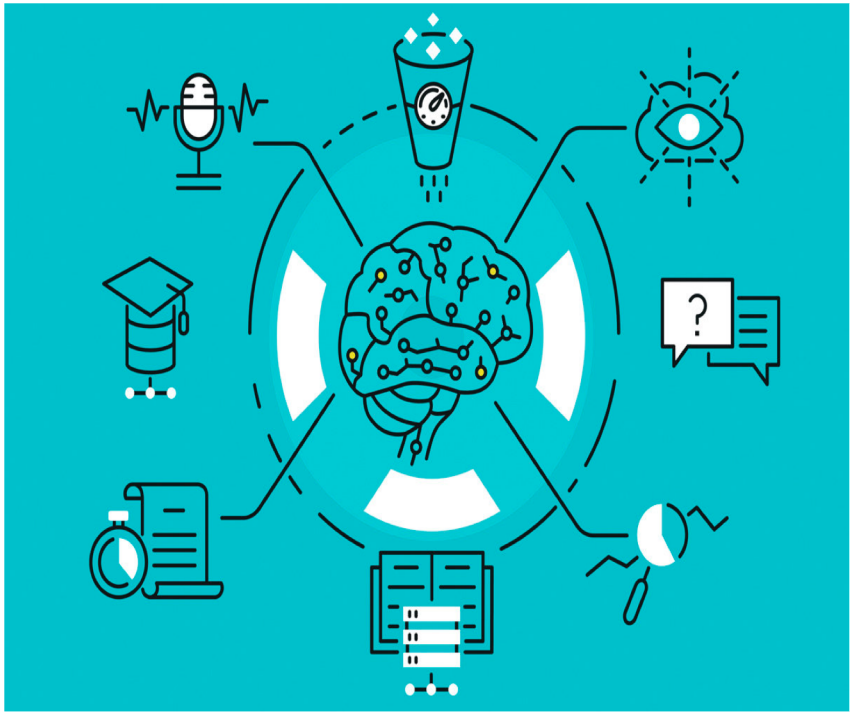
An Introduction to Machine Learning/Data Science

By Supreet Takkar
Data Scientist, RBC



Register here: https://bit.ly/RBC-CUTC-2020

AGENDA

01 **What is AI, Machine Learning and Deep Learning?**

How we use them in our daily lives, what those terms mean

02 **Data Science**

What I do as a Data Scientist, cool things *you (!)* can do with Data Science

03 **Machine Learning Problem Types**

Types, differences

04 **Steps in a DS problem**

Collection, EDA, Preparation, Modelling, Evaluation

05 **Statistics in the midst of this**

How Statistics relates to Data Science

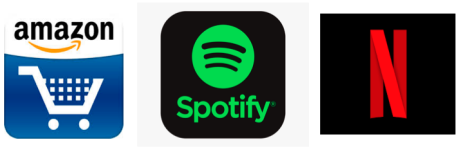
06 **Data at scale**

Working with Big Data

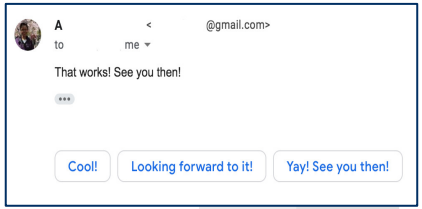
AI in our Daily Lives

01 Alexa, Siri, Google Assistant

02 Shopping and Music, Shows, Movies Recommendations

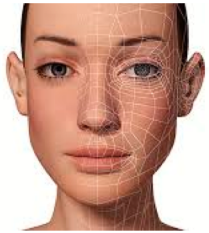


03 Smart Replies/Spam Filtration in Gmail



AI in our Daily Lives

04 Facial Recognition



05 Chatbots

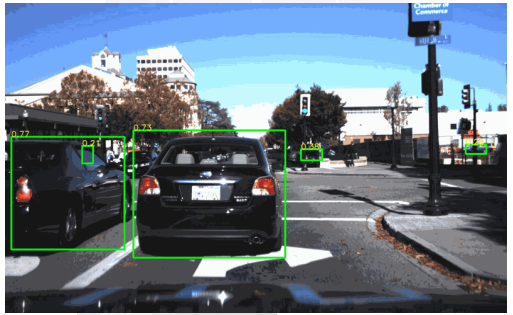
finn.ai

RBC

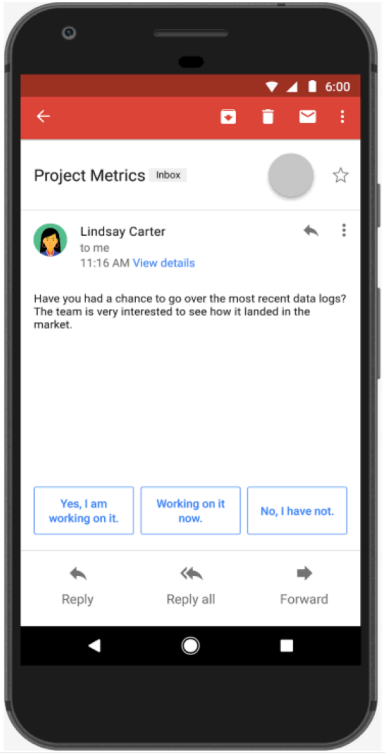
06 Navigation and Travel, Self-Driving Cars

