TUX GLOBAL
INSTITUTE

Data science and Analytics

Duration	: 45 Hours
Monday	: 6:00pm to 8:00pm
Wednesday	: 6:00pm to 8:00pm
Thursday	: 6:00pm to 8:00pm

Know your Presenter: TENG CHANTO

Data Science Researcher

Areas of Expertise:

- Data Analytics
- Machine Learning
- Deep Learning
- Data Management
- Business Intelligence
- Researcher
- Lecturer

TENG CHANTO is a former Data Science Researcher and has worked with prestigious companies like Banking , software company in the past 4 years.

Professional Experience

- Worked on credit Risk Analytics(CLM & Pricing), p2p landing(Credit Risk),Banking , Student feedback Analysis and factors to chose digital skill in Cambodia.
- Hands on experience in Banking sector work as data Management at Chipmong Bank using SQL ,Excel and Oracle BI
- Hands on analytical techniques including classical & machine learning algorithms including regression, instance based, regularization, Decision tree, Bayesian, clustering, and Ensemble algorithms
- Worked on Data Analytics research on Stock prediction using machine learning (Soramithsu khmer)
- Have used different statistical flat forms like Excel, Python, Machine learning, SQL, Excel, PowerBI, MongoDB, Java, Php
- Joined various work shop and training on data science with local and international.
- Achievements :
  - winning BS.C Computer Science Scholarship ICCR (2010) in India
  - winning MS.C Computer Science Scholarship ICCR (2016) In India

Academic Credentials

- Master of Computer science(Machine learning/Statistics): Bangalore, India

Course Assessment

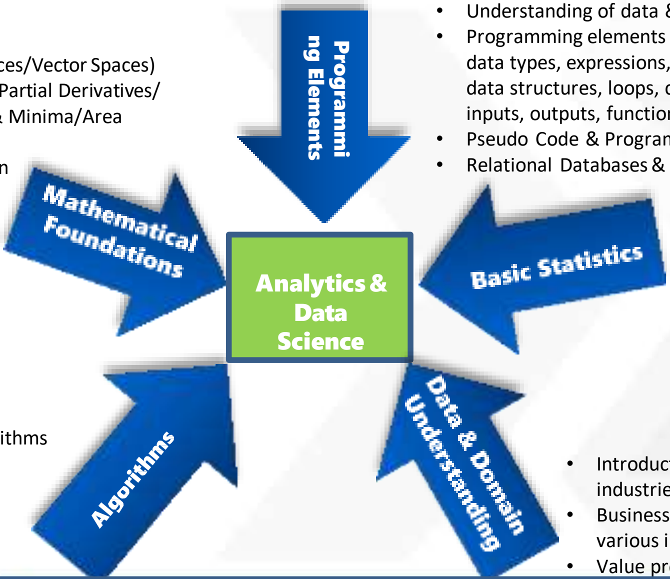
Project	30%
Attendance	15%
Class Activities	15%
Exam	40%

Introduction to Data Science

Foundations of Analytics & Data Science

- Introduction to Basic Analytics Tools: Excel
- Understanding of data & storage
- Programming elements (variables, constants, data types, expressions, keywords, comments, data structures, loops, conditional statements, inputs, outputs, functions etc…)
- Pseudo Code & Programming Languages
- Relational Databases & SQL

- Linear Algebra (Matrices/Vector Spaces)
- Calculus (Derivatives/Partial Derivatives/ Integration/Maxima & Minima/Area Under the curve
- Theory of Optimization



- **Basic Statistics**
- **Measures of central tendencies/ variance /Frequency/Rank**
- **Probability, Distributions,**
- **Conditional Probability**
- **Relationships**
- **Others: CLT, Confidence Intervals, Hypothesis testing etc..**

- Design of Algorithms
- Various types of Algorithms

- Introduction to data models in various industries & functions
- Business problems (Pain points) in various industries
- Value proposition of Analytics across

What is Data Science?

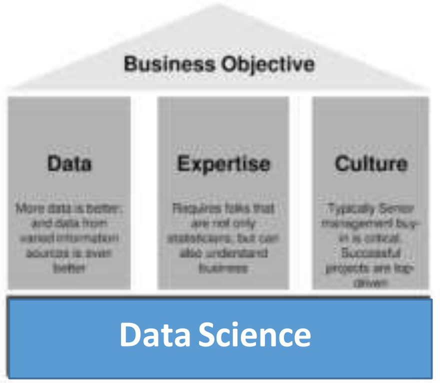
> *“To gain insights into data through computation,statistics,and visualization.”*

QuoraThreads for Expert Definitions

- What is Data Science?
- What does a Data Scientist do?

