PADERBORN UNIVERSITY

data analytics group

PREDICTIVE ANALYTICS

Kick-off

AGENDA

Today's Agenda

I. What is Predictive Analytics?

II. Taxonomy of Predictive Analytics
- PA by Methodology
- PA by Time Horizon
- PA by Input Data

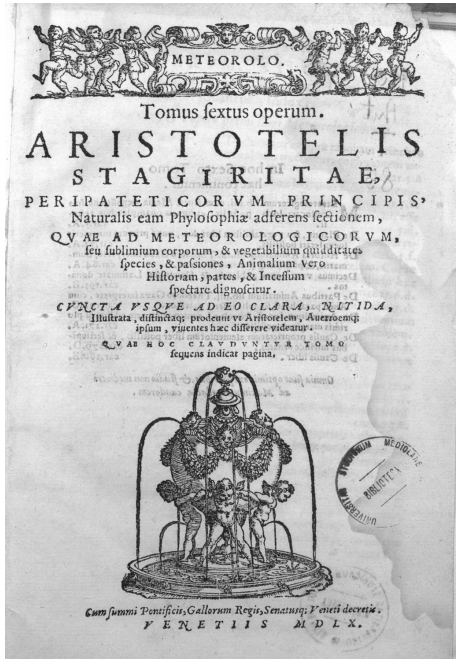
III. About this Course
- Your Task

IV. Kaggle
- The Challenges

WHAT IS PREDICTIVE ANALYTICS?

WHAT IS PREDICTIVE ANALYTICS?

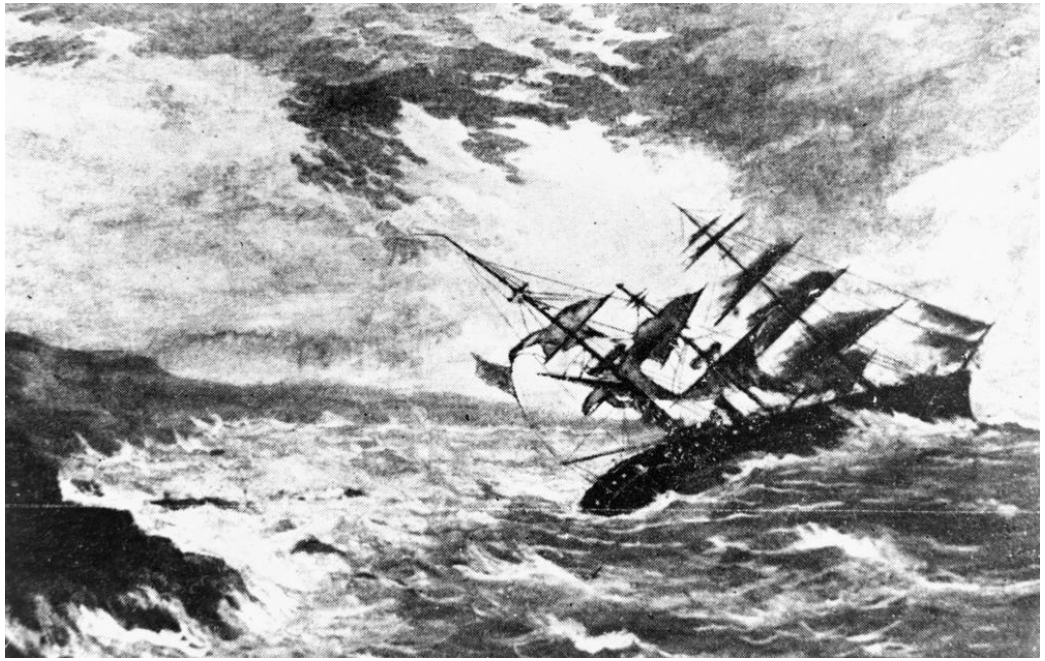
Weather Forecasting


METEOROLO.

Tomus ſextus operum.

ARISTOTELIS
STAGIRITAE,
PERIPATETICORVM PRINCIPIS,
Naturalis eam Philoſophiæ adferens ſectionem,
QVAE AD METEOROLOGICORVM,
ſeu ſublimium corporum, & vegetabilium quidditates
ſpecies, & paſsiones, Animalium vero
Hiſtoriam, partes, & Inceſſum
ſpectare dignoſcitur.
CVNCTA VSQVE AD EO CLARA, NITIDA,
Illuſtrata, diſtincta: prodeunt vt Ariſtotelem, Auerroemq;
ipſum, viuentes hæc differre videantur.
QVAE HOC CLAVDVNTVR TOMO
ſequens indicat pagina.

Cum ſummi Pontificis, Gallorum Regis, Senatusq; Venieti decreto.
VENETIIS MDLX.

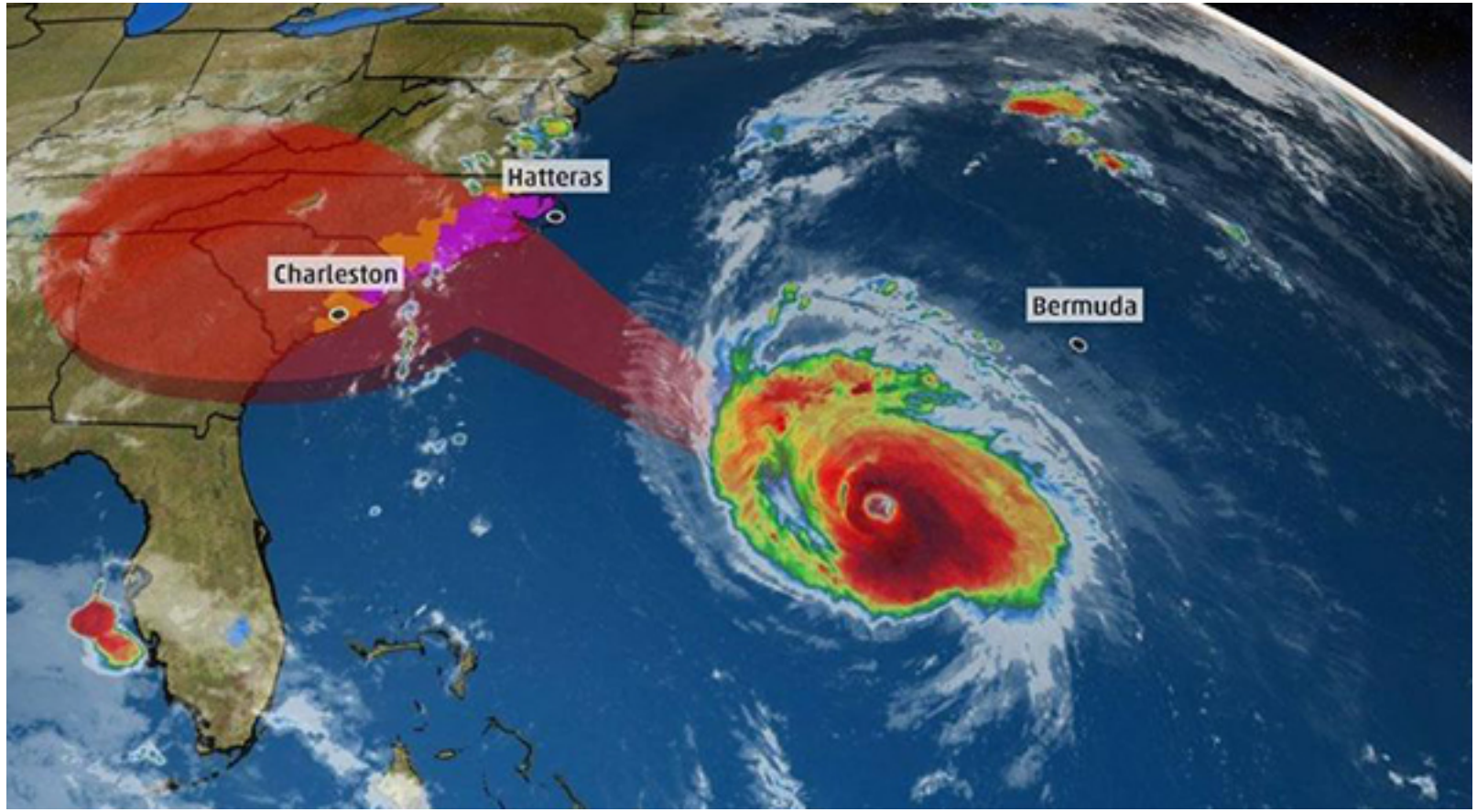


Meteorologica by Aristoteles,
350 BCE

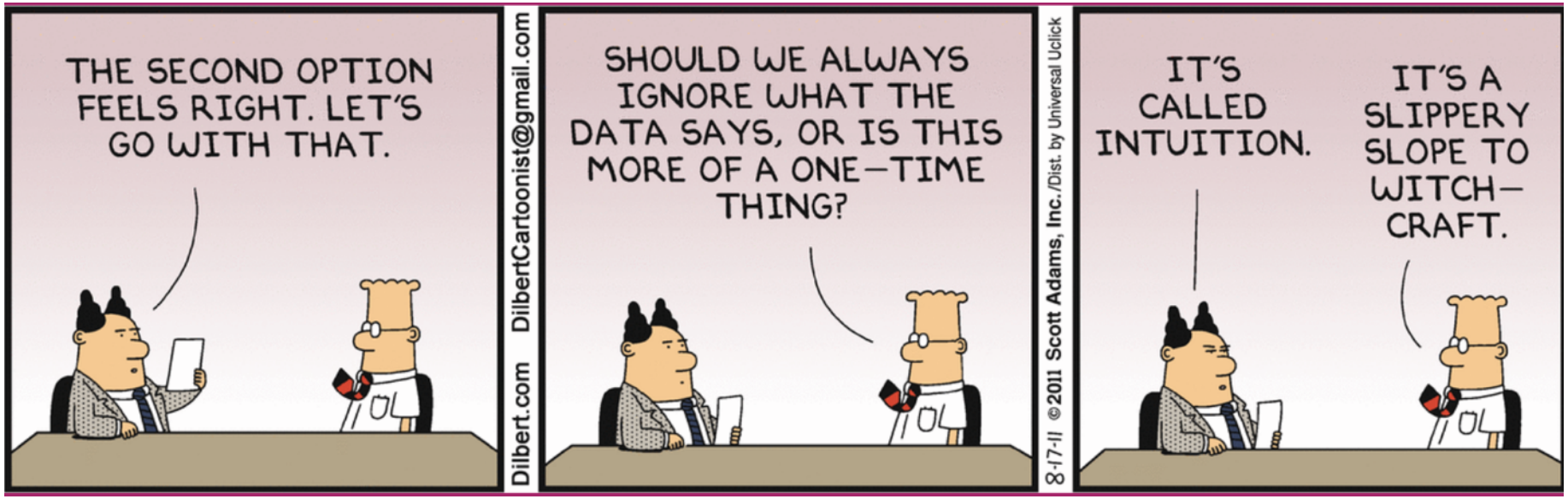
The Royal Charter sank in an 1859 storm,
stimulating the establishment of
modern weather forecasting

WHAT IS PREDICTIVE ANALYTICS?

Weather Forecasting



WHAT IS PREDICTIVE ANALYTICS?



WHAT IS PREDICTIVE ANALYTICS?

