Data Science for Beginners

Sep 26, 2018

Viji Krishnamurthy Ph.D.

Agenda

Part 1

- 6:30-6:45– Overview of data science
- 6:45-7:00 – Data science in business context
- 7:00-7:15 – Exercise: Data science end-to-end example

Part 2

- 7:15-7:30 – Data: cleaning, manipulating, scaling, dimensionality reduction
- 7:30-7:45 – Algorithms and data science process
- 7:45-8:00 – Let us solve a problem

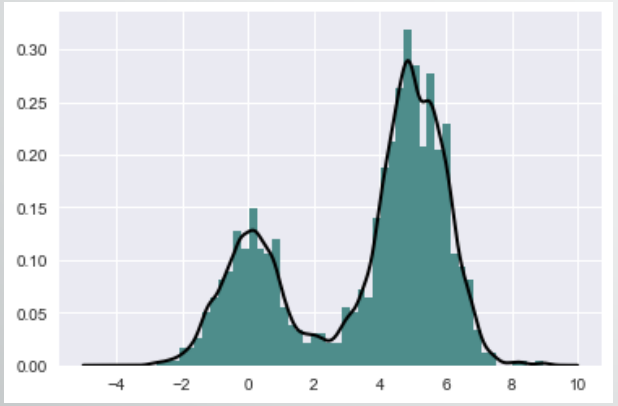
Definition and Examples

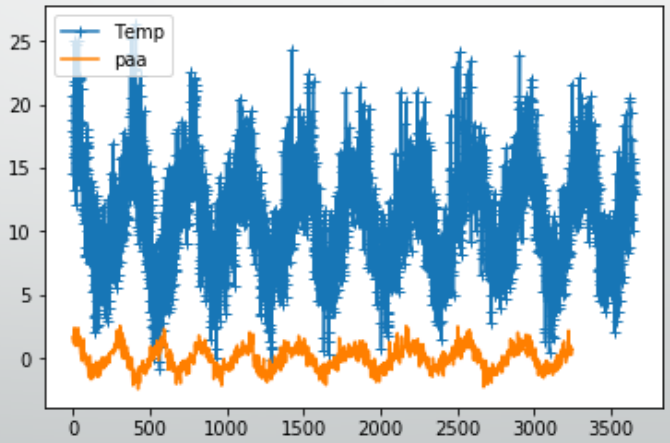
- Data Science is the process of extracting insights from data.

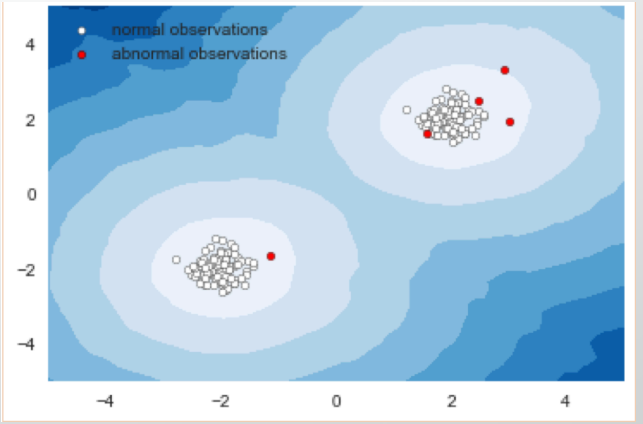
- Examples:
  - Flu epidemic prediction based on search query data: Link
  - Facebook's Global Migration: Link
  - Facebook's NFL Fans: Link

Concepts: Data Exploration and Visualization

- Visualization has two goals: To explore data and to communicate data/insights







Concepts: Algebra

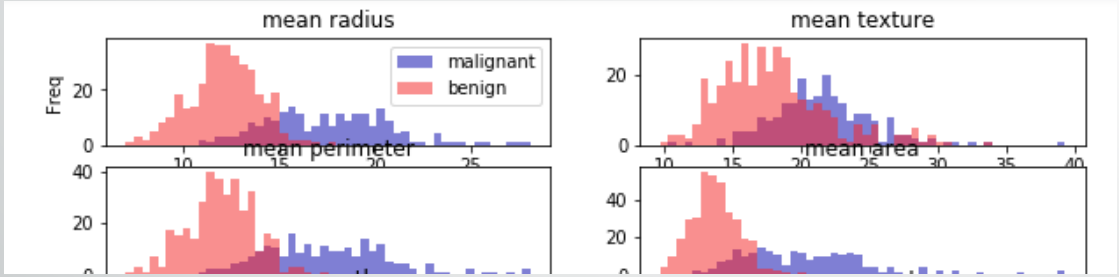
- Algebra helps to organize information where certain mathematical structures are present.
- All data science algorithms use algebra.
  - Data is represented as vectors, matrices, groups etc.
  - Algorithm uses linear / non-linear functions to determine relationship between data and response.

Trained Model: Price= 36.98 +(-0.12 * CRIM) +(0.04 * ZN) +(-0.01 * INDUS) +(2.39 * CHAS) +(-15.63 * NOX) +(3.76 * RM) +(-0.01 * AGE) +(-1.44 * DIS) +(0.24 * RAD) +(-0.01 * TAX) +(-0.99 * PTRATIO) +(0.01 * B) +(-0.50 * LSTAT)

Trained Model with polynomial degree 2: Intercept = 31.65; Coeff 1 -402.75 Coeff 2 -50.07 Coeff 3 -133.32 Coeff 4 -12.00 Coeff 5 -12.71 Coeff 6 28.31 Coeff 7 54.49 …. Coeff 100 34.47 Coeff 101 11.22 Coeff 102 1.14 Coeff 103 3.74 Coeff 104 31.38

