https://goo.gl/wfHwza

Machine Learning at Scale: Logistics and Introduction

James G. Shanahan[1,2]

[1]Church and Duncan Group Inc., [2]iSchool UC Berkeley, CA,
EMAIL: James_DOT_Shanahan_AT_gmail_DOT_com

Unit 1: Lecture 1

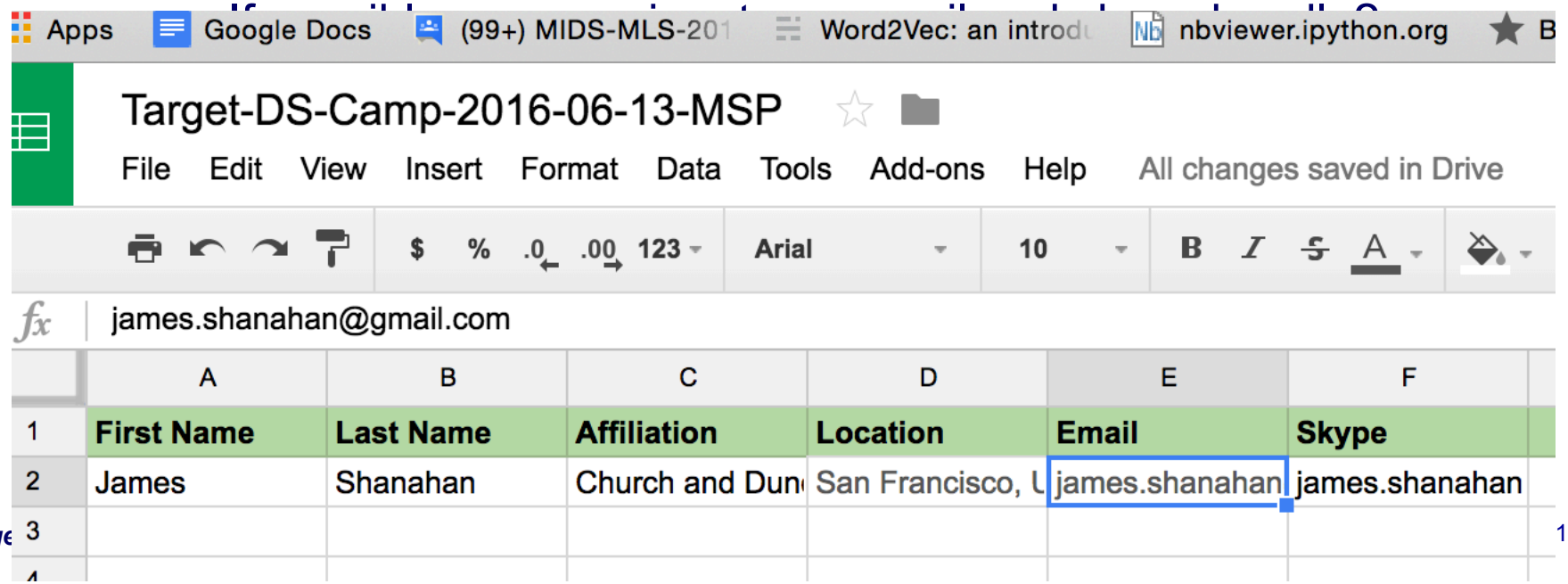
June 13, 2016

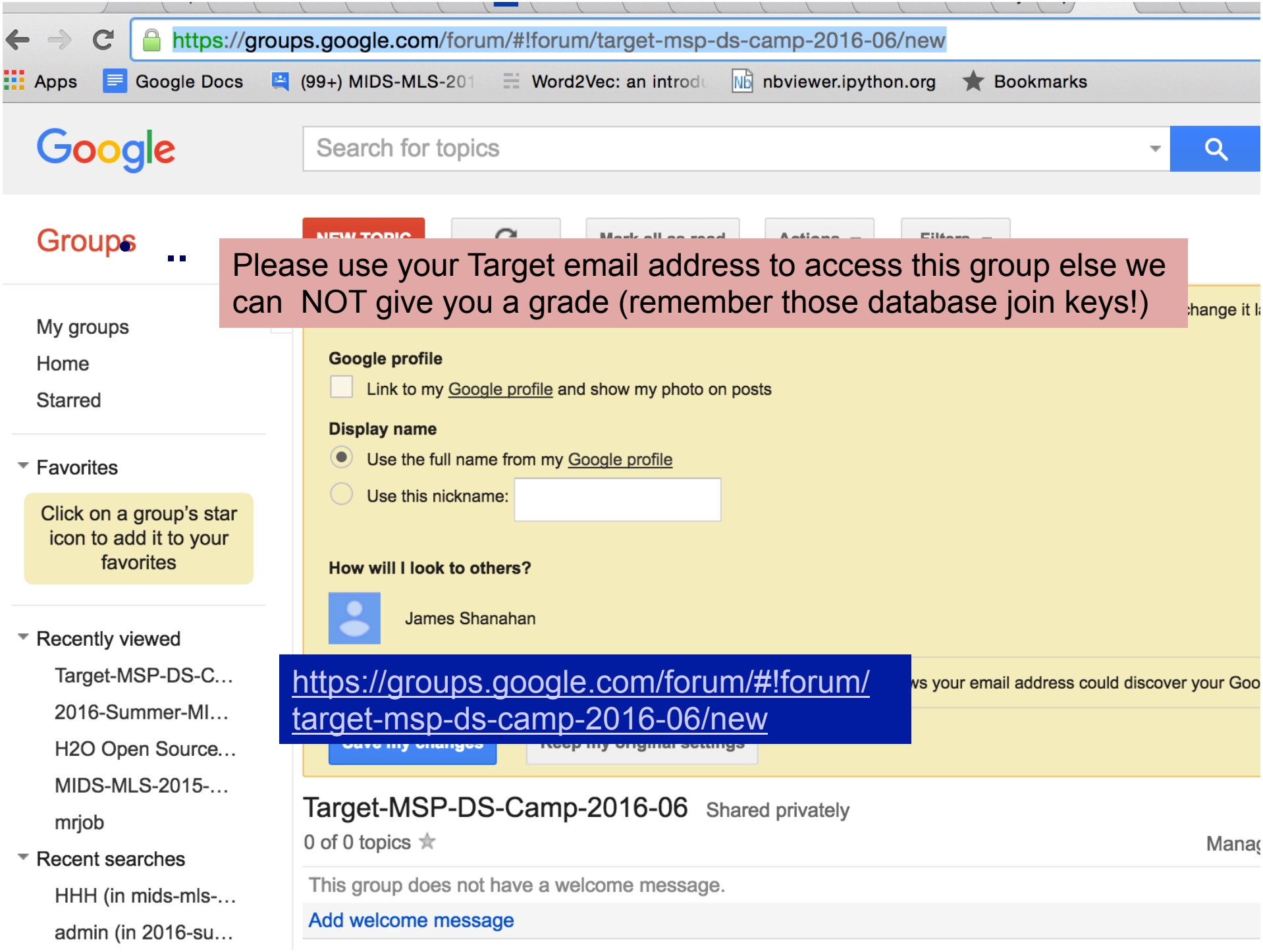
Lecture Outline

- Google Doc and Group
- Welcome & Class Introductions
- Big Data and Applications
- Course introduction
- Class logistics
- Systems (part 1 of N)

For communication during Tutorial

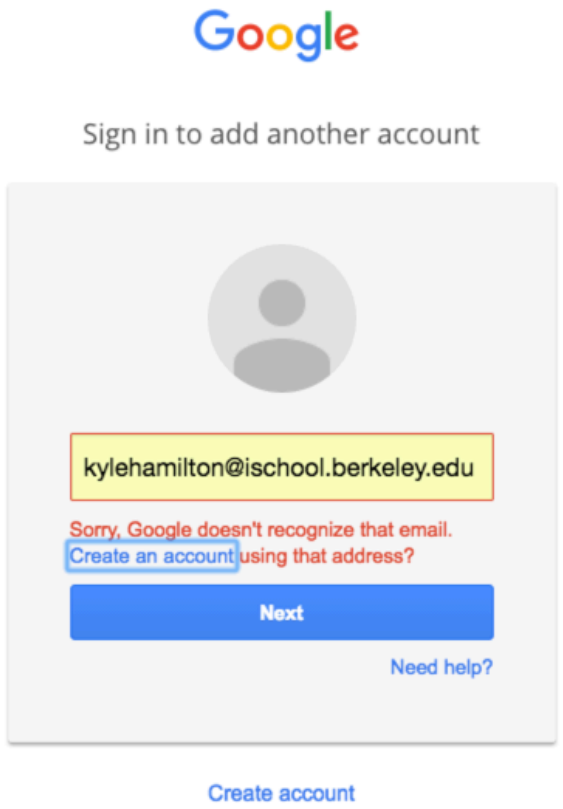
- **Google Doc**
  - https://goo.gl/wfHwza
  - https://docs.google.com/spreadsheets/d/1IsQwtdVdR977rUGt5P3sxJMx7M6RAEvl-ejidbR2gWM/edit?usp=sharing




https://groups.google.com/forum/#!forum/target-msp-ds-camp-2016-06/new
Apps
Google Docs
(99+) MIDS-MLS-201
Word2Vec: an introd
nbviewer.ipython.org
Bookmarks
Google
Search for topics
Groups
Please use your Target email address to access this group else we can NOT give you a grade (remember those database join keys!)
My groups
Home
Starred
Favorites
Click on a group's star icon to add it to your favorites
Recently viewed
Target-MSP-DS-C…
2016-Summer-MI…
H2O Open Source…
MIDS-MLS-2015-…
mrjob
Recent searches
HHH (in mids-mls-…
admin (in 2016-su…
Google profile
Link to my Google profile and show my photo on posts
Display name
Use the full name from my Google profile
Use this nickname:
How will I look to others?
James Shanahan
https://groups.google.com/forum/#!forum/target-msp-ds-camp-2016-06/new
Target-MSP-DS-Camp-2016-06
Shared privately
0 of 0 topics
This group does not have a welcome message.
Add welcome message

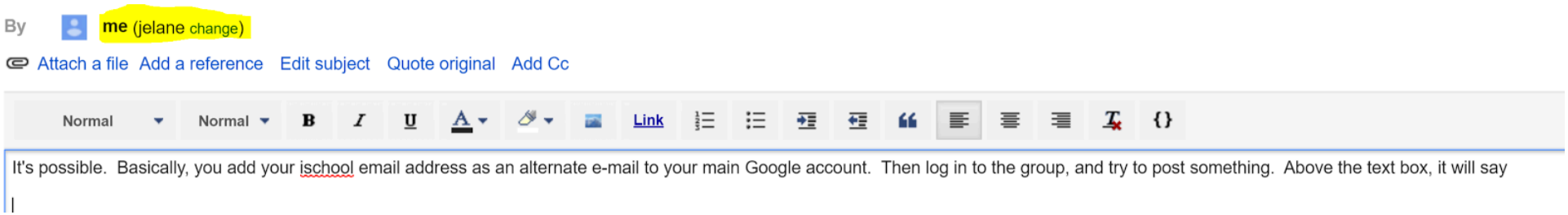
Accessing Google Groups

- **You have to attempt to sign-in with the ischool email at which point it gives you the option of creating a Google account with just your iSchool email. Here's a screenshot that Kyle used to show everyone on slack.**

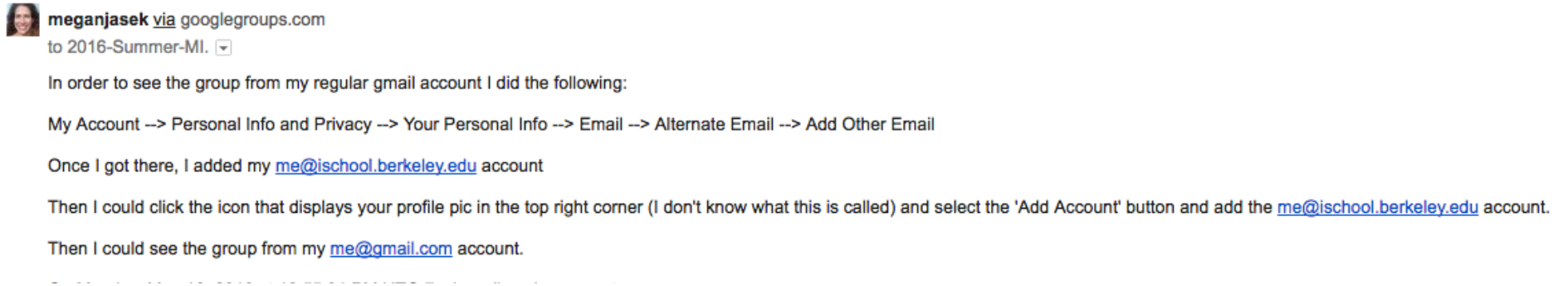


Consolidate your emails

- Basically, you add your ischool email address as an alternate e-mail to your main Google account. Then log in to the group, and try to post something. Above the text box, it will let you change which e-mail address you use"



In order to see the group from my regular gmail account



meganjasek via googlegroups.com
to 2016-Summer-MI.

In order to see the group from my regular gmail account I did the following:

My Account --> Personal Info and Privacy --> Your Personal Info --> Email --> Alternate Email --> Add Other Email

Once I got there, I added my me@ischool.berkeley.edu account

Then I could click the icon that displays your profile pic in the top right corner (I don't know what this is called) and select the 'Add Account' button and add the me@ischool.berkeley.edu account.

Then I could see the group from my me@gmail.com account.

Course Schedule

Row	Date	Units	Description
1	May12-June 13	Unit 0	Probability Theory, Python, R
2	6/13 (Mon) -6/16 (Thur)	Units 1-7	Class Time
3	*6/28 Tues 11AM, MSP Time*	*Unit 8: Exam*	*2 hours Exam*
4	*7/2 Sat.*	*Unit 8 : All Homework*	
5	7/12 (Tues) -7/15 (Fri)	Units 9-15	Class Time
6	*7/26 Tues, 11AM MSP Time*	*Unit 16: Exam*	*2 Hours Exam*
7	*7/30*	*Unit 16: All Homework*	
8	8/1-10/1	Work or a Target project	
9	Week of October 3	1-2 days wrapup	A half day workshop where results will be presented to colleagues along with prep before

Lecture Outline

- Google Doc and Group
- Welcome & Class Introductions
- Big Data and Applications
- Course introduction
- Class logistics
- Systems (part 1 of N)

Audience Participation is encouraged!

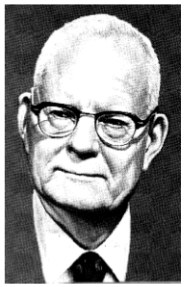
Please share and help your colleagues!

Data Analysis Has Been Around for a While

1935: "The Design of Experiments"

R.A. Fisher 

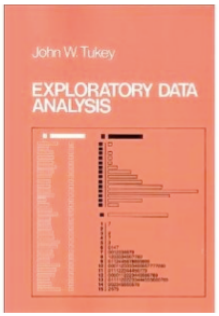
1939: "Quality Control"

W.E. Demming 

1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"



1989: "Business Intelligence"

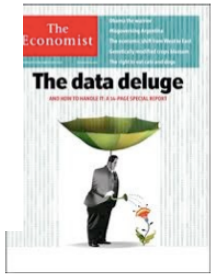
Howard Dresner 

1997: "Machine Learning"

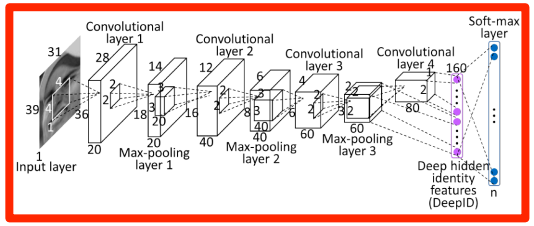
1997 Google

Google

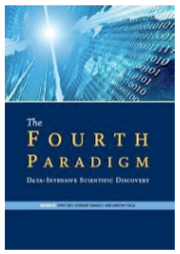
2010:"The Data Deluge"



2012: Deep Learning



2007:"The Fourth Paradigm"

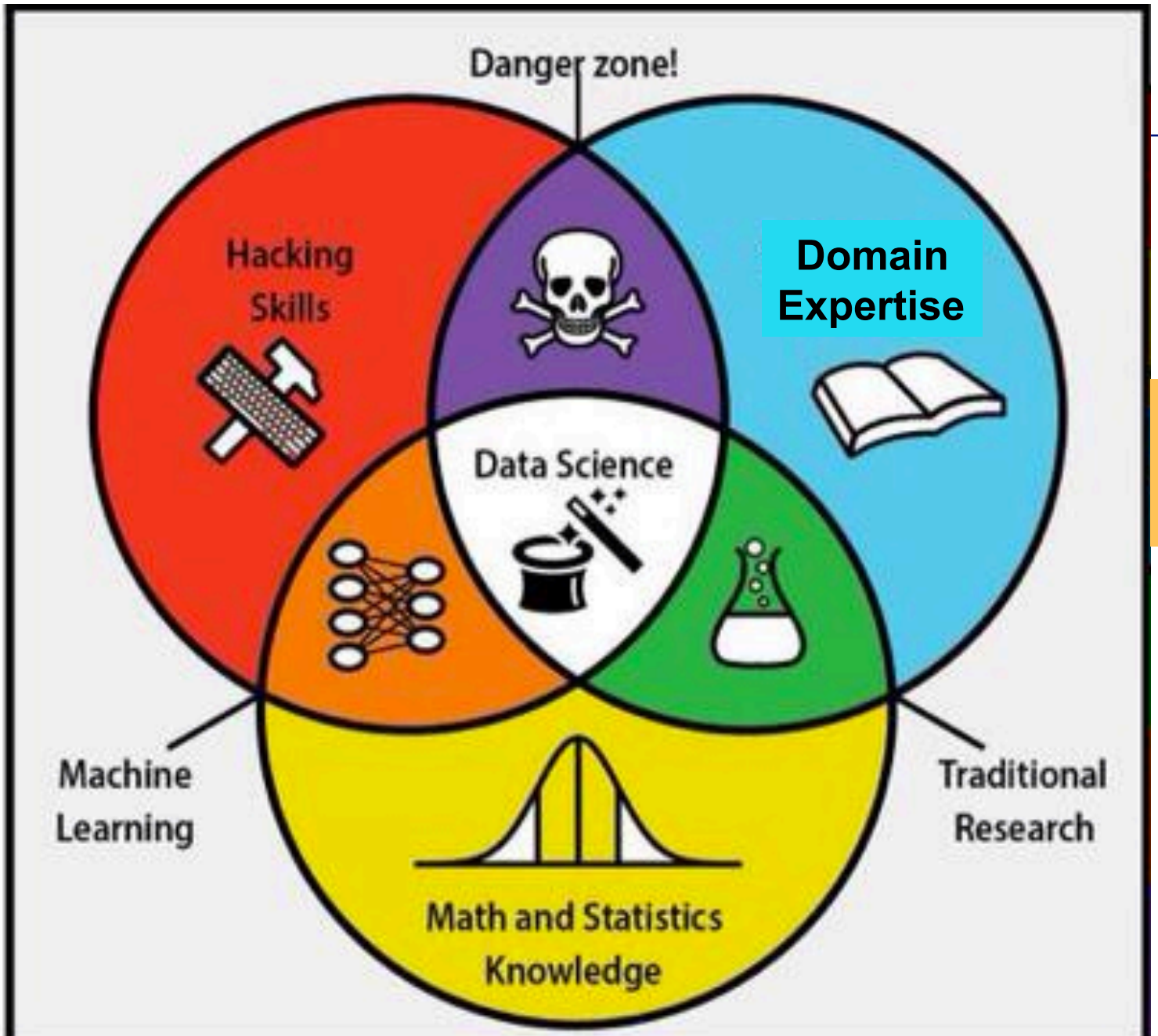


2009:"The Unreasonable Effectiveness of Data"

2013: Spark

Spark

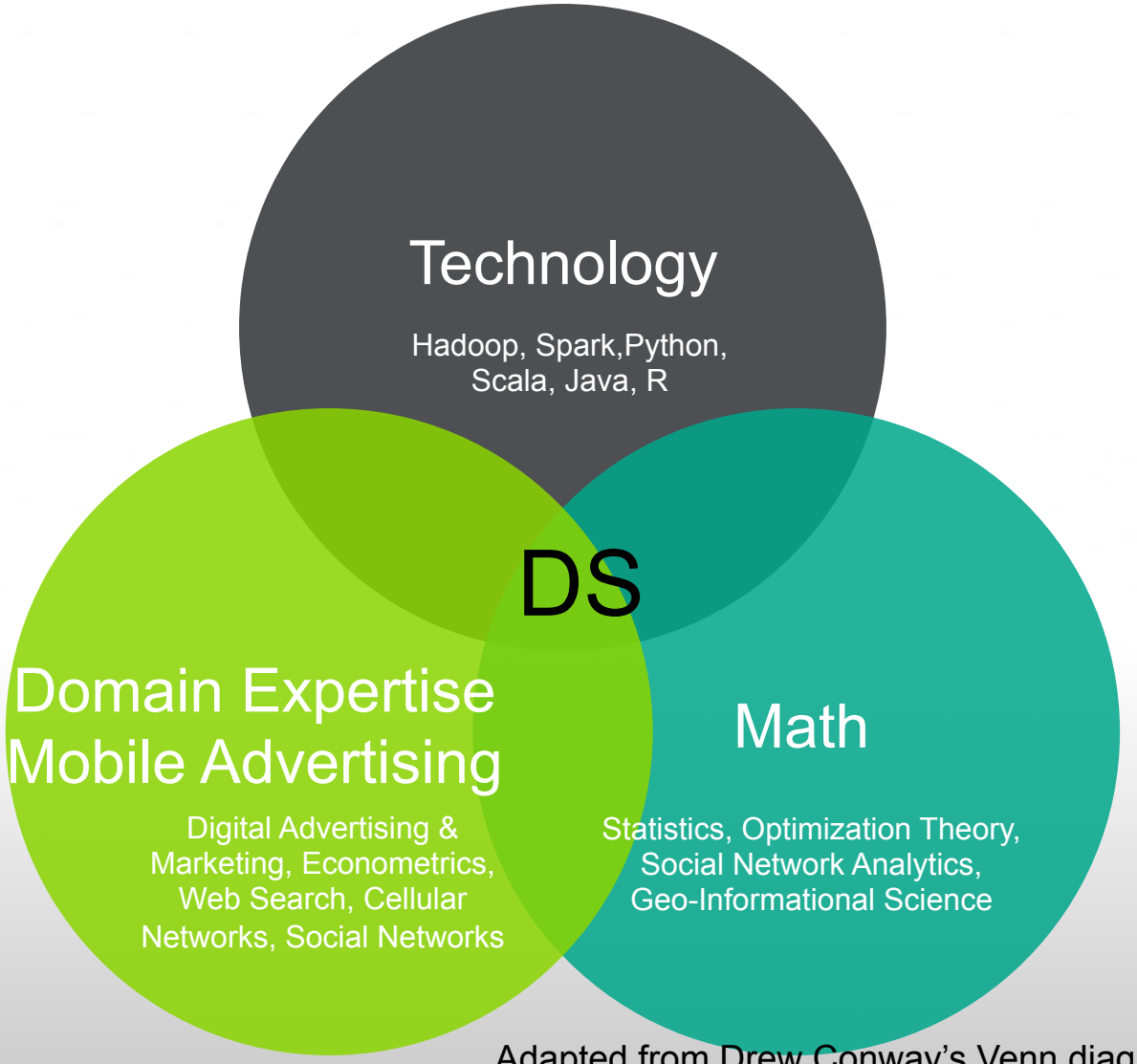
DS Skillset



A venn diagram with a Danger Bearing

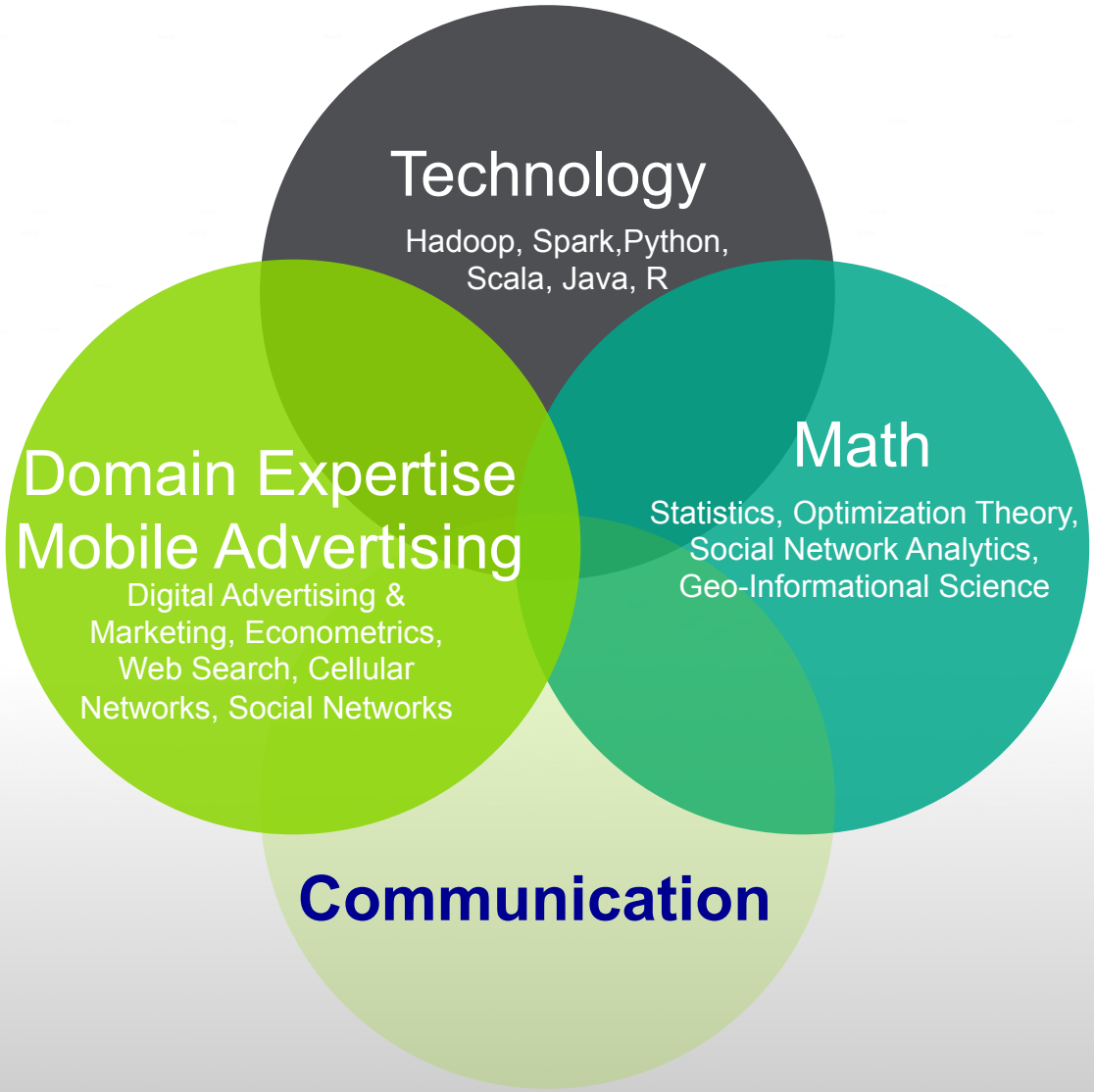
[adapted from Drew Conway]

Data Science

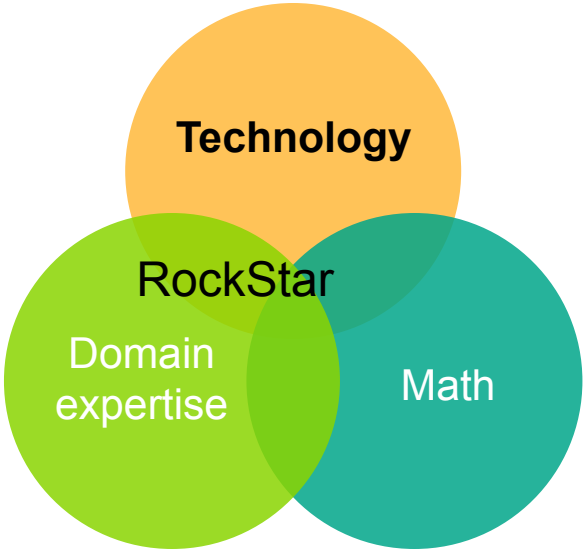


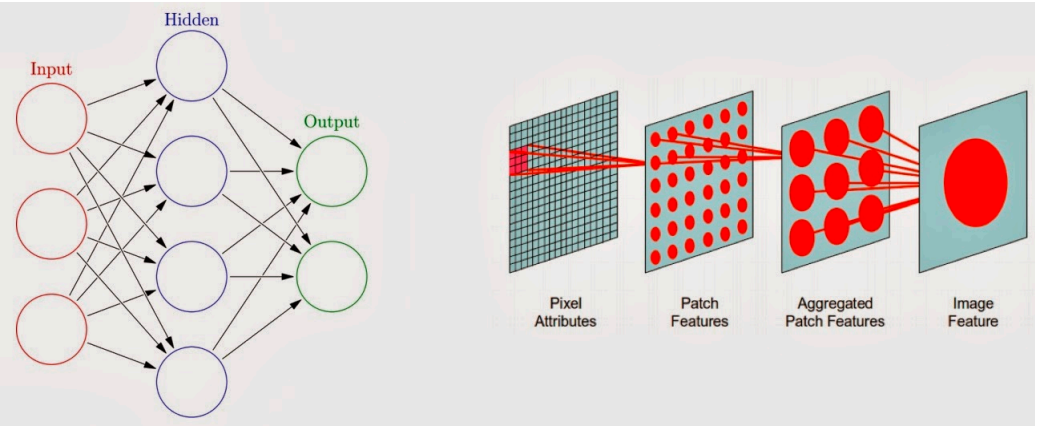
Adapted from Drew Conway's Venn diagram of data science

