INTRODUCTION & CLASS OVERVIEW

Prof. Mohammad Hajiaghayi & Arefeh Nasr
Wiki & Linkedin: @Mohammad Hajiaghayi
Twitter:@MTHajiaghayi
YouTube:@hajiaghayi [PLEASE SUBSCRIBE]
Instagram:@mhajiaghayi
Original slides prepared by Prof. Amol Deshpande

Lectures #1 and #2

DATA/MSML602: Principles of Data Science
TuTh6:00pm – 8:30pm

TODAY'S CLASS

- What is Data Science?
  - Data Lifecycle
  - Where is it headed?
- Introductions
- DATA/MSML602 Details
- Few Data Science Success Stories and Cautionary Tales
- Basic Technology Stack and Best Practices
  - Python, Jupyter Notebook, GitHub
  - Cloud Computing, Containers (e.g., Docker)

WHO AM I?

- PhD, MIT 2005
- Professor at UMD since 2010
- Research Interests: (BIG Data) Algorithms, Game Theory, Social Networks, Cloud Computing, Machine Learning, System Design
- Major Awards: ACM Fellow (2018), Guggenheim Fellow (2019), IEEE Fellow (2020), EATCS Fellow (2020), Blavatnik Honoree (2020) Received ONR Young Investigator Award, NSF CAREER Award, European Theoretical Computer Science Award, ACM ICPC Programming Coach Award, Silver Medal IOI'97.
- Industry Experience: 12 years of industry experience at IBM, Microsoft, AT&T, Google, Amazon, Overstock and consulting activities with Uber, Lyft, and Airbnb

MOTIVATION

- Explosion of data, in pretty much every domain
  - Sensing devices and sensor networks that can monitor everything 24/7 from temperature to pollution to vital signs
  - Increasingly sophisticated smart phones
  - Internet, social networks makes it easy to publish data
  - Scientific experiments and simulations ➔ astronomical data volumes
  - Internet of Things
  - Dataification: taking all aspects of life and turning them into data (e.g., what you like/enjoy has been turned into a stream of your "likes")
- How to handle that data? How to extract interesting actionable insights and scientific knowledge?
- Data volumes expected to get much worse

FOUR V'S OF BIG DATA

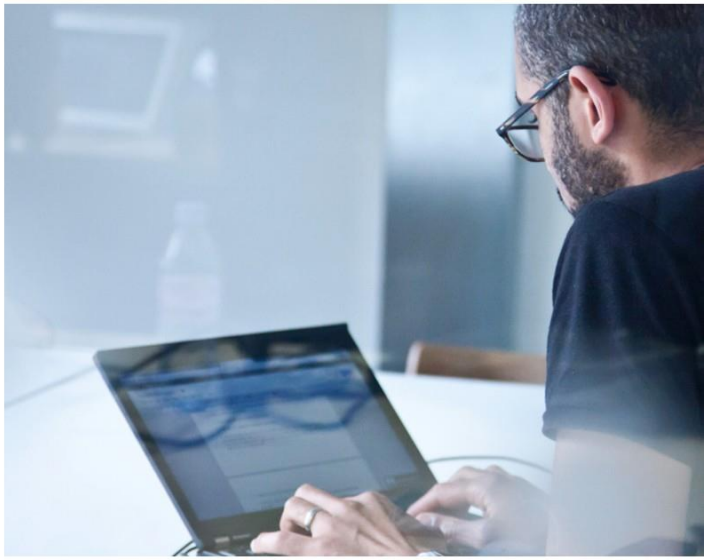
- Increasing data Volumes
  - Scientific data: 1.5GB per genome -- can be sequenced in .5 hrs
  - 500M tweets per day (as of 2013)
  - As of 2012: 2.5 Exabytes of data created every day
- Variety:
  - Structured data, spreadsheets, photos, videos, natural text, …
- Velocity
  - Sensors everywhere -- can generate high-rate "data streams"
  - Real-time analytics requires data to be consumed as fast as it is generated
- Veracity
  - How do you decide what to trust? How to remove noise? How to fill in missing values?

BIG DATA/DATA SCIENCE TO THE RESCUE

- Terms increasingly used synonymously: also data analytics, data mining, business intelligence
  - "Data science is the application of computational and statistical techniques to address or gain [managerial or scientific] insight into some problem in the real world"; Zico Kolter; CMU
- Data scientist called the sexiest job of the 21st century
  - But the term has becoming very muddled at this point

BUSINESS INSIDER

1. Data scientist



Shutterstock

Overall job score (out of 5.0): 4.8

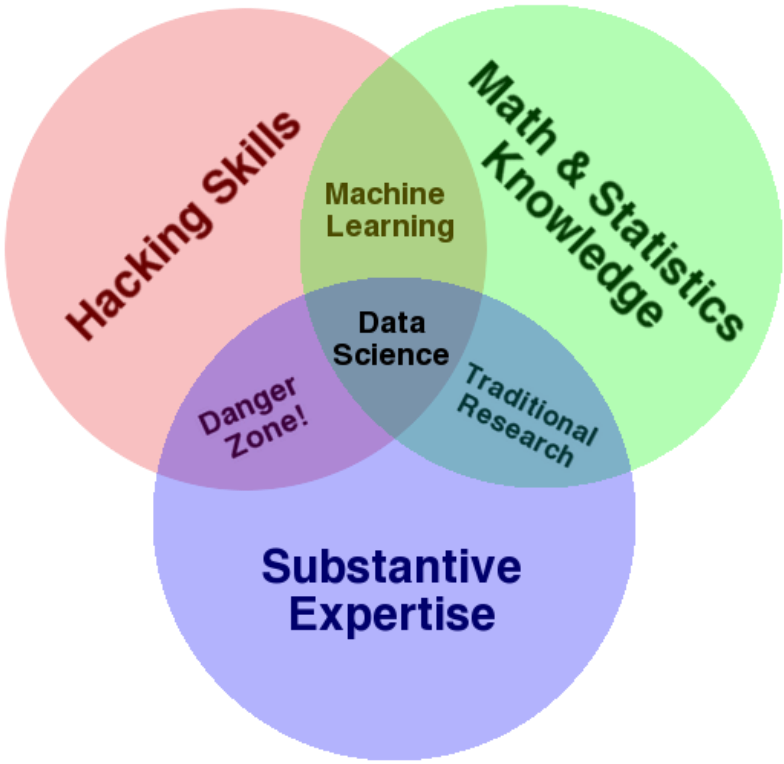
Job satisfaction rating (out of 5.0): 4.4

