Lecture 2 – Data Mining Process

Introduction to the Data Mining Process, Explanation vs. Prediction

Information Systems for Sustainable Society (is3) | WiSo Faculty | Univ.-Prof. Dr. Wolfgang Ketter | October 9th 2023

Agenda

- To Explain or To Predict?
- CRISP - Data Mining as a Process
- Common Data Science Tasks and Terminology

In data science you often look at the same problem from different angles – **Example Customer Default**

Let us look at the example of customer default on a payment

1. You may want to **analyze what drives** customer defaults is by asking:
   - **What** are the **key significant factors** that **determine** whether a **customer defaults?**
2. You may also be interested in **understanding the cause** of a default by asking:
   - **Why does** an average **customer default**?
3. Finally, when assessing whether to accept a new customer, you may be interested in the **likelihood of default**:
   - **Will this new customer pay** his/her bill or will he/she default (not pay)?

These are three perfectly sensible data science angles to take – We refer to them as **Descriptive**, **Explanatory** and **Predictive Modeling**

Descriptive Modeling

What are the **key significant factors** that determine whether a customer defaults?

Explanatory Modeling

Why does an average **customer default**?

Predictive Modeling

Will this new customer pay his/her bill or will he/she default?

Explanatory Modeling is most prevalent in the social sciences such as ecnomics and management

A typical Explanatory Model

Parameter of interest

$$Y_i|X_i = \beta_0 + \boxed{\beta_1} X_i + \boldsymbol{\beta}_2 X_{controls} + \epsilon_i$$

Controls to reduce omitted variable bias

Error term – danger of endogeneity if correlated with dependent variable

- Definition
  - Theory-based, statistical **testing of causal hypotheses**
  - **Explanatory power** is measured in terms of **strength of relationship** in statistical model, e.g. magnitude and significance of paramters
- Scientific Goal
  - **Test/quantify causal effect** between constructs for **average unit** in population
  - **Reduce bias** (selection bias, omitted variable bias, etc.) as much as possible to obtain unconfounded estimates of the causal effect

Source: Galit Shmueli

Philosophy of Science

“Explanation and prediction have the same logical structure”

Hempel & Oppenheim, 1948

“It becomes pertinent to investigate the possibilities of predictive procedures autonomous of those used for explanation”

Helmer & Rescher, 1959

“Theories of social and human behavior address themselves to two distinct goals of science: (1) prediction and (2) understanding”

Dubin, *Theory Building*, 1969

In explanatory modeling you usually start with theory, which you try to proof

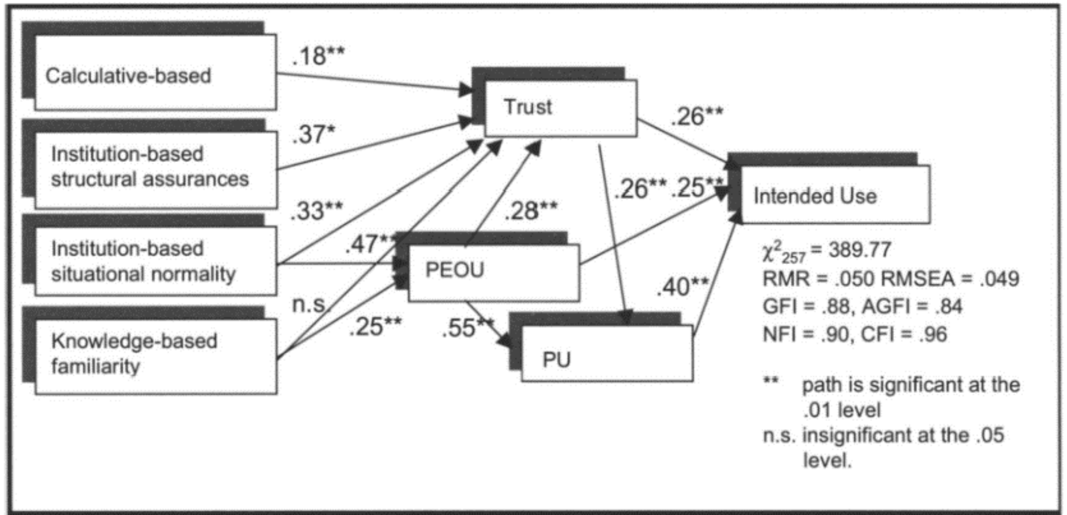
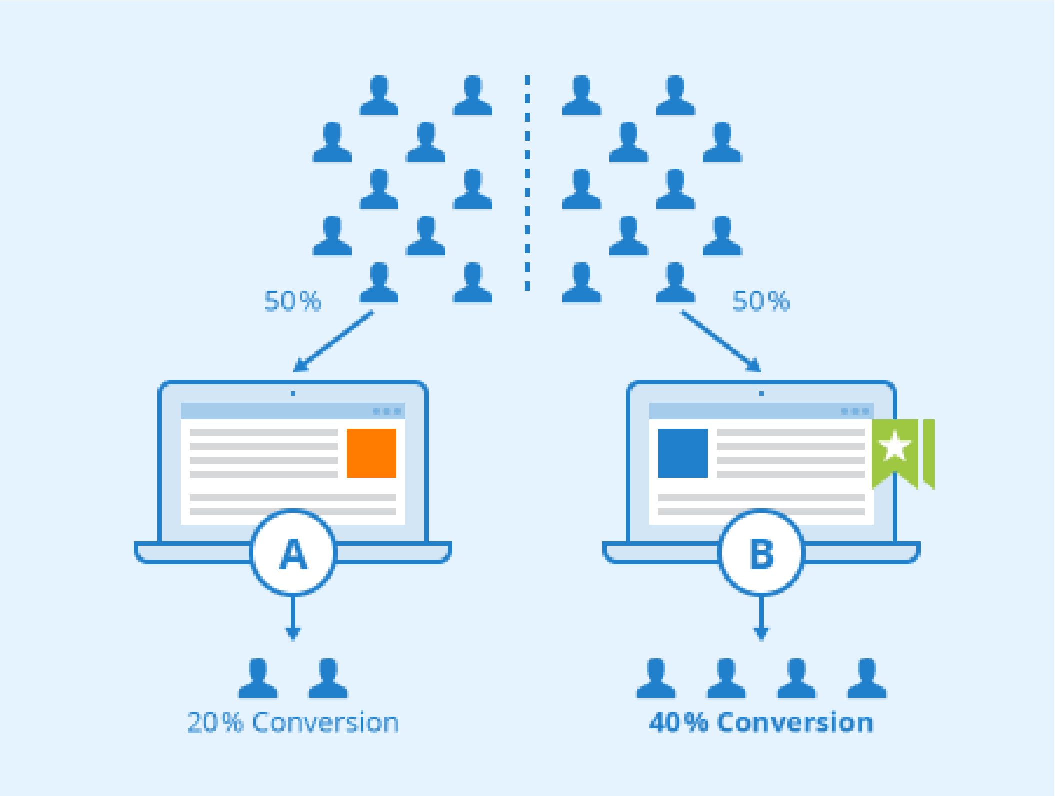


Figure 2. Standardized LISREL Solution

- Start with a causal theory
- Generate causal hypotheses on constructs
- Operationalize constructs → Measurable variables
- Fit statistical model
- Statistical inference → Causal conclusions
- A typical IS journal paper is an excellent example of Explanatory Modeling

But we also see examples of Explanatory Modeling in practice – The case of AB testing in marketing

AB testing methodology



Source: seobility.net

- **A/B tests** (sometimes referred to as **split tests**) **compares two versions of a website/app/interface**
  - 1 base version (untreated)
  - 1 new version including a variation (the treated version)
- The **hypothesis** is that the **induced variation will** be **beneficial** for reaching some predefined business goal
- Using a **randomized controlled trial (RCT)** design, a **certain percentage of users is channeled** to the new „treated“ website and outcomes (such as conversion) is observed over a certain period of time
- **Causal modelling** can **identify** an **unbiased causal treatment effect** of the induced variation, i.e. it allows for testing the hypothesis that the variation is beneficial

Causal inference is at the heart of explanatory analytics

Sources of correlation between two variables

$X \rightarrow Y$	$X \leftarrow Y$	$Z \rightarrow X$ $Z \rightarrow Y$	$X \rightarrow Y$ $Y \rightarrow X$
X causes Y („causality“)	Y causes X („reverse causality“)	Z causes X and Y („common cause“)	X causes Y and Y causes X („simultaneous equations“)

Causal inference - Informal examples of causal expressions

1. “My headache went away because I took an aspirin.”
2. “She got a good job last year because she went to college.”
3. “She has long hair because she is a girl.”

Such causal expressions:

- are often informed by observations on past exposures,
- involve informal statistical analyses, drawing conclusions from associations

Causal inference - Causality is tied to an action applied to a unit

- A **unit** can be a physical object, a firm, an individual, a market etc. at a particular point in time. I.e., the same object or person at a different time is a different unit
- Although a unit was subject to a particular **action** (or treatment), the same unit could be exposed to an **alternative action**. E.g., you could take an aspirin to relieve a headache, or you could not take an aspirin.

Causal inference - Causality is tied to an action applied to a unit

- Every unit-action pair has a **potential outcome**
- If there is one unit and two possible actions (or treatments) there are two possible outcomes
- The causal effect of the action is the difference in potential outcomes
- But we can only observe one possible outcome, for the action actually taken
- The other potential outcome is missing data

Causal inference - Informal examples of causal expressions

1. „My headache went away because I took an aspirin.“

   Action: Taking aspirin.

