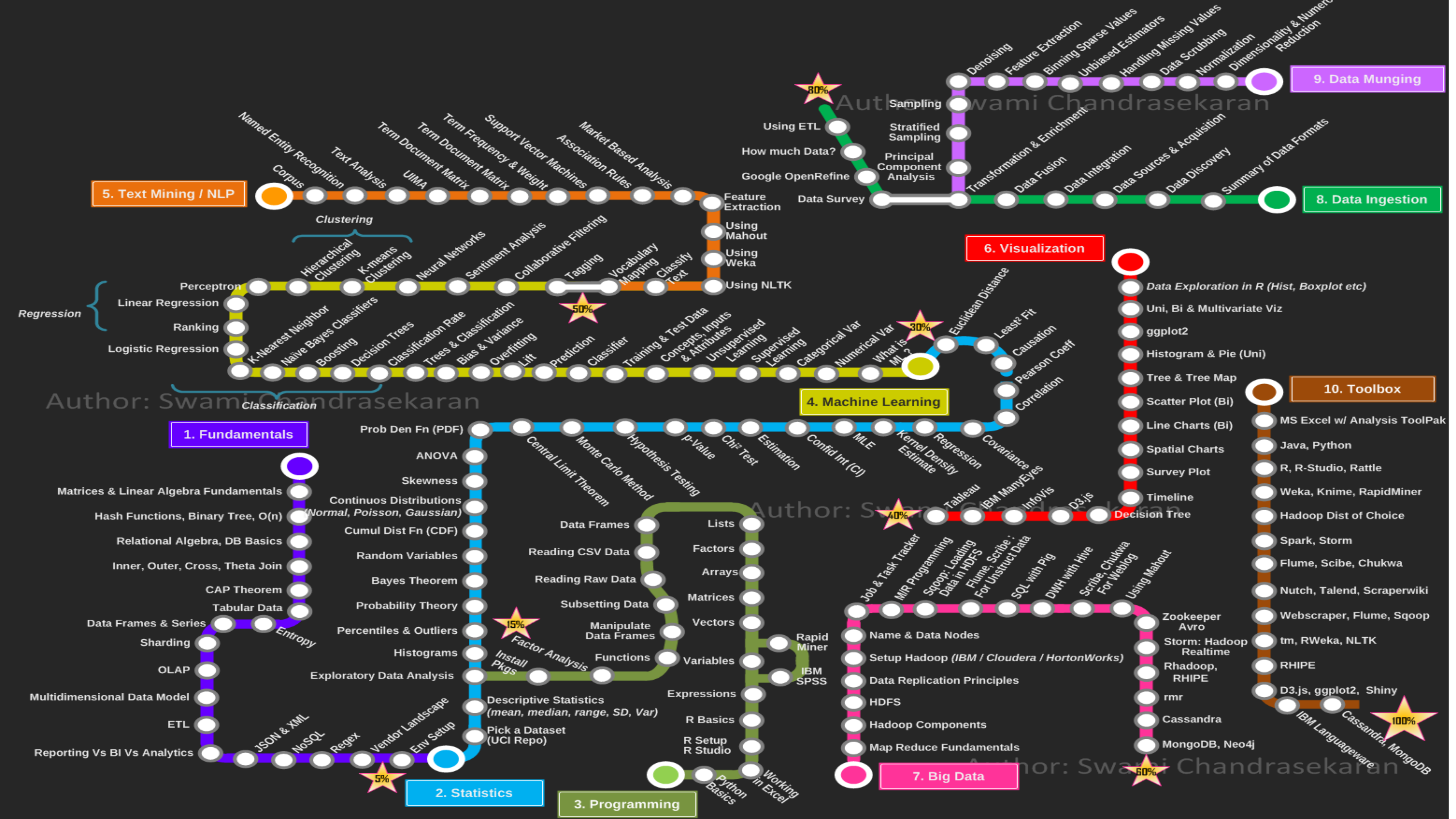
