Introduction to Data Science

Introduction to Data Science

https://sherbold.github.io/intro-to-data-science

Outline

- Introduction to Big Data
- Data Science definition
- The Skillset of Data Scientists
- AI
- ML
- Summary

What is „Big Data“?!?

Is this really about size?

Naive Definition

- Naive definition:
  - Big data only depends on the data size
  - 1 Gigabyte? 1 Terabyte? 1 Petabyte?

- Naive interpretation misses important aspects
  - Time:
    - Analyzing 1 Gigabyte of data per day is different from analyzing 1 Gigabyte of data per second
  - Diversity:
    - Analyzing spread sheets with numeric data is different from analyzing Web pages that contain a mixture of text and images
  - Distribution:
    - Analyzing data from a single source is different from analyzing data from multiple sources

Definition of Big Data

- Following Gartner‘s IT Glossary:
  - Big data is high-**volume**, high-**velocity** and/or high-**variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

- The three Vs
  - Volume
  - Velocity
  - Variety

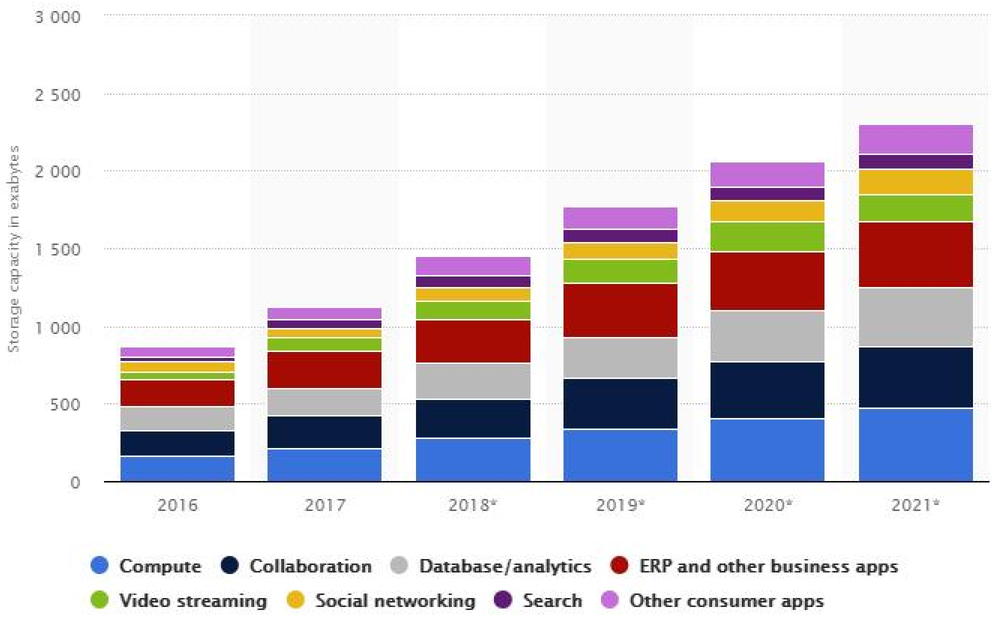
Some people actually use 10 Vs to define big data!

- Variability
- Veracity
- Validity
- Vulnerability
- Volatility
- Visualization
- Value

The 3 Vs: Volume

- Scale of the data must be „big“
  - No clear definition
  - „that demand […] innovative forms of information processing“ (Gartner)

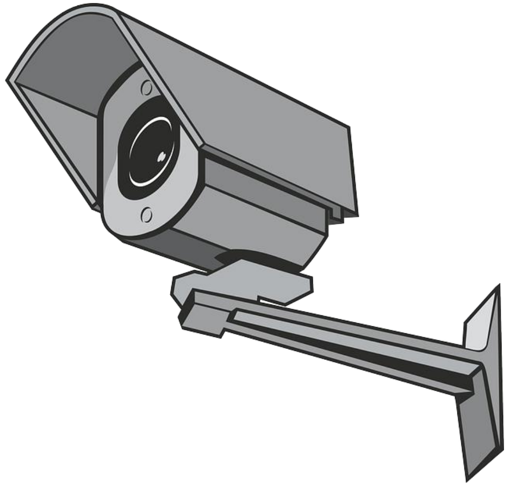
Data center storage worldwide



© Statista 2018

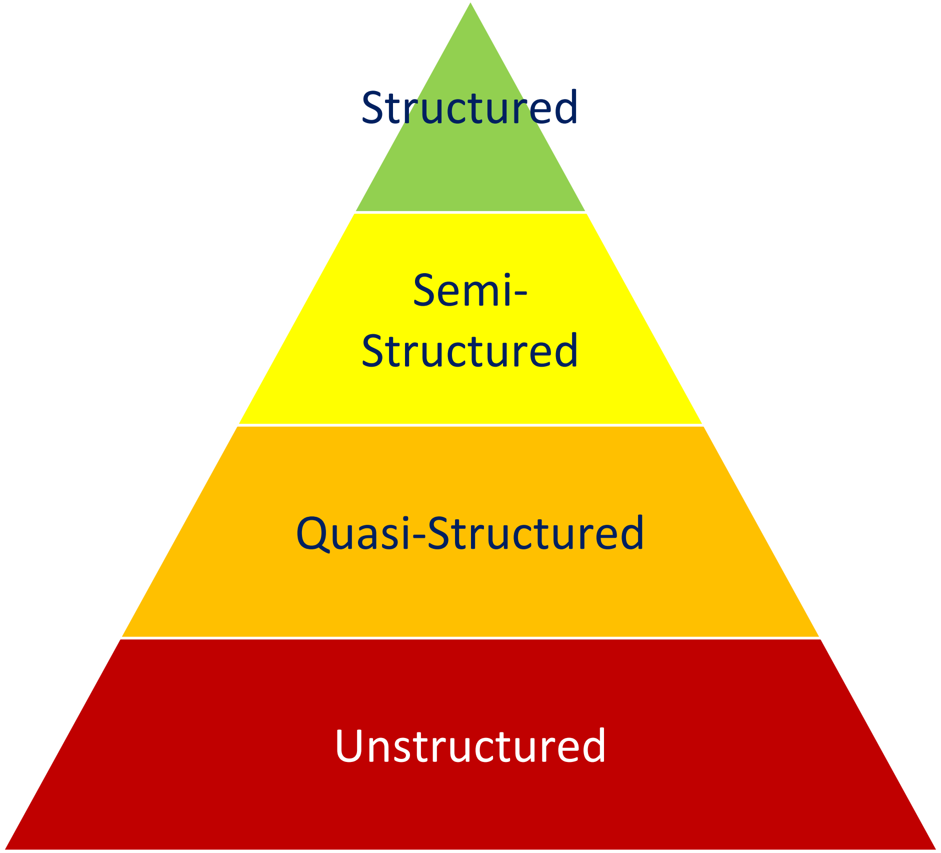
The 3 Vs: Velocity

- Speed at which new data is created
- Speed at which data must be processed and analyzed
  - Often close to real-time



The 3 Vs: Variety

- Diversity in data types and data sources



- Data with defined types and structure
- Example: comma separated values

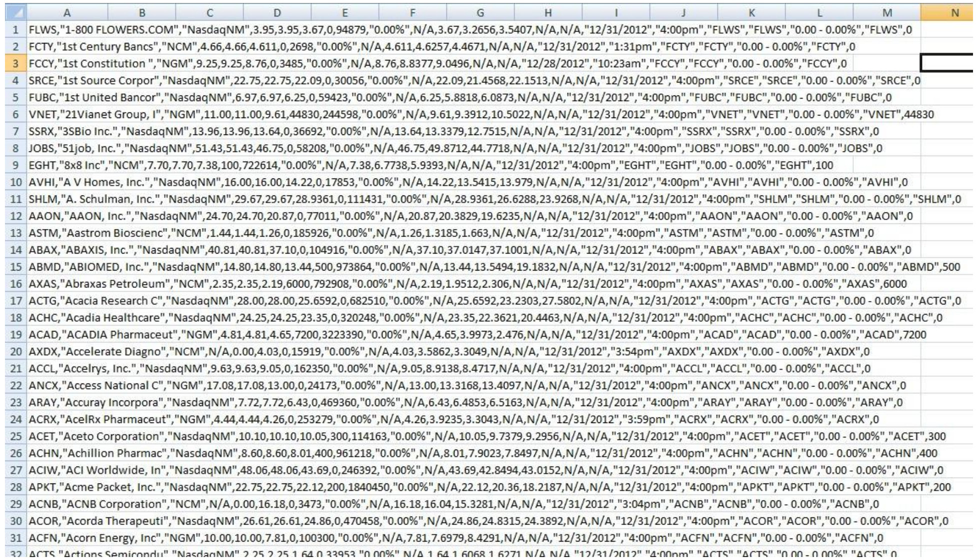
- Textual data with parseable pattern
- Example: XML files with schema

- Textual data with erratic formats that can be formated with effort
- Example: Clickstream data

- Data that has no inherent structure, often with multiple formats
- Example: Web site, videos

