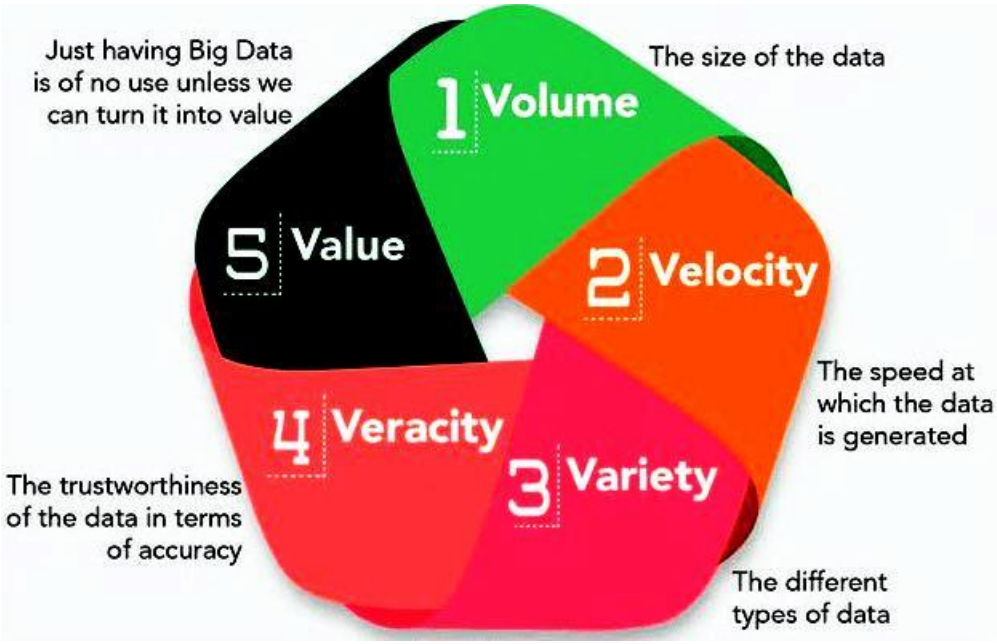
ARISTOTLE
UNIVERSITY of
THESSALONIKI

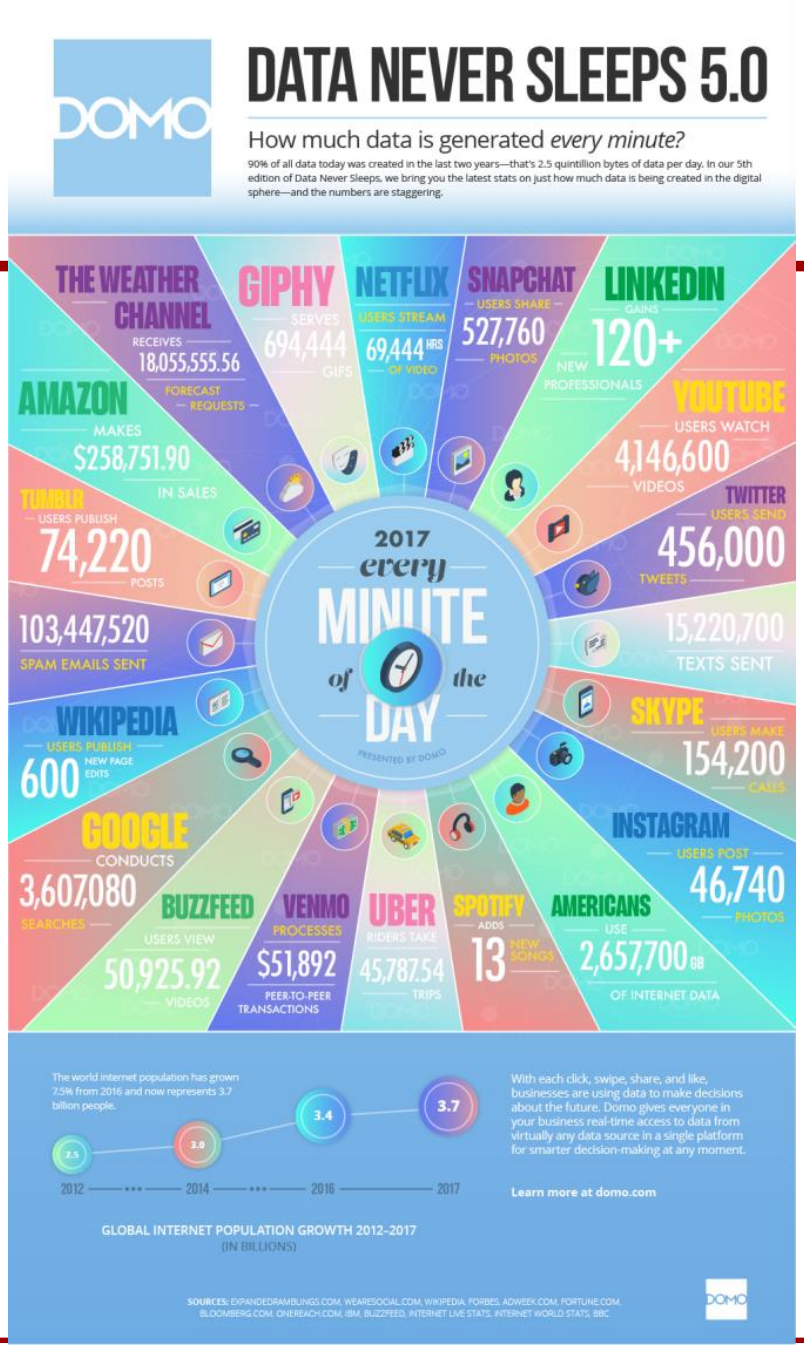
SCHOOL OF INFORMATICS

# Applying big data analytics in practice

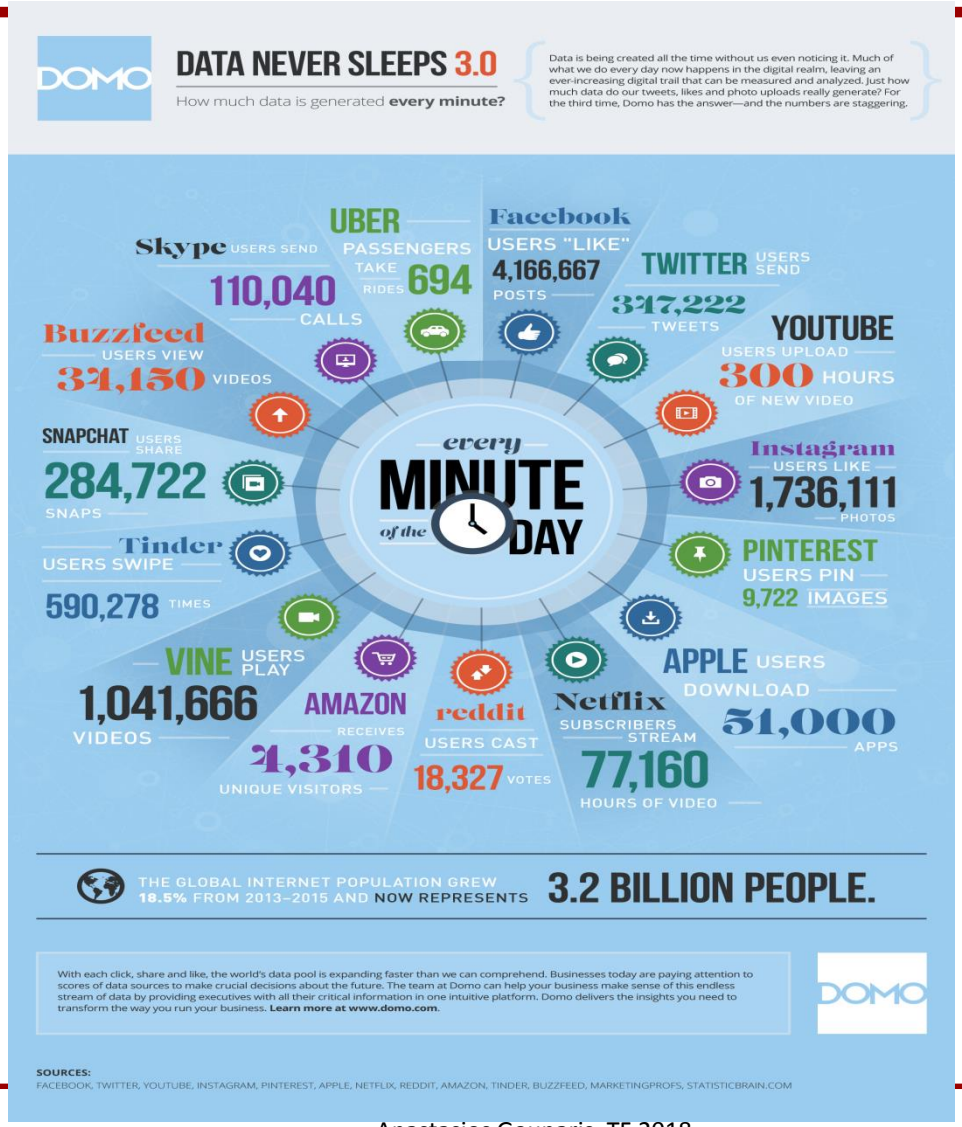
Anastasios Gounaris
School of Informatics
Aristotle University of Thessaloniki
*datalab.csd.auth.gr/~gounaris*
*email: gounaria@csd.auth.gr*

# New data every 1 min



# What was happening 2 years before (Nov. 2015)




**Big Data is becoming bigger and bigger!!**

# Big Data Sizes

Sizes:

- Tiny → 0s
- Small → 1000s fitting in memory
- Medium → 1000000 (millions, may not fit in memory)
- Large → 1000000000 (billions)
- Huge → 1000000000000 ++ (trillions ++)

*From Graefe's "New algorithms for join and grouping operations", 2011.*

**Big data**

1) **is there**
2) **is becoming bigger**
3) **applies to many entities**

"**driverless vehicle systems will create up to 1GB of data per second**"
(https://datafloq.com/read/how-autonomous-cars-will-make-big-data-even-bigger/1795)

