

Welcome to Data Science

Ivan Corneillet

Data Scientist

Data Scientists: The Sexiest Job of the 21st Century





DS

What is Data Science?










Elections 2016

The New York Times

New Hampshire Primary Results



FEB. 10, 2016, 2:45 PM ET

Republican Primary

CANDIDATES	VOTE	PCT.	DELEGATES
 Donald J. Trump ✓	100,406	35.3% <div><div></div></div>	10
 John Kasich	44,909	15.8 <div><div></div></div>	4
 Ted Cruz	33,189	11.7 <div><div></div></div>	3
 Jeb Bush	31,310	11.0 <div><div></div></div>	3
 Marco Rubio	30,032	10.6 <div><div></div></div>	3
 Chris Christie	21,069	7.4 <div><div></div></div>	-
 Carly Fiorina	11,706	4.1 <div><div></div></div>	-
 Ben Carson	6,509	2.3 <div><div></div></div>	-
 Jim Gilmore	133	0.0	-
Other	4,857	1.7 <div><div></div></div>	-

284,120 votes, 100% reporting (300 of 300 precincts)

Democratic Primary

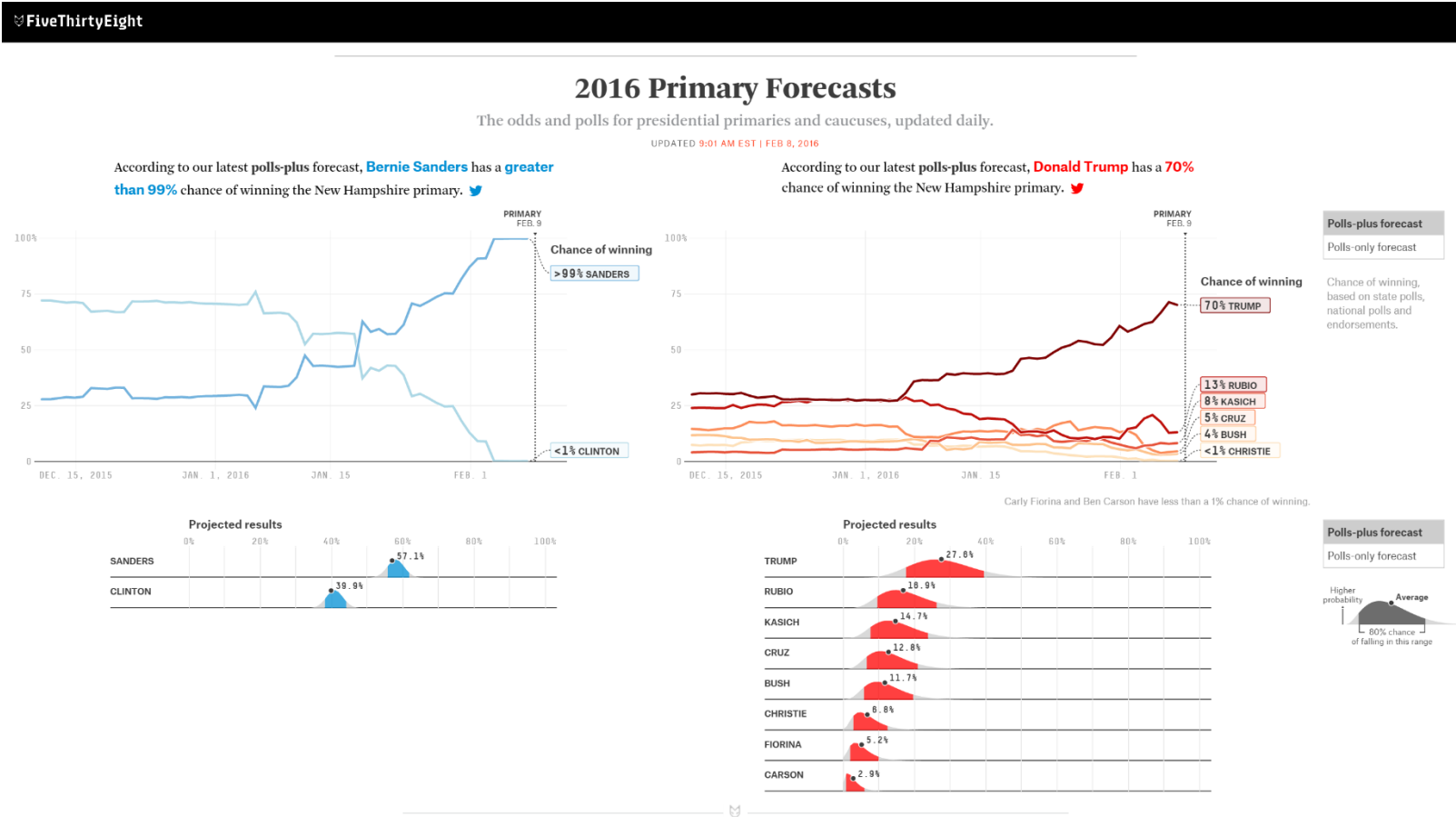
CANDIDATES	VOTE	PCT.	DELEGATES
 Bernie Sanders ✓	151,578	60.4% <div><div></div></div>	15
 Hillary Clinton	95,249	38.0 <div><div></div></div>	9
Other	4,147	1.7 <div><div></div></div>	-

250,974 votes, 100% reporting (300 of 300 precincts)

FiveThirtyEight



Data Science by FiveThirtyEight



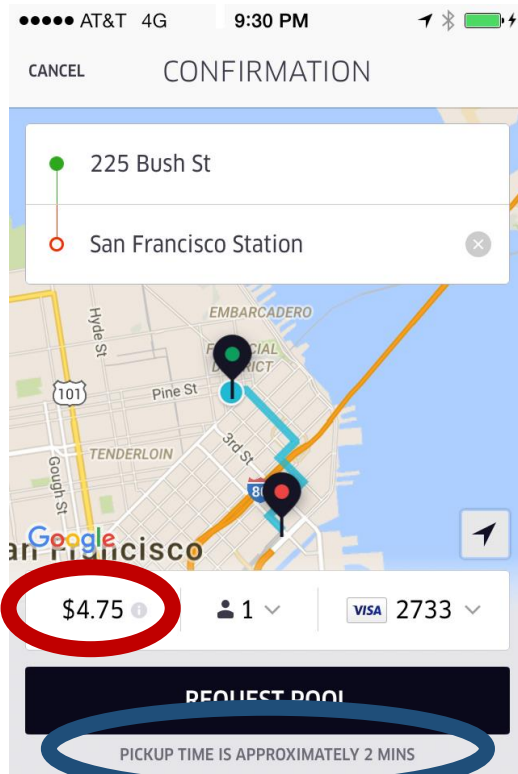
Über

U B E R

Data Science by Uber

For Riders...

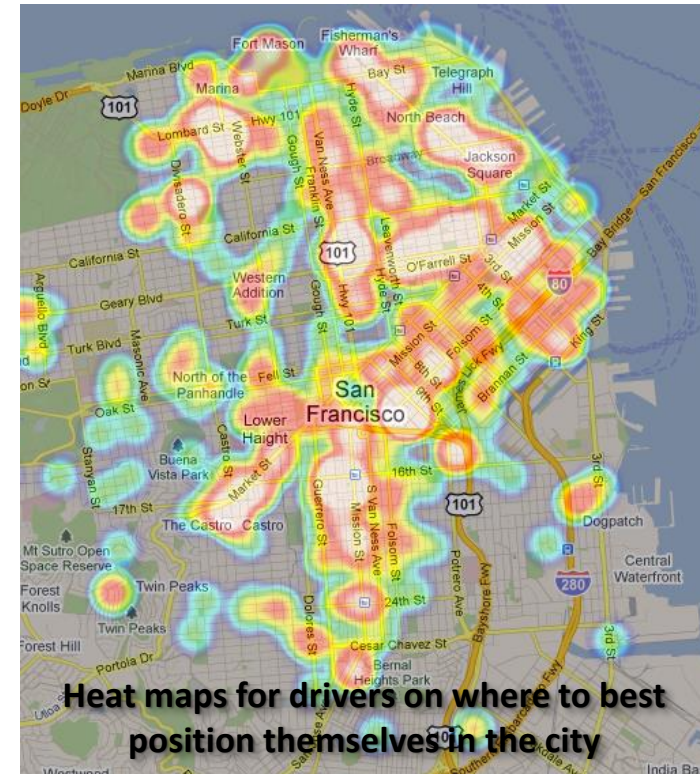
**Fare Estimator and
Dynamic Pricing
(e.g., Surge Pricing)
Algorithms**



ETA (Estimated Time of Arrival) Algorithms

U B E R

and Drivers!



