Business Analytics

Fabrice Rossi

CEREMADE
Université Paris-Dauphine

2021

What is Business Analytics?

A BA story...

Target

- 8th-largest retailer in the United States
- collects data about its customers
  - (as everybody!)
  - unique customer identifier
  - personal information (email, mailing address, etc.)
  - complete shopping history

A BA story...

Target

- 8th-largest retailer in the United States
- collects data about its customers
  - (as everybody!)
  - unique customer identifier
  - personal information (email, mailing address, etc.)
  - complete shopping history
- *baby-shower registry*
  - pregnancy score from key products
  - due date estimation
  - adapted coupon program
  - for details

TARGET

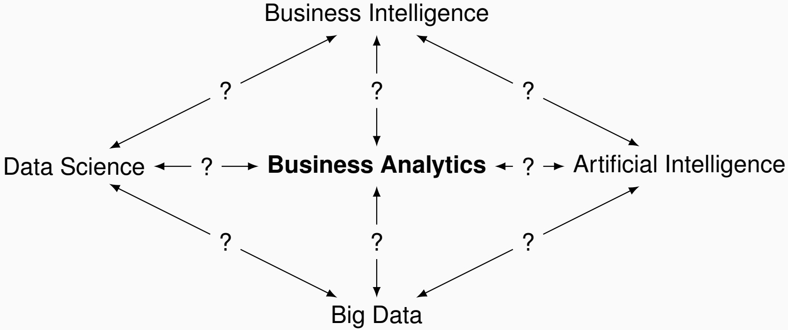
Business Intelligence

Data Science

Business Analytics

Artificial Intelligence

Big Data


Business Intelligence
?
?
?
Data Science
?
Business Analytics
?
Artificial Intelligence
?
?
?
Big Data

Data Science

Definition

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data (Wikipedia).

Data Science

Definition

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data (Wikipedia).

Data

Data is a set of values of subjects with respect to qualitative or quantitative variables (Wikipedia).

Data Science

Definition

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data (Wikipedia).

Data

Data is a set of values of subjects with respect to qualitative or quantitative variables (Wikipedia).

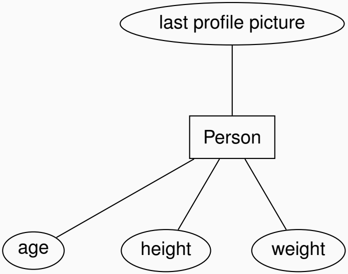
Person

- age
- height
- weight
- last profile picture

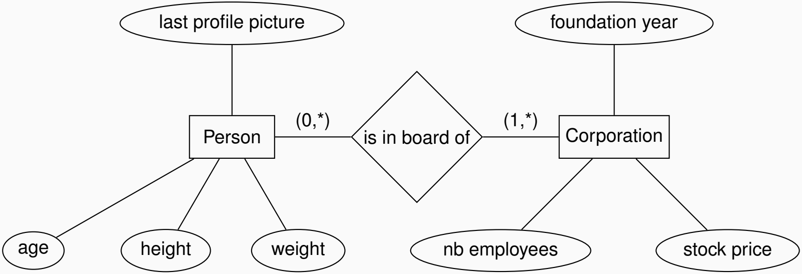
Corporation

- board of directors
- number of employees
- stock price
- foundation year

Entity relationship diagram



Entity relationship diagram



Example: marketing campaign

age	gender	employment	csp_42	family	diploma	code_insee	target
53	Female	ce_2_1	csp_2_2	m_4_1	d_1_7	01004	failure
85	Female	NA	csp_7_7	m_1_2	d_1_2	01004	failure
55	Male	ce_1_6	csp_4_8	m_4_1	d_1_3	01010	success
45	Male	ce_2_1	csp_4_3	m_4_1	d_1_6	01032	failure
54	Male	ce_1_6	csp_6_7	m_4_1	d_1_3	01046	success
32	Male	NA	csp_8_5	m_4_4	d_1_3	01053	success
41	Male	NA	csp_6_2	m_1_1	d_1_3	01053	failure
18	Male	NA	csp_8_4	m_4_1	d_1_3	01053	failure
45	Male	NA	csp_4_7	m_1_1	d_1_6	01053	failure
65	Female	NA	csp_7_5	m_4_4	d_1_3	01053	failure
49	Female	ce_1_6	csp_4_5	m_4_1	d_1_7	01105	failure
59	Female	NA	csp_7_5	m_4_4	d_1_6	01116	failure
25	Female	ce_2_2	csp_2_1	m_1_2	d_1_6	01118	failure
49	Female	ce_1_5	csp_5_6	m_4_1	d_1_3	01135	failure
86	Male	NA	csp_7_8	m_4_4	d_0_2	01136	failure
46	Female	ce_1_6	csp_5_2	m_4_1	d_1_3	01149	failure
30	Male	NA	csp_6_3	m_3_1	d_1_3	01158	failure
20	Female	ce_1_1	csp_5_4	m_4_1	d_1_4	01160	failure
70	Female	NA	csp_7_8	m_1_2	d_0_3	01160	failure
33	Male	ce_1_6	csp_3_8	m_4_1	d_1_8	01160	failure

From data to knowledge

Data $\neq$ information

Hierarchy:

1. data
2. information/insights
3. knowledge

From data to knowledge

Data $\neq$ information

Hierarchy:

1. data
2. information/insights
3. knowledge

SoundHound/Shazam

1. data: sound (digital recording)
2. information (low level): fingerprint of the recording
3. information (high level): metadata of the song associated to the fingerprint
4. knowledge: musical genre, band history, etc.

Example

age	gender	family	diploma	csp_42	code_insee	target
53	Female	m_4_1	d_1_7	csp_2_2	01004	failure
85	Female	m_1_2	d_1_2	csp_7_7	01004	failure
55	Male	m_4_1	d_1_3	csp_4_8	01010	success
45	Male	m_4_1	d_1_6	csp_4_3	01032	failure
54	Male	m_4_1	d_1_3	csp_6_7	01046	success
32	Male	m_4_4	d_1_3	csp_8_5	01053	success
41	Male	m_1_1	d_1_3	csp_6_2	01053	failure
18	Male	m_4_1	d_1_3	csp_8_4	01053	failure
45	Male	m_1_1	d_1_6	csp_4_7	01053	failure
65	Female	m_4_4	d_1_3	csp_7_5	01053	failure
49	Female	m_4_1	d_1_7	csp_4_5	01105	failure
59	Female	m_4_4	d_1_6	csp_7_5	01116	failure
25	Female	m_1_2	d_1_6	csp_2_1	01118	failure
49	Female	m_4_1	d_1_3	csp_5_6	01135	failure
86	Male	m_4_4	d_0_2	csp_7_8	01136	failure
46	Female	m_4_1	d_1_3	csp_5_2	01149	failure
30	Male	m_3_1	d_1_3	csp_6_3	01158	failure
20	Female	m_4_1	d_1_4	csp_5_4	01160	failure
70	Female	m_1_2	d_0_3	csp_7_8	01160	failure
33	Male	m_4_1	d_1_8	csp_3_8	01160	failure
15	Female	m_4_4	d_0_3	csp_8_5	01173	failure
67	Female	m_1_2	d_1_1	csp_7_7	01180	failure
66	Female	m_4_4	d_1_4	csp_7_7	01192	failure
44	Female	m_4_1	d_1_8	csp_3_8	01194	failure
56	Female	m_1_2	d_1_3	csp_5_2	01195	success
43	Female	m_1_2	d_1_4	csp_5_2	01236	failure
60	Male	m_3_1	d_1_6	csp_3_8	01244	failure
59	Female	m_4_1	d_1_3	csp_4_7	01281	success
53	Male	m_4_1	d_1_7	csp_4_6	01283	failure
32	Female	m_4_2	d_1_6	csp_8_4	01283	failure

Insights

- Who are the successful respondents?
- How do they differ from the general population?
- What variables can be used to predict the answer (if any)?
- etc.

