Co-funded by the Erasmus+ Programme of the European Union

Pathway in Enterprise Systems Engineering (PENS)

(Big) Data, Data Analytics, Business Intelligence

Opportunities and Risks

Giorgio Giacinto

July, 25th 2018
Acalà de Henares

Big data – Smart *

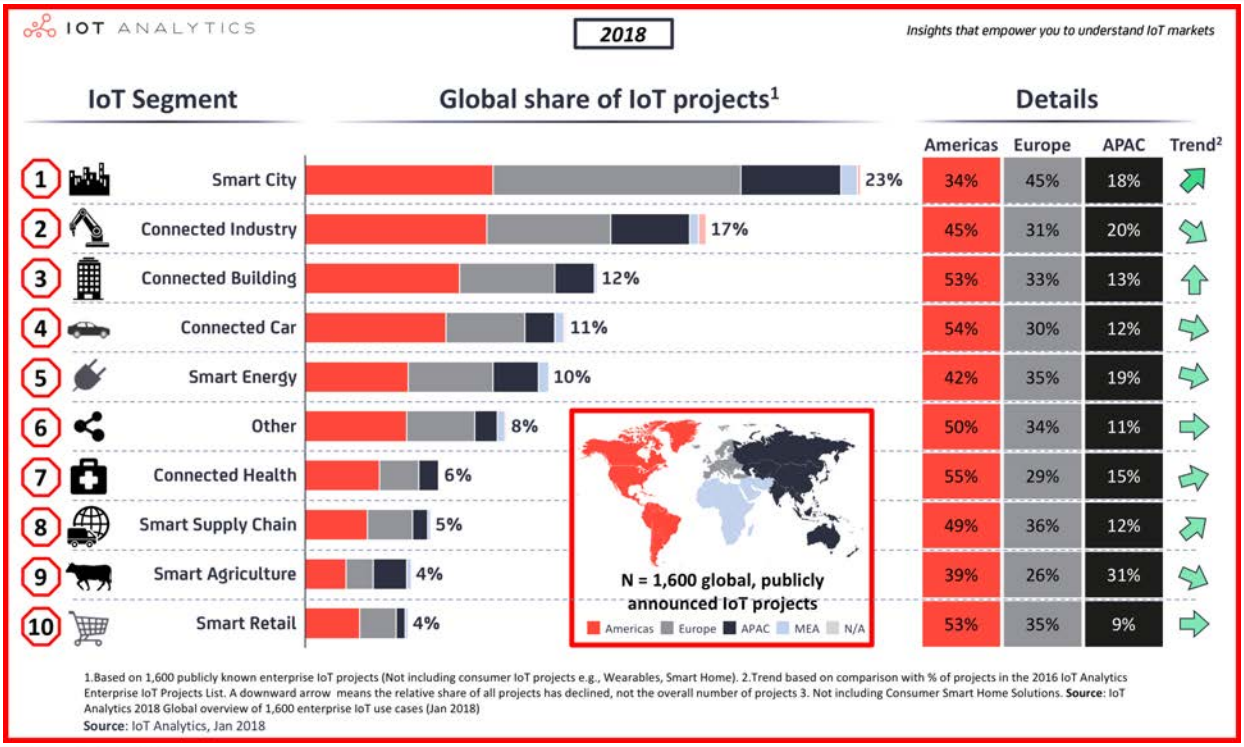
History of Big Data

- Nice overview at the WinShuttle website
  - **'70s** - The widespread adoption of relational DBMS for data storage and management
  - **'80s** - The growth of communication capabilities
    - text and multimedia data
  - **early '90s** - Enterprises start adopting ERPs, CRMs, as well as developing Decision Support Systems
  - **late '90s** - The explosion of the WWW
    Data Mining
  - **early 2000** – Explosion of social media
    Business Intelligence
  - **late 2000** – Virtualisation & cloud & IoT
    Advanced Business Intelligence

Data sources

- **Business processes**
  - sales, revenues, enterprise systems (ERP, CRM, etc.)
- **The web**
  - text, multimedia
- **IoT**
  - connection with the physical world
- **Wereables**
  - sense our body, what we do
- **Cloud**
  - enables the connection of anything to anything

