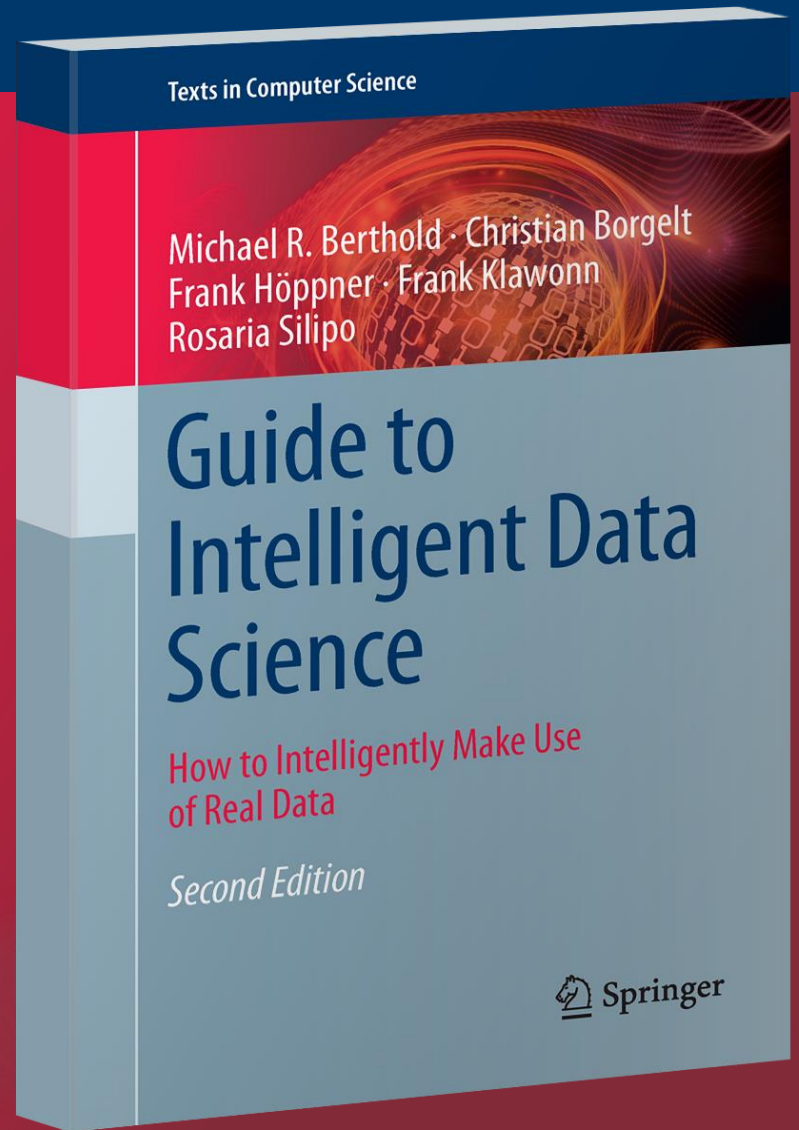


# Project & Data Understanding



*„... the goal of the project understanding phase is to assess the main objective, the potential benefits, as well as the constraints, assumptions, and risks”*

How do we identify the main objective of a project, and plan the approach?

*\*This lesson refers to chapter 3 and part of chapter 4 of the GIDS book*

## Content of this lesson

- Some Classic Use Cases
- Project Understanding
- ETL: Extraction, Transformation Loading
- Data Understanding
- Describing your Data
- Finding Patterns
- Finding Models
- Finding Predictors
- A tiny bit of History
- One final word of Warning: Correlation vs. Causality

# Some Classic Use Cases

- Churn Prediction: will a customer quit the contract?



- CRM System  
Data about your customer
- Demographics
  - Behavior
  - Revenues



Model



- Churn Prediction
- Upselling Likelihood
- Product Propensity /NBO
- Campaign Management
- Customer Segmentation
- ...

- Customer Segmentation: which groups of customers am I serving?



- CRM System  
Data about your customer
- Demographics
  - Behavior
  - Revenues

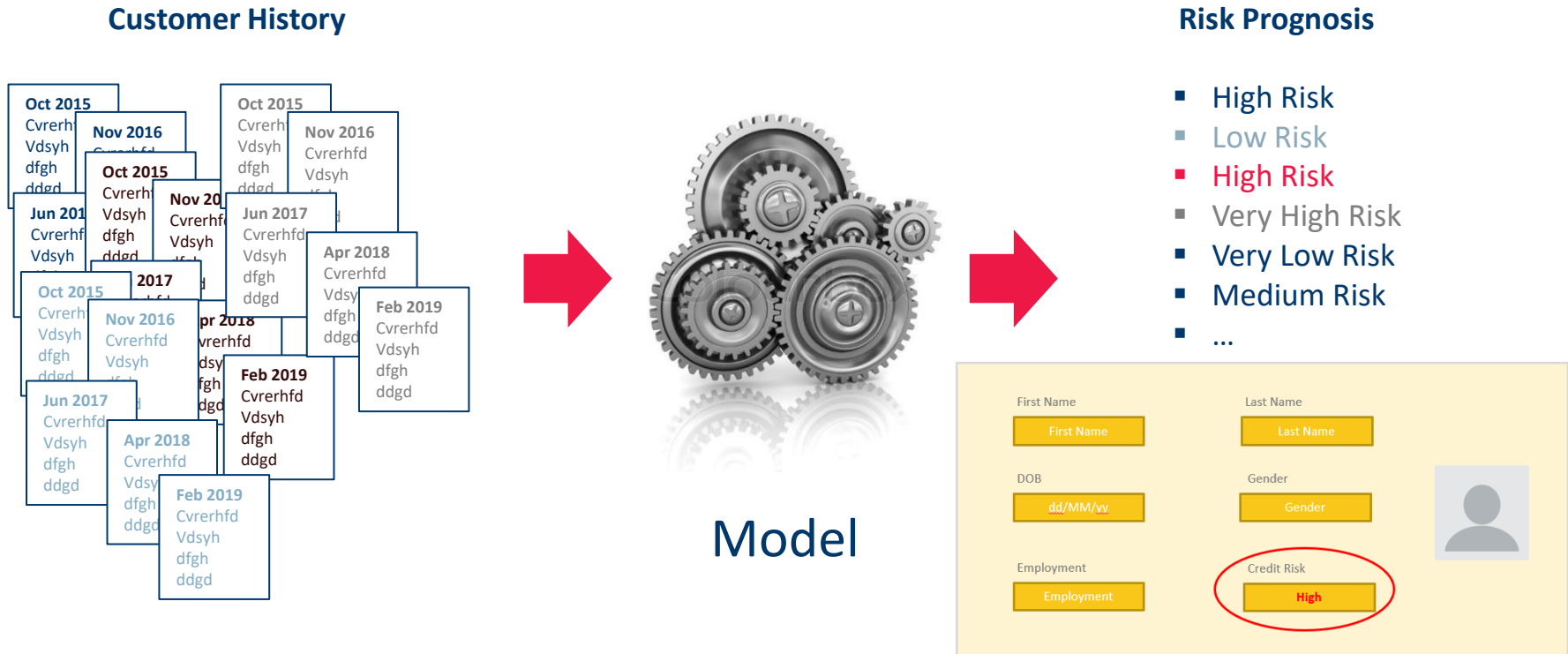


Model



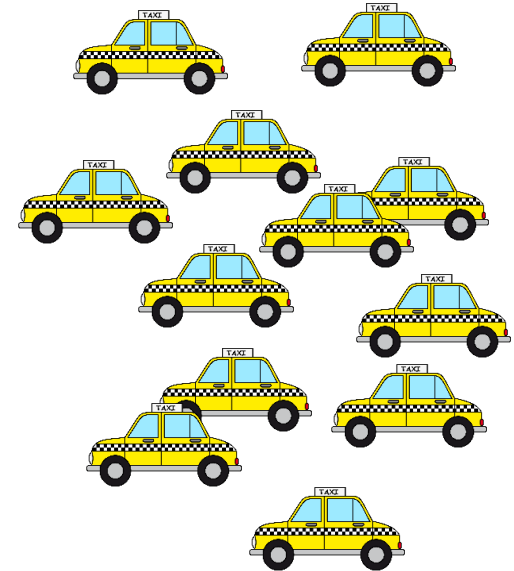
- Churn Prediction
- Upselling Likelihood
- Product Propensity /NBO
- Campaign Management
- **Customer Segmentation**
- ...

- Risk Assessment: is this person going to repay the loan?