Data Science is Process

- ✓ Ask an interesting question
- ✓ Get the data
- ✓ Explore the data
- ✓ Model the data
- ✓ Communicate and visualize your results



Business Objective

Data

More data is better, and data from varied information sources is even better

Expertise

Requires folks that are not only statisticians, but can also understand business

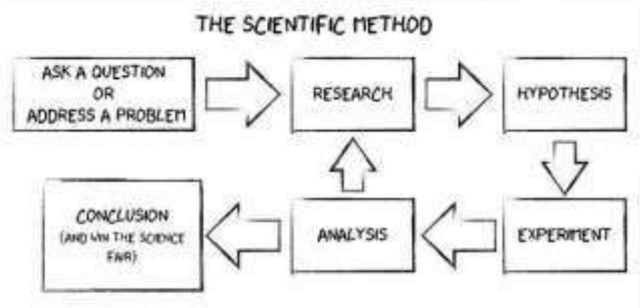
Culture

Typically Senior management buy-in is critical. Successful projects are top-driven

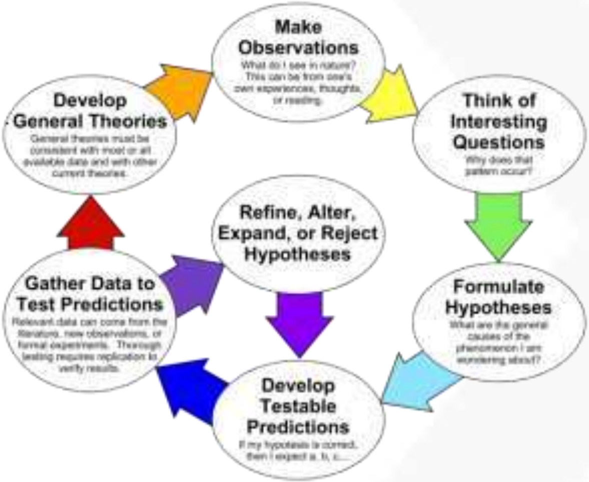
Data Science

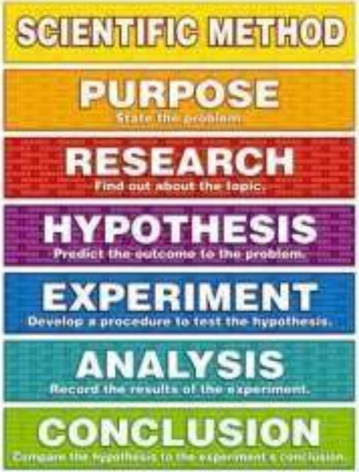
Data Science is Multidisciplinary

- The Scientific Method (wiki)
- Programming
- Databases
- Statistics
- Machine Learning
- Domain Knowledge