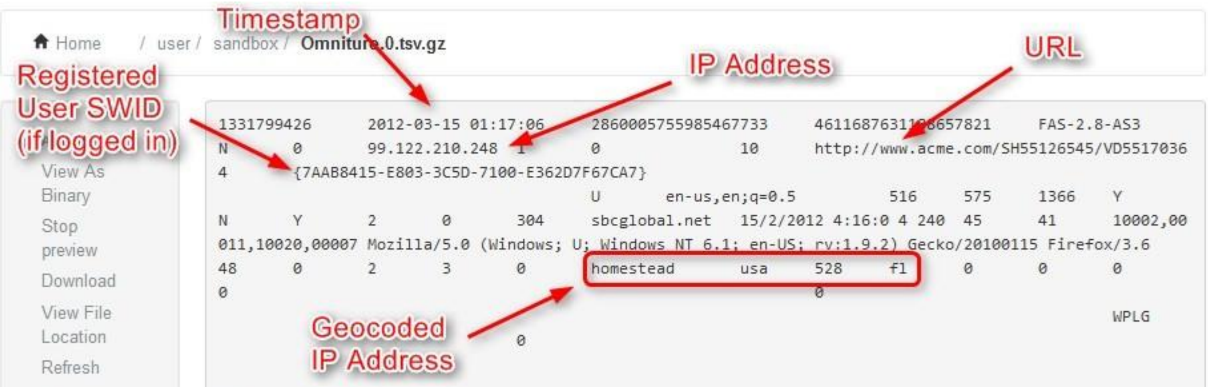
Examples for data types

Structured



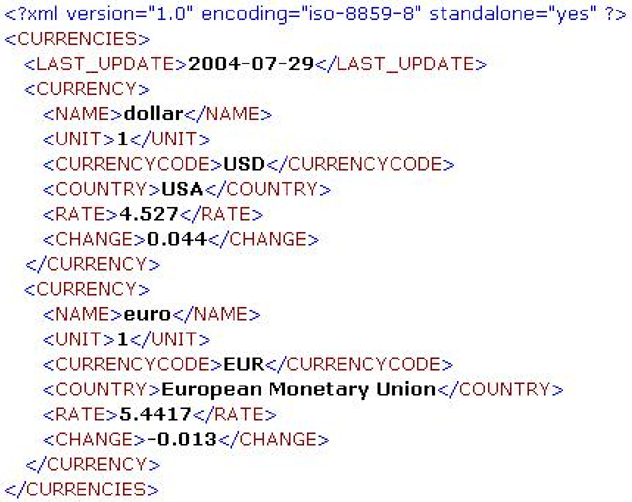
```
1  FLWS,"1-800 FLOWERS.COM","NasdaqNM",3.95,3.95,3.67,0,94879,"0.00%",N/A,3.67,3.2656,3.5407,N/A,N/A,"12/31/2012","4:00pm","FLWS","FLWS","0.00 - 0.00%","FLWS",0
2  FCTY,"1st Century Bancs","NCM",4.66,4.66,4.611,0,2698,"0.00%",N/A,4.611,4.6257,4.4671,N/A,N/A,"12/31/2012","1:31pm","FCTY","FCTY","0.00 - 0.00%","FCTY",0
3  FCCY,"1st Constitution ","NGM",9.25,9.25,8.76,0,3485,"0.00%",N/A,8.76,8.8377,9.0496,N/A,N/A,"12/28/2012","10:23am","FCCY","FCCY","0.00 - 0.00%","FCCY",0
4  SRCE,"1st Source Corpor","NasdaqNM",22.75,22.75,22.09,0,30056,"0.00%",N/A,22.09,21.4568,22.1513,N/A,N/A,"12/31/2012","4:00pm","SRCE","SRCE","0.00 - 0.00%","SRCE",0
5  FUBC,"1st United Bancor","NasdaqNM",6.97,6.97,6.25,0,59423,"0.00%",N/A,6.25,5.8818,6.0873,N/A,N/A,"12/31/2012","4:00pm","FUBC","FUBC","0.00 - 0.00%","FUBC",0
6  VNET,"21Vianet Group, I","NGM",11.00,11.00,9.61,44830,244598,"0.00%",N/A,9.61,9.3912,10.5022,N/A,N/A,"12/31/2012","4:00pm","VNET","VNET","0.00 - 0.00%","VNET",44830
7  SSRX,"3SBio Inc.","NasdaqNM",13.96,13.96,13.64,0,36692,"0.00%",N/A,13.64,13.3379,12.7515,N/A,N/A,"12/31/2012","4:00pm","SSRX","SSRX","0.00 - 0.00%","SSRX",0
8  JOBS,"51job, Inc.","NasdaqNM",51.43,51.43,46.75,0,58208,"0.00%",N/A,46.75,49.8712,44.7718,N/A,N/A,"12/31/2012","4:00pm","JOBS","JOBS","0.00 - 0.00%","JOBS",0
9  EGHT,"8x8 Inc","NCM",7.70,7.70,7.38,100,722614,"0.00%",N/A,7.38,6.7738,5.9393,N/A,N/A,"12/31/2012","4:00pm","EGHT","EGHT","0.00 - 0.00%","EGHT",100
10 AVHI,"A V Homes, Inc.","NasdaqNM",16.00,16.00,14.22,0,17853,"0.00%",N/A,14.22,13.5415,13.979,N/A,N/A,"12/31/2012","4:00pm","AVHI","AVHI","0.00 - 0.00%","AVHI",0
11 SHLM,"A. Schulman, Inc.","NasdaqNM",29.67,29.67,28.9361,0,111431,"0.00%",N/A,28.9361,26.6288,23.9268,N/A,N/A,"12/31/2012","4:00pm","SHLM","SHLM","0.00 - 0.00%","SHLM",0
12 AAON,"AAON, Inc.","NasdaqNM",24.70,24.70,20.87,0,77011,"0.00%",N/A,20.87,20.3829,19.6235,N/A,N/A,"12/31/2012","4:00pm","AAON","AAON","0.00 - 0.00%","AAON",0
13 ASTM,"Aastrom Bioscienc","NCM",1.44,1.44,1.26,0,185926,"0.00%",N/A,1.26,1.3185,1.663,N/A,N/A,"12/31/2012","4:00pm","ASTM","ASTM","0.00 - 0.00%","ASTM",0
14 ABAX,"ABAXIS, Inc.","NasdaqNM",40.81,40.81,37.10,0,104916,"0.00%",N/A,37.10,37.0147,37.1001,N/A,N/A,"12/31/2012","4:00pm","ABAX","ABAX","0.00 - 0.00%","ABAX",0
15 ABMD,"ABIOMED, Inc.","NasdaqNM",14.80,14.80,13.44,500,973864,"0.00%",N/A,13.44,13.5494,19.1832,N/A,N/A,"12/31/2012","4:00pm","ABMD","ABMD","0.00 - 0.00%","ABMD",500
16 AXAS,"Abraxas Petroleum","NCM",2.35,2.35,2.19,6000,792908,"0.00%",N/A,2.19,1.9512,2.306,N/A,N/A,"12/31/2012","4:00pm","AXAS","AXAS","0.00 - 0.00%","AXAS",6000
17 ACTG,"Acacia Research C","NasdaqNM",28.00,28.00,25.6592,0,682510,"0.00%",N/A,25.6592,23.2303,27.5802,N/A,N/A,"12/31/2012","4:00pm","ACTG","ACTG","0.00 - 0.00%","ACTG",0
18 ACHC,"Acadia Healthcare","NasdaqNM",24.25,24.25,23.35,0,320248,"0.00%",N/A,23.35,22.3621,20.4463,N/A,N/A,"12/31/2012","4:00pm","ACHC","ACHC","0.00 - 0.00%","ACHC",0
19 ACAD,"ACADIA Pharmaceut","NGM",4.81,4.81,4.65,7200,3223390,"0.00%",N/A,4.65,3.9973,2.476,N/A,N/A,"12/31/2012","4:00pm","ACAD","ACAD","0.00 - 0.00%","ACAD",7200
20 AXDX,"Accelerate Diagno","NCM",N/A,0.00,4.03,0,15919,"0.00%",N/A,4.03,3.5862,3.3049,N/A,N/A,"12/31/2012","3:54pm","AXDX","AXDX","0.00 - 0.00%","AXDX",0
21 ACCL,"Accelrys, Inc.","NasdaqNM",9.63,9.63,9.05,0,162350,"0.00%",N/A,9.05,8.9138,8.4717,N/A,N/A,"12/31/2012","4:00pm","ACCL","ACCL","0.00 - 0.00%","ACCL",0
22 ANCX,"Access National C","NGM",17.08,17.08,13.00,0,24173,"0.00%",N/A,13.00,13.3168,13.4097,N/A,N/A,"12/31/2012","4:00pm","ANCX","ANCX","0.00 - 0.00%","ANCX",0
23 ARAY,"Accuray Incorpora","NasdaqNM",7.72,7.72,6.43,0,469360,"0.00%",N/A,6.43,6.4853,6.5163,N/A,N/A,"12/31/2012","4:00pm","ARAY","ARAY","0.00 - 0.00%","ARAY",0
24 ACRX,"AcelRx Pharmaceut","NGM",4.44,4.44,4.26,0,253279,"0.00%",N/A,4.26,3.9235,3.3043,N/A,N/A,"12/31/2012","3:59pm","ACRX","ACRX","0.00 - 0.00%","ACRX",0
25 ACET,"Aceto Corporation","NasdaqNM",10.10,10.10,10.05,300,114163,"0.00%",N/A,10.05,9.7379,9.2956,N/A,N/A,"12/31/2012","4:00pm","ACET","ACET","0.00 - 0.00%","ACET",300
26 ACHN,"Achillion Pharmac","NasdaqNM",8.60,8.60,8.01,400,961218,"0.00%",N/A,8.01,7.9023,7.8497,N/A,N/A,"12/31/2012","4:00pm","ACHN","ACHN","0.00 - 0.00%","ACHN",400
27 ACIW,"ACI Worldwide, In","NasdaqNM",48.06,48.06,43.69,0,246392,"0.00%",N/A,43.69,42.8494,43.0152,N/A,N/A,"12/31/2012","4:00pm","ACIW","ACIW","0.00 - 0.00%","ACIW",0
28 APKT,"Acme Packet, Inc.","NasdaqNM",22.75,22.75,22.12,200,1840450,"0.00%",N/A,22.12,20.36,18.2187,N/A,N/A,"12/31/2012","4:00pm","APKT","APKT","0.00 - 0.00%","APKT",200
29 ACNB,"ACNB Corporation","NCM",N/A,0.00,16.18,0,3473,"0.00%",N/A,16.18,16.04,15.3281,N/A,N/A,"12/31/2012","3:04pm","ACNB","ACNB","0.00 - 0.00%","ACNB",0
30 ACOR,"Acorda Therapeuti","NasdaqNM",26.61,26.61,24.86,0,470458,"0.00%",N/A,24.86,24.8315,24.3892,N/A,N/A,"12/31/2012","4:00pm","ACOR","ACOR","0.00 - 0.00%","ACOR",0
31 ACFN,"Acorn Energy, Inc","NGM",10.00,10.00,7.81,0,100300,"0.00%",N/A,7.81,7.6979,8.4291,N/A,N/A,"12/31/2012","4:00pm","ACFN","ACFN","0.00 - 0.00%","ACFN",0
32 ACTS,"Actions Semicondu","NasdaqNM",2.25,2.25,1.64,0,33953,"0.00%",N/A,1.64,1.6068,1.6271,N/A,N/A,"12/31/2012","4:00pm","ACTS","ACTS","0.00 - 0.00%","ACTS",0
```

Quasi-Structured



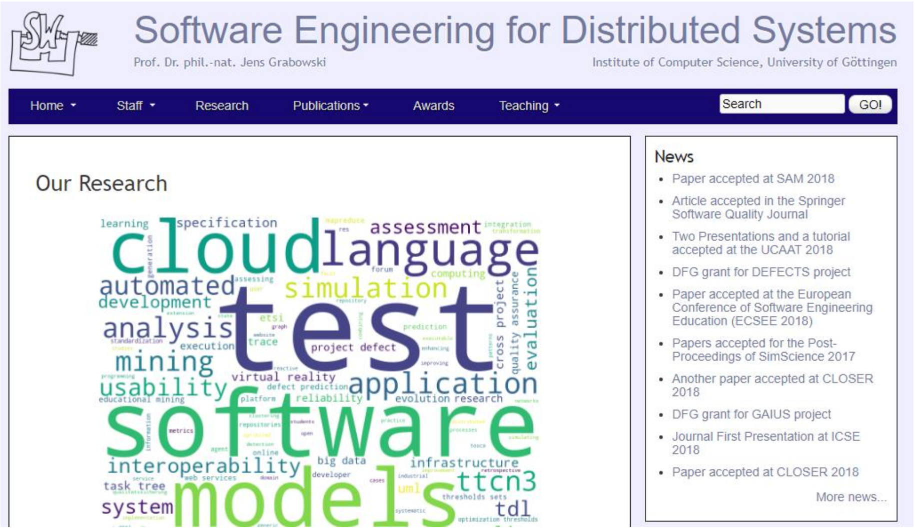
```
1331799426      2012-03-15 01:17:06     2860005755985467733     4611687631108657821     FAS-2.8-AS3
N       0       99.122.210.248  1       0               10      http://www.acme.com/SH55126545/VD5517036
4       {7AAB8415-E803-3C5D-7100-E362D7F67CA7}
                                        U       en-us,en;q=0.5          516     575     1366    Y
N       Y       2       0       304     sbcglobal.net   15/2/2012 4:16:0 4 240  45      41      10002,00
011,10020,00007 Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6
48      0       2       3       0       homestead       usa     528     fl      0       0       0
0                                                               0
                                                                                        WPLG
                                0
```

Semi-Structured



```
<?xml version="1.0" encoding="iso-8859-8" standalone="yes" ?>
<CURRENCIES>
  <LAST_UPDATE>2004-07-29</LAST_UPDATE>
  <CURRENCY>
    <NAME>dollar</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>USD</CURRENCYCODE>
    <COUNTRY>USA</COUNTRY>
    <RATE>4.527</RATE>
    <CHANGE>0.044</CHANGE>
  </CURRENCY>
  <CURRENCY>
    <NAME>euro</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>EUR</CURRENCYCODE>
    <COUNTRY>European Monetary Union</COUNTRY>
    <RATE>5.4417</RATE>
    <CHANGE>-0.013</CHANGE>
  </CURRENCY>
</CURRENCIES>
```

Unstructured



Software Engineering for Distributed Systems

Prof. Dr. phil.-nat. Jens Grabowski

Institute of Computer Science, University of Göttingen

Our Research

News

- Paper accepted at SAM 2018
- Article accepted in the Springer Software Quality Journal
- Two Presentations and a tutorial accepted at the UCAAT 2018
- DFG grant for DEFECTS project
- Paper accepted at the European Conference of Software Engineering Education (ECSEE 2018)
- Papers accepted for the Post-Proceedings of SimScience 2017
- Another paper accepted at CLOSER 2018
- DFG grant for GAIUS project
- Journal First Presentation at ICSE 2018
- Paper accepted at CLOSER 2018

More news...