Concepts: Statistics

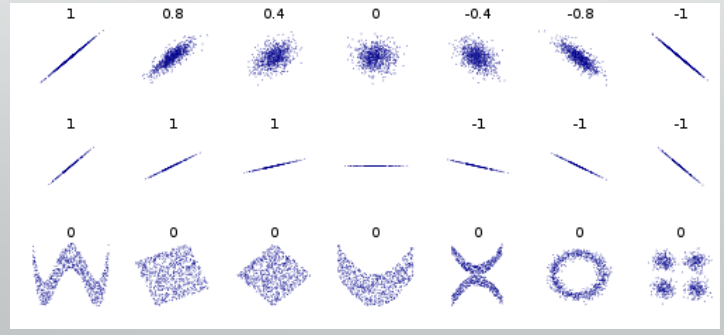
- Data science includes gathering data, analyzing and forming conclusions. Statistics is a critical component of these steps.
- Let us look at 3 specific concepts: Sampling, Correlation, Simpson's Paradox



- This cancer dataset has 30 features with target classes of "malignant" and "benign"
- There are 212 rows for "malignant" and 357 for "benign". Improper sampling will make the model skew to one or the other

Correlation can help and hurt:

- between features and target
- between features



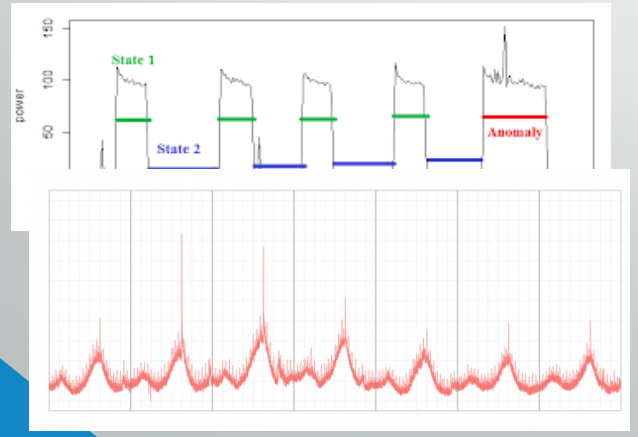
Simpson

coast	degree	# of members	avg. # of friends
West Coast	PhD	35	3.1
East Coast	PhD	70	3.2
West Coast	no PhD	66	10.9
East Coast	no PhD	33	13.4

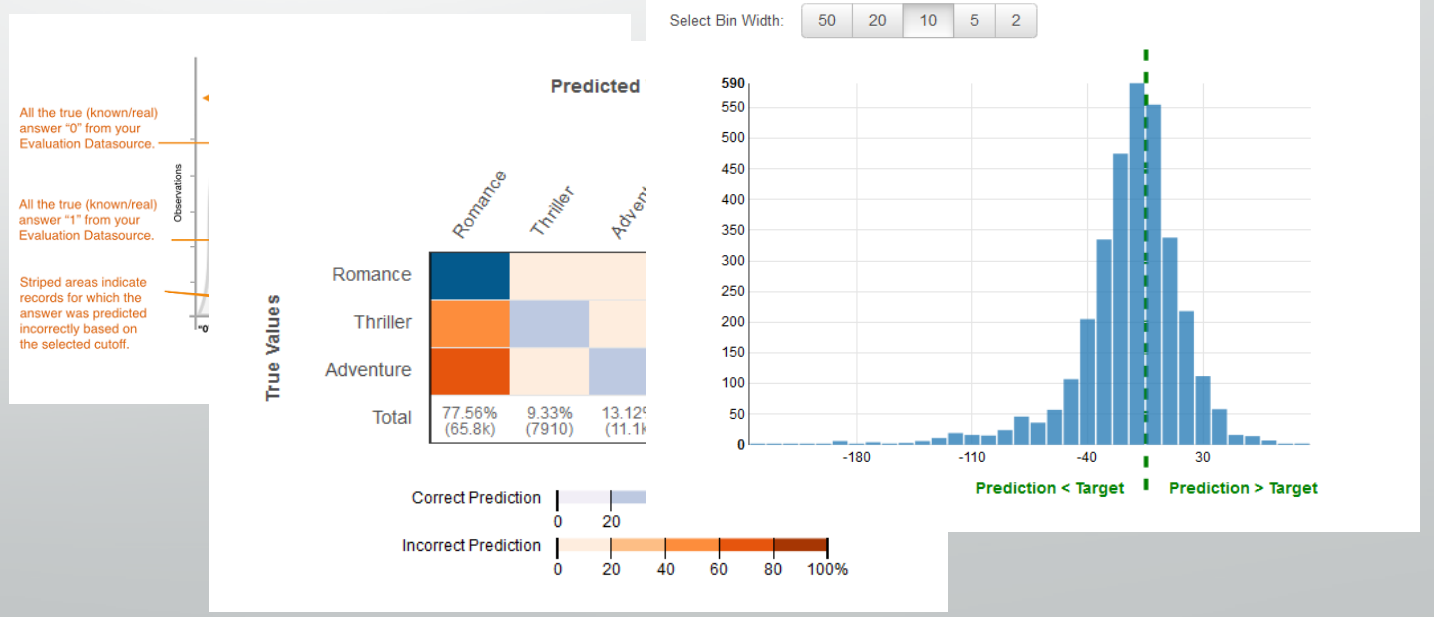
Concepts: Probability

- Data science uses probability for quantifying uncertainty. Probability is used in building models as well as evaluating models.
- Let us look at 2 concepts: Distribution, Accuracy. Other concepts include Dependence/Independence, Conditional probability, Random variables, Hypothesis testing.

Identifying the distributions help in multiple data science efforts – example: Anomaly detection

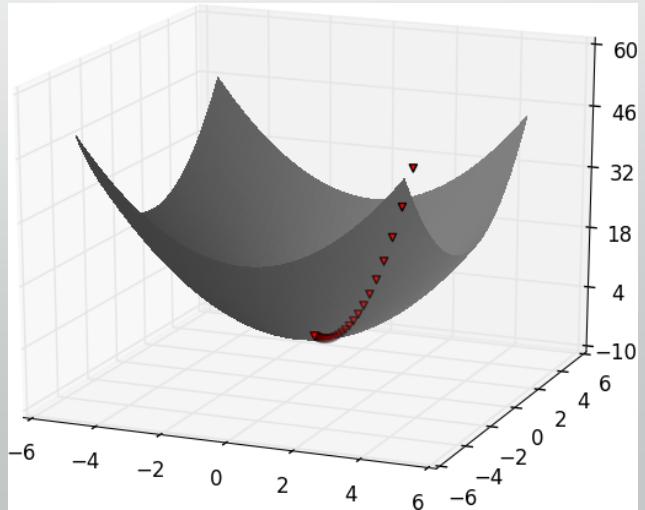


Results accuracy measurement



Concepts: Gradient Descent

- Data science is about finding the best model for the given data. "Best" represents the one that has minimum error or maximum likelihood of the data.
- To find the best, we have to solve optimization problem in which we will use gradient descent.
- Gradient descent gives the input direction in which the function most quickly increases.



