Data Science in 8 steps: introduction

Ann Van Eyken

#welcome

Ann Van Eyken

NCOI LEARNING

THE MASTER LABS


Mentimeter
Go to www.menti.com

Let's get started.


$E=mc^2$
Mentimeter
Go to www.menti.com

Understand the question


Not all problems can be solved
with data science.


Make decisions
Compare options
Find relationships
Make predictions
Summarize & understand
Remember things
So what can we do?

When was the last order placed?

Who placed the last order?

Who, what, where or when did something happen?

How did customer A find my website?

What browser is a particular user using to browse the site?

Make decisions

Compare options

Find relationships

Make predictions

Summarize & understand

Remember things

How many outstanding orders do we have?

How many orders where placed last month?

Can you summarize what happened?

Who are my 10 largest customers?

What browsers do my users tend to use?

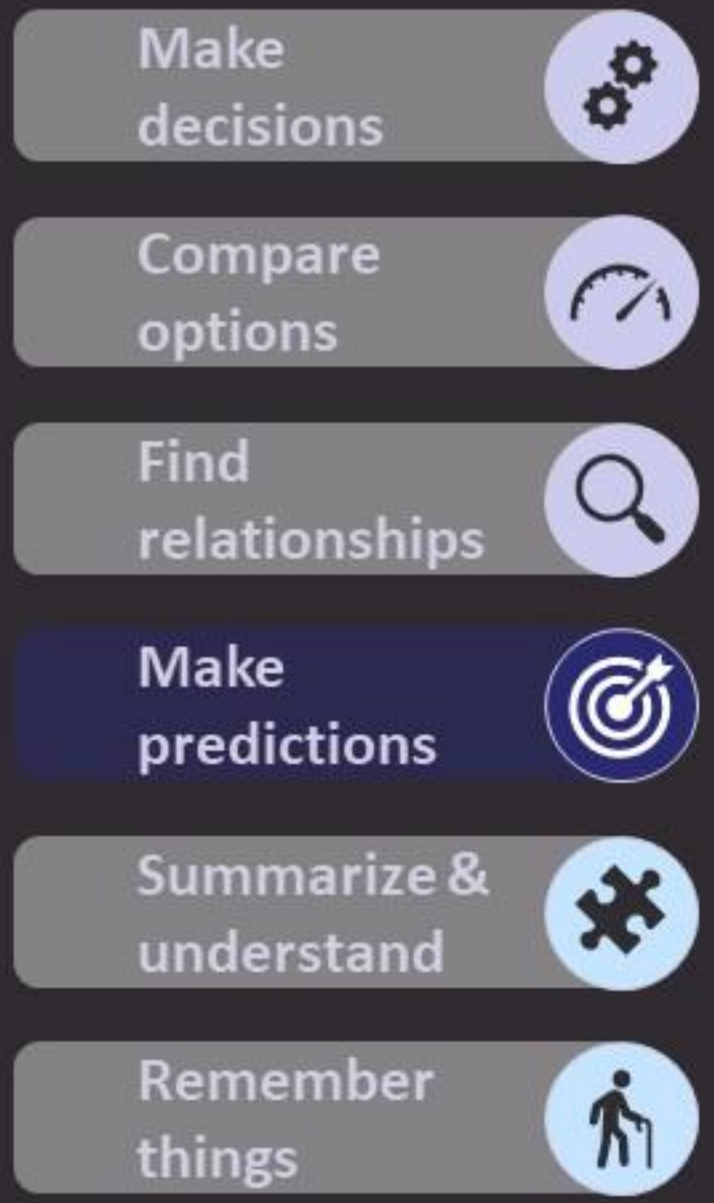
What is the probability that this air conditioner will fail in the next year?

Will this air conditioner fail in the next three years, yes or no?

What happens when … ?

Is there a relationship between time spent under the sun and height of the plant?

Is this bank transaction fraudulent?



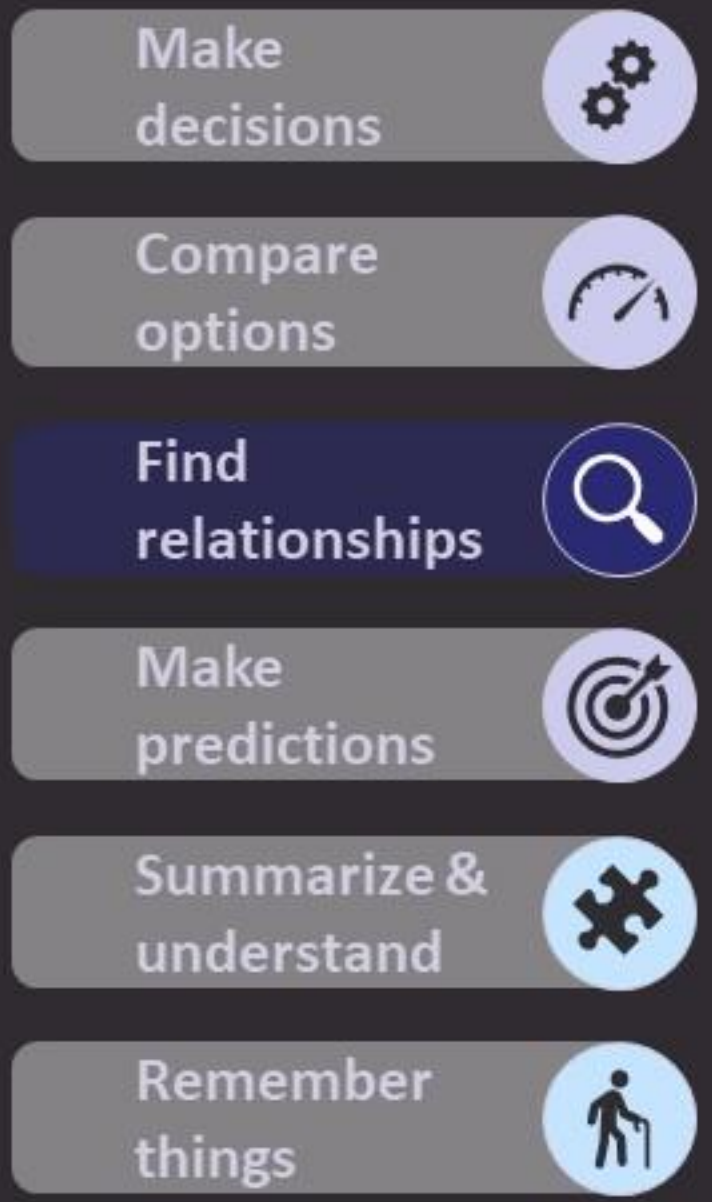
Which viewers like the same kind of movies?

What are the key differences between apples and oranges?

What are the key parts and relationships of ...?

What factors best predict demand for electricity?