Data Scientist

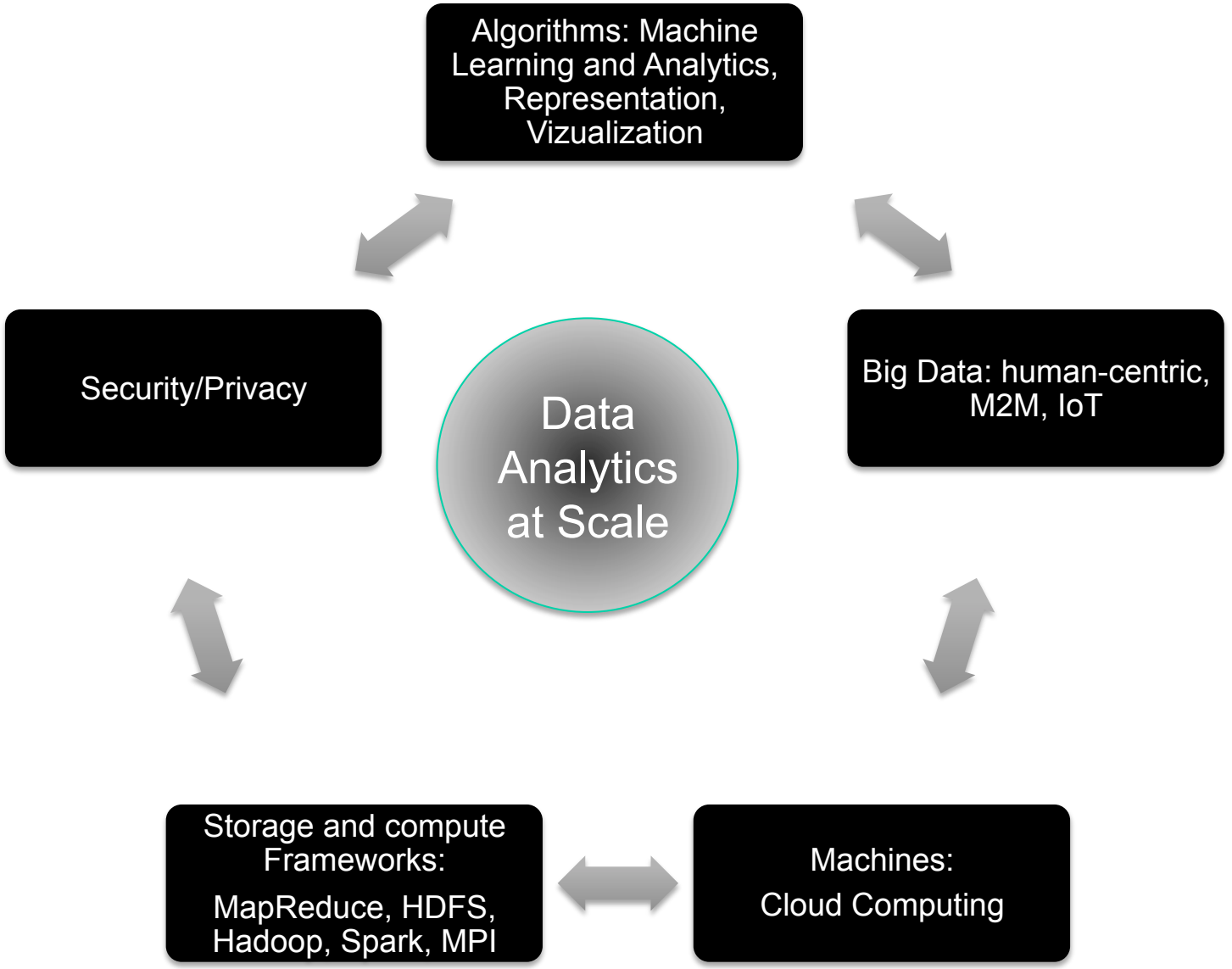


RockStars and Super Models

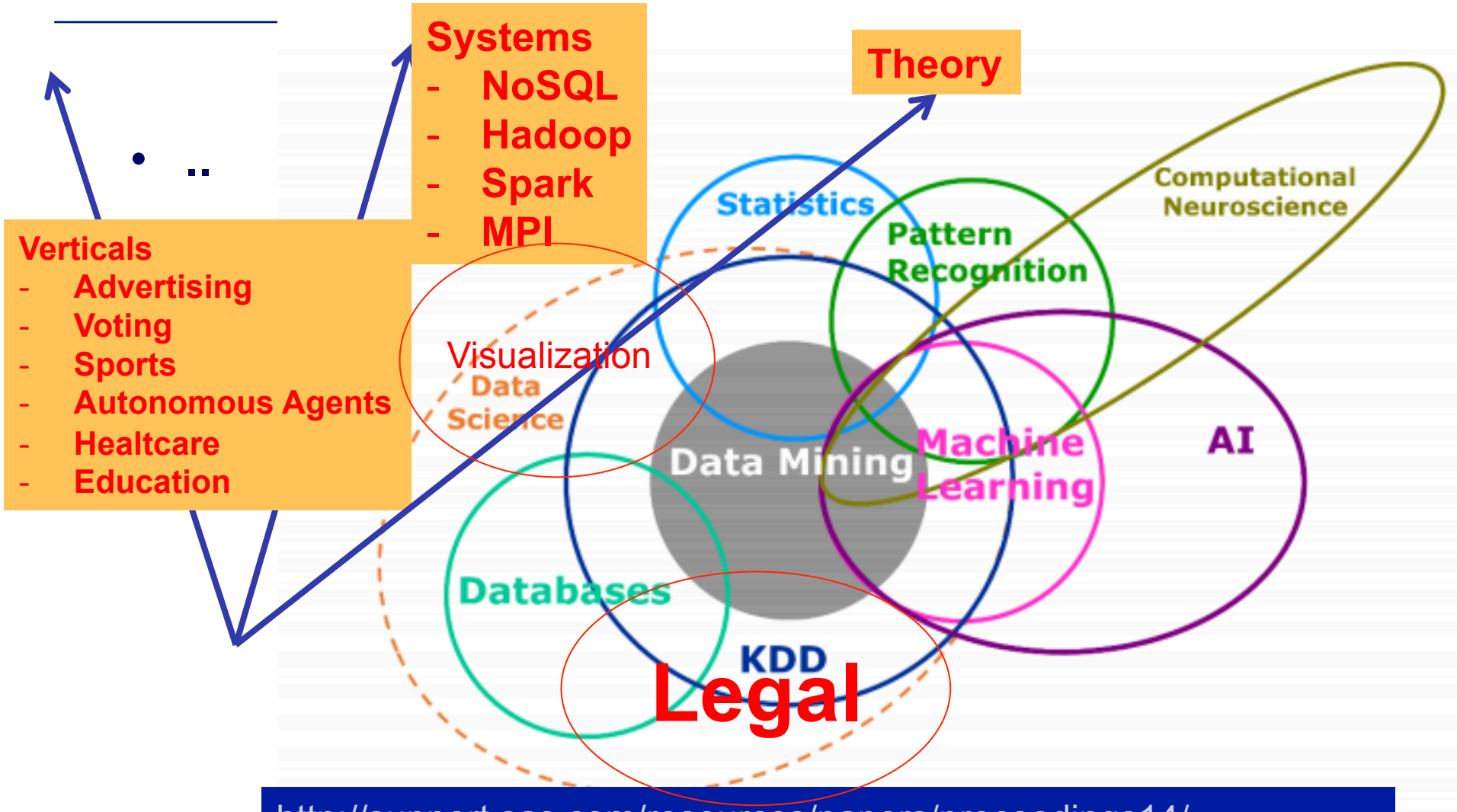




Data Analytics at Scale



DS is Systems + Theory + Verticals

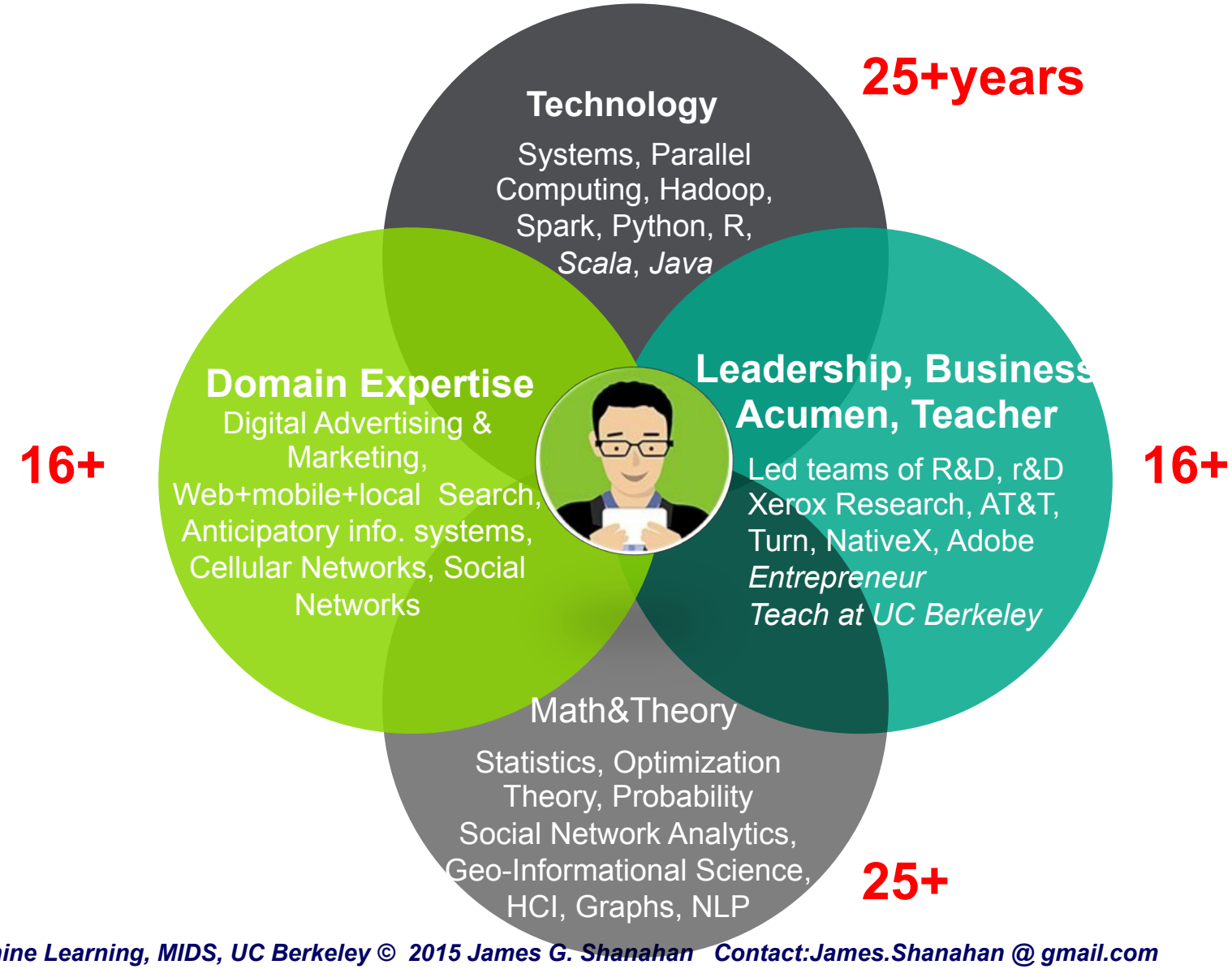


http://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf

Data Science-centric Self introduction

- **Name, Affiliation, which part of the world?**
- **Student/researcher/practitioner**
- **Familiarity with (novice/intermediate/experienced) :**
  - Domain expertise (Areas of research/application)
  - Systems: MapReduce/Hadoop/Spark
  - Theory: Machine learning Experience
- **Engineer/Scientist/FullStack**
- **Hardware/Software**
  - Windows
  - Mac
  - Linux
  - All

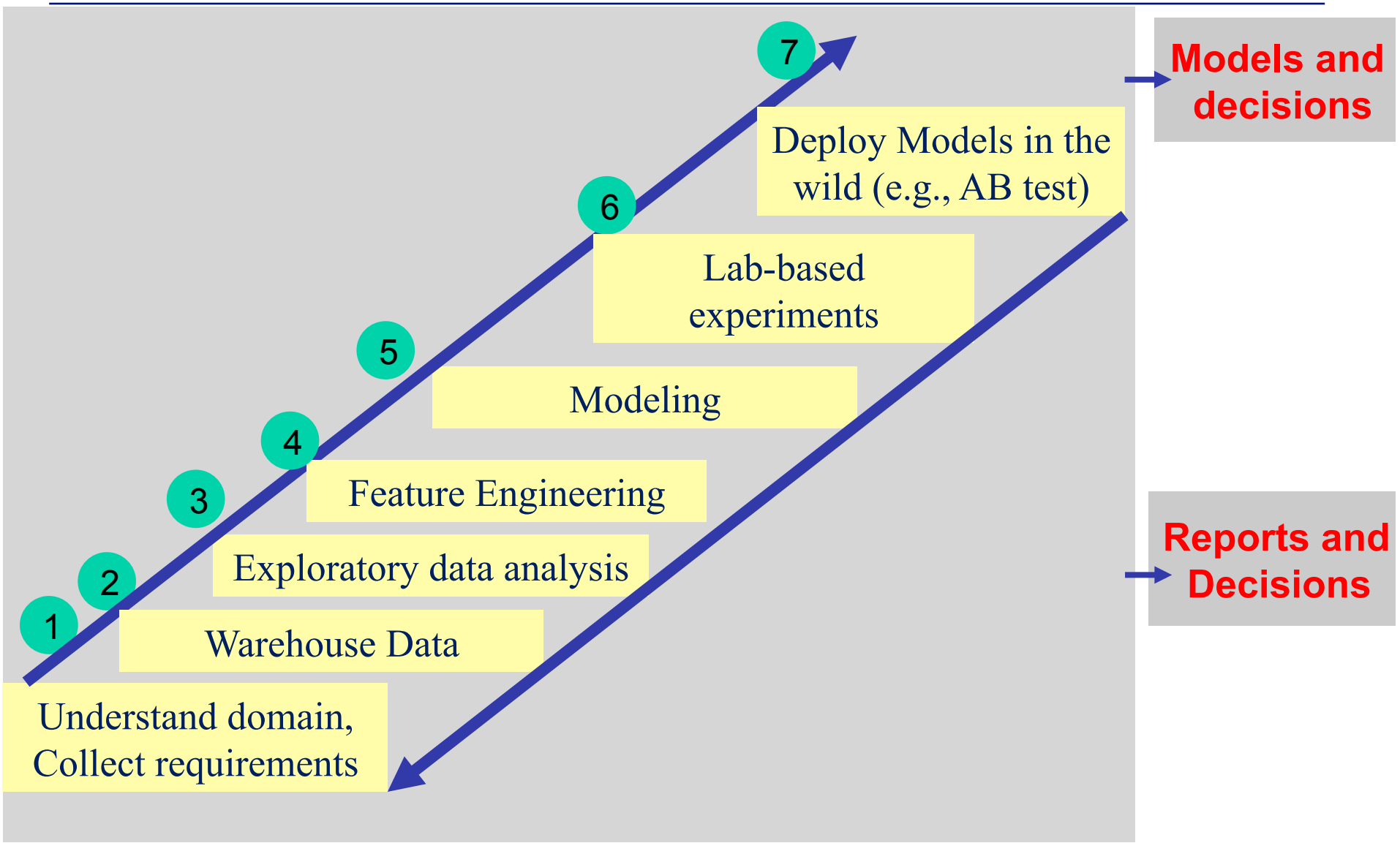
James G. Shanahan 25+ years in data science



James G. Shanahan

- **25+ years in data science**
- **Currently**
  - Principal and Founder, Data Science Consultancy
    - Clients: Adobe, Akamai, Ancestry, AT&T, Nokia Siemens, SearchMe, …
  - Teaching
    - Co-creator of UC Berkeley MIDS program; curriculum development
    - Teach Large Scale Machine Learning (Fall 2014,2015,2016)
    - Teach Machine Learning and Optimization Theory at University of California Santa Cruz (UCSC), TIM 206, TIM 209, TIM 250, TIM 251 (since 2008)
  - Advising: Quixey, InferSystems, Knotch
- **Previously**
  - NativeX: SVP of Data Science, Chief Scientist, and board member
  - Founding Chief Scientist, Turn Inc.
  - Principal Scientist, Clairvoyance Corp (CMU spinoff; sister lab to JRC)
  - Research Scientist, Xerox Research;
  - Entrepreneur: Cofounder of Document Souls and RTB Fast
- **Education:** PhD in ML, University of Bristol, UK; B.Sc. CS, Uni. of Limerick, Ireland

Typical Abstract Data Analytics Pipeline



Self Introductions

- **Paste the following into the Google Docs**
  - Your Location
  - Background (paste 100 word bio into chat)
  - What you want to get out of this class

Unit 0	Bootcamp: Probability Theory and Python
Lecture 1	Introduction and Motivation for Machine Learning at Scale
References	Richard E. Neapolitan , Learning Bayesian Networks, Pearson, 2003 Introduction to Python (on the web from O'Reilly)

Topics
 	
Quiz
	Probability Theory
	
	• **Probability Basics**
	  ○ Probability Axioms
	  ○ Conditional probabilities
	  ○ Product Rule, Chain Rule, Bayes Rule
	• **Bayes Rule**
	  ○ Learning
	  ○ Independence
	  ○ Conditional independence
	
	Programming in Python
	
	• Variables; assignments; loops; functions;
	• Object oriented programming
	  ○ Classes; methods;
	• Iterators; lambda functions; closures

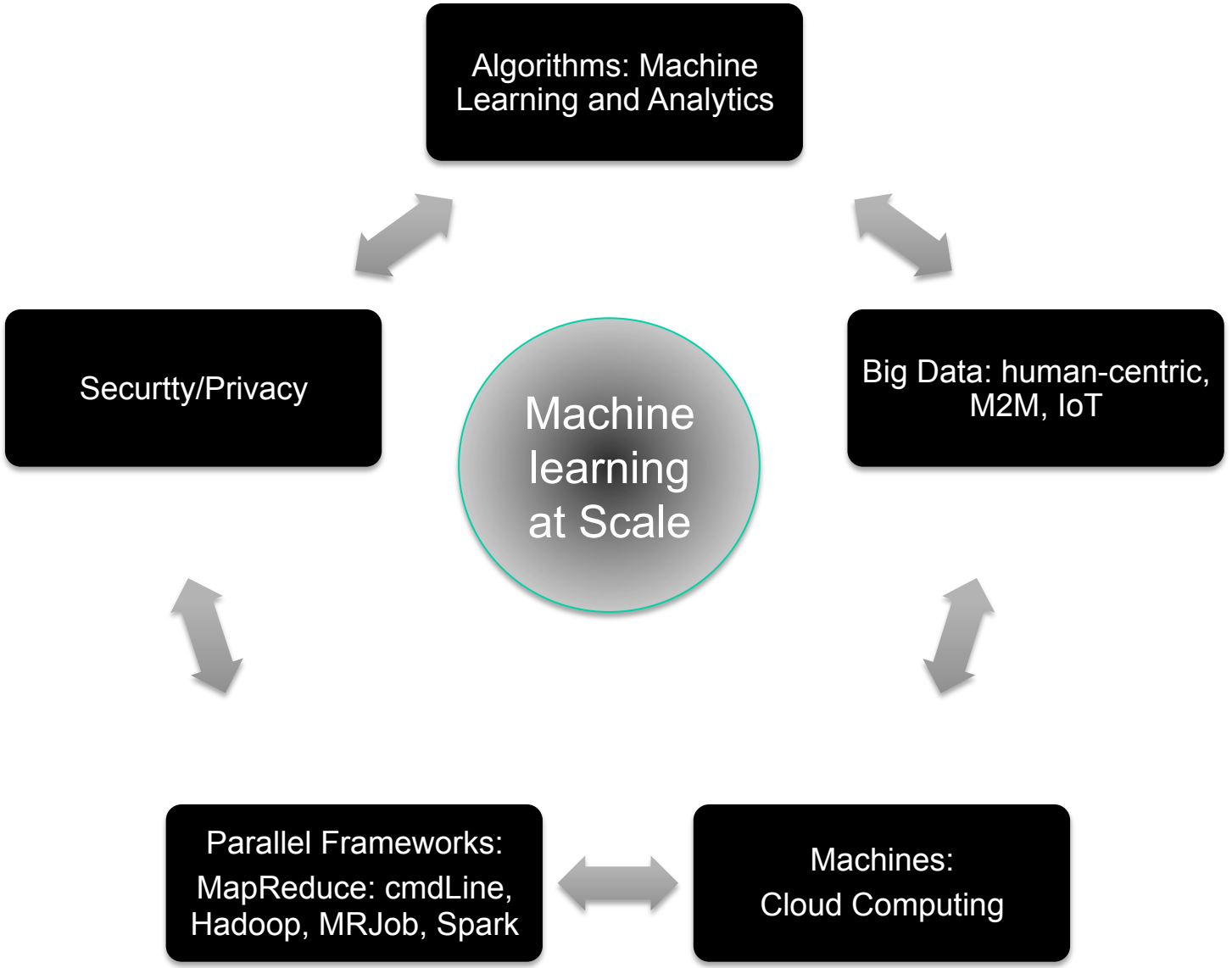
Lecture Outline

- Google Doc and Group
- Welcome & Class Introductions
- Big Data and Applications
- Course introduction
- Class logistics
- Systems (part 1 of N)

Implementing algorithms from scratch

- **Focus on theory, implementation and practice**
- **There are several different reasons why implementing algorithms from scratch can be useful:**
  - it can help us to understand the inner workings of an algorithm
  - Let's us figure how to parallize
  - we can add new features to an algorithm or experiment with different variations of the core idea
  - we want to invent new algorithms or implement algorithms no one has implemented/shared yet
  - we are not satisfied with the API and/or we want to integrate it more "naturally" into an existing software library
  - we could try to implement an algorithm more efficiently
  - we circumvent licensing issues (e.g., Linux vs. Unix) or platform restrictions

Machine learning at Scale

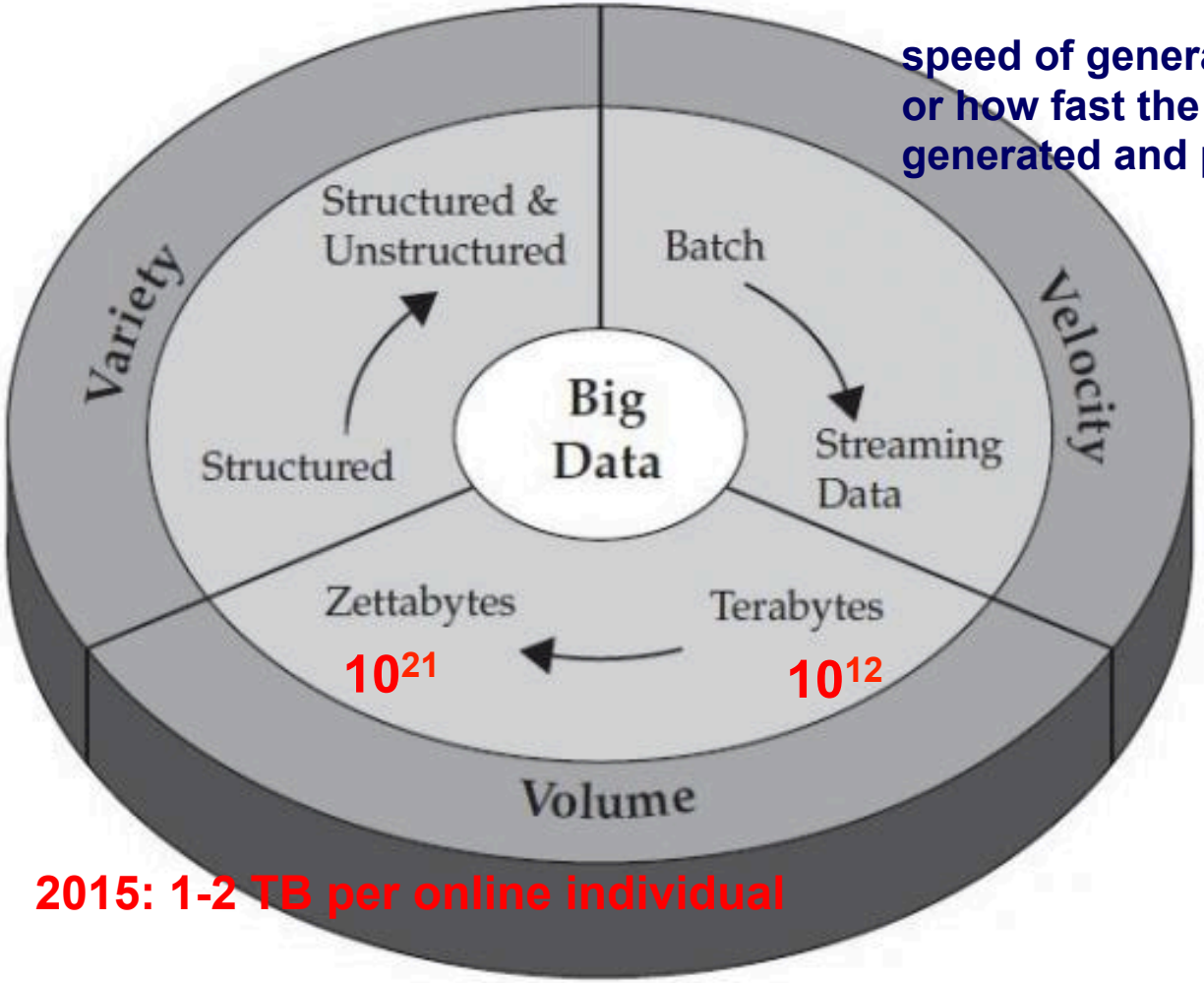


Big data Definition: use

- **Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.**
  - PROCESSING:
    - Think of your laptop that gets overwhelmed with 3-4 gig of data (disk space is 1TB)
  - STORAGE:
    - Laptop : 1 TB
  - THROUGH-PUT
    - 1TB would take 3 hours to read it using your laptop
- **Challenges**
  - Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, security, and information privacy.

- **In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."[18]**

Big Data: V^3



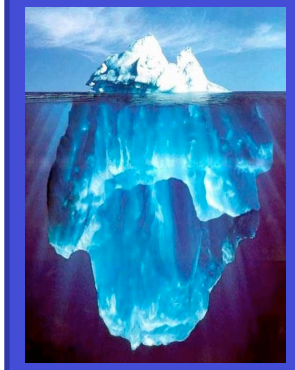
speed of generation of data or how fast the data is generated and processed

2015: 1-2 TB per online individual

Figure 1-1 *IBM characterizes Big Data by its volume, velocity, and variety—or simply, V^3.*

Sources Driving Big Data

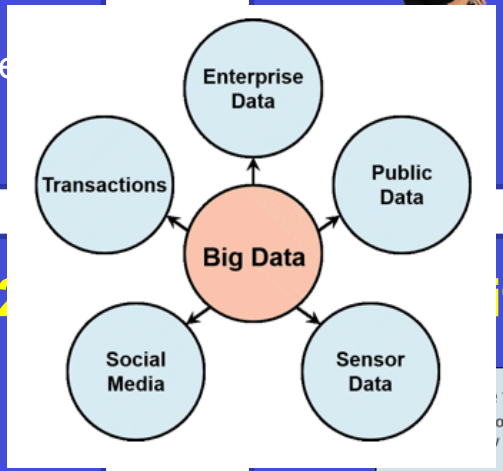
It's All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Friend Request
Transaction
Network message
Fault
…

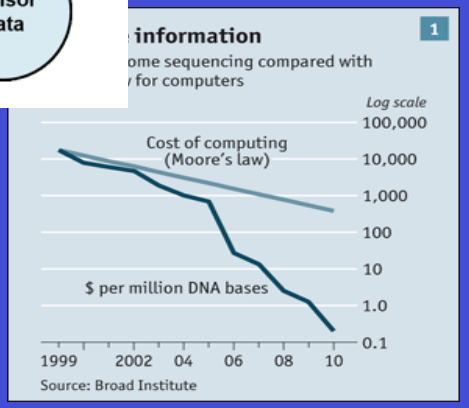
User Generated (Web, Social & Mobile) Quantified Self

…
..



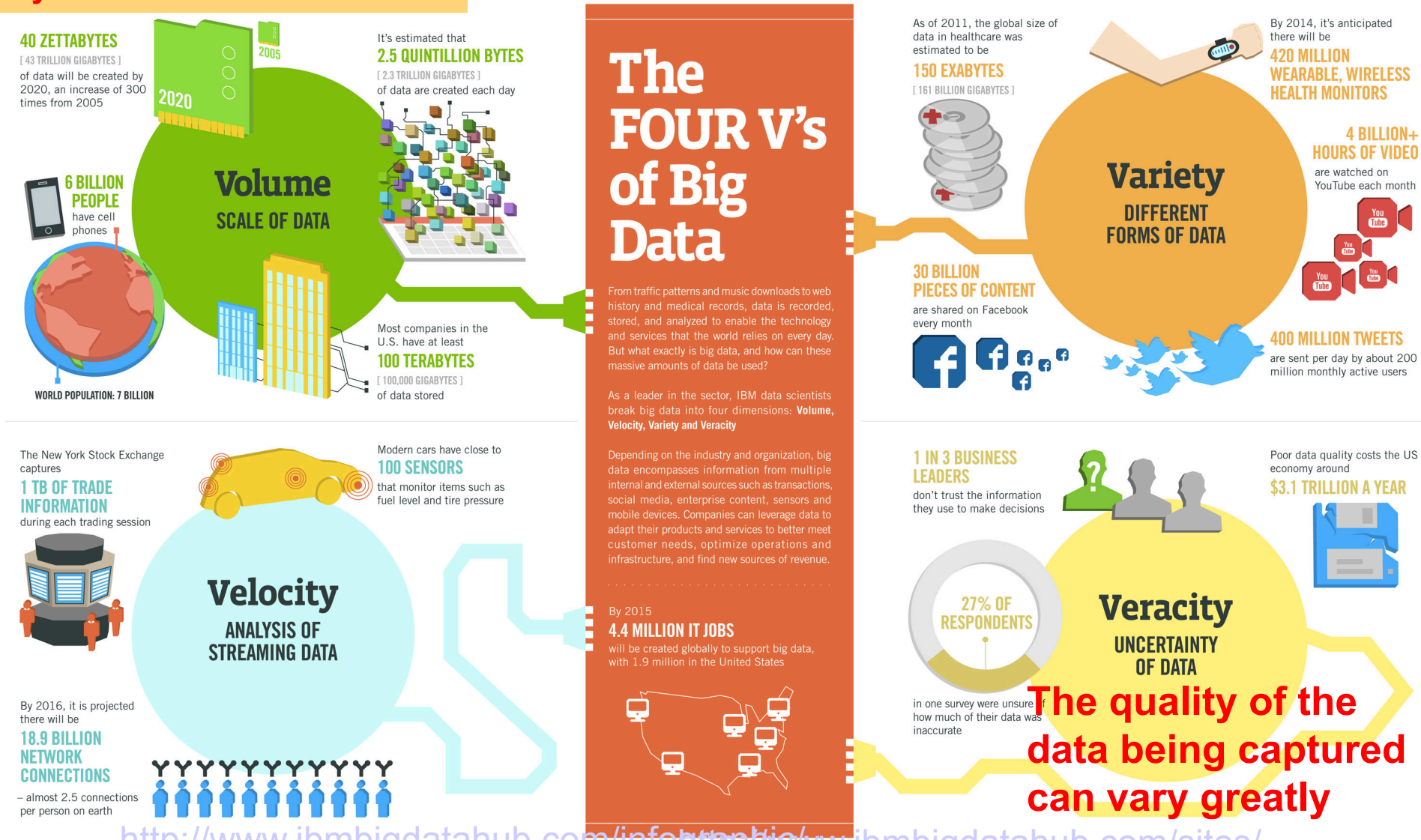
Internet of Things / M2

ific Computing



By 2005 we had 120* 10^{18}
By 2007 we had 280*20^{18}
By 2020 we will have 40* 10^{21}

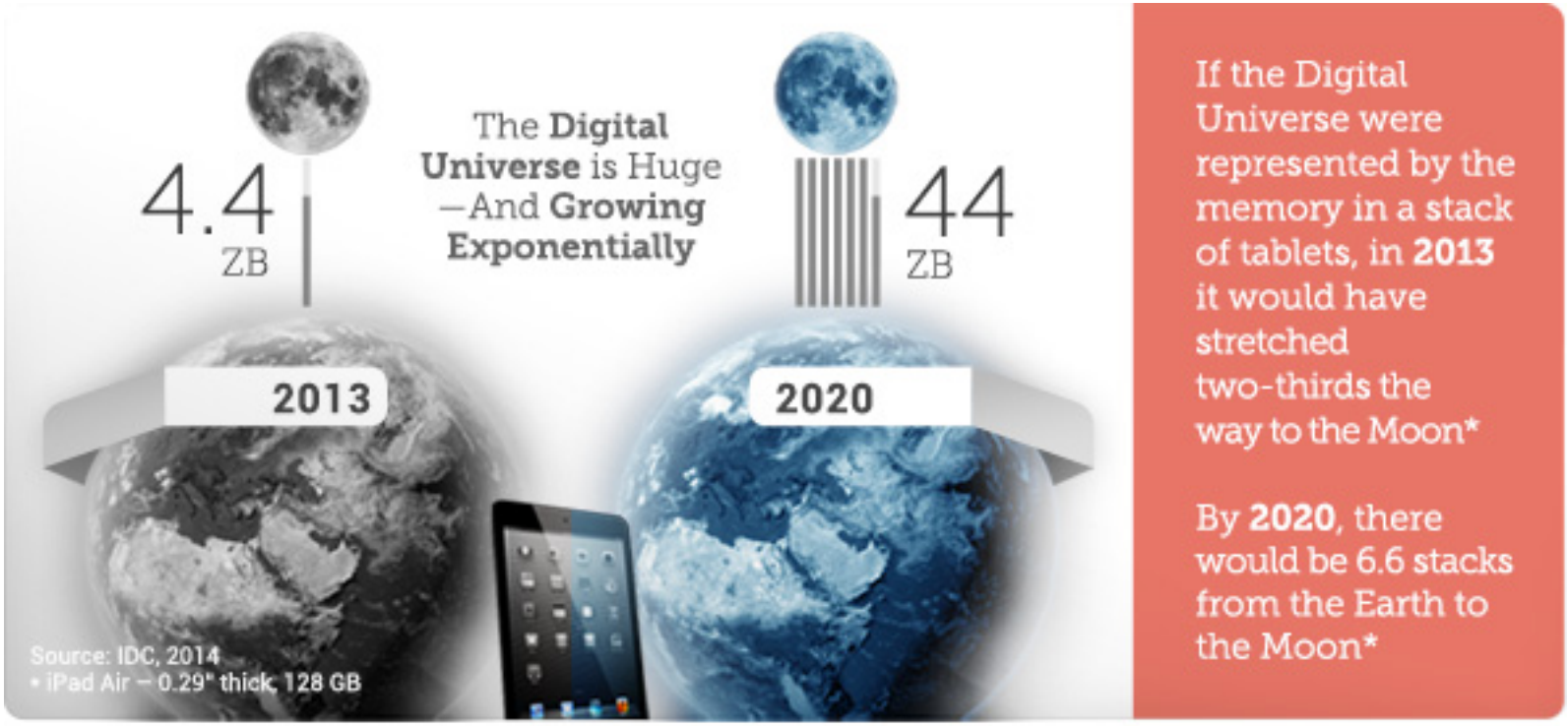
Big Data Infographic



The quality of the data being captured can vary greatly

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

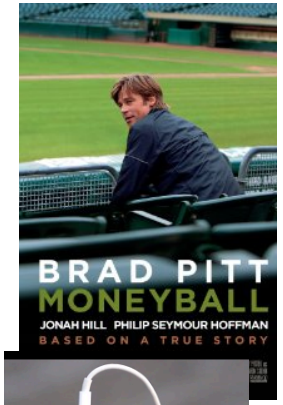
http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

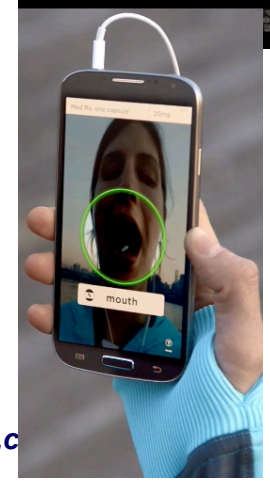


http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Why all the excitement?

- **Government:**
  - Obama used 80 pieces of information on each person; 4 year history (versus Romney)
  - Nate Silver used Bayesian techniques to publish analyses and predictions related to the 2008 and 2012 United States presidential election
- **Sports:**
  - Oakland Athletics baseball team and its manager Billy Beane
- **Transportation (e.g., Autonomous Vehicles)**
- **HCI: Speech Recognition and Translation**
- **Healthcare**
  - AI Cure: Do you know if your patients are taking their meds?
- **Digital Advertising**
- **Search (web, local, mobile)**





3 Vs of Big Data

1-2 TB per person today2014/2015

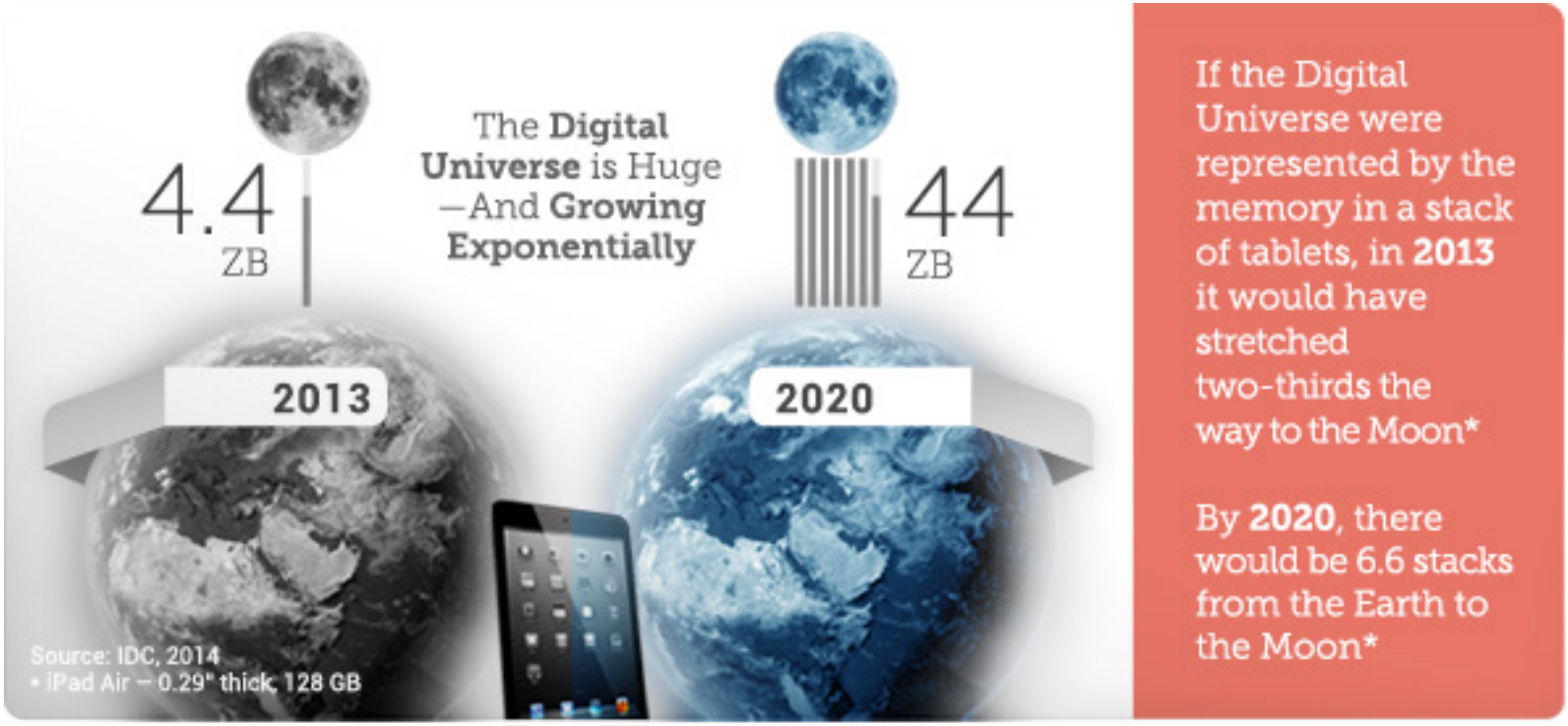
The data from these sources has a number of features that make it a challenge for a data warehouse:

Exponential Growth. An estimated 2.8ZB of data in 2012 is expected to grow to 40ZB by 2020. 85% of this data growth is expected to come from new types; with machine-generated data being projected to increase 15x by 2020. (Source IDC)

40TB per person by 2020

Varied Nature. The incoming data can have little or no structure, or structure that changes too frequently for reliable schema creation at time of ingest.