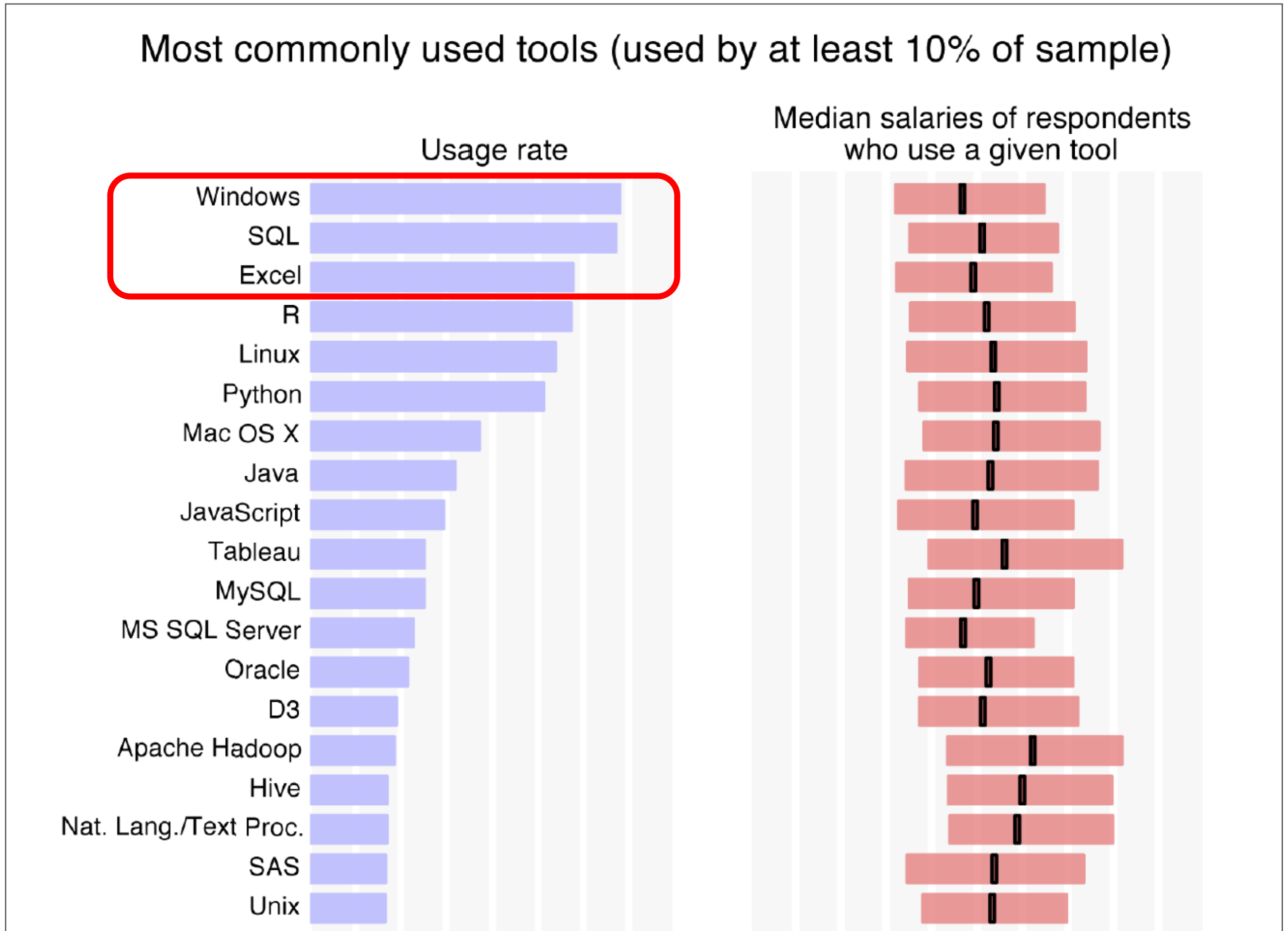
Data Science

