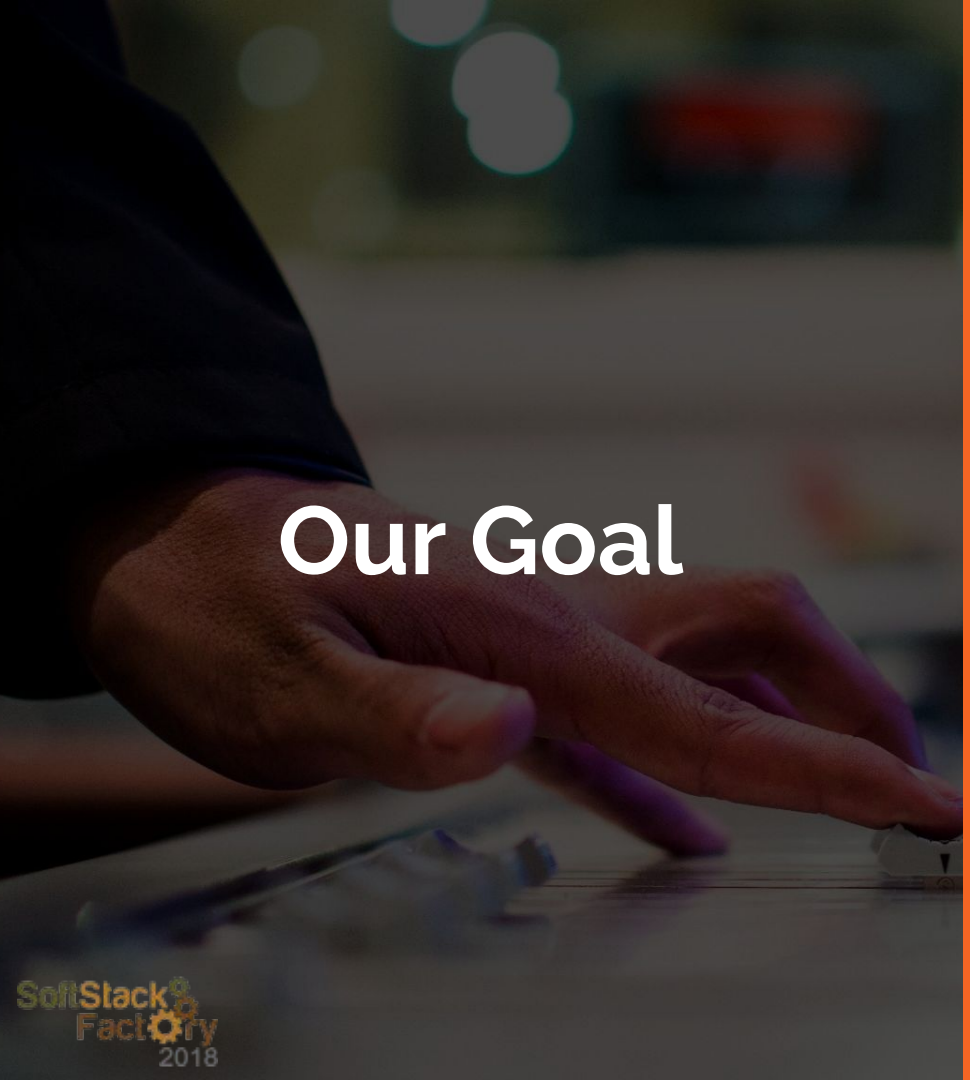
Data Science and Machine Learning Primer

