METIS

Intro to Data Science

Learning Objectives & Agenda

METIS

Learning objectives

Be able to

- Describe data science and explain its different facets
- Explain the differences between statistics and machine learning
- Explain the major branches of machine learning and the types of problems they solve
- Describe special topics within data science

Agenda

1. A Brief History of Data Science
2. Basics of Data Science
3. Analytics and Statistics
4. Statistics and Machine Learning
5. Machine Learning and Artificial Intelligence
6. Special Topics
7. Course Structure

A BRIEF HISTORY OF DATA SCIENCE

METIS

Definition

Data science is the practice of extracting useful and actionable information from data, which is then used to create value

This is achieved through a combination of analysis, statistics, machine learning, artificial intelligence, and programming

With these tools, we can use computers to answer questions and achieve results that were previously untenable

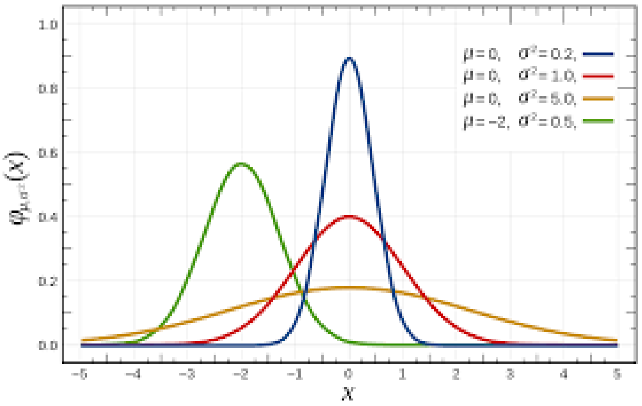
BASICS OF DATA SCIENCE

METIS

Major Components

Analytics: the discovery of patterns in data and their application to decision making

Statistics: branch of mathematics focusing on uncovering meaning in data and randomness

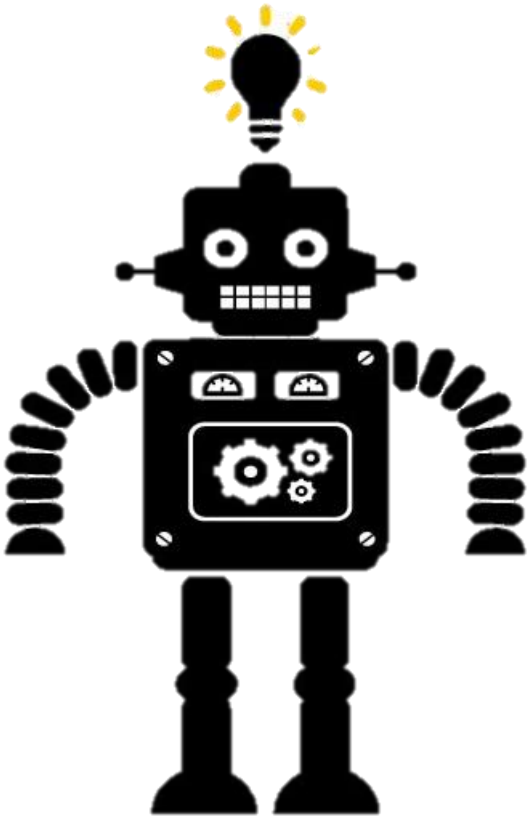

$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \quad \sigma^2=0.5,$
$\varphi_{\mu,\sigma^2}(X)$
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5
X

Major Components

Machine Learning: the study of algorithms and statistical models to improve task performance

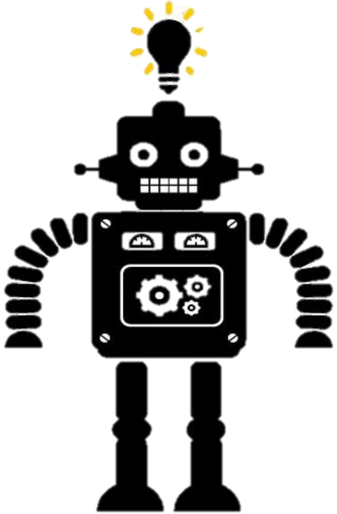
Computer Science: the study of algorithms and computation

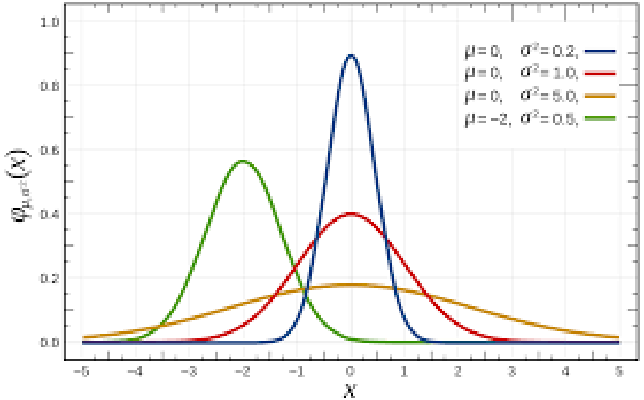
Artificial Intelligence (AI): No agreed upon definition



Major Components

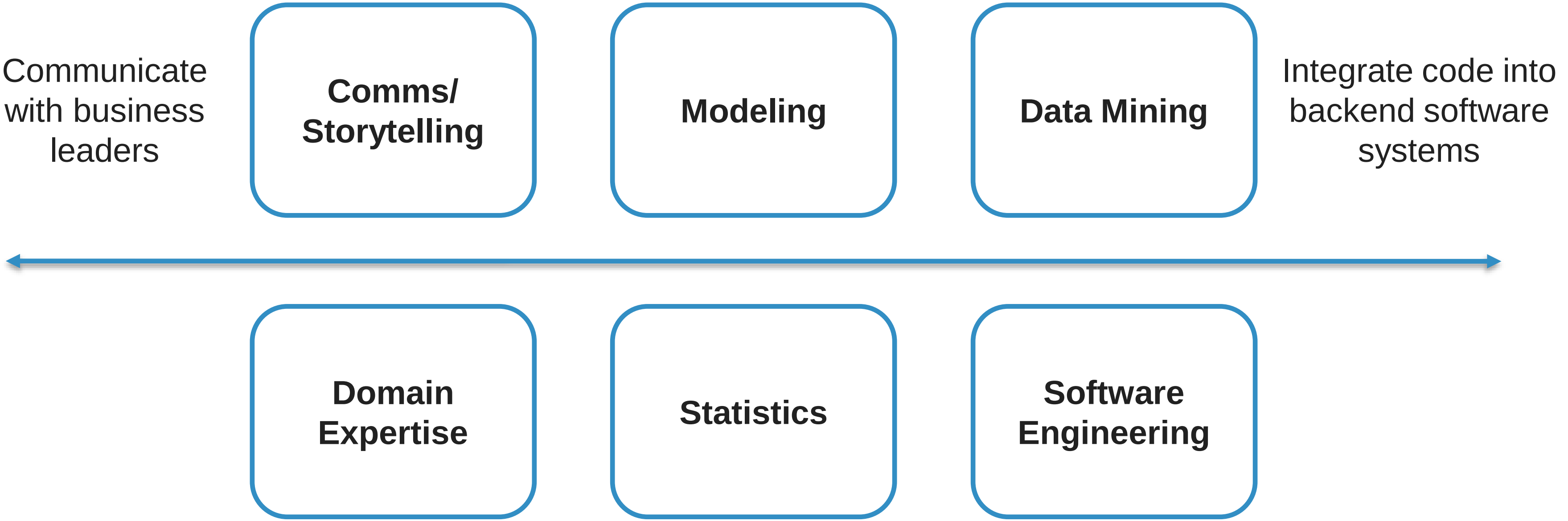
- There is no hard cut line between any of these components
- They cannot stand independent of each other




