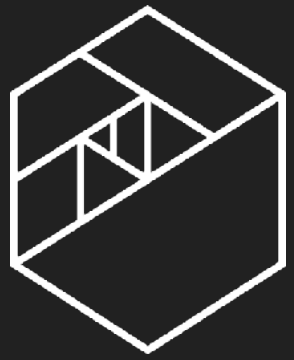



# Intro to Data Science

# 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

A Brief History of Data Science

Basics of Data Science

Analytics and Statistics

Statistics and Machine Learning

Machine Learning and Artificial Intelligence

Special Topics

# A BRIEF HISTORY OF DATA SCIENCE

METIS

# History

In 2008, Jeff Hammerbacher at Facebook and DJ Patil at Linkedin needed a new title to describe the growing responsibilities of their data and analytics teams: "Data Scientist" was born

"Data Science", as a term, has been around much longer

# History

Data science rose to prominence circa 2012, when DJ Patil and Thomas H. Davenport wrote “Data Scientists: Sexiest Job of the 21st Century” for Harvard Business Review

# Definition

Data science is the practice of extracting useful and actionable information from data that is 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 not possible

# BASICS OF DATA SCIENCE

METIS

# Ambiguity

The details of what qualifies as data science are still up for dispute

The field is young, so its boundaries are naturally soft

"Data Science" is now often used as a catchall term for **advanced analytics**

# Ambiguity

Different organizations use the term “data science” and “data scientist” to refer to vastly different functions and roles:

- Product Analytics
- Research and development
- Statistics
- And many others

# Ambiguity

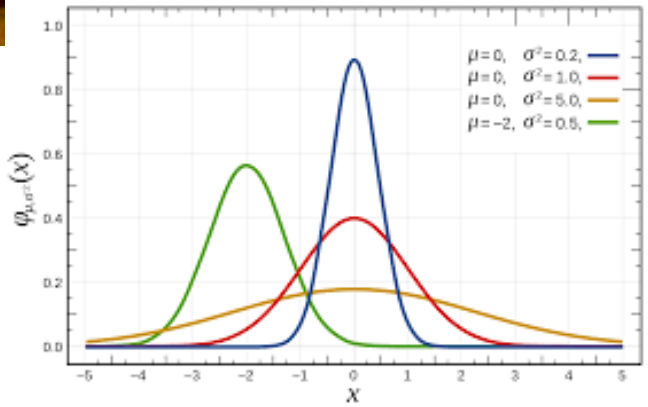
Data science generally includes some combination of analysis, statistics, machine learning, artificial intelligence, and programming

Typical languages used by data scientists include Python, SQL, R

# 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

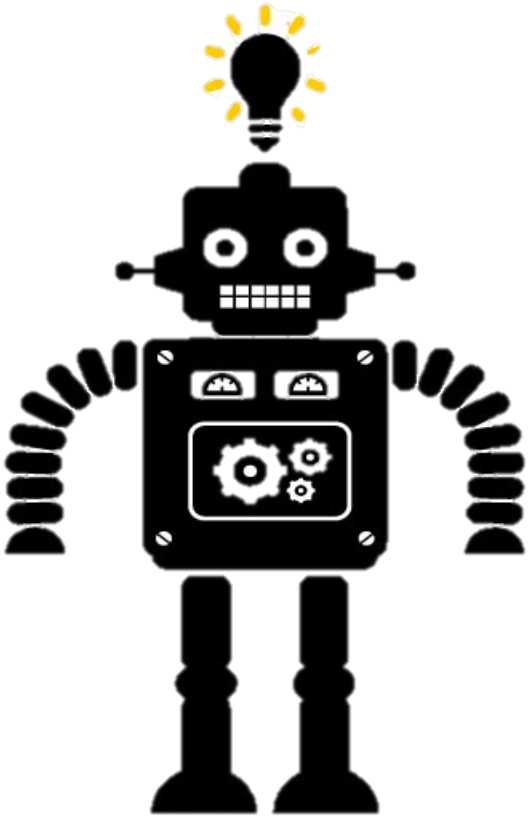


# Major Components

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

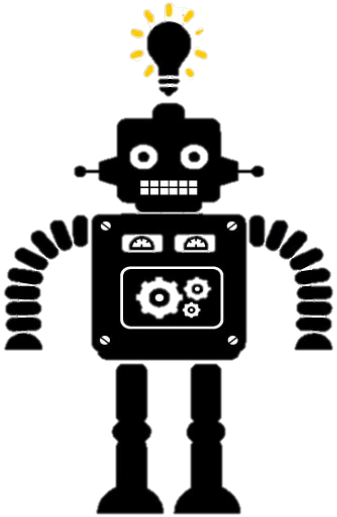
**Computer Science**: the study of algorithms and computation

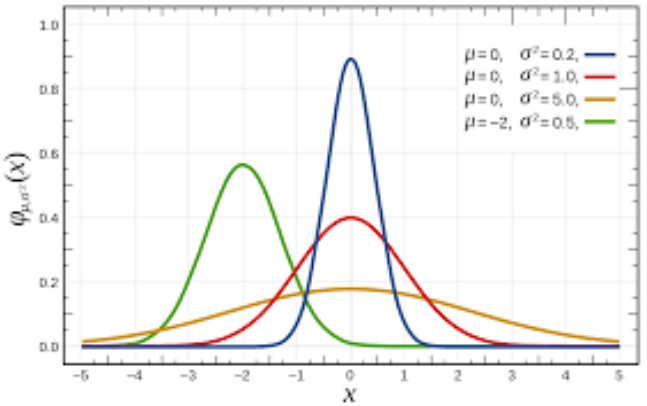
**Artificial Intelligence (AI)**: No agreed upon definition



# Major Components

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





# Problem Types

**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

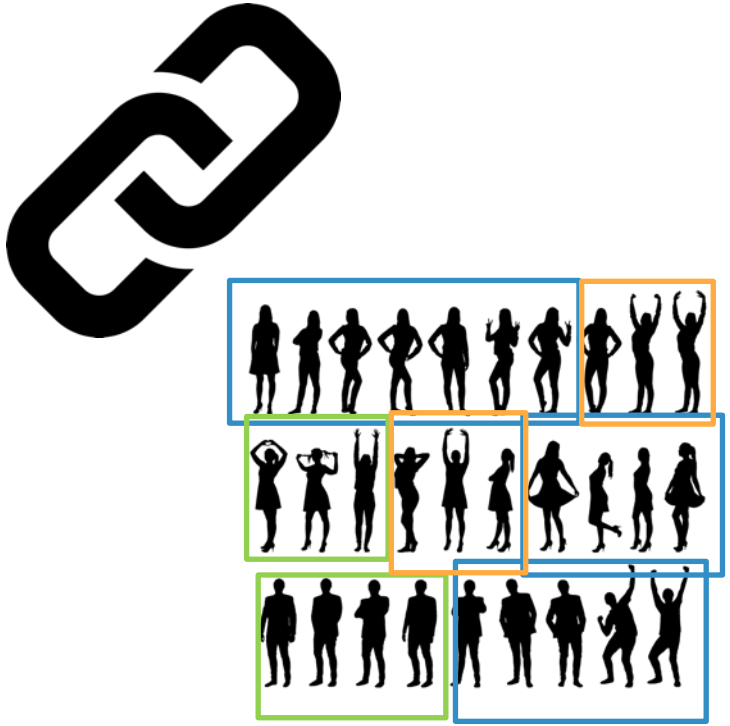
# Problem Types

**Relationships**: If one variable changes, is this other variable likely to?

- Correlation, linear coefficients, “links”

**Organization**: Is there natural structure to the data? Clear groups?

- Clustering, dimension reduction, customer segmentation



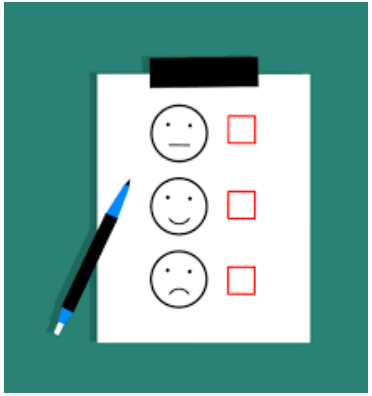
# Problem Types

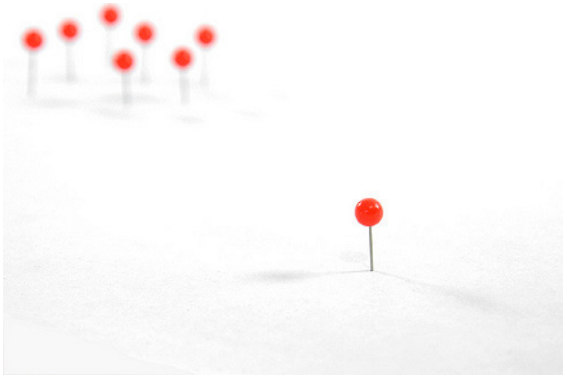
**Identification**: Can we identify unknown values?