Business Intelligence

Data science applied to business data

Business Intelligence

Data science applied to business data

Specificity

- business data:
  - large scale
  - frequently unstructured
  - byproduct of the business rather than collected on purpose
- goals:
  - profit (cost-benefit trade-off)
  - decision and optimization

Specificity

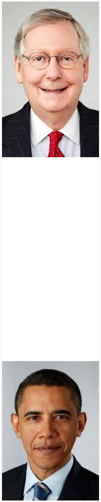
Data collection

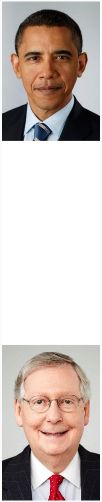
- limited control over the sampling process
  - by essence clients have a specific demographics
  - vocal clients are not representative
- feedback loop effect
  - positive and negative
  - e.g. rejected loan applications

BI use

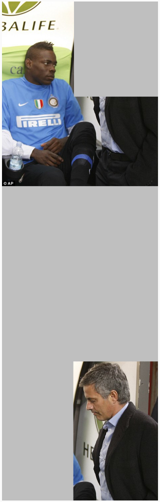
- patterns discovery can frighten customers (e.g. pregnancy detection by Target)
- communication effects (e.g. Twitter “racial bias”, Apple “sexist credit card”)

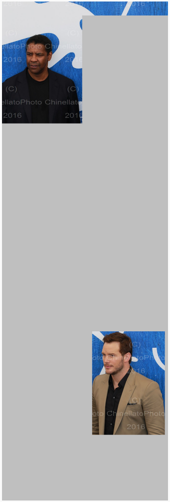
Twitter autocrop feature



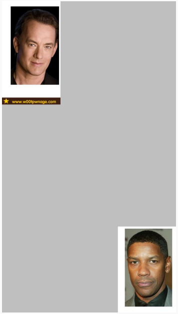


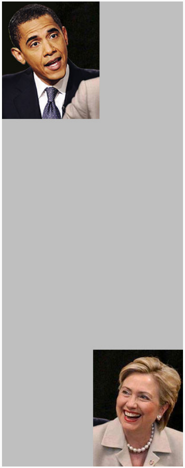
Twitter autocrop feature





Twitter autocrop feature





Data Science Pipeline

From data to knowledge

1. data collection/gathering
   - generation
   - pre-processing
   - transmission
2. data storage and querying
   - integration
   - indexing
3. data analysis
   - visualization
   - data mining
   - predictive models

Data Science Pipeline

From data to knowledge

1. data collection/gathering
   - generation
   - pre-processing
   - transmission
2. data storage and querying
   - integration
   - indexing
3. data analysis
   - visualization
   - data mining
   - predictive models

Business Intelligence

- Data Science *value chain*
- each step should increase the *value* of the data
  - compression: less storage
  - indexing and integration: faster and easier access
  - information extraction: direct business related results
  - etc.

Business Analytics

A possible definition

In the context of Business Intelligence, that is when data science is applied to business data, Business Analytics is the data analysis part of the data science pipeline

Business Analytics

A possible definition

In the context of Business Intelligence, that is when data science is applied to business data, Business Analytics is the data analysis part of the data science pipeline

Tasks

- visualization and reports: dashboard, scorecards, etc.
- data mining: clustering, frequent pattern analysis, etc.
- predictive models: applied to sales, churn, etc.

Business Analytics

A possible definition

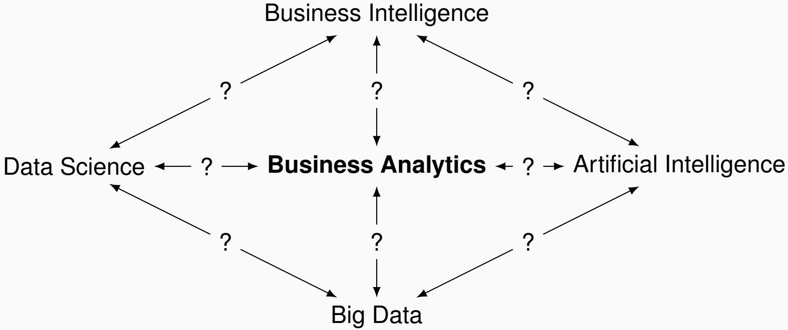
In the context of Business Intelligence, that is when data science is applied to business data, Business Analytics is the data analysis part of the data science pipeline

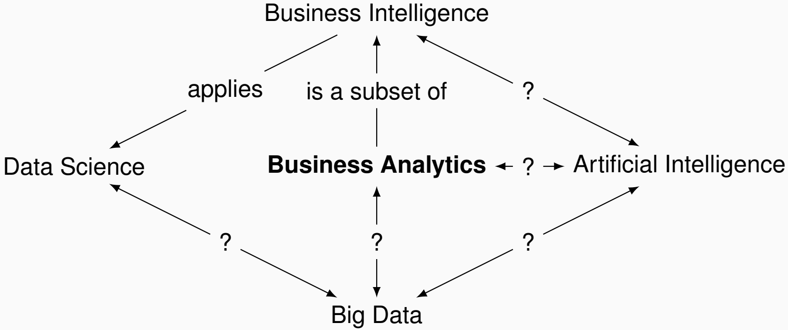
Tasks

- visualization and reports: dashboard, scorecards, etc.
- data mining: clustering, frequent pattern analysis, etc.
- predictive models: applied to sales, churn, etc.

Evolving vocabulary

- regular confusion between BI and BA
- blurry boundaries


Business Intelligence
?
?
?
Data Science
?
Business Analytics
?
Artificial Intelligence
?
?
?
Big Data


Business Intelligence
applies
is a subset of
?
Data Science
Business Analytics
?
Artificial Intelligence
?
?
?
Big Data

Big Data

A (very bad) definition of Big Data

Big Data = data centric

A (very bad) definition of Big Data

Big Data = data centric

Correct definition

Big Data = a data set that is too large to be processed on a single computer

Big Data

~~A (very bad) definition of Big Data~~

~~Big Data = data centric~~

Correct definition

Big Data = a data set that is too large to be processed on a single computer

Business data

- are frequently very large
- tend to grow

BI and BA use Big Data oriented methods on a regular basis

X Vs

Doug Laney's 3 Vs

- Doug Laney was an analyst at the META group (now Gartner)
- He proposed in 2001 the 3 V's:
  1. Volume: data size
  2. Velocity: streaming context
  3. Variety: text, image, video, etc.
- frequently used as "characteristics of big data" (but Laney did not use the terms!)
- complemented by other Vs such as Veracity (data quality, confidence in the results)

X Vs as typical corporate BS

Volume

is the only acceptable characterization

X Vs as typical corporate BS

Volume

is the only acceptable characterization

Velocity

- can be Volume when data is stored
- induces completely different challenges in a true streaming context (when data is thrown away!)
- is related to drifting and other advanced *standard* data science problems

X Vs as typical corporate BS

Volume
is the only acceptable characterization

Velocity

- can be Volume when data is stored
- induces completely different challenges in a true streaming context (when data is thrown away!)
- is related to drifting and other advanced *standard* data science problems

Variety and Veracity
have been part of data science since almost its beginning!

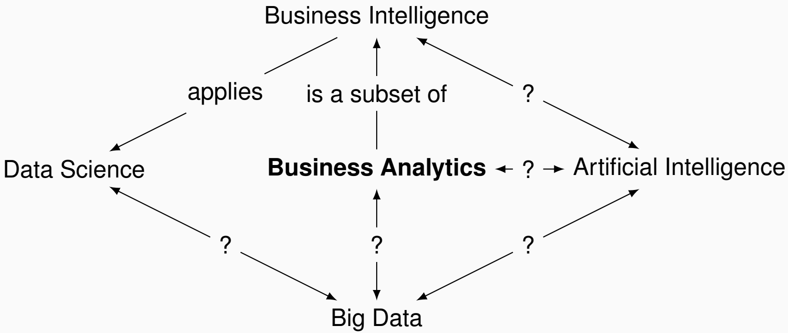
Processing Big Data

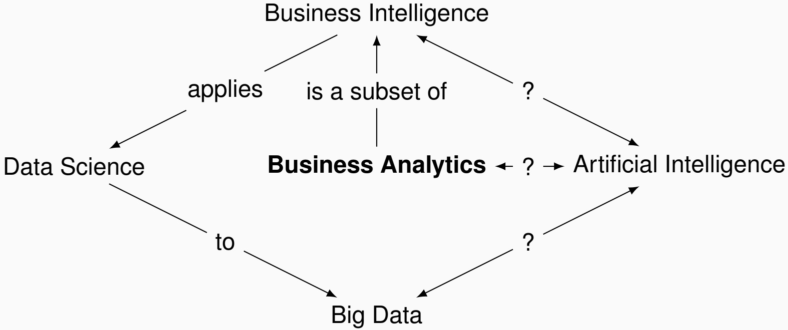
Distributed systems

- too large for a single computer: use several computers!
- mainly a set of computer science/engineering problems
- standard open source solutions:
  - Apache Hadoop
  - Apache Spark
- relevant for BI less for BA

Specific methods

- modified algorithms adapted to large scale data
- e.g. stochastic gradient descent
- relevant for BA (e.g. understand the limitations)


Business Intelligence
applies
is a subset of
?
Data Science
Business Analytics
?
Artificial Intelligence
?
?
?
Big Data


Business Intelligence
applies
is a subset of
?
Data Science
Business Analytics
?
Artificial Intelligence
to
?
Big Data

Artificial Intelligence

Definition

In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans (Wikipedia).

Definition

In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans (Wikipedia).

Artificial Intelligence

Definition

Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving". (Wikipedia).

Artificial Intelligence

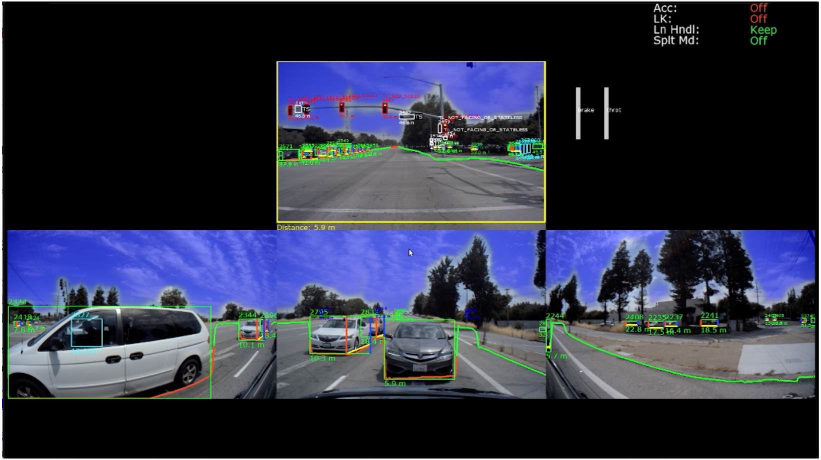
Definition

Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving". (Wikipedia).

