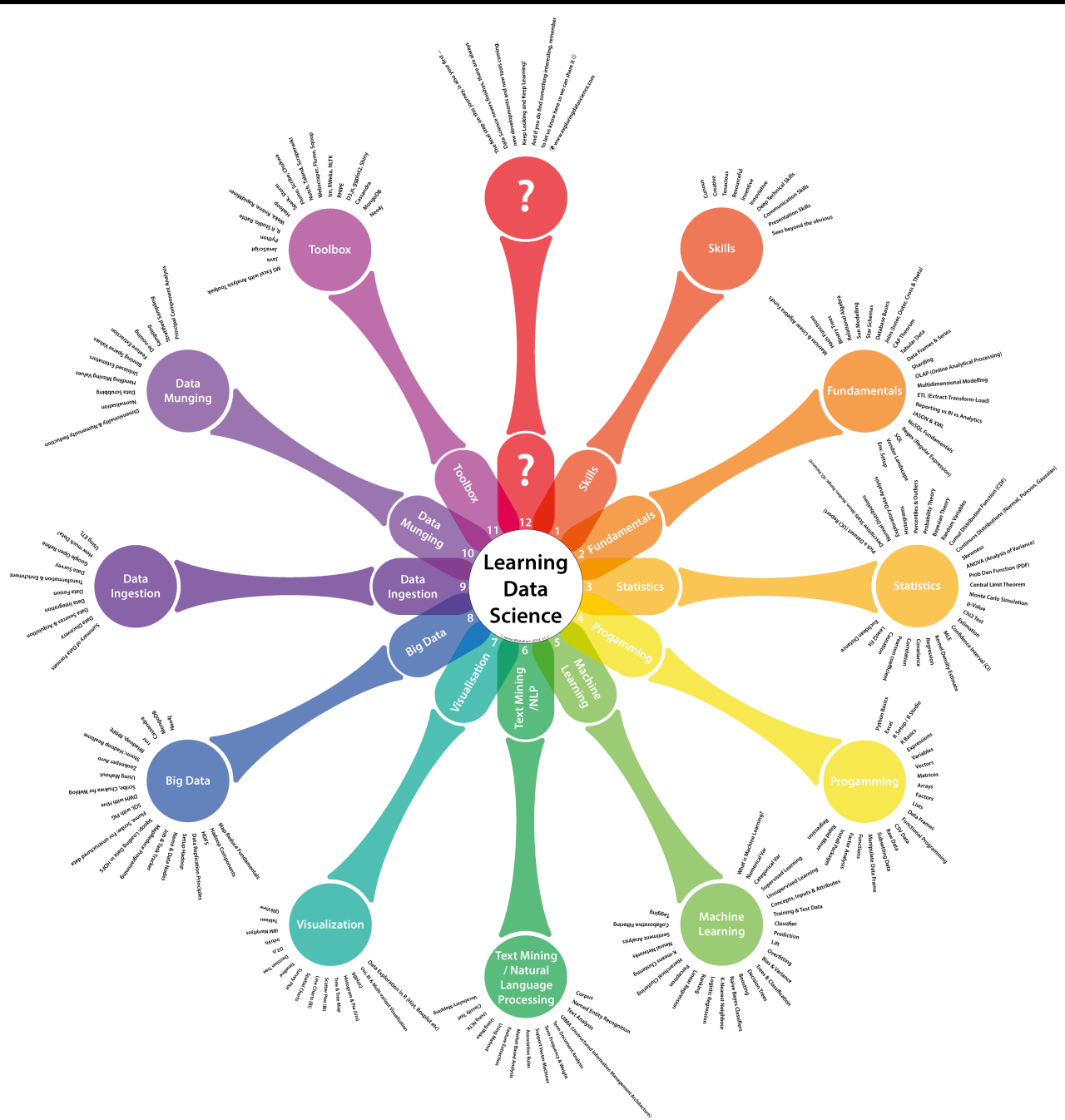
به‌نام آنکه به شماره موجودات آگاه است

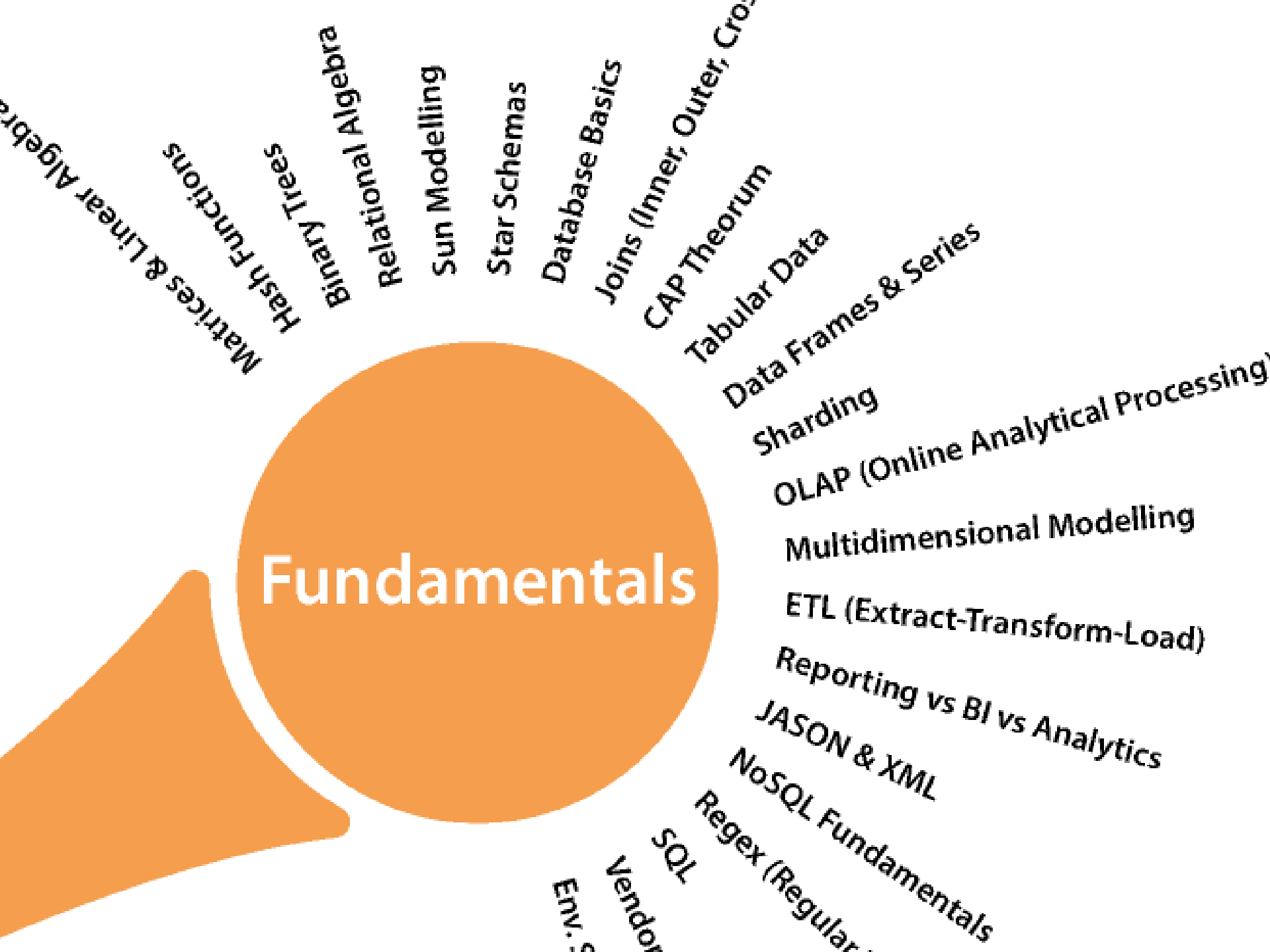
آمار و رشته‌ی علوم داده‌ها

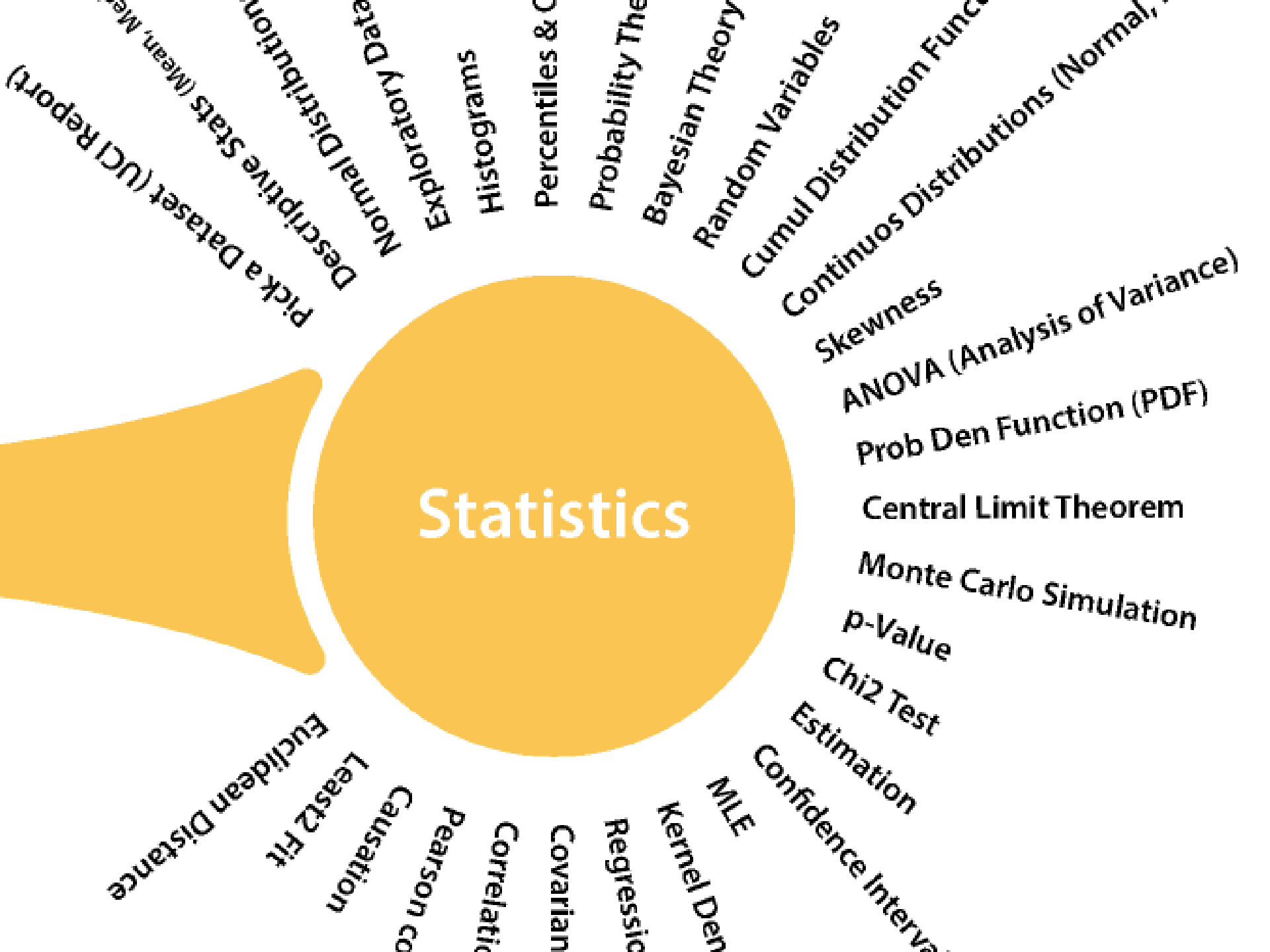
عادل محمدپور
دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)
AdelM.ir
adel@aut.ac.ir


Learning Data Science
1 Fundamentals
2 Statistics
3 Progamming
4 Machine Learning
5 Text Mining /NLP
6 Visualisation
7 Big Data
8 Data Ingestion
9 Data Munging
10 Toolbox
11 ?
12 Skills
Fundamentals
Statistics
Progamming
Machine Learning
Text Mining / Natural Language Processing
Visualization
Big Data
Data Ingestion
Data Munging
Toolbox
?
Skills


Learning Data Science
© Jamie Whitehorn 2014 v1.1
?
1 Skills
2 Fundamentals
3 Statistics
4 Progamming
5 Machine Learning
6 Text Mining /NLP
7 Visualisation
8 Big Data
9 Data Ingestion
10 Data Munging
11 Toolbox
12