$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \quad \sigma^2=0.5,$
$\varphi_{\mu,\sigma^2}(x)$
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5
x

Data Science Team Skills

To be successful, data science teams need a variety of skills



Communicate with business leaders

Comms/ Storytelling

Modeling

Data Mining

Integrate code into backend software systems

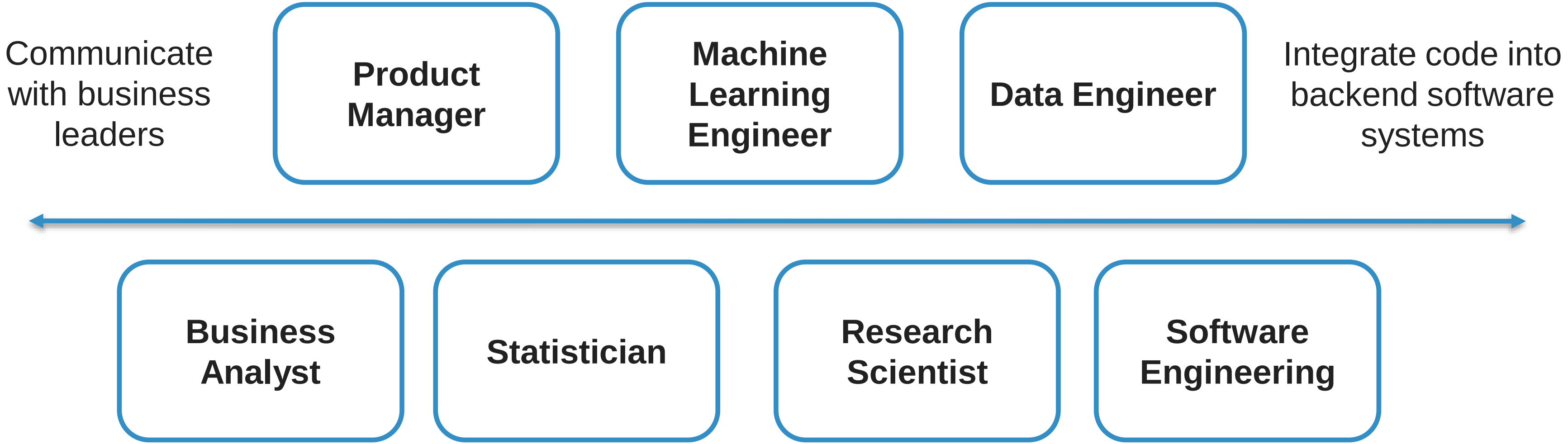
Domain Expertise

Statistics

Software Engineering

Data Science Team Roles

To support the needed skills and achieve impact, data science teams need a diverse set of roles



Communicate with business leaders

Product Manager

Machine Learning Engineer

Data Engineer

Integrate code into backend software systems

Business Analyst

Statistician

Research Scientist

Software Engineering

Data Science Project Workflow

Data science projects have predictable steps, but iterate on and revisit them often

Problem Statement	What problem are you trying to solve?
Data Collection	What data do you need to solve it?
Data Exploration & Preprocessing	Do you understand your data? Will your model?
Modeling	Build a model to solve your problem
Validation	Did I solve the problem?
Decision Making & Deployment	Communicate to stakeholders or put into production

ANALYTICS & STATISTICS

METIS

Types of Analytics Techniques

Descriptive: What *did* happen?

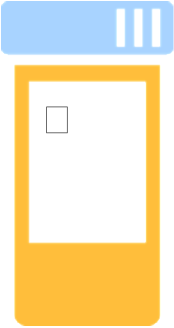
- Mean, median, distribution, max

Predictive: What *will* (likely) happen?

- Stock price prediction, estimated probability of churn

Prescriptive: What *should* we do?

- Pricing, resource allocation



Analytics

Answers direct, clear questions with deterministic answers

Monitors changes in business and informs decision makers

Leans heavily on business rules

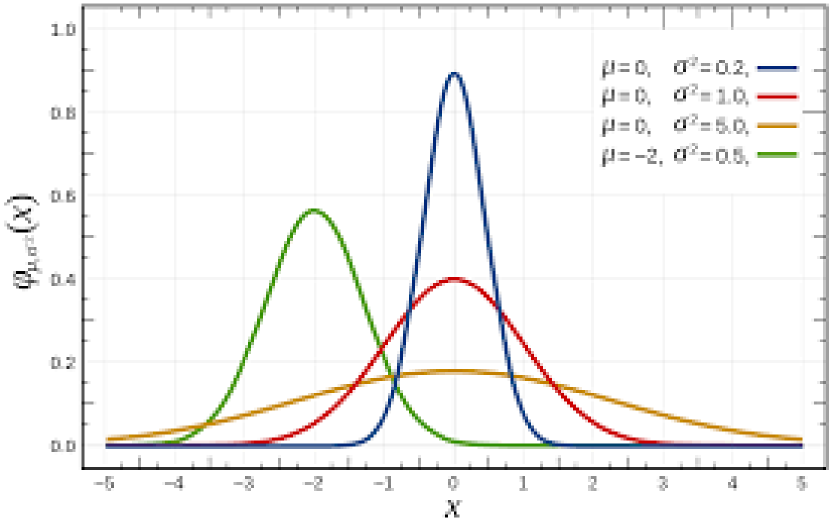
Statistics

A field of mathematics dedicated to interpreting patterns in data and making inferences about them

Two major branches: frequentist (standard) and Bayesian (new & exciting)

