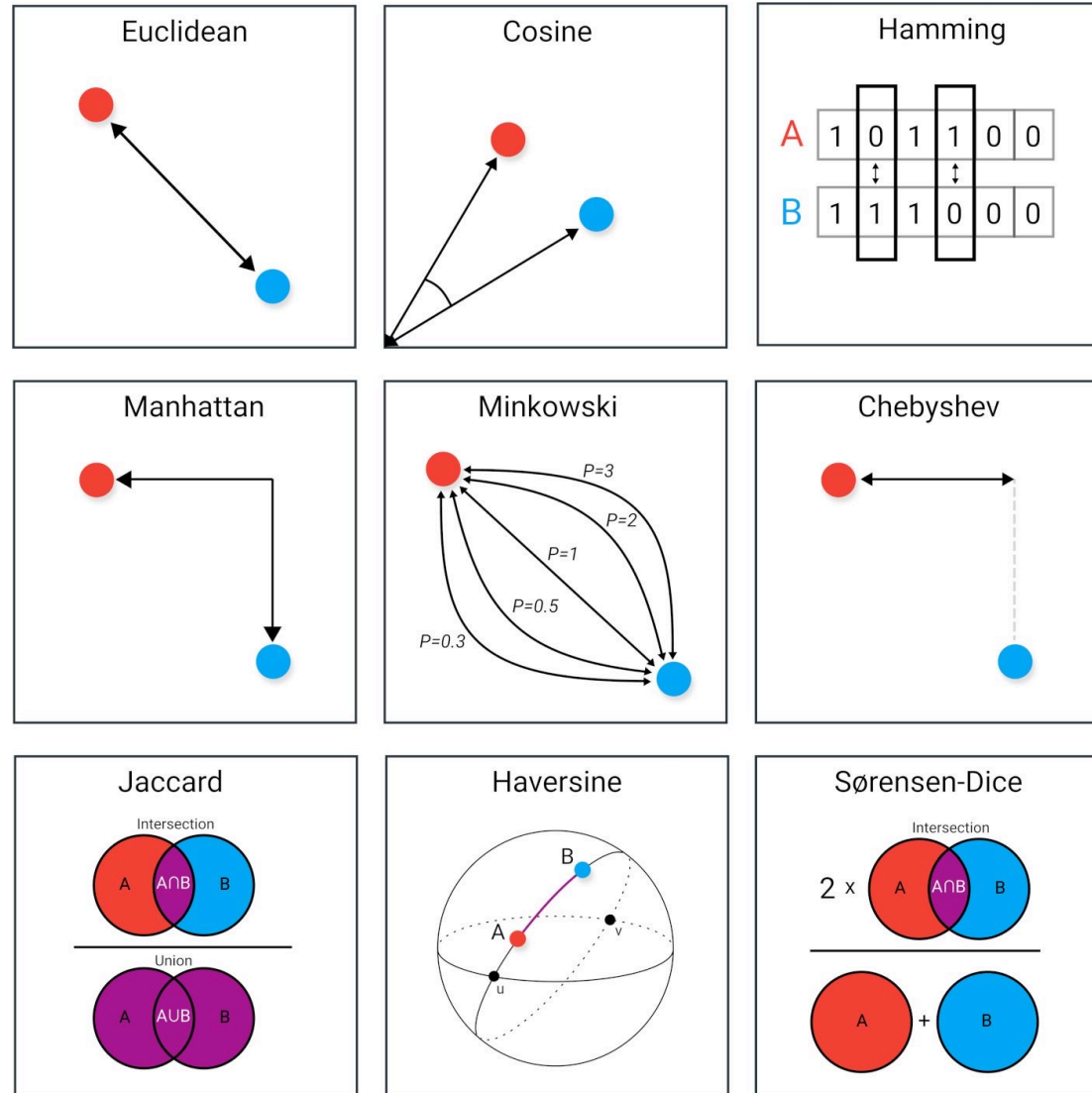


Προετοιμασία των δεδομένων για ανάλυση (5)



Γνωριμία με τα δεδομένα:

- ...
- Εντοπισμός τυχόν θορύβου (outliers) → πώς μετράμε (αν)ομοιότητα ή «απόσταση»



Προετοιμασία των δεδομένων για ανάλυση (6)



- **Ομοιότητα (similarity)**

- Αναλόγως του τύπου των δεδομένων...