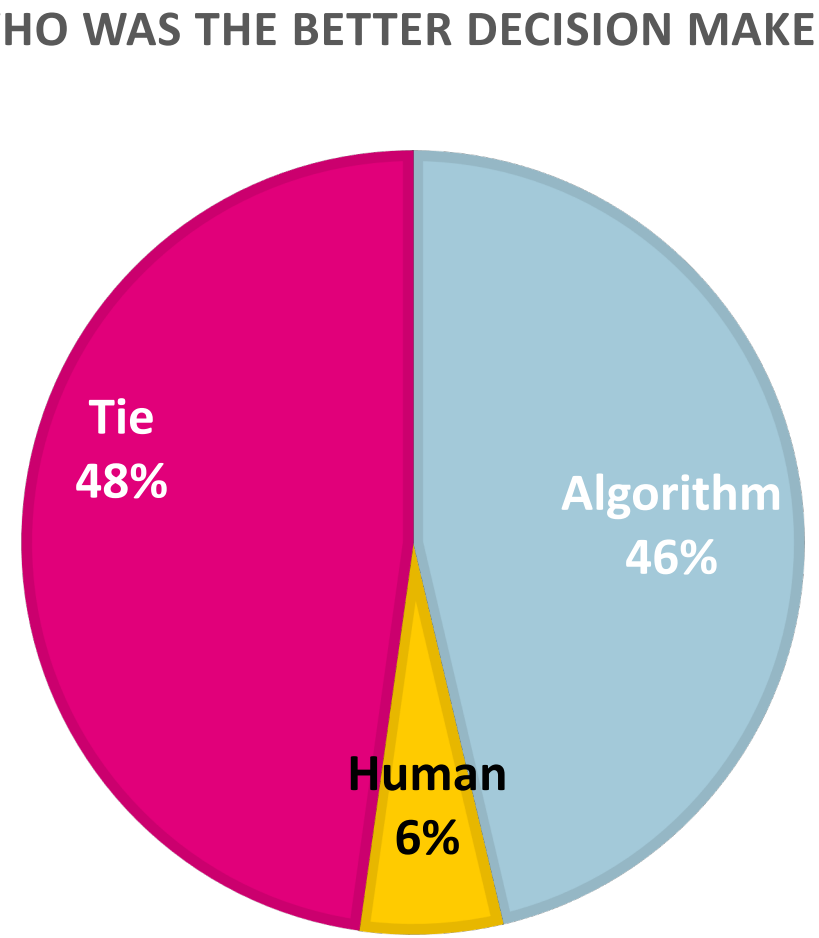
Literature Review by Grove et al. (2010)

- Extension of Meehl's (1954) seminal work

- Grove and colleagues analyzed 136 studies comparing the accuracy of human and algorithmic prediction, e.g.,
  - Medical and psychiatric diagnosis
  - Criminal behavior
  - Employee selection
  - Job performance
  - Student performance
  - Business failure
  - …

WHAT IS PREDICTIVE ANALYTICS?

Literature Review by Grove et al. (2010)



On average, **algorithmic prediction was about 10% more accurate** than human prediction.

Superiority of the algorithm was consistent regardless of task, type of humans, amounts of experience, or types of data being combined