   Alternative action: Not taking the aspirin

2. „She got a good job last year because she went to college.“

   Action: Going to college

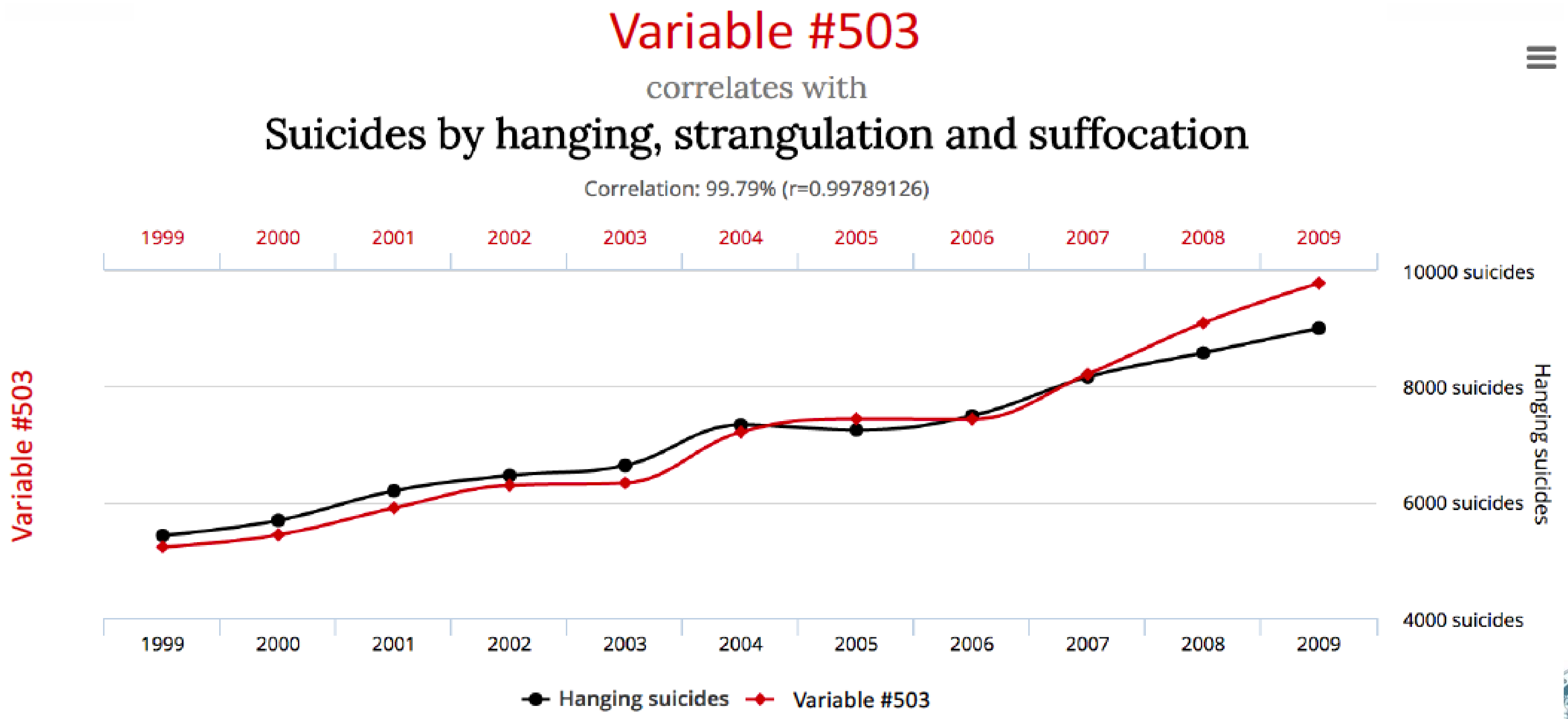
   Alternative action: Doing another job (?), not enrolling in college (?)

3. “She has long hair because she is a girl.“

   Action: Being a girl (?)

   Alternative action: Being a boy (?), what is the action?

The fallacy of correlation vs. causation (1/2)



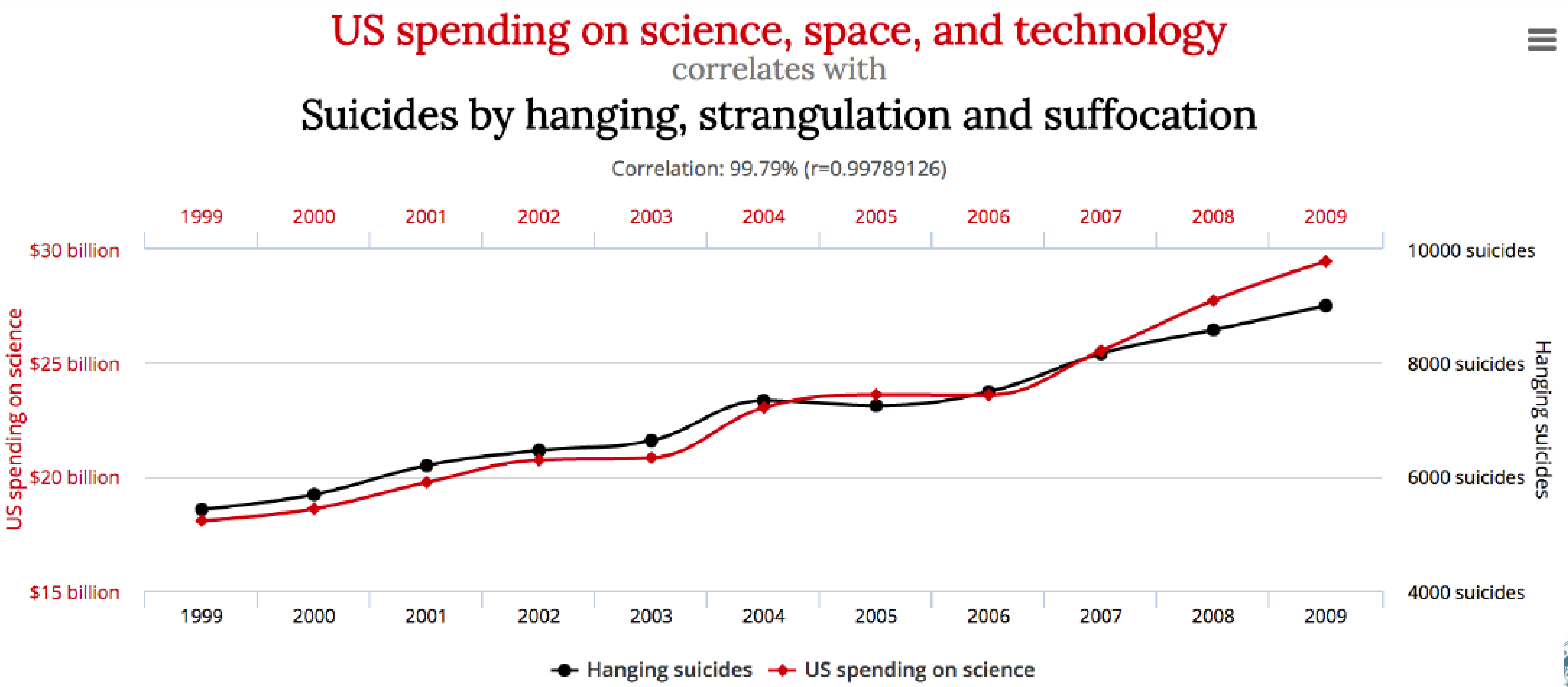
Variable #503

correlates with

Suicides by hanging, strangulation and suffocation

Correlation: 99.79% (r=0.99789126)

The fallacy of correlation vs. causation (2/2)



Descriptive Modeling is what statisticians do – Descriptive analysis is much more cautious about making causal statements

A typical Descriptive Model

Parameters of interest for inference

$$Y_i|X_i = \beta_0 + \boxed{\beta_1} X_{1i} + \cdots + \boxed{\beta_p} X_{pi} + \epsilon_i$$

Chosen because of correlation with Y, only retain if statistically significant

Error term for residual analysis (e.g. heteroscedasticity)

- Definition
  - Statistical model for approximating a distribution or relationship
  - **Descriptive power measured** in terms of **goodness of fit**, generalizable to population (e.g. R^2)
- Scientific Goal
  - Test/quantify distribution or correlation structure for measured "average" unit in population

Source: Galit Shmueli

Predictive Modeling is concerned with predicting new instances for an individual unit based on observable covariates

A typical Predictive Model

Quantity of interest for new instance i (prediction)

$$\boxed{Y_i|X_i} = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

Chosen because of possible correlation with Y; only retain if improves predictive power

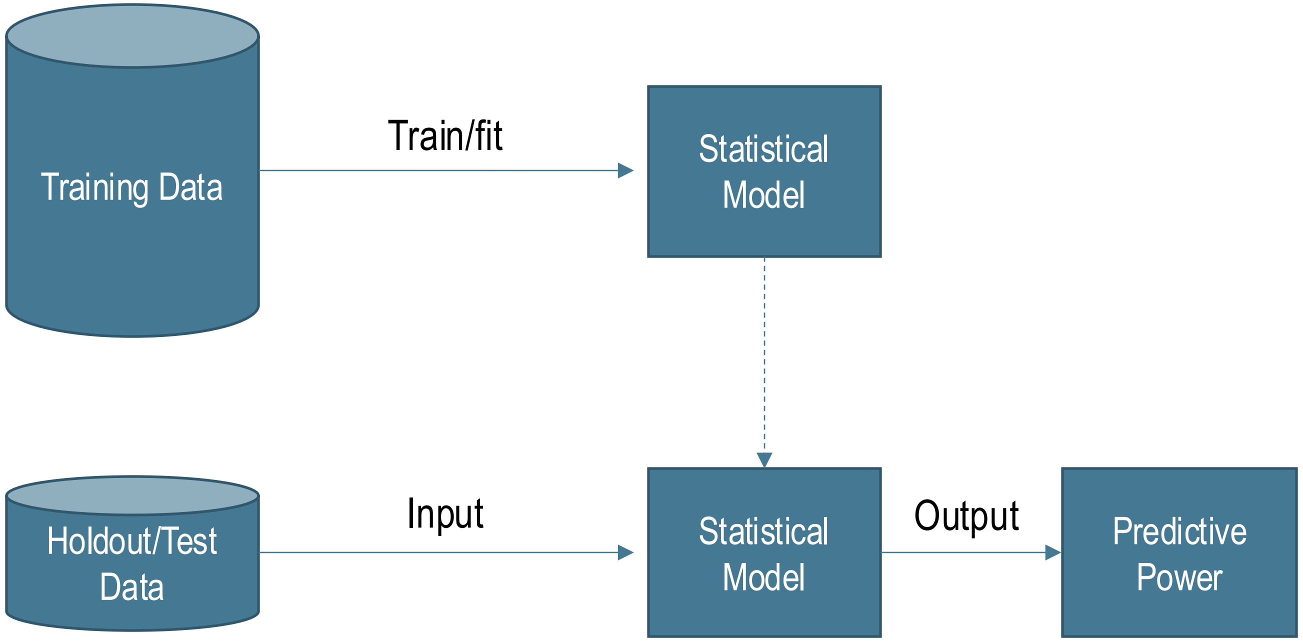
Error term for evaluating generalizability and overfitting on holdout set

- Definition
  - Empirical method for predicting new observations
  - **Explanatory power** is measured in terms of ability to **accurately predict new observations** (e.g. test set performance)
- Scientific Goal
  - Predict values for new/future individual units based on a set of known covariates (also termed features)

Source: Galit Shmueli

Machine Learning has mostly focused on predictive analytics so far, neglecting description and explanations – Predictive modeling always starts with data

Typical Predictive Modeling Procedure



- A predictive model is learned (or trained) on a set of training data
- The same model is used to predict instances of Y of a (usually smaller) holdout set
- Comparing predictions against actual realizations of Y allows for an appraisal of the predictive power of the statistical model

Which of the following statements is correct?

a) In explanatory modeling the goal is to reduce the variance
b) In descriptive modeling we start with a causal theory
c) In descriptive modeling the goal is to quantify the effect on average unit in population
d) In predictive modeling we predict values for average unit in population

Which of the following statements is correct?

