Introduction to ML

DS-3000A/DS-9000

Fall 2023

Alireza Fazeli

Linkedin

Email

WSC-272

Why we are here?

- Get acquainted with machine learning and statistical methods for data analysis through applied examples.

- Get exposed to topics related to statistical learning such as, Linear Regression, Logistic Regression, Discriminant Analysis, Model Selection and Regularization, Cross Validation, Tree Based Methods, and Clustering.

Caveat!

Are you also taking (or have taken) CS 4414 / CS9637A/ CS9114A / SE4460B / SS 3850G ?

This course is identical to those above. We are improving data science teaching across all faculties and these courses are merged now. If you need both (i.e., you are majoring in Data Science), you can choose any other course taught in CS or SS at 3000 level or above. A suggestion:

Artificial Intelligence II: More advanced models and more in depth with the algorithmic side of data science!

Support Channel

Please post your enquiries to the most relevant topic that you see on OWL Forums.

I and the TA's will monitor the Forum and try to answer your questions as soon as we can.

We have learnt from the past that it is the most efficient way of providing support.

Also, there will be TA and instructor office hours (TBD).

Evaluation

10 (or may be 9) weekly assignments:

30% of final course grade. Posted every Friday with submission deadline on the next Friday.

Midterm exam:

30% of final course grade
Open-book coding

Final exam:

40% of final course grade
Part 1: Closed-book multiple-choice
Part 2: Open-book coding

For the coding parts you will be uploading a Jupyter notebook to OWL. And yes, you do need a personal laptop throughout the course (even for the midterm and final exams).

Class Format

- The attempt will be to have Tuesday sessions devoted to lectures and Thursday sessions to computer labs. Exceptions could happen.

  Lectures and labs run in the same room, *i.e.*, UCC-56
  3:30 – 5:30PM

Self-Assessment

You are expected to already know programming in Python and some relevant libraries such as Pandas and Numpy as well as a sufficient level of Linear Algebra, Calculus, and Statistics. To check this, we strongly encourage you to take the self-assessment test. We will not grade it. This is just to help you decide whether to stay in the course.

You can access it at OWL/Lessons/Introduction and Preliminaries

Python Recap.

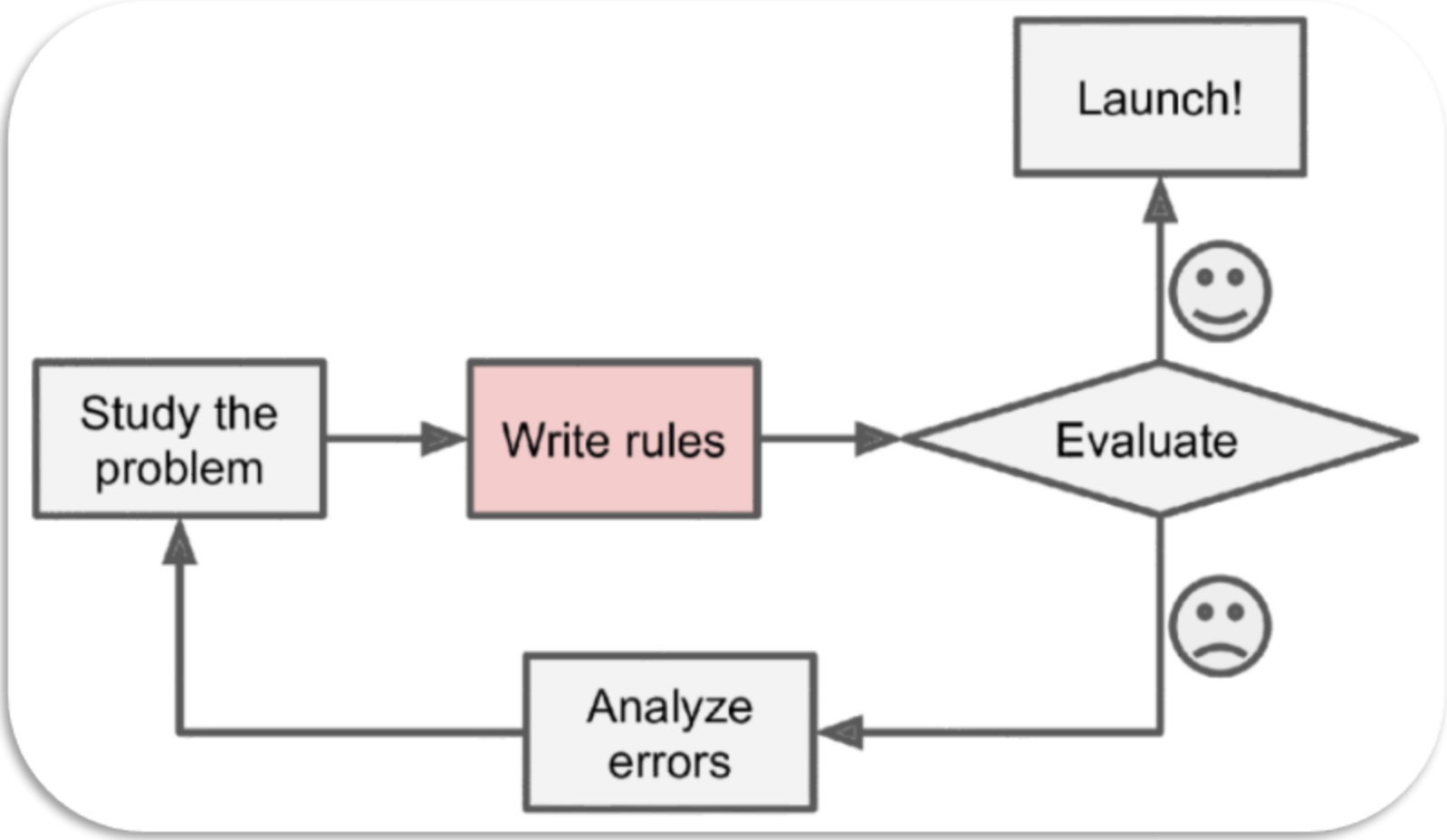
We strongly recommend that you also practice the Python recap notebooks accessible at:

OWL/Lessons/Introduction and Preliminaries

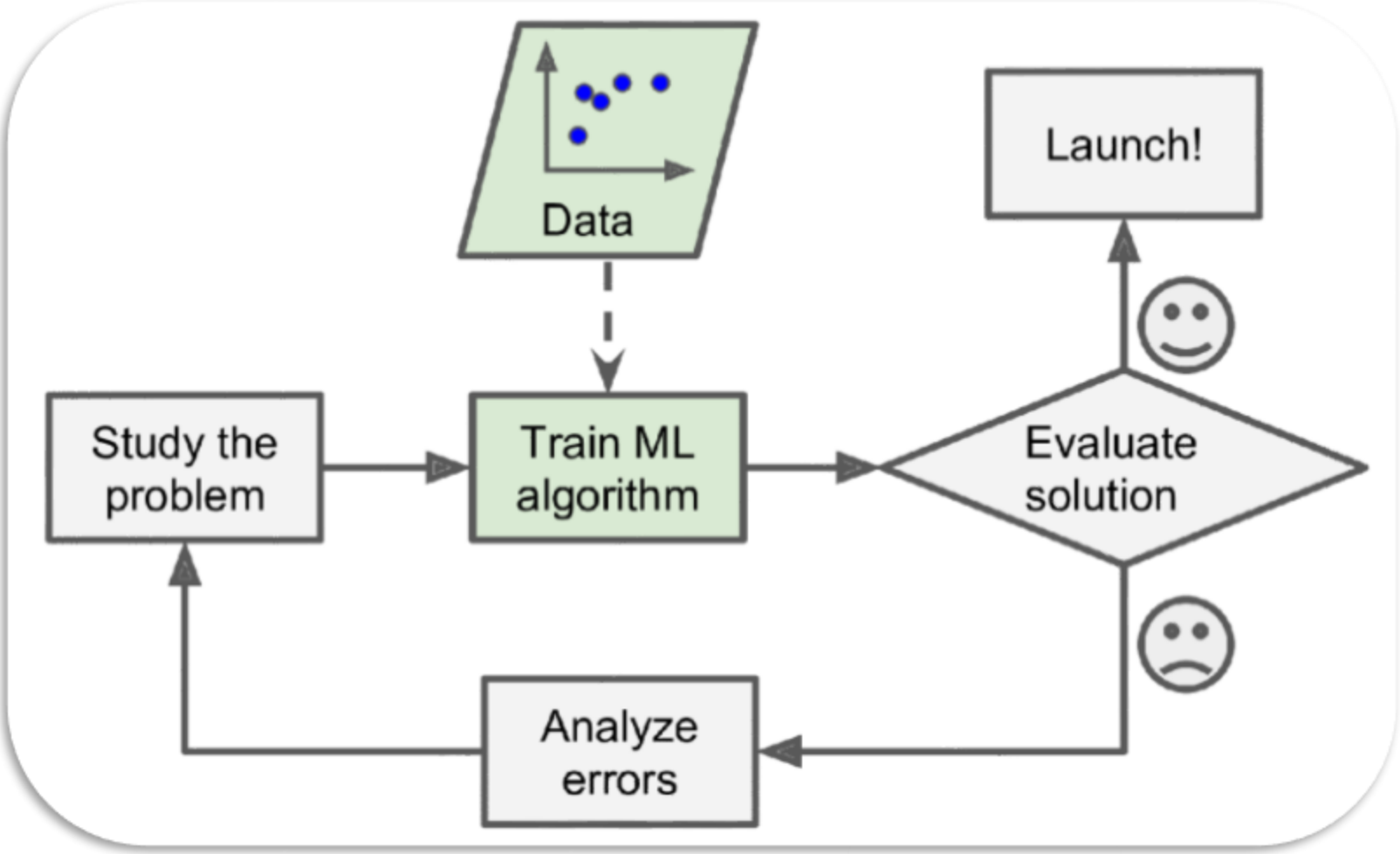
What is ML?