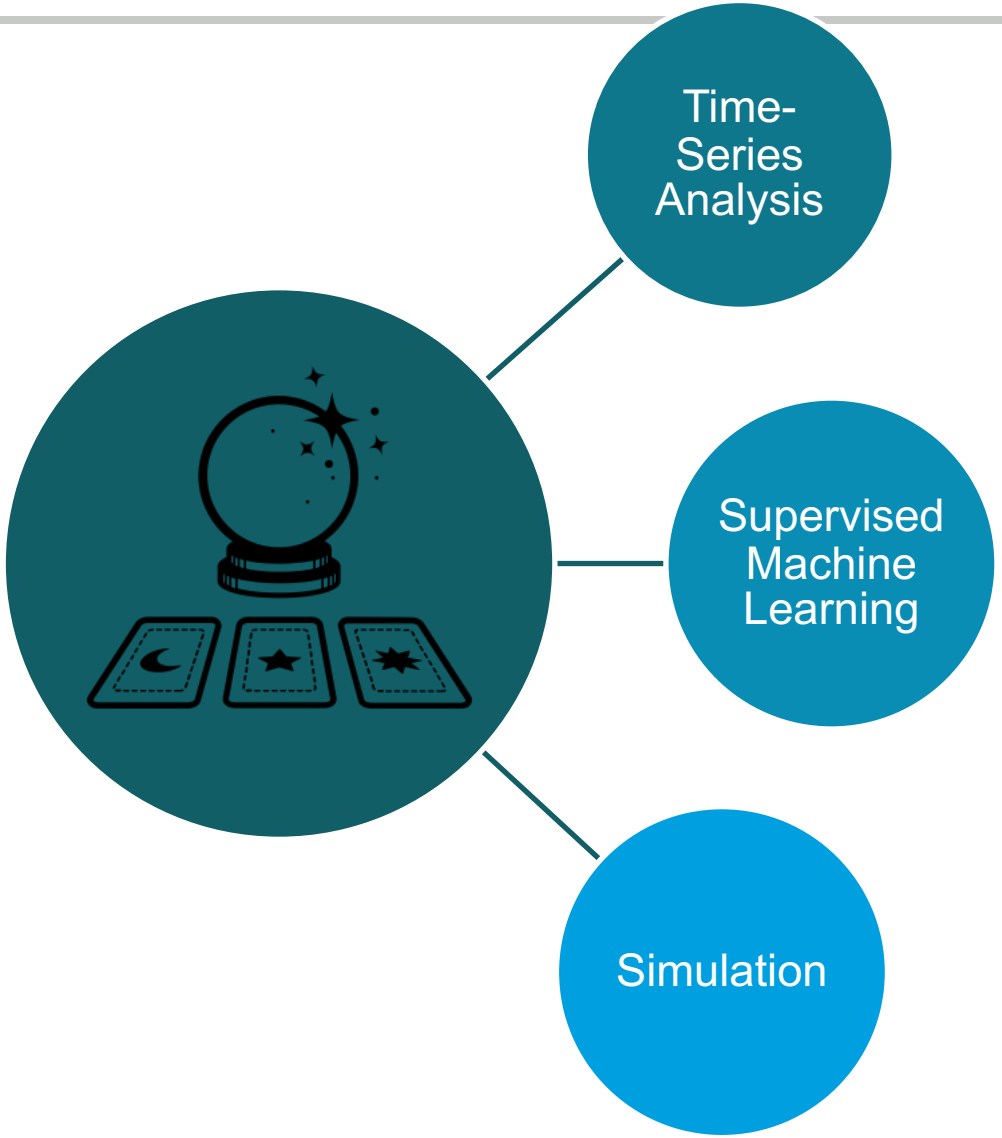
TAXONOMY OF PREDICTIVE ANALYTICS

TAXONOMY OF PREDICTIVE ANALYTICS

by Methodology

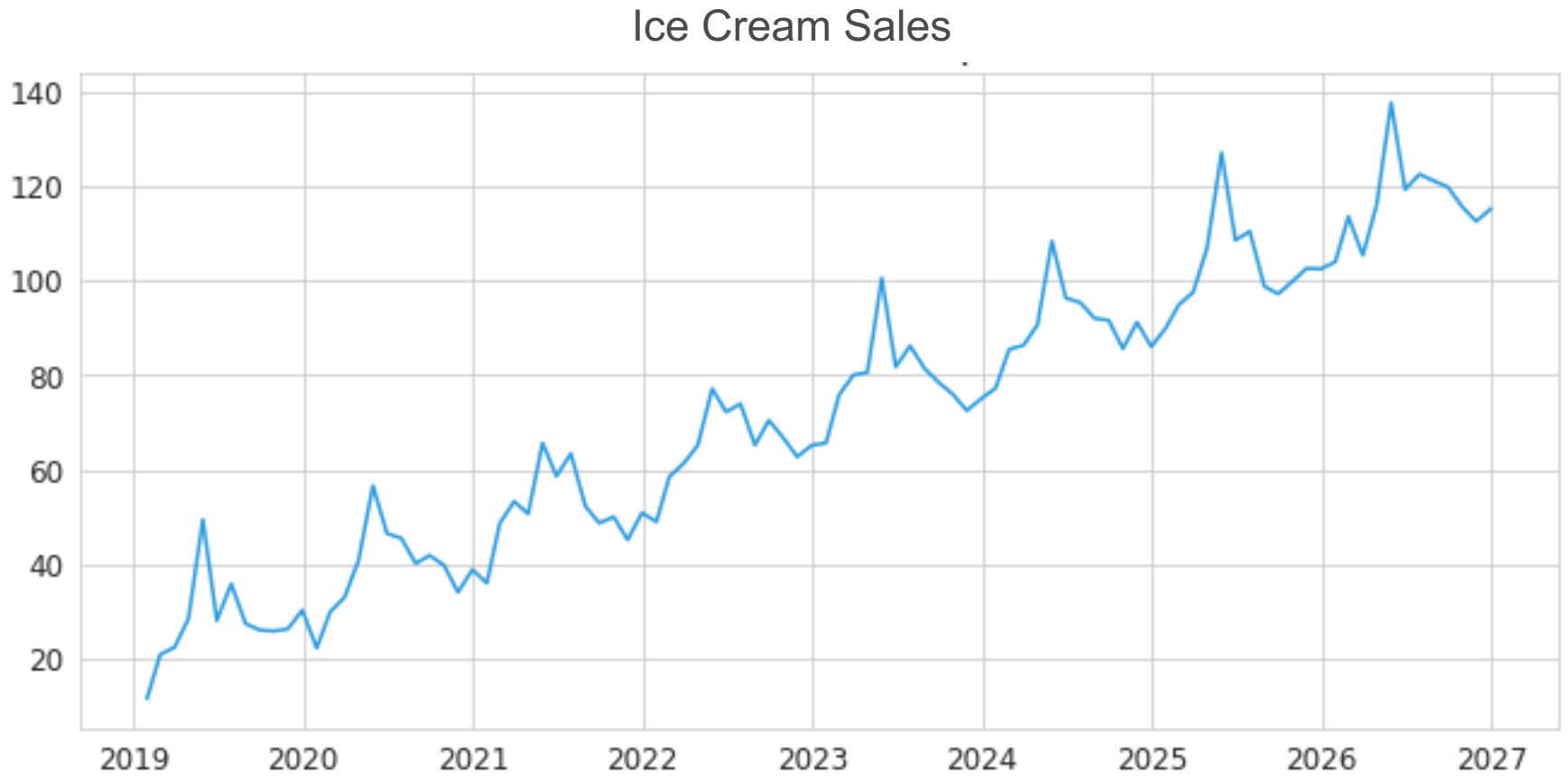
TAXONOMY: METHODOLOGY

Overview



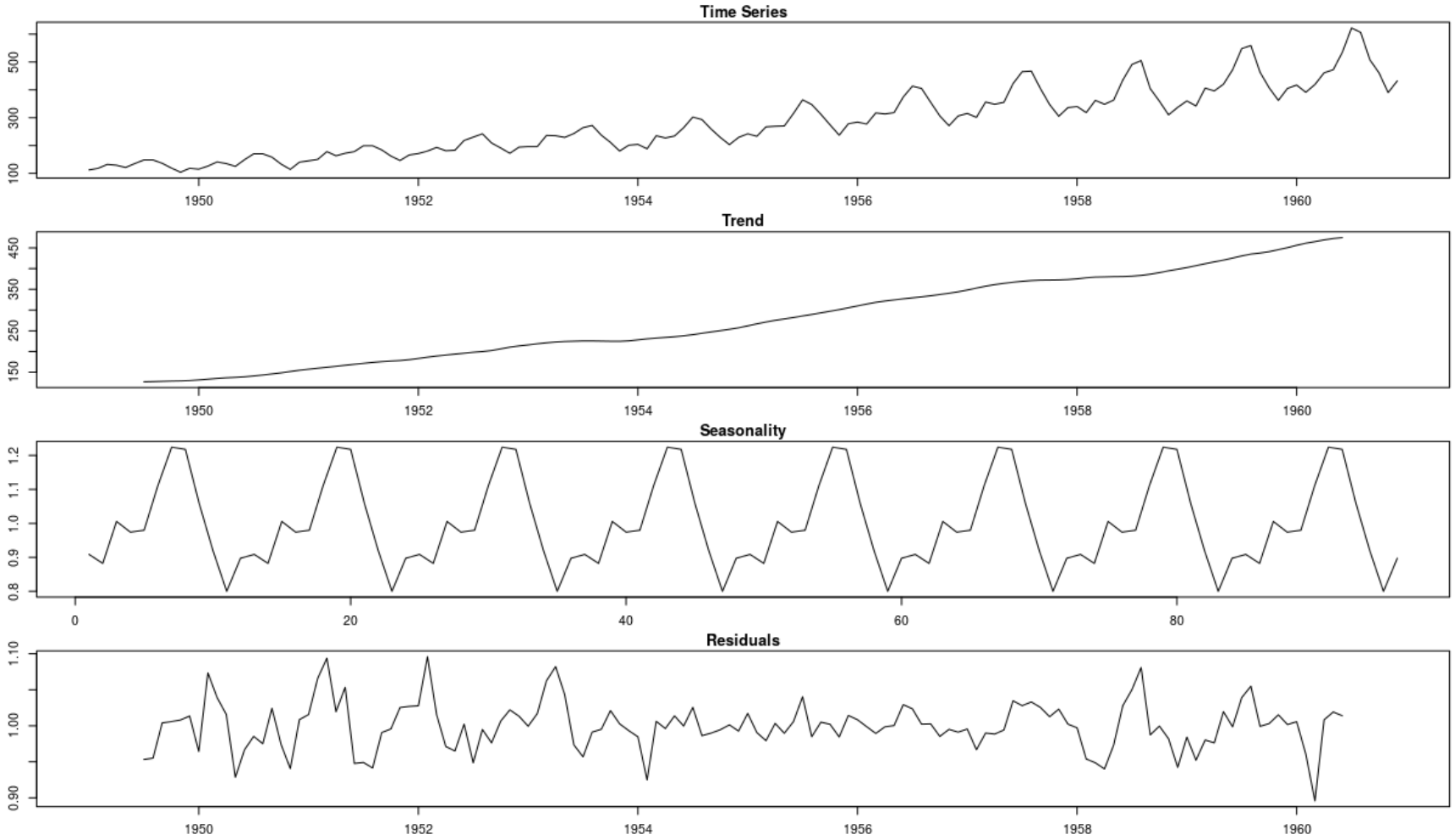
TAXONOMY: METHODOLOGY

Time-Series Analysis



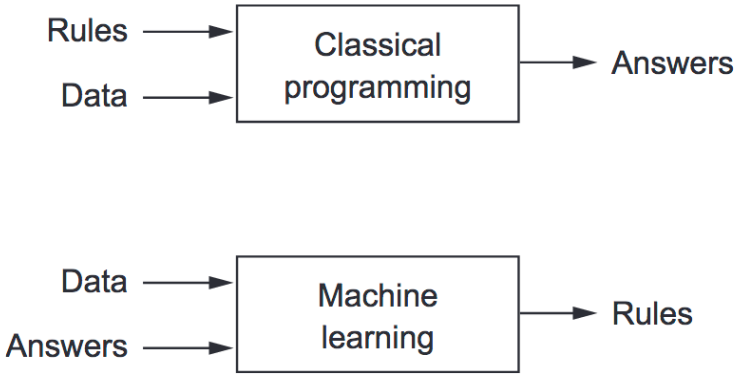
TAXONOMY: METHODOLOGY

Time-Series Analysis



TAXONOMY: METHODOLOGY

Supervised Machine Learning

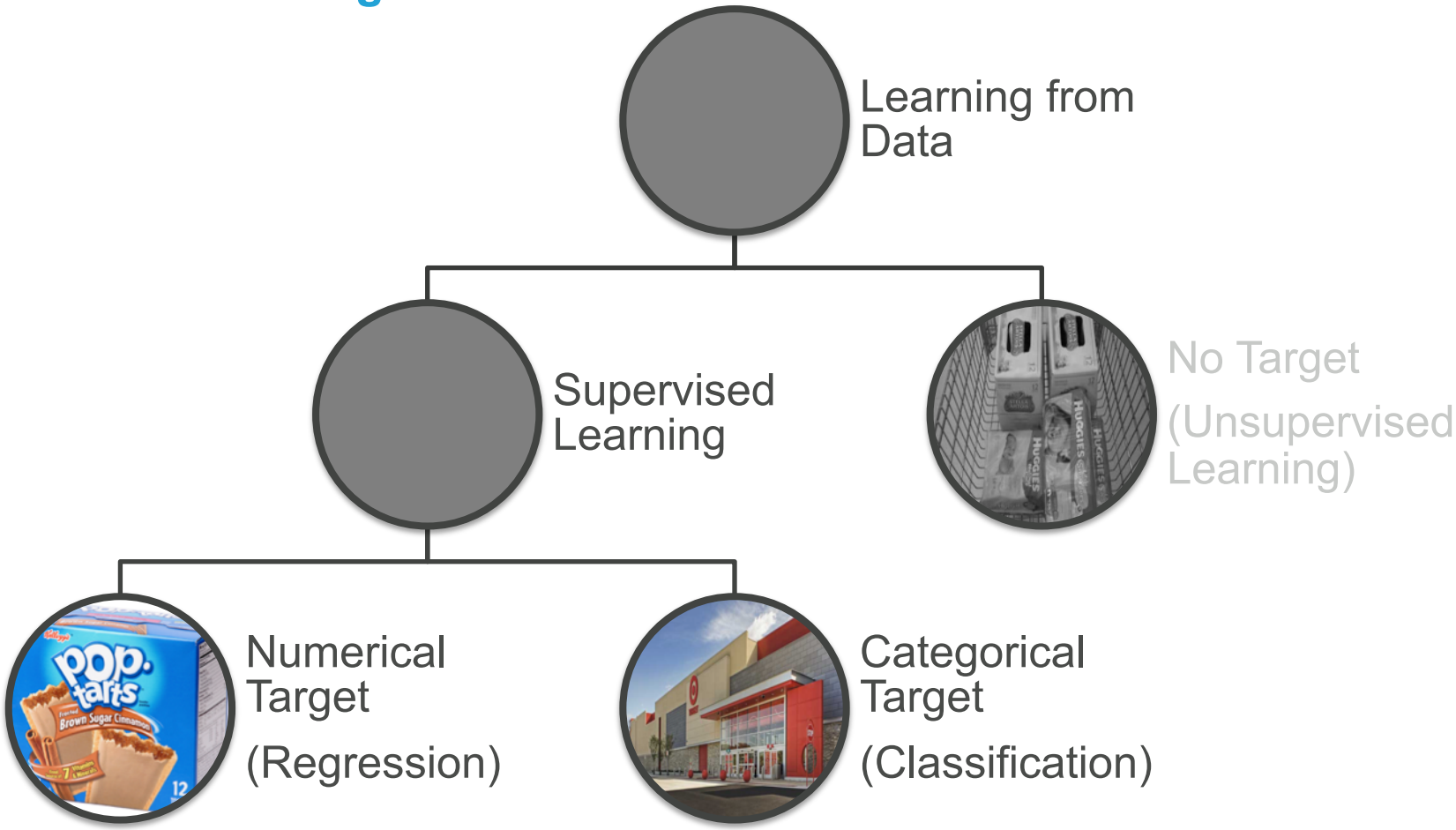


Supervised Machine Learning

$$Y = f(X) + \varepsilon$$