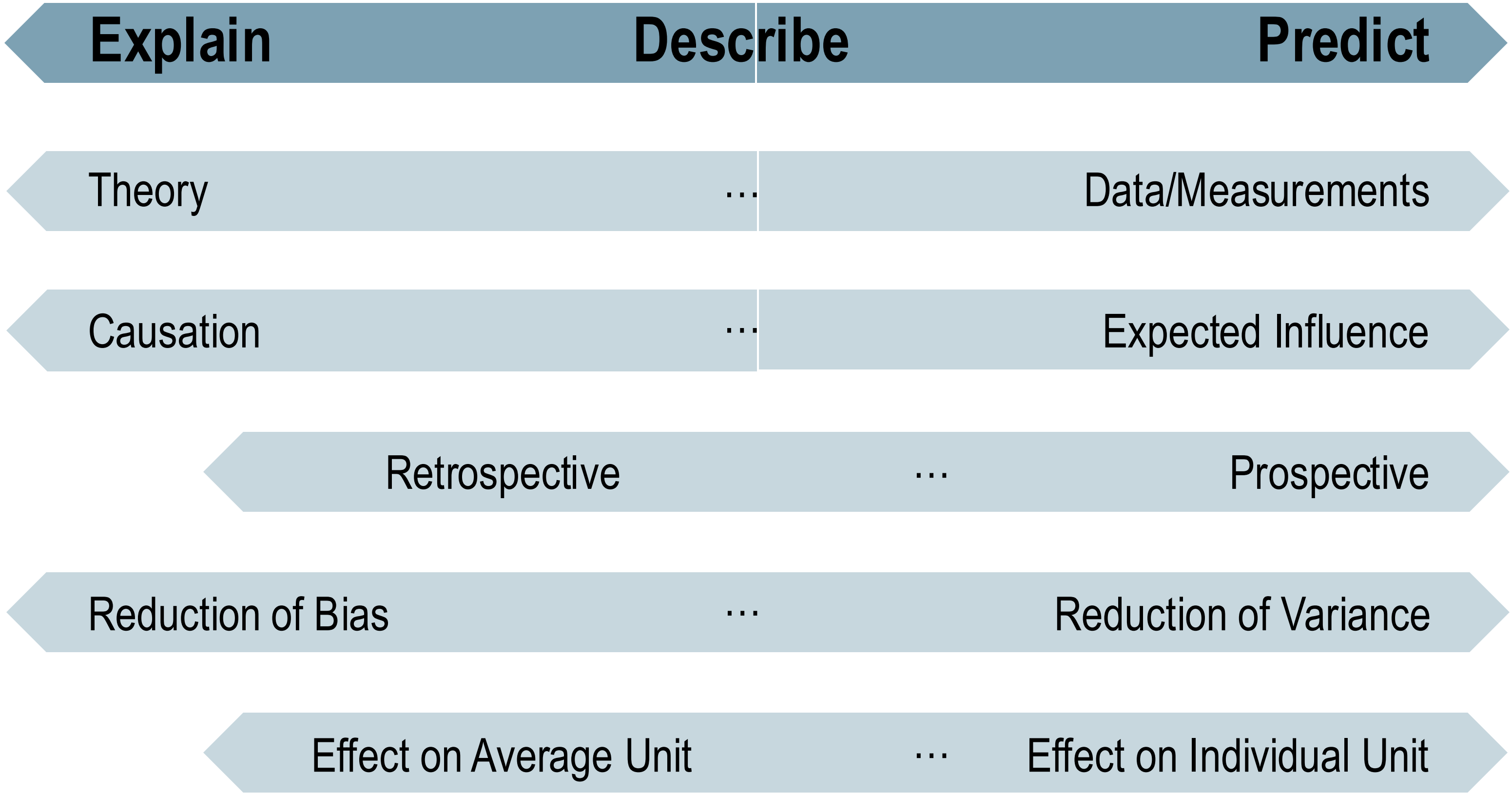
a) In explanatory modeling the goal is to reduce the variance

b) In descriptive modeling we start with a causal theory

c) In descriptive modeling the goal is to quantify the effect on average unit in population

d) In predictive modeling we predict values for average unit in population

In summary explanatory, descriptive and predictive modeling are **distinct across multiple dimensions** – An overview



Dimension	Explain	Describe	Predict
Starting Point	Theory	…	Data/Measurements
Parameter Interpretation	Causation	…	Expected Influence
Temporal Viewpoint	Retrospective	…	Prospective
Goal	Reduction of Bias	…	Reduction of Variance
Outcomes	Effect on Average Unit	…	Effect on Individual Unit

Source: Galit Shmueli

There are two common misconceptions which we need to clear up (1/2)

#1

“The best explanatory model is also the best descriptive/predictive model and vice versa”

- **Social Sciences** & Management often **build explanatory models** and **use them to predict**
- **Engineering & CS** build **predictive models** and **use them to explain**
- **Both approaches** are **equally flawed** as both modelling approaches set out to satisfy different objectives
- While **some features** may be **good for descriptive or explanatory** modeling they may be **useless for improving predictive power** and vice versa

Predict ≠ Explain

Netflix Price

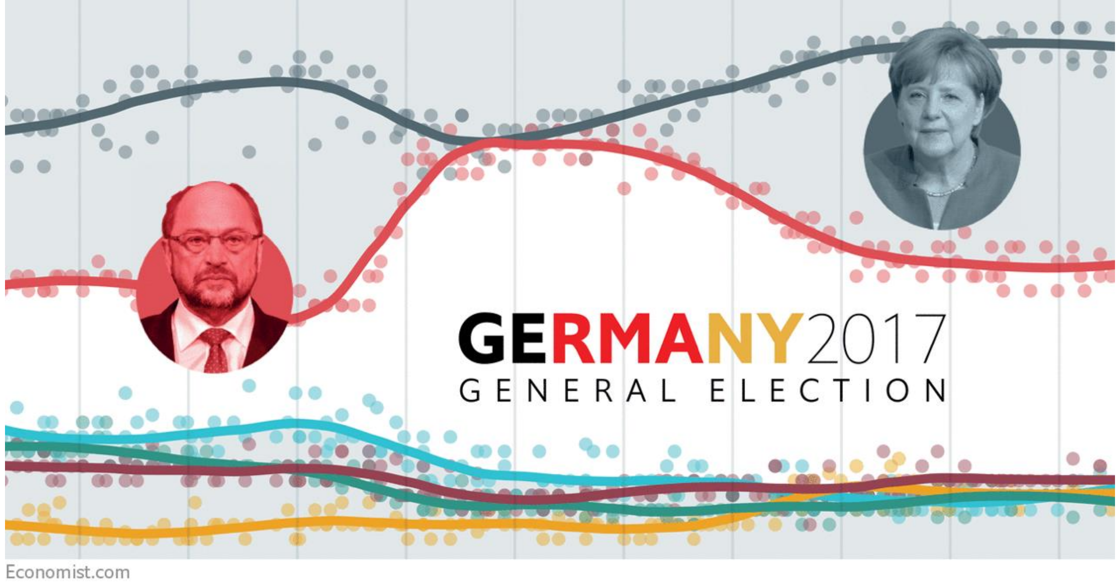
```
STRUCT GROUP_INFO INIT_GROUPS = { .USAGE = ATOMIC_INIT(2) };
STRUCT GROUP_INFO *GROUPS_ALLOC(INT GIDSETSIZE){
    STRUCT GROUP_INFO *GROUP_INFO;
    INT NBLOCKS;
    INT I;

    NBLOCKS = (GIDSETSIZE + NGROUPS_PER_BLOCK - 1) / NGROUPS_PER_BLOCK;
    /* MAKE SURE WE ALWAYS ALLOCATE AT LEAST ONE INDIRECT BLOCK POINTER */
    NBLOCKS = NBLOCKS ? : 1;
    GROUP_INFO = KMALLOC(SIZEOF(*GROUP_INFO) + NBLOCKS*SIZEOF(GID_T *), GFP_USER);
    IF (!GROUP_INFO)
        RETURN NULL;
    GROUP_INFO->NGROUPS = GIDSETSIZE;
    GROUP_INFO->NBLOCKS = NBLOCKS;
```

- *“We tried to benefit from an extensive set of attributes describing each of the movies in the dataset. Those **attributes** certainly carry a significant signal and can **explain some of the user behavior**. **However**… they **could not help** at all for **improving** the [predictive] **accuracy**.”*

Bell et al., 2008

Predict ≠ Describe

Election Polls



- *“There is a **subtle, but important, difference** between **reflecting current public sentiment** and **predicting the results of an election**. Surveys have focused largely on the former… [as opposed to] survey based prediction models [that are] focused entirely on analysis and projection”*

Kenett, Pfefferman & Steinberg, 2017

Source: The Economist

There are two common misconceptions which we need to clear up (2/2)

#2

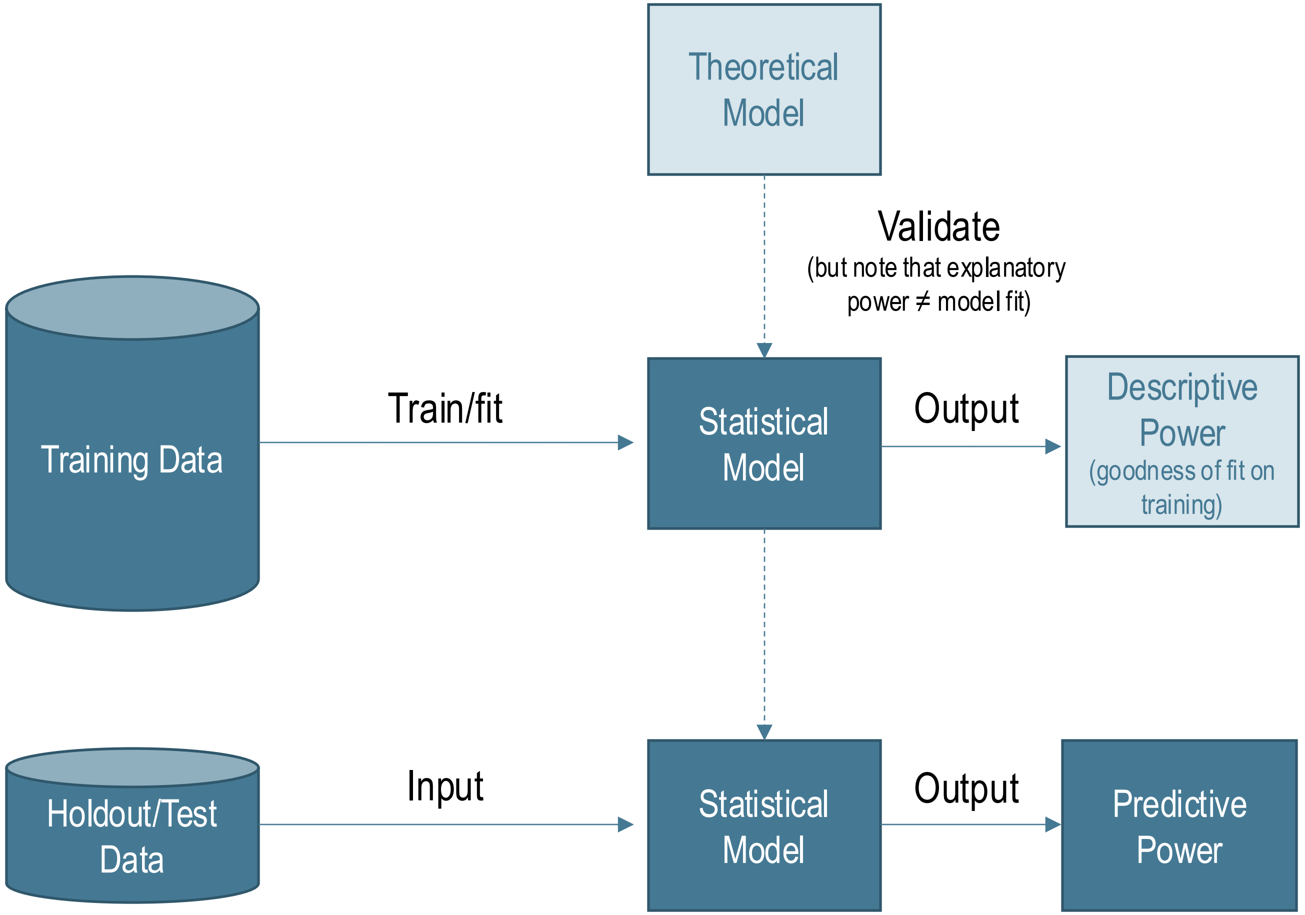
“Explanatory modeling is more relevant than descriptive/predictive modeling or vice versa”