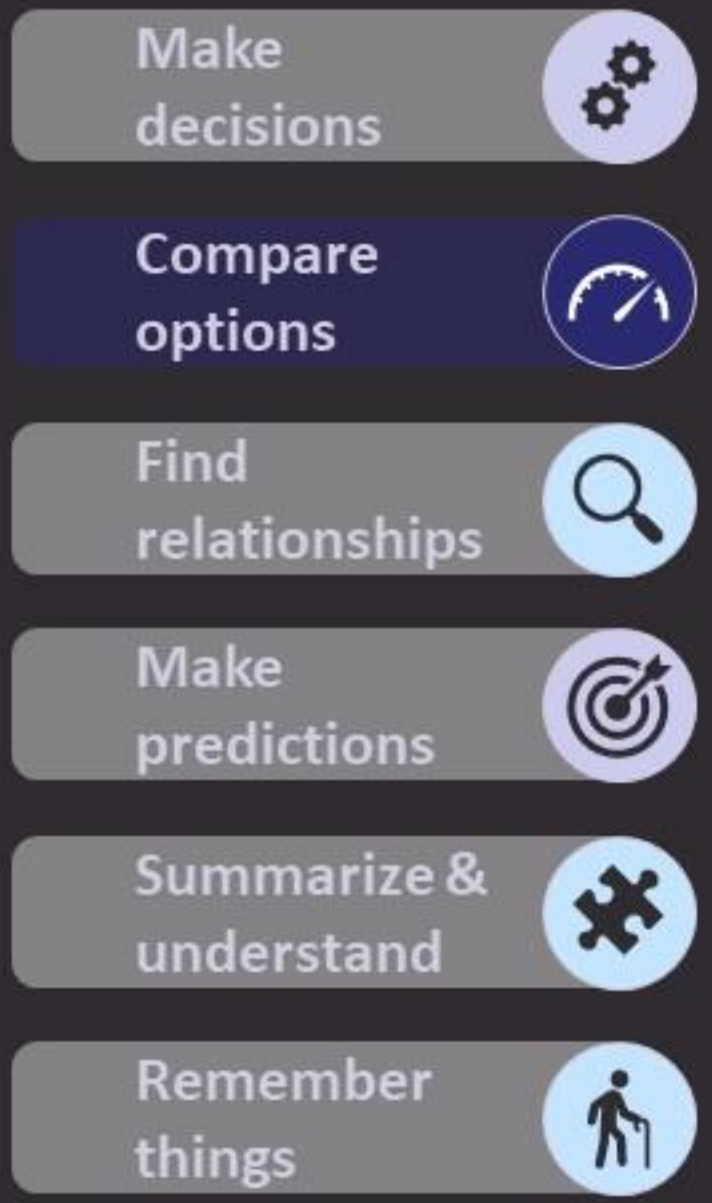
What combination of sensors best displays the overall health of the system?



What should we change to our website to get a higher conversion rate?

Is this the best approach?

Can we save money by pricing different products better?



Where on my website should I place this ad so that viewers are most likely to click it?

Where should we set up our new location?

Can you predict what will happen to … under new conditions?

What route should my delivery truck take?

Should my automated cooling and heating system adjust the temperature higher, lower or leave it where it is?

