TMA Innovation Center

Data Science Level 1

Hung Nguyen

TIC Data Team Lead

Course objective

Theory

Deeply understand about Data Science:

- What is DS
- What DS [can] do
- How to be a DS

Practice

- Build an end-to-end predictive model

Course syllabus

Session 1

- Data Science introduction
- Data science project cycle
- Basic concepts of statistics
- Intro duction to Python

Session 2

- Data manipulation with Python - **Pandas**
- Data Visualization with Python - **Matplotlib**

Session 3

- Machine Learning with **Scikit-Learn** – build model
- Model evaluation
- Basic algorithms
- PoCs planning

Session 4

- PoCs review:

➢ Prediction with linear regression

➢ Classification with logistic regression

TMA Innovation Center

Data Science Level 1

-- Session 1--

Introduction

Hung Nguyen

TIC Data Team Lead

Data
Science
0101
1010
0010

What is Data Science?

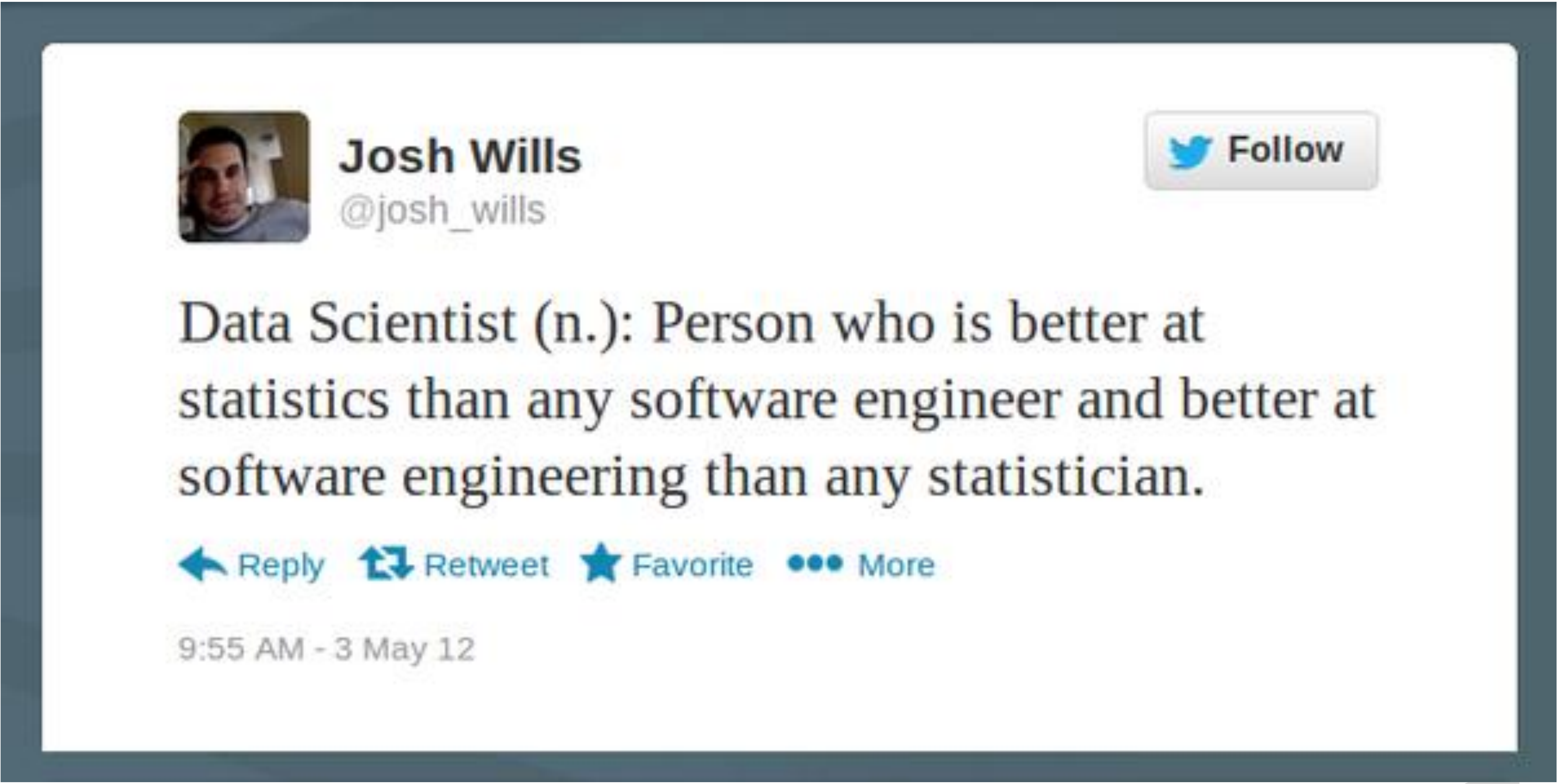
- Data science is a **multi-disciplinary** field
- Uses scientific methods, processes, algorithms and systems
- Extracts **knowledge** and **insights** from *data in various forms*, both structured and unstructured

HBR.ORG
Harvard Business Review
OCTOBER 2012
The Big Idea
The True Measures Of Success
Michael J. Mauboussin
International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz
Leadership
What Ever Happened To Accountability?
Thomas E. Ricks
GETTING CONTROL OF
BIG DATA
How vast new streams of information are changing the art of management

Harvard Business Review

DATA SCIENTIST

The Sexiest Job of the 21st Century



Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

9:55 AM - 3 May 12

What Data Scientist do?