**Heat maps for drivers on where to best
position themselves in the city**

Data Science Based-Business Models is the New Normal

 **FiveThirtyEight**

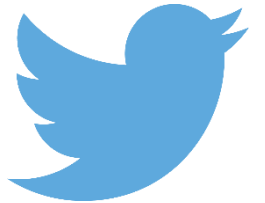
NETFLIX



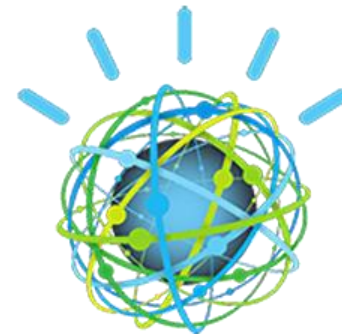
Walmart 



amazon 



Google



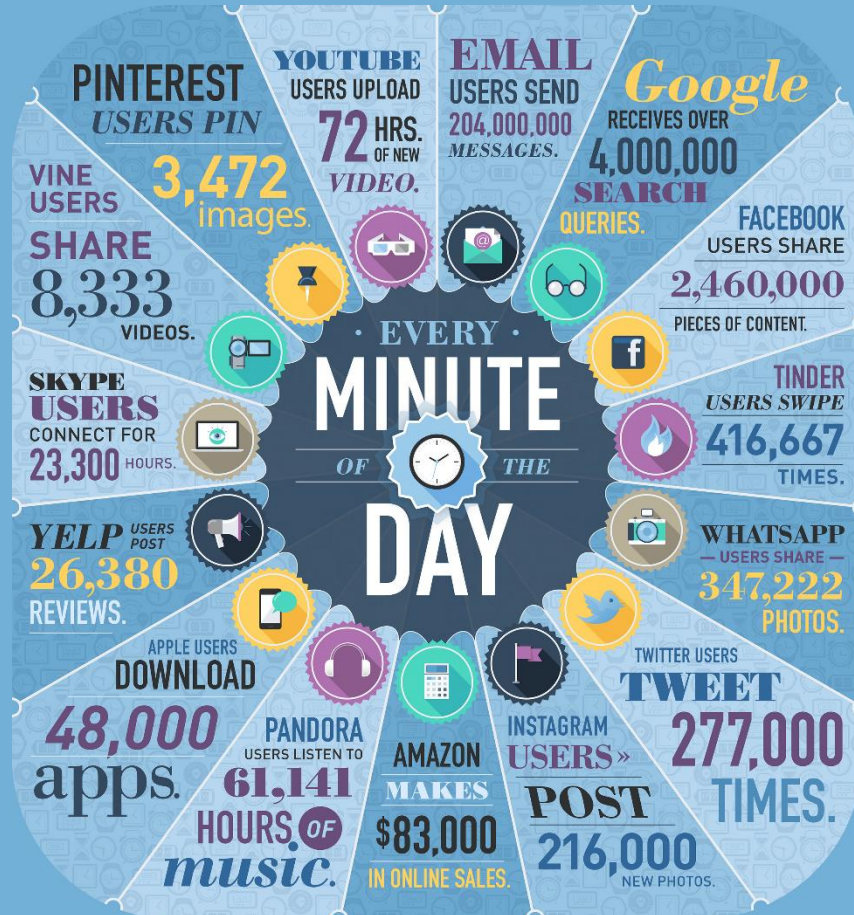
IBM Watson

facebook

UBER

Linked in

The Data Never Sleeps



DATA NEVER SLEEPS 2.0

How Much Data is Generated Every Minute?

Data is being created every minute of every day without us even noticing it. Given how much information is floating around these days, it's tempting to talk about big data only in terms of size. Big data describes the massive avalanche of digital activity pulsating through cables and airwaves, but it also describes all the things we were never able to measure before. With every status we share, every article we read or every photo we upload, we are creating a digital trail that tells a story. Below, we explore how much data is generated in one minute.



THE GLOBAL INTERNET POPULATION GREW **14.3%** FROM 2011 - 2013 AND NOW REPRESENTS

2.4 BILLION PEOPLE.

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. Learn more at www.domo.com.

SOURCES:

BITS.BLOGS.NYTIMES.COM, INTEL.COM, APPLE.COM, TIME.COM, DAILYMAIL.CO.UK, SKYPE.COM, STATISTICBRAIN.COM

DOMO



DS

Who are Data Scientists?

Who are Data Scientists?



EXERCISE

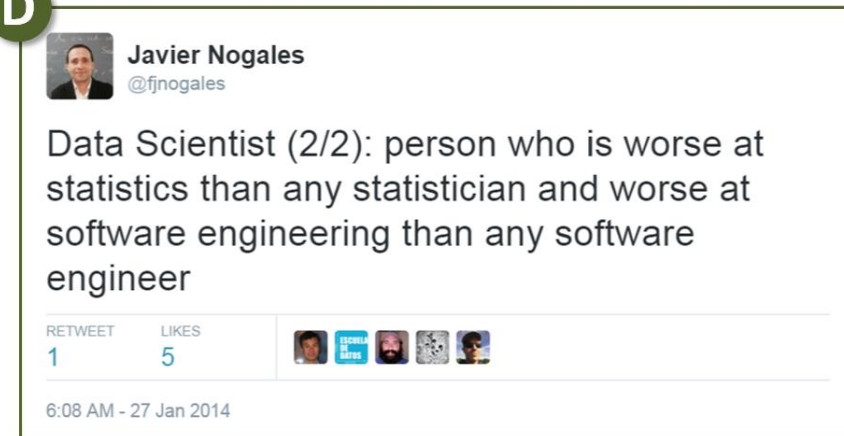
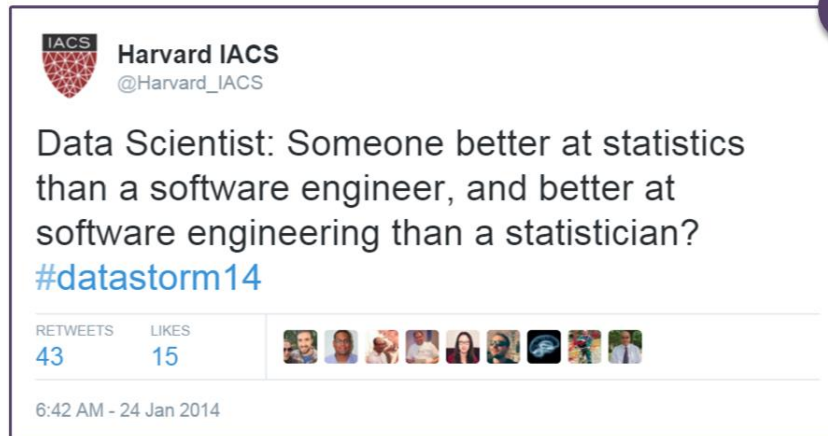
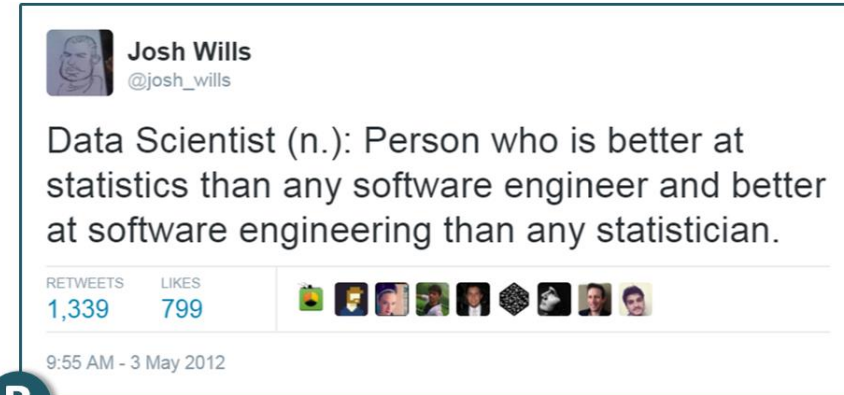
DIRECTIONS (10 minutes)

1. Form groups of 3
2. Answer the following questions
 - a. Who are Data Scientists?
 - b. How do Data Scientists add value?
 - c. What makes a good Data Scientist?

