Machine learning – introduction

Till Bärnighausen

DS-I Africa Short course

UKZN, Durban

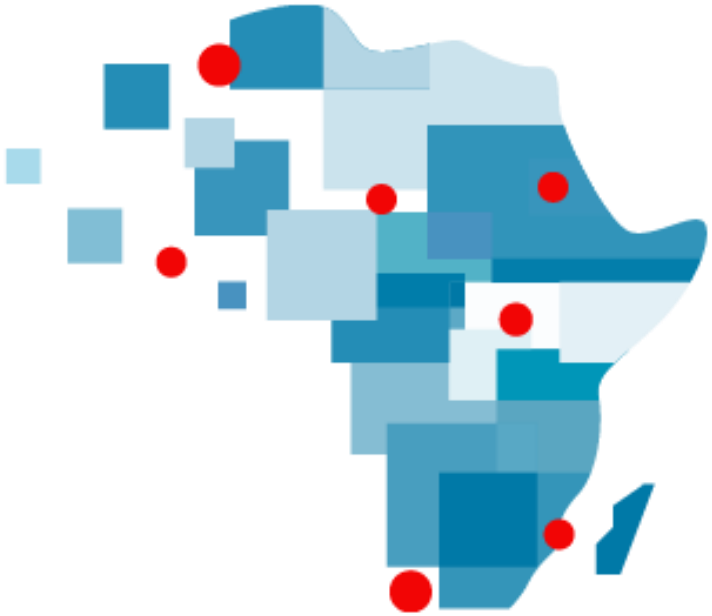
23 January 2023

HEIDELBERG UNIVERSITY HOSPITAL

HEIDELBERG FACULTY OF MEDICINE

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

Alexander von Humboldt Stiftung/Foundation

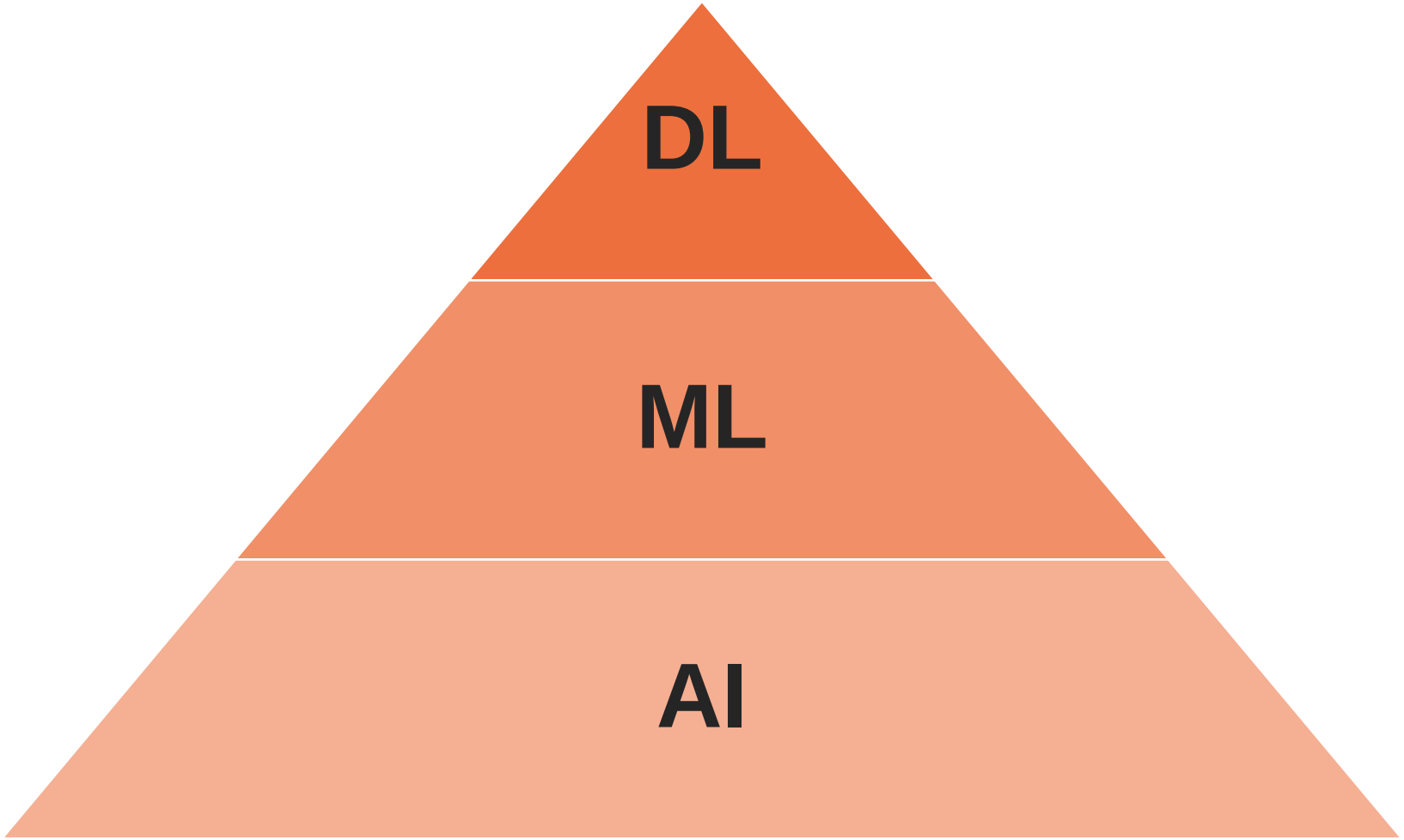


DS-I Africa

Data Science for Health Discovery and Innovation in Africa

Machine learning is a type of AI