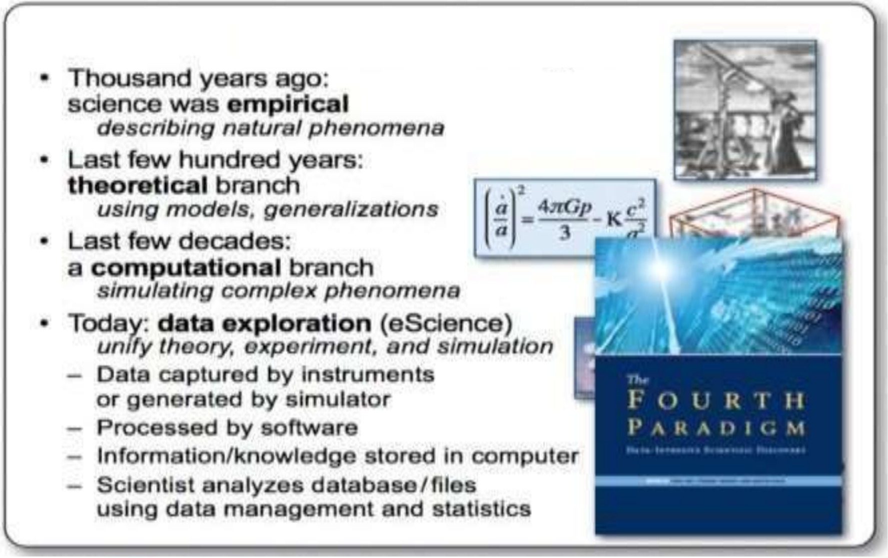

THE SCIENTIFIC METHOD
ASK A QUESTION OR ADDRESS A PROBLEM
RESEARCH
HYPOTHESIS
EXPERIMENT
ANALYSIS
CONCLUSION (AND WIN THE SCIENCE FAIR)

Data Science is Multidisciplinary


Make Observations
What do I see in nature? This can be from one's own experiences, thoughts, or reading.
Think of Interesting Questions
Why does that pattern occur?
Formulate Hypotheses
What are the general causes of the phenomenon I am wondering about?
Develop Testable Predictions
If my hypothesis is correct, then I expect a, b, c....
Gather Data to Test Predictions
Relevant data can come from the literature, new observations, or formal experiments. Thorough testing requires replication to verify results.
Refine, Alter, Expand, or Reject Hypotheses
Develop General Theories
General theories must be consistent with most or all available data and with other current theories.


SCIENTIFIC METHOD
PURPOSE
State the problem.
RESEARCH
Find out about the topic.
HYPOTHESIS
Predict the outcome to the problem.
EXPERIMENT
Develop a procedure to test the hypothesis.
ANALYSIS
Record the results of the experiment.
CONCLUSION
Compare the hypothesis to the experiment's conclusion.

Science Paradigm



- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
  or generated by simulator
  – Processed by software
  – Information/knowledge stored in computer
  – Scientist analyzes database/files
  using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - \mathrm{K}\frac{c^2}{a^2}$$

Why Data Science?

- The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and **ubiquitous data**."

- – Hal Varian

Who is Data Scientist?

“A data scientist… excels at **analyzing data**, particularly large amounts of data, to help a business gain a competitive edge.”

“The analysis of data using the **scientific method**”

“A data scientist is an individual, organization or application that performs statistical analysis, data mining and retrieval processes on a large amount of data to **identify trends, figures and other relevant information**.”

WHO’S A DATA SCIENTIST

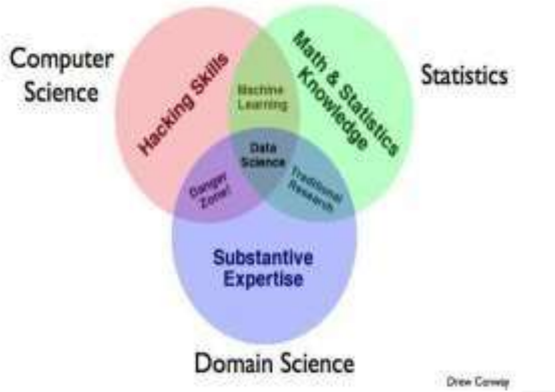
- “A data scientist is someone who knows morestatistics than a computer scientist and more computer sciencethan a statistician.”

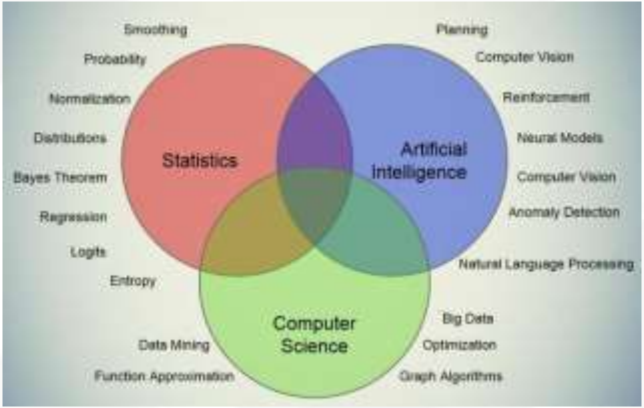
  - Josh Blumenstock

“Data Scientist = statistician + programmer + coach + storyteller + artist”