Number of job openings: 4,184

Median base pay: $110,000

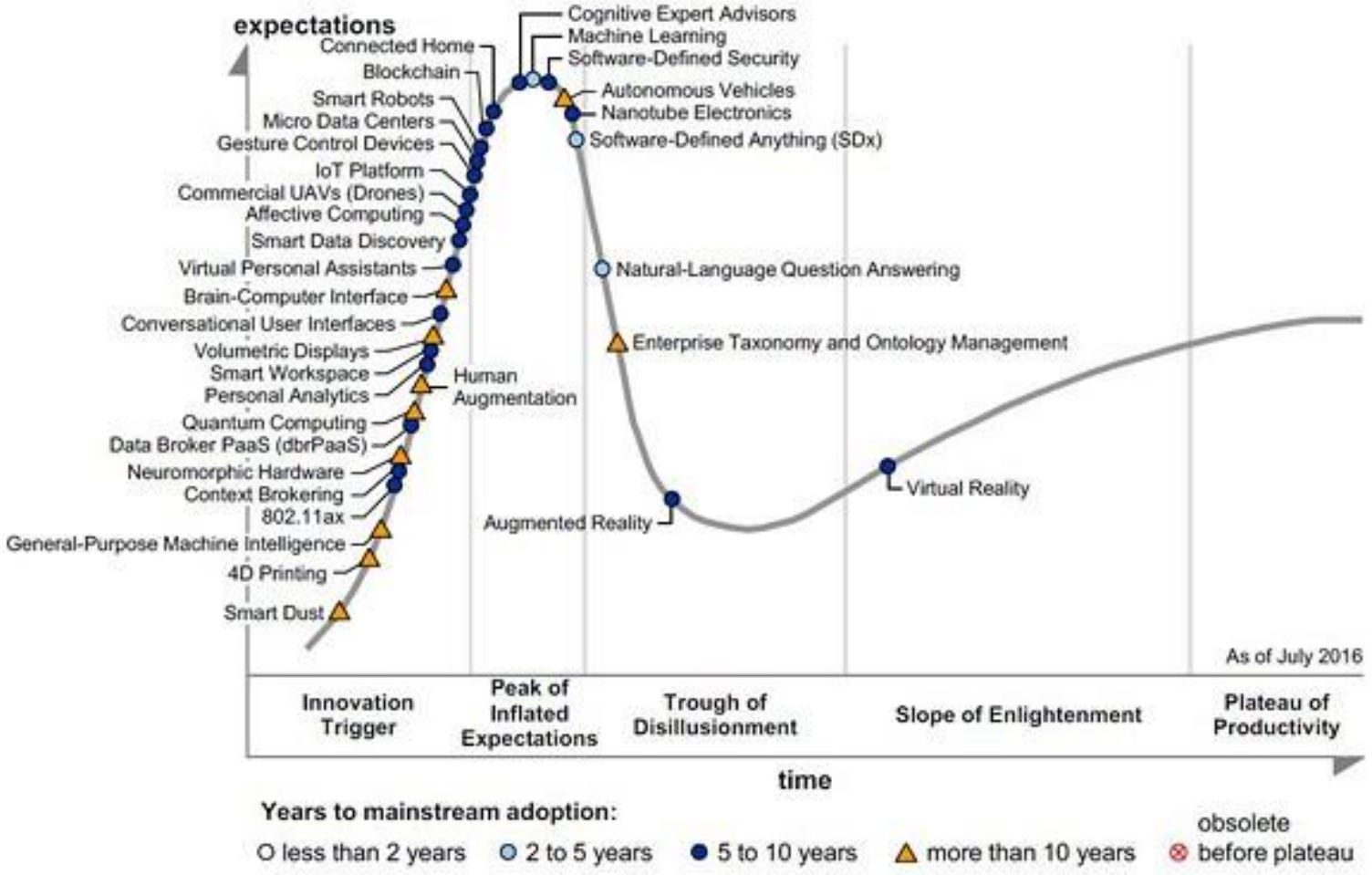
MANY DEFINITIONS

- Broad: necessarily larger than a single discipline
- Interdisciplinary: statistics, computer science, operations research, statistical and machine learning, data warehousing, visualization, mathematics, information science, …
- Insight-focused: grounded in the desire to find insights in data and leverage them to inform decision making



Drew Conway
CEO, Alluvium
(analytics company)

ALL HYPE?



Source: Gartner (July 2016)

ALL HYPE?

- How Big Data Will Change …
  - Same goes for Machine Learning
- Extracting insights and knowledge from data very important, and will continue to increase in importance
  - Revolutionizing things in Education, Food Supply, Disease Epidemics, ..
  - But: it is not much different from what we, especially statisticians, have been doing for many years
- What is different?
  - Much more data is digitally available than was before
  - Inexpensive computing + Cloud + Easy-to-use programming frameworks = Much easier to analyze it
  - Often: large-scale data + simple algorithms > small data + complex algorithms
    - Changes how you do analysis dramatically