# Demand Prediction

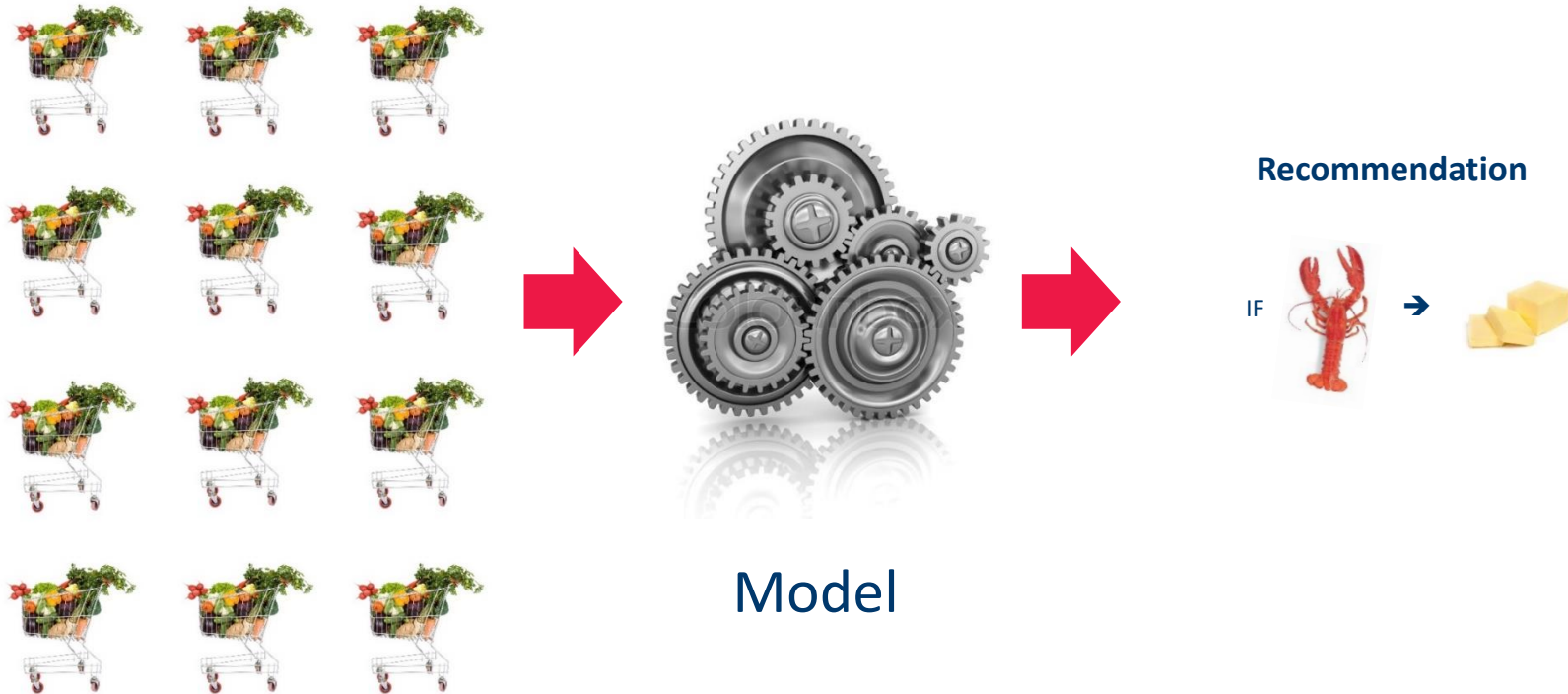
- How many taxis do I need in NYC on Wednesday at noon?
- Or how many kW will be required tomorrow at 6am in London?
- Or how many customers will come tonight to my restaurant?



Model



- Recommendation Engines: People who bought this item were often interested in this other items.



- Fraud Detection: Is this transaction legitimate or is it a fraud?



## Transactions

- Trx 1
- Trx 2
- Trx 3
- Trx 4
- Trx 5
- Trx 6
- ...



## Model

## Suspicious Transaction

Trans ID	Trans Date	Trans Time	Trans Type	Trans Amount	Trans Status	Trans Description
1001	2011-01-01	10:00:00	Payment	100.00	Success	Payment
1002	2011-01-01	10:05:00	Payment	200.00	Success	Payment
1003	2011-01-01	10:10:00	Payment	300.00	Success	Payment
1004	2011-01-01	10:15:00	Payment	400.00	Success	Payment
1005	2011-01-01	10:20:00	Payment	500.00	Success	Payment
1006	2011-01-01	10:25:00	Payment	600.00	Success	Payment
1007	2011-01-01	10:30:00	Payment	700.00	Success	Payment
1008	2011-01-01	10:35:00	Payment	800.00	Success	Payment
1009	2011-01-01	10:40:00	Payment	900.00	Success	Payment
1010	2011-01-01	10:45:00	Payment	1000.00	Success	Payment
1011	2011-01-01	10:50:00	Payment	1100.00	Success	Payment
1012	2011-01-01	10:55:00	Payment	1200.00	Success	Payment
1013	2011-01-01	11:00:00	Payment	1300.00	Success	Payment
1014	2011-01-01	11:05:00	Payment	1400.00	Success	Payment
1015	2011-01-01	11:10:00	Payment	1500.00	Success	Payment
1016	2011-01-01	11:15:00	Payment	1600.00	Success	Payment
1017	2011-01-01	11:20:00	Payment	1700.00	Success	Payment
1018	2011-01-01	11:25:00	Payment	1800.00	Success	Payment
1019	2011-01-01	11:30:00	Payment	1900.00	Success	Payment
1020	2011-01-01	11:35:00	Payment	2000.00	Success	Payment
1021	2011-01-01	11:40:00	Payment	2100.00	Success	Payment
1022	2011-01-01	11:45:00	Payment	2200.00	Success	Payment
1023	2011-01-01	11:50:00	Payment	2300.00	Success	Payment
1024	2011-01-01	11:55:00	Payment	2400.00	Success	Payment
1025	2011-01-01	12:00:00	Payment	2500.00	Success	Payment
1026	2011-01-01	12:05:00	Payment	2600.00	Success	Payment
1027	2011-01-01	12:10:00	Payment	2700.00	Success	Payment
1028	2011-01-01	12:15:00	Payment	2800.00	Success	Payment
1029	2011-01-01	12:20:00	Payment	2900.00	Success	Payment
1030	2011-01-01	12:25:00	Payment	3000.00	Success	Payment
1031	2011-01-01	12:30:00	Payment	3100.00	Success	Payment
1032	2011-01-01	12:35:00	Payment	3200.00	Success	Payment
1033	2011-01-01	12:40:00	Payment	3300.00	Success	Payment
1034	2011-01-01	12:45:00	Payment	3400.00	Success	Payment
1035	2011-01-01	12:50:00	Payment	3500.00	Success	Payment
1036	2011-01-01	12:55:00	Payment	3600.00	Success	Payment
1037	2011-01-01	13:00:00	Payment	3700.00	Success	Payment
1038	2011-01-01	13:05:00	Payment	3800.00	Success	Payment
1039	2011-01-01	13:10:00	Payment	3900.00	Success	Payment
1040	2011-01-01	13:15:00	Payment	4000.00	Success	Payment
1041	2011-01-01	13:20:00	Payment	4100.00	Success	Payment
1042	2011-01-01	13:25:00	Payment	4200.00	Success	Payment
1043	2011-01-01	13:30:00	Payment	4300.00	Success	Payment
1044	2011-01-01	13:35:00	Payment	4400.00	Success	Payment
1045	2011-01-01	13:40:00	Payment	4500.00	Success	Payment
1046	2011-01-01	13:45:00	Payment	4600.00	Success	Payment
1047	2011-01-01	13:50:00	Payment	4700.00	Success	Payment
1048	2011-01-01	13:55:00	Payment	4800.00	Success	Payment
1049	2011-01-01	14:00:00	Payment	4900.00	Success	Payment
1050	2011-01-01	14:05:00	Payment	5000.00	Success	Payment
1051	2011-01-01	14:10:00	Payment	5100.00	Success	Payment
1052	2011-01-01	14:15:00	Payment	5200.00	Success	Payment
1053	2011-01-01	14:20:00	Payment	5300.00	Success	Payment
1054	2011-01-01	14:25:00	Payment	5400.00	Success	Payment
1055	2011-01-01	14:30:00	Payment	5500.00	Success	Payment
1056	2011-01-01	14:35:00	Payment	5600.00	Success	Payment
1057	2011-01-01	14:40:00	Payment	5700.00	Success	Payment
1058	2011-01-01	14:45:00	Payment	5800.00	Success	Payment
1059	2011-01-01	14:50:00	Payment	5900.00	Success	Payment
1060	2011-01-01	14:55:00	Payment	6000.00	Success	Payment
1061	2011-01-01	15:00:00	Payment	6100.00	Success	Payment
1062	2011-01-01	15:05:00	Payment	6200.00	Success	Payment
1063	2011-01-01	15:10:00	Payment	6300.00	Success	Payment
1064	2011-01-01	15:15:00	Payment	6400.00	Success	Payment
1065	2011-01-01	15:20:00	Payment	6500.00	Success	Payment
1066	2011-01-01	15:25:00	Payment	6600.00	Success	Payment
1067	2011-01-01	15:30:00	Payment	6700.00	Success	Payment
1068	2011-01-01	15:35:00	Payment	6800.00	Success	Payment
1069	2011-01-01	15:40:00	Payment	6900.00	Success	Payment
1070	2011-01-01	15:45:00	Payment	7000.00	Success	Payment
1071	2011-01-01	15:50:00	Payment	7100.00	Success	Payment
1072	2011-01-01	15:55:00	Payment	7200.00	Success	Payment
1073	2011-01-01	16:00:00	Payment	7300.00	Success	Payment
1074	2011-01-01	16:05:00	Payment	7400.00	Success	Payment
1075	2011-01-01	16:10:00	Payment	7500.00	Success	Payment
1076	2011-01-01	16:15:00	Payment	7600.00	Success	Payment
1077	2011-01-01	16:20:00	Payment	7700.00	Success	Payment
1078	2011-01-01	16:25:00	Payment	7800.00	Success	Payment
1079	2011-01-01	16:30:00	Payment	7900.00	Success	Payment
1080	2011-01-01	16:35:00	Payment	8000.00	Success	Payment
1081	2011-01-01	16:40:00	Payment	8100.00	Success	Payment
1082	2011-01-01	16:45:00	Payment	8200.00	Success	Payment
1083	2011-01-01	16:50:00	Payment	8300.00	Success	Payment
1084	2011-01-01	16:55:00	Payment	8400.00	Success	Payment
1085	2011-01-01	17:00:00	Payment	8500.00	Success	Payment
1086	2011-01-01	17:05:00	Payment	8600.00	Success	Payment
1087	2011-01-01	17:10:00	Payment	8700.00	Success	Payment
1088	2011-01-01	17:15:00	Payment	8800.00	Success	Payment
1089	2011-01-01	17:20:00	Payment	8900.00	Success	Payment
1090	2011-01-01	17:25:00	Payment	9000.00	Success	Payment
1091	2011-01-01	17:30:00	Payment	9100.00	Success	Payment
1092	2011-01-01	17:35:00	Payment	9200.00	Success	Payment
1093	2011-01-01	17:40:00	Payment	9300.00	Success	Payment
1094	2011-01-01	17:45:00	Payment	9400.00	Success	Payment
1095	2011-01-01	17:50:00	Payment	9500.00	Success	Payment
1096	2011-01-01	17:55:00	Payment	9600.00	Success	Payment
1097	2011-01-01	18:00:00	Payment	9700.00	Success	Payment
1098	2011-01-01	18:05:00	Payment	9800.00	Success	Payment
1099	2011-01-01	18:10:00	Payment	9900.00	Success	Payment
1100	2011-01-01	18:15:00	Payment	10000.00	Success	Payment