Make decisions

Compare options

Find relationships

Make predictions

Summarize & understand

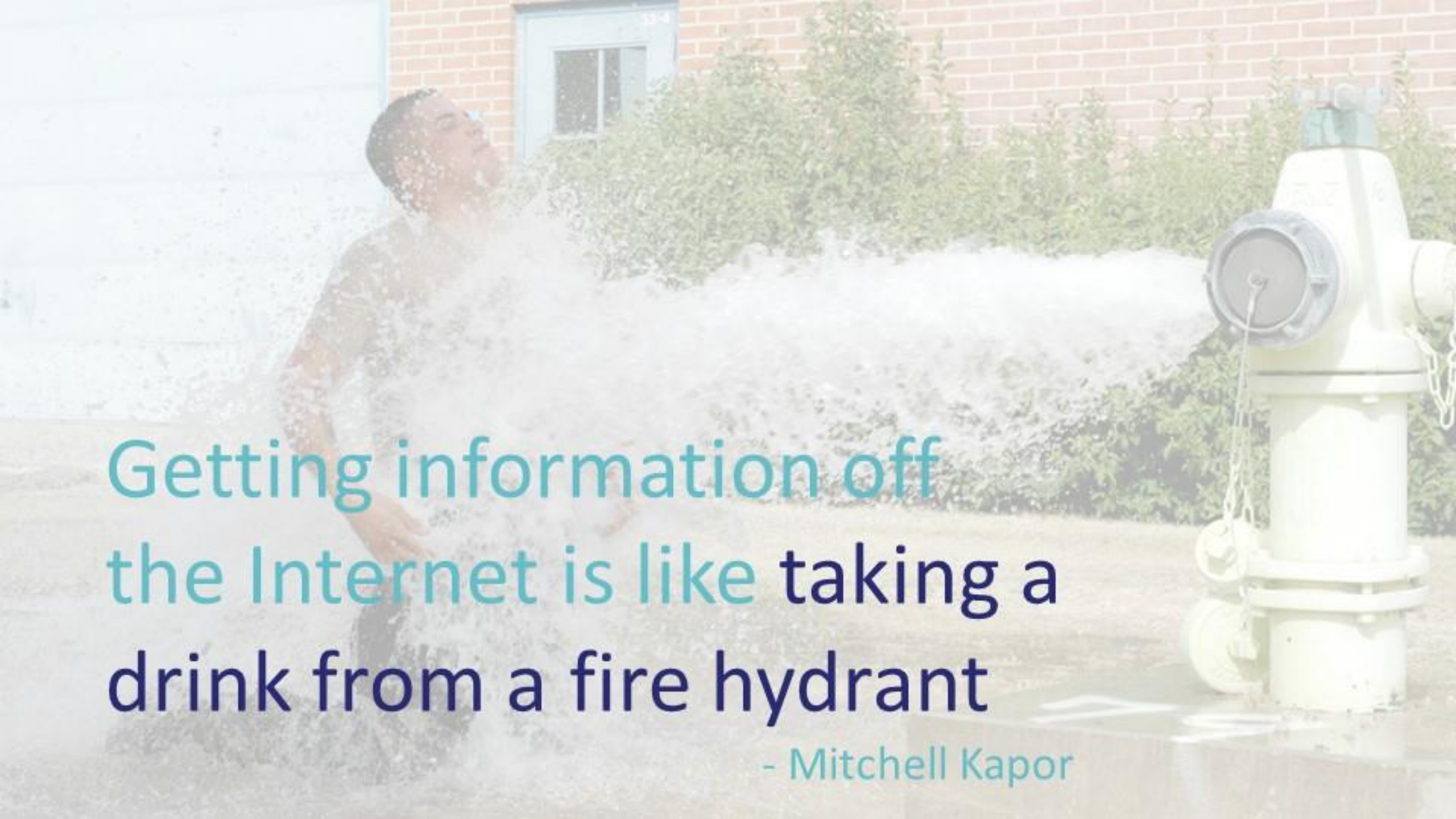
Remember things

Understand the question


Mentimeter
Go to www.menti.com


Find the data
Understand the question


33-4



Getting information off the Internet is like taking a drink from a fire hydrant

- Mitchell Kapor

www

Google

Find the data


Mentimeter
Go to www.menti.com

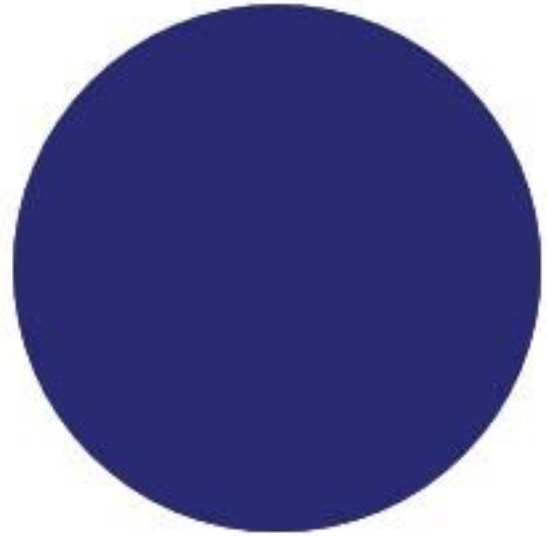

Store the data
Find the data
Understand the question

OLAP vs OLTP

OLAP

OLTP

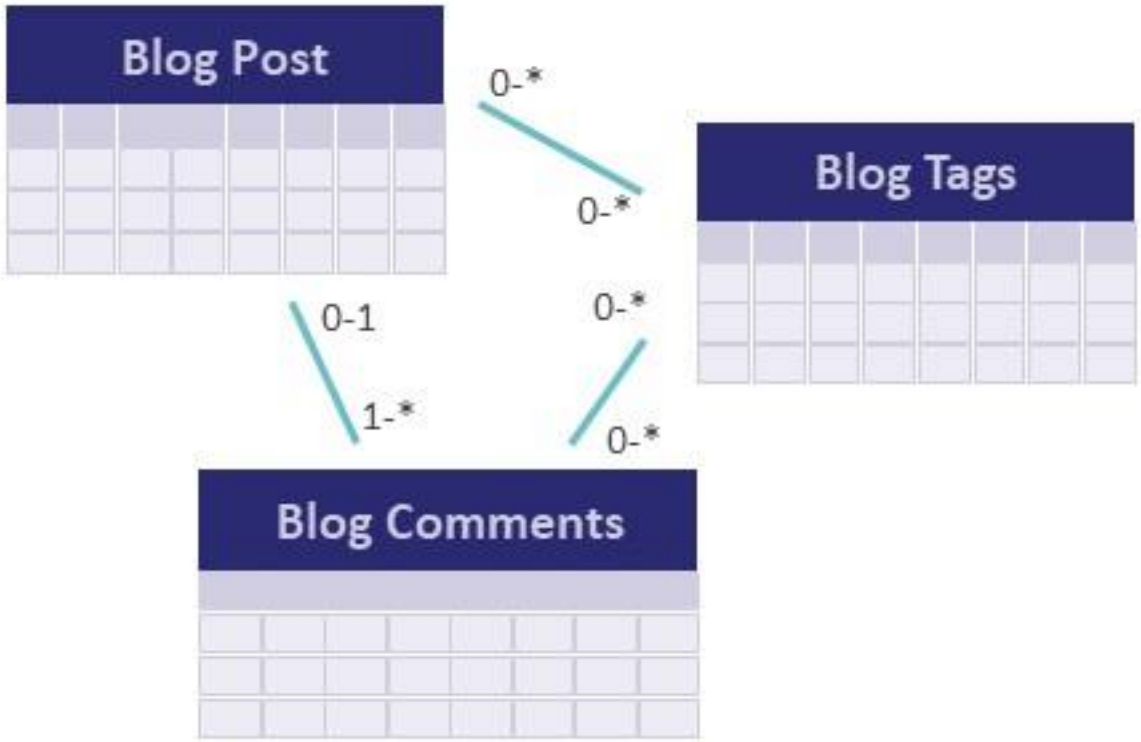
RDBMS vs. NoSQL




RDBMS
Relational database management systems (SQL)
NoSQL
non-relational or distributed databases

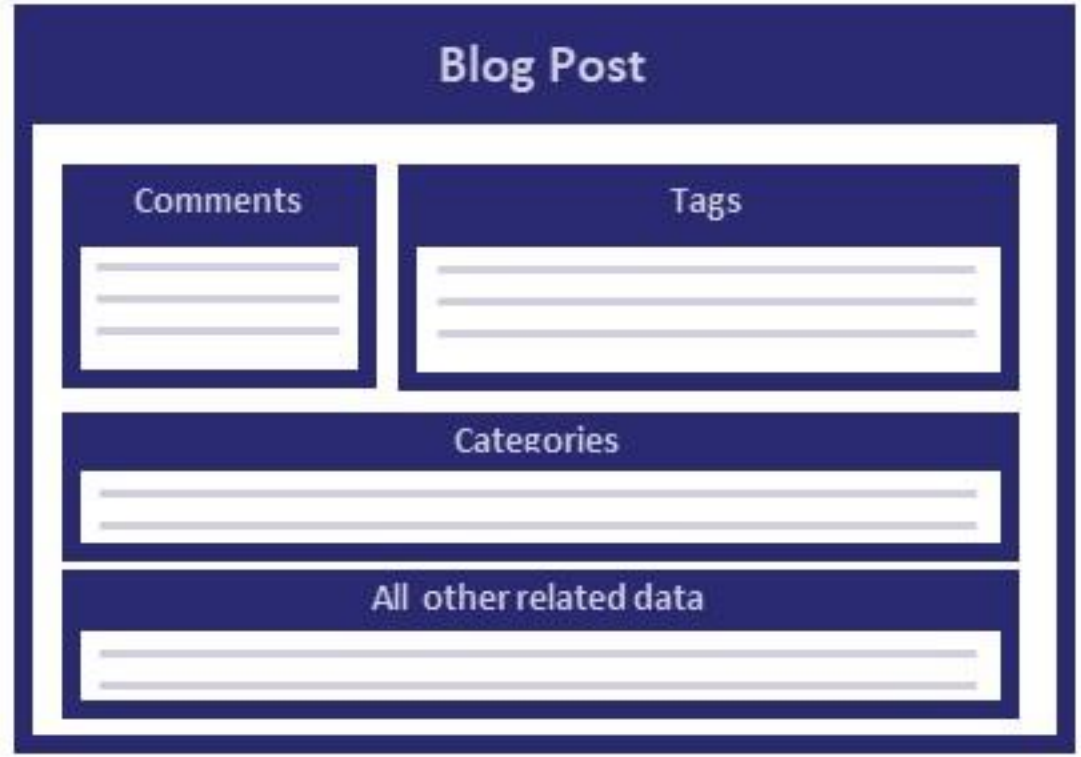
RDBMS

Relational database management systems (SQL)



NoSQL

non-relational or distributed databases



RDBMS

Relational database management systems (SQL)

Predefined schema

For structured data

Vertical scalability

SQL Databases are scaled by increasing the horse-power of the hardware. (Better CPU, RAM, SSD, etc on a single server)

Table-based

Excellent support

for all SQL databases from their vendors

ACID Properties

Enforced data integrity

Normalisation

NoSQL

non-relational or distributed databases

Horizontal scalability

NoSQL Databases are scaled by increasing the number of database servers in the pool of resources to reduce the load (just add a few servers)

Large datasets

Community support

Limited outside experts are available to support large scale NoSQL deployments

Dynamic schema

For unstructured data

Graph store

Key-Value store

Document store

Column Family store

Denormalisation

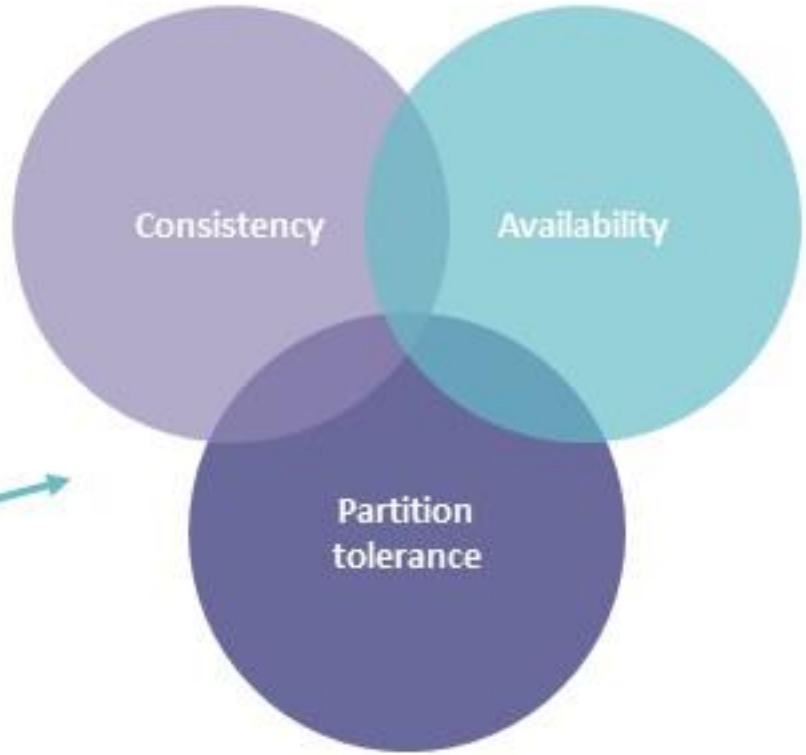
CAP Theorem

Eventual data integrity

CAP Theorem

It is impossible for a distributed data store to simultaneously provide more than two out of the following three guarantees: Consistency, Availability, Partition Tolerance