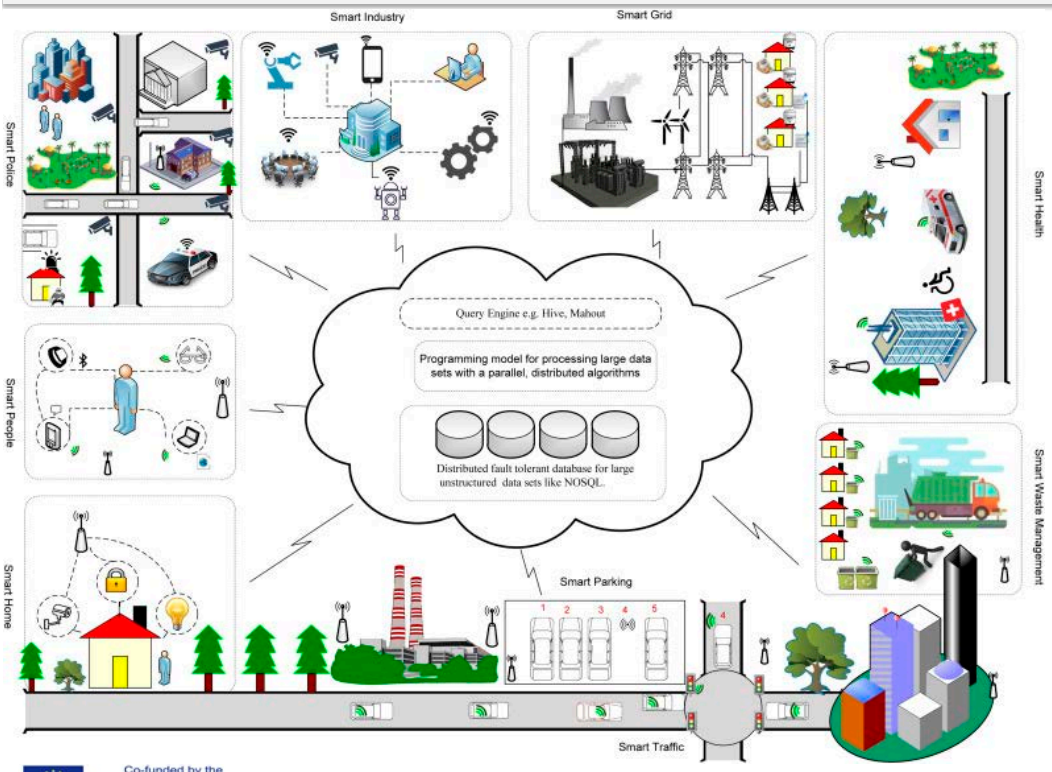
Big Data - The IoT driver



Co-funded by the Erasmus+ Programme of the European Union

https://iot-analytics.com/top-10-iot-segments-2018-real-iot-projects/

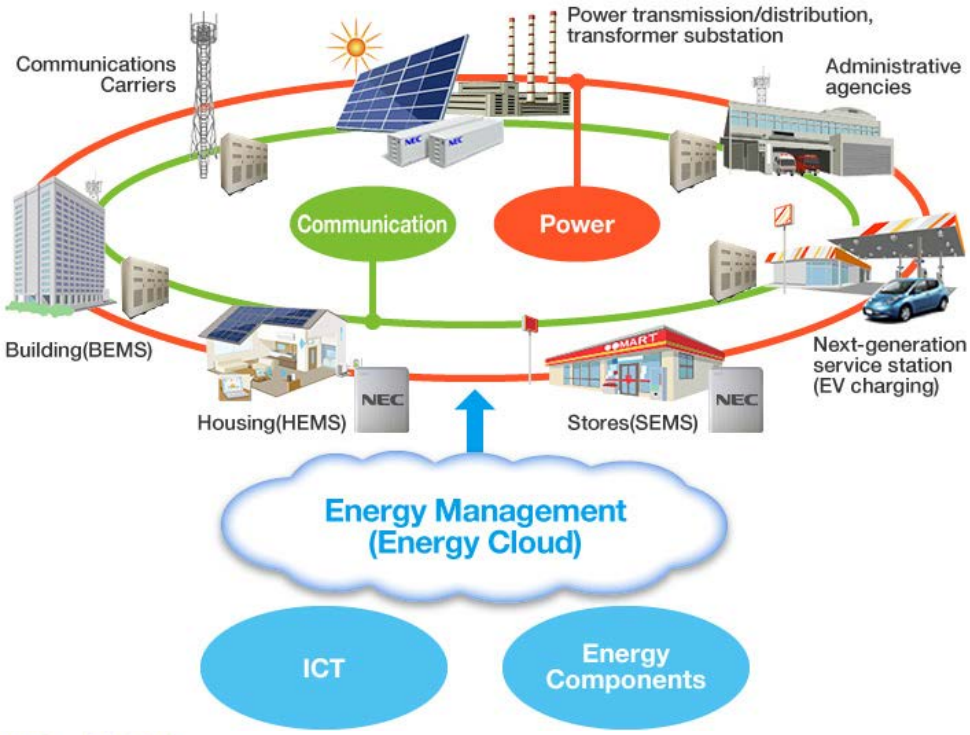
Big Data and Smart City



The role of big data in smart city, International Journal of Information Management, 2016.

Co-funded by the Erasmus+ Programme of the European Union

Smart Energy



Manipulating Big Data

- **Business Analytics**
  Tools to explore past data to gain insight into future business decisions.
- **Business Intelligence (BI)**
  Tools and techniques to turn data into meaningful information.
- **Big Data**
  data sets that are so large or complex that traditional data processing applications are inadequate.
- **Data Mining**
  Tools for discovering patterns in large data sets.

Operational Vs. BI and Analytics

Operational Systems	BI and Analytics
Normalized models are standard for OLTP	Dimensional models are standard for BI and OLAP
Highly volatile	Generally not updated
Transaction throughput (updating and maintaining numerous records) is critical	Query performance (gathering and aggregating large sets of records) is critical
Characteristics supporting use of normalized models	**Characteristics supporting use of dimensional models**
Minimal redundancy (normalization)	Increased redundancy (denormalization)
Limited index use	Increased index use
Efficient use of storage space	Increased storage space
Eliminate inconsistent data	Consolidate inconsistent data
Few maintenance concerns	Increased maintenance issues

Rick Sherman, “Business Intelligence Guidebook”, Morgan Kaufmann, 2014

NoSQL distilled by Sadalage and Fowler
https://www.martinfowler.com/articles/nosqlKeyPoints.html

Data models

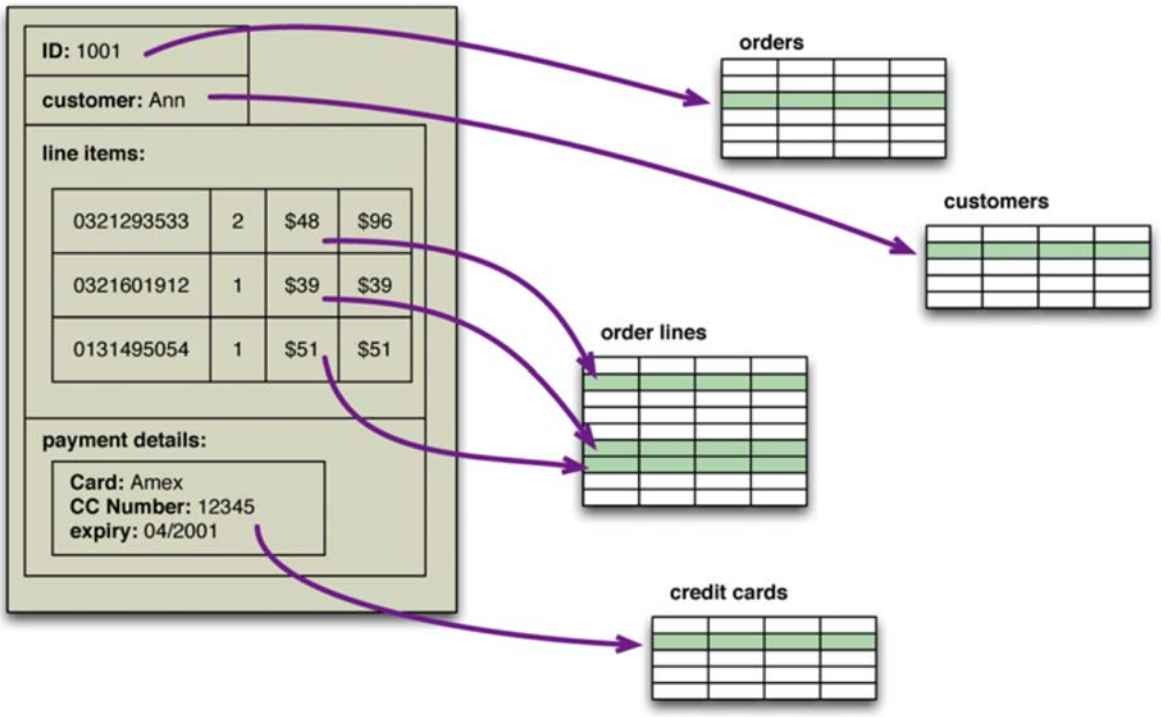
New data models

- Google, Amazon, Facebook the big players
- **Between 2000 and 2010** they were investigating new ways to store and query the vast amount of data they were collecting
  - Google ➔ BigTable
  - Amazon ➔ DynamoDB
  - Facebook ➔ Cassandra
- The **NoSQL** movement is officially established in 2009 NotOnlySQL