DELIVERABLE

Answers to the above questions

Data Scientists in ≤ 140 characters





DS

Welcome to GA and DS

WELCOME TO GA!





Today

- Welcome to GA and DS!
- Setting You up for Success
 - Learning Objectives
 - Logistics
 - Meet Your Team
 - Typical Class
 - Course Philosophy
 - Road to Success
- What is Data Science?
- The Data Science Workflow
- Setting Up the Development Environment
- Assigned
 - Unit Project 1 (due in 1 week)
 - Final Project 1 (due in 3.5 weeks)

A black circle containing the white letters "DS".

DS

Setting You Up for Success

Learning Objectives

After this lesson, you should be able to:

- Describe the roles and components of a successful learning environment
- Define data science and the data science workflow
- Apply the data science workflow to meet your classmates
- Setup your development environment and review Python basics

Logistics

- Instructor

- Ivan Corneillet (ivan+GA@paspeur.com)

- Experts-in-Residence

- Jeremiah Gaw (jeremiah.gaw@gmail.com)
 - Azi Hussain (asjedhussain@gmail.com)

- Course Producer

- Vanessa Ohta (vanessa@generalassemb.ly)

- Class

- February 25 – May 3, Tuesdays and Thursdays, 6:30PM – 9:30PM

- Classroom 7

- GitHub

- <https://github.com/ga-students/SF-DAT-21>

- Slack

- <https://sf-dat-21.slack.com>

What can I expect to learn?


Research Design and Data Analysis	Research Design	Data Visualization in Pandas	Statistics	Exploratory Data Analysis in Pandas
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and Random Forest	Time Series Data	Natural Language Processing	Databases

Projects

<div>Unit Project</div> <div>You will design a research project, perform exploratory data analysis and build a logistic model to determine what factors affect admission the most</div>	Research Design		Exploratory Data Analysis		Logistic Modeling	Executive Summary with Findings
<div>Final Project</div> <div>Using a dataset of your choosing, you will design a project, build a data science model and present their finding to the course</div>	Lightning Presentation	Experimental Write-up	Exploratory Analysis	Notebook Draft	Final Presentation	

Past-Student Projects

(<https://gallery.generalassemb.ly/DS>)



Data Science


All Cities

LOGIN

THE GALLERY


A COLLECTION OF STUDENT-UPLOADED PROJECTS

DS




READ LIKE YOU TWEET NYC
by Karsten Kreis

DS




RETHINKING THE COMMUTE SF
by Gregory Naifeh

DS



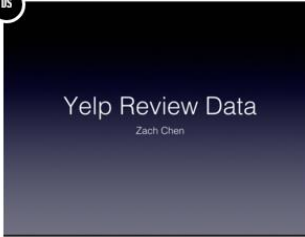
PREDICTING REDDIT POPULARITY NYC
by Nick Pico

DS




PREDICT DISCOUNTS: 2POINTB DC
by Deron Hogsans

DS



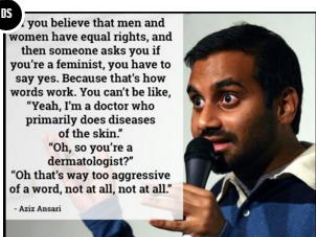
YELP REVIEW DATA ANALYSIS NYC
by Zegru Chen

DS



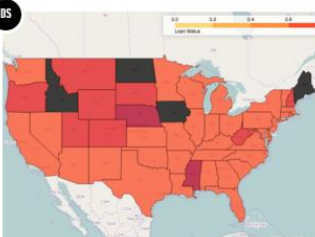
TALK TO ME NYC
by Inna Starikova

DS



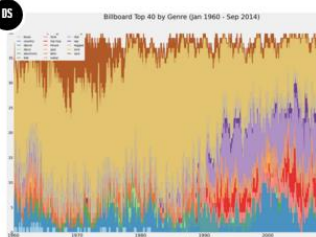
TWITTER & SENTIMENT ANALYSIS LA
by Heather Cohen

DS




P2P LOAN ANALYSIS NYC
by Osh Dubno

DS



BILLBOARD TOP 40 ANALYSIS NYC
by Matthew Lentz

DS



HATER NEWS NYC
by Kevin McAlear

Detailed Class Schedule

Tuesday						Thursday					
February							25	U1	P1	L1	What is Data Science
March	Research Design and Pandas	L2		1		U1	3	U2		L3	Statistics Fundamentals I
	Statistics Fundamentals II	L4		8		U2	10	U3		L5	<i>Flexible Class Session</i>
	Introduction to Regression	L6		15			17			L7	Evaluating Model Fit
	Introduction to Classification	L8	P1	22	P2	U3	24	U4		L9	Introduction to Logistic Regression
	Communicating Model Results	L10		29		U4	31			L11	<i>Flexible Class Session</i>
April	Decision Trees and Random Forests	L12		5			7			L13	Natural Language Processing
	Dimensionality Reduction	L14	P2	12	P3		14			L15	Time Series Data I
	Time Series Data II	L16	P3	19	P4		21			L17	Database Technology
	Where to Go Next	L18	P4	26	P5		28			L19	<i>Flexible Class Session</i>
May	Final Project Presentations	L20	P5	3							
Unit 1 – Research Design & Data Analysis						Px – Final Project Assigned		Ux – Unit Project Assigned			
Unit 2 – Foundations of Modeling						Px – Final Project Due		Ux – Unit Project Due			
Unit 3 – Data Science in the Real World											