Outline

- Introduction to Big Data

- Data Science definition

- The Skillset of Data Scientists
- AI
- ML

- Summary

Defining Data Science

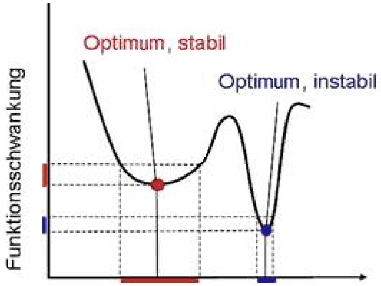
- Unfortunately, there is no clear definition (yet?)
- Goal is the extraction of knowledge from data
- Combination of techniques from different disciplines
- Scientific principles guide the data analysis

What is „Data Science“?!?

Tools? Big Data?
Machine Learning?

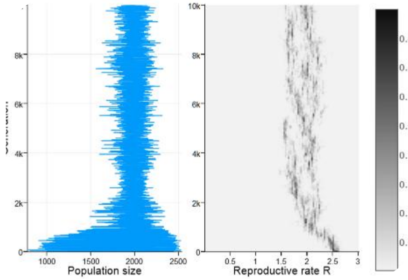
Mathematical Aspects

Computational Geometry



Optimization

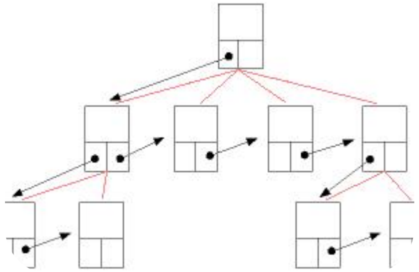
Stochastics



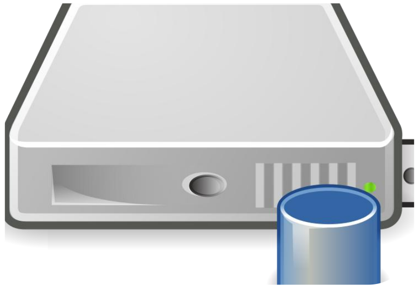
Scientific Computing

Machine Learning

Computer Science Aspects

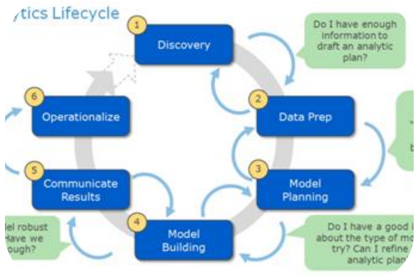


Data Structures and Algorithms

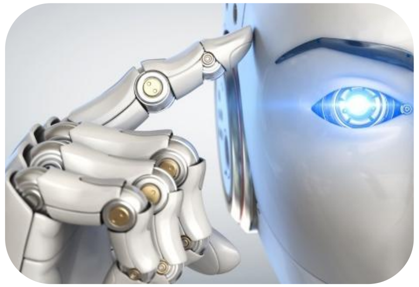


Databases

Distributed Computing



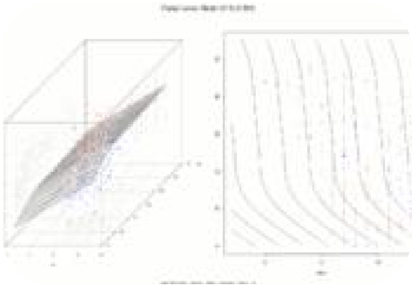
Software Engineering



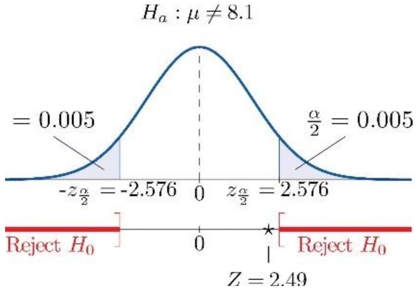
Artificial Intelligence

Machine Learning

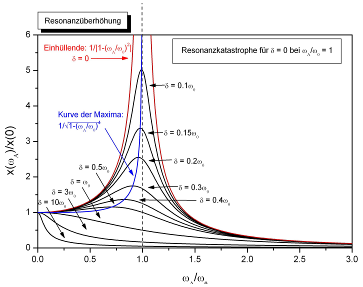
Statistical Aspects



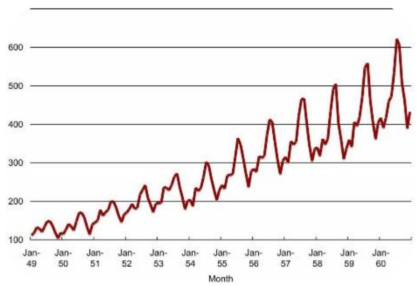
Linear Models



Statistical Tests



Inference



Time Series Analysis

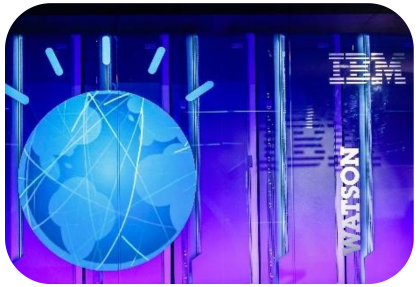
Machine Learning

Applications

Intelligent Systems

Robotics

Marketing



Medicine

Autonomous Driving

Social Networks

Outline

- Introduction to Big Data
- Data Science definition
- The Skillset of Data Scientists
- AI
- ML
- Summary

What are Data Scientists?

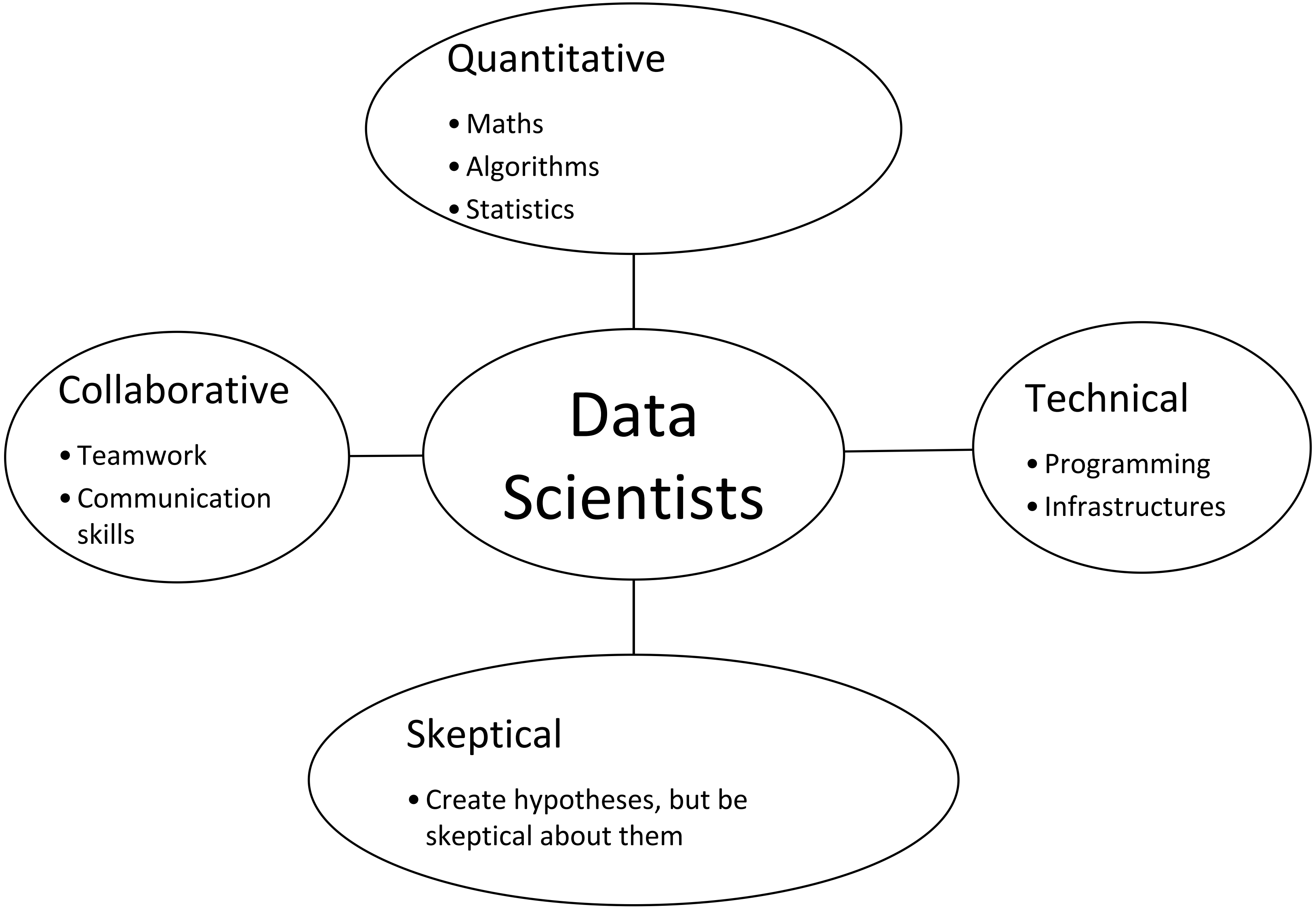
- Not computer scientists
    - But should know about databases, data structures, algorithms, etc.

- Not mathematicians
    - But should know about optimization, stochastics, etc.

- Not statisticians
    - But should know about regression, statistical tests, etc.

- Not domain experts
    - But must work together with them

Skills of Data Scientists



Data Scientists

Quantitative

- Maths
- Algorithms
- Statistics

Collaborative

- Teamwork
- Communication skills

Technical

- Programming
- Infrastructures

Skeptical

- Create hypotheses, but be skeptical about them

A bit of everything

… but actually as much as possible of everything

Different types of Data Scientists

- According to Microsoft Research:
  - Polymath
    - „Do it all“
  - Data Evangelist
    - Data analysis, disseminating and acting on insights
  - Data Preparer
    - Querying existing data, preparing data for analysis
  - Data Shapers
    - Analyzing and preparing data
  - Data Analyzer
    - Analyzing data
  - Platform Builder
    - Collect data and create infrastructures
  - Moonlighters (50%/20%)
    - „Spare time“ data scientists
  - Insight Actors
    - Use the outcome and act on insights.