UBER

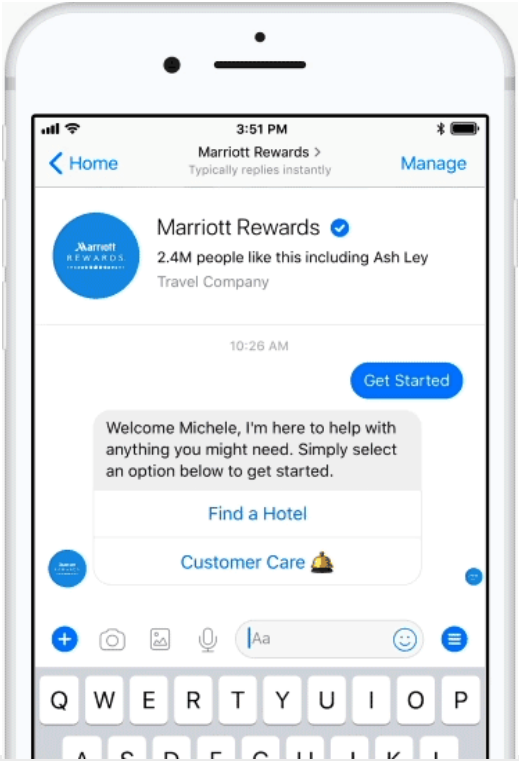
lyft



LIVE: Smart Replies

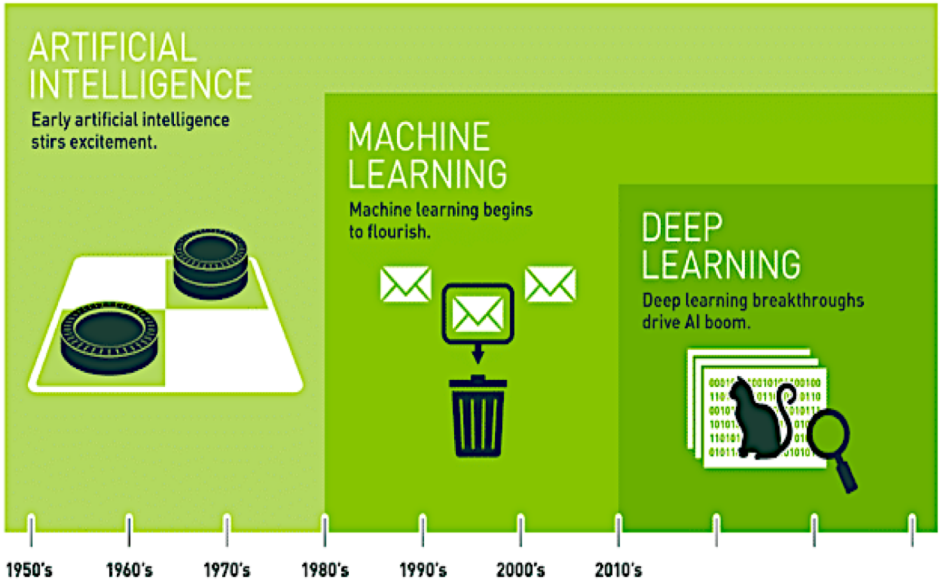


LIVE: Chatbots



Evolution of Artificial Intelligence

- In the 1940s and 1950s, a handful of scientists from a variety of fields began to discuss the possibility of creating an artificial brain.
- The first was widely known for the creation in 1952 of the self-learning **Checkers-playing program**.
- The field of artificial intelligence research was founded as an academic discipline in 1956.
- Before 1949 computers lacked the technology to perform such “intelligence”. In the early 1950s, the cost of leasing a computer ran up to $200,000 a month, so only prestigious universities and big technology companies could afford to have them.
- From 1957 to 1974, AI flourished. Computers could store more information and became faster, cheaper, and more accessible. Machine learning algorithms also improved and people got better at knowing which algorithm to apply to their problem.
- Over the next 30 - 40 years, through the efforts of academically oriented scientists, Machine Learning was turned into an independent mathematical discipline.
- Deep learning technology, which is meant to simulate biological neural networks in brains and was a special branch of ML, gained more popularity in 2010s.



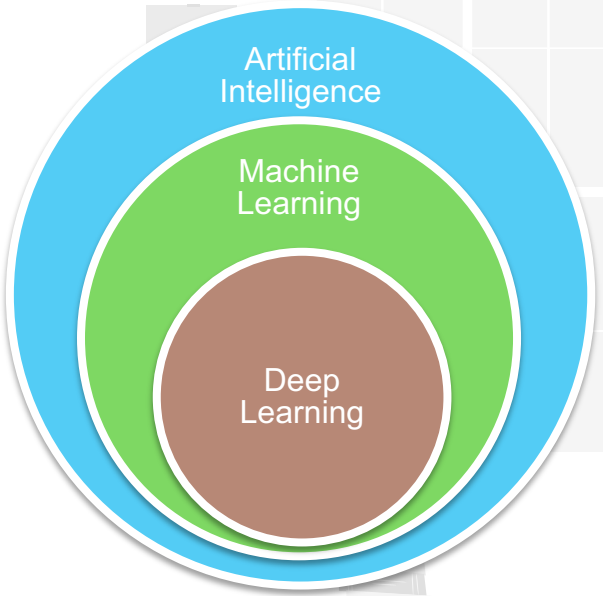
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Artificial Intelligence, Machine Learning and Deep Learning

- **Artificial intelligence** can be considered the all-encompassing umbrella.
- It implies the computer's ability to be able "think", behave and do things as human beings might do them.