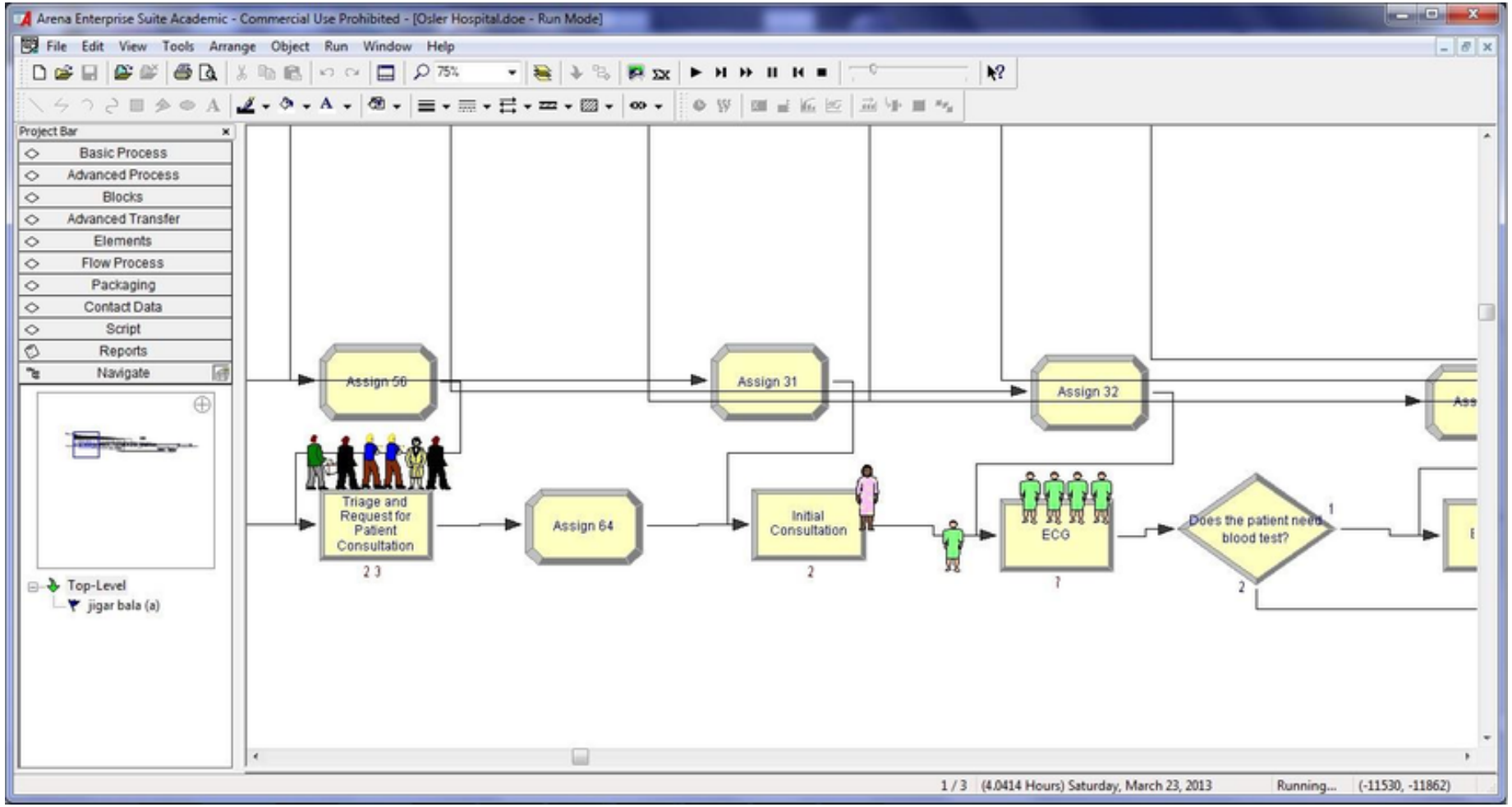
TAXONOMY: METHODOLOGY

Supervised Machine Learning



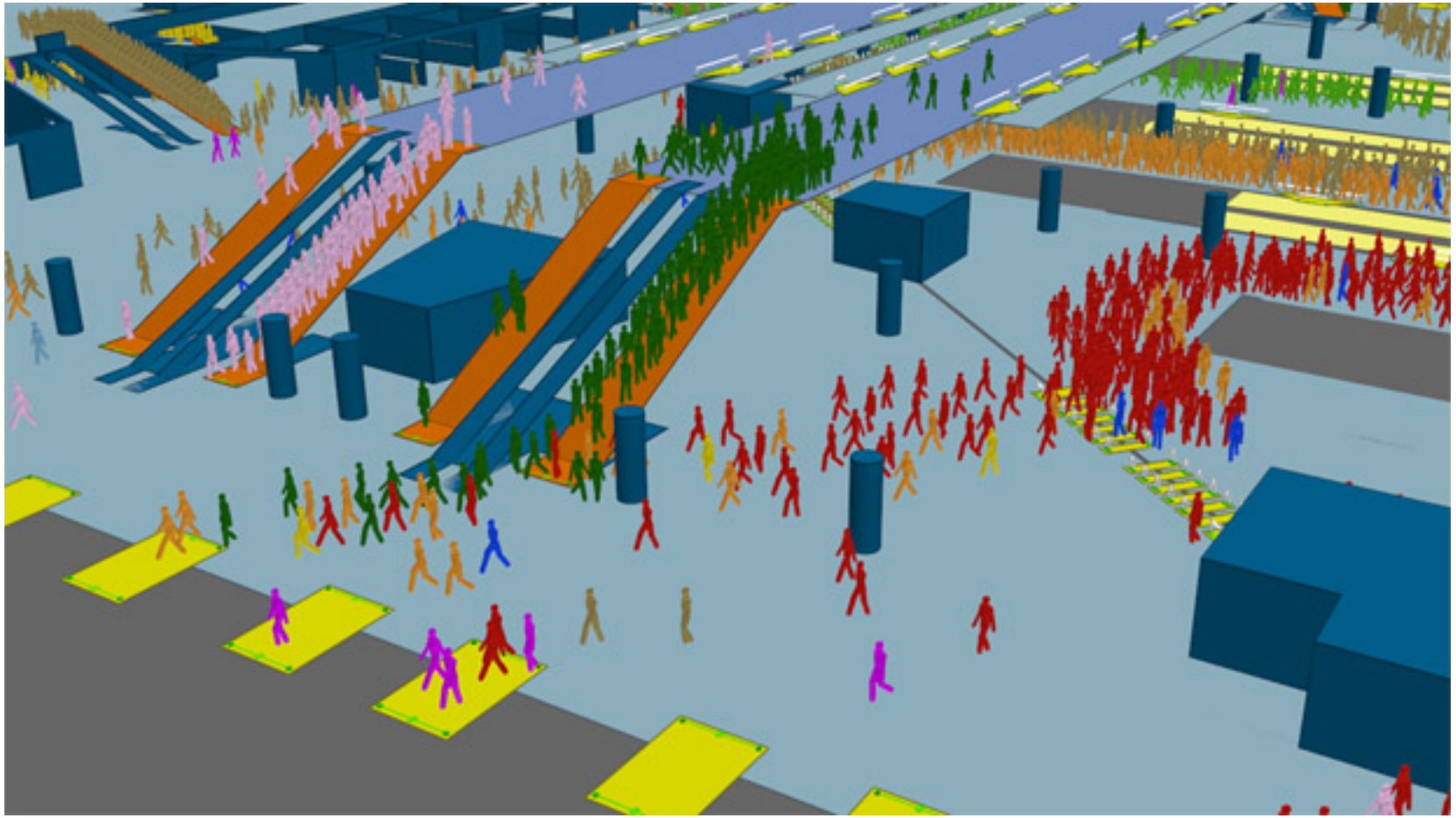
TAXONOMY: METHODOLOGY

Simulation: Processes and Events



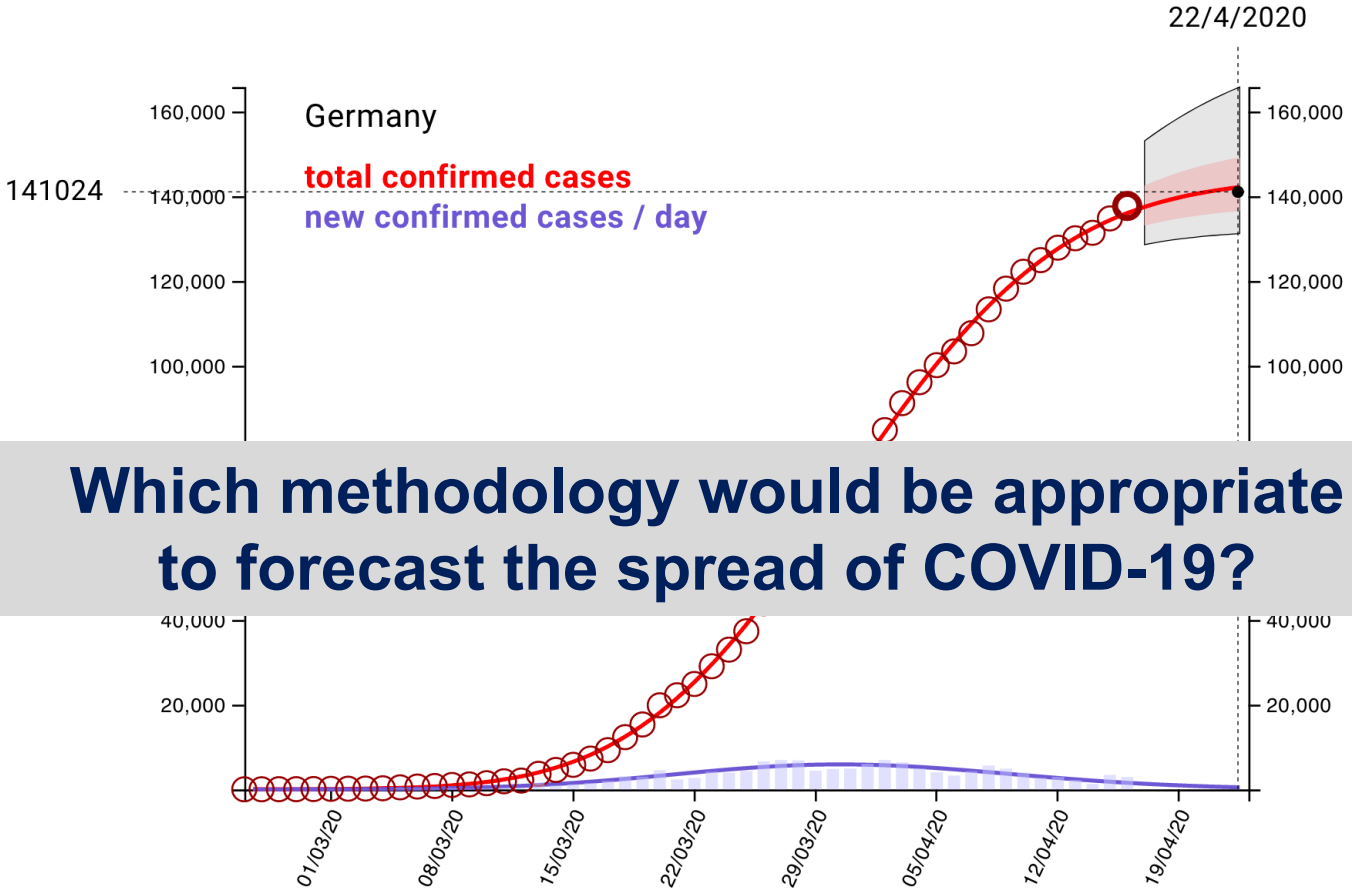
TAXONOMY: METHODOLOGY

Simulation: Agent-based Modeling



TAXONOMY: METHODOLOGY

Example: COVID-19



Which methodology would be appropriate to forecast the spread of COVID-19?

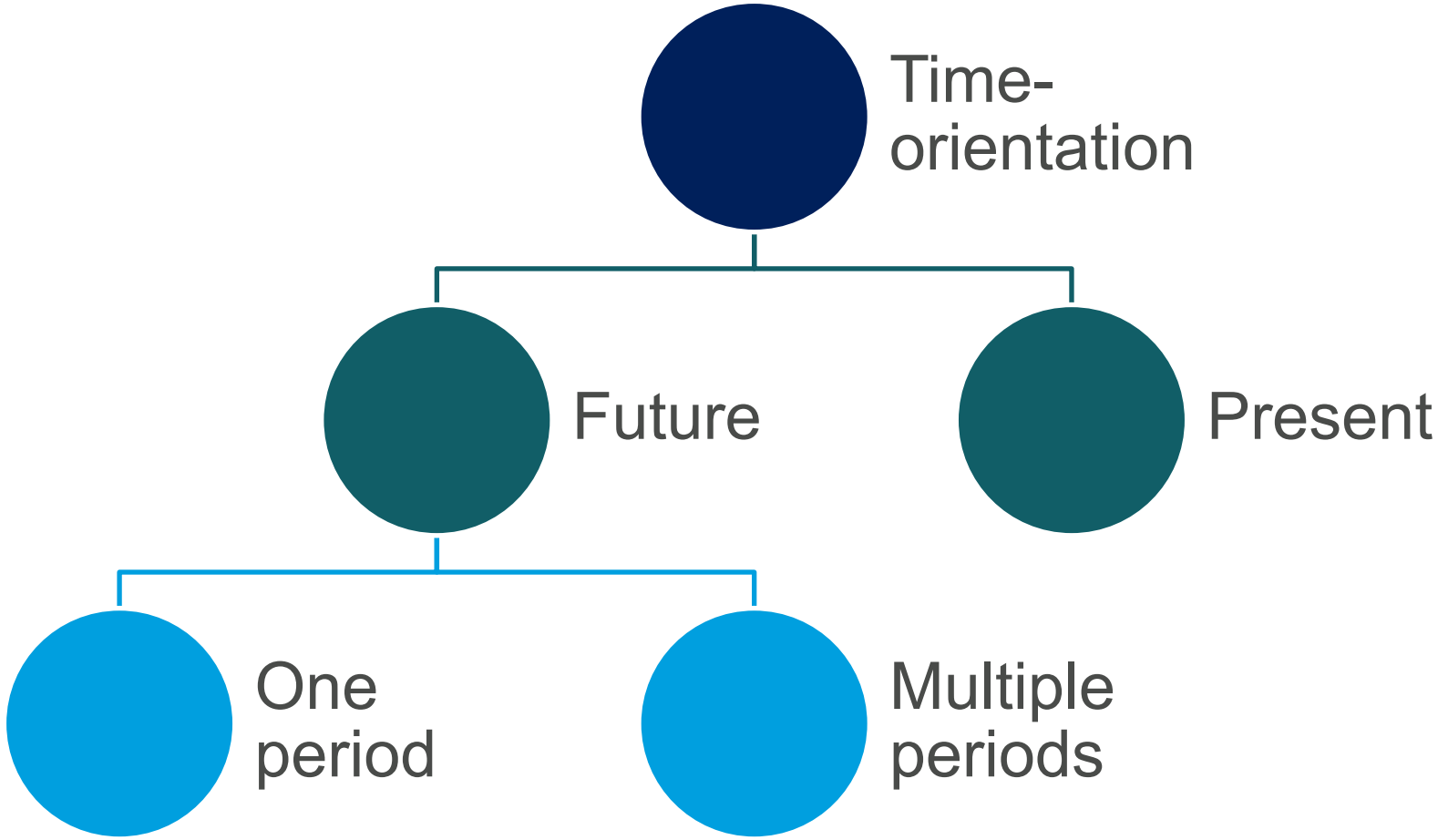
Source: http://rocs.hu-berlin.de/corona/docs/forecast/results_by_country/