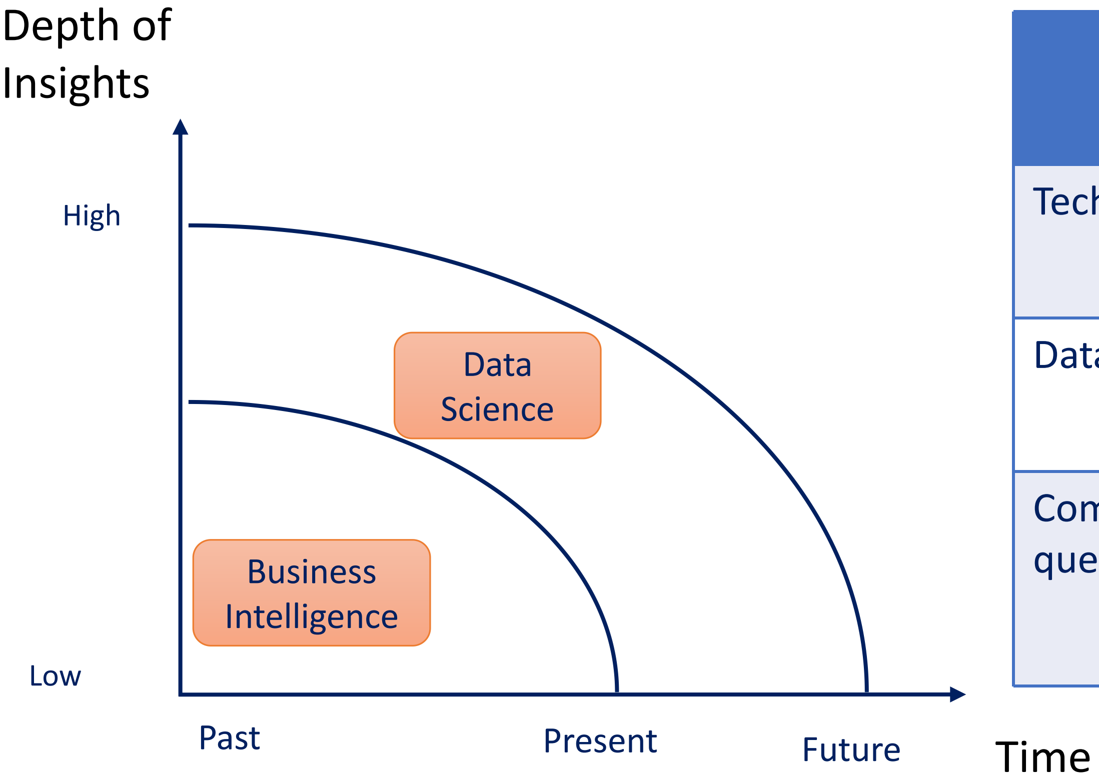
Miyung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel: Data Scientists in Software Teams: State of the Art and Challenges, IEEE Transactions on Software Engineering (Online First)

Data Science Definition

- Data science is a multidisciplinary field that uses scientific methods, algorithms, processes, and systems to extract valuable insights and knowledge from structured and unstructured data.
- It combines elements of statistics, computer science, domain knowledge, and data visualization to analyze large and complex datasets, uncover patterns, make predictions, and inform decision-making.

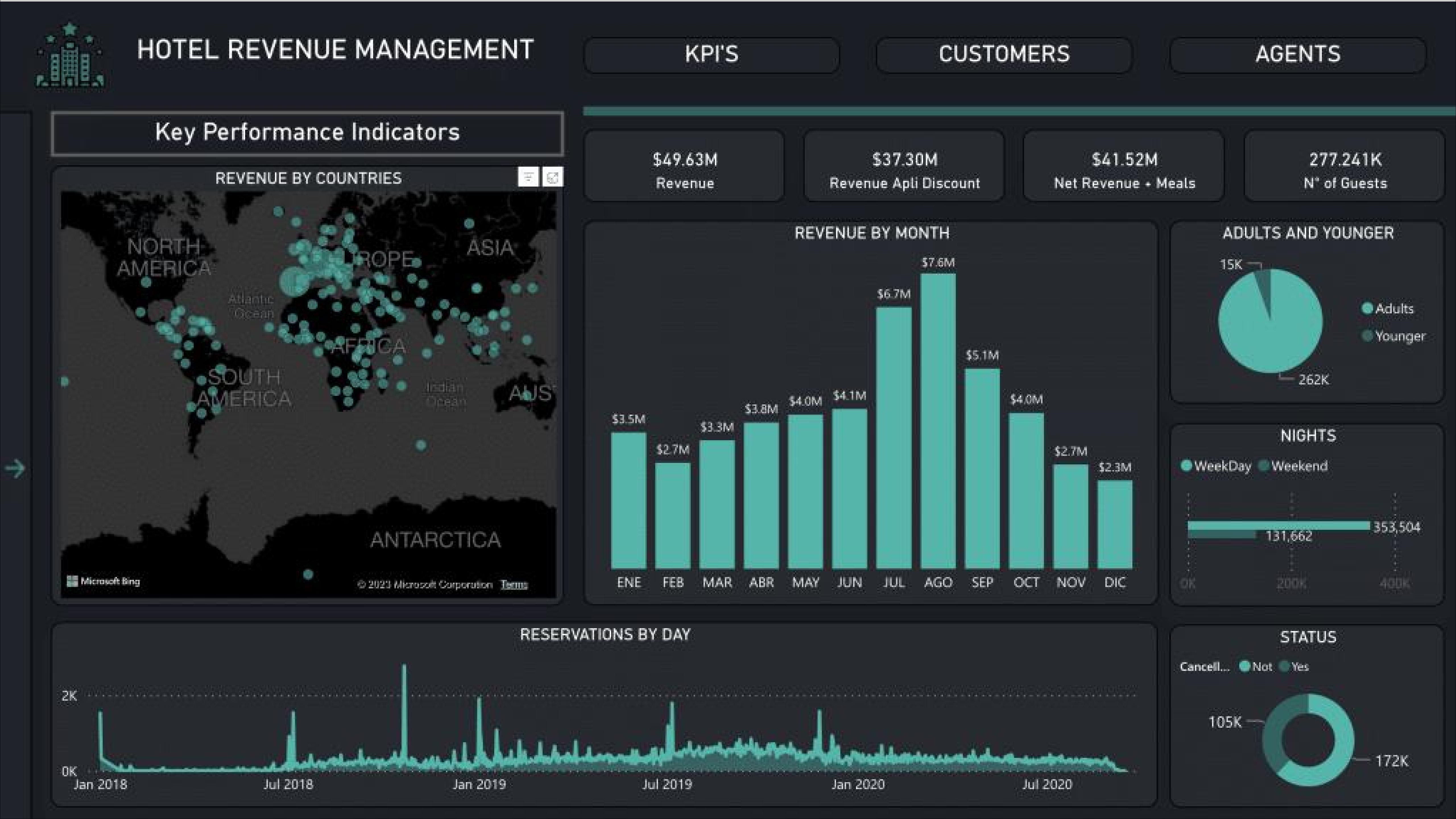
Data Science vs. Business Intelligence

- Business Intelligence (Gartner IT Glossary)
  - […] best practices that enable access to and analysis of information to improve and optimize decisions and performance.



	Business Intelligence	Data Science
Techniques	Dashboards, alerts, queries	Optimization, predictive modelling, forecasting
Data Types	Structured, data warehouses	Any kind, often unstructured
Common questions	What happened…? How much did…? When did…?	What if…? What will…? How can we…?