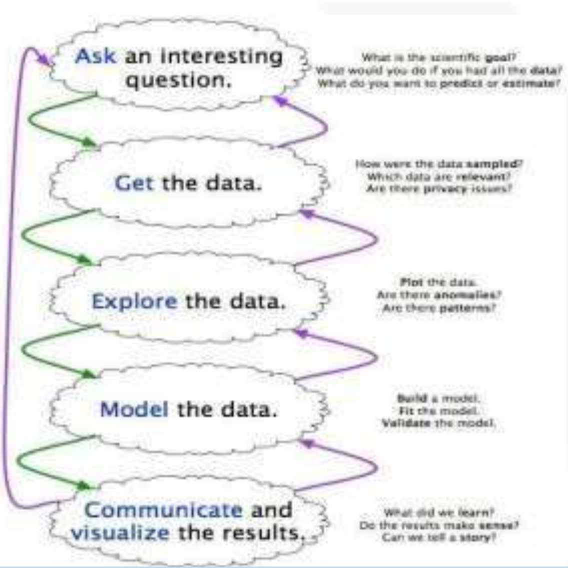
- Shlomo Aragmon

WHO’S A DATA SCIENTIST





Who is Data Scientist?



Ask an interesting question.

What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?

Get the data.

How were the data sampled?
Which data are relevant?
Are there privacy issues?

Explore the data.

Plot the data.
Are there anomalies?
Are there patterns?

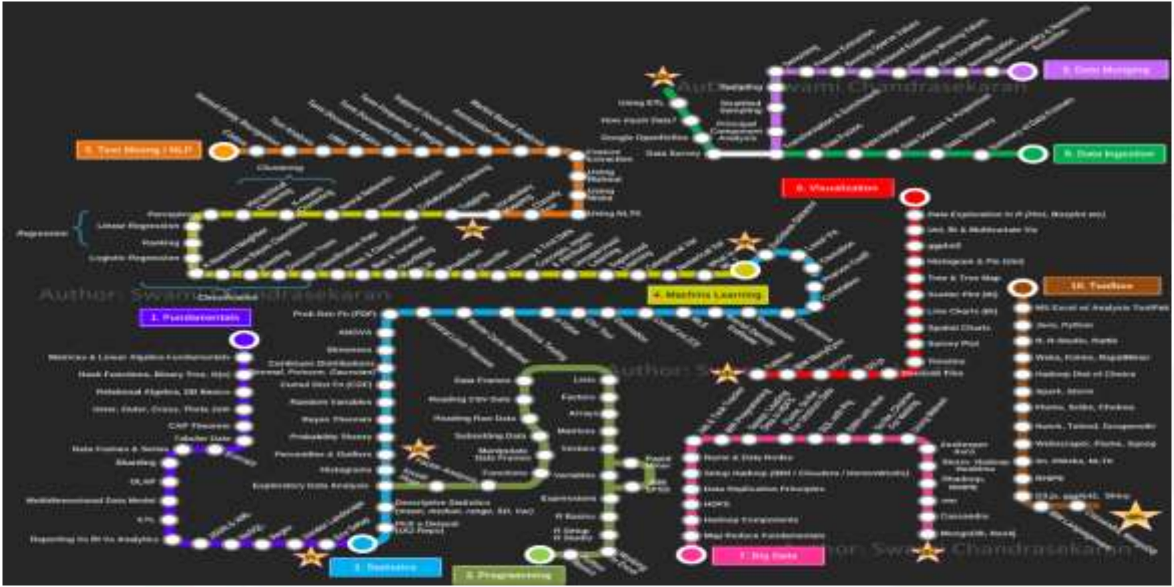
Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we learn?
Do the results make sense?
Can we tell a story?

Who’s a Data Scientist?



What does a Data Scientist Do?

BUILD
DATA
PRODUCTS

tools built with data
to inform decision making

DESCRIPTIVE
PREDICTIVE
PRESCRIPTIVE

OSEMN Things!