- **Απόσταση (distance)**

- Ειδική περίπτωση: μετρική (metric), π.χ. Ευκλείδεια απόσταση

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = x - y / (n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

ΟΡΙΣΜΟΣ. Εστω X οποιοδήποτε μη-κενό σύνολο. Ονομάζουμε **μετρική** στο X κάθε συνάρτηση d ορισμένη στο καρτεσιανό γινόμενο $X \times X$ και με πραγματικές τιμές

$$d : X \times X \rightarrow \mathbb{R}$$

με τις παρακάτω ιδιότητες:

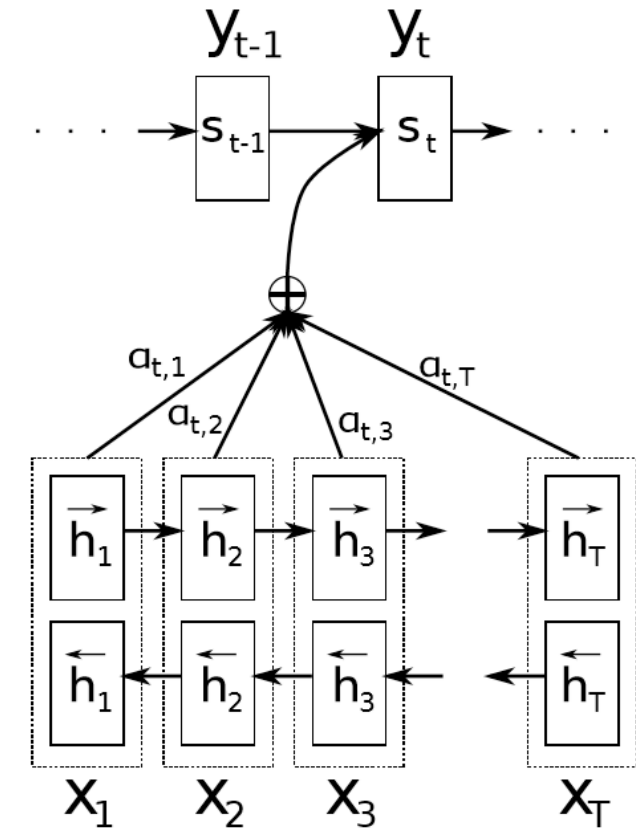
- (i) $d(x, y) \geq 0$ για κάθε $x, y \in X$.
- (ii) Για κάθε $x, y \in X$ ισχύει: $d(x, y) = 0$ αν και μόνο αν $x = y$.
- (iii) $d(x, y) = d(y, x)$ για κάθε $x, y \in X$.
- (iv) $d(x, y) \leq d(x, z) + d(z, y)$ για κάθε $x, y, z \in X$.

πηγή: http://fourier.math.uoc.gr/~papadim/analysis_2/5-5-14.pdf

Μηχανική μάθηση βασισμένη σε NNs (1)



- **Νευρωνικά Δίκτυα σε προβλήματα ανάλυσης δεδομένων**
 - Autoencoders για τη μείωση διαστάσεων
 - Embeddings για την αναπαράσταση δεδομένων
 - Μονοδιάστατες και δισδιάστατες συνελκτικές αρχιτεκτονικές (1D, 2D convolutional architectures)
 - Μηχανισμοί προσοχής (attention mechanisms)
 - Παραδείγματα ταξινόμησης, αναπαράστασης και μετασχηματισμών δεδομένων

