Introduction to Data Science
SKKU University, Summer 2015

Prof. Jevin D. West
University of Washington
Lecture 2 – June 30, 2015

Example of data science…

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

323 comments: 169 called-out + Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents to be at that crucial moment before they turn into

Target has got you in its aim

"[Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole's colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date."

"My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

Example of data science…

nature

Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

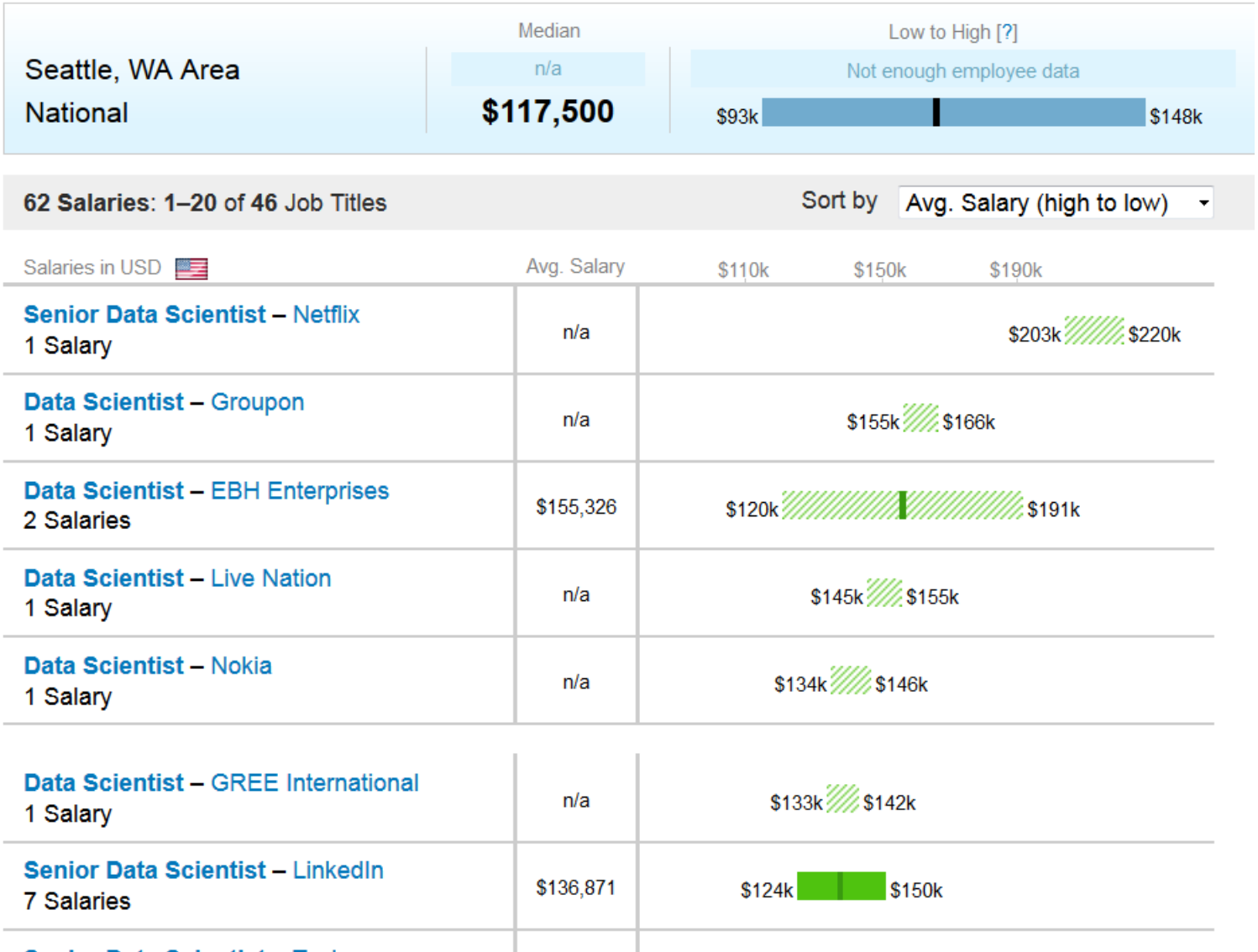
Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \varepsilon$, where

Why Data Science?



