General Approach to Data Science

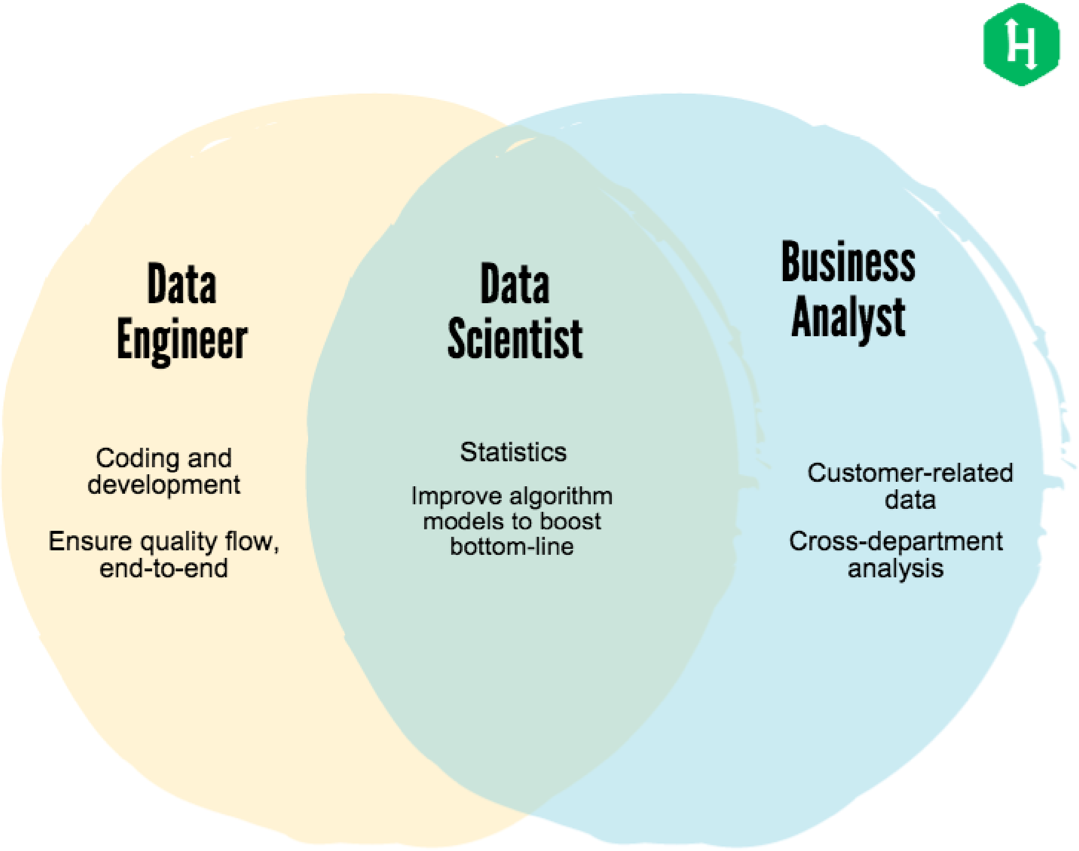
Bhuvan M S

msbhuvanbhuvi@gmail.com

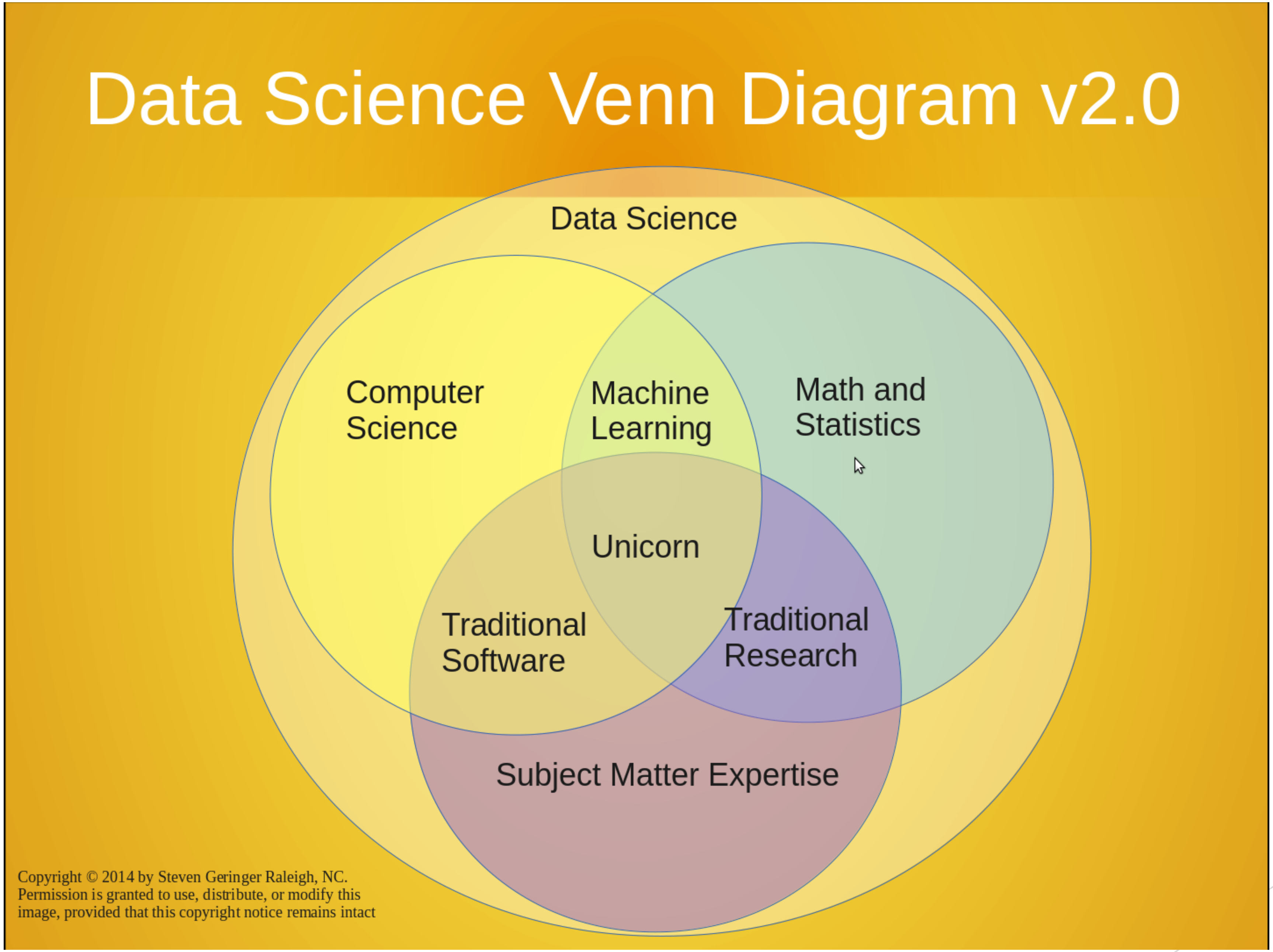
Contents

- Scope of Data Science
- Skills for a Data Scientist
- How to approach towards a data centric solution to a problem?
  - Analysis Pipeline
- Tools for implementing
- Deployment Framework
- Note on NLP specific approach

