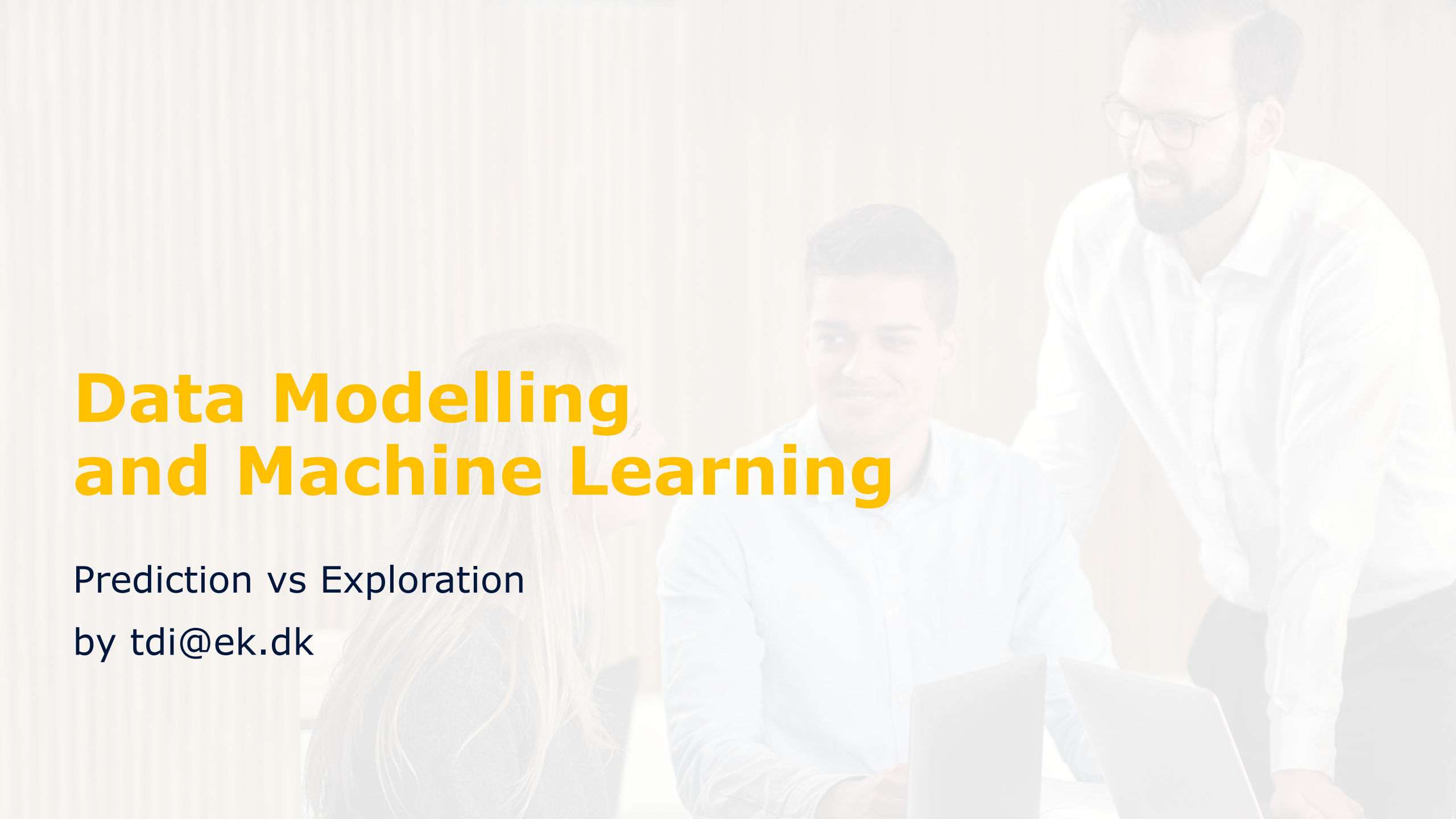


# Data Modelling and Machine Learning

Prediction vs Exploration

by tdi@ek.dk


Define Business Requirement
Data Collection
Data Cleaning
Data Exploration & Analysis
Data Modelling
Data Validation
Deployment & Optimization