Business Problems: Descriptive Analytics

- Descriptive analytics presents the data in a format for understanding "What has happened" or "What is happening?"
- These use data aggregation and data mining methods. Typically, a BI software is used

How many cars are sold last year?

How many patients were diagnosed with disease A last year?

How many security violations have occurred by region?

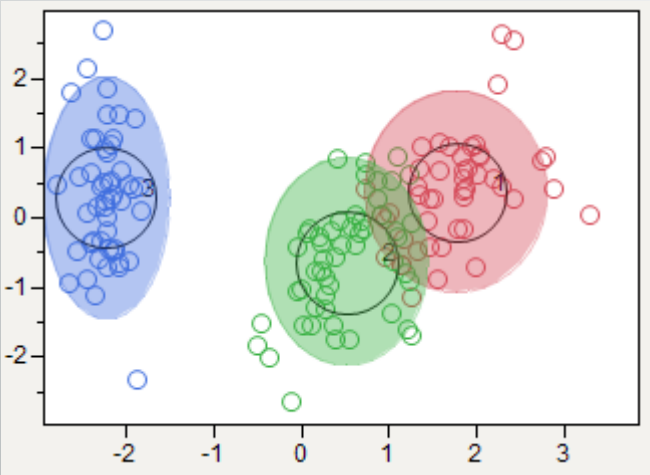
How many shipments are behind schedule?

How many and which machines are down?

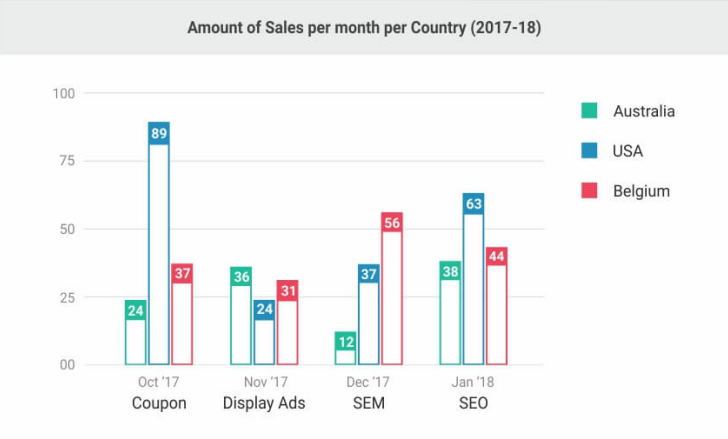
Business Problems: Exploratory Analytics

- Exploratory Analytics presents the data in a format for understanding main characteristics. Typically, a BI software is used

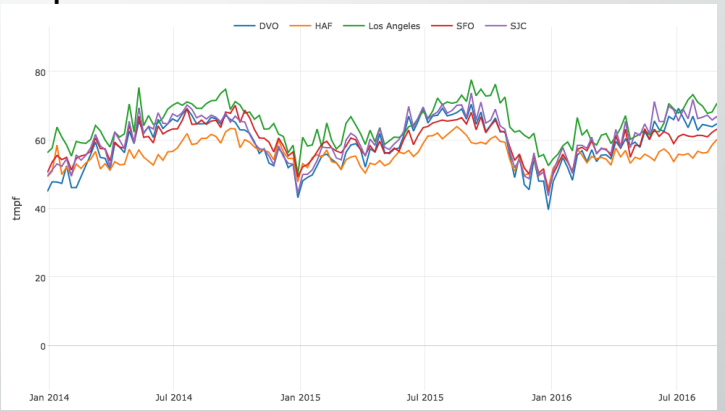
Segment: Do we have product-region groups within customers?



Trend: Are there trends that describe specific behavior?