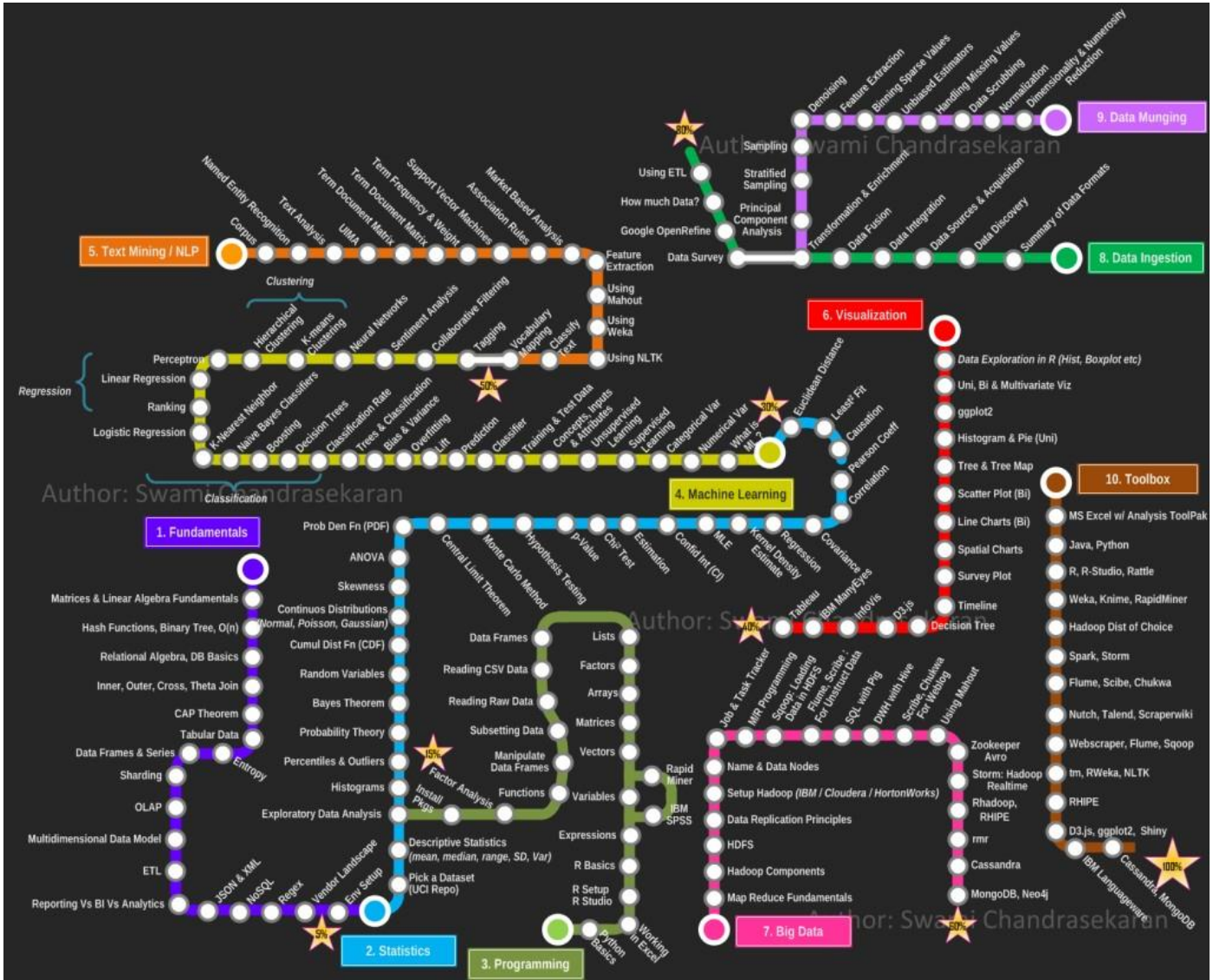
SOME KEY SHITS

- From: Rise of Big Data
- Curated, clean, small samples --> large, uncurated, messy datasets
  - Statistics based on small, carefully collected random samples
  - Costly; hard to do fine-grained analysis; careful planning
  - Today: Collect huge volume of data, feed it into algorithms, usually the signal is strong enough to overcome the noise
- From Causation to Correlation
  - Goal of analysis often to figure out what caused what
  - Causation very hard to figure out
  - Today: give up causation for correlation -- we can find out two things are correlated and often that's enough
- "Dataification": Process of converting abstract things into concrete data
  - e.g., what you like represented as a stream of your likes; your "sitting posture" captured using 100's of sensors placed in a car seat

"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—that'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."

Hal Varian
Chief Economist at Google

WHAT A DATA SCIENTIST SHOULD KNOW?



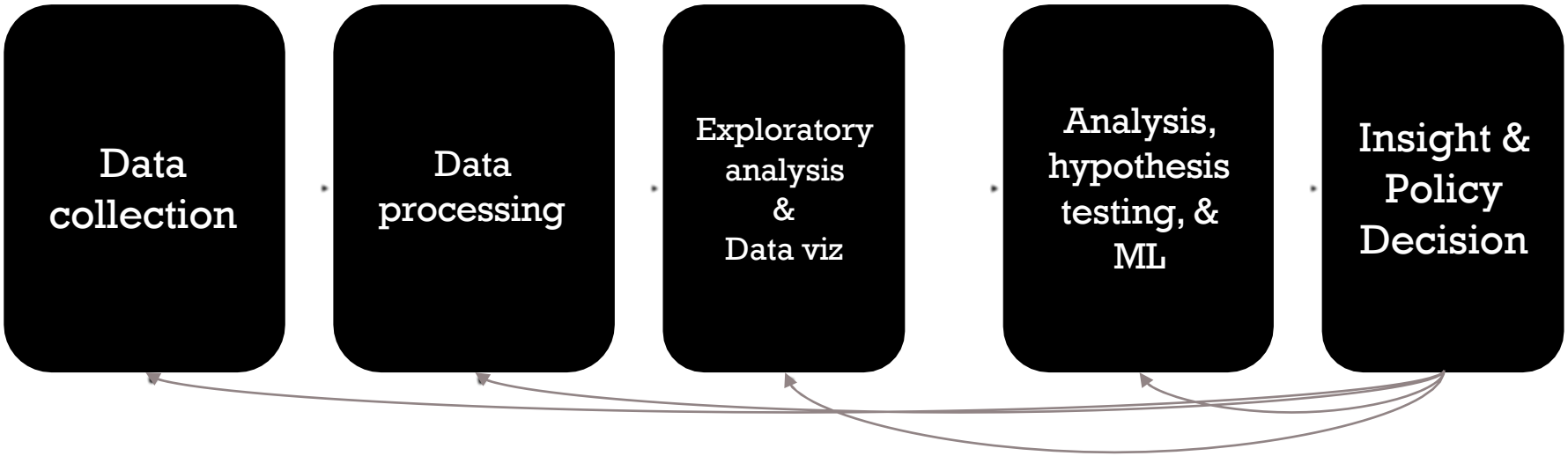
WHAT A DATA SCIENTIST SHOULD KNOW?

- From: how to hire a data scientist
- Data grappling/wrangling skills: how to move data around and manipulate it with some programming langauge
  - Scripting languages like Python, Ruby; Data storage tools like relational databases, key-value Stores; Programming frameworks like SQL, Hadoop, etc.
- Data viz experience: how to draw informative pictures of data
  - Many tools (e.g., D3.js, plotting libraries)
  - Harder question is knowing what to draw
- Knowledge of stats, errorbars, confidence intervals
  - Python libraries; Matlab; R
- Experience with forecasting and prediction, both general and specific
  - Basic Machine Learning techniques
- Great communication skills: to communicate the findings