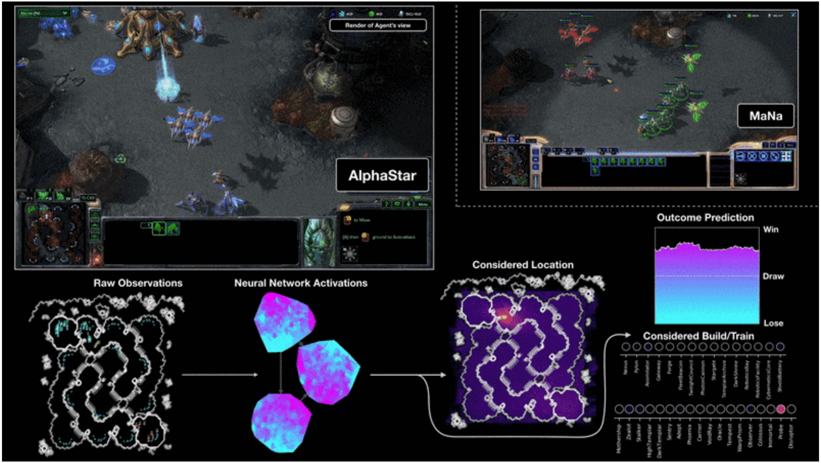
Examples

- playing games (chess, go, StarCraft)
- driving (autonomous vehicles)
- understanding human language (home assistants, IBM watson)
- understanding images (face recognition)
- creating art (style transfer)

Autonomous cars



StarCraft


Render of Agent's view
AlphaStar
MaNa
Outcome Prediction
Win
Draw
Lose
Raw Observations
Neural Network Activations
Considered Location
Considered Build/Train

AI and Data Science

Machine Learning

- construct a program that solves a task using examples of solving this task
- typical application: predictive models

Natural Language Processing

- understand human language
- typical application: knowledge extraction

AI and BI&A

AI based analytics

- predictive models (machine learning)
- recommender system (machine learning)
- text mining and text generation (NLP)

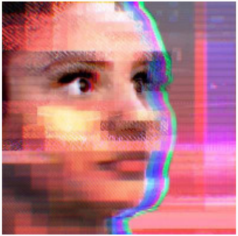
Analytics improved by AI

- smart graphics
- text generation
- fully automated machine learning

Image generation

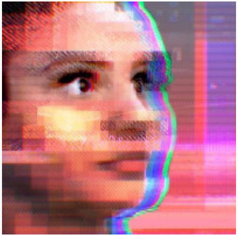
Those persons do not exist

Very difficult to control



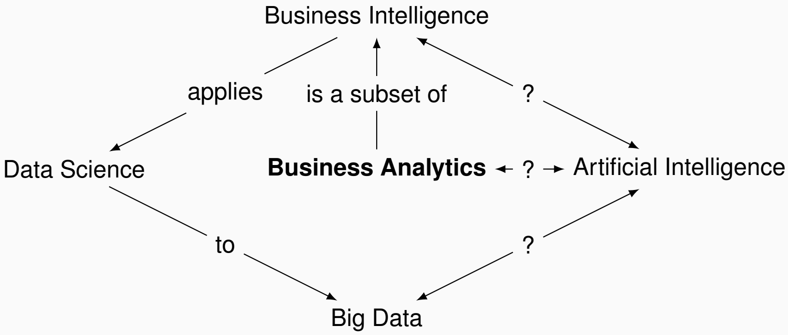
Tay: conversational agent test by Microsoft on 23/03/2016

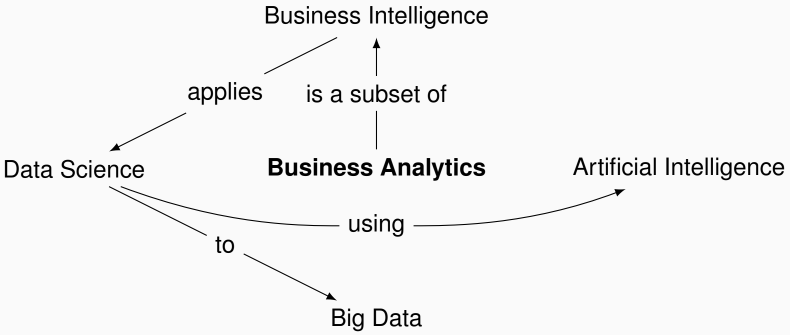
Very difficult to control



Tay: conversational agent test by Microsoft on 23/03/2016

A few hours of evolution...


Business Intelligence
applies
is a subset of
?
Data Science
Business Analytics
?
Artificial Intelligence
to
?
Big Data


Business Intelligence
applies
is a subset of
Data Science
Business Analytics
Artificial Intelligence
using
to
Big Data

Analysis vs Analytics

Management oriented distinction

- analysis: describe what happened
- analytics:
  - explain why something happened
  - predict what will happen

Mostly artificial

- BA makes heavy use of data analysis methods (e.g. clustering) that do no provide explanation
- causal inference remains a very difficult task
- should be considered as an encouragement to switch from descriptive analysis to hypotheses building (and testing)

Some BI&A use cases

Customer Segmentation

Goals

- identify segments of customers with similar characteristics:
  - demographic characteristics
  - behavioral characteristics
  - etc.
- segment specific marketing

Data and Tools

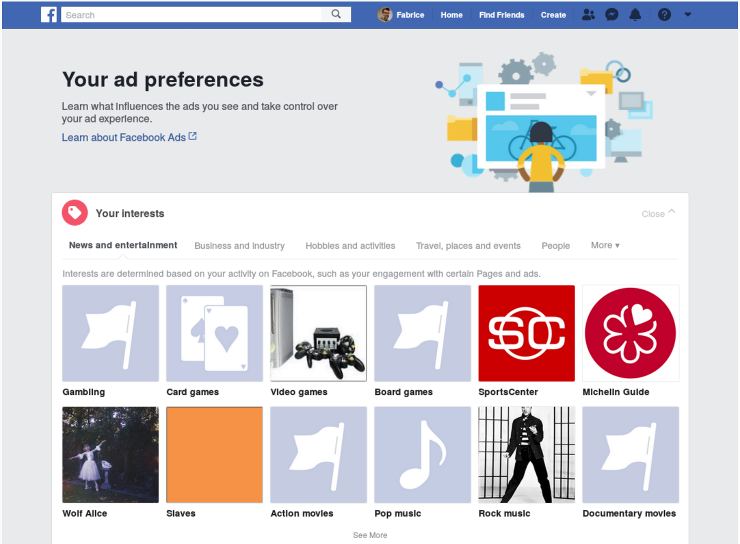
- data:
  - transaction records
  - service usage log
  - customer survey
  - external data
- tools:
  - clustering algorithms
  - mixture models

Examples

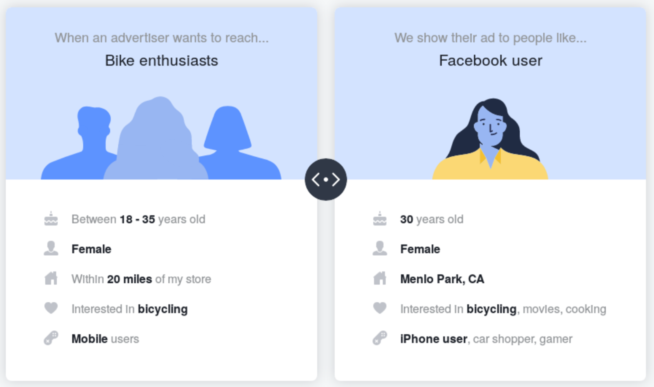
Targeted Advertisement

- market products to a specific audience
- core business of Facebook
  - user direct information sharing (gender, age, location, etc.)
  - user indirect information sharing (likes and connections)
  - analytics based
    - user interests
    - based on “meaningful interaction” (on Facebook)
- core business of Google
  - user information sharing: customer match (upload your customer list to google ads)
  - analytics
    - click stream based (including google search)
    - content based (displaying ads from others): NLP oriented
- others actors, such as Criteo

Facebook Ad Preferences



Facebook Ad Explanation



source: `https://www.facebook.com/ads/about/?entry_product=ad_preferences`

Churn Prevention/Prediction

Goals

- estimate the churn probability of a user
- trigger specific offers/actions to reduce the churn risk

Data and Tools

- data:
  - transaction records
  - service usage log
  - direct interaction (e.g. chat)
- tools:
  - classical statistical models (logistic regression)
  - advanced machine learning techniques if needed

Examples

All Subscription Based Business!