Curious
Creative
Tenacious
Resourceful
Inventive
Innoviative
Deep Technical Skills
Communication Skills
Presentation Skills
Sees beyond the obvious
Skills


Fundamentals
Matrices & Linear Algebra
Hash Functions
Binary Trees
Relational Algebra
Sun Modelling
Star Schemas
Database Basics
Joins (Inner, Outer, Cro
CAP Theorum
Tabular Data
Data Frames & Series
Sharding
OLAP (Online Analytical Processing)
Multidimensional Modelling
ETL (Extract-Transform-Load)
Reporting vs BI vs Analytics
JASON & XML
NoSQL Fundamentals
Regex (Regular
SQL
Vendor
Env.


Statistics
Pick a Dataset (UCI Report)
Descriptive Stats (Mean, Me
Normal Distribution
Exploratory Data
Histograms
Percentiles &
Probability The
Bayesian Theory
Random Variables
Cumul Distribution Func
Continuos Distributions (Normal,
Skewness
ANOVA (Analysis of Variance)
Prob Den Function (PDF)
Central Limit Theorem
Monte Carlo Simulation
p-Value
Chi2 Test
Estimation
Confidence Interva
MLE
Kernel Den
Regressio
Covarian
Correlatio
Pearson Co
Causation
Least2 Fit
Euclidean Distance

Progamming

Python B

Excel

R Setup / R St

R Basics

Expressions

Variables

Vectors

Matrices

Arrays

Factors

Lists

Data Frames

Functional Programming

CSV Data

Raw Data

Subsetting Data

Manipulate Dat

Functions

Factor Analysis

Install Packages

Rapid Miner

Regression

rning

ing

Machine Learning

What is
Numerical
Categorical Va
Supervised Learning
Unsupervised Learning
Concepts, Inputs & Attributes
ages
Training & Test Data
Classifier
Prediction
Lift
Overfitting
Bias & Variance
Trees & Classification
Decision Trees
Boosting
Naive Bayes Class
K-Nearest Neigh
Logistic Regressi
Ranking
Linear Regression
Perceptron
Hierarchical Clustering
k-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filter
Tagging

Text Mining / Natural Language Processing

Corpus

Named Entity Recognition

Text Analysis

UIMA (Unstructured Information Management Architec

Term Document Analysis

Term Frequency & Weight

Support Vector Machines

Association Rules

Market Based Analysis

Feature Extraction

Using Mahout

Using Weka

Using NLTK

Classify Text

Vocabulary Map

al Clustering

ns Clustering

al Networks

Visualization

- Data Exploration in R (Hist, Boxplot etc)
- Uni, Bi & Multi-variant Visualisation
- 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
- QlikView

Big Data

- Map Reduce Fundamentals
- Hadoop Components
- HDFS
- Data Replication Pr
- Setup Hadoop
- Name & Data Nodes
- Job & Task Tracker
- MapReduce programming
- Sqoop: Loading Data in HDFS
- Flume, Scribe: For unstructured data
- SQL with PIG
- DWH with Hive
- Scribe, Chukwa for Weblog
- Using Mahout
- Zookeeper Avro
- Storm: Hadoop Realtime
- RHadoop, RHIPE
- rmr
- Cassandra
- Mongo
- Neo4

Data
Ingestion
Summary of Data Formats
Data Discovery
Data Sources & Acquisition
Data Integration
Data Fusion
Transformation & Enrichment
Data Survey
Google Open Refine
How much Data?
Using ETL

Data Munging

Dimensionality Reduction

Normalisation

Data Scrubbing

Handling Missing Values

Unbiased Estimators

Binning Sparse Values

Feature Extraction

De-noising

Sampling

Stratified Sampling

Principal Component Analysis

nalysis Toolpak

Toolbox
MS Excel with
Java
JavaScript
Python
R, R Studio, Rattle
Weka, Knime, RapidMiner
Hadoop
Spark, Storm
Flume, Scribe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
Cassandra
MongoDB
Neo4j

The final step on this journey, is also your first...

Data science nevers finishes, there are always

new developments and new tools coming

Keep Looking and Keep Learning!

And if you do find something interesting

to let us know here so we can share it ☺

www.exploringdatascience.com

?