Data Science With Python

Mosky

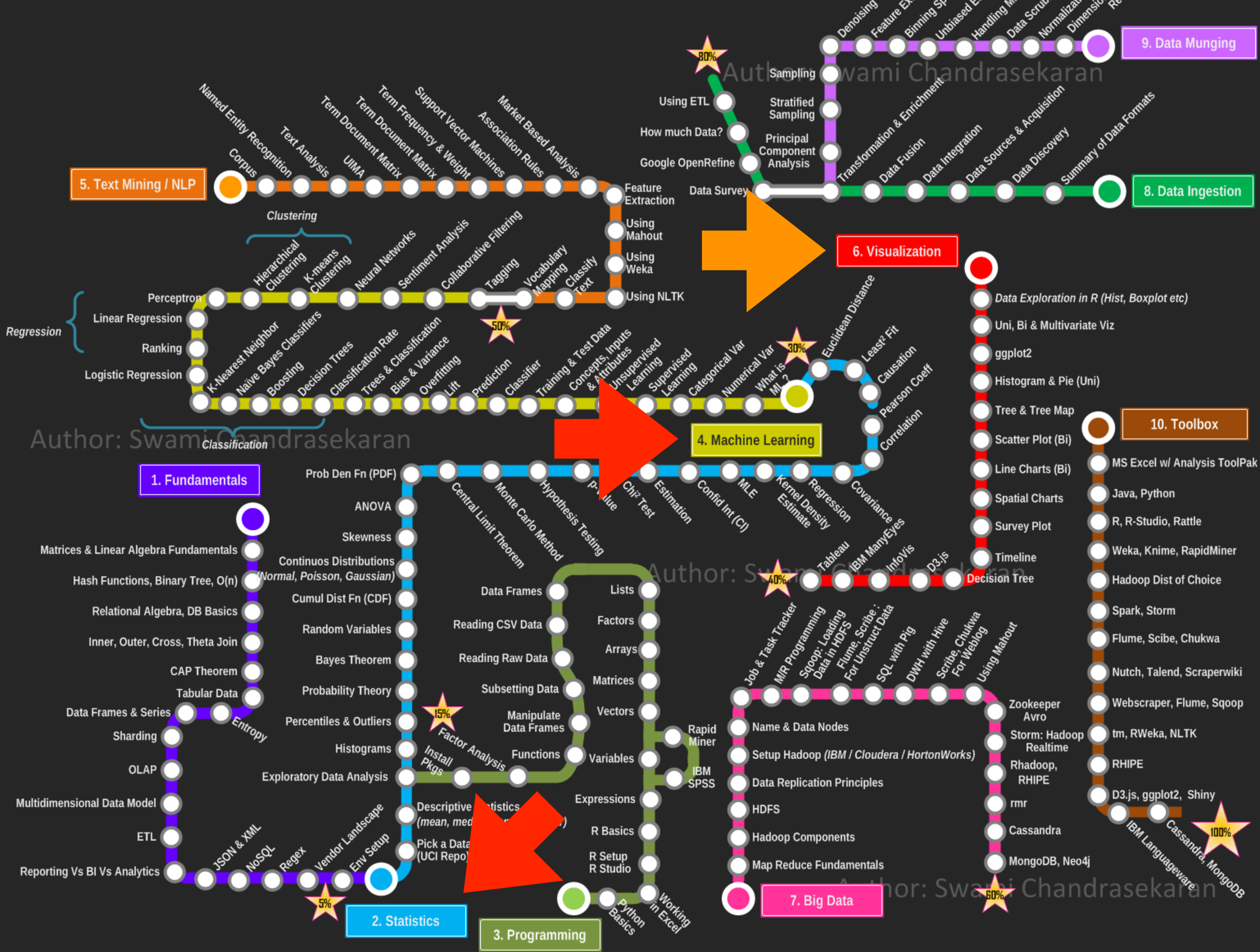
Data Science

- = Extract knowledge or insights from data.
- Data science includes:
  - Visualization
  - Statistics
  - Machine learning
  - Deep learning
  - Big data
  - And related methods
- ≈ Data mining

Data Science

- = Extract knowledge or insights from data.
- Data science includes:
  - Visualization
  - Statistics
  - Machine learning
  - Deep learning
  - Big data
  - And related methods
- ≈ Data mining

We will introduce.