- Classification, sentiment analysis

**Anomalies**: Where are the strange data points?

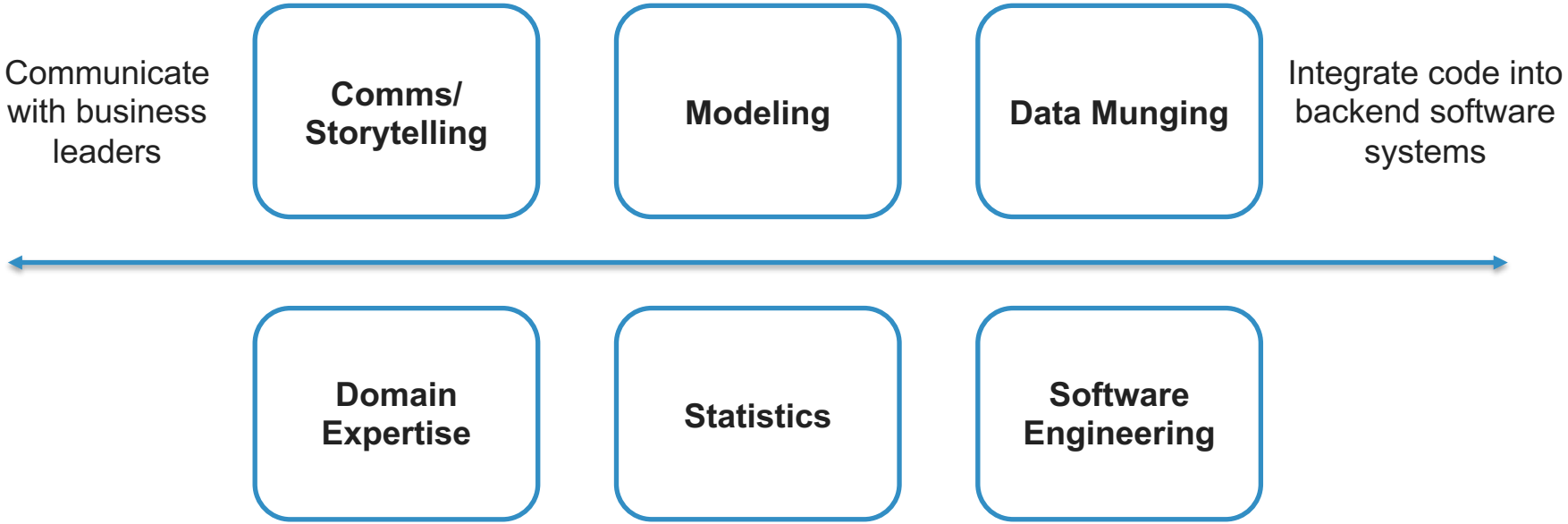
- Anomaly detection, outliers





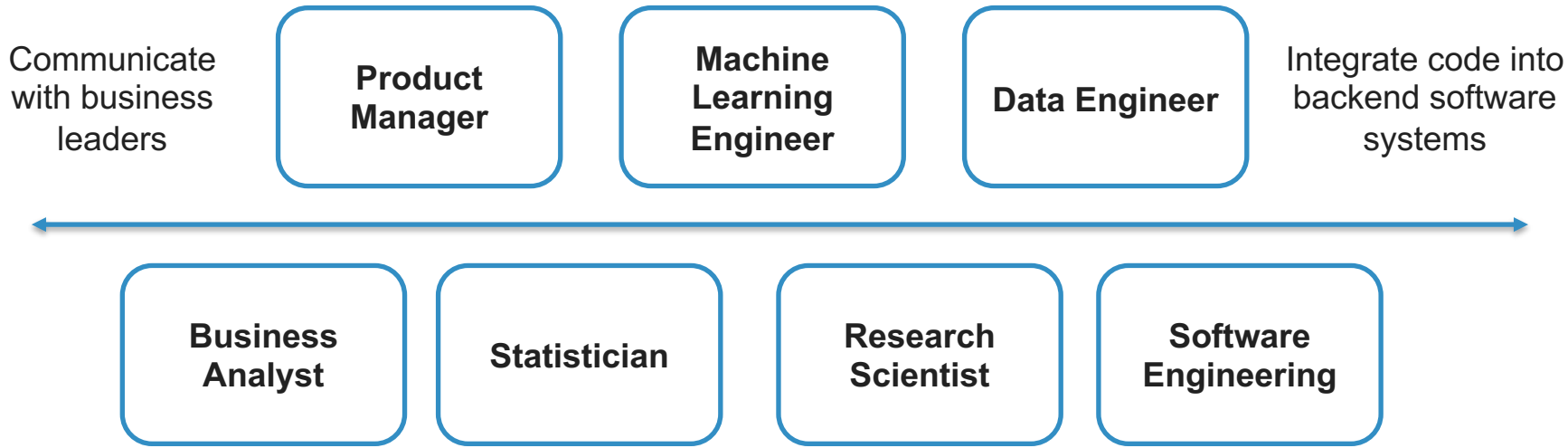
# Data Science Team Skills

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



# Data Science Team Roles

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



# 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

# Analytics

Answers direct, clear questions with deterministic answers

Monitors changes in business and informs decision makers

Leans heavily on business rules

# Analytics

Often specializing in one branch of business, e.g. product analytics, financial analytics, marketing analytics

Combine some level of subject matter expertise (SME) with data know-how

Tools: SQL, Excel, Tableau

# Analytics

Most commonly answer *descriptive* questions

Sometimes answers predictive questions

Results inform decision makers to help prescribe action

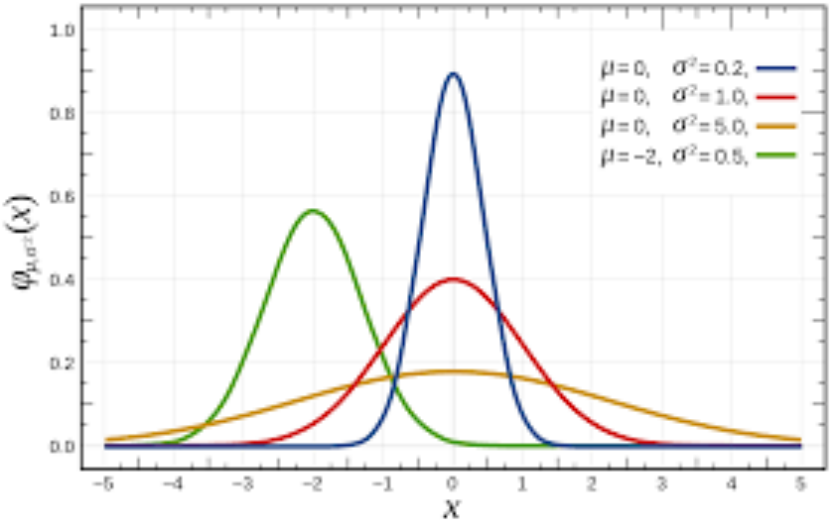
# Statistics

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

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

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

"Backbone" of modern science

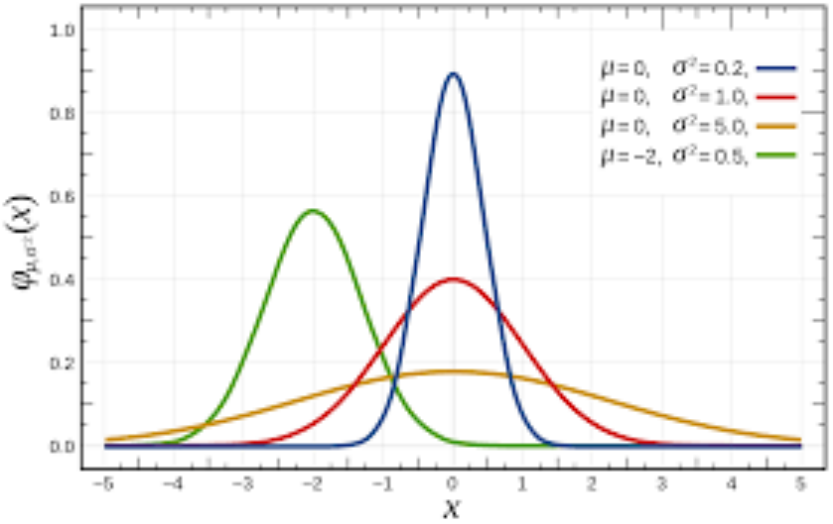


# Statistics

High overlap with other fields, e.g. operations research, applied math, computer science, econometrics, etc.