Specialized subfields, e.g. time series analysis, experimental design

”Backbone” of modern science

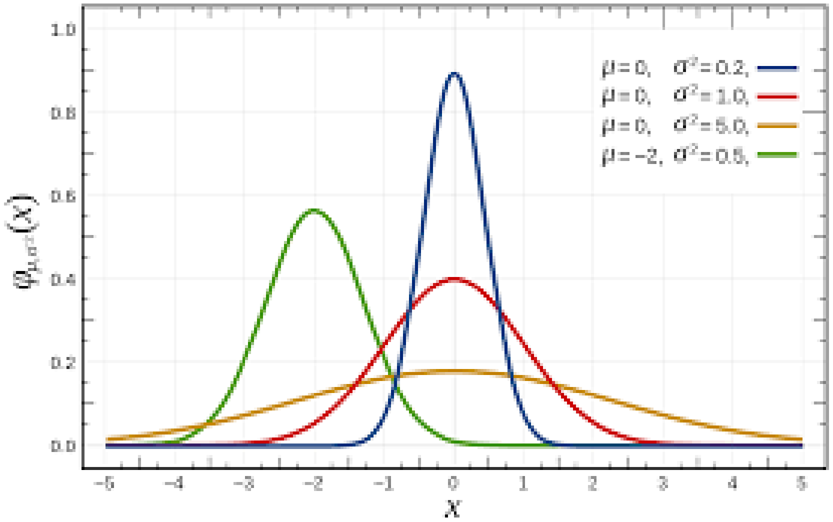


Statistics

Answers descriptive, predictive, and relationship questions

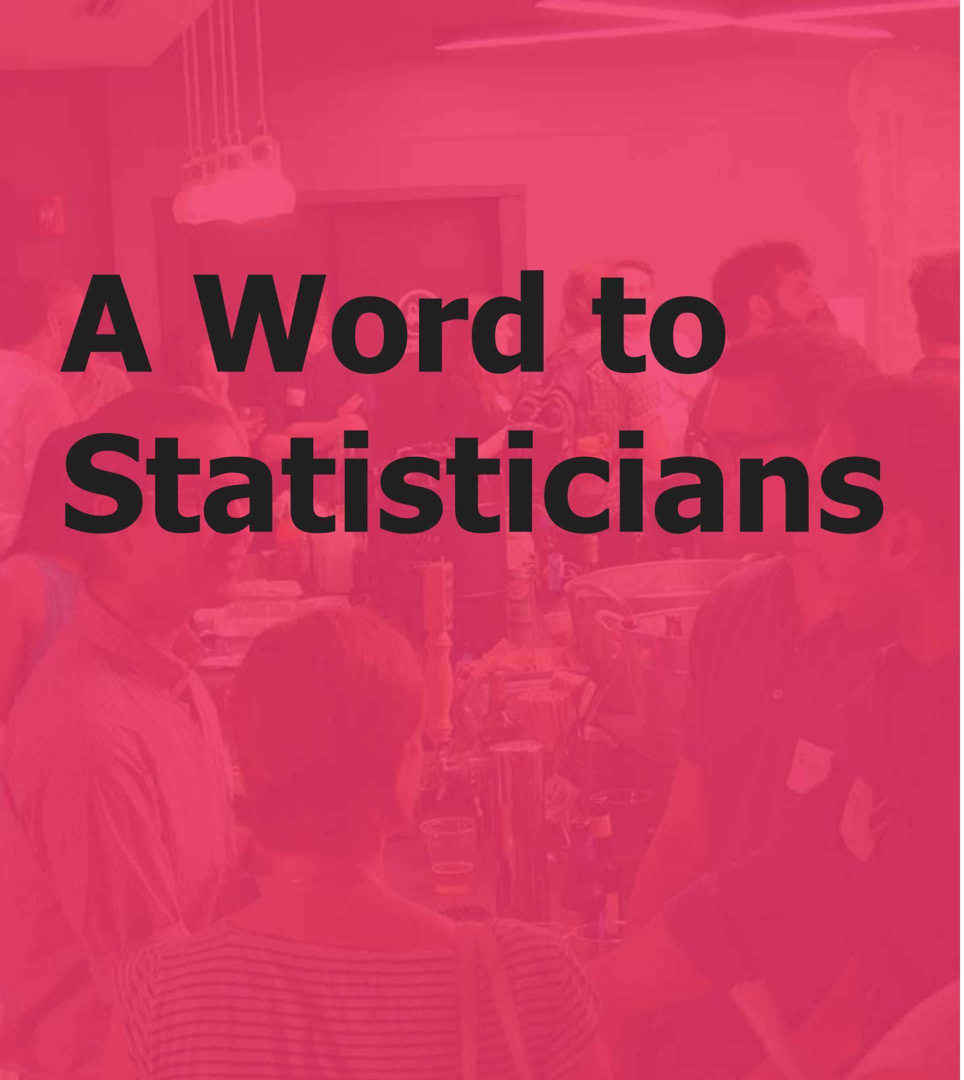
Probability and mathematical guarantees

Concerned with the *distribution* of numbers & metrics



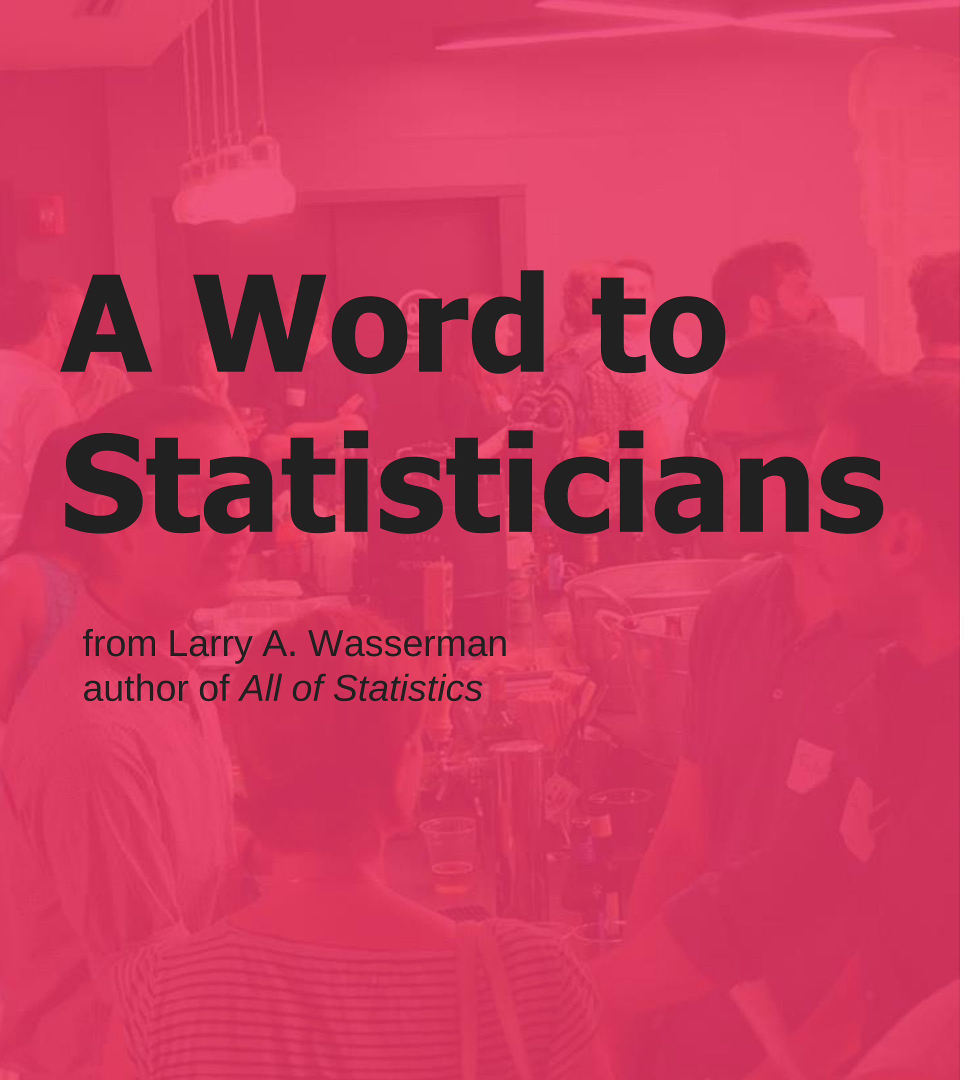
STATISTICS & MACHINE LEARNING

METIS



A Word to Statisticians

All models are wrong,
but **some** are useful"



A Word to Statisticians

from Larry A. Wasserman
author of *All of Statistics*

No Free Lunch .