- standard tools for churn prevention
  - special offers
  - perks
  - news
- analytics
  - enable to trigger offers at crucial times
  - enable to tailor offers to specific users (e.g. new shows recommendation for Netflix)
- also used as an event detection technique
  - can trigger an offer for a payed service (Amazon Prime, Linkedin premium)
  - or for an upgraded service (Metal card for Revolut)

From Analysis to Analytics

Churn Prevention

1. base analysis
   - measure churn rate
   - cost analysis (customer lifetime value)
2. refined analysis
   - frequent churn pattern
   - churner characteristics
3. base analytics
   - predictive models
   - prevention
4. refined analytics
   - prevention prediction

From Analysis to Analytics

Churn Prevention

1. base analysis
   - measure churn rate
   - cost analysis (customer lifetime value)
2. refined analysis
   - frequent churn pattern
   - churner characteristics
3. base analytics
   - predictive models
   - prevention
4. refined analytics
   - prevention prediction

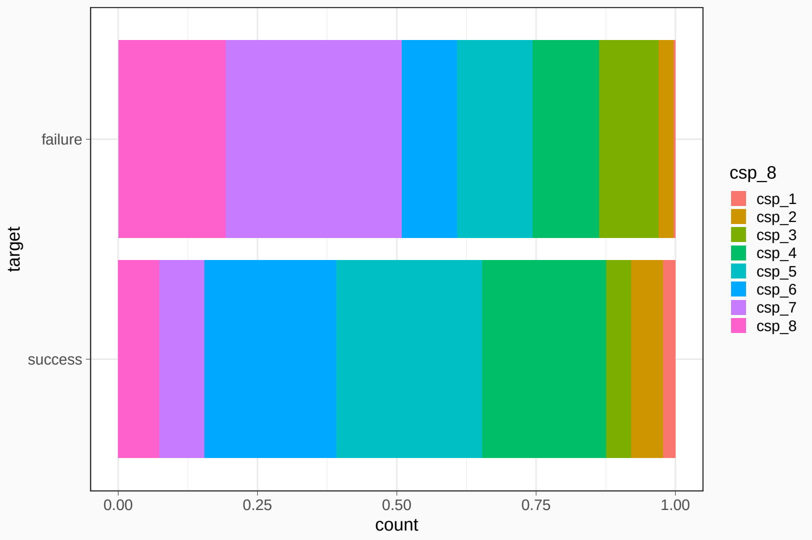
Explanation vs Prediction

	Stays	*Quits*
Active	850	20
Nonactive	50	80

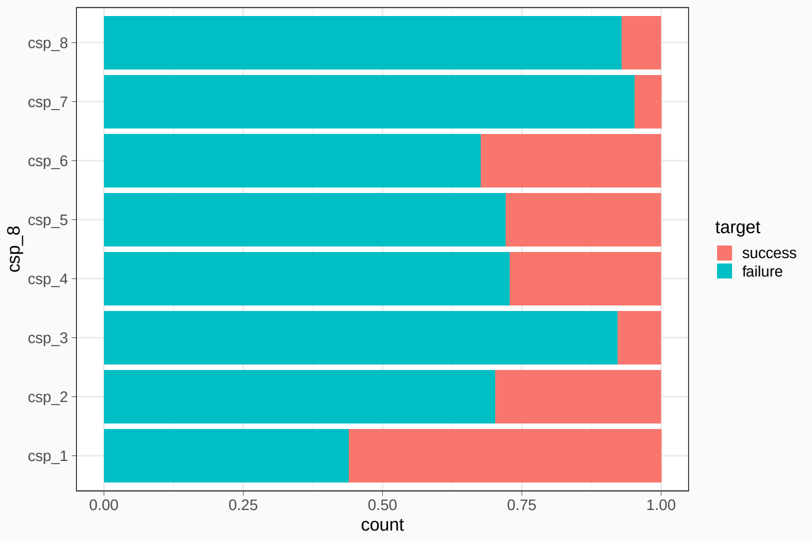
Example: marketing campaign

age	gender	employment	csp_42	family	diploma	code_insee	target
53	Female	ce_2_1	csp_2_2	m_4_1	d_1_7	01004	failure
85	Female	NA	csp_7_7	m_1_2	d_1_2	01004	failure
55	Male	ce_1_6	csp_4_8	m_4_1	d_1_3	01010	success
45	Male	ce_2_1	csp_4_3	m_4_1	d_1_6	01032	failure
54	Male	ce_1_6	csp_6_7	m_4_1	d_1_3	01046	success
32	Male	NA	csp_8_5	m_4_4	d_1_3	01053	success
41	Male	NA	csp_6_2	m_1_1	d_1_3	01053	failure
18	Male	NA	csp_8_4	m_4_1	d_1_3	01053	failure
45	Male	NA	csp_4_7	m_1_1	d_1_6	01053	failure
65	Female	NA	csp_7_5	m_4_4	d_1_3	01053	failure
49	Female	ce_1_6	csp_4_5	m_4_1	d_1_7	01105	failure
59	Female	NA	csp_7_5	m_4_4	d_1_6	01116	failure
25	Female	ce_2_2	csp_2_1	m_1_2	d_1_6	01118	failure
49	Female	ce_1_5	csp_5_6	m_4_1	d_1_3	01135	failure
86	Male	NA	csp_7_8	m_4_4	d_0_2	01136	failure
46	Female	ce_1_6	csp_5_2	m_4_1	d_1_3	01149	failure
30	Male	NA	csp_6_3	m_3_1	d_1_3	01158	failure
20	Female	ce_1_1	csp_5_4	m_4_1	d_1_4	01160	failure
70	Female	NA	csp_7_8	m_1_2	d_0_3	01160	failure
33	Male	ce_1_6	csp_3_8	m_4_1	d_1_8	01160	failure

Example: marketing campaign



Example: marketing campaign



Online Reputation

Goals

- monitor the online reputation of a brand
- act to improve it

Data and Tools

- data:
  - social networks
  - reviews
  - direct interaction and feedback
- tools:
  - natural language processing
  - network analysis method

Examples

Ubiquitous

- major part of brand strategies (Chief Reputation Officer!)
- integrated for instance in trading software but still mostly non automated
- monitoring (social media oriented, online)
  - name disambiguation and entity detection
  - topic detection
  - polarity (sentiment analysis)
  - central actor detection (network analysis)
  - word of mouth and information propagation (network analysis)
- profiling (static analysis)
  - media summary report
  - benefits from clustering for instance

Fraud Detection

Goals

- detect fraud attempts
- more generally detect potential risks (such as loan default)

Data and Tools

- data:
  - client personal data
  - transaction records
  - service usage log
- tools:
  - supervised machine learning (predictive models)
  - unsupervised machine learning (outlier detection)
  - network analysis

Examples

Fraud Detection as a Service

- e.g. Wordline: real time fraud detection for online payment
- machine learning with network analysis

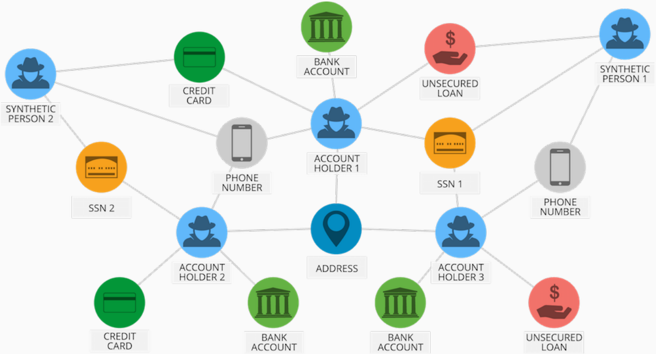
Fraud Detection for Compliance

- e.g. Revolut: money laundering prevention
- machine learning

Risk Assessment

- as part of Basel accords implementation, banks must estimate their risk exposure
- probabilities of default of various entities are estimated via statistical learning techniques

Fraud Detection with Graphs



source: https://neo4j.com/blog/financial-services-neo4j-fraud-detection/

Product Recommendation

Goals

- recommend products to customers
- might be part of churn prevention techniques
- can be used internally to assess substitution risks between products

Data and Tools

- data:
  - transaction records
  - consumer personal data
  - object data (e.g. reviews)
- tools:
  - supervised machine learning such as k-nearest neighbors
  - specific approximation techniques (non negative matrix factorization)

Examples

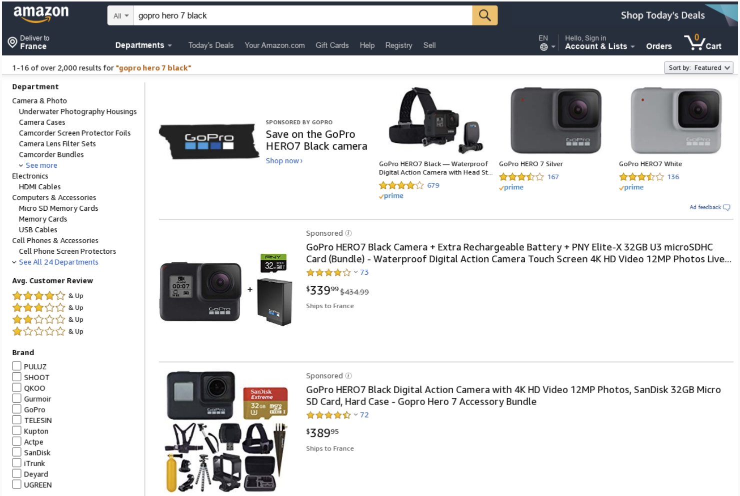
Recommender System

- general matching between two entities
- “mandatory” component of numerous business