Value at High Volumes. The incoming data can have little or no value as individual, or small groups of records. But high volumes and longer historical perspectives can be inspected for patterns and used for advanced analytic applications.



http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Personal; society; M2M; crowdsourcing

- **Society**
  - Graphs: Social, professional;
  - Quantified self: Eating; Sleeping; exercising
  - Voting
  - Education
  - Healthcare…. Economics, shopping, etc.
- **Internet of things**
  - Tracking Wildebeests in Serengeti, Tanzania (not just with GPS tags, but also with cameras at key strategic locations through out the Serengeti
    - Population changes in species; Scheduling safaris
  - 1 Billion smart meters by 2020;
    - 1 Petabyte of data per day? 10^9 =10^12 10^15
    - 1 Billion smart meters (One megabye of data per device per day; Poll meter 1000 times per day; 1000 bytes of data each time
  - Smart cities
- **Etc.**

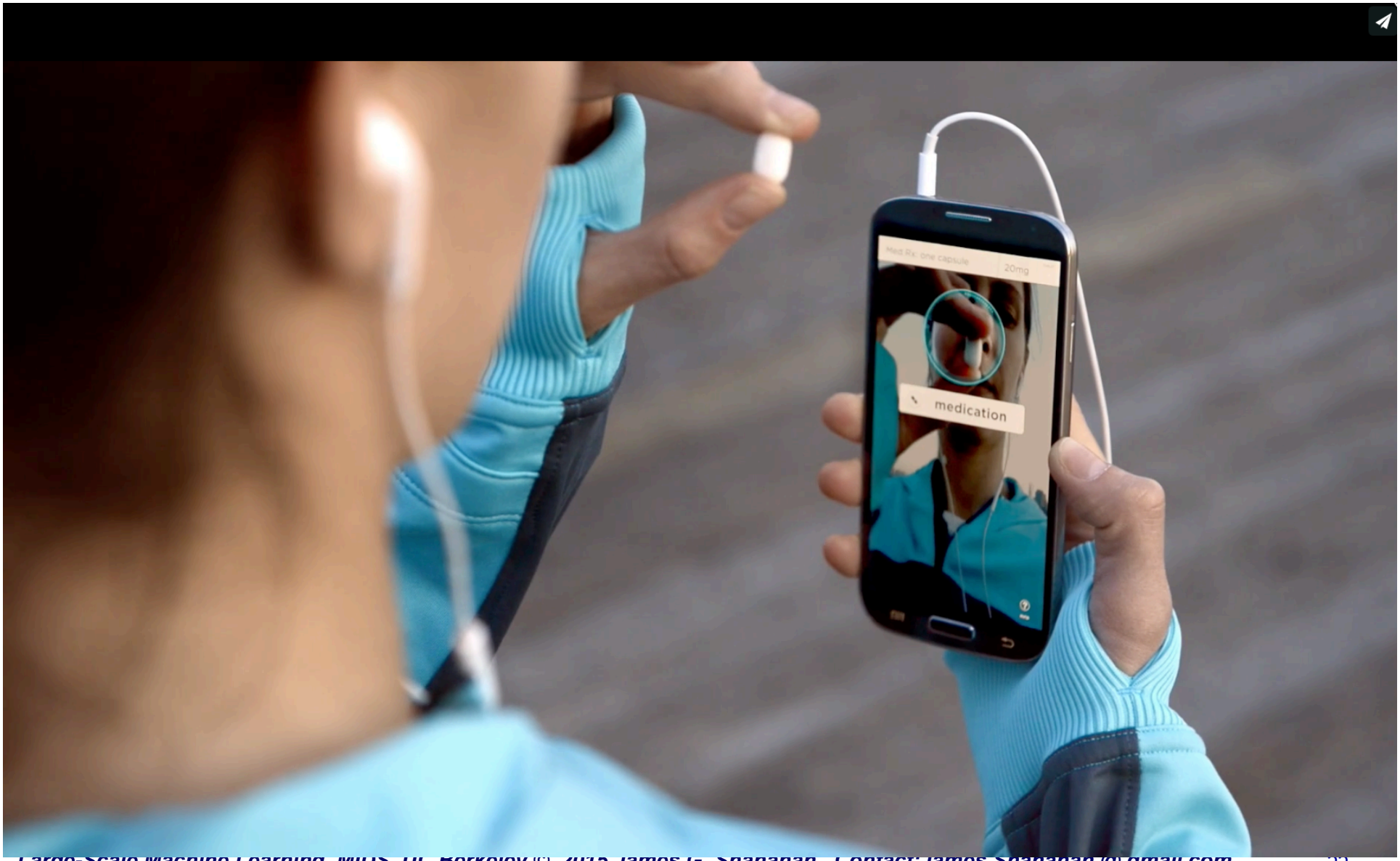
Data Science in Ecommerce

Defining Product Strategy for the optimum product mix

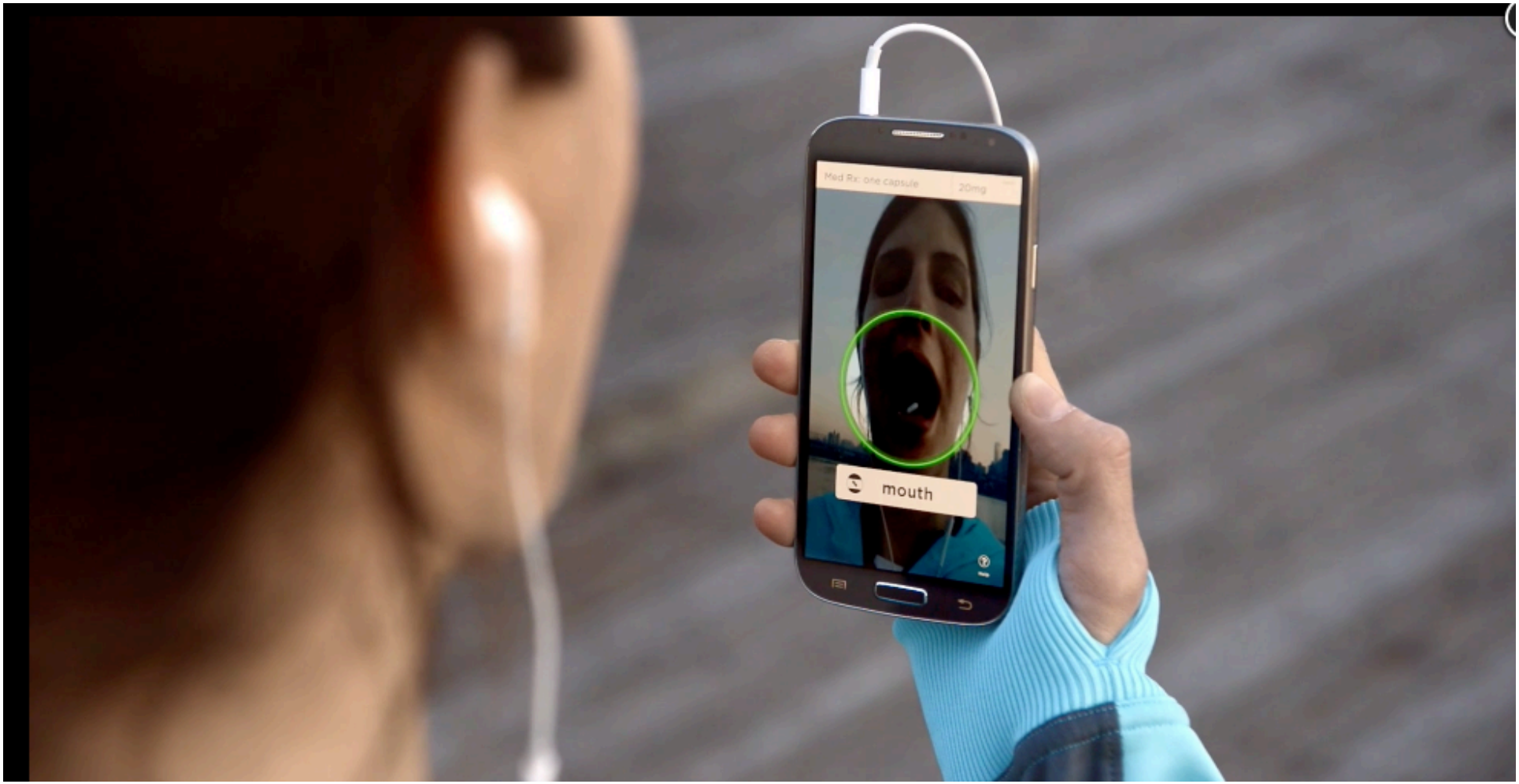
- **Ecommerce, and bricks and mortar businesses**
  - What products should they sell?
  - What price should be offered for the products and when?
- **Data science algorithms help ecommerce businesses define and optimize the product mix.**
  - Every ecommerce business has a product team that looks into the design process where data science algorithms can help the business with forecasting like-
    - What are the loopholes in the product mix?
    - What should they make?
    - How many quantities should be ordered as initial batch from the factory outlet?
    - When should they halt the supply of those products?
    - When should they sell?
- **Data scientists versus Data Analysts**
  - work on advanced predictive and prescriptive analytics
  - whereas data analysts will merely look into the retrospective analysis like

- https://www.aicure.com/

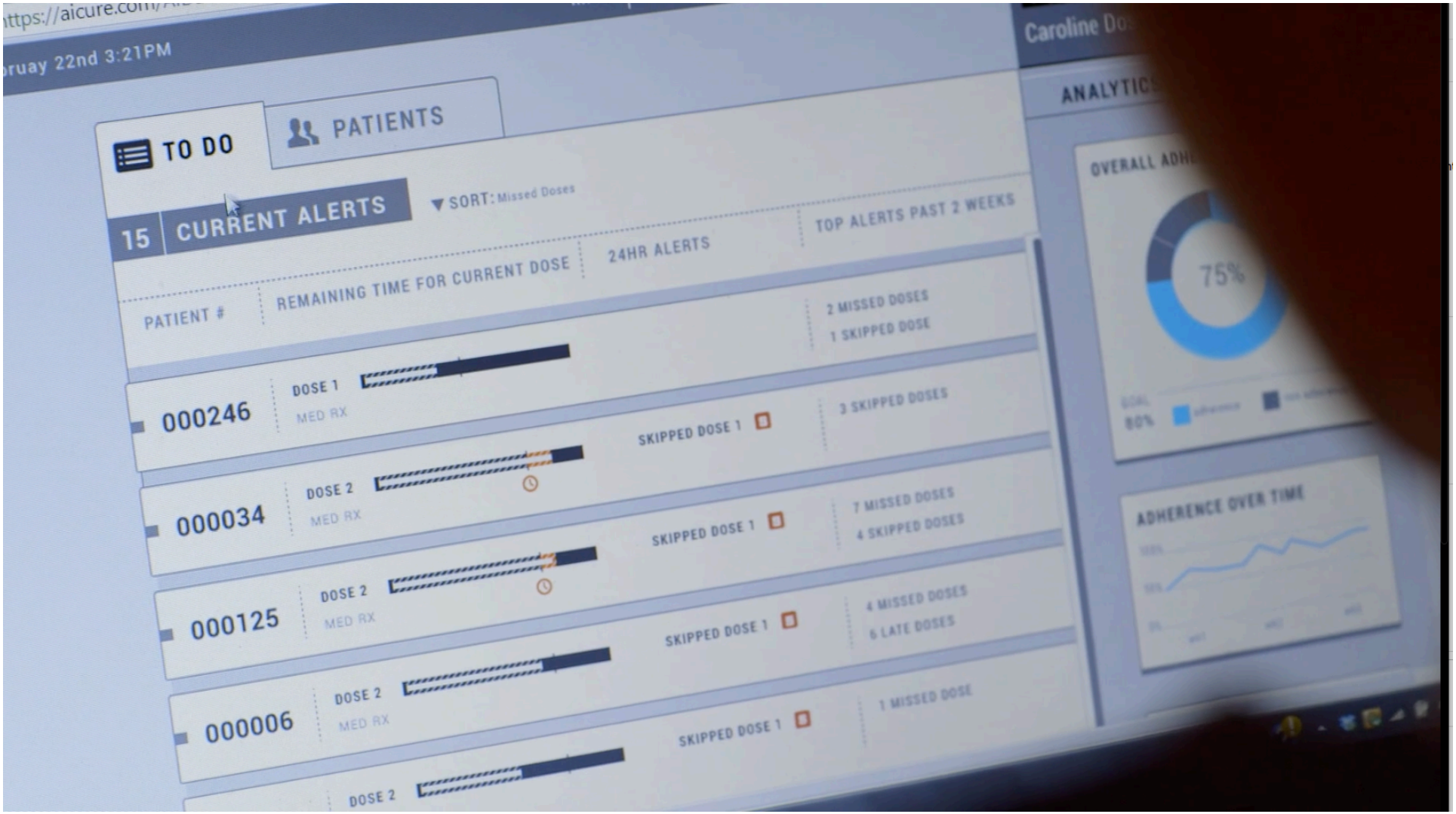
Do you know if your patients are taking their meds?



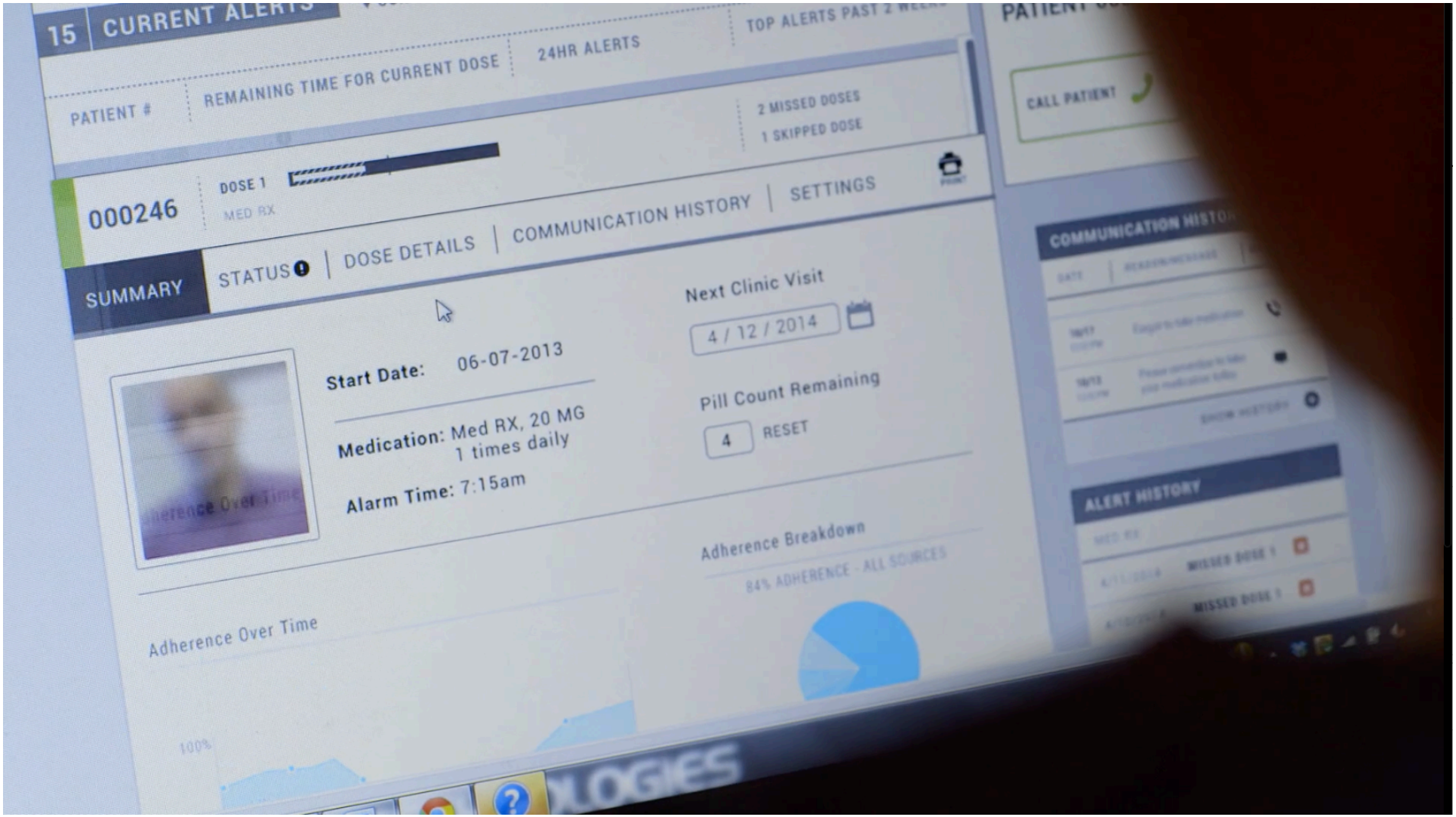
Trust but verify!



Rank patients



Alerts


15 CURRENT ALERTS
PATIENT #
REMAINING TIME FOR CURRENT DOSE
24HR ALERTS
TOP ALERTS PAST 2
2 MISSED DOSES
1 SKIPPED DOSE
000246
DOSE 1
MED RX
SUMMARY
STATUS
DOSE DETAILS
COMMUNICATION HISTORY
SETTINGS
Start Date: 06-07-2013
Medication: Med RX, 20 MG 1 times daily
Alarm Time: 7:15am
Next Clinic Visit
4 / 12 / 2014
Pill Count Remaining
4
RESET
Adherence Breakdown
84% ADHERENCE - ALL SOURCES
Adherence Over Time
100%
CALL PATIENT
COMMUNICATION HISTO
ALERT HISTORY
MISSED DOSE 1
MISSED DOSE 1

As Data Scientist

- **As an early stage startup, what types of tasks (data science centric) would you embark on?**

As a data scientist

- **What would you do?**
- **Infrastructure**
  - Build data pipelines to capture consumer behavior
  - Understand what data is collected? Identify anomalies
  - Instrument and collect right data
- **EDA: 3 times versus 4 times. Side-effects**
- **Classifier for patients who take and don't take**
  - Explain using the independent variables
- **Join other thirdparty data sources (fitbit)**
- **Proactive communication with patiences: Predictive classifiers: detect and notify with reminders (phone call intervention)**
- **Tracking right person:**
  - finger print;
  - Use camera to identify and person
- **Other compliance**
  - Different types of medication: inhaler versus pills (recommendation system); suggest different medical plans
- **Data export: ownership (patient)**
- **Measures to secure data**
- **AB Testing**
- **Wearables strategy (intelligent pill box); sprectrometry**

Text Processing

Deep Learning based

CNN
RNN

http://aylien.com/

Finding Friends

Linking other things such as groups

- **Growing body of research captures dynamics of social network graphs**

  [Latanzi, Sivakumar '08] [Zheleva, Sharara , Getoor '09] [Kumar, Novak, Tomkins '06] [Kossinets, Watts '06] [L., Kleinberg, Faloutsos '05]

- **What links will occur next?** [LibenNowell, Kleinberg '03]
  - **Networks + many other features:**

    Location, School, Job, Hobbies, Interests, etc.

facebook Search Home Profile Account

Find friends from different parts of your life

Use the checkboxes below to discover people you know from your hometown, school, employer and more.

Hometown
- [ ] Indianapolis, Indiana

Enter another city

Current City
- [ ] Indianapolis, Indiana

Enter another city

High School
- [ ] North Central High School

Enter another high school

Mutual Friend

Enter a name

College or University
- [ ] Martin University

Enter another college

Employer
- [ ] ARIES GRAPHIC DESIGN

Enter another employer


Judy Pyles
36 mutual friends
Add Friend

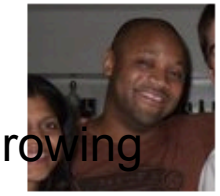

Rocky Campbell
41 mutual friends
Add Friend


Laura White
12 mutual friends
Add Friend

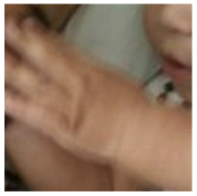

King Ro Conley
59 mutual friends
Add Friend


Dillon Rhodes
43 mutual friends
Add Friend

Rhonda Landrum
54 mutual friends
Add Friend


David Corbitt
90 mutual friends
Add Friend


Eric Bettis
15 mutual friends
Add Friend


Eric Hughes
110 mutual friends
Add Friend


Marki Ann
26 mutual friends
Add Friend


Michael Pugh
21 mutual friends
Add Friend


Lisa Williams
22 mutual friends
Add Friend


LouieBaur Digg
39 mutual friends
Add Friend


LaTonya Mayberry Bynum
51 mutual friends
Add Friend


Durece Johnson
2 mutual friends
Add Friend


Kendale Adams
64 mutual friends
Add Friend


Bruce T. Caldwell
143 mutual friends
Add Friend


Angela Blackwell Miller
61 mutual friends
Add Friend


Landon Montel

Kevin Brown


Stanley F. Henry

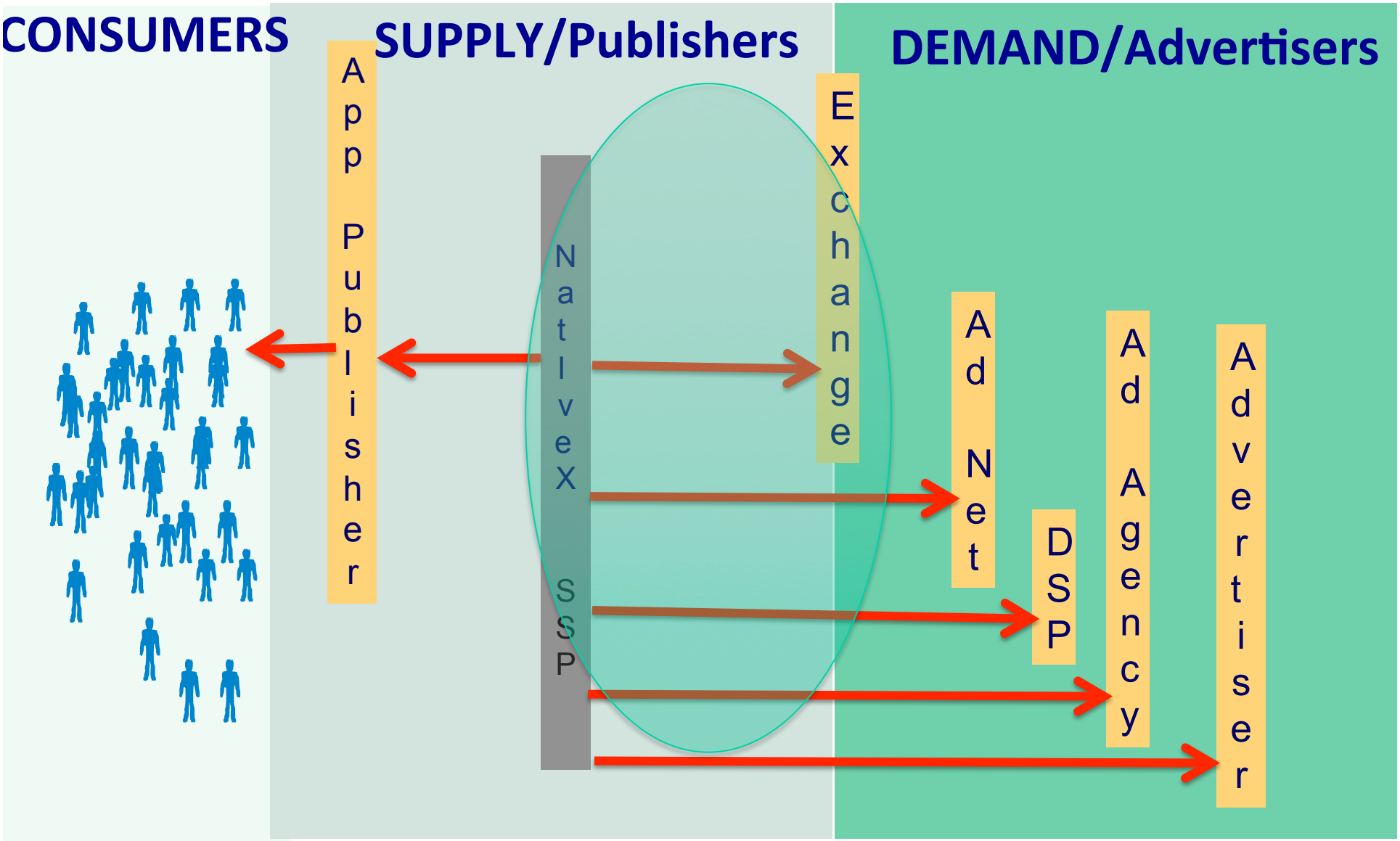

Saundria Mccrackin


Ebonye X-Endsley

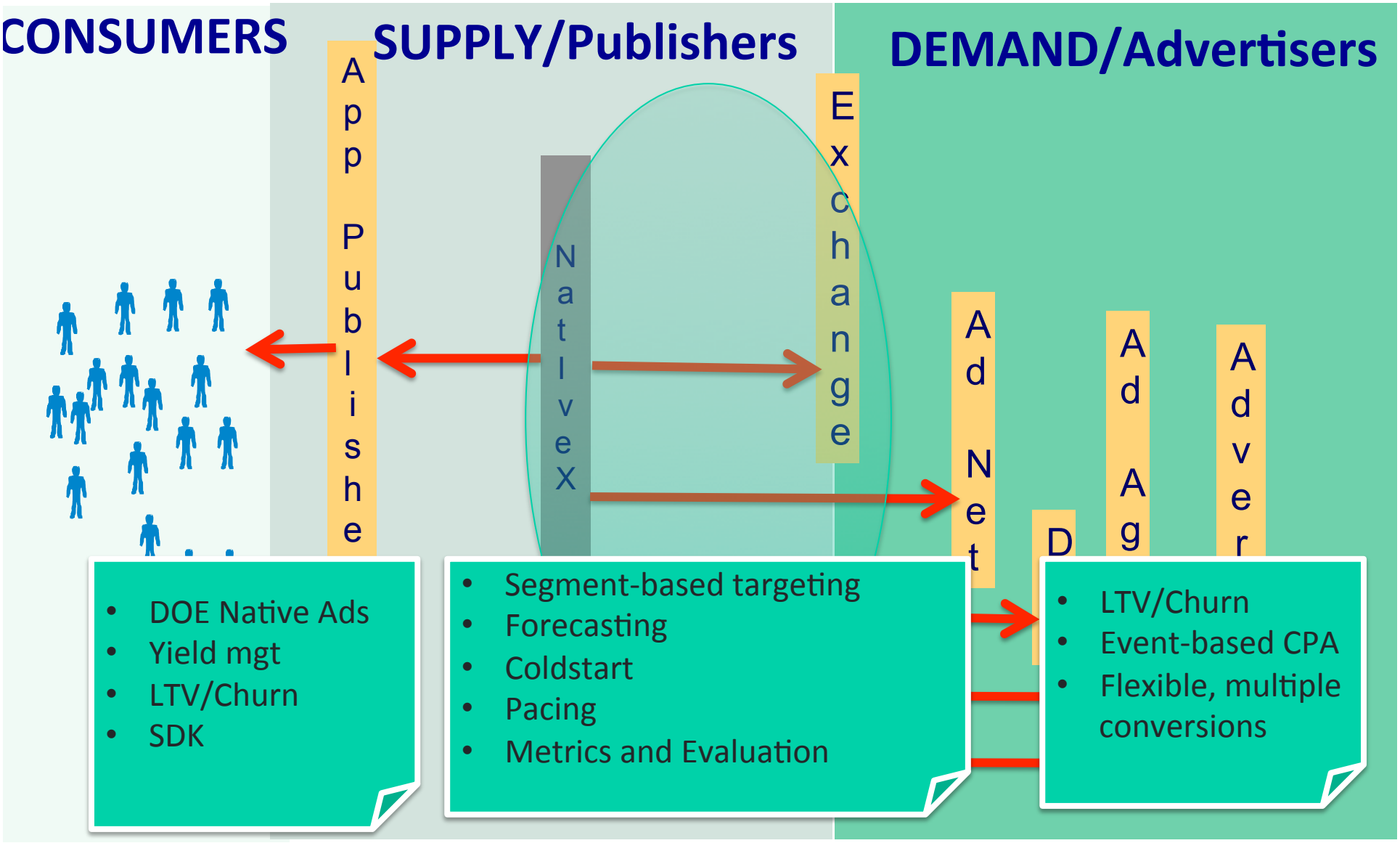

Anita Hawkins

Growing

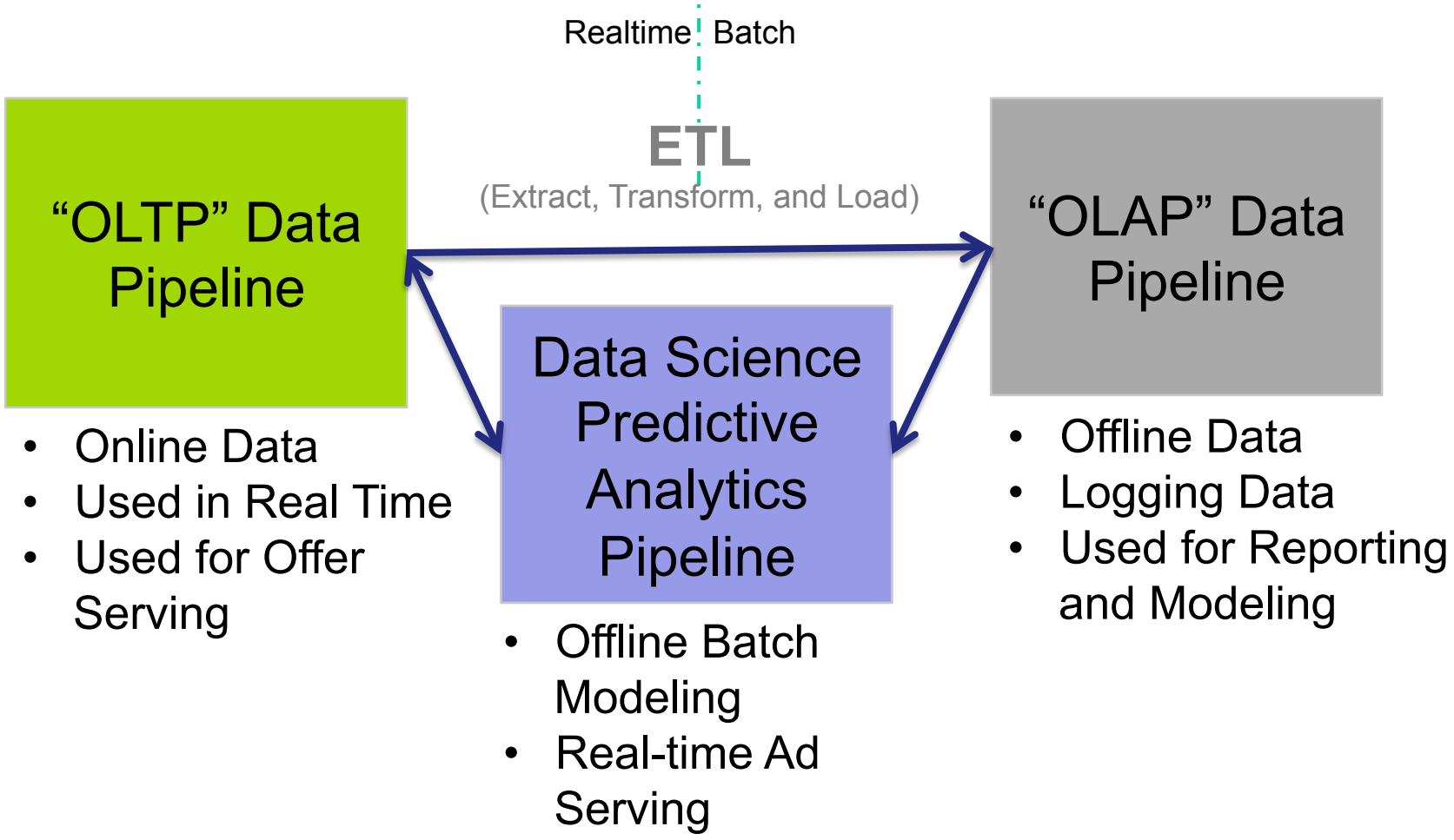
NativeX: Art and Science of Native Mobile Advertising



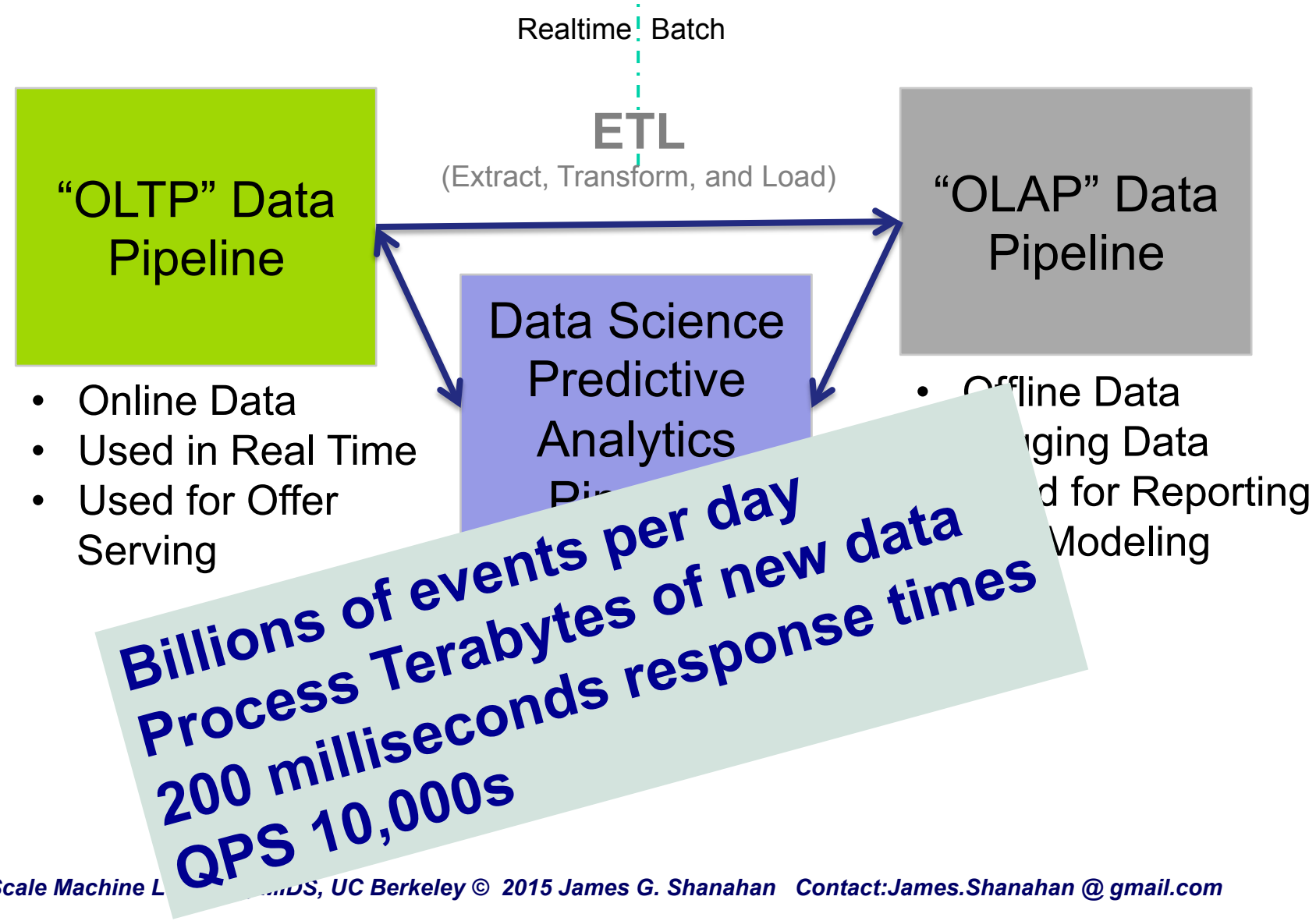
NativeX: Art and Science of Native Mobile Advertising