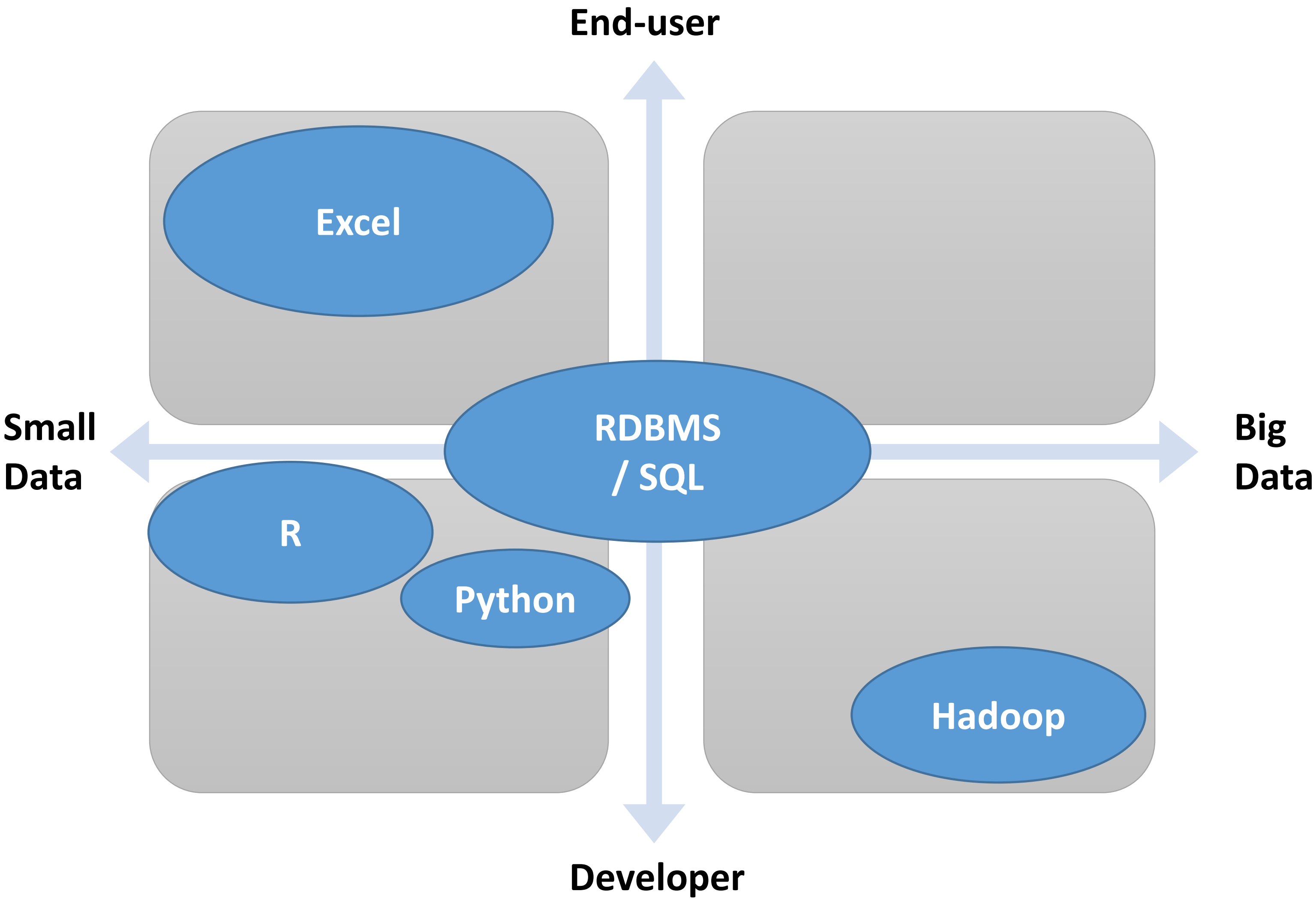

1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Entropy
Data Frames & Series
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
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 ?
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
Naïve 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
30%
50%
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
Classify Text
Vocabulary Mapping
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
Summary of Data Formats
Data Discovery
Data Sources & Acquisition
Data Integration
Data Fusion
Transformation & Enrichment
Data Survey
Google OpenRefine
How much Data?
Using ETL
80%
9. Data Munging
Dimensionality & Numerosity Reduction
Normalization
Data Scrubbing
Handling Missing Values
Unbiased Estimators
Binning Sparse Values
Feature Extraction
Denoising
Sampling
Stratified Sampling
Principal Component Analysis
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%

Data Science Tool Usage Survey (2014/O'Rielly)

