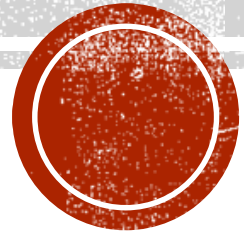
DATA SCIENCE AND MACHINE LEARNING

How to build a career

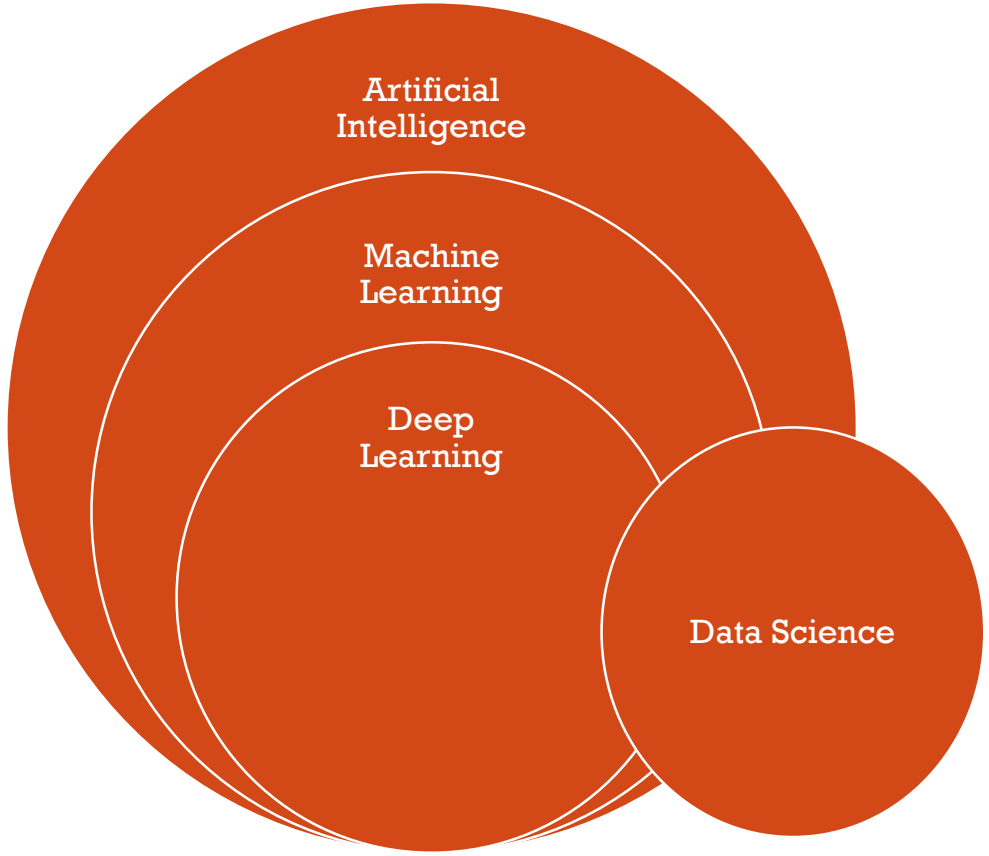


V. Partovi Nia, Ph.D.
Senior Machine Learning Scientist, Noah's Ark Research Lab of Huawei Technologies
Adjunct Professor, Ecole Polytchnique de Montreal
September 19, 2019