# Intended Learning Outcomes

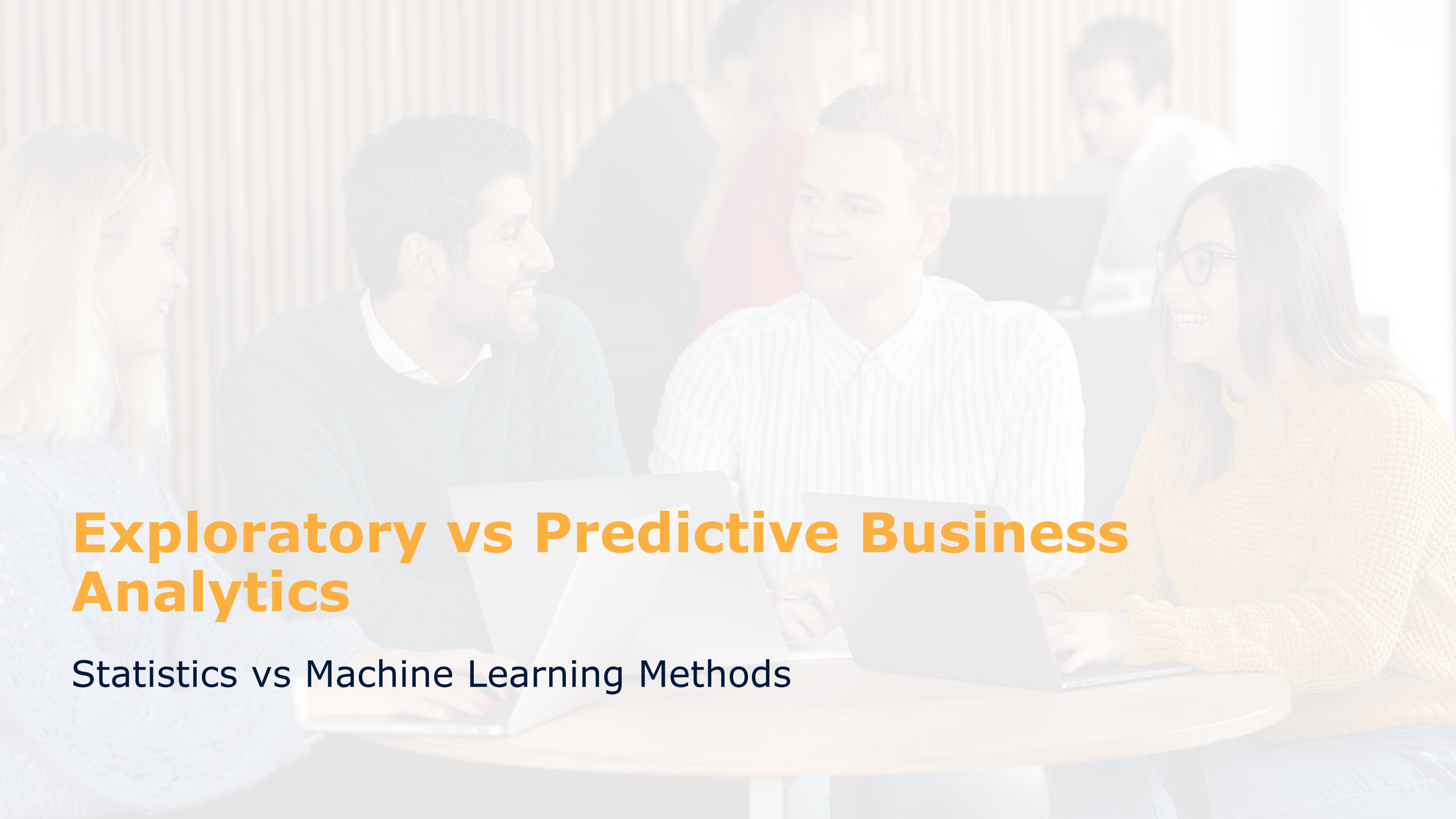
- To get acquainted with Machine Learning as BI engine
  - to get familiar with the terminology
  - to understand the machine learning tasks, processes and implementations
- To be able to make difference between exploratory and predictive analytics

tdi@ek.dk

# Agenda

What did we learn so far?