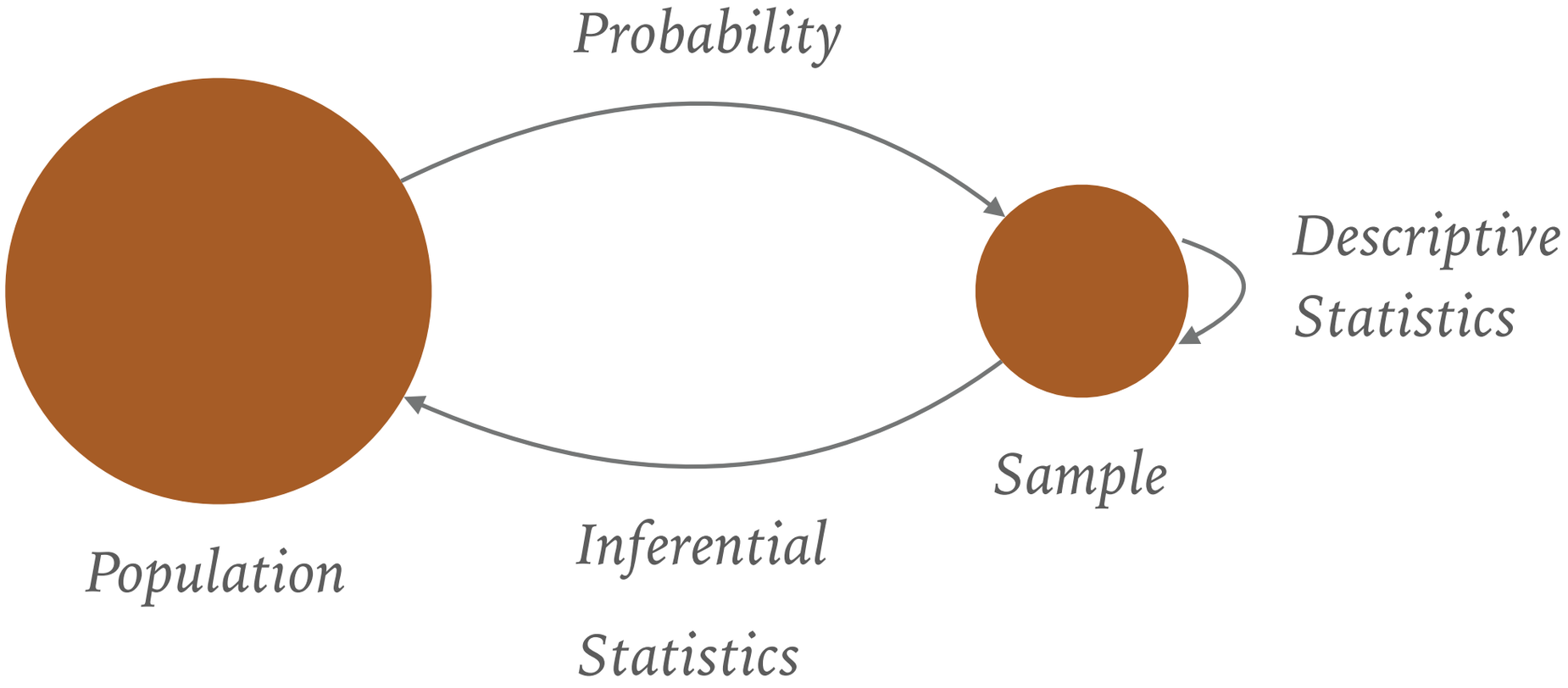

Author: Swami Chandrasekaran
1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics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

- It's kind of outdated, but still contains lot of keywords.
- MrMimic/data-scientist-roadmap – GitHub
- Becoming a Data Scientist – Curriculum via Metromap

Statistics vs. Machine Learning

- Machine learning = statistics - checking of assumptions 😆
- But does resolve more problems. 👍🏻
- Statistics constructs more solid inferences.
- Machine learning constructs more interesting predictions.

Probability, Descriptive Statistics, and Inferential Statistics



Machine Learning vs. Deep Learning

- Deep learning is the most renowned part of machine learning.
  - A.k.a. the “AI”.
- Deep learning uses artificial neural networks (NNs).
- Which are especially good at:
  - Computer vision (CV) 👀
  - Natural language processing (NLP) 📖
    - Machine translation
  - Speech recognition
- Too costly to simple problems.

Big Data

- The “size” is constantly moving.
  - As of 2012, ranges from 10n TB to n PB, which is 100x.
- Has high-3Vs:
  - Volume, amount of data.
  - Velocity, speed of data in and out.
  - Variety, range of data types and sources.
- A practical definition:
  - A single computer can't process in a reasonable time.
  - Distributed computing is a big deal.

Today.

- “Models” are the math models.
- “Statistical models”, emphasize inferences.
- “Machine learning models”, emphasize predictions.
- “Deep learning” and “big data” are gigantic subfields.

- We won't introduce.
- But the learning resources are listed at the end.