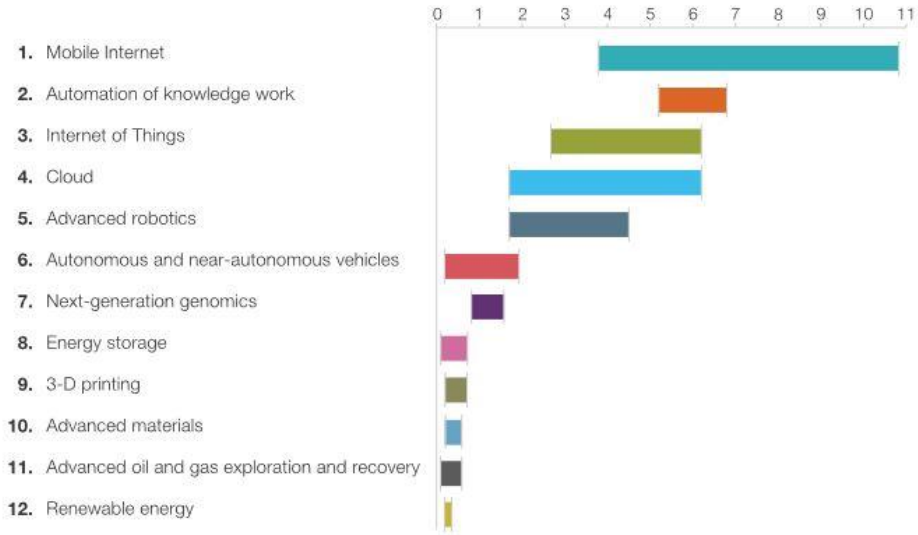

"**the volume of the data generated has exceeded 1000 exabytes (or 1 billion terabytes) annually, since 2015** "
(Shen Yin, Okyay Kaynak:Big Data for Modern Industry: Challenges and Trends [Point of View]. Proc. of the IEEE 103(2): 143-146 (2015))

# On big data value

Exhibit 1 of 13

## A gallery of **disruptive technologies**

**Estimated potential economic impact of technologies across sized applications in 2025**, $ trillion, annual




SOURCE: McKinsey Global Institute

Notes on sizing: These economic impact estimates are not comprehensive and include potential direct impact of sized applications only. They do not represent GDP or market size (revenue), but rather economic potential, including consumer surplus. The relative sizes of technology categories shown do not constitute a "ranking," since our sizing is not comprehensive. We do not quantify the split or transfer of surplus among or across companies or consumers, since this would depend on emerging competitive dynamics and business models. Moreover, the estimates are not directly additive, since some applications and/or value drivers are overlapping across technologies. Finally, they are not fully risk- or probability-adjusted.

# Some quotes

**61% of companies acknowledge that Big Data is now a driver of revenues in its own right and is becoming as valuable to their businesses as their existing products an**
*Capgemini report*

**The Big Data market, $84B in 2026**
*Forbes report*

**Companies that fully 50% improvement in overall efficiencies ir**
*Forbes report*

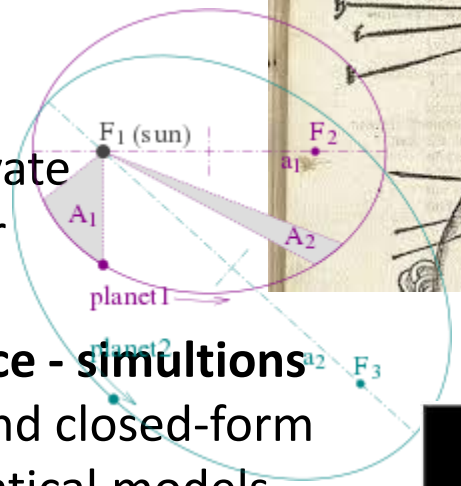
**0.5% of all the data is ever analyzed**
*MIT Technology Review (2013)*

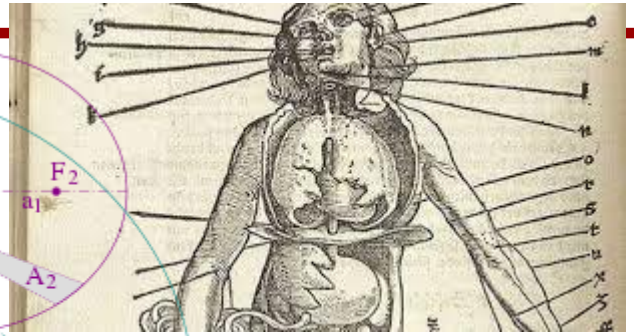
**Through 2015, 85% of Fortune 500 organizations will be unable to exploit Big Data for competitive advantage**
*Gartner report*

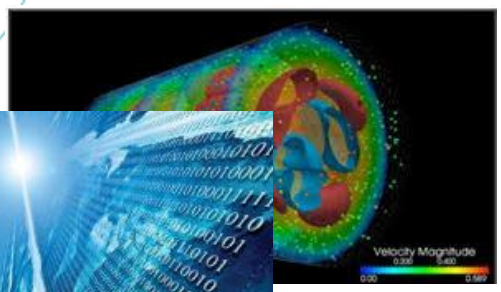
**Big data analytics is profitable yet difficult and challenging**