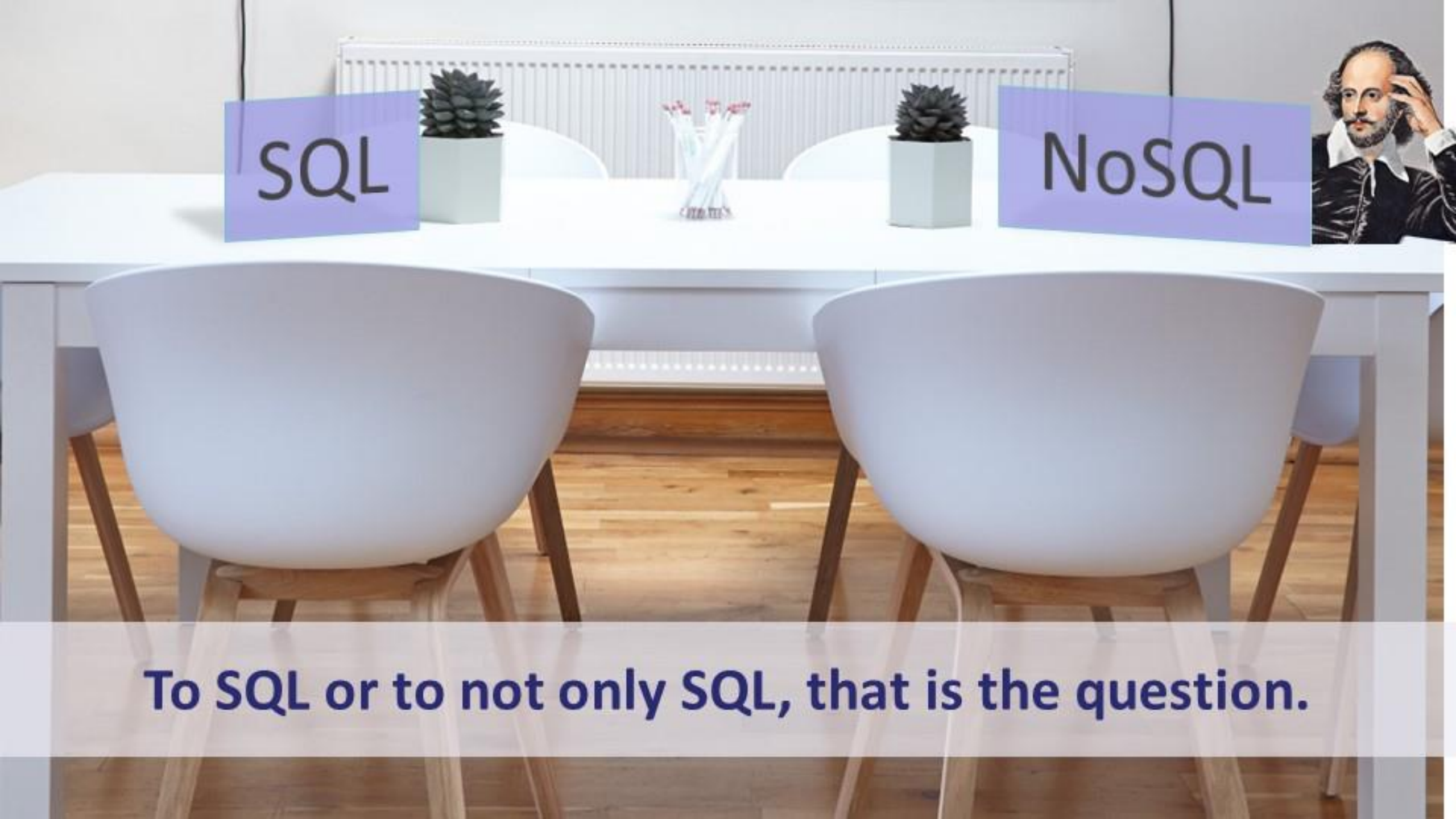
Every read receives the most recent write or an error



Every request receives a (non-error) response, without guarantee that it contains the most recent write

Pick two

The system continues to operate despite an arbitrary number of messages being dropped or delayed by the network between nodes


SQL
NoSQL
To SQL or to not only SQL, that is the question.

Store the data


Clean the data
Store the data
Find the data
Understand the question

DATA IN THE REAL WORLD

INCOMPLETE

NOISY

INCONSISTENT


Mentimeter
Go to www.menti.com

What to do when Pre-processing

What to do when Pre-processing

F_Name	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	Male
Jens	“1 M”	9,3	04/13/2012	Male
Johan	“1500”	8,5	April 4th 2012	Male

Merging & appending datasets

F_Name	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	Male
Jens	"1 M"	9,3	04/13/2012	Male
Johan	"1500"	8,5	April 4th 2012	Male
Michelle	6,2 k	10,0	3 June 2015	Female
Niels	0.36 M		19/08/12	Male
Matthias	5 M	9,4	Mar-04-14	Male
Robin	158620	100	04-17-2017	Male

What to do when

Pre-processing

Merging & appending datasets

F_Name	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	Male
Jens	"1 M"	9,3	04/13/2012	Male
Johan	"1500"	8,5	April 4th 2012	Male
Michelle	6,2 k	10,0	3 June 2015	Female
Niels	0.36 M		19/08/12	Male
Matthias	5 M	9,4	Mar-04-14	Male
Robin	158620	100	04-17-2017	Male

NCOI LEARNING

Rename variables

FirstName	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	Male
Jens	“1 M”	9,3	04/13/2012	Male
Johan	“1500”	8,5	April 4th 2012	Male
Michelle	6,2 k	10,0	3 June 2015	Female
Niels	0.36 M		19/08/12	Male
Matthias	5 M	9,4	Mar-04-14	Male
Robin	158620	100	04-17-2017	Male

FirstName	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	Male
Jens	“1 M”	9,3	04/13/2012	Male
Johan	“1500”	8,5	April 4th 2012	Male
Michelle	6,2 k	10,0	3 June 2015	Female
Niels	0.36 M		19/08/12	Male
Matthias	5 M	9,4	Mar-04-14	Male
Robin	158620	100	04-17-2017	Male

What to do when

Pre-processing

Merging & appending datasets

Renaming variables

NCOI LEARNING

Data type conversion

FirstName	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	Male
Jens	1 M	9,3	04/13/2012	Male
Johan	1500	8,5	April 4th 2012	Male
Michelle	6,2 k	10,0	3 June 2015	Female
Niels	0.36 M		19/08/12	Male
Matthias	5 M	9,4	Mar-04-14	Male
Robin	158620	100	04-17-2017	Male

FirstName	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	Male
Jens	1 M	9,3	04/13/2012	Male
Johan	1500	8,5	April 4th 2012	Male
Michelle	6,2 k	10,0	3 June 2015	Female
Niels	0.36 M		19/08/12	Male
Matthias	5 M	9,4	Mar-04-14	Male
Robin	158620	100	04-17-2017	Male