Dashboard examples



HOTEL REVENUE MANAGEMENT

KPI'S | CUSTOMERS | AGENTS

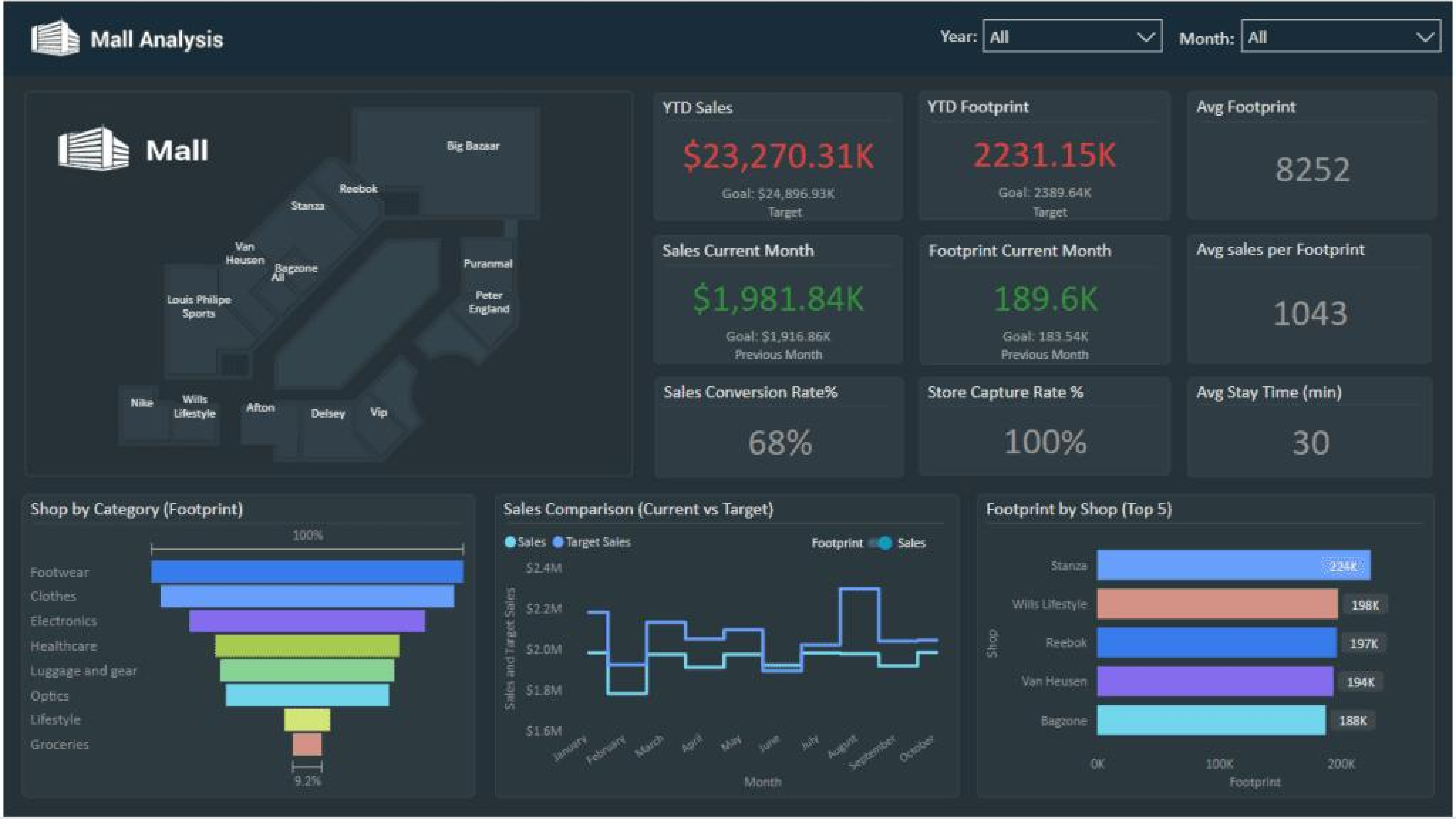
Key Performance Indicators

$49.63M Revenue

$37.30M Revenue Apli Discount

$41.52M Net Revenue + Meals

277.241K N° of Guests



Mall Analysis

Year: All

Month: All

YTD Sales
$23,270.31K
Goal: $24,896.93K
Target

YTD Footprint
2231.15K
Goal: 2389.64K
Target

Avg Footprint
8252

Sales Current Month
$1,981.84K
Goal: $1,916.86K
Previous Month

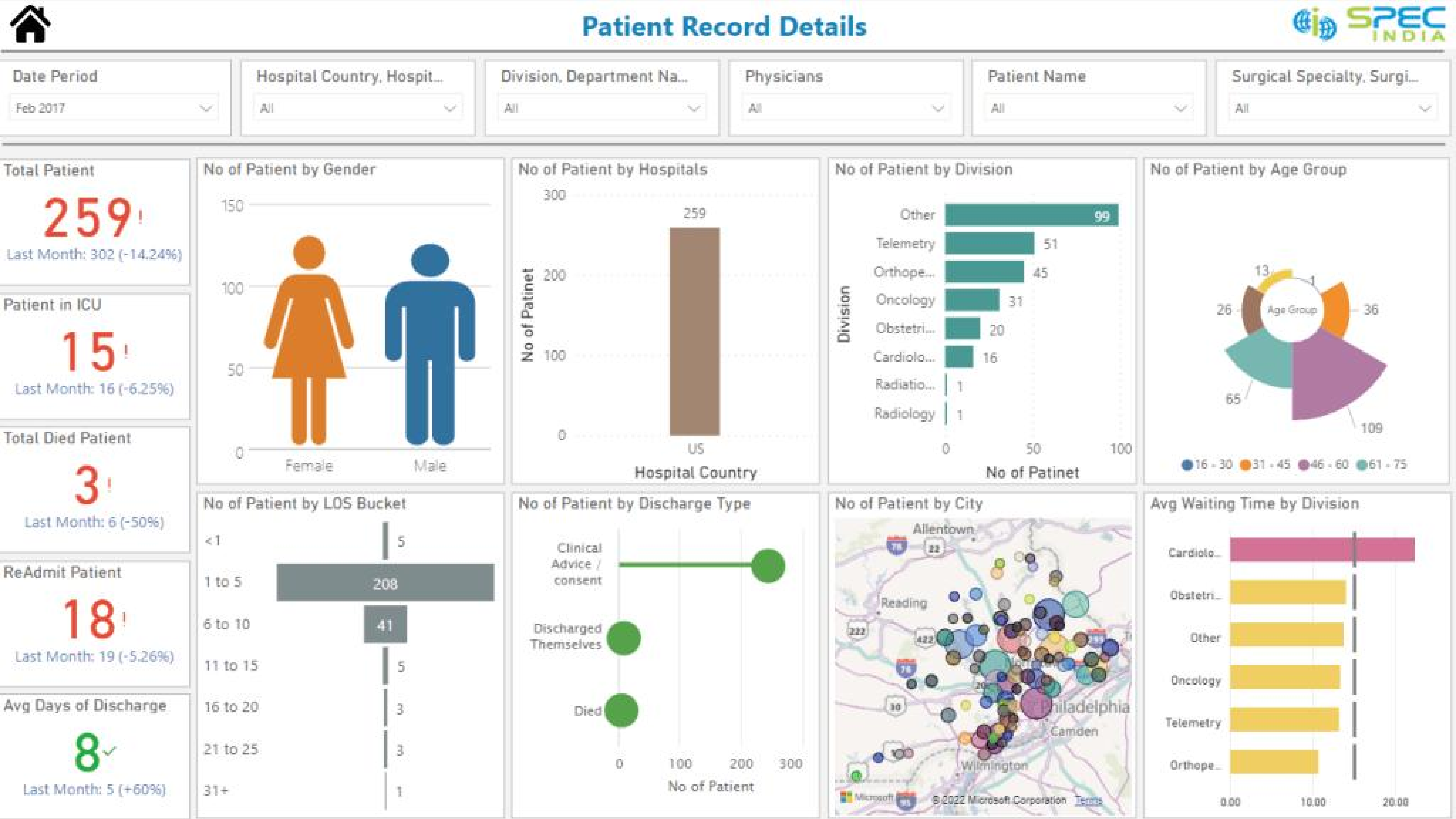
Footprint Current Month
189.6K
Goal: 183.54K
Previous Month

Avg sales per Footprint
1043

Sales Conversion Rate%
68%

Store Capture Rate %
100%

Avg Stay Time (min)
30



Patient Record Details

SPEC INDIA

Date Period: Feb 2017

Hospital Country, Hospit...: All

Division, Department Na...: All

Physicians: All

Patient Name: All

Surgical Specialty, Surgi...: All

Total Patient
259 !
Last Month: 302 (-14.24%)

Patient in ICU
15 !
Last Month: 16 (-6.25%)

Total Died Patient
3 !
Last Month: 6 (-50%)

ReAdmit Patient
18 !
Last Month: 19 (-5.26%)

Avg Days of Discharge
8 ✓
Last Month: 5 (+60%)

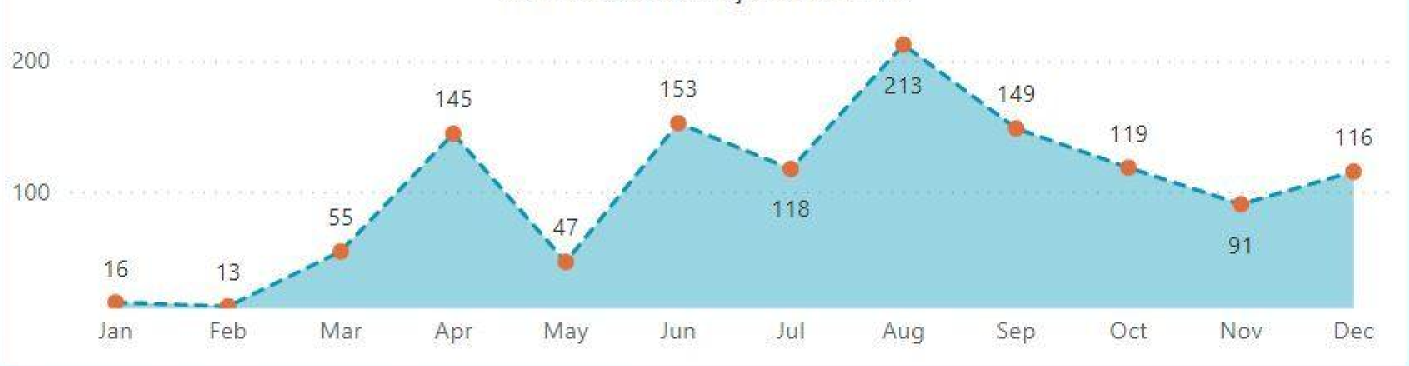
LinkedIn

Kamil.M.S (BI Consultant)

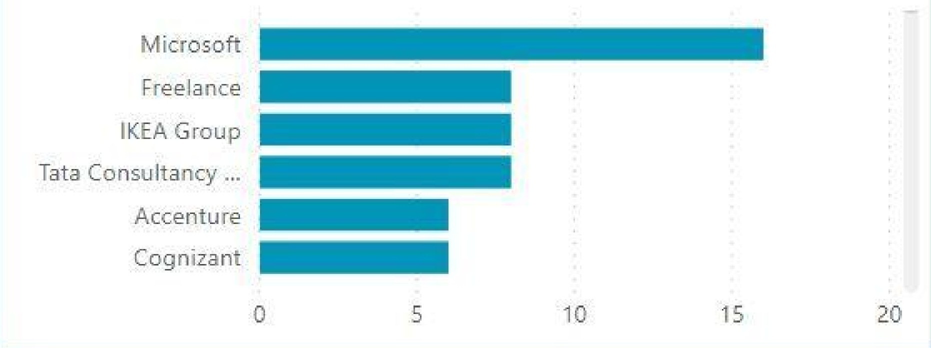
□ 2019
□ 2020

Connections	Companies	Invitations Received	Invitations Sent	Reactions
1235	1086	504	322	2008