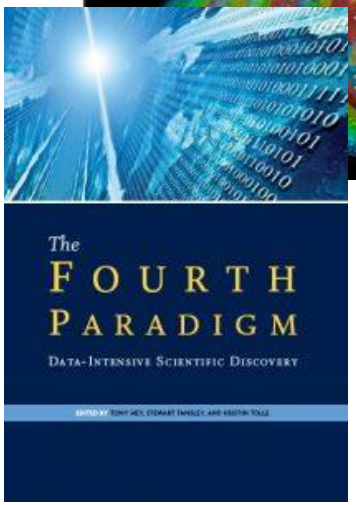
# Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science - simultions**
  - It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science - Data-Intensive Scientific Discovery**
  - Advanced **data analytics/mining** is both a major new challenge and the driver for new scientific discoveries!






*Material from "Data Mining: Concepts and Techniques"*

# What Is Data Analytics?

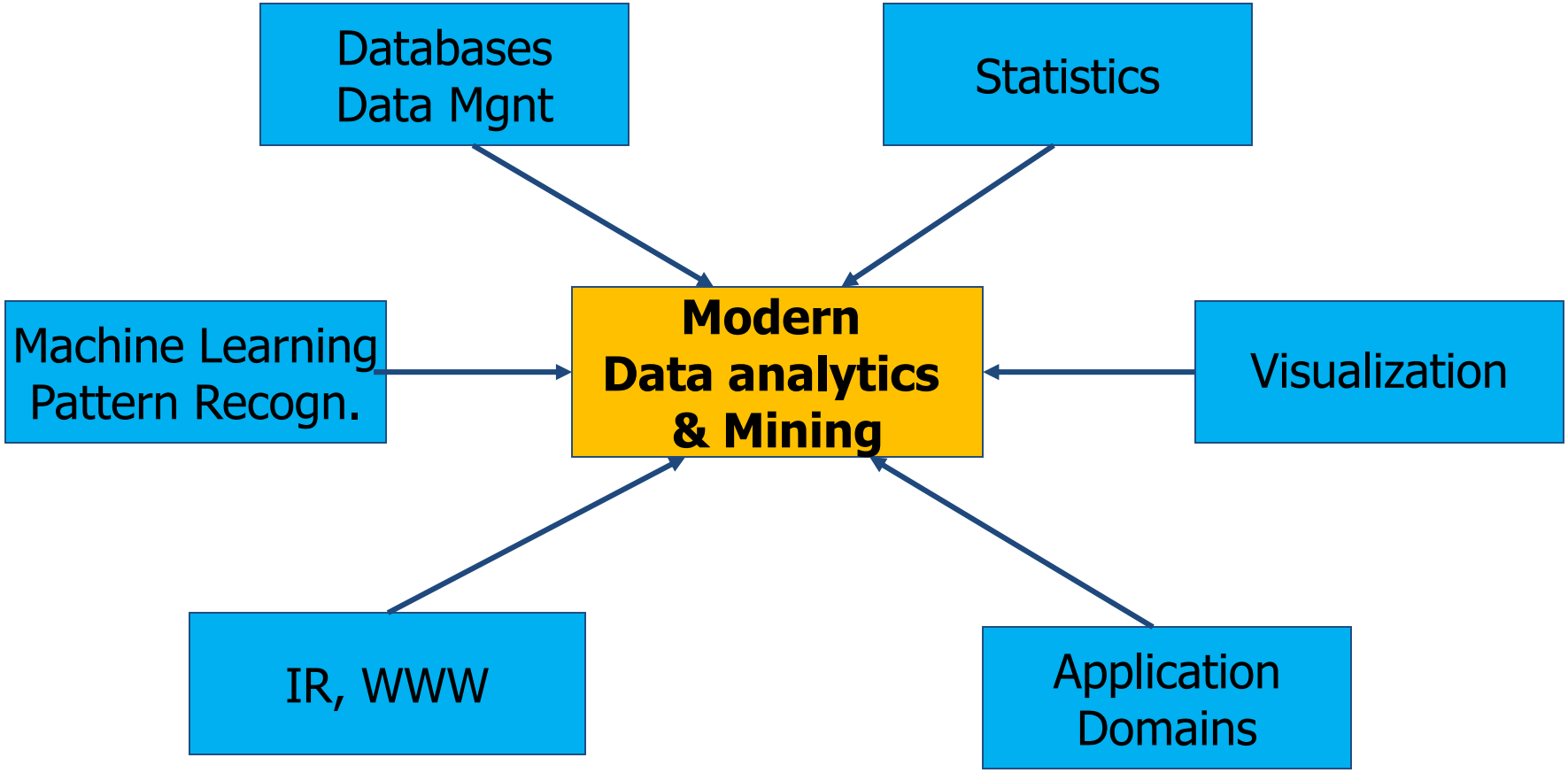
Types:

- Descriptive vs. predictive

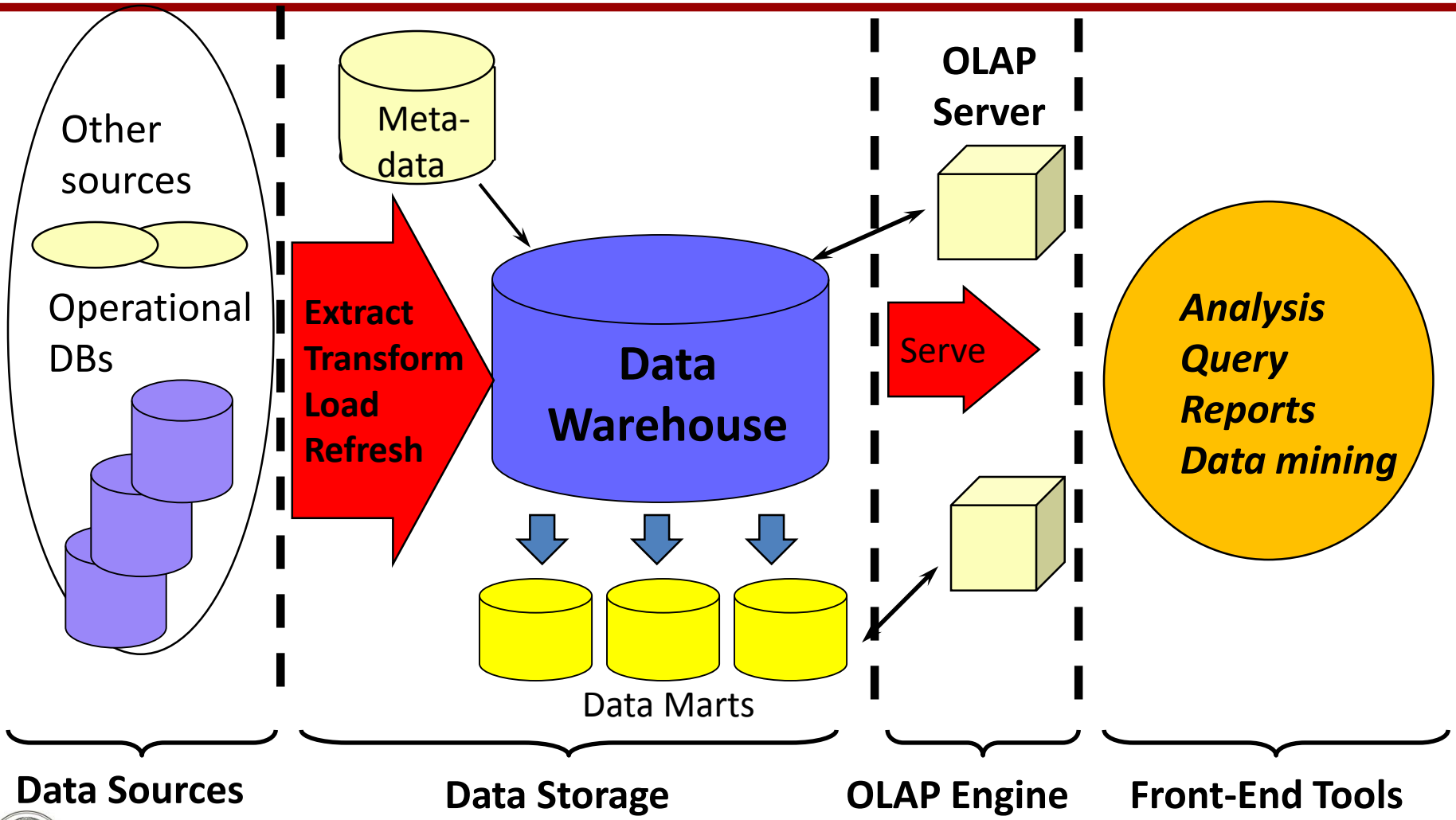
- Multiple Functions
  - Applied Statistics
  - Regression / Time series Forecasting
  - Data Warehousing
  - Data Flows
  - Association Rules
  - Classification
  - Clustering
  - Outlier Detection
  - Causality