- Still dominated by simple tools…

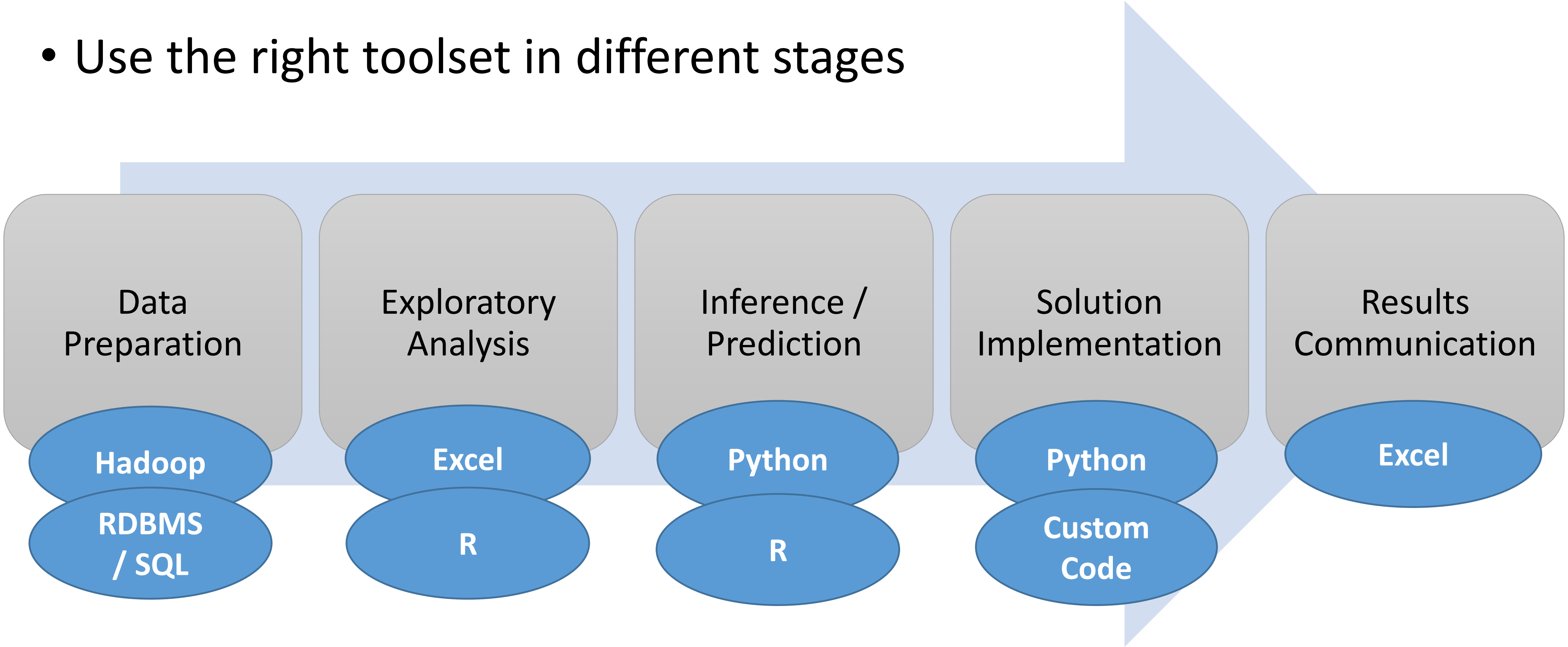


Choosing Tools for Data Science



Chaining Tools for Data Science

- Use the right toolset in different stages



What is Data Science?

- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

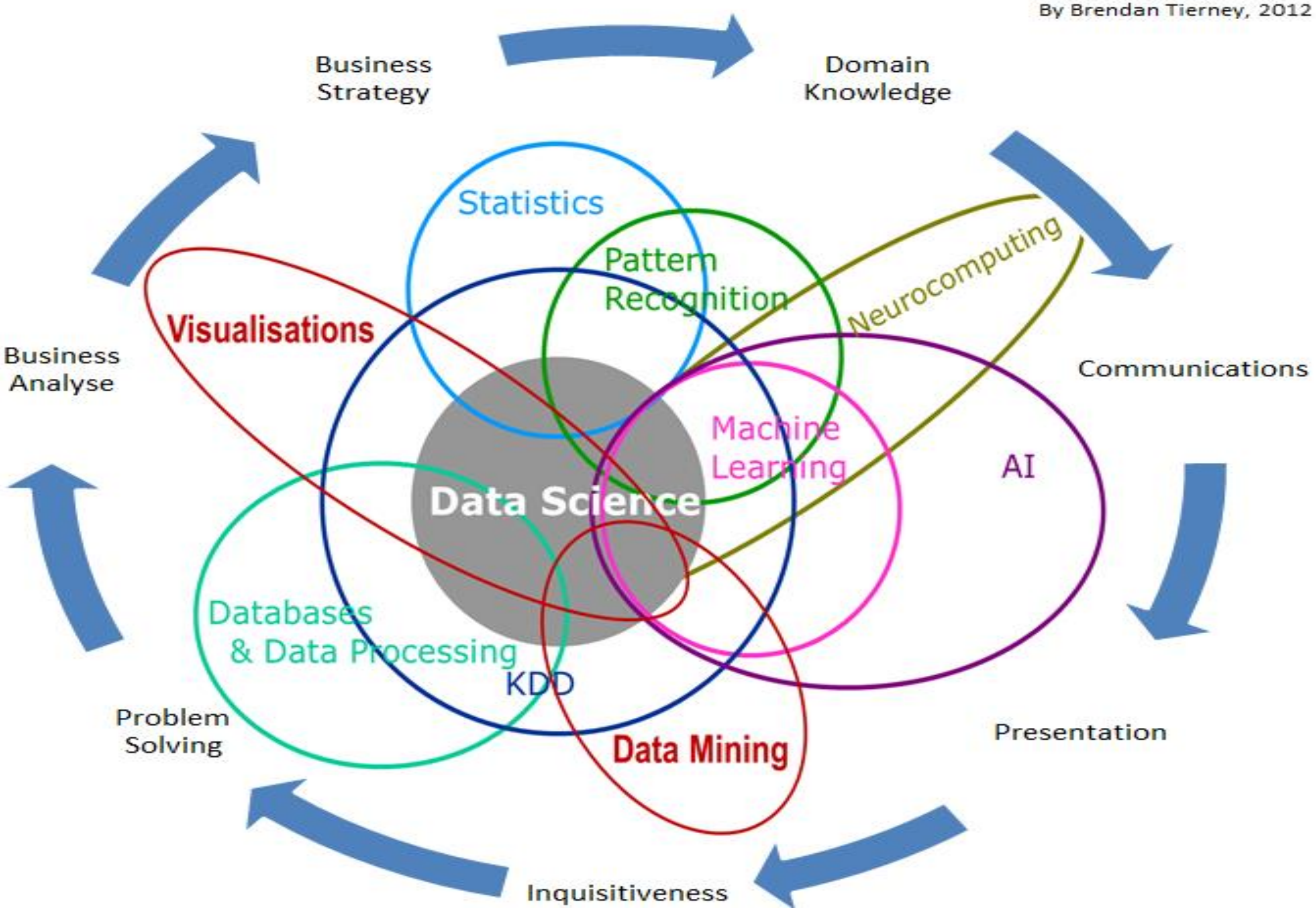
https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

What is Data Science?