What a relational data model is good at

- Data persistence
- Same data model for multiple applications
- Transaction support
- Easy Portability and Maintenance thanks to the reliance on the SQL standard

Example of the impedance mismatch with the relational model



Relational Vs. non-relational models

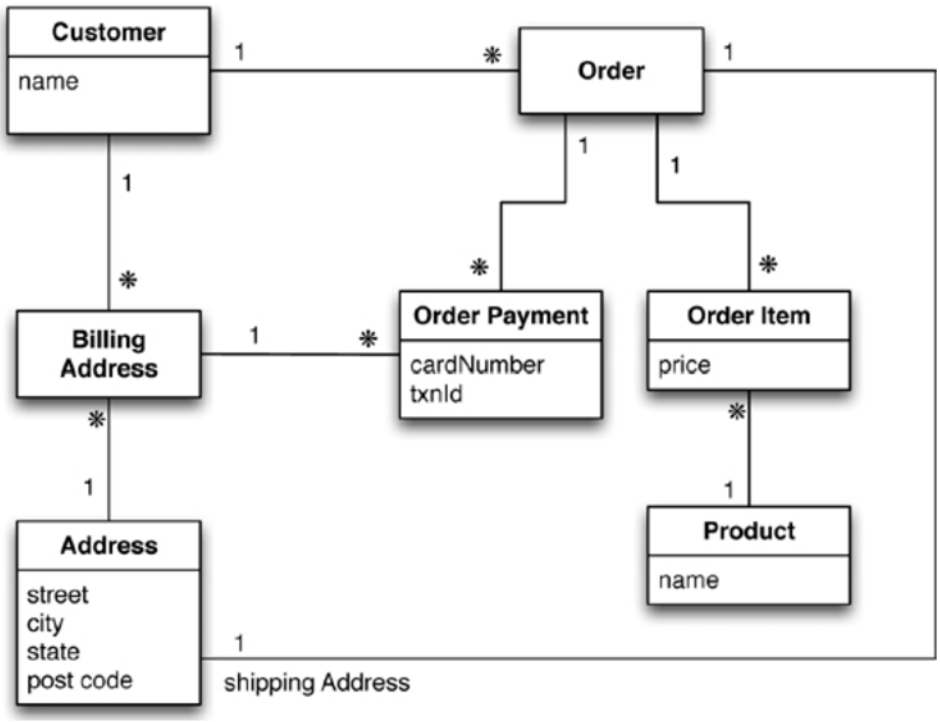
- The **relation** model is good at **integration**
  - One database for many different applications
- Application oriented databases
  - The storage model is optimised against one application
  - Data exchange between different application via XML and JSON documents
  - Consistency has to be verified at the application level

Aggregate data models

- The relational model represents entities as tuple of atomic attribute values
- An aggregate is a collection of objects in relationship with each other
  - the aggregate is considered as a unique object