Total Connections by Month Name



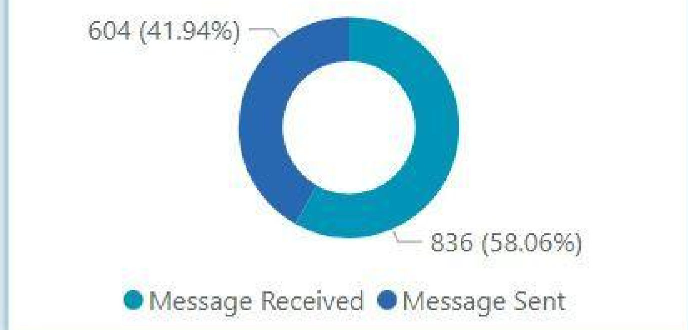
Total Connections by Company



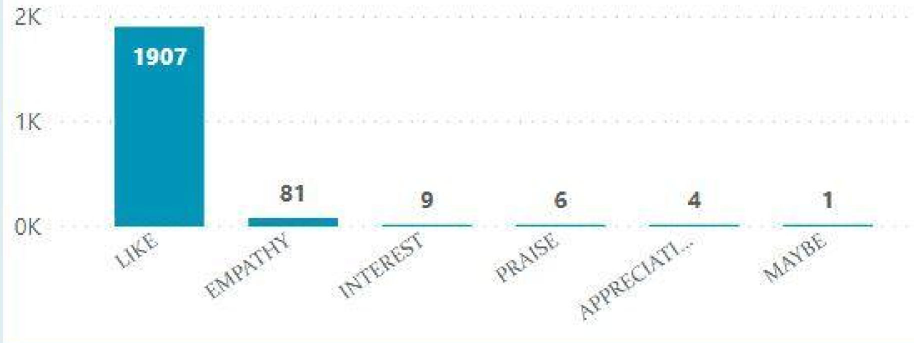
Total Connections by Position

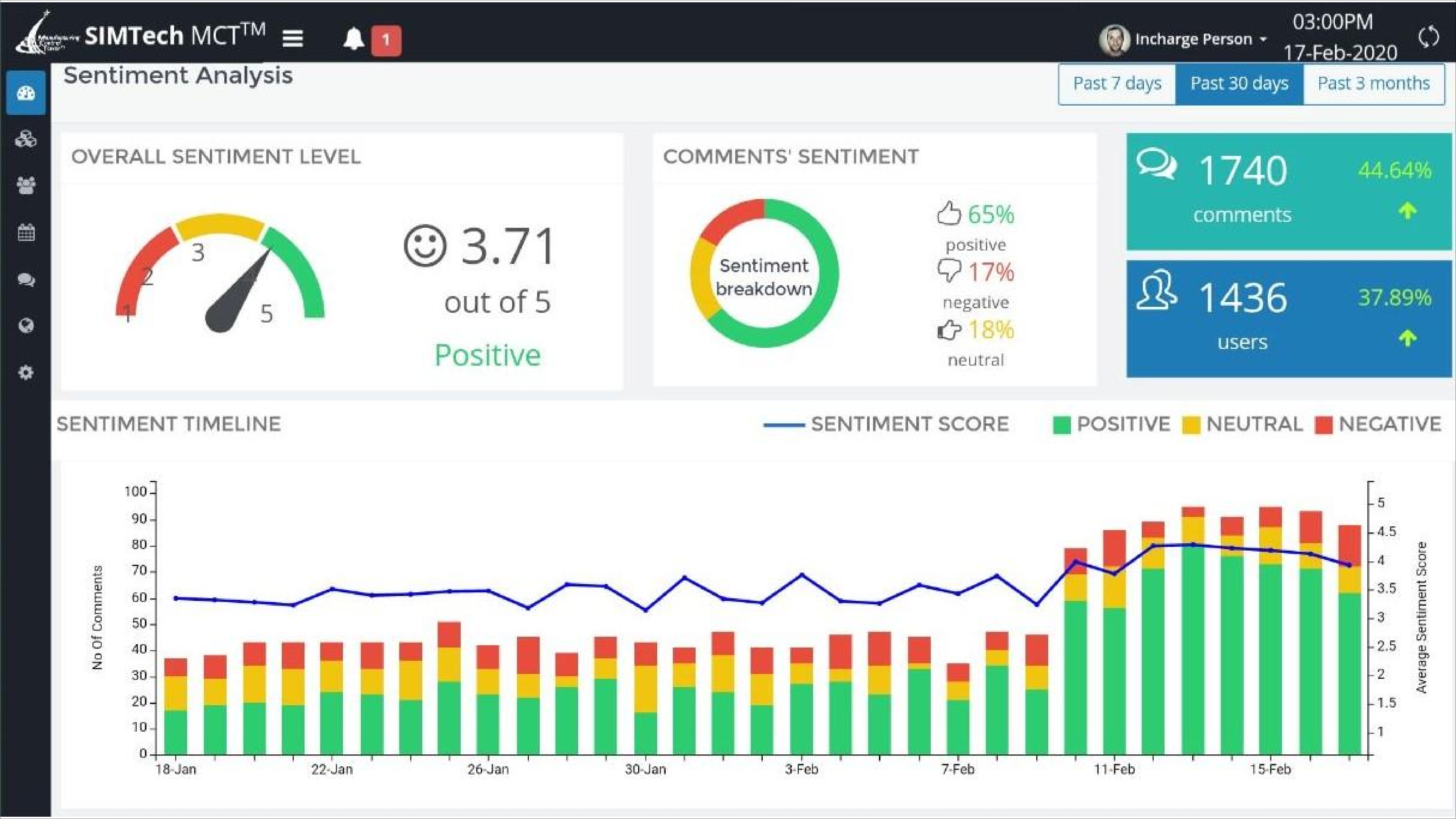
Developer
Manager
Senior
Analyst
Consultant
Business
Data
Intelligence
Engineer
Analytics
Specialist
BI
Recruitment
Power
Director
Architect
Microsoft
Support
Officer

Messages Received and Messages Sent



Reactions by Type





Sentiment Analysis

OVERALL SENTIMENT LEVEL

☺ 3.71

out of 5

Positive

COMMENTS' SENTIMENT

65% positive

17% negative

18% neutral

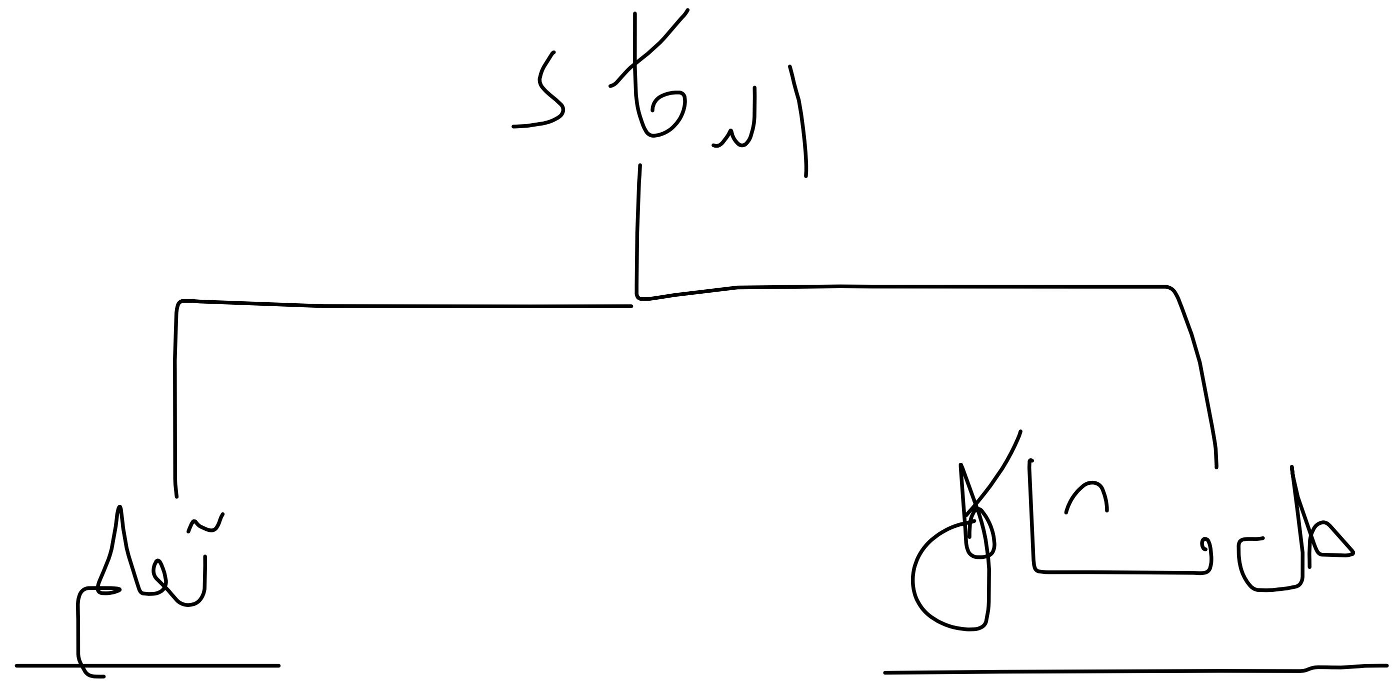
1740 comments 44.64% ↑

1436 users 37.89% ↑

SENTIMENT TIMELINE

AI definition

- Artificial Intelligence (AI) refers to the simulation of human intelligence in machines or computer systems.
- It involves the development of algorithms, software, and hardware that enable computers to perform tasks that typically require human intelligence, such as understanding natural language, recognizing patterns, making decisions, and solving problems.



المواد
مخاليط
نقية

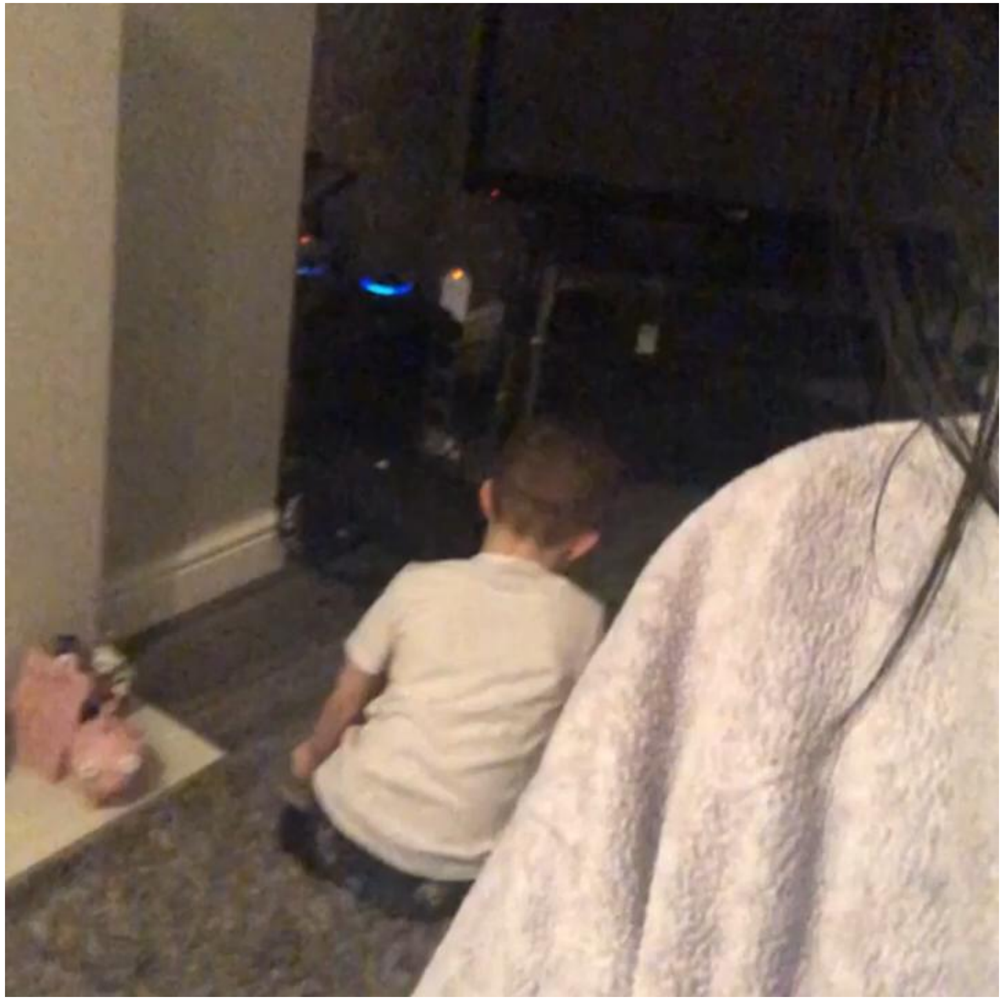
State of the Art: Human vs Robot

clideo.com

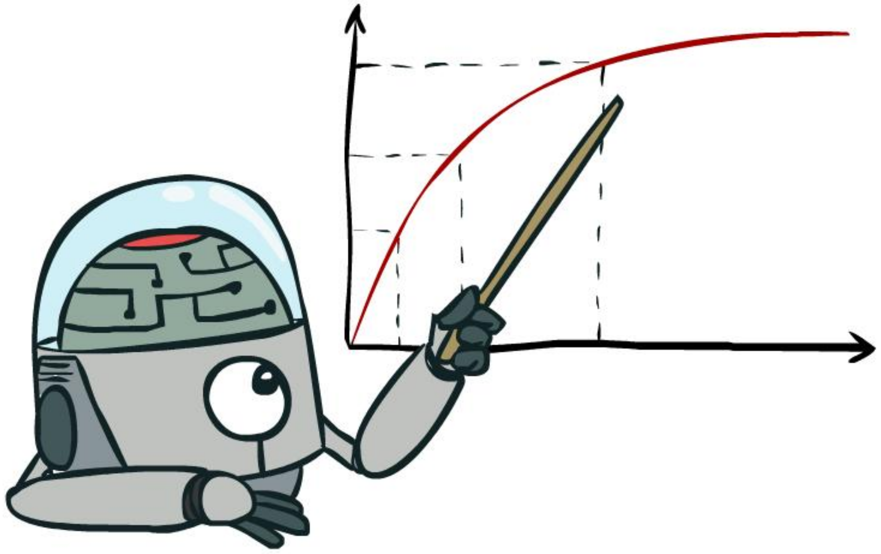
Types of AI

1. Narrow AI (Weak AI): This type of AI is designed for specific tasks or domains. It excels at performing functions, such as voice recognition, image classification, or playing board games like chess or Go. Narrow AI systems do not possess general intelligence or consciousness and are trained for specific applications.
2. General AI (Strong AI): General AI represents machines or systems that possess human-like intelligence, including the ability to understand, learn, and apply knowledge across a wide range of tasks and domains. Achieving general AI is a long-term goal of AI research and development and is yet to be realized.