Overview

SoftStack Factory 2018

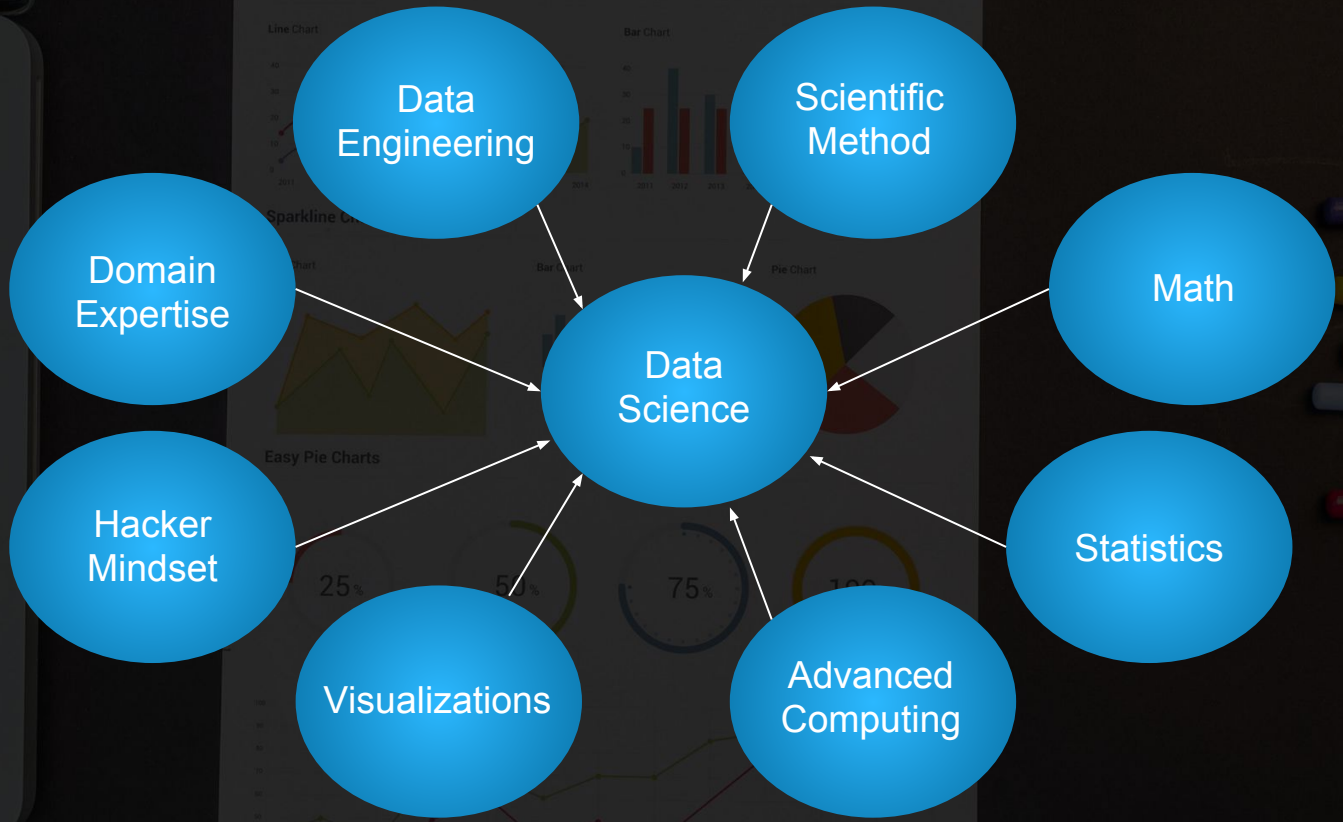


Our Goal

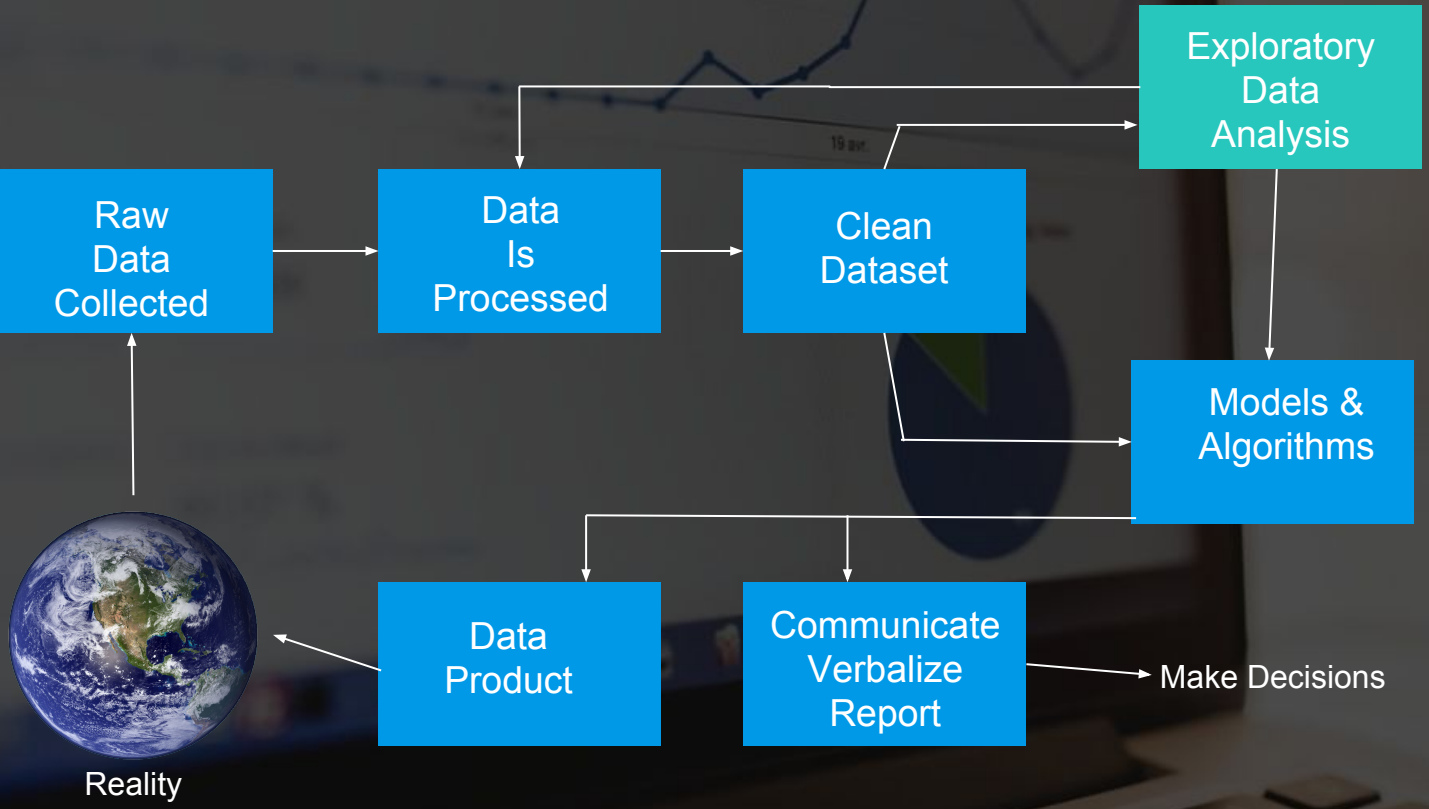
To provide an introductory understanding of machine learning, techniques and tools to further investigate and create predictive solutions.

- This is a comprehensive field.
- Tip of the Iceberg
- Do not expect to be an expert
- We are covering a small portion in this class.
- Discover what you will need to learn.

What is Data Science?



Data Science Process



What is a Data Scientist?

- Leverages existing data sources.
- Create new data features/sources as needed.
- Extracts meaningful information and actionable insights.
- Their insights drive business decisions to achieve business objectives.

What is Machine Learning?

“Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world.” – Nvidia

“Machine learning is the science of getting computers to act without being explicitly programmed.” – Stanford

What is Machine Learning?

“Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.” – University of Washington

“The field of Machine Learning seeks to answer the question “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” – Carnegie Mellon University

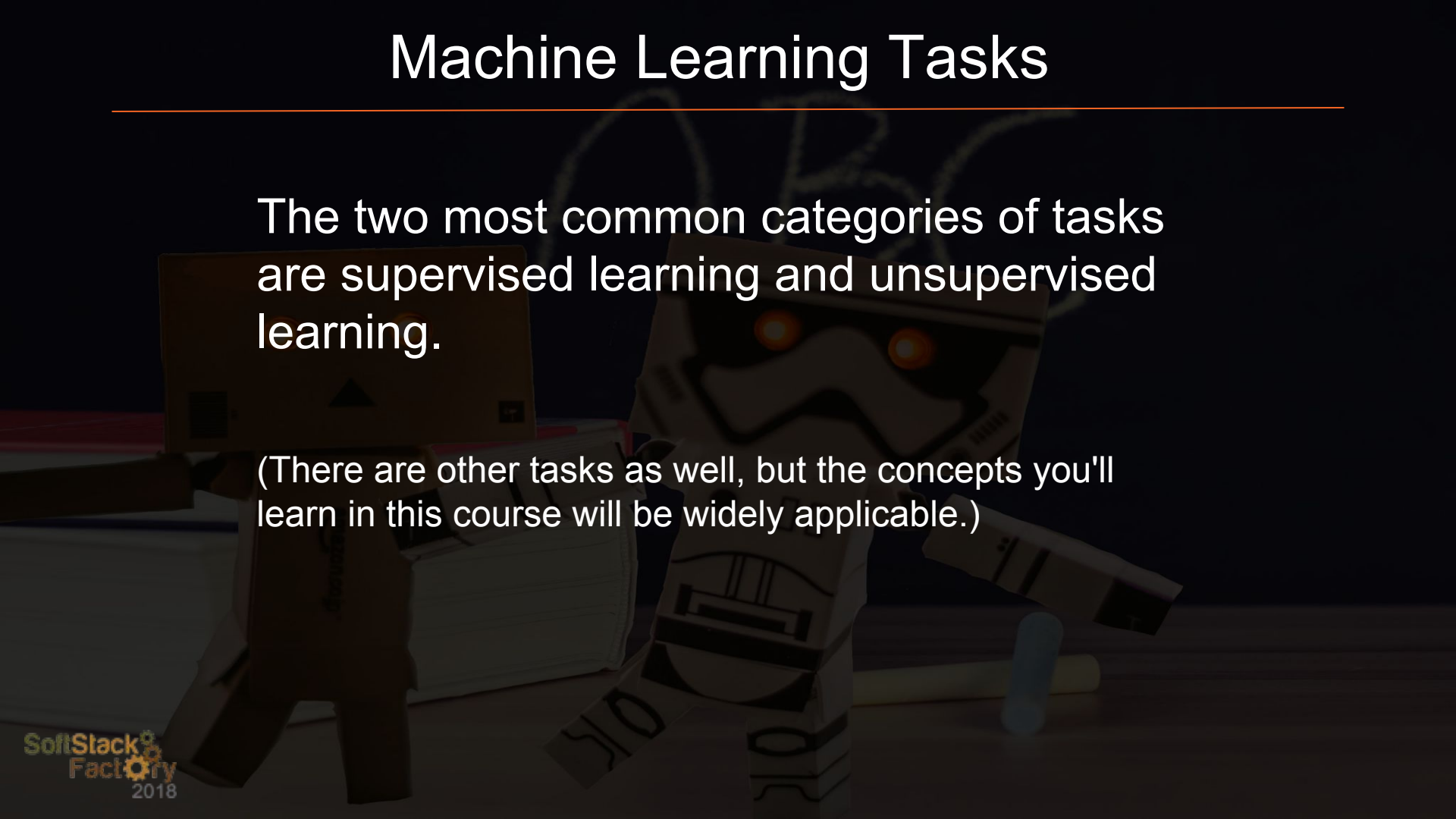
What is Machine Learning?

Machine Learning is the science of getting computers to learn and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.

Machine Learning Tasks

The two most common categories of tasks are supervised learning and unsupervised learning.

(There are other tasks as well, but the concepts you'll learn in this course will be widely applicable.)



Supervised Learning

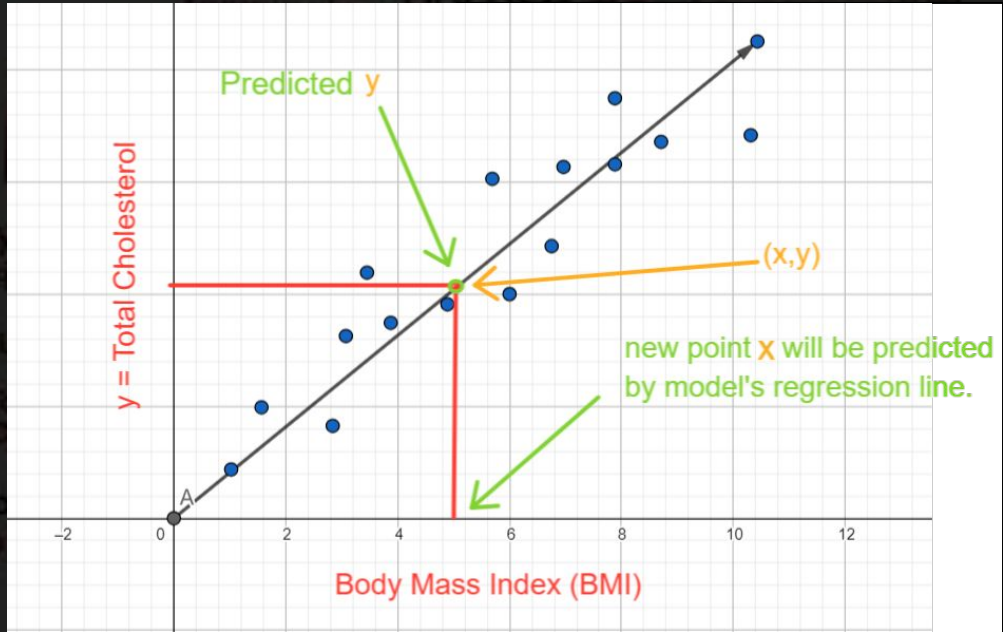
- In practice, it's often used as an advanced form of *predictive modeling*.
- Each observation must be labeled with a "correct answer."
- Only then can you build a predictive model because you must tell the algorithm what's "correct" while training it (hence, "supervising" it).
- **Regression** is the task for modeling *continuous target variables*.
- **Classification** is the task for modeling *categorical (a.k.a. "class") target variables*.