Scope of Data Science



Data Engineer
Coding and development
Ensure quality flow, end-to-end
Data Scientist
Statistics
Improve algorithm models to boost bottom-line
Business Analyst
Customer-related data
Cross-department analysis



Data Science Venn Diagram v2.0

Data Science

Computer Science

Machine Learning

Math and Statistics

Unicorn

Traditional Software

Traditional Research

Subject Matter Expertise

DATA SCIENTIST

'AS RARE AS UNICORNS'

Languages

R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents

- ✓ *Distributed computing*
- ✓ *Predictive modeling*
- ✓ *Story-telling and visualizing*
- ✓ *Math, Stats, Machine Learning*

Role

Cleans, massages and organizes (big) data

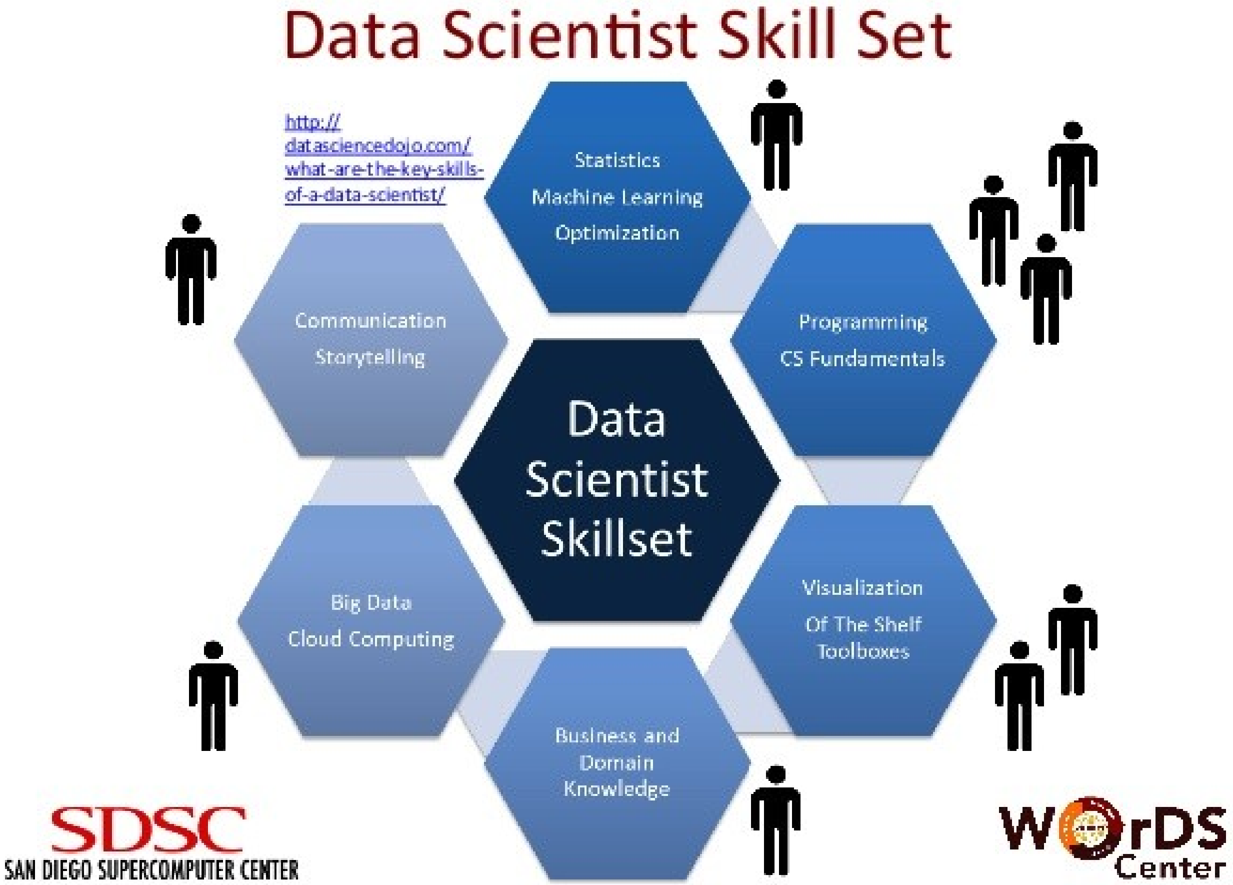
Mindset

Curious data wizard

HIRED BY

Google Microsoft Adobe

Skills of a Data Scientist



Data Scientist Skill Set

http://datasciencedojo.com/what-are-the-key-skills-of-a-data-scientist/

Data Scientist Skillset

Statistics
Machine Learning
Optimization

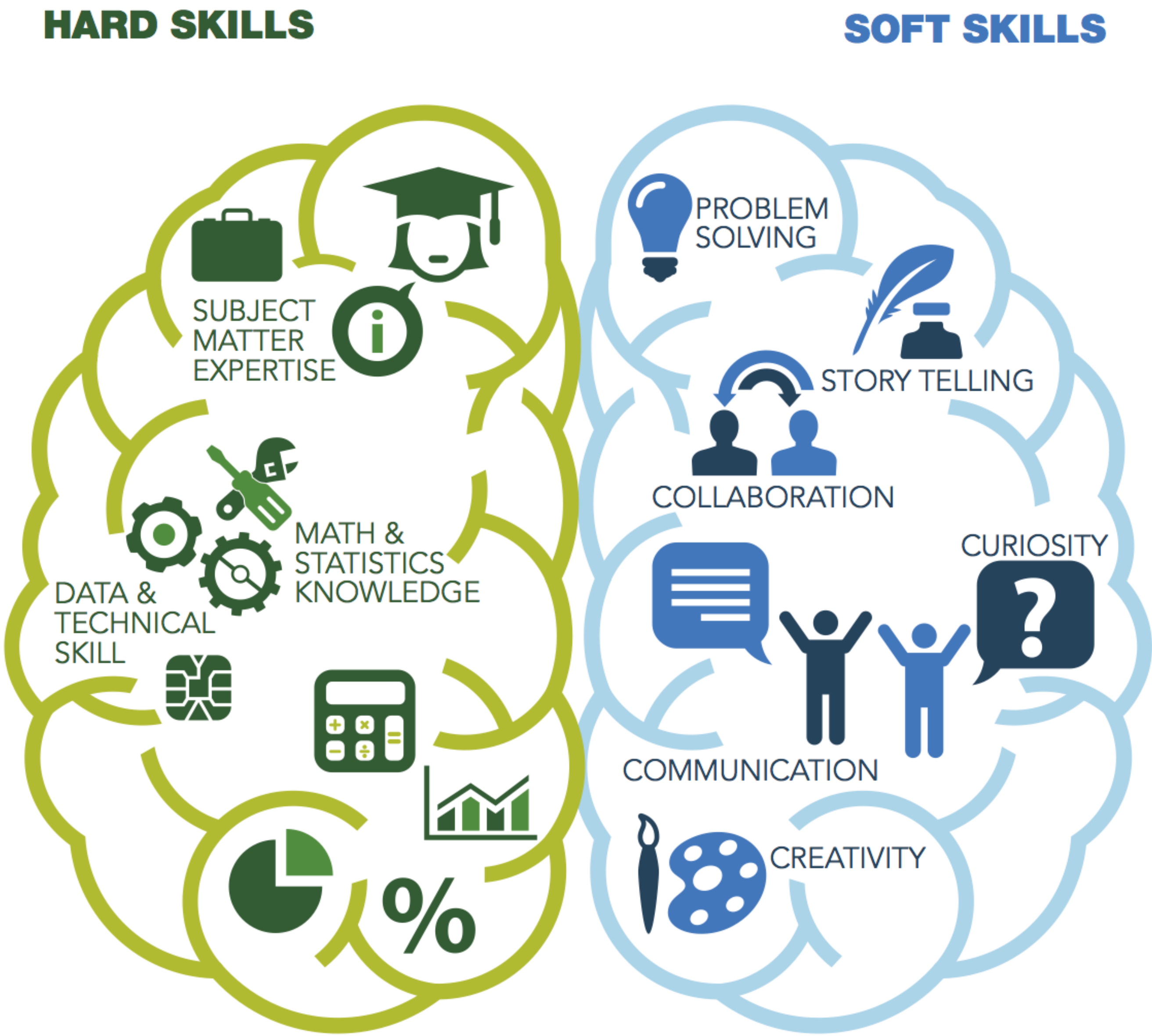