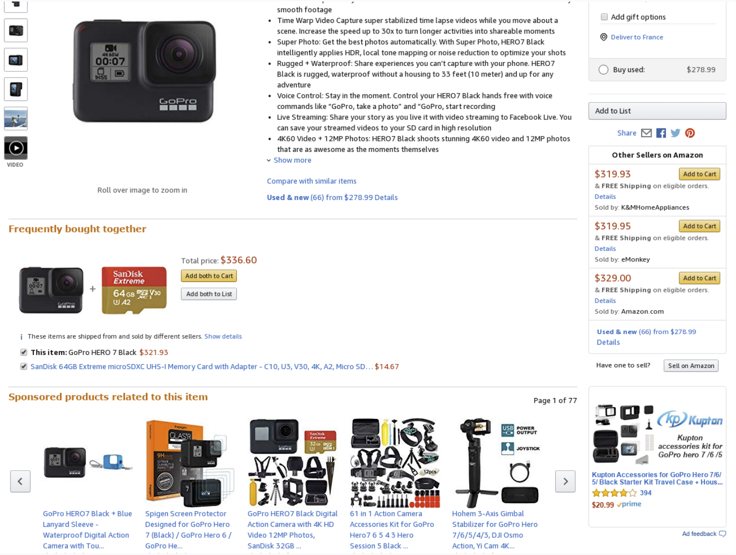
Ubiquitous

- online sellers (Amazon, Best Buy. etc.)
- subscription based streaming services (Netflix, Spotify, etc.)
- Ad based systems (e.g. youtube)
- Hotels (e.g. Accor)
- Online Dating (for matchmaking)

Amazon search

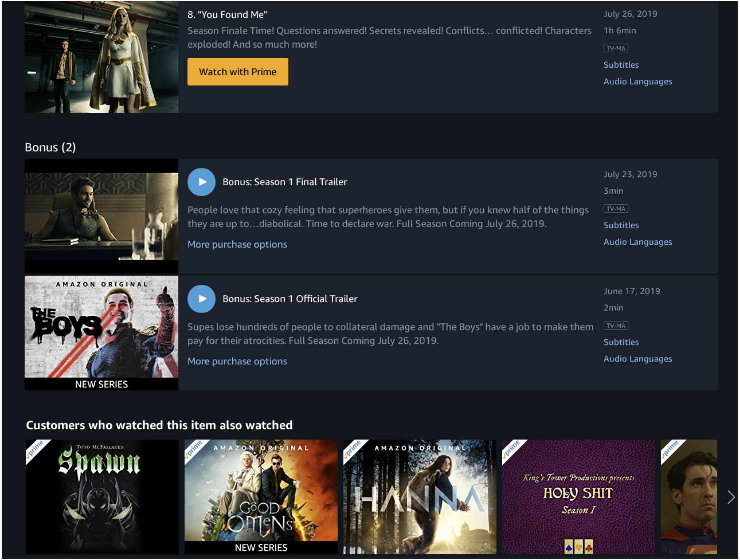


Amazon product



Amazon product

Amazon Prime



Analysis and Services

Analysis for internal use

Classical BI&A use cases

- customer segmentation
- churn prediction
- online reputation

Services

"Modern" use cases (data based business)

- external: product recommendation
- internal/regulatory: fraud detection and risk assessment

From BI&A to Data Science

Traditional use

- BI as Data Science on business data
- Trends
  - larger coverage of business data
  - more sophisticated methods (machine learning, NLP)
  - external services

Creative use

Data Science related techniques used to optimize internal processes without relying only on business data

- Revolut uses “machine learning" to compare a selfie and an official ID for account registration
- Chatbots are everywhere
  - front facing customers
  - internally for e.g. HR

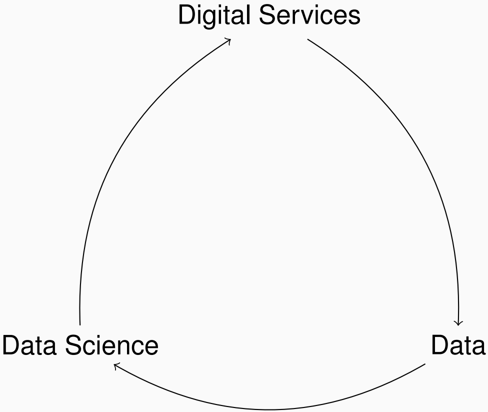
Data Science Revolution

Digital Services

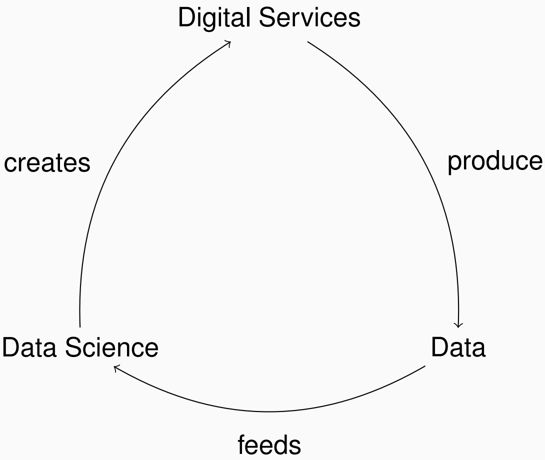
Data Science

Data

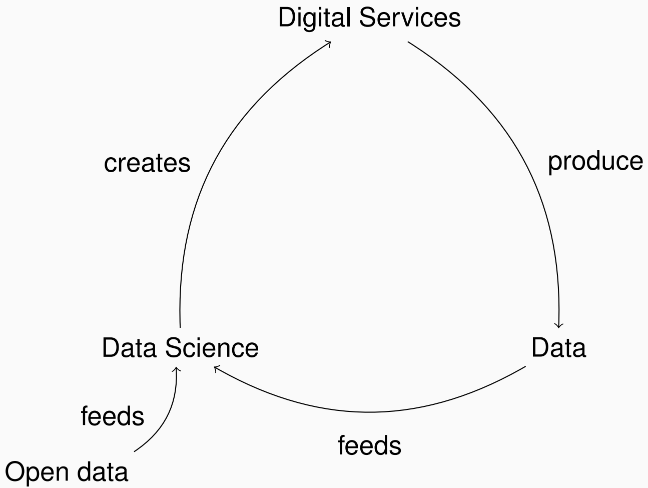
Data Science Revolution



Data Science Revolution



Data Science Revolution



Outline

BA data

Data Sources

Internal sources

- core databases
  - consumer data
  - transaction records
  - consumer relationship
- consumer activity log
  - web browsing
  - click stream
- byproducts
  - internal emails
  - memos, notes, etc.
  - internal reporting (accounting, hr, etc.)

Data Sources

External sources

- web monitoring
  - tweets
  - online discussions
  - reviews
- public data
  - governmental open data
  - easy to scrap web data (e.g. wikipedia)
- outsourced data
  - customer related database
  - specific surveys
  - product database

Data Types

Structured data

- tabular data
  - spreadsheet model
  - fixed variables that describe entities
- relational data
  - several data tables (or relations)
  - constraints that enable (among other things) to link tables
  - by far the most common type of exploitable data

Retailer data

- user table (uid, first name, last name, gender, email, phone number, ...)
- address table (aid, uid, ...)
- product table (pid, description, price, unit in stock, ...)
- order table (oid, uid, aid, pid, quantity)

Data Types

Semi-structured data

- data with some form of “variable” structure
- typical format: XML and JSON

```json
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
  ]
}
```

```xml
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <phoneNumbers>
    <phoneNumber>
      <type>home</type>
      <number>212 555-1234</number>
    </phoneNumber>
    <phoneNumber>
      <type>fax</type>
      <number>646 555-4567</number>
    </phoneNumber>
  </phoneNumbers>
</person>
```

Data Types

Unstructured data

- all the rest!
- mainly texts, sometimes images
- frequently associated to (semi)structured metadata

Limitations

Storage/Query level

- numerous possibly inconsistent data sources
- strong need for a Data Warehouse
  - data integration
  - data history
  - analytics views

Analytics level

- the vast majority of analytics methods work only on tabular data
- specific methods for some data types
  - text with NLP methods
  - network data with graph oriented methods

GDPR

General Data Protection Regulation

- EU law implemented in may 2018
- restricts personal data collection and processing
  - explicit collection and redistribution
  - collect only what is needed
  - data protection (anonymization)
  - explicit consent
  - right to withdraw consent, right of access, right of portability, right to be forgotten

Impacts

- limits personal data collection and processing
- but clarifies and simplifies certain aspects
- long term consequences are unclear
- ongoing very active research on privacy preserving data science

BA Methods

Multidimensional Analysis (MDA)