Regression

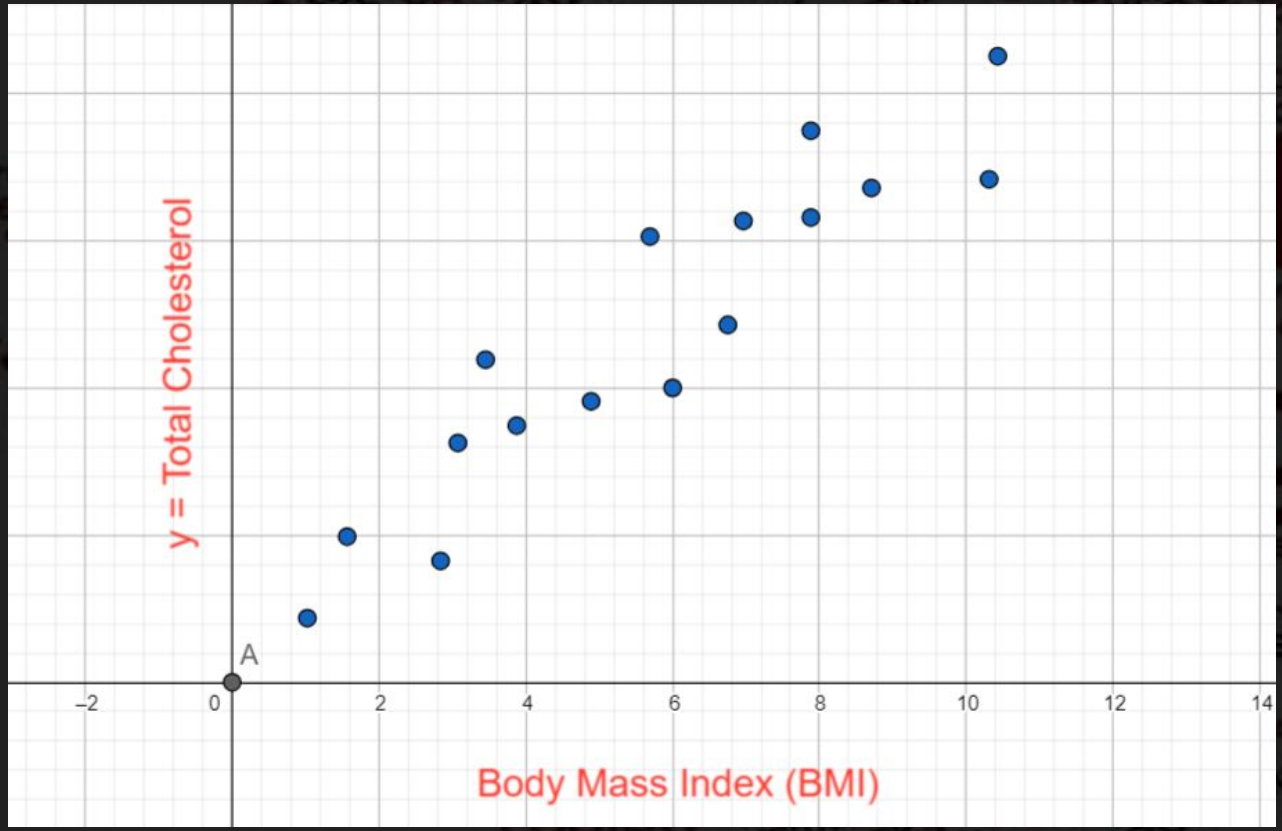
Machine learning technique primarily used for predicting a numerical value.

For Example:

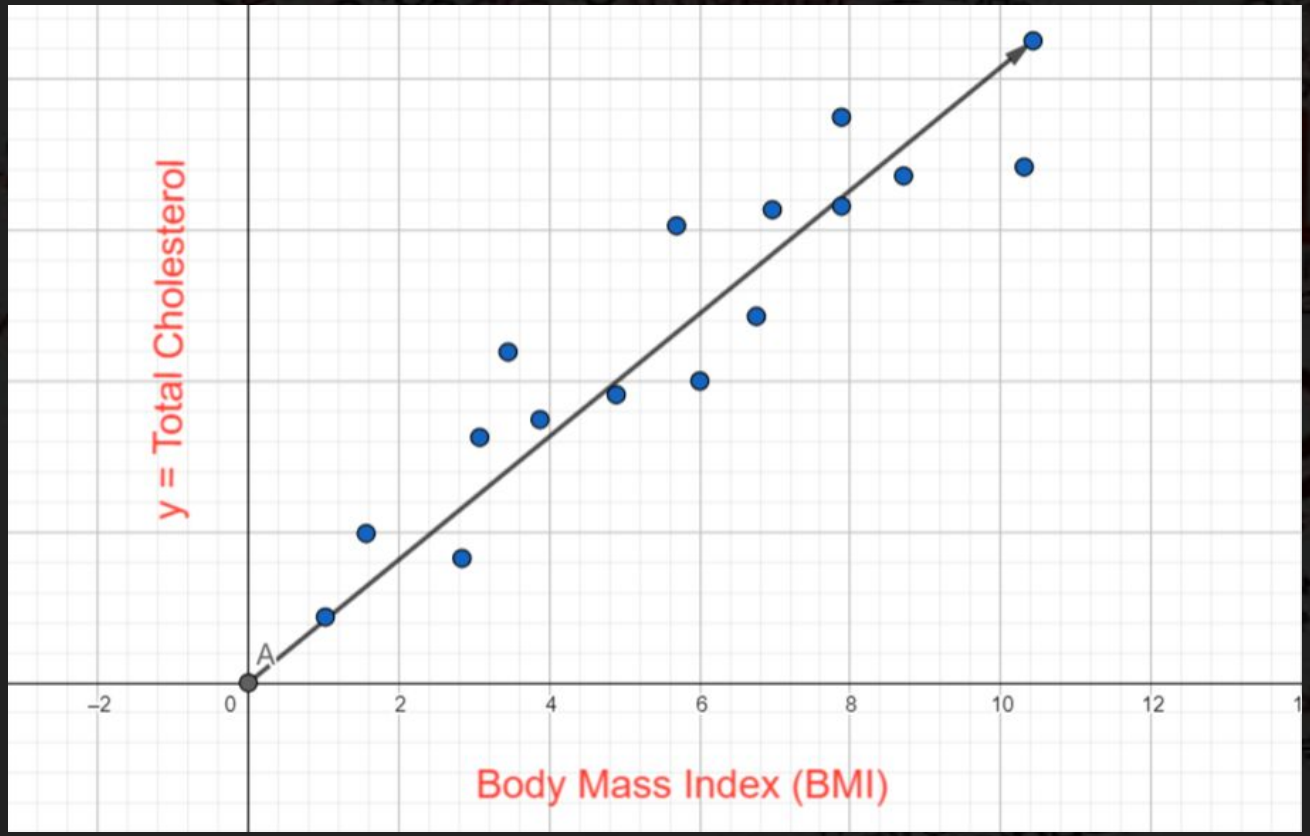
1. Given a set of BMI values, X.
2. Predict total cholesterol y, for any particular x in X.