TAXONOMY OF PREDICTIVE ANALYTICS

by Time Horizon

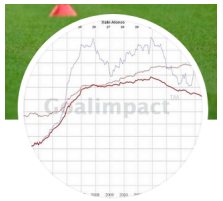
TAXONOMY: TIME HORIZON

Overview



TAXONOMY: TIME HORIZON

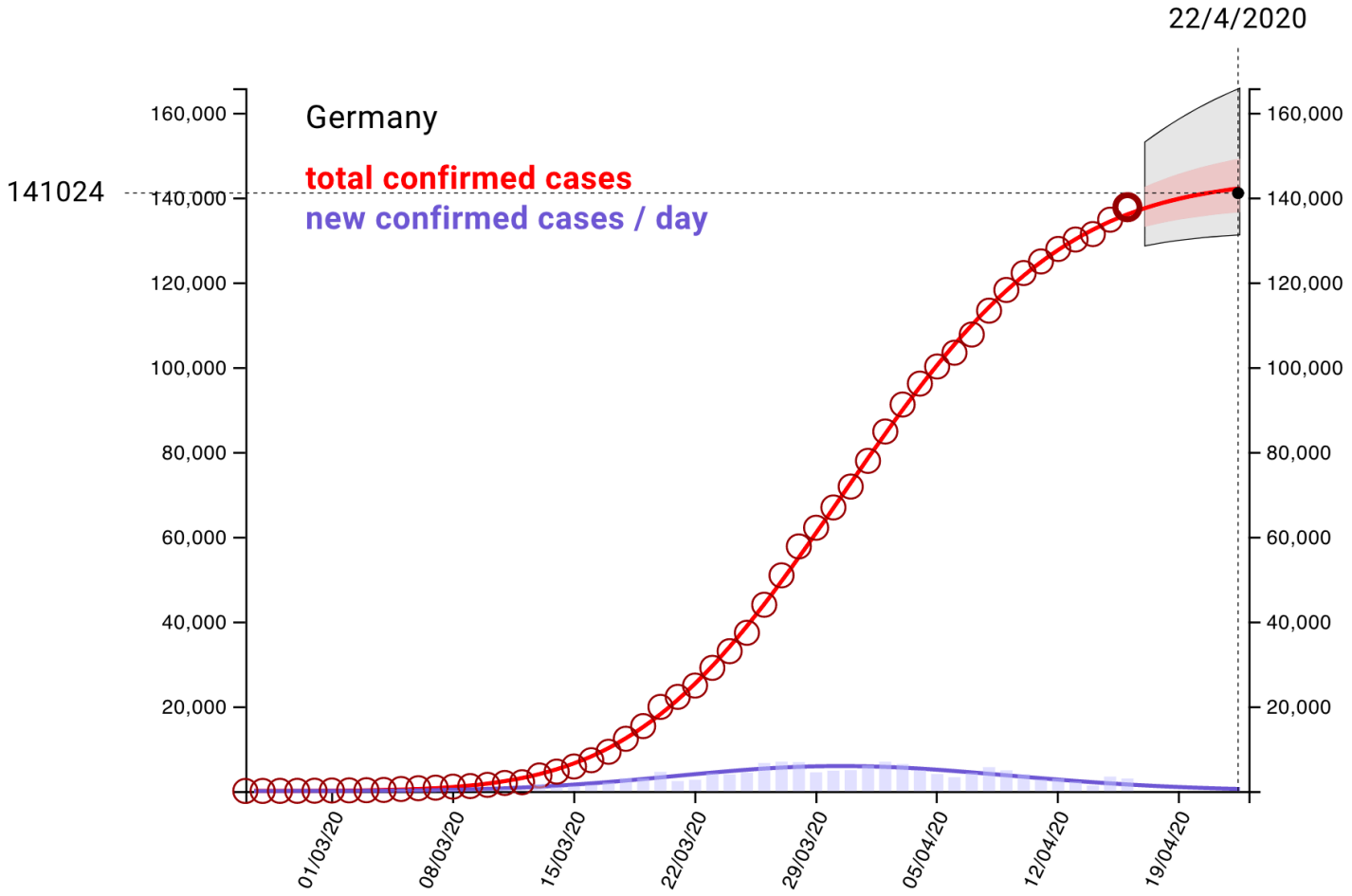
Future: One Period



	08.03.2020	Expected Points	97,8	76,4	64,9	61,4	59,1	57,3	55,7	54,2	52,7	51,2	49,7	47,9	46,0	43,9	41,6	39,7	38,1	36,5	34,6	31,6	
Expected Rank	Goalimpact	Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Expected Points
1,0	164,5	Liverpool FC	100,0%																				97,8
2,0	170,2	Manchester City		99,0%	0,9%	0,1%																	76,4
3,7	140,2	Leicester City		0,7%	60,4%	22,2%	9,4%	4,3%	1,8%	0,8%	0,3%	0,1%	0,0%										63,4
4,8	154,9	Chelsea FC		0,2%	22,0%	32,6%	19,2%	11,5%	6,9%	4,0%	2,2%	1,0%	0,4%	0,1%	0,0%								60,7
5,9	148,3	Manchester United		0,0%	8,5%	17,7%	22,2%	17,9%	12,8%	9,2%	5,7%	3,3%	1,7%	0,7%	0,3%	0,0%	0,0%						57,7
6,8	149,2	Arsenal FC		0,0%	3,3%	9,9%	16,0%	18,1%	17,0%	14,4%	10,4%	6,0%	3,2%	1,2%	0,3%	0,1%	0,0%						56,3
7,5	148,8	Tottenham Hotspur		0,0%	2,0%	6,7%	11,9%	15,4%	16,2%	15,3%	12,6%	9,0%	6,0%	3,2%	1,2%	0,3%	0,0%	0,0%					54,9
7,7	139,9	Wolverhampton Wandere			1,5%	5,7%	10,6%	14,6%	16,6%	16,0%	13,7%	10,1%	6,5%	3,4%	1,1%	0,2%	0,0%						54,8
8,6	130,8	Sheffield United			1,1%	3,2%	6,1%	9,1%	12,5%	15,1%	17,2%	15,0%	10,8%	6,3%	2,7%	0,7%	0,1%	0,0%					53,5
9,6	148,6	Everton FC			0,3%	1,3%	3,0%	5,2%	8,6%	11,5%	14,9%	17,9%	16,4%	11,2%	6,2%	2,6%	0,7%	0,2%	0,1%	0,0%			51,7
10,8	132,4	Crystal Palace			0,1%	0,4%	1,0%	2,3%	4,0%	6,7%	10,2%	15,5%	19,7%	20,4%	12,5%	5,4%	1,5%	0,4%	0,1%	0,0%	0,0%		49,9
11,1	135,6	Burnley FC			0,0%	0,1%	0,6%	1,4%	3,1%	5,4%	9,2%	14,6%	20,1%	22,5%	14,1%	6,3%	2,0%	0,6%	0,1%	0,0%	0,0%		49,4
13,1	135,7	Newcastle United				0,0%	0,1%	0,2%	0,5%	1,1%	2,4%	4,8%	8,9%	15,9%	26,2%	20,8%	10,1%	5,1%	2,5%	1,1%	0,4%	0,0%	46,1
14,0	134,7	Southampton FC					0,0%	0,1%	0,2%	0,5%	1,1%	2,3%	4,6%	9,4%	19,3%	27,7%	16,5%	9,3%	5,3%	2,6%	1,0%	0,2%	44,3
16,3	135,5	Brighton & Hove Albion							0,0%	0,0%	0,0%	0,2%	0,5%	1,8%	4,6%	10,0%	17,8%	19,7%	17,6%	14,2%	9,7%	3,8%	39,0
16,4	137,3	West Ham United							0,0%	0,0%	0,0%	0,1%	0,6%	1,7%	4,4%	9,5%	18,1%	18,5%	16,9%	14,1%	10,9%	5,1%	39,0
17,1	136,5	Aston Villa						0,0%	0,0%	0,0%	0,0%	0,1%	0,4%	1,1%	3,1%	6,3%	11,9%	14,8%	16,3%	17,9%	18,5%	9,7%	37,9
17,3	134,8	Watford FC									0,0%	0,0%	0,2%	0,6%	2,0%	5,1%	10,4%	14,2%	17,5%	19,6%	19,6%	10,7%	37,4
17,4	133,5	AFC Bournemouth								0,0%	0,0%	0,0%	0,2%	0,5%	1,8%	4,5%	9,3%	13,9%	18,1%	20,4%	21,2%	10,0%	37,4
19,2	142,5	Norwich City												0,0%	0,1%	0,4%	1,5%	3,2%	5,6%	10,0%	18,7%	60,5%	32,9

TAXONOMY: TIME HORIZON

Future: Multiple Periods



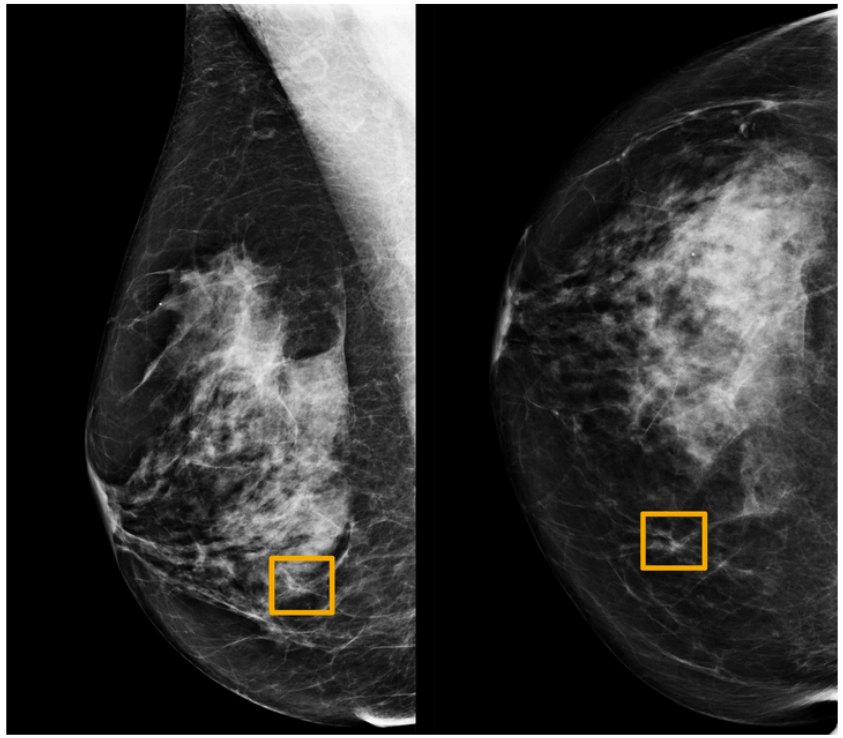
TAXONOMY: TIME HORIZON

Present

The New York Times

A.I. Is Learning to Read Mammograms

Computers that are trained to recognize patterns and interpret images may outperform humans at finding cancer on X-rays.



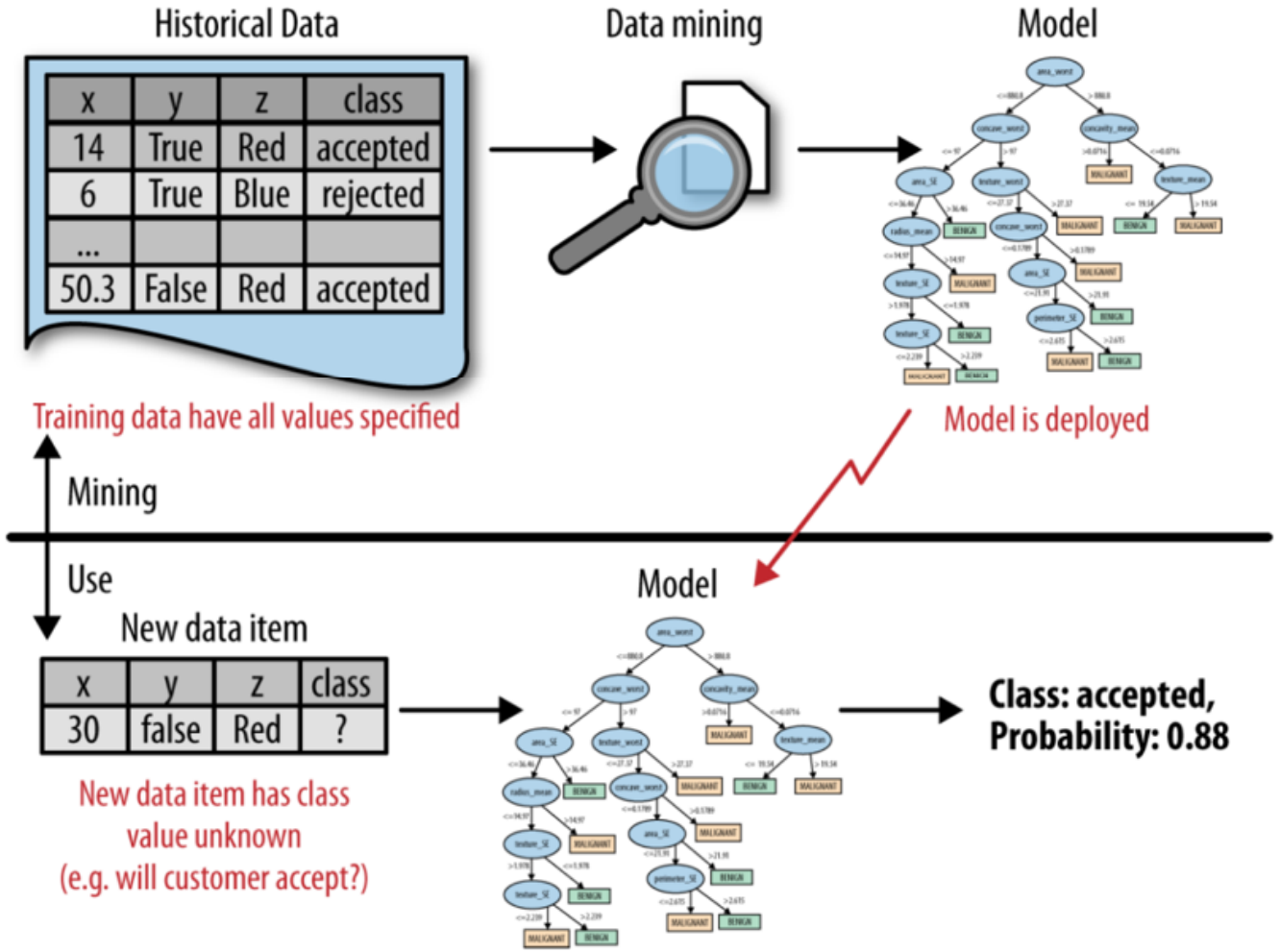
A yellow box indicates where an A.I. system found cancer hiding inside breast tissue. Six previous radiologists failed to find the cancer in routine mammograms. Northwestern University