Alexa helps to solve assignments ☺

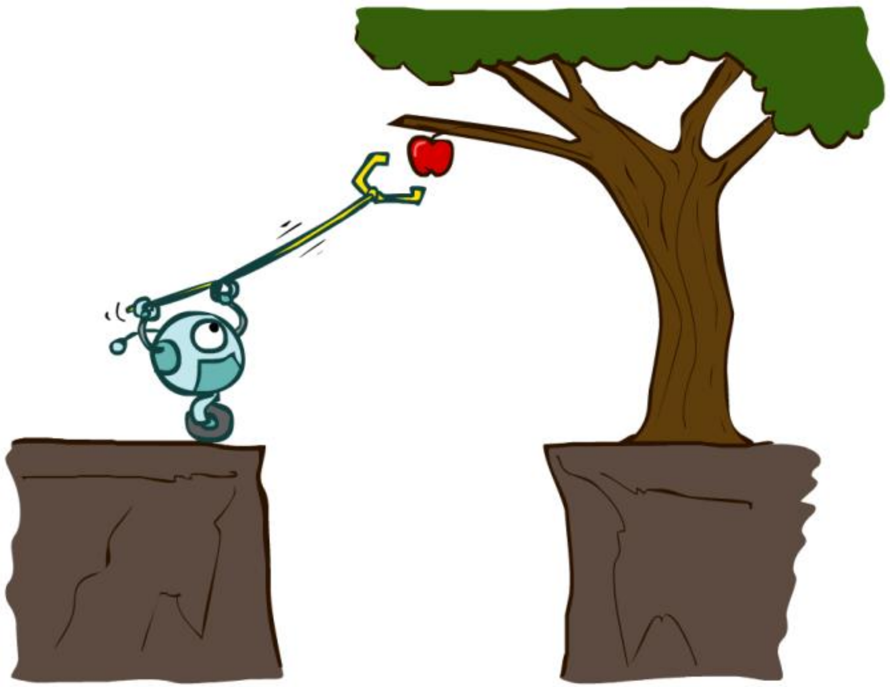


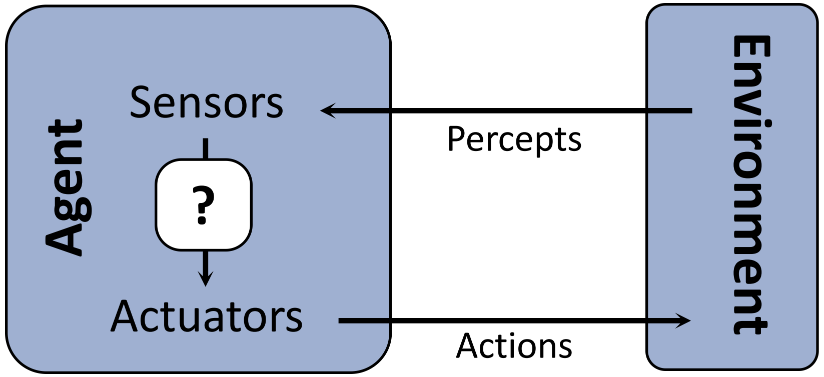
Maximize Your Expected Utility



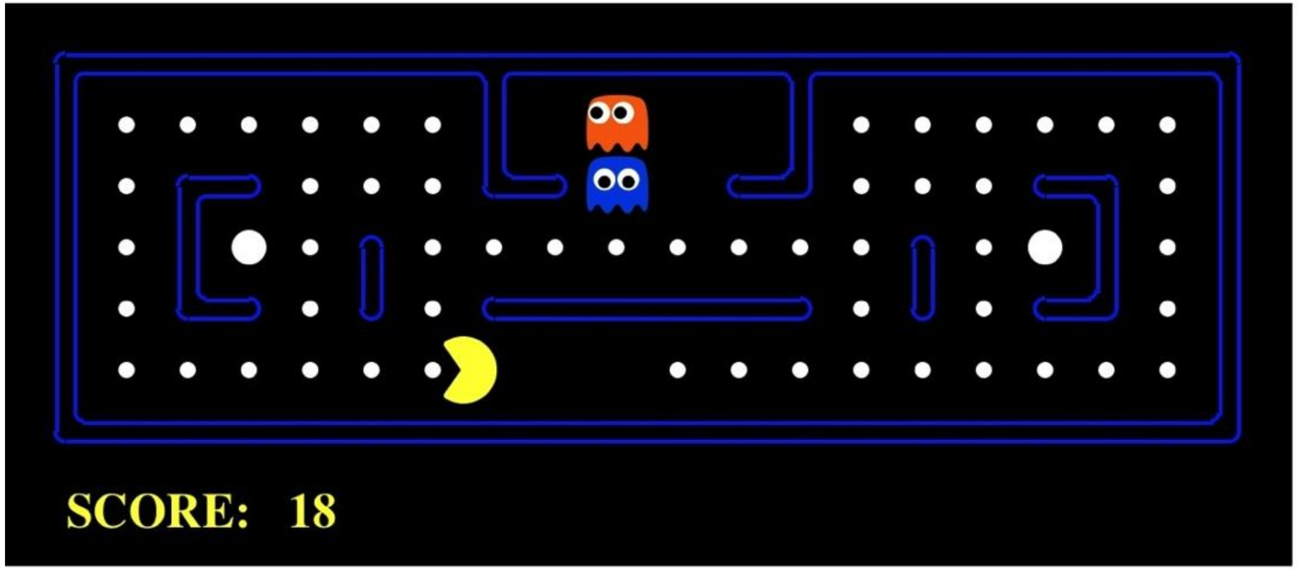
Designing Rational Agents

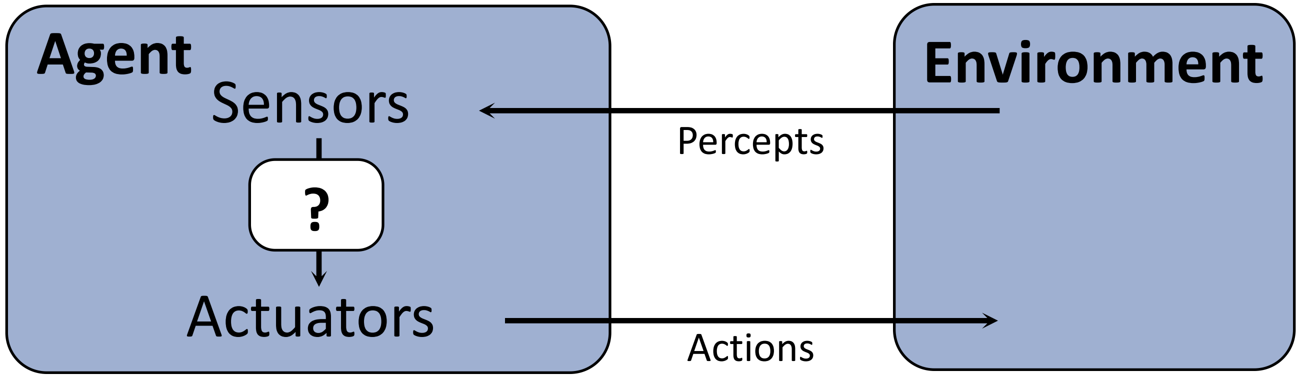
- An **agent** is an entity that *perceives* and *acts*.
- A **rational agent** selects actions that maximize its (expected) **utility**.
- Characteristics of the **percepts, environment,** and **action space** dictate techniques for selecting rational actions





Pac-Man as an Agent





Agent

Sensors

?

Actuators

Percepts

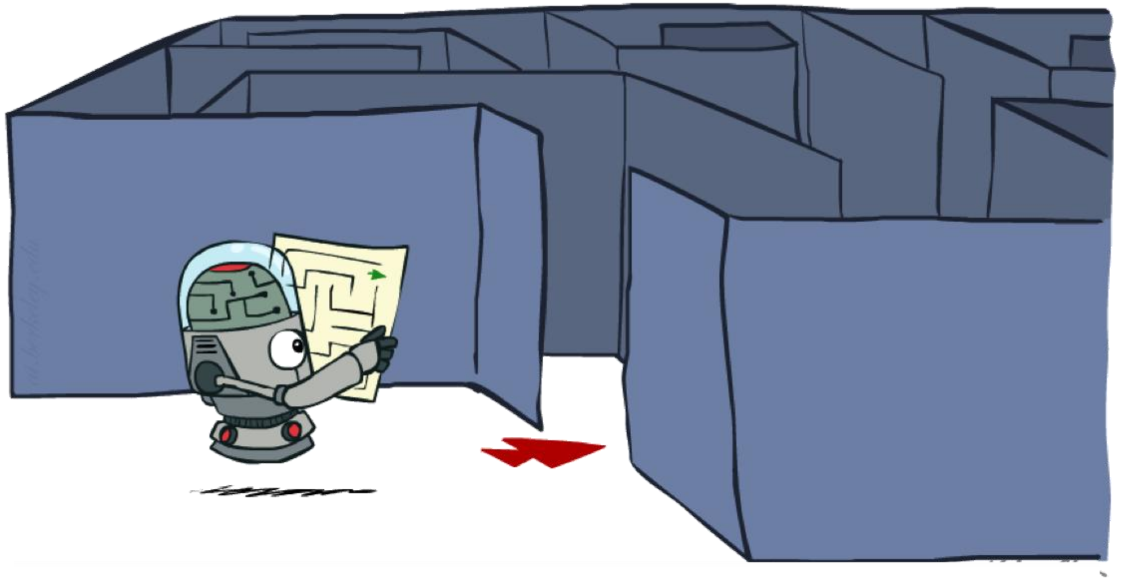
Actions

Environment

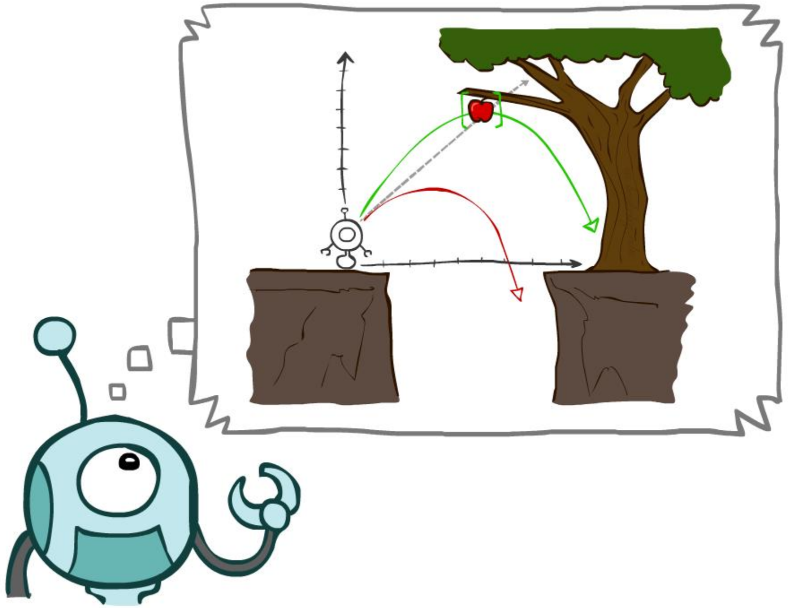
Pac-Man is a registered trademark of Namco-Bandai Games, used here for educational purposes