Mosky

- Python Charmer at Pinkoi.
- Has spoken at
  - PyCons in
    TW, MY, KR, JP, SG, HK,
    COSCUPs, and TEDx, etc.
- Countless hours
  on teaching Python.
- Own the Python packages:
  - ZIPCodeTW,
    MoSQL, Clime, etc.
- http://mosky.tw/

The Outline

- “Data”
- The Analysis Steps
- Visualization
- Preprocessing
- Dimensionality Reduction
- Statistical Models
- Machine Learning Models
- Keep Learning

The Packages

- `$ pip3 install jupyter numpy scipy sympy matplotlib ipython pandas seaborn statsmodels scikit-learn`
- Or
- `> conda install jupyter numpy scipy sympy matplotlib ipython pandas seaborn statsmodels scikit-learn`

Common Jupyter Notebook Shortcuts

`Ctrl-Enter`	Run the cell.
`B`	Insert cell below.
`D, D`	Delete the current cell.
`M`	To Markdown cell.
`Cmd-/`	Comment the code.
`Esc`	Edit mode → command mode.
`H`	Show keyboard shortcuts.
`P`	Open the command palette.

Checkpoint: The Packages

- Open *00_preface_the_packages.ipynb* up.
- Run it.

"Data"

"Data"

- = Variables
- = Dimensions
- = Labels + Features

Data in Different Types

Discrete	Nominal		{male, female}
	Ordinal Ranked	↑ & can be ordered.	{great > good > fair}
Continuous	Interval	↑ & distance is meaningful.	temperatures
	Ratio	↑ & 0 is meaningful.	weights

Data in the X–Y Form

y	x
dependent variable	independent variable
response variable	explanatory variable
regressand	regressor
endogenous variable | endog	exogenous variable | exog
outcome	design
label	feature

- Confounding variables:
  - May affect y, but not x.
  - May lead erroneous conclusions, “garbage in garbage out”.
    - Controlling, e.g., fix the environment.
    - Randomizing, e.g, choose by computer.
    - Matching, e.g., order by gender and then assign group.
    - Statistical control, e.g., BMI to remove height effect.
    - Double-blind, even triple-blind trials.

Get the Data

- Logs
- Experiments
  - Strongly recommend Research Methods Knowledge Base.
- Existent datasets
  - The Datasets Package – StatsModels
  - Kaggle

The Analysis Steps

The Three Steps

1. Define Assumption
2. Validate Assumption
3. Validated Assumption?

1. Define Assumption

- Specify a *feasible* objective.
  - "Use the AI to get the moon!"
- Write an *formal* assumption.
  - "The users will buy 1% items from our recommendation." rather than "The users will love our recommendation!"
- Note the *dangerous* gaps.
  - "All the items from recommendation are free!"
  - "Correlation does not imply causation."
- Consider the *next* actions.
  - "Release to 100% of users." rather than "So great!"

2. Validate Assumption

- Collect *potential* data.
- List *possible* methods.
  - A plotting, median, or even mean may be good enough.
  - Selecting Statistical Tests – Bates College
  - Choosing a statistical test – HBS
  - Choosing the right estimator – Scikit-Learn
- Evaluate the metrics of methods with data.

3. Validated Assumption?

- Yes → Congrats! Report fully and take the actions! 🎉
- No → Check:
  - The *hypotheses* of methods.
  - The *confounding variables* in data.
  - The *formality* of assumption.
  - The *feasibility* of objective.

Iterate Fast While Industry Changes Rapidly

- Resolve the small problems first.
- Resolve the high impact/effort problems first.
- One week to get a quick result and improve
  rather than one year to get the may-be-the-best result.
- Fail fast!

Checkpoint: Pick up a Method

- Think of an interesting problem.
  - E.g., revenue is higher, but is it random?
- Pick one method from the cheatsheets.
  - Selecting Statistical Tests – Bates College
  - Choosing a statistical test – HBS
  - Choosing the right estimator – Scikit-Learn
- Remember the analysis steps.

Visualization

Visualization

- Make Data Colorful – Plotting
  - *01_1_visualization_plotting.ipynb*
- In a Statistical Way – Descriptive Statistics
  - *01_2_visualization_descriptive_statistics.ipynb*