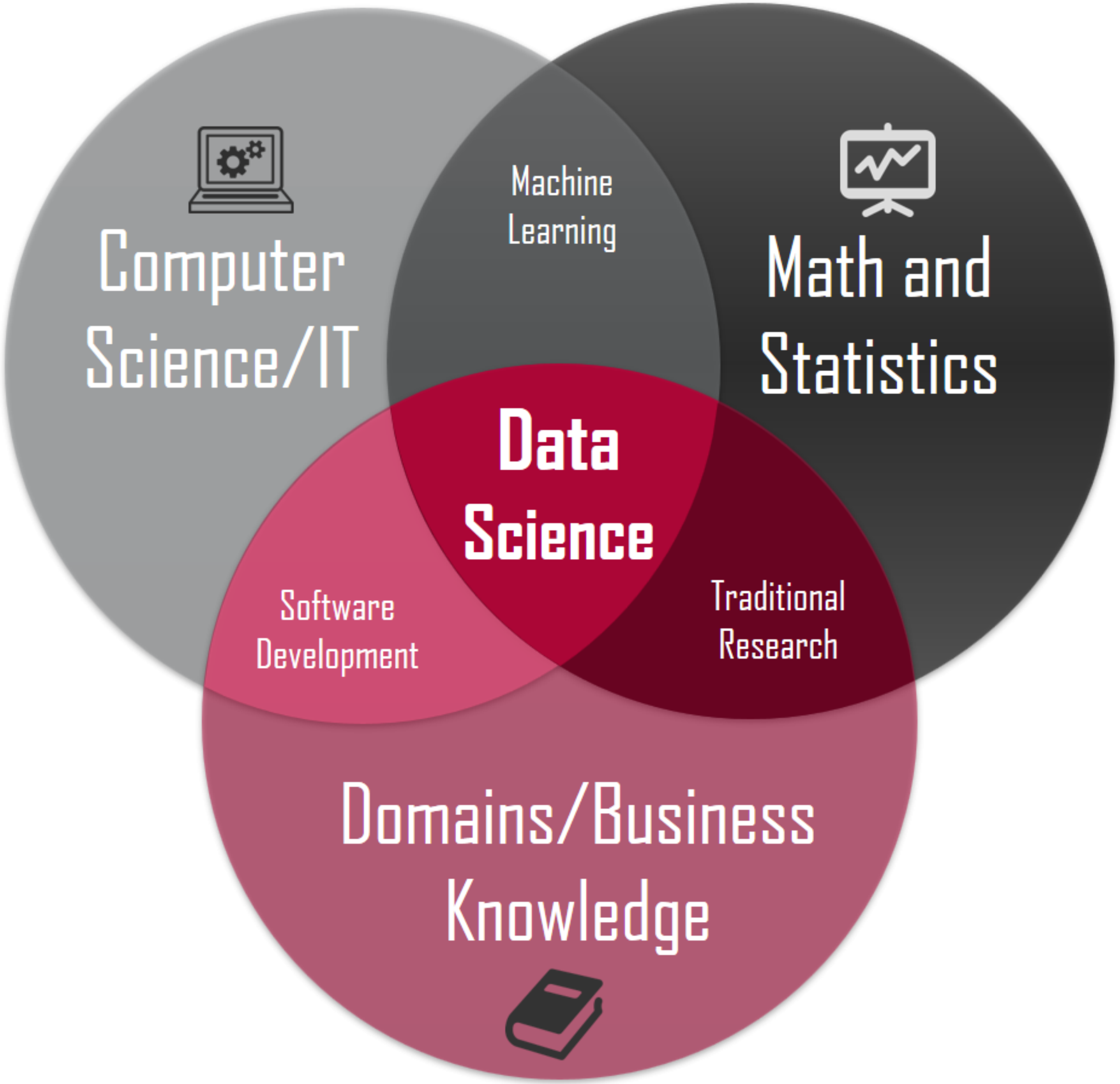
Product and Research

- Find **better** techniques
- Make sure ML system was **“smart” enough**
- **Help** the **engineering** team
- Work with product managers to **incorporate new improvements**
- Develop **new** products or features

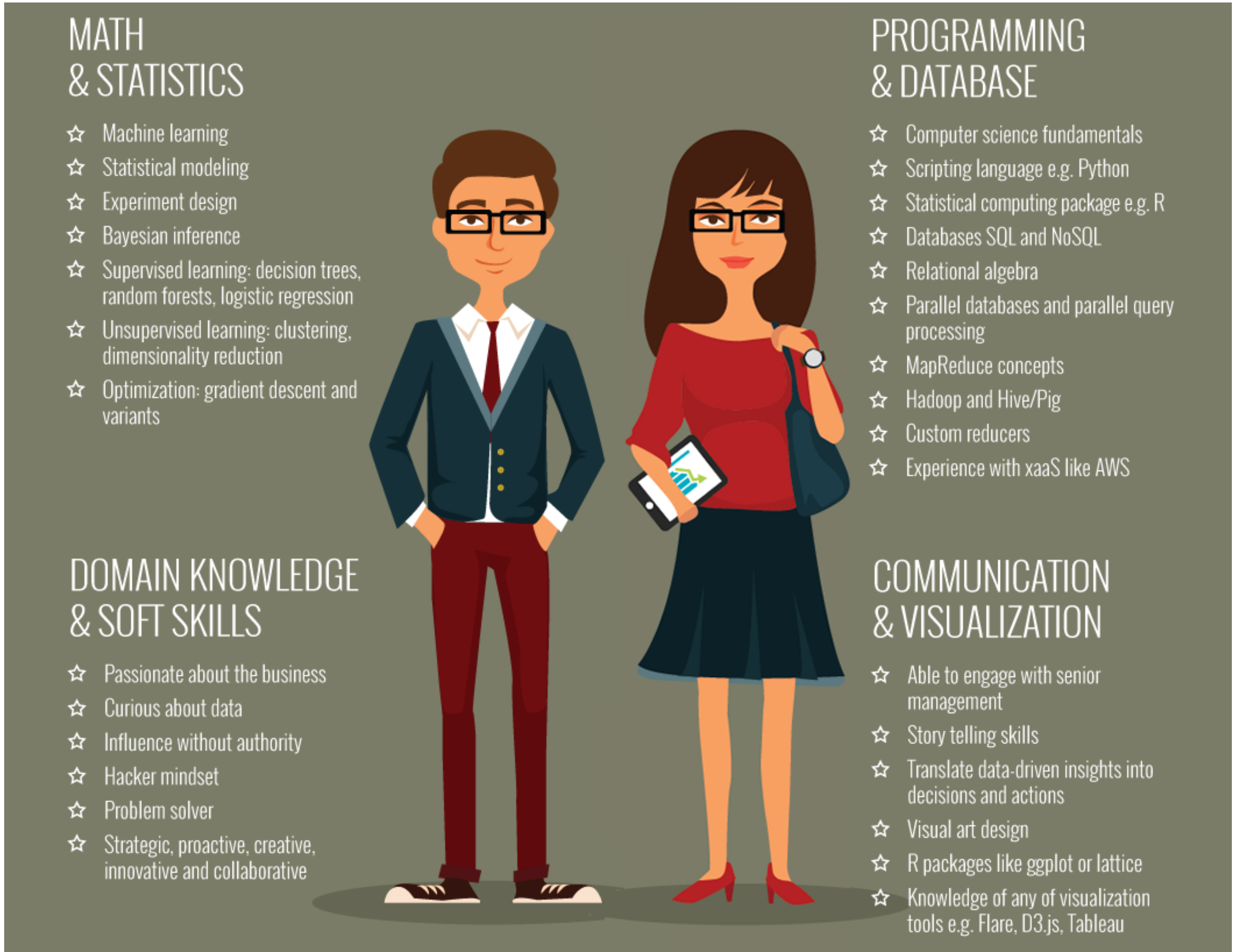
Sales and Marketing

- **Explain** technology aspects to **potential clients**
- Work **with data set** from a potential client
- **Share** knowledge
- Create **visualization**

