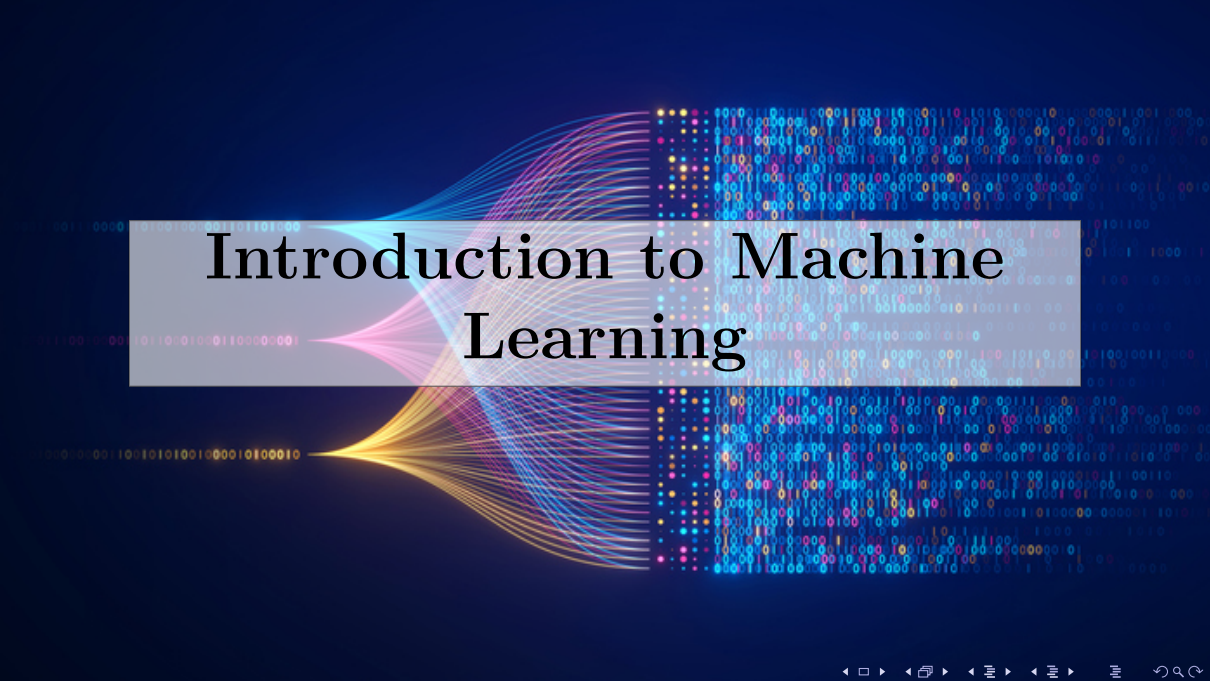
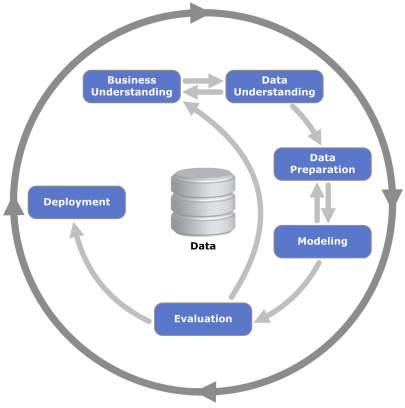

Introduction to Machine Learning

Learning Outcomes



The CRISP-DM process.

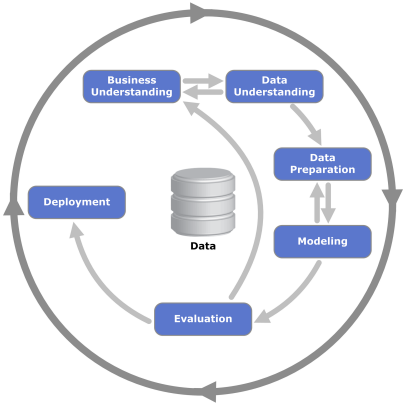
- Define machine learning and compare with AI;
- Determine if a given problem is regression or classification;
- Define three machine learning models;
- Identify two key steps needed before machine learning.

Public service announcement:

Not every data science project leads to machine learning!