What to do when Pre-processing

Merging & appending datasets

Renaming variables

Data type conversions

NCOI LEARNING

Encoding values

FirstName	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	M
Jens	1 M	9,3	04/13/2012	M
Johan	1500	8,5	April 4th 2012	M
Michelle	6,2 k	10,0	3 June 2015	F
Niels	0.36 M		19/08/12	M
Matthias	5 M	9,4	Mar-04-14	M
Robin	158620	100	04-17-2017	M

FirstName	Amount	Score	Date	Sex
David	5 k	8,2	19/03/04	M
Jens	1 M	9,3	04/13/2012	M
Johan	1500	8,5	April 4th 2012	M
Michelle	6,2 k	10,0	3 June 2015	F
Niels	0.36 M		19/08/12	M
Matthias	5 M	9,4	Mar-04-14	M
Robin	158620	100	04-17-2017	M

What to do when

Pre-processing

Merging & appending datasets

Renaming variables

Data type conversions

Encoding values

Converting units

FirstName	Amount	Score	Date	Sex
David	€ 5.000	8,2	19/03/04	M
Jens	€ 1.000.000	9,3	04/13/2012	M
Johan	€ 1.500	8,5	April 4th 2012	M
Michelle	€ 6.200	10,0	3 June 2015	F
Niels	€ 360.000		19/08/12	M
Matthias	€ 5.000.000	9,4	Mar-04-14	M
Robin	€ 158.620	100	04-17-2017	M

Converting units

FirstName	Amount	Score	Date	Sex
David	€ 5.000	8,2	19/03/04	M
Jens	€ 1.000.000	9,3	04/13/2012	M
Johan	€ 1.500	8,5	April 4th 2012	M
Michelle	€ 6.200	10,0	3 June 2015	F
Niels	€ 360.000		19/08/12	M
Matthias	€ 5.000.000	9,4	Mar-04-14	M
Robin	€ 158.620	100	04-17-2017	M

Converting units

FirstName	Amount	Score	Date	Sex
David	€ 5.000	8,2	19/03/04	M
Jens	€ 1.000.000	9,3	13/04/12	M
Johan	€ 1.500	8,5	04/04/12	M
Michelle	€ 6.200	10,0	03/06/15	F
Niels	€ 360.000		19/08/12	M
Matthias	€ 5.000.000	9,4	04/03/14	M
Robin	€ 158.620	100	17/04/17	M

FirstName	Amount	Score	Date	Sex
David	€ 5.000	8,2	19/03/04	M
Jens	€ 1.000.000	9,3	13/04/12	M
Johan	€ 1.500	8,5	04/04/12	M
Michelle	€ 6.200	10,0	03/06/15	F
Niels	€ 360.000		19/08/12	M
Matthias	€ 5.000.000	9,4	04/03/14	M
Robin	€ 158.620	100	17/04/17	M

What to do when

Pre-processing

Merging & appending datasets

Renaming variables

Data type conversions

Encoding values

Converting units

NCOI LEARNING

Handling missing data

FirstName	Amount	Score	Date	Sex
David	€ 5.000	8,2	19/03/04	M
Jens	€ 1.000.000	9,3	13/04/12	M
Johan	€ 1.500	8,5	04/04/12	M
Michelle	€ 6.200	10,0	03/06/15	F
Niels	€ 360.000	Average of scores?	19/08/12	M
Matthias	€ 5.000.000	9,4	04/03/14	M
Robin	€ 158.620	100	17/04/17	M

What to do when Pre-processing

Merging & appending datasets

Renaming variables

Data type conversions

Encoding values

Converting units

Handling missing data

FirstName	Amount	Score	Date	Sex
David	€ 5.000	8,2	19/03/04	M
Jens	€ 1.000.000	9,3	13/04/12	M
Johan	€ 1.500	8,5	04/04/12	M
Michelle	€ 6.200	10,0	03/06/15	F
Niels	€ 360.000	Average of scores	19/08/12	M
Matthias	€ 5.000.000	9,4	04/03/14	M
Robin	€ 158.620	100	17/04/17	M

What to do when Pre-processing

Merging & appending datasets

Renaming variables

Data type conversions

Encoding values

Converting units

Handling missing data

NCOI LEARNING

Handling anomalous data

FirstName	Amount	Score	Date	Sex
David	€ 5.000	8,2	19/03/04	M
Jens	€ 1.000.000	9,3	13/04/12	M
Johan	€ 1.500	8,5	04/04/12	M
Michelle	€ 6.200	10,0	03/06/15	F
Niels	€ 360.000	Average of scores	19/08/12	M
Matthias	€ 5.000.000	9,4	04/03/14	M
Robin	€ 158.620	10,0	17/04/17	M

5 TASKS IN DATA PRE-PROCESSING

CLEANING

INTEGRATION

TRANSFORMATION

REDUCTION

DISCRETIZATION

Clean the data


Clean the data
Explore the data
Store the data
Find the data
Understand the question


Mentimeter
Go to www.menti.com

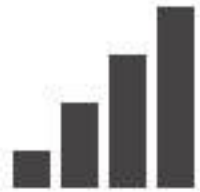
It's all about
gaining insight

... by asking questions about the data.

- Amount of data
- Data types
- Outliers
- Missing values
- Distributions
- Connections between data