- A new paradigm in computing so machines can learn from data
- A field of study that gives computers the ability to learn without being explicitly programmed. (Arthur Samuel, 1959)
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. (Tom Mitchell, 1997)

Traditional Computing Paradigm

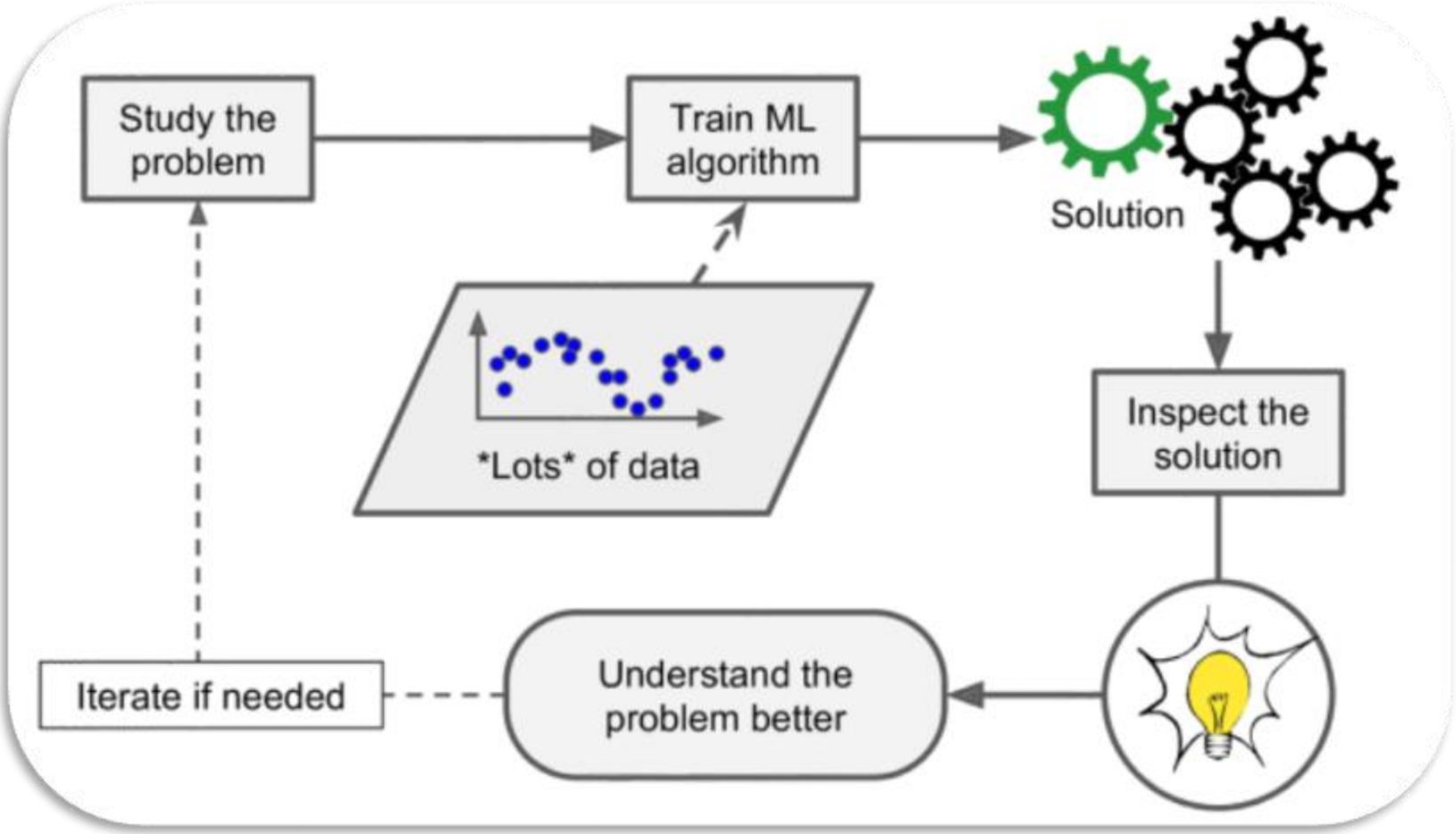


Machine Learning Paradigm

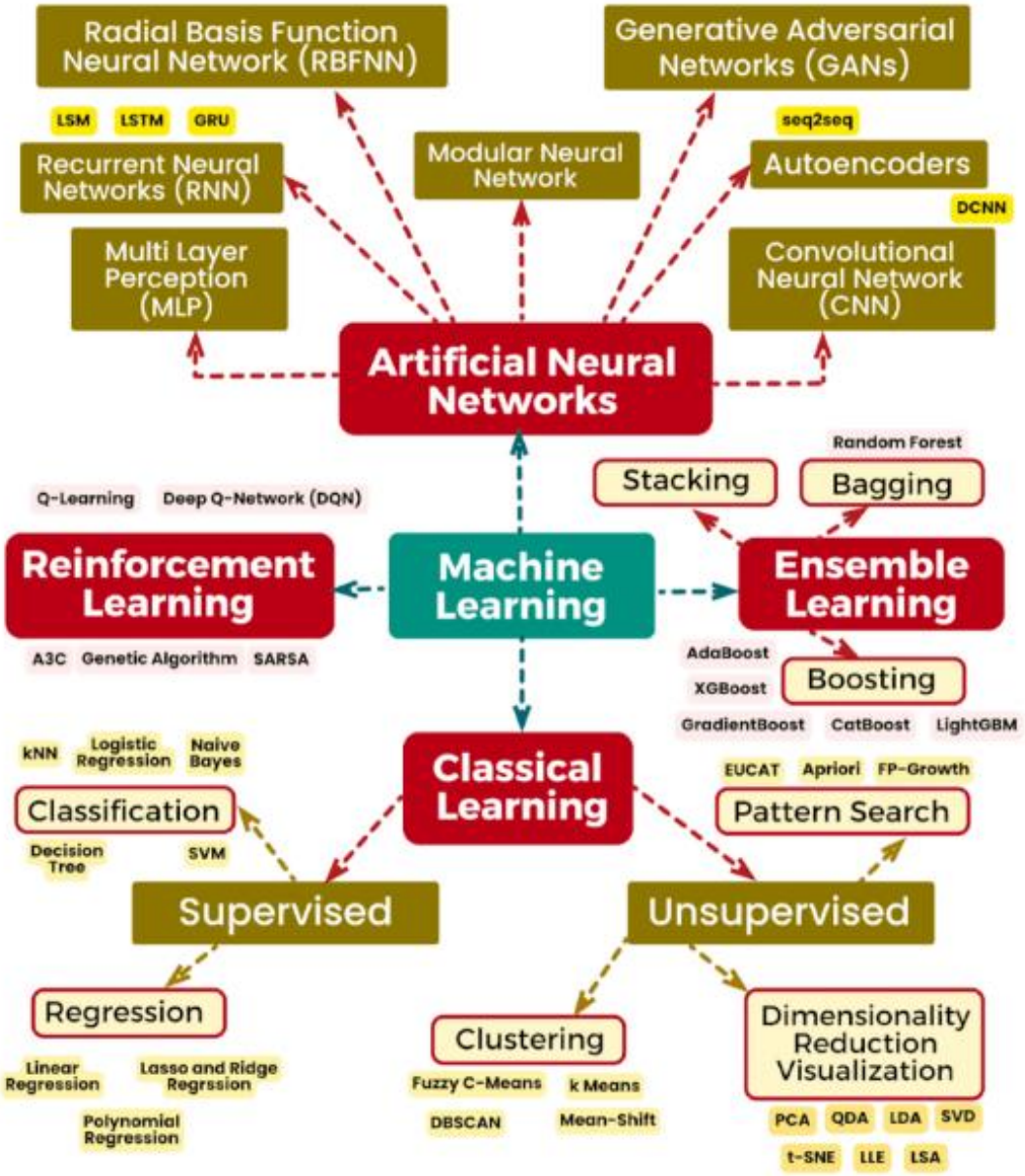


Data Mining

ML can help humans learn.