- Sentiment Analysis: how can I know what people are thinking?



Samsung

Samsung Galaxy S7 Edge G935A 32GB Unlocked - Gold Platinum

★★★★☆ 125 customer reviews | 606 answered questions

★★★★★ **Beautiful phone from a wonderful seller!**

By \_\_\_\_\_ on May 29, 2017

Color: Gold | **Verified Purchase**

This practically new beautiful phone well exceeded my expectations!



★☆☆☆☆ **One Star**

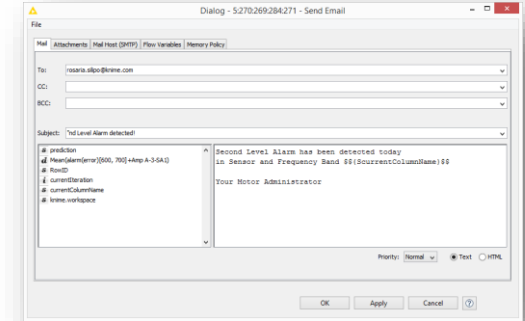
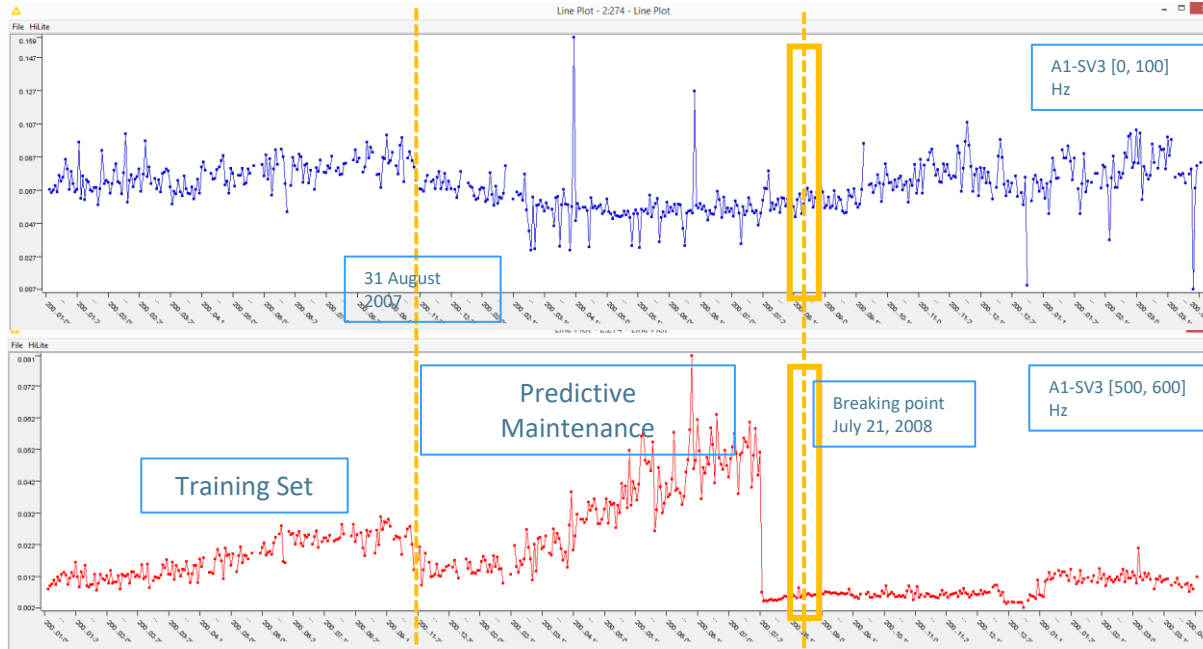
By \_\_\_\_\_ on August 3, 2016

Color: Black Onyx | **Verified Purchase**

Very bad experience



## Predicting mechanical failure as late as possible but before it happens

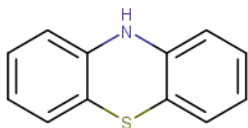


via REST

Only some Spectral Time Series show the break down

# Compound Search

Are there other compounds having this substructure and being a dopaminergic antagonist?



Enter Search Options

Enter a SMILES to search ChEBI for compounds (substructure based search):  
N1C2=CC=CC=C2SC2=CC=CC=C12

Type in a role of your interest and select using the autocomplete function a valid term:  
dopamine|

- dopamine receptor D2 antagonist
- EC 1.14.17.1 (dopamine beta-monoxygenase) inhibitor
- dopaminergic agent
- dopaminergic antagonist
- dopamine uptake inhibitor
- dopamine agonist

Reset Apply Close



Selected Compound

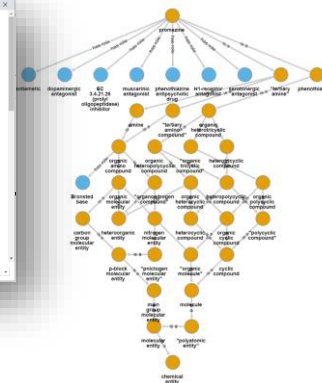
Network View

List of compounds matching your query: Select only one!