Exploratory data analysis

standard deviation



minimum

maximum

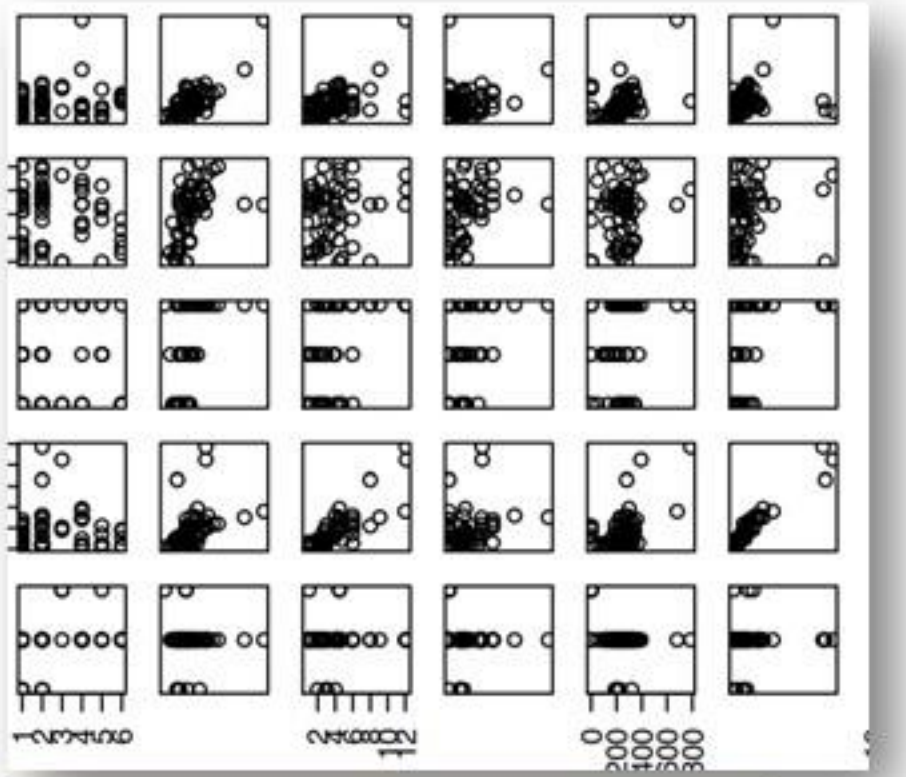
median

average

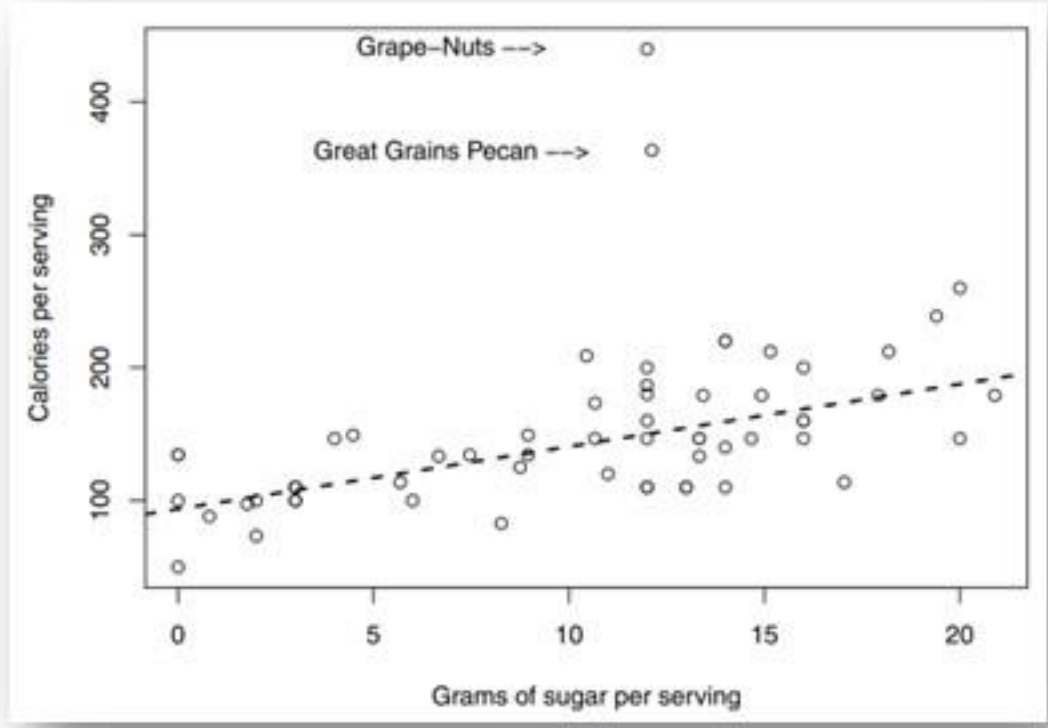
count

Exploratory vs explanatory

data analysis



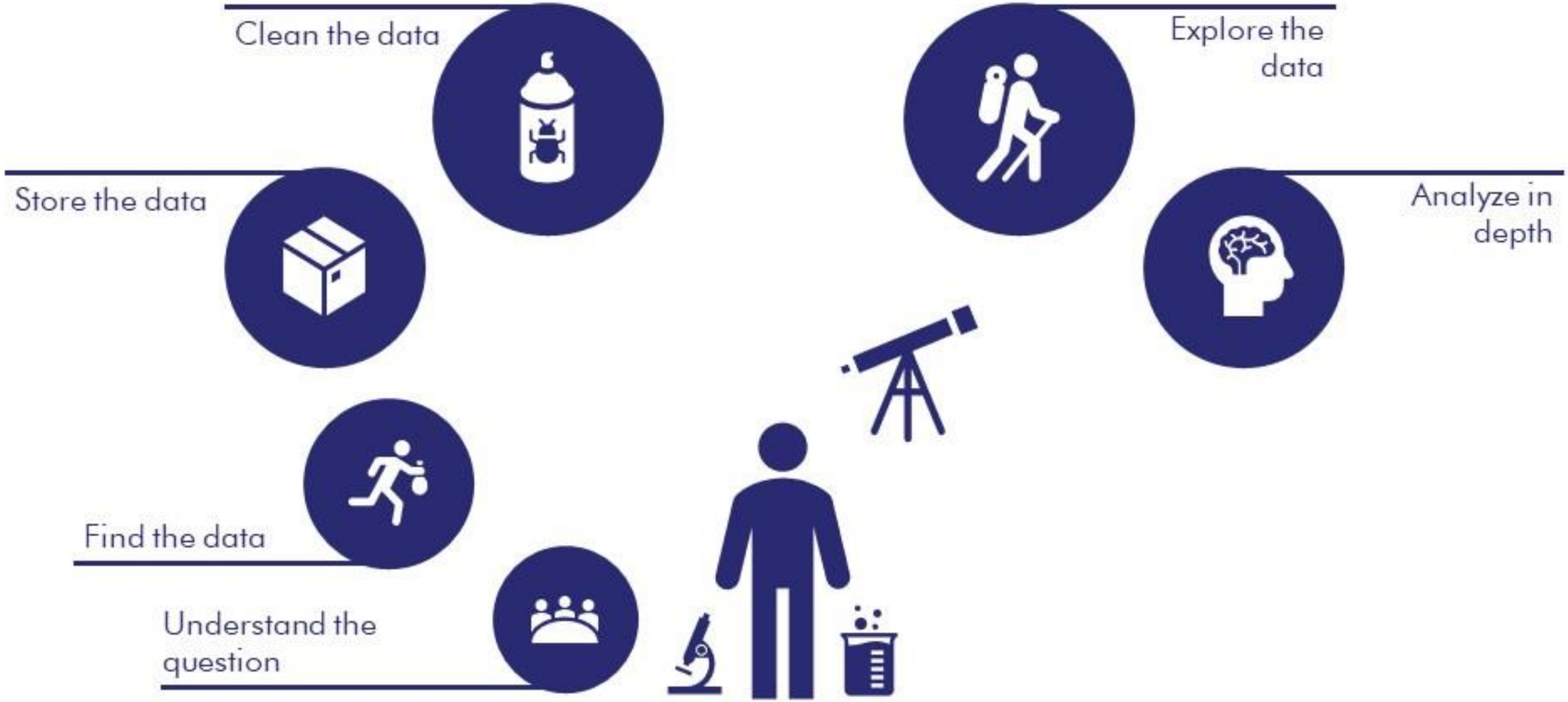
Get a sense of what's inside



Convey conclusions to others

Explore the data


Mentimeter
Go to www.menti.com


Clean the data
Explore the data
Store the data
Analyze in depth
Find the data
Understand the question

Explore the data

What do we have here?

Analyze in depth

What will we do with it?