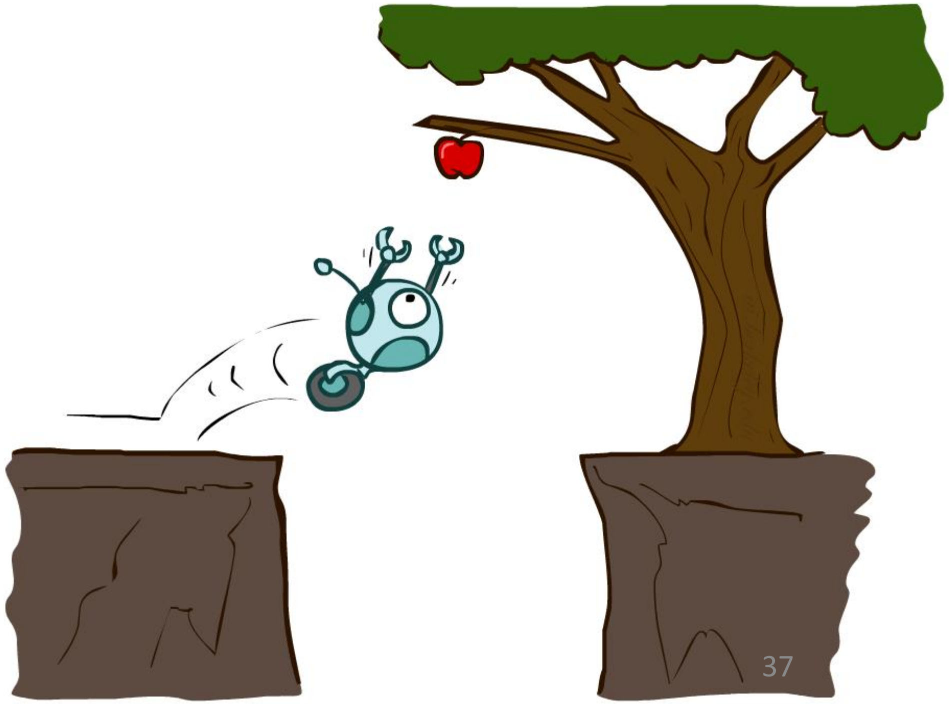
Demo1: pacman-l1.mp4

Search



How do we formulate a search problem?





Machine Learning (ML) definition

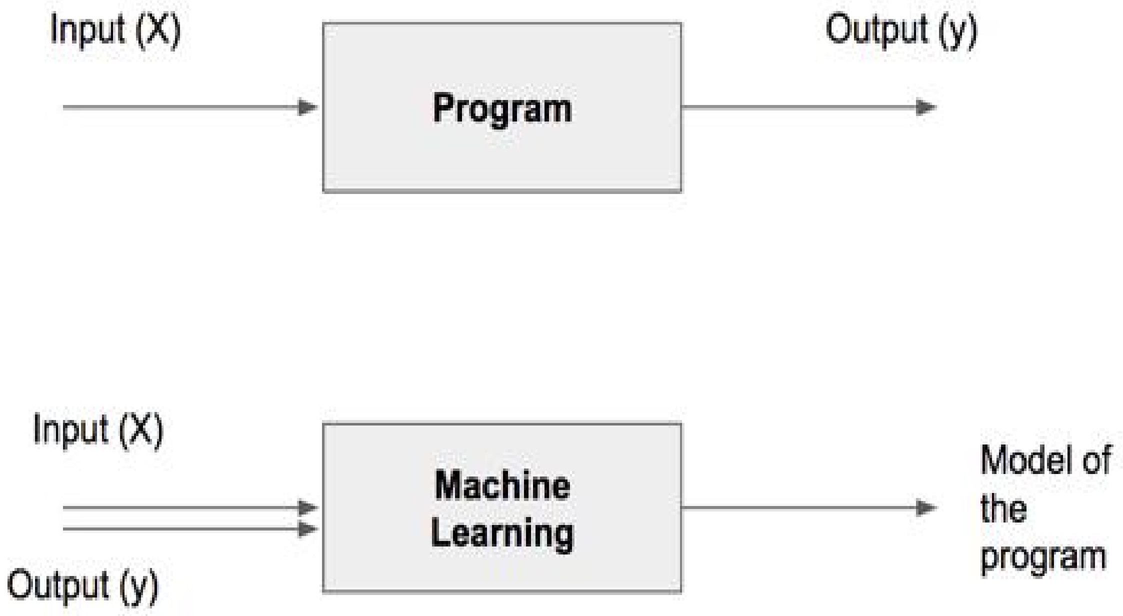
Machine Learning (ML) is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task through learning from data, without being explicitly programmed

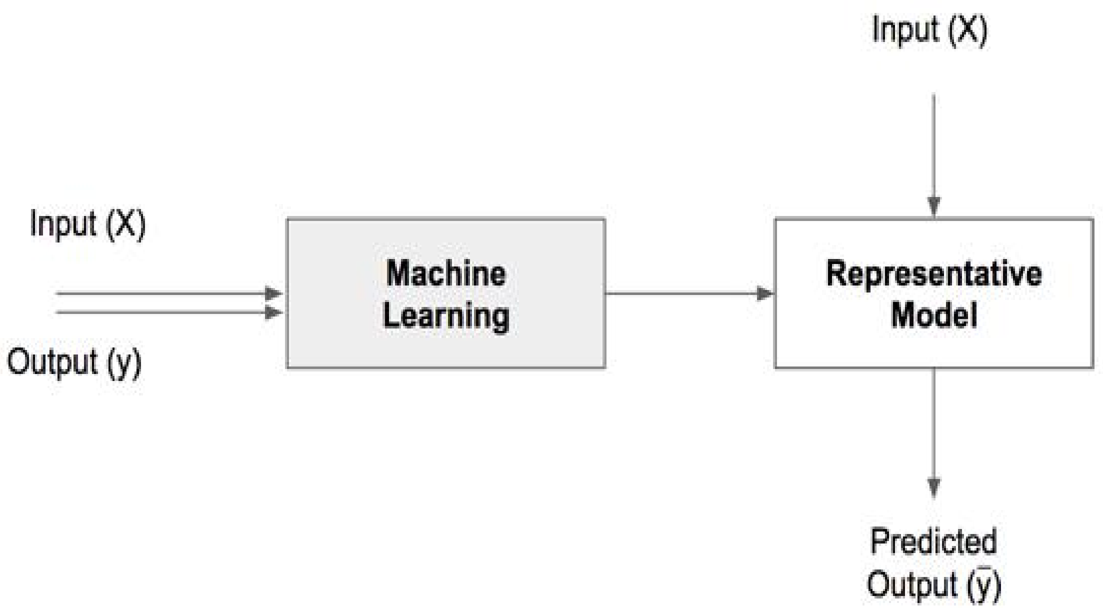
ML allows computers to automatically discover patterns, make predictions, or take actions based on past experiences or examples.

ML vs Traditional Programming

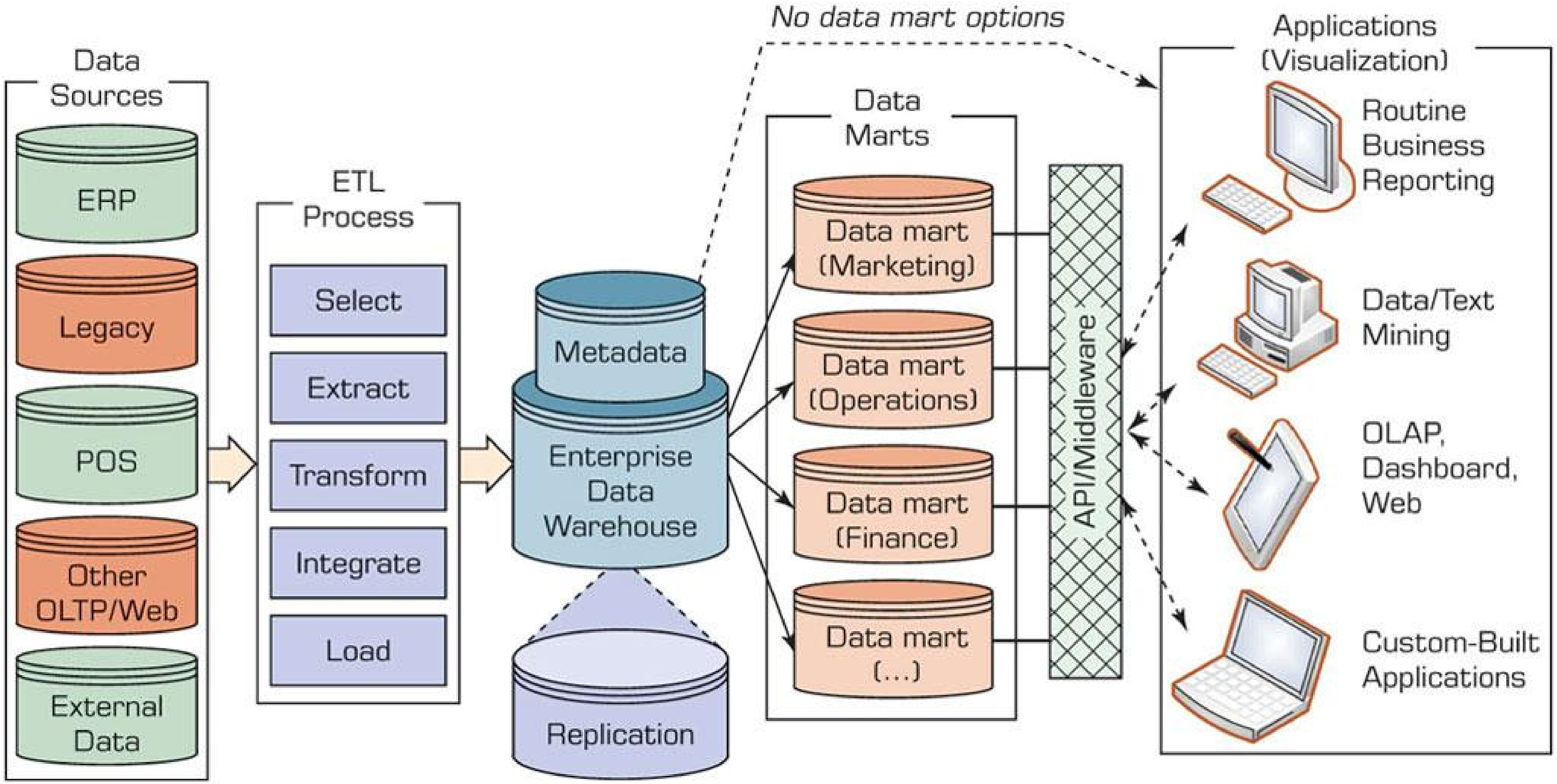
Aspect	Machine Learning (ML)	Traditional Programming
Purpose	Learning from data to make predictions, recognize patterns, and automate decision-making.	Executing predefined instructions and algorithms to achieve specific tasks.
Data-driven	Relies on data for training and learning patterns.	Not data-driven; instructions are explicitly programmed.
Flexibility	Adaptable to changing data and can improve with more examples.	Less adaptable and requires manual code changes for modifications.
Problem Complexity	Suited for complex problems with large datasets or uncertain environments.	Effective for well-defined, deterministic tasks.
Expertise Required	Requires knowledge of data preprocessing, algorithm selection, and model tuning.	Requires expertise in programming languages, algorithms, and problem-solving.
Maintenance	May require periodic retraining and adjustment as data changes.	Maintenance involves debugging, updating, and code optimization.

ML models vs programs





Data Pipeline



Some Data Science Tasks

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density-based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.

Summary

- Big data has a high volume, velocity, and variety

- Different data structures
  - Structured, semi-structured, quasi-structured, unstructured

- Data science is a very diverse discipline
  - Maths, computer science, statistics, applications

→ Data scientists require a diverse skillset