POSITION



AI = artificial intelligence, ML = machine learning, DL = deep learning

With layoffs, tech companies are quickening the robot revolution

Google, Microsoft and others in Big Tech have announced massive job cuts, as the companies pivot to artificial intelligence projects. What does it mean?

Jan. 21, 2023, 2:36 AM EAT

By Ja'han Jones

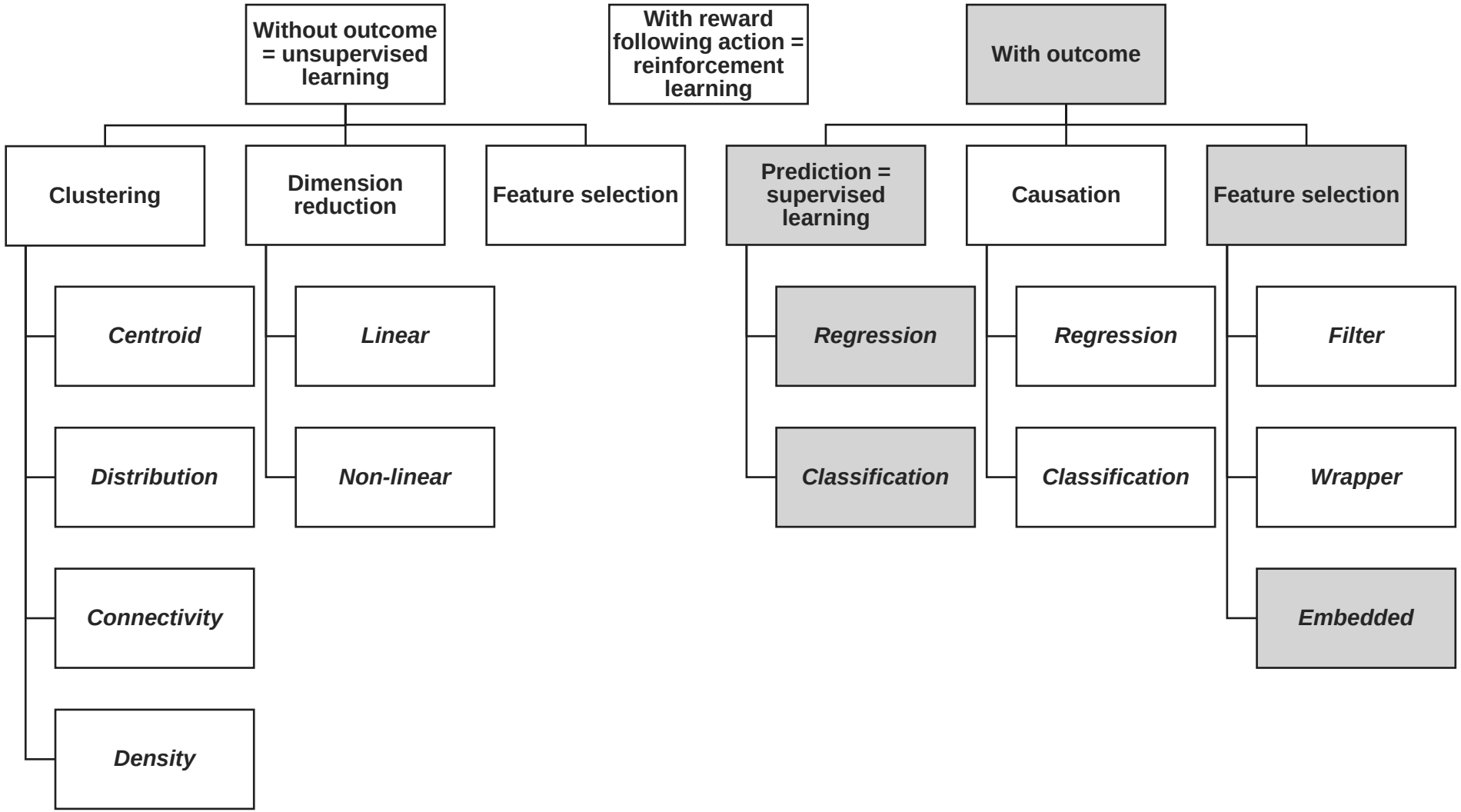
The robot revolution appears to be in full swing.