- Prediction vs Exploration
- Introduction to Machine Learning - tasks, processes and implementations
- Implementation of regression in machine learning



# Exploratory vs Predictive Business Analytics

Statistics vs Machine Learning Methods

# Exploratory and Predictive Analytics

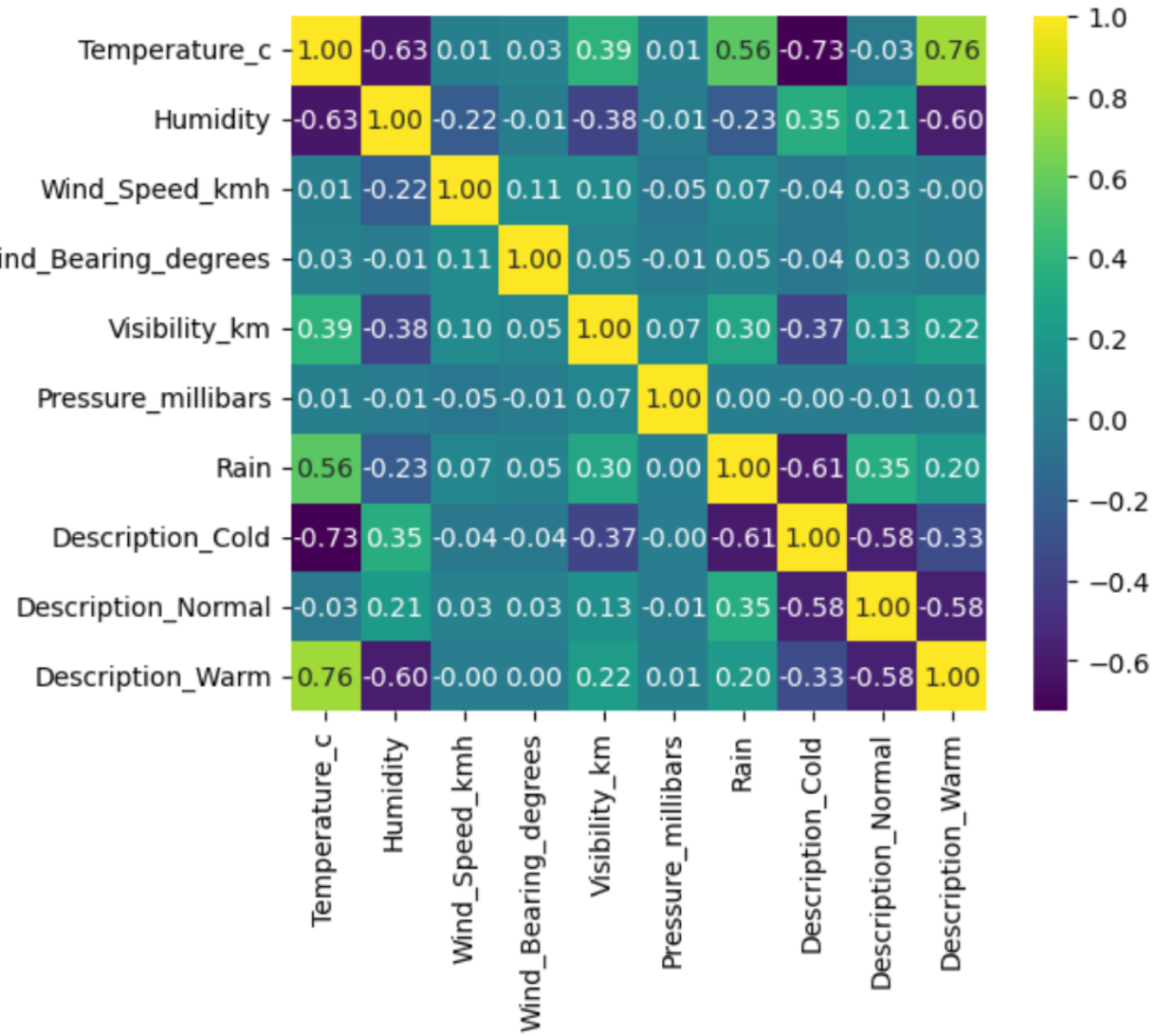
- Exploratory Analysis
  - trying to explain what had happened by processing the collected data
  - Examples
    - dependent and independent variables
    - correlation
- Predictive Analytics
  - trying to create models of future behaviour
  - Example
    - how the correlation influences future observations

# Correlation

- Covariance determines the relationships of two data sets – the differences in their variance

$$\sigma(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Correlation is normalised covariance – a coefficient between -1 and +1