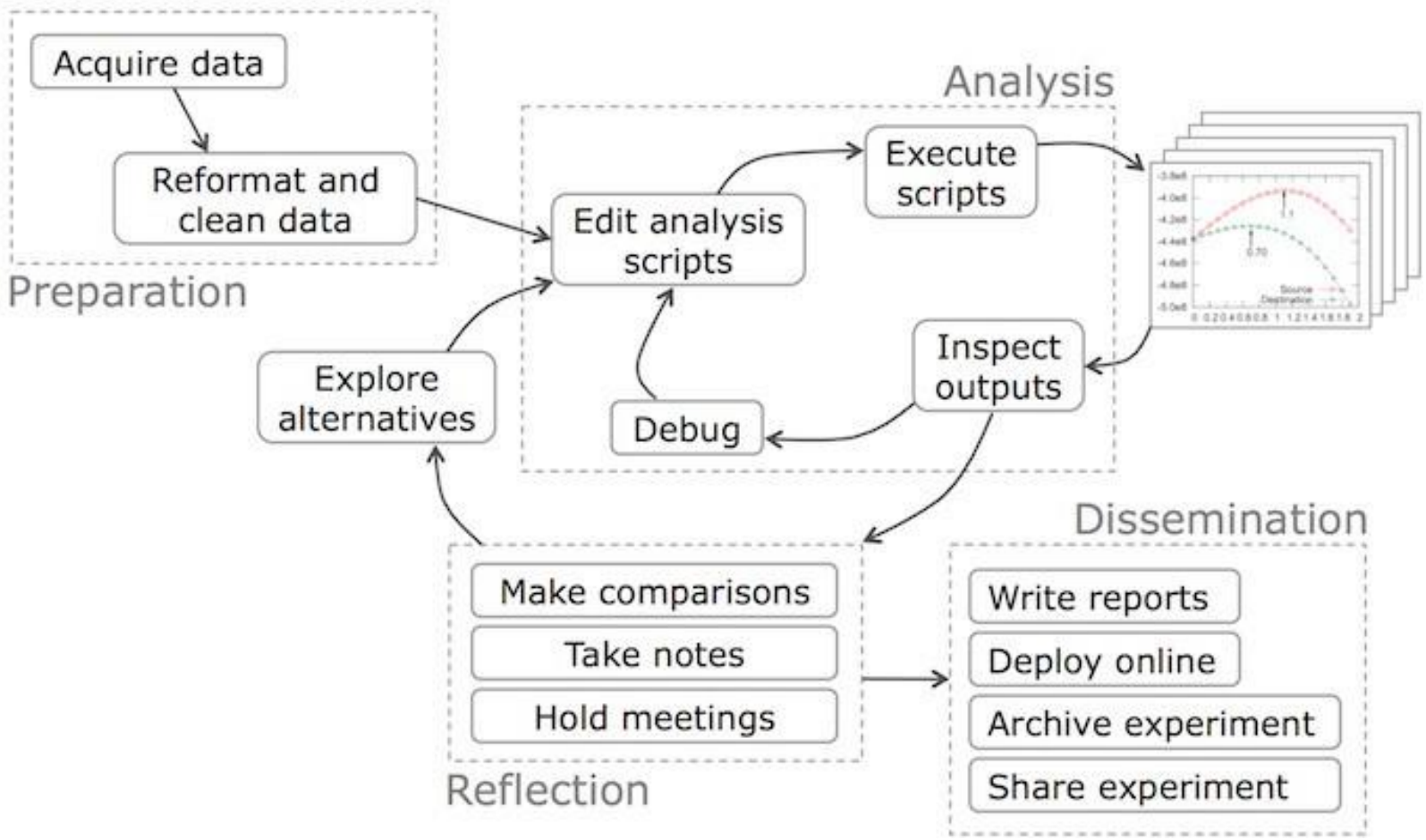
TODAY'S CLASS

- What is Data Science?
  - Data Lifecycle
  - Where is it headed?
- Introductions
- DATA/MSML602 Details
- Few Data Science Success Stories and Cautionary Tales
- Basic Technology Stack and Best Practices
  - Python, Jupyter Notebook, GitHub
  - Cloud Computing, Containers (e.g., Docker)

THE DATA LIFECYCLE



TYPICAL DATA SCIENCE WORKFLOW



Source: Data Science Workflow; Phillip Guo

WHERE DATA SCIENTIST SPENDS MOST TIME?

- 'Janitor Work' in Data Science; Research Directions in Data Wrangling
  - Estimates that 80-90% of the work is in data cleaning and wrangling