Programming
CS Fundamentals

Visualization
Of The Shelf Toolboxes

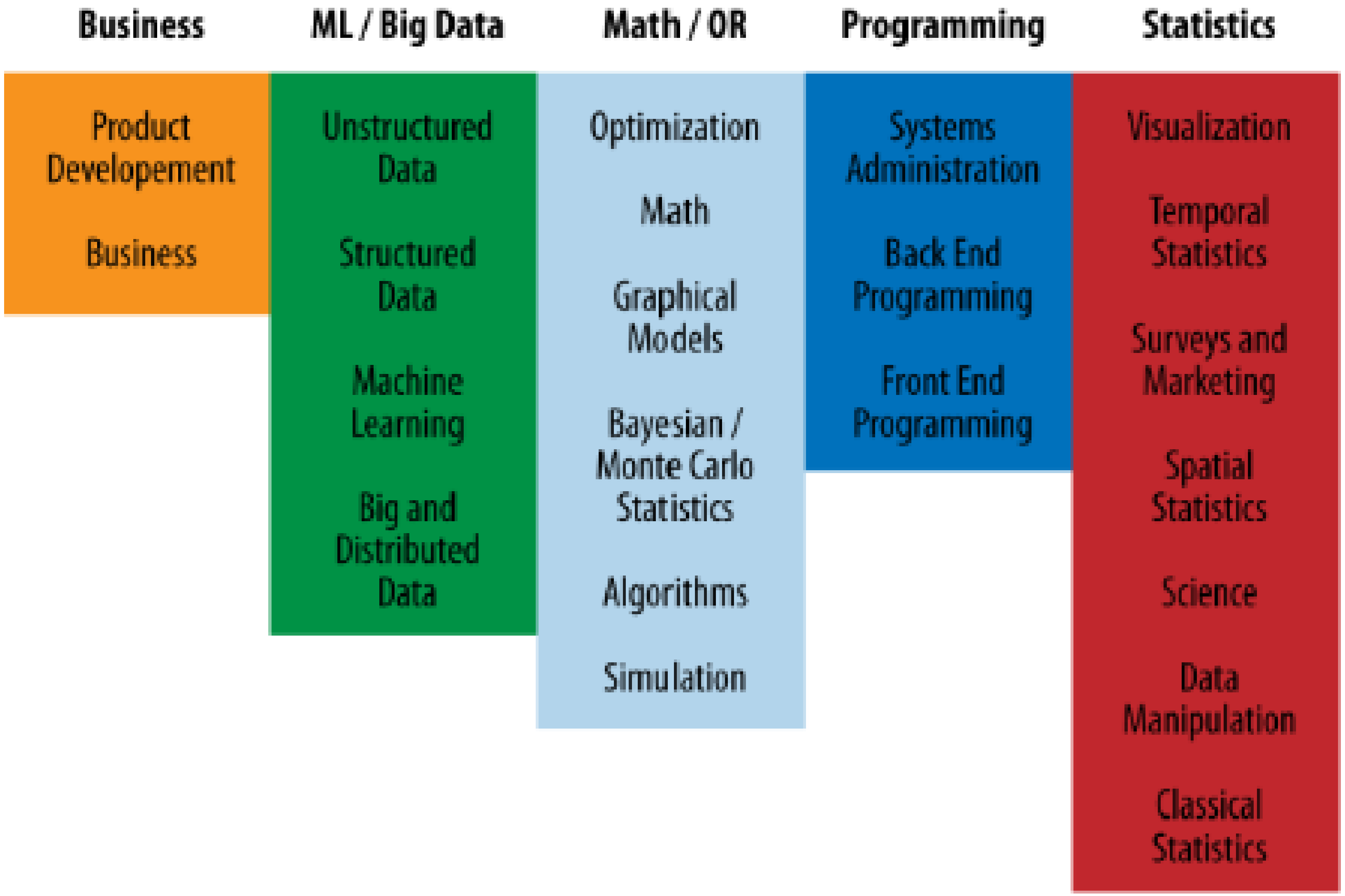
Business and Domain Knowledge

Big Data
Cloud Computing

Communication
Storytelling

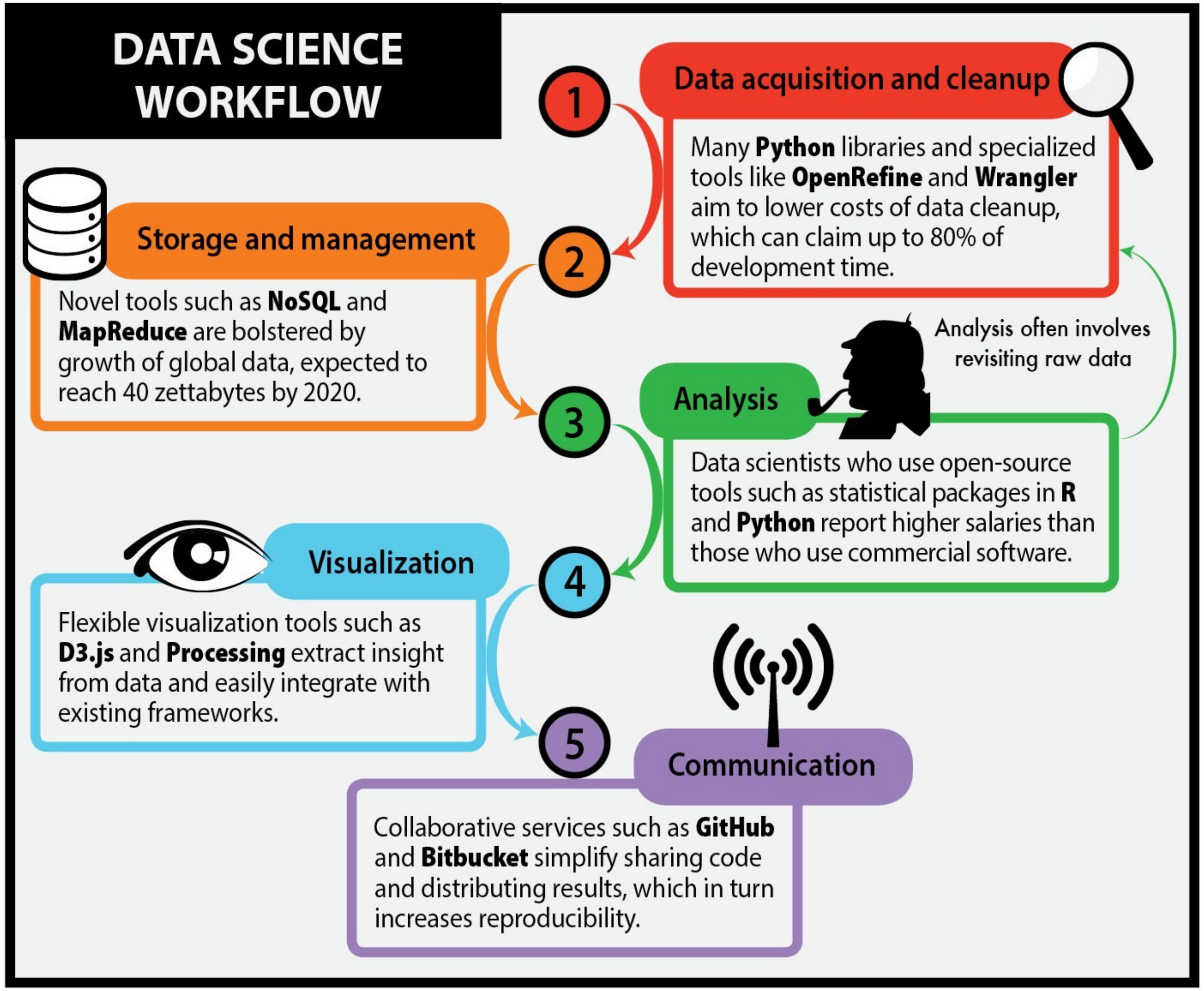


HARD SKILLS
SUBJECT MATTER EXPERTISE
MATH & STATISTICS KNOWLEDGE
DATA & TECHNICAL SKILL
%
SOFT SKILLS
PROBLEM SOLVING
STORY TELLING
COLLABORATION
CURIOSITY
?
COMMUNICATION
CREATIVITY



Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

The Process and Workflow



DATA SCIENCE WORKFLOW

1 Data acquisition and cleanup

Many **Python** libraries and specialized tools like **OpenRefine** and **Wrangler** aim to lower costs of data cleanup, which can claim up to 80% of development time.

2 Storage and management

Novel tools such as **NoSQL** and **MapReduce** are bolstered by growth of global data, expected to reach 40 zettabytes by 2020.

3 Analysis

Data scientists who use open-source tools such as statistical packages in **R** and **Python** report higher salaries than those who use commercial software.