- **All** three have their **purpose and rightful place**, hence a **ranking** is **not helpful**
- They **cannot** actually **be compared** as they set out **doing very different things** and optimize for different goal:
  - **Explanation**: **Test theory** using statistical data
  - **Description**: **Test covariates** for significance
  - **Prediction**: Optimize **predictive power** on unseen data
- In many data science cases you will have to use all three!

Universität zu Köln

Description, Explanation and Prediction all have their use cases and **you may have to draw on all three** in your data science project!

Typical Predictive Modeling Procedure



- A predictive model is learned (or trained) on a set of training data
- The same model is used to predict instances of Y of a (usually smaller) holdout set
- Comparing predictions against actual realizations of Y allows for an appraisal of the predictive power of the statistical model

What are the key factors of interest of a descriptive model?

a) Dependent variable y
b) All parameters
c) All parameters that are significant
d) Parameters of the covariate of interest

What are the key factors of interest of a descriptive model?

a) Dependent variable y
b) All parameters
c) All parameters that are significant
d) Parameters of the covariate of interest

Is the best descriptive model also the best predictive model, and why?

a) Yes, as it provides the best statistical fit for the available data
b) Yes, as only the most significant covariates are used in a descriptive model
c) No, good fit does not ensure best performance on test set

Is the best descriptive model also the best predictive model, and why?

a) Yes, as it provides the best statistical fit for the available data

b) Yes, as only the most significant covariates are used in a descriptive model

c) No, good fit on past / training data (descriptive) does not ensure best performance on test set (predictive)

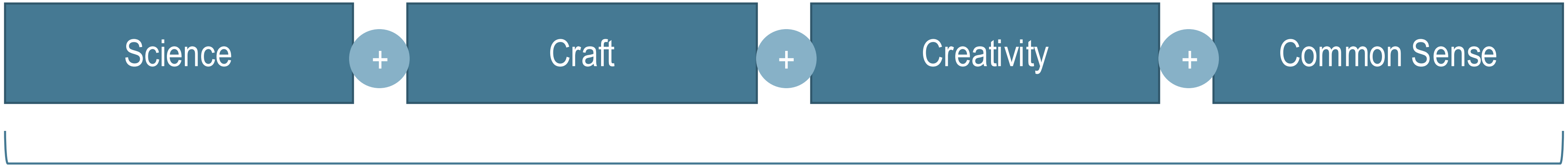
Agenda

To Explain or To Predict?

CRISP - Data Science as a Process

Common Data Mining Tasks and Terminology

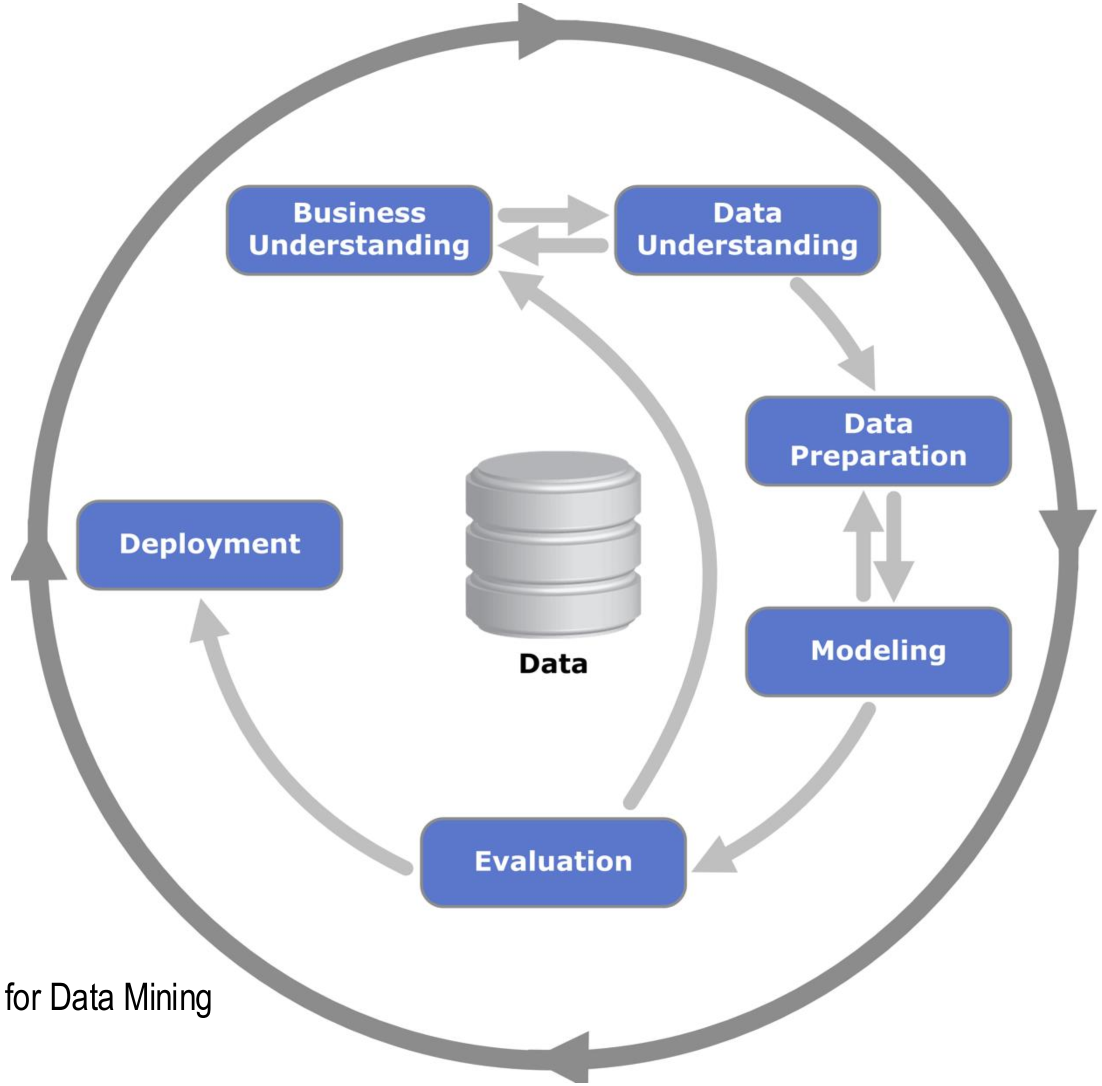
Business data science is a process that combines different core attributes