MSOE UNIVERSITY

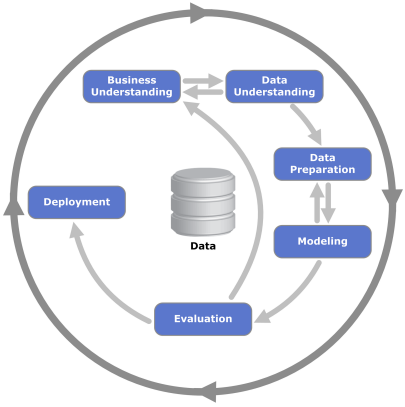
Where are we going?



The CRISP-DM process.

During the second half of the term we will delve into **machine learning**!

Machine Learning in the data science process



The CRISP-DM process.

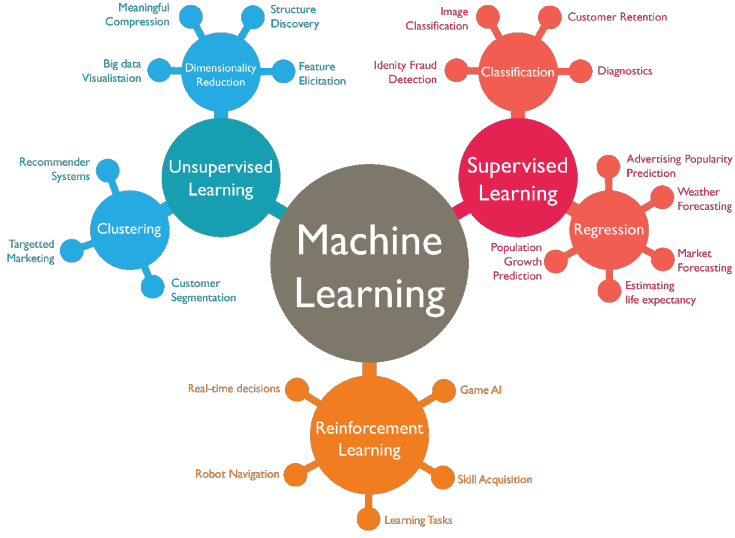
During the second half of the term we will delve into **machine learning**!

During the **Modeling** phase we typically:

1. Preprocess;
2. Devise experimental structure, e.g. split test/train;
3. Apply data transformations;
4. Train model;
5. Evaluate model.

MSOE UNIVERSITY

Three types of Machine Learning



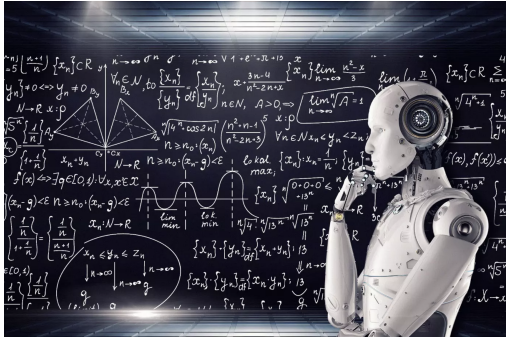
The source for this graphic.

Quick aside – what is AI?

Quick aside – what is AI? Well, what are we?

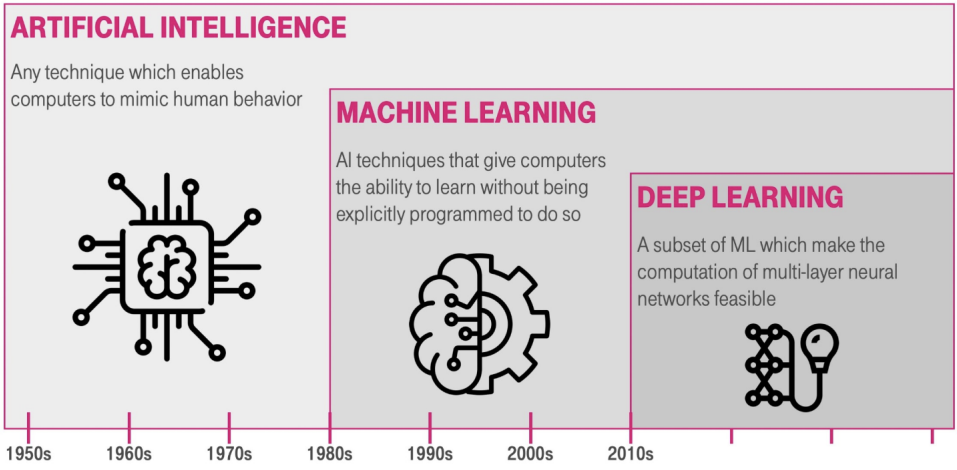
Decision makers.