Analysis often involves revisiting raw data

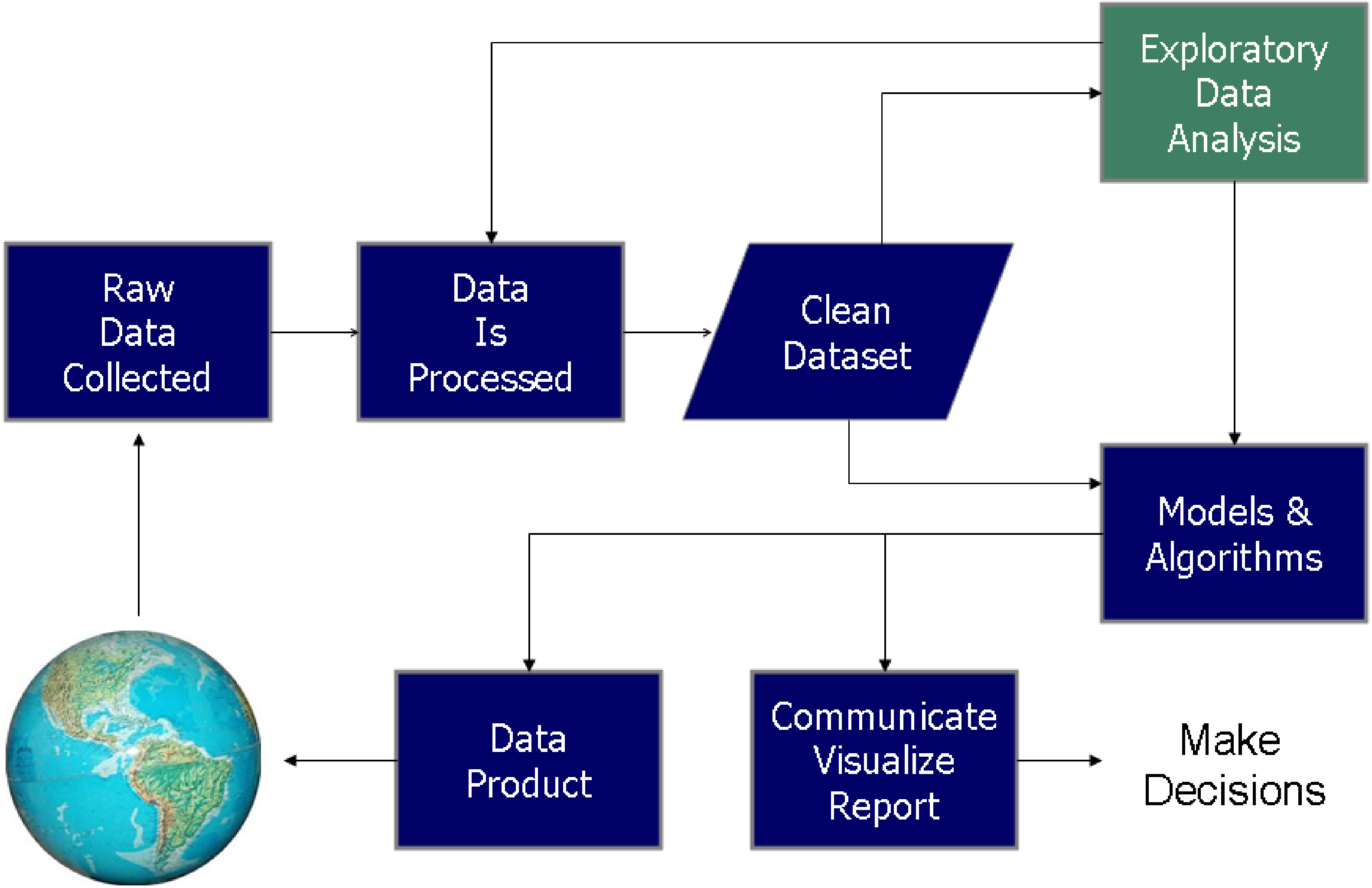
4 Visualization

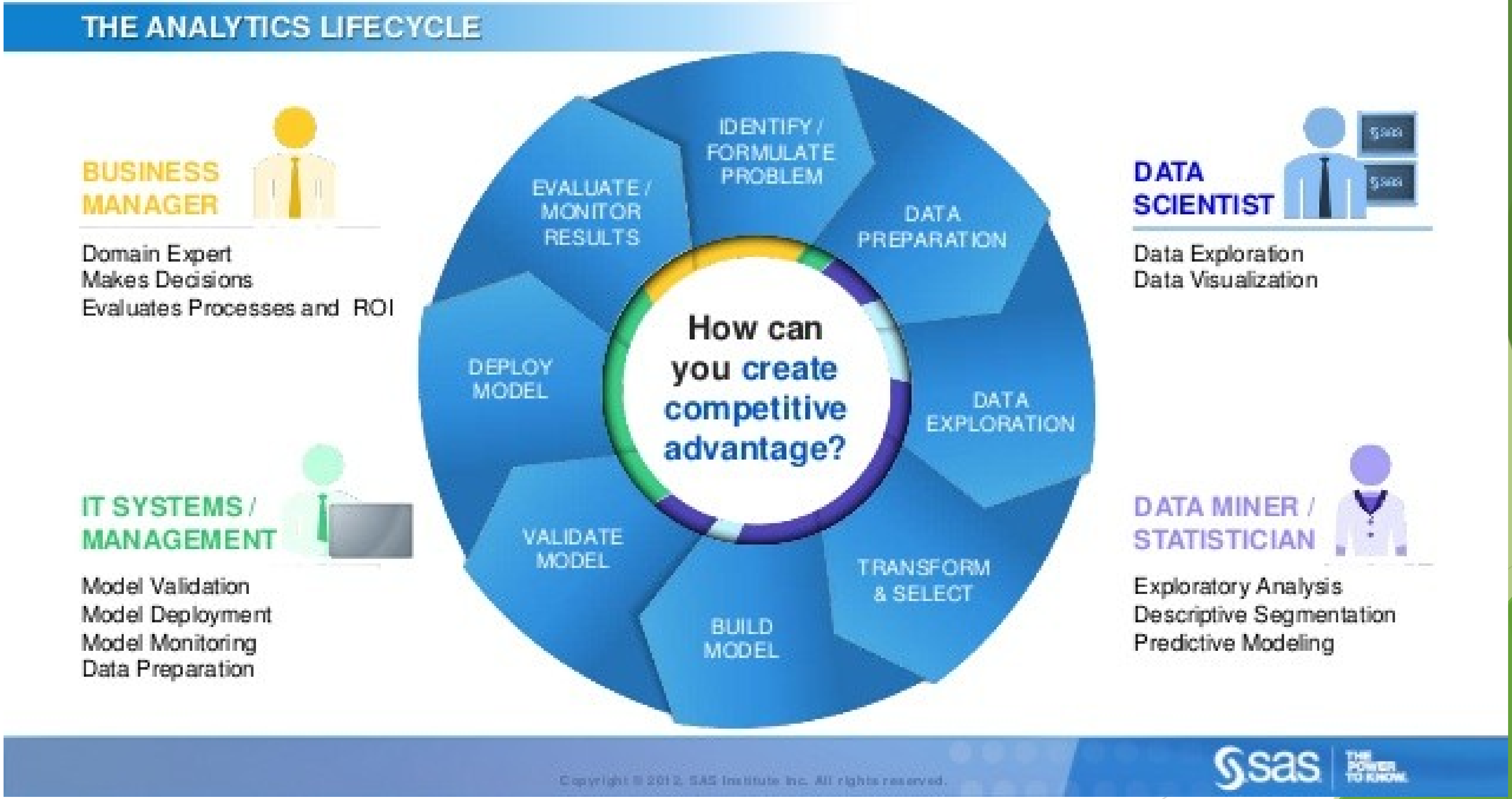
Flexible visualization tools such as **D3.js** and **Processing** extract insight from data and easily integrate with existing frameworks.

5 Communication

Collaborative services such as **GitHub** and **Bitbucket** simplify sharing code and distributing results, which in turn increases reproducibility.

Data Science Process