Co-funded by the Erasmus+ Programme of the European Union

A possible relational model for an e-commerce website



Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering

RDBMS for an e-commerce website

Customer

Id	Name
1	Martin

Order

Id	CustomerId	ShippingAddressId
99	1	77

Product

Id	Name
27	NoSQL Distilled

BillingAddress

Id	CustomerId	AddressId
55	1	77

OrderItem

Id	OrderId	ProductId	Price
100	99	27	32.45

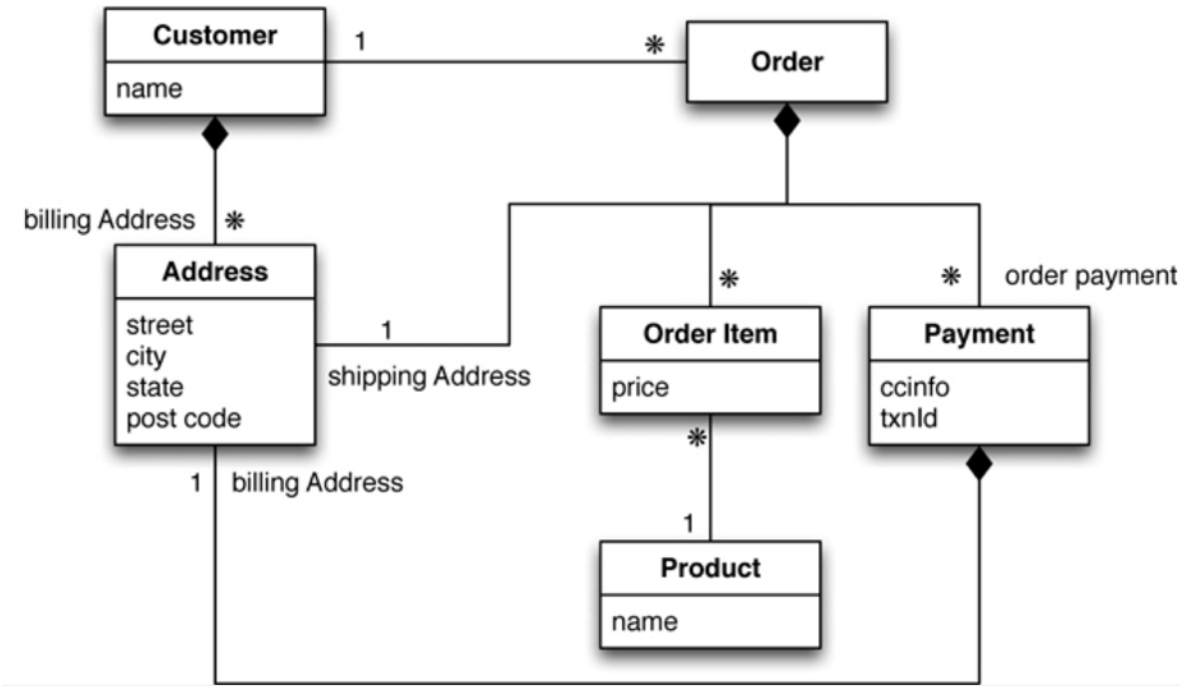
Address

Id	City
77	Chicago

OrderPayment

Id	OrderId	CardNumber	BillingAddressId	txnId
33	99	1000-1000	55	abelif879rft

A possible model based on aggregation for the e-commerce website



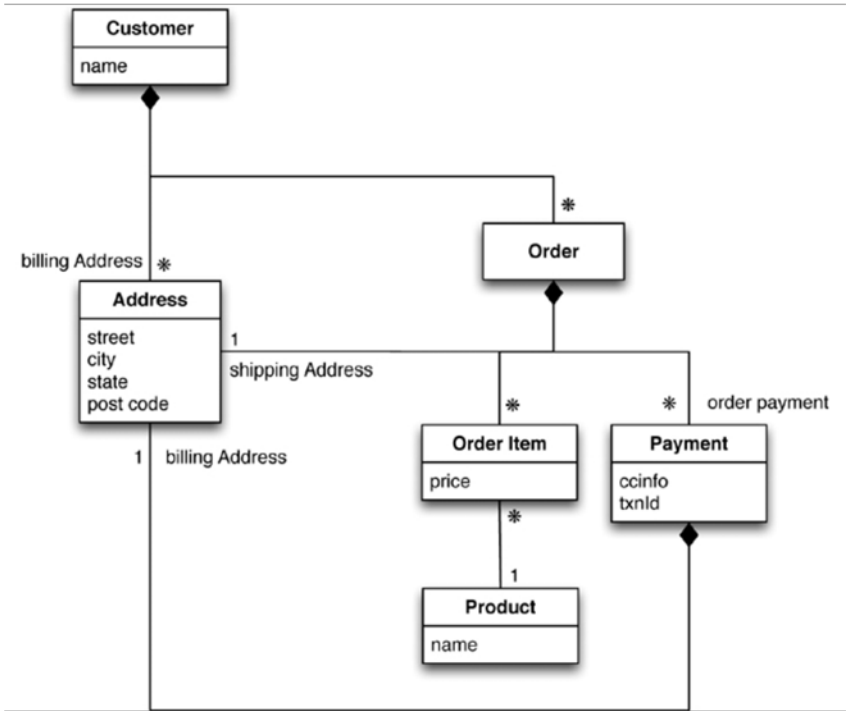
JSON definition

redundancies

```
// in customers
{
"id":1,
"name":"Martin",
"billingAddress":[{"city":"Chicago"}]
}

// in orders
{
"id":99,
"customerId":1,
"orderItems":[
  {
  "productId":27,
  "price": 32.45,
  "productName": "NoSQL Distilled"
    }
  ],
"shippingAddress":[{"city":"Chicago"}]
"orderPayment":[
  {
    "ccinfo":"1000-1000-1000-1000",
    "txnId":"abelif879rft",
    "billingAddress": {"city": "Chicago"}
  }
  ],
}
```


Another possible model based on aggregation

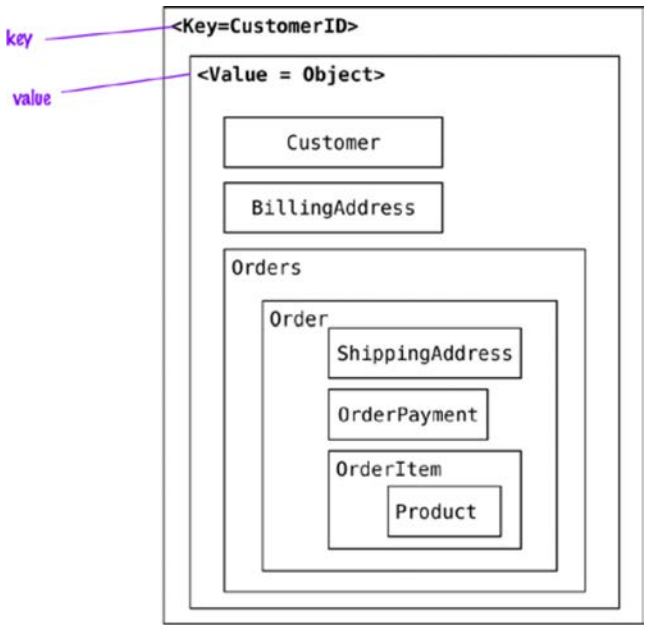


JSON definition

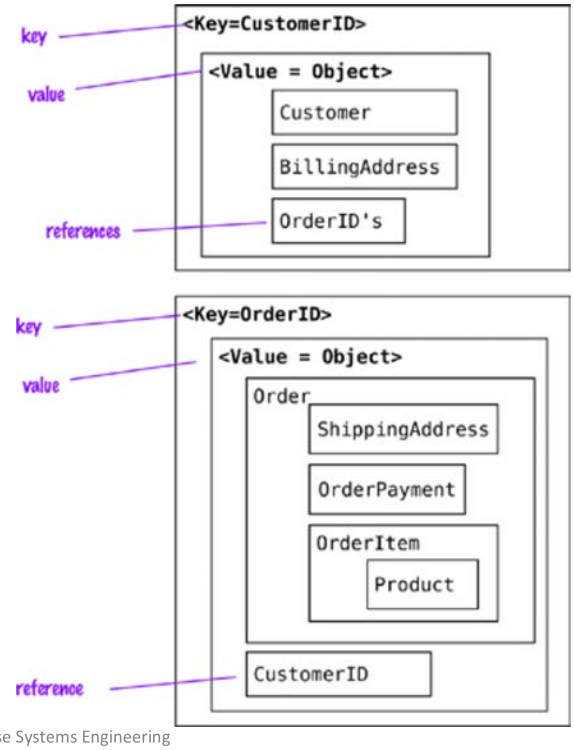
```
// in customers
{
"customer": {
"id": 1,
"name": "Martin",
"billingAddress": [{"city": "Chicago"}],
"orders": [
  {
    "id":99,
    "customerId":1,
    "orderItems":[
    {
    "productId":27,
    "price": 32.45,
    "productName": "NoSQL Distilled"
    }
  ],
  "shippingAddress":[{"city":"Chicago"}]
  "orderPayment":[
    {
    "ccinfo":"1000-1000-1000-1000",
    "txnId":"abelif879rft",
    "billingAddress": {"city": "Chicago"}
    }],
}]
}
}
```


Queries

All objects in one aggregate



Two aggregates for orders and customers



The relational data model Vs. aggregate data models

- The relational model represents the aggregation property through *foreign keys*
  - This model allows designing any kind of query
  - The optimizer finds the best way to produce the results

- An aggregate data model is tailored to specific queries
  - low efficiency in running arbitrary queries

Four NoSQL data models

- **Key-Value**
  - A *blob* of data can be retrieved by specifying the value of one *key* attribute
- **Document**
  - Structured documents, allowing for simple queries on individual attributes.
- **Column family**
  - The most similar to the relational data model
- **Graph**
  - Relationships are represented by a graph