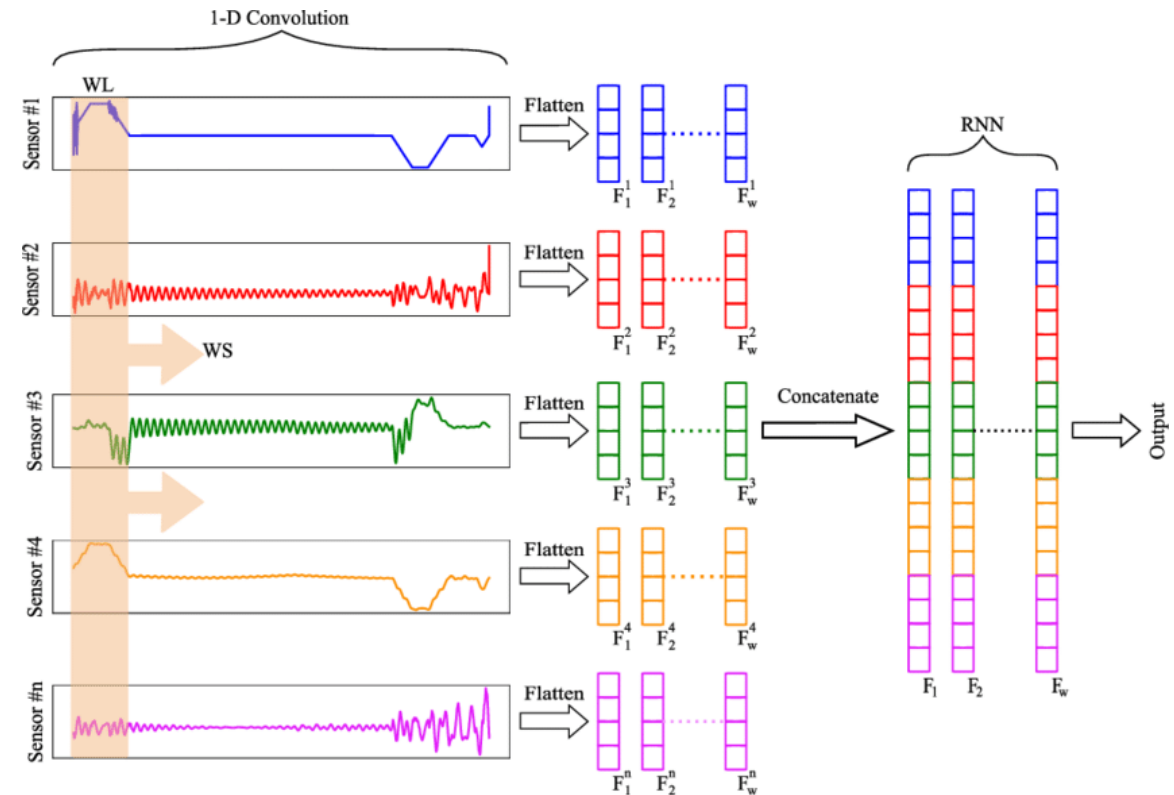


πηγή εικόνας: Original Attention Mechanism (D. Bahdanau, 2014)

Μηχανική μάθηση βασισμένη σε NNs (2)



- **Αναλυτική Χρονοσειρών (time-series analytics)**
 - Αναδρομικές (recurrent) και συνελκτικές (convolutional) αρχιτεκτονικές και συνδυασμοί τους
 - Sequence-to-Sequence μοντέλα
 - Δίκτυα Transformers
 - Παραδείγματα forecasting, classification, anomaly detection



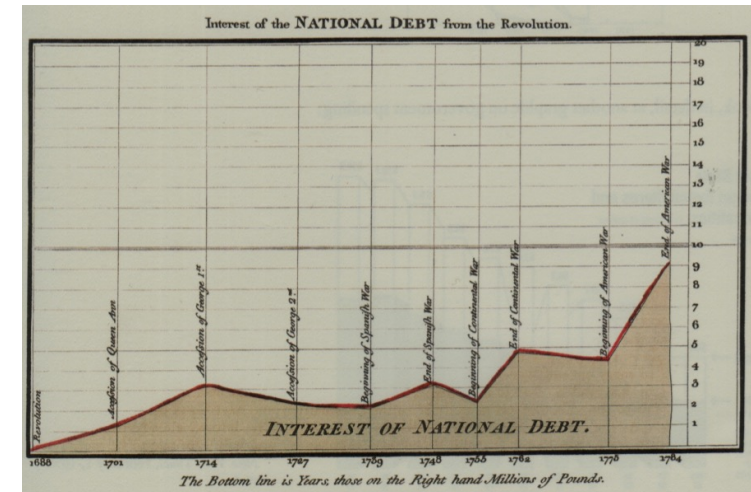
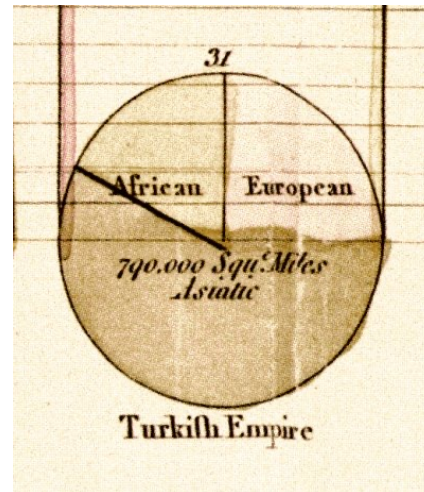
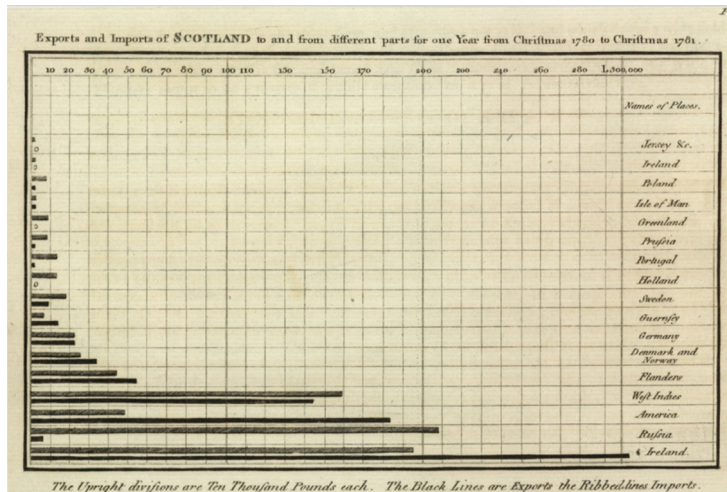
πηγή εικόνας: Multi-Head CNN-RNN for Multi-Time Series Anomaly Detection (M. Canizo et al., 2019)