THE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation

How can you create competitive advantage?

IDENTIFY / FORMULATE PROBLEM

DATA PREPARATION

DATA EXPLORATION

TRANSFORM & SELECT

BUILD MODEL

VALIDATE MODEL

DEPLOY MODEL

EVALUATE / MONITOR RESULTS

DATA SCIENTIST

Data Exploration
Data Visualization

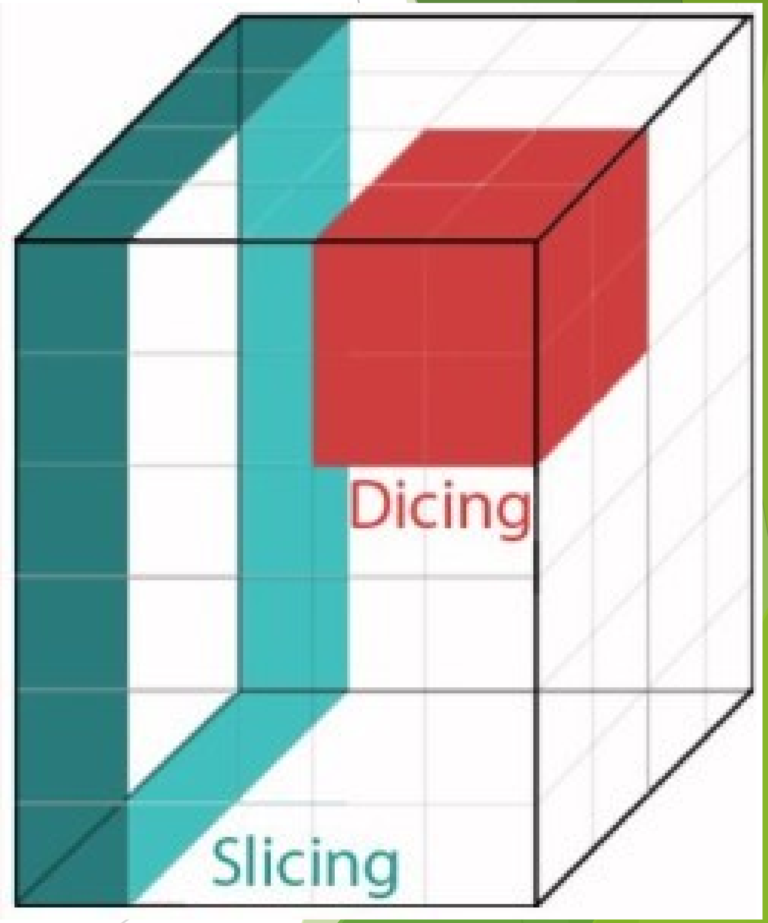
DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Exploratory Data Analysis

- Slicing and Dicing

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments.