Data Scientist skill set



Data Scientist skill set



MATH
& STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING
& DATABASE

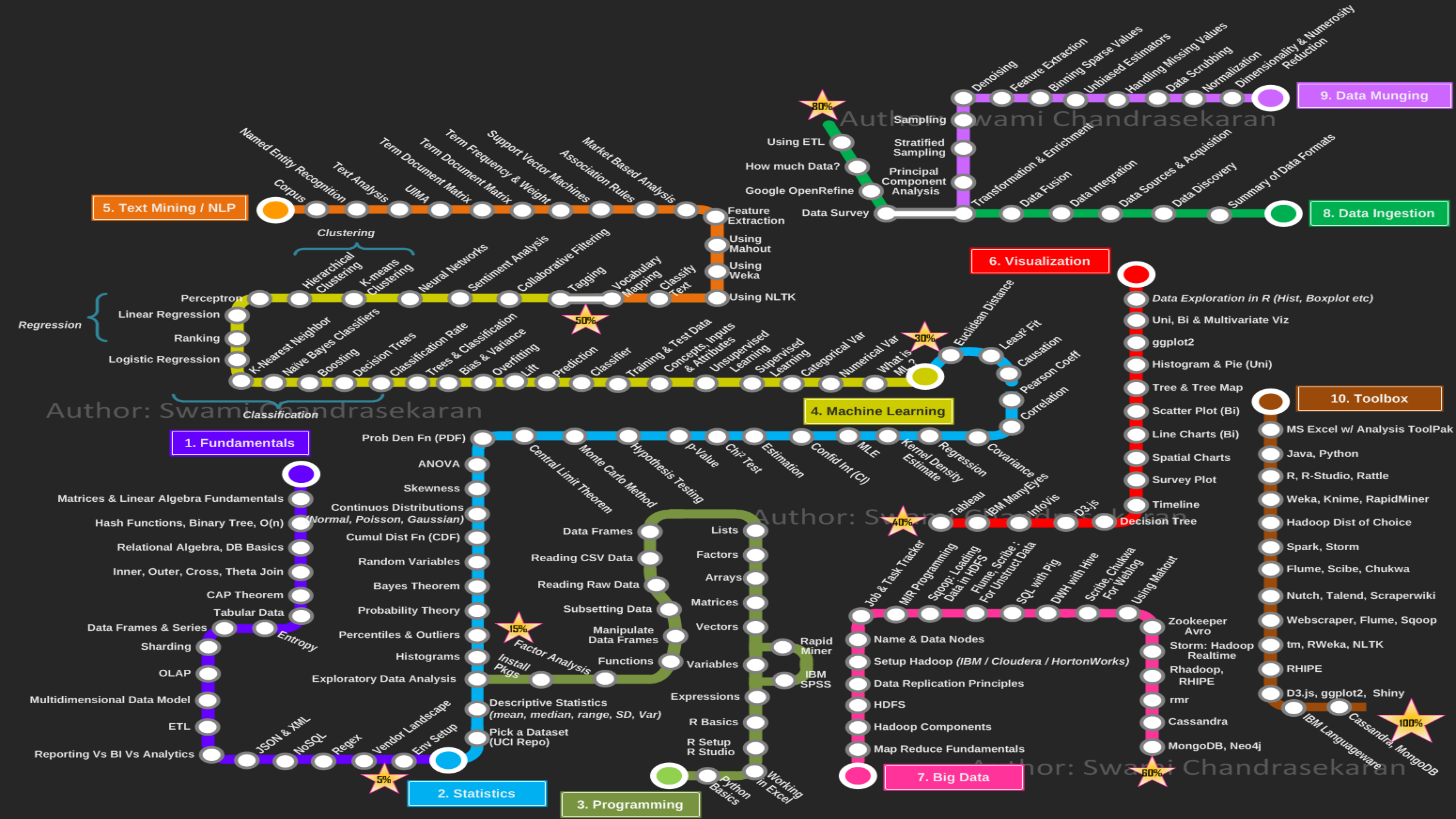
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE
& SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION
& VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Entropy
Data Frames & Series
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Pick a Dataset (UCI Repo)
Descriptive Statistics (mean, median, range, SD, Var)
Exploratory Data Analysis
Histograms
Percentiles & Outliers
Probability Theory
Bayes Theorem
Random Variables
Cumul Dist Fn (CDF)
Continuos Distributions (Normal, Poisson, Gaussian)
Skewness
ANOVA
Prob Den Fn (PDF)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
Pearson Coeff
Causation
Least² Fit
Euclidean Distance
3. Programming
Python Basics
Working in Excel
R Setup R Studio
R Basics
Expressions
Variables
IBM SPSS
Rapid Miner
Vectors
Matrices
Arrays
Factors
Lists
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Factor Analysis
Install Pkgs
15%
4. Machine Learning
What is ML?
30%
Numerical Var
Categorical Var
Supervised Learning
Unsupervised Learning
Concepts, Inputs & Attributes
Training & Test Data
Classifier
Prediction
Lift
Overfitting
Bias & Variance
Trees & Classification
Classification Rate
Decision Trees
Boosting
Naive Bayes Classifiers
K-Nearest Neighbor
Classification
Logistic Regression
Ranking
Linear Regression
Regression
Perceptron
Hierarchical Clustering
K-means Clustering
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
50%
Vocabulary Mapping
Classify Text
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
D3.js
InfoVis
IBM ManyEyes
Tableau
40%
7. Big Data
Map Reduce Fundamentals
Hadoop Components
HDFS
Data Replication Principles
Setup Hadoop (IBM / Cloudera / HortonWorks)
Name & Data Nodes
Job & Task Tracker
M/R Programming
Sqoop: Loading Data in HDFS
Flume, Scribe: For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Data Survey
Google OpenRefine
How much Data?
Using ETL
80%
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
Principal Component Analysis
Stratified Sampling
Sampling
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran

Data Science Level 1

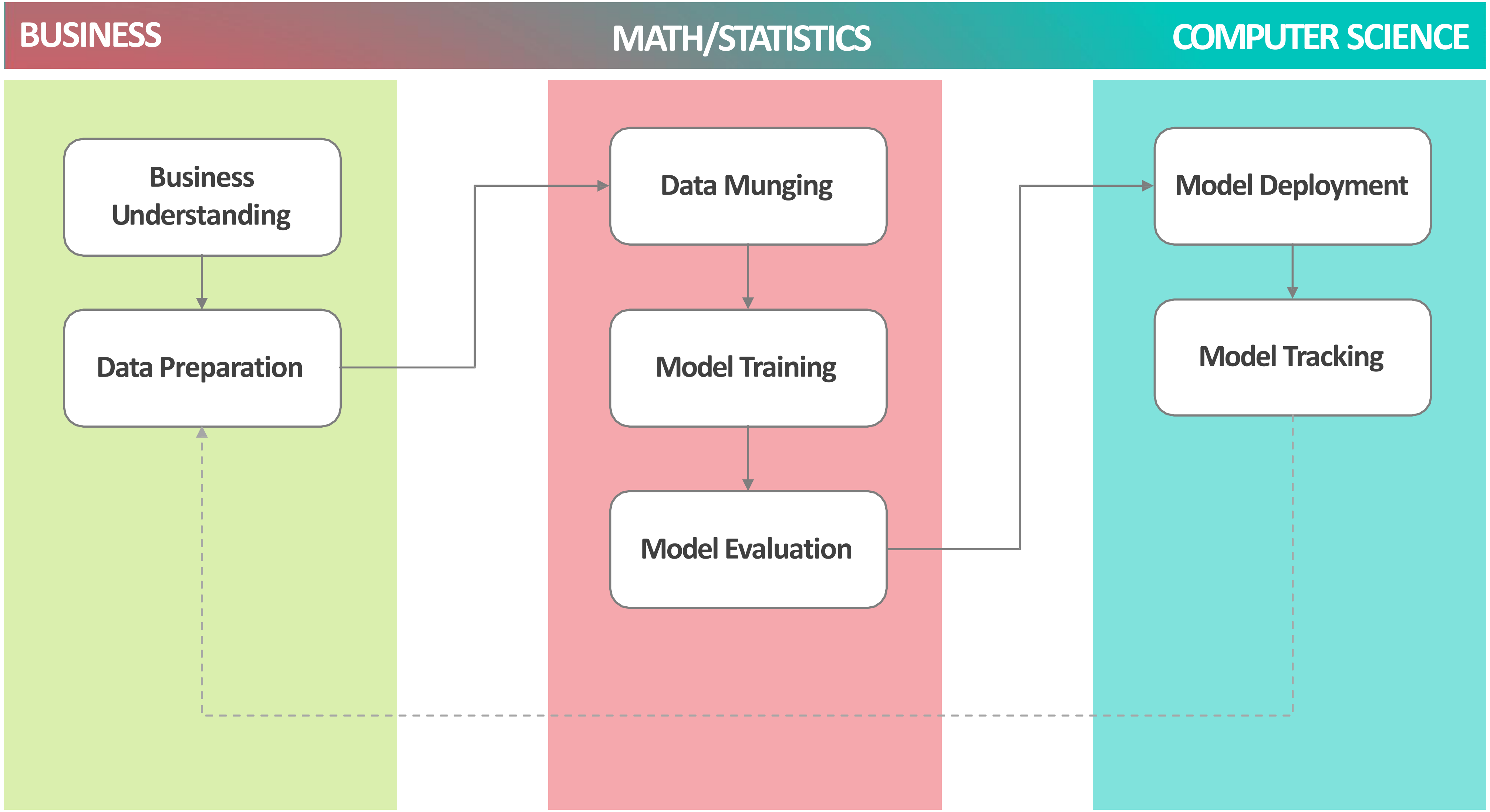
-- Session 1--

Data science project cycle

Hung Nguyen

TIC Data Team Lead

The Data Science Process



Business understanding

Determine

What does the client want to achieve?

- ✓ Reduce attrition
- ✓ Customized targeting
- ✓ Plan future media spend
- ✓ Prevent fraud
- ✓ Recommend Products

Understand

- **Understand** success criteria.
- **List** assumptions, constraints, and important factors.
- **Identify** secondary or competing objectives.
- **Study** existing solutions (if any).

Map

- **Business Objective** → **Technical Objective**
  - ✓ State the project objective in **technical terms**.
  - ✓ How data science project will help
  - ✓ **Successful** scenarios.

Data Preparation

IDENTIFY

- Data **sources, formats**
- Entity **Relationship** Diagram
- Identify **relevant** data
- Record **unavailable** data.
- How long

COLLECT

- Access or acquire all relevant data in **a central location**
- Quality control **checks** and **tests**

ASSESS

- Get **familiar** with the data.
- Study **seasonality**.
- Detect **mistakes**.
- Check **assumptions**.
- Review **distributions**.

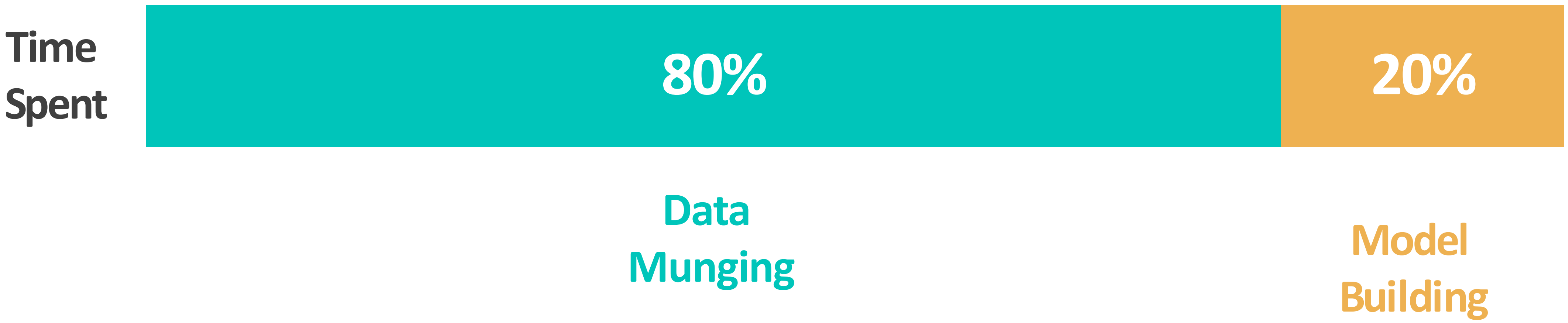
VECTORIZE

- Create the Analysis **Dataset**

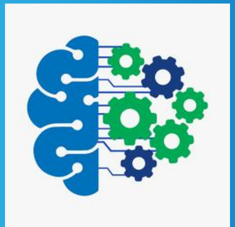
Data munging (preprocessing)

- **Descriptive** statistics
- **Correlation** analysis
- Impute **missing** values
- Trim **extreme** values
- Process **categorical** attributes

- **Transformations** (square, log, etc.)
- Multicollinearity: **reduce** redundancy
- Create **additional feature**
- **Interactions**
- **Normalization** (scaling)

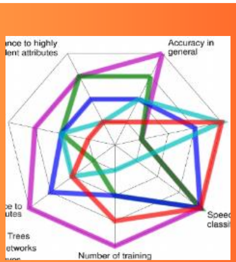

Time Spent
80%
20%
Data Munging
Model Building

Model training

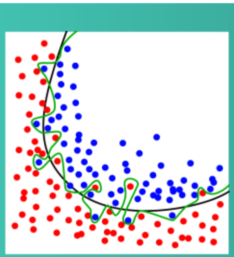


Try **more than one** machine learning technique.

Fine-tune parameters.



Assess model **performance**.



Avoid **Over-fitting**.

Model Evaluation

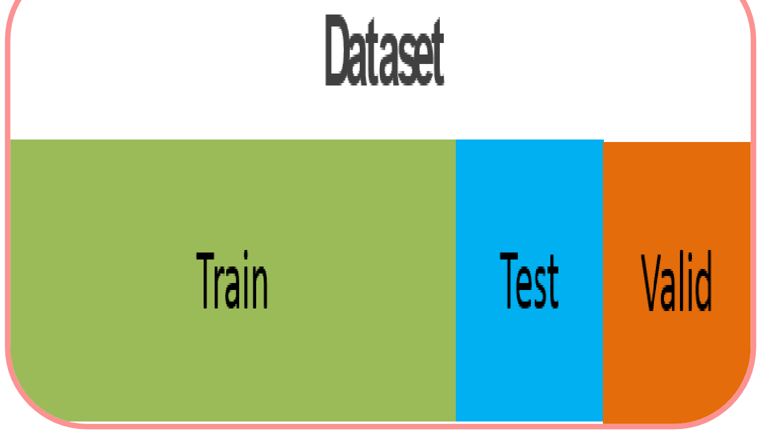
MODEL SELECTION

- Law of Parsimony:

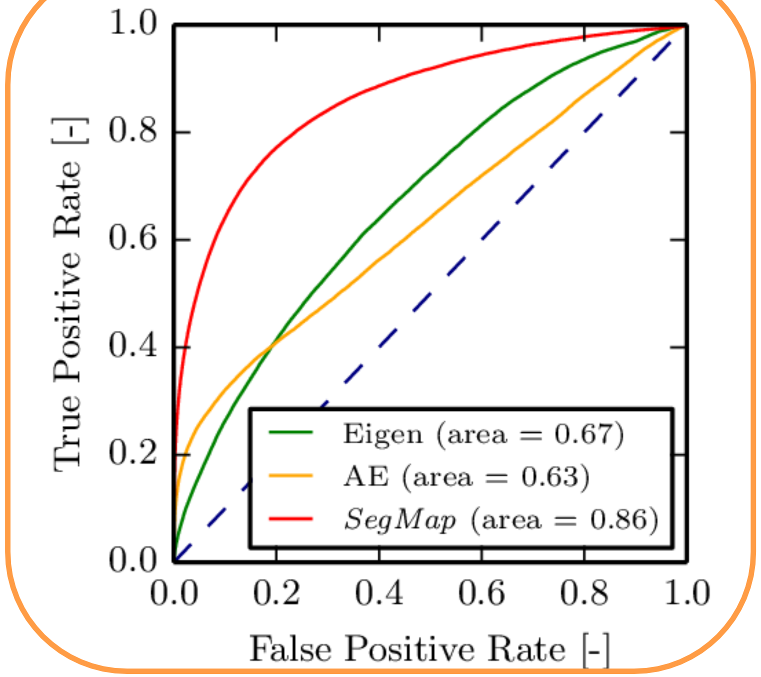
simple is better

- Execution time
- Deployment complexity

ASSESSMENT



PRESENTATION



Model Deployment

- Model production cycle
- Scoring code, or publish model as a web service
- Model Documentation (Technical Specifications)
- Reproducibility
- Model Persistence vs. Model Transience

Model Tracking

MONITOR

- Model performance over time
- Predictor distribution

MAINTAIN

- Model maintenance plan
- Adding new data sources
- Version control

TEST

- Experimental Design (A/B tests, Fractional Factorial)

Data Science Process: Recap

Business Understanding	Data Preparation	Data Munging	Model Training	Model Evaluation	Model Deployment	Model Tracking
Determine	Identify	Impute	Train	Evaluate	Deploy	Monitor
Understand	Collect	Transform	Assess	Peer Review	Document	Maintain
Map	Assess	Reduce	Select	Present		Test
	Vectorize					
DISCUSS	COLLATE	WRANGLE	PERFORM	COMMUNICATE	EXECUTE	TRACK

Data Science Level 1

-- Session 1--

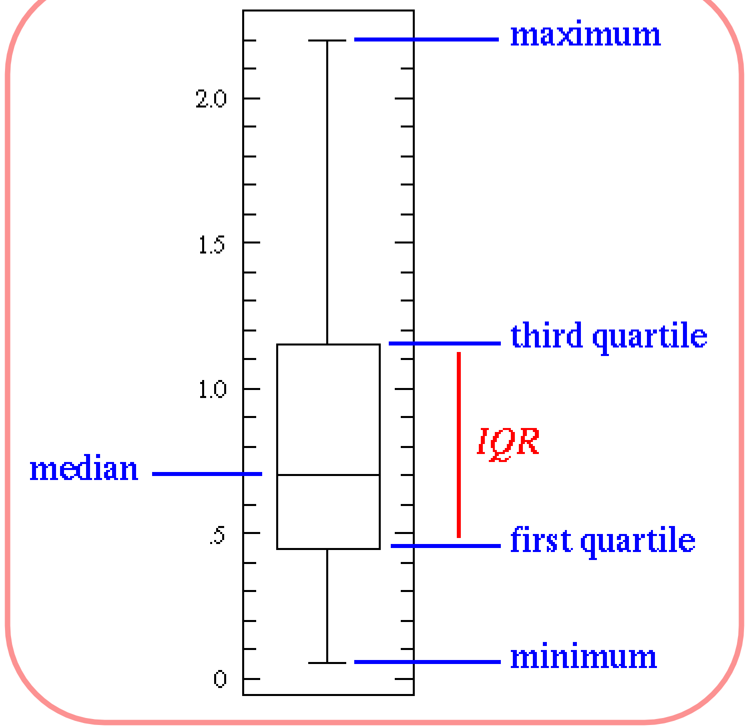
Basic statistical concepts

Hung Nguyen

TIC Data Team Lead

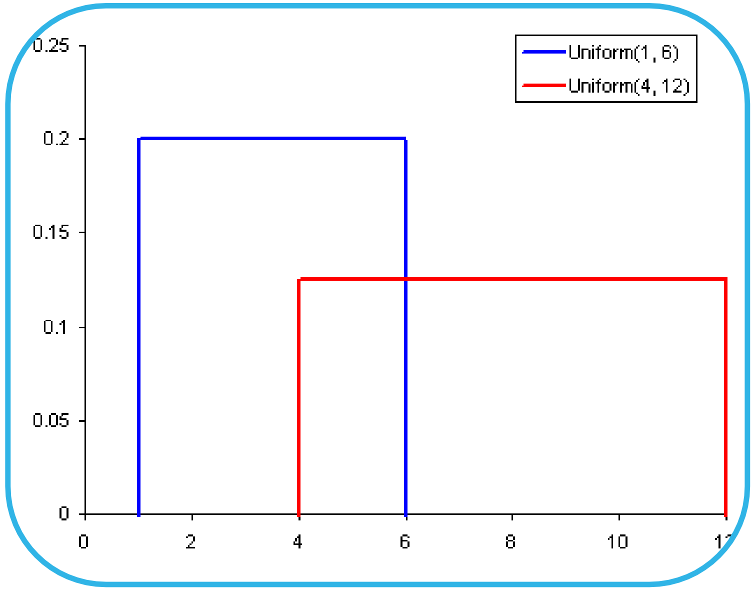
Statistical Features

- min, max, mean, median, mode, standard deviation, first & third quartiles
- the first stats technique to apply when exploring a dataset
- easy to understand and implement in code