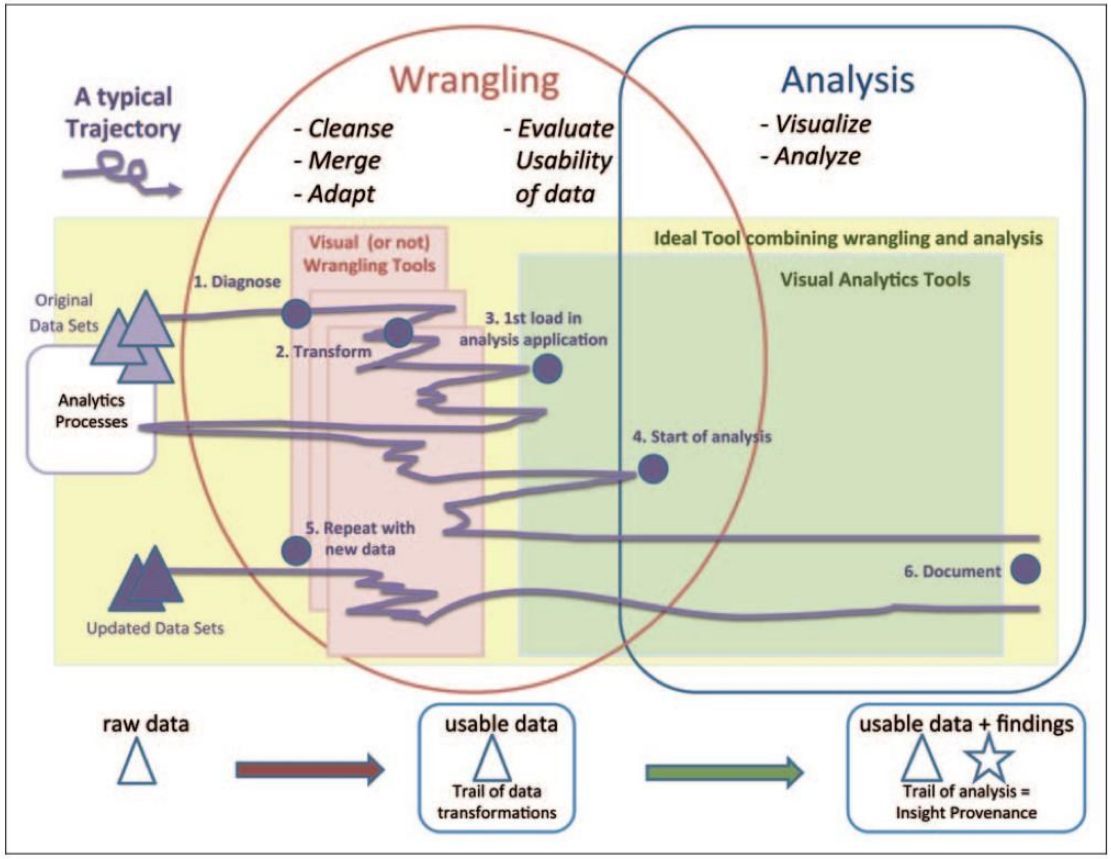


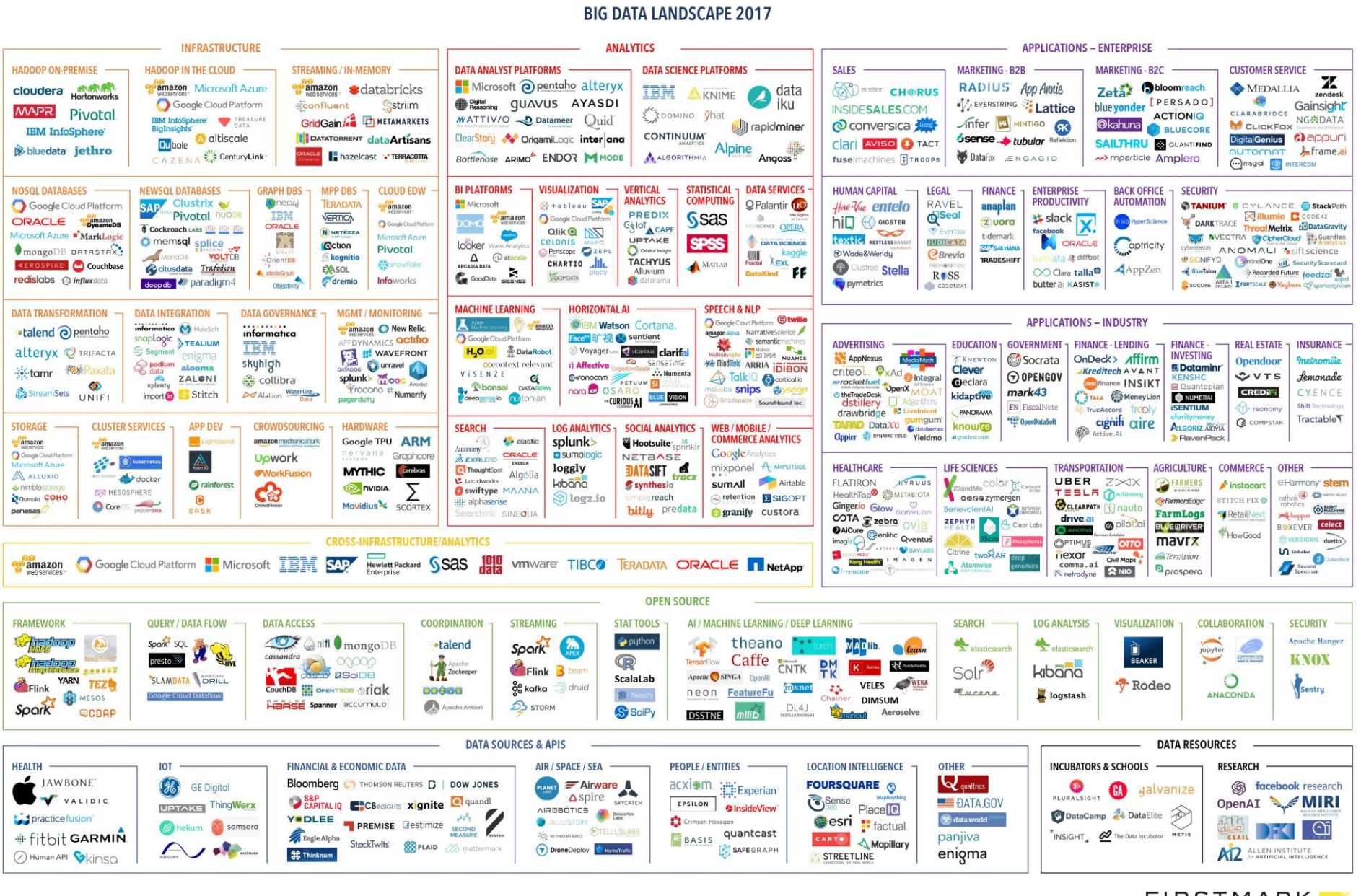
Figure 1. The iterative process of wrangling and analysis. One or more initial data sets may be used and new versions may come later. The wrangling and analysis phases overlap. While wrangling tools tend to be separated from the visual analysis tools, the ideal system would provide integrated tools (light yellow). The purple line illustrates a typical iterative process with multiple back and forth steps. Much wrangling may need to take place before the data can be loaded within visualization and analysis tools, which typically immediately reveals new problems with the data. Wrangling might take place at all the stages of analysis as users sort out interesting insights from dirty data, or new data become available or needed. At the bottom we illustrate how the data evolves from raw data to usable data that leads to new insights.

TODAY'S CLASS

- What is Data Science?
  - Data Lifecycle
  - Where is it headed?
- Introductions
- DATA/MSML602 Details
- Few Data Science Success Stories and Cautionary Tales
- Basic Technology Stack and Best Practices
  - Python, Jupyter Notebook, GitHub
  - Cloud Computing, Containers (e.g., Docker)

WHERE ARE THINGS HEADED?

- Toolkits have matured dramatically over the last 5 years
  - Big data analysis frameworks (Spark, Flink, …)
  - Machine learning toolkits (scikit, R, Tensorflow, …)
  - Cloud computing; Virtualization; "Containerization"
- Easy to use complex ML techniques at very large scale
- "Big Data provides pipes; AI provides the smarts" (Matt Turck)
  - With Cloud providing the third leg in most cases
- Enterprise data increasingly moving to cloud
- Overwhelming number of tools out there
  - Ultra-specialization
  - Not sustainable



Matt Turck, Jim Hao and FirstMark Capital, http://mattturck.com/bigdata2017/

WHERE ARE THINGS HEADED?

- Acquiring ”good” data (veracity) remains a big challenge
- “Interpretability”, “ethics”,“accountability”,“privacy” starting to come to the forefront
- Trends:
  - Real-time analytics becoming more crucial
  - Data virtualization/integration across many sources starting to become mainstream
  - Data Governance and Security gaining more importance
  - SQL is officially back !
  - Rise of “data engineers”
- Ability to “understand” what you are doing going to be much more critical

TODAY'S CLASS

- What is Data Science?
  - Data Lifecycle
  - Where is it headed?
- Introductions
- DATA/MSML602 Details
- Few Data Science Success Stories and Cautionary Tales
- Basic Technology Stack and Best Practices
  - Python, Jupyter Notebook, GitHub
  - Cloud Computing, Containers (e.g., Docker)

THIS MASTER PROGRAM

- You'll learn to take data:
  - Process it
  - Visualize it
  - Understand it
  - Communicate it
  - Extract value from it

THIS MASTER PROGRAM

- DATA 6XX: Principles of Data Science
  - An introduction to the data science pipeline, i.e., the end-to-end process of going from unstructured, messy data to knowledge and actionable insights. Provides a broad overview of what data science means and systems and tools commonly used for data science, and illustrates the principles of data science through several case studies.
- DATA 6XX: Big Data Systems
- DATA 6XX: Machine Learning and Data Mining
- DATA 6XX: Algorithms for Data Science

THIS MASTER PROGRAMcc

- DATA 6XX: Principles of Data Science
  - An introduction to the data science pipeline, i.e., the end-to-end process of going from unstructured, messy data to knowledge and actionable insights. Provides a broad overview of what data science means and systems and tools commonly used for data science, and illustrates the principles of data science through several case studies.
- DATA 6XX: Big Data Systems
- DATA 6XX: Machine Learning and Data Mining
- DATA 6XX: Algorithms for Data Science

THIS MASTER PROGRAM

- DATA 6XX: Principles of Data Science

- DATA 6XX: Big Data Systems
  - An overview of data management systems for performing data science on large volumes of data, including relational databases, and NoSQL systems. The topics covered include: different types of data management systems, their pros and cons, how and when to use those systems, and best practices for data modeling.
- DATA 6XX: Machine Learning and Data Mining

- DATA 6XX: Algorithms for Data Science

THIS MASTER PROGRAM

- DATA 6XX: Principles of Data Science
- DATA 6XX: Big Data Systems
- DATA 6XX: Machine Learning and Data Mining
  - Provides a broad overview of key machine learning and data mining algorithms, and how to apply those to very large datasets. Topics covered include decision trees, linear models for classification and regression, support vector machines, neural networks and deep learning, online learning, recommendation systems, clustering and dimensionality reduction, and systems for large-scale machine learning.
- DATA 6XX: Algorithms for Data Science

THIS MASTER PROGRAM

- DATA 6XX: Principles of Data Science
- DATA 6XX: Big Data Systems
- DATA 6XX: Machine Learning and Data Mining
- DATA 6XX: Algorithms for Data Science
  - Provides an in-depth understanding of some of the key data structures and algorithms essential for advanced data science. Topics include random sampling, graph algorithms, network science, data streams, and optimization.