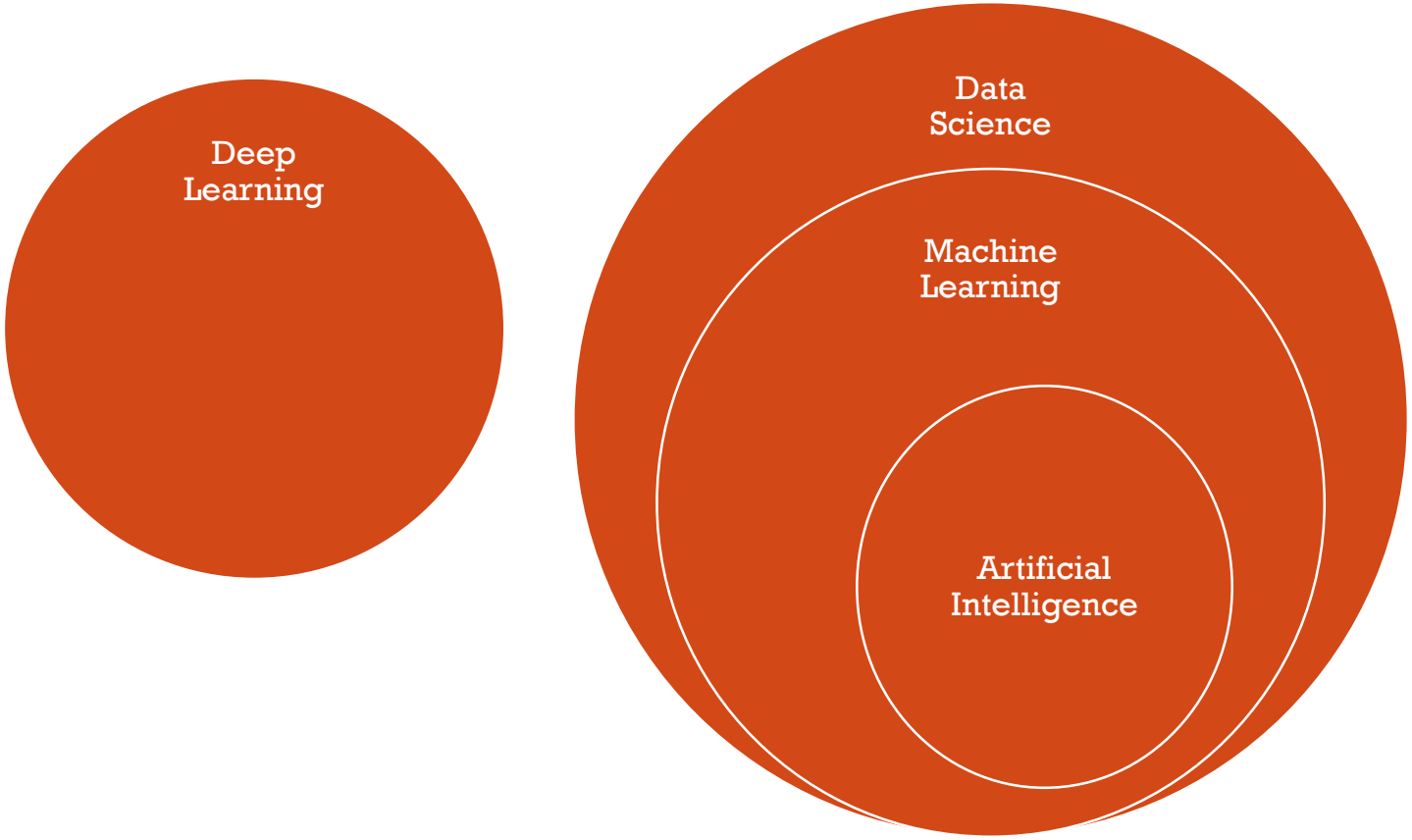
WHAT DO THEY HAVE IN COMMON?

- Data
  - Data intuition
  - Domain knowledge
  - Data visualization
  - Distributed computation
  - Programming
- Algorithm
  - Efficient computation
  - Computing hardware (CPU, GPU, Cluster, Cloud, etc)
  - Computational complexity
  - Professional coding (recursive methods, grasp the coding culture, commenting codes, reading others' codes, share with github, use svn, etc)