Analyze in depth

Business Intelligence

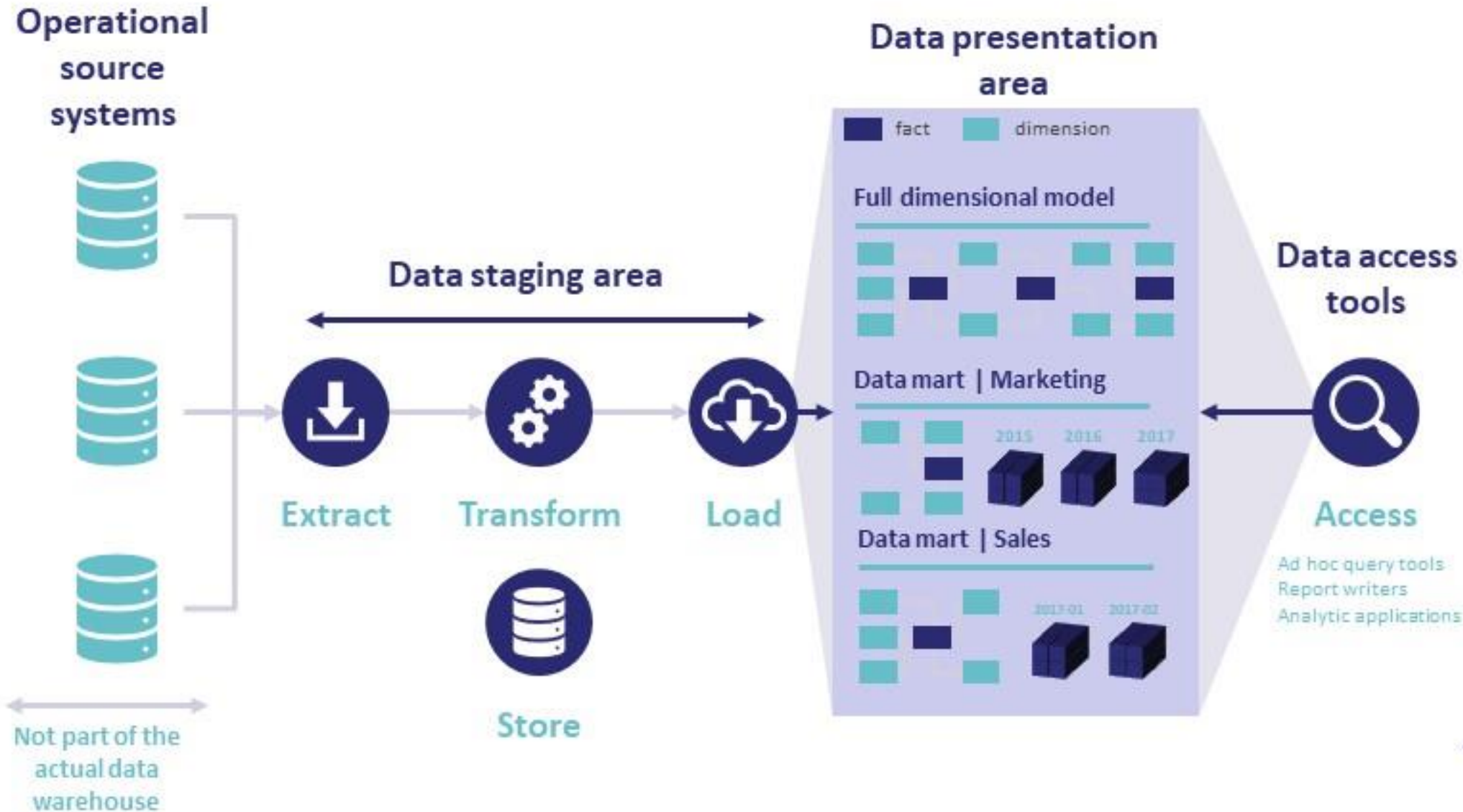
Data Mining

Analyze in depth

Business Intelligence

Data Mining

Components of a data warehouse



Analyze in depth

Business Intelligence

Data Mining

Analyze in depth

Business Intelligence

Data Mining

What do we mean when we say data mining?

What is data mining?

- Data mining is a process...
- ... that aims to find patterns in data, moving from raw data to insight and knowledge...
- ... by using information technology, creativity, business knowledge and common sense...
- ... in support of business decision-making.

These patterns are

True	Knowledge can be generalised
New	Knowledge is not yet known
Useful	Knowledge can be used to take action

These words refer to the same thing

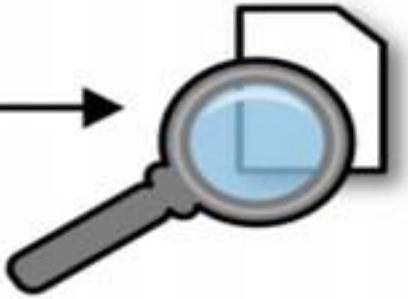
- Knowledge Discovery in Databases KDD
- Information Harvesting
- Information Discovery
- Knowledge Extraction

Historical Data

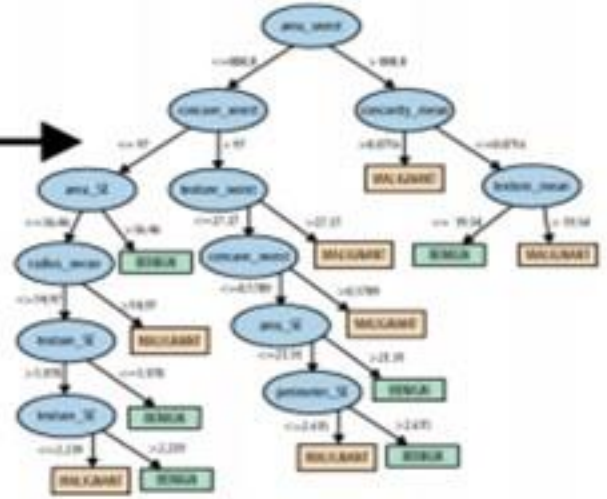
x	y	z	class
14	True	Red	accepted
6	True	Blue	rejected
...			
50.3	False	Red	accepted

Training data have all values specified

Data mining



Model



Model is deployed

Mining

Use

New data item

x	y	z	class
30	false	Red	?

New data item has class value unknown (e.g. will customer accept?)

Model

Class: accepted, Probability: 0.88

Example: classification

Let's try

Make a decision tree based on this data

The goal is to decide **whether or not to grant a loan.**

Monthly income	Requested loan	Loan granted
€ 200	€ 88.000	No
€ 1.100	€ 80.000	Yes
€ 1.400	€ 75.000	Yes
€ 500	€ 50.000	No
€ 2.200	€ 180.000	No
€ 2.500	€ 105.000	No
€ 4.000	€ 99.000	Yes

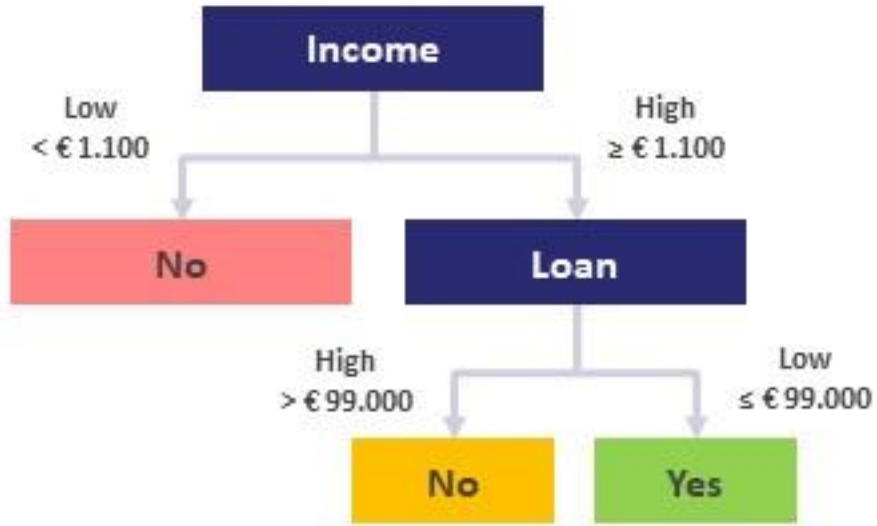
Example: classification

Let's try

Make a decision tree based on this data

The goal is to decide **whether or not to grant a loan.**

Monthly income	Requested loan	Loan granted
€ 200	€ 88.000	No
€ 1.100	€ 80.000	Yes
€ 1.400	€ 75.000	Yes
€ 500	€ 50.000	No
€ 2.200	€ 180.000	No
€ 2.500	€ 105.000	No
€ 4.000	€ 99.000	Yes



Monthly income	Requested loan	Loan granted
€ 1.500	€ 75.000	?
€ 500	€ 80.000	?