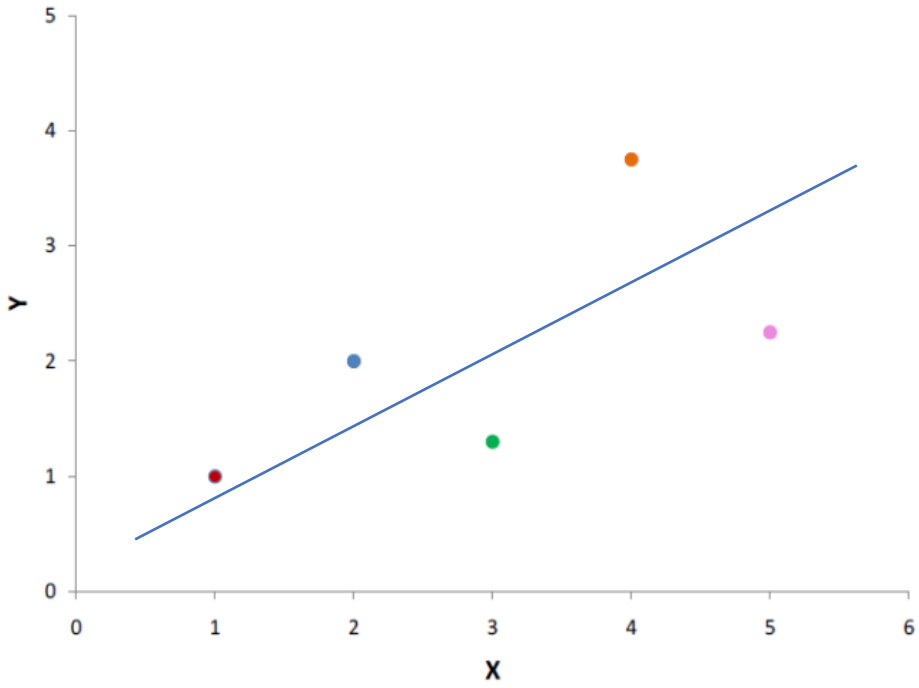


# Example of Two Variables

Can you calculate **r** for **X** and **Y** sets?

Can you predict Y if X=6?

| X | Y |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 1.3 |
| 4 | 3.75 |
| 5 | 2.25 |




scatter plot

Online Stats, pages 462-464

# Machine Learning

- AI technology dealing with predicting the future based on the past, *for instance*,
  - predict the weather at specific location based on historical data
  - predict how much a person will like a movie that she hasn't seen, based on her ratings of movies that she has seen in the past
  - predict the future sells and prices of a product

- Making informed guesses about some unobserved property of an object, based on observed properties of that object

- Called machine learning, because the computer program 'learns' from the available observed data and later implements the learning outcomes for prediction even in a new, unobserved situation

# Machine Learning Foundation

- The idea is that the unobserved feature - the *Output* (dependant variable), is an unknown function *F()* of the observed features - the *Input* (independent variables)

$$Output = F(Input)$$

- The goal is to reveal this function and later use it for predictions

Input => Features

Output <= Label

- ML algorithms are used to extract patterns from data for the purpose of empowering the computers to predict and draw inferences