	Median	Low to High [?]
Seattle, WA Area	n/a	Not enough employee data
National	**$117,500**	$93k – $148k

62 Salaries: **1–20** of **46** Job Titles — Sort by: Avg. Salary (high to low)

Salaries in USD	Avg. Salary	$110k $150k $190k
Senior Data Scientist – Netflix 1 Salary	n/a	$203k – $220k
Data Scientist – Groupon 1 Salary	n/a	$155k – $166k
Data Scientist – EBH Enterprises 2 Salaries	$155,326	$120k – $191k
Data Scientist – Live Nation 1 Salary	n/a	$145k – $155k
Data Scientist – Nokia 1 Salary	n/a	$134k – $146k
Data Scientist – GREE International 1 Salary	n/a	$133k – $142k
Senior Data Scientist – LinkedIn 7 Salaries	$136,871	$124k – $150k

Introduction to Data Science
SKKU University, Summer 2015

Prof. Jevin D. West
University of Washington
Lecture 2 – June 30, 2015

Agenda

- 9:30 – 10:00 Introductions
- 10:00 – 10:15 Syllabus and Schedule
- 10:15 – 10:30 Introduction to R
- 10:30 – 10:45 Break
- 10:45 – 11:15 Working with R
- 11:15 – 12:00 Version Control Systems

Learning Objectives

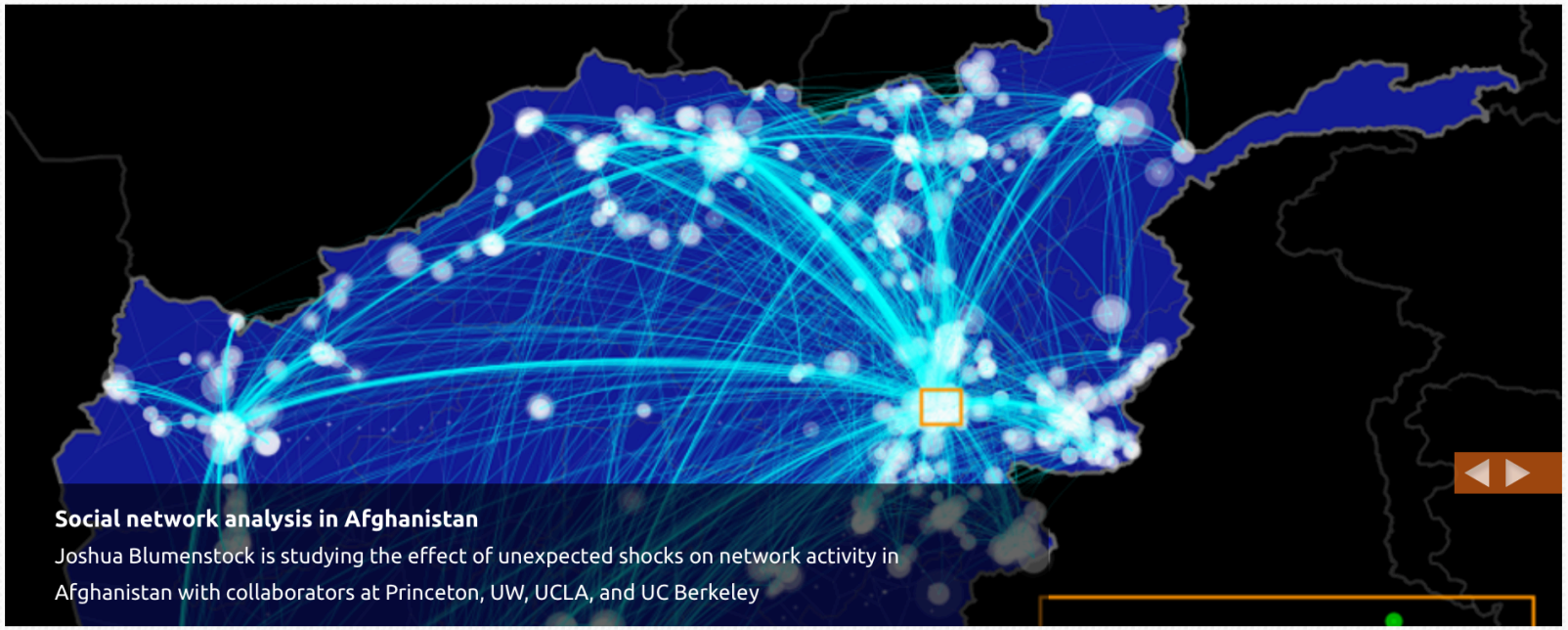
1. Introduction to R
   - Install R Studio
   - R scripts
   - Basic operations
   - Data import
   - Basic statistics
   - Basic plotting
2. Introduction to Version Control Systems
   - Git and mercurial
   - Github and Bitbucket
   - Setting up a repository
   - Basic commands (git)