Quick aside – what is AI?

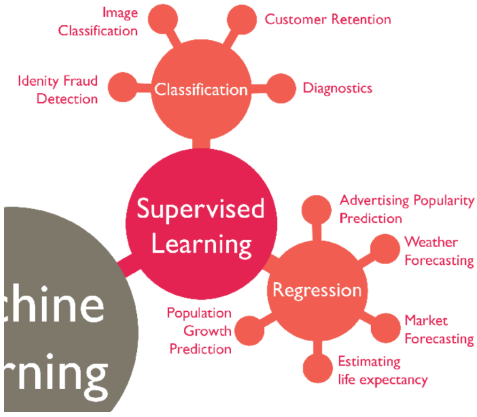


An **automated** decision maker.

How does AI compare to Machine Learning?



Dependent variable structure is everything

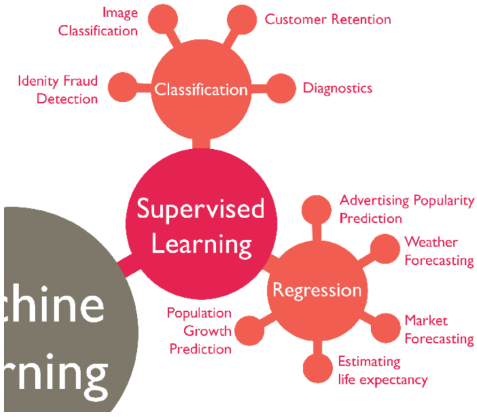


Let's focus on supervised ML for now...

- Many of the decisions we make with data will be driven by how that data is structured, e.g. the right...
  - ...Visualization;
  - ...Hypothesis test;
  - ...ML model;

MSOE UNIVERSITY

Dependent variable structure is everything



Let's focus on supervised ML for now...

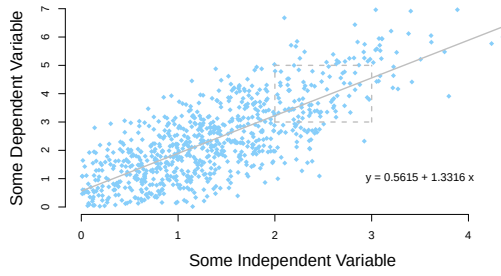
- Many of the decisions we make with data will be driven by how that data is structured, e.g. the right...
  - ...Visualization;
  - ...Hypothesis test;
  - ...ML model;

- Two key structures:
  1. **Regression:** the dependent variable y is **numerical**, i.e. numbers;
  2. **Classification:** the dependent variable y is **categorical**, i.e. NOT numbers.

Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- y is a numerical dependent variable;
- x's are the features;
- β's are weights learned from data;
- Super easy to interpret, workhorse ML model for regression.

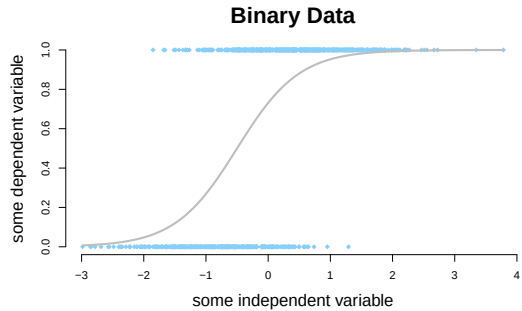


Linear regression, AKA OLS!

Logistic Regression

$$p(y = 1) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\}}$$

- y is a binary categorical dependent variable;
- x's are the features;
- β's are weights learned from data;
- Super easy to interpret, workhorse ML model for classification.



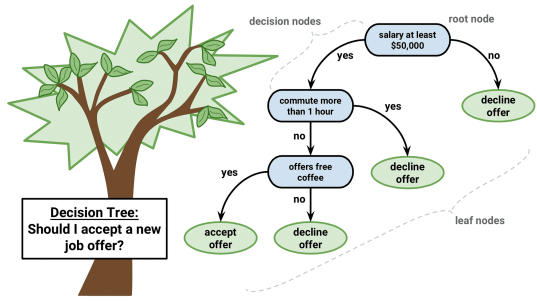
People who actually do logistic regression call this logit!