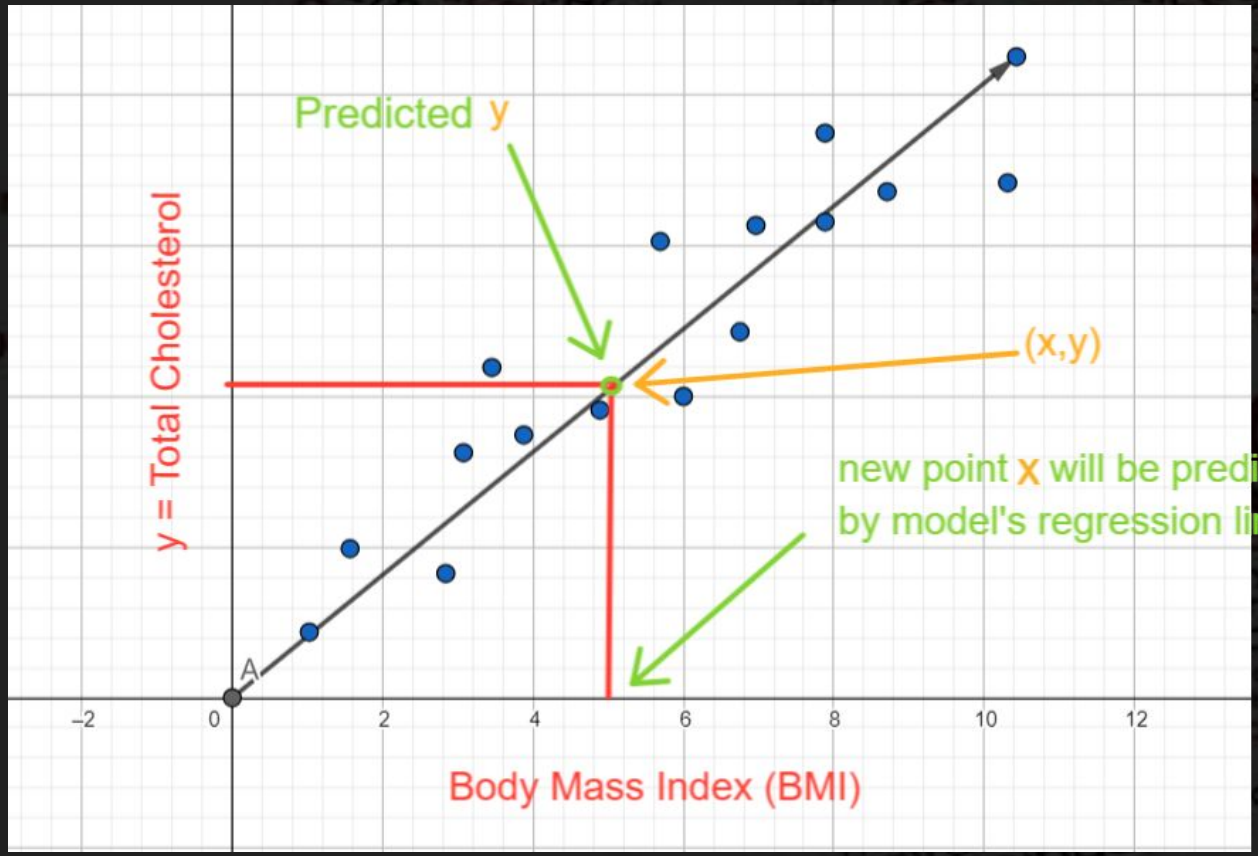
Regression



Regression



Regression

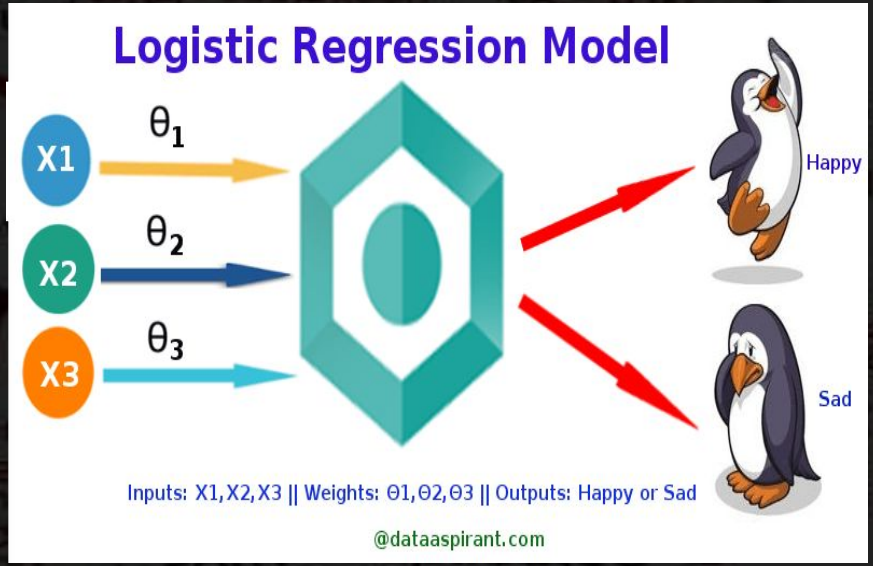


Classification In a Nutshell

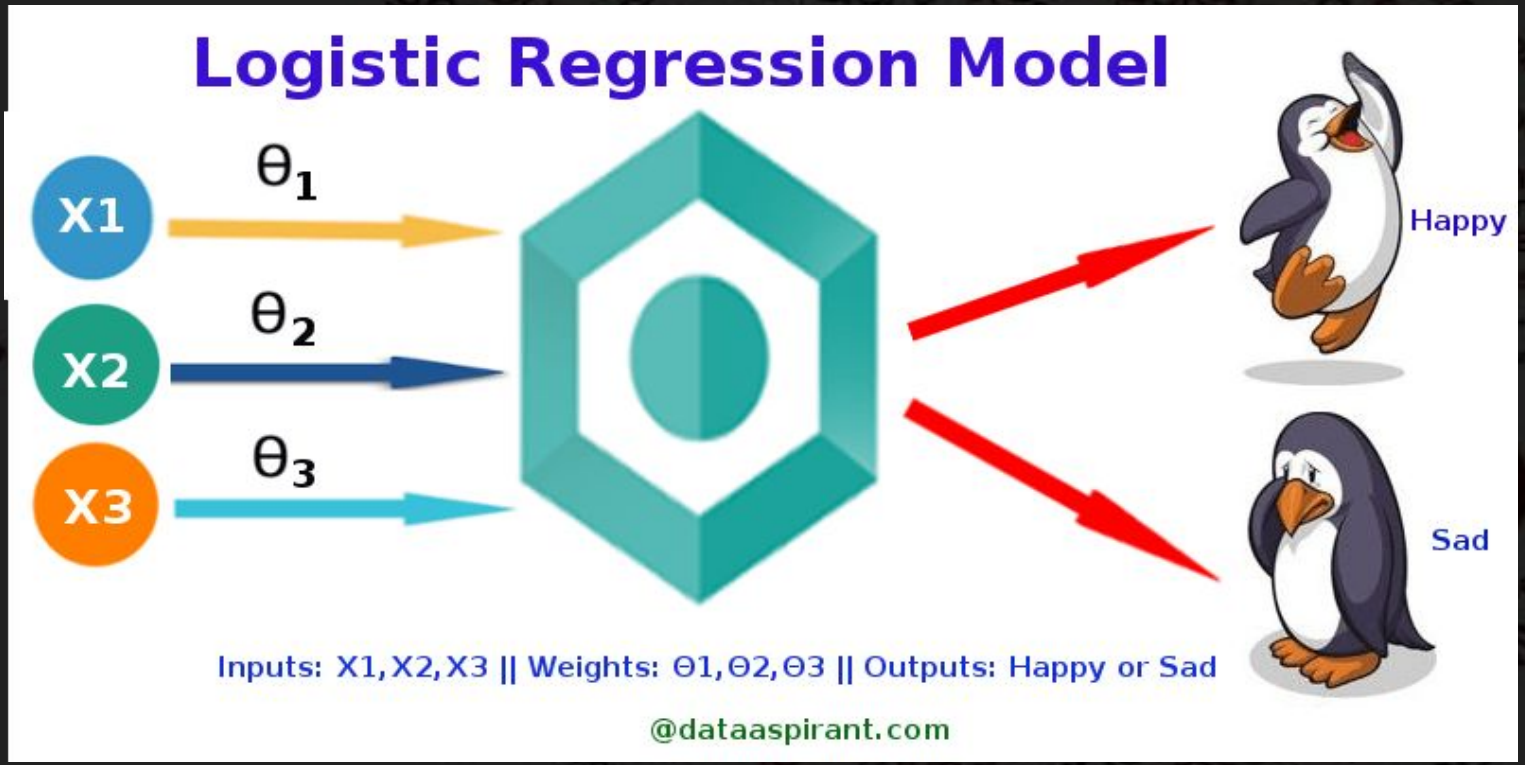
Machine learning technique primarily used for predicting the class that a data point belongs to.

For Example:

1. Given a set of penguins X.
2. Predict class “happy” or “sad”, for any particular penguin (X1, X2, X3 … , Xn).



Classification In a Nutshell



Unsupervised Learning

- In practice, it's often used either as a form of *automated data analysis* or *automated signal extraction*.
- Unlabeled data has no predetermined "correct answer."
- You'll allow the algorithm to directly learn patterns from the data (without "supervision").
- **Clustering** is the most common unsupervised learning task, and it's for *finding groups* within your data.

The First Big Data Success Story

Payment Fraud Prevention

SoftStack Factory 2018

Before Falcon - 1992

- Cards were authorized by phone calls when they exceeded a certain value.
- Clerks had books with card numbers that were no longer good and then would take them away if you tried to use them.
- Carbon imprints of your card made it easier for fraudster
- Credit card swipe terminals made it simpler for merchants and initially safer, but not for long.

1992 - Today

- Early 1990's Fraudsters were defeating physical fraud-prevention elements
- The solution to this problem turned out to be analytics
- The modern age of payment fraud prevention started in 1992 when HNC Software introduced Falcon Fraud Manager
- Anti-fraud applications run up to 15,000 calculations in a matter of milliseconds
- They leverage intelligence from trillions of transactions -- to determine the likelihood of fraud

Incredible Analytic Innovations in Payment Fraud Prevention

- Large data-set modeling
- Where analytic techniques are used to take complex, large data sets and translate the data into a format that is usable for real-time scoring.
- Transaction profiling (1992) summarizes complex transaction histories
- Neural network models (1993) work like the human brain to understand non-linear interactions between variables (e.g., transaction amount and location)

Incredible Analytic Innovations in Payment Fraud Prevention

- Sparse data-set modeling - Solves problems where large sets of historical data do not exist, leading to innovations in self-learning and auto-calibrating models.
- E-commerce fraud modeling (1999) protects merchants from card-not-present fraud.
- Outlier models (2005) increase precision by examining unusual payment incidents within the context of other outlier transactions.
- First-party fraud modeling (2006)
- Self-calibrating technology (2008) allows anti-fraud software to fine-tune itself