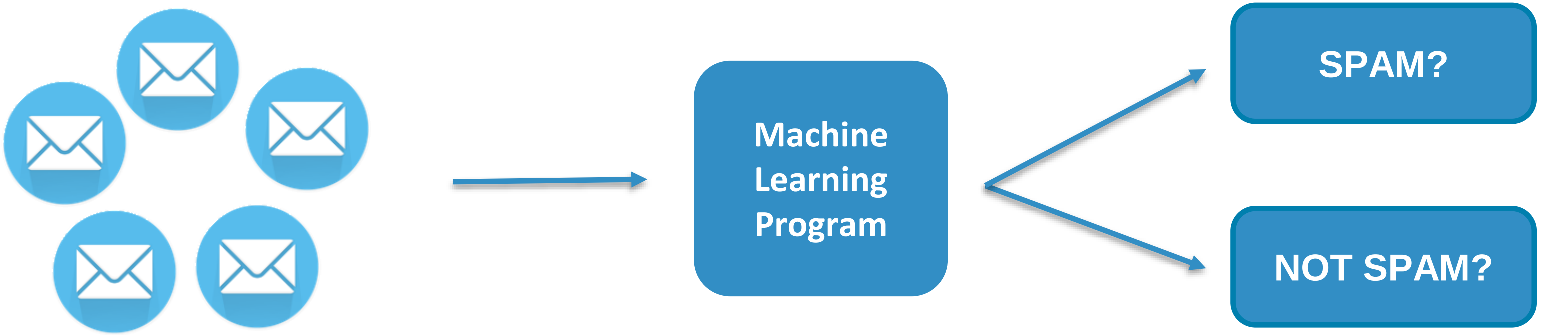
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

METIS

Machine Learning (ML)

Machine learning allows computers to learn and infer from data

These programs learn from repeatedly seeing data, rather than being explicitly programmed by humans



Machine Learning Program

SPAM?

NOT SPAM?

Emails are labeled as spam vs. not

The more emails the program sees…

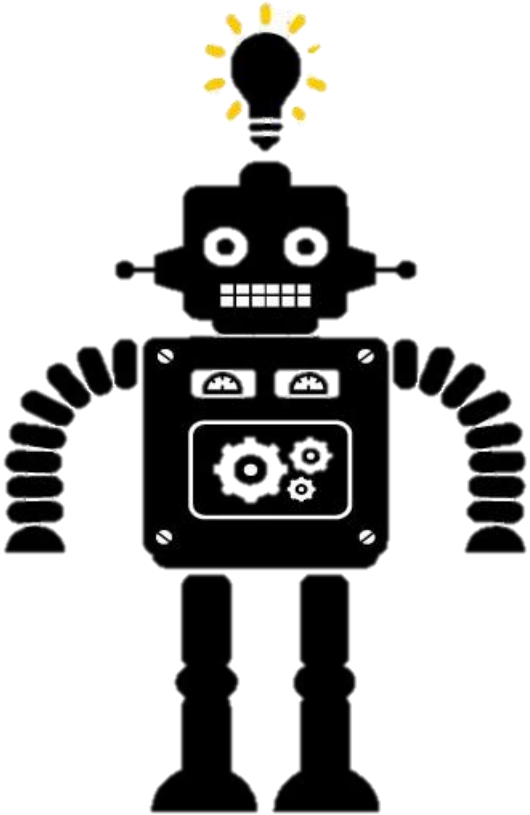
…the better it gets at classification

Machine Learning (ML)

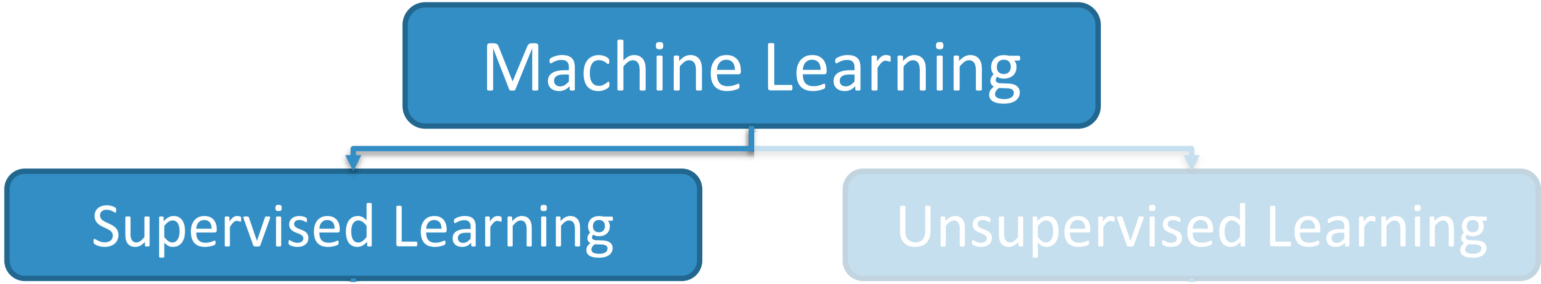
- Algorithms and statistical models that enable computers to uncover patterns in data

- High overlap with statistics; some classic statistical models are also referred to as machine learning models, e.g. linear regression

- Two main branches of algorithms: **supervised and unsupervised**



Machine Learning

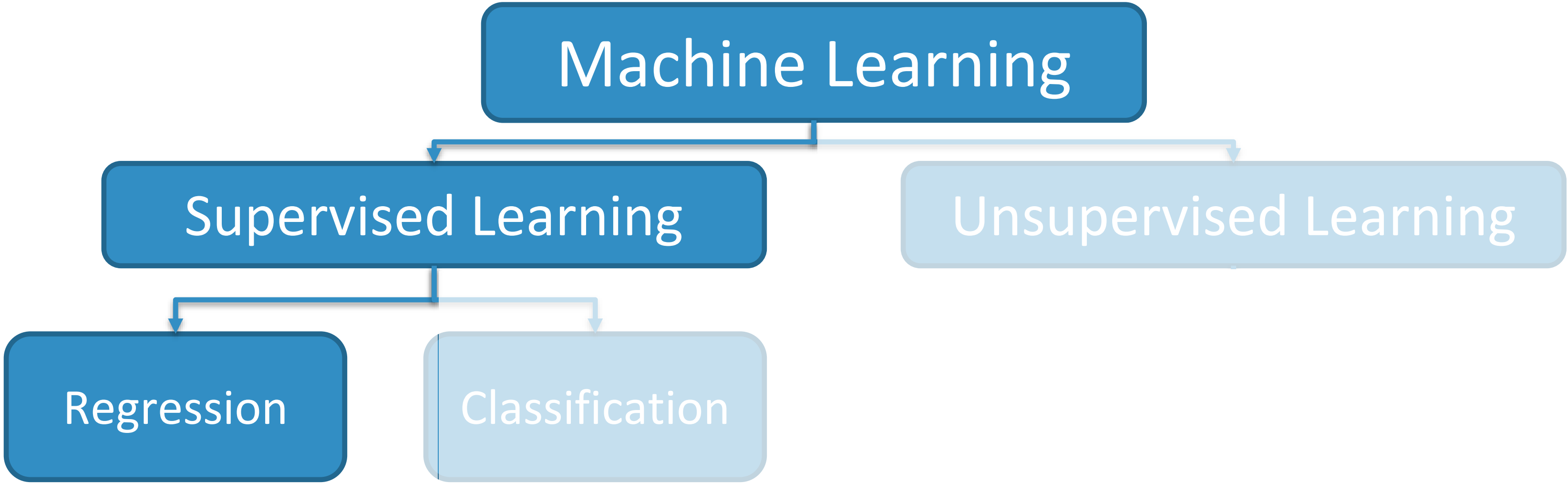


Supervised Learning

Supervised Learning

- Machine learning with **labels**
- Label: also known as target, y, output, class
- Two major flavors: **regression** and **classification**