Random Forest Classification

each scoring function is a decision tree represented as a pair (T_j, L_j) where T_j is a particular partition of the data and $L_j : X \to T_j$ is a function mapping from an observation in the feature space to an element of the partition.[4] The scoring functions would be:

$$s_j(x, y) \equiv \frac{1}{|L_j(x)|} \sum_{x', y' \in L_j(x)} I(y = y')$$

- y is a dependent variable;
- x's are the features;
- Algo learns how to split the data using the features;
- Less easy to interpret but still doable – perhaps the best off-the-shelf model.



Random forests have MANY of these.

There are a lot of machine learning models...

- Linear (hinge loss) SVM regression/classification;
- Kernelized SVM;
- Linear Discriminant Analysis;
- Decision Trees;
- KNN;
- Naive Bayes;
- Gradient boosting (xgboost, lightGBM, etc,);
- Bayesian Additive Regression Trees;
- Lasso/AdaptiveLasso/Ridge/Elastic Net Regression;
- Multilayer Perceptrons/Deep learning;
- Dozens of reinforcement learning algos;
- Etc., etc., etc...

Two key things to do BEFORE ML...

- **Make data ready for model ingest** – the ML models we'll be using will be looking for a data set of numbers in a table;
  - Deal with missing values;
  - Feature engineering;
  - Scaling;

- **Choose a model** – even in small data sets there are many models you could choose to build;
  - This problem is called 'feature selection' or 'model specification';
  - Address via visualization, hypothesis testing, or an algo.

Example: Titanic

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...

Example: Titanic

- Suppose we use logit to model survived (dep var) in terms of Pclass, Age, and Cabin – then for row 1 we would have:

$$p(\text{Survived}) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}}$$

Problems:

Example: Titanic

- Suppose we use logit to model survived (dep var) in terms of Pclass, Age, and Cabin – then for row 1 we would have:

$$p(\text{Survived}) = \frac{\exp\{\beta_0 + \beta_1 * \text{Pclass} + \beta_2 * \text{Age} + \beta_3 * \text{Cabin}\}}{1 + \exp\{\beta_0 + \beta_1 * \text{Pclass} + \beta_2 * \text{Age} + \beta_3 * \text{Cabin}\}}$$

Problems:

Example: Titanic

- Suppose we use logit to model survived (dep var) in terms of Pclass, Age, and Cabin – then for row 1 we would have:

$$p(\text{Survived}) = \frac{\exp\{\beta_0 + \beta_1 * 3 + \beta_2 * 22 + \beta_3 * NaN\}}{1 + \exp\{\beta_0 + \beta_1 * 3 + \beta_2 * 22 + \beta_3 * NaN\}}$$

Problems:

Example: Titanic

- Suppose we use logit to model survived (dep var) in terms of Pclass, Age, and Cabin – then for row 2 we would have:

$$p(\text{Survived}) = \frac{\exp\{\beta_0 + \beta_1 * 1 + \beta_2 * 38 + \beta_3 * \text{C85}\}}{1 + \exp\{\beta_0 + \beta_1 * 1 + \beta_2 * 38 + \beta_3 * \text{C85}\}}$$

Problems:

Example: Titanic

- Suppose we use logit to model survived (dep var) in terms of Pclass, Age, and Cabin – then for row 2 we would have:

$$p(\text{Survived}) = \frac{\exp\{\beta_0 + \beta_1 * 1 + \beta_2 * 38 + \beta_3 * \text{C85}\}}{1 + \exp\{\beta_0 + \beta_1 * 1 + \beta_2 * 38 + \beta_3 * \text{C85}\}}$$

Problems: Is Pclass really numeric? Cabin certainly is not!

Example: Titanic

- Suppose we use logit to model survived (dep var) in terms of Pclass, Age, and Cabin – then for row 2 we would have:

$$p(\text{Survived}) = \frac{\exp\{\beta_0 + \beta_1 * 1 + \beta_2 * 38 + \beta_3 * \text{C85}\}}{1 + \exp\{\beta_0 + \beta_1 * 1 + \beta_2 * 38 + \beta_3 * \text{C85}\}}$$

  Problems: Is Pclass really numeric? Cabin certainly is not!

- Taking the Titanic data there are 11 features and so 2047 models;

PassengerID	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
PassengerID										
PassengerID	Pclass									
PassengerID	Pclass	Name								
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Upcoming lectures

In the upcoming weeks we will build the tools that we need to manage these problems and prepare data for machine learning!

- Data types;
- Visualization;
- Probability review;
- Hypothesis testing;
- Exploratory data analysis;
- Feature engineering...