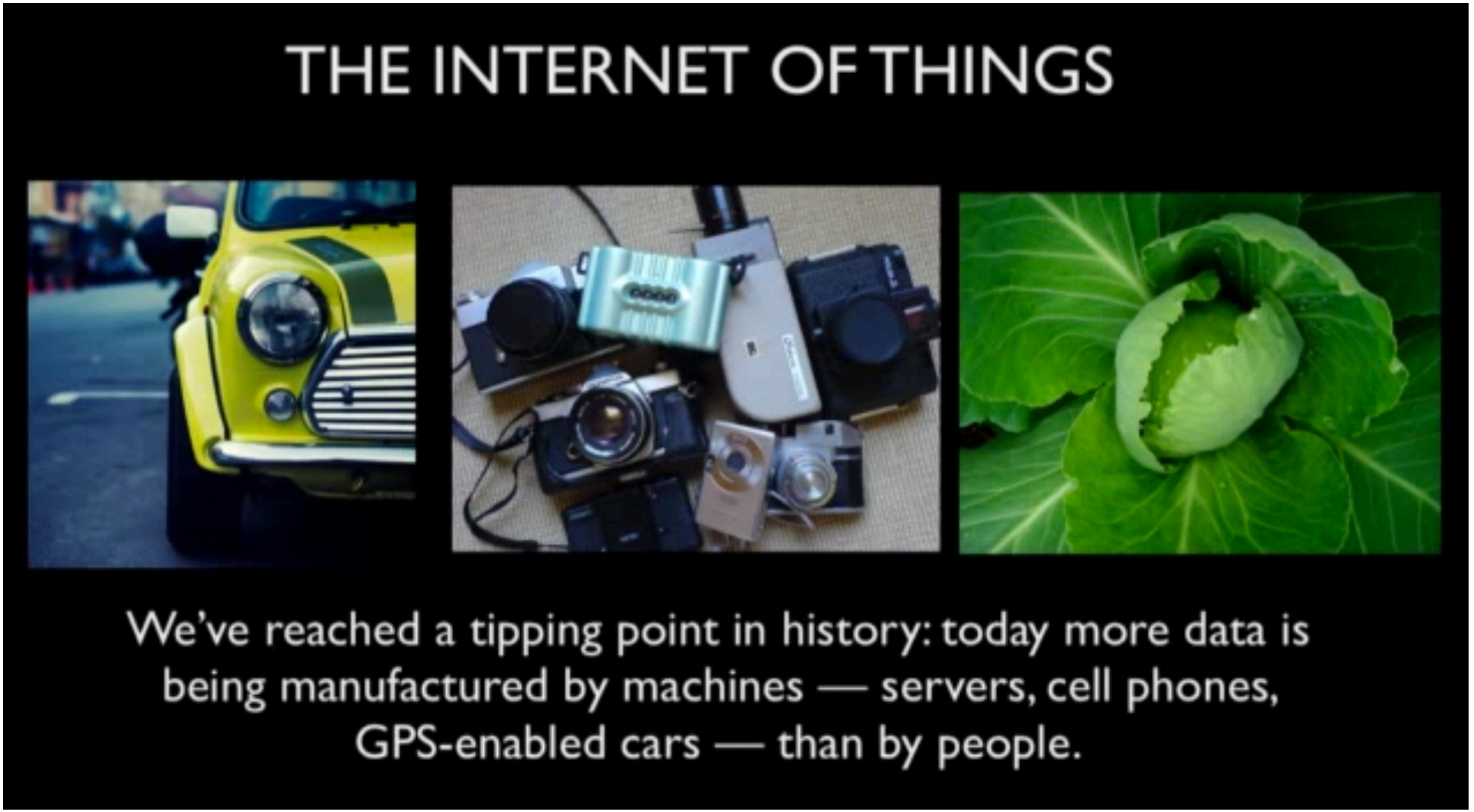
Ad serving data pipelines



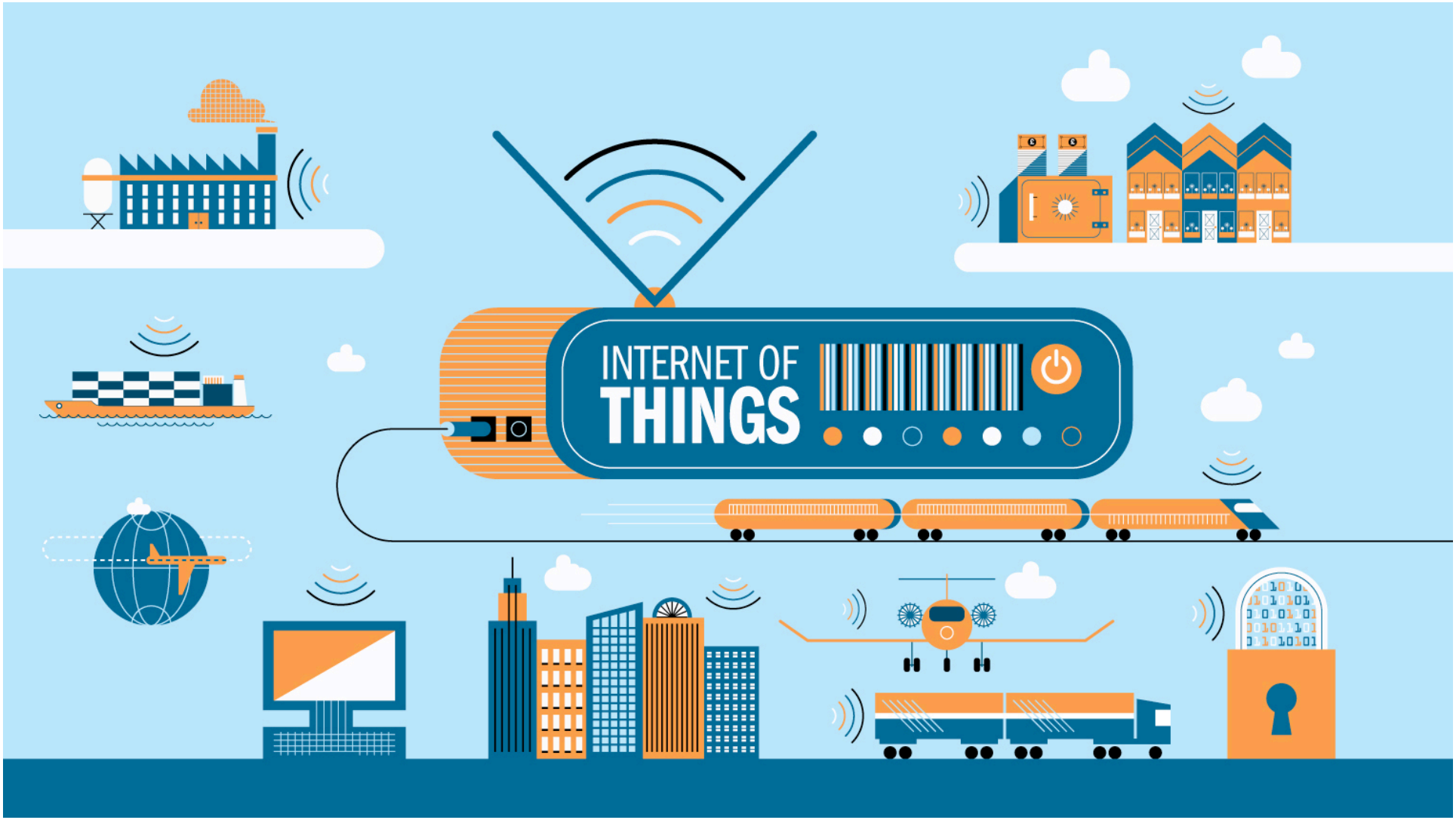
Ad serving data pipelines



Tipping point: Humans no longer the center to the data universe



IoT



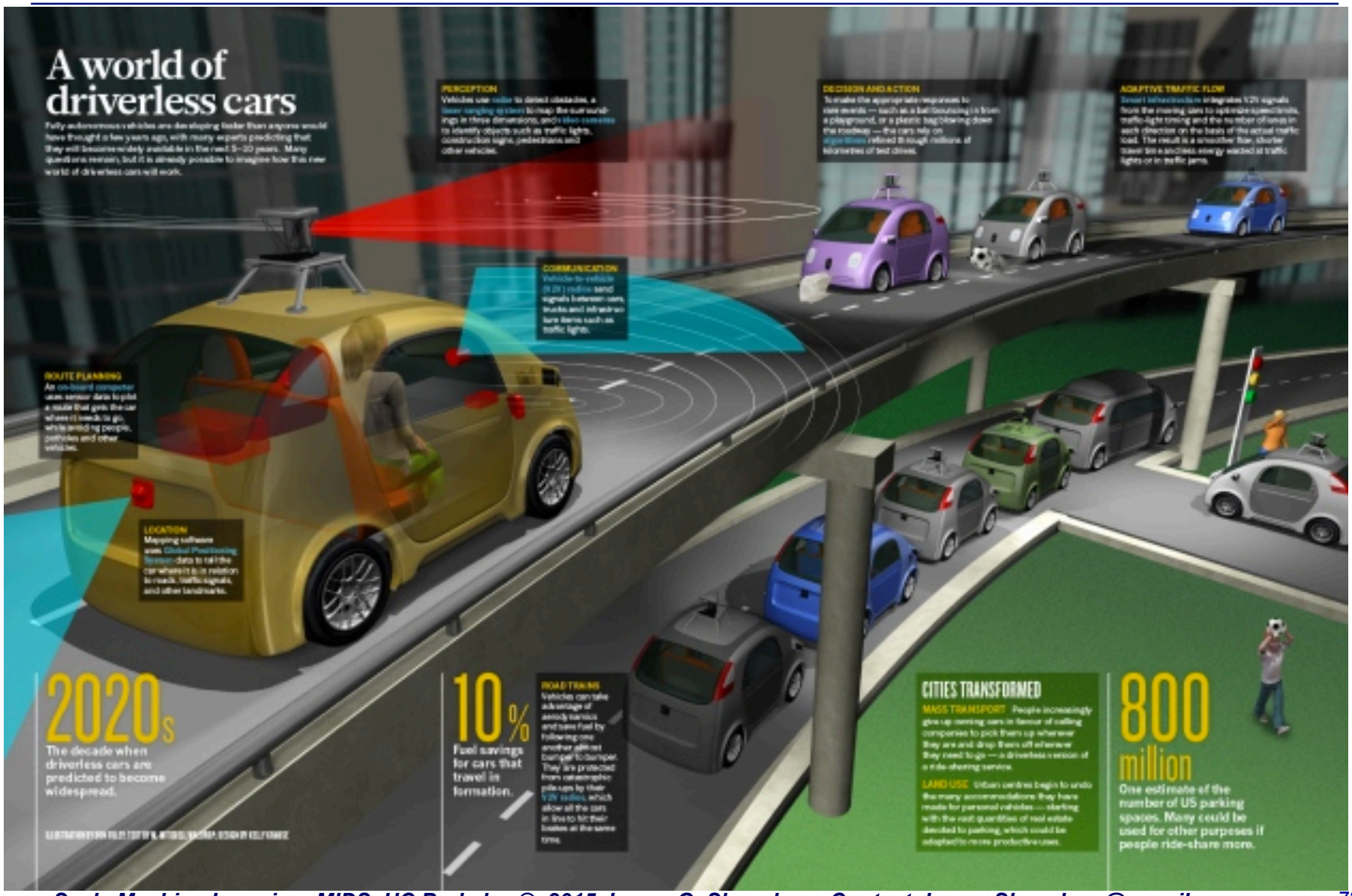

smart home

Target Distribution Center

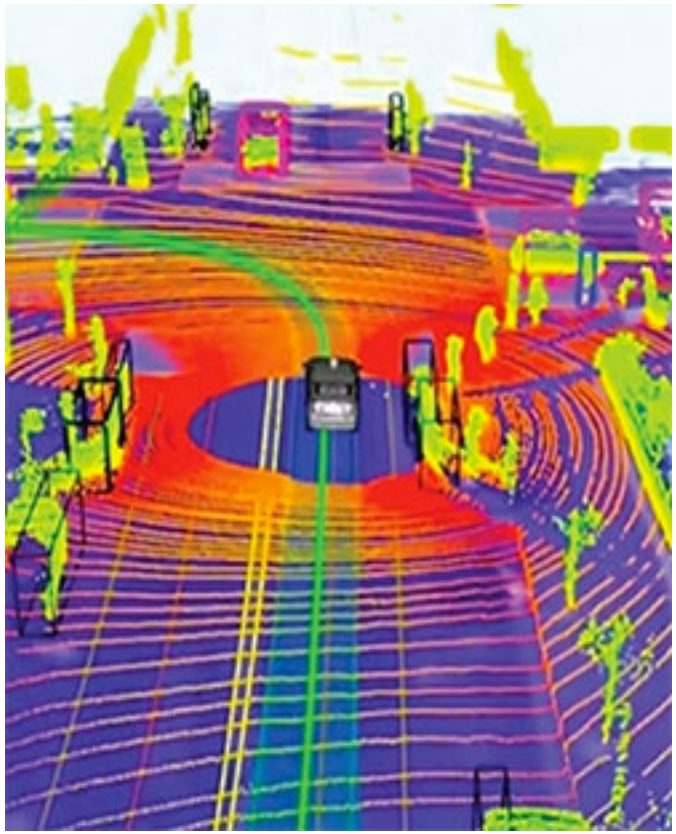
- 

Products are labeled and palletized, then placed on an assembly line where they eventually end up on shelves to be pulled for orders at Target Distribution Center, Friday April 12, 2013, in Denton.

Autonomous Vehicles



Autonomous Vehicles



http://www.rand.org/pubs/research_briefs/RB9755.html

An image of what Google's self-driving car sees when it makes a left turn.

DITCH THE DRIVER

1.24 million
traffic fatalities every year worldwide

90%
of all accidents are due to driver error

4 US states
and the District of Columbia have passed laws to allow driverless cars on their roads

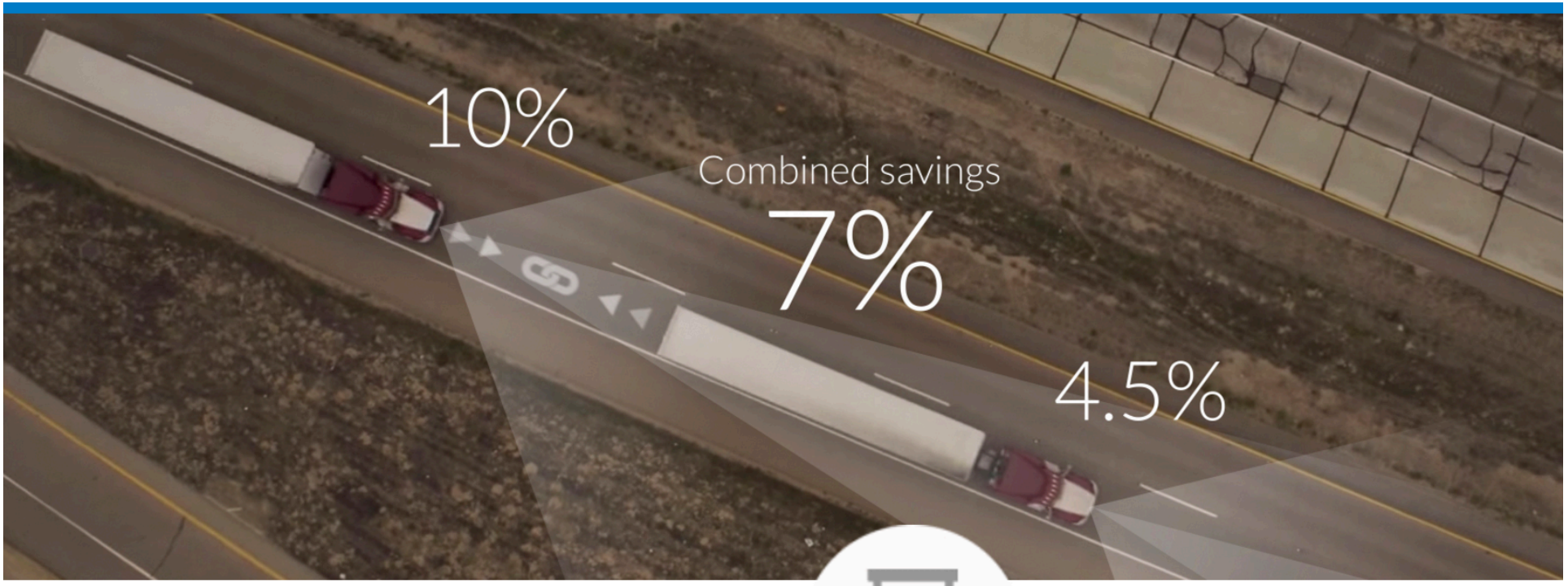
800Million parking spots in US

http://www.nature.com/news/autonomous-vehicles-no-drivers-required-1.16832

http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics

Save fuel, Safer logistics

Peloton



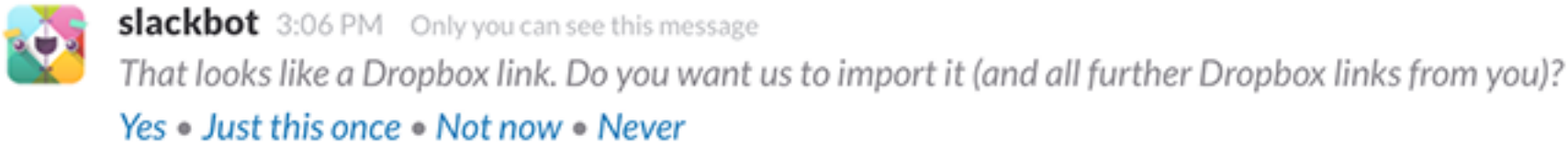
http://peloton-tech.com/

Conversational UI

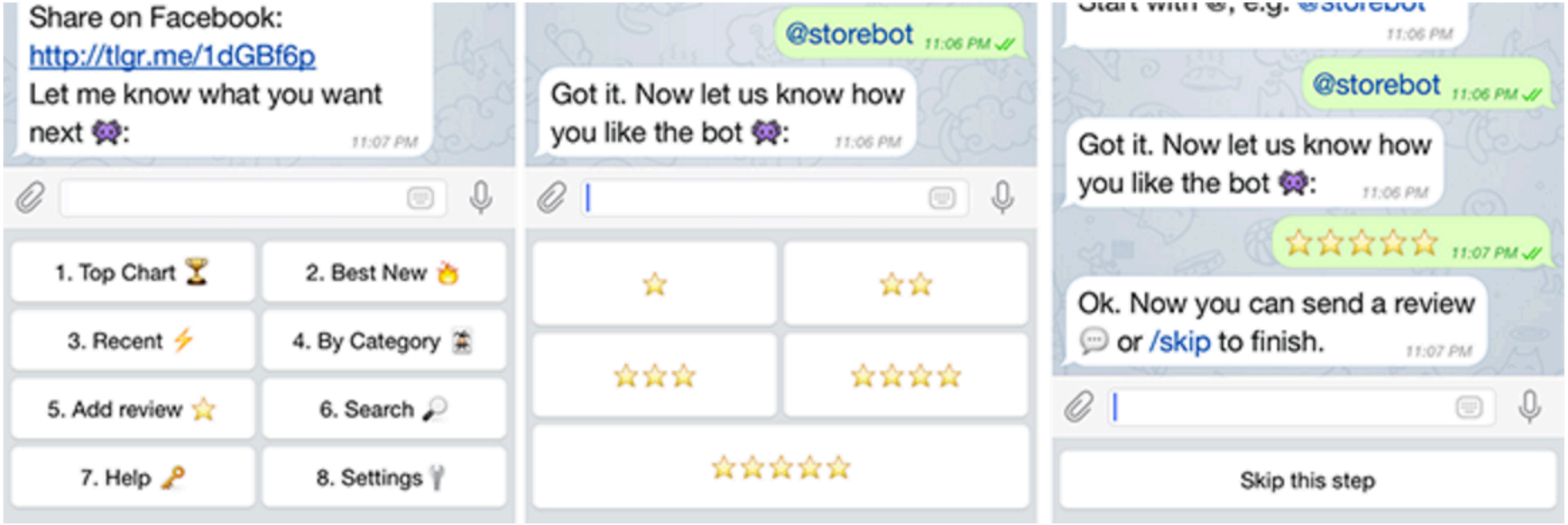
- We're witnessing an explosion of applications that no longer have a graphical user interface (GUI).
- They've actually been around for a while, but they've only recently started spreading into the mainstream.
- They are called bots, virtual assistants, invisible apps.
- They can run on Slack, WeChat, Facebook Messenger, plain SMS, or Amazon Echo.
- They can be entirely driven by artificial intelligence, or there can be a human behind the curtain.

Conversational UI

- ..



With all the advantages of a conversational interface, some tasks (like multiple selections, document browsing, and map search) are better performed with a pointing device and buttons to click. There's no need to insist on a purely conversational interface if your platform gives you a more diverse toolbox. When the flow you present to your user gets narrowed down to a specific action, a simple button can work better than typing a whole line of text.



Telegram uses pop-up buttons for discovery and for shortcuts.

Conversational UI

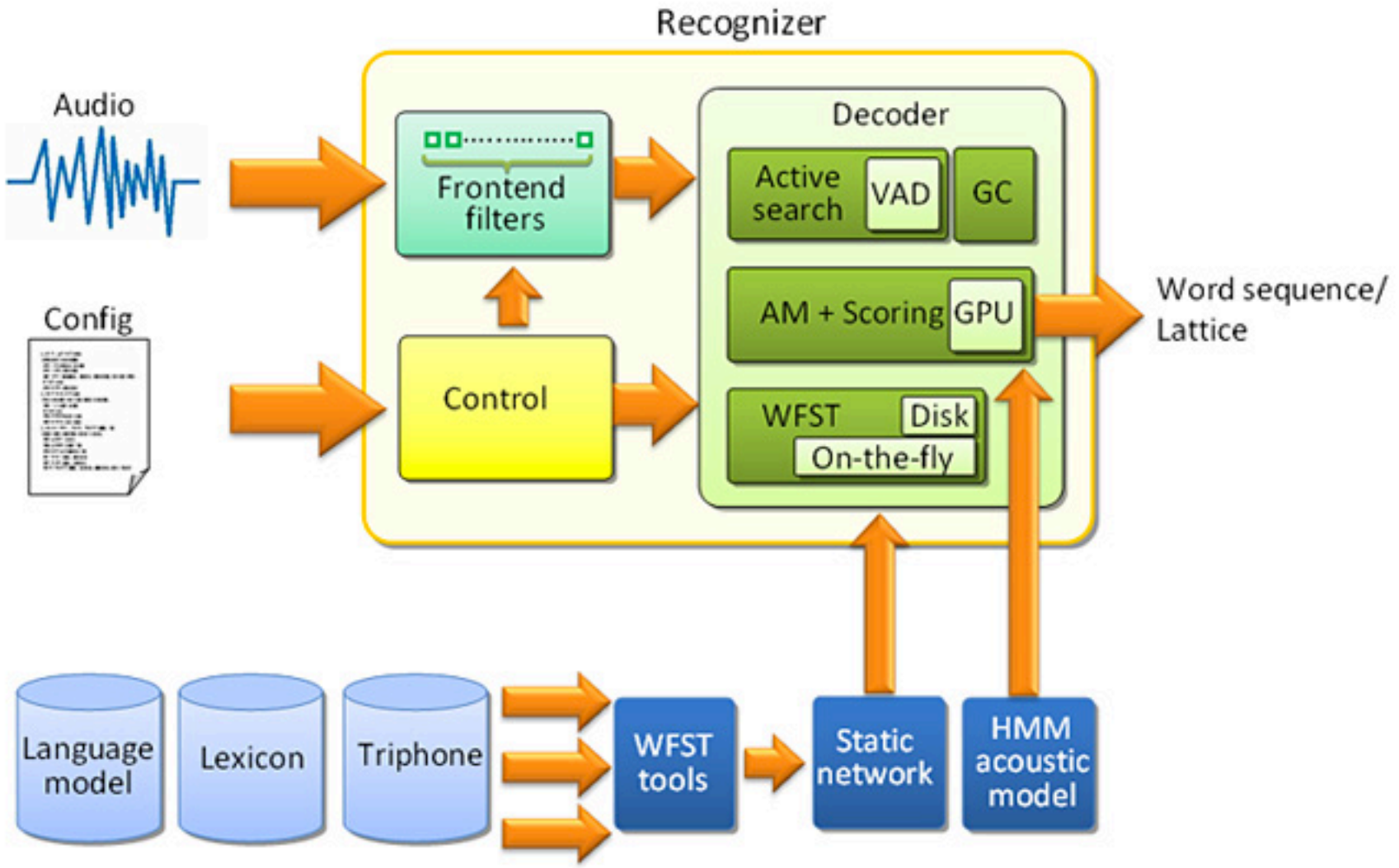
- **Amazon Echo is controlled by voice, but has a companion app.**

Speech Recognition Breakthrough for the Spoken, Translated Word

- **Published on Nov 8, 2012**
- **Chief Research Officer Rick Rashid demonstrates a speech recognition breakthrough via machine translation that converts his spoken English words into computer-generated Chinese language. The breakthrough is patterned after deep neural networks and significantly reduces errors in spoken as well as written translation.**
- **For more information on Speech Recognition and Translation, visit**
  - http://www.microsoft.com/translator/skype.aspx
- **Excellent Video (please watch all this video!)**
  - https://www.youtube.com/watch?v=Nu-nlQqFCKg (Minute 7:11)
  - English text (ASR) → Chinese Text → Text to speech system (sound like english speaker)

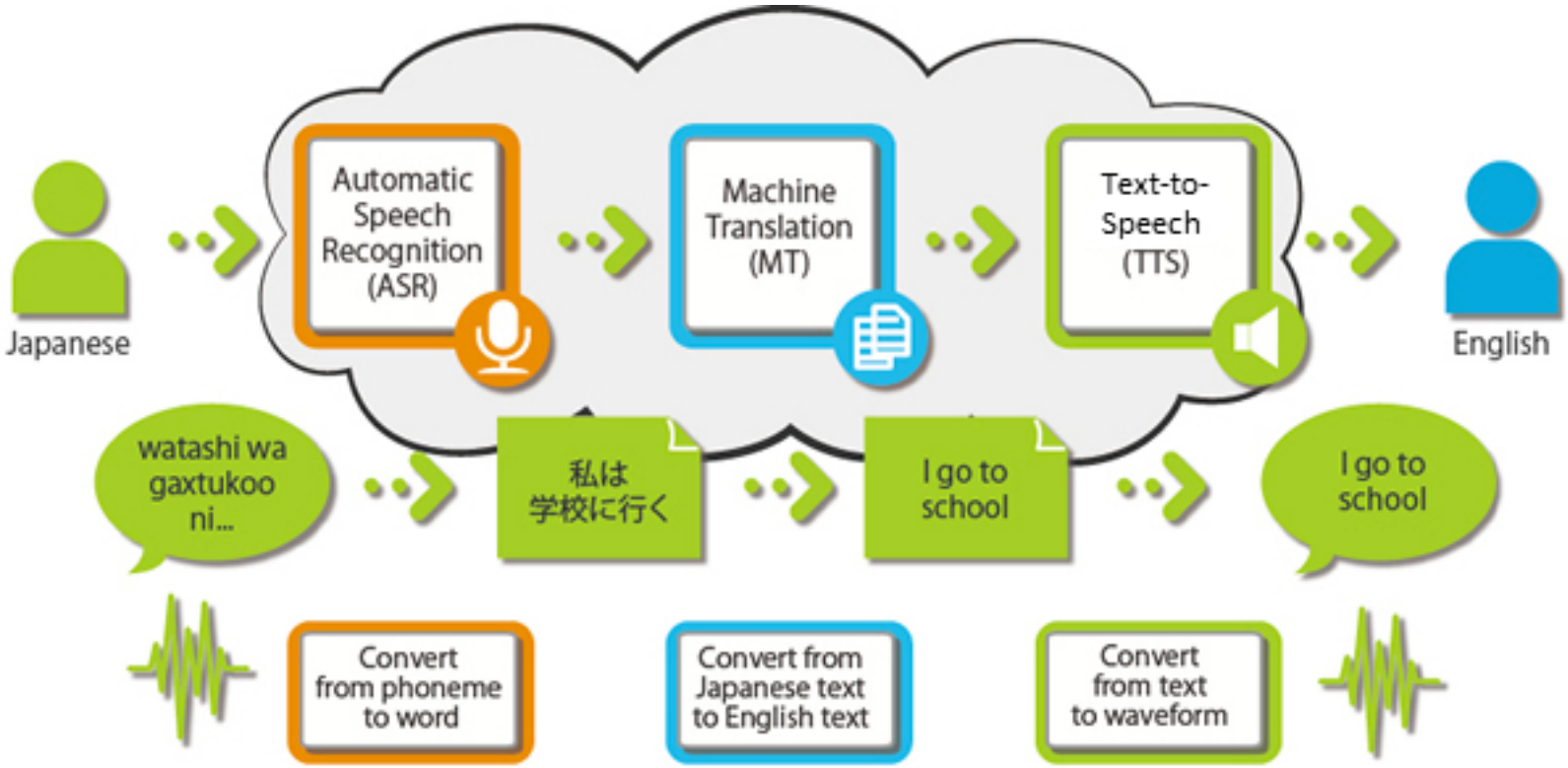

https://www.youtube.com/watch?v=Nu-nlQqFCKg
(4) MIDS-MLS-2015
nbviewer.ipython.org
Stanford Machine Le
Getting Started
Statistical Analysis
eBay/Google 2013:
Tube Red
English text (ASR) → Chinese Text → Text to speech system (sound like english speaker)
I'm speaking in english and hopefully you'll hear me speaking chinese in my own voice.
Again, the results are not perfect.
They are in fact quite a few
我是说英语，希望你能听我说中国人自己的声音。
再次，结果并不完美。

ASR (Audio signal → word sequence)



HMM, Deep Learning, Language models

Japanese to English

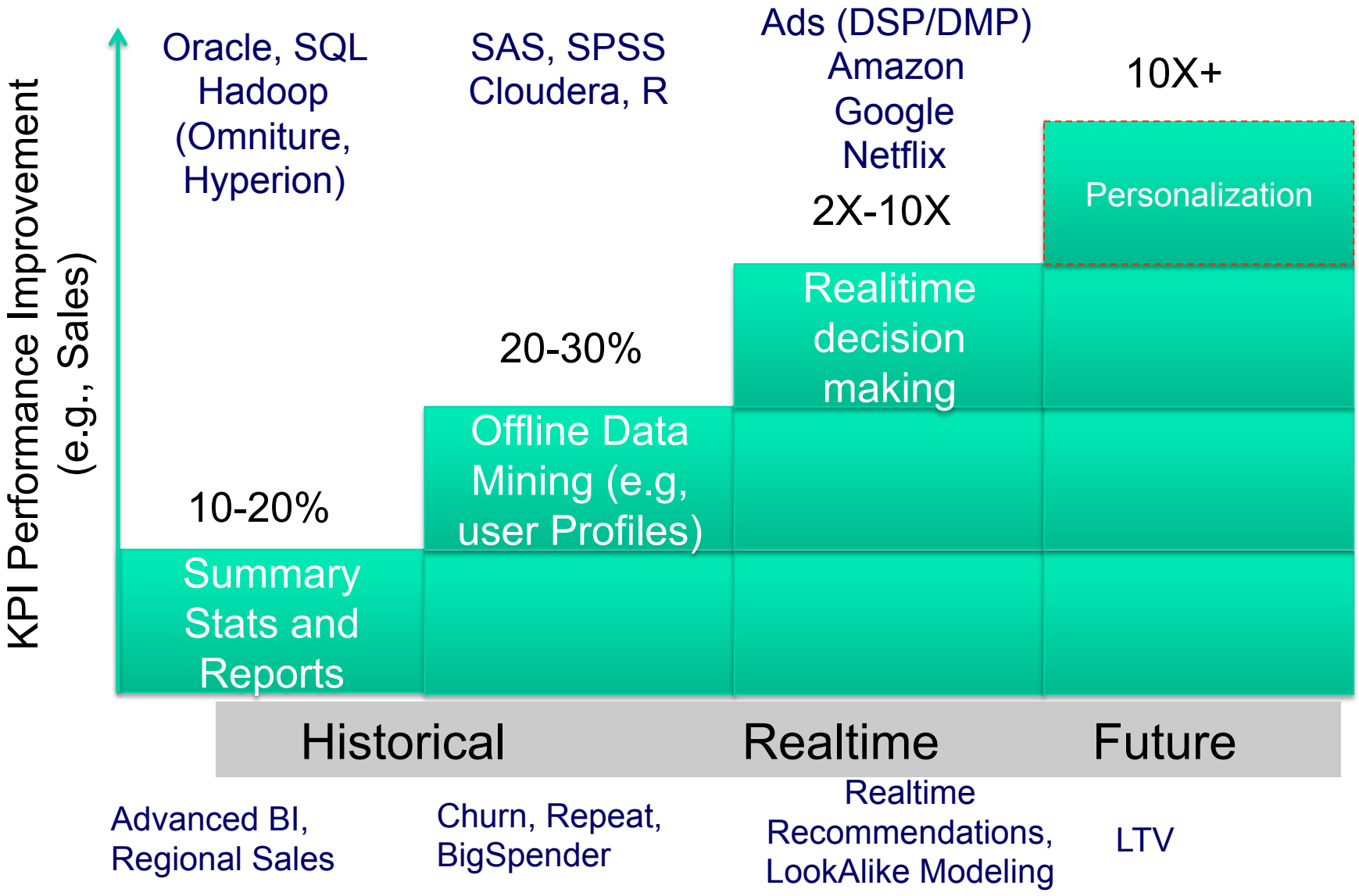


http://www.ustar-consortium.com/research.html

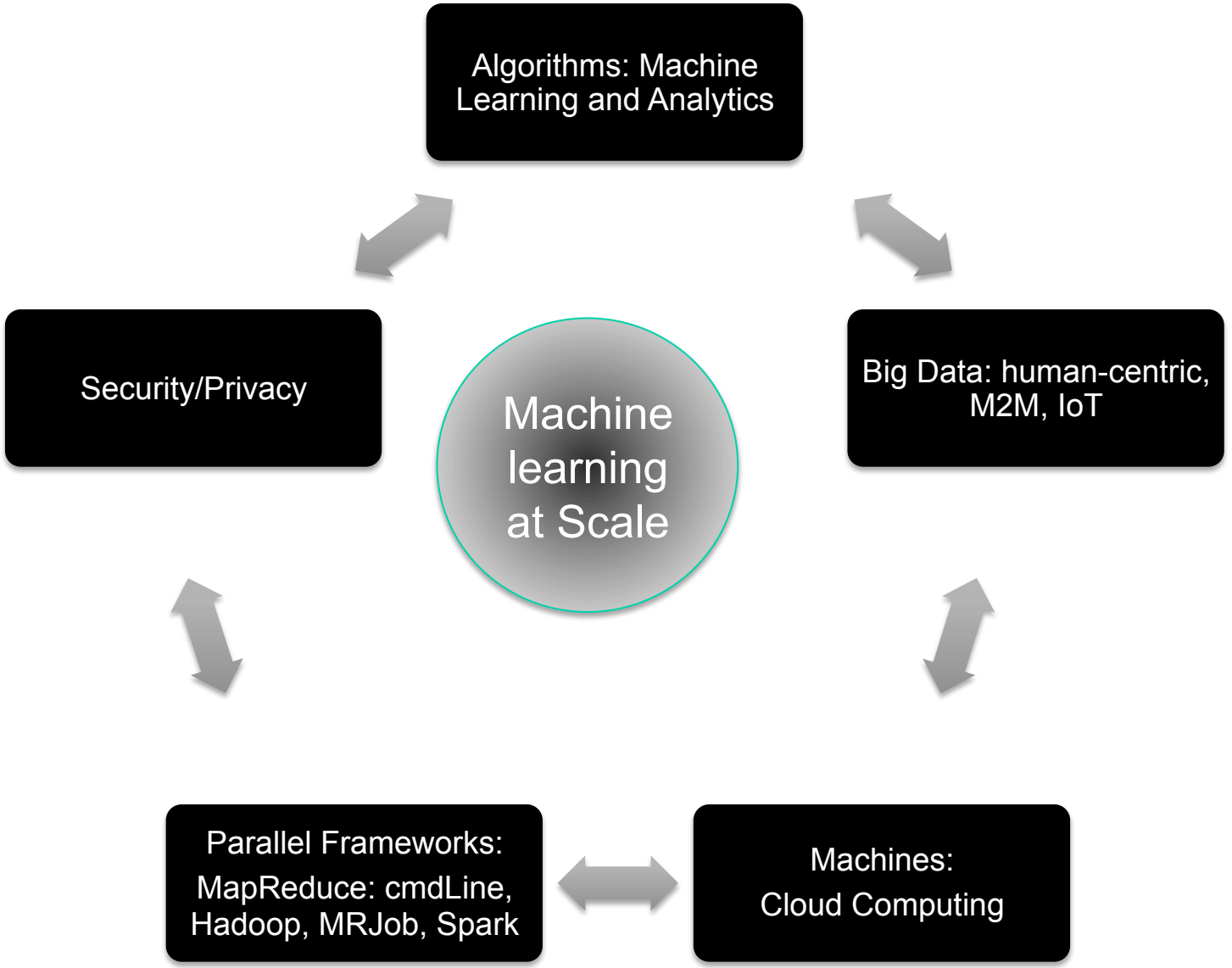
Impact of deep learning in speech technology

Skype to get 'real-time' translator

Managers and CEOs see the value of DA
Data (Science) improves KPIs dramatically



Machine learning at Scale



150,000 Data Scientists needed in US

With such enormous potential to change the world, it will come as no surprise that data scientists are in huge demand

[McKinsey Report on Big Data 2011]

Top 10 Best Jobs in the US as of 2/2016

THE BEST JOBS IN AMERICA RIGHT NOW

What makes a job not just good but great? According to Glassdoor.com, it's a combination of how much you make, the demand for your skills, and how easily you can advance in your field.