Methods include: regression, generalized linear models, p-values

Tools: R, SAS

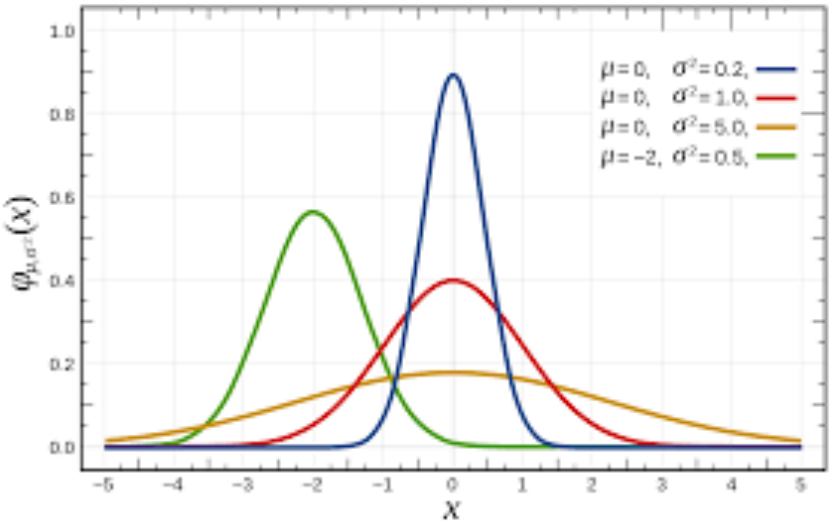


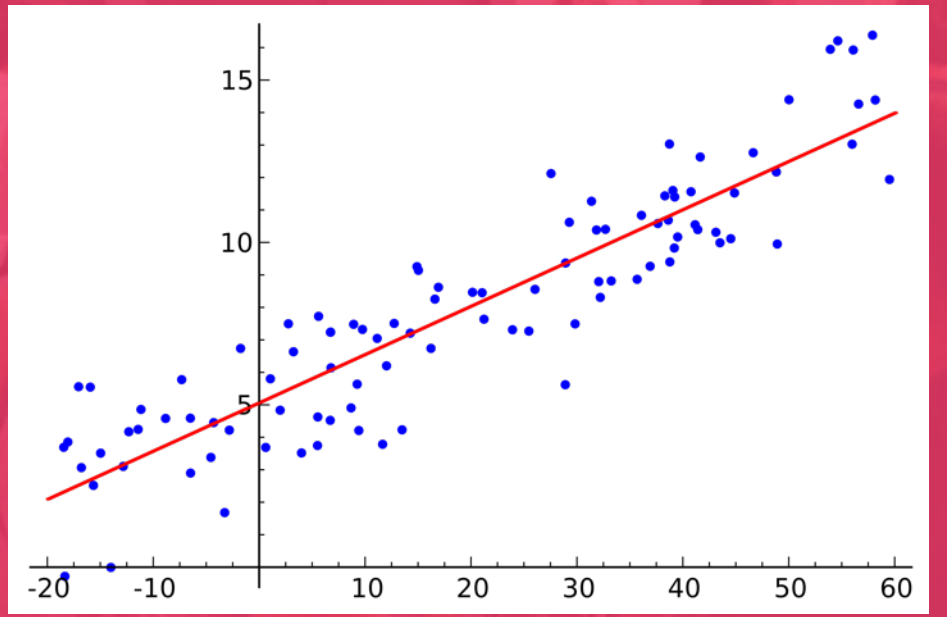
# Statistics

Answers descriptive, predictive, and relationship questions

Probability and mathematical guarantees

Concerned with the *distribution* of numbers & metrics






## Know the Jargon

Linear Model:

$$y = \alpha + \beta x + \epsilon$$

$y$ *dependent variable, response*

$\alpha$ *intercept*

$\beta$ *coefficient*

$x$ *independent variable*
*covariate, predictor*

$\epsilon$ *(random) error, noise*

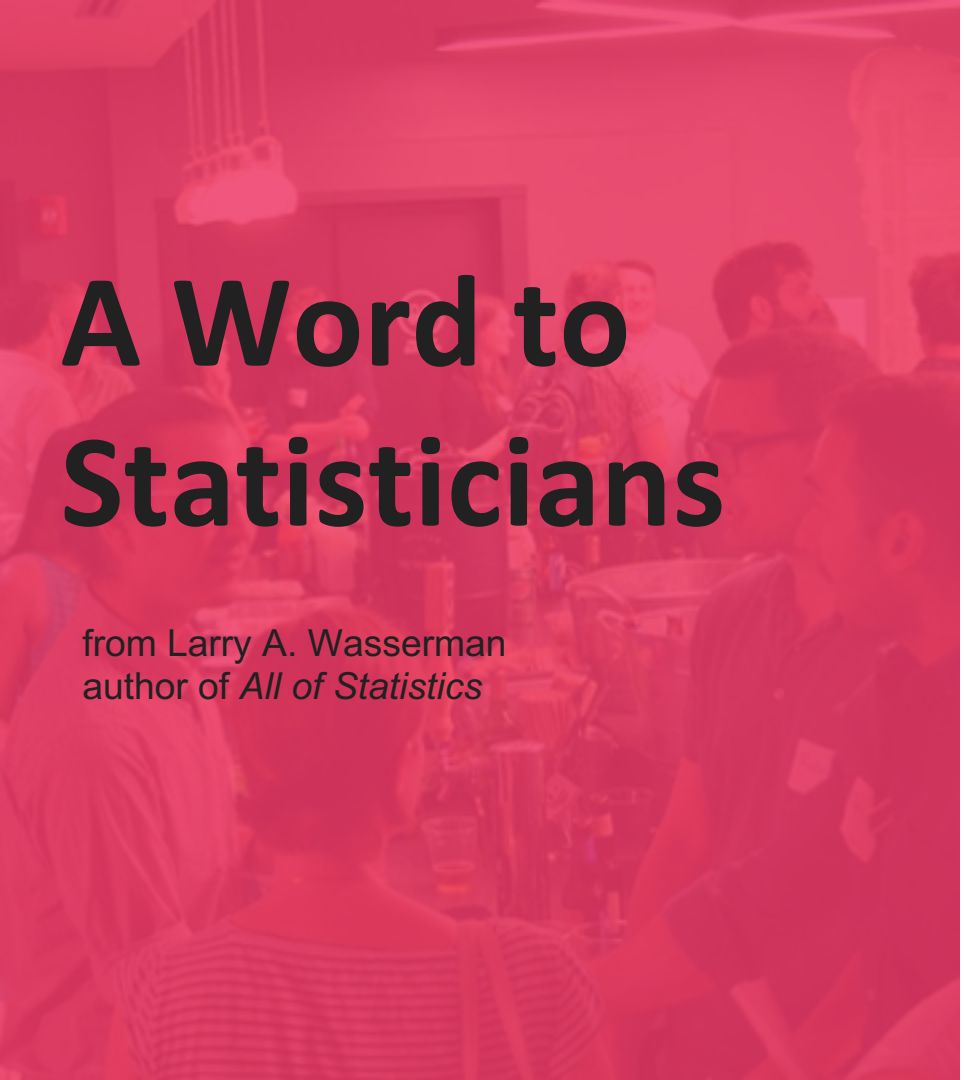
# STATISTICS & MACHINE LEARNING

METIS

# A Word to Statisticians

Rather notoriously, the line between statistics and machine learning is blurry

And, statistics came first



# A Word to Statisticians

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

Statistics emphasizes

- formal statistical inference (confidence intervals, hypothesis tests, optimal estimators) in low dimensional problems.

Machine Learning emphasizes

- high dimensional prediction problems.

"But this is a **gross over-simplification**."

Statistics pays more attention to

- survival analysis, spatial analysis, multiple testing, minimax theory, deconvolution, semiparametric inference, bootstrapping, time series.

Machine Learning pays more attention to

- online learning, semisupervised learning, manifold learning, active learning, boosting

**"But the differences become blurrier all the time."**

# A Word to Statisticians

from Robert Tibshirani
coauthor of *Elements of Statistical Learning*

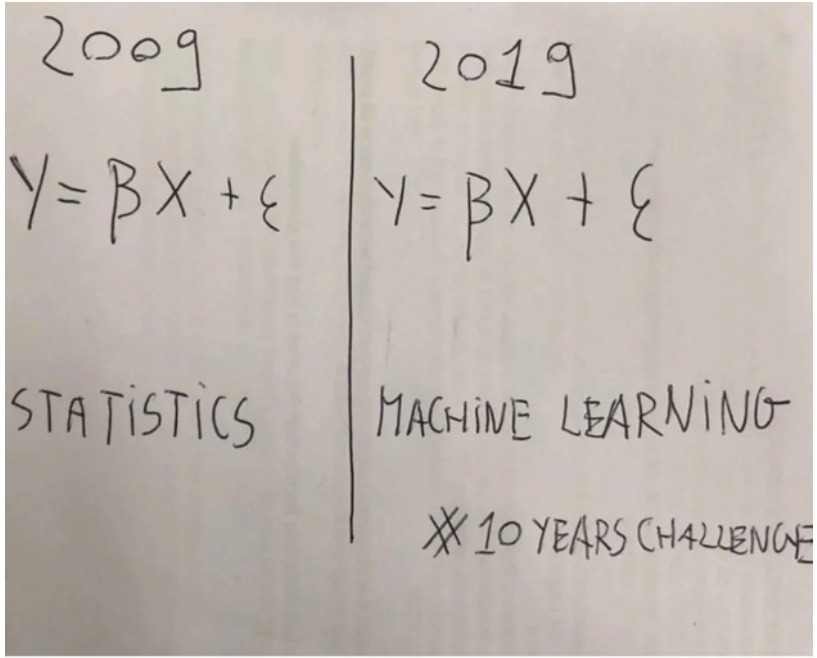
## Machine learning is “glorified statistics”

**Glossary**

| Machine learning | Statistics |
|---|---|
| network, graphs | model |
| weights | parameters |
| learning | fitting |
| generalization | test set performance |
| supervised learning | regression/classification |
| unsupervised learning | density estimation, clustering |
| large grant = $1,000,000 | large grant= $50,000 |
| nice place to have a meeting: Snowbird, Utah, French Alps | nice place to have a meeting: Las Vegas in August |