Motivation

- native data representations are seldom adapted for analysis
- aggregation and reorganization is needed
- MDA reorganizes "flat" data into multidimensional data mostly via aggregation

Principle

- standard data table: each object is described by some variables
- some nominal/categorical variables are chosen as "dimensions"
- a numerical variable is summarized conditionally to the chosen dimensions

Example

Flat table

	age	job	marital	education	balance	housing	loan
1	30	unemployed	married	primary	1787	no	no
2	33	services	married	secondary	4789	yes	yes
3	35	management	single	tertiary	1350	yes	no
4	30	management	married	tertiary	1476	yes	yes
5	59	blue-collar	married	secondary	0	yes	no
6	35	management	single	tertiary	747	no	no
7	36	self-employed	married	tertiary	307	yes	no
8	39	technician	married	secondary	147	yes	no
9	41	entrepreneur	married	tertiary	221	yes	no
10	43	services	married	primary	-88	yes	yes

MDA

- possible dimensions: job, marital, education, housing and loan (and age)
- aggregation targets: age and balance

Example

Mean balance vs marital status and education level

marital/education	primary	secondary	tertiary	unknown
divorced	1072.72	891.18	1437.90	1849.33
married	1371.64	1272.91	1860.72	1725.55
single	2065.75	1154.01	1754.71	1562.17

Mean balance vs job and education level

job/education	primary	secondary	tertiary	unknown
admin.	390.59	1269.68	1053.29	1590.47
blue-collar	1072.21	1068.59	2385.50	1032.88
entrepreneur	383.92	1276.17	2585.90	328.18
housemaid	1807.11	2011.89	2392.55	4282.40
management	2685.41	1250.10	1776.34	2386.26
retired	2744.60	2089.10	2476.74	1265.14
self-employed	1471.73	1164.55	1615.97	506.00
services	1107.32	998.88	1894.88	3058.00
student	1787.50	1610.43	1175.68	1754.88
technician	2593.00	1153.61	1631.63	1780.00
unemployed	873.19	1025.16	1224.78	3919.50
unknown	360.29	1229.00	2497.75	1648.60

Example: marketing campaign

age	gender	employment	csp_42	family	diploma	code_insee	target
53	Female	ce_2_1	csp_2_2	m_4_1	d_1_7	01004	failure
85	Female	NA	csp_7_7	m_1_2	d_1_2	01004	failure
55	Male	ce_1_6	csp_4_8	m_4_1	d_1_3	01010	success
45	Male	ce_2_1	csp_4_3	m_4_1	d_1_6	01032	failure
54	Male	ce_1_6	csp_6_7	m_4_1	d_1_3	01046	success
32	Male	NA	csp_8_5	m_4_4	d_1_3	01053	success
41	Male	NA	csp_6_2	m_1_1	d_1_3	01053	failure
18	Male	NA	csp_8_4	m_4_1	d_1_3	01053	failure
45	Male	NA	csp_4_7	m_1_1	d_1_6	01053	failure
65	Female	NA	csp_7_5	m_4_4	d_1_3	01053	failure
49	Female	ce_1_6	csp_4_5	m_4_1	d_1_7	01105	failure
59	Female	NA	csp_7_5	m_4_4	d_1_6	01116	failure
25	Female	ce_2_2	csp_2_1	m_1_2	d_1_6	01118	failure
49	Female	ce_1_5	csp_5_6	m_4_1	d_1_3	01135	failure
86	Male	NA	csp_7_8	m_4_4	d_0_2	01136	failure
46	Female	ce_1_6	csp_5_2	m_4_1	d_1_3	01149	failure
30	Male	NA	csp_6_3	m_3_1	d_1_3	01158	failure
20	Female	ce_1_1	csp_5_4	m_4_1	d_1_4	01160	failure
70	Female	NA	csp_7_8	m_1_2	d_0_3	01160	failure
33	Male	ce_1_6	csp_3_8	m_4_1	d_1_8	01160	failure

Examples

Age versus target and gender

gender/target	success	failure
Female	52	50
Male	52	46

Percentage of success versus gender and csp

CSP/gender	Female	Male
csp_1	0.82	0.47
csp_2	0.42	0.23
csp_3	0.14	0.04
csp_4	0.35	0.19
csp_5	0.33	0.14
csp_6	0.43	0.30
csp_7	0.06	0.03
csp_8	0.11	0.02

Pivot Table

- Spreadsheet oriented implementation of MDA
- introduced by Lotus Improv (1991) in a general strategy to separate data from their view
- standard feature of all modern spreadsheet programs (as well as data science and database oriented software)
- interactive filtering possibilities

OLAP

Online analytical processing

- de facto standard for efficient MDA
- a data set is composed of OLAP Cubes (hypercubes in fact)
  - dimensions and measures
  - a simple pivot table is a 2 dimensional “cube”

Example

- Bank example
- Dimensions: job, marital, education, housing, loan
- Measures: age and balance
- A cell (unemployed, married, primary education, no housing and no loan) contains the average age and the average balance for the persons with the specified values on the dimensions

OLAP

Hierarchies

- dimensions have frequently a hierarchical structure:
  - time: year, quarter, month
  - geographical: country, state, district
  - etc.

Standard operations

- Roll up: summarize the cube by climbing up in the hierarchy of a dimension (e.g. from district level sales to state level sales)
- Drill down: reverse of Roll up
- Slice/Dice: remove some dimensions by selection the value to keep for each of them

Summary

- analysis oriented view of the data
  - multiple views
  - interactive process
  - somewhat adapted to big data (aggregated views)
- implemented by a data warehouse
  - historical data (as opposed to live data)
  - integrated data
  - efficiency aspects
- MDA is the entry point of Business Intelligence

Visualization

Report or Discovery?

- visual representations are used routinely to convey business related results
- graphics can also be used to help inference and analysis
- interactive graphics provide better discovery capabilities

BI&A parlance

- Dashboards
  - collection of graphical representation of important data
  - dynamic (connected to the data warehouse)
  - interactive
- Scorecards
  - focused dashboards (non consensual definition)
  - KPI (key performance indicators)
  - monitoring oriented (KPI are associated to target values)

Examples

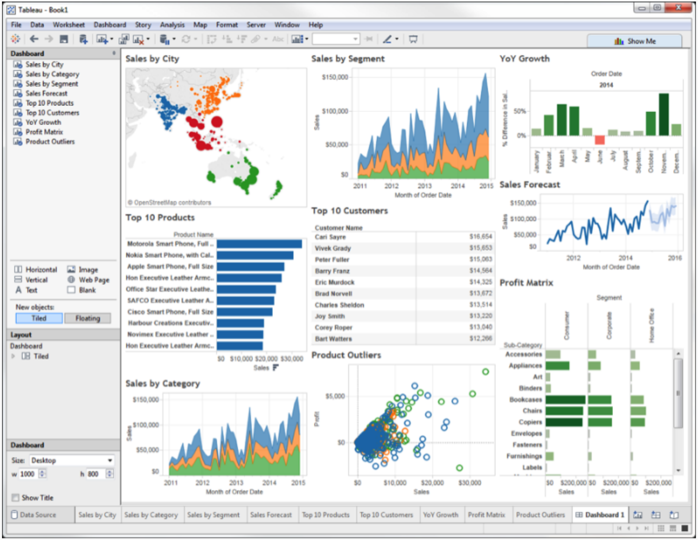
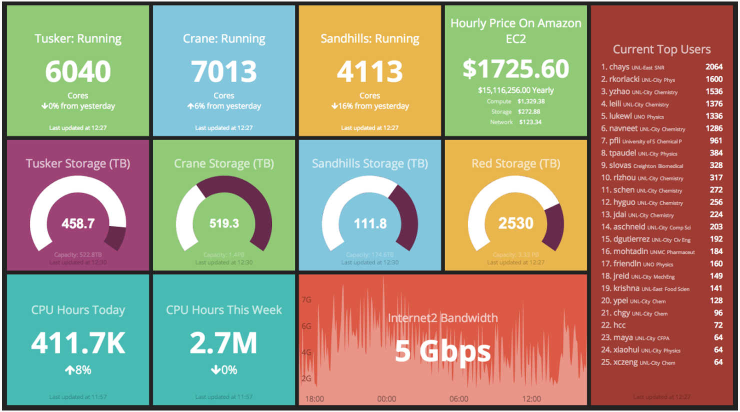
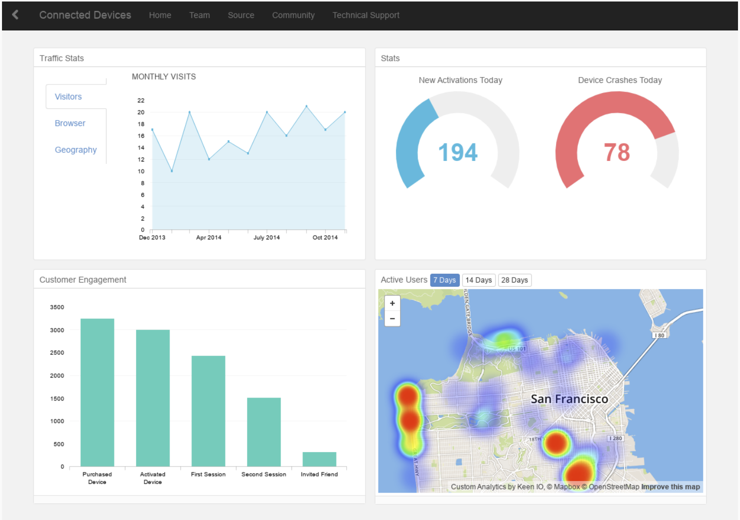