By those measures, Glassdoor's 2016 best job in America is data scientist, with a median base salary of nearly $117,000 and more than 1,700 openings right now. Technology jobs in general have a robust presence on the list, as do finance jobs.

Here is the top 10 list: (1) data scientist, (2) tax manager, (3) solutions architect, (4) engagement manager, (5) mobile developer, (6) HR manager, (7) physician assistant, (8) product manager, (9) software engineer, and (10) audit manager. –M.C.W.

16 MONEY.COM

ATT HARRISON CLOUGH

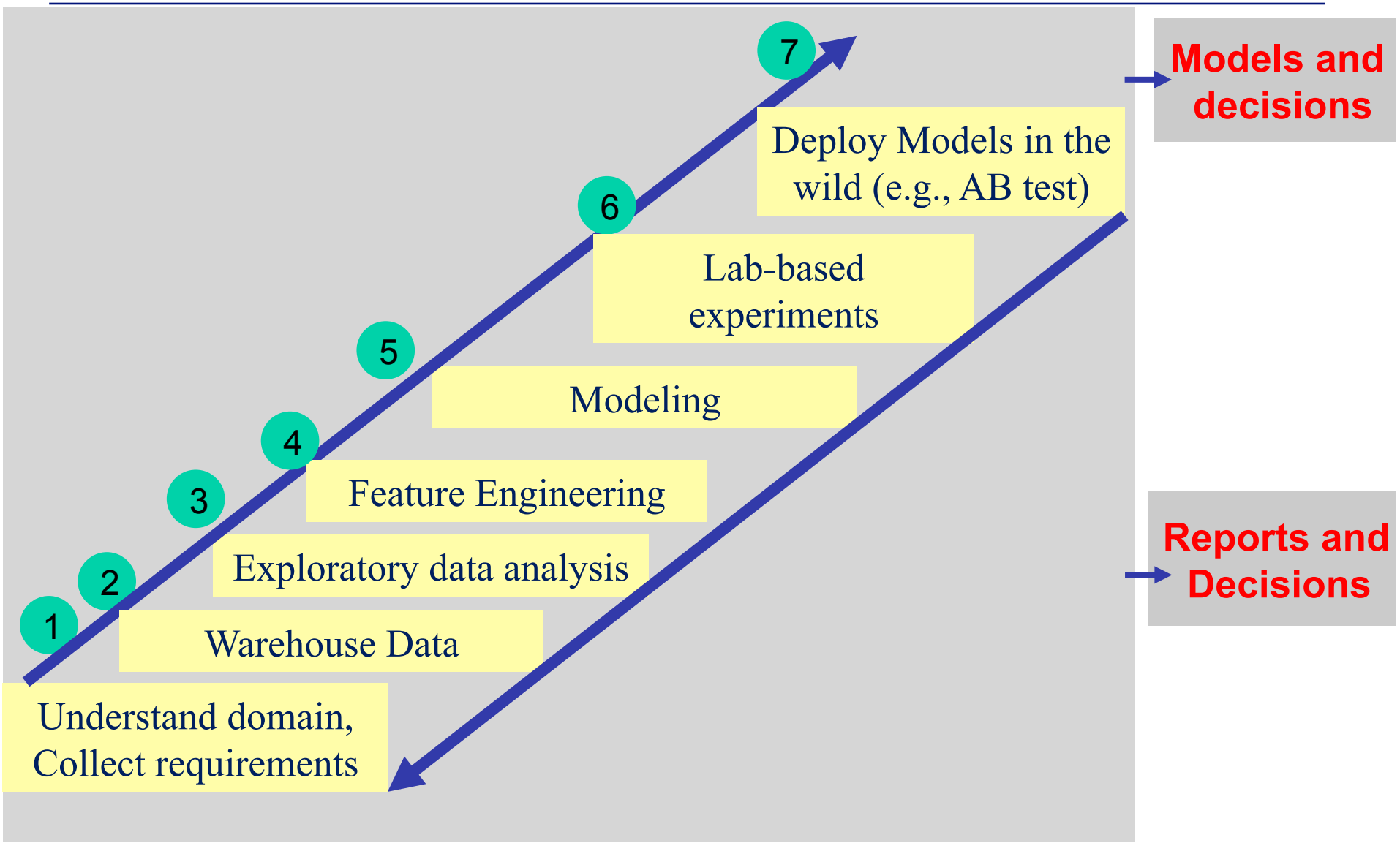
How much you make
The demand for your skills
How easily you can advance
117K Median salary
1,700 openings right now

Lecture Outline

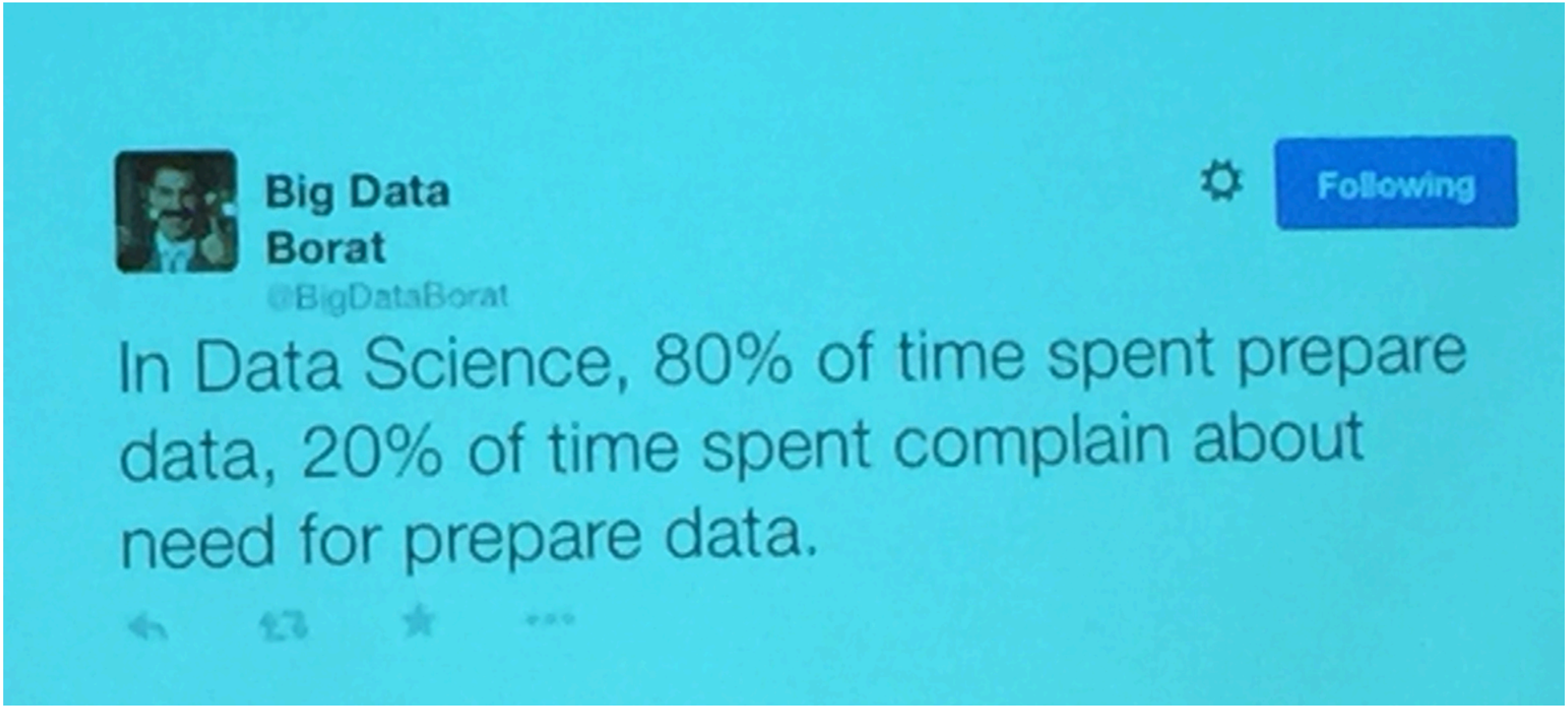
- Google Doc and Group
- Welcome & Class Introductions
- Big Data and Applications
- Course introduction
- Class logistics
- Systems (part 1 of N)

- **Big data applications described in the previous sections have generated new opportunities and business needs**
- **This presents a literally a massive opportunity and challenge for machine learning**
- **ML does not live in a vacuum (although it used for me in grad school)**
- **These days ML requires a sophisticated ecosystem.**

Typical Abstract Data Analytics Pipeline



Where does time go in large scale machine learning?



- **A popular question today in the advent of big data**

- **More data scientists versus more data (that another way to ask about do you want a model bias-variance)**

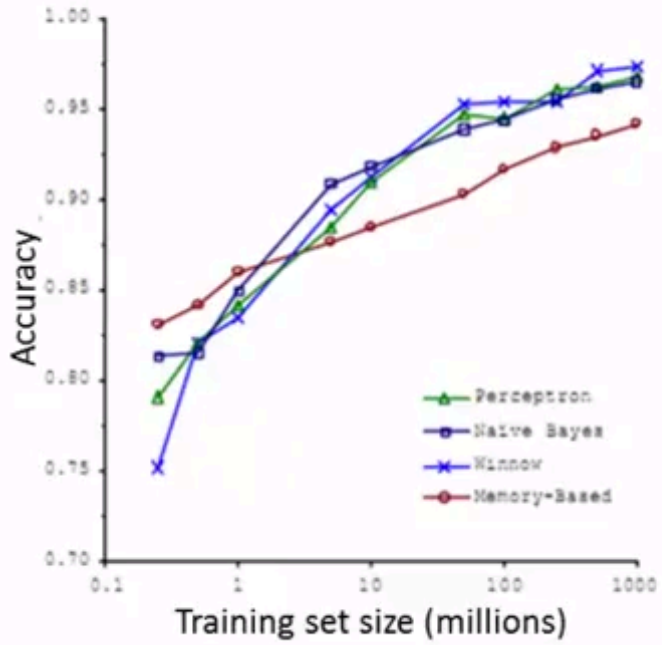
- **Empirical studies suggest more data….But…..**

More data or more data science?

Machine learning and data

Classify between confusable words.
E.g., {to, two, too}, {then, than}.

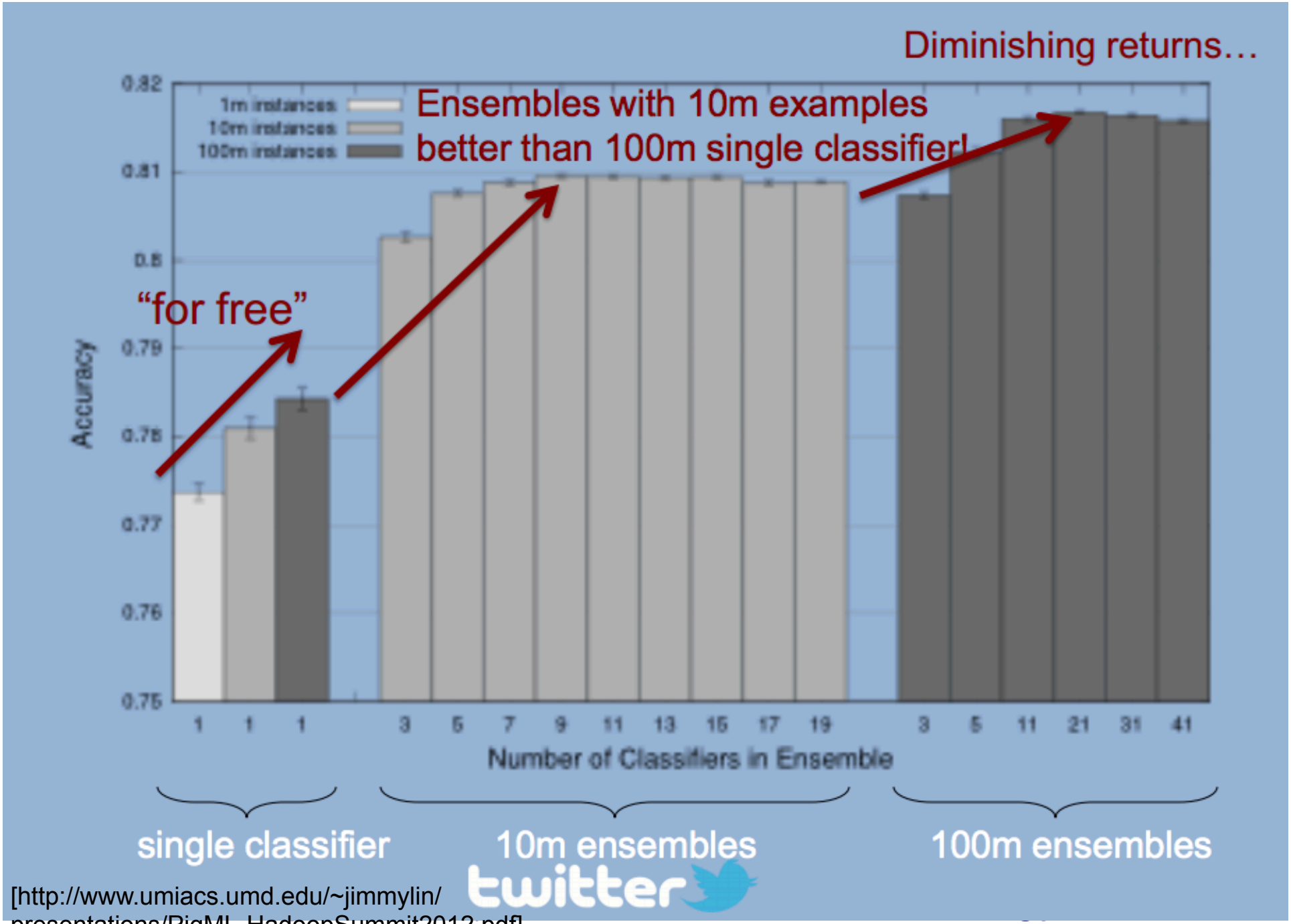
For breakfast I ate _____ eggs.



"It's not who has the best algorithm that wins.
It's who has the most data."

[Figure from Banko and Brill, 2001]

Andrew Ng



[http://www.umiacs.umd.edu/~jimmylin/presentations/PigML-HadoopSummit2012.pdf]

More data or more data scientists

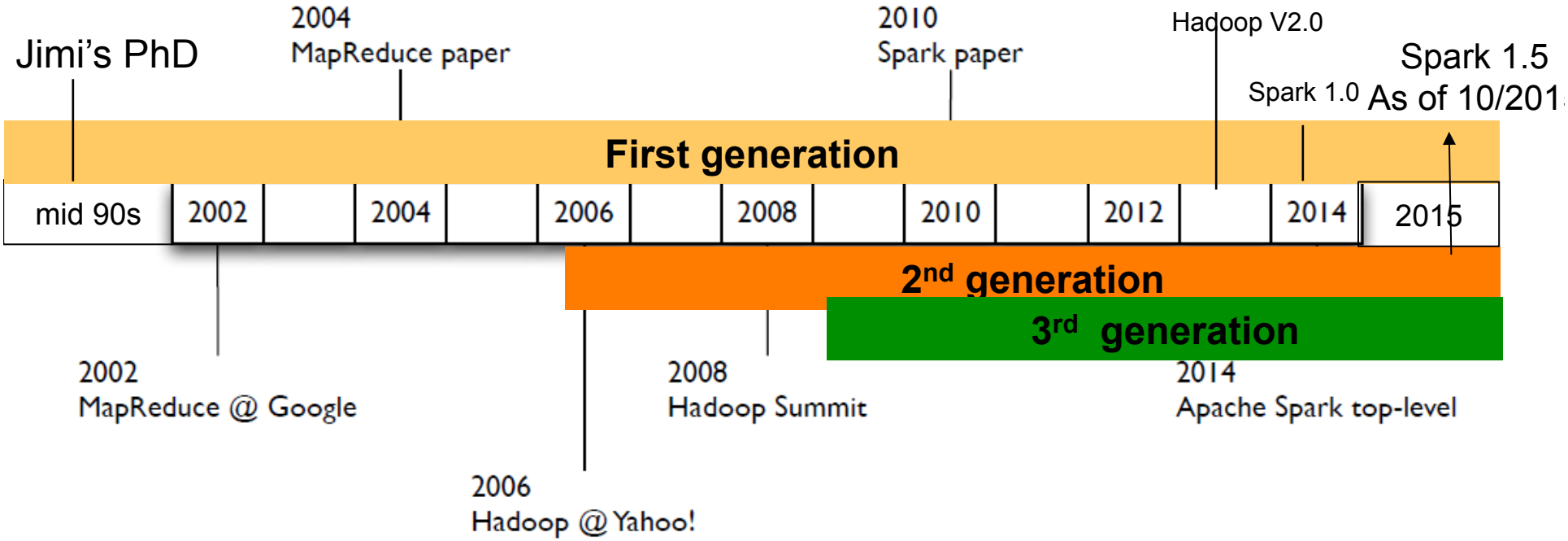
- **Reduce Bias (more scientists → increased variance (overfit))**
  - Manage bias via models and features
  - Linear model on nonlinear data?
  - Polynomial model on linear data
- **Variance (more data → reduce variance)**
  - Manage variance with data

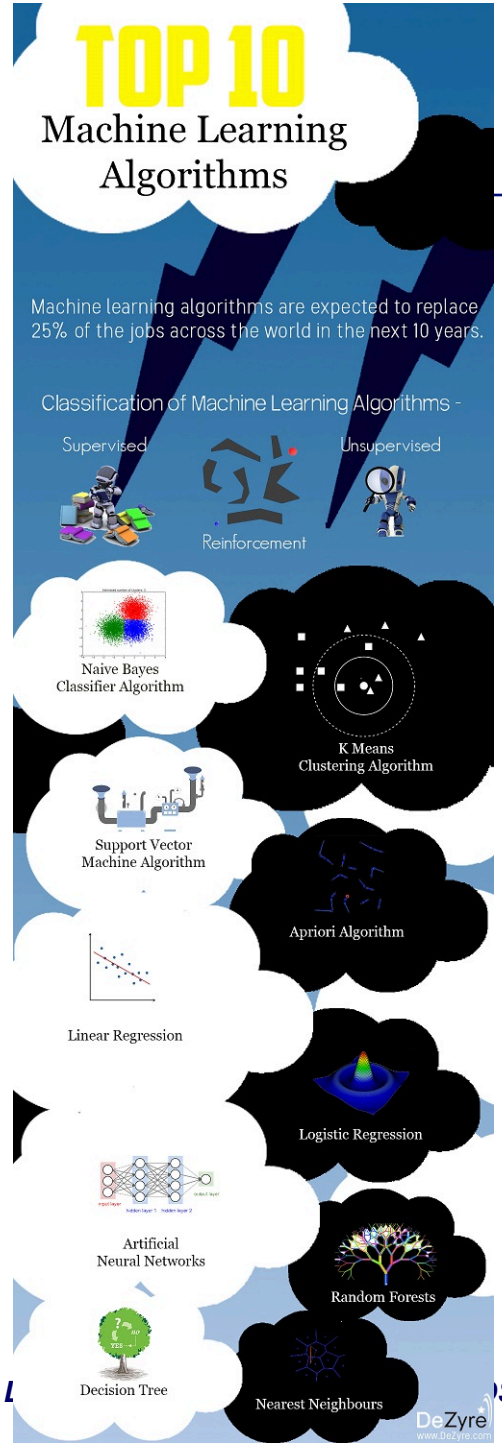
- **Look at a more formal characterization of this question in terms of bias-variance**

Three generations of machine learning

- **First generation: dataset that fits in memory**
  - Single node learning summary statistics and some batch modeling (at small scale); SQL, R
  - Down sampling the data
- **Second generation: General purpose clusters and frameworks**
  - Distributed frameworks that allows us to divide and conquer problems
  - Learning using general purpose frameworks such as hadoop big data analysis offline, realtime decision making, homegrown specialist systems (Hadoop for analysis and modeling;), Hadoop, R
  - In-house purpose built systems; specialist sport
- **Third generation: Purpose-built libraries and frameworks**
  - Built for iterative algorithms that are common place in ML
  - huge scale realtime analysis and decision making systems
  - Specialized frameworks for large scale manipulation the type of data you are workign with.
  - For example, Machine learning libraries like MLLib in Spark, graph processing libraries like Apache Giraph or GraphX in Spark

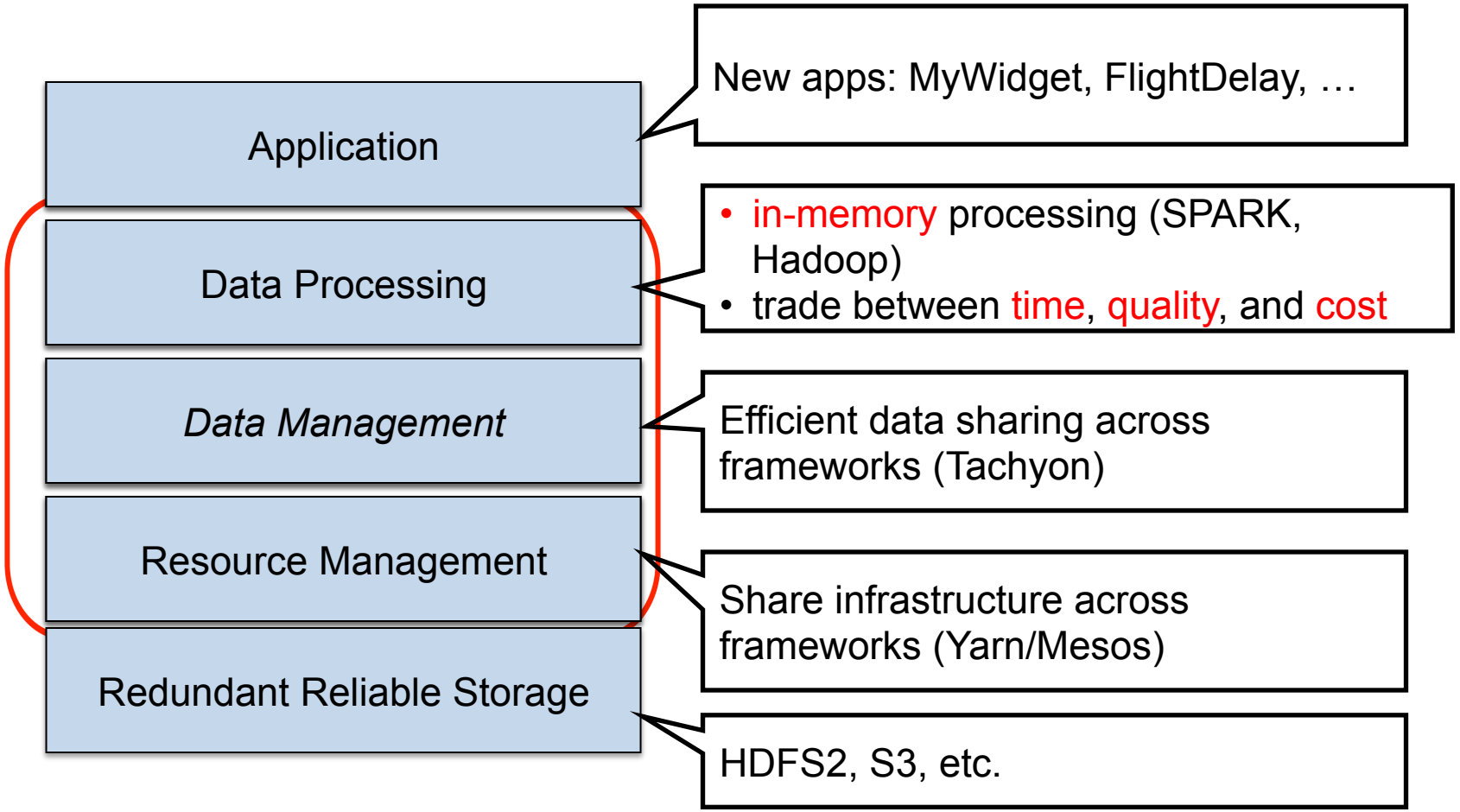
Evolution of Map-Reduce frameworks for big data processing





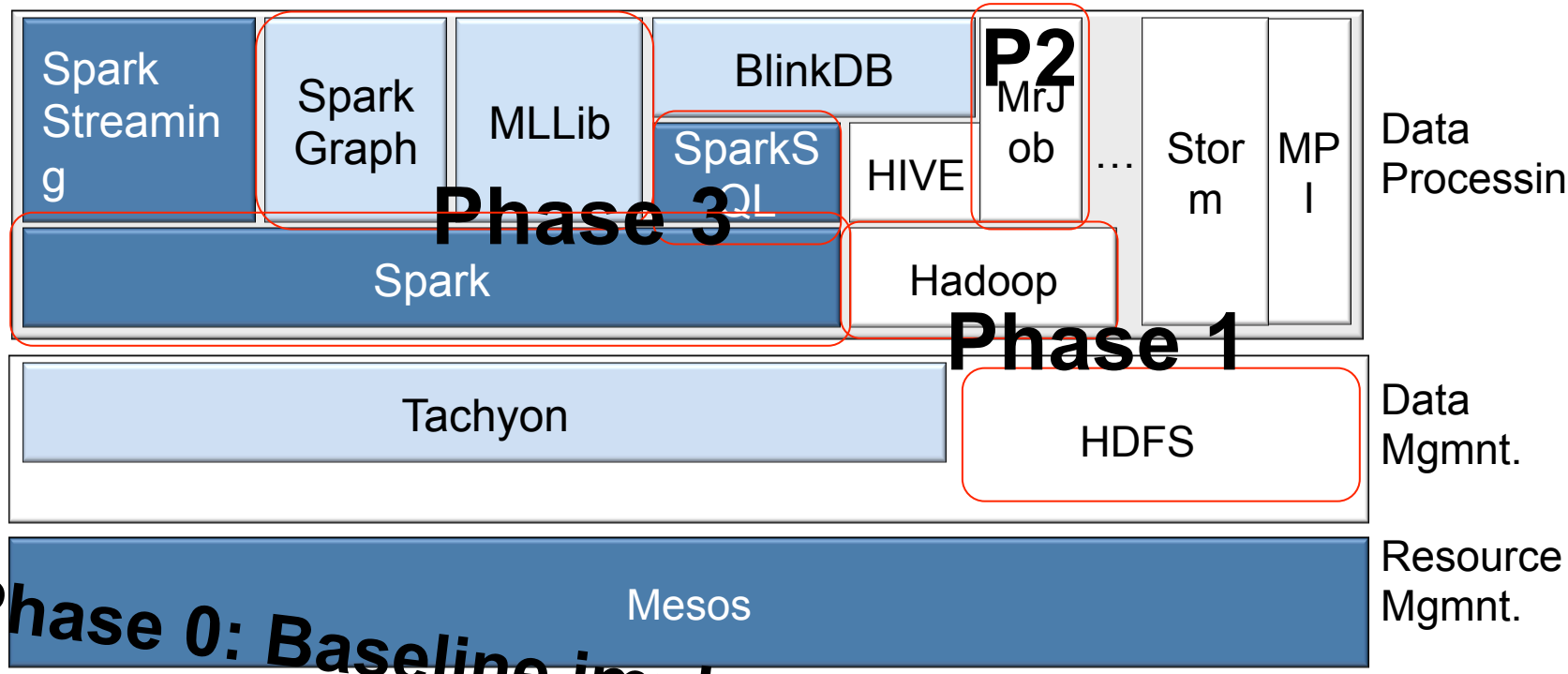
https://www.dezyre.com/article/top-10-machine-learning-algorithms/202

Berkeley Data Analytics Stack (BDAS)



Machine Learning at Scale: Class Phases

- **Focus on distributed computation using functional programming over very large datasets (highlighted in RED)**
- **Develop scalable ML algorithms**
- **Phase 0: Command line; Phase 1: Hadoop/HDFS; Phase 2: MrJob; Phase 3: Spark**

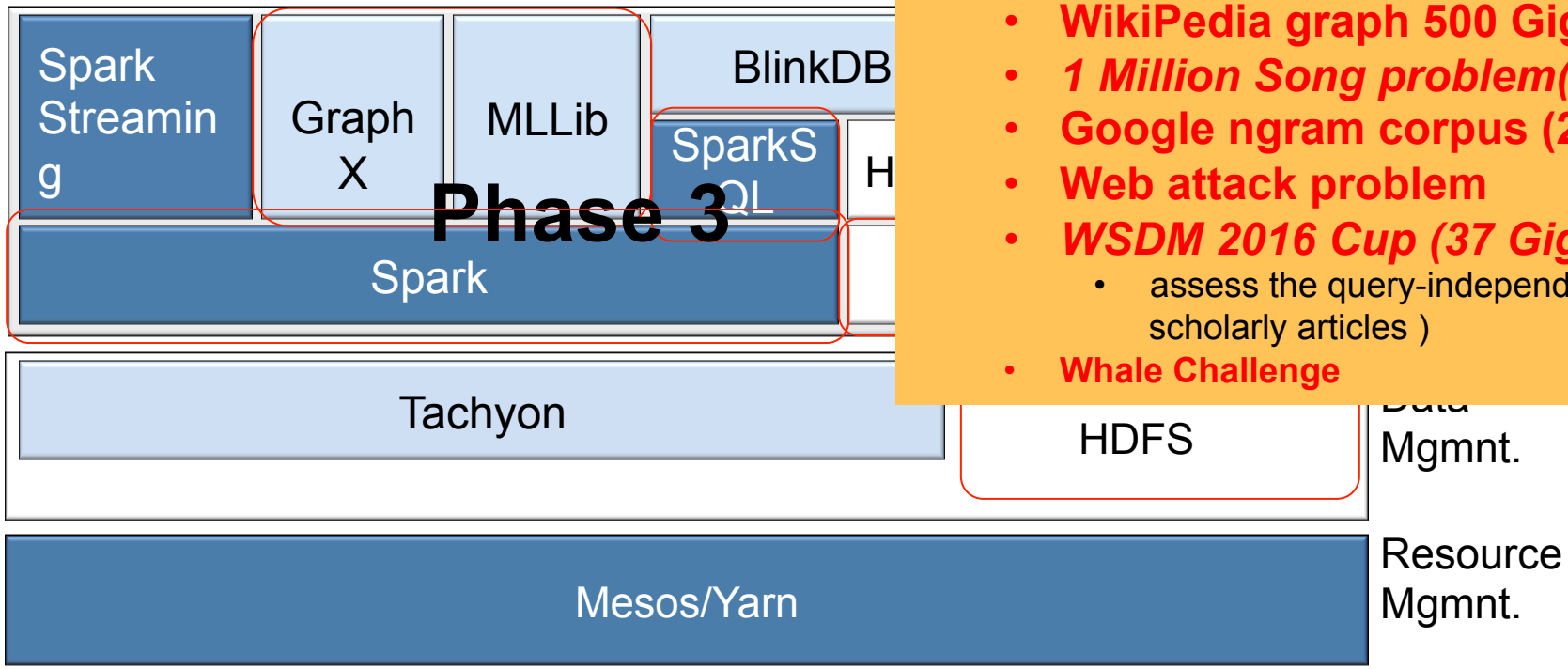


Machine Learning at Scale: Class Phases

- **Distributed computation using functional programming-like paradigm PLUS scalable ML algorithms PLUS Applications**
- **Phase 0: Poorman MapReduce (Unit 1)**
- **Phase 1: Hadoop/HDFS (Units 2 – 3);**
- **Phase 2: MrJob; (Units 4 – 9)**
- **Phase 3: Spark (unit 10 – 14)**

+

- **ML, Graphs,**
- **Case Studies**
  - **$100 Billion problem CTR pre**
  - **WikiPedia graph 500 Gig**
  - ***1 Million Song problem(300G***
  - **Google ngram corpus (2.2 TB**
  - **Web attack problem**
  - ***WSDM 2016 Cup (37 Gig)***
    - assess the query-independent imp scholarly articles)
  - **Whale Challenge**



ML + Systems + CaseStudies == ML at Scale

- **Parallel frameworks for big data**
  - Unix, Hadoop, MrJob, Spark
- **Supervised Machine Learning**
  - Convex optimization, gradient descent, linear regression, decision trees, ensembles of models, support vector machines
- **Unsupervised**
  - Expectation maximization, matrix multiplication, alternating least squares
- **Graphs**
  - random walks, PageRank, graph search algorithms such as BFS, shortest path
- **Hybrid algos**
  - Supervised ML + Random walks
- **Applications**
  - Digital advertising, social media, healthcare, ecommerce, entertainment

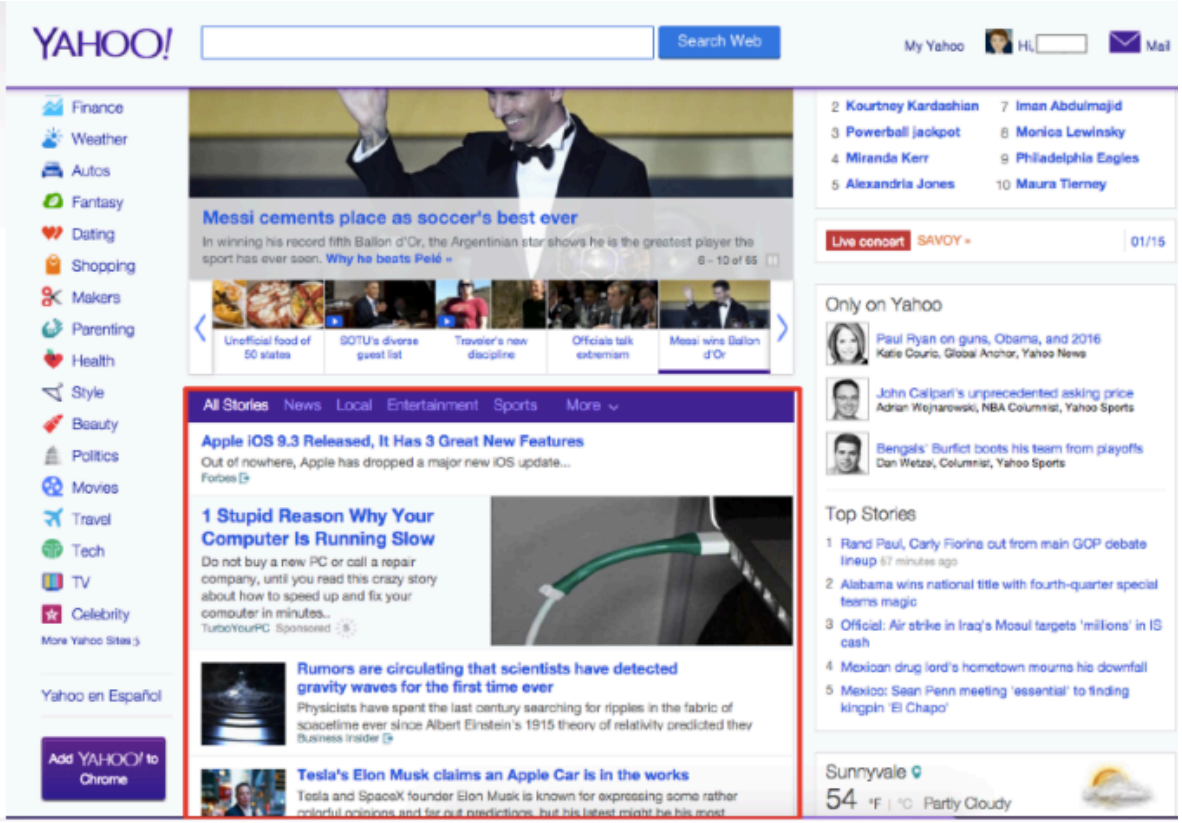
Plus
- Metrics
- Statistics

Datasets and Case studies

- **Enron SPAM Data (100k articles**
- **Google ngram corpus (2.2 TB)**
- **WikiPedia graph (500 Gig)**
- **$100 Billion problem CTR prediction (20-30 Gig)**
- **Web attack (severely sparse)**
- ***1 Million Song problem (300Gig)***
  - *Suggest songs for 100K from 1Million users use of 1Million Songs*
- ***WSDM 2016 Cup (37 Gig)***
  - *assess the query-independent importance of scholarly articles)*
- **13.5 TB Dataset from Yahoo**
  - contains interactions from about 20 million users from February 2015 through May 2015, including those that took place on the Yahoo homepage, Yahoo News, Yahoo Sports, Yahoo Finance, and Yahoo Real Estate

13.5 TB Dataset from Yahoo

- http://techcrunch.com/2016/01/14/yahoo-releases-its-biggest-ever-machine-learning-dataset-to-the-research-community/



Yahoo Dataset: 13.5 TB compressed

- Yahoo announced this morning that it's making the largest-ever machine learning dataset available to the academic research community through its ongoing program, Yahoo Labs Webscope. The new dataset measures a whopping 13.5 TB (uncompressed) in size, and consists of anonymized user interaction data. Specifically, it contains interactions from about 20 million users from February 2015 through May 2015, including those that took place on the Yahoo homepage, Yahoo News, Yahoo Sports, Yahoo Finance, and Yahoo Real Estate.