# A Word to Statisticians

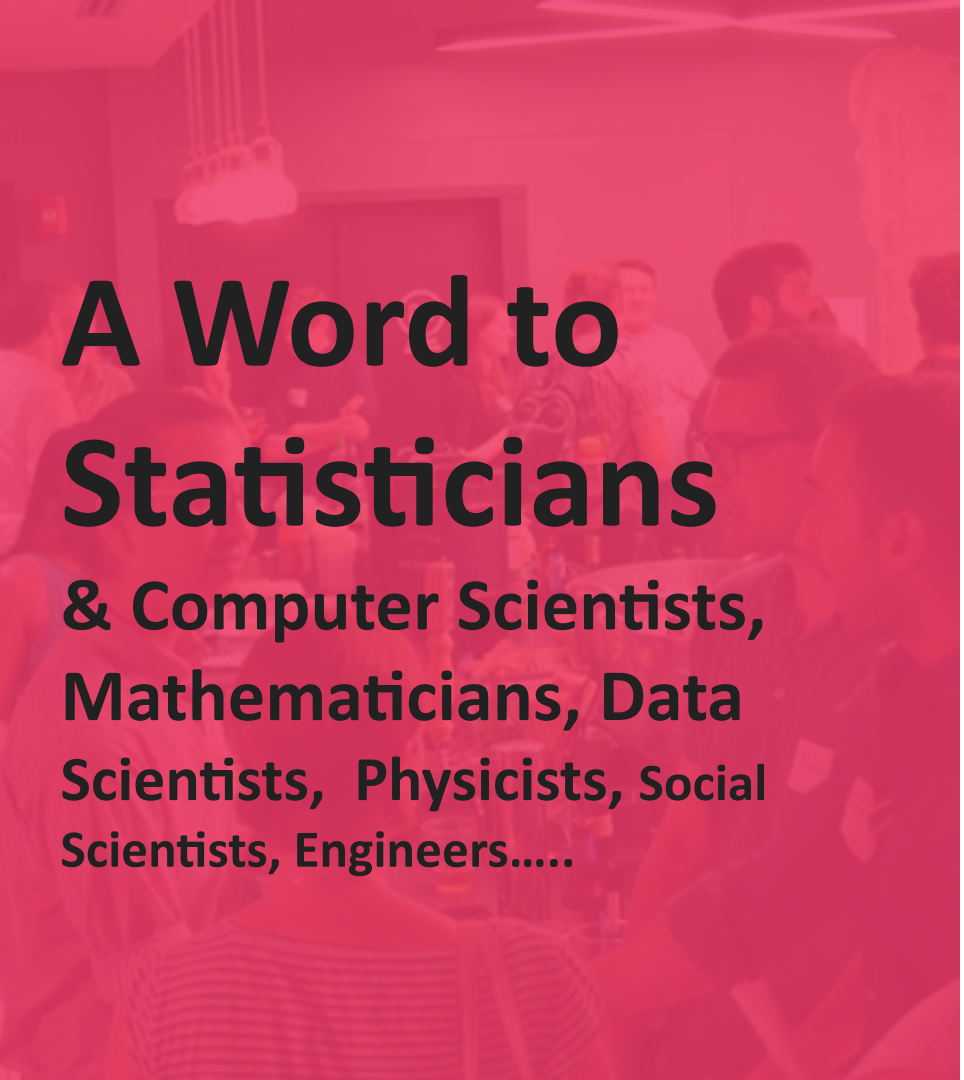
from u/keymado
redditor on r/datascience



# A Word to Statisticians

The concepts are often familiar, although the words are sometimes different

Data science leans towards ML jargon, but both are used



# A Word to Statisticians

## & Computer Scientists, Mathematicians, Data Scientists, Physicists, Social Scientists, Engineers…..

We have many words for the same thing

And sometimes, we have the same word for many things 🤷‍♀️

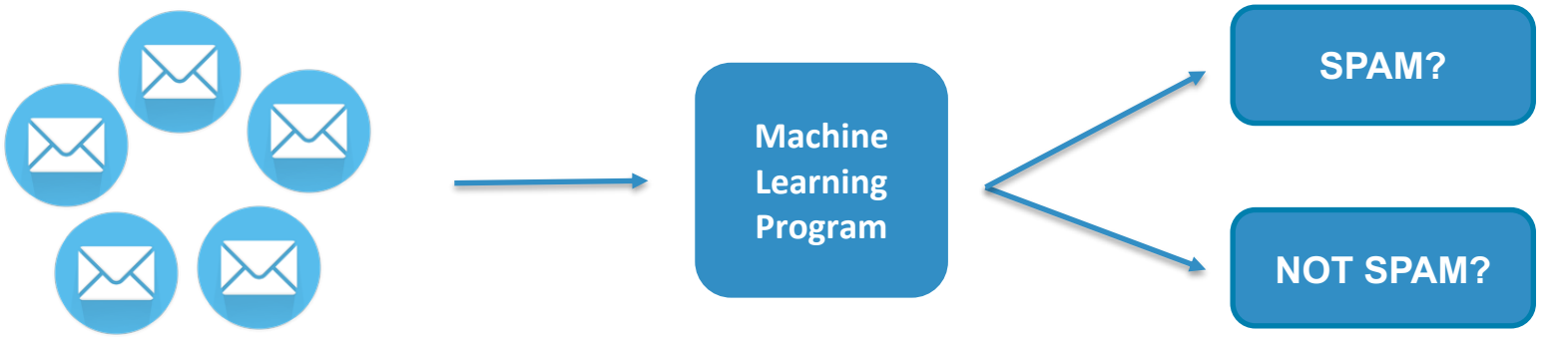
# MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

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



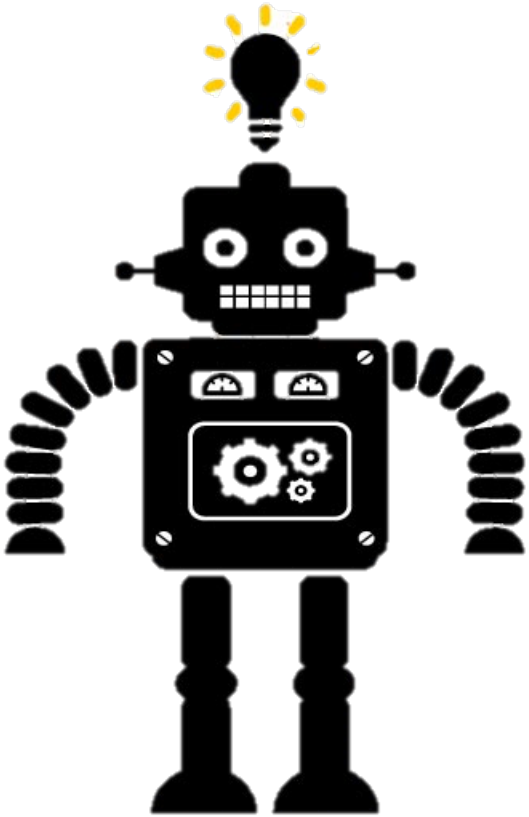

*Emails are labeled as spam vs. not*

*The more emails the program sees…*

*…the better it gets at classification*

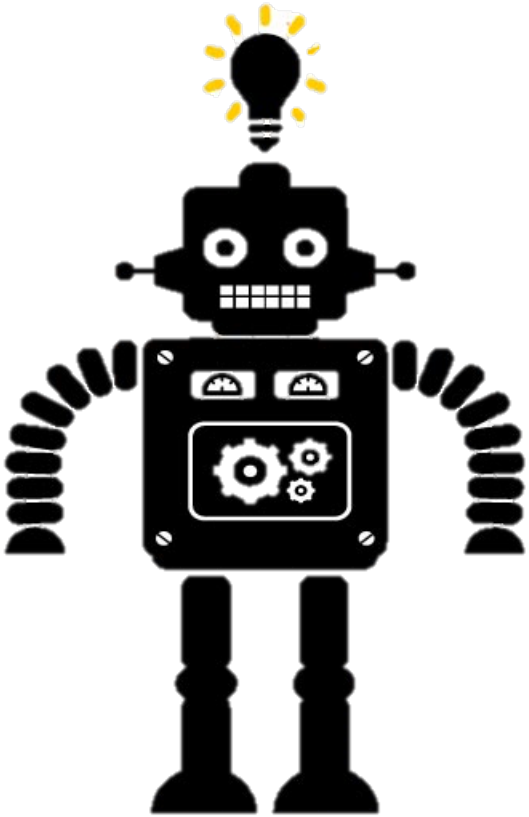
# Machine Learning (ML)