Η οπτικοποίηση ως αναλυτική μέθοδος (1)



Information visualization: απεικόνιση πληροφορίας που δεν έχει ξεκάθαρη αναπαράσταση σε 2D/3D χώρο

- «οι γραφικές παραστάσεις μεταφέρουν το νόημα καλύτερα από ό,τι κάνουν τα δεδομένα» (William Playfair, 1759-1823)

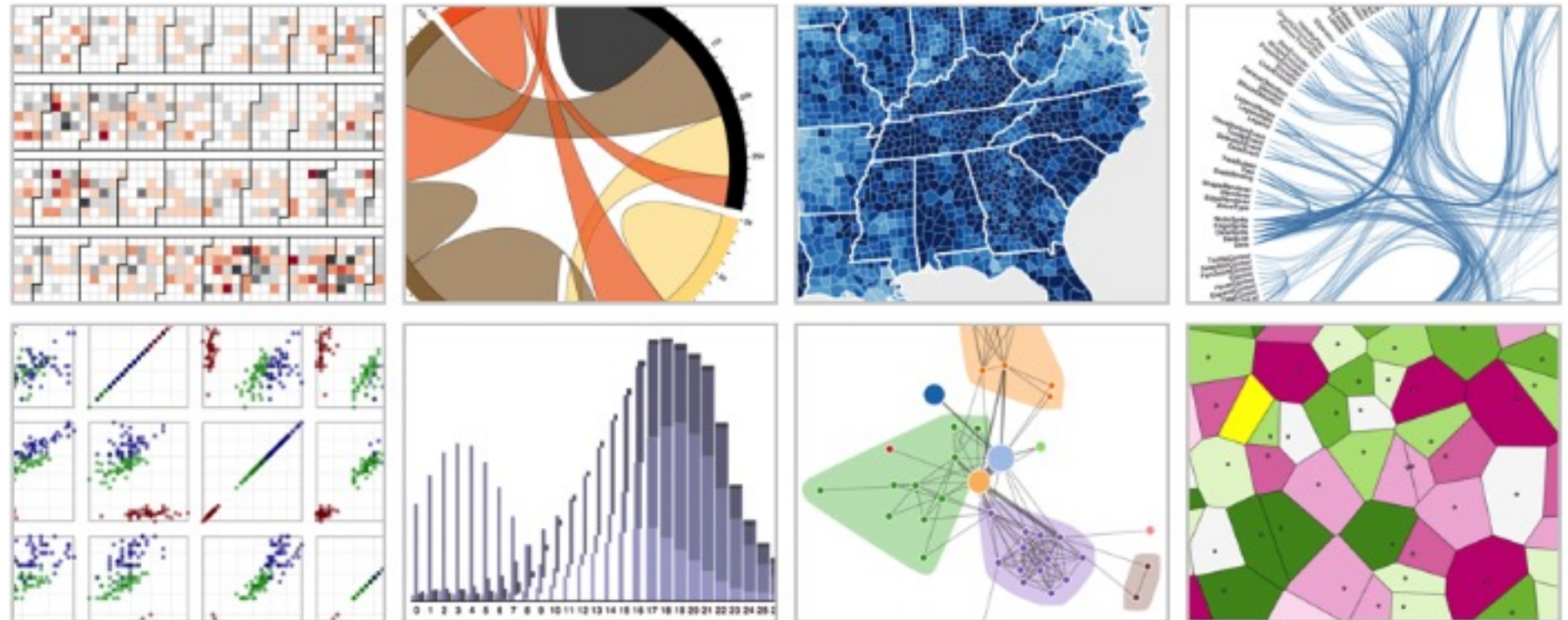