Science + Craft + Creativity + Common Sense

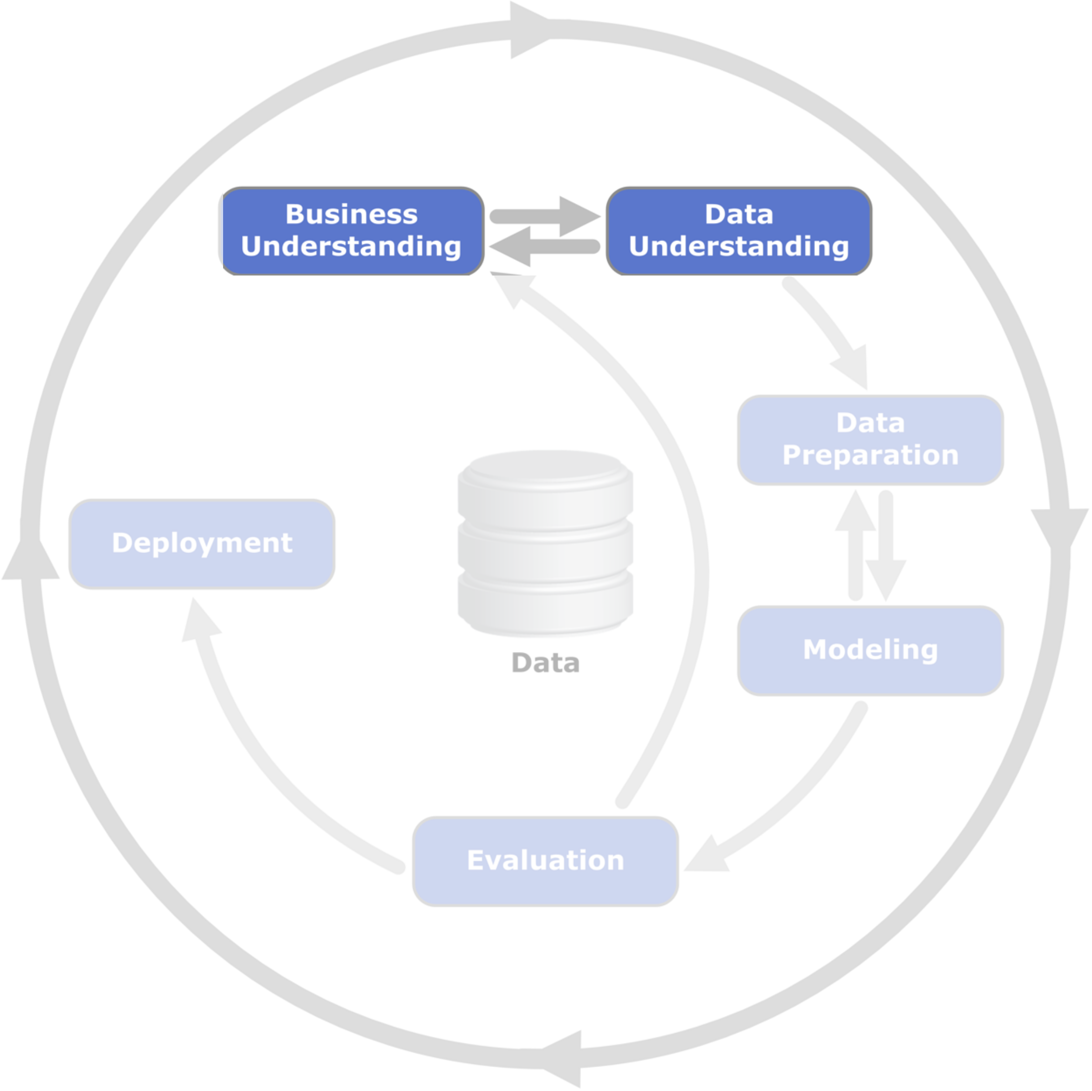
Data Science as a process

The CRISP* Data Science Process is a common way of describing this process



*CRISP – Cross Industry Standard Process for Data Mining

Step 1 & 2: Business and Data Understanding



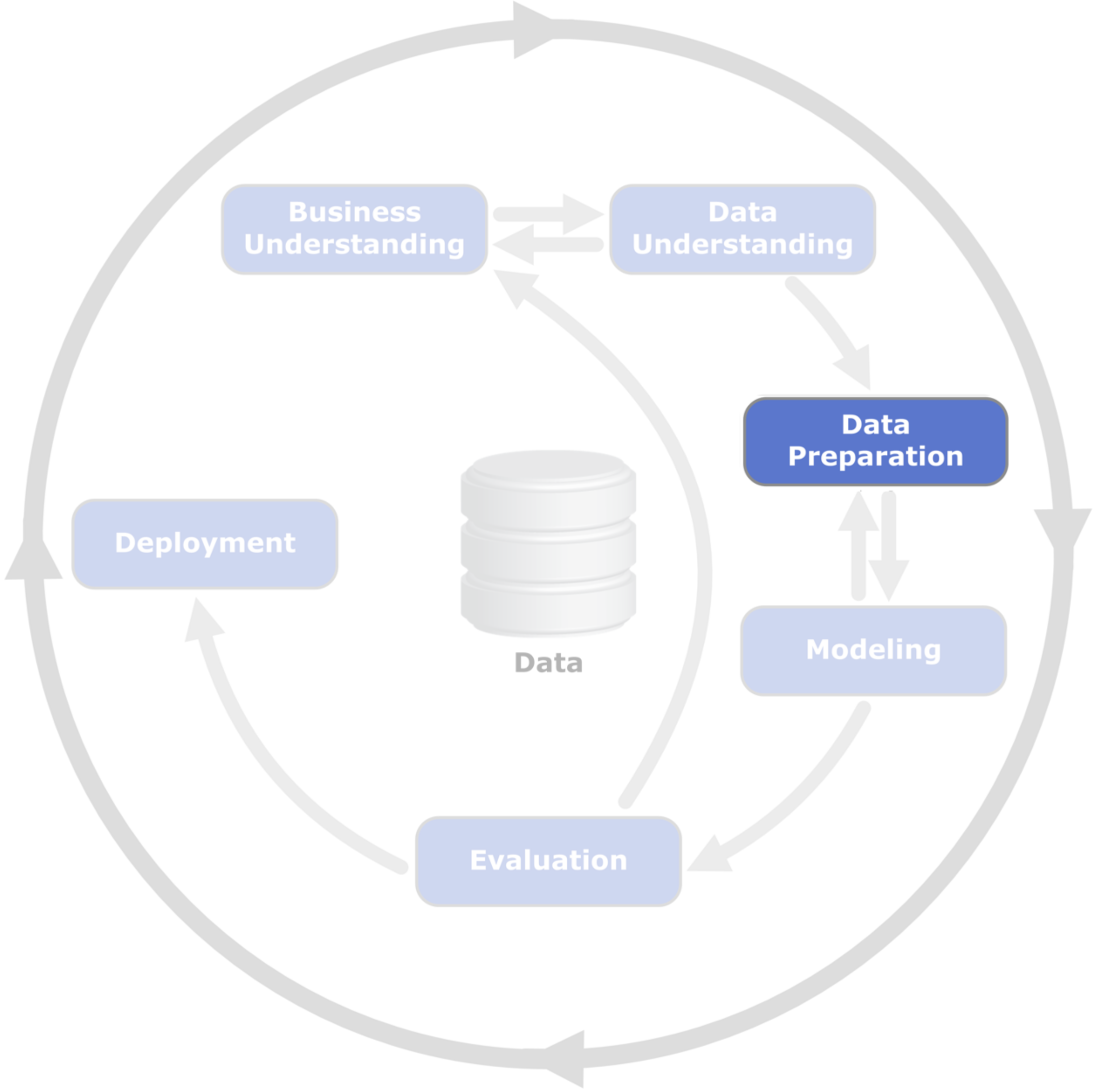
Every project begins with **business understanding**.

- What are **project objectives**?; How do you define "success" and how can you measure it?
- Do we fully **understand the domain** we are operating in?

From the business understanding **data understanding** is informed and vice versa

- Which **analytics approach** should be employed (regression, classification, etc.)?
- For this approach, what are **data requirements** and how can **data collection** be organized?
- Descriptive statistics and visualization combined with business understanding facilitate **data understanding**

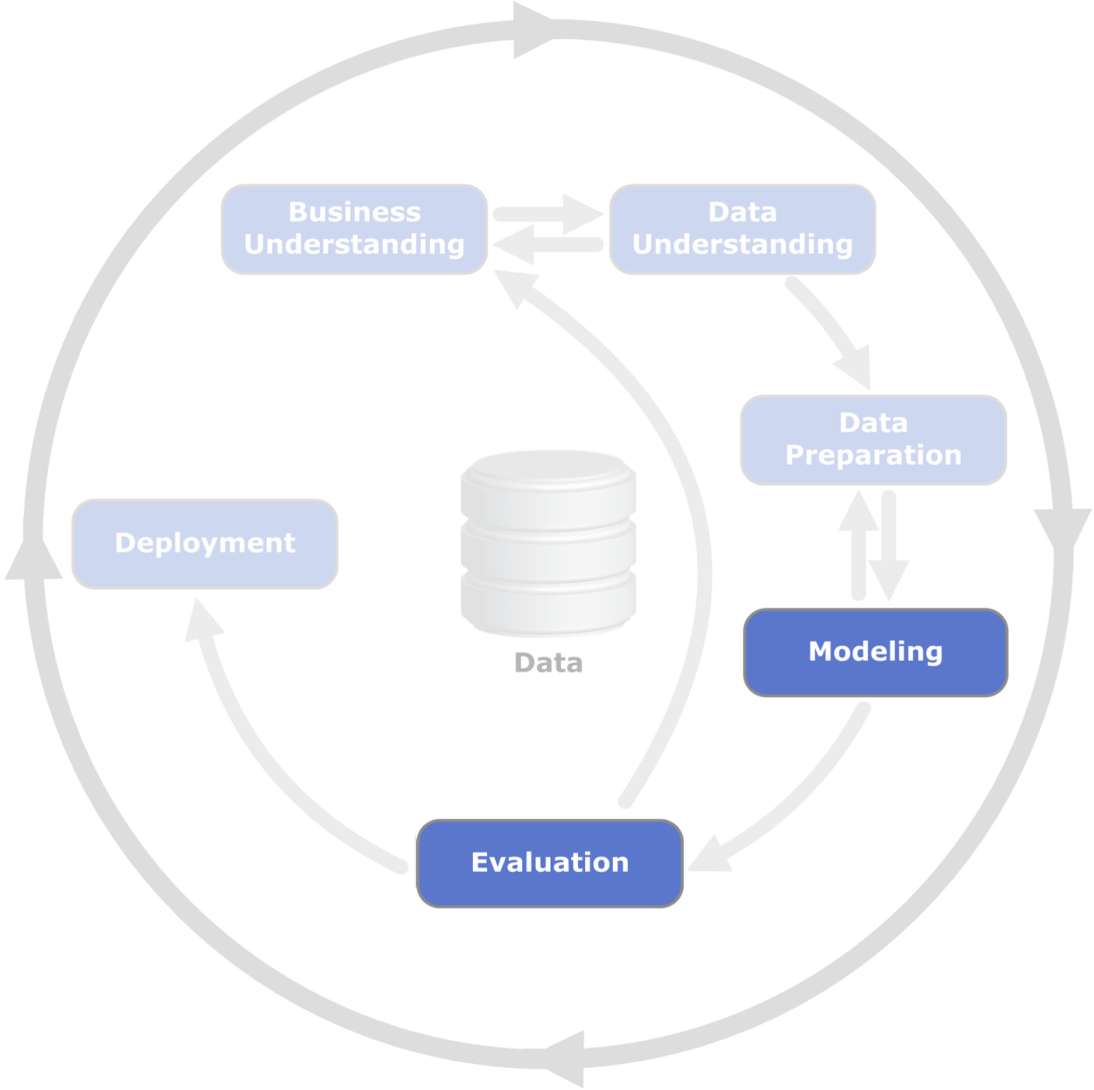
Step 3: Data Preparation



Data preparation encompasses all activities to construct and clean the data set.