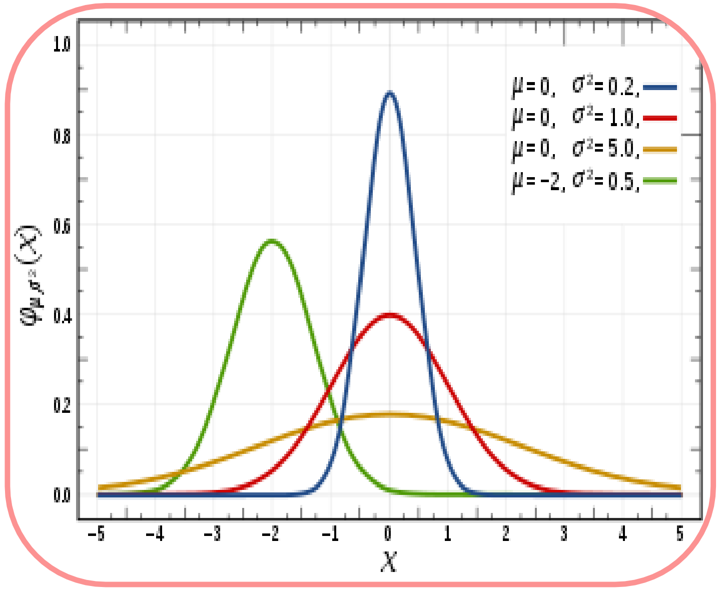


Probability Distributions

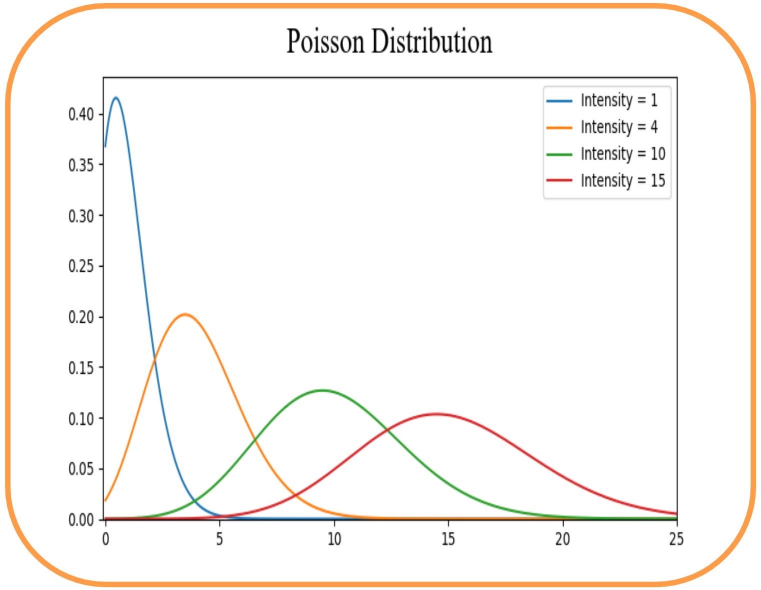
Uniform



Gaussian

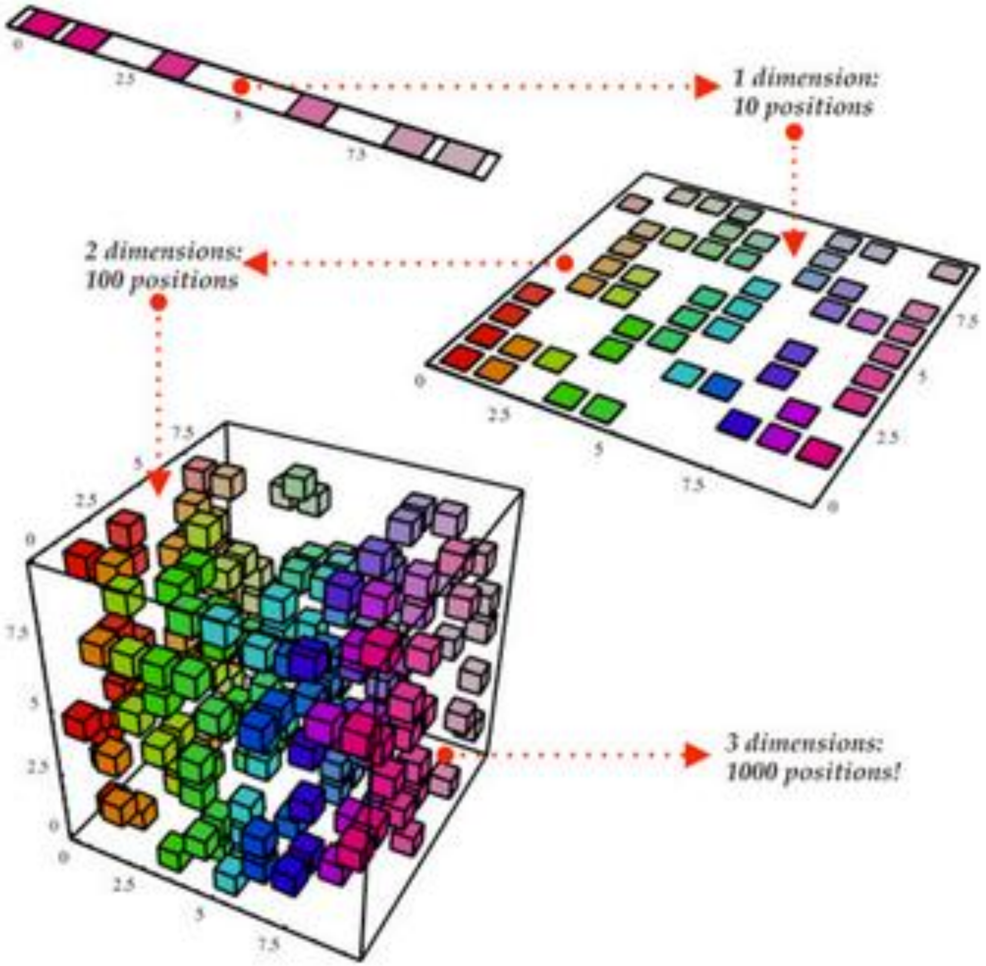


Poisson



Probability - percent chance that some event will occur

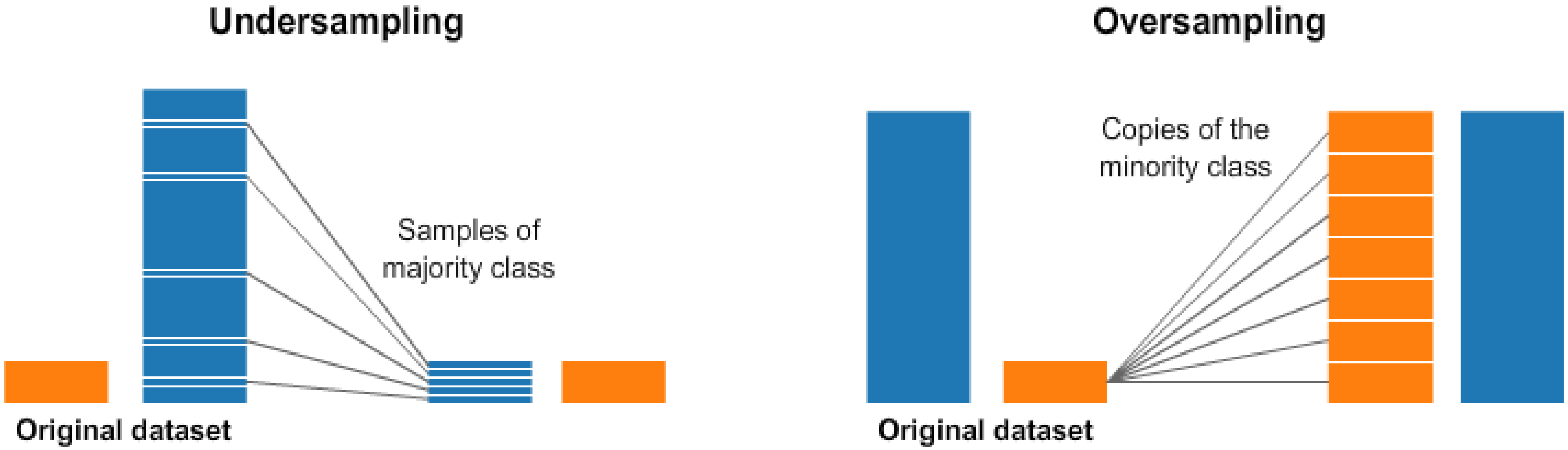
Dimensionality Reduction



Techniques:

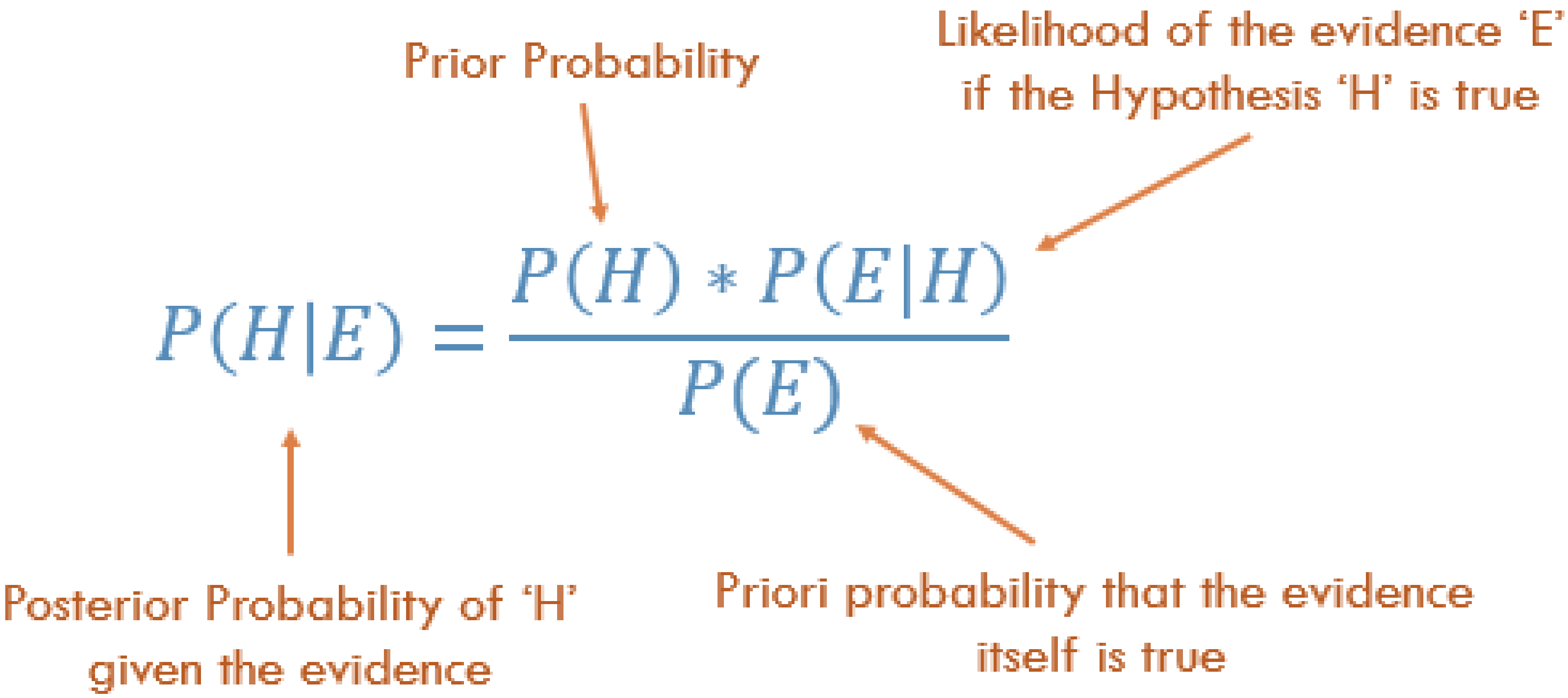
- Feature pruning
- Principal component analysis (PCA)

Over and Under Sampling



usually use in **imbalanced classification** problem

Bayesian Statistics



Prior Probability

Likelihood of the evidence 'E' if the Hypothesis 'H' is true

$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Posterior Probability of 'H' given the evidence

Priori probability that the evidence itself is true

- Example 1 : the probability of a certain medical test being positive is 90%, if a patient has disease D. 1% of the population have the disease, and the test records a false positive 5% of the time. If you receive a positive test, what is your probability of having D?

- P(+|D)=0.9, P(D)=0.01, P(+|no D)=0.05, we want P(D|+)

- $P(D|+) = \frac{P(+|D)\;P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\mathrm{no}D)P(\mathrm{no}D)}$