- **Machine Learning's** intention is to enable the machines to "learn" by themselves using the provided data *without* being explicitly programmed to do it.
- It is a technique to realize AI

- **Deep Learning** is the next evolution of Machine Learning – it is a technique for realizing ML.
- DL algorithms are roughly inspired by the information processing patterns found in the human brain.
- Just like we use our brains to identify patterns and classify various types of information, deep learning algorithms can be taught to accomplish the same tasks for machines.



Data Science

Data science is a multidisciplinary term for a whole set of tools and techniques of data inference and algorithm development to solve complex analytical problems.

It makes use of scientific processes, methods, and algorithms to make it happen. Initially, the goal was to identify hidden patterns in raw data to help a business to enhance and expand their profits.

The term Data Science become a buzzword when Harvard Business referred to it as **"The Sexiest Job of the 21st Century"**.

Applicable across multiple industries – Healthcare, Technology, Finance, and so on.

Formula

- Problem Solving
- Ability to extract insights from Data/Data Mining
- Machine Learning
- Statistics
- Math - Derivatives
- Data Visualization
- Communication of ideas, research, results
- Ability to code (Python, R, Java, etc.)
- (Optional) Dealing with Big Data – Spark

- Domain knowledge!

ML Components

Data

- Want to detect spam? Get samples of spam messages. Want to forecast stocks? Find the price history. Want to find out user preferences? Parse their activities on Facebook
- The more diverse the data, the better the result.
- Data collection techniques: Manual (few errors, but time consuming) and automatic (cheaper; gather what you can approach)

Features

- Parameters/variables
- Those could be car mileage, user's gender, stock price, word frequency in the text.
- That's why *Feature Selection* is important and time consuming! Also, its a source of error.

Algorithms

- The method you choose affects the precision, performance, and size of the final model. There is one important nuance though: if the data is crappy, even the best algorithm won't help.

Caution!
Garbage in → Garbage out!

Machine Learning Problem Types

Supervised Learning

Describes a class of problem that involves learning from a correct input-output pairs that we use to train the algorithm.

Classification: Predict a category

- Assign observations into discrete categories
- Will the customer leave us for the competitor?
- Does a patient have cancer?
- Is *this* transaction fraudulent?
- *Algorithms: Logistic Regression, Decision Trees*

Regression: Predict a numerical label.

- What will be the price of *this* house?
- What will be the price of *this* stock?
- What will be the temperature tomorrow?
- *Algorithms: Linear/Polynomial Regression*

Unsupervised Learning

- Finds patterns
- The right answers are not known.
- The "right answers" are unobservable, or infeasible to obtain, or maybe for a given problem, there isn't even a "right answer".
- Example – what types of customers do I have?
- Anomaly detection – which transactions fraudulent?

- Algorithms: Clustering, Dimension Reduction, etc.

Reinforcement Learning

- The agent learns to achieve a goal in an uncertain, potentially complex environment by itself (no direct instructions)
- The artificial intelligence faces a game-like situation. It then employs trial and error to come up with a solution to the problem.
- To get the machine to do what the programmer wants, the artificial intelligence gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward.
- Leverages the power of search and many trials, and can run parallel gameplays when provided with resources.

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

Regression: Linear and polynomial regression

Regression

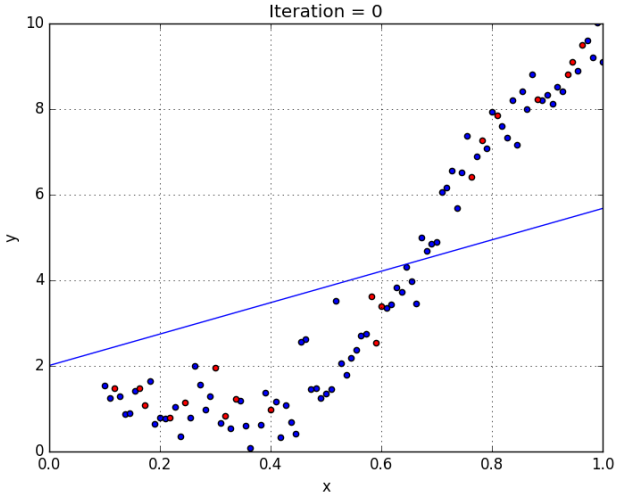
Using the relationship between variables to find the best fit line or the regression equation that can be used to make predictions

Linear Regression

Defines the relationship between two variables by fitting a linear equation to the data.

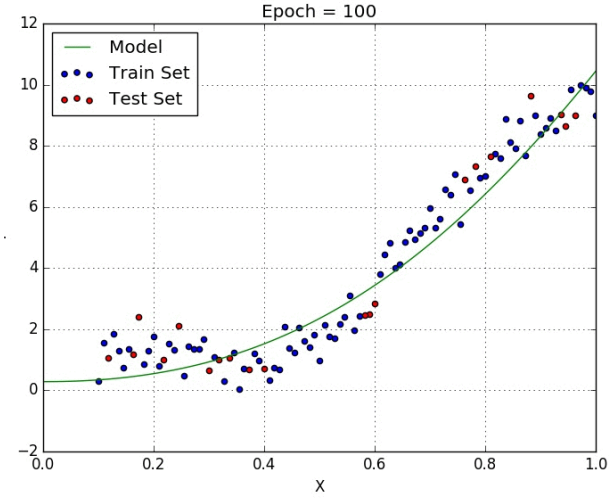
Polynomial Regression:

A polynomial term: a quadratic (squared) or cubic (cubed) term turns a linear regression model into a curve.



Plot a line on the original scatter-plot by running every value of x through the linear equation:

$y = b_0 + b_1 * x_1$



More flexible than Linear Regression.
Plot a flexible line using powers of the variable, e.g., x^2, x^3, etc.