THIS COURSE

- End-to-end data science lifecycle
- Acquiring, wrangling, cleaning, and integrating data; Setting up pipelines for ETL
- Data modeling
- Information Visualization
- Ethics, Privacy, and Reproducibility
- Feel free to tell me if there are topics that you think we should cover…

PREREQUISITE KNOWLEDGE

- Accessible to anyone with programming experience and mathematical maturity
- We do not assume:
  - Experience with Python, pandas, scikit-learn, matplotlib, etc …
  - Deep statistics or any ML knowledge
  - Database or distributed systems knowledge
- We do assume:
  - You want to be here!

TODAY'S CLASS

- What is Data Science?
  - Data Lifecycle
  - Where is it headed?
- Introductions
- DATA/MSML602 Details
- Few Data Science Success Stories and Cautionary Tales
- Basic Technology Stack and Best Practices
  - Python, Jupyter Notebook, GitHub
  - Cloud Computing, Containers (e.g., Docker)

TODAY'S CLASS

- What is Data Science?
  - Data Lifecycle
  - Where is it headed?
- Introductions
- DATA/MSML602 Details (Course Agenda)
- Few Data Science Success Stories and Cautionary Tales
- Basic Technology Stack and Best Practices
  - Python, Jupyter Notebook, GitHub
  - Cloud Computing, Containers (e.g., Docker)

COURSE STRUCTURE

- First 3 lectures: intro & primers, basic stats
- Next 6-7 lectures: data collection & management
  - Data modeling, data wrangling, data cleaning
  - Data integration, information extraction
  - Different types of data management systems
  - Setting up data science pipelines; deploying models
- Next 2-3 lectures: information visualization
- Next 2 lectures: data science ethics, reproducibility, …

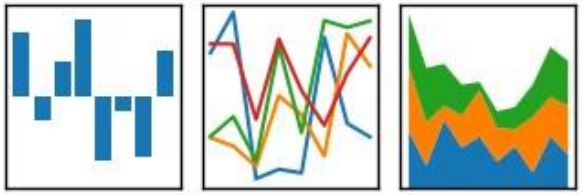
SOME TECHNOLOGIES WE WILL USE

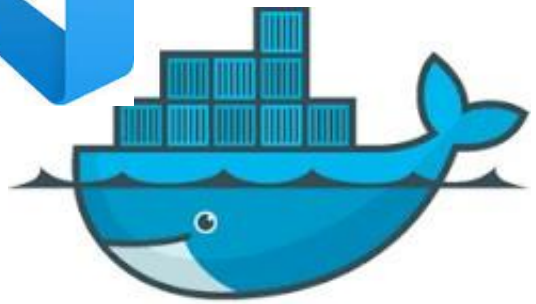
python™

ANACONDA
Powered by Continuum Analytics

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



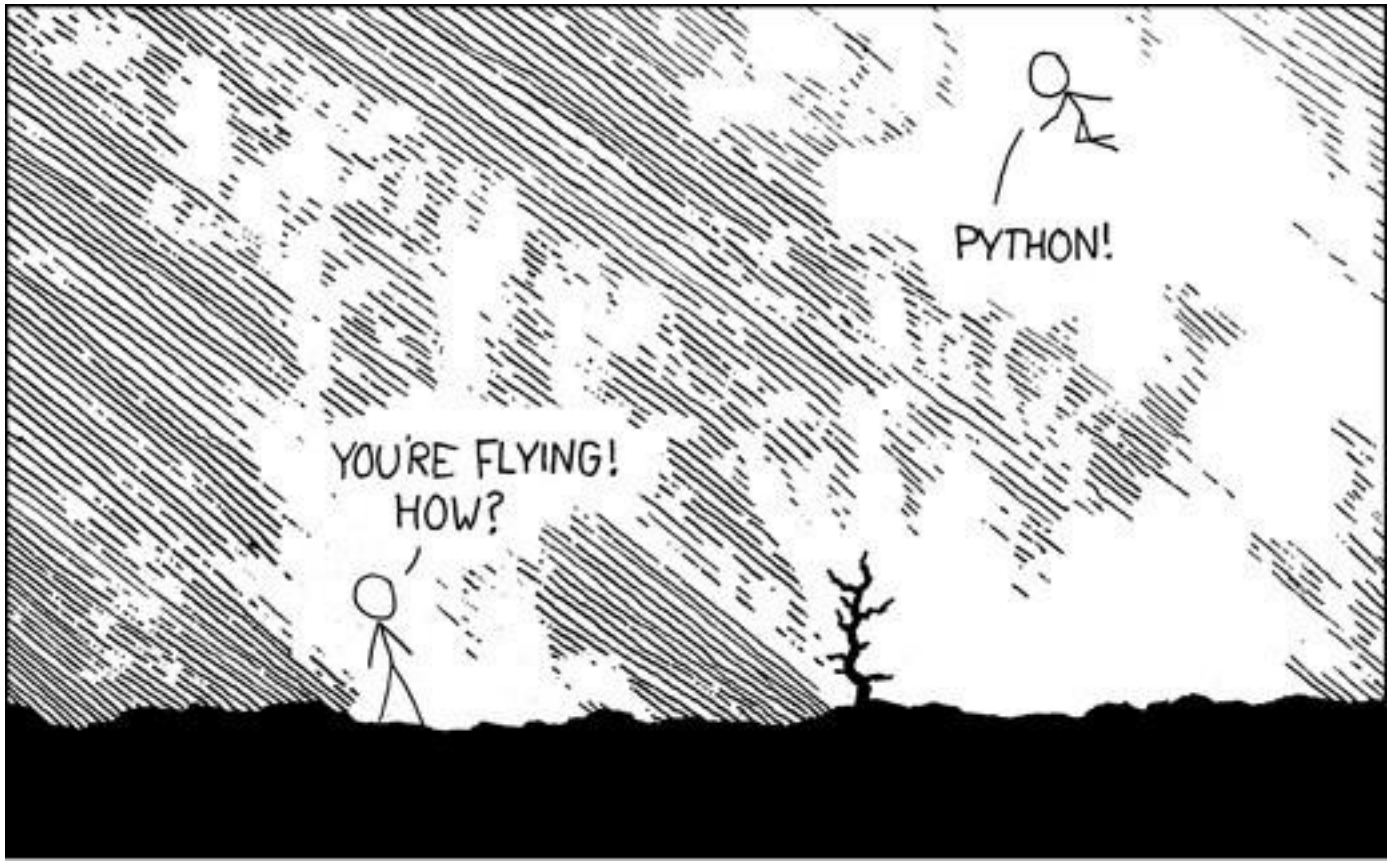


docker

jupyter

scikit learn

C++20


PYTHON!
YOU'RE FLYING!
HOW?