- In addition to the user interaction data, the dataset also includes demographic information like age range, gender, and generalized geographic data, while items in the dataset include title, summary, and key phrases of the news article in question, plus local timestamps, and partial device information.

- Explains Suju Rajan, Director of Personalization Science at Yahoo Labs, "Data is the lifeblood of research in machine learning. However, access to truly large-scale datasets is a privilege that has been traditionally reserved for machine learning researchers and data scientists working at large companies – and out of reach for most academic researchers."

- As you may imagine, the inability to test against "real-world" data can hamper innovation. And, in turn, can slow down progress.

Challenge

- **Example: predict popular articles?**
  - What makes a popular article

Syllabus

https://www.dropbox.com/s/dyftxnycynu281v/Data-Analytics-and-Machine-Learning-2016-04-22.docx?dl=0

Unit 1 | Introduction / Motivation for Machine Learning at Scale

Show Contents

Unit 2 | Parallel Computing, MapReduce, and Hadoop (Data Storage and Algorithms)

Show Contents

Unit 3 | MapReduce Algorithm Design

Show Contents

Unit 4 | MRJob, Unsupervised Learning at Scale: Clustering, Canopy-Based K-Means, and Expectation Maximization

Show Contents

Unit 5 | Big Data Pipelines

Show Contents

Unit 6 | Distributed Supervised Machine Learning Part 1

Show Contents

Part 1

	A	B	C	D	E	F	G	H	I	J
1	DAY 1								Day 1 Objectives: 2	
2	6/13/16	09:00-10:30	Course Intro: Data Science at Scale						DS Intro, Poormans M	
3	6/13/16	10:30-11:00	COFFEE BREAK						Hadoop, WordCount	
4	6/13/16	11:00-12:30	Machine Learning basics						Metrics; Bias Variance; Linear Regression notebook	
5	6/13/16	12:30-01:30	LUNCH						NiaveBayes	
6	6/13/16	01:30-03:00	Map-Reduce: Cmd line; Hadoop						Basic Linear Regressi	
7	6/13/16	03:00-03:30	COFFEE BREAK						Bias-Variance	
8	6/13/16	03:30-05:00	Problem solving in Hadoop			Word count variations, Naïve Bayes				
9		5:00	Homework							
10	DAY 2									
11	6/14/16	09:00-10:30	MapReduce Algorithm Design and Design Patterns				Sorts, combiners			
12		10:30-11:00	COFFEE BREAK							
13		11:00-12:30	Unsupervised Learning: Pairs and Strips, Apriori sequence mining							
14		12:30-01:30	LUNCH							
15		01:30-03:00	Unsupervised Learning: Clustering, K-Means, Expectation Maximization						Single core machines	
16		03:00-03:30	COFFEE BREAK							
17		03:30-05:00	Unsupervised Learning @ Scale: Clustering, K-Means, Expectation Maximization							
18		5:00	Homework							
19	Day 3									
20	6/15/16	09:00-10:30	Clustering part 2: GMM							
21		10:30-11:00	COFFEE BREAK							
22		11:00-12:30	Big Data Pipelines							
23		12:30-01:30	LUNCH							
24		01:30-03:00	Supervised ML: Linear Regression						Single core machines	
25		03:00-03:30	COFFEE BREAK							
26		03:30-05:00	Supervised ML: Regression Diagnostics, Experiments, and Extensions of Linear Regression							
27		5:00	Homework							
28	DAY 4									
29	6/16/16	09:00-10:30	Feature engineering and data engineering							
30		10:30-11:00	COFFEE BREAK							
31		11:00-12:30	Predicting the length of stay of a patient (Healthcare)							
32		12:30-01:30	LUNCH							

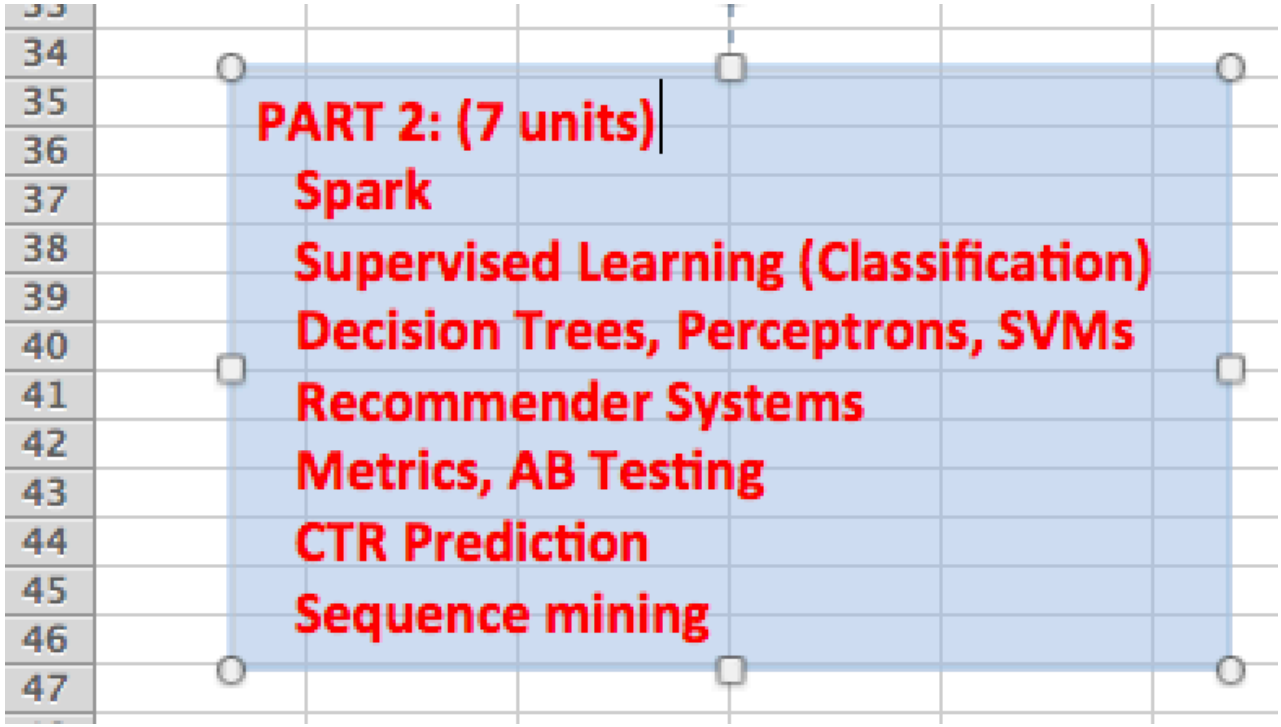
Intro: map-reduce/Hadoop
Naive Bayes/ Linear Regression

Unsupervised Learning

Unsupervised + Supervised Learning (Regression)

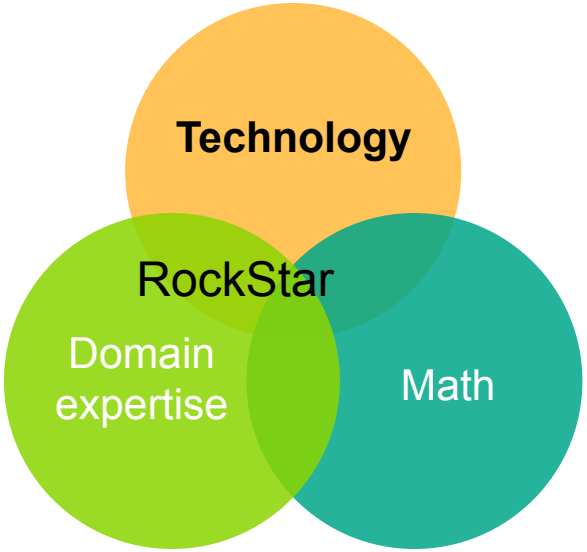
Feature/Data engineering Project

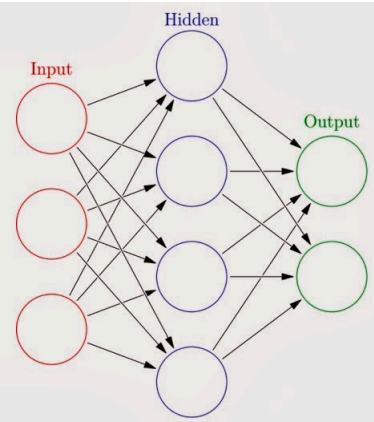
Part 2

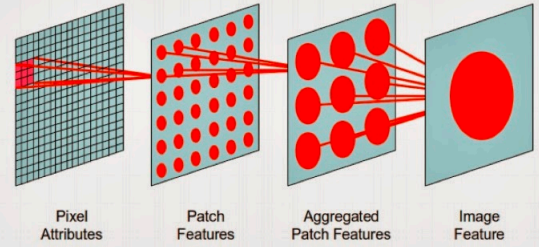

34
35
36
37
38
39
40
41
42
43
44
45
46
47
PART 2: (7 units)
Spark
Supervised Learning (Classification)
Decision Trees, Perceptrons, SVMs
Recommender Systems
Metrics, AB Testing
CTR Prediction
Sequence mining

Deep Learning via Neural Networks at scale

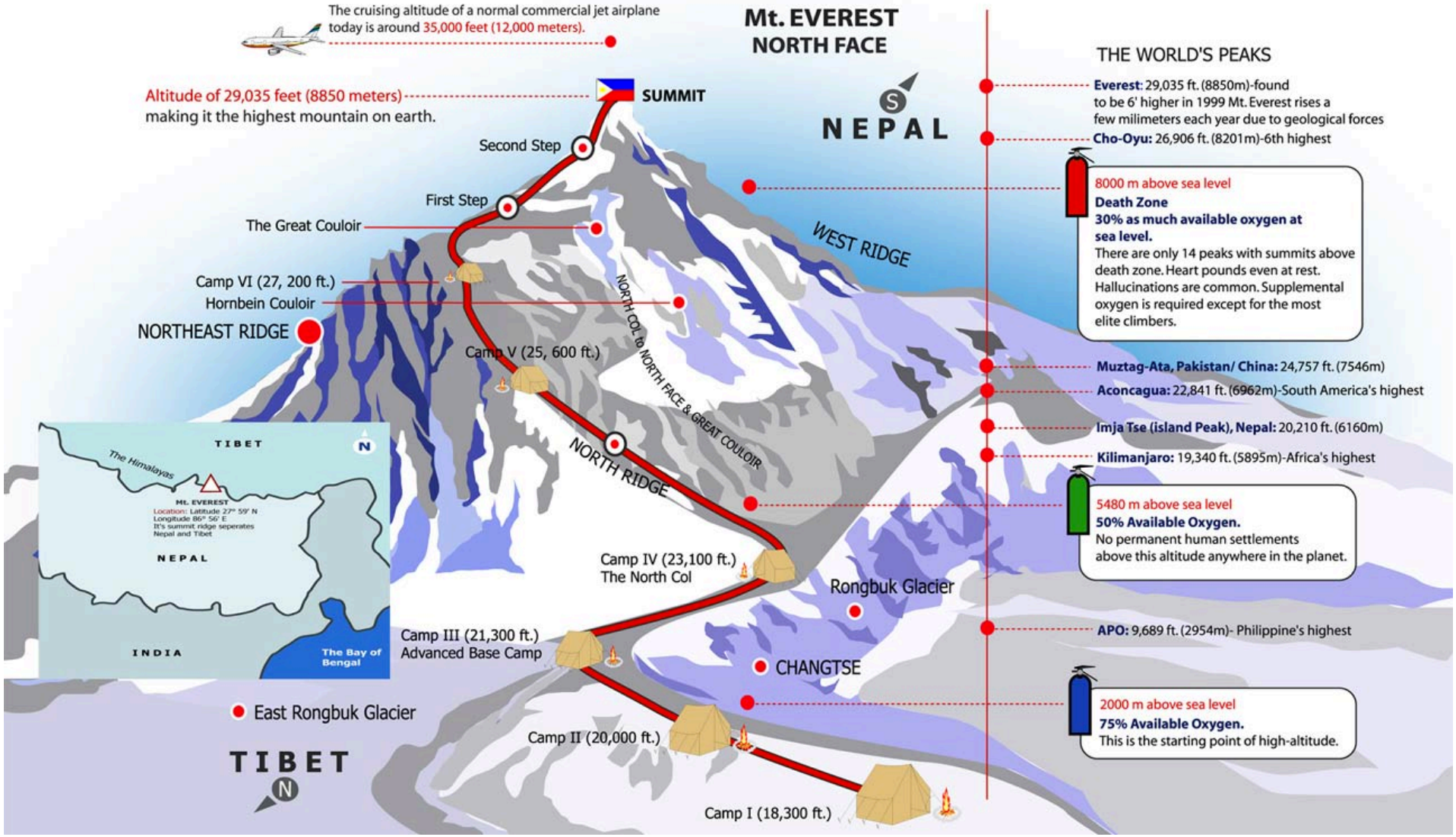
RockStars and Super Models Lecture





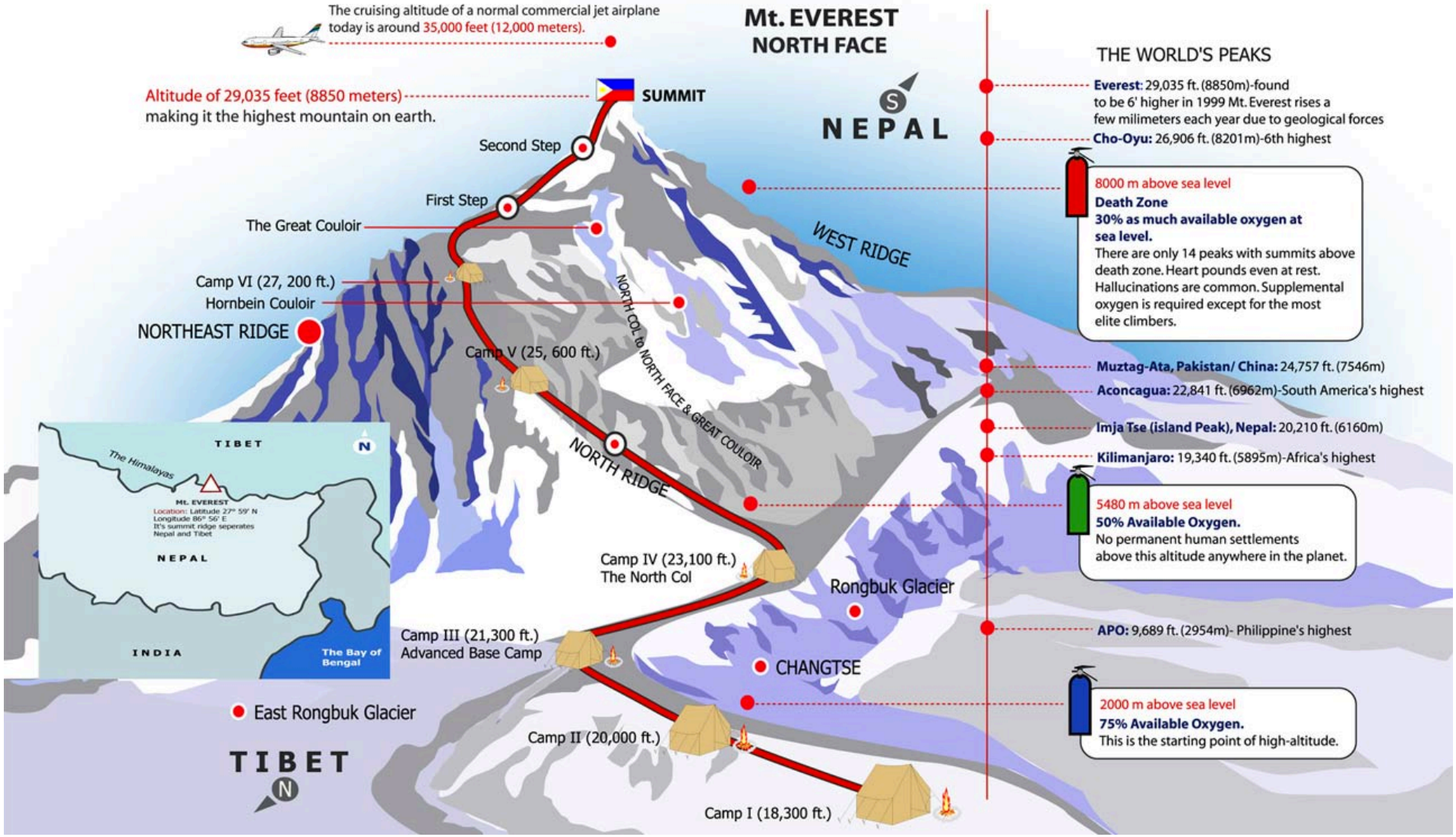


Unit 9/10 Summit Mt Everest



After Unit10

……..you will be cruising at high octane levels



Machine Learning @ Scale: Type II Fun

- **Getting lost, getting cold, getting hungry, getting wet, getting scared, and coming out on top; that's the stuff you remember. That's Type II fun.**
- **Even if you've never heard about the fun scale, you will probably understand it pretty intuitively. On this highly scientific spectrum,**
  - Type I is the easy, fun-while-it's-happening stuff—mellow powder skiing, lazy cragging, afternoon hiking. You're bummed when it's over, but you'd be hard-pressed to remember more than a few specific examples.
  - Type II Tough going, character building and
  - Type III fun resides at the other end of the scale—miserable while it's happening, still miserable when it's over and just as miserable to think about later.
    - Anything that ends with you eating your own shoes, being evacuated by helicopter, or featuring prominently in a non-fiction bestseller likely classifies as Type III.

Fun: intellectual suntan

Fun?, Grueling, muscle memory

Total Disaster

- **- See more at: http://www.backcountry.com/explore/type-ii-fun#sthash.CZankeqe.dpuf**

After this course you will …

- **Role**
  - Individual contributors: R&D, r&D, R&d, D
  - Managers/leaders
- **Focus**
  - Research
    - Continue to study and get a PhD on subject
    - Teach and shape future generations of data scientists
    - Theoretical and applied research
    - Applications
  - Development- infrastructure: Build infrastructure
  - Development
    - Architects of big data pipelines and large scale ML
    - Full stack people
    - Build Apps (on a fully supported framework)
    - Manage teams who this

This class will be demanding

- But it is a high **RoI** class
- Plan on spending 20 hours +/-10 hours per week on this class

Lecture Outline

- Google Doc and Group
- Welcome & Class Introductions
- Big Data and Applications
- Course introduction
- Class logistics
- Systems (part 1 of N)

Course Schedule

14 Lectures, 1 Midterm, End of term exam, plus homework and projects

Row	Date	Units	Description
1	May12-June 13	Unit 0	Probability Theory, Python, R
2	6/13 (Mon) -6/16 (Thur)	Units 1-7	Class Time
3	*6/28 Tues 11AM, MSP Time*	*Unit 8: Exam*	*2 hours Exam*
4	*7/2 Sat.*	*Unit 8 : All Homework*	
5	7/12 (Tues) -7/15 (Fri)	Units 9-15	Class Time
6	*7/26 Tues, 11AM MSP Time*	*Unit 16: Exam*	*2 Hours Exam*
7	*7/30*	*Unit 16: All Homework*	
8	8/1-10/1	Work or a Target project	
9	Week of October 3	1-2 days wrapup	A half day workshop where results will be presented to colleagues along with prep before

Mid term and end of term exams

- **Will consist of Multiple choice questions**
- **Exams are open book but no communication is allowed between students (exams will be proctored via an online proctoring system).**

Logistics/Performance Evaluation

High RoI class: opens up new worlds;

14 Lectures with 2 exam type weeks

- Mid term exam and End of term
  - ~2 hours

% of Grade	Tasks
20%	Unit quizzes
10%	Class participation (fact-to-face and in the online forums; please answer each others questions; collaborate; communicate)
15%	Midterm Exam
15%	End term exam
30%	Projects
10%	Project writeup and presentation

Homework, Projects

- **Each day you will have quizzes, homework**
- **Homework will consist of lecture content based questions and project-based questions**

- **Part 1: 20-30 hours**
  - HW each day (2, 4, 6, Project)
- **Part 2: 20-30 hours**
  - HW each day (10, 12, Project)

Attendance at each session is required

- **Please send me a request when will not be available to attend a live session**
  - **–No request. Zero grade.**
- **10% of your grade goes to attendance and class participation**
- **NOTE:**
  - If you miss too many sessions, then I may request you to drop the class

DATASCI W261: Machine Learning at Scale

Nick Hamlin
nickhamlin@gmail.com
Time of Initial Submission: 9:21 PM EST, Monday, January 18, 2016
Time of **Resubmission**: 8:38 AM EST, Friday, January 22, 2016 W261-3, Sp
Week 1 Homework

http://nbviewer.jupyter.org/urls/dl.dropbox.com/s/j8x7qcck7o4d09u/MIDS-W261-2015-HWK-Week01-HHHHHH

Submission Notes:

- For each problem, I've included a summary of the question as posed in [illegible] submission as uncluttered as possible. For reference, I've included a link to the original instructions in the "Useful Reference" below.
- Problem statements are listed in *italics*, while my responses are shown in plain text.
- I have written driver functions for each problem where a solution is provided in pure Python. For simplicity, I have omitted them for the sections that use Bash commands either directly or to create files.

Useful References:

- **Original Assignment Instructions**
- Wikipedia explaination of Naive Bayes document classification
- Original paper describing the background of the Enron email corpus
- Documentation for Scikit-Learn implementation of Naive Bayes
- Stanford NLP Group's explaination of Naive Bayes algorithm

HW1.0.0.

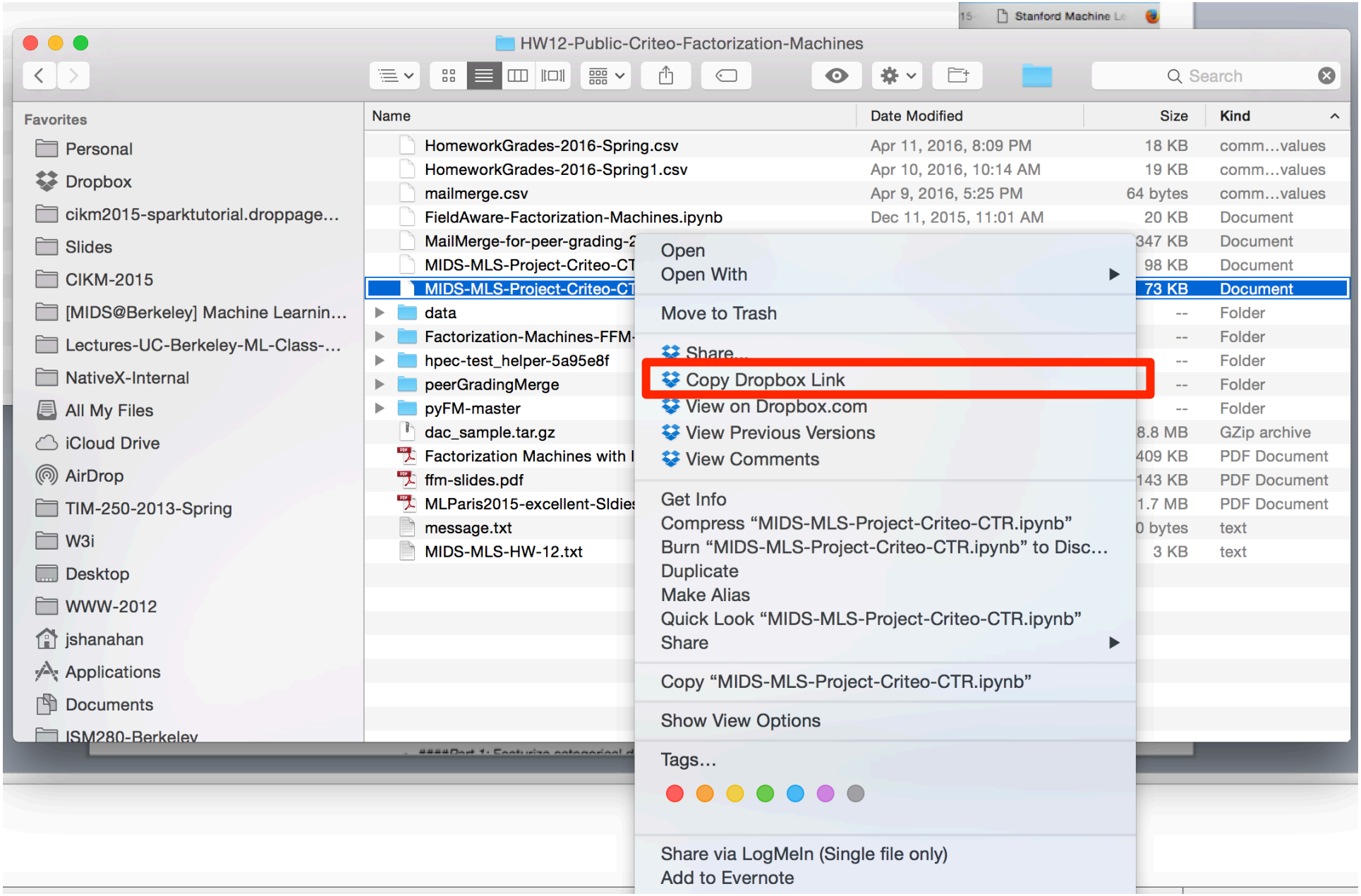
Define big data. Provide an example of a big data problem in your domain of expertise.

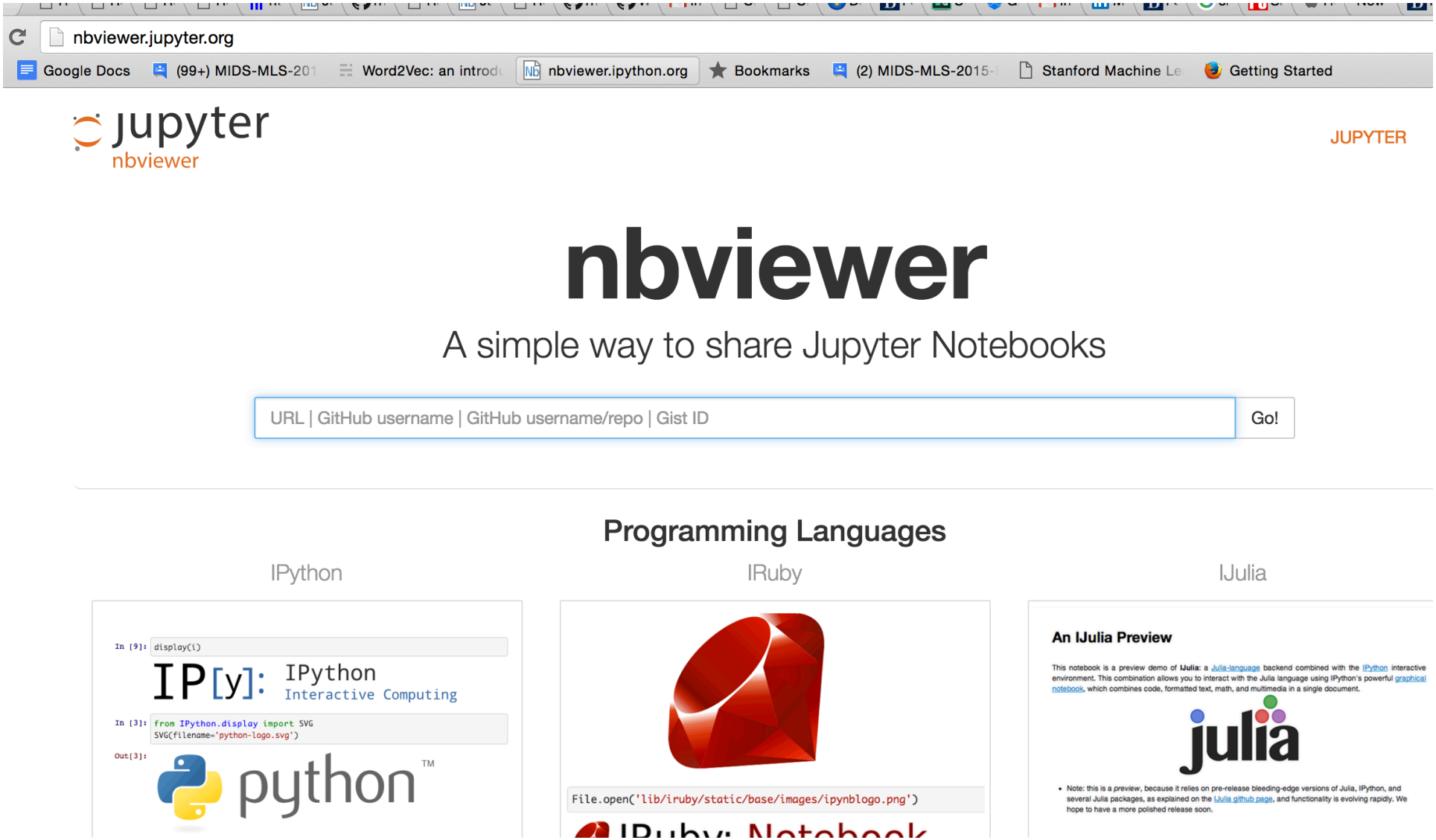
Big data is data with high volumn, velocity, or variety. This data typically represents terabytes or petabytes worth of storage, and is often too much for a single computer in terms of both processing and throughput. For example, a personal laptop with 1TB of storage space is typically only able to effectively process 3-4GB of data at once, orders of magnitude smaller than many "big datasets". As a result, parallel solutions become essential tools for extracting meaning at scale. A big data problem I encounter in my role is in aggregating information about all the individual visitors and their daily activity on the website that my organization maintains. Logging every click, page view, email, call, etc. creates a large, diverse set of data that must be stored and processed effectively at scale for us to be able to derive insights from it.

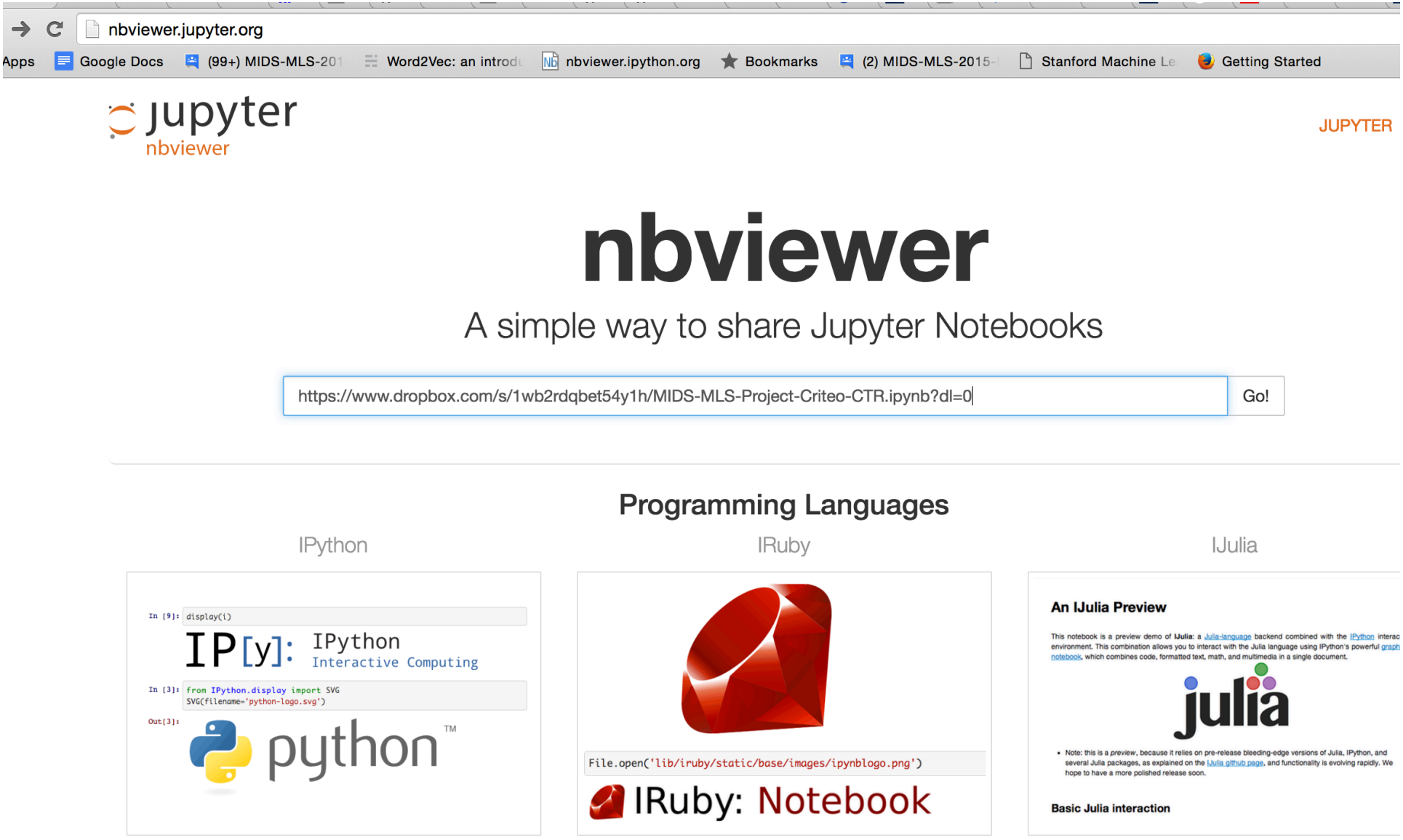
HW1.0.1.