- Data cleaning and preparation routines include, e.g.
  - Missing or invalid values elimination or imputation
  - Eliminating duplicate rows
  - Aligning formatting
  - Combining multiple data sources
  - Transforming and normalizing data (e.g. categorical to encoded features)
  - Engineering new features (e.g. via NLP, etc.)
- „Arguably the most time-consuming step of the entire DS process is data cleaning and preparation„
- Accelerate data preparation by automating common steps

Step 4 & 5: Modeling and Evaluation



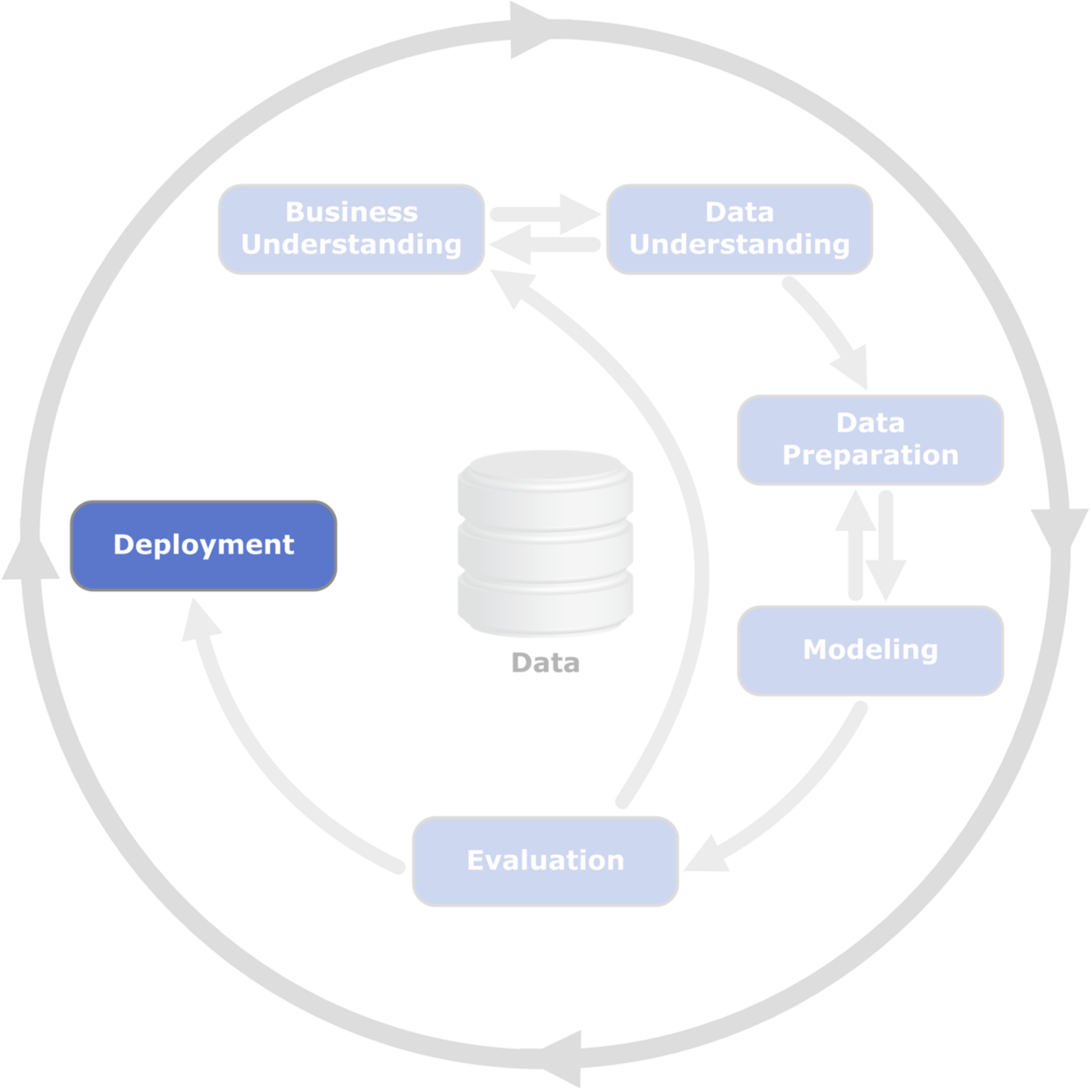
Modeling builds on the prepared dataset

- Developing predictive or descriptive models
- Modeling is often a highly iterative process in which different features and models are tried

Model **evaluation** is performed during model development and before model deployment

- Assess the model’s quality and it‘s performance in the real world – How reliable is it?
- Use statistical tests and common test metrics (R^2, RMSE, etc.) to compare model performance
- Ensure that the model properly addresses the business problem
- Refine model as needed

Step 6: Deployment



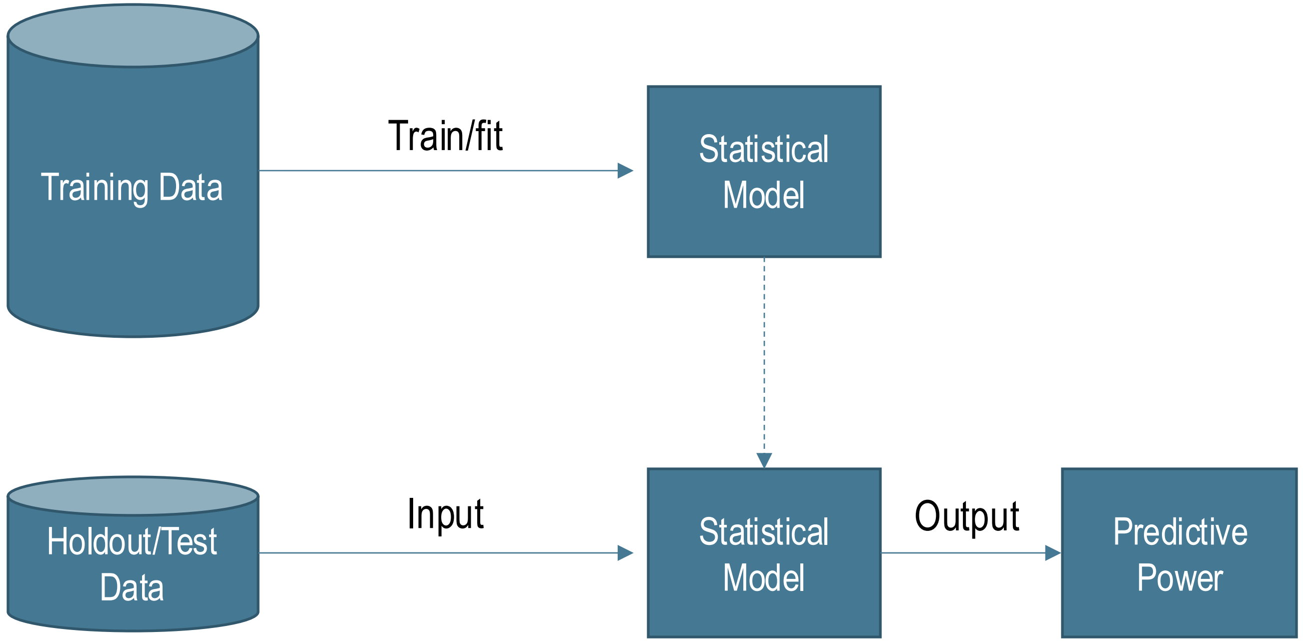
Once finalized, the model is **deployed** into a production environment.

- It is advisable to start the roll-out in a secure test environment first
- Key stakeholder roles must be involved throughout the roll-out process. These may include:
  - Solution owner
  - Marketing
  - Application developers
  - IT administration
- Continuously monitor and appraise model performance in the real world**:**
  - How well did the model perform?
  - If required, refine model and re-deploy

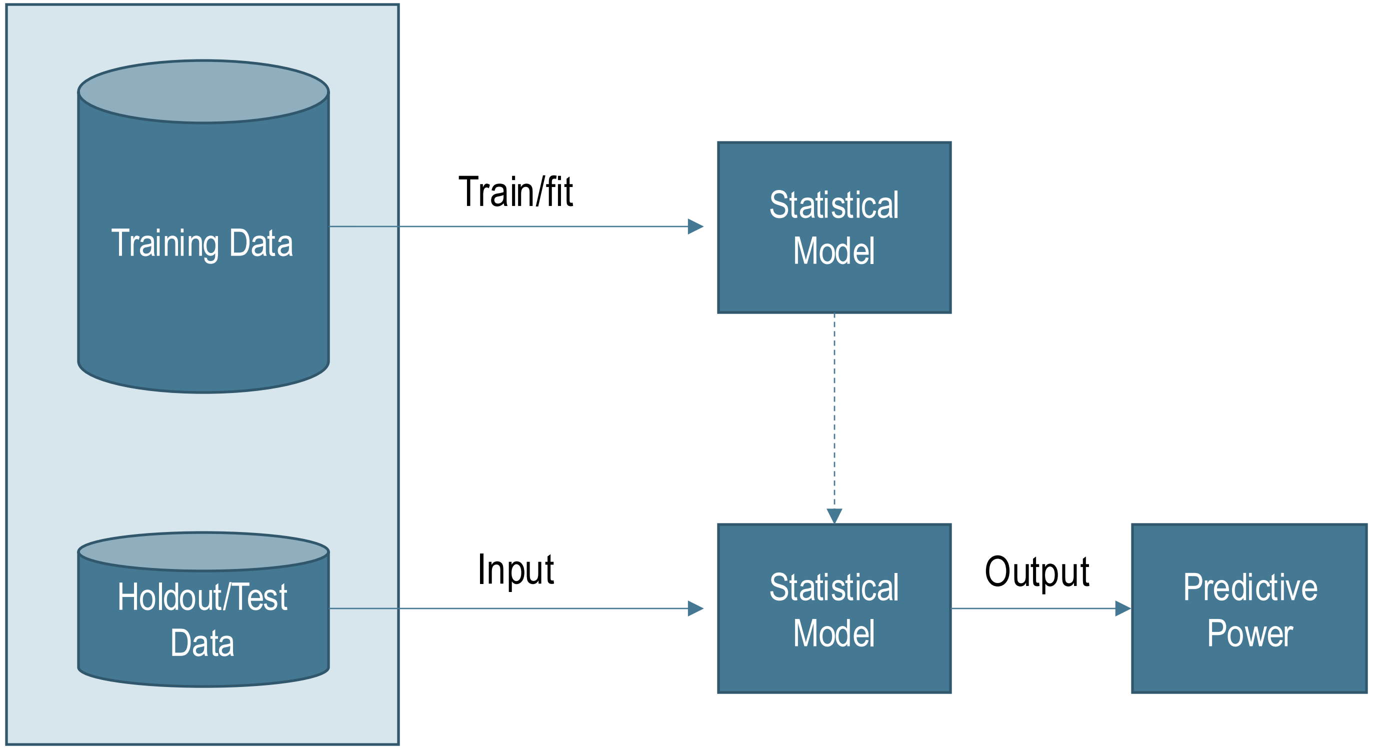
Agenda

- To Explain or To Predict?
- CRISP - Data Science as a Process
- Common Data Science Tasks and Terminology

Let‘s go back to our very general predictive modeling procedure and specify some nomenclature



Let‘s go back to our very general predictive modeling procedure and specify some nomenclature



A simple example: Peak Electrical Power

- One of the challenges in the electricity system is satisfying electricity demand at all times – Especially also during peak times
- Suppose **you want to predict** what **tomorrow's peak electricity demand** will be during the day for some area
- This is actually a very **important problem from a planning perspective**: electricity generators, which for the most part are based on boiling water to move turbines (for now!), cannot turn on instantly, so in order to guarantee that we have enough power to supply a given area, a system operator typically needs to have some excess generation always waiting in the wings.
- The **better we can forecast** future demand, the **smaller our excess stand-by capacity** can be, leading to **increased efficiency** of the entire electrical grid.

Peak Electrical Power

- What data do you need use?; How would they be used?
- The power consumption tomorrow depends on many factors:
  - temperature
  - day of week
  - season
  - holiday events, etc.
- Not to mention some inherent randomness that we don't expect to even predict with perfect accuracy. However, even for someone working in the area, it would be very difficult to come up with a model for electrical demand based solely upon "first principles", thinking about the nature of electricity consumption or the devices people may use, in an attempt to predict future consumption.