Introductions

DataLab

Data Science and Analytics Lab

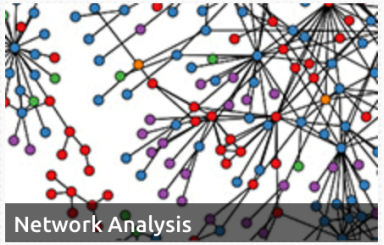
University of Washington Information School



Social network analysis in Afghanistan

Joshua Blumenstock is studying the effect of unexpected shocks on network activity in Afghanistan with collaborators at Princeton, UW, UCLA, and UC Berkeley

Research Focus Areas



News and Updates

28 **Blumenstock at Population Association of America**

What we do

The DataLab is the nexus for research on Data Science and Analytics at the UW iSchool. We study **large-scale, heterogeneous human data** in an

GORDON AND BETTY MOORE FOUNDATION

DATA-DRIVEN DISCOVERY

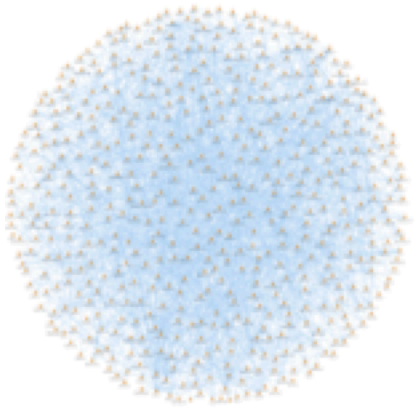
Data Science Environments

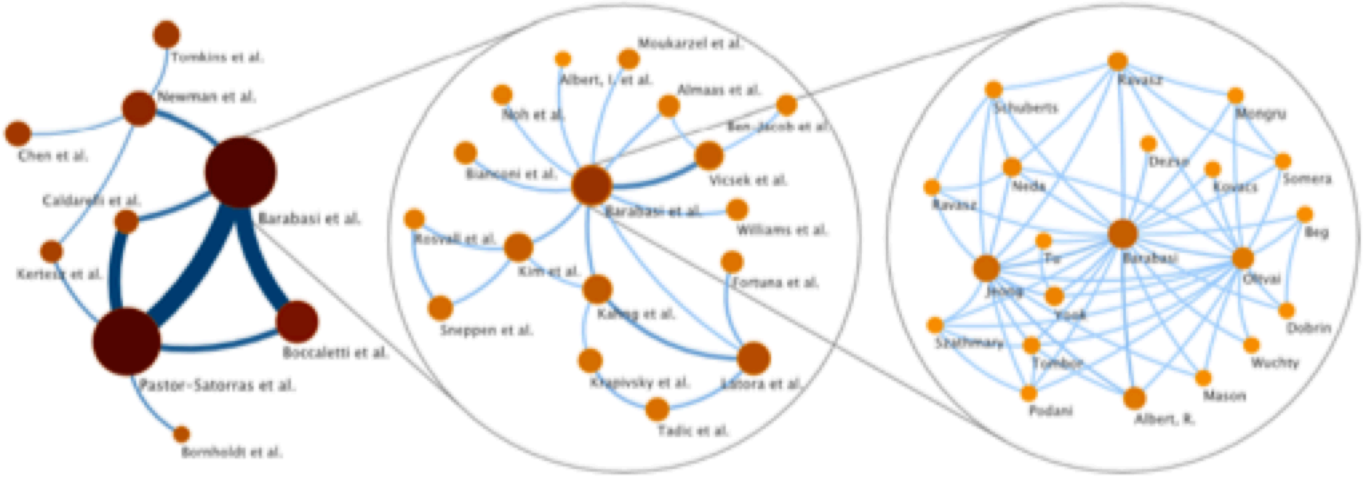
Jevin West

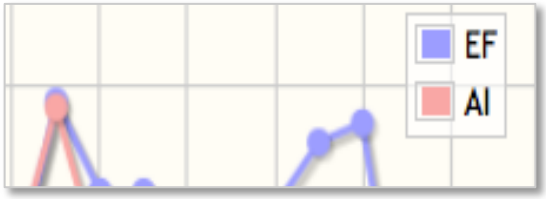
Assistant Professor | iSchool

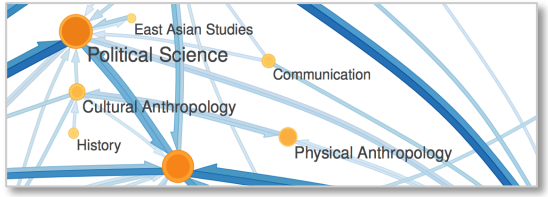
eigenFACTOR.org®

RANKING AND MAPPING SCIENTIFIC KNOWLEDGE









Ranking

Mapping

Navigating

DataLab

Data Science and Analytics Lab

Information School

UNIVERSITY of WASHINGTON

Introductions

Why are you interested in Data Science?

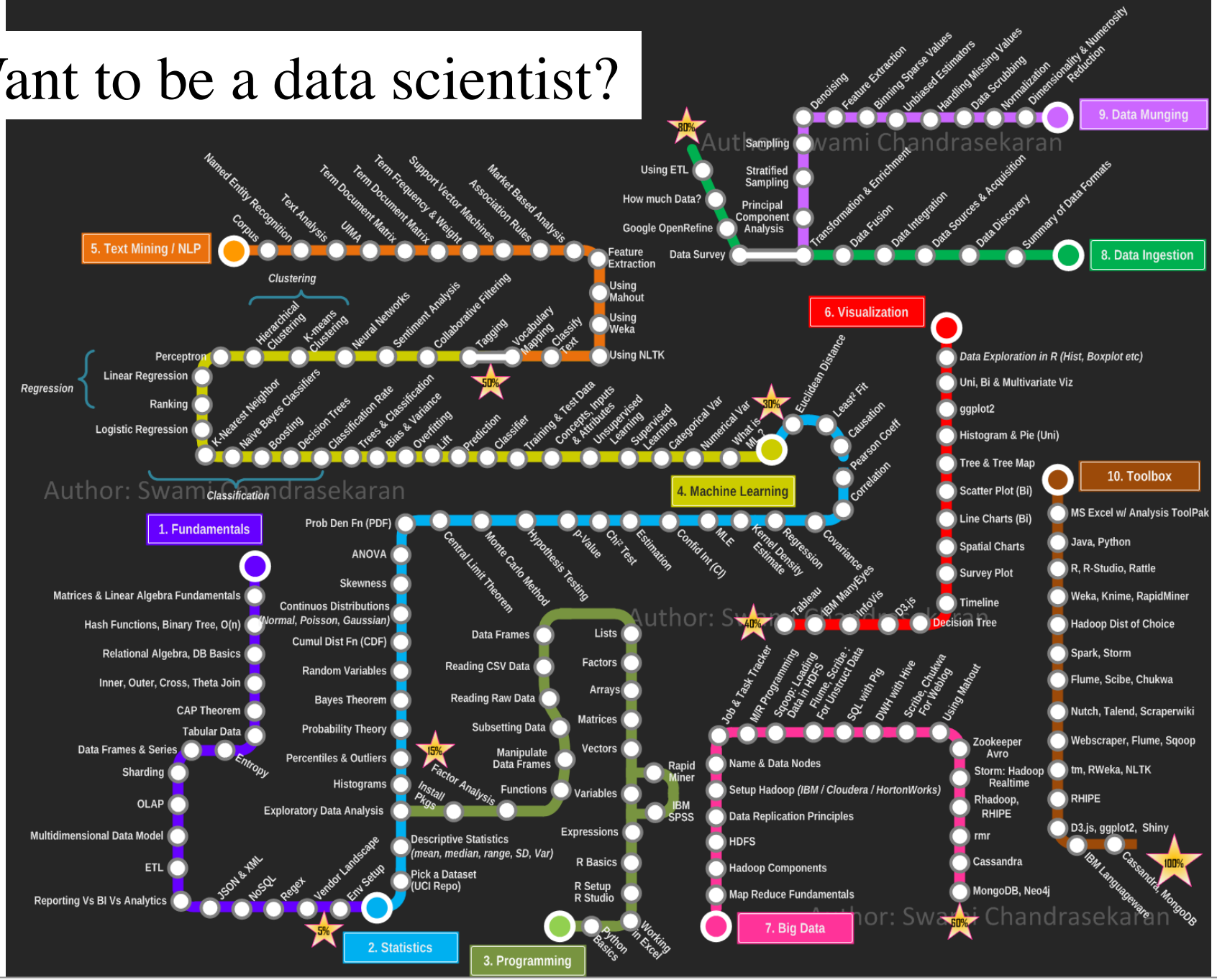
What is Data Science?