Machine Learning



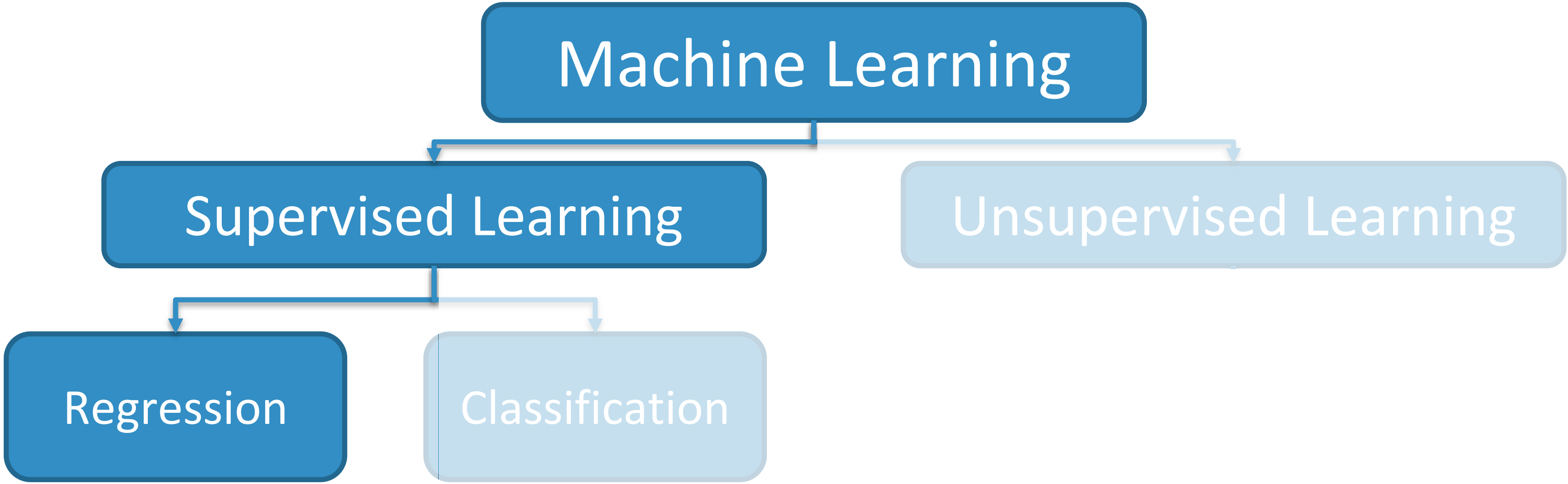
Supervised Learning: Regression

Answers questions like:

- How much profit will we make next year?
- How long will a reader stay on our site?

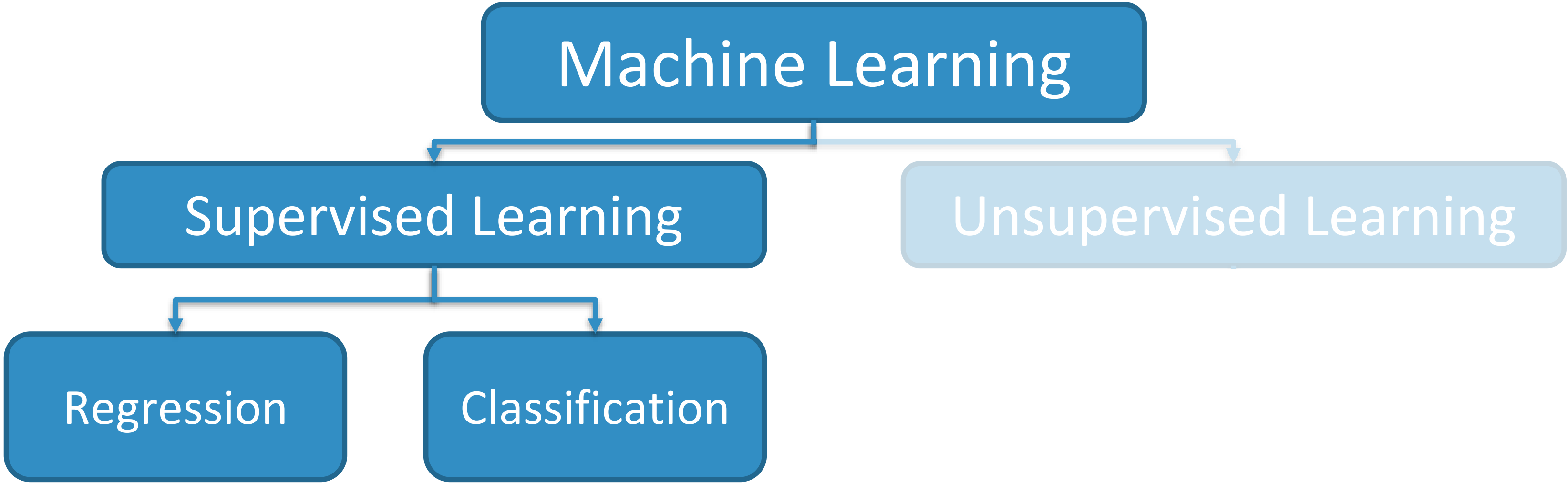
Applications: demand forecasting, predicting stock prices, customer lifetime value

Machine Learning



- Demand forecasting
- Lifetime value

Machine Learning



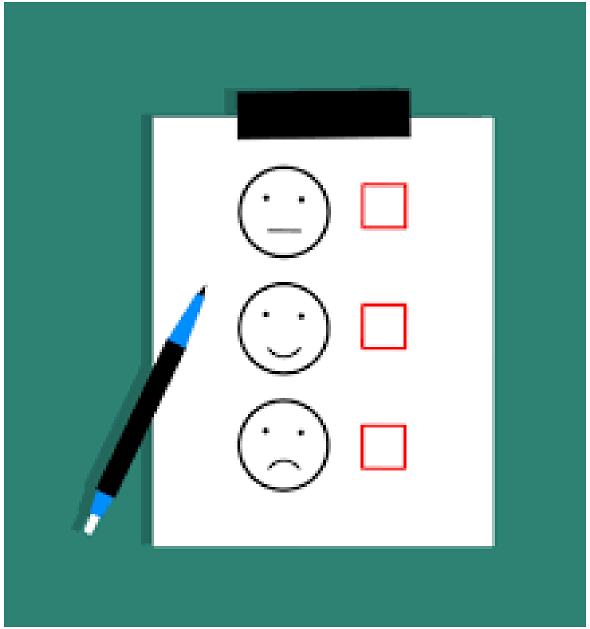
- Demand forecasting
- Lifetime value

Supervised Learning: Classification

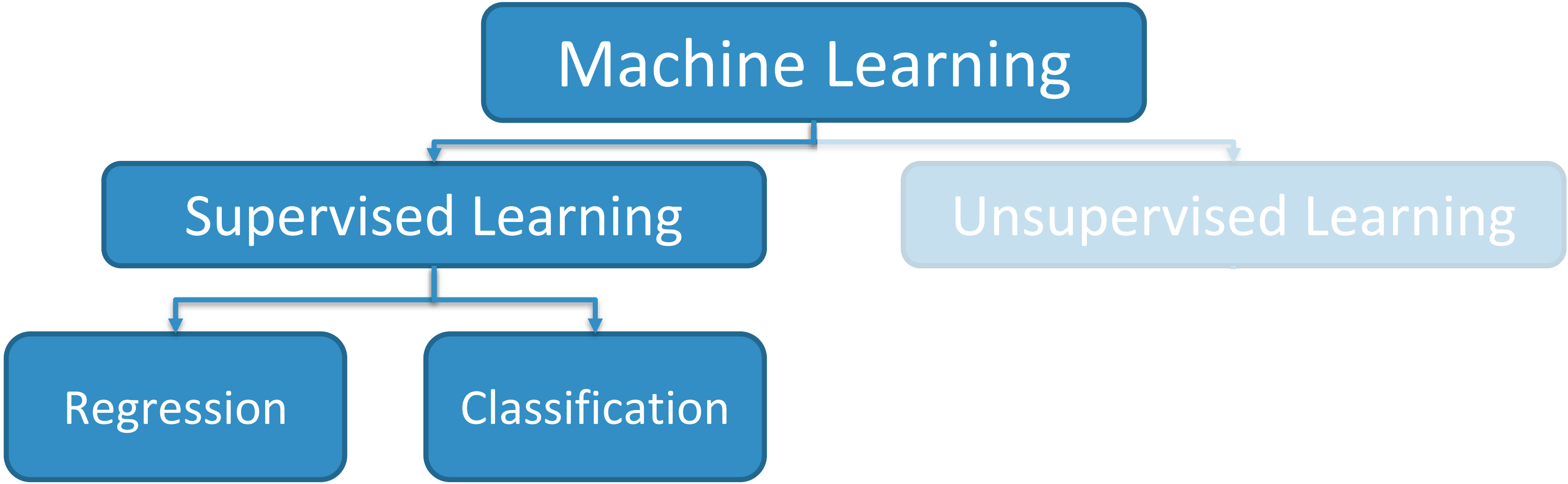
Labels are class or group, e.g. 1 or 0, “churned” or “not churned”