Data mining tasks

This is what we want to do

Classification & class probability estimation

Attempt to predict, for each individual in a population, to which of a (small) set of classes this individual belongs

> Among all our customers, which are likely to respond to a given offer?

Regression or value estimation

Attempt to estimate or predict, for each individual, the numerical value of some variable for that individual

> How much money will a given customer pay to use our service?

Causal modelling

Attempt to help us understand what events or actions actually influence others

> Did X happen because of Y or was this just a coincidence?

Clustering

Attempt to group individuals in a population together by their similarity, but not driven by any specific purpose

> Do our customers form natural groups or segments?

Co-occurrence grouping

Attempt to find associations between entities based on transactions involving them

> What items are commonly purchased together?

Profiling

Attempt to characterize the typical behavior of an individual, group, or population

> What is the typical cell phone usage of this customer segment?

Link prediction

Attempt to predict connections between data items, usually by suggesting that a link should exist, and possibly estimating the strength of the link

> Since you and Karen share 10 friends, maybe you'd like to be Karen's friend?

Data reduction

Attempt to take a large set of data and replace it with a smaller set of data that contains much of the important information in the larger set

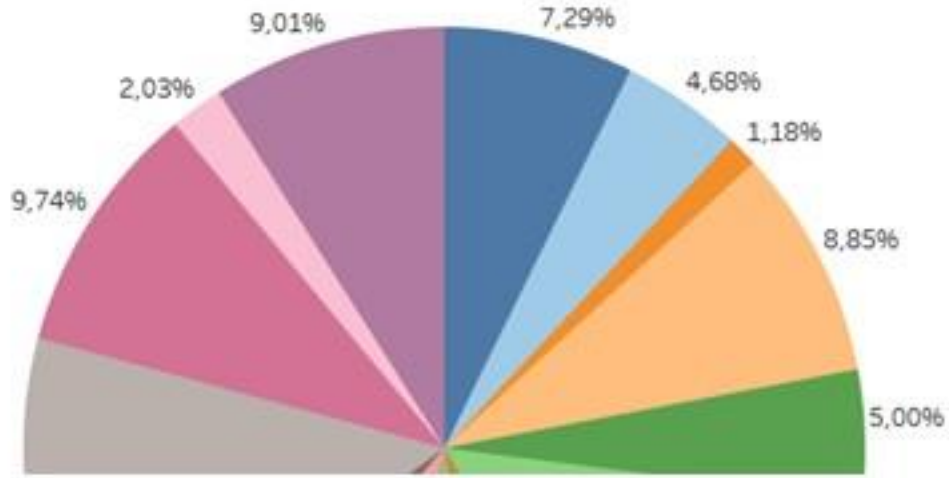
Similarity matching

Attempt to identify similar individuals based on data known about them

> Which companies are similar to our best customers? Let's focus our sales efforts on those.

NCOI LEARNING

Analyze in depth



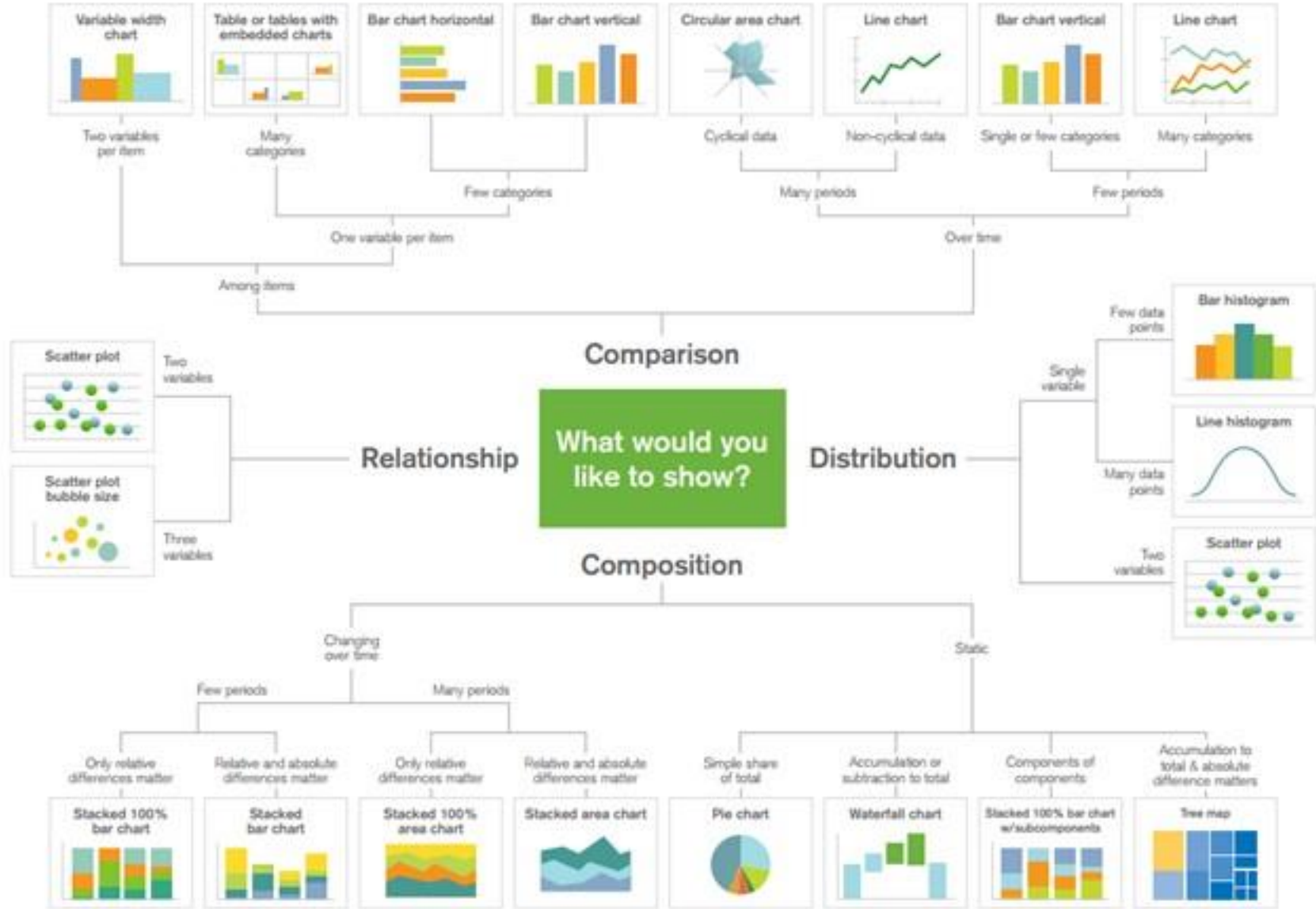
Sub-Category
- Accessories
- Appliances
- Art
- Binders
- Bookcases
- Chairs
- Copiers
- Envelopes
- Fasteners
- Furnishings
- Labels
- Machines
- Paper
- Phones
- Storage
- Supplies
- Tables

Mentimeter

Go to **www.menti.com**


Clean the data
Explore the data
Store the data
Analyze in depth
Find the data
Visualize the results
Understand the question

Visualizations



Source: ©A. Abela, 2010 www.ExtremePresentation.com

Visualize the results


Clean the data
Explore the data
Store the data
Analyze in depth
Find the data
Visualize the results
Understand the question
Tell the story

Situation

Complication

Question

Answer

Tell the story


PROGENEX

Let's get started.

Thank you

Ann Van Eyken

ann.vaneyken@themasterlabs.com

NCOI LEARNING

THE MASTER LABS