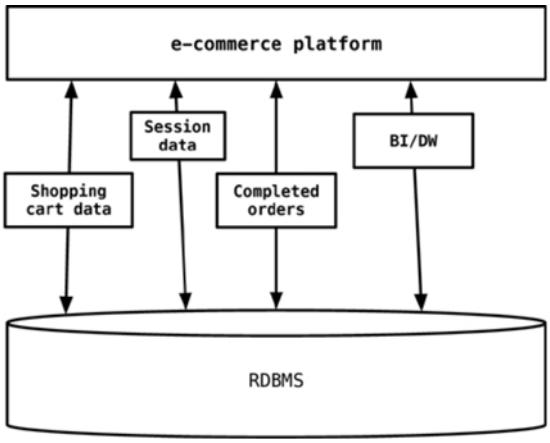
Summary on Data models

- The **relational** data model
  - Fixed schema
  - Allows data correlation
  - Powerful query language

- **Non relational** data models
  - Good for easy storage and retrieval
  - Heterogeneous data schemas are allowed

Co-funded by the Erasmus+ Programme of the European Union

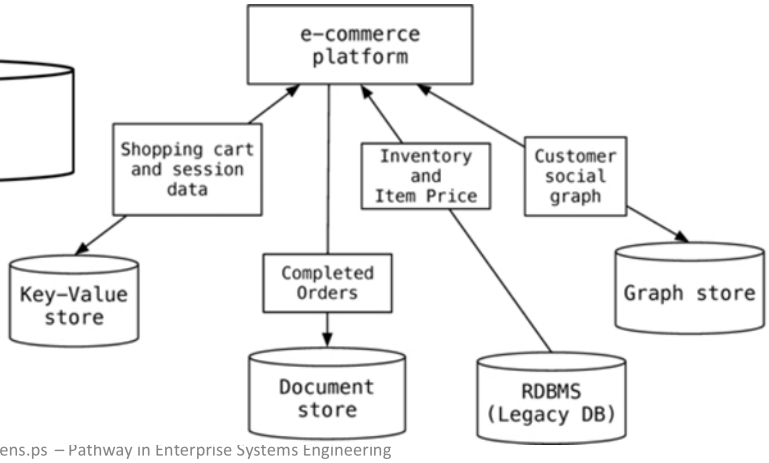
Polyglot persistence



All data stored in a Relational DBMS

Polyglot persistence: different data models according to their use

Redundancy



Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering

Gartner 2013
Magic Quadrant Operational DBMS

Co-funded by the Erasmus+ Programme of the European Union

Gartner 2014
Magic Quadrant Operational DBMS

Co-funded by the Erasmus+ Programme of the European Union

Gartner 2015
Magic Quadrant Operational DBMS

Co-funded by the Erasmus+ Programme of the European Union

Gartner 2016
Magic Quadrant Operational DBMS

Co-funded by the Erasmus+ Programme of the European Union

Gartner 2017
Magic Quadrant Operational DBMS

Co-funded by the Erasmus+ Programme of the European Union

Rick Sherman, “Business Intelligence Guidebook”
Morgan Kaufmann, 2014

Business Intelligence

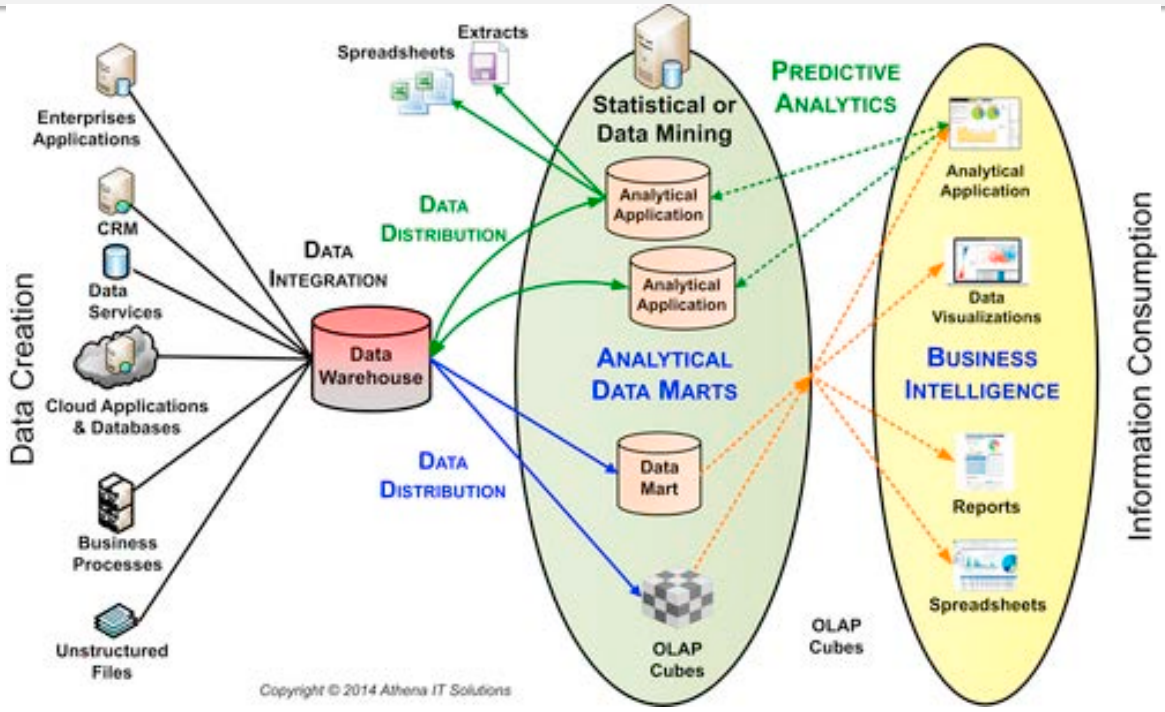
Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering

Scope and define the predictive analytics project

- What business outcomes are you trying to effect?
- What business processes, external events, and factors, such as economic or demographics, will you analyze as part of the initiative?
- Who (people) and how (business processes) will the predictive models be used?

The Process



Explore and profile your data

- Considerable effort is required to determine the **data** that is needed for the project
  - where it is **stored**
  - whether it is readily **accessible**
  - its **completeness** and **quality**
- *It is common to use data that is incomplete or has quality issues simply because it is the best that can be obtained.*
- The data scientist may need to **adjust the models** to **compensate for the data quality**

Issues in data completeness and quality

- Different databases related to the same ground truth may exhibit
  - different **data types** for the same attributes
  - different **values** for the same instances
  - different **semantics** for the same values

- **Missing values**
  - One data source may miss values that are available in other data sources

Gather, cleanse, and integrate the data

- Once the necessary data is located and evaluated, then it has to be turned into a **clean**, **consistent** and **comprehensive** set of information
  - Integration of heterogeneous data sources
  - Use of unstructured or semi-structured data
- Sometimes data need to be **synthesized** or created to be used as input to the predictive models.
  - The data scientist may need to create separate predictive models to generate the input data needed for the primary predictive model.