$y = a + bx + c\,x^2$

Classification: Decision Tree Example

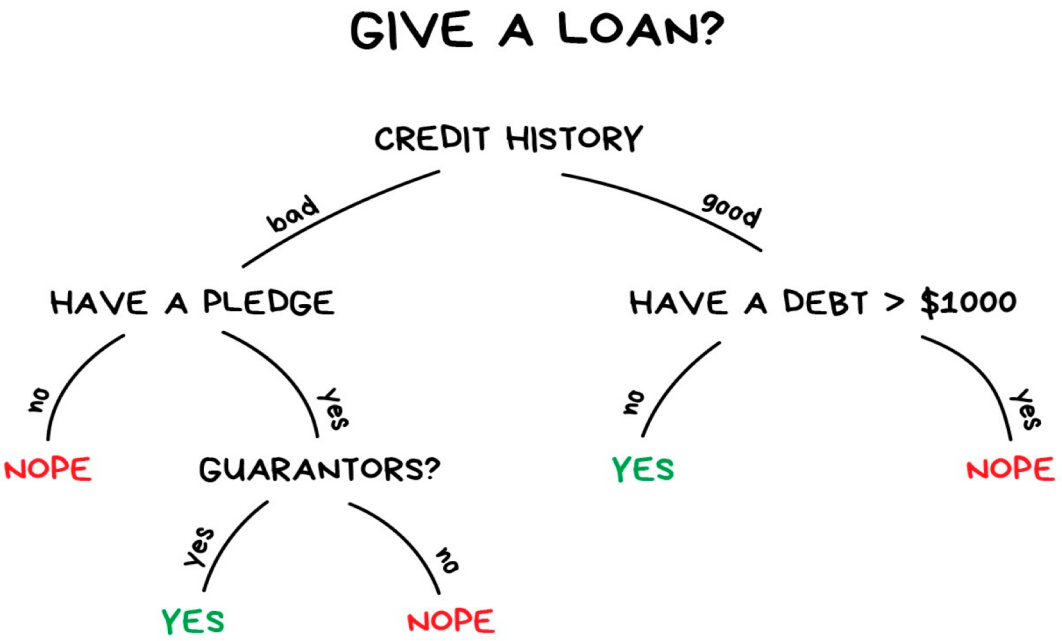
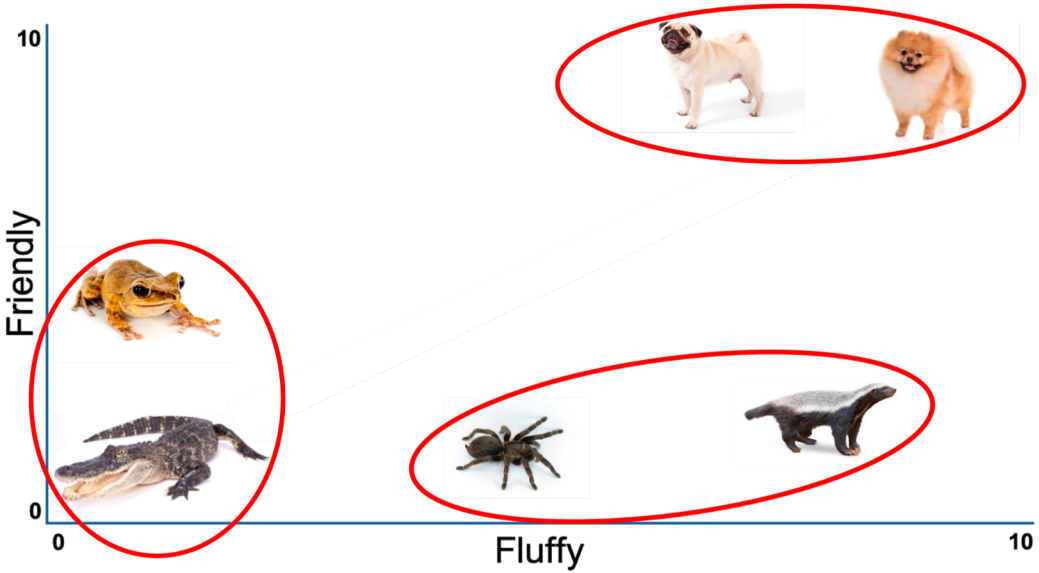


Image Classification Problem: is It a dog?