Meet Your Team

▸ Ivan Corneillet, Instructor



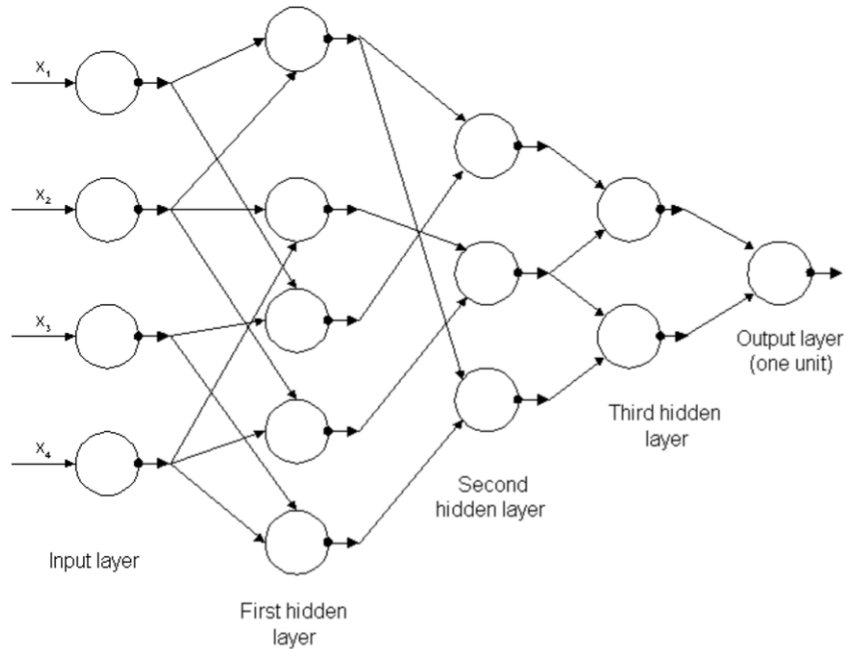
▸ Azi Hussain, Expert-in-Residence

▸ Jeremiah Gaw, Expert-in-Residence

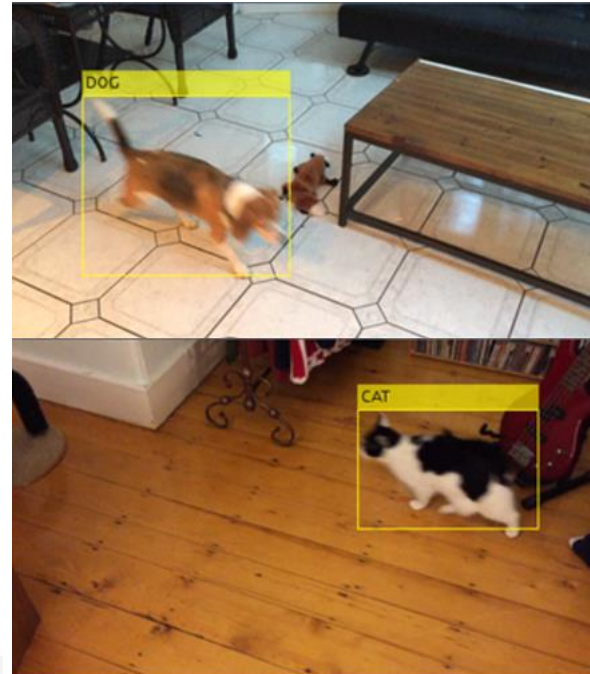


▸ Vanessa Ohta, Course Producer

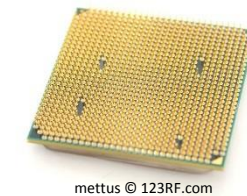
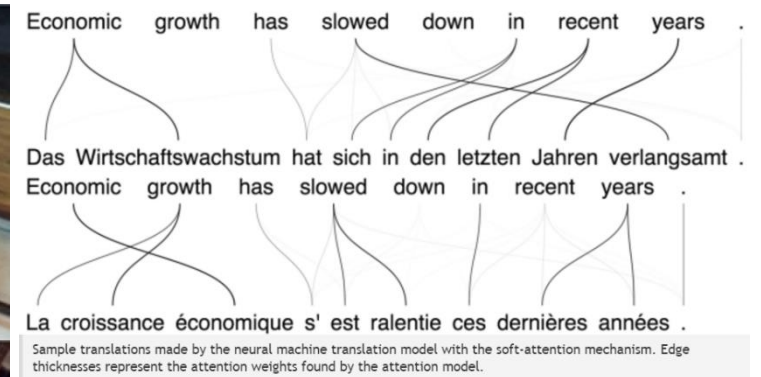
Deep Learning and High-Performance Computing



The architecture of the first known deep network which was trained by Alexey Grigorevich Ivakhnenko in 1965. The feature selection steps after every layer lead to an ever-narrowing architecture which terminates when no further improvement can be achieved by the addition of another layer.



Pet detection and recognition system.



Typical Class

- Recap of last time
- Series of mini-lectures and practices (exercises and codealongs)
- Lab

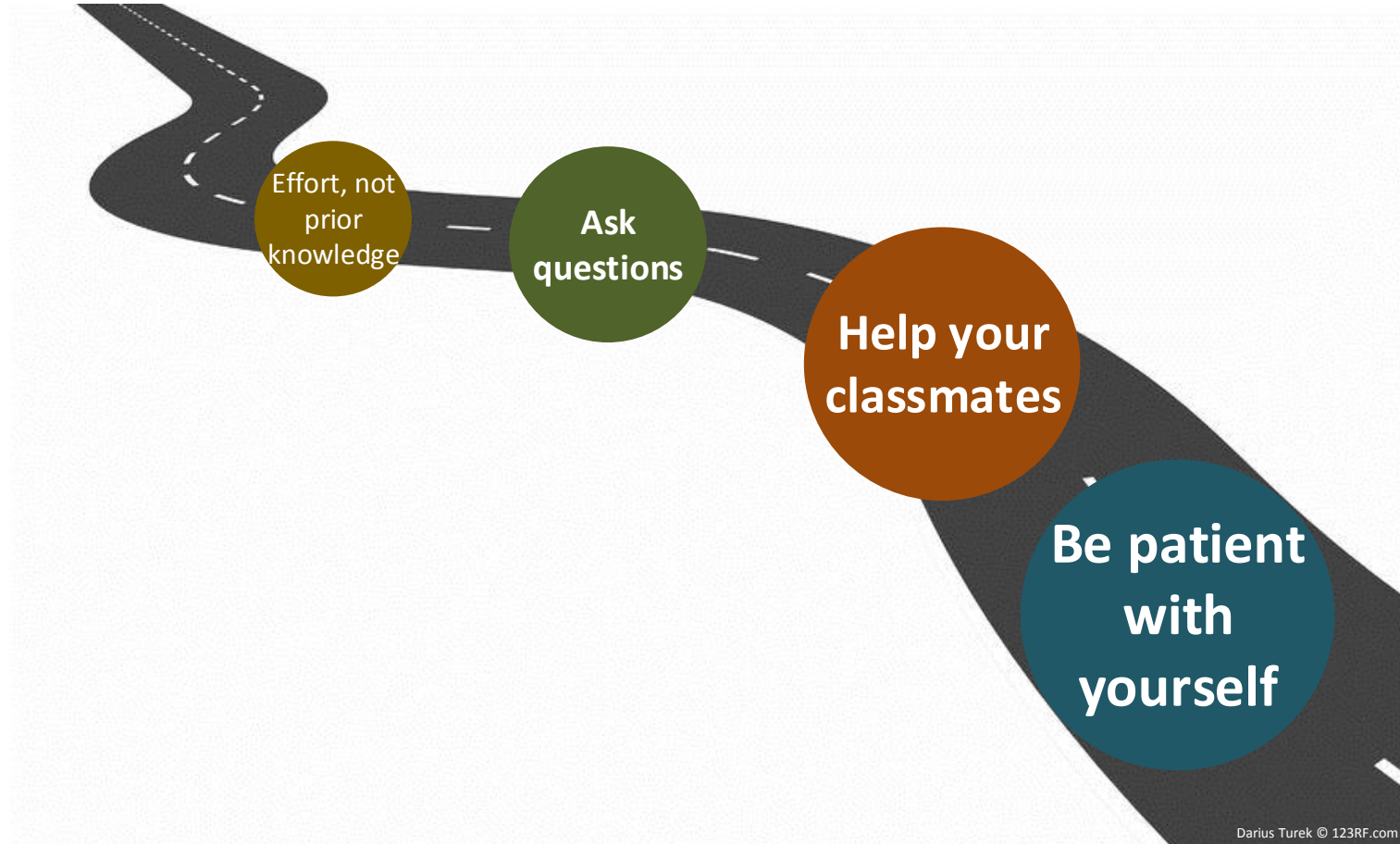
Course Philosophy

- Embrace diversity
- Seek an optimal pace
- Communicate early and often
- Success is not a grade
- Application-based approach
- Understand key principles
- Balance depth with breadth
- Course project

Your 4 Steps to Graduation



Your DS Road to Success



Your GA Road to Success





DS

Q & A

DS

Pre-Work

Pre-Work

Before this lesson, you should already be able to:

- Define basic data types used in object-oriented programming
- Recall the Python syntax for lists, dictionaries, and functions
- Create files and navigate directories using the command line interface