Linear and nonlinear models

Algorithms include k-nearest neighbors, logistic regression, decision trees, SVMs



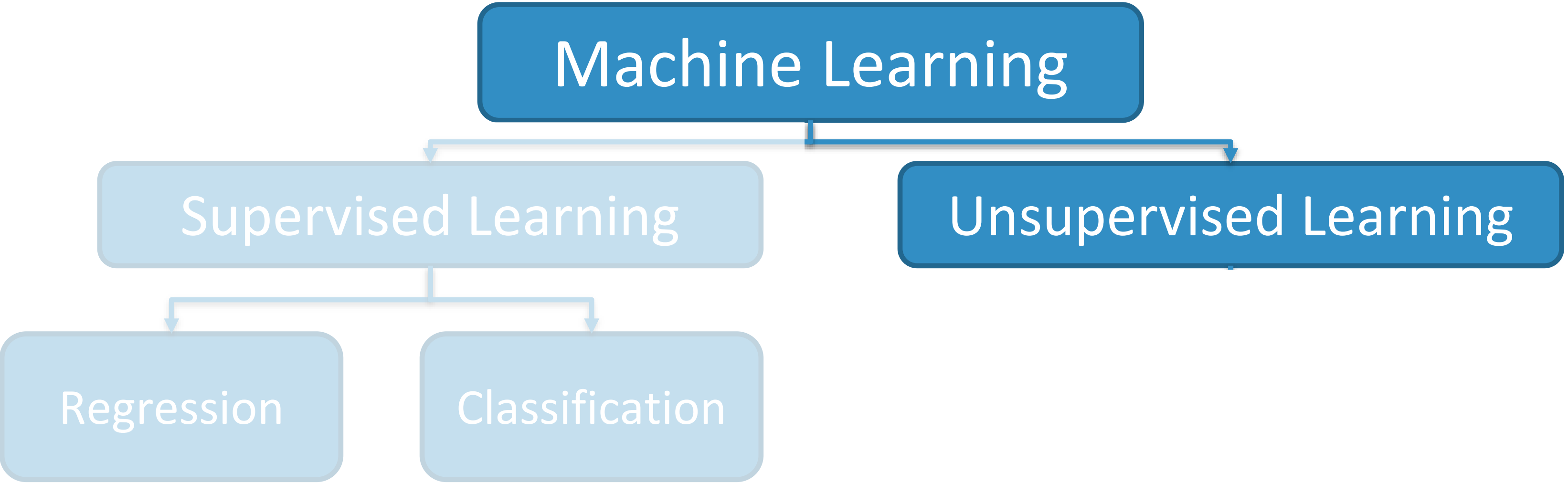
Machine Learning



- Demand forecasting
- Lifetime value

- Churn
- Fraud

Machine Learning



- Demand forecasting
- Lifetime value

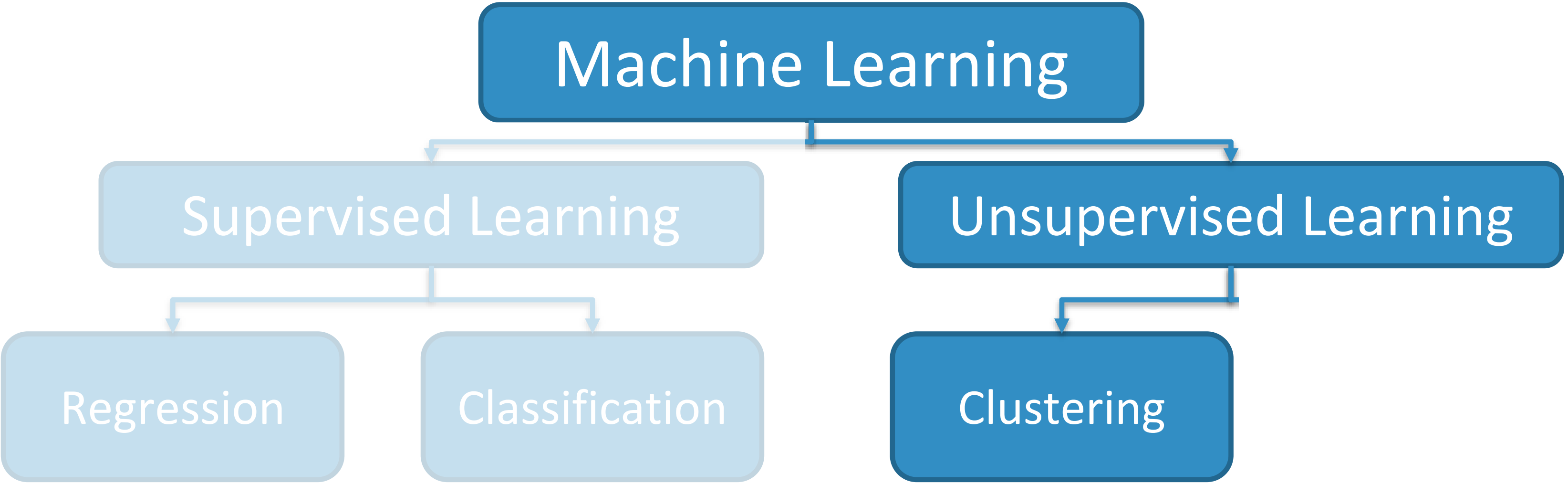
- Churn
- Fraud

Unsupervised Learning

Unsupervised Learning

- Machine learning ***without*** labels
- Uncover the underlying structure of data
- Two major branches: **clustering** and **dimension reduction**