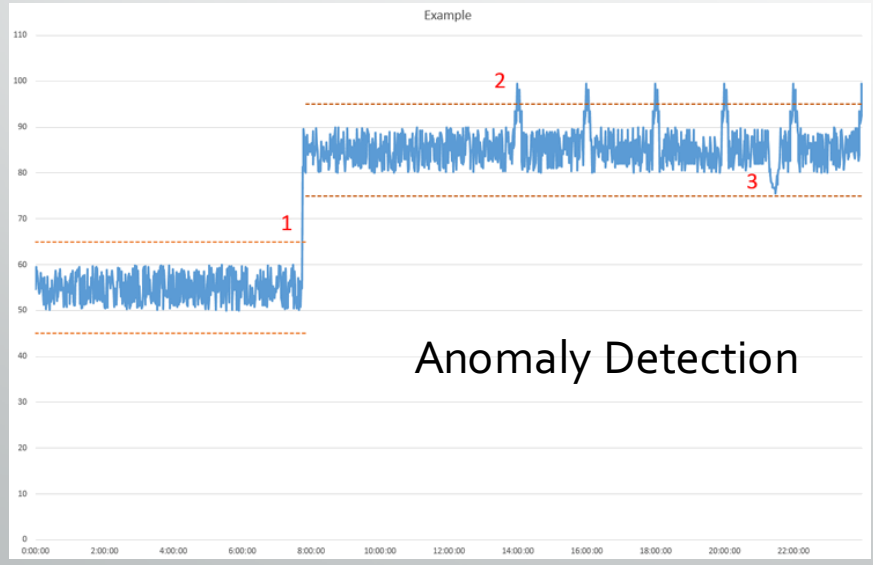


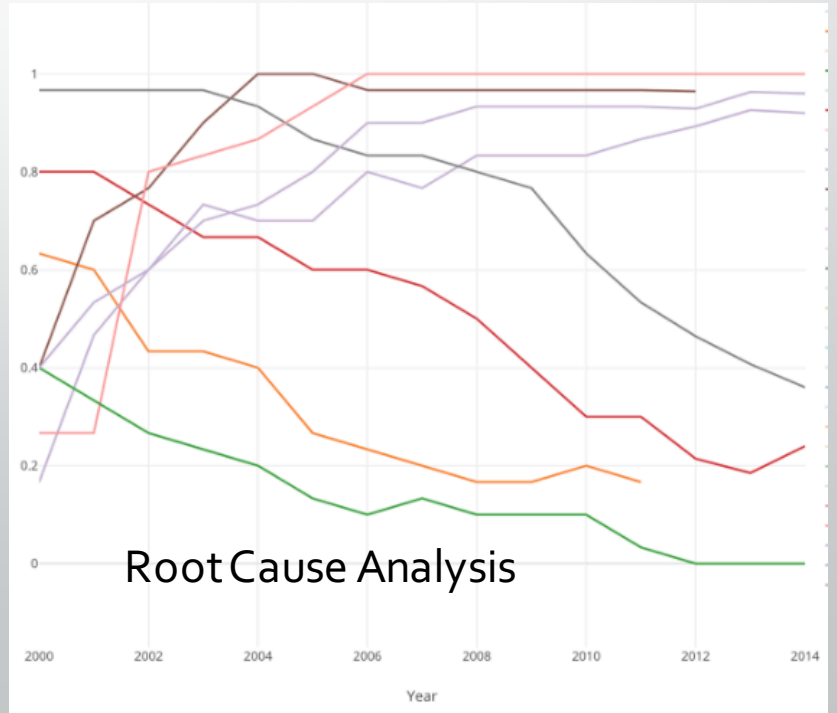
Correlation: Are there dependencies in the data?



Business Problems: Diagnostic Analytics

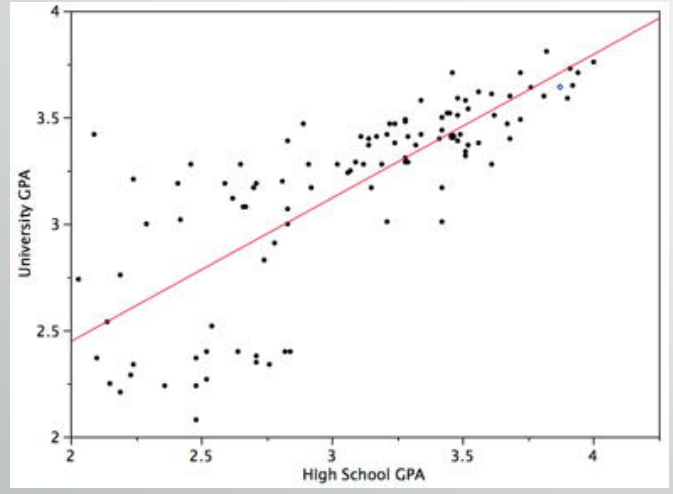
- Diagnostic analytics typically answers "Why did it happen?"
- Examples:
  - Why did we sell more cars last year compared to year before?
  - Why did we have more machines down?
  - What caused shipment delays?

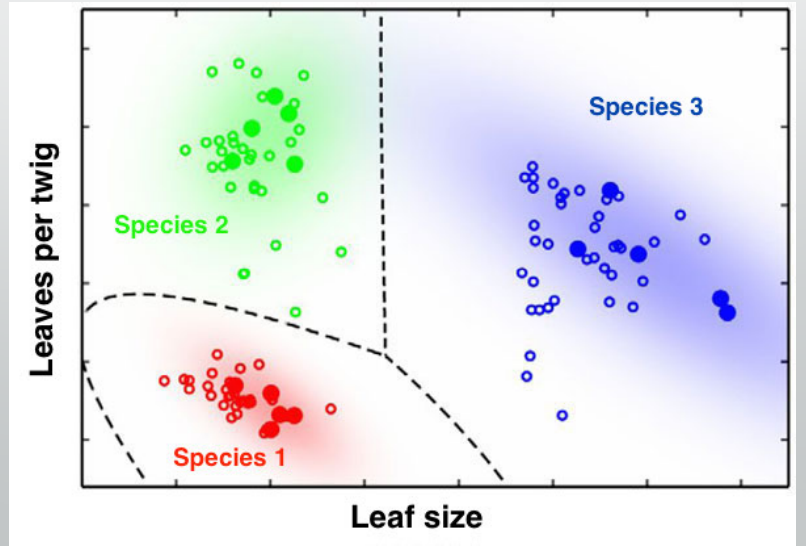


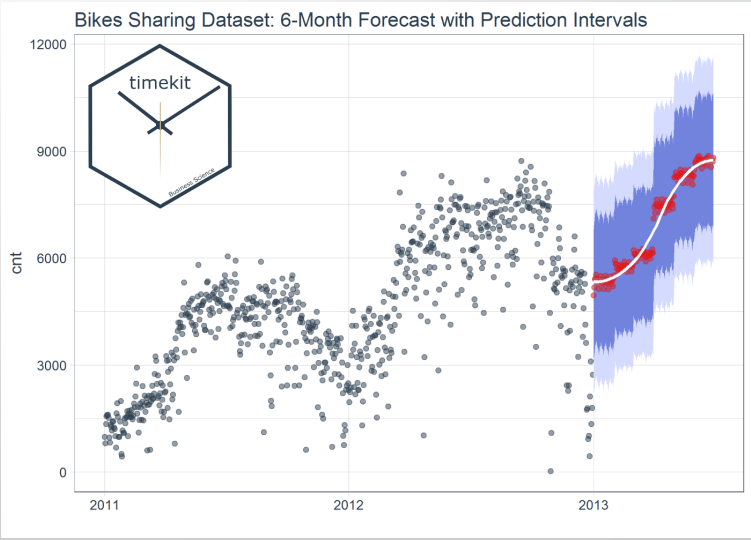


Business Problems: Predictive Analytics

- Predictive analytics typically uses data science (ML/AI) techniques to answer "What could happen in the future?"
- Predictive analytics can use regression, classification or forecasting methods.