 SMILES: <chem>C1=CC=C(C=C1)C2=CC=CC=C2N3C=CC=CC3</chem>	 SMILES: <chem>C1=CC=C(C=C1)C2=CC=CC=C2N3C=CC=CC3</chem>	 SMILES: <chem>C1=CC=C(C=C1)C2=CC=CC=C2N3C=CC=CC3</chem>	 SMILES: <chem>C1=CC=C(C=C1)C2=CC=CC=C2N3C=CC=CC3</chem>
 SMILES: <chem>C1=CC=C(C=C1)C2=CC=CC=C2N3C=CC=CC3</chem>	 SMILES: <chem>C1=CC=C(C=C1)C2=CC=CC=C2N3C=CC=CC3</chem>	 SMILES: <chem>C1=CC=C(C=C1)C2=CC=CC=C2N3C=CC=CC3</chem>	 SMILES: <chem>C1=CC=C(C=C1)C2=CC=CC=C2N3C=CC=CC3</chem>

Selected compound has following roles assigned:

ChEBI LABEL	relationship	parent LABEL	definition_of parent LABEL
dopaminergic antagonist	has role	EC 1.14.17.1 (dopamine beta-monoxygenase) inhibitor	any EC 1.14.17.1 (dopamine beta-monoxygenase) inhibitor that interferes with the action of beta-monoxygenase (EC 1.14.17.1).
dopaminergic antagonist	has role	D2 receptor antagonist	D2 receptor antagonist are the drugs that selectively bind to but do not activate dopamine D2 receptors, thereby blocking the actions of endogenous dopamine.
dopaminergic antagonist	has role	antidote	A drug acting against a disease or condition, as antidotes may act to a wide range of mechanisms, it might affect the necessary cellular entry, the working order and the downstream trigger (and/or) affect the peripheral receptors.
dopaminergic antagonist	has role	dopaminergic antagonist	A drug that binds to but does not activate dopamine receptors, thereby blocking the actions of dopamine or dopamine agonists.
dopaminergic antagonist	has role	dopaminergic antagonist	A drug that binds to but does not activate dopamine receptors, thereby blocking the actions of endogenous dopamine or dopamine agonists.



# Project Understanding

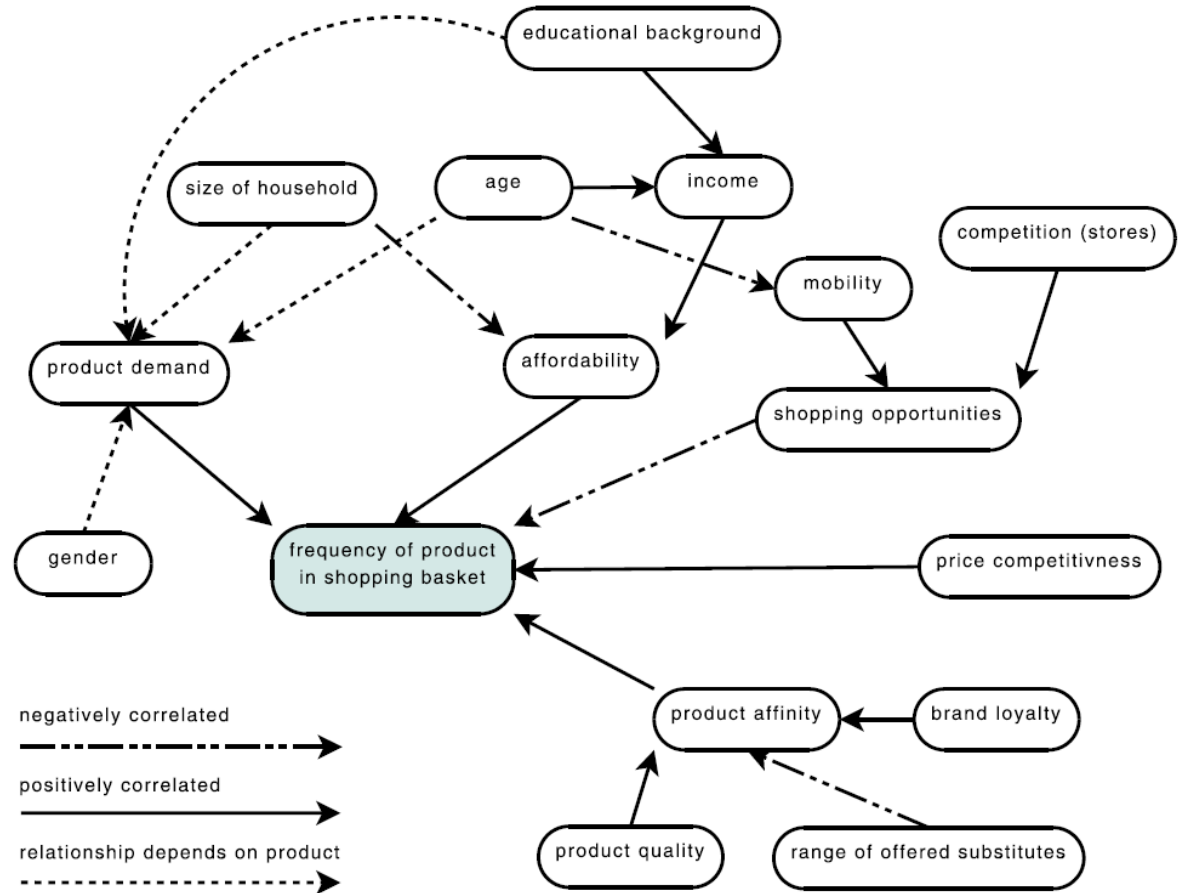
## Determine the Project Objective

- What is the primary objective?
- What are the criteria for success?
- These are difficult to define
  - The project owner & the analysis *speak different languages*

Problem source	Project owner perspective	Analyst perspective
Communication	Project owner does not understand the technical terms of the analyst	Analyst does not understand the terms of the domain of the project owner
Lack of understanding	Project owner was not sure what the analyst could do or achieve Models of analyst were different from what the project owner envisioned	Analyst found it hard to understand how to help the project owner
Organization	Requirements had to be adopted in later stages as problems with the data became evident	Project owner was an unpredictable group (not so concerned with the project)

# Cognitive maps

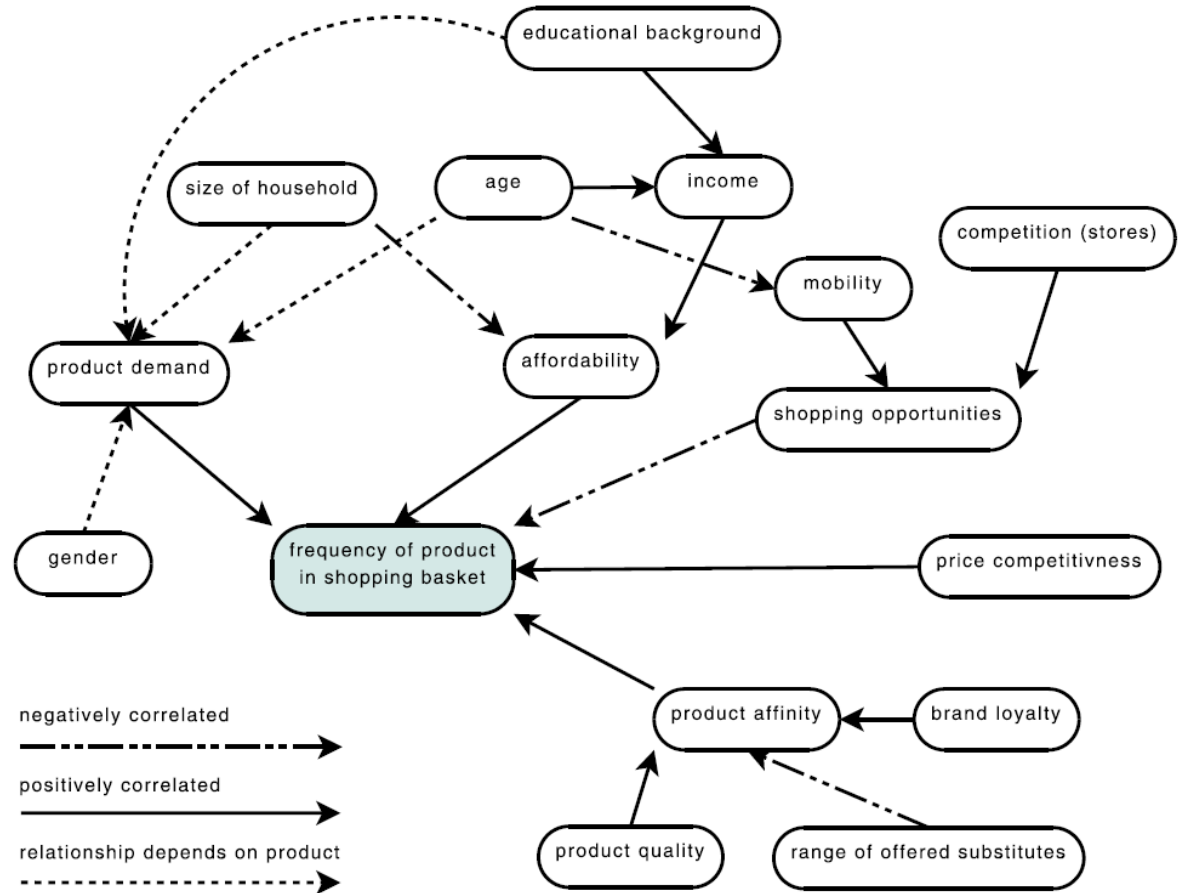
- Tool to sketch
  - Beliefs
  - Experiences
  - Known factors
  - How they influence each other





# Cognitive maps

- How often will a certain product be found in a basket
  - Directly influenced by factors around it
    - E.g., affordability
  - Indirectly influenced by other factors
    - E.g., size of household
  - Positive or negative correlation



## Clarifying the Primary Objectives

- Once the solution is identified
  - Explore advantages & disadvantages
- Is the goal
  - Precise enough?
  - Actionable?