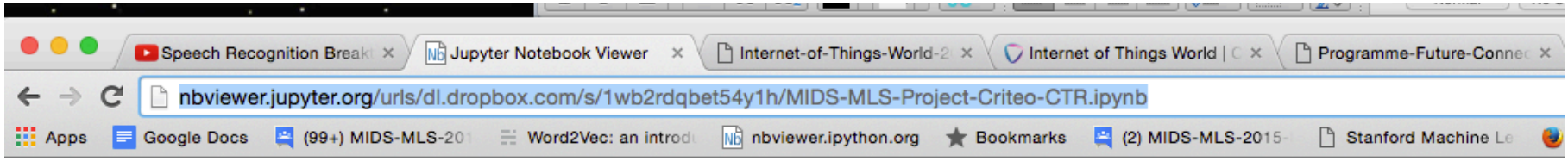
In 500 words (English or pseudo code or a combination) describe how to estimate the bias, the variance, the irreducible error for a test dataset T when using

Copy dropbox link




nbviewer.jupyter.org
Google Docs
(99+) MIDS-MLS-201
Word2Vec: an introd
nbviewer.ipython.org
Bookmarks
(2) MIDS-MLS-2015-I
Stanford Machine Le
Getting Started
jupyter
nbviewer
JUPYTER
nbviewer
A simple way to share Jupyter Notebooks
URL | GitHub username | GitHub username/repo | Gist ID
Go!
Programming Languages
IPython
IRuby
IJulia
In [9]: display(i)
IP[y]: IPython
Interactive Computing
In [3]: from IPython.display import SVG
SVG(filename='python-logo.svg')
Out[3]:
python
File.open('lib/iruby/static/base/images/ipynblogo.png')
An IJulia Preview
This notebook is a preview demo of IJulia: a Julia-language backend combined with the IPython interactive environment. This combination allows you to interact with the Julia language using IPython's powerful graphical notebook, which combines code, formatted text, math, and multimedia in a single document.
julia
Note: this is a preview, because it relies on pre-release bleeding-edge versions of Julia, IPython, and several Julia packages, as explained on the IJulia github page, and functionality is evolving rapidly. We hope to have a more polished release soon.


nbviewer.jupyter.org
Apps
Google Docs
(99+) MIDS-MLS-201
Word2Vec: an introdu
nbviewer.ipython.org
Bookmarks
(2) MIDS-MLS-2015-
Stanford Machine Le
Getting Started
jupyter
nbviewer
JUPYTER
nbviewer
A simple way to share Jupyter Notebooks
https://www.dropbox.com/s/1wb2rdqbet54y1h/MIDS-MLS-Project-Criteo-CTR.ipynb?dl=0
Go!
Programming Languages
IPython
IRuby
IJulia
In [9]: display(i)
IP[y]: IPython
Interactive Computing
In [3]: from IPython.display import SVG
SVG(filename='python-logo.svg')
Out[3]:
python
File.open('lib/iruby/static/base/images/ipynblogo.png')
IRuby: Notebook
An IJulia Preview
julia
Basic Julia interaction



jupyter nbviewer

JU

DATASCI W261: Machine Learning at Scale

W261-1 Fall 2015
Week 12: Criteo CTR Project
November 14, 2015

Student name **INSERT STUDENT NAME HERE**

Click-Through Rate Prediction Lab

This lab covers the steps for creating a click-through rate (CTR) prediction pipeline. You will work with the Criteo Lab Kaggle competition.

This lab will cover:

- ####Part 1: Featurize categorical data using one-hot-encoding (OHE)

La

Google Form

XXXX LINK

Please enter your ischool email addrees *

Please enter first name (formal name) *

Please enter family name (formal name) *

Link to Notebook (use a NBViewer link or Github link) *

Please render your notebook in incognito mode on your browser to make sure everything is visible

Link to Notebook (PDF Note timestamp will be checked) *

Homework for Week *

Select a week: Select HW1 for homework for week 1

Group Number *

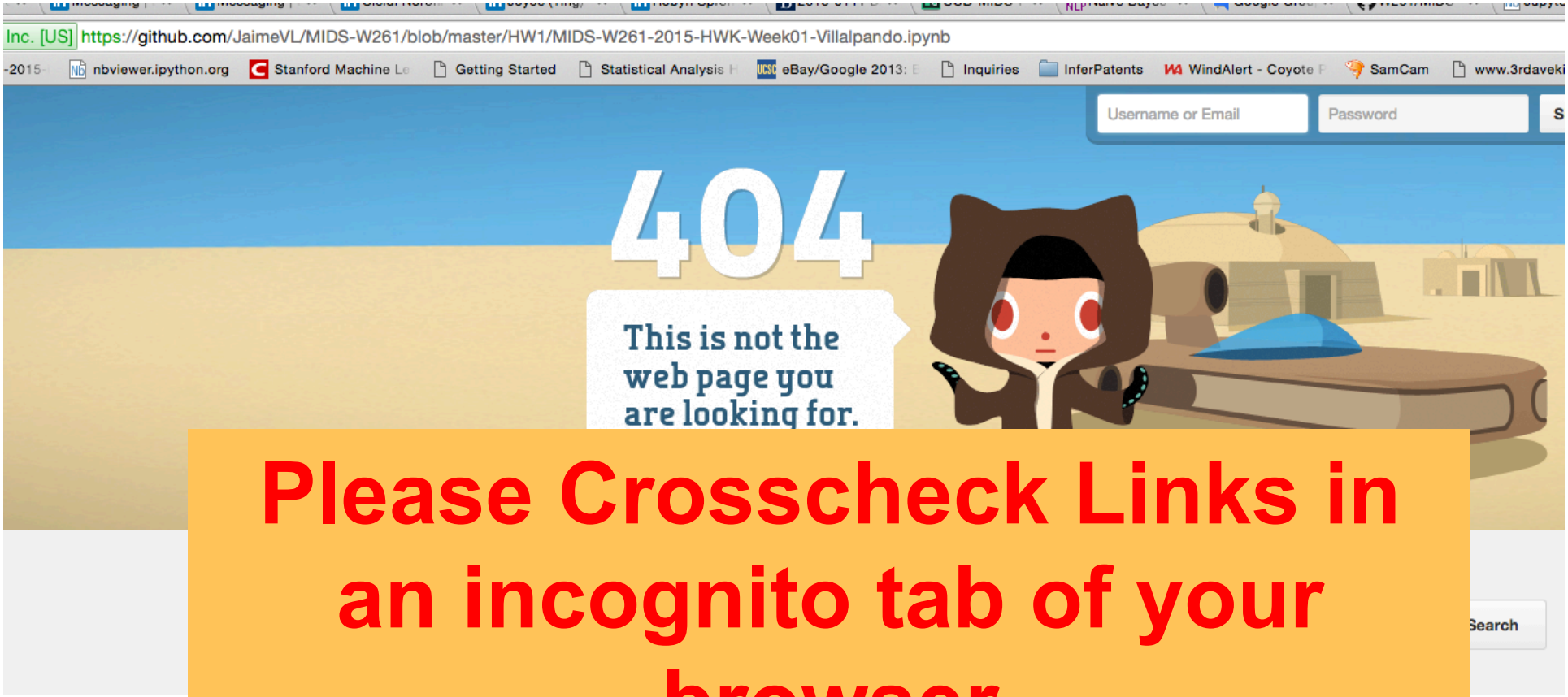
Select a week: Select HW1 for homework for week 1

Homework Submissions

- **Name, date stamp of submission, HW,**
- **Notebook submissions on ISVC**
  - Upload Notebook and its PDF
  - PDF (dropbox/Github)
  - If using Dropbox please provide a **NBViewer link** (http://nbviewer.ipython.org/) and the raw dropbox link
- **Submit on time!**

Please Crosscheck Links in an incognito tab of your browser

Please Avoid This problem



Please Crosscheck Links in an incognito tab of your browser

Please Avoid This problem

Big data Definition: use

Definition

- **Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.**

Plus intuition

  - PROCESSING:
    - Think of your laptop that gets overwhelmed with 3-4 gig of data (disk space is 1TB)
  - STORAGE:
    - Laptop : 1 TB
  - THROUGH-PUT
    - 1TB would take 3 hours to read it using your laptop

- **Challenges**
  - Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, security, and information privacy.

Local machine versus Cloud

- Units 1-4 Local computer based
- From Unit 5 move to the cloud

Weekly Office Hours: on demand

- **Office hours on demand: Monday of the next week at 4PM West Coast time**
- **You can attend office hours if you submit a question(s) 24 hours before the start of office hours to the Google Group**
- **Use the STAR methodology to present you Questions/Answers**
  - Knowledge base questions
  - Problem solving

Lecture Outline

- Google Doc and Group
- Welcome & Class Introductions
- Big Data and Applications
- Course introduction
- Class logistics
- Systems (part 1 of N)

Install

- Jupyter notebook (via Anaconda)
- Install Hadoop on your local machines

- **Installing Hadoop directly on your machine (and not your virtual machine)**

Installing Hadoop

- **Mac:**
  - http://amodernstory.com/2014/09/23/installing-hadoop-on-mac-osx-yosemite/
  - This link is for haddoop 2.6. I follow the instructions and easily get hadoop installed.

- **Windows:**
  - Hadoop 1.0
  - http://saphanatutorial.com/hadoop-installation-on-windows-7-using-cygwin/

  - Hadoop 2.0 (Hortonworks Data Platform 2.0 for Windows)
  - http://hortonworks.com/blog/install-hadoop-windows-hortonworks-data-platform-2-0/

- **Linux (Ubuntu):**
  - http://www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_single_node_cluster.php

On Mac install HomeBrew

- **Install HomeBrew**
  - Download it from the website at http://brew.sh/ or simply paste the script inside the terminal
  - $ ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"

On Windows install CygWin

- **Cygwin is:**
  - a large collection of GNU and Open Source tools which provide functionality similar to a Linux distribution on Windows.

Current Cygwin DLL version

The most recent version of the Cygwin DLL is 2.2.1. Install it by running setup-x86.exe (32-bit installation) or setup-x86_64.exe (64-bit installation).

Use the setup program to perform a fresh install or to update an existing installation.

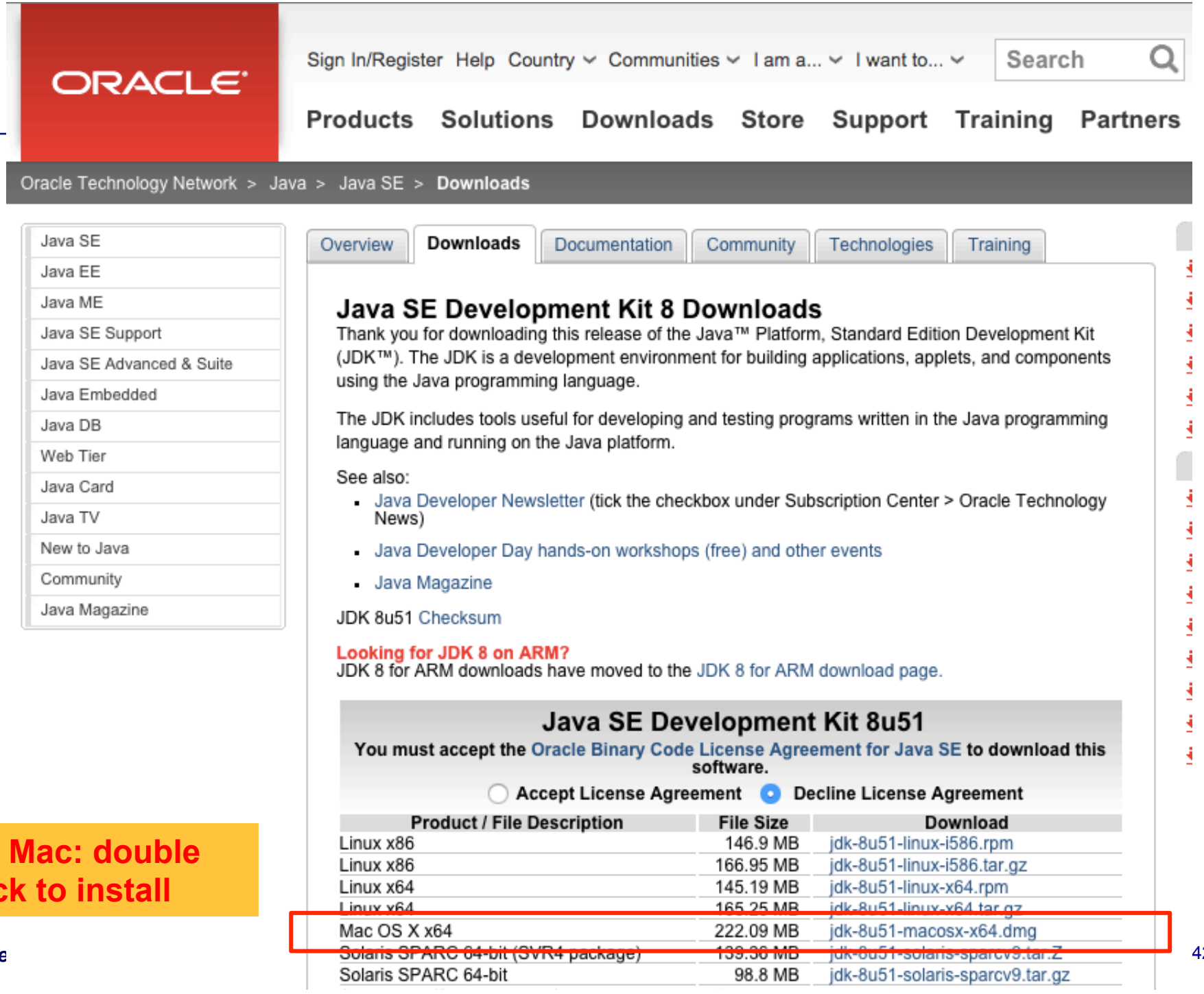
Note that individual packages in the distribution are updated separately from the DLL so the Cygwin DLL version is not useful as a general Cygwin release number.

Make sure Java JDK is installed

- **Click here:** http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

- **On windows machine set up $JAVA_HOME**

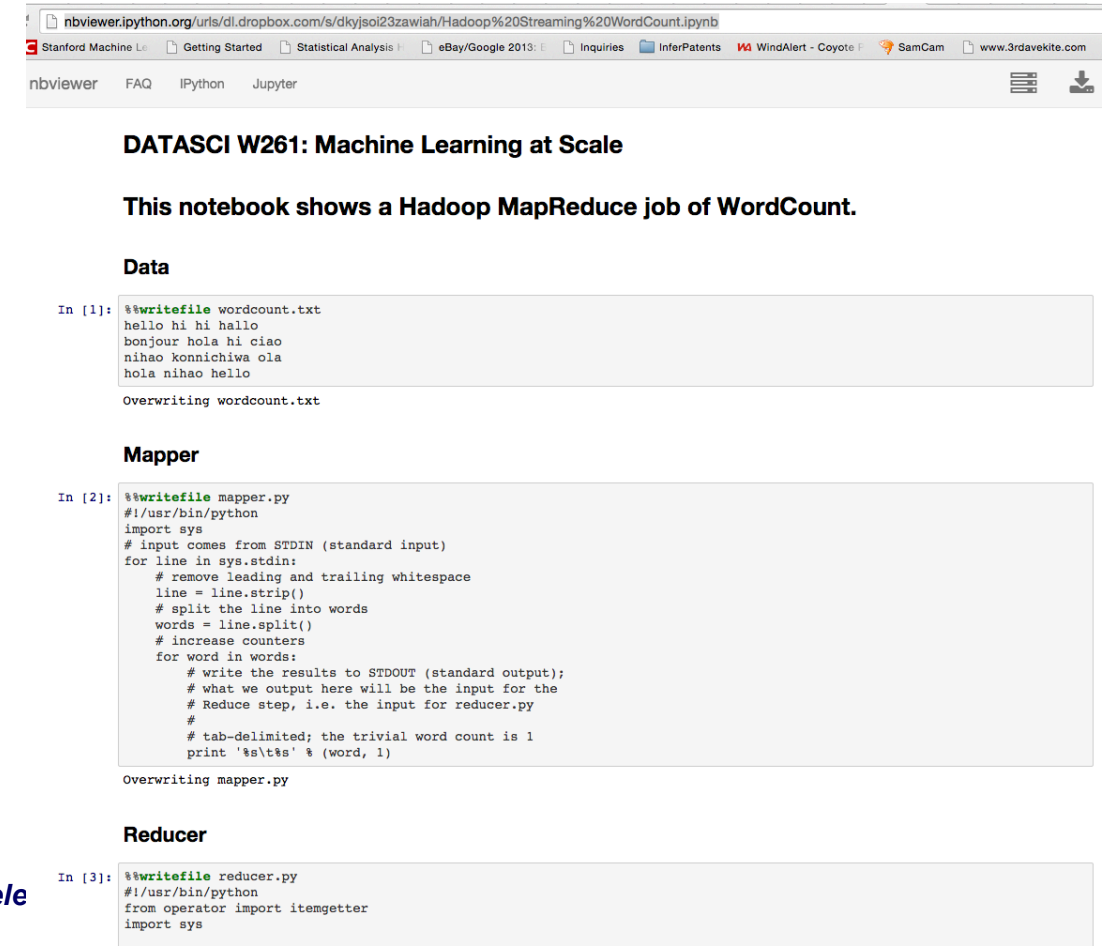
- **Right-click the My Computer icon on your desktop and select Properties**
- **Click the Advanced tab**
- **Click the Environment Variables button**
- **Under System Variables, click New**
- **Enter the variable name as JAVA_HOME**
- **Enter the variable value as the installation path for the Java Development Kit**



On Mac: double click to install

Word Count Notebook

http://nbviewer.ipython.org/urls/dl.dropbox.com/s/dkyjsoi23zawiah/Hadoop%20Streaming%20WordCount.ipynb



nbviewer.ipython.org/urls/dl.dropbox.com/s/dkyjsoi23zawiah/Hadoop%20Streaming%20WordCount.ipynb

Stanford Machine Le | Getting Started | Statistical Analysis | eBay/Google 2013: E | Inquiries | InferPatents | WindAlert - Coyote | SamCam | www.3rdavekite.com

nbviewer FAQ IPython Jupyter

DATASCI W261: Machine Learning at Scale

This notebook shows a Hadoop MapReduce job of WordCount.

Data

In [1]:
```
%%writefile wordcount.txt
hello hi hi hallo
bonjour hola hi ciao
nihao konnichiwa ola
hola nihao hello
```

```
Overwriting wordcount.txt
```

Mapper

In [2]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; the trivial word count is 1
        print '%s\t%s' % (word, 1)
```