What will peak power consumption be in Pittsburgh tomorrow?

Data Terminology (1/2)

Features/Predictors

Date	Average demand	Peak demand	High temperature	Average temperature
01.01.2013	1.598524	1.859947	0	-1.68
02.01.2013	1.809347	2.054215	-3.9	-6.58
03.01.2013	1.832822	2.04955	0.6	-6.12
04.01.2013	1.812699	2.008168	0	-1.95

Target/Outcome/Response

- Covariates (i.e. the **independent variables**) are commonly referred to as **"features"**)
- The **dependent variable** Y is referred to as the **"target"** (only available for supervised tasks, more on that later)
- One row represents an **instance/example/ observation/sample** (all synonyms)

Data Terminology (2/2)

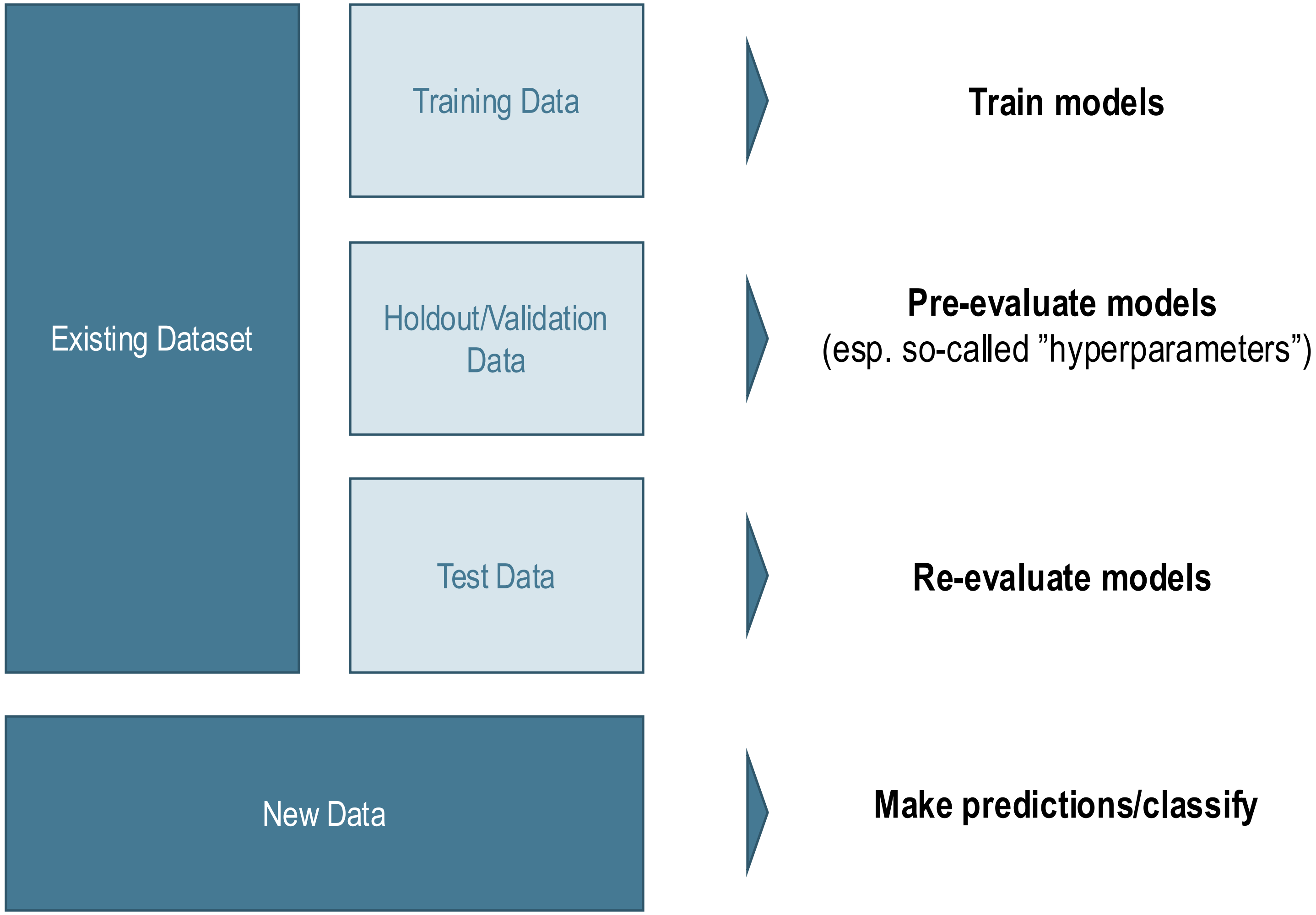
Features/Predictors

Date	Average demand	Peak demand	High temperature	Average temperature
01.01.2013	1.598524	1.859947	0	-1.68
02.01.2013	1.809347	2.054215	-3.9	-6.58
03.01.2013	1.832822	2.04955	0.6	-6.12
04.01.2013	1.812699	2.008168	0	-1.95

Target/Outcome/Response

- **Dimensionality** of a dataset is the sum of the feature dimensions, i.e. the **sum of the number of numeric features** and the **number of values of categorical features**
  - **Numeric**: can take any continuous value
  - **Categorical**: can take values from a pre-defined set only (e.g. gender)

In machine learning it is common to split a dataset into multiple parts – For now it is important to know the following



Existing Dataset

Training Data → **Train models**

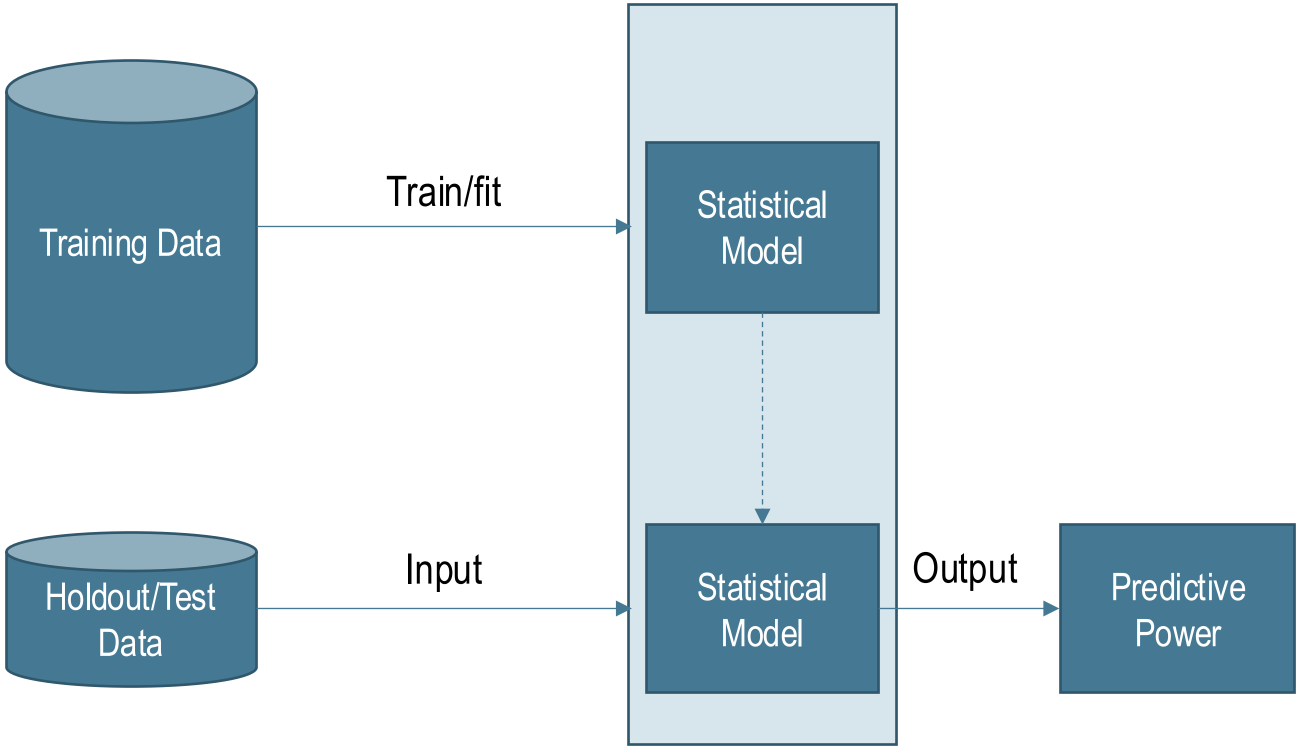
Holdout/Validation Data → **Pre-evaluate models** (esp. so-called ”hyperparameters”)

Test Data → **Re-evaluate models**

New Data → **Make predictions/classify**

- The **existing data** is typically **divided into a various subsets** on which the model is learned and evaluated (more on this later)
- The **model can then be used** to **make predictions** for **new data** instances

Let‘s go back to our very general predictive modeling procedure and specify some nomenclature



What is a model?

"A **simplified representation** of **reality** created for a specific purpose – based on some assumptions"

Examples

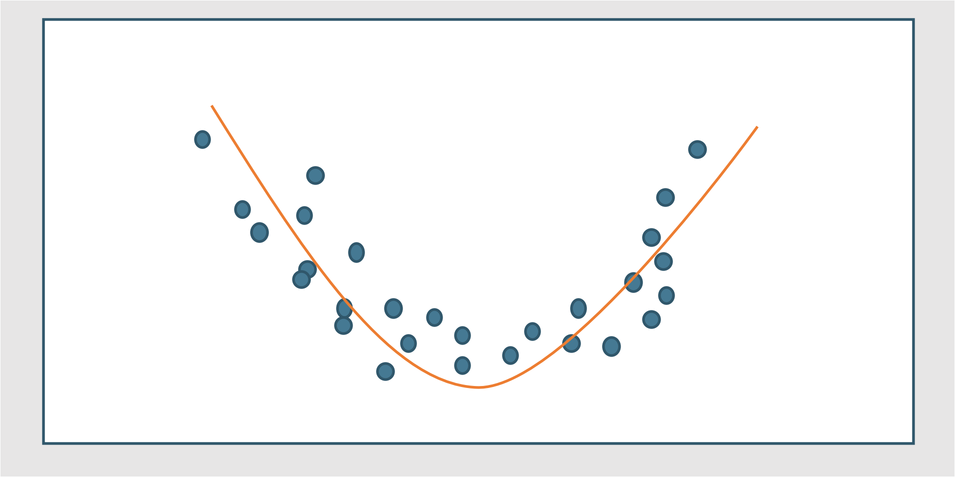
- Geographical map,
- Prototype of a car
- Power TAC, etc.
- "Formula" for predicting probability of customer attrition at contract expiration

Some model-related Terminology

- **Algorithm:**
  - A procedure used to implement a particular data science task (classification tree, linear regression, etc.)
  - A model in a data science context is an algorithm applied to a specific problem
- **Predictive Model:**
  - A formula for estimating the unknown value of interest: the target
  - The formula can be mathematical, logical statement (e.g., rule), etc.
- **Prediction:**
  - Estimate an unknown value (i.e. the target)