πηγή εικόνας: https://commons.wikimedia.org/wiki/William_Playfair

Η οπτικοποίηση ως αναλυτική μέθοδος (2)



Interactive visualization: ο χρήστης των εργαλείων visual analytics – VA (data scientist? domain expert?) έχει τον πρώτο ρόλο



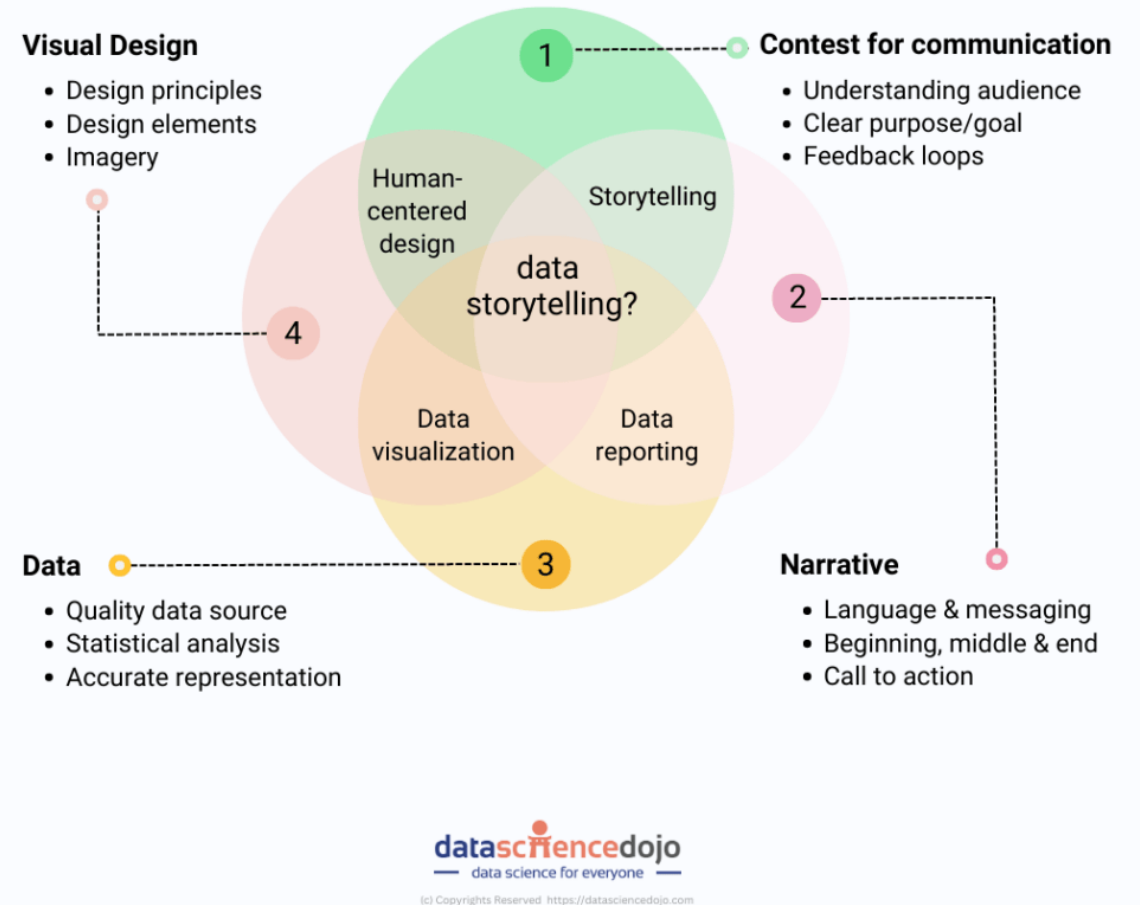
πηγή εικόνας: <http://idl.cs.washington.edu/papers/d3>