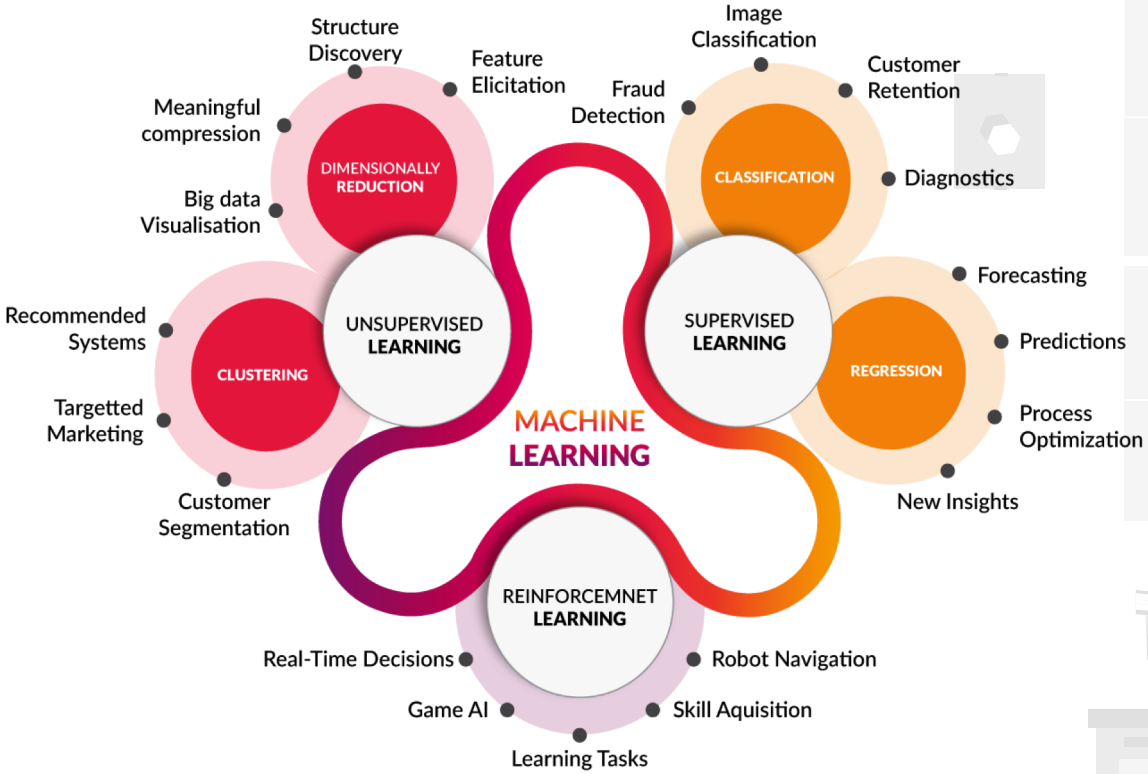
- Identifying/Classifying an object in an image – dog or not a dog.
- Confusing images should be distinguishable

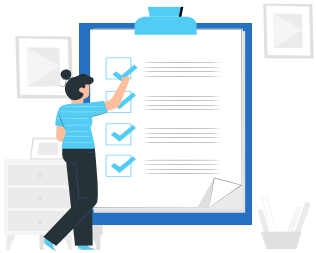
Clustering



- Animals classified on an x-y axis based on 2 features: their friendliness and their fluffiness.
- A frog is a bit friendlier than an alligator, but it's not at all fluffier.
- A Pug isn't quite as fluffy as a Pomeranian, but it is is a bit friendlier. And then you have the honey badger and tarantula, which are a little fluffy, but not particularly friendly.
- Based on their scores, k-means clustering aims to figure out how best to group them.

Machine Learning Problem Types





Steps to 'solve' a Data Science Problem

Data Collection

Manual/Automatic: Quantity and Quality determine results

Exploring the Data

Exploratory Data Analysis

Data Preparation

Cleaning, Feature Engineering

Modelling

Choose a model, train
Tweak the assumed parameters

Evaluation

Use test data to evaluate the results

Prediction

Using the model to predict future observations

Kaggle example

Data Collection

```
<?xml version="1.0" encoding="UTF-8"?>
<Print_Records>
<form1>
   <Name>Ego ille</Name>
   <Address>345 Park Aven</Address>
   <City>San Jose</City>
   <State>CA</State>
   <ZipCode>94087</ZipCode>
   <Country>USA</Country>
</form1>
<form1>
   <Name>Johnson</Name>
   <Address>1 Almaden Blvd</Address>
   <City>San Jose</City>
   <State>CA</State>
   <ZipCode>94089</ZipCode>
   <Country>USA</Country>
</form1>
</Print_Records>
```

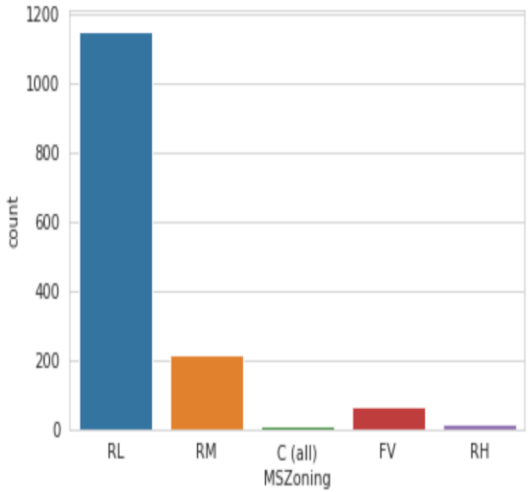
Data types

Structured	Organized, tabular data.
Unstructured	Not organized in a pre-defined manner Simple text files, images, video
Semi-structured	Does not reside in a relational database But that have some organizational properties that make it easier to analyze XMLs