Feature Engineering

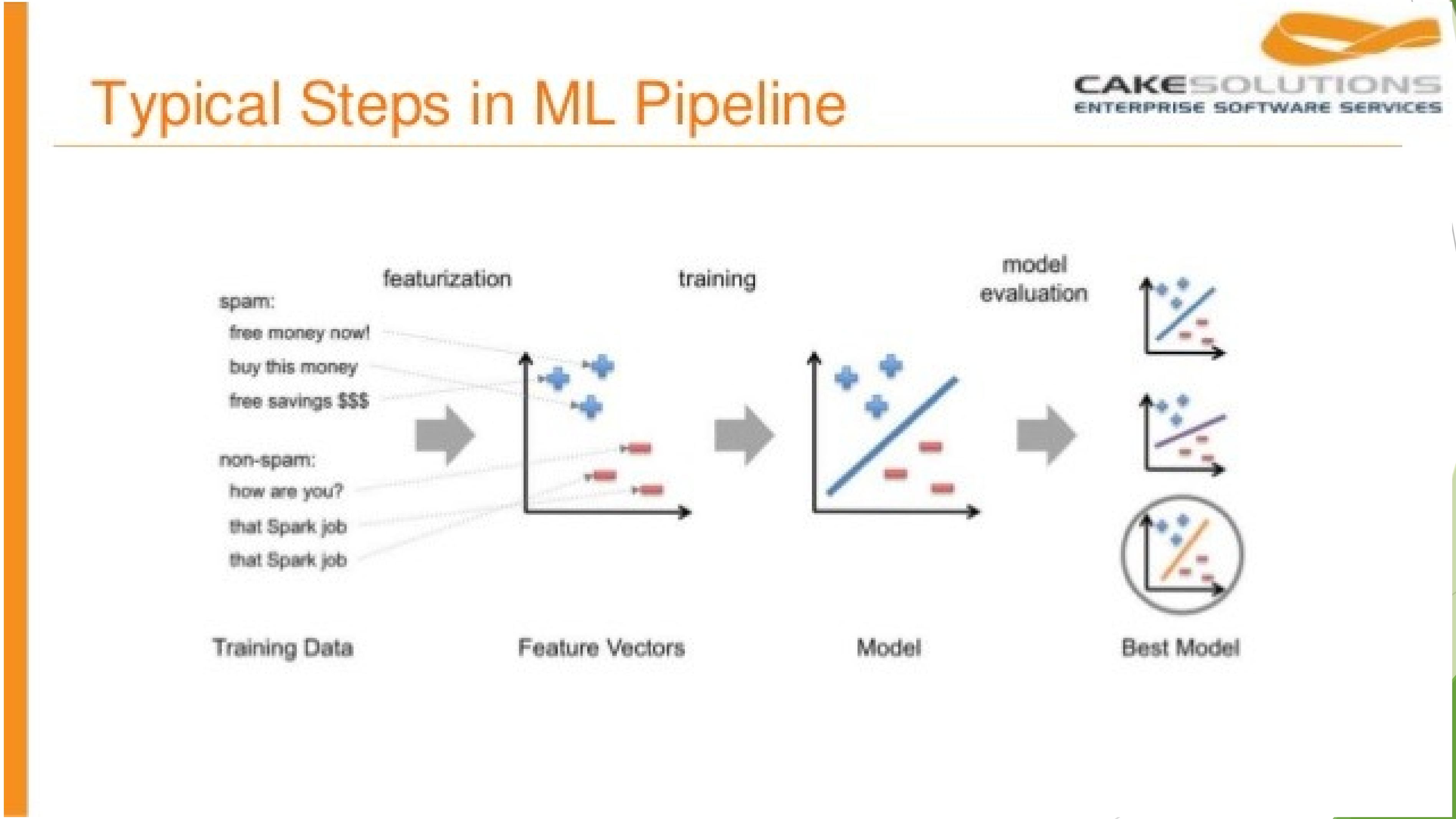
- Features: The dimensions of the data!
- Data Types: Binary, Numeric, Categorical, Ordinal
- Features Identification
- Features Extraction

Data Preprocessing

- Data Distribution and Data Scale observation
- Data Integration
- Data cleaning
  - Missing values
  - Noise
- Dimensionality reduction
  - PCA
  - Correlation Analysis
- Data Transformation: Normalization
  - Data Type specific

Machine Learning

ML pipeline



Typical Steps in ML Pipeline

ML Techniques

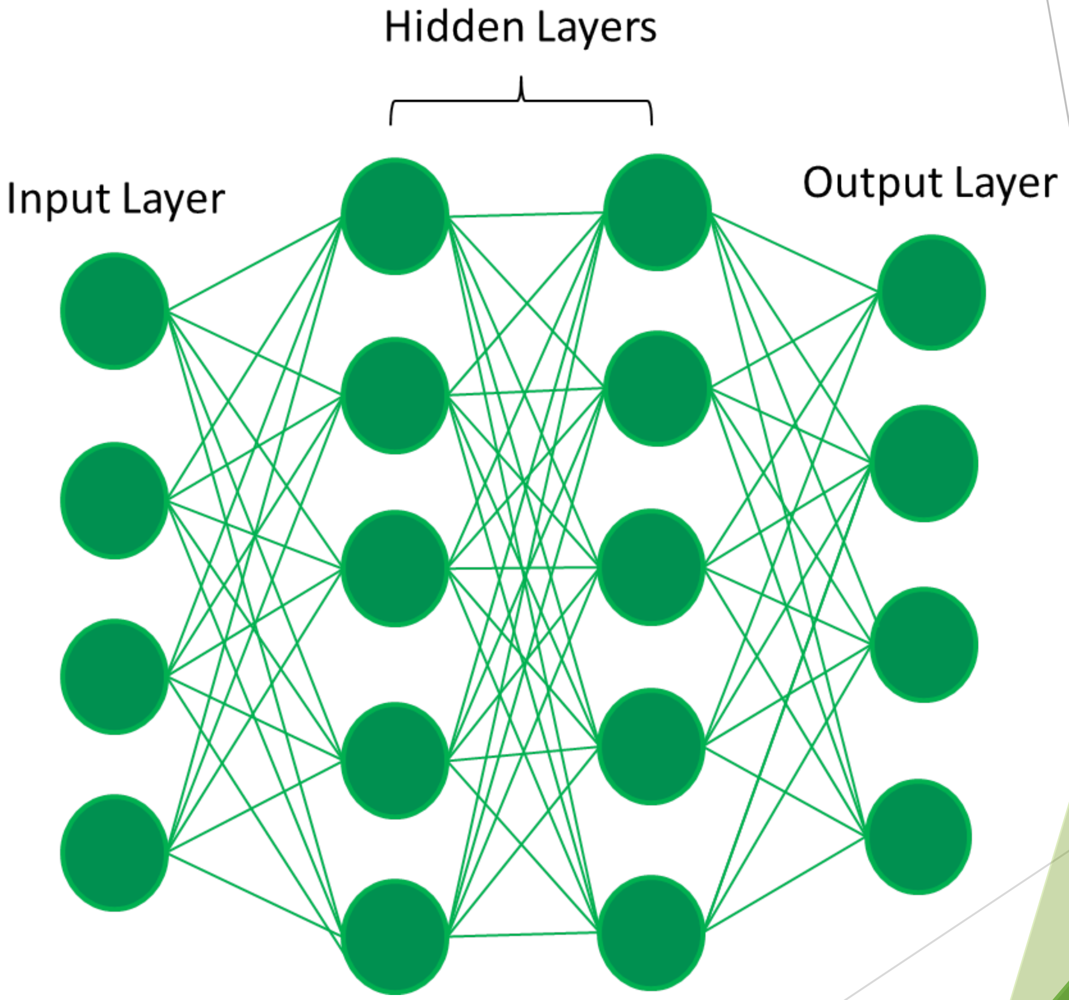
- Supervised:
  - Classification
- Unsupervised
  - Rule Based Classification
  - Clustering
- Association Rule Mining