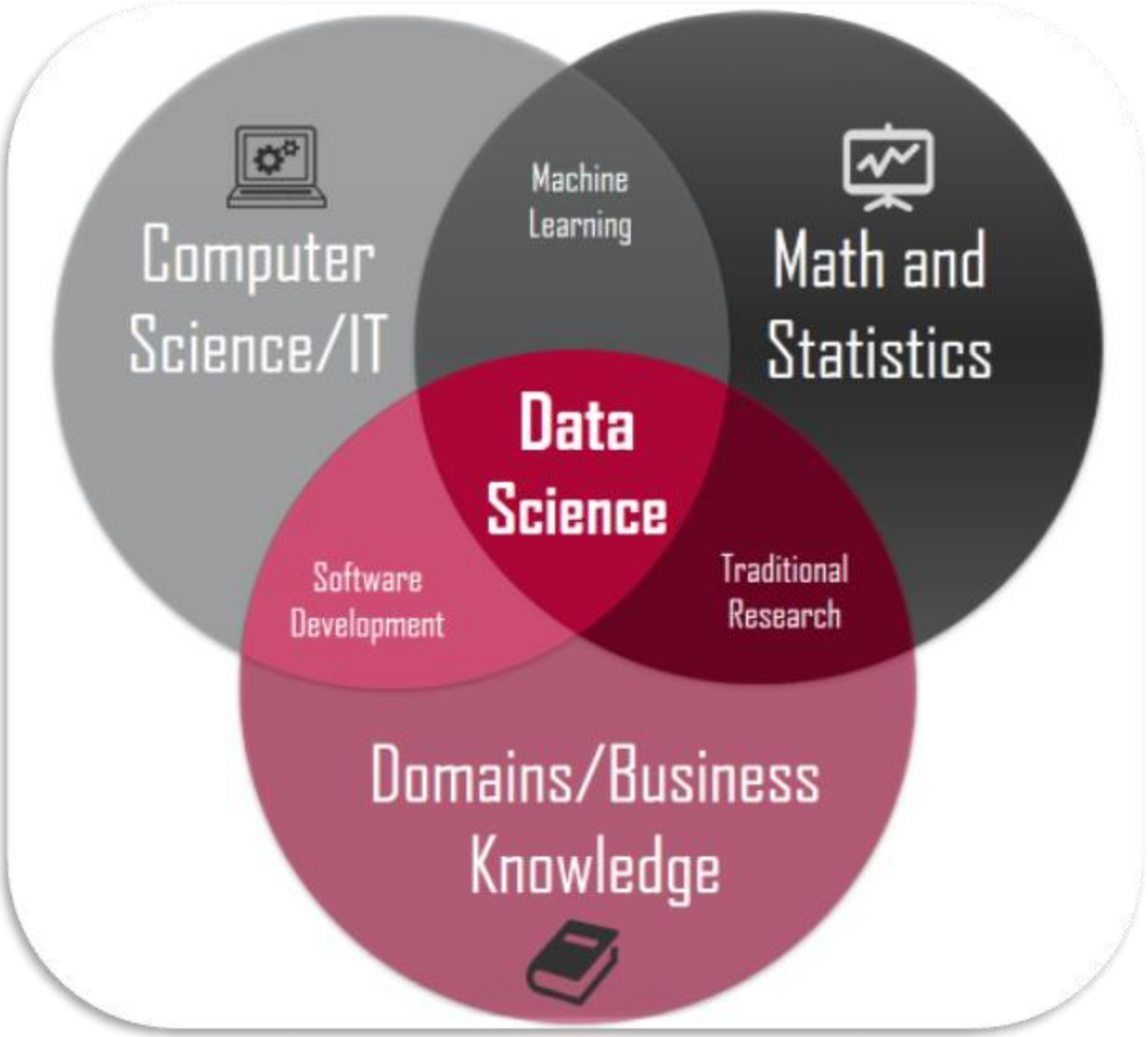
ML Algorithms



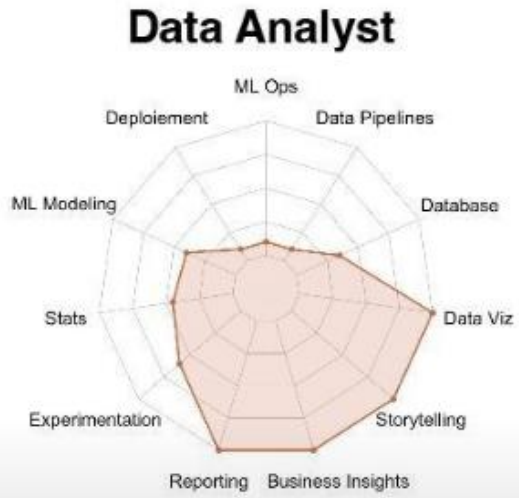
Topics

- Supervised Learning and Model Fitting
- Statistics, Prediction, and Maximum Likelihood
- Introduce test set/out-of-sample idea
- Classification, Evaluation, Logistic regression Regularization, Multi-class problems
- Estimating Performance, Quantifying Uncertainty on parameter estimates and on model predictions
- Test error, Cross-validation, Model Selection, Bias-Variance tradeoff
- Feature Selection and Regularization (L1 and L2)
- Trees, Random Forest
- Neural Networks
- Dimensionality reduction
- Clustering
- Fairness
- Accelerated DS