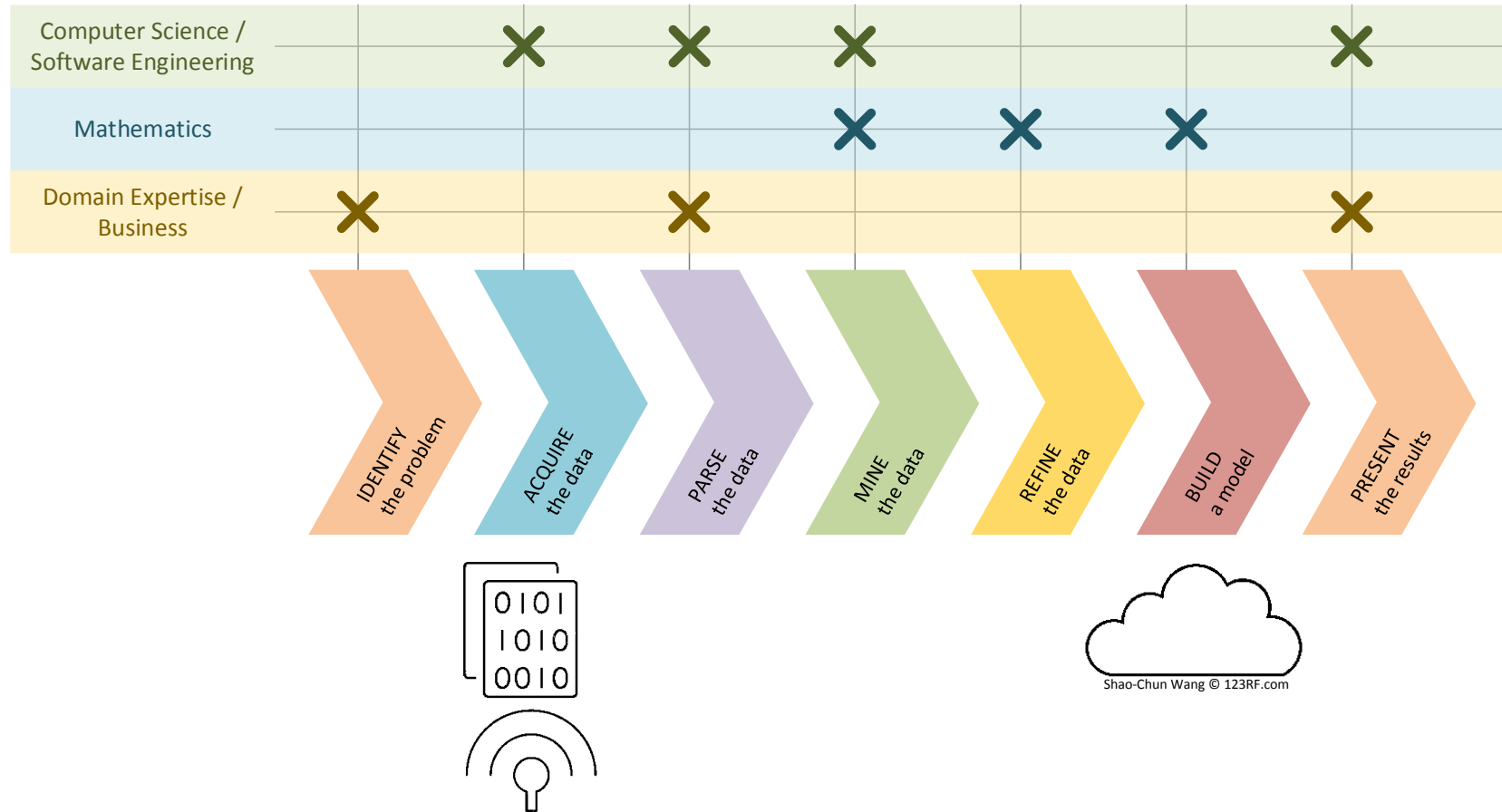
DS

What is Data Science?

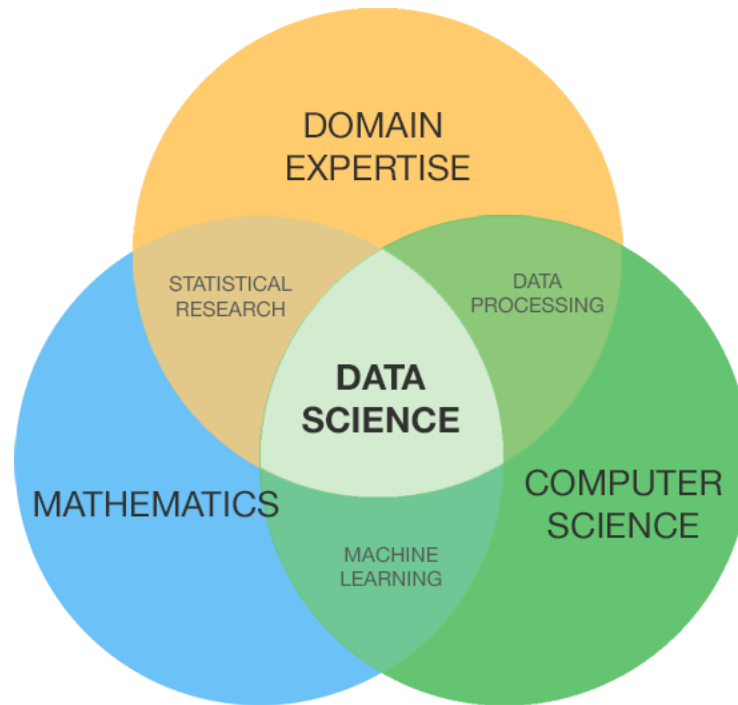
Data is the New Oil of the Digital Economy and IoT, Big Data, DS, and Cloud relate to one another



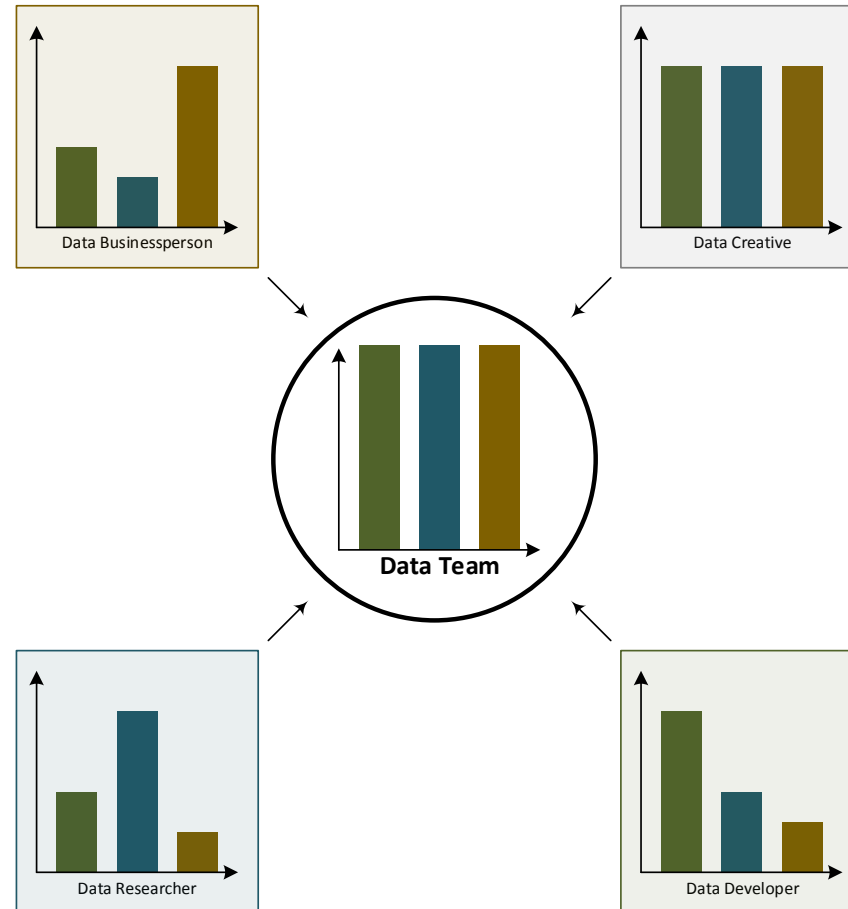
Data Science involves a variety of skillsets, not just one



Data Science involves a variety of skillsets, not just one (cont.)



Data Scientists have different roles that prioritize different skillsets but all roles involve some part of each skillset to form strong data teams



To sum it up

- Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms
- An (ideal) data scientist is “someone who has the both the engineering skills to acquire and manage large data sets, and also has the statistician’s skills to extract value from the large data sets and present that data to a large audience” – John Rauser



DS

Data Science Baseline

Data Science Baseline Quiz



EXERCISE

DIRECTIONS (10 minutes)

1. Form groups of 3
2. Answer the following questions
 - a. True or False: Gender (coded male=0, female=1) is a continuous variable.
 - b. According to the table on the next slide, BMI is the _____
 - i. Outcome
 - ii. Predictor
 - iii. Covariate
 - c. Draw a normal distribution
 - d. True or False: Linear regression is an unsupervised learning algorithm.
 - e. What is a hypothesis test?