Google is just the latest company to announce major job cuts said to align with its prioritization of artificial intelligence.

Industries far and wide are pivoting to AI, which – shameless plug – you may have anticipated if you read my end-of-year ReidOut Blog post on 2023 being an important year for the development of artificial intelligence technology.

Where is our course located?

A FRAMEWORK OF MACHINE LEARNING METHODS



Clustering analyses can identify types

OVERVIEW

- ***Centroid*** – k means, k medoids, k prototypes …
- ***Distribution*** – EM, GMM, BMM, …
- ***Connectivity*** – hierarchical, …
- ***Density*** – DBSCAN, ADBSCAN, HDBSCAN, …

EM = expectation-maximization, GMM = Gaussian mixture model, BMM = Bernoulli mixture model, DBSCAN = density-based spatial clustering of applications with noise, ADBSCAN = adaptive DBSCAN, HDBSCAN = hierarchical DBSCAN

Dimension reduction can reduce complexity and identify latent constructs and

OVERVIEW

- ***Linear*** – PCA, SVM, LDA, …
- ***Non-linear*** – kernel PCA, FAMD, t-SNE, …

PCA = principal component analysis, SVM = support vector machine, LDA = linear discrimnant analysis, FAMD = factor analysis for mixed data, t-SNE = t-distributed stochastic neighbor embedding

Feature selection can reduce complexity and cost

OVERVIEW

- **Supervised**
  - ***Filter*** – information gain, correlation, chi-squared, …
  - ***Wrapper*** – forward, backward, stepwise, exhaustive
  - ***Embedded*** – RF, regularization, …
- **Unsupervised –** variance, multicollinearity, incompleteness, …

RF = random forest

Causal analyses is the traditional mainstay of epidemiology

OVERVIEW

	Control of unobserved confounding	Assumptions
• **Experiments**	Complete	Weak
• **Quasi-experiments**		
– Instrumental variable approaches – Regressions discontinuity	Complete	Less weak
– Difference-in-differences approaches – Fixed effects approaches	Partial	Less weak
• **Non-experiments**		
– Regression – Matching – Stratification	None	Strong

Bärnighausen et al. *Journal of Clinical Epidemiology* 2017

Our course focuses on prediction

OVERVIEW

- ***Regression***
  - SVM
  - Decision trees
  - RF
  - kNN
  - Neural networks
  - …
- ***Classification***
  - kNN
  - Logistic regression
  - LDA
  - QDA
  - Naïve Bayesian
  - Neural networks
  - …

SVM = support vector machine, RF = random forest, kNN = k nearest neighbors, LDA = linear discriminant analysis, QDA = quadratic discriminant analysis