IMPORTANT WALLS OF TEXT

Common Sense!

ANTI-HARASSMENT

(Adapted from ACM SIGCOMM's policies)

- The open exchange of ideas and the freedom of thought and expression are central to our aims and goals. These require an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, we are dedicated to providing a harassment-free experience for participants in (and out) of this class.

- Harassment is unwelcome or hostile behavior, including speech that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation, in a conference, event or program.

Common Sense!

ACADEMIC INTEGRITY

(Text unironically stolen from Hal Daumé III)

- Any assignment or exam that is handed in must be your own work (unless otherwise stated). However, talking with one another to understand the material better is strongly encouraged. Recognizing the distinction between cheating and cooperation is very important. If you copy someone else's solution, you are cheating. If you let someone else copy your solution, you are cheating (this includes *posting solutions online in a public place)*. If someone dictates a solution to you, you are cheating.

- Everything you hand in must be in your own words, and based on your own understanding of the solution. If someone helps you understand the problem during a high-level discussion, you are not cheating. We strongly encourage students to help one another understand the material presented in class, in the book, and general issues relevant to the assignments. When taking an exam, you must work independently. Any collaboration during an exam will be considered cheating. Any student who is caught cheating will be given an F in the course and referred to the University Office of Student Conduct. Please don't take that chance – if you're having trouble understanding the material, please let me know and I will be more than happy to help.

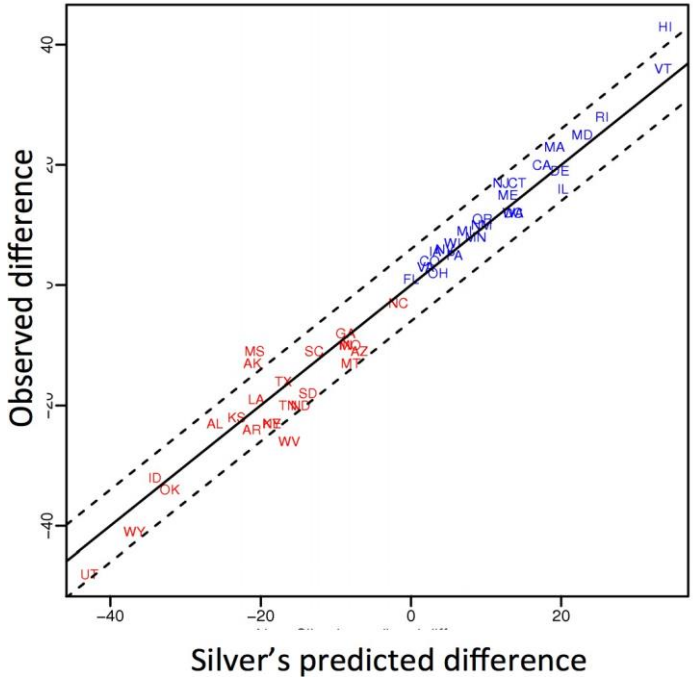
(A FEW) DATA SCIENCE SUCCESS STORIES & CAUTIONARY TALES

POLLING: 2008 & 2012

- Nate Silver uses a simple idea – taking a principled approach to aggregating polling instead of relying on punditry – and:
  - Predicts 49/50 states in 2008
  - Predicts 50/50 states in 2012

  - (He is also a great case study in creating a brand.)

https://hbr.org/2012/11/how-nate-silver-won-the-2012-p



Democrat (+) or Republican (-) in 2012

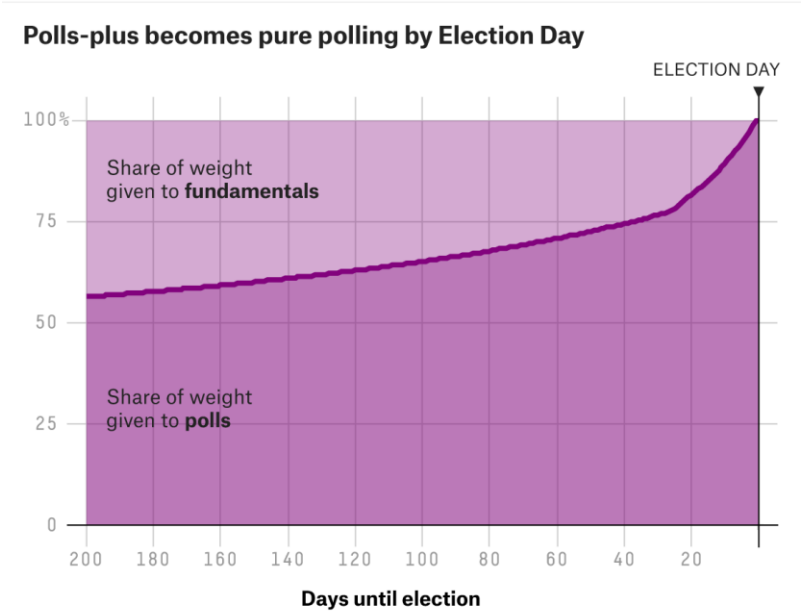
POLLING: 2016

POLITICS

Nate Silver Is Unskewing Polls — All Of Them — In Trump's Direction

The vaunted 538 election forecaster is putting his thumb on the scales.

- HuffPo:"He may end up being right, but he's just guessing. A "trend line adjustment" is merely political punditry dressed up as sophisticated mathematical modeling."
- 538: Offers quantitative reasoning for re-/under-weighting older polls, & changing as election approaches



http://www.huffingtonpost.com/entry/nate-silver-election-forecast_us_581e1c33e4b0d9ce6fbc6f7f

https://fivethirtyeight.com/features/a-users-guide-to-fivethirtyeights-2016-general-election-forecast/

POLLING: 2016

Dr Sam Wang, a polling expert who said he would eat a bug if Donald Trump won more than 240 electoral votes during Tuesday's election has made good on that promise on Saturday

- Many other less-famous poll aggregators who got 2008/2012 right
  - Including Andrew Tannenbaum (electoral-vote.org)
- Princeton Election Consortium
  - Even more accurate than Nate Silver for earlier ones
  - http://election.princeton.edu/2016/11/08/final-mode-projections-clinton-323-ev-51-di-senate-seats-gop-hou

AD TARGETING

- Pregnancy is an expensive & habit-forming time
  - Thus, valuable to consumer-facing firms
- 2012:
  - Target identifies 25 products and subsets thereof that are commonly bought in early pregnancy
  - Uses purchase history of patrons to predict pregnancy, targets advertising for post-natal products (cribs, etc)
  - Good: increased revenue
  - Bad: this can expose pregnancies – as famously happened in Minneapolis to a high schooler

http://www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2

AUTOMATED DECISIONS OF CONSEQUENCE

[Sweeney 2013, Miller 2015, Byrnes 2016, Rudin 2013, Barry-Jester et al. 2015]

textio

DOXA

gild

GJ GapJumpers

zest finance

LendingClub

Kabbage

PREDPOL

NORTHPOINTE
Advancing Justice. Embracing Community.