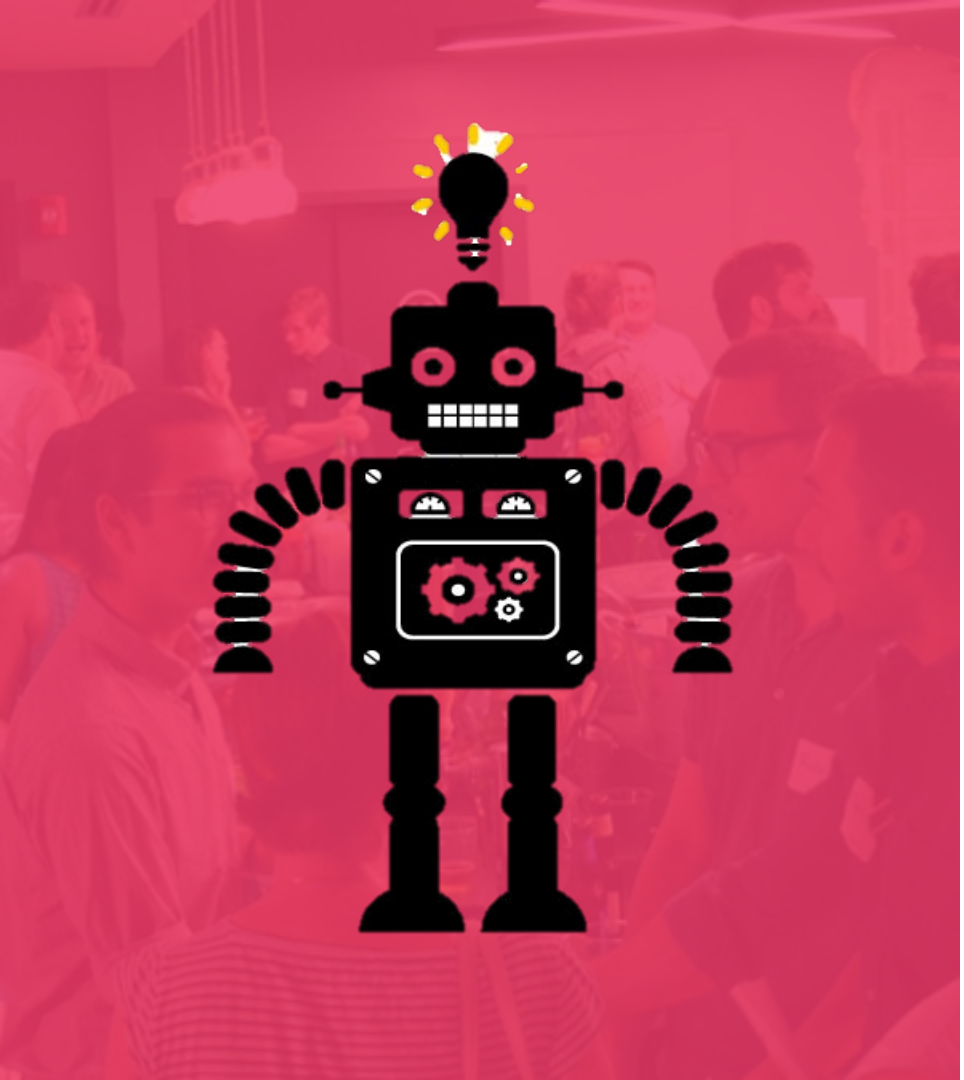
- Algorithms and statistical models that enable computers to uncover patterns in data
- High overlap with statistics; some classic statistical models are also referred to as machine learning models, e.g. linear regression
- Two main branches of algorithms: **supervised and unsupervised**



# Artificial Intelligence (AI)

- No agreed upon definition
- May refer to anything from machine learning to self-driving cars to Ex Machina's Ava
- It's not magic--it's just math
- AI effect: "AI is whatever hasn't been done yet." ~ Douglas Hofstadter





## Know the Jargon

ML Model:

$$y \sim x$$

$y$ $target, label, output$

$x$ $features, input$

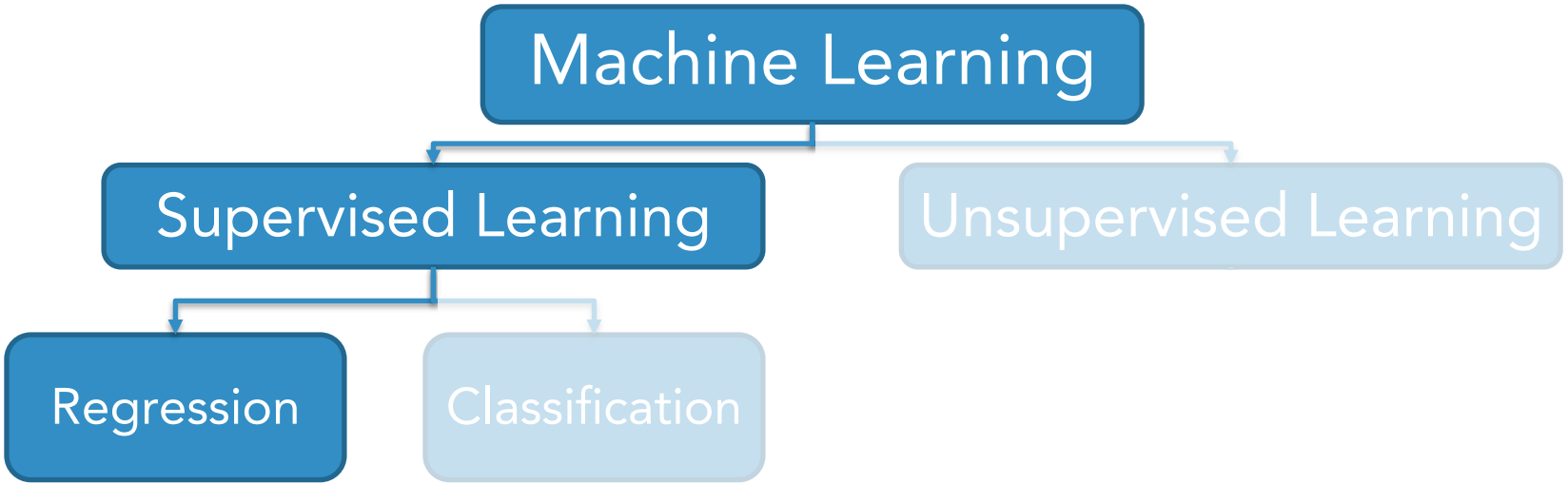
# Machine Learning

# 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

Labels are numbers, e.g. 74.2, .00053, 32.0

Linear and nonlinear models

Algorithms include linear regression, SVMs, random forests

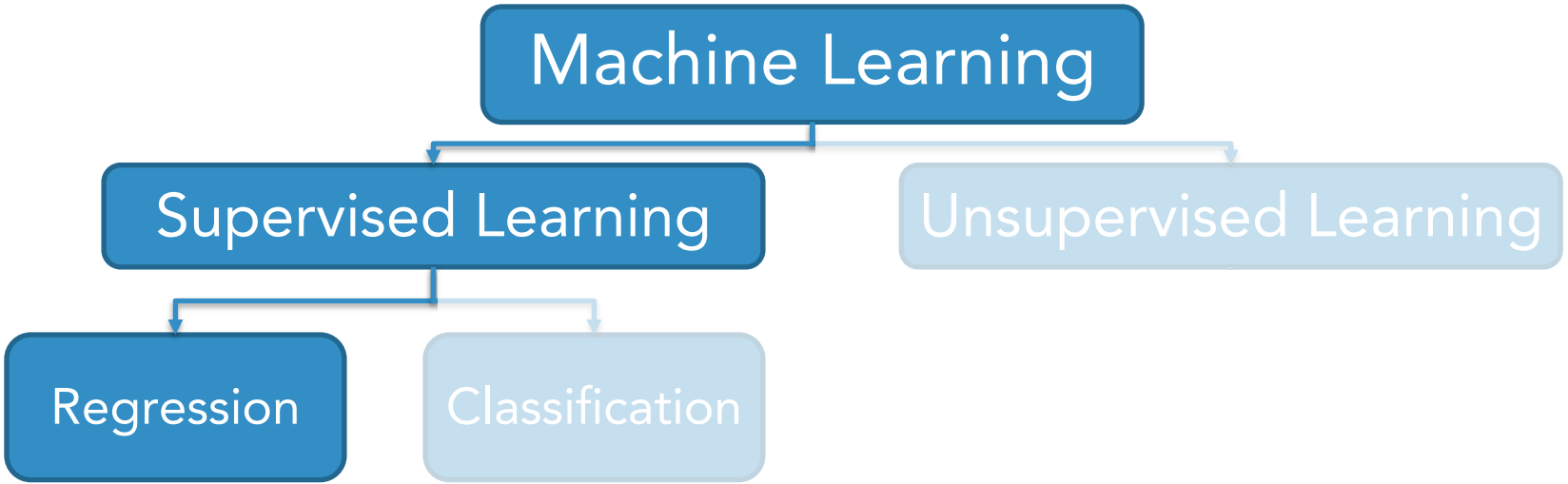
# Supervised Learning: Regression

Answer questions like:

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

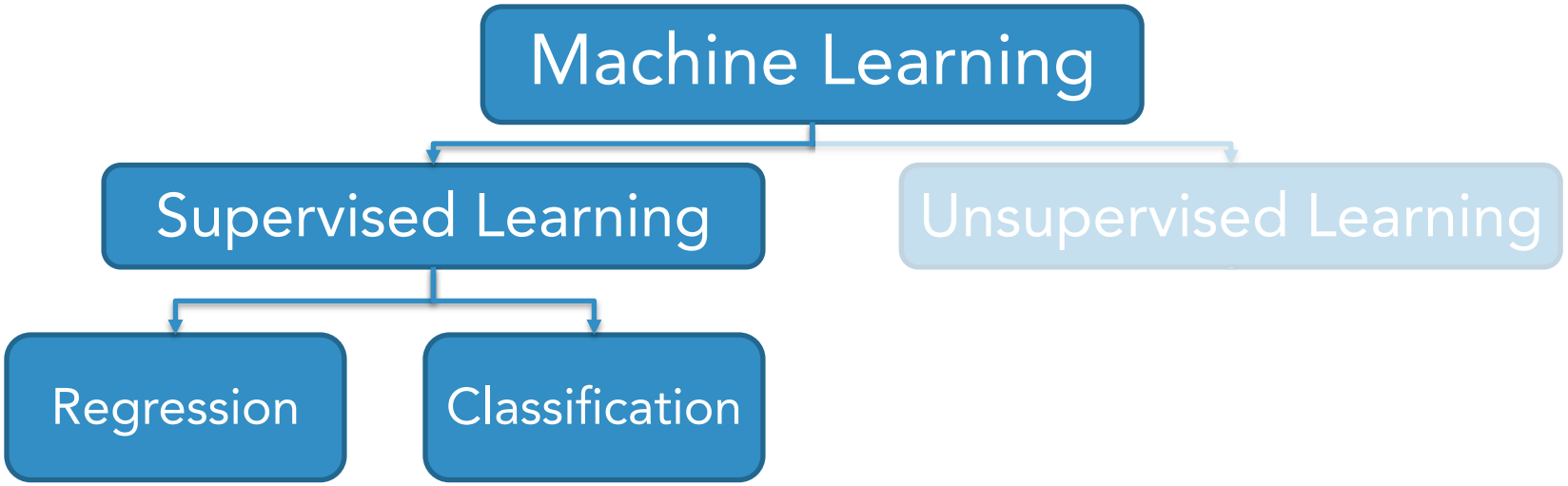
**Applications:** demand forecasting, predicting stock prices, customer lifetime value