Incredible Analytic Innovations in Payment Fraud Prevention

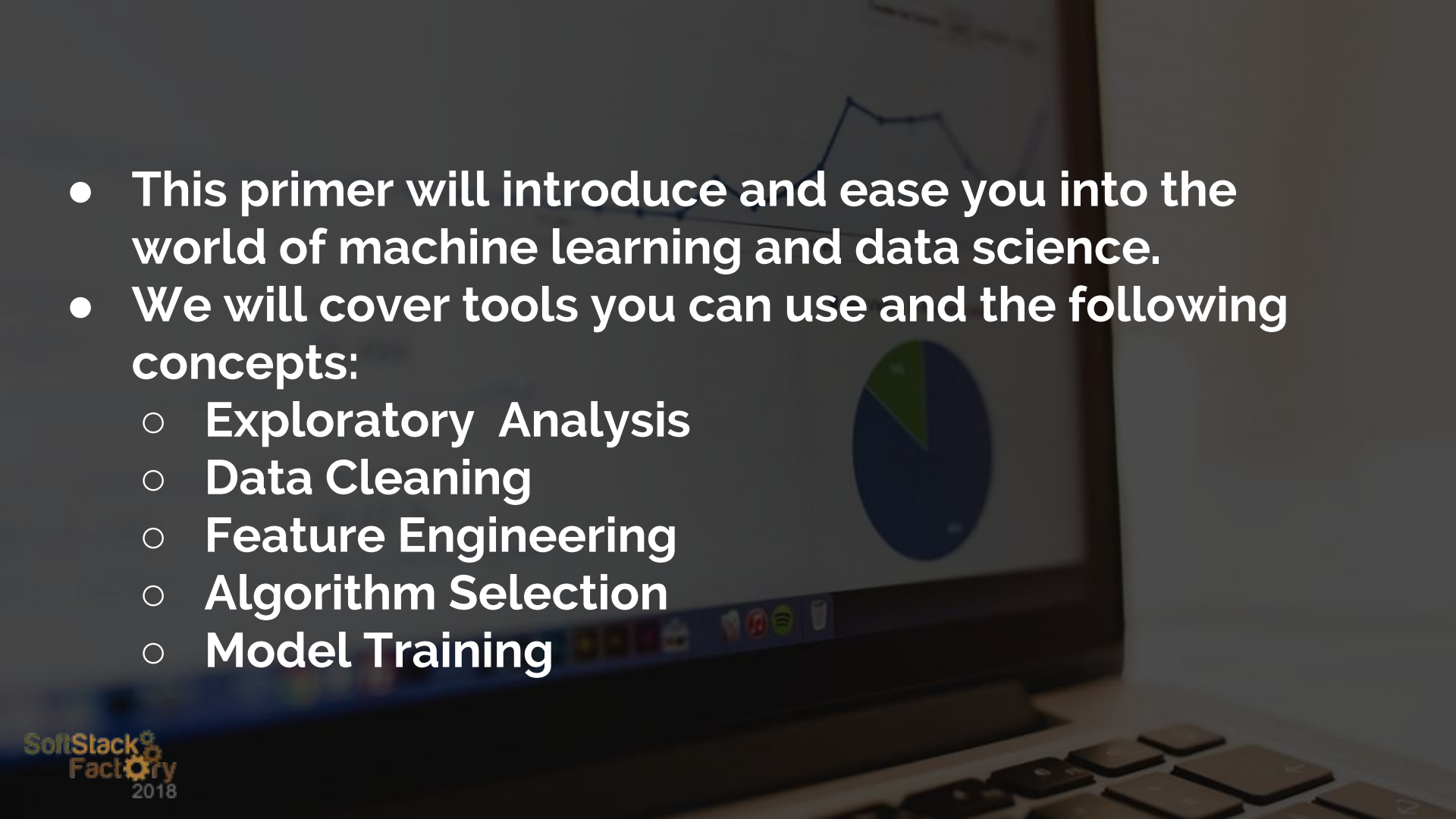
- False-positive reduction - Analytic techniques that focus on the “cost” of anti-fraud efforts by reducing the number of non-fraudulent transactions stopped or investigated, and improving the customer experience at the point-of-sale
- Global intelligent profiles (2009) identify the riskiest ATMs, merchants and regions so extra scrutiny can be applied where risk is greatest.
- Adaptive analytics (2010) enables analytic software to adjust models as fraud patterns change
- Behavior sorted lists (2013) improve the ability to identify suspicious transactions

Source: T.J. Horan

https://www.kdnuggets.com/2014/03/evolution-fraud-analytics-inside-story.html

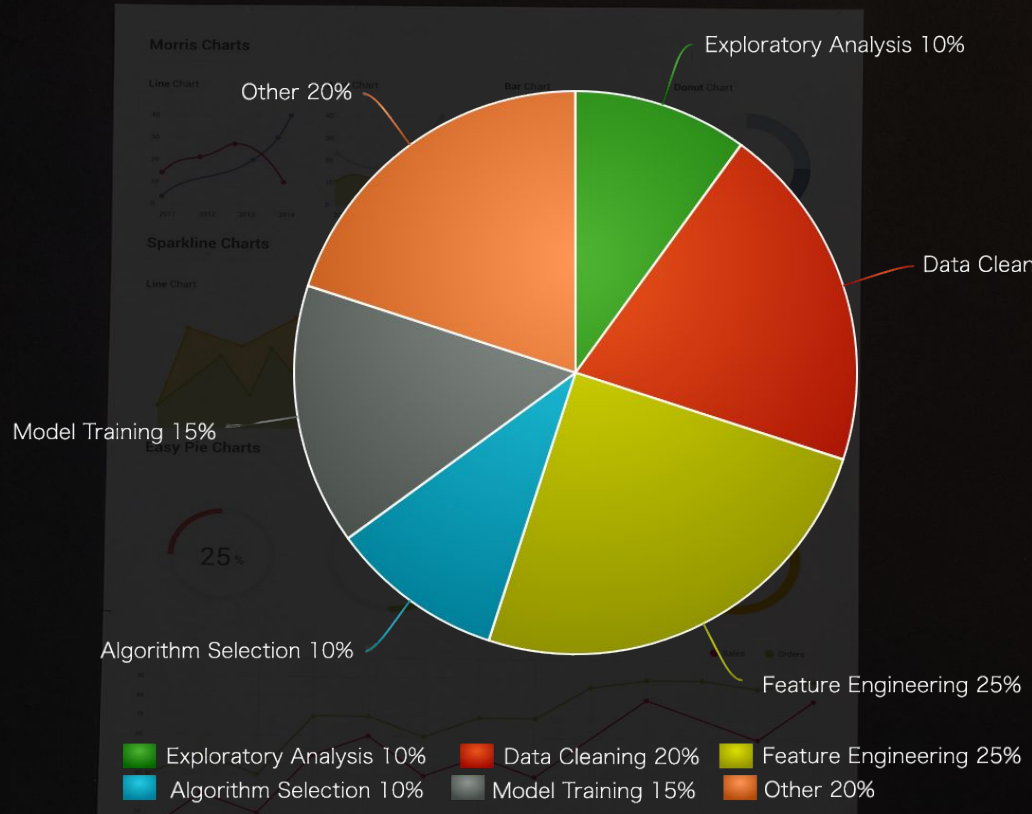
Other Real World Examples

- Amazon
  - Advertising
  - Alexa Smart Home Devices
  - Supply Chain Optimization
- Netflix
  - Recommendations