- Substituting in the numbers : P(D|+) = 0.15

TMA Innovation Center

Data Science Level 1

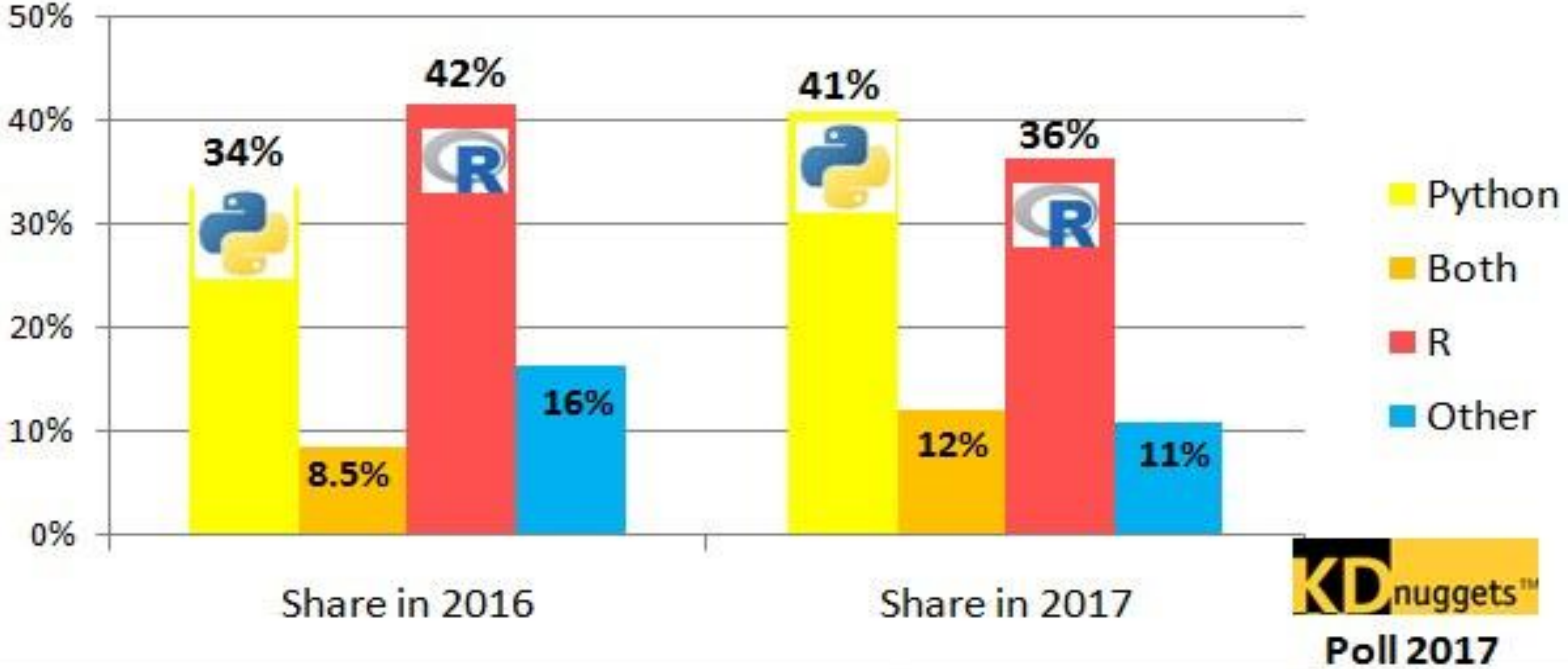
-- Session 1--

Introduction to Python

Hung Nguyen

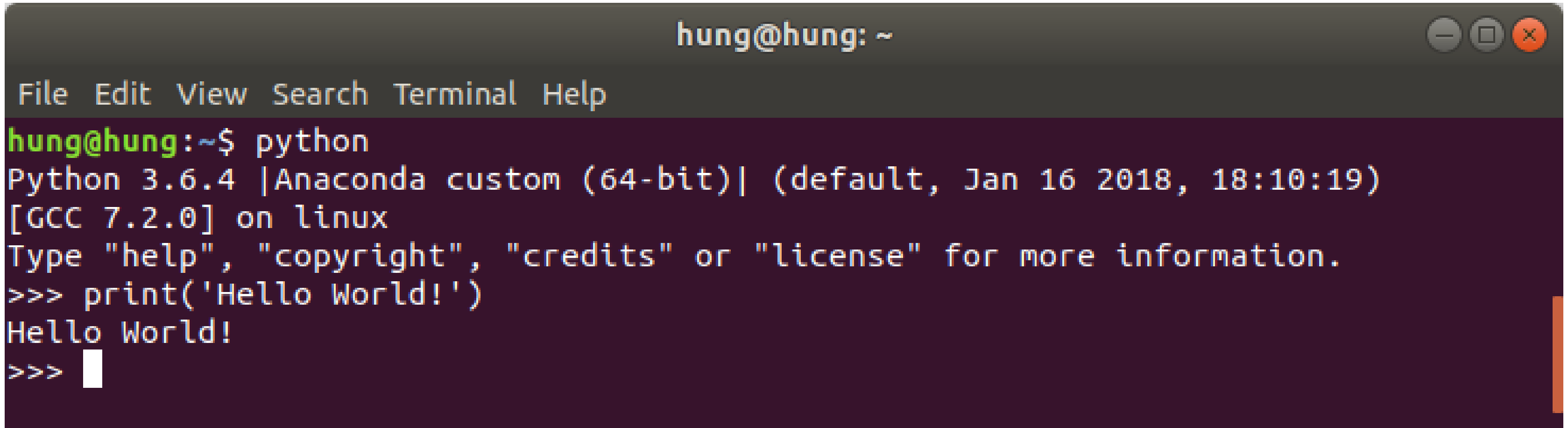
TIC Data Team Lead

Python, R, Both, or Other platforms for Analytics, Data Science, Machine Learning



Why Python for Data Science?

- Python’s inherent readability and simplicity make it relatively **easy to pick up**
- Large **amount** of dedicated analytical **libraries** and **open-source communities**
- **Millions** of users who are happy to **offer advice** or suggestions
- Supports object-oriented programming, structured programming, and functional programming patterns



```
hung@hung:~$ python
Python 3.6.4 |Anaconda custom (64-bit)| (default, Jan 16 2018, 18:10:19)
[GCC 7.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> print('Hello World!')
Hello World!
>>> 
```

Python basic libraries for Data Science

```
Install:    $ pip install <library_name>
Use:        >>> import    <library_name>
```

SciPy

NumPy

matplotlib

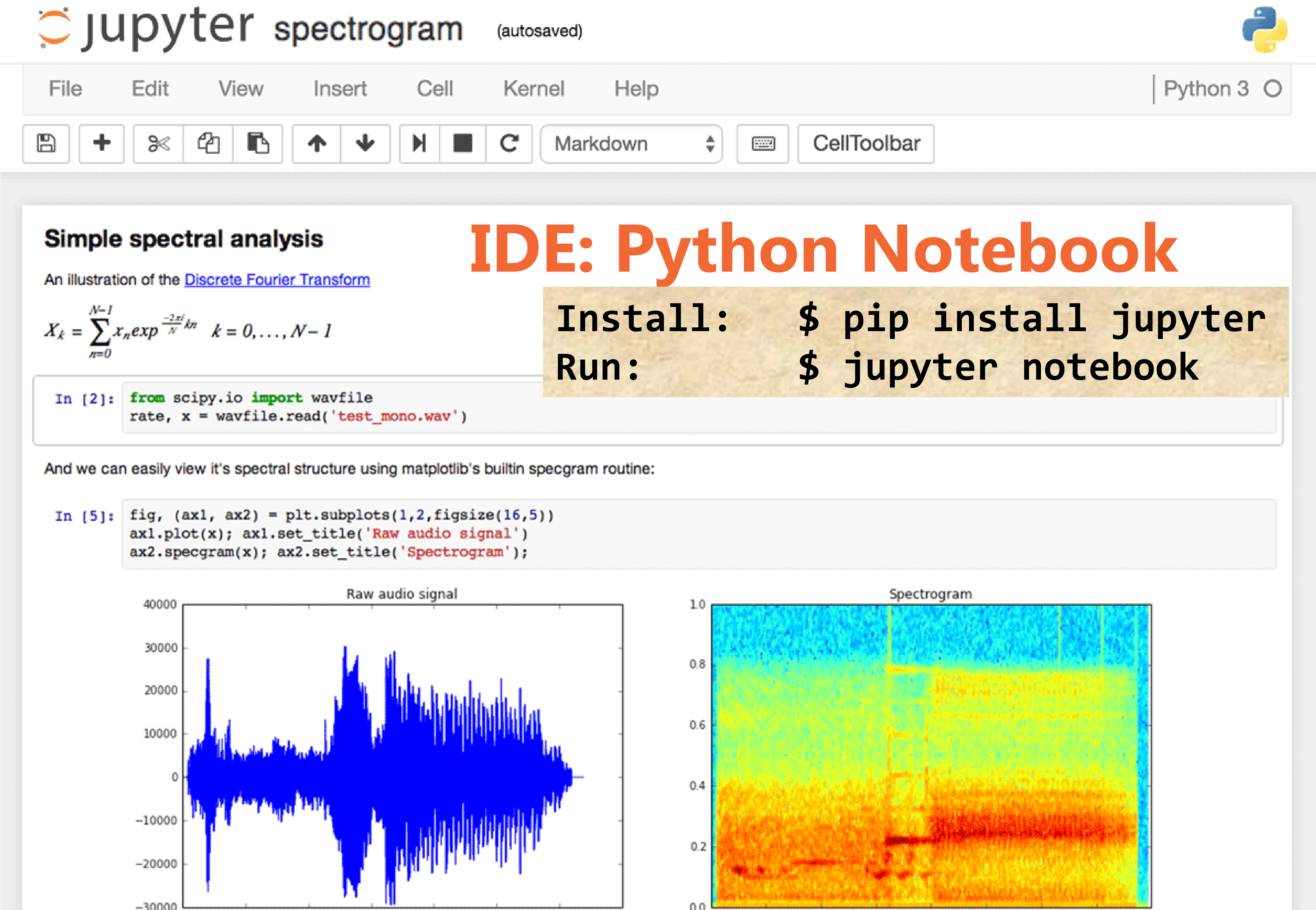
IP[y]: IPython
Interactive Computing

pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

scikit learn

IDE: Python Notebook

```
Install:    $ pip install jupyter
Run:        $ jupyter notebook
```



Simple spectral analysis

An illustration of the Discrete Fourier Transform

$$X_k = \sum_{n=0}^{N-1} x_n exp^{\frac{-2\pi i}{N}kn} \quad k = 0,\ldots,N-1$$

```
In [2]: from scipy.io import wavfile
        rate, x = wavfile.read('test_mono.wav')
```

And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

```
In [5]: fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```

Libraries/APIs for Big Data & Deep Learning

THANK YOU!

Vietnam:	**84-2839-951-059**
North America:	**+1 844 224 4188**
Australia:	**+61 414 734 277**
Japan:	**+81 364 324 994**
Website:	**http://tma-innovation.center**
Email:	**innovation@tma.com.vn**