Exploring the Data

```
#descriptive statistics summary
df_train['SalePrice'].describe()
```

```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```



In [7]:
```
ds_cat['MSZoning'].unique()
```

Out[7]:
```
array(['RL', 'RM', 'C (all)', 'FV', 'RH'], dtype=object)
```

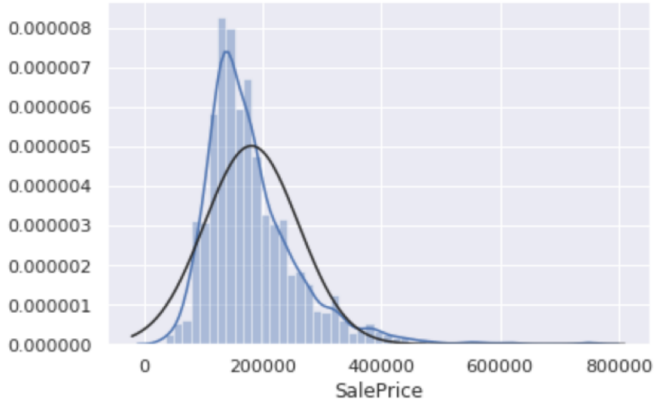
The critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

- Unique counts in each category
- What languages do you speak? - English, French, Spanish, Others
- Types of data – numerical, categorical, etc.
- Find nulls/holes in data
- Find outliers – plot histograms, different distributions, boxplots
- Feature contributions
- Correlation heatmaps
- Understand the problem!

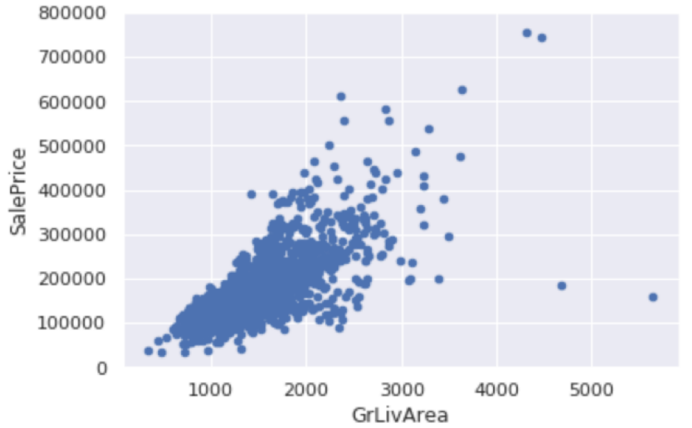
Exploring the Data

In [20]:

```
#histogram and normal probability plot
sns.distplot(df_train['SalePrice'], fit=norm);
fig = plt.figure()
res = stats.probplot(df_train['SalePrice'], plot=plt)
```