# Machine Learning




- Demand forecasting
- Lifetime value

# Machine Learning



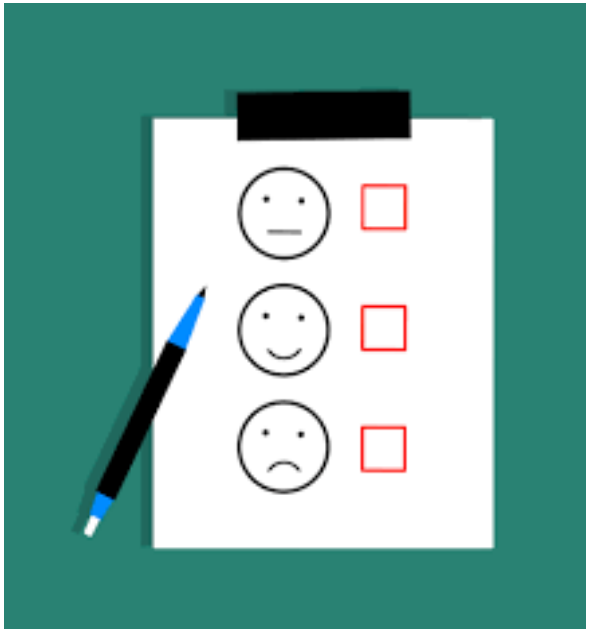

- Demand forecasting
- Lifetime value

# 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

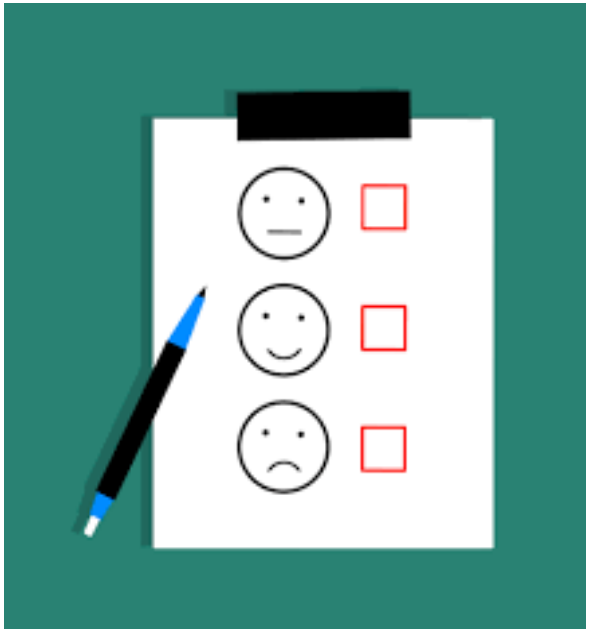


# Supervised Learning: Classification

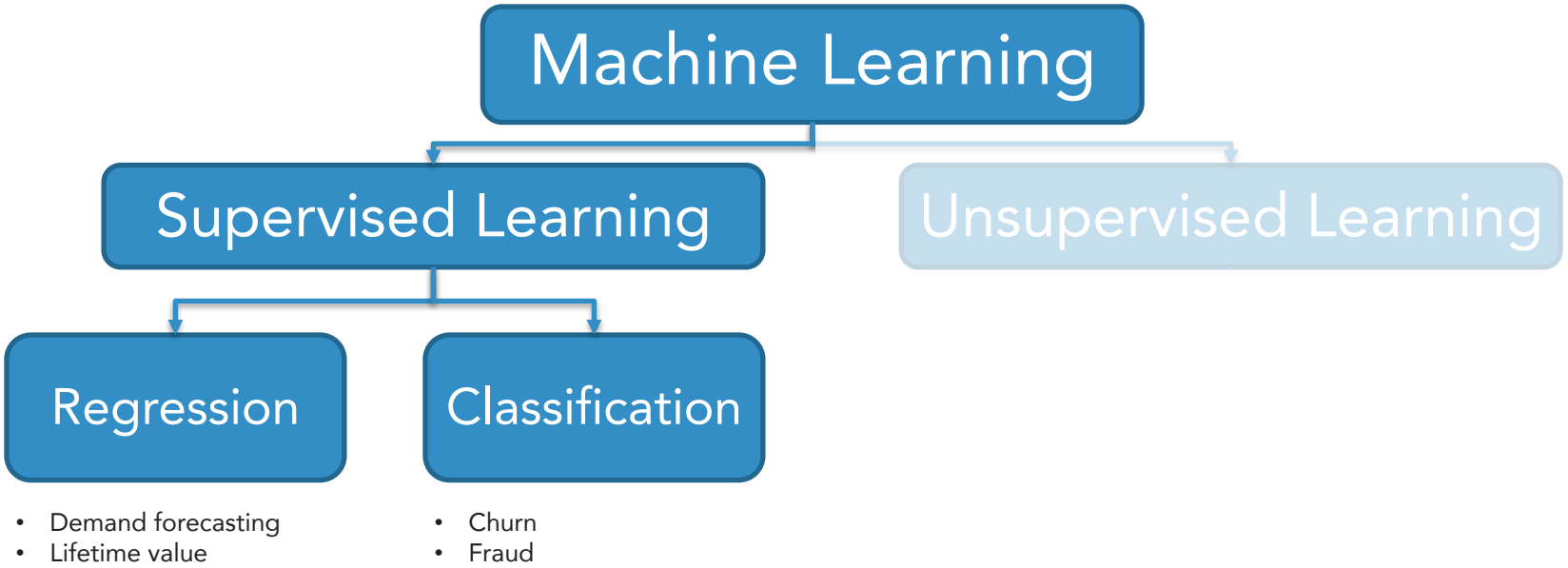
Answer questions like:

- Is this a preferred member?
- Is this purchase fraudulent?

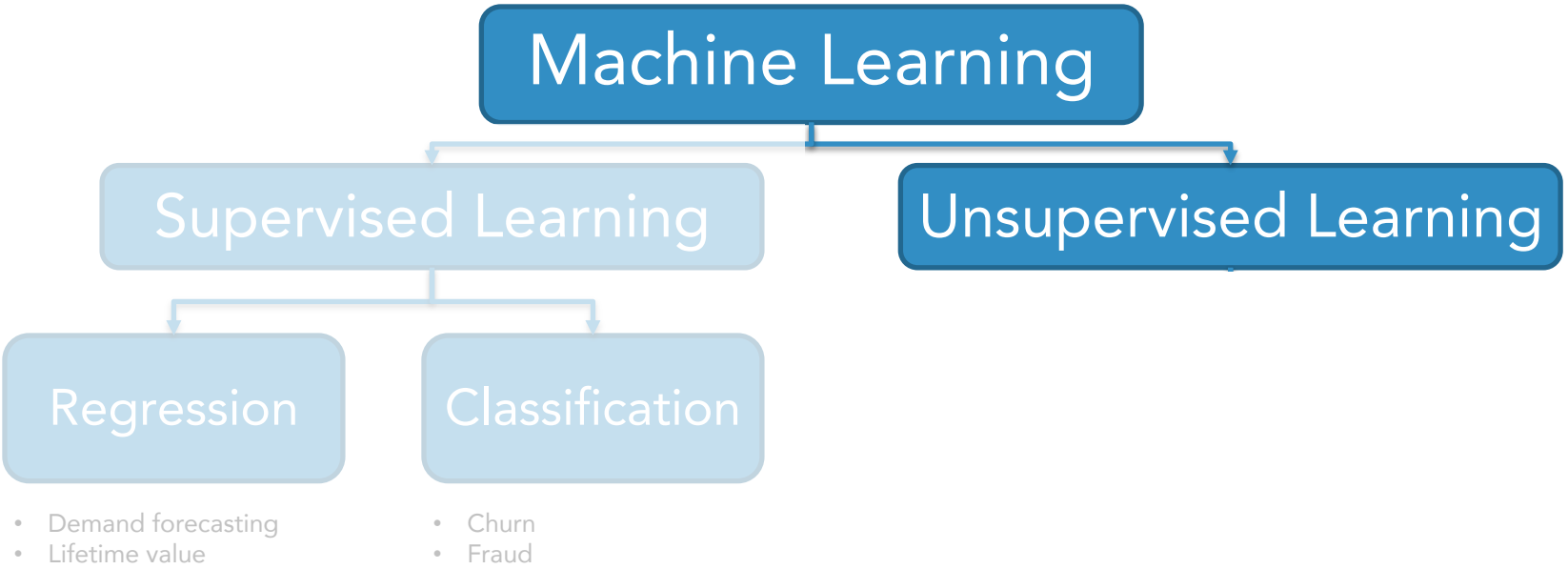
**Applications:** churn modelling, fraud detection



