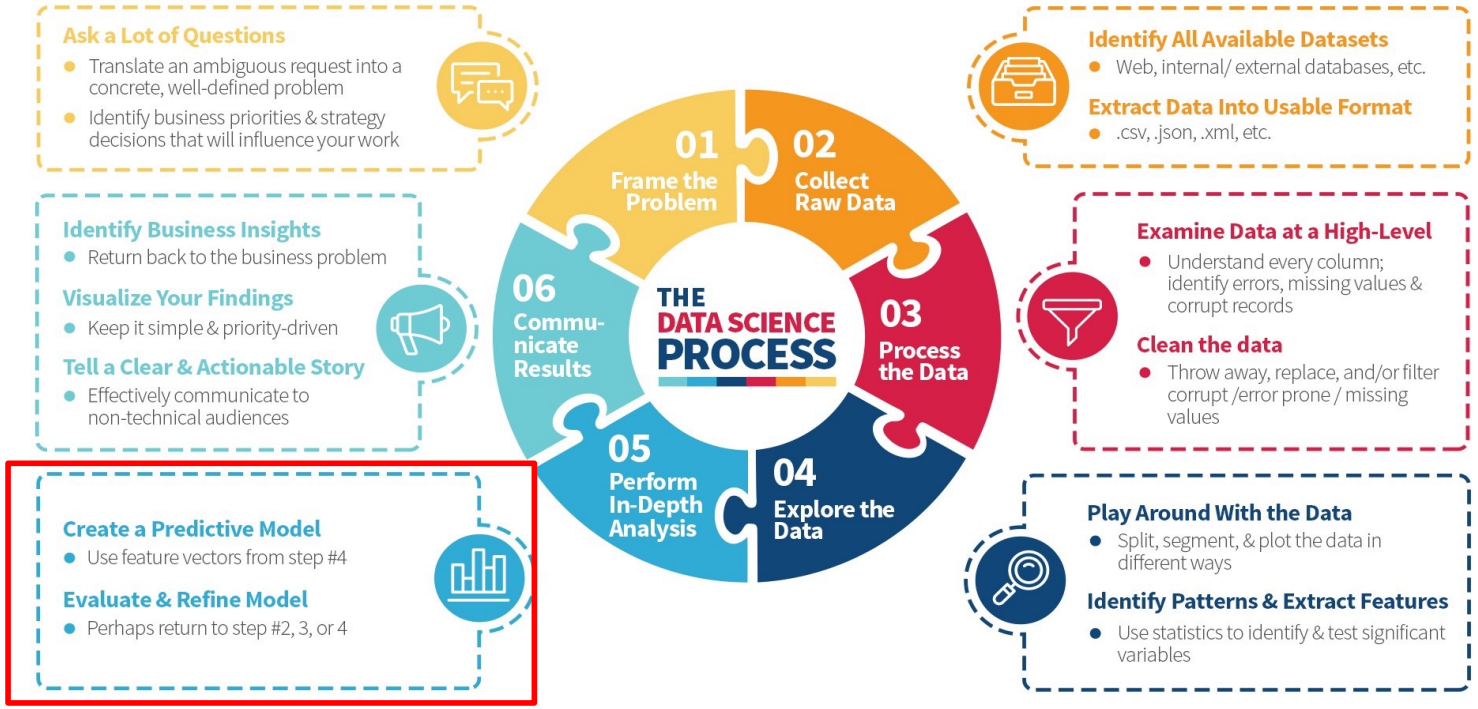

ubiqum code academy
M1T2: Where Are We?