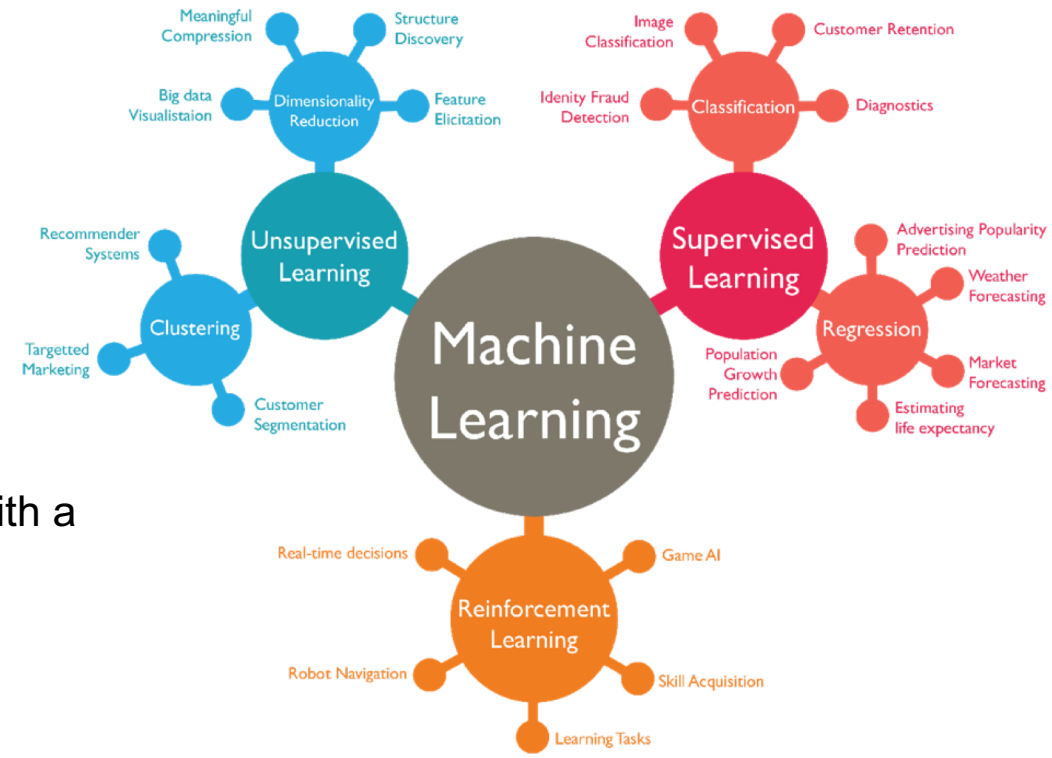
# Model

- What is model?
  - realistic and relatively precise representation of real object or event
- Types of models
  - business model
  - physical model
  - mathematical model
- In machine learning
  - mathematical models
  - function telling how the output variable depends on the input variable/s

$$\begin{cases} u(x) = g_1(x) + \int_0^1 (\frac{1}{2}t^2)u^2(t)dt + \int_0^1 xv(t)dt, \\ v(x) = g_2(x) + \int_0^1 u(t)dt + \int_0^1 \frac{1}{4}v^3(t)dt. \end{cases}$$