Causal and predictive analysis require very different approaches

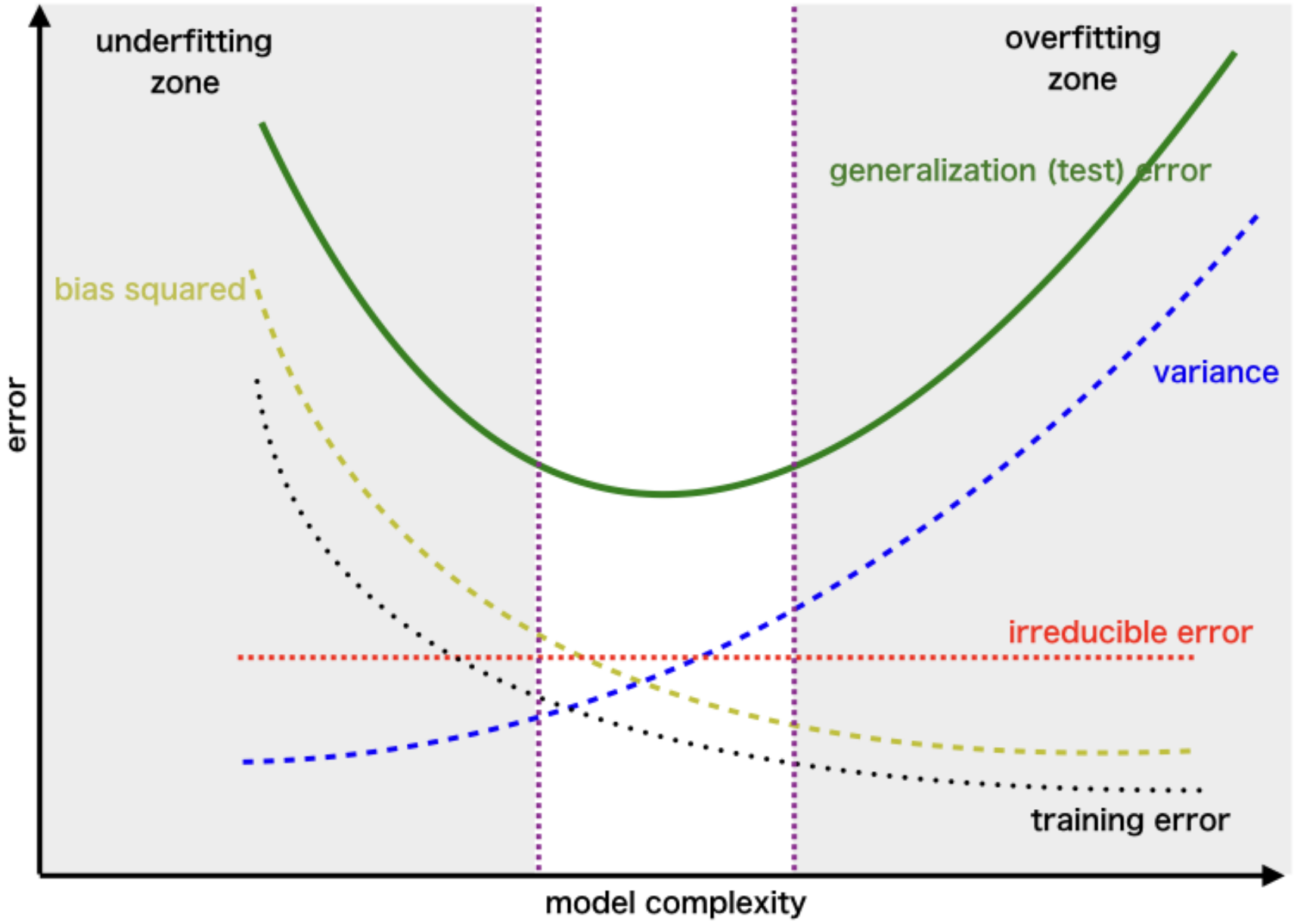
COMPARISON

	Causation	Prediction
Disciplines	• Epidemiology • Economics	• Machine learning • Computer science
Foundation	• Theory-based • Hypothesis testing	• Data-driven
Purposes	• Understanding • Policy guidance • Regulatory approval	• Intervention targeting • Intervention tailoring • Now- and forecasting
Goal of approach	• Minimize/eliminate bias	• Optimize bias-variance trade-off
Approach	• Estimation	• Training-(validation)-testing • Complexity reduction
Example	• Ordinary multiple regression • PSM • IV and RDD	• Regularized multiple regression • kNN • Neural networks

PSM = propensity score matching, IV = instrumental variable analysis, RDD = regression discontinuity design, kNN = k nearest neighbors