# Machine Learning



# Machine Learning

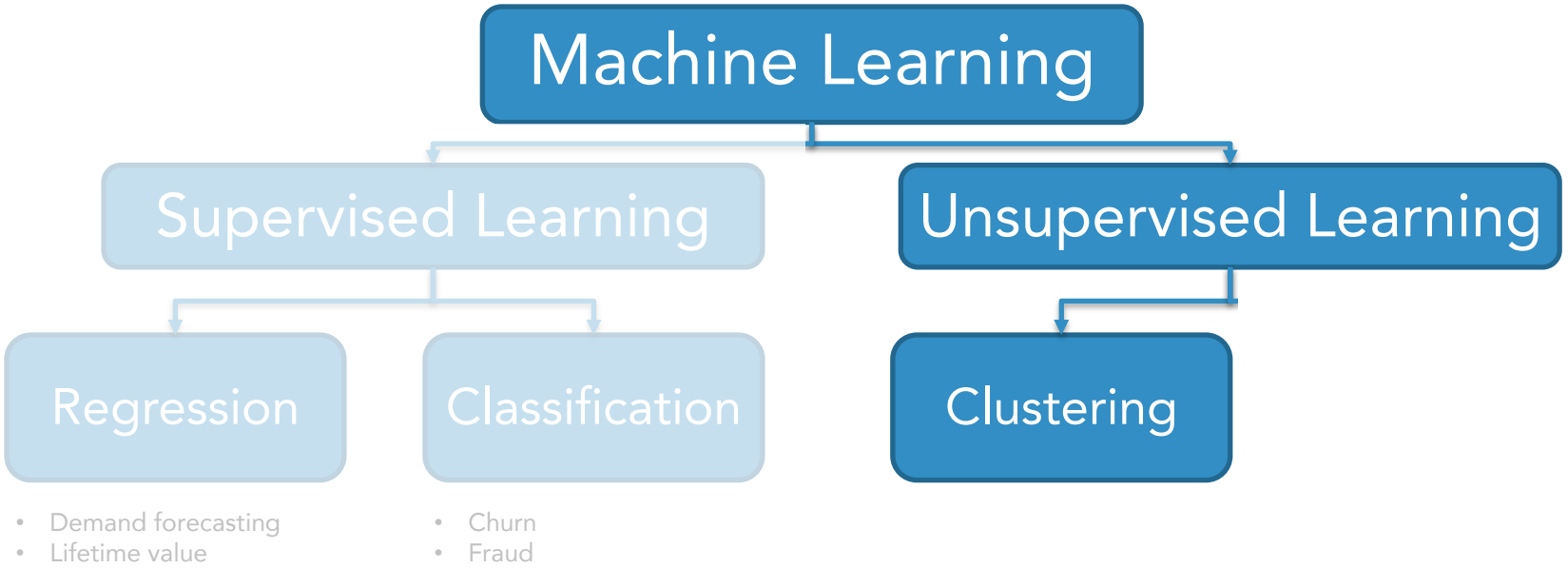


# Unsupervised Learning

Unsupervised Learning

- Machine learning **with*out*** labels
- Uncover the underlying structure of data
- Two major branches: **clustering** and **dimension reduction**

# Machine Learning



# Unsupervised Learning: Clustering

Goal is to uncover naturally occurring groups within the data

Algorithms include k-means, hierarchical agglomerative clustering, DBSCAN

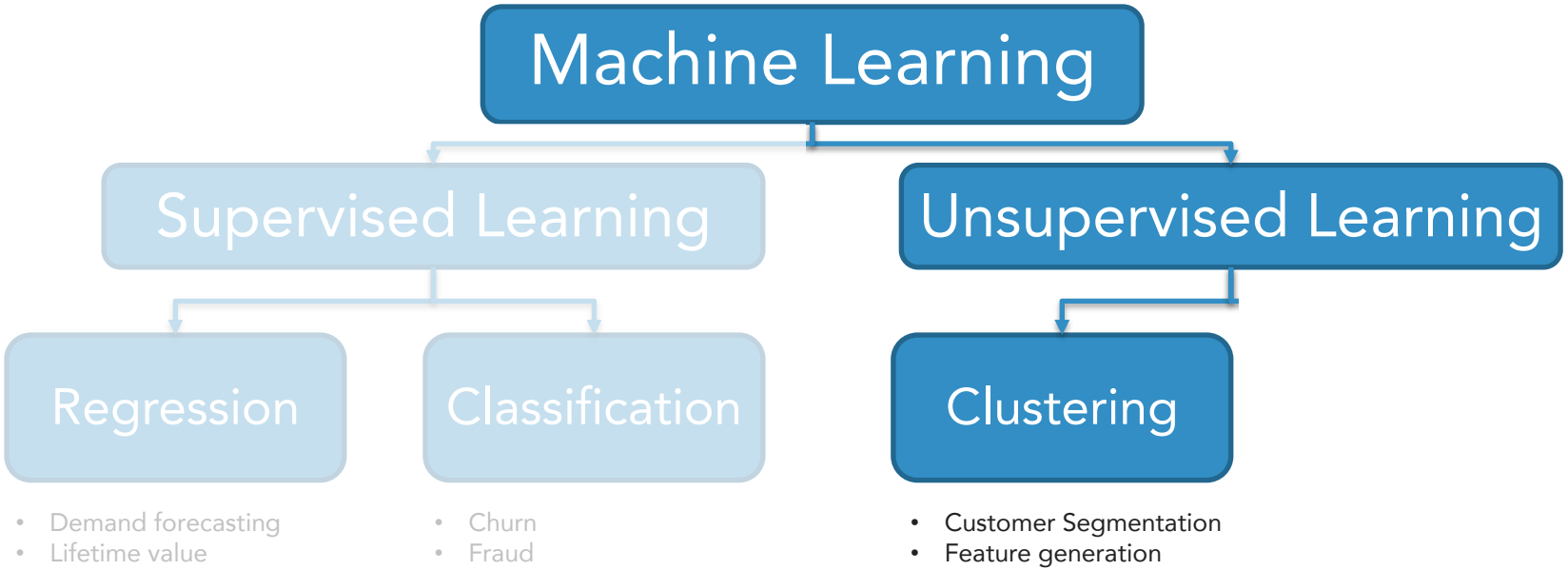
# Unsupervised Learning: Clustering

Answer questions like:

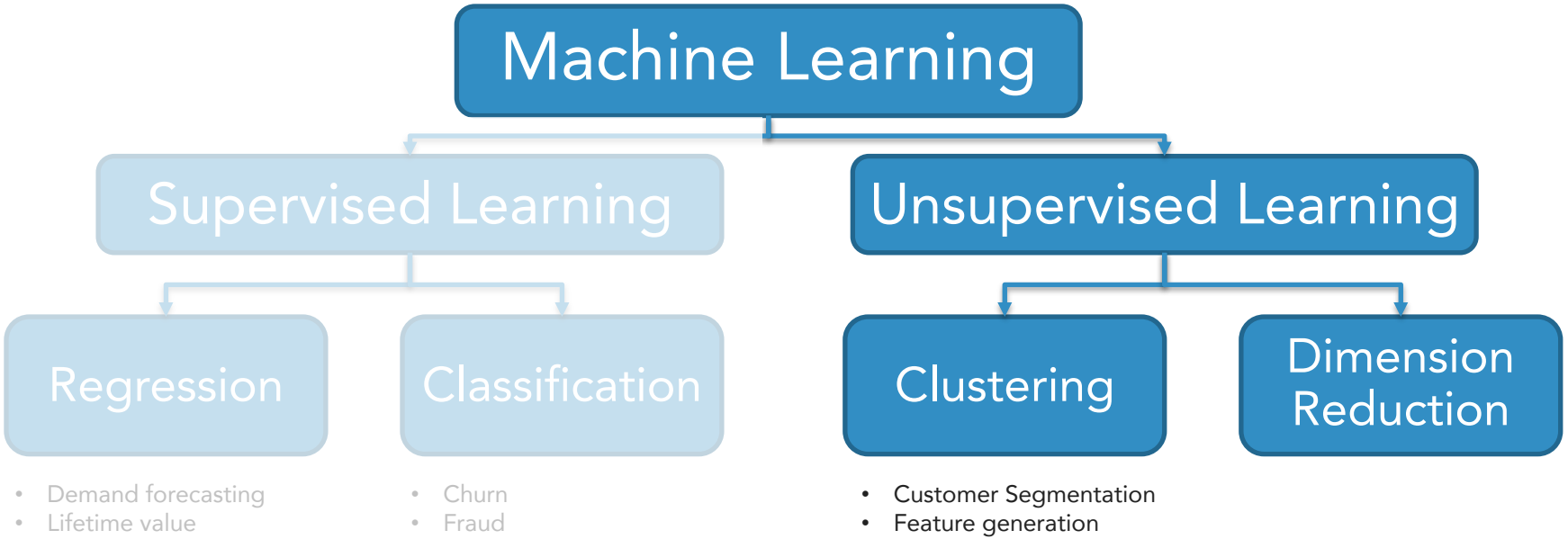
- Do we have groups of similar drivers?
- Can we organize our archives into similar articles?