tdi@ek.dk

# Building Machine Learning Models




**Supervised**:

knowing the labels before the learning phase

**Unsupervised**:

not knowing any labels before the training

**Classification**

categorisation, association with a predefined class

- predicts qualitative values

**Regression**

estimation of a new, unknown real value

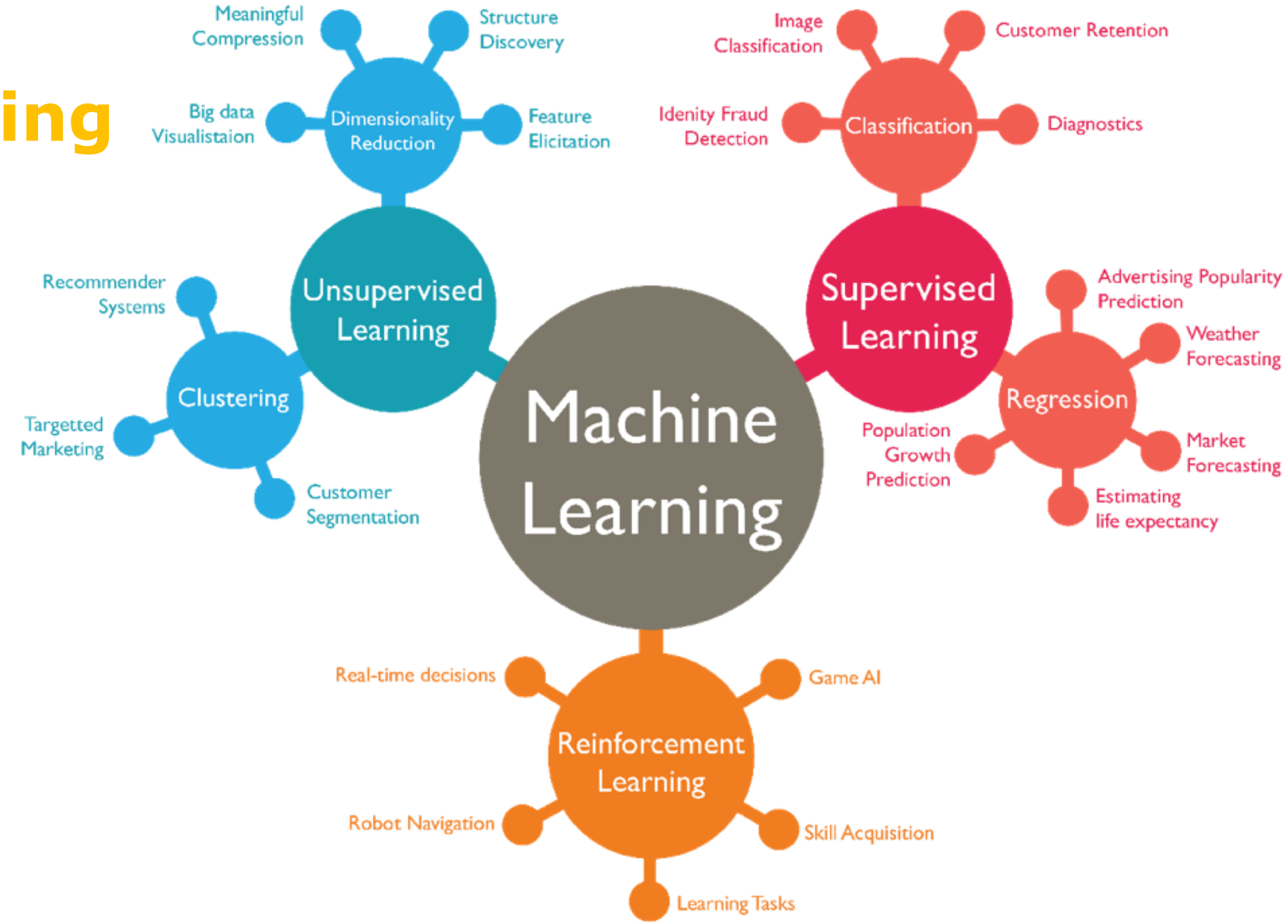
- predicts quantitative values

**Clustering**

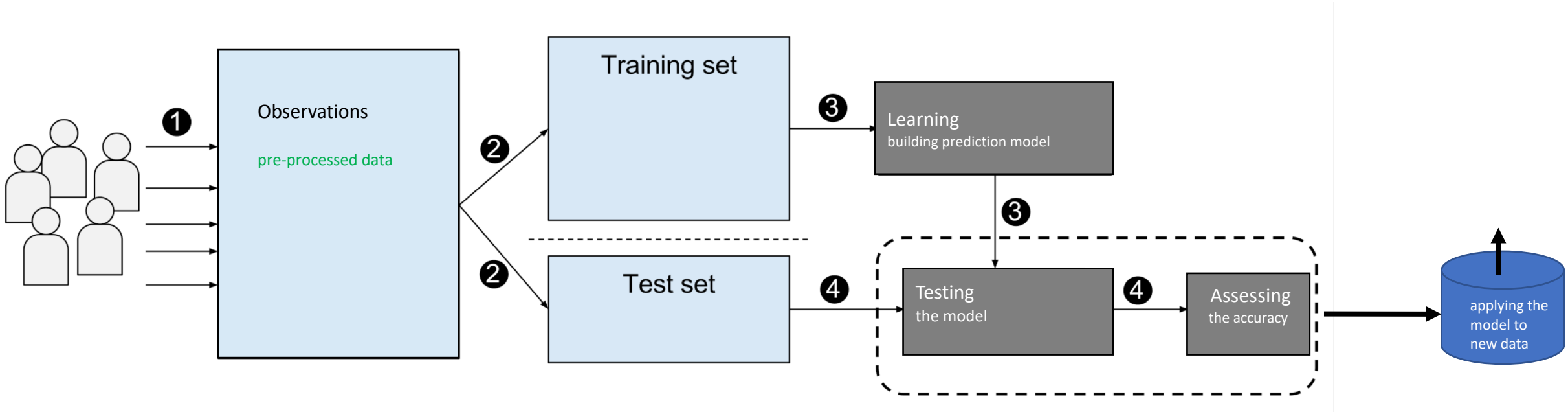
categorisation, association with newly defined groups, clusters

*https://blogs.oracle.com/datascience/types-of-machine-learning-and-top-10-algorithms-everyone-should-know-v2*

# Building Machine Learning Models



# Learning Process




tdi@ek.dk

## Training Machine Learning Model

- Formulate the problem
- Ingest the relevant data
- Clean it
- Wrangle it
- Explore it by statistical analysis
- Split it into train and test subset
- Train the train subset
- Test prediction on the test subset
- Validate the results with unknown data

tdi@ek.dk