Checkpoint: Plot the Variables

- We have three datasets.
- Star98
  - `star98_df = sm.datasets.star98.load_pandas().data`
- Fair
  - `fair_df = sm.datasets.fair.load_pandas().data`
- Howell1
  - ```
    howell1_df = pd.read_csv(
        'dataset_howell1.csv', sep=';')
    ```
- Or your own datasets.
- Plot the variables that interest you among them.
    ```

Preprocessing

Feed the Data That Models Like

- Preprocess data for:
  - Hard requirements, e.g.,
    - corpus → vectors
    - "What kind of news will be voted down on PTT?"
  - Assumptions, e.g.,
    - Samples are normally distributed (t-test).
    - Assuming features are centered around zero (SVM).
  - More representative features, e.g., total / units.
- Note that different models have different tastes.

Preprocessing

- The Dishes – Containers
  - *02_1_preprocessing_containers.ipynb*
- A Cooking Method – Standardization
  - *02_2_preprocessing_standardization.ipynb*
- Watch Out for Poisonous Data Points – Removing Outliers
  - *02_3_preprocessing_removing_outliers.ipynb*

Checkpoint: Preprocess the Variables

- Try to standardize and compare.
- Try to trim the outliners.

Dimensionality Reduction

The Model Sicks Up!

- Let's reduce the variables.
- Feed a subset → feature selection.
  - Feature Selection – Scikit-Learn
- Feed a transformation → feature extraction.
  - PCA, FA, etc.
  - Scikit-Learn defines it like non-numbers → numbers.

Dimensionality Reduction

- Principal Component Analysis
  - *03_1_dimensionality_reduction_principal_component_analysis.ipynb*
- Factor Analysis
  - *03_2_dimensionality_reduction_factor_analysis.ipynb*

Checkpoint: Reduce the Variables

- Try to *PCA(all variables)* → *the better components*, or FA.
- And then plot n-dimensional data onto 2-dimensional plane.

Statistical Models

Statistical Models

- Identify Boring or Interesting – Hypothesis Testings
  - *04_1_statistical_models_hypothesis_testings.ipynb*
  - "Hypothesis Testing With Python"
- Identify X-Y Relationships – Regression
  - *04_2_statistical_models_regression_anova.ipynb*

More Regression Models

- If y is not linear,
  - Logit or Poisson Regression | Generalized Linear Models, GLMs
- If y is correlated,
  - Linear Mixed Models, LMMs | Generalized Estimating Equation, GEE
- If x has multicollinearity,
  - Lasso or Ridge Regression
- If error term is heteroscedastic,
  - Weighted Least Squares, WLS | Generalized Least Squares, GLS
- If x is time series – predict x_0 by x_{-1}, not predict y by x,
  - Autoregressive Integrated Moving Average, ARIMA

Checkpoint: Apply a Statistical Method

➤ Try to apply the analysis steps with a statistical method.

1. Define Assumption
2. Validate Assumption
3. Validated Assumption?

Machine Learning Models

Machine Learning Models

- Apple or Orange? – Classification
  - *05_1_machine_learning_models_classification.ipynb*
- Without Labels – Clustering
  - *05_2_machine_learning_models_clustering.ipynb*
- Predict the Values – Regression
- Who Are the Best? – Model Selection

Confusion matrix, where $A = OO_2 = C[0, 0]$

	predicted - AC	predicted + BD
actual - AB	true - A	false + B
actual + CD	false - C	true + D

Common "rates" in confusion matrix

- precision = D / BD
- recall = D / CD
- sensitivity = D / CD = recall = observed power
- specificity = A / AB = observed confidence level
- false positive rate = B / AB = observed α
- false negative rate= C / CD = observed β

Ensemble Models

- Bagging
  - N independent models and average their output.
  - e.g., the random forest models.
- Boosting
  - N sequential models, the n model learns from n-1's error.
  - e.g., gradient tree boosting.

Checkpoint: Apply a Machine Learning Method

- Try to apply the analysis steps with a ML method.

    1. Define Assumption
    2. Validate Assumption
    3. Validated Assumption?

Keep Learning

Keep Learning

- Statistics
  - Seeing Theory
  - Biological Statistics
  - scipy.stats + StatsModels
  - Research Methods
- Machine Learning
  - Scikit-learn Tutorials
  - Standford CS229
  - Hsuan-Tien Lin
- Deep Learning
  - TensorFlow
  - Standford CS231n
  - Standford CS224n
- Big Data
  - Dask
  - Hive
  - Spark
  - HBase
  - AWS

The Facts

- ∵
  - You can't learn *all* things in the data science!
- ∴
  - “Let's learn to do” ❌
  - “Let's do to learn” ✅

The Learning Flow

1. Ask a question.
    - ➤ “How to tell the differences confidently?”
2. Explore the references.
    - ➤ “T-test, ANOVA, ...”
3. Digest into an answer.
    - ➤ Explore by the breadth-first way.
    - ➤ Write the code.
    - ➤ Make it work, make it right, finally make it fast.

Recap

- Let's do to learn, *not* learn to do.
- What is your objective?
- For the objective, what is your assumption?
- For the assumption, what method may validate it?
- For the method, how will you evaluate it with data?
- Q & A