Obtain data

Scrub data

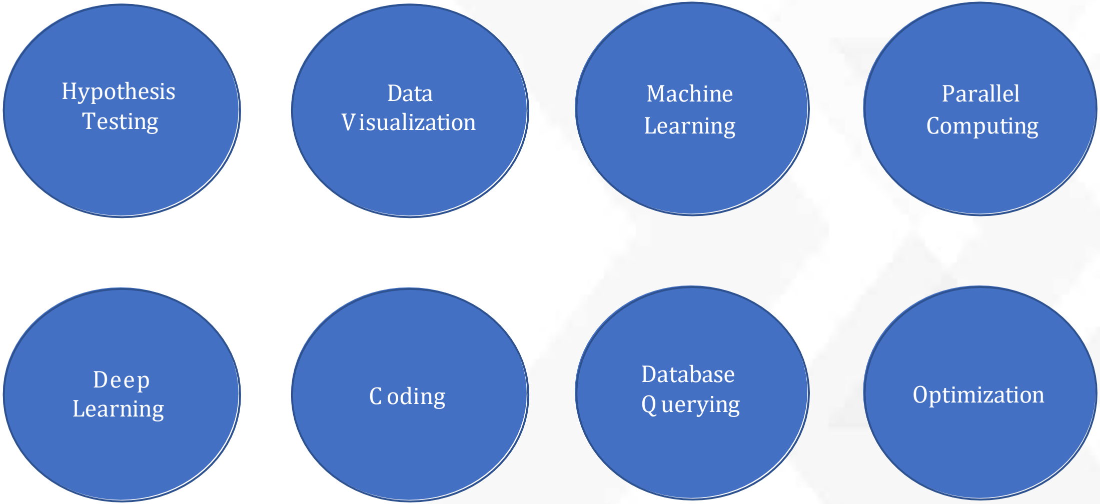
Explore data

Build Models

iNterpret results

Hence the acronym
O-S-E-M-N
(pronounced, ‘awesome’)

.. And This



Hypothesis Testing

Data Visualization

Machine Learning

Parallel Computing

Deep Learning

Coding

Database Querying

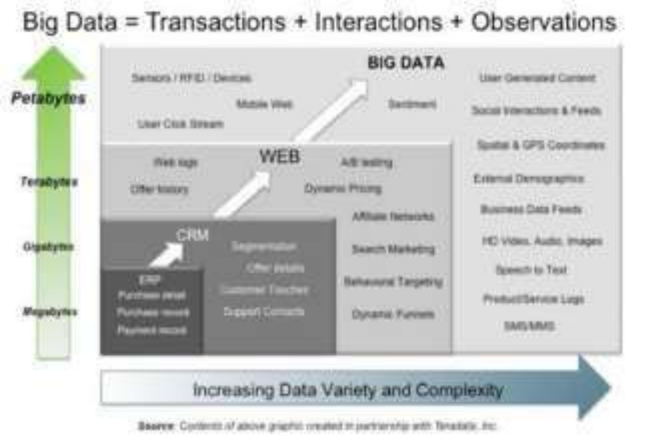
Optimization

Key Concepts

- *use many data sources*
- *understand how the data were collected* (sampling is essential)
- *weightthe data thoughtfully* (not all polls are equally good)
- *use statistical models* (not just hacking around in Excel)
- *understand correlations* (e.g., states that trend similarly)
- *think like a Bayesian, check like a frequentist* (reconciliation)
- *have good communication skills*(What does a 60% probability even mean?
- *visualize, validate, and understand the conclusions*

Common Challenges

- *Big (massive) data* (millions of users, billions of events)
- *curse of dimensionality* (hundreds of variables)
- *missing data* (*not* missing at random)
- *need to avoid overfitting* (test data vs. training data)



Source: Contents of above graphic created in partnership with Teradata, Inc.