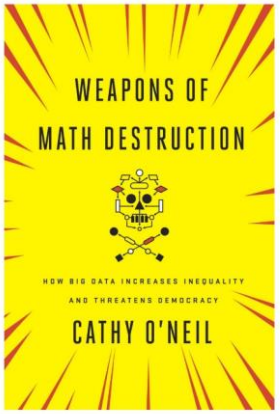
Hiring

Lending

Policing/ sentencing

Search for minority names ➔
ads for DUI/arrest records

Female cookies ➔
less freq. shown professional job opening ads



"… a lot remains unknown about how big data-driven decisions may or may not use factors that are proxies for race, sex, or other traits that U.S. laws generally prohibit from being used in a wide range of commercial decisions … What can be done to make sure these products and services–and the companies that use them treat consumers fairly and ethically?"

- FTC Commissioner Julie Brill [2015]

NETFLIX PRIZE I

- Recommender systems: predict a user's rating of an item

	Twilight	Wall-E	Twilight II	Furious 7
User 1	+1	-1	+1	?
User 2	+1	-1	?	?
User 3	-1	+1	-1	+1

Netflix Prize: $1MM to the first team that beats our in-house engine by 10%

- Happened after about three years
- Model was never used by Netflix for a variety of reasons
  - Out of date (DVDs vs streaming)
  - Too complicated / not interpretable

NETFLIX PRIZE II

- Netflix initially planned a follow-up competition
- In 2007, UT Austin managed to deanonymize portions of the original released (anonymized) Netflix dataset:
  - ??????????
  - Matched rating against those made publicly on IMDb
- Why could this be bad?
- 2009—2010, four Netflix users filed a class-action lawsuit against Netflix over

TODAY'S CLASS

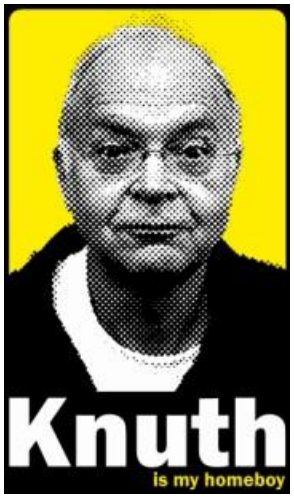
- What is Data Science?
  - Data Lifecycle
  - Where is it headed?
- Introductions
- DATA/MSML602 Details
- Few Data Science Success Stories and Cautionary Tales
- Basic Technology Stack and Best Practices
  - Python, Jupyter Notebook, VS Code, GitHub
  - Cloud Computing, Containers (e.g., Docker)

PYTHON

- Python is an interpreted, dynamically-typed, high-level, garbage-collected, object-oriented-functional-imperative, and widely used scripting language.
  - Interpreted: instructions executed without being compiled into (virtual) machine instructions*
  - Dynamically-typed: verifies type safety at runtime
  - High-level: abstracted away from the raw metal and kernel
  - Garbage-collected: memory management is automated
  - OOFI: you can do bits of OO, F, and I programming
- Not the point of this class!
  - Python is fast (developer time), intuitive, and used in industry!

*you can compile Python source, but it's not required

LITERATE PROGRAMMING



- Literate code contains in one document:
  - the source code;
  - text explanation of the code; and
  - the end result of running the code.
- Basic idea: present code in the order that logic and flow of human thoughts demand, not the machine-needed ordering
  - Necessary for data science!
  - Many choices made need textual explanation, ditto results.

IP[y]: IPython Interactive Computing

jupyte

Visual Studio Code

JUPYTER PROJECT

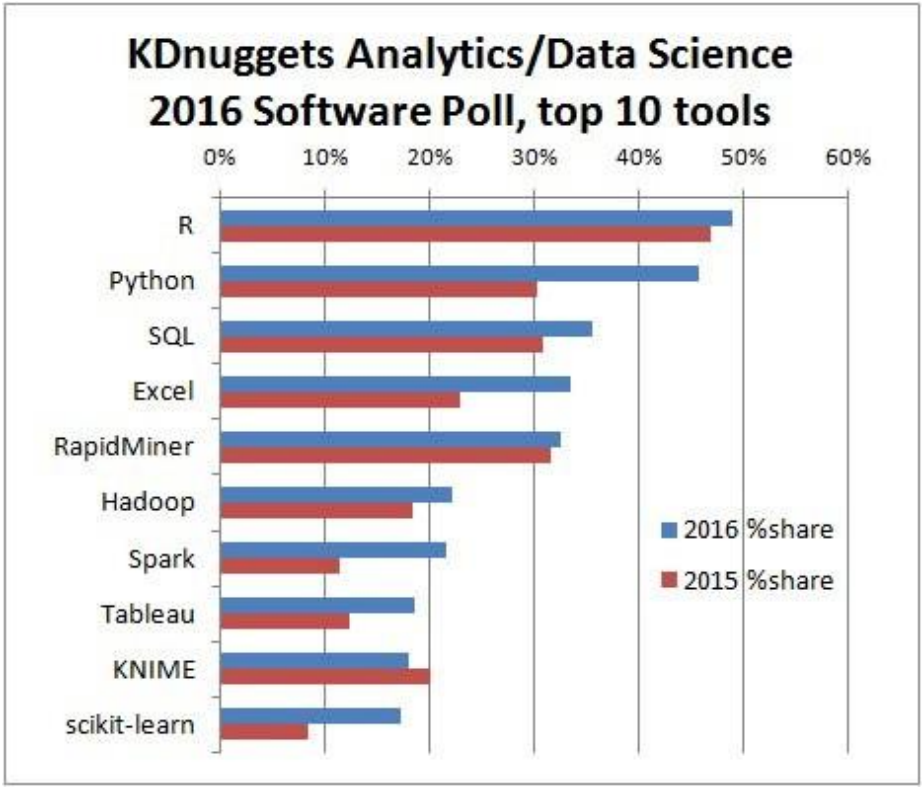
- Started as IPython Notebooks, a web-based frontend to the IPython Shell
  - Notebook functionality separated out a few years ago
  - Now supports over 40 languages/kernels
  - Notebooks can be shared easily
  - Can leverage big data tools like Spark

- Apache Zeppelin:
  - https://www.linkedin.com/pulse/comprehensive-comparison-jupyter-vs-zeppelin-hoc-q-phan-mba-

- Several others including RStudio (specific to R)
- Spyder and more recently VS Code (VS Code works for almost everything)

PYTHON VS R (FOR DATA SCIENTISTS)

- There is no right answer here!
  - Python is a "full" programming language – easier to integrate with systems in the field
  - R has a more mature set of pure stats libraries …
  - … but Python is catching up quickly …
  - … and is already ahead specifically for ML.
- You will see Python more in the tech industry.



WRAP-UP SO FAR

- Welcome to Data Science Master Program !!
- Start learning several topics mentioned in these first lectures
- Go through Python Notebooks as well as VS Code to (re-) familiarize with Python
- If you are interested in more programing, go through C++ to (re-)familiarize with C++20