Build the predictive models

- Models are created and the underlying hypotheses **tested** through steps
  - **including** and **ruling out** different variables and factors
  - **back-testing** the models against historical data
  - determining the **potential business value** of the analytical results produced by the models
- This is a highly iterative process.
  The modeler may uncover the **need for additional data** and **data integration** to develop a more robust model

Monitor the models and measure their business results

- Predictive models need to **adapt** to changing business conditions and data.
- The results of predictive models need to be **tracked** to know
  - which models are providing **the most value** to your organization
  - which **model's value** starts to **decline**.
- With increasing data sources and volume, predictive model performance data, and additional business insights, **new or modified models** are likely to emerge

Data Mining tasks

- **Association rules**
  - e.g., products that are frequently bought together
- **Behavioral sequential models**
  - e.g., forecasting sequential purchases
- **Classification trees**
  - e.g., customer profiling according to the frequency they go shopping, the quantities of products, etc-A

Examples of BI processes

Technique	Example
Statistics	Use for customer segmentation.
Predictive modeling	Create fraud detection models for credit cards.
Forecasting	Create sales forecasts for each product category and country including seasonality and weather.
Data mining	Determine college freshmen retention rates based on demographic and academic attributes.
Descriptive modeling	Split customers into categories by their product preferences and stage of life (age, children, marital status, working, etc.)
Econometrics	Determine impact of economy and US Federal Reserve's bond buying policy on job postings and hires.
Operations research	Determine the flow of raw materials in supply chain using variability of demand and supply.
Optimization	Determine the best routes for delivery trucks.
Simulation	Determine what the impact is to customers' loyalty and market share from pricing changes.
Textual analytics	Perform a sentiment analysis on the introduction of product line extensions and new categories.

Co-funded by the Erasmus+ Programme of the European Union

Visualization and Business Intelligence

Data Analytics

Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering

Gartner 2018
Magic Quadrant Data Analytics and BI

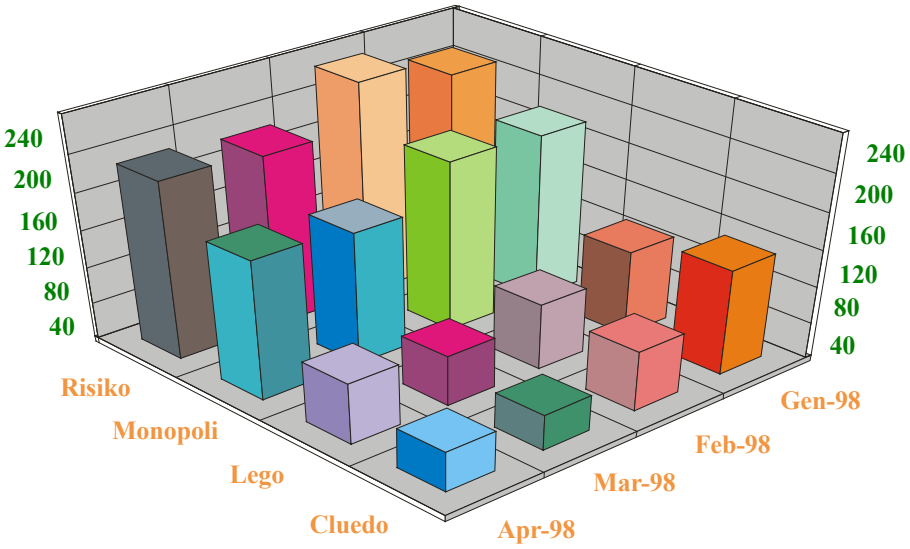
Data visualisation

- Powerful tool
- Usually referred to “data analytics”
- Allows getting a more quick and deep understanding of the data
  - The human mind is more clever with graphs, colors etc. than with words

- Different commercial and open-source tools available

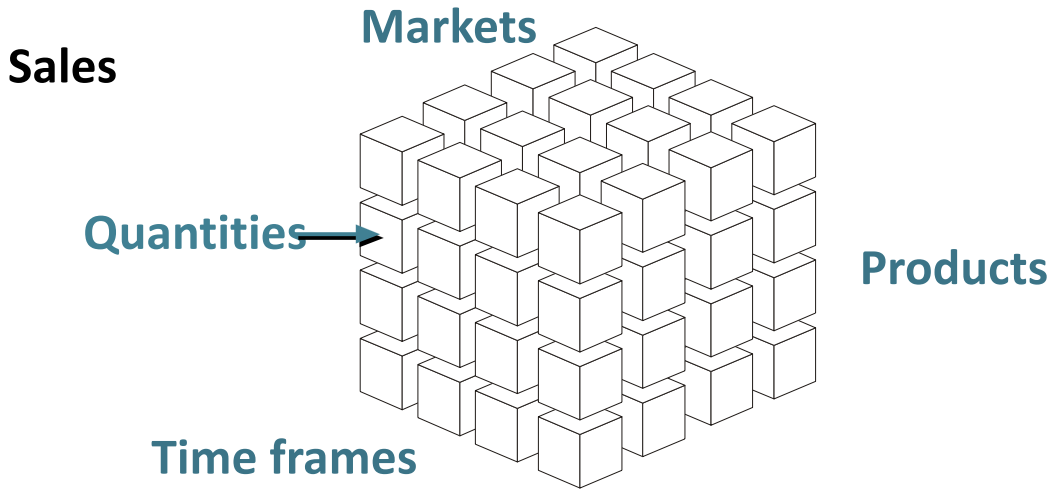
Simple visualisation tools

Monthly sales of games in Rome



Co-funded by the Erasmus+ Programme of the European Union

Multidimensional analysis

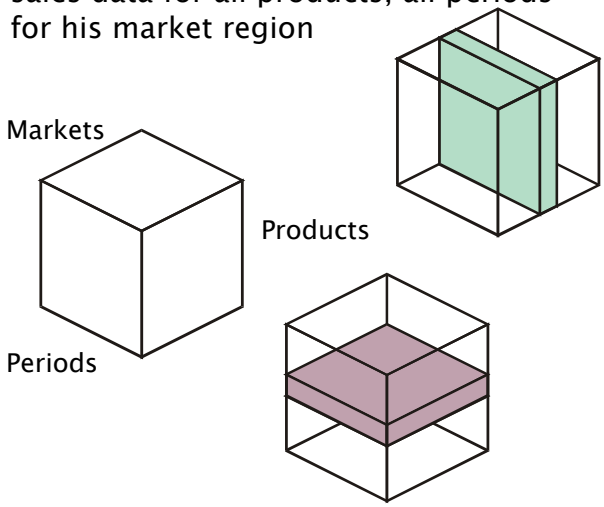


Co-funded by the Erasmus+ Programme of the European Union

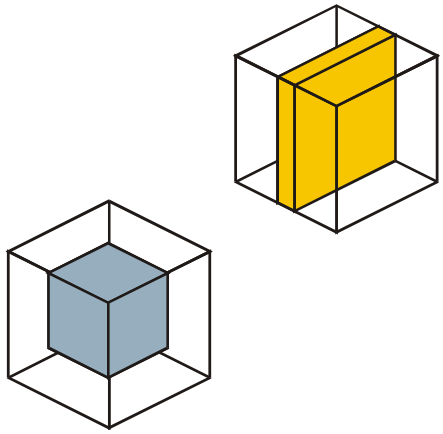
http://www.pens.ps – Pathway in Enterprise Systems Engineering