```
#bivariate analysis saleprice/grlivarea
var = 'GrLivArea'
data = pd.concat([df_train['SalePrice'], df_train[var]], axis=1)
data.plot.scatter(x=var, y='SalePrice', ylim=(0,800000));
```



Data Preparation: The scrubbing

- Also called Data preparation, cleaning, pre-processing, cleansing, wrangling
- Most of the time data you have can't be used straight away for your analysis: it will usually require some manipulation and adaptation

Common Data Cleaning Procedures:

1. Removing Missing Values: Data value that is not stored – blanks/nulls

- Can have a significant effect on the conclusions that can be drawn from the data.

2. Outliers: An outlier is a data point that is distant from other similar points. They may be due to variability in the measurement or may indicate experimental errors.

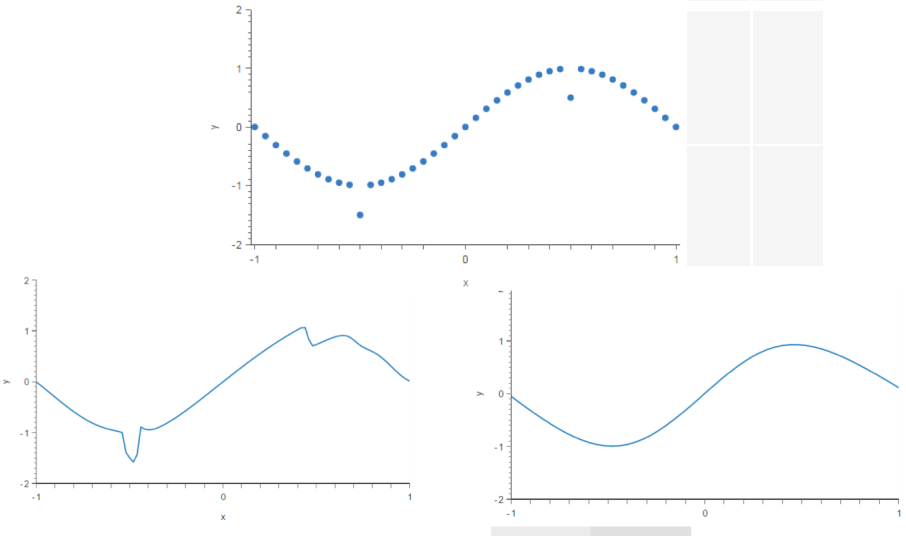
- If possible, outliers should be excluded from the data set.

and Feature Scaling, Encoding, Dimensionality reduction, balancing the data, etc.

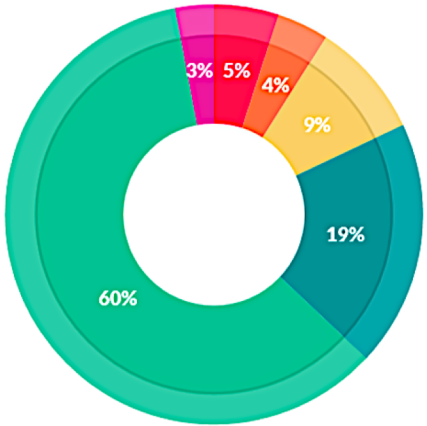
	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0



Perception vs Reality



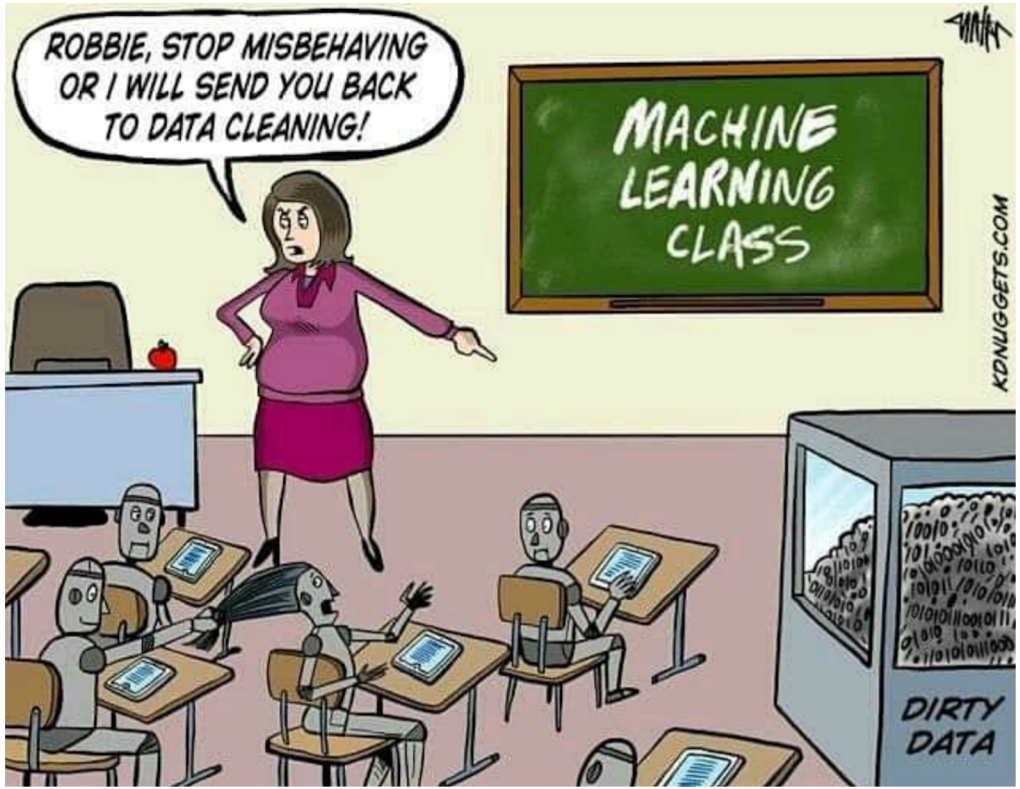
What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

79%

79% of Machine Learning is **Working with Data**

Cleaning data is the most time consuming task!


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

Modelling: Where magic happens!

Hold-out dataset

- Split data - The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen data
- Untouchable data.
- 80%-20%, 75%-25% Split
- Less data – problem!

Training

- "Training" or the learning phase
- The machine learning model learns to perform:

1. Classification to solve the problem using the provided data: Differentiating the emails you received as "*Inbox*" and "*Spam*" using logistic regressions.
2. We can also forecast values using linear regressions. We can also use modelling to group data to understand the logic behind those clusters.

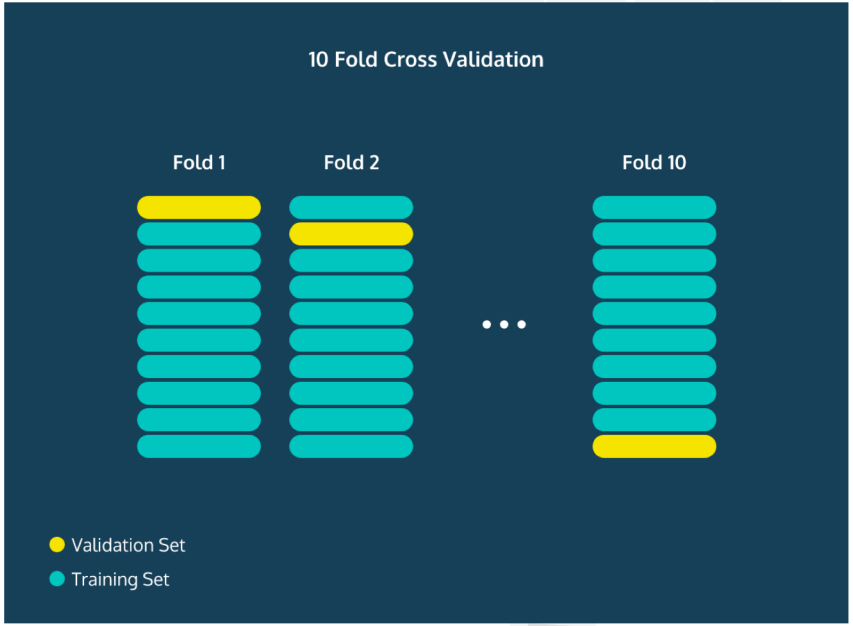
- Many algorithms already have an implementation – sometimes, you need to code from scratch
- Parameter tuning: grid search; cross-validation
- Document results

Validation

Cross-validation during training

- The dataset is randomly split up into ‘k’ groups. One of the groups is used as the test set and the rest are used as the training set.
- The model is trained on the training set and scored on the test set. Then the process is repeated until each unique group as been used as the test set.

- For example, for 10-fold cross validation, the dataset would be split into 10 groups, and the model would be trained and tested 10 separate times so that each group would get a chance to be the test set. It does this for all combinations and averages the result on each instance.
- The advantage is that all observations are used for both training and validation, and each observation is used once for validation.