- This primer will introduce and ease you into the world of machine learning and data science.
- We will cover tools you can use and the following concepts:
  - Exploratory Analysis
  - Data Cleaning
  - Feature Engineering
  - Algorithm Selection
  - Model Training

What Goes Into A Successful Model?



Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the first step in your data analysis process. Here, you make sense of the data you have and then figure out what questions you want to ask and how to frame them, as well as how best to manipulate your available data sources to get the answers you need.

The purpose of exploratory analysis is to "**get to know**" the dataset. Doing so upfront will make the rest of the project much smoother, in 3 main ways:

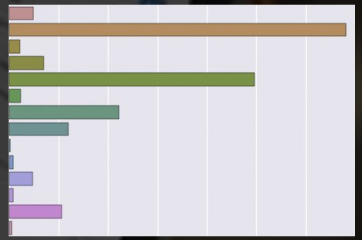
1. You'll gain valuable hints for Data Cleaning (which can make or break your models).
2. You'll think of ideas for Feature Engineering (which can take your models from good to great).
3. You'll get a "**feel**" for the dataset, which will help you communicate results and deliver greater impact.

Exploratory Data Analysis - Topics Covered

Display example observations from the dataset using excel or pandas just to name a few.

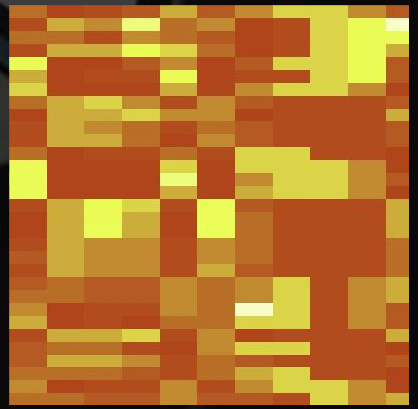
	tx_price	beds	baths	sqft	year_built	lot_size	property_type	exterior_walls	roof	basement	restaurants	groceries	nightlife
0	295850	1	1	584	2013	0	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	107	9	30
1	216500	1	1	612	1965	0	Apartment / Condo / Townhouse	Brick	Composition Shingle	1.0	105	15	6
2	279900	1	1	615	1963	0	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	183	13	31
3	379900	1	1	618	2000	33541	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	198	9	38
4	340000	1	1	634	1992	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	149	7	22

Plot categorical distributions using bar plots



Plot Numerical Distributions

Study Correlations



Data Cleaning

Data cleaning is one those things that everyone does but no one really talks about. Proper data cleaning can make or break your project. Professional data scientists usually spend a very large portion of their time on this step.

Better data beats fancier algorithms. In other words… garbage in gets you garbage out.

Data Cleaning - Topics Covered

1. Removing unwanted observations such as duplicate or irrelevant observations.
2. Fix structural errors such as typos or inconsistent capitalizations. Check for mislabeled classes.
3. Filter unwanted outliers. If you have a legitimate reason to remove an outlier it will help your models performance. However, **outliers are innocent until proven guilty**, never remove an outlier just because it a large number.
4. Handle missing data by either dropping observations that have missing values or imputing the missing values based on other observations. How to handle missing values for categorical versus numerical data.

Feature Engineering

Feature engineering is about creating new input features from your existing ones.

In general, you can think of data cleaning as a process of subtraction and feature engineering as a process of addition.

Feature engineering has limitless possibilities and is a skill that will naturally improve over time.

Feature Engineering - Topics Covered

1. Domain knowledge (area of expertise).
2. Creating interaction features (combination of two or more features).
3. Combining sparse classes.
4. Adding dummy variables.
5. Removing unused features.

Algorithm Selection

Instead of giving you a long list of algorithms, our goal is to explain a few essential concepts (e.g. regularization, ensembling, automatic feature selection) that will teach you why some algorithms tend to perform better than others.

Algorithm Selection - Topics Covered

1. How to pick ML Algorithms
2. Why Linear Regression is Flawed
3. Regularization in Machine Learning
4. Regularized Regression Algos
5. Decision Tree Algos
6. Tree Ensembles

Model Training

It might seem like it took us a while to get here, but professional data scientists actually spend the bulk of their time on the steps leading up to this one:

1. Exploring the data.
2. Cleaning the data.
3. Engineering new features.

Again, that's because **better data beats fancier algorithms**.

In this lesson, you'll learn how to set up the entire modeling process to maximize performance while **safeguarding against overfitting**.

Model Training - Topics Covered

1. How to Train ML Models
2. Split Dataset
3. What are Hyperparameters?
4. What is Cross-Validation?
5. Fit and Tune Models
6. Select Winning Model

Installing Python

- Miniconda (http://conda.pydata.org/miniconda.html)
  - Gives you the Python interpreter along with a command-line tool called conda that is a package manager geared toward installing Python packages.
  - Lightweight version requiring less disk space.
  - Any packages included with Anaconda can also be installed manually on top of Miniconda.
- Anaconda (https://www.continuum.io/downloads)
  - Includes both Python, conda, and additionally bundles a suite of other pre installed packages geared toward scientific computing.
  - User friendly UI (Anaconda Navigator).

Key Terminology

- Model - A set of patterns learned from data.
- Algorithm - A specific ML process used to train a model.
- Training data - The dataset from which the algorithm learns the model.
- Test data - A new dataset for reliably evaluating model performance.
- Features - Variables (columns) in the dataset used to train the model.
- Target variable - A specific variable you're trying to predict.
- Observations - Data points (rows) in the dataset.