TAXONOMY OF PREDICTIVE ANALYTICS

by Input Data

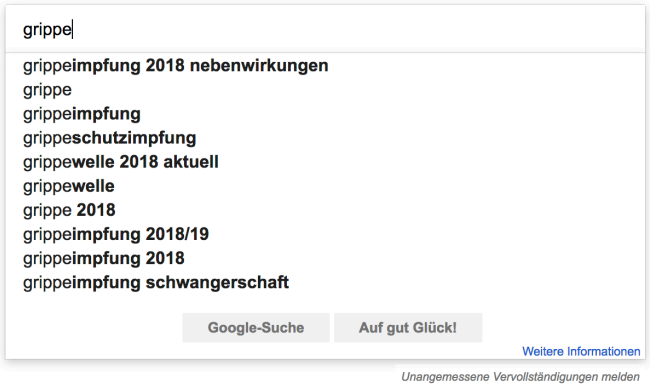
TAXONOMY: INPUT DATA

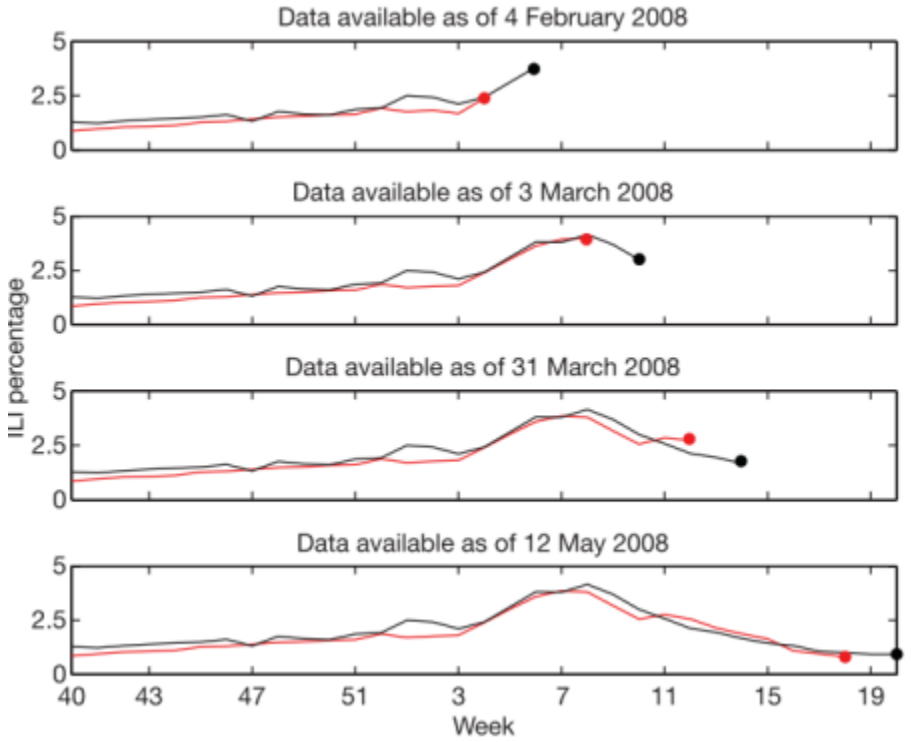
Tabular Data



TAXONOMY: INPUT DATA

Text Data





Source: Ginsberg et al. (2009)

"The final model was validated on 42 points per region of previously untested data from 2007 to 2008, which were excluded from all previous steps. Estimates generated for these 42 points obtained a **mean correlation of 0.97** (min: 0.92, max: 0.99, n: 9 regions) with the CDC-observed ILI percentages.

TAXONOMY: INPUT DATA

Text Data

$X_{1..n}$ Y

	Date	Location	"coughing"	„soar throat"	"cold"	...	ILI prevalence
Observation 1	01.01.2019	NYC	2321	3441	5513	...	0.020
Observation 2	01.01.2019	LA	1968	3201	4236	...	0.008
Observation 3	02.01.2019	NYC	2331	3446	5657	...	0.021
...	...	...	...	...	...	...	

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

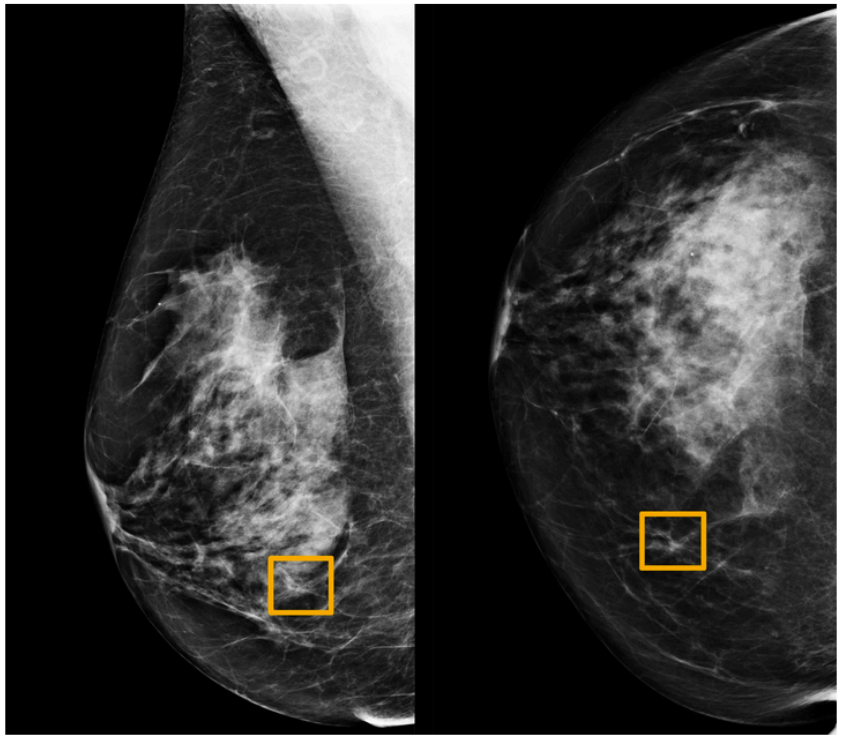
TAXONOMY: INPUT DATA

Image Data

The New York Times

A.I. Is Learning to Read Mammograms

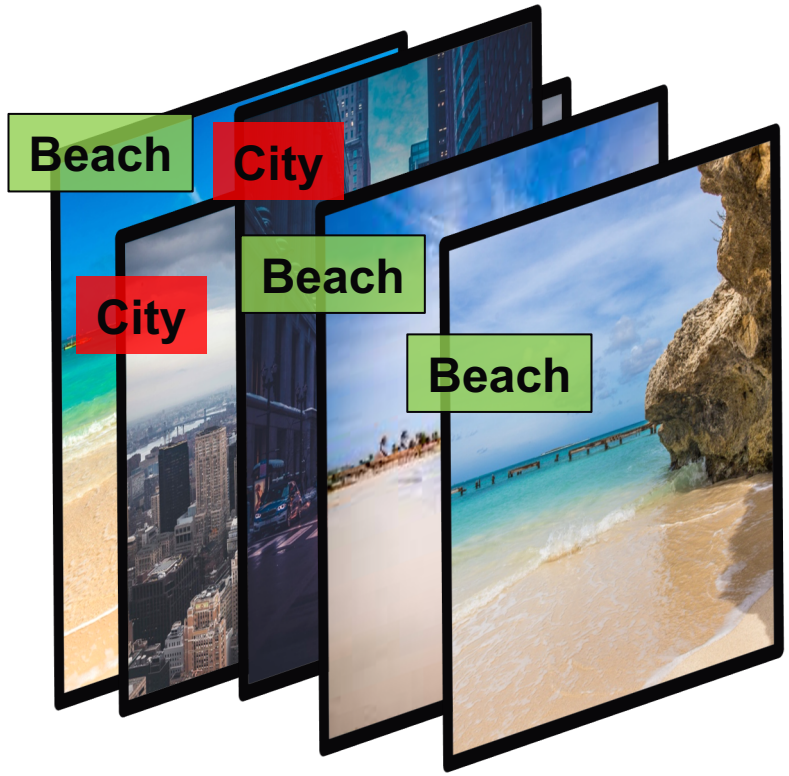
Computers that are trained to recognize patterns and interpret images may outperform humans at finding cancer on X-rays.

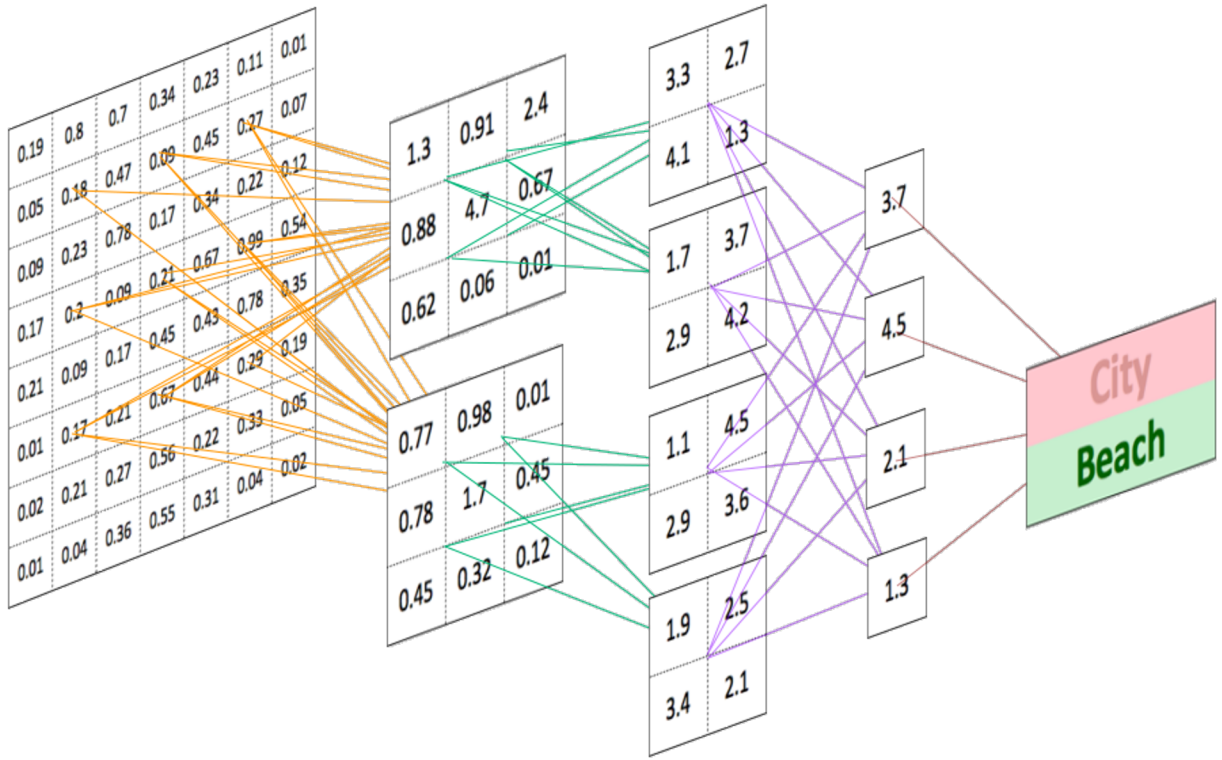


A yellow box indicates where an A.I. system found cancer hiding inside breast tissue. Six previous radiologists failed to find the cancer in routine mammograms. Northwestern University