Business Problems: Prescriptive Analytics

- Prescriptive analytics generates recommendation for "What should we do?"
- These typically use simulation, prediction and optimization methods.

What specific discount program need to be announced for improving sales?

Which product mix should be produced in each factory to meet shipment schedules?

What should be the maintenance schedule to avoid predicted failures?

What security measures should be implemented to improve campus security?

What mix of products to place in store racks to improve sales?

Data Science end-to-end example

- Classification example: Based on patient data, build a model to identify malignant and benign cancer.
- Regression example: Based on housing data, build a model to determine price of a house based on given attributes.
- Forecasting example: Build a model to forecast shampoo sales.

You are managing a fleet of 5 trucks. You operate them between western states.

- Business goals could be:
  - Maximize utilization of truck and its space; Maximize revenue per unit truck space volume
  - Minimize breakdowns / Minimize shipment delays
  - Minimize empty or lower than threshold miles
- Data science problems would include:
  - Demand forecasting
  - Truck failure prediction
  - Optimal pricing
  - ETA prediction

You are managing a retail store

- Business goals could be:
  - Maximize store revenue
  - Maximize loyalty program participation
  - Maximize time spent at store
  - Minimize operating cost
- Data science problems would be:
  - Forecast demand for each product / product family
  - Determine products that are most likely bought together
  - Optimize location of products in shop
  - Optimize support staff
  - Optimize pricing

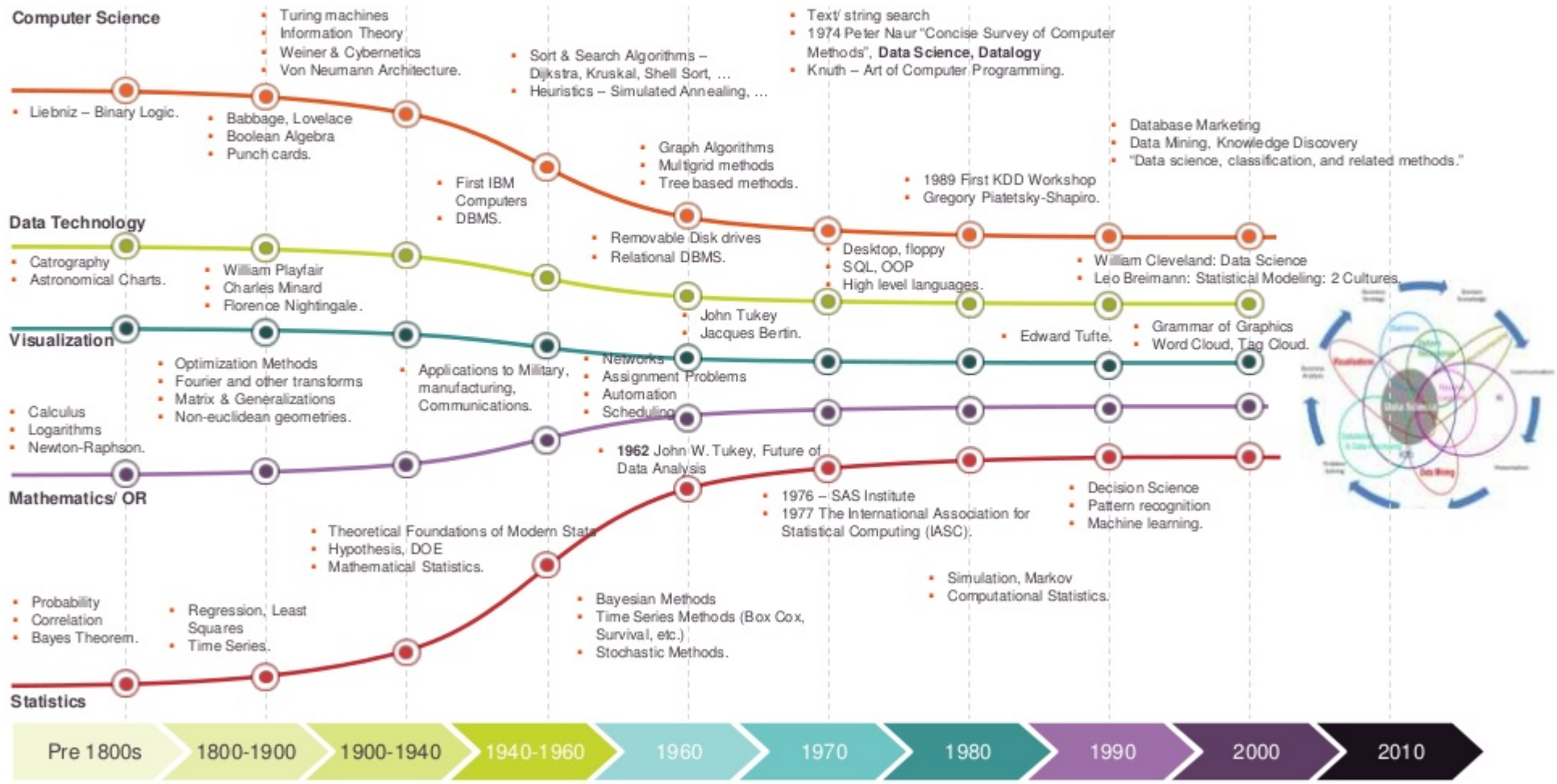
You are managing a factory

- Business goals could be:
  - Maximize production
  - Minimize breakdowns
  - Minimize quality failures
  - Minimize product cost
- Data science problems would be:
  - Forecast demand
  - Predict machine failures
  - Predict product quality
  - Optimize product mix
  - Optimize maintenance schedule

Data: Why focus on it?