Feature Importance Mining

- Feature Weights
- Ablation Study
- Random Forests

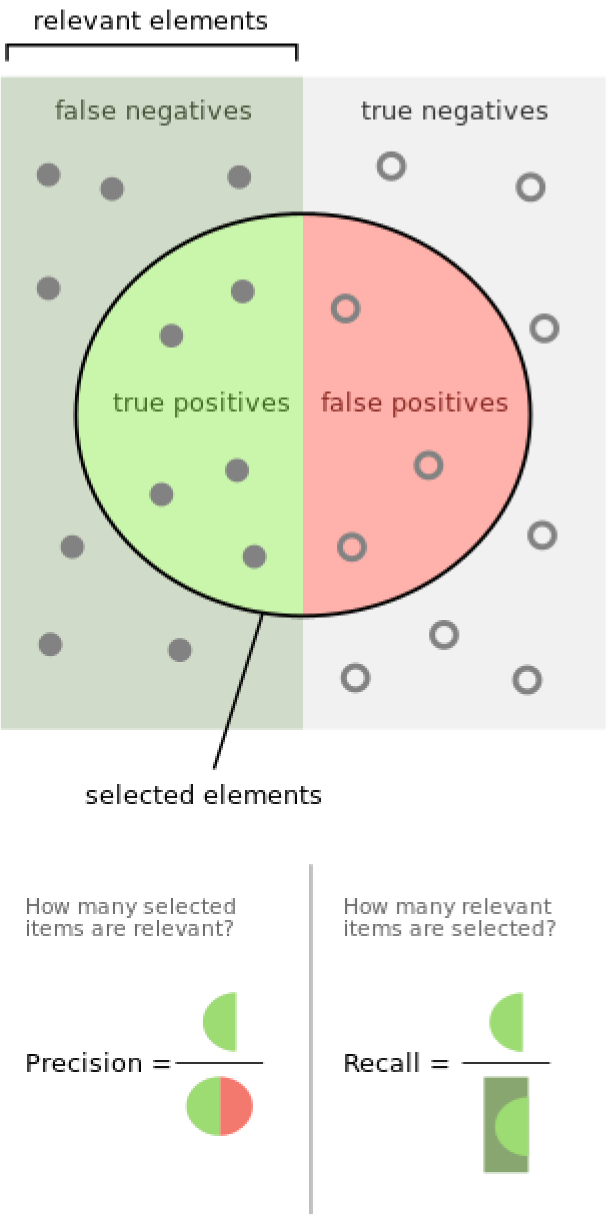
Deep Learning

- Neural Networks
  - Classification
- Auto-encoders
  - Automatic Feature Learning
- Self Organizing Maps
  - Clustering

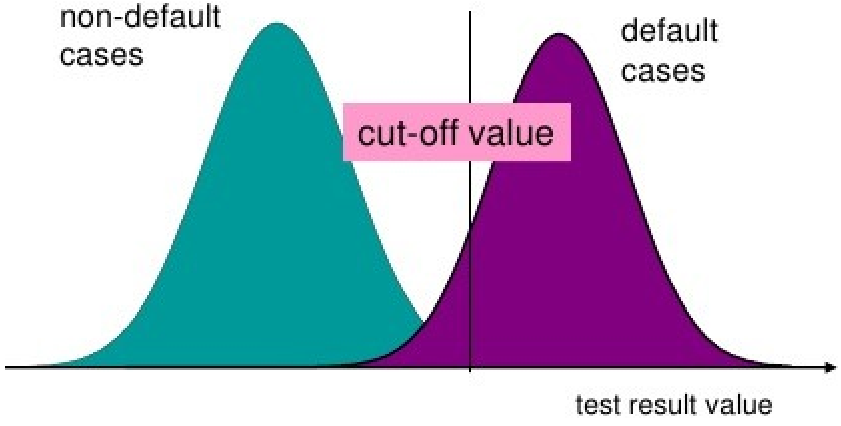


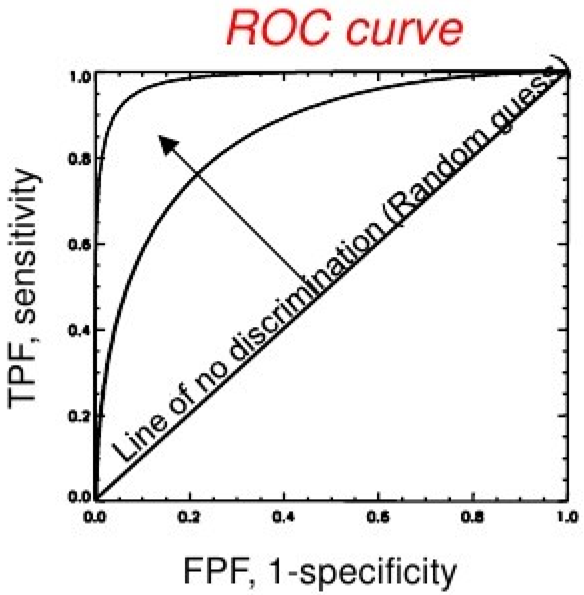
Evaluation and Tuning

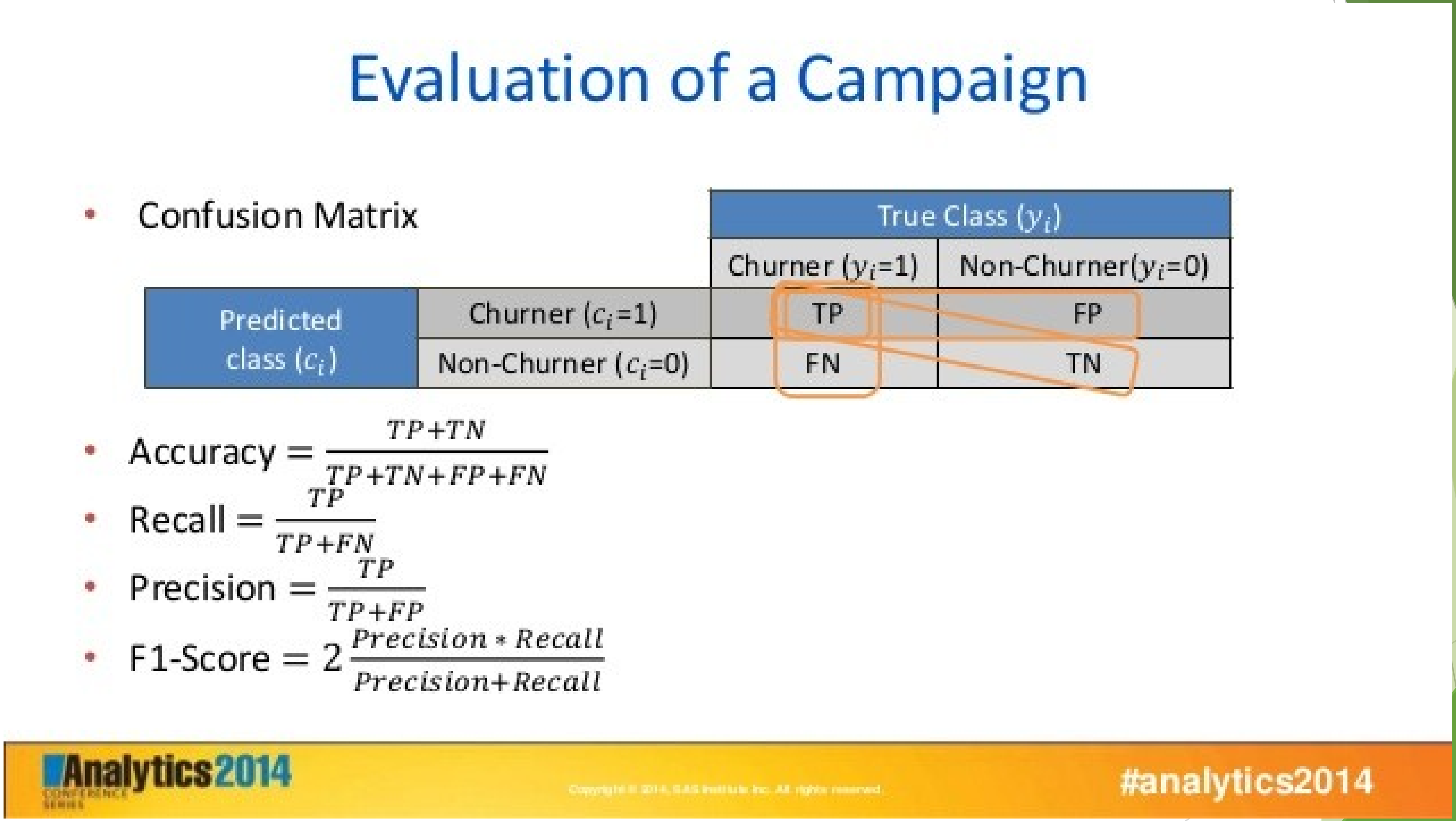
- Accuracy
- RMSE
- F1 measure
- Precision and Recall
- Receiver Operating Characteristics (ROC)