Applications: customer segmentation, feature generation

# Machine Learning



# Machine Learning



# Unsupervised Learning: Dimension Reduction

Use underlying structure of data to represent it more simply

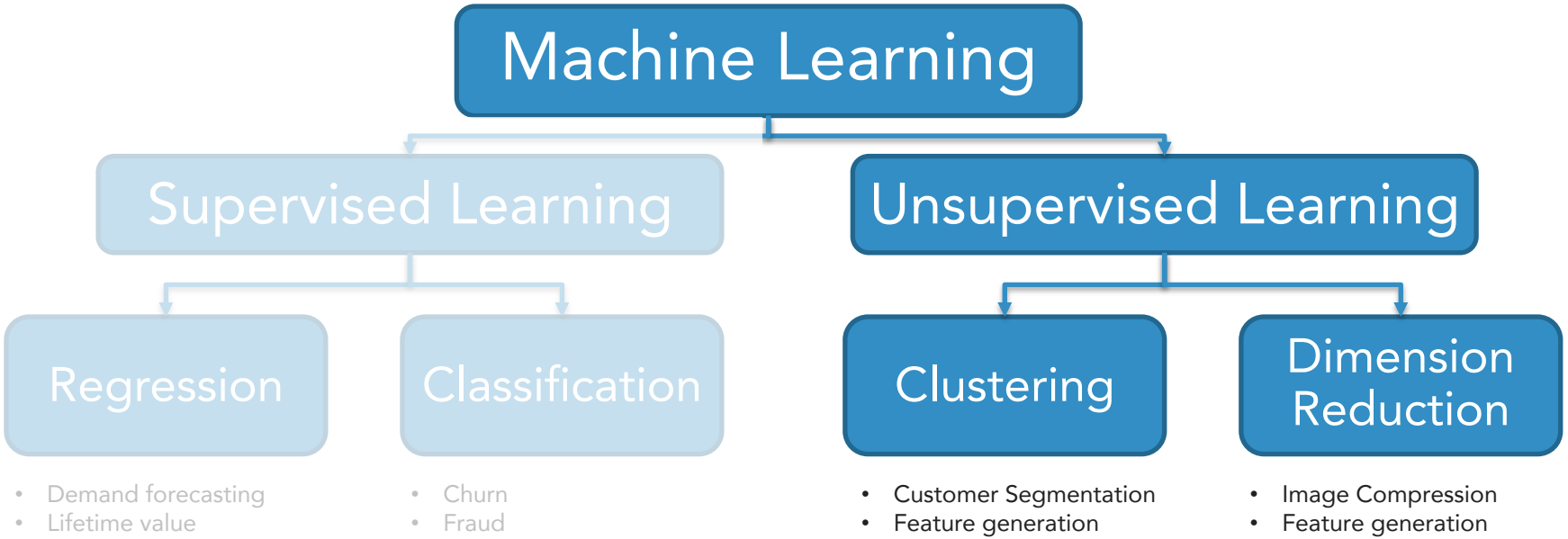
Loses some specificity

Algorithms include PCA, Non-negative Matrix Factorization

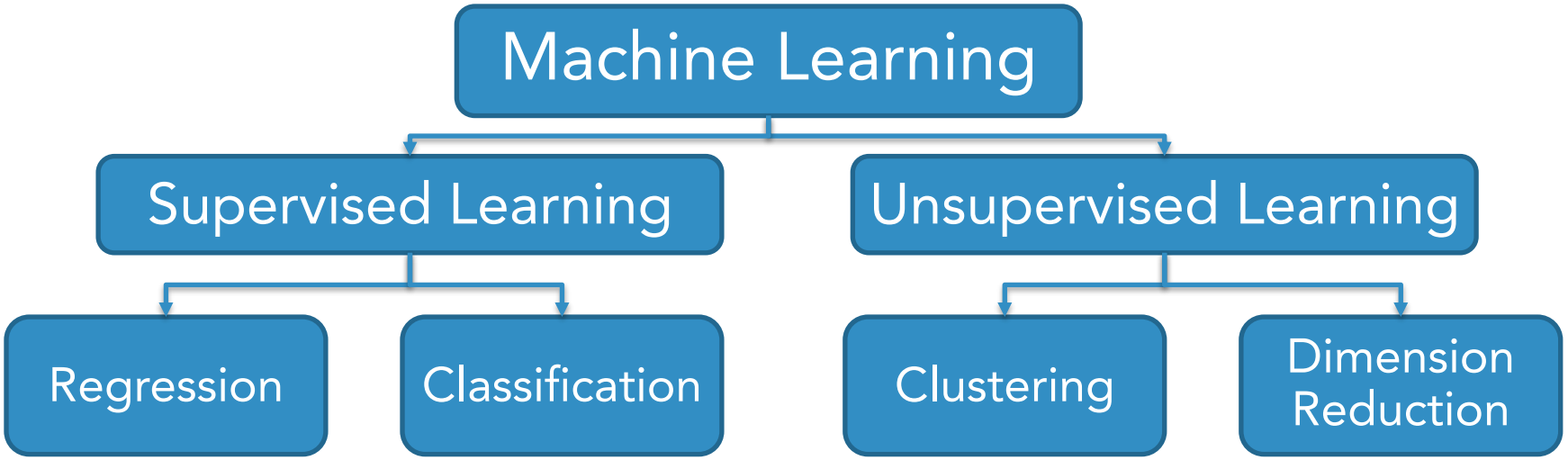
# Unsupervised Learning: Dimension Reduction

Applications: image compression, feature engineering, noise reduction

# Machine Learning



# Machine Learning



# SPECIAL TOPICS

METIS

# Special Topics

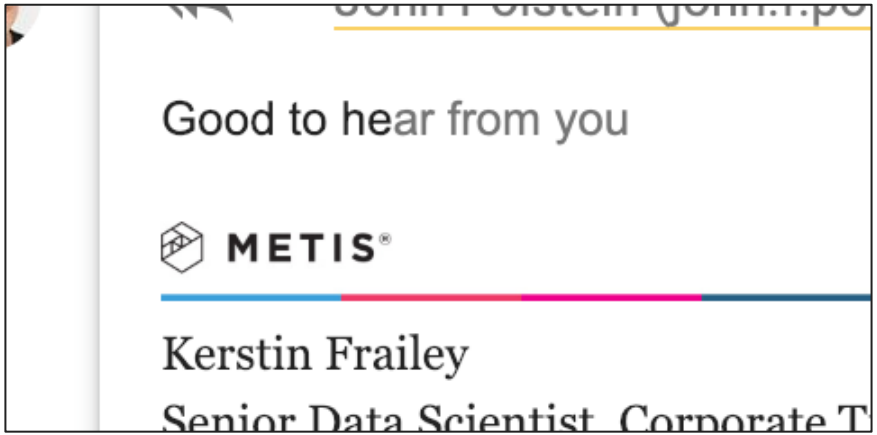
**A/B Testing**: running an "experiment" to test two (or more) alternatives against each other

- 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

- E.g. sentiment analysis, topic modelling
- Everyday application: autocomplete, chatbot, hiring



# 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

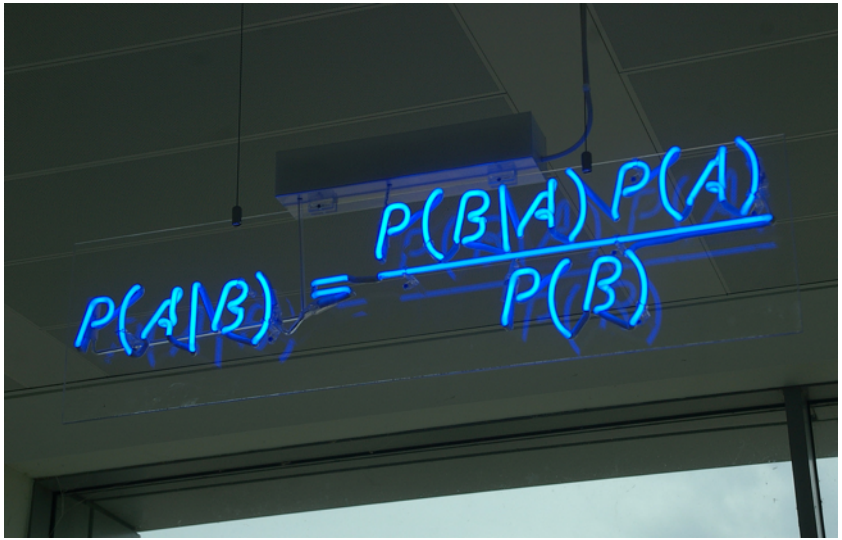
**Computer Vision**: a field of study on how computers can gain information about an environment through images

- 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 include naïve Bayes



# Special Topics

**Big Data**: extremely large data sets; data that cannot be adequately stored and analyzed on a high-performance personal computer

- Not a well-defined term
- Term is no longer en vogue

# 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

# Agenda

A Brief History of Data Science

Basics of Data Science

Analytics and Statistics

Statistics and Machine Learning

Machine Learning and Artificial Intelligence

Special Topics

# 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