Experience?

Agenda

- 9:00 – 9:30 Introductions
- 9:30 – 10:00 Syllabus and Schedule
- 10:00 – 10:30 Introduction to R
- 10:30 – 10:45 Break
- 10:45 – 11:15 Working with R
- 11:15 – 12:00 Version Control (github)

Want to be a data scientist?



Topics

- Opportunities in data science
- The language of data science (R and Python)
- Version Control Systems (VCS)
- Cloud computing and distributed computing
- Data management and data ingestion
- Experimental Design and Empirical Frameworks
- Basic probability and statistical inference
- Basic machine learning
- Network analysis
- Information Visualization
- Data Privacy and Ethics

Schedule

** The schedule will update regularly so make sure to check it regularly*

- Week 1
  - Introduction to R (and python)
  - Version Control Systems
  - Opportunities in Data Science
  - Data Ingestion
  - **Assignments:** Quiz #1, Assign. #1, Group Project (Data Set Identified)
- Week 2
  - Cloud Computing
  - Experimental Design
  - Basics in probability and statistics
  - **Assignments:** Quiz #2, Assign. #2, Group Project (Question, Preliminary Stats)
- Week3
  - Basics in machine learning
  - Network analysis
  - Information visualization
  - Data Ethics
  - **Assignments:** Quiz #3, Assign. #3, Group Project (Final Paper, Final Presentation)

Grades

- Participation (20%)
  - Discussion
  - In-class exercises
- Assignments and Quizzes (40%)
- Group Project (40%)

Expectations

Questions

Agenda

- 9:00 – 9:30 Introductions
- 9:30 – 10:00 Syllabus and Schedule
- 10:00 – 10:30 Introduction to R
- 10:30 – 10:45 Break
- 10:45 – 11:15 Working with R
- 11:15 – 12:00 Version Control Systems

Introduction to 

* See **Introduction_to_R_Raza.pdf** for more details.

Version Control Systems

GitHub

Bitbucket

GIT commands

- sudo apt-get install git
- git init
- git remote add origin [https or ssh]
- git add file.txt
- git commit -m "first commit”
- git push origin master

Homework

- Quiz on readings
  - Kramer, Adam DI, Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences* 111.24 (2014): 8788-8790.
  - Executive summary of: McKinsey Global Institute (May 2011). *Big data: The next frontier for innovation, competition, and productivity*, Executive Summary, pgs 1-15
- Organize a group for the class project
- Install R Studio
- Sign up for Github
- [Optional] Read Chapter 1 of Torgo, *Data Mining with R*

Data Science

www.youtube.com/watch?v=jbkSRLYSojo

Prof. Jevin West

jevinw@uw.edu