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many selected items are relevant?
How many relevant items are selected?
Precision =
Recall =

We evaluate performance of the model







Evaluation of a Campaign

- Confusion Matrix

		True Class (y_i)	
		Churner (y_i=1)	Non-Churner(y_i=0)
Predicted class (c_i)	Churner (c_i=1)	TP	FP
	Non-Churner (c_i=0)	FN	TN

- Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- Recall $= \frac{TP}{TP+FN}$
- Precision $= \frac{TP}{TP+FP}$
- F1-Score $= 2\frac{Precision * Recall}{Precision+Recall}$

Analytics2014 CONFERENCE SERIES

#analytics2014

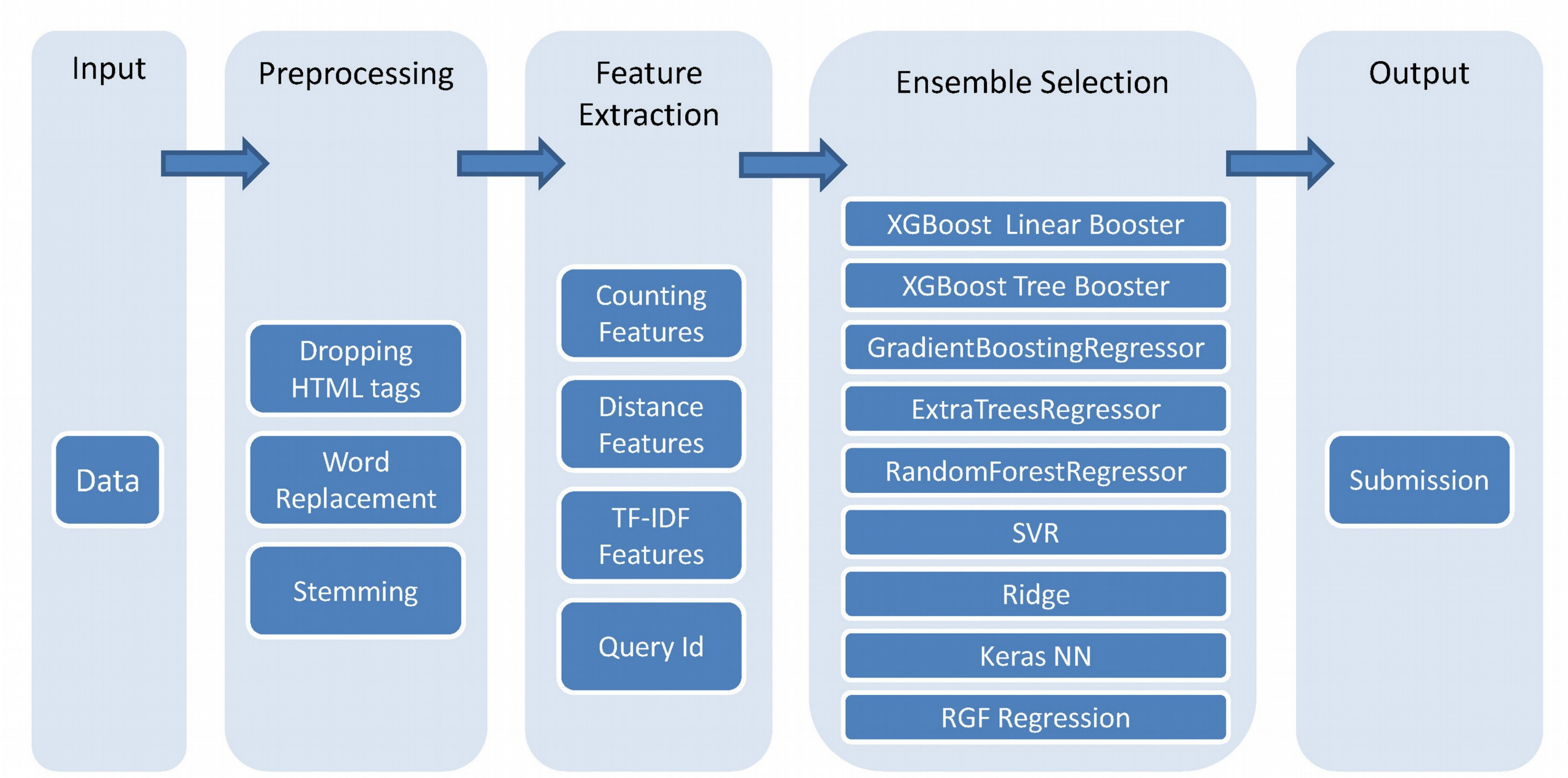
Tools

- Programming
  - Python: Scikit-Learn, Py-Weka, NLTK, PyBrain
  - PySpark: mllib (distributed)
  - R
  - Matlab, Neural Network Toolkit, Image Processing Toolkit
- Experimentation
  - Weka: Explorer, Experimenter, Knowledge Flow
- Visualization
  - Python: matplotlib
  - Tableau
  - D3 js

Deployment Frameworks

- Database: Scalable, Distributed
  - Graph Based: Neo4j
  - Document Store: MangoDB
  - Other: HDFS, Spark (RDD)
- Handle Big Data – Map Reduce Programming Paradigm
  - Apache Spark
    - MLLIB
    - Streaming
  - Apache Storm
    - Topology – Spout, Bolt

NLP – Natural Language Processing



Input

Data

Preprocessing

Dropping HTML tags

Word Replacement

Stemming

Feature Extraction

Counting Features

Distance Features

TF-IDF Features

Query Id

Ensemble Selection

XGBoost Linear Booster

XGBoost Tree Booster

GradientBoostingRegressor

ExtraTreesRegressor

RandomForestRegressor

SVR

Ridge

Keras NN

RGF Regression

Output

Submission

Questions?

Thank you