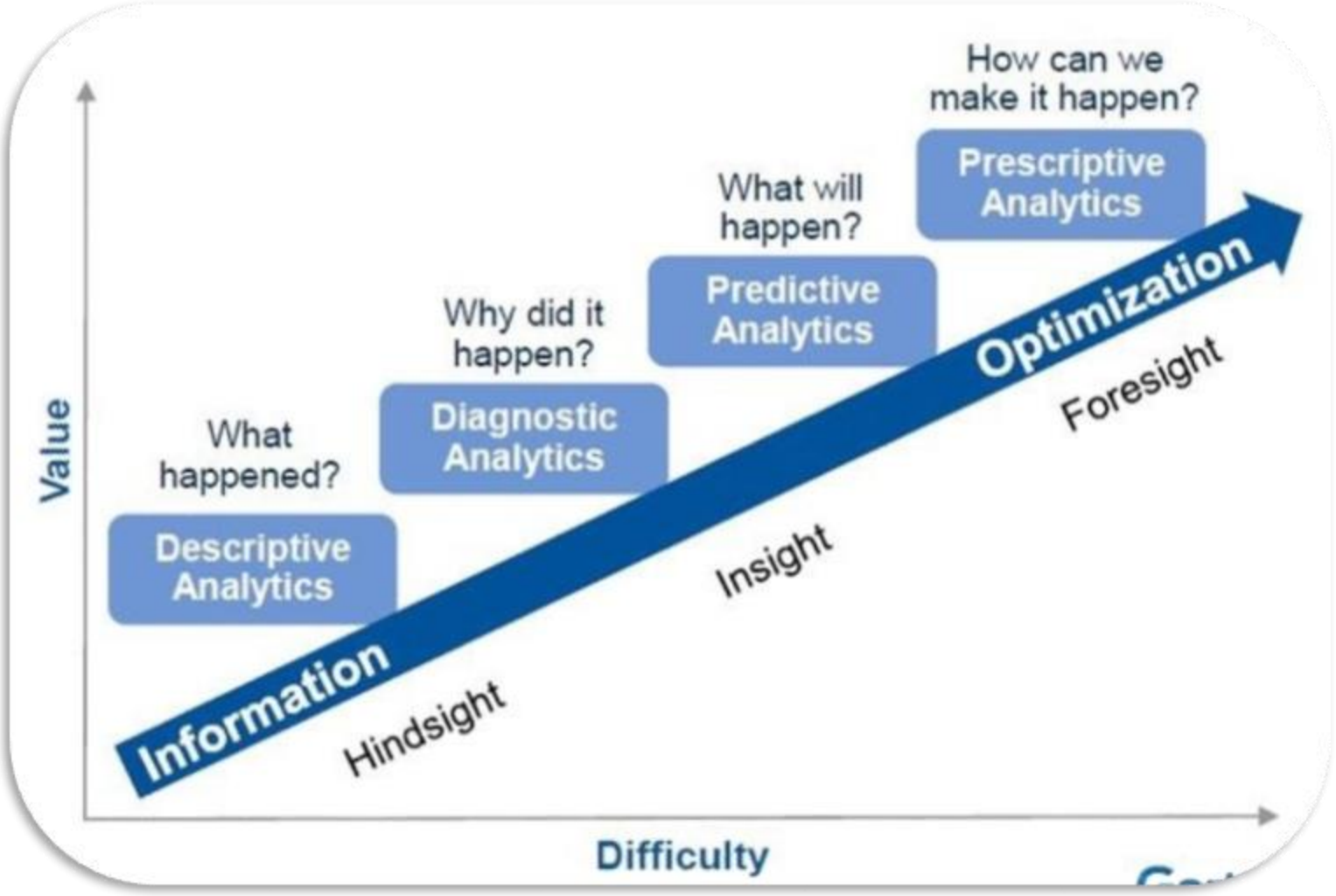
Data Science



Data Professionals



Business Analytics



A Standard

CRISP-DM

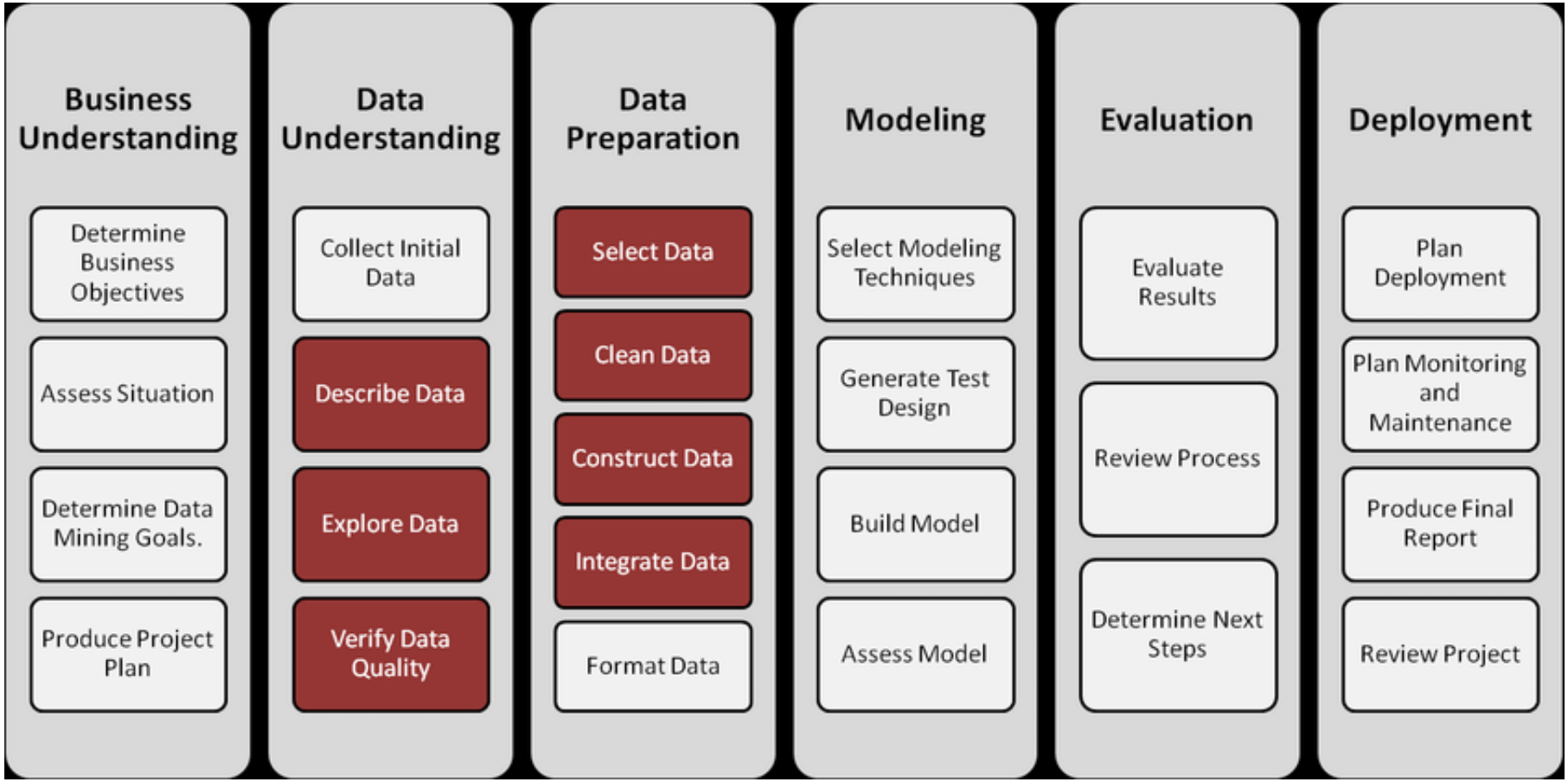
Cross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

- A process model that provides a complete blueprint for conducting a data mining project.
- Breaks down the life cycle of a data mining project into six high-level phases to describe the analytics process.