From a task perspective…

Data Science Deconstructed



We are here

SKILLS REQUIRED

We are here

01 FRAME THE PROBLEM

- **Domain Knowledge** (needs)
- **Product Intuition** (metrics)
- **Business Strategy** (priorities)
- **Teamwork** (people & resources)

02 COLLECT RAW DATA

- **Database Management**
  - Systems: mySQL, postgreSQL, Oracle, MongoDB
- **Querying Structured Databases**
  - SQL
- **Retrieving Unstructured Info**
  - Informational Retrieval / Text Mining
- **Distributed Storage**
  - Hadoop HDFS, Spark, Flink

03 PROCESS THE DATA

- **Scripting Language**
  - Python or R
- **Data Wrangling & Cleaning**
  - Python "Pandas" library
- **Distributed Processing**
  - Hadoop MapReduce / Spark

04 EXPLORE THE DATA

- **Scientific Computing**
  - Python: numpy, matplotlib, scipy, pandas
- **Inferential Statistics**
  - hypothesis testing
  - correlation vs. causation
- **Experimental Design**
  - A/B tests, controlled trials

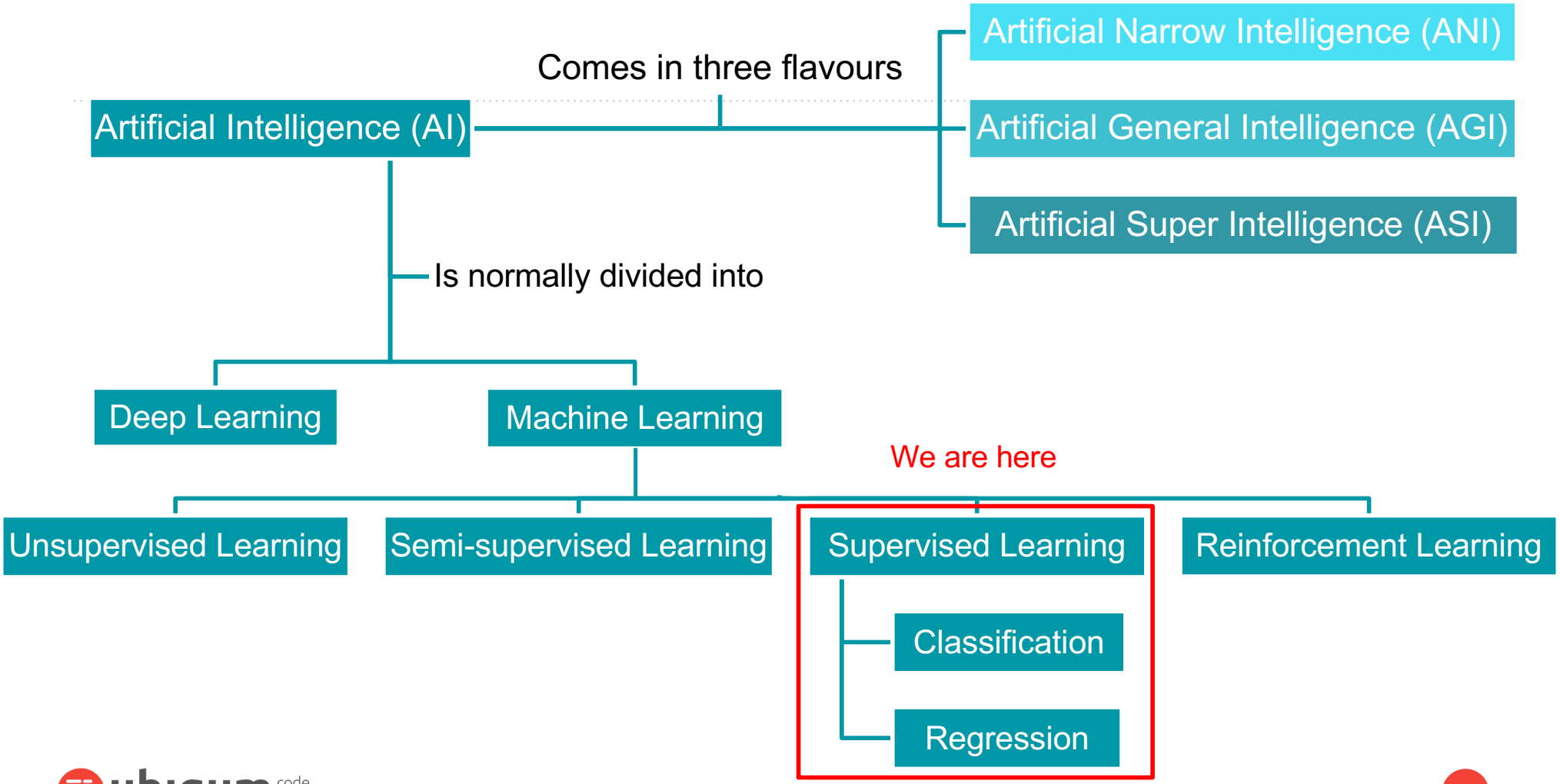
05 PERFORM IN-DEPTH ANALYSIS

- **Machine Learning**
  - Supervised / Unsupervised algorithms
  - Contextual pros/cons)
- **ML Tools Library**
  - Python: scikit-learn
- **Advanced Math**
  - Linear Algebra & Multivariate Calculus

06 COMMUNICATE RESULTS

- **Business Acumen**
  - Non-technical terminology
- **Data Visualization Tool(s)**
  - Tableau, D3.js, Google visualize, matplotlib, ggplot, seaborn
- **Data Storytelling**
  - presenting & speaking
  - reporting & writing

… but where are we exactly, within the field of Artificial Intelligence?



Supervised Learning

Supervised learning (SL) is the **machine learning** task of **learning** a function that maps an input to an output, based on example labelled input-output pairs.