TAXONOMY: INPUT DATA

Image Data





ABOUT THIS COURSE

ABOUT THIS COURSE

Teams of 1 or 2 students

Prüfungsleistungen *(examinations)*			
Art der Modulprüfung *(type of modul examination)*: Modulprüfung			
	Art der Prüfung *(type of examination)*	**Umfang** *(extent)*	**Gewichtung** *(weighting)*
a)	Hausarbeit mit Präsentation	ca. 15 Seiten	60.00 %
b)	Präsentation	20-30 Minuten	40.00 %

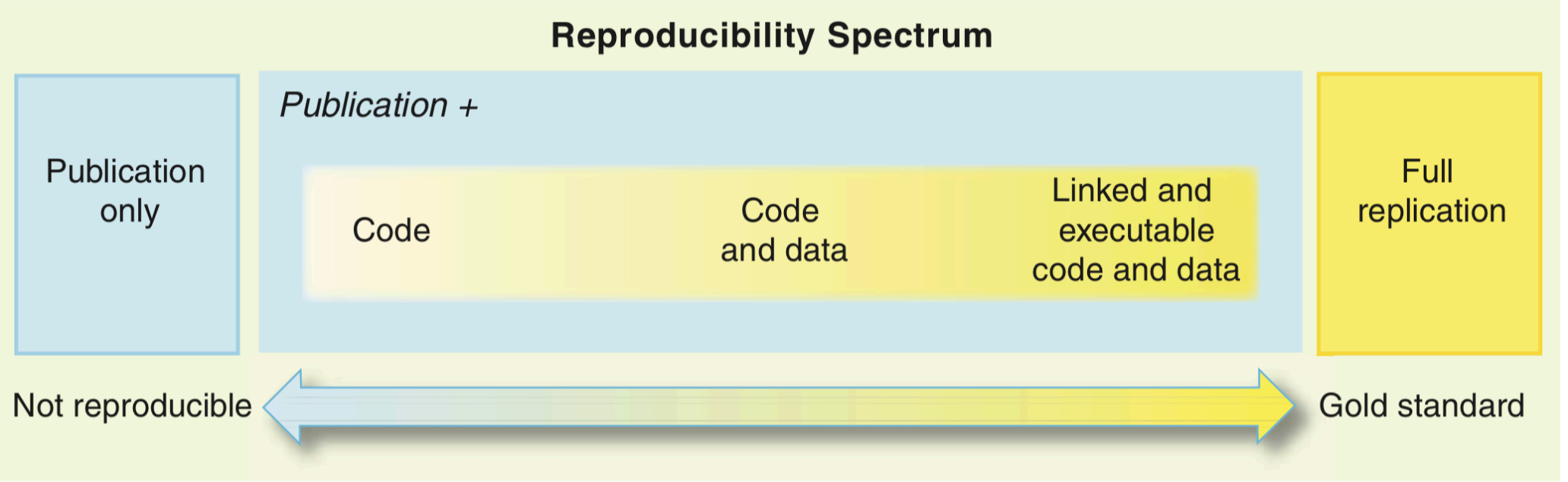
Hausarbeit: 26. Feb. 2021
Päsentation: 19. März 2021

Expectations

Implementation | Replication | Innovation

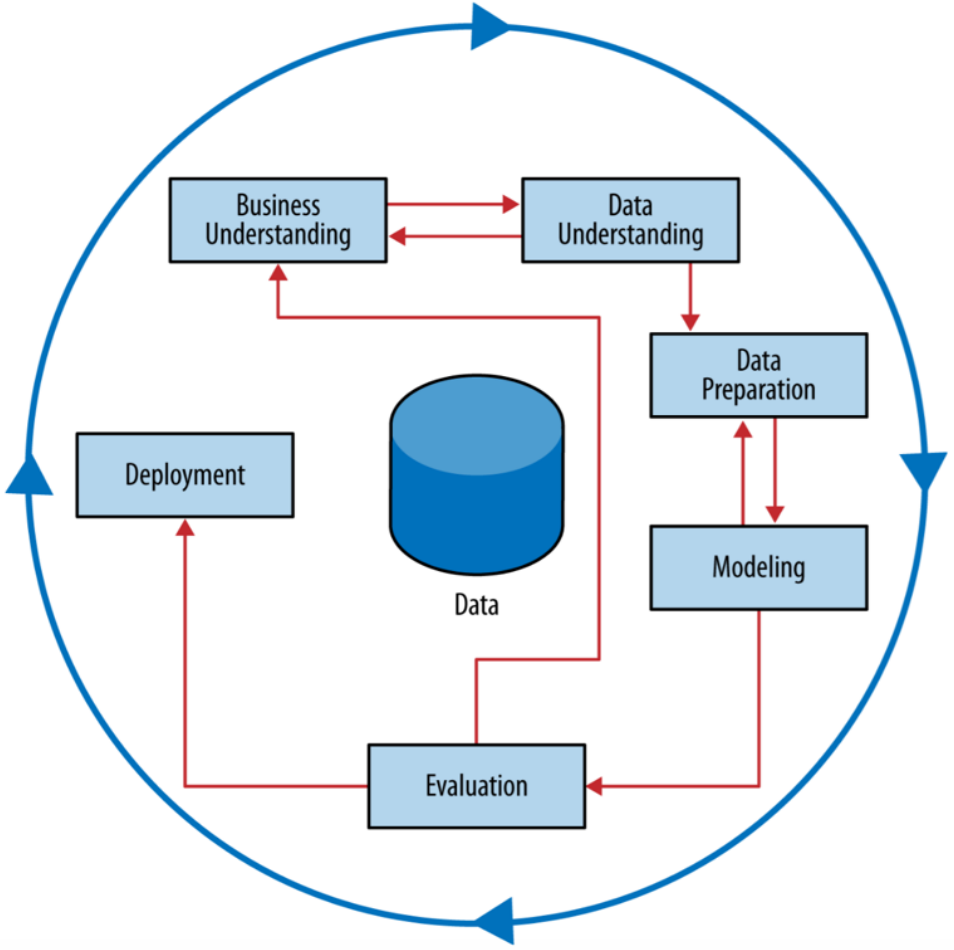
	Seminar Papers	Bachelor Thesis	Master Thesis
Topic/ Research Question	Given by supervisor	Find your own topic and RQ, based on existing literature	Find your own topic and RQ, based on real-life problem
Dataset	Given by supervisor	Select your own dataset (e.g., Kaggle)	Collect a “new” dataset
Structure	CRISP-DM	Like a scientific paper	Like a scientific paper
Literature Review	Business understanding, Modelling	+ Motivation & Research gap + Discussion	+ Implications for research and practice

Reproducibility



Source: Peng (2011)

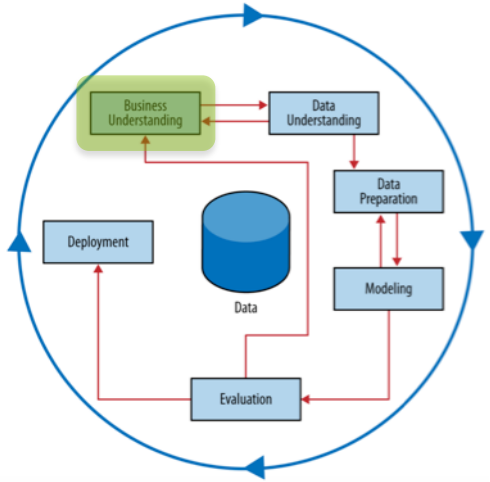
Cross-Industry Process for Data Mining (CRISP-DM)



Source: Shearer (2000)

Business Understanding

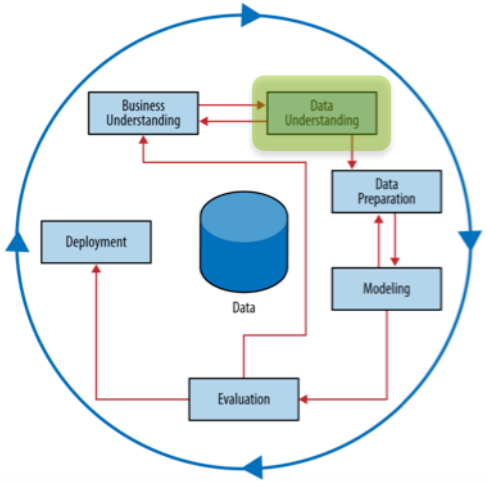
- Define business problem/question
- Restructure business problem as a data mining problem
  - Divide and conquer
  - Classification/regression/clustering/…
- Define success criteria



Source: Shearer (2000)

Data Understanding

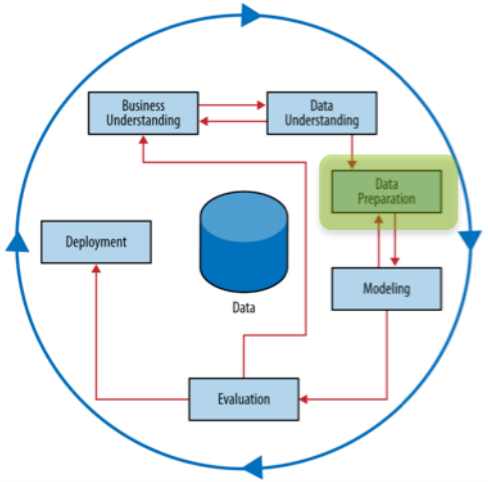
- Collect initial dataset
- Understand data structures
- Exploratory data analysis (incl. visualization)
- Assess data quality
- Acquire additional data



Source: Shearer (2000)

Data Preparation

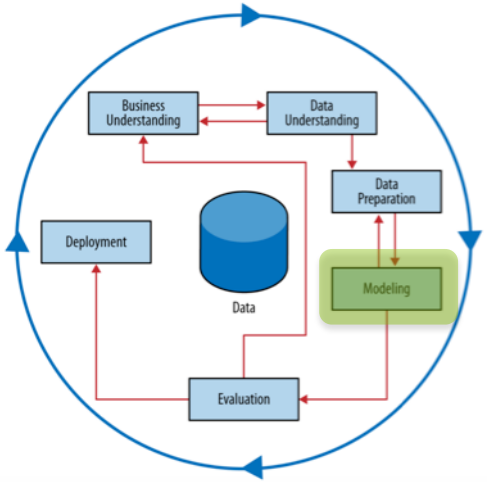
- Clean data
- Integrate data
- Reformat data
- Construct new data
  - new variables
  - new records

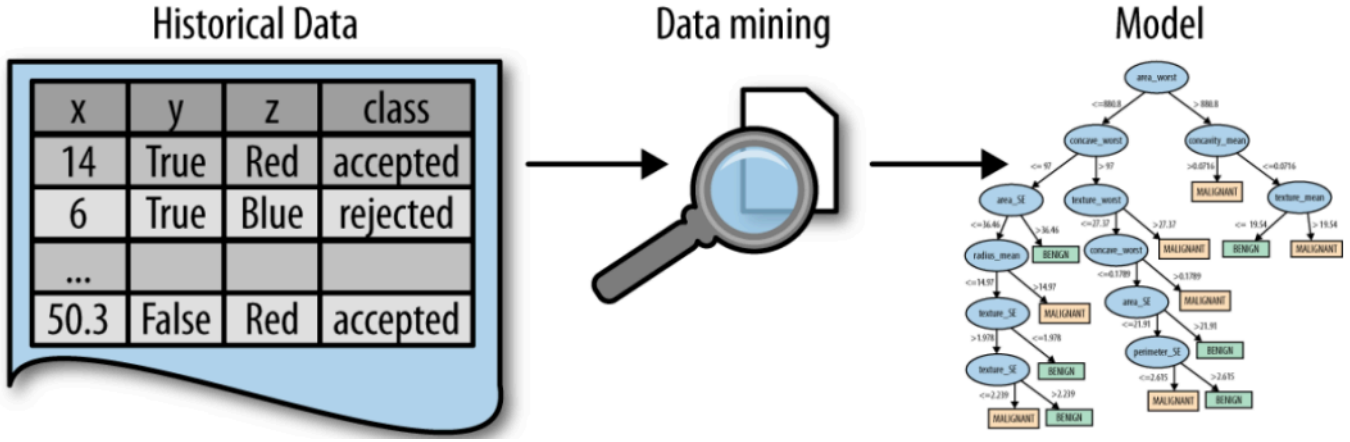


Source: Shearer (2000)