Tableau - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Show Me
Dashboard
Sales by City
Sales by Category
Sales by Segment
Sales Forecast
Top 10 Products
Top 10 Customers
YoY Growth
Profit Matrix
Product Outliers
Horizontal
Vertical
Text
Image
Web Page
Blank
New objects:
Tiled
Floating
Layout
Dashboard
Tiled
Size: Desktop
w 1000
h 800
Show Title
Sales by City
© OpenStreetMap contributors
Sales by Segment
Month of Order Date
YoY Growth
Order Date
2014
% Difference in Sal..
Sales Forecast
Top 10 Products
Product Name
Motorola Smart Phone, Full ..
Nokia Smart Phone, with Cal..
Apple Smart Phone, Full Size
Hon Executive Leather Armc..
Office Star Executive Leathe..
SAFCO Executive Leather A..
Cisco Smart Phone, Full Size
Harbour Creations Executiv..
Novimex Executive Leather ..
Hon Executive Leather Armc..
Top 10 Customers
Customer Name
Cari Sayre $16,654
Vivek Grady $15,653
Peter Fuller $15,063
Barry Franz $14,564
Eric Murdock $14,325
Brad Norvell $13,672
Charles Sheldon $13,514
Joy Smith $13,220
Corey Roper $13,040
Bart Watters $12,266
Profit Matrix
Segment
Consumer
Corporate
Home Office
Sub-Category
Accessories
Appliances
Art
Binders
Bookcases
Chairs
Copiers
Envelopes
Fasteners
Furnishings
Labels
Sales by Category
Product Outliers
Profit
Sales
Data Source
Dashboard 1

Examples


Tusker: Running
6040
Cores
0% from yesterday
Last updated at 12:27
Crane: Running
7013
Cores
6% from yesterday
Last updated at 12:27
Sandhills: Running
4113
Cores
16% from yesterday
Last updated at 12:27
Hourly Price On Amazon EC2
$1725.60
$15,116,256.00 Yearly
Compute $1,329.38
Storage $272.88
Network $123.34
Current Top Users
1. chays UNL-East SNR 2064
2. rkorlacki UNL-City Phys 1600
3. yzhao UNL-City Chemistry 1536
4. leili UNL-City Chemistry 1376
5. lukewl UNO Physics 1336
6. navneet UNL-City Chemistry 1286
7. pfli University of S Chemical P 961
8. tpaudel UNL-City Physics 384
9. slovas Creighton Biomedical 328
10. rlzhou UNL-City Chemistry 317
11. schen UNL-City Chemistry 272
12. hyguo UNL-City Chemistry 256
13. jdai UNL-City Chemistry 224
14. aschneid UNL-City Comp Sci 203
15. dgutierrez UNL-City Civ Eng 192
16. mohtadin UNMC Pharmaceut 184
17. friendln UNO Physics 160
18. jreid UNL-City MechEng 149
19. krishna UNL-East Food Scien 141
20. ypei UNL-City Chem 128
21. chgy UNL-City Chem 96
22. hcc 72
23. maya UNL-City CFPA 64
24. xiaohui UNL-City Physics 64
25. xczeng UNL-City Chem 64
Tusker Storage (TB)
458.7
Capacity: 522.8TB
Last updated at 12:30
Crane Storage (TB)
519.3
Capacity: 1.4PB
Last updated at 12:30
Sandhills Storage (TB)
111.8
Capacity: 174.6TB
Last updated at 12:30
Red Storage (TB)
2530
Capacity: 3.33 PB
Last updated at 12:27
CPU Hours Today
411.7K
8%
Last updated at 11:57
CPU Hours This Week
2.7M
0%
Last updated at 11:57
Internet2 Bandwidth
5 Gbps
7G
6G
4G
2G
18:00
00:00
06:00
12:00

Examples



Connected Devices
Home
Team
Source
Community
Technical Support
Traffic Stats
MONTHLY VISITS
Visitors
Browser
Geography
Dec 2013
Apr 2014
July 2014
Oct 2014
Stats
New Activations Today
194
Device Crashes Today
78
Customer Engagement
Purchased Device
Activated Device
First Session
Second Session
Invited Friend
Active Users
7 Days
14 Days
28 Days
San Francisco
Custom Analytics by Keen IO, © Mapbox © OpenStreetMap Improve this map

Visualization

Difficult

- highly dependent to what is shown (MDA, KPI, etc.)
- GIGO: garbage in garbage out
- information visualization is difficult (meaningful vs beautiful)
- active ongoing research

Report or Discovery?

- mostly report oriented
- discovery is strongly related to the level of interactivity (filtering, linked views, etc)
- “programming” is generally needed

Clustering

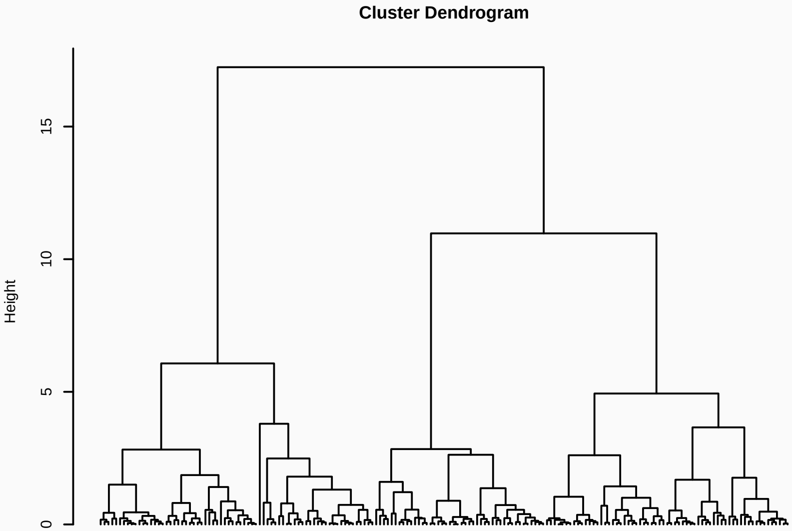
Definition

Clustering: grouping objects in such a way that objects in a given group are more similar to each other than to objects in other groups.

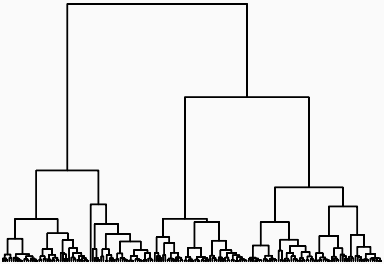
Numerous algorithms

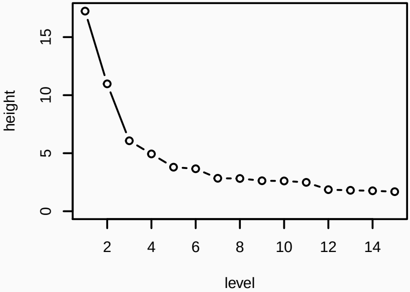
- hierarchical clustering:
  - start with as many clusters as there are objects
  - merge closest clusters
- k-means and other prototype based clustering methods:
  - start with random prototypes
  - assign objects to closest prototypes
  - update the prototypes
- and others...

Example of clustering result



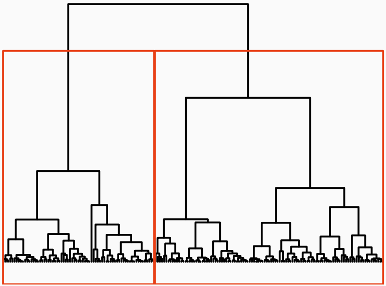
Hierarchical clustering

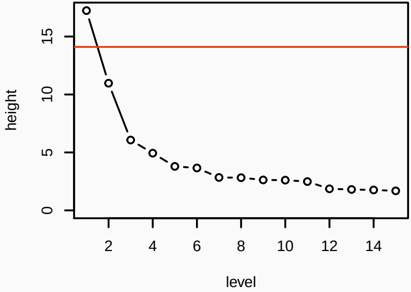




- look for "gaps" between levels: potential candidates for interesting partitions
- local partitions (i.e. branches) might also be interesting

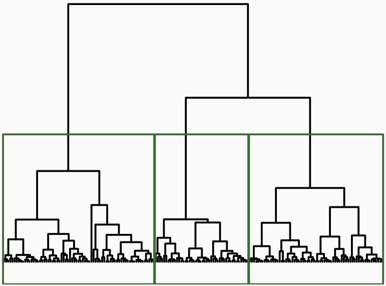
Hierarchical clustering

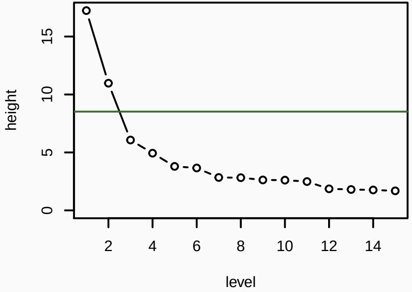




- look for "gaps" between levels: potential candidates for interesting partitions
- local partitions (i.e. branches) might also be interesting