Views on multidimensional data

The regional manager analyses sales data for all products, all periods for his market region



The financial manager analyses sales data for all products and markets related to the current period and compares to the previous period



The product manager analyses the sales of one product for all periods and all markets

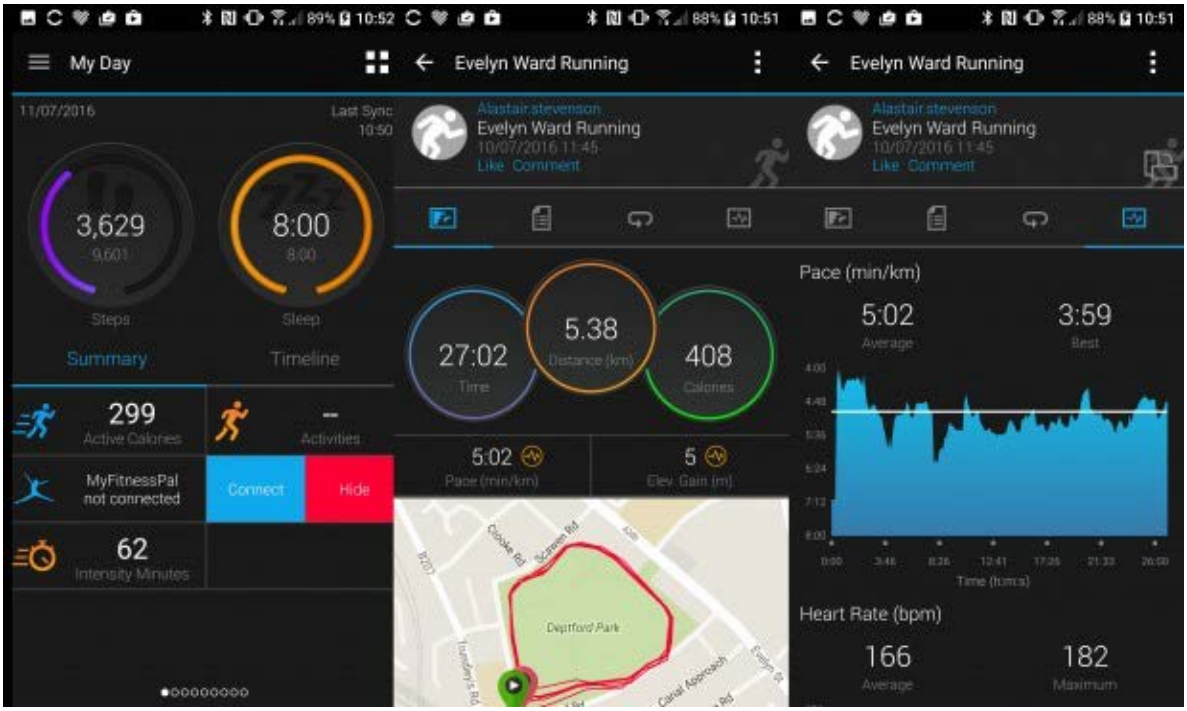
The strategic manager focuses on a product category, a geographical area and a range of periods

Kibana

Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering

Sport apps



Machine Learning

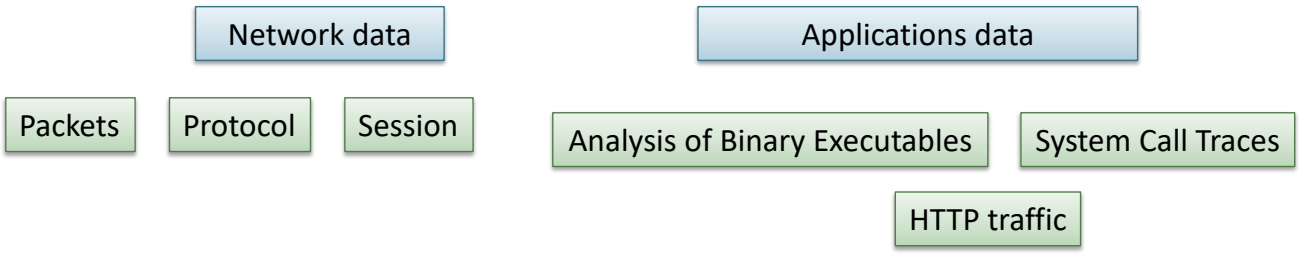
Machine Learning

- Supervised techniques
  - Labelled data
- Unsupervised techniques
  - clustering

What's Machine Learning?

- Building machines that can automatically perform *tedious* classification tasks *with high accuracy*.

- Learning to classify… what does it mean?
  - Sensing



What's Machine Learning?

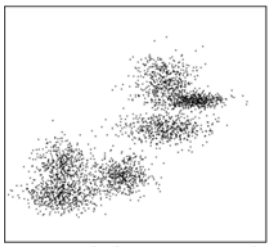
- Building machines that can automatically perform *tedious* classification tasks *with high accuracy*.

- Learning to classify… what does it mean?
  - Sensing
  - Measurements extraction
    - Packet/Byte statistics
    - System Call statistics
    - Session Analysis
    - HTTP/DNS request rates
    - …

What's Machine Learning?

- Building machines that can automatically perform *tedious* classification tasks *with high accuracy*.

- Learning to classify… what does it mean?
  - Sensing
  - Measurements extraction
  - Choice of a data model
    - e.g., patterns (i.e., packets, sessions, binary executables, etc.) are represented as points of a multidimensional feature space



Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering

What's Machine Learning?

- Building machines that can automatically perform *tedious* classification tasks *with high accuracy*.

- Learning to classify… what does it mean?
  - Sensing
  - Measurements extraction
  - Choice of a data model
  - Automatically extract the *decision boundaries* in the data space
    - This is the *learning* step – we need a set of *training* data
    - Statistical Decision theory (Bayes)
    - Optimization

What's Machine Learning?