- Between data and algorithms, most of the recent achievements in data science are due to data.
  - 2011 success of IBM's Jeopardy uses algorithm from 20 years earlier – but, with 8.6M data from Wikipedia, Wikitionary, Wikiquote etc.
  - 2014 imagenet success of Google uses CNN from 25 years earlier – but, with ImageNet data that is published in 2010.
  - 2016 success of Google's DeepMind in playing Atari games uses Q-learning algorithm from 23 years earlier – but, with Arcade Learning Environment dataset from 2014.

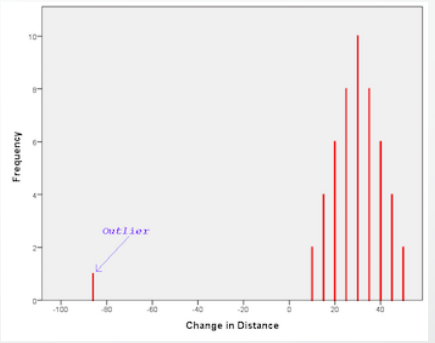
A brief history of Data Science



Data Engineering

- Many names for this step: data preparation, cleaning, munging, wrangling, validation, merging etc.
- Having the representative dataset is key for accuracy of data science. Preparing that data for analysis is often where majority of time is spent.
- Topics:
  - Data sourcing, transformation, merging to collect right dataset – broad area that includes connected things/IoT, data storage, merging contextual data etc.
  - Data discovery, exploration and visualization - broad area for understanding the data and its relevance to business goal
  - Data cleaning
  - Feature engineering
  - Data representation

Data Cleaning



- Real world data is rarely clean.
- Typical cleaning includes
  - Conversion of data type – example: string to float
  - Parsing – example: CA 95054 to State=CA, Zip=95054
  - Removing outliers
  - Handling null values
  - Imputing missing values – last known value, moving average, value estimate
  - Frequency matching
  - Encoding – one hot, dummy, effect, feature hashing, bin counting

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

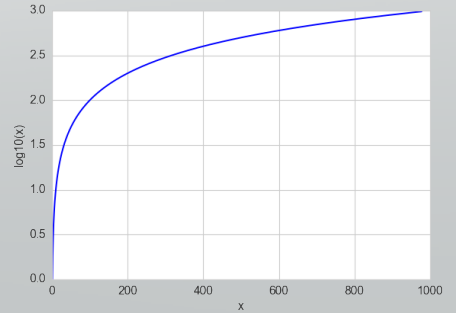
→ One Hot Encoding

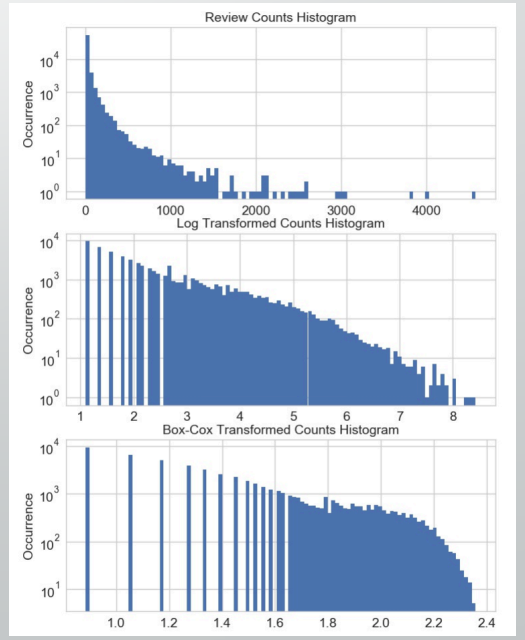
Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Airport	e_1	e_2
SFO	1	0
NYC	0	1
MAA	0	0

Feature Engineering

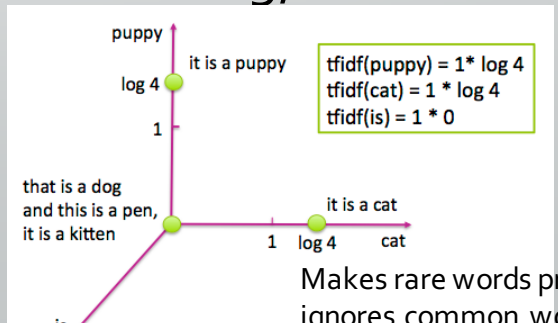
- Feature engineering is the process of creating features based on domain knowledge of data so that algorithms can be applied to derive insights.
- Typical feature engineering work includes:
  - Decision on data and feature space – example: Bob likes Burger; Katie likes Pizza. Person=Data; Food preference = Feature?
  - Binning – grouping to scale – example: song preference
  - Transformations:
    - Log transformation –
      - compresses range of large and expands range of small numbers
    - Power transformation
  - Scaling – normalization





Feature Engineering – Contd.

- Singleton features Vs. Interaction features – example: Product A sales is in zip 95054, by 18-25 year old, along with Product B and T
- Feature selection
  - Filtering – pre-process to remove the features that won't be useful in the model
  - Wrapper methods – procedures that wraps the algorithms to determine the accuracy with various combinations of features. Typically, expensive unless the wrapper is well designed
  - Embedded methods – feature selection is part of training such as regularization methods
- Text data feature engineering – broad topic that includes counting, collocation, phrase detection, tf-idf



Makes rare words prominent; ignores common words

Feature Engineering – Contd.