Machine Learning


Machine Learning

Supervised Learning

Unsupervised Learning

Regression

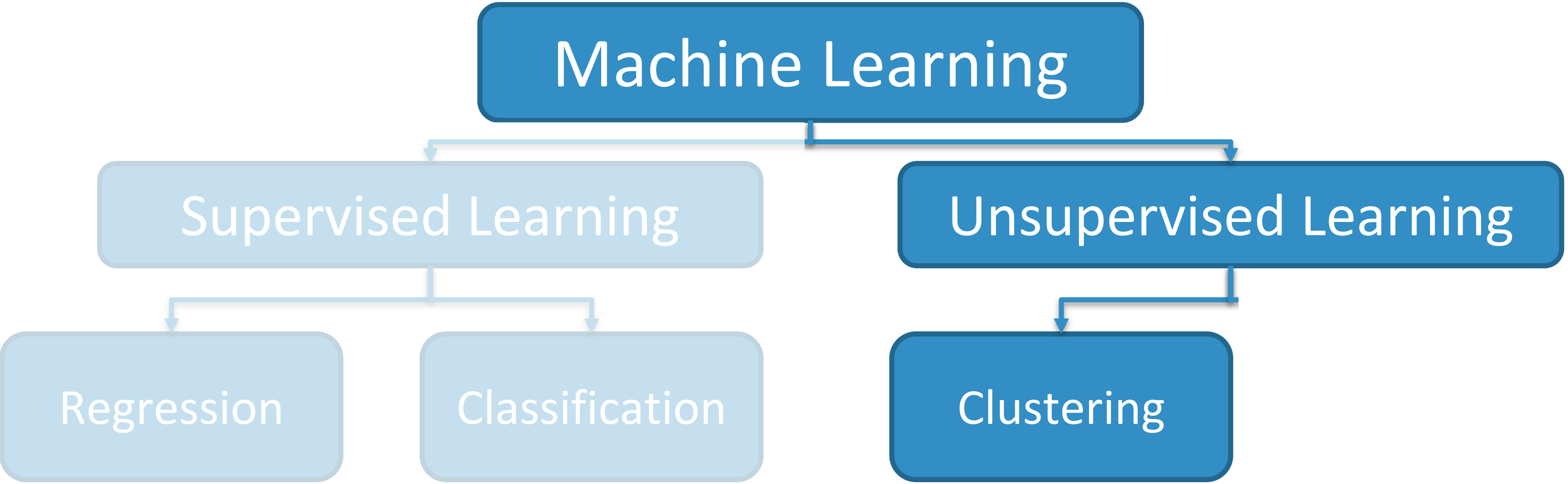
Classification

Clustering

- Demand forecasting
- Lifetime value

- Churn
- Fraud

Machine Learning

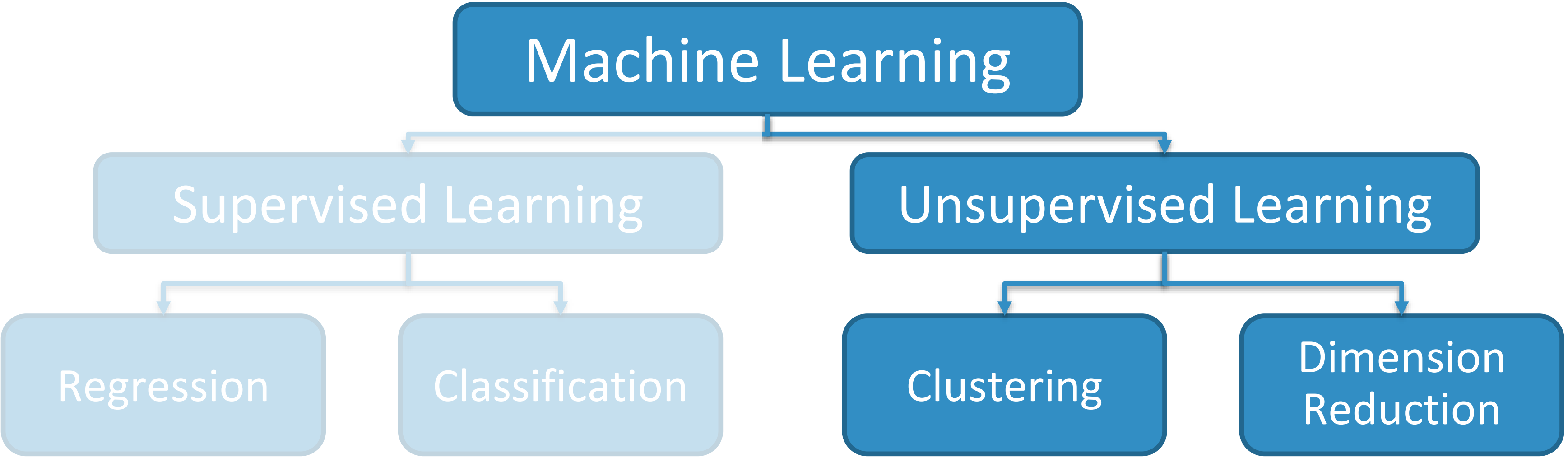


- Demand forecasting
- Lifetime value

- Churn
- Fraud

- Customer Segmentation
- Feature generation

Machine Learning

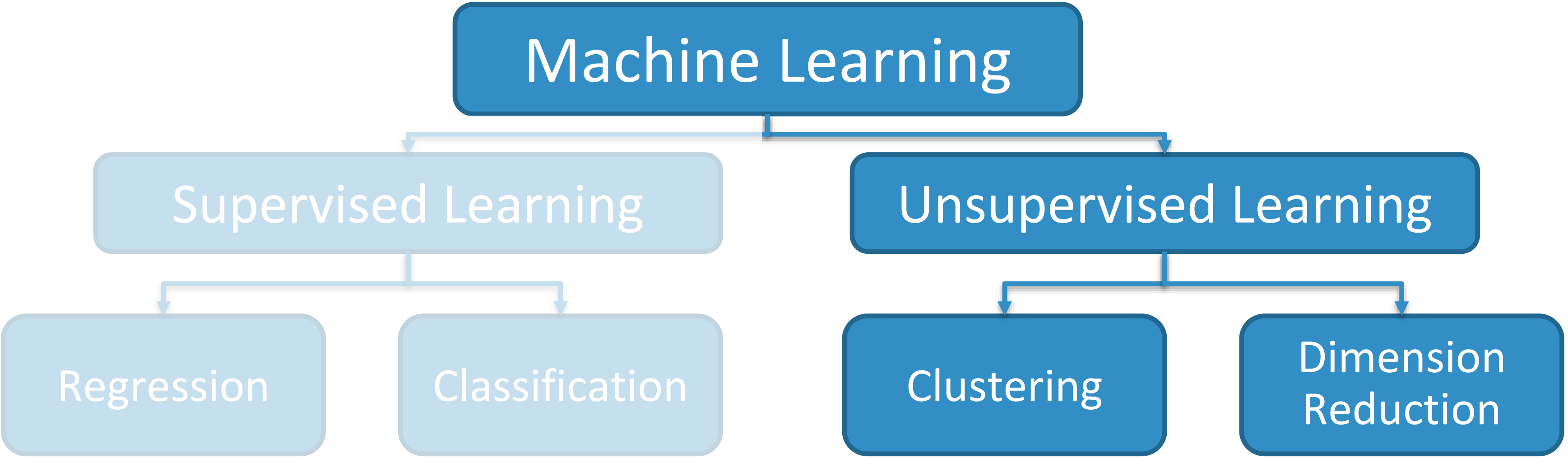


- Demand forecasting
- Lifetime value

- Churn
- Fraud

- Customer Segmentation
- Feature generation

Machine Learning



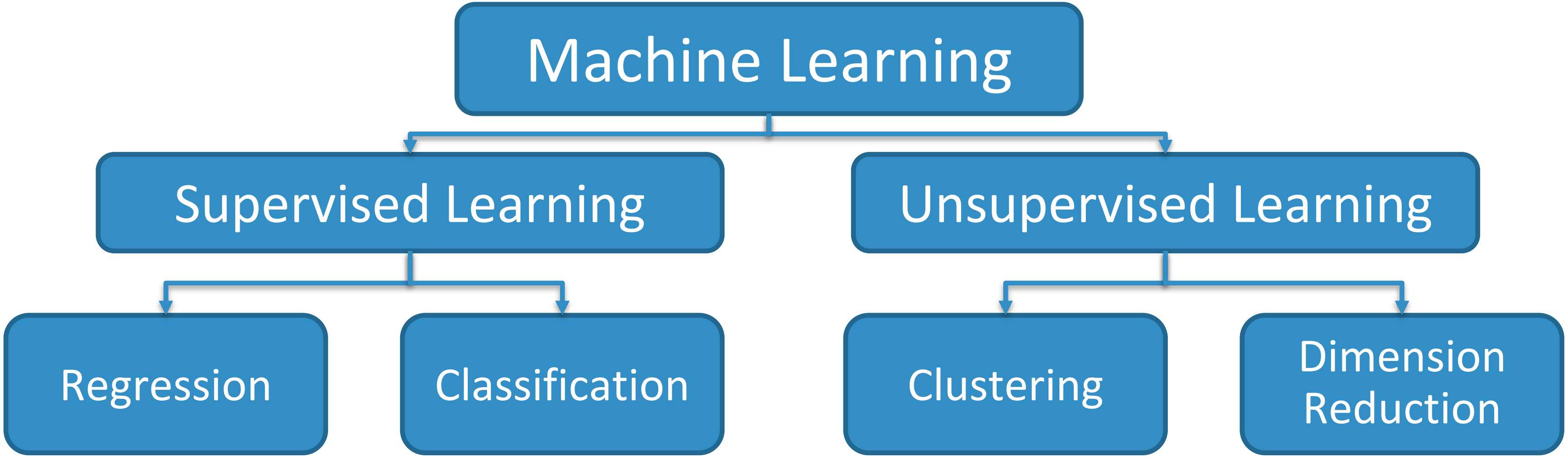
- Demand forecasting
- Lifetime value

- Churn
- Fraud

- Customer Segmentation
- Feature generation

- Image Compression
- Feature generation

Machine Learning



Machine Learning

Supervised Learning

Unsupervised Learning

Regression

Classification

Clustering

Dimension Reduction

Regression:

- Demand forecasting
- Lifetime value

Classification:

- Churn
- Fraud

Clustering:

- Customer Segmentation
- Feature generation

Dimension Reduction:

- Image Compression
- Feature generation

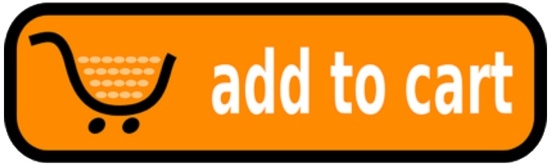
SPECIAL TOPICS

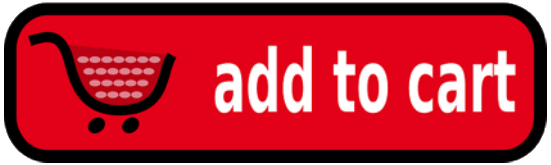
METIS

Special Topics

A/B Testing: running an "experiment" to test two (or more) alternatives against each other

- Common in marketing and online sales
- Everyday application: button color testing

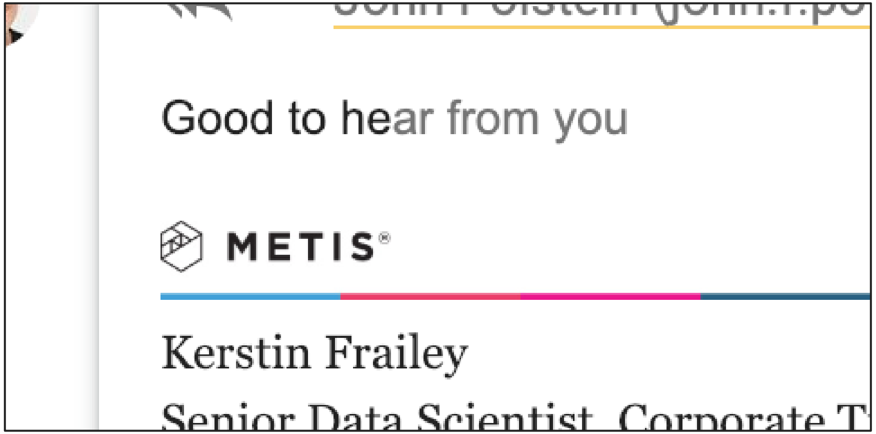




Special Topics

NLP (Natural Language Processing): analysis of human language by computers; machine learning and AI applied to text

- Methods: sentiment analysis, topic modelling, etc.
- Everyday application: autocomplete,



Good to hear from you

Kerstin Frailey

Senior Data Scientist, Corporate T

Special Topics

Time Series Analysis: applying statistical and machine learning techniques to find patterns in and predict with time-indexed data

- Common in financial markets