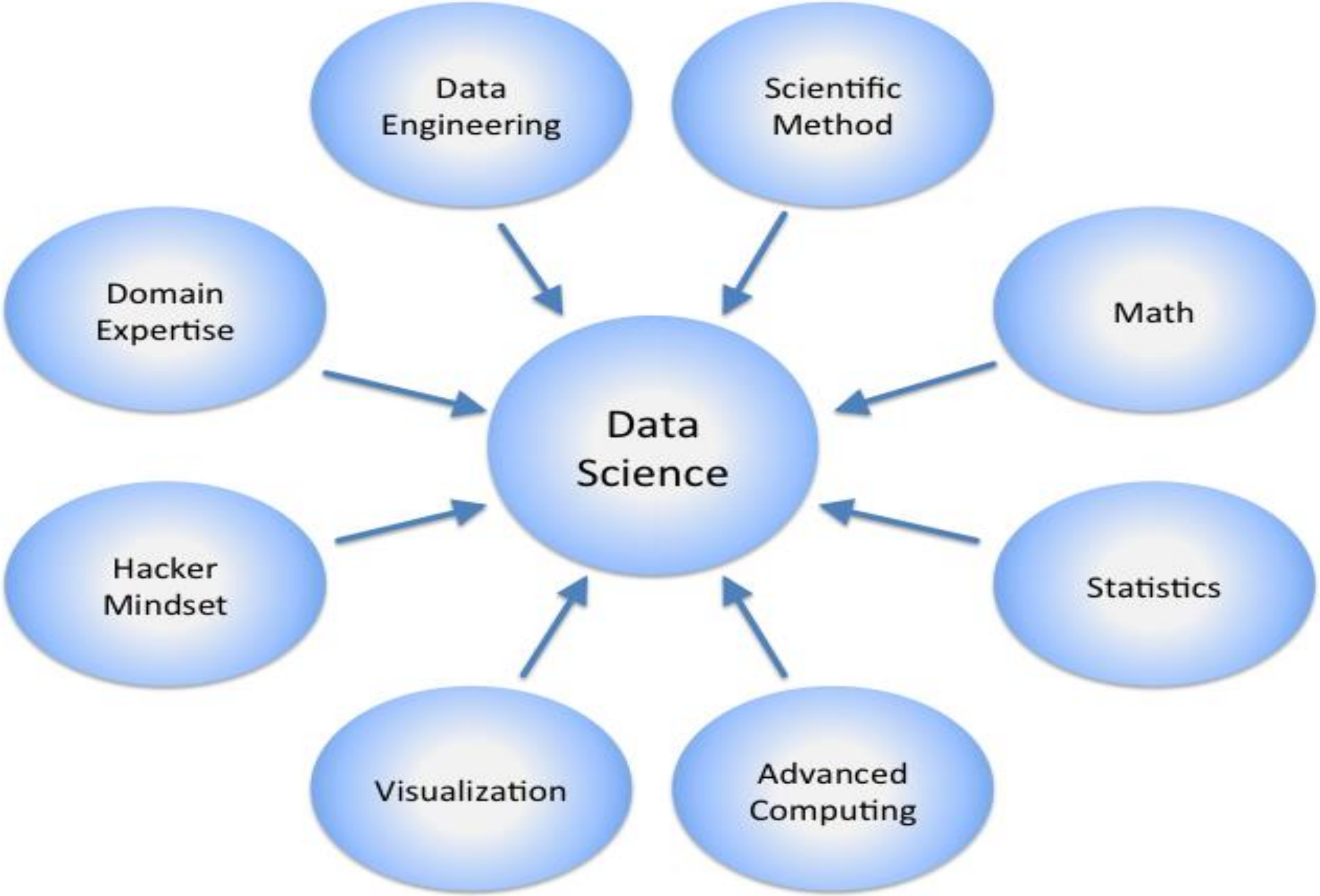
- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