<b>Objective</b>	Increase revenues (per campaign and/or per customer) in direct mailing campaigns by personalized offer and individual customer selection
<b>Deliverable</b>	Software that automatically selects a specified number of customers from the database to whom the mailing shall be sent, runtime max. half-day for database of current size
<b>Success criteria</b>	Improve order rate by 5% or total revenues by 5%, measured within 4 weeks after mailing was sent, compared to rate of last 3 mailings

- Will this be a successful data analysis project?
- Examine the following:
  - **Requirements and constraints**
    - Model requirements (e.g., explanatory model)
    - Ethical, political, and legal issues (e.g., must exclude gender, race, and/or age)
    - Technical constraints
  - **Assumptions**
    - Representativeness (the sample represents the whole population)
    - Informativeness (influencing factors should be included in the model)
    - Good data quality
    - Presence of external factors

### Select models and techniques with the following properties

- **Interpretability**
  - The model can be understood / interpreted
- **Reproducibility / stability**
  - Similar model performance every time the analysis is carried out
- **Model flexibility / adequacy**
  - The model can adapt to more complicated situations
- **Runtime**
  - Strict runtime requirements may limit computationally intensive approaches
- **Interestingness / use of expert knowledge**
  - Experts may already know the findings from the analysis

# ETL: Extraction, Transformation, Loading

Getting the data in not always easy:

- Different resources: flat files, different databases, excel spreadsheets, ...
- Integration is cumbersome: Missing/not unique IDs, wrong entries, ...
- Sometimes also privacy concerns (not all data in one location)

Data needs to be transformed:

- Type conversions
- Missing value correction/clean up/imputation
- Generation of new values (e.g. convert year of birth into age)

- Three files:
  - customers,
  - products,
  - shopping baskets.
  
- Can we load these file and create a new attribute “age”?
  
- Can we find out:
  - how often each customer went shopping
  - how much (s)he bought together (and on average)

- Database issues
- More details regarding pre-processing later:
  - Normalization
  - Binning
  - Feature (and Data!) Reduction
  - ...

## **The 80% Rule**

Over 80% of data analysts' time is spent on loading and cleaning data.



# Data Understanding

- **Goal of the Data Understanding phase**
  - Gain general insights about the data that will potentially be helpful for the further steps in the data analysis process
- **Reasons**
  - Never trust any data as long as you have not carried out some simple plausibility checks.
- **Results**
  - At the end of the data understanding phase, we know much better whether the assumptions we made during the project understanding phase concerning representativeness, informativeness, data quality, and the presence or absence of external factors are justified

# Attribute Understanding

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

**Attributes**, features, variables...

**Instances**, records, data objects, entries...

- Data can usually be described in terms of table or matrices
- Sometimes data are spread among different table that need to be **joined**

# Attribute Understanding

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Diagram illustrating the classification of attributes in the table above:

- Categorical**: Sex, Blood pr., Drug
- Ordinal**: Blood pr. (low, normal, high)
- Numeric**: Age, Height

Additional classification labels below the table:

- Numeric**: Age, Height
- Categorical**: Sex, Blood pr., Drug

- Attributes differ for their **scale type**, according to the type of values that they can assume
- Three scale types:
  - Categorical / Nominal
  - Ordinal
  - Numeric

# Categorical Attributes

Categorical

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Categorical

- Categorical (or Nominal) attributes have a finite set of possible values
- Granularity must be taken into account
  - Hierarchical structure of the categories
  - e.g. shallow subdivision: *food, non-food, drinks...*
  - further subdivision for drinks: *water, beer, wine...*
  - Which level of granularity is appropriate?
- Dynamic Domain
  - Some attributes have a fixed domain (e.g. months)
  - For other attributes the domain can change over time (e.g. the products in a catalogue)
  - Those attributes must be identified and handled

Ordinal

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

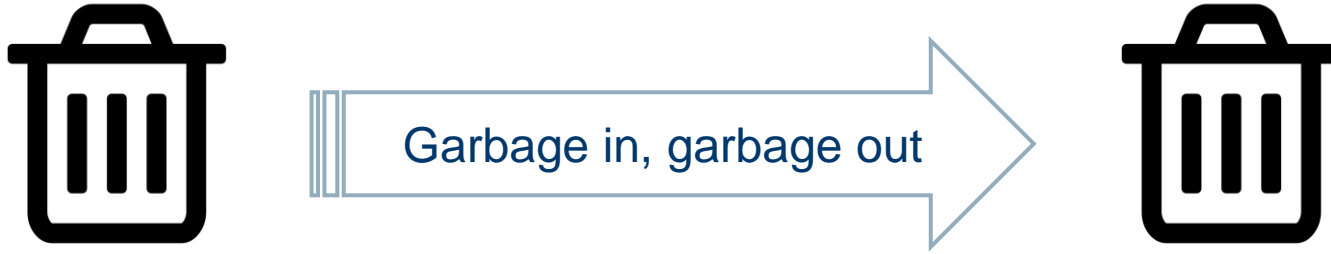
- Ordinal attributes have an additional linear ordering offered by the domain
- The ordering does not provide the distance between two object
- e.g. for an attribute containing university degrees, we can state that a *Ph.D* is an higher degree than a *M.Sc.* and that this is higher than a *B.Sc.*.

Numeric continuous

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Numeric discrete

- The domain of numerical attributes are numbers. They can be
  - **Discrete**
    - e.g. age, count...
    - Represented as integer values
  - **Continuous**
    - e.g. height, weight, distance...
    - Represented as real values
    - Precision (rounding) has to be handled
- The scale of numeric attributes can be:
  - Interval e.g. date
  - Ratio Scale e.g. distance, with a canonical zero value
  - Absolute Scale e.g. counting



- Data quality refers to how well the data fit their intended use
- There are various data quality dimensions
  - Accuracy
  - Completeness
  - Unbalanced Data
  - Timeliness



**Accuracy** is defined as the closeness between the value in the data and the true value.

## Syntactic

- The value might not be correct but it belongs at least to the domain of the corresponding attribute
- Easy to spot: verify values lying in the domain

e.g. “female” for the attribute Gender and “-15” for the attribute Weight violate the syntactic accuracy

## Semantic

- The value might be in the domain of the corresponding attribute, but it is not correct
- Hard or impossible to spot: double check with other sources or check “business rules”

e.g. “2090” for the attribute *YearOfBirth* is (at least at the moment) surely incorrect, therefore violates the semantic accuracy

- Completeness with respect to **attributes**
  - All the attributes have a value associated
  - i.e. Missing Values (coming soon in next lessons)
  - Missing values might not always be explicitly marked
  
- Completeness with respect to **records**
  - The data set contains the necessary information required for the analysis
  - Some rows might have been lost for various reasons (e.g. during DB migration)
  - Sometimes data about a certain situation simply does not exist (e.g. data about a failure that has never –yet- occurred)
  - It is hard to obtain a reasonably wide dataset containing all the possible combinations of data

### Unbalanced Data

- Data regarding a certain situation might be underrepresented
- E.g. machine quality control: parts produced with flaws are – hopefully – lower than the correct ones, therefore the corresponding data will be way less

### Timeliness

- Available data are too old to provide up to date information
- Often a problem in dynamically changing domains, where older data might indicate trends that have vanished

# Describing your Data

## Familiarize yourself with the data

- Identify trends
- strange patterns
- outliers
- ...

## Types of views

- Basic Statistics
- 1D: Histograms
- 2D: Scatterplots, Scatter Matrix, Multi Dimensional Scaling
- 3D Scatterplots
- 3D: Parallel Coordinates

- Let's look at our data
- Can we find some connections between age and shopping cart size?
- Anything else that looks a bit odd? (...the age distribution, maybe?)
- Visualizations are a good way for first sanity checks
- Interactivity on a plot or among plots is very helpful

- Simple statistical descriptors, such as:
  - range
  - mean/median
  - standard deviation
  - nominal values and their frequencies
  - ...
- can help to sanity check your data (and find dependencies that otherwise might surprise you quite a bit afterwards!)
- Can we look at the range and other simple 1D descriptors?
- How about 2D correlations between attributes?