- Everyday application: demand forecasting

Special Topics

Neural Network: a type of machine learning vaguely inspired by the workings of neurons in a brain; composed of an input layer, output layer, and “hidden” layers

Deep Learning: a type of neural net with many hidden layers

- Common in image recognition, NLP
- Everyday application: speech recognition

Special Topics

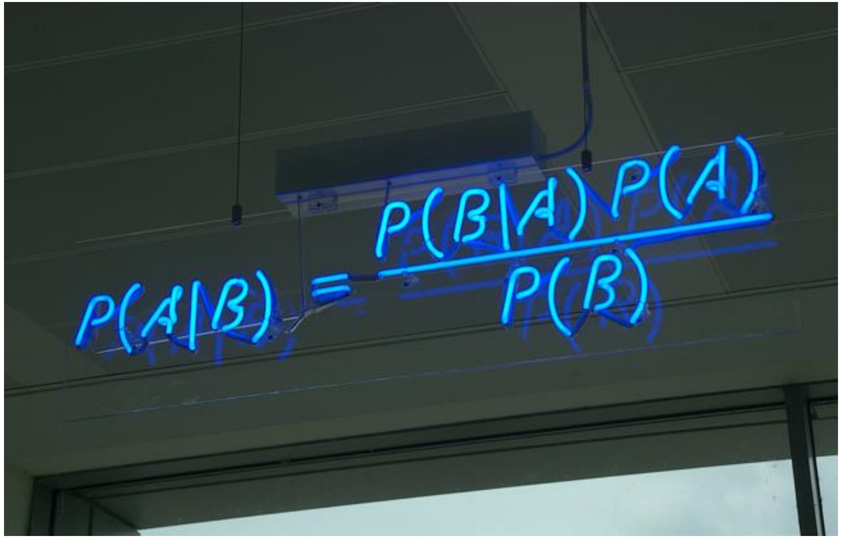
Computer Vision: a field of study on how computers can gain information about an environment through images

- Machine learning and neural networks are often applied for image recognition
- Everyday application: goofy video filters

Special Topics

Bayesian Statistics: a theory in statistics which takes the approach that probability expresses a “degree of belief”

- Results in different assumptions and underlying math
- Machine learning methods naïve Bayes



Course Structure

METIS

Course Structure

- Module 1: Basic Python & Math (weeks 1 and 2)
- Module 2: Exploratory Data Analysis (weeks 3 and 4)
- Module 3: Regression (weeks 5 and 6)
- Module 4: Classification (weeks 7 and 8)
- Module 5: Unsupervised Learning & NLP (weeks 9 and 10)
- Module 6: Deep Learning (weeks 11 and 12)

Recap

METIS

Learning objectives

Be able to

- Describe data science and explain its different facets
- Explain the differences between statistics and machine learning
- Explain the major branches of machine learning and the types of problems they solve
- Describe special topics within data science

Takeaways

- Data science means different things at different places, but it generally involves, analytics, statistics, machine learning, artificial intelligence, and programming.
- Supervised and unsupervised learning are the two main branches of machine learning
- Statistics and machine learning have a large overlap
- Artificial Intelligence is not well defined

QUESTIONS?

METIS