- Data representation - Dimensionality reduction / PCA
- Clustering based featurization
- Automated feature engineering – lots of work in this for image feature engineering

Algorithms Categories

1. Supervised learning – Example: Classification or Regression
   - Determine the probability of car failure in the next week based on historical operation and failure data.
   - Determine the price of a home in a neighborhood based on historical sales data.
   - Classify email as spam or not-spam

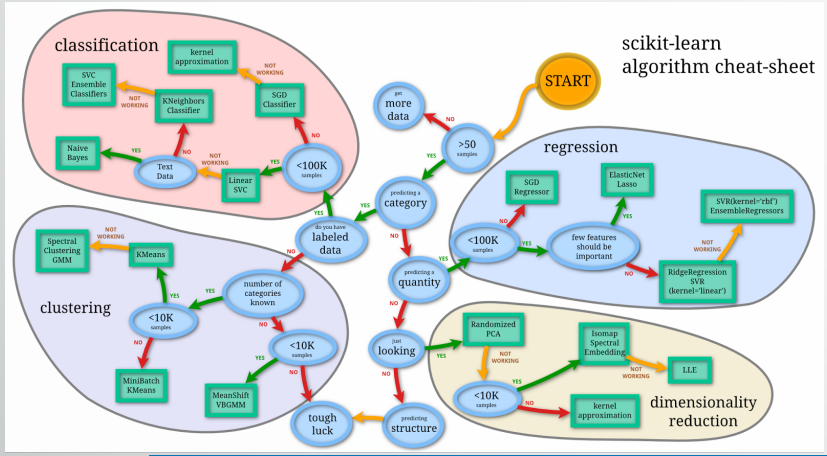
2. Unsupervised learning – Example: Anomaly detection, Association rule learning
   - Fraud detection
   - Determine the anomalous power consumption of HVAC.
   - Determine which products the customers will buy if s/he buys milk.

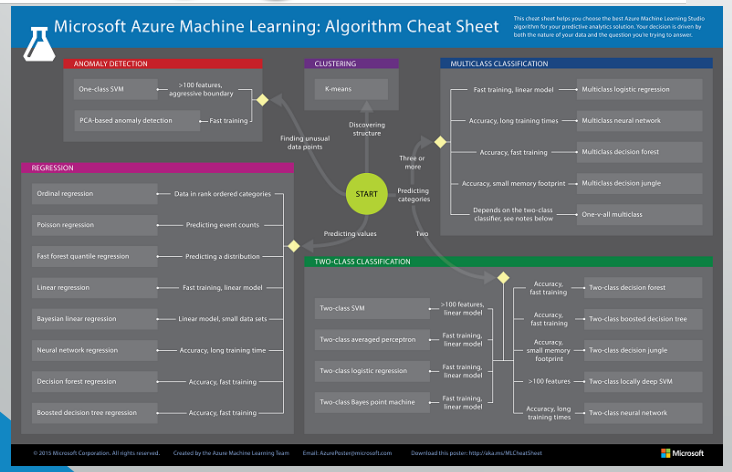
3. Reinforcement learning
   - Teaching robots tasks – machine calibration, quality control, process optimization
   - AGV programming - stay within limits, perform specific tasks at locations that have high reward etc.
   - Automated inventory management between manufacturers, DC/warehouses, retailers
   - Automated energy management in data centers