- Building machines that can automatically perform *tedious* classification tasks *with high accuracy*.

- Learning to classify… what does it mean?
  - Sensing
  - Measurements extraction
  - Choice of a data model
  - Automatically extract the *decision boundaries* in the data space
  - Generalization
    - The ability of the machine to classify never-seen-before data

Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering

When Machine Learning is used

- Automatic classification
  - High speed
  - High accuracy
- Big data
  - A lot of data to analyze
  - Data described by a very large number of features
    - Pattern discrimination requires complex rules

Machine Learning is suited for performing some Computer Security tasks!

Machine Learning for Computer Security

- Network traffic analysis and classification
  - Intrusion Detection
  - Detection of malware generated HTTP traffic
  - Analysis of DNS query/response pairs for botnet detection
- Classification of binary files for virus detection
- Common characteristics
  - Large number of parameters used to describe patterns
  - Classification is performed by non-trivial rules
- Machine Learning techniques can cope with polymorphism
- …ML techniques can be a (easy) target for attackers
  - To mislead or evade detection

Machine Learning Approaches

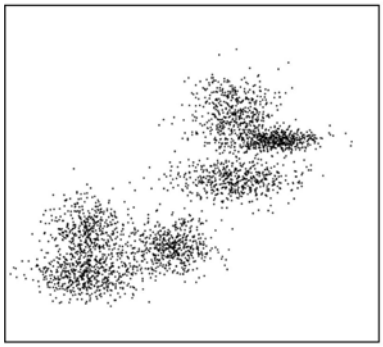
- Unsupervised classification
  - Clustering, i.e., detecting natural grouping of patterns in the feature space
    - No training set is needed

- Supervised classification
  - A (training) set of labeled patterns from different data classes is used to learn a discriminating function
  - You must to reliably label a significant number of patterns

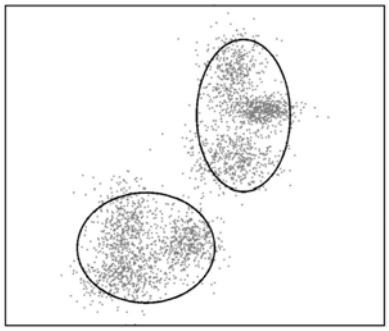
Co-funded by the Erasmus+ Programme of the European Union

Clustering

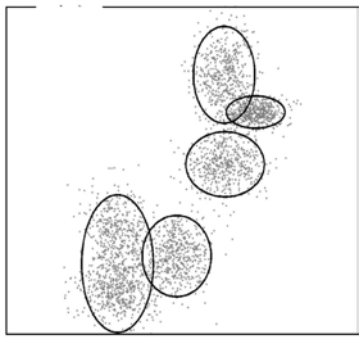
Jain, Data Clusetering: 50 years beyond K-means, Patter Recognition Letters, 2010



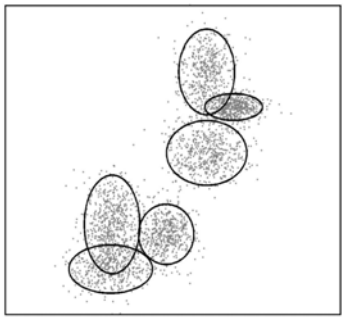
(a) Input data



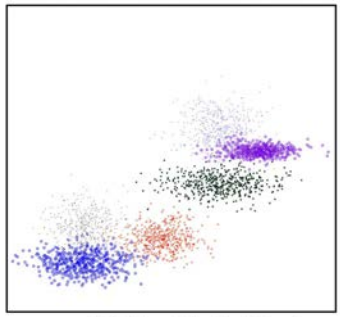
(b) GMM (K=2)



(c) GMM (K=5)



(d) GMM (K=6)



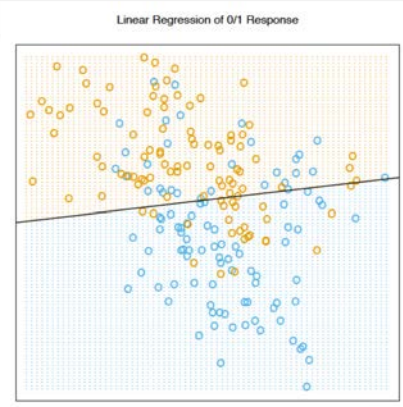
(e) True labels, K = 6

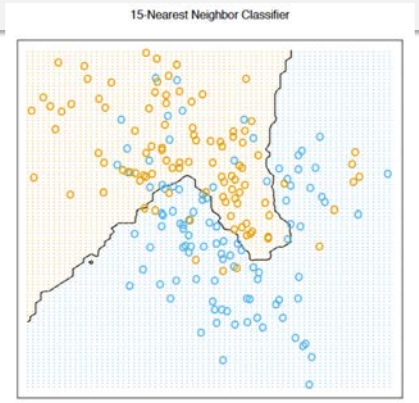
Co-funded by the Erasmus+ Programme of the European Union

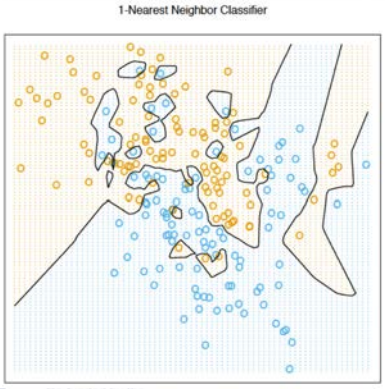
http://www.pens.ps – Pathway in Enterprise Systems Engineering

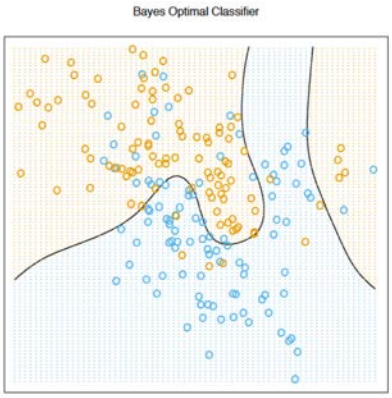
Supervised classification

Hastie, Tibshirani, Friedman, The Elements of Statistical Learning, Springer, 2008









Conclusion

Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering

What's next

- Data is available almost for free
- The risk is to consider any kind of data as a *fact*
- Data science is emerging as one key factors in business development
- Understanding the basics of clustering, classification, and prediction models, allows emphasizing the role of data collection, cleaning and aggregation.

Co-funded by the Erasmus+ Programme of the European Union

email@provider.com

Thank you for your attention!

Co-funded by the Erasmus+ Programme of the European Union

http://www.pens.ps – Pathway in Enterprise Systems Engineering