Η οπτικοποίηση ως αναλυτική μέθοδος (3)

Data storytelling: πώς «επικοινωνούμε» το αποτέλεσμα της ανάλυσης στο κοινό.
Συνδυασμός:

- διήγησης μιας ιστορίας (storytelling)
- αναφορών δεδομένων (data reporting)
- οπτικοποίησης δεδομένων (data visualization)
- ανθρωποκεντρικής σχεδίασης (human-centered design)

What is data storytelling?



πηγή εικόνας: <https://datasciencedojo.com/blog/data-storytelling-in-action/>

All-in-one: “Datathon”



A **datathon** is an event where participants gather to solve practical problems through the application of data science tools and techniques, by working together in teams to generate insights and potential solutions (πηγή: <https://www.datacamp.com/blog/how-to-plan-a-successful-datathon>)



*“Tell me and I will forget;
show me and I may remember;
involve me and I will understand”*
– Confucius

Παράδειγμα: 2022 Datathon @ Data Science Lab

www.datastories.org

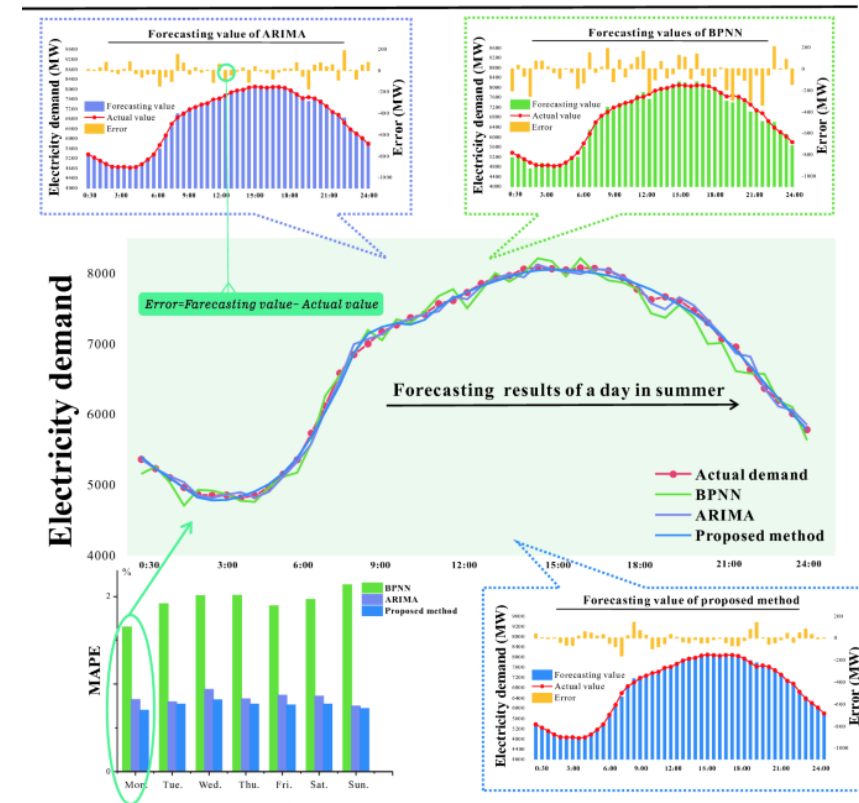


Challenge: **Energy consumption forecasting**

As a team of data scientists in a power plant operator, your goal is to analyze the information received by the smart energy meters of households in a region ...

Data Science pipeline (3 steps):

- **Data collection and preprocessing** -- Collect and pre-process (cleanse, transform, etc.) the required data as well as other datasets you consider relevant (weather? demographic? news posts?) ...
- **Data analytics** -- through short- and long-term forecasting (using ML techniques), evaluate whether the energy consumption will increase or not; short-term means within next 3 hours; long-term means within next 3 days; ...
- **Data storytelling** -- Present / visualize the findings of ML tasks in a form appropriate for a non-technical audience ...



πηγή εικόνας: https://doi.org/10.1007/978-3-030-39431-8_18

Καλό εξάμηνο!



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

Josh Wills, Director of Data Engineering at Slack