For best prediction, we trade-off bias and variance

CONCEPTS

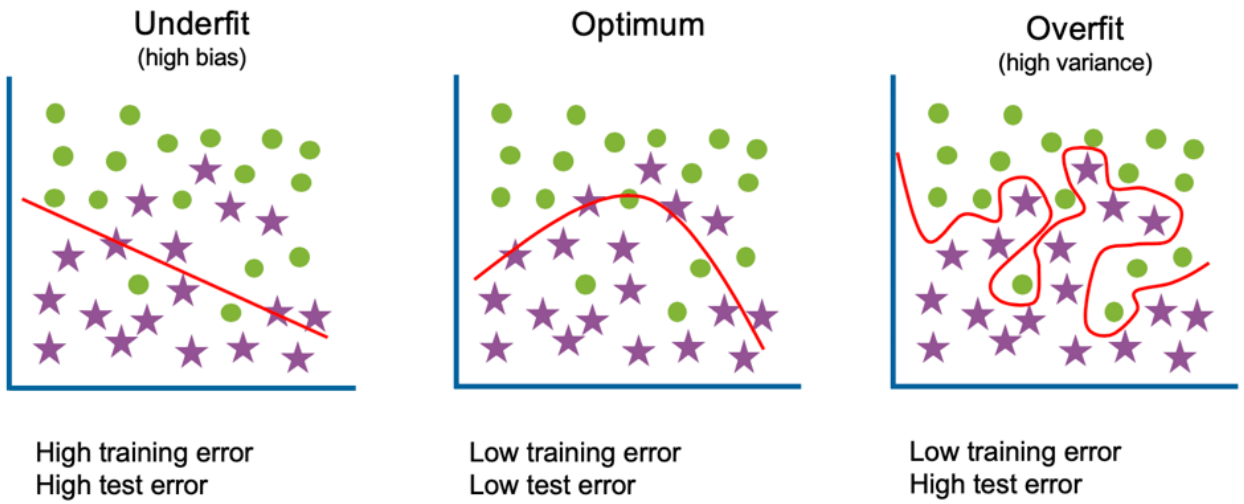


Source: https://www.geeksforgeeks.org/ml-bias-variance-trade-off/

How can we reduce overfitting in ML?

APPROACHES

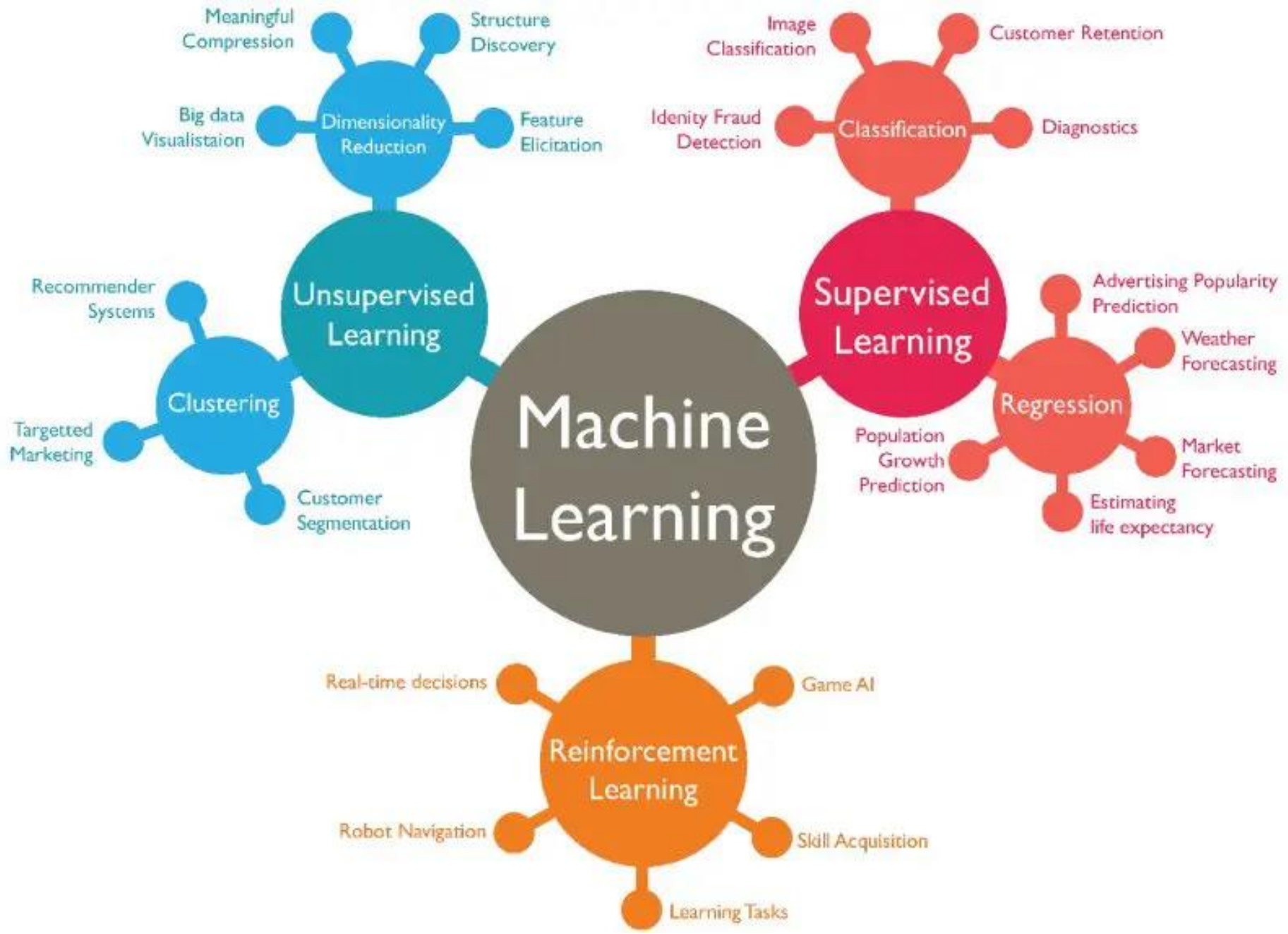
- ***Data***
  - Adding training data
  - Data augmentation
  - Adding noise
- ***Features***
  - Feature selection
  - Dimension reduction
  - Regularization (Ridge, Lasso)
- ***Algorithm-specific***
  - Pruning (trees)
  - Bagging of weakly correlated trees (forests)
  - Dropping out layers (neural networks)