*Material from "Data Mining: Concepts and Techniques"*

# Confluence of intertwined disciplines

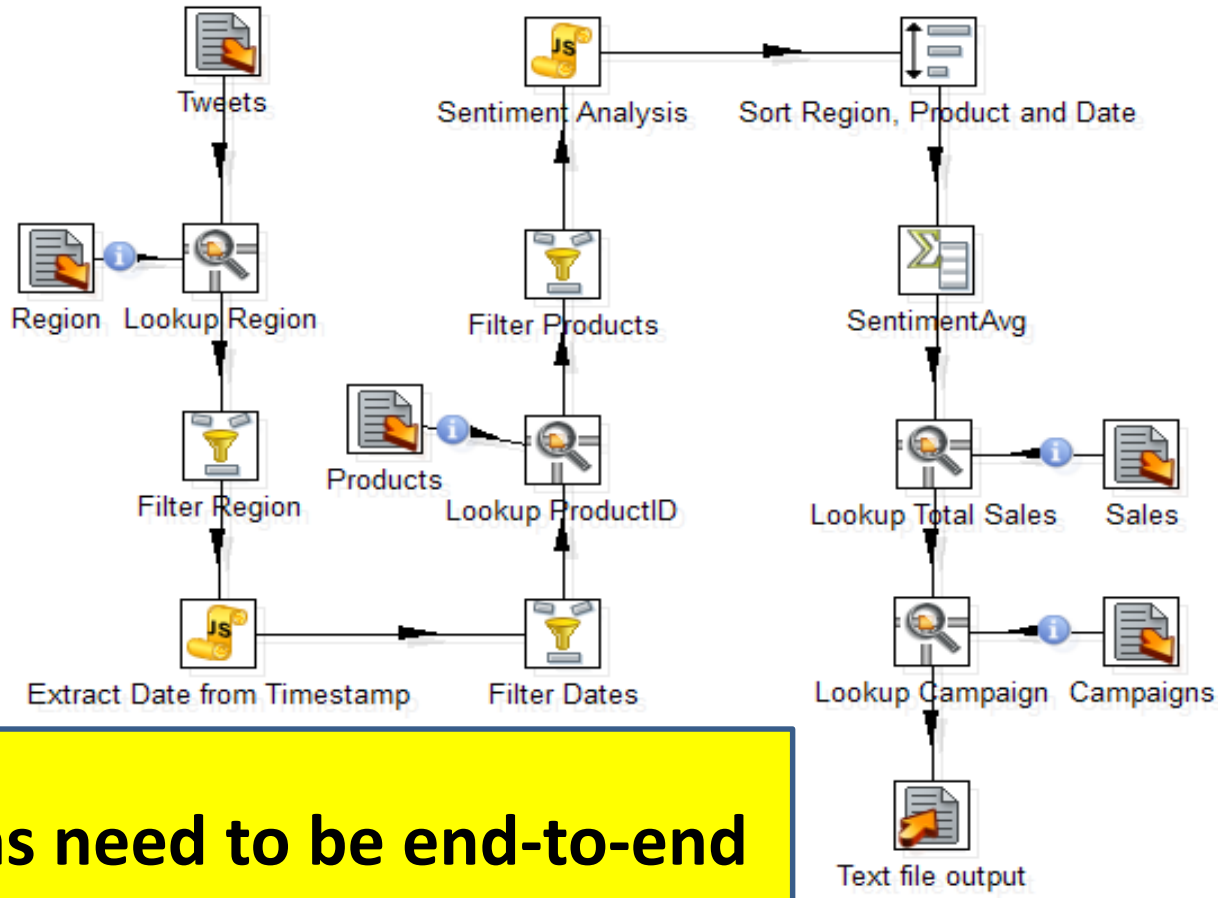


# Applying BDA in practice – tip 1




*from "Data Mining: Concepts and Techniques"*

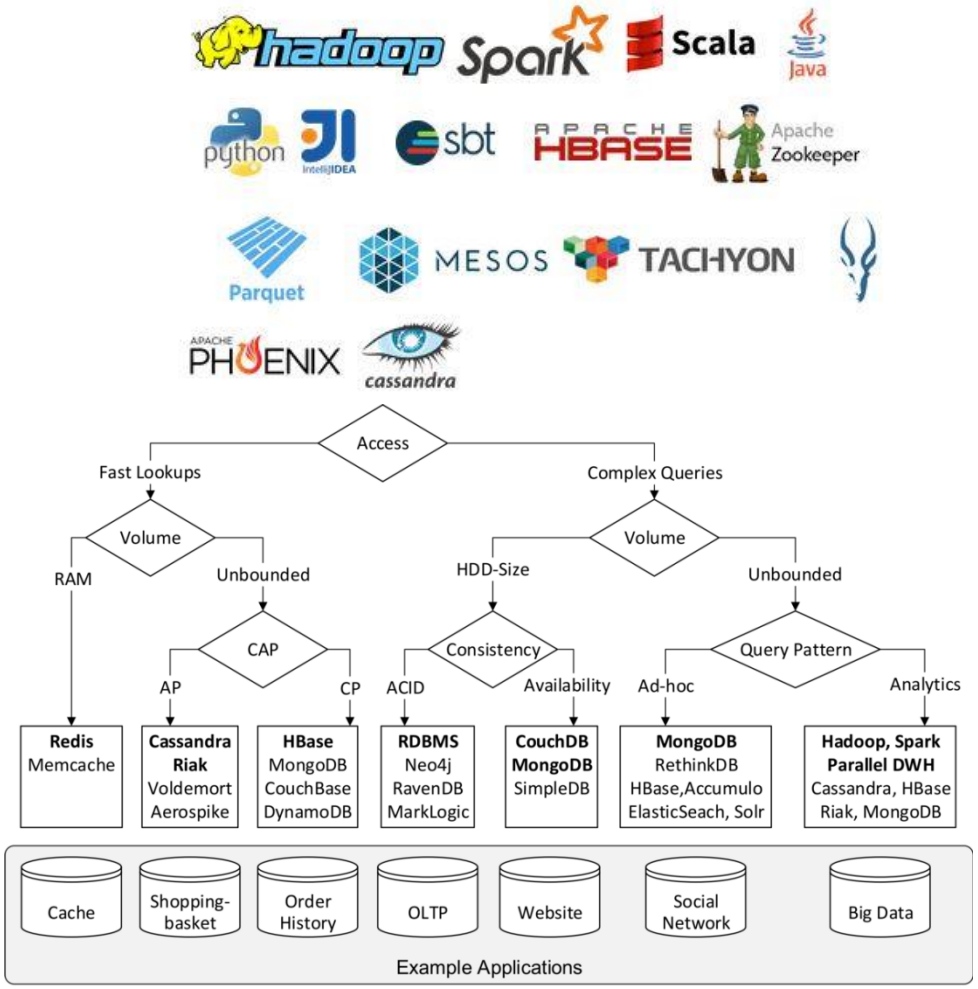
# Applying BDA in practice – tip 1




**Solutions need to be end-to-end**

# Applying BDA in practice

- ✓ Tip1: End-to-end solutions
- ✓ Tip2: Algorithms
  - ❖ Rapid prototyping
  - ❖ Sound and
  - ❖ scalable solutions
- ✓ Tip3: System dev/code
- ✓ Tip4: Take a HPC approach




*from https://adtmag.com/*

# Forbes: τα 50 Καλύτερα Επαγγέλματα του 2017

1. **Data Scientist**
2. **Software Development and Information Technology Operations**
3. **Data Engineer**
4. Tax Manager
5. **Analytics Manager**
6. HR Manager
7. **Database Administrator**
8. Strategy Manager
9. UX Designer
10. Solutions Architect
11. Marketing Manager
12. Occupational Therapist
13. Audit Manager
14. Electrical Engineer
15. Nurse Practitioner
16. **Software Engineer**
17. Corporate Recruiter
18. Supply Chain Manager
19. Finance Manager
20. Mechanical Engineer
21. Communications Manager
22. QA Manager
23. Controls Engineer
24. Nurse Manager
25. Compliance Manager
26. **Mobile Developer**
27. **Systems Administrator**
28. Executive Assistant
29. Technical Account Manager
30. Hardware Engineer
31. Intelligence Analyst
32. Product Manager
33. Professor
34. Physician
35. Compensation Analyst
36. Civil Engineer
37. Research Engineer
38. Accounting Manager
39. Information Security Engineer
40. Manufacturing Engineer
41. Security Analyst
42. Business Operations Manager
43. UI Designer
44. Pharmacy Manager
45. Supplier Quality Engineer
46. National Sales Manager
47. Business Intelligence Developer
48. Dental Hygienist
49. Physical Therapist
50. Construction Project Manager

Forbes Magazine

(http://www.forbes.com)