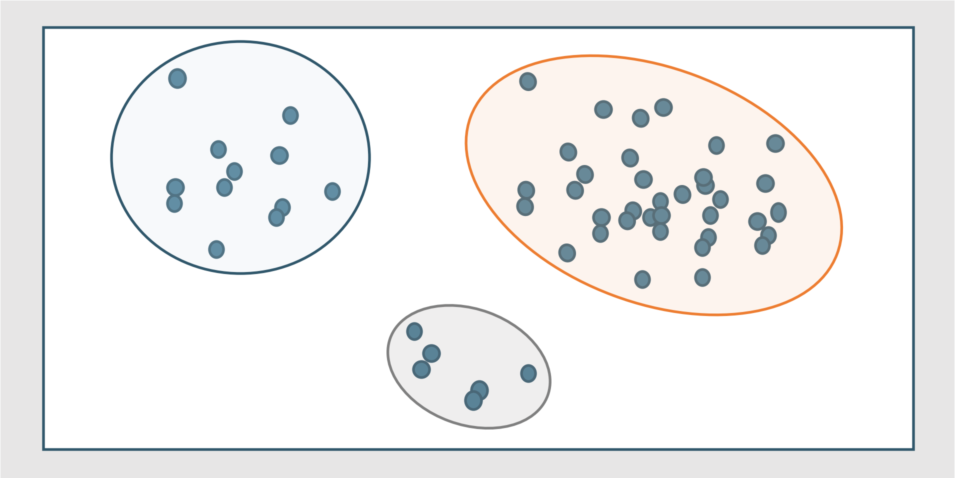
For now we will differentiate between two fundamental Machine Learning modeling techniques

Supervised Learning



- Availability of **labeled** data
- Goal to learn a model that describes the relationship of **input features** and label
- Differentiation between **regression** (i.e. typically continuous targets) and **classification**
- Model performance **relatively easy** to evaluate

Unsupervised Learning



- Data **without** labels
- Goal to find certain structural **patterns** within the data
- Find **clusters** in data with similar characteristics
- Model performance **hard** to evaluate

What general modeling technique is underlying the Energy Demand Example?

a) Supervised Learning
b) Unsupervised Learning

What general modeling technique is underlying the Energy Demand Example?

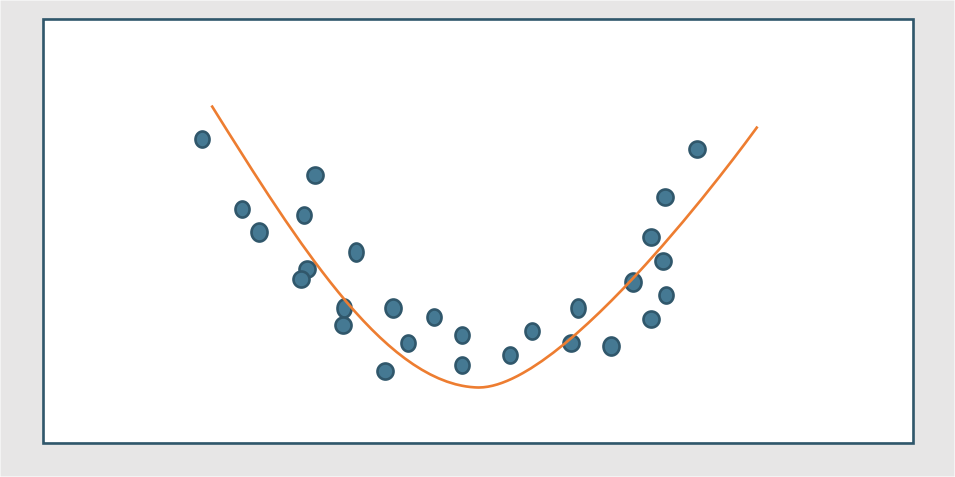
a) **Supervised Learning**
b) Unsupervised Learning

Some more examples – What general modeling technique is underlying the different data science tasks?

	Data Science Task	Supervised	Unsupervised
1)	How likely is a consumer to respond to our marketing campaign?	✓	
3)	Do my customers form natural groups?		✓
4)	How much will the customer tip the Uber driver?	✓	
5)	What items are commonly purchased together?		✓
6)	What does “normal behavior” look like? (e.g., as a baseline to detect fraud)	✓	(✓)

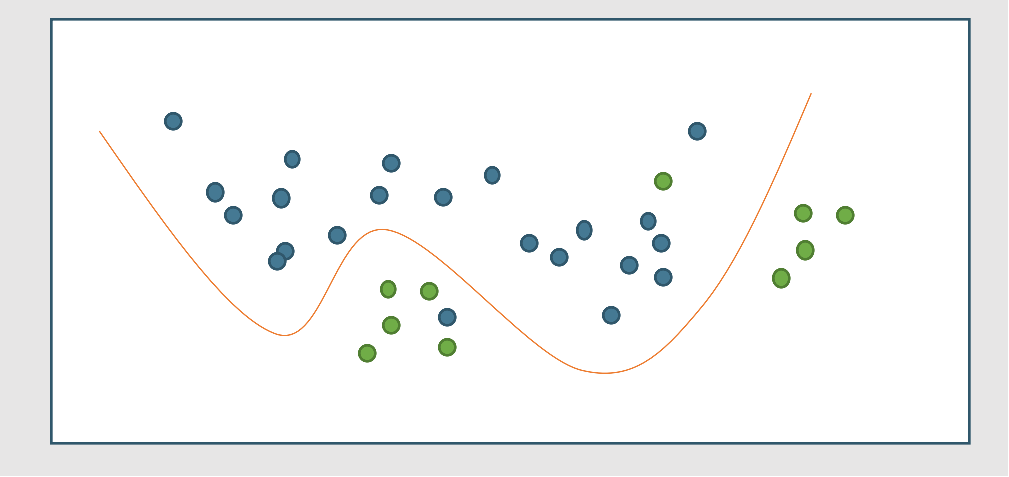
Within supervised learning we further differentiate between classification (class prediction) and regression (continuous value prediction) – More on this later

Regression



- Predict **continuous target value** based on input features
- **Linear and non-linear relationship** between input and target possible

Classification

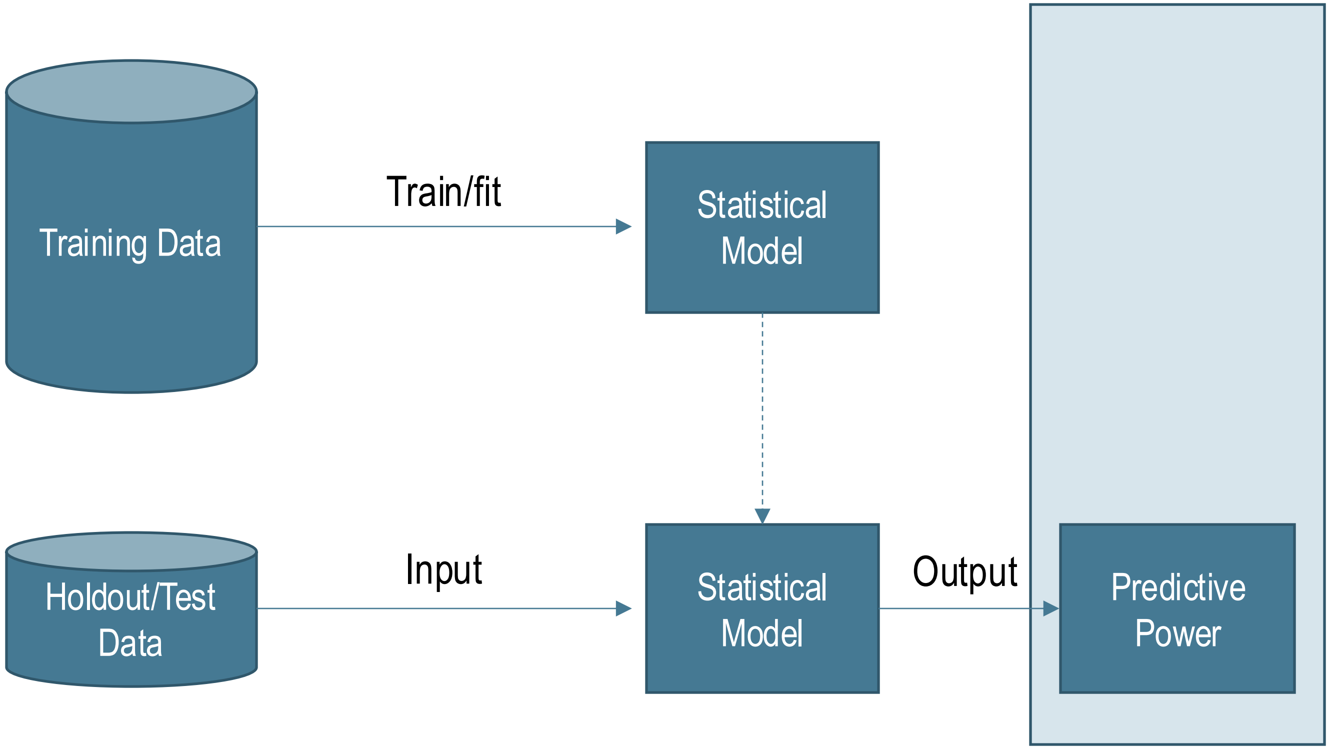


- Predict **discrete target class** based on input features
- Borders between classes can be linear and non-linear

Regression or Classification?

	Data Science Task	Regression	Classification
1)	Predict peak power consumption next week?	✓	
2)	Predict whether a cancer cell is benign or malignant?		✓
3)	Predict demand for carsharing in the city centers?	✓	
4)	Predict if customers will buy a certain product?		✓
5)	Predict the destination of a carsharing trip?		✓
6)	Make a forecast for tomorrow's sales volumes of toilet paper?	✓	

Let‘s go back to our very general predictive modeling procedure and specify some nomenclature



How to measure performance in a **regression setting**?

“**How far off** are our predictions **compared to** the **true realizations** as observed in the real world?”

Typical distance measures

- Root-mean-squared-error (RMSE)
- Mean absolute error (MAE)
- Mean absolute percentage error (MAPE)

How to measure performance in a **classification setting**?

"**How accurate** are our **classifications**, i.e. what is the **ratio** of **correctly classified** vs. **incorrectly classified** instances as compared to the true realizations?"

Typical accuracy measures

- Precision
- Recall
- Confusion Matrix
- …

Contact

For general questions and enquiries on **research**, **teaching**, **job openings** and new **projects** refer to our website at www.is3.uni-koeln.de

For specific enquiries regarding this course contact us by sending an email to the **IS3 teaching** address at is3-teaching@wiso.uni-koeln.de

To help us process your request efficiently, use the following subject line format:

[AA] <Your request subject>