Source:
https://www.linkedin.com/pulse/overfitting-underfitting-machine-learning-ml-concepts-com

What are uses of ML?

EXAMPLES

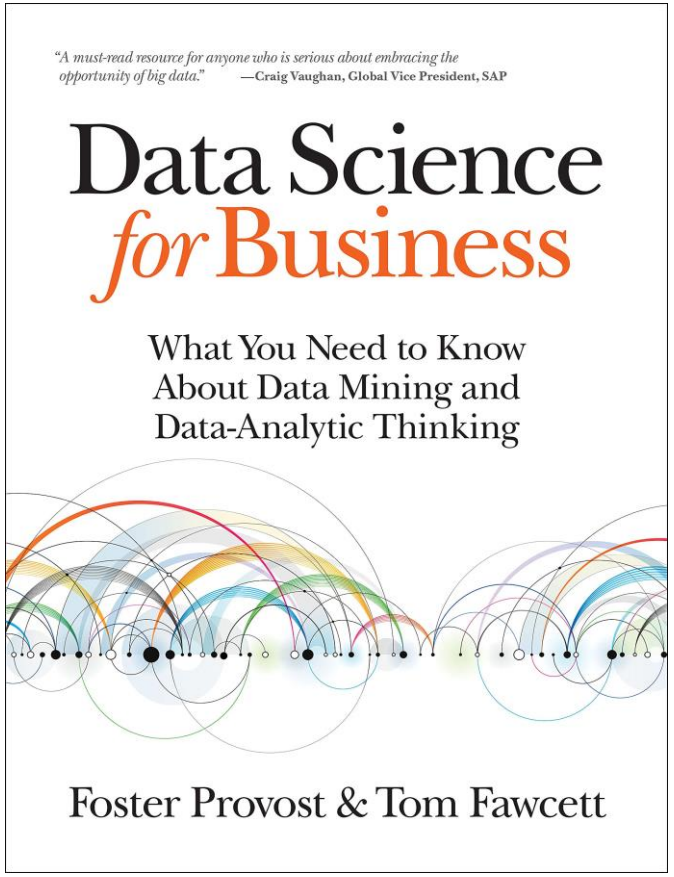


Source: https://datasciencedojo.com/blog/machine-learning-101/

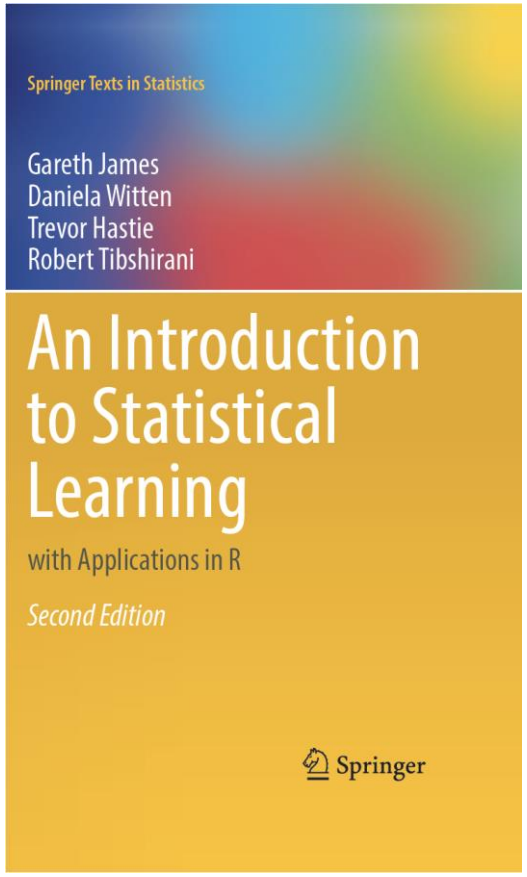
There are excellent introductory textbooks …

EXAMPLES

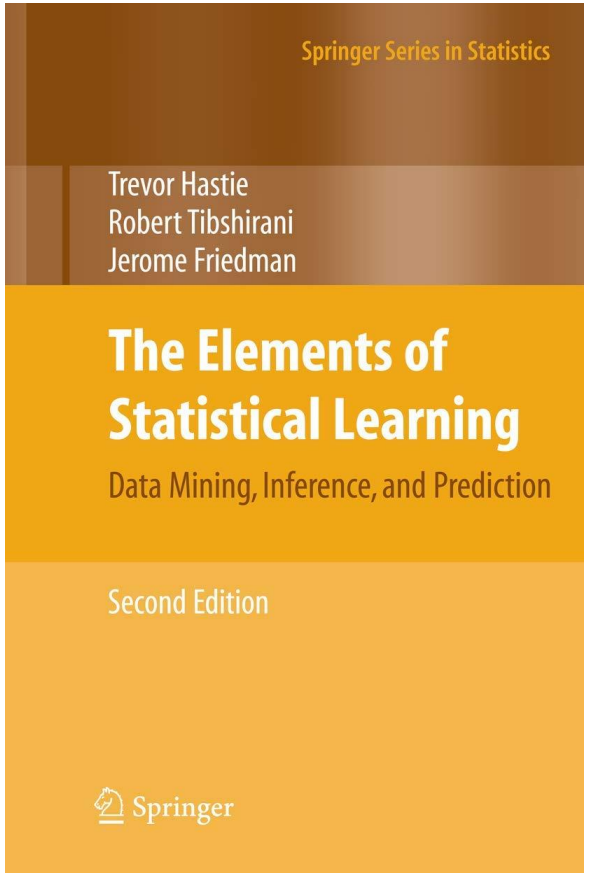
Conceptual



Basic



Basic to intermediary



… and online resources for machine learning

EXAMPLES

In this short course, we will use the R programming language

PROGRAMMING LANGUAGES

- ***Procedural*** – C, Java
- ***Functional*** – R, Scala
- ***Object-oriented*** – Python, Java
- ***Scripting*** – Python, Ruby
- ***Logic*** – Prolog

In the coming two weeks, we will cover ML concepts, concrete methods, and applications

COURSE OVERVIEW

- ***General approaches***
  - Feature selection
  - Validation / cross-validation
  - Parameter tuning to reduce overfitting
  - Performance metrics
- ***Concrete methods***
  - Regression / logistic regression
  - Ridge and lasso regression
  - kNN
  - Naïve Bayes
  - LDA / QDA
  - SVM
  - Trees and forests
  - Boosting

In the coming two weeks, we will cover both ML general approaches and concrete predictive methods

COURSE OVERVIEW

- ***General approaches***
  - Feature selection
  - Validation / cross-validation
  - Parameter tuning to reduce overfitting
  - **Performance metrics**

- ***Concrete predictive methods***
  - Regression / **logistic regression**
  - Ridge and lasso regression
  - **kNN**
  - **Naïve Bayes**
  - **LDA / QDA**
  - SVM
  - Trees and forests
  - Boosting