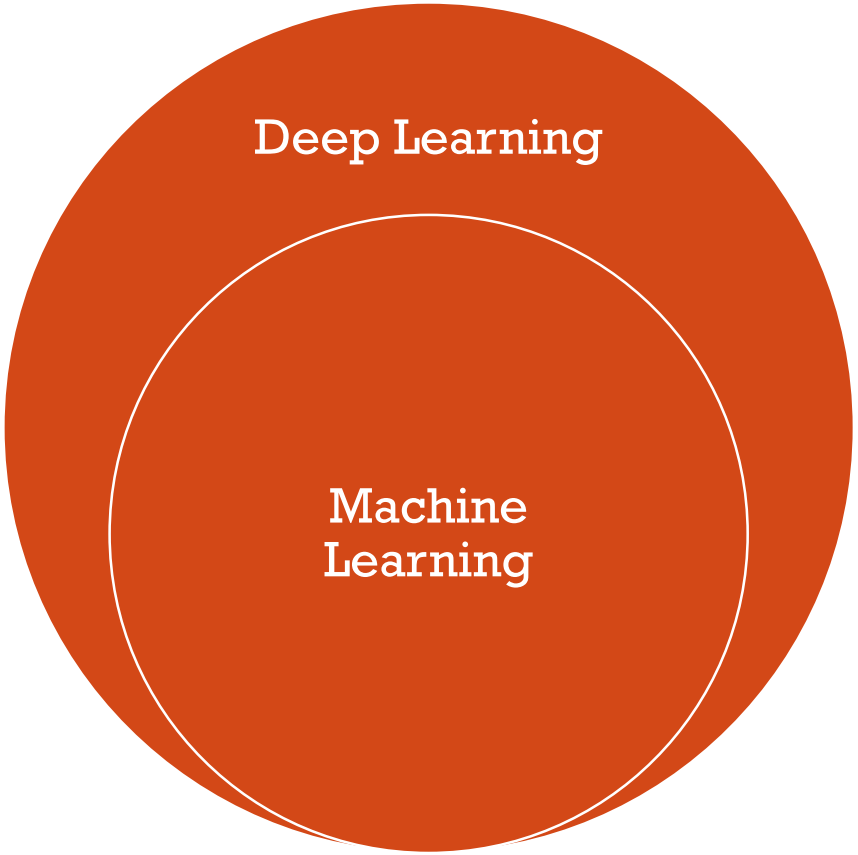
WHAT PEOPLE SEE IN ACADEMIA



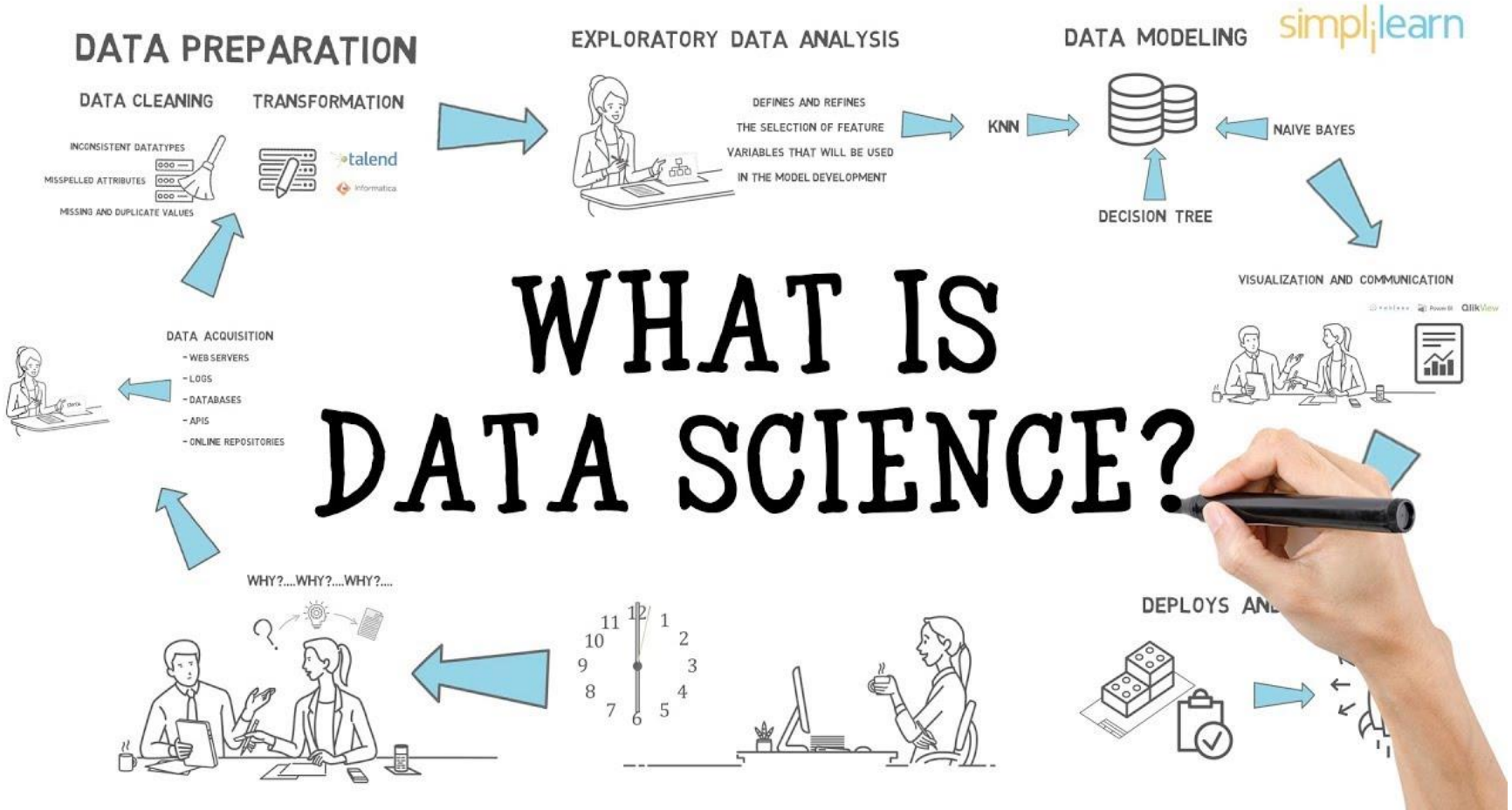
WHAT HAPPENS IN PRACTICE



HOW YOU SEE IT



DATA SCIENCE PIPELINE



WHAT IS DATA SCIENCE?

- The fields that provide the foundation
  - Statistics
  - Computer science
  - Applied mathematics (Optimization)
  - Physics
- The fields that provide concrete skills
  - Computer engineering
  - Electrical engineering
  - Biomedical engineering
  - etc

HOW TO GET STRATED?

- If "bachelor's degree" == False then stop
- If "bachelor's degree"=="qualitative" then stop.
- If "bachelor's degree" == "quantitative" then "master's degree".

Master's program in Montreal:

- MILA: course-based master's
- UdeM: DIRO department
- McGill: computer science department
- McGill: electrical and computer engineering department
- Polytechnique: MAGI department (mathematics program)
- Concordia: computer science and software engineering department

PhD program in Montreal

- Find a good advisor who recently published in top AI/ML/DS conferences: ICDM, Neurips, ICML, AAAI, ACM, ICLR, etc.

WHAT SKILLS DO I NEED FOR DATA SCIENCE?