CRISP-DM: Process

https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

CRISP-DM: Tasks



Must Read

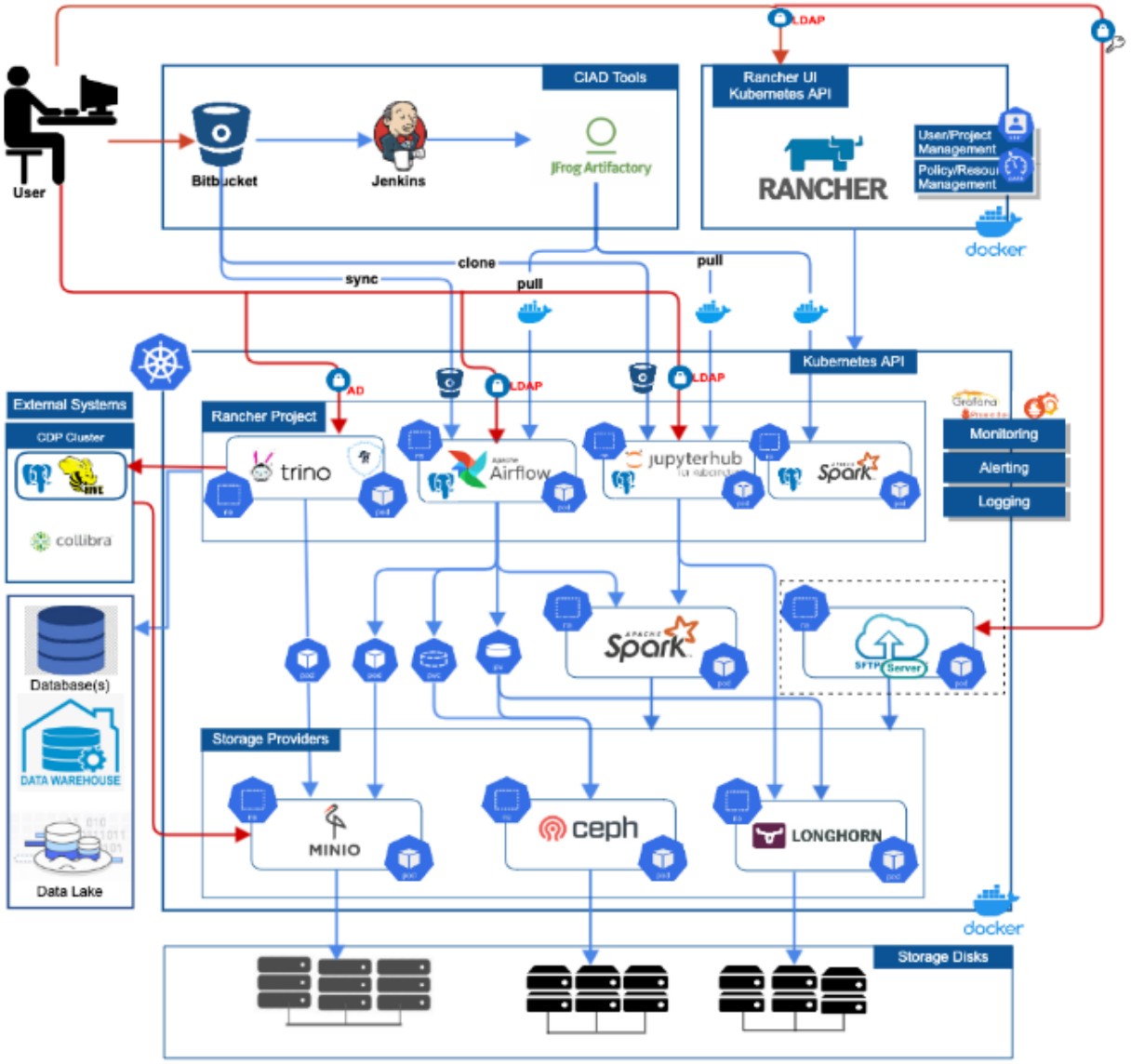
2012

Data Scientist: The Sexiest Job of the 21st Century

2022

Is Data Scientist Still the Sexiest Job of the 21st Century?

Enterprise Architecture



Course Outline

Please review the course outline on OWL.

You will find the answers to many of your current and future question in it.

For example, what happens if I miss the midterm, any relief?