Data Science Is Multidisciplinary

By Brendan Tierney, 2012

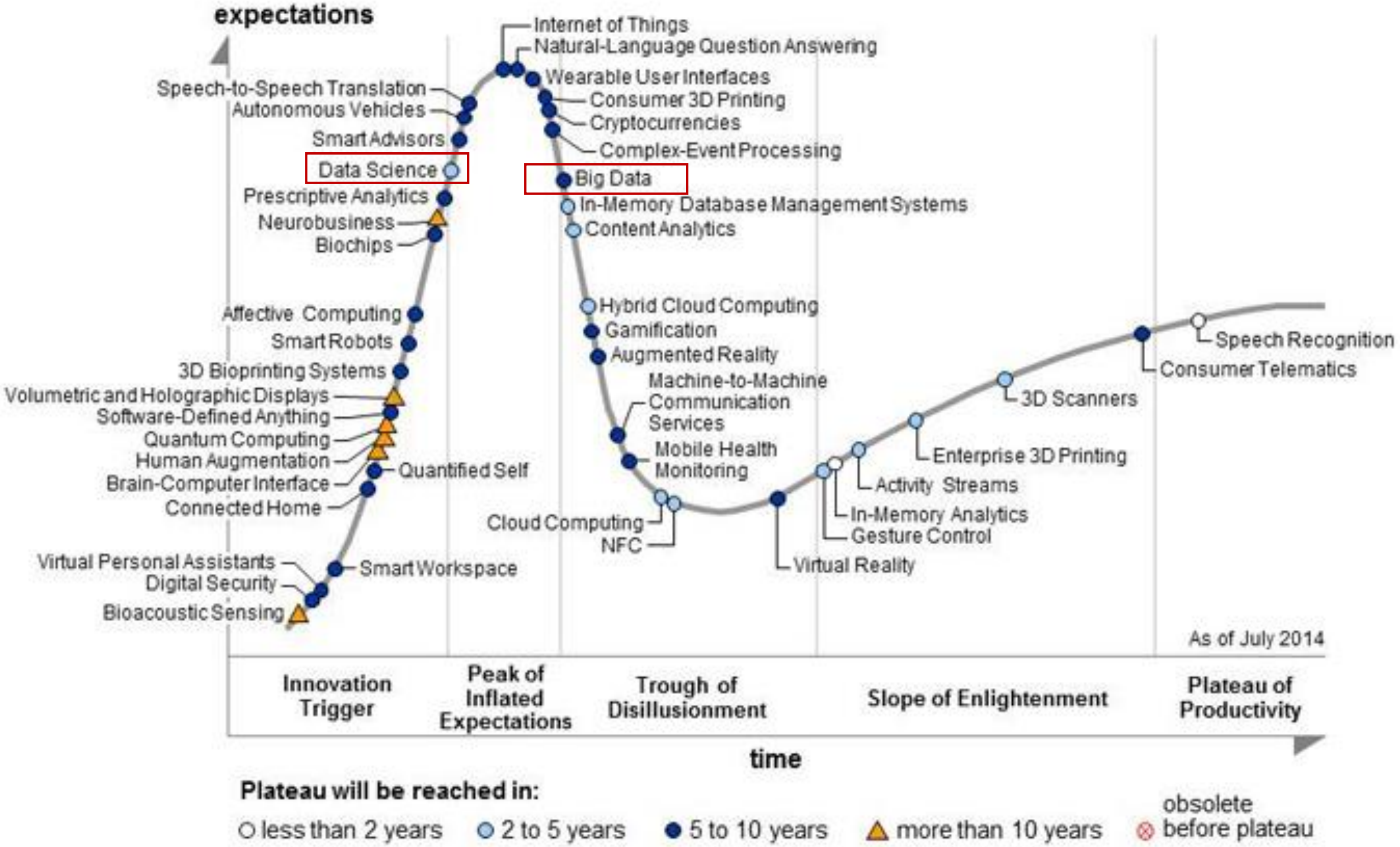


Data Science



Why is it sexy?

- Gartner's



Data Scientists

- Data Scientist
  - The Sexiest Job of the 21st Century
- They find stories, extract knowledge. They are not reporters