DELIVERABLE

Answers to the above questions

Data Science Baseline Quiz

EXERCISE

Table 3. Adjusted mean¹ (95% confidence interval) of body mass index and serum concentration of metabolic biomarkers in American adults by categories of weekly frequency of fast-food or pizza meals, NHANES 2007-2010

BMI or serum Biomarker	Weekly frequency of fast food or pizza meals				P ²
	0 time	1 time	2-3 times	≥4 times	
BMI³, kg/m²					
All N=8169	27.5 (27.1, 27.8)	27.9 (27.6, 28.2)	28.9 (28.4, 29.4)	28.8 (28.3, 29.2)	<0.0001
Men n=4002	27.9 (27.4, 28.3)	28.0 (27.6, 28.4)	28.5 (28.0, 29.0)	28.6 (28.2, 29.0)	0.05
Women n=4167	27.2 (26.8, 27.6)	27.7 (27.3, 28.1)	29.3 (28.6, 29.9)	29.0 (28.1, 29.8)	<0.0001
Total cholesterol, mg/dL N=8236	199 (197, 202)	198 (196, 200)	199 (196, 201)	198 (196, 201)	0.5
HDL-cholesterol³, mg/dL					
All n=8236	54 (53, 55)	53 (52, 54)	52 (51, 53)	51 (50, 52)	<0.0001
Men n=4042	48 (47, 49)	48 (47, 49)	48 (46, 49)	46 (45, 47)	0.003
Women n=4194	60 (59, 61)	58 (57, 60)	56 (55, 57)	56 (54, 58)	0.001
LDL-cholesterol⁴, mg/dL					
All n=3604	113 (111, 116)	117 (113, 120)	113 (110, 116)	114 (110, 118)	0.6
<50 y n=2151	107 (105, 110)	112 (109, 116)	111 (107, 114)	108 (104, 112)	0.8
≥50 y n=1453	123 (118, 129)	126 (121, 131)	118 (113, 123)	129 (122, 137)	0.5
Triglycerides, mg/dL n=3659	103 (98, 109)	103 (99, 108)	110 (106, 115)	110 (104, 117)	0.2
Fasting glucose³, mg/dL					
All n=3668	99 (98, 100)	99 (98, 100)	99 (98, 100)	99 (98, 100)	0.5
Men n=1750	102 (101, 104)	102 (101, 104)	101 (99, 102)	101 (99, 102)	0.1
Women n=1918	97 (95, 98)	95 (94, 97)	97 (96, 99)	98 (96, 101)	0.2
Glycohemoglobin, % N=8234	5.42 (5.39, 5.44)	5.39 (5.36, 5.42)	5.39 (5.36, 5.42)	5.40 (5.37, 5.44)	0.2

¹Adjusted means were computed from multiple linear regression models with each biomarker as a continuous dependent variable. All biomarkers (except BMI, total and HDL cholesterol) were log-transformed for analysis; therefore, the back-transformed values for LDL-cholesterol, triglycerides, fasting glucose, and glycohemoglobin are geometric means and their 95% confidence intervals. Independent variables included: frequency of fast food meals (0, 1, 2-3, ≥4 times), age (20-39, 40-59, ≥60), sex, race/ethnicity (non-Hispanic white, non-Hispanic black, Mexican-American, other), Poverty income ratio (≤1.3, >1.3 to 3.5, ≥3.5, unknown), years of education (<12, 12, some college, ≥college), serum cotinine (continuous), hours of fasting before phlebotomy, (continuous), physical activity (none, tertiles of MET minutes/week), alcohol drinking status (never drink, former drinker, current drinker, unknown). N refers to observations used in the regression model for each biomarker. ²P value for the Satterthwaite-adjusted F test for frequency of fast food meals as a continuous variable. ³Significant interaction of fast-food meals with sex ($P_{\text{interaction}} < 0.05$), thus the results are stratified by sex. ⁴Significant interaction of frequency of fast-food meals with age ($P_{\text{interaction}} < 0.05$), thus the results are stratified by age categories.



DS

The Data Science Workflow

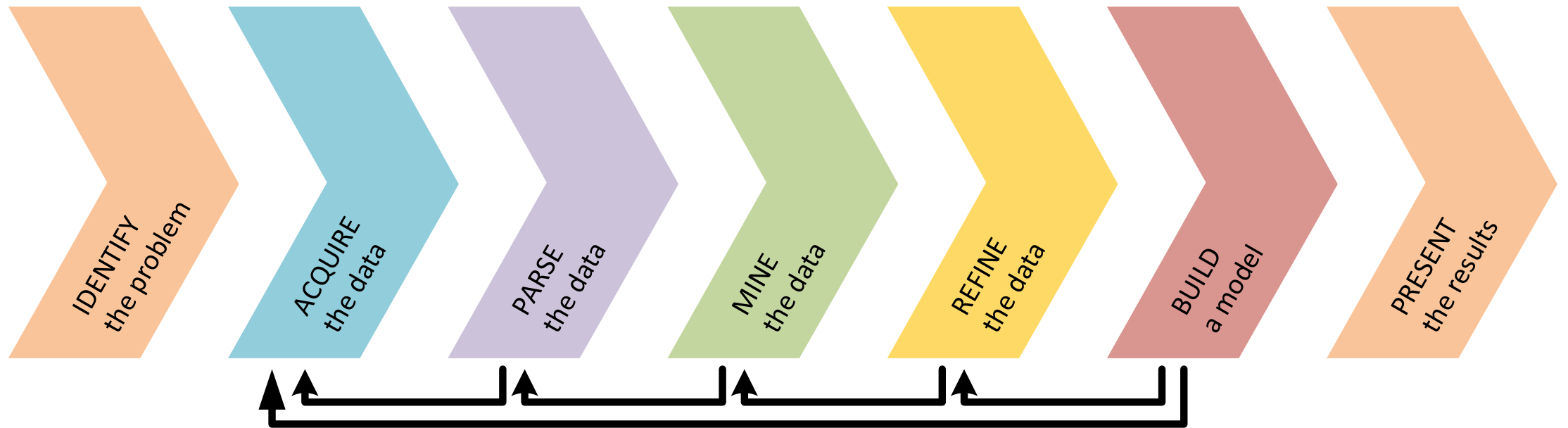
What is the Data Science Workflow for?

- A methodology for Data Science to produce *reliable* and *reproducible* results
 - **Reliable:** Accurate findings
 - **Reproducible:** Others can follow your steps and get the same results
- Similar to the scientific method

The scientific method:

- Ask a Question
- Do Background Research
- Construct a Hypothesis
- Test Your Hypothesis by Doing an Experiment
- Analyze Your Data and Draw a Conclusion
- Communicate Your Results

The Data Science Workflow (also called the Data Science Pipeline)



The Data Science Workflow is at the core of this course (sessions and projects)

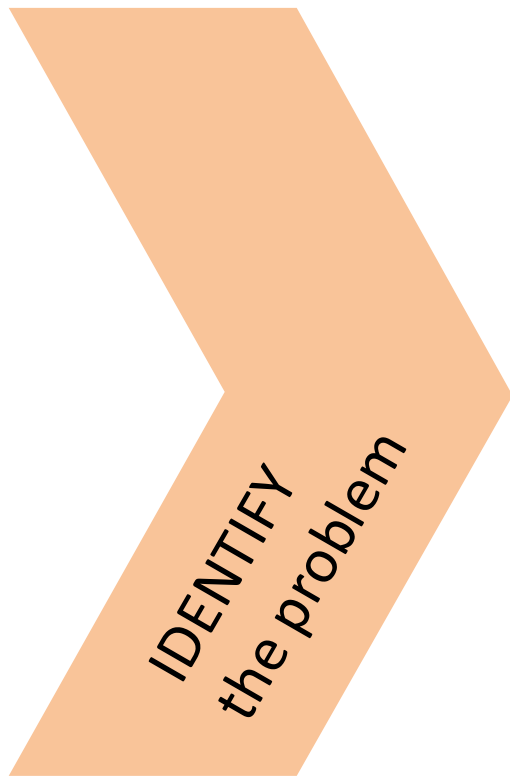
Step	Sessions	Unit Project	Final Project
IDENTIFY	2	1	1 – 2
ACQUIRE	2	1	1 – 2
PARSE	2 – 4	2	3
MINE	3 – 4	2	3
REFINE	3 – 4	3	4
BUILD	6 – 10, 12 – 16	3	4
PRESENT	10, 20	4	5

A dark blue circle containing the white text "DS".

DS

Q & A

Identify the Problem



- Identify the Problem
 - Identify business/product objectives
 - Identify and hypothesize goals and criteria for success
 - Create a set of questions for identifying correct dataset

Neonatal Infections on “Preemies”

CASE
STUDY



Identify the Problem (cont.)