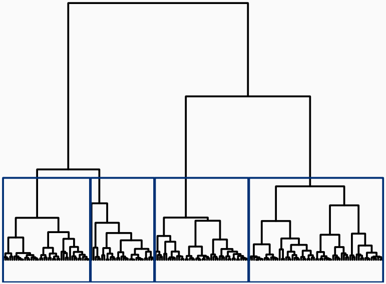
Hierarchical clustering

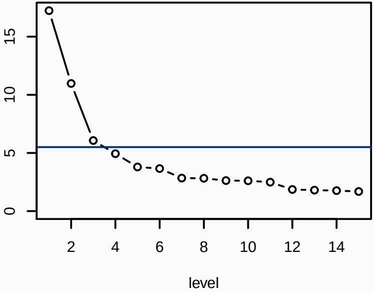




- look for "gaps" between levels: potential candidates for interesting partitions
- local partitions (i.e. branches) might also be interesting

Hierarchical clustering





- look for “gaps” between levels: potential candidates for interesting partitions
- local partitions (i.e. branches) might also be interesting

Clustering Results

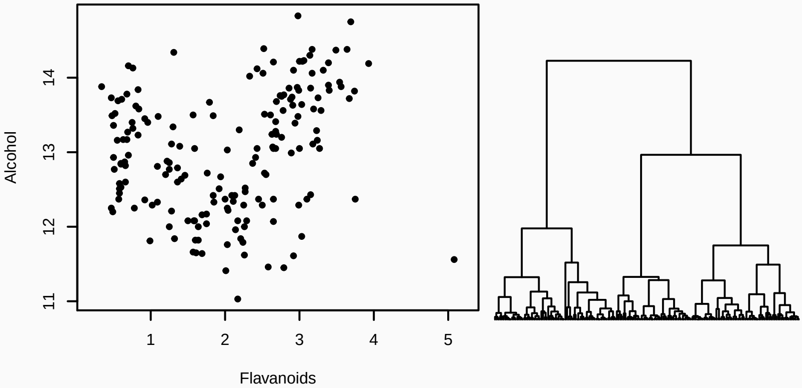
Expert cluster analysis

- how to interpret the clusters?
  - making sense of a list of objects
  - easier with prototype based methods: a central “typical” object per cluster (its prototype)
- explanation vs prediction

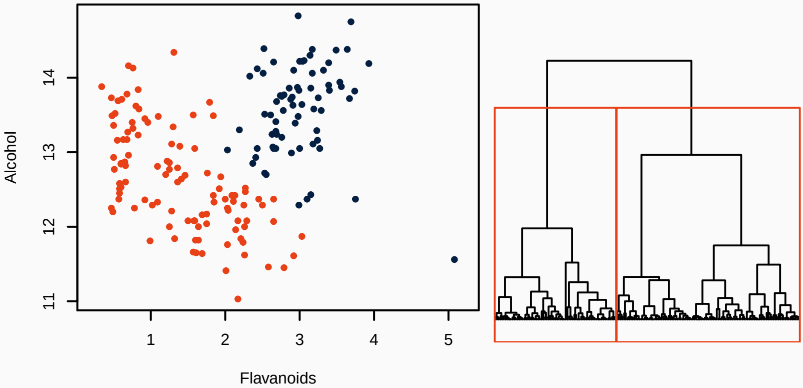
Clustering as an art

- results are highly depend on parameters (cluster number, dissimilarities, etc.)
- “artificial” clusters

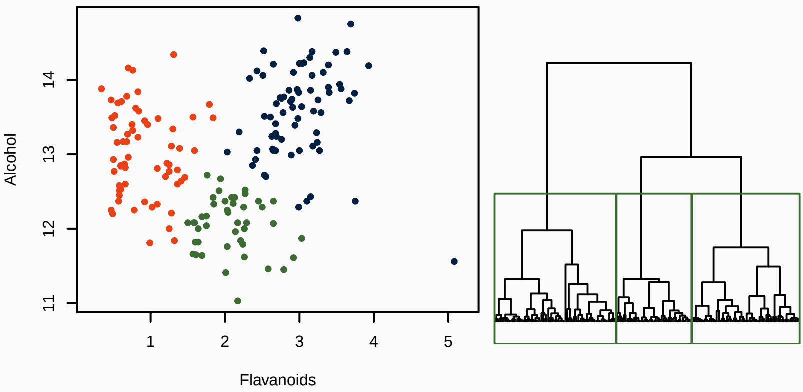
Example



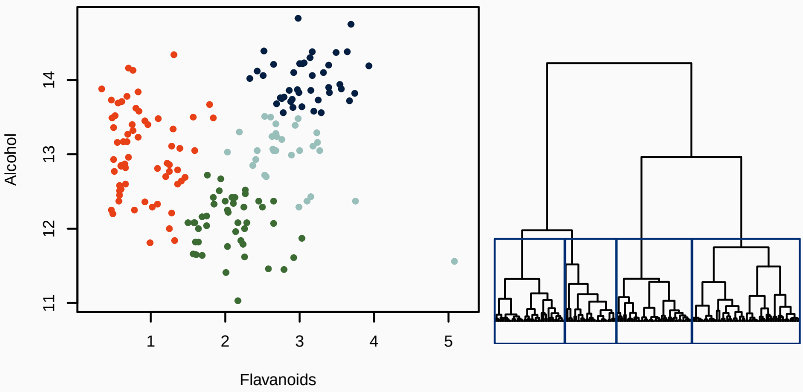
Example



Example



Example



Clustering

In practice

- used massively as a summary tool
  - with a high number of clusters
  - prototype based method: analyze the prototypes
- used to extract knowledge but
  - time consuming process (no consensual automatic way of selecting “optimal” parameters)
  - cluster understanding is difficult
  - alignment with business interests is not guaranteed

Frequent Pattern Mining

Principle

- to detect frequent associations between events in transactions
  - objects in a shopping basket
  - actions in an app
  - web pages in a navigation
- applications
  - recommend objects (amazon)
  - improve apps and web sites (recommend sections, reorganize navigation)
  - fraud detection
  - etc.
- natural extension to sequences (sequential pattern mining)

Frequent Pattern Mining

Monotony principle

- we look for frequent itemsets (events that occur frequently together)
- if a set S of items is frequent all its subsets are frequent
  - if $S = \{A, B, C\}$ is frequent, than a “sufficient” number of transactions contain A, B and C, and possibly of other items
  - and the number of transactions that contain A is at least equal to to the number of transactions containing all items in S
  - and thus $\{A\}$ is frequent!
- efficient algorithms are based on this principle
- original algorithm: APriori

Frequent/Sequential Pattern Mining

In practice

- very efficient for some practical applications, e.g.
  - recommendation
  - process mining
  - monitoring
- but with some limitations
  - computational efficiency
  - spurious pattern discovery in large data sets
  - very large outputs (too many patterns)

Predictive Methods

Goal

- statistics parlance: use some (explanatory) variables to “guess” the values of others (target) variables
- reveal hidden/unknown information
- assumes some form of dependence

Examples

- churn prediction
  - explanatory: consumer profile (including logs)
  - target: churn next month?
- used car market value
  - explanatory: car profile (age, custom parts, mileage)
  - target: market value of the car

Supervised/machine Learning

Principle

- use past values to build a predictive model
- circular situation
  - we need to know the unknown information to build a model!
  - major difficulty: “labelled” data
- links with artificial intelligence
  - learning aspect (learn to infer missing information from examples)
  - human based labelling in complex examples (e.g. image recognition)

This is not econometrics

- machine learning: best prediction
- econometrics: best explanation

Methods

Numerous methods are available

- linear/logistic regression
- decision trees
- ensemble methods (random forest, boosting)
- support vector machines and kernel methods
- artificial neural networks (and deep learning)

State of the art

- impressive results in some cases (above human performances for image classification for instance)
- poor results in others
- well establish methodologies (computationally intensive)
- major difficulty: access to labelled data!

Business Analytics

In summary

- Business Analytics analyses business data with data science tools
- Business data is stored in data warehouses
- Multidimensional analysis (using OLAP) provides aggregated expert views of the raw data
- visualization, data mining and machine learning is applied to MDA tables or to raw tables

Applications

- consumer relationship management
  - market segmentation
  - churn detection
  - recommendation
  - social media monitoring
- and many more!

Sources

- Target logo: https://commons.wikimedia.org/wiki/File:Target_logo.svg
- Captain Obvious image: https://imgur.com/gallery/PazzF
- Facebook ads: https://www.facebook.com/ads/about/?entry_product=ad_preferences
- Fraud ring as a graph: https://neo4j.com/blog/financial-services-neo4j-fraud-detection/

Licence

Version

Last git commit: 2021-01-18
By: Fabrice Rossi (Fabrice.Rossi@apiacoa.org)
Git hash: 98a933c9c4319cadc4fe5eaceeafabd892986b07