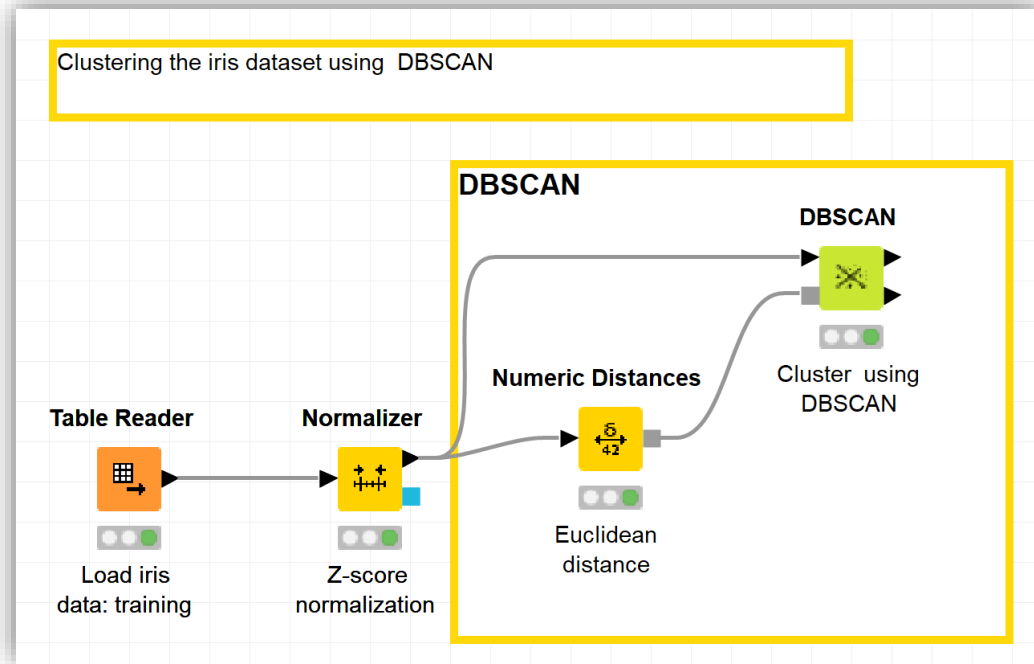
# Finding Patterns



- Finding (significant?) patterns in data may reveal interesting connections:
- Global patterns: groups of customers or products
  - Clusters
- Local patterns: connections between products, sub populations of customers (recommendation engines!)
  - Subgroups
  - Association Rules

- Can we find groups of similar customers?
- (and what does similarity mean, anyway?)
  
- **Similarity**
- Finding the right similarity metric is an art.
- (and what is a cluster anyway?)
- Distance based methods in high dimensions offer all sorts of interesting surprises...

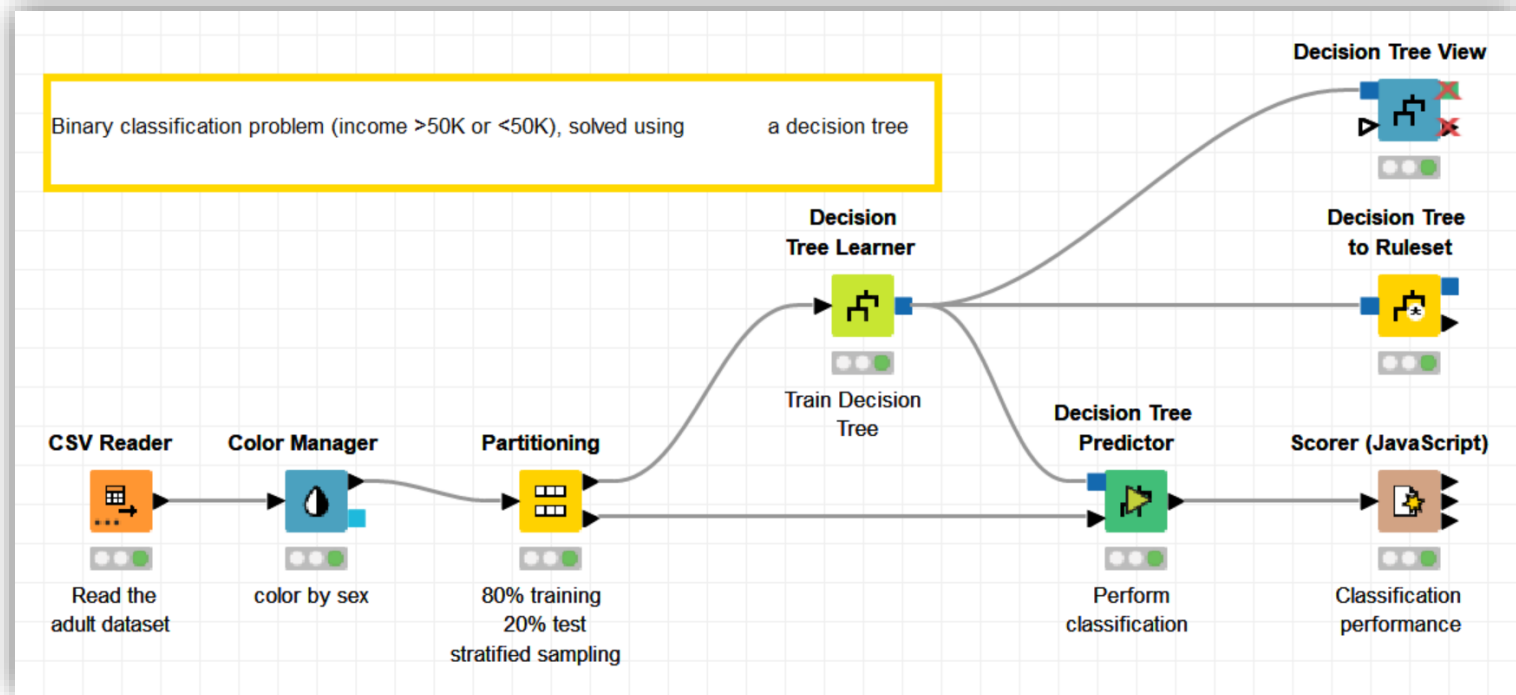
## – Screenshot of KNIME workflow with clustering



# Finding Models

- Deriving models that describe (aspects of) the data:
  - Rules
  - Trees
  - Typical (or really odd!) examples
  - ...
- Models attempt to describe what is going on in the system that “generated” the data.
- Example:
  - Can we find a decision tree describing why certain customers buy so much?

## – Screenshot of KNIME workflow with decision tree



# Finding Predictors

- Sometimes we want to find a model which we can use to later predict the target variable(s):
  - Predict future shopping behaviour
  - Predict credit risk
  - Predict activity of a chemical compound
  - Predict tomorrow's weather, stock market, ...
- And we may not care too much about actually understanding the model itself.



## Brute Force Predictors

Very simple: look at your closest neighbour

- Case based reasoning works that way
- Depends heavily on your distance function
- Does not work well with outliers/noise

Slightly better: look at a few of your neighbors

- K Nearest Neighbor
- Works pretty well
- But pretty expensive to compute...

Even better: look at all neighbors, but weight them

- Weighted K Nearest Neighbor
- Works even better
- Even more expensive...