```
Overwriting mapper.py
```

Reducer

In [3]:
```
%%writefile reducer.py
#!/usr/bin/python
from operator import itemgetter
import sys
```

- **End of lectures**

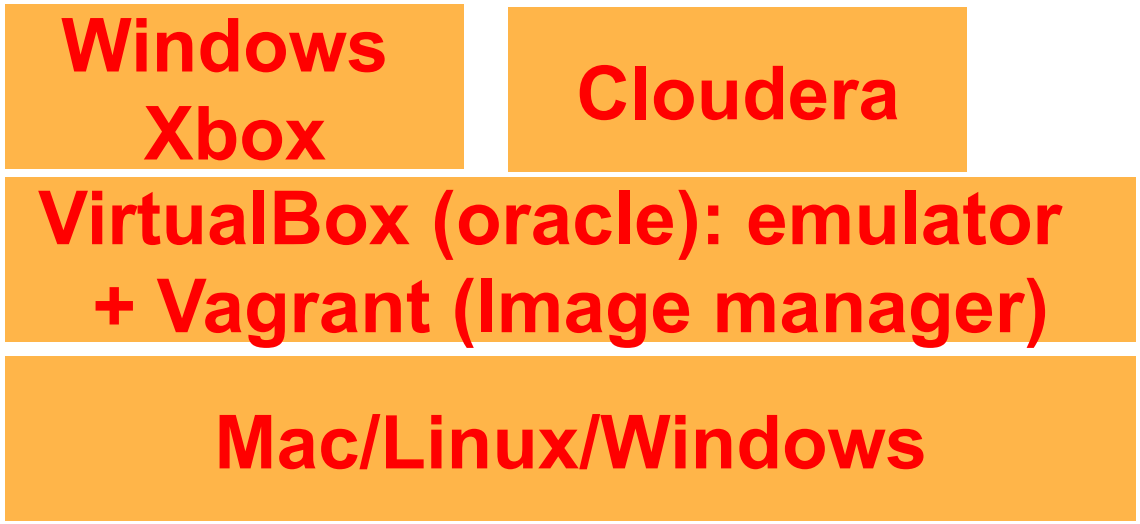
Live Session Outline

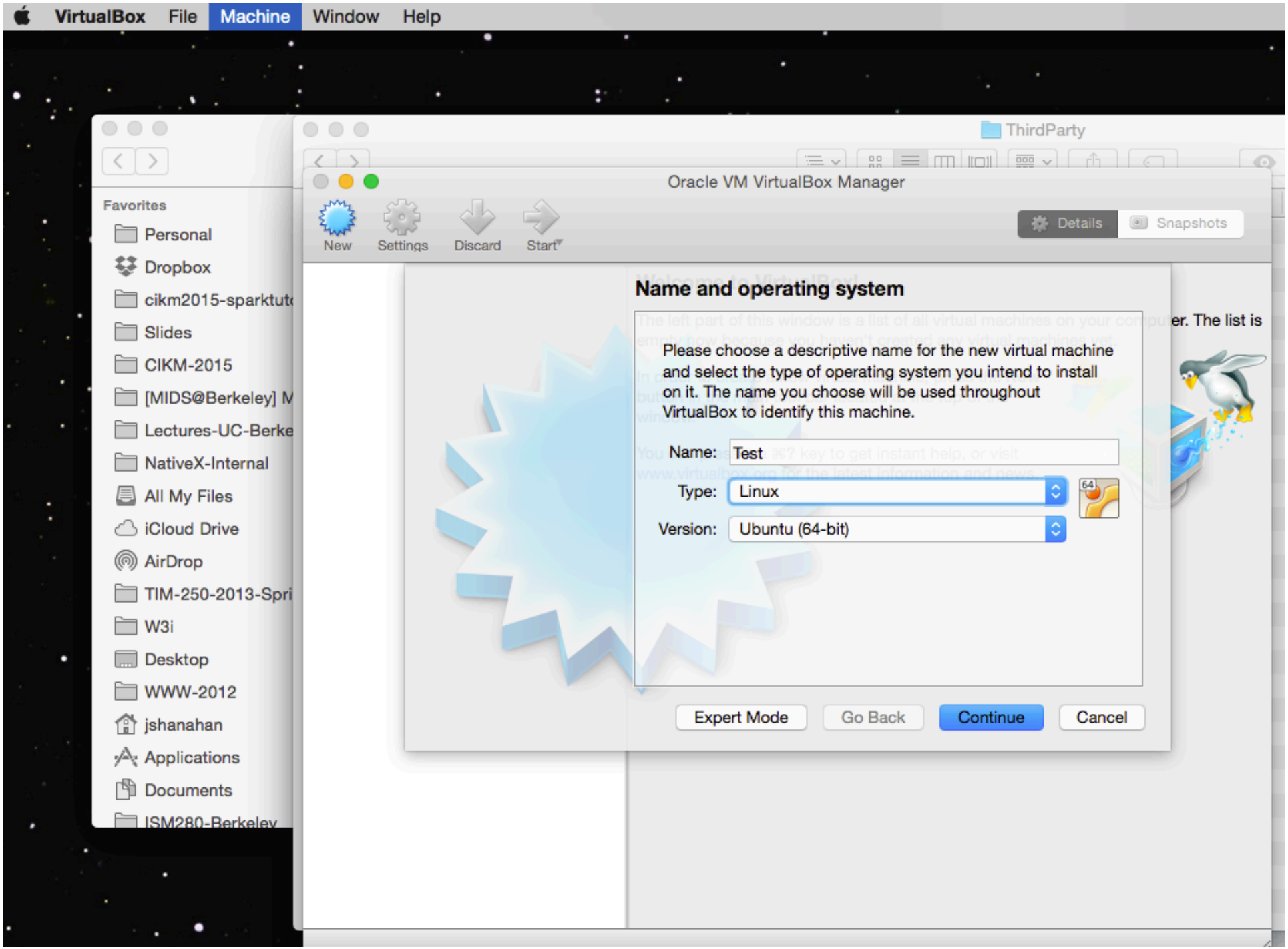
- **Welcome & Class Introductions**
  - Please mute you microphones
  - Start RECORDING (bonus points for reminding me!)
  - Class, homework, project Logistics + Office hours
  - Self-introductions (Bios + WWK01: Q1)
- **W261 introduction**
- **Class logistics**
- **Q&A (WK01):**
- **Homework 1:**
  - Enron SPAM Dataset + Wordcount
- **Open Mike**
- ***Naive Bayes***
  - *Basic derivation*
  - *Various Naïve Bayes Flavours (Live Session #2)*
- **Wrapup**

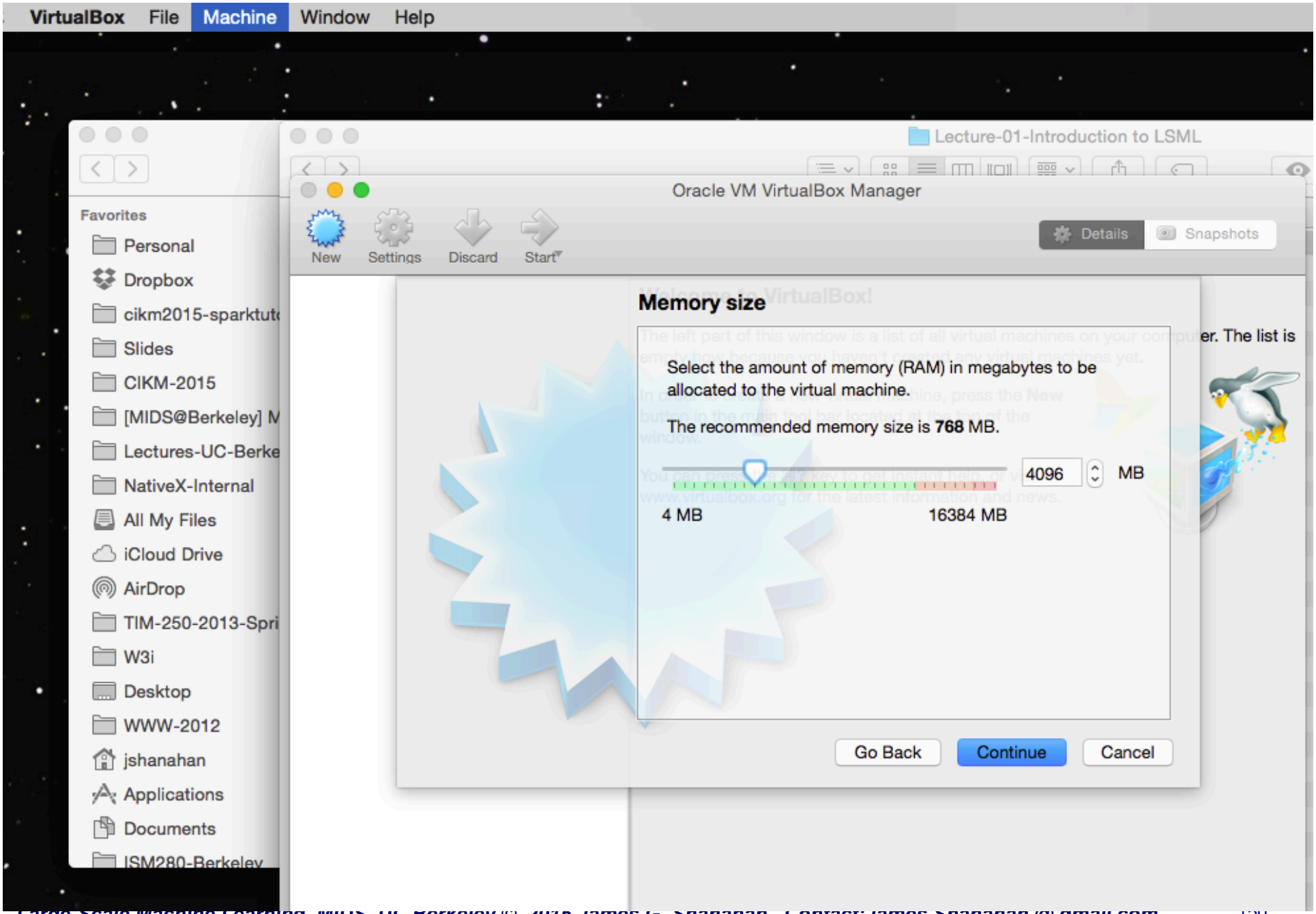
- **Cygwin**
- **Virtual machine**

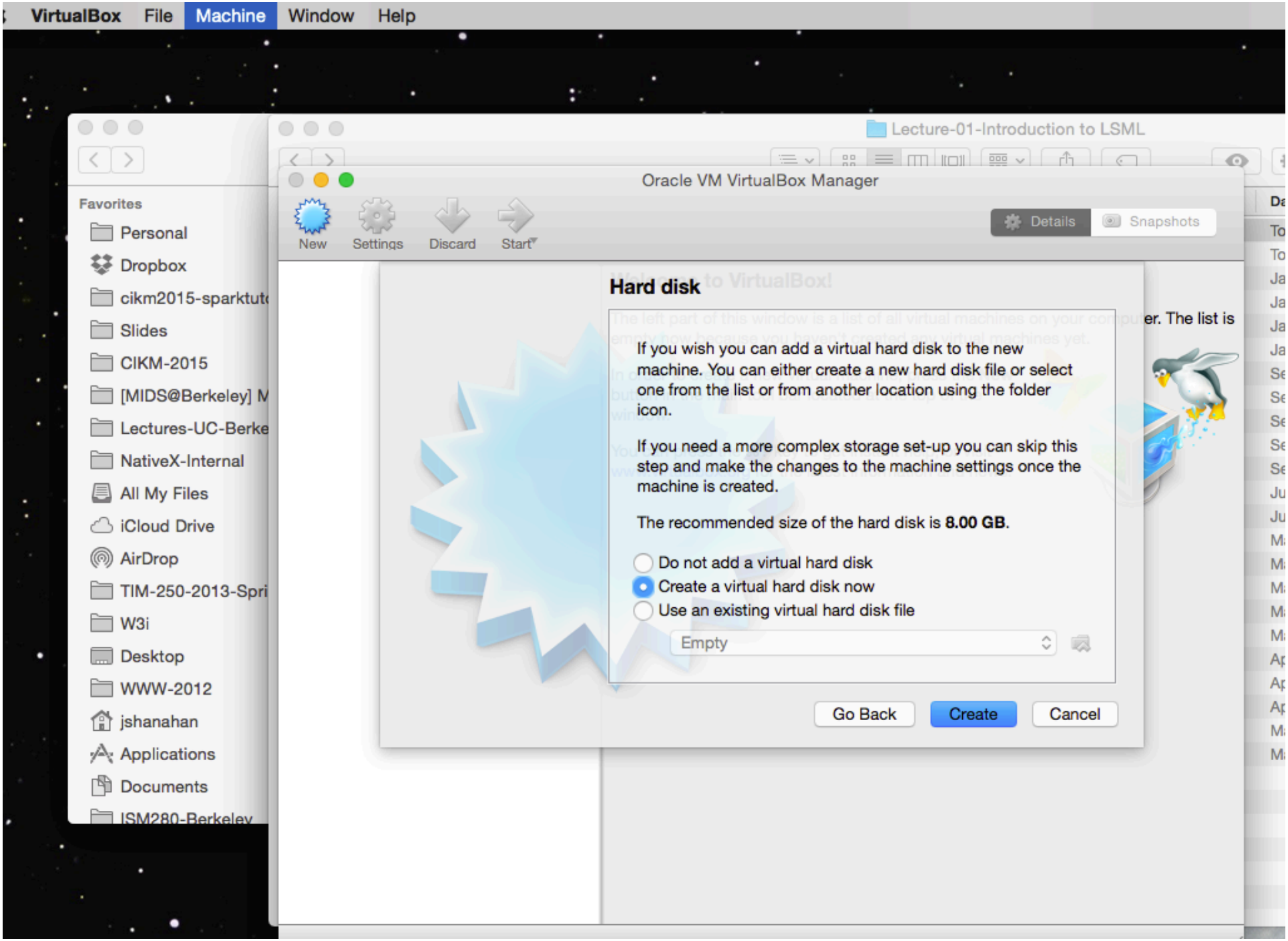
Standard setup: Virtual Box

A virtual machine is a software computer that, like a physical machine, runs an operating system and applications. A virtual machine uses the physical resources of the physical machine on which it runs, which is called the host system. Virtual machines have virtual devices that provide the same functionality as physical hardware, but with the additional benefits of portability, manageability, and security. A virtual machine has an operating system and virtual resources that you manage in much the same way that you manage a physical computer. For example, you install an operating system in a virtual machine in the same way that you install an operating system on a physical computer. You must have a CD-ROM, DVD, or ISO image that contains the installation files from an operating system vendor.




VirtualBox File Machine Window Help
ThirdParty
Favorites
Personal
Dropbox
cikm2015-sparktut
Slides
CIKM-2015
[MIDS@Berkeley] M
Lectures-UC-Berke
NativeX-Internal
All My Files
iCloud Drive
AirDrop
TIM-250-2013-Spri
W3i
Desktop
WWW-2012
jshanahan
Applications
Documents
ISM280-Berkeley
Oracle VM VirtualBox Manager
New
Settings
Discard
Start
Details
Snapshots
Name and operating system
Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.
Name: Test
Type: Linux
Version: Ubuntu (64-bit)
Expert Mode
Go Back
Continue
Cancel

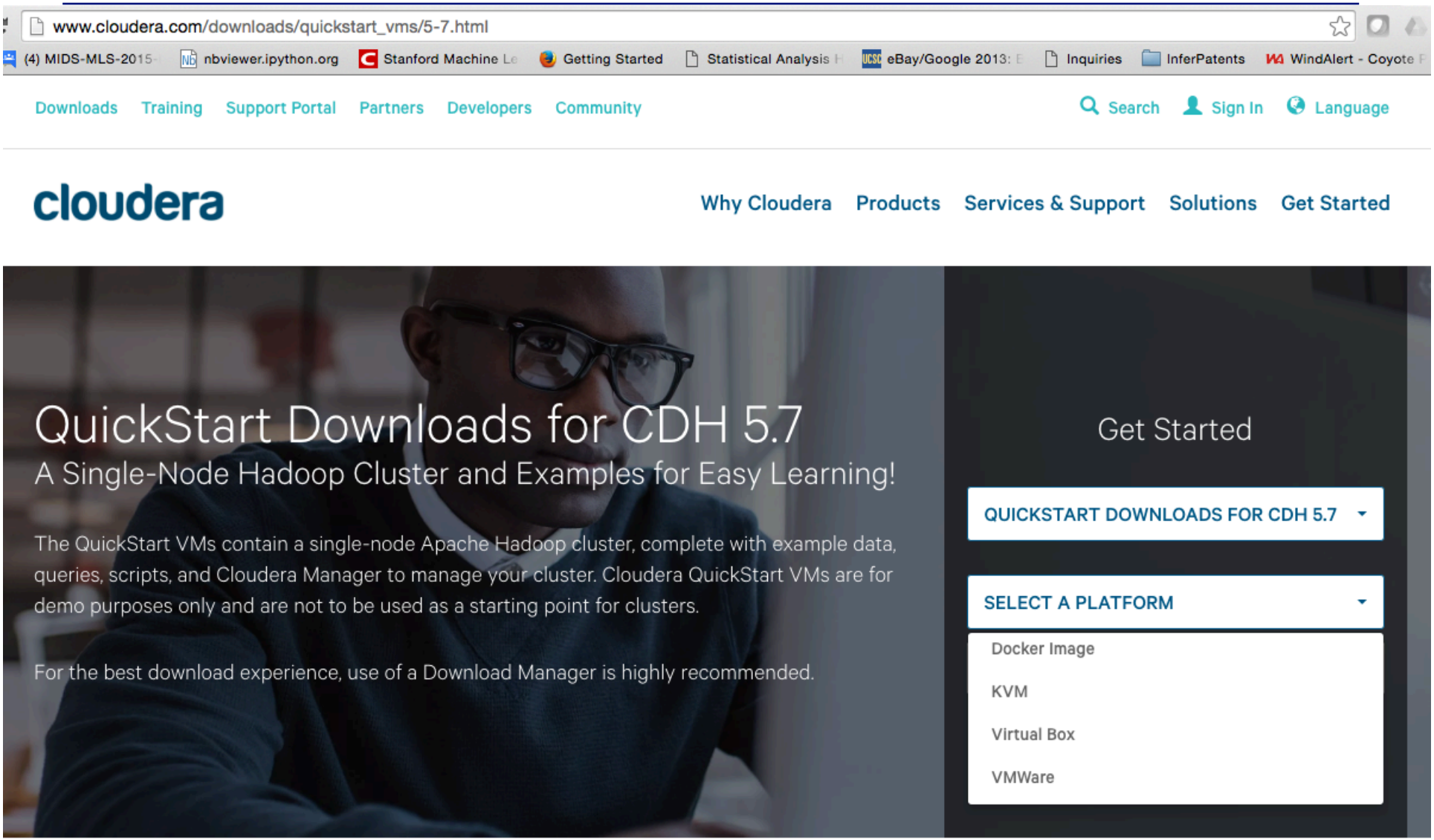

VirtualBox File Machine Window Help
Lecture-01-Introduction to LSML
Oracle VM VirtualBox Manager
New Settings Discard Start
Details Snapshots
Favorites
Personal
Dropbox
cikm2015-sparktut
Slides
CIKM-2015
[MIDS@Berkeley] M
Lectures-UC-Berke
NativeX-Internal
All My Files
iCloud Drive
AirDrop
TIM-250-2013-Spri
W3i
Desktop
WWW-2012
jshanahan
Applications
Documents
ISM280-Berkeley
Memory size
Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.
The recommended memory size is 768 MB.
4096
MB
4 MB
16384 MB
Go Back
Continue
Cancel

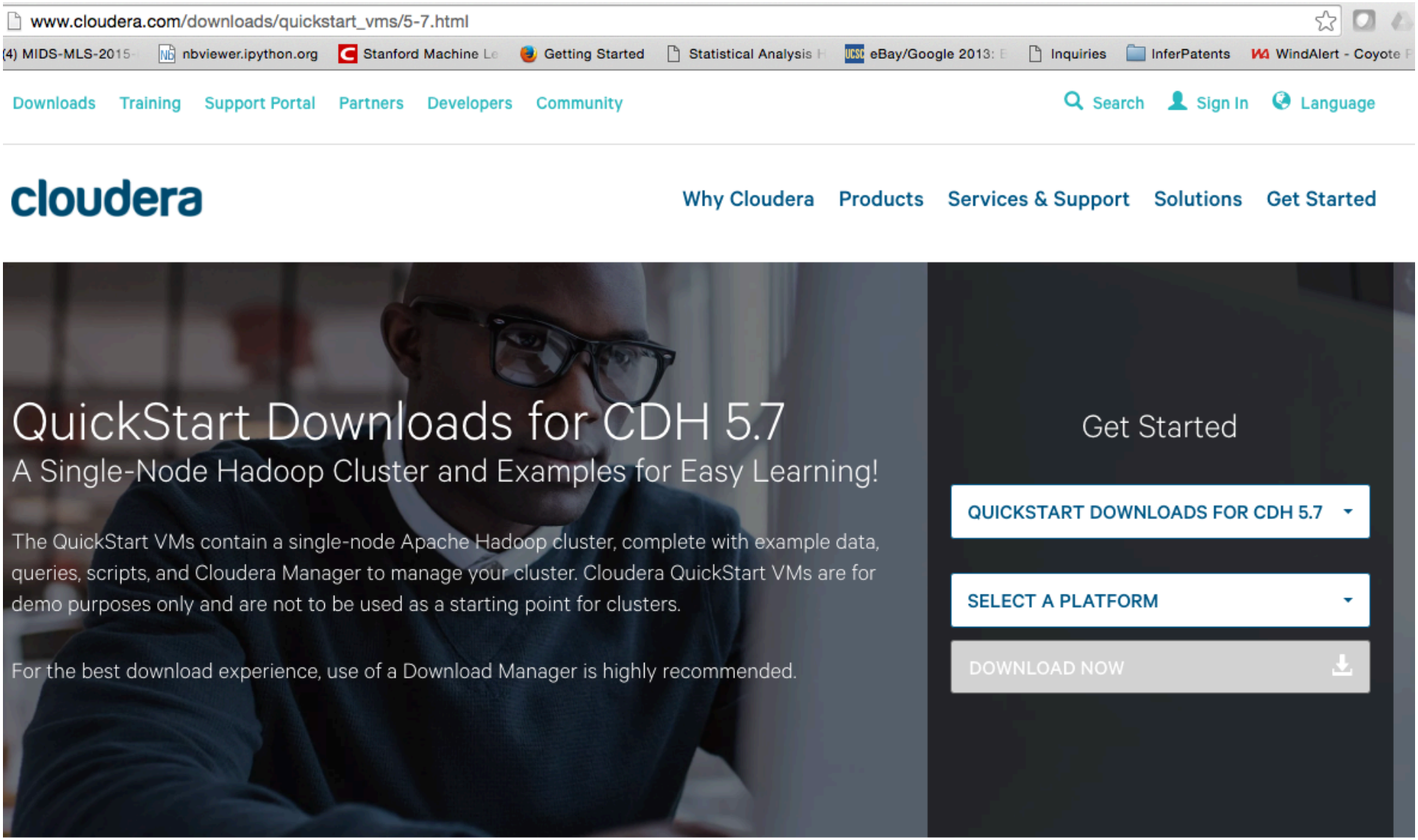

VirtualBox File Machine Window Help
Lecture-01-Introduction to LSML
Favorites
Personal
Dropbox
cikm2015-sparktut
Slides
CIKM-2015
[MIDS@Berkeley] M
Lectures-UC-Berke
NativeX-Internal
All My Files
iCloud Drive
AirDrop
TIM-250-2013-Spri
W3i
Desktop
WWW-2012
jshanahan
Applications
Documents
ISM280-Berkeley
Oracle VM VirtualBox Manager
New
Settings
Discard
Start
Details
Snapshots
Hard disk
If you wish you can add a virtual hard disk to the new machine. You can either create a new hard disk file or select one from the list or from another location using the folder icon.
If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.
The recommended size of the hard disk is 8.00 GB.
Do not add a virtual hard disk
Create a virtual hard disk now
Use an existing virtual hard disk file
Empty
Go Back
Create
Cancel

Install iPython on VirtualBox

- **How run code on a unix machine (when I have a windows)?**
  - Set up a virtual machine (e.g., a linux box) on your Windows/Mac machines
    - Install VirtualBox and then install Vagrant (this will install iPython etc on the virtual box)
  - Follow all steps in the following slides to install VirtualBox and Vagrant
    - https://www.dropbox.com/s/z8wri7bghh4js9h/Simplified-IPython-Installation-V2-onLinux-VirtualBox.pdf?dl=0
  - **Vagrant** (Image manger: AMI)
    - Instead of building a virtual machine from scratch, which would be a slow and tedious process, Vagrant uses a base image to quickly clone a virtual machine. These base images are known as "boxes" in Vagrant, and specifying the box to use for your Vagrant environment is always the first step after creating a new Vagrantfile.

Installing Hadoop (via cloudera)




www.cloudera.com/downloads/quickstart_vms/5-7.html
(4) MIDS-MLS-2015
nbviewer.ipython.org
Stanford Machine Le
Getting Started
Statistical Analysis H
eBay/Google 2013: E
Inquiries
InferPatents
WindAlert - Coyote F
Downloads
Training
Support Portal
Partners
Developers
Community
Search
Sign In
Language
cloudera
Why Cloudera
Products
Services & Support
Solutions
Get Started
QuickStart Downloads for CDH 5.7
A Single-Node Hadoop Cluster and Examples for Easy Learning!
The QuickStart VMs contain a single-node Apache Hadoop cluster, complete with example data, queries, scripts, and Cloudera Manager to manage your cluster. Cloudera QuickStart VMs are for demo purposes only and are not to be used as a starting point for clusters.
For the best download experience, use of a Download Manager is highly recommended.
Get Started
QUICKSTART DOWNLOADS FOR CDH 5.7
SELECT A PLATFORM
DOWNLOAD NOW

Cloudera QuickStart virtual machines (VMs) include everything you need to try CDH, Cloudera Manager, Cloudera Impala, and Cloudera Search. The VM uses a package-based install. This allows you to work with or without Cloudera Manager. Parcels do not work with the VM unless you first migrate your CDH installation to use parcels. On your production systems, Cloudera recommends that you use parcels.

Prerequisites

- These 64-bit VMs require a 64-bit host OS and a virtualization product that can support a 64-bit guest OS.
- To use a VMware VM, you must use a player compatible with WorkStation 8.x or higher:
  - Player 4.x or higher
  - Fusion 4.x or higher

  Older versions of WorkStation can be used to create a new VM using the same virtual disk (VMDK file), but some features in VMware Tools are not available.
- The amount of RAM required varies by the run-time option you choose:

CDH and Cloudera Manager Version	RAM Required by VM
CDH 5 (default)	4+ GiB*
Cloudera Express	8+ GiB*
Cloudera Enterprise (trial)	10+ GiB*

*Minimum recommended memory. If you are running workloads larger than the examples provided, consider allocating additional memory.

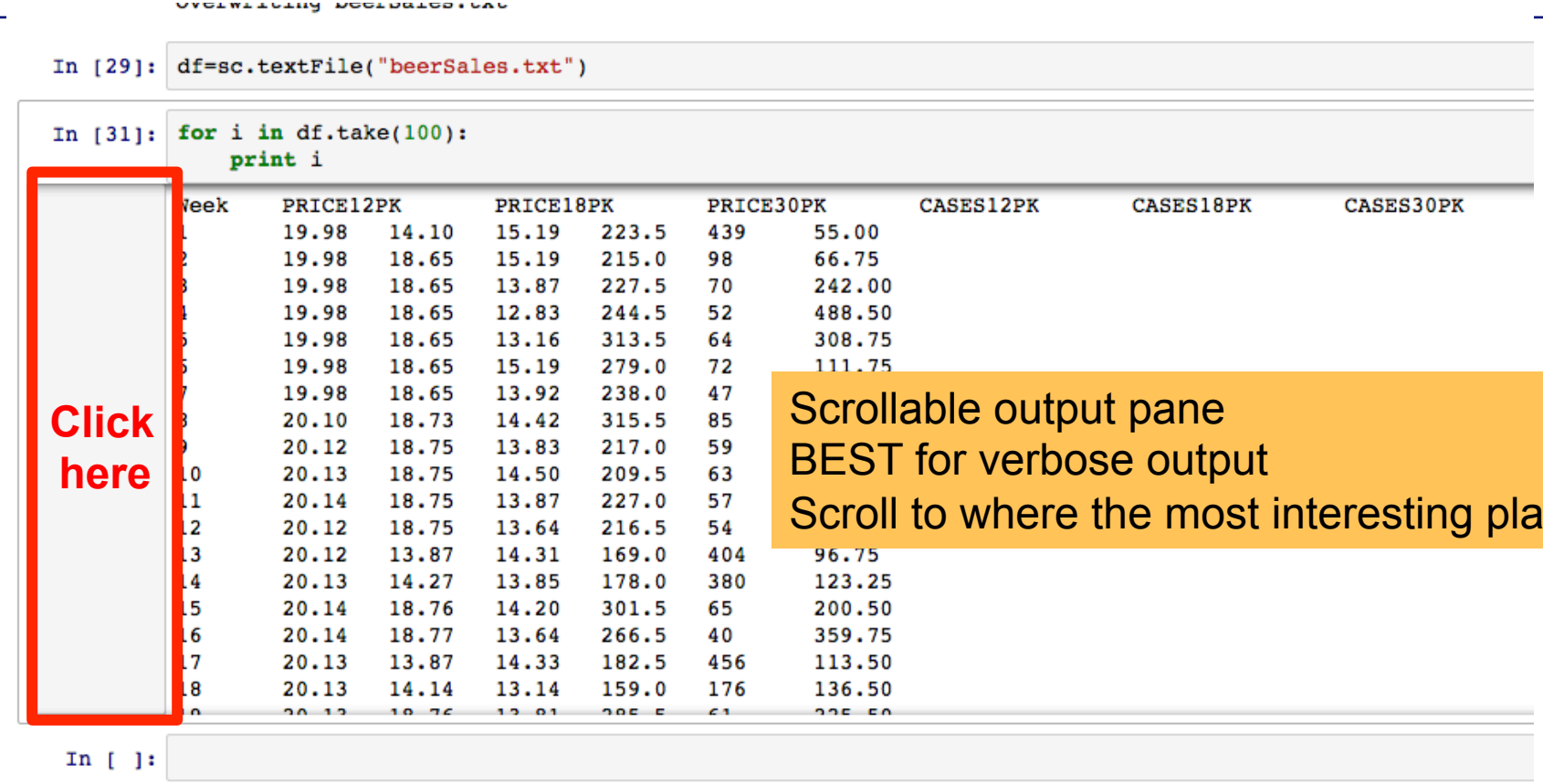
Install python, ipython

- **Brew python**

Independent input variables that are weakly predictive of the target variable

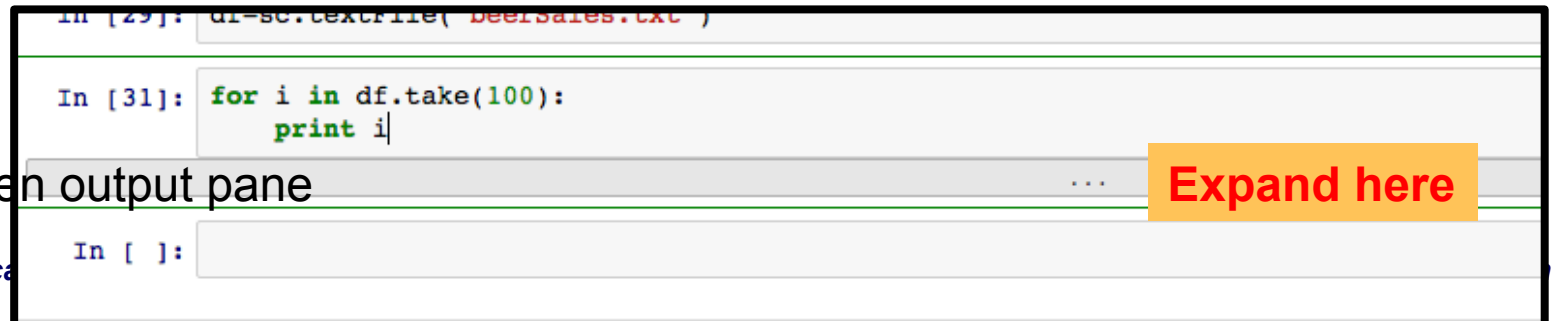
- What makes this interesting is that the singular values tell us something about the importance of the features represented by the left and right singular vectors (the vectors are the rows of U and T). In particular, the singular values tell us the extent to which the corresponding feature vectors are independent.
- Consider the implication of interdependent or covariant features. Or to make it a bit easier, imagine that two features, A and B, are identical. Once feature A has been considered by the model, feature B has nothing to contribute. It contains no new information.
- As builders of predictive models, the features we want are independent, and each one is at least a weak predictor of our target. If we have many many weak predictors, so long as their predictions are better than random, in combination they gain strength. But this phenomenon, the ensemble effect, only works when features are independent.

Ipython: Control output: choose style



Click here

Scrollable output pane
BEST for verbose output
Scroll to where the most interesting place



Hidden output pane

Expand here

```
In [29]: df=sc.textFile("beerSales.txt")
```

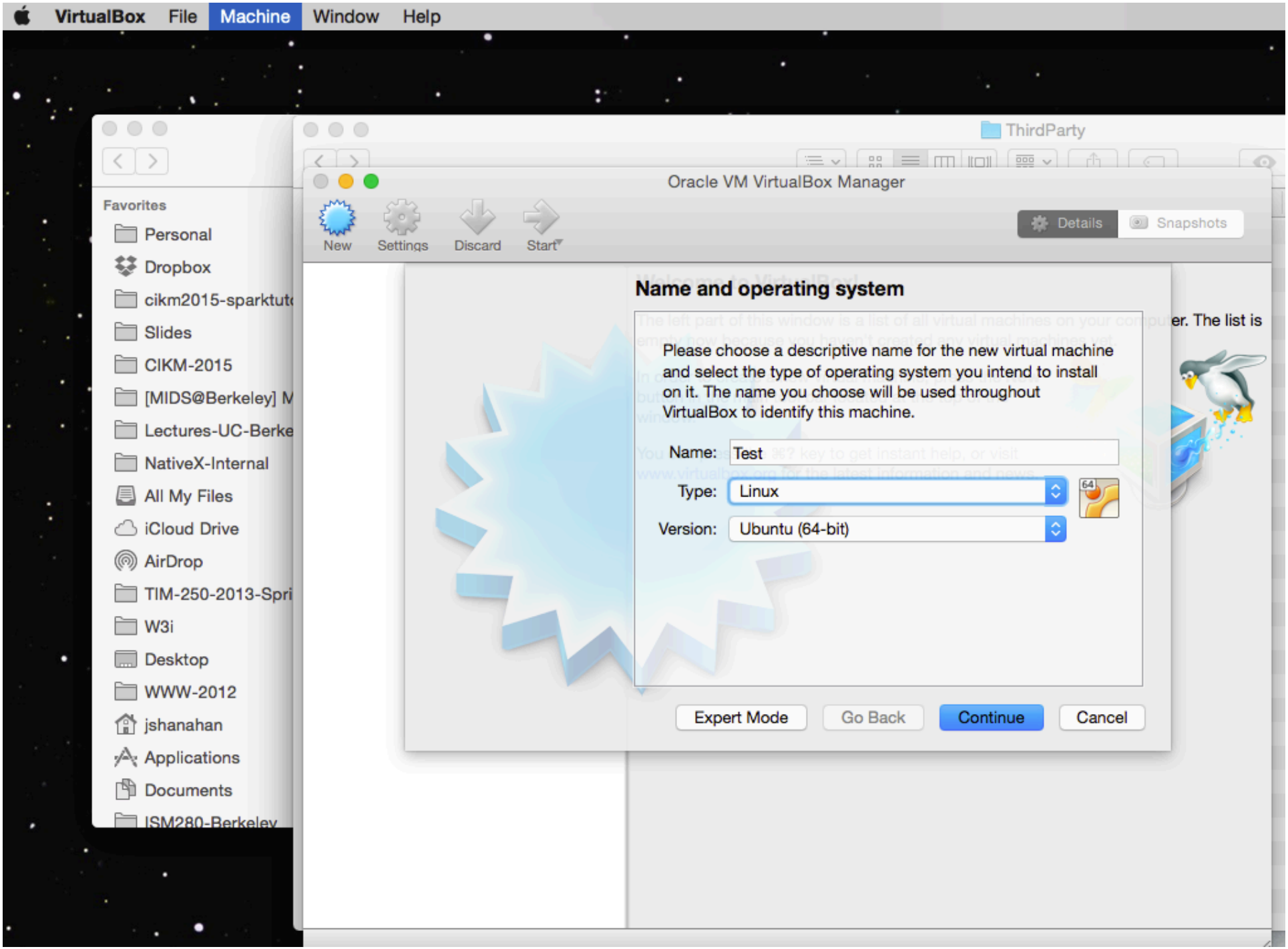
```
In [31]: for i in df.take(100):
             print i
```

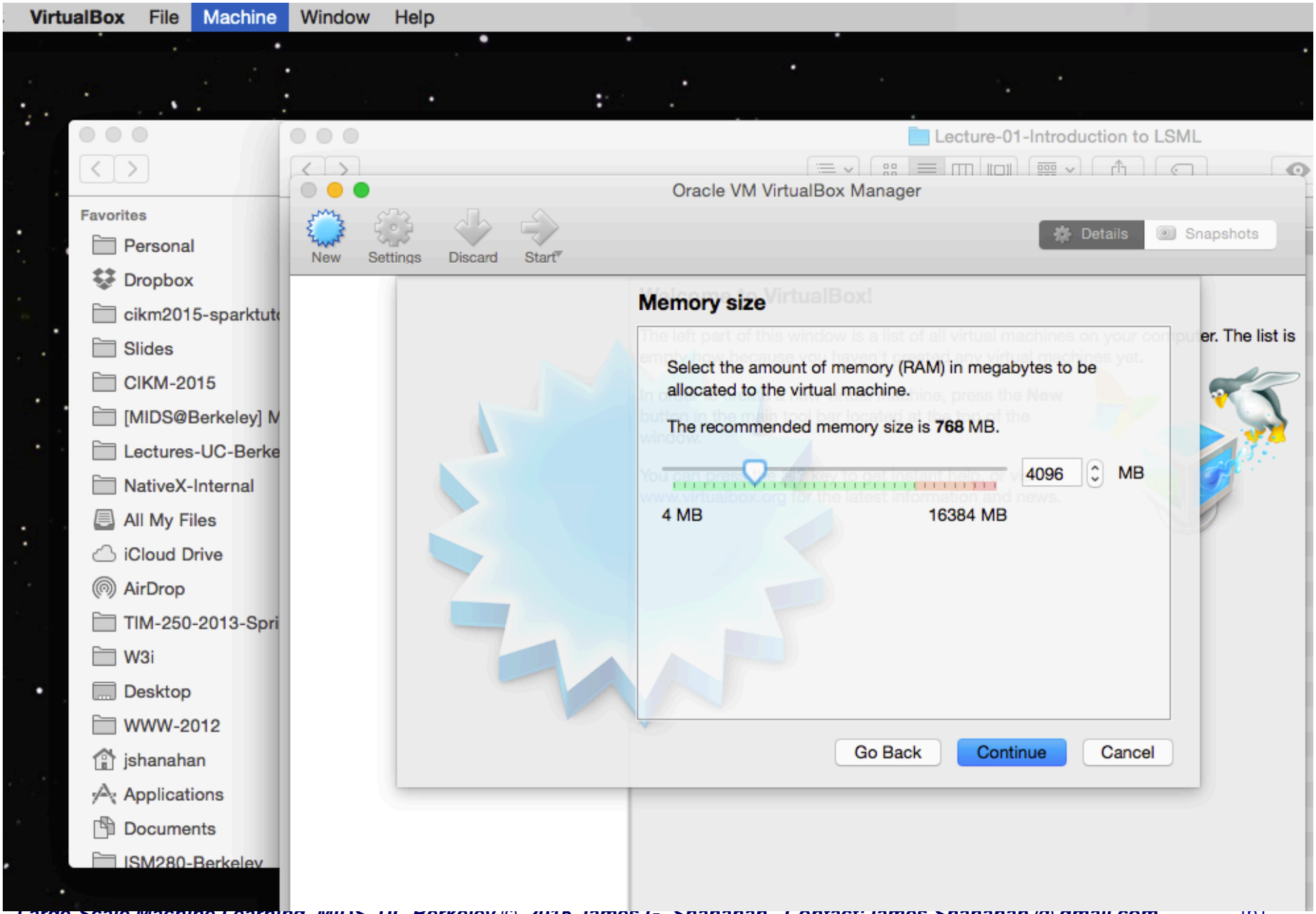
```
Week    PRICE12PK       PRICE18PK       PRICE30PK       CASES12PK       CASES18PK       CASES30PK
1       19.98   14.10   15.19   223.5   439     55.00
2       19.98   18.65   15.19   215.0   98      66.75
3       19.98   18.65   13.87   227.5   70      242.00
4       19.98   18.65   12.83   244.5   52      488.50
5       19.98   18.65   13.16   313.5   64      308.75
6       19.98   18.65   15.19   279.0   72      111.75
7       19.98   18.65   13.92   238.0   47      252.50
8       20.10   18.73   14.42   315.5   85      221.25
9       20.12   18.75   13.83   217.0   59      245.25
10      20.13   18.75   14.50   209.5   63      148.50
11      20.14   18.75   13.87   227.0   57      229.75
12      20.12   18.75   13.64   216.5   54      312.00
13      20.12   13.87   14.31   169.0   404     96.75
14      20.13   14.27   13.85   178.0   380     123.25
15      20.14   18.76   14.20   301.5   65      200.50
16      20.14   18.77   13.64   266.5   40      359.75
17      20.13   13.87   14.33   182.5   456     113.50
18      20.13   14.14   13.14   159.0   176     136.50
19      20.13   18.76   13.81   285.5   61      225.50
20      20.13   18.72   15.19   360.0   91      122.25
21      20.13   18.76   13.13   263.0   59      443.75
22      19.18   18.76   13.63   443.5   83      322.75
23      14.78   18.74   15.19   1101.5  41      53.00
24      16.04   18.75   13.89   814.0   47      140.75
25      20.12   18.75   14.28   365.0   84      210.75
26      19.75   18.75   15.19   510.0   85      110.50
27      19.65   18.75   13.12   580.5   116     568.25
28      19.69   13.79   13.78   251.0   544     115.50
29      20.12   13.49   15.19   237.0   890     58.75
30      20.12   14.89   15.19   302.5   371     77.25
31      20.13   13.94   15.19   229.5   557     66.25
32      20.14   13.67   15.19   188.5   775     50.00
33      15.14   14.43   15.19   795.5   236     46.50
34      14.33   18.75   15.19   1556.5  43      65.75
35      16.24   18.22   13.14   807.5   63      252.75
36      19.93   14.06   13.45   243.0   469     179.00
37      21.06   14.43   13.00   201.5   335     226.25
38      21.19   19.48   13.60   294.0   75      288.50
39      21.23   15.15   14.46   220.5   461     114.25
40      20.12   13.79   14.94   255.5   817     70.00
41      14.73   14.31   15.19   920.5   200     47.75
42      14.57   19.50   15.19   730.0   32      98.75
43      15.94   13.85   15.19   262.5   460     77.00
44      20.70   14.23   13.43   209.5   751     160.50
45      19.57   19.31   14.37   283.0   70      143.50
46      19.60   19.29   15.19   262.5   80      133.00
47      19.94   13.76   15.19   310.0   523     68.75
48      21.28   13.45   15.19   278.5   741     81.75
49      14.56   15.13   15.19   741.5   130     56.25
50      14.39   19.43   15.19   1316.0  69      68.75
51      16.81   13.26   15.19   449.0   493     49.25
52      19.86   13.92   15.19   505.0   814     76.50
```

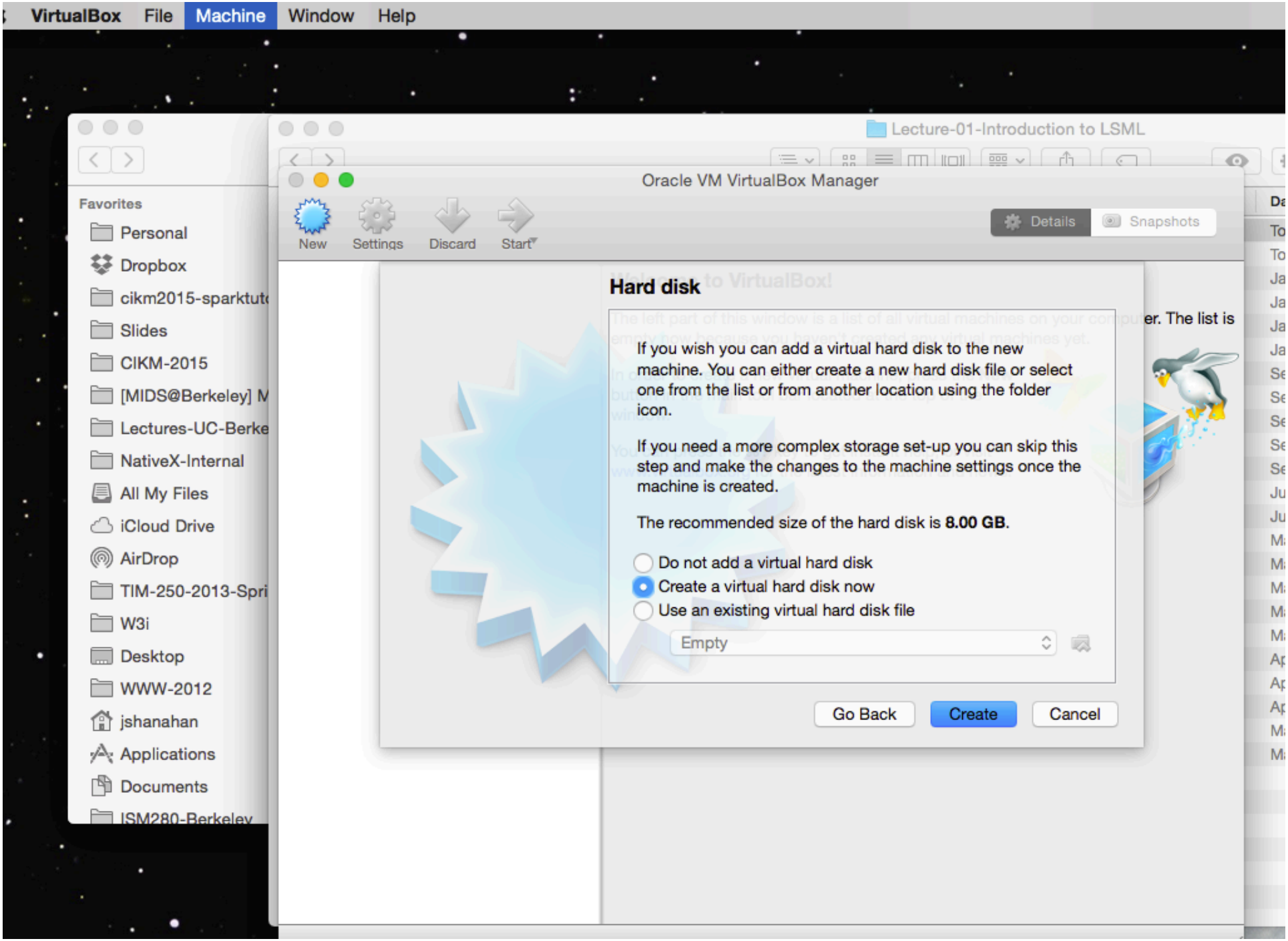
```
In [ ]:
```

Show full output

man tail
man head


VirtualBox File Machine Window Help
ThirdParty
Favorites
Personal
Dropbox
cikm2015-sparktut
Slides
CIKM-2015
[MIDS@Berkeley] M
Lectures-UC-Berke
NativeX-Internal
All My Files
iCloud Drive
AirDrop
TIM-250-2013-Spri
W3i
Desktop
WWW-2012
jshanahan
Applications
Documents
ISM280-Berkeley
Oracle VM VirtualBox Manager
New
Settings
Discard
Start
Details
Snapshots
Name and operating system
Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.
Name: Test
Type: Linux
Version: Ubuntu (64-bit)
Expert Mode
Go Back
Continue
Cancel


VirtualBox File Machine Window Help
Lecture-01-Introduction to LSML
Favorites
Personal
Dropbox
cikm2015-sparktut
Slides
CIKM-2015
[MIDS@Berkeley] M
Lectures-UC-Berke
NativeX-Internal
All My Files
iCloud Drive
AirDrop
TIM-250-2013-Spri
W3i
Desktop
WWW-2012
jshanahan
Applications
Documents
ISM280-Berkeley
Oracle VM VirtualBox Manager
New
Settings
Discard
Start
Details
Snapshots
Memory size
Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.
The recommended memory size is 768 MB.
4096
MB
4 MB
16384 MB
Go Back
Continue
Cancel


VirtualBox File Machine Window Help
Lecture-01-Introduction to LSML
Favorites
Personal
Dropbox
cikm2015-sparktut
Slides
CIKM-2015
[MIDS@Berkeley] M
Lectures-UC-Berke
NativeX-Internal
All My Files
iCloud Drive
AirDrop
TIM-250-2013-Spri
W3i
Desktop
WWW-2012
jshanahan
Applications
Documents
ISM280-Berkeley
Oracle VM VirtualBox Manager
New
Settings
Discard
Start
Details
Snapshots
Hard disk
If you wish you can add a virtual hard disk to the new machine. You can either create a new hard disk file or select one from the list or from another location using the folder icon.
If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.
The recommended size of the hard disk is 8.00 GB.
Do not add a virtual hard disk
Create a virtual hard disk now
Use an existing virtual hard disk file
Empty
Go Back
Create
Cancel

Install iPython on VirtualBox

- **How run code on a unix machine (when I have a windows)?**
  - Set up a virtual machine (e.g., a linux box) on your Windows/Mac machines
    - Install VirtualBox and then install Vagrant (this will install iPython etc on the virtual box)
  - Follow all steps in the following slides to install VirtualBox and Vagrant
    - https://www.dropbox.com/s/z8wri7bghh4js9h/Simplified-IPython-Installation-V2-onLinux-VirtualBox.pdf?dl=0
  - Vagrant
    - Instead of building a virtual machine from scratch, which would be a slow and tedious process, Vagrant uses a base image to quickly clone a virtual machine. These base images are known as "boxes" in Vagrant, and specifying the box to use for your Vagrant environment is always the first step after creating a new Vagrantfile.

Install python, ipython

- **Brew python**