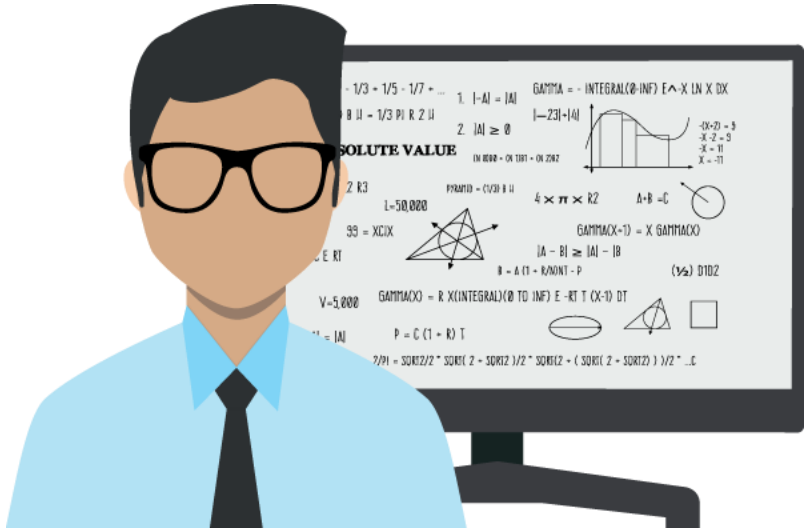
- Understand Linear Algebra:
  - what is SVD?
  - How to solve a system of linear equations?
  - What is a linear hyperplane?
  - What is a linear manifold?
- Understand Data:
  - What is Pearson correlation?
  - What is scale, ordinal, nominal data?
  - What is the difference between covariance and correlation?
  - What is discrete correlation?
  - What is linear regression?
  - What is logistic regression?
  - What is kernel density?
  - What is PCA?
  - What are different data structures? (arrays, trees, graphs, hash tables, etc)

- Understand Algorithm
  - What is k-means?
  - What is dendrogram?
  - What is classification and regression tree (CART)?
  - What is random forest?
  - What is bagging?
  - What is boosting?
  - What is bootstrapping?
  - How do you fit logistic regression using only linear regression?
  - How do you fit logistic regression using coordinate descent?
  - How do you fit logistic regression using gradient descent?
  - How do you fit logisti regression using stochastic gradient descent?
  - How do you adjust bias in observational data?
- Access Data
  - Database (SQL, Oracle, etc)
  - Web crawling, pattern matching "grep", etc
  - Read CSV file
  - Manipulate data (create tables out of tables, tidyverse, what must be in row what must be in column?)
  - Guess how much does it take!
  - Print a counter once a while!

WHAT ARE THE MAJOR SKILLS IN INDUSTRY?

- Data understanding! (not taught in any course), you learn by spending time in a certain domain for along time.
- Data manipulation! (not taught in any course), you learn by analyzing many data sets of various domains.
- Data visualization (not taught in any course), you learn by reading blogs and technical news online.
- Data analysis (partially taught in few courses), you learn from your senior.
- Machine learning (many resources), you learn in courses, you apply them if you prepared the right data.
- Programming (online resources, github). Main languages are Python (numpy, scipy, matplotlib, sklearn, pandas, statsmodels), R, Julia, scala.

DATA SCIENTIST



- Gather data
- Prepare data (row-column)
- Visualize
- Analyze
- Argue you find a new correlation
- Argue this is not a correlation and is a causation
- Find an improvement strategy to improve data quality
- Automate (give it to a machine learning engineer)
- Change your dataset!

MACHINE LEARNING ENGINEER

- Gather Data
- Prepare Data (row-column)
- Run Algorithm
- Optimize Hyperparameters
- Automate the process
- Change to a new project

MACHINE LEARNING RESEARCHER

- Read papers
- Implement state-of-the-art
- Evaluate on benchmark
- Run on company's data
- Change the state-of-the-art algorithm
- Patent your algorithm
- Re-run on benchmark
- Show a slight improvement
- Publish in an AI/ML conference

MACHINE LEARNING RESEARCHER SKILLS

- Understanding of foundation
  - What is least squares?
  - What is sigmoid?
  - What is cross-entropy?
  - Why you must not invert a matrix?
  - Why deep learning is popular?
  - Where deep learning fails?
  - What is the loss function?
  - What is regularization?
  - What is over/under fitting?
- Implementation skills
  - Knows how to work with github/gitlab
  - Knows at least Python (numpy, pandas, scipy)
  - Implemented some of the state-of-the-art
  - Knows PyTorch or Tensorflow

I WANT TO CHANGE MY JOB TO DS/ML

- Free your time
- Get some coursera courses (but do not trust them)
- Test your skills on kaggle.com (bring yourself in ¼ top). Start with the "titanic challenge", then increase the complexity.
- Try McGill Continuing studies program
  - Data Science,
  - Artificial Intelligence

Do exercises of these books!