Data Scientists

- Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions

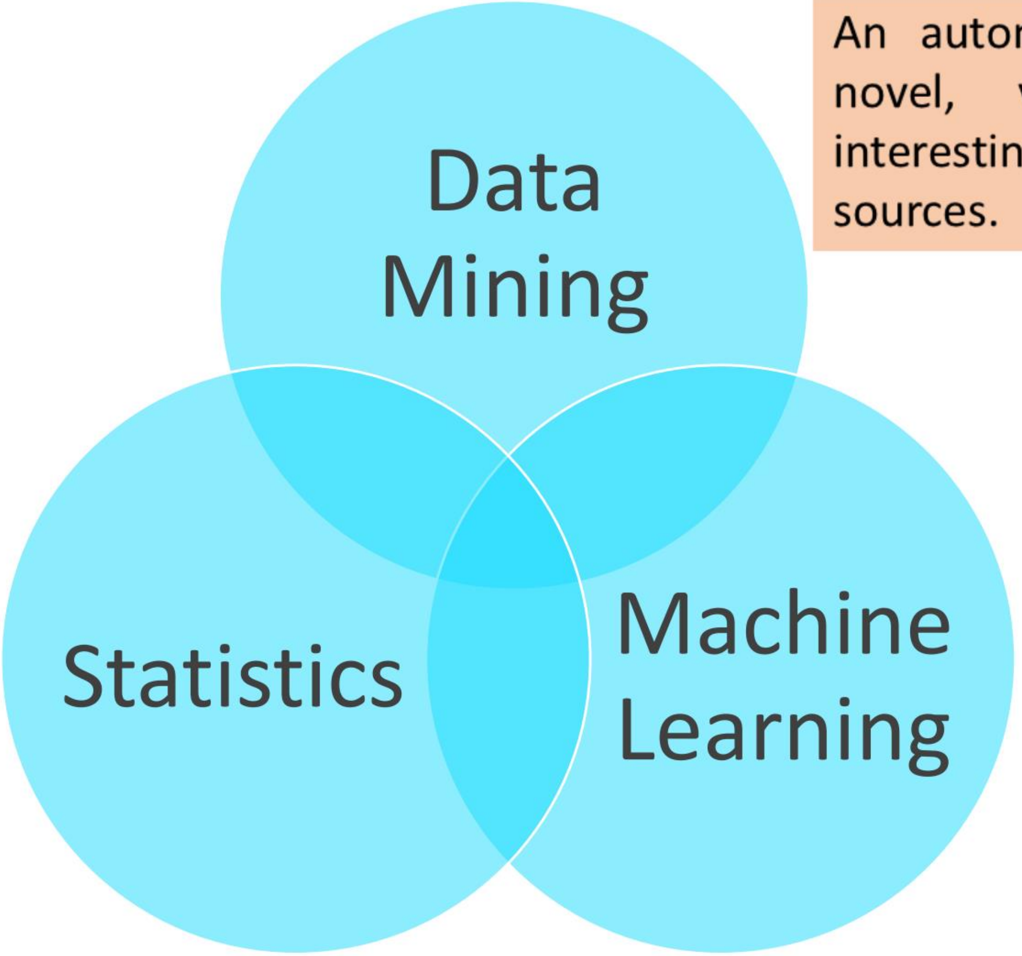
Concentration in Data Science

- Mathematics and Applied Mathematics
- Applied Statistics/Data Analysis
- Solid Programming Skills (R, Python, Julia, SQL)
- Data Mining
- Data Base Storage and Management
- Machine Learning and discovery

Machine Learning Problems

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

COMPLEMENTARIES AND DIFFERENCES



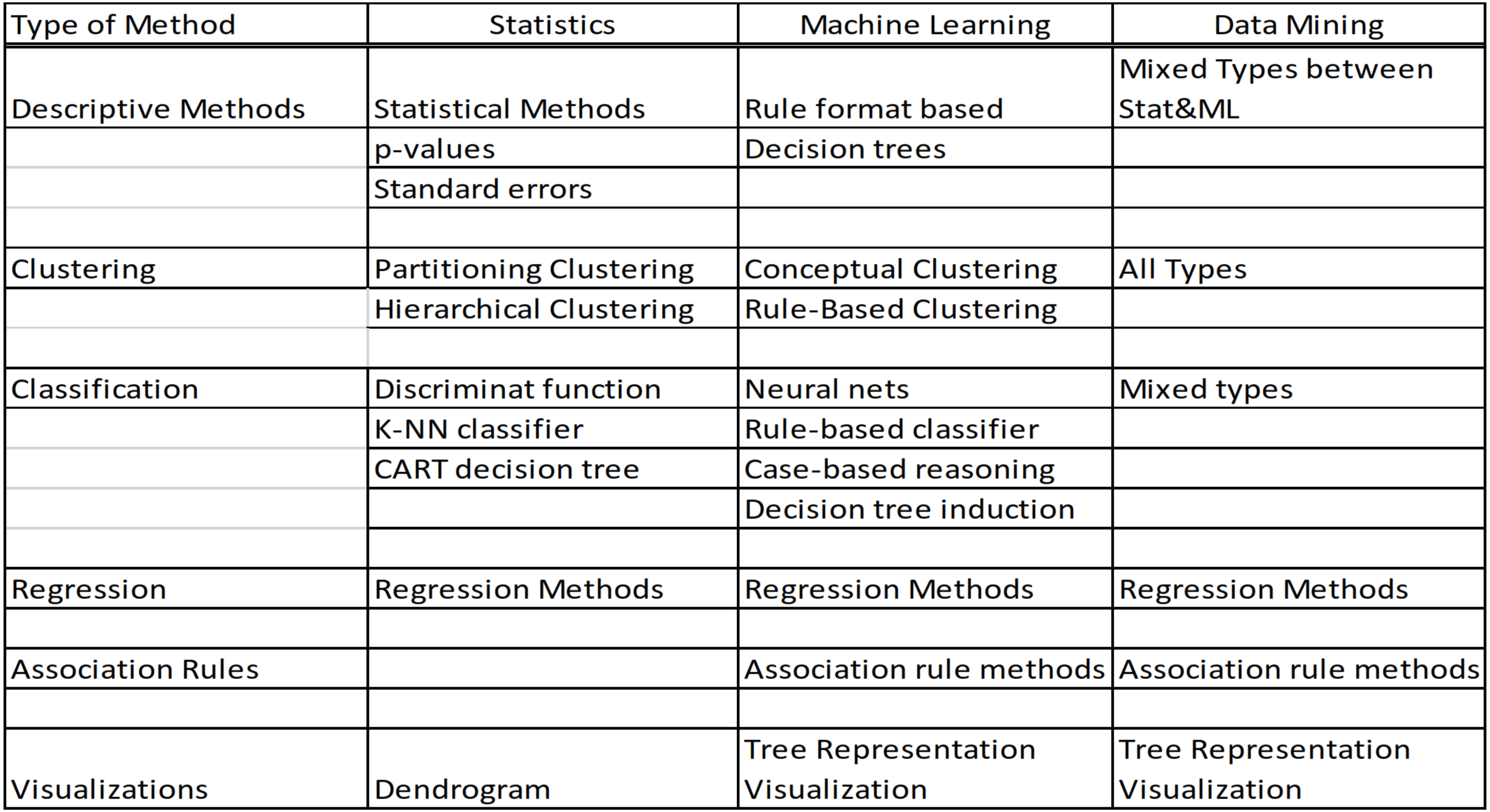
An automated process used to discover novel, valid, useful and potentially interesting knowledge from large data sources.