Common Tasks

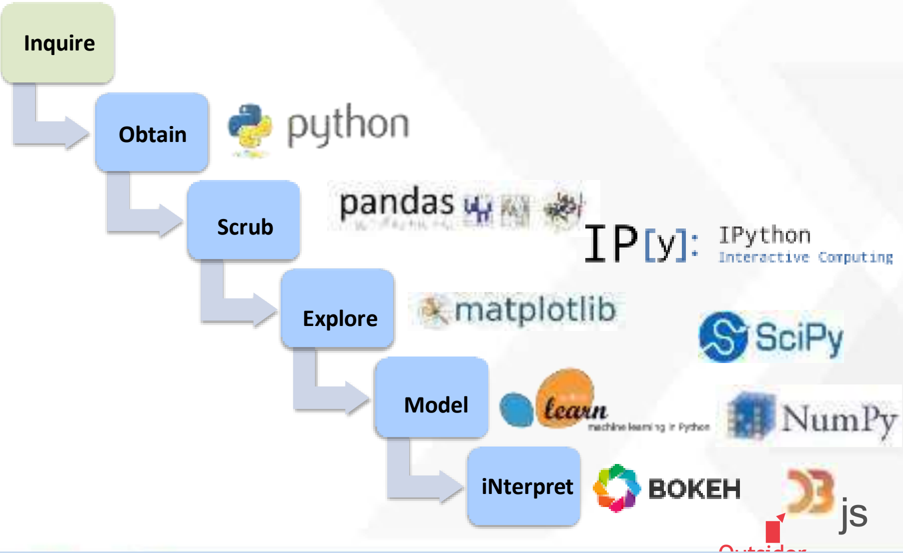
- data munging/scraping/sampling/cleaning in order to get an informative, manageable data set;
- data storage and management in order to be able to access data quickly and reliably during subsequent analysis;
- exploratory data analysis to generate hypotheses and intuition about the data;
- prediction based on statistical tools such as regression, classification, clustering, forecasting and optimization; and
- communication ofresults through visualization, stories, and interpretable summaries.

Tools for the course

Tools for the course - Python

IP[y]: IPython
Interactive Computing
pandas
scikits
learn
machine learning in Python
NumPy
SciPy.org
matplotlib

Python Is IOSEMN



Inquire

Obtain

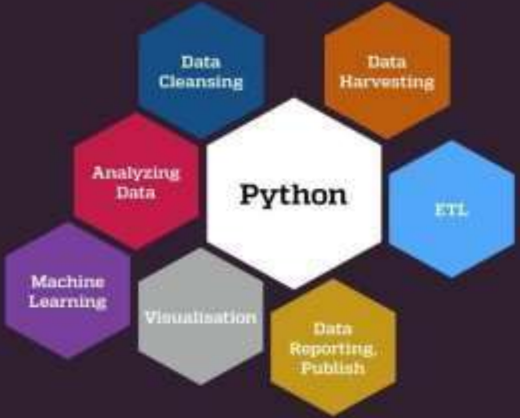
Scrub

Explore

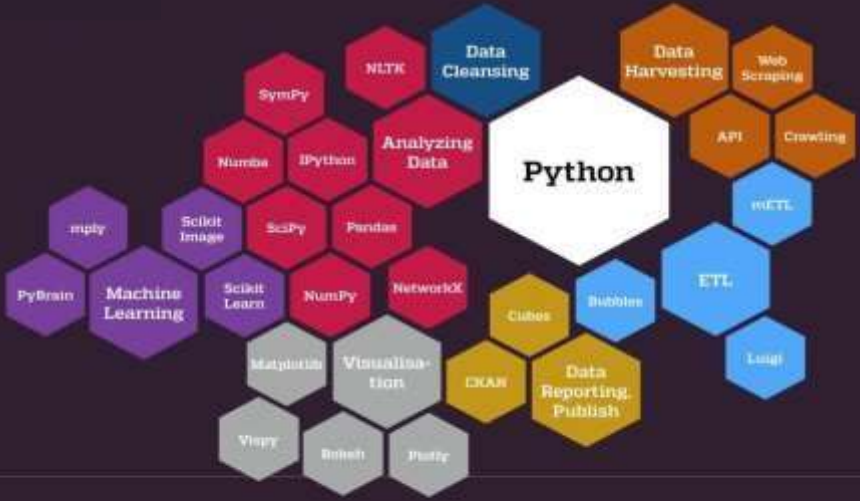
Model

iNterpret

Python Data Science Ecosystem



Python Data Science Ecosystem



Packages - Data Manipulation

NumPy ✎Low level array operations

pandas

- Data tables and in-memory manipulation

Dask

- Parallel out-of-core array manipulation

Blaze

- High level interface for databases and different computational backends

Packages - Modelling

SciPy

- FFTs, integration, other general algorithms

StatsModels

- Statistical distributions and tests

learn

- Machine Learning pipelines

PyMC3

- Bayesian Probabilistic Programming

Packages - Visualisation

matplotlib

- Widely used and powerful plotting package

seaborn

- Opinionated but beautiful data visualisations

Bokeh

- Interactive plotting with server option

plotly

- Graphics API with translation between languages (e.g. Python -> D3)

IPython Notebooks

IP[y]: IPython Interactive Computing

jupyter

Packages - Description

- **NumPy**

  NumPy is a low level library written in C (and FORTRAN) for high level mathematical functions. NumPy cleverl y overcomes the problem of running slower algorithms on Python by using multidimensional arra ys and functions that operate on arra ys . Any algorithm can then be expressed as a function on arra ys , allowing the algorithms to be run quickly.