- Identify the Problem

- Identify business/product objectives
- Identify and hypothesize goals and criteria for success
- Create a set of questions for identifying correct dataset

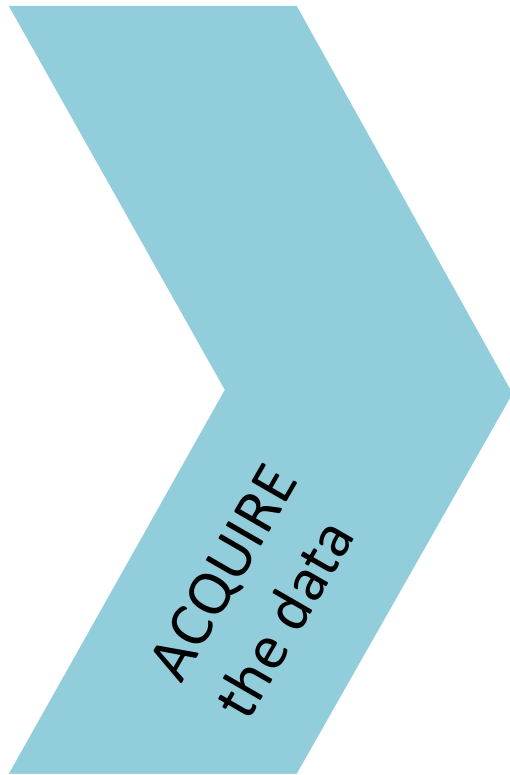
- Background

- Premature babies or “preemies” are at high risk of neonatal infections (infections within the first four weeks after birth)
- Neonatal intensive care involves sophisticated measurement of the preemies’ temperature, respiration, cardiac function, oxygenation, and brain activity
- However most of that vast stream of data (1,260 data points per second) is tossed away

- Objective

- Can we better use these measurements to predict the onset of an infection before overt symptoms appear and get the babies treated earlier?

Acquire the Data



- Acquire the Data
 - Identify the “right” dataset(s)
 - Import data and set up local or remote data structure
 - Determine most appropriate tools to work with data

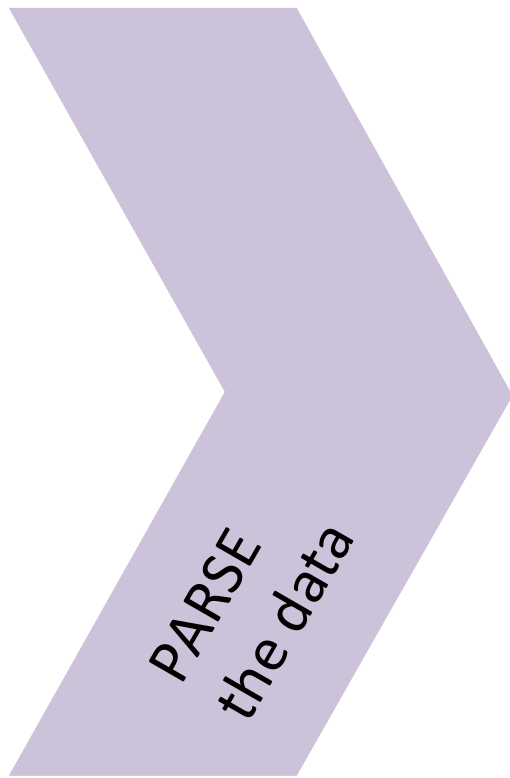
Acquire the Data (cont.)

- Acquire the Data

- Identify the “right” dataset(s)
 - Import data and set up local or remote data structure
 - Determine most appropriate tools to work with data

- This step can be very simple (e.g., simply reading a readily available text file) or extremely complicated (i.e., trying to glean useful data out of a large system)

Parse the Data



- Parse the Data
 - Read any documentation provided with the data
 - Perform exploratory data analysis
 - Verify the quality of the data

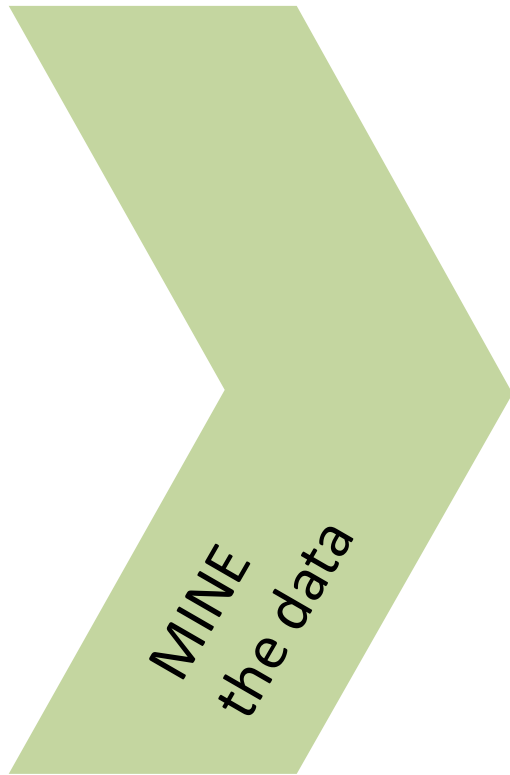
Parse the Data (cont.)

- Parse the Data

- Read any documentation provided with the data
 - Perform exploratory data analysis
 - Verify the quality of the data

- The incubator software tracks 16 different data streams, such as heart rate, respiration rate, temperature, blood pressure, and blood oxygen level; every data point being a number
- More generally, data needs to be formatted, i.e., broken along its individual parts and each piece of data is converted to its useful format (e.g., string, number, date)

Mine the Data



- Mine the Data
 - Determine sampling methodology and sample data
 - Format, clean, slice, and combine data in Python
 - Create necessary derived columns from the data (new data)

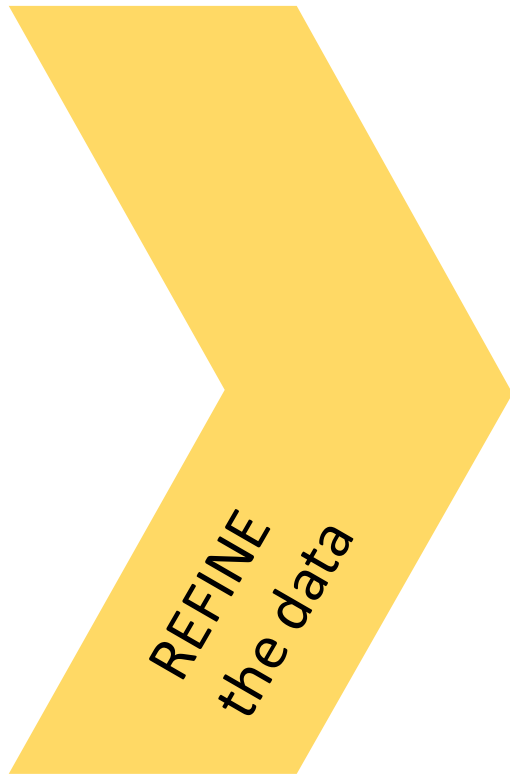
Mine the Data (cont.)

- Mine the Data

- Determine sampling methodology and sample data
 - Format, clean, slice, and combine data in Python
 - Create necessary derived columns from the data (new data)

- The incubator generates a vast stream of data (around 1,260 data points per second) but we will aggregate the data at the minute level

Refine the Data



- Refine the Data
 - Identify trends and outliers
 - Apply descriptive and inferential statistics
 - Document and transform data

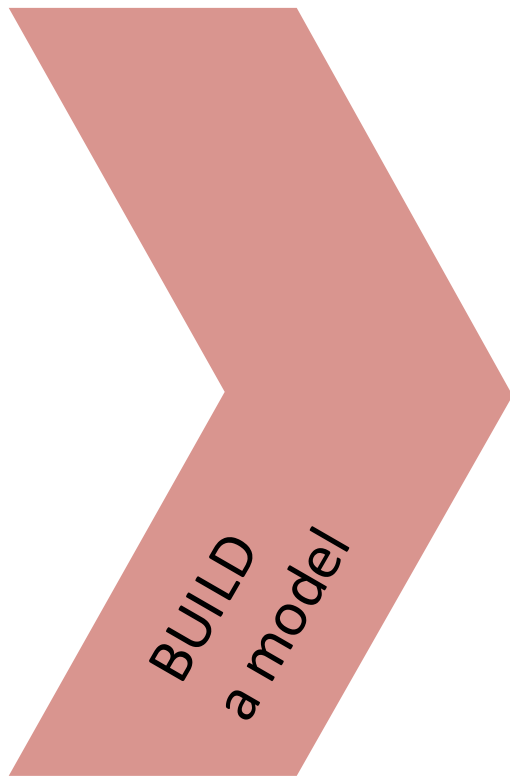
Refine the Data (cont.)