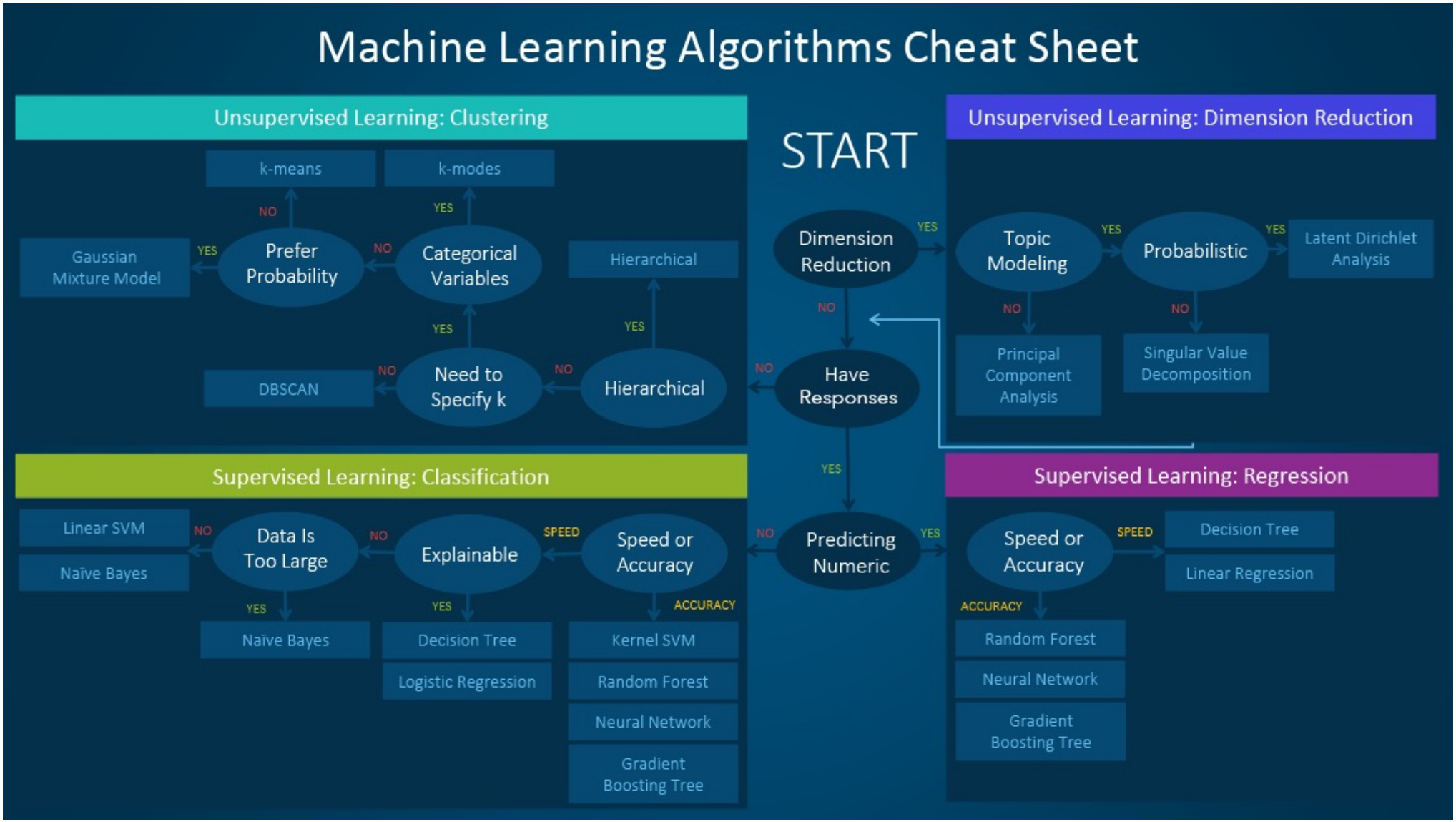
Classification VS Regression

They are two major types of prediction problems Machine Learning tries to solve. Fundamentally, they differ because:

- **classification** is about predicting a label (true/false, low/medium/high, etc)
- **regression** is about predicting a quantity (age, price, height, etc)

Main Algorithms

REGRESSION	CLASSIFICATION
Decision Tree	Decision Tree
Linear Regression	Logistic Regression
Random Forest	Random Forest
Neural Network	Neural Network
Gradient Boosting Tree	Gradient Boosting Tree
Support Vector Machine	Support Vector Machine
	Kernel SVM
	Naïve Bayes
	K-Nearest Neighbours


Machine Learning Algorithms Cheat Sheet
START
Unsupervised Learning: Clustering
k-means
k-modes
NO
YES
Gaussian Mixture Model
YES
Prefer Probability
NO
Categorical Variables
Hierarchical
YES
YES
DBSCAN
NO
Need to Specify k
NO
Hierarchical
Unsupervised Learning: Dimension Reduction
Dimension Reduction
YES
Topic Modeling
YES
Probabilistic
YES
Latent Dirichlet Analysis
NO
NO
NO
Principal Component Analysis
Singular Value Decomposition
NO
Have Responses
YES
Supervised Learning: Classification
Linear SVM
NO
Data Is Too Large
NO
Explainable
SPEED
Speed or Accuracy
Naïve Bayes
YES
YES
ACCURACY
Naïve Bayes
Decision Tree
Kernel SVM
Logistic Regression
Random Forest
Neural Network
Gradient Boosting Tree
NO
Predicting Numeric
YES
Supervised Learning: Regression
Speed or Accuracy
SPEED
Decision Tree
Linear Regression
ACCURACY
Random Forest
Neural Network
Gradient Boosting Tree

Tips:

- As you are being bombarded with a multitude of new exciting concepts and ideas, it is very easy to get lost. Therefore, try to stay focused and explore only the concepts/tools you need for task 2.
- Run your fist model ASAP and write down the first set of metrics. Once that's done, you start reiterating the process, over and over, until you get the best possible results.

Resources:

- https://scikit-learn.org/stable/
- https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/
- https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html
- https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

Thanks.