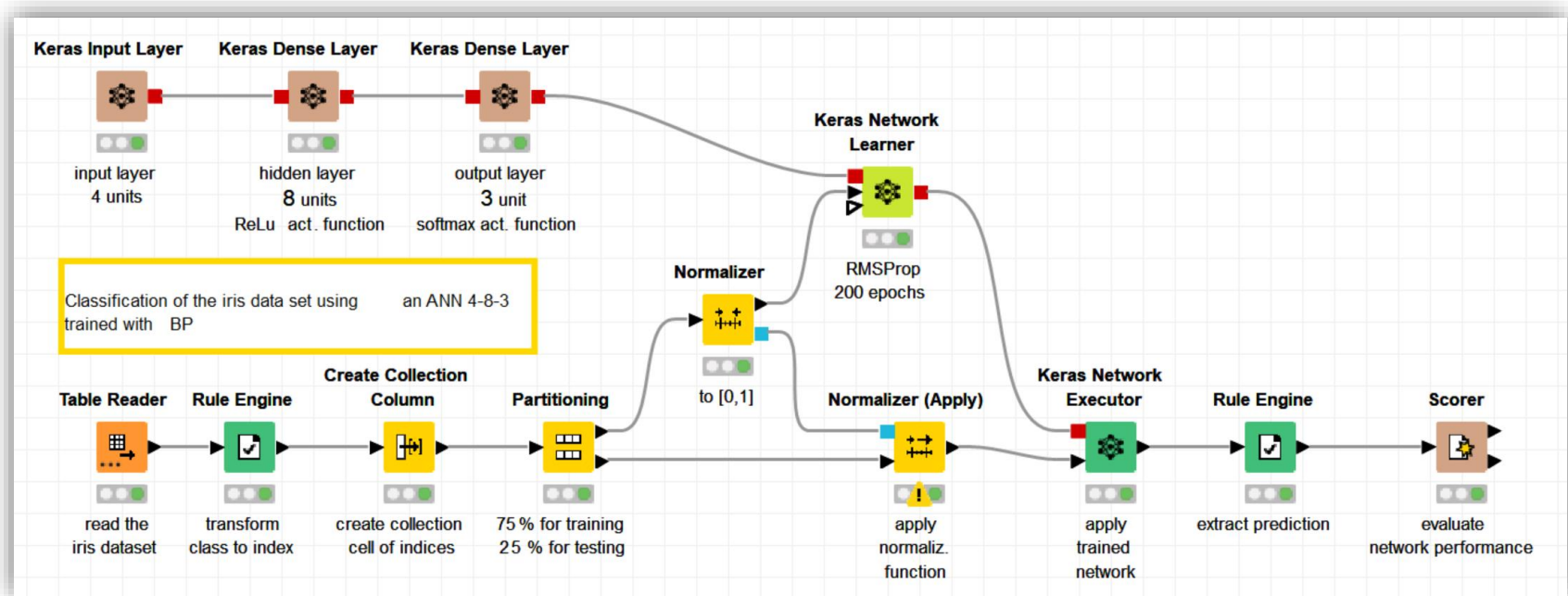
- Decision Trees, Rules, ... (all of our models!)
- (Naïve) Bayes Classifiers
- Regression
- (Artificial) Neural Networks
- Support Vector Machines (Kernel Methods)

Can we predict the size of shopping-cart?

- Brute force: look at a (few) neighbor(s).
- Use our decision tree?...

What's wrong with that approach?

## – Screenshot of KNIME workflow with a neural network



## What kind of systems do we need?

- easy to use (also by non Data Mining Expert!)
- simple knowledge representation (understandable!)
- mergers of disciplines (machine learning, stats, databases, ...)
- (partial) automation of feedback (“Intelligent” Data Science!)
- quick turn-around (interactive!)

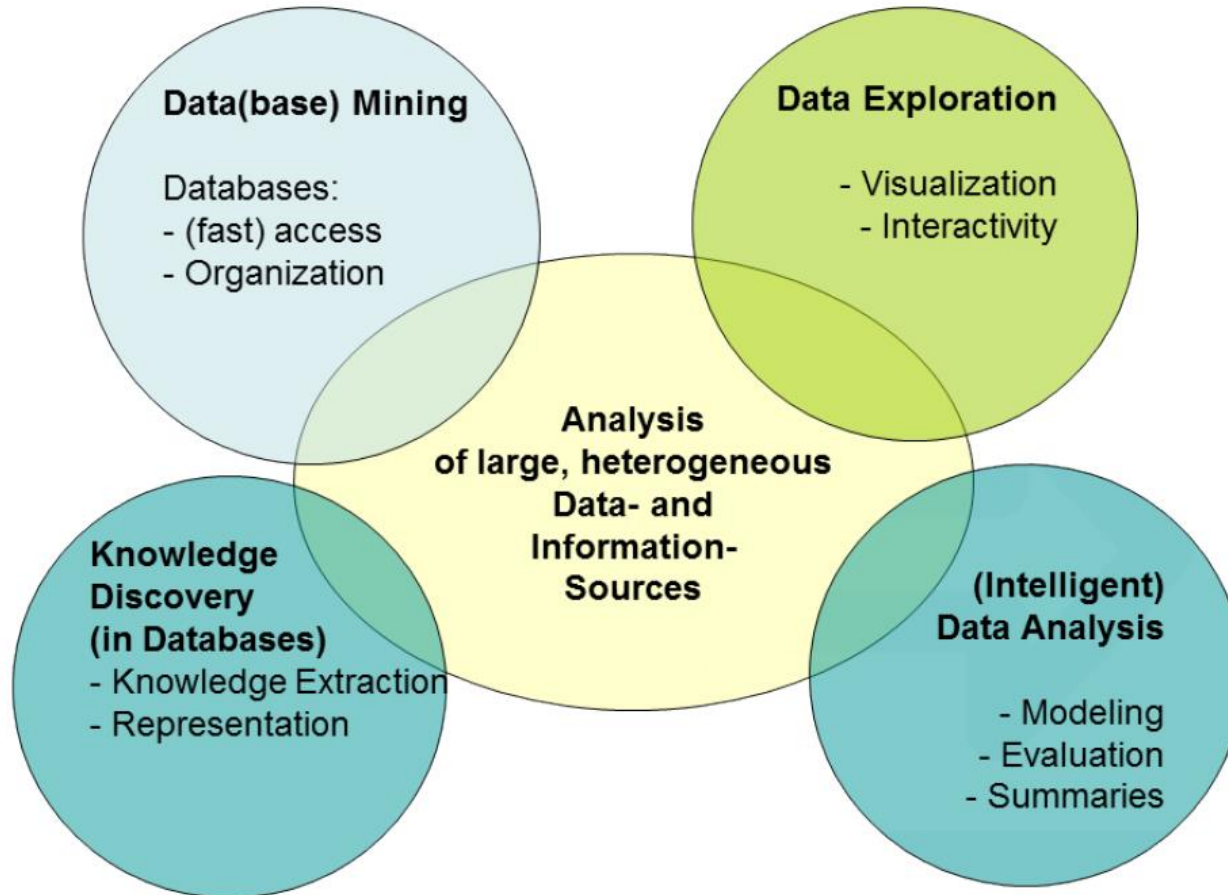
# A tiny bit of History

- **History: Classical Data Analysis**
- Small, usually manually recorded data sets
- Calculation of correlation measures and statistical significance measures.
- Calculations done with minimal to no compute support.
- Calculations later supported by basic calculation equipment

- **History: Table based Analysis**
- Data points are stored in tables, often recorded in spread sheets
- Simple analyses performed automatically on demand (calculate mean, add columns, ...)
- Visicalc, ...



- **Today: Large Scale Mining**
- Data in various formats and from various sources
- manual analysis impossible
- efficient compute support essential
- analysis still question driven:
  - find patterns of this type
  - check correlations
  - build model to predict this behaviour



# One final Word of Warning

## Correlation $\not\Rightarrow$ Causality

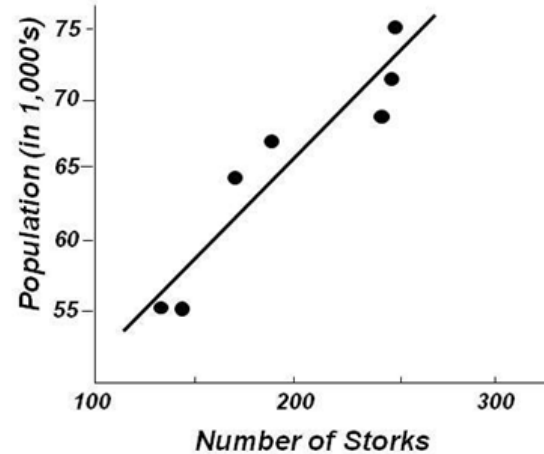
Hypothesis: Storks bring babies

And the data?

Hypothesis: Storks bring babies

And the data?

*Population of Oldenburg, Germany, at Year's End  
vs. Number of Storks Observed Each Year  
(1930 – 1936)*



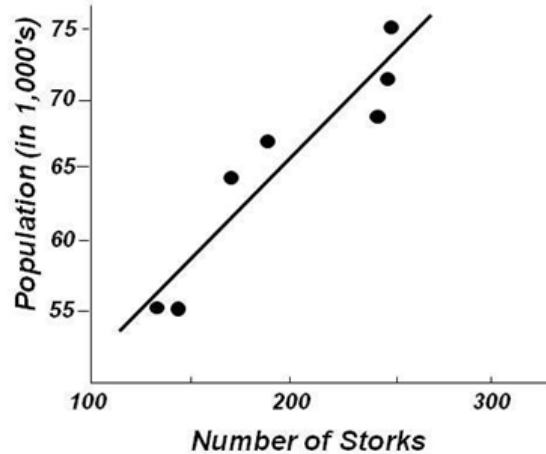
*Source: Statistics for Experimenters,  
by Box, Hunter & Hunter*

Correlation is significant and positive!