Modelling

- Select algorithm
- Generate training and test sets
- Fit model to data
- Assess model accuracy

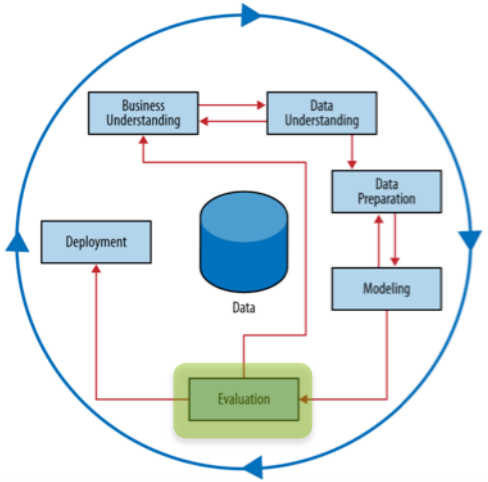




Source: Shearer (2000), Provost & Fawcett (2013)

Evaluation

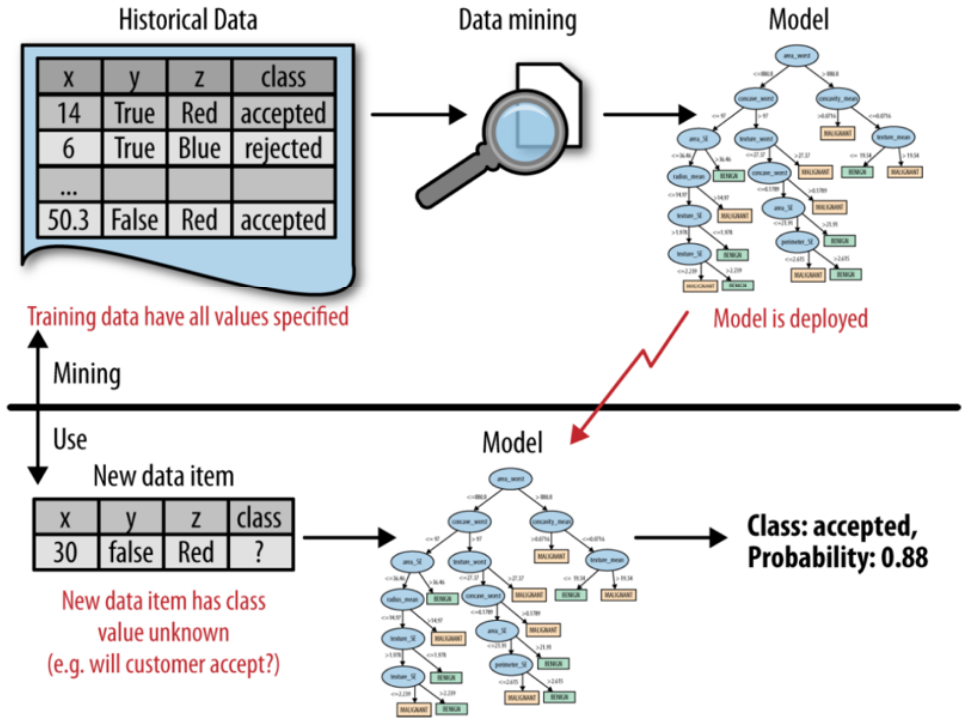
- Evaluate data mining results with respect to business goals and success criteria
- Assess legal and ethical considerations
- “In vivo” testing
- “Sign off” from relevant stakeholders

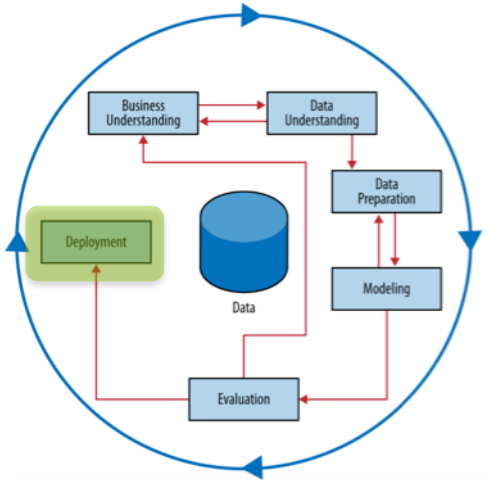


Source: Shearer (2000)

Deployment

- Develop production system
- Use system to solve new cases
- Continuous monitoring and maintenance





Source: Shearer (2000), Provost & Fawcett (2013)

KAGGLE

Competitions

Documentation | InClass

General | InClass | Hosted

Sort by Grouped

All Categories | Search competitions

3 Entered Competitions

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started · Ongoing · tutorial, tabular data, binary classification

Knowledge
11,173 teams

House Prices: Advanced Regression Techniques
Predict sales prices and practice feature engineering, RFs, and gradient boosting
Getting Started · Ongoing · tabular data, regression

Knowledge
4,362 teams

Predict Future Sales
Final project for "How to win a data science competition" Coursera course
Playground · 8 months to go

Kudos
2,874 teams

15 Active Competitions

2σ
TWO SIGMA

Two Sigma: Using News to Predict Stock Movements
Use news analytics to predict stock price performance
Featured · Kernels Competition · 3 months to go · news agencies, time series, finance, money

$100,000
2,927 teams

ZILLOW PRICE

- Goal: Improve the algorithm that changed the world of real estate
- Source: https://www.kaggle.com/c/zillow-prize-1
- Features:
  - Size
  - Quality
  - Location
  - Age
  - …
- Target:
  - Difference between Zestimate and sales price

HOME CREDIT DEFAULT RISK

- Goal: Predict whether a customer can payback a home loan
- Source: https://www.kaggle.com/c/home-credit-default-risk
- Features:
  - Income
  - Demographics
  - Family status
  - Information about desired loan
  - …
- Target:
  - Payback (yes/no)

ROSSMANN STORE SALES

- Goal: Forecast sales using store, promotion, and competitor data.
- Source: https://www.kaggle.com/c/rossmann-store-sales
- Features:
  - Store
  - Time
  - Promotions
  - Distance to next competitor's store
  - …
- Target:
  - Sales

NEW YORK CITY TAXI FARE PREDICTION

- Goal: Predict taxi fare based on pick-up and drop-off locations
- Source: https://www.kaggle.com/c/new-york-city-taxi-fare-prediction
- Features:
  - Pickup datetime
  - Pickup location (longitude+latitude)
  - Dropoff location (longitude+latitude)
  - Passenger count
- Target:
  - Fare amount

NEW YORK CITY TAXI TRIP DURATION

- Goal: Predict the total ride duration of taxi trips in New York City
- Source: https://www.kaggle.com/c/nyc-taxi-trip-duration
- Features:
  - Pickup datetime
  - Pickup location (longitude+latitude)
  - Dropoff location (longitude+latitude)
  - Passenger count
- Target:
  - Trip duration

ASHRAE - GREAT ENERGY PREDICTOR III

- Goal: Predict how much energy a building will consume.
- Source: https://www.kaggle.com/c/ashrae-energy-prediction
- Features:
  - Size, age, type of building
  - Time
  - Weather
  - …
- Target:
  - Energy usage

KOBE BRYANT SHOT SELECTION

- Goal: Which shots did Kobe sink?
- Source: https://www.kaggle.com/c/kobe-bryant-shot-selection
- Features:
  - Location (X and Y coordinates) of shot
  - Circumstances of shot (e.g., game, opponent, time on clock, ….)
- Target:
  - Shot successful (yes/no)

REAL OR NOT? NLP WITH DISASTER TWEETS

- Goal: Predict which Tweets are about real disasters and which ones are not
- Source: https://www.kaggle.com/c/nlp-getting-started
- Features:
  - Text of Tweets
  - Location of Tweet
  - Keywords
- Target:
  - Real or fake

SAN FRANCISCO CRIME CLASSIFICATION

- Goal: Predict the category of crimes that occurred in San Francisco
- Source: https://www.kaggle.com/c/sf-crime
- Features:
  - Time
  - Location
- Target:
  - Category of crime (multi-class)

SEE CLICK PREDICT FIX

- Goal: Predict which 311 issues are most important to citizens
- Source: https://www.kaggle.com/c/see-click-predict-fix
- Features:
  - Time the issue originated
  - Text description of the issue
  - Issue location (longitude+latitude)
  - …
- Target:
  - Number of votes, comments, and views

COSTA RICAN HOUSEHOLD POVERTY LEVEL PREDICTION

- Goal: Predict households with the highest need for social welfare assistance
- Source: https://www.kaggle.com/c/costa-rican-household-poverty-prediction
- Features:
  - Monthly rent payment
  - Overcrowding by bedrooms
  - Number of persons living in the household
  - …
- Target:
  - Groups of income levels (1-5)

WALMART RECRUITING STORE SALES FORECASTING

- Goal: Predict store sales based on historical markdown data
- Source: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting
- Features:
  - Dates + sales
  - Temperature
  - Store size
  - …
- Target:
  - Weekly sales

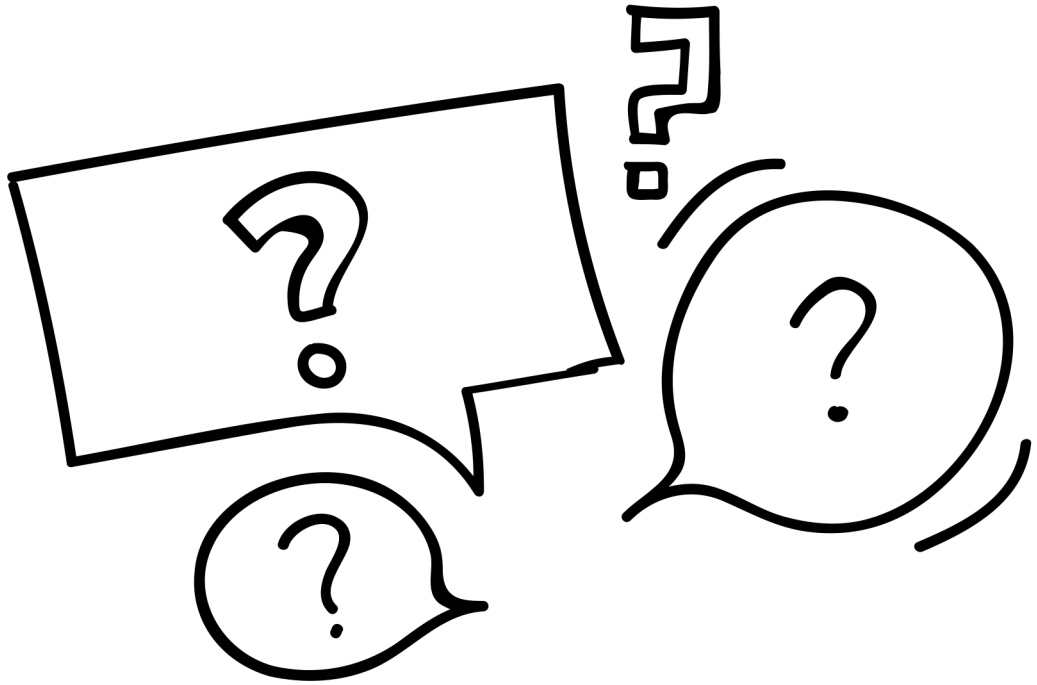
TMDB BOX OFFICE PREDICTION

- Goal: Predict a movie's worldwide box office revenue
- Source: https://www.kaggle.com/c/tmdb-box-office-prediction
- Features:
  - Genre
  - Popularity
  - Original language
  - Budget
  - …
- Target:
  - Revenue

DON'T GET KICKED!

- Goal: Predict if a car purchased at auction is a lemon
- Source: https://www.kaggle.com/c/DontGetKicked
- Features:
  - Vehicle build year
  - Model
  - Size
  - …
- Target:
  - Is bad buy? (yes/no)

THE END



Prof. Dr. Oliver Müller
Lehrstuhl für Wirtschaftsinformatik, insb. Data Analytics
Universität Paderborn
Warburger Str. 100, 33098 Paderborn
R: Q2.457
E: oliver.mueller@uni-paderborn.de
T: +49-5251-605100
W: https://wiwi.uni-paderborn.de/dep3/mueller/