  NumPy is part of the Sci Py project, and is released as a separate library so people who only need the basic requirements can use it without installing the rest of Sci Py.

  NumPy is compatible with Python versions 2.4 through to 2.7.2 and 3.1+

- **SciPy**

  SciPy is a library that uses NumPy for more mathematical functions. Sci Py uses NumPy arra ys as the basic data structure, and comes with modules for va ri ous commonly used tasks in scientific programming, including linear algebra, integration (calculus), ordinary differential equation solvingand signal processing.

- **Numba**

  Numba is a NumPy aware Python compiler (just-in-time (JIT) specializing compiler) which compiles annotated Python (and NumPy) code to LLVM (Low Level Virtual Machine) through special decorators. Briefl y, Numba uses a sys tem that compiles Python code with LLVM to code which can be nativelyexecuted at runtime.

Packages - Description

- **scikit-learn**

  scikit-learn is a Python module for machine learning built on top of SciPy and distributed under the 3-Clause BSD license.

- **Pandas**

  Pandas is data manipulation library based on Numpy which provides many useful functions for accessing, indexing, merging and grouping data easily. The main data structure (DataFrame) is close to what could be found in the R statistical package; that is, heterogeneous data tables with name indexing, time series operations and auto-alignment of data.

- **Matplotlib**

  Matplotlib is a flexible plotting library for creating interactive 2D and 3D plots that can also be saved as manuscript-quality figures. The API in many ways reflects that of MATLAB, easing transition of MATLAB users to Python. Many examples, along with the source code to re-create them, are available in the matplotlib gallery.

Packages - Description

- **Rpy2**

  Rpy2 is a Python binding for the R statistical package allowing the execution of R functions from Python and passing data back and forth between the two environments. Rpy2 is the object oriented implementation of the Rpy bindings.

- **PsycoPy**

  PsychoPy is a library for cognitive scientists allowing the creation of cognitive psychology and neuroscience experiments. The library handles presentation of stimuli, scripting of experimental design and data collection.

Packages - Description

- **datetime (or) time**

  Date and time functions to manage date and time data
- **math**

  Core math functions and the constants likepi, e etc.
- **pickle**

  Serializes objects to file
- **os (or) os.path**

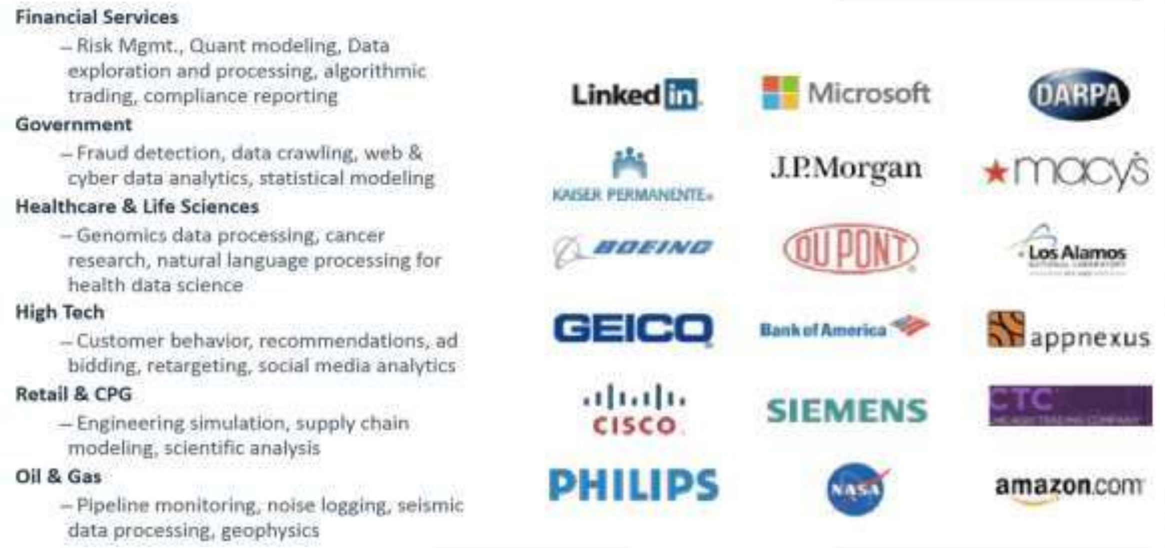
  Operatingsystem interfaces.
- **re**

  A library of perl-like regularexpression operations
- **string**

  Useful constants and classes related to strings.
- **sys**

  System parametersand functions

Who is using Python?