- Refine the Data

- Identify trends and outliers
 - Apply descriptive and inferential statistics
 - Document and transform data

- Plot the data to identify outliers
- We will start by calculating the mean and standard deviation of the different measurements to identify trends

Build a Model



- Build a Model
 - Select appropriate model
 - Build model
 - Evaluate and refine model

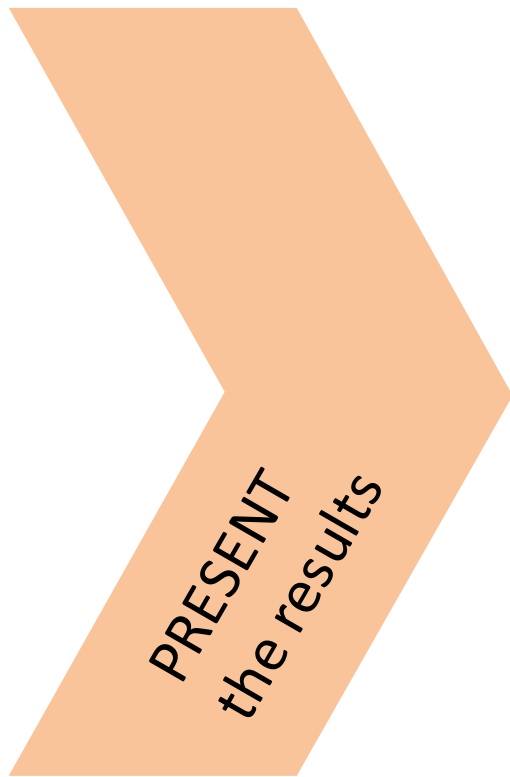
Build a Model (cont.)

- Build a Model

- Select appropriate model
 - Build model
 - Evaluate and refine model

- We could for example, first build a model based on decision trees
- Then build another model using logistic regression
- Refine them, compare them, and select the best one based on its predictability power

Present the Results



- Present the Results
 - Summarize findings with narrative, storytelling techniques
 - Present limitations and assumptions of your analysis
 - Identify follow up problems and questions for future analysis

Present the Results (cont.)

- Present the Results

- Summarize findings with narrative, storytelling techniques
 - Present limitations and assumptions of your analysis
 - Identify follow up problems and questions for future analysis

- The model can detect subtle changes in the preemies' condition that may signal the onset of infection 24 hours before overt symptoms appear

A Note About Iteration

- Iteration is an important part of *every* step in the Data Science Workflow. At any given point in the process, you may find yourself repeating or going back and re-doing elements in order to better understand your data, clarify your model, and refine your presentation
- For example, after presenting your findings, you may want to:
 - Identify follow-up problems and questions for future analysis
 - Create a visually effective summary or report
 - Consider the needs of different stakeholders and how your report might be changed for them
 - Identify the limitations of your analysis
 - Identify relationships between visualizations

Multiple variants exist but they are pretty much all doing the same thing

- Jeff Hammerbacher

- Identify problem
- Instrument data sources
- Collect data
- Prepare data (integrate, transform, clean, impute, filter, aggregate)
- Build model
- Evaluate model

- Ben Fry

- Acquire
- Parse
- Filter
- Mine
- Represent
- Refine
- Interact

- Peter Huber

- Inspection
- Error checking
- Modification
- Comparison
- Modeling and model fitting
- Simulation
- What-if analyses
- Interpretation
- Presentation of conclusions

- Dataists

- Obtain
- Scrub
- Explore
- Model
- Interpret

- Colin Mallows

- Identify data to collect and its relevance to your problem
- Statistical specification of the problem
- Method selection
- Analysis of method
- Interpret results for non-statisticians

- Jim Gray

- Capture
- Curate
- Communicate

- Ted Johnson

- Assemble an accurate and relevant dataset
- Choose the appropriate algorithm

Data Science Workflow Activity



EXERCISE

DIRECTIONS (25 minutes)

1. Divide into 4 groups, each located at a whiteboard
2. Identify: Each group should develop 1 research question they would like to know about their classmates. Create a hypothesis to your question. Don't share your question yet! (5 minutes)
3. Acquire: Rotate from group to group to collect data for your hypothesis. Have other students write or tally their answers on the whiteboard (10 minutes)
4. Present: Communicate the results of your analysis to the class (10 minutes)
 - a. Create a narrative to summarize your findings
 - b. Provide a basic visualization for easy comprehension
 - c. Choose one student to present for the group

DELIVERABLE

Presentation of the results

Today's Closing Thought



FROM THE DEAN

Applying Uber's Business Analytics Lesson

The source of success at Uber and Airbnb is their use of business analytics. What lessons can all business leaders learn from their examples? BY GEOFFREY GARRETT

I had a conversation recently with an early Uber investor who made something very clear to me: Uber is a data and analytics company. The same is true for the other tech valuation megafunds, Airbnb.

Investors are wowed by the fact that Uber's fixed-cost base is so low. It doesn't own its fleet of vehicles, and most people who drive its cars aren't permanent employees.

Critics worry about how well insured Uber's drivers and their vehicles are, and whether it is a good thing Uber seems so hard to regulate. But talking with people who really know Uber, the key to its success is how much the company knows about its market-in-particular, where the customers and its drivers are and what they want.

The genius of the market, economists tell us, is that its "invisible hand" magically connects buyers and sellers and also figures out how much goods or services are worth. In Uber's case, the magic is analytics.

Uber's customers can find out before they book a ride how many vehicles are in their area and how long before one gets to them. On the other side of the transaction, Uber drivers know whether it's worth their while to get into their cars because they know how many customers there are, where they are and how much they are willing to pay. Substitute houses and apartments for cars and Uber, and you have the Airbnb business model.

It is a cliché of the big data era that information is everywhere. But

as Uber and Airbnb show, the key to business success is knowing what to do with it.

That is why Wharton Online has today launched its new Business Analytics Specialization on Coursera. The first course is in Customer Analytics, to be followed by People Analytics, Accounting Analytics and Operations Analytics.

The specialization will finish up with a capstone project where you can apply what you've learned to a real business problem at Yahoo.

You don't have to be a math whiz or a genius computer scientist to take and benefit from our specialization. You will learn simple but powerful tools that will help you make better business decisions by seeing the strategic horizon through the blizzard of information that is all around you.

Who knows? Maybe your idea will turn into the next tech "unicorn," powered by your business analytics savvy. I hope you'll enroll in the specialization and see for yourself.

Geoffrey Garrett is dean and Endurance Professor of Management and Practice Professor at the Wharton School of the University of Pennsylvania.

Editor's note: Read more about the Business Analytics Specialization in our article on p. 16.

((ON THE WEB))
Cover photo: Shutterstock/Wharton
Dean Geoffrey Garrett's thought process by following him on LinkedIn and on Twitter at @geoffg.

WINTER 2016 | WHARTON MAGAZINE | 5

- It is a cliché of the big data era that information is everywhere. But as Uber and Airbnb show, the key to business success is knowing what to do with it – Geoffrey Garrett, Dean of The Wharton School



DS

Review

Review

You should now be able to answer the following questions:

- What is Data Science?
- What is the Data Science workflow?
- How can you have a successful learning experience at GA?

DS

Q & A



DS

Setting Up the Development Environment

DS Pre-Course Checklist

- GitHub Account
 - Create a GitHub account, if you don't already have one (<https://github.com/join>)
- Python and Anaconda Software
 - Install Python 2.7.x (<https://www.python.org/downloads>)
 - We will be using Python 2.7 in this course, NOT Python 3, which is significant different from 2.7
 - Install Anaconda for Python 2.7 (<https://www.continuum.io/downloads>)
 - Test for successful installation. For example, on a Mac you can by open a Terminal window (using the Anaconda Launcher app, if needed) and type “ipython notebook”. In a few moments, your browser should open to a window titled “Jupyter”
 - Install pip, the recommended package manager for Python (<http://pip.readthedocs.org/en/stable/installing>)
 - Note: Most versions of Python come with pip pre-installed. Check by opening your Terminal and running `pip install -U pip` to see if you have the latest version



DS

Before Next Class

Before Next Class

- Start looking the two newly assigned projects
 - Unit Project 1 (due 1 week from now on 3/3)
 - Final Project 1 (due 3.5 weeks from now on 3/22)
- Pre-Work for the next session
 - Have completed Python pre-work
 - Open and create iPython notebooks



DS

Exit Ticket

Don't forget to fill out your exit ticket [here](#)

Sources

- Slide #2 – Harvard Business Review
- 4 – The New York Times
- 6 – FiveThirtyEight
- 8 – Ivan Corneillet; Wired
- 10 – DOMO
- 13 – Twitter
- 15, 16, and 23 – General Assembly
- 26 – NVIDIA
- 38 – Data Science for the C-suite
- 43 – NHANES 2007-2010
- 50 – Big Data: A Revolution That Will Transform How We Live, Work, and Think; Amico
- 67 – Wharton Magazine