- Statistical analysis outputs: p-values, standard errors, regression models, principal components, discriminant score functions, ANOVA tables, control charts, descriptive statistics etc…
- translate statistical results into relevant information, careful formulation of findings is required

- Deals with representation and generalization
- Representation of data instances and functions evaluated on these instances
- Generalization is the property that the system will perform well on unseen instances

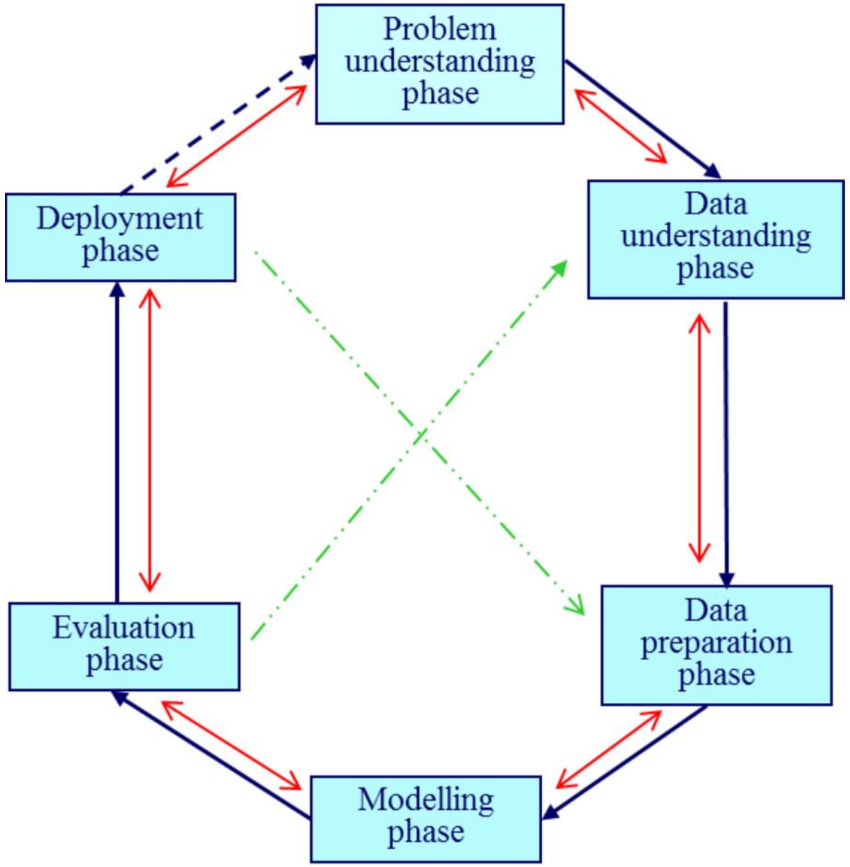
An Example of Machine Learning: Google’s Self-Driving Car

COMPLEMENTARIES AND DIFFERENCES



Type of Method	Statistics	Machine Learning	Data Mining
Descriptive Methods	Statistical Methods	Rule format based	Mixed Types between Stat&ML
	p-values	Decision trees	
	Standard errors		
Clustering	Partitioning Clustering	Conceptual Clustering	All Types
	Hierarchical Clustering	Rule-Based Clustering	
Classification	Discriminat function	Neural nets	Mixed types
	K-NN classifier	Rule-based classifier	
	CART decision tree	Case-based reasoning	
		Decision tree induction	
Regression	Regression Methods	Regression Methods	Regression Methods
Association Rules		Association rule methods	Association rule methods
Visualizations	Dendrogram	Tree Representation Visualization	Tree Representation Visualization

THE DATA MINING CYCLE AND SOME TYPICAL APPLICATION DOMAINS



Problem understanding phase

Data understanding phase

Data preparation phase

Modelling phase

Evaluation phase

Deployment phase

- Database Marketing
- Finance
- Manufacturing
- Health Insurance
  Medicine/Patient records
  Bioinformatics
- Homeland Security
- BioDefense
- Text Mining
- WWW mining
  Crowdsourcing
  Twitter

Optimization algorithms

- Example: $f = x_1 + x_2 + x_1{}^2 + x_1 e^{-x_2} + x_2{}^2 e^{-x_1}$

Stationary points:

$$\nabla f = \mathbf{0} \Rightarrow \begin{cases} \dfrac{\partial f}{\partial x_1} = 1 + 2x_1 + e^{-x_2} - x_1 x_2{}^2 e^{-x_1} = 0 \\ \dfrac{\partial f}{\partial x_2} = 1 - x_1 x_2 e^{-x_2} + 2x_2 e^{-x_1} = 0 \end{cases}$$