Financial Services

- Risk Mgmt., Quant modeling, Data exploration and processing, algorithmic trading, compliance reporting

Government

- Fraud detection, data crawling, web & cyber data analytics, statistical modeling

Healthcare & Life Sciences

- Genomics data processing, cancer research, natural language processing for health data science

High Tech

- Customer behavior, recommendations, ad bidding, retargeting, social media analytics

Retail & CPG

- Engineering simulation, supply chain modeling, scientific analysis

Oil & Gas

- Pipeline monitoring, noise logging, seismic data processing, geophysics

Why should I become a Data Scientist?

DEMAND & SUPPLY

"We projecta needfor 1.5 millionadditional managers andanalysts in the United States who can ask the right questions and consumethe results of the analysis of Big Data effectively."

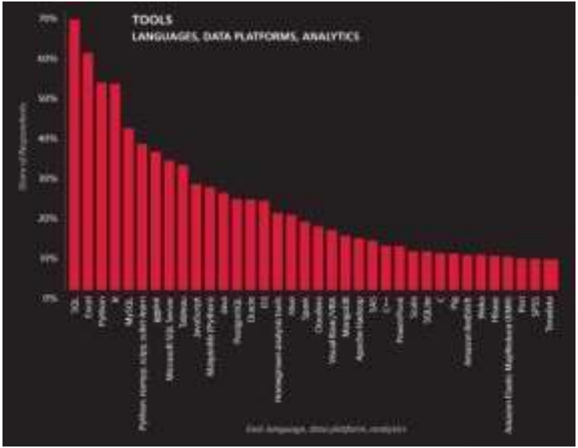
"A significantconstraint onrealizing value from BigData will be a shortage of talent, particularly of people with deepexpertise in statistics and machine learning, and the managers andanalysts whoknowhowto operate companies by using insights from Big Data."

Big data: The next frontier for innovation, competition, and productivity, McKinsey report

"By 2018 the United States will experience a shortage of 190,000 skilleddata scientists, and 1.5 million managers andanalysts capable of reaping actionableinsights fromthe big data deluge."

Game changers: Five opportunities for US growth and renewal,McKinsey report

OK. How so do I become a Data Scientist?



Read books on

- Statistics
- Machine Learning
- Programming
- Databases

Take University courses

Apply for internships to workon real-life projects

Spend hours debugging on

StackOverflow

Participate in Data Hackathons/Data Driven competitions

What is Python?

- Programming language
- You write instructions to the computer
- Python “interpreter” runs those instructions

Why python?

- It’s awesome and popular!
- Free and Open Source language.
- Readablesyntax.
- Great for interactive work
- Easyto learn and has an active community.
- Large amount of libraries.
- High level, general purpose.
- Backed up with fast C & Fortran numerical libraries

Python - Applications

Python is a powerful multi-paradigm computer programming language. With Python, we can do many things. Beloware some of the things that canbe achieved using Python.

- ✓ **Systems Programming:** Python's built-in interfaces to operating-system services make it ideal for writing portable, maintainable system-administration tools and utilities (sometimes called shell tools). Python programs can search files and directory trees, etc.

- ✓ **GUIs:** Python's simplicity and rapid turnaround makesit a good match for graphical user interface programming on the desktop. Python comes with a standardobject-oriented interface to the Tk GUI API called tkinter (Tkinter in 2.X) that allows Python programs to implement portable GUIswith a native look and feel.

- ✓ **Internet Scripting:** Python comes with standard Internet modules that allow Python programs to performa wide varietyof networking tasks in client and server modes.

- ✓ **Database Programming:** For traditional database demands, there are Python interfaces to all commonly used relationaldatabase systems like Sybase, Oracle, Informix, ODBC, MySQL, PostgreSQL, SQLite, and more.