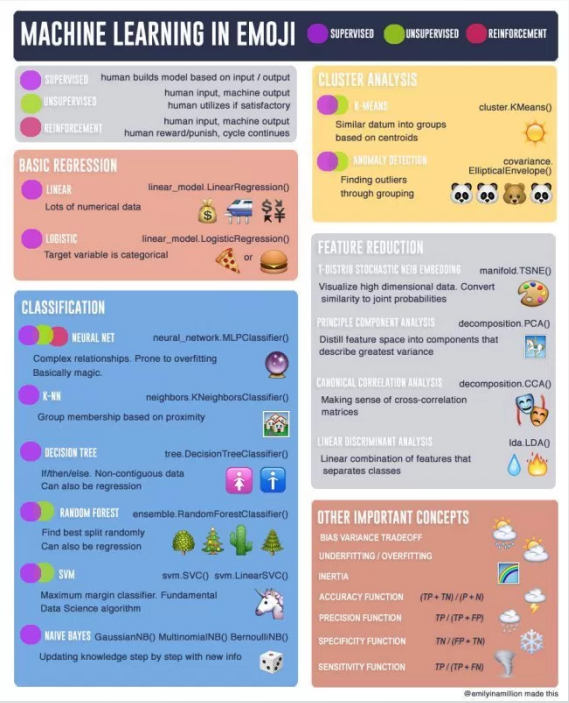
Algorithms - There are many…

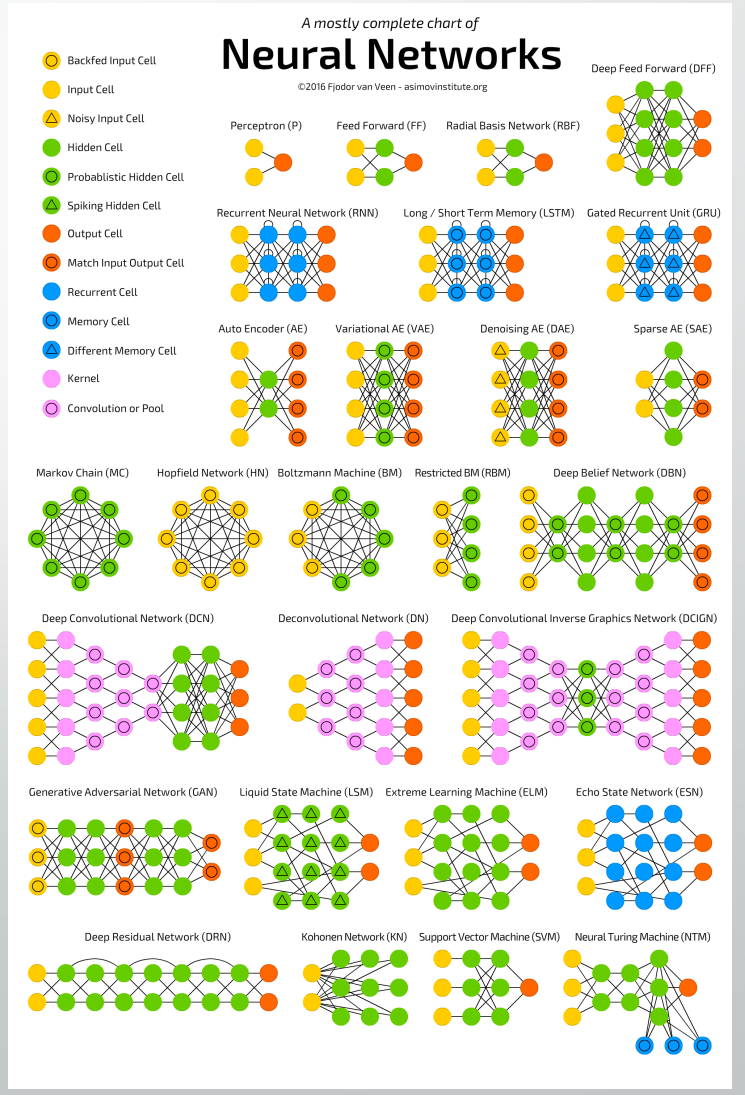
Algorithms - Contd

- Cheat sheet of algorithms

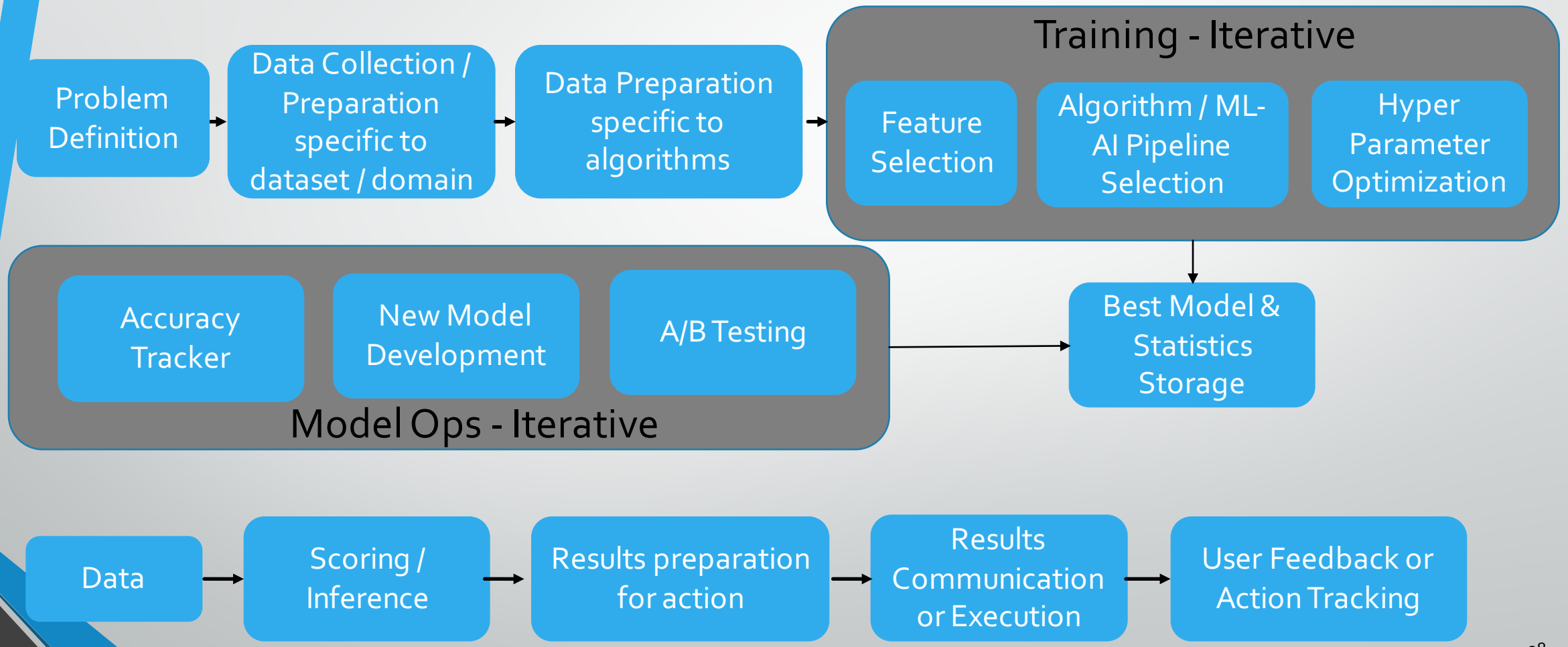








Typical Data Science Process



Business Goal and Problem Definition

- You are the head of marketing at Tesla. Your goal is to make sure positive stories on Tesla is communicated broadly.
- Today's task is to design a system that will re-tweet positive tweets on Tesla.
- Problem definition: Identify positive tweets on Tesla. Re-tweet them.

Data Collection

- We need to collect tweets on Tesla. Say, we collect tweets from #tsla, @tesla, @elonmusk, @teslalife, etc.
- Given the subjective sentiments of "Positive" and "Negative", let us say, we will label the tweets. Note that, this means we will use supervised learning method for classifying tweet.
- We need to make sure we have sufficient representation of "Positive" and "Negative" so that we can build model with high accuracy for identifying sentiments.

Data Preparation / Feature Engineering

- We split the dataset into Training and Test dataset. Training dataset will be used for building the model. Testing dataset will be used for verifying accuracy of the model. We need to split it such that both datasets have balanced representation of sentiments.

- As tweets would contain some common words that doesn't help in classifying tweet, we need to remove those from features. We also need to remove symbols such as #, @ etc. Then, we will extract the remaining word features.

Training / Model Building

- We will build models with multiple algorithms:
  - Naive Bayes Classifier
  - Decision Tree Classifier
  - Maxent Classifier
- We will check their accuracy on how well they identified sentiments of tweets in test dataset

Let us solve the problem

Take Away

1. We are collecting more data annually now than what we have collected in the entire history of human race. Deriving insights from those data are essential. Data science is omnipresent.
2. Data science isn't one thing – it encompasses multiple mathematical methods, requires multiple business and technical skills and, serves varied purposes. So, everyone has a part to play in a data science project.
3. Data (and, the business goal) is more important than algorithms. We have multiple open source algorithms that we can start with and fine-tune for specific problem. Focusing on business goal helps us to shape the data collection, preparation and action execution.