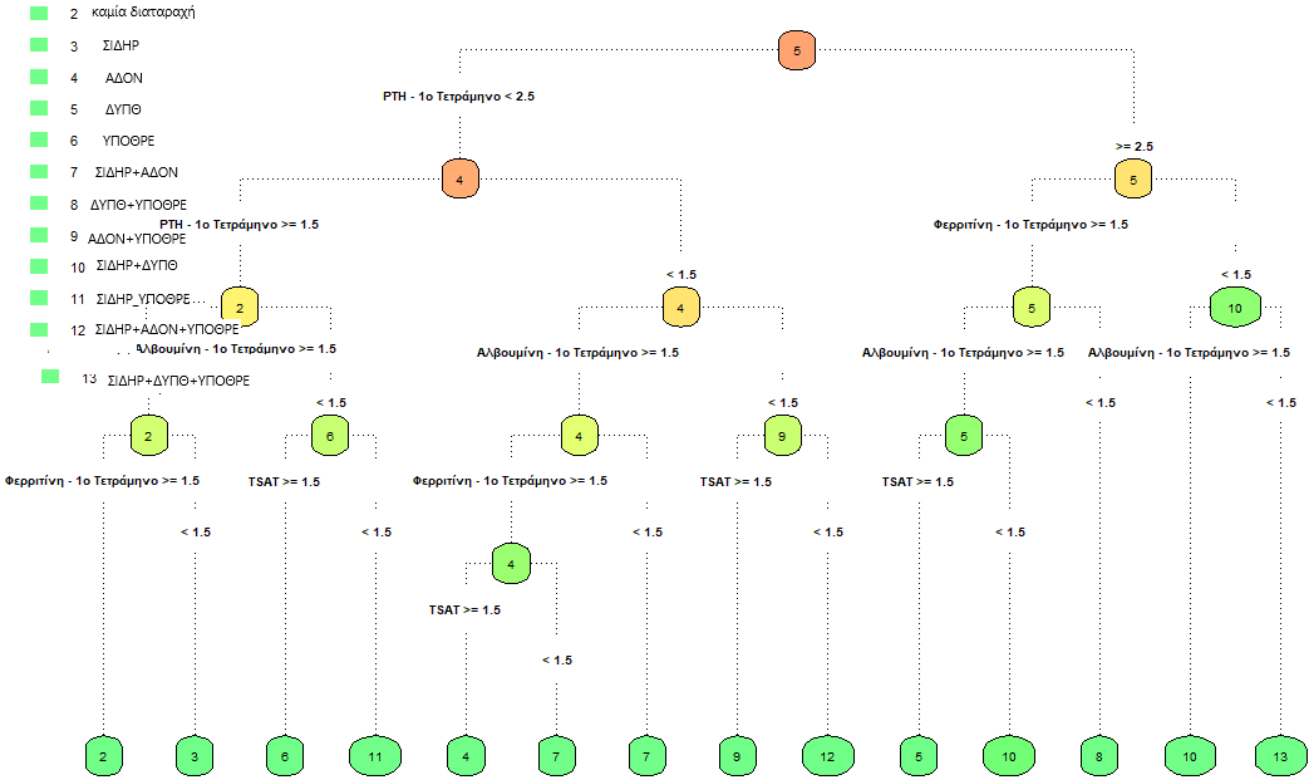
# Case study 1: e-shop recommendations

- You may or may not have a profile of customers.
- More purchases do not necessarily mean higher profit.
- Issues:
  - Initial data
  - Offline vs online processing
  - Multiple Algorithms (existing plus extensions)
  - Low budget

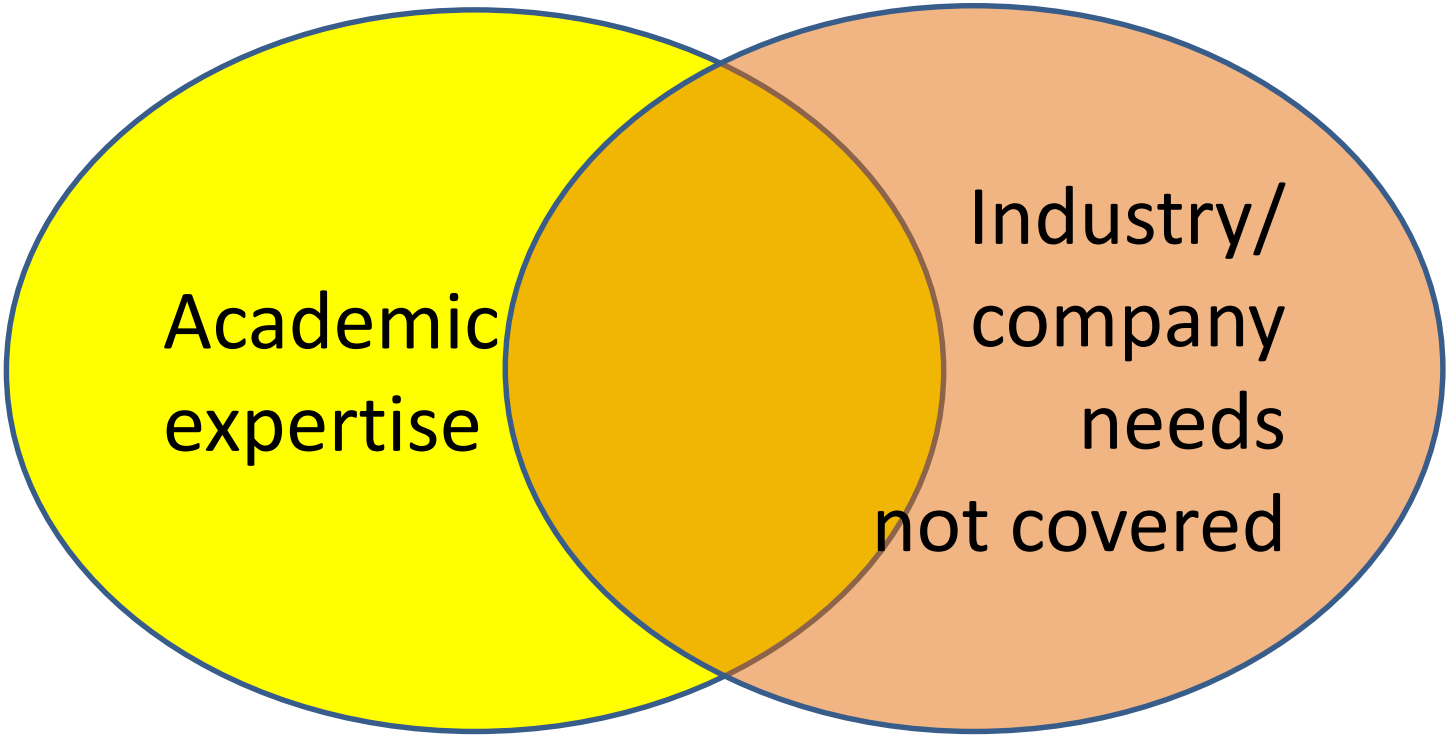
# Case study 2: logs analysis

- Online report of arbitrary event combinations.
- Millions of interesting events per day.
- Additional Issues:
  - Manage to perform/run $O(n^2)$ and $O(n^3)$ algorithms in a practical manner

# Case study 3: chronic kidney disease

- Trade-offs between quality in terms of accuracy and interpretability and usability.
- 

# My view about the role of academia



# Final/Summary comments

- Big data analytics is a not trivial multi-objective problem.
- Practical solutions need to account for end-to-end data processing.
  - Combining small solutions to each isolated step does not give the full solution.
- No one size fits all (refers to both algorithms/techniques and tools).
  - A solution may not comprise one size only.
- Academia can offer a lot to interested industries/companies.

# Thank you

Questions?

*Email: gounaria@csd.auth.gr*