Evaluation & Prediction

If you can't measure it, you can't improve it!
Estimate an unbiased generalization performance using the (never before seen) hold-out dataset.
Hold-out dataset purpose: no leaking of training data in the evaluation procedure.

Evaluation Metrics

- Accuracy, F-score, precision-recall, confusion matrix, Area under ROC curve (AUC)
- Statistical techniques also exist to the select a "better-performing method" machine Learning model out of many.

Prediction

Actually predict on unseen, future values!

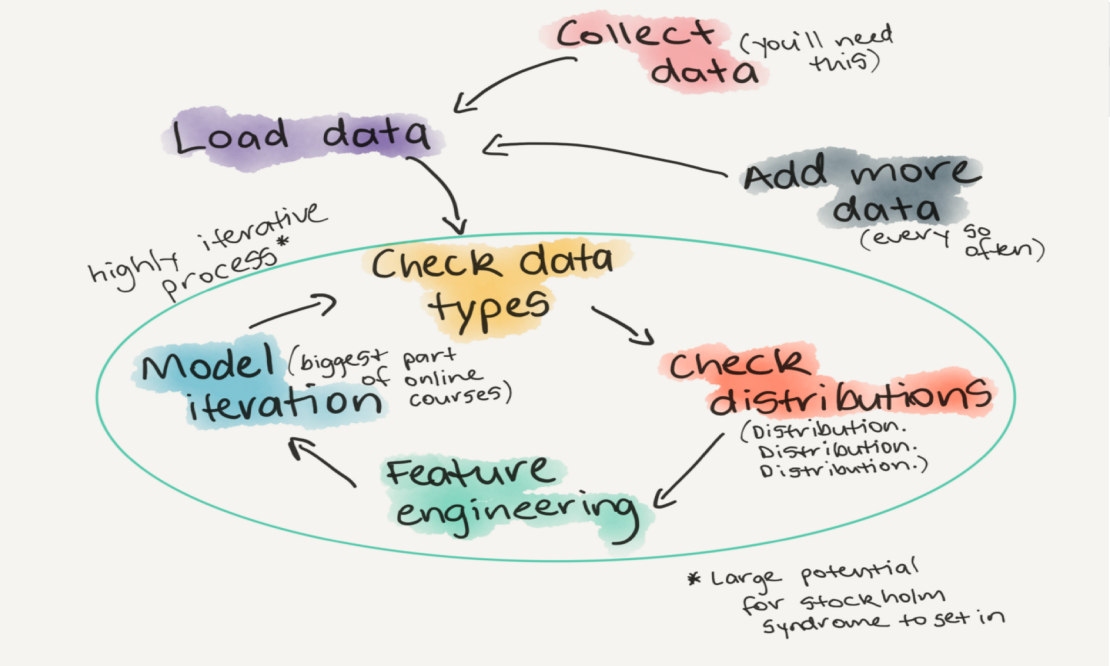
Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Precision = $\frac{\text{True Positive}}{\text{Actual Results}}$ or $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

Recall = $\frac{\text{True Positive}}{\text{Predicted Results}}$ or $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$

Accuracy = $\frac{\text{True Positive + True Negative}}{\text{Total}}$

The iterative nature of Modelling



And things change!

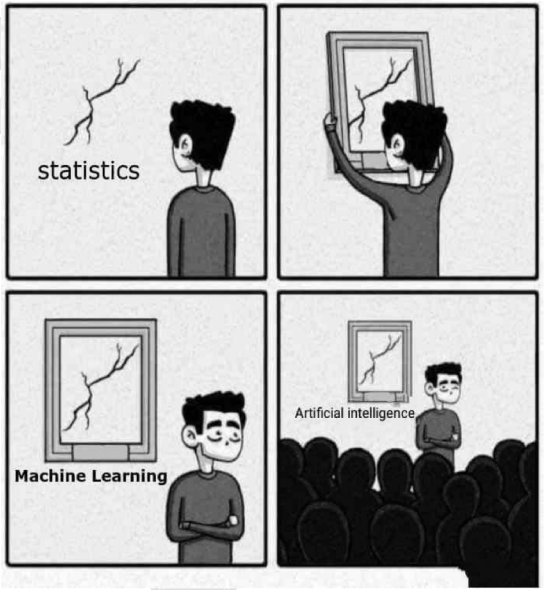
- Track feature drift
- Re-train!

Statistics in the midst of this

- This is caused in part by the fact that Machine Learning has adopted many of Statistics' methods.
- Machine Learning is largely a hybrid field, taking its inspiration and techniques from all manner of sources.
- Many answers have been given, ranging from the neutral or dismissive:

"Machine learning is glorified statistics"
v/s
"Machine learning is for Computer Science majors who couldn't pass a Statistics course."

- Machine learning is almost universally presented to beginners assuming that the reader has some background in statistics.
- It should be clear that these two approaches are different in their goal, despite using similar means to get there.
- The assessment of the machine learning algorithm uses a test set to validate its accuracy. Whereas, for a statistical model, analysis of the regression parameters via confidence intervals, significance tests, and other tests can be used to assess the model's legitimacy.



Data at scale

- **“Big Data”** refers to data that is so large, fast or complex that it’s difficult or impossible to process and manage using traditional methods.
- Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services.
- With the advent of the **Internet of Things** (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance.
- The emergence of machine learning has produced still more data.

Advantages

More data is good!

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data, can lead to a completely different (possibly, better) approach to tackling problems.

Challenges

- Data storage
- Processing – cleaning, filtering, analyzing
- Changing at a rapid pace

Big Data Technologies

Demo

- Kaggle example

THANKS!

Do you have any questions/comments?

Supreet Takkar
supreet.kt@gmail.com
linkedin.com/in/supreetkt/
Register here: https://bit.ly/RBC-CUTC-2020