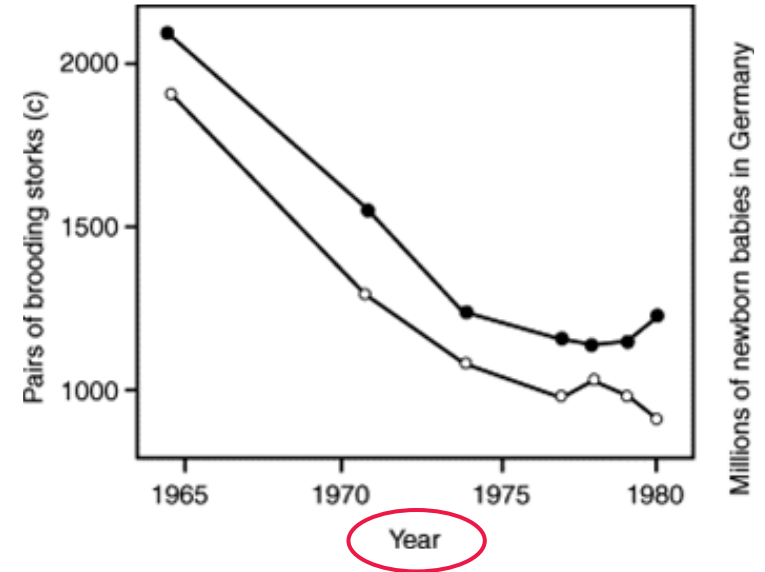
Hypothesis: Storks bring babies

And the data?

Population of Oldenburg, Germany, at Year's End  
vs. Number of Storks Observed Each Year  
(1930 - 1936)



Source: *Statistics for Experimenters*,  
by Box, Hunter & Hunter

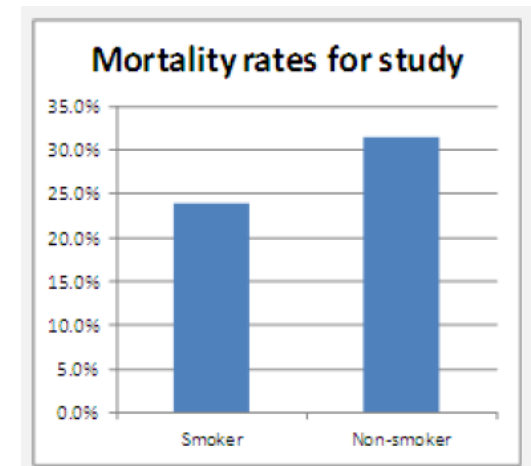


Correlation is significant and positive!

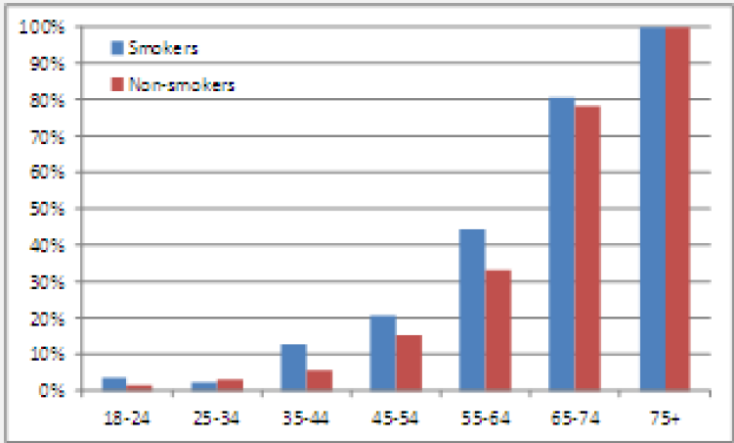
- Should I start smoking to live longer?
- **Mortality Rate Study**

	Died	Survived	Total	Rate
Smokers	139	443	582	23.9%
Non Smokers	230	502	732	31.4%
Total	369	945	1314	28.1%

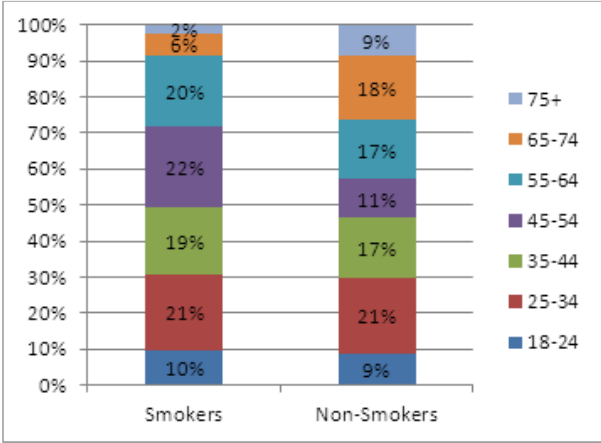
Credit: <http://www.significancemagazine.org/details/webexclusive/2671151/>



## Mortality Rates by Age



## Distribution of Age by Smoking Status



Credit: <http://www.significancemagazine.org/details/webexclusive/2671151/>

[Simpsons-Paradox-A-Cautionary-Tale-in-Advanced-Analytics.html](http://Simpsons-Paradox-A-Cautionary-Tale-in-Advanced-Analytics.html)



# Simpson's Paradox

Adjusted gross income	Tax Rate		% of total income	
	1974	1978	1974	1987
Under \$5000	0.054	0.035	4.73	1.60
\$5000 - \$9999	0.093	0.072	16.63	9.89
\$10000 - \$14999	0.111	0.100	21.89	13.83
\$15000 - \$99999	0.160	0.159	53.40	69.62
\$100000 and more	0.384	0.383	3.34	5.06
<b>Total</b>	<b>0.141</b>	<b>0.152</b>	<b>100</b>	<b>100</b>

Table Credit: Counting for Something by William S. Peters

... does the overall tax rate go up, while all individual rates go down?

and what about Chocolate and Nobel prices?

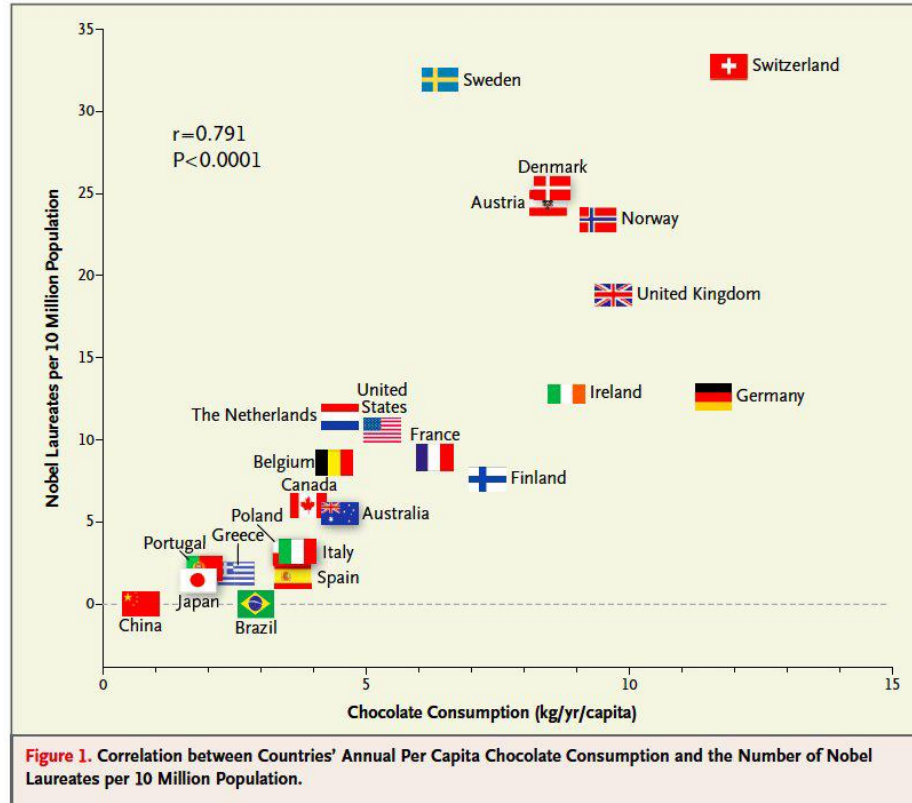


Image Credit: <http://www.nejm.org/doi/full/10.1056/NEJMon1211064>

## **Tymans's Law**

Any statistic that appears interesting is almost certainly a mistake.

- The different kind of projects
  - Common Use Cases
  - Search strategies
- The steps in project understanding
- The different kinds of datasets
- The steps in data understanding
  - ETL
  - Describing your Data
  - Finding Patterns
  - Finding Models
  - Finding Predictors
- A tiny bit of History
- Correlation vs. Causality

# Thank you

Guide to Intelligent Data Science Second Edition, 2020