Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges

Sean Kross and Philip Guo
May 9
CHI 2019
Glasgow, UK

UC San Diego
The Design Lab


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
Entropy
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup R Studio
Rapid Miner
IBM SPSS
Python Basics
Working in Excel
4. Machine Learning
Regression
Perceptrons
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
50%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Using NLTK
Feature Extraction
Using Mahout
Using Weka
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
80%
Using ETL
How much Data?
Google OpenRefine
Data Survey
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
9. Data Munging
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
8. Data Ingestion
6. Visualization
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
30%
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
7. Big Data
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%

UNIVERSITY of WASHINGTON
eScience Institute
DATA SCIENCE FOR SOCIAL GOOD

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

MIDAS
MICHIGAN INSTITUTE
FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN

Department of
Biomedical
Data Science
AT STANFORD MEDICINE

Carnegie Mellon University
Statistics & Data Science

What are data science educators actually doing?

- **What challenges** do data science instructors face?
- **How do they cope** with these challenges?
- **What are the differences** between data science ed and computer science ed?

ID	Gender	Age	Degree	Field	Sector	Workplace	Teaching setting(s)	Students
P1	F	25–34	PhD	Biostatistics	Academia	R1 university	workshops, online	1000+
P2	M	25–34	PhD	Biostatistics	Academia	R1 university	workshops, online	1000+
P3	F	25–34	MS	Genomics	Industry	R&D nonprofit	workshops, online	1000+
P4	F	25–34	PhD†	Education	Industry	Startup company	online	350
P5	F	25–34	PhD	Genetics	Academia	R1 university	ugrad/grad courses	20
P6	F	25–34	MPH	Medical stats	Academia	Medical school	workshops	20
P7	F	35–44	PhD	Marine biology	Academia	Research institute	workshops	15
P8	M	25–34	PhD	Statistics	Academia	R1 university	grad course, workshops	20
P9	F	35–44	PhD	Neuro/genomics	Academia	R1 university	grad course, workshops	20
P10	M	25–34	PhD†	Biostatistics	Academia	R1 university	grad course	20
P11	F	35–44	PhD	Psychology	Academia	Medical school	grad course, online	1000+
P12	F	45–54	MS	Psychology	Industry	Coding bootcamp	bootcamp	30
P13	F	35–44	BS	Sci/tech studies	Industry	Mid-sized company	workshops	20
P14	F	25–34	PhD	Statistics	Academia	Liberal arts college	ugrad course, workshops	30
P15	F	35–44	PhD	Statistics	Academia	R1 university	ugrad course, workshops	30
P16	M	25–34	PhD	Neuroscience	Industry	Pro sports franchise	online video livestreams	20
P17	M	25–34	BS	Math/business	Industry	Startup company	online	1000+
P18	F	25–34	MS	Library sci.	Academia	R1 university	ugrad/grad courses	15
P19	F	25–34	BS	English/stats	Industry	Mid-sized company	workshops	20
P20	F	45–54	MS	Management	Industry	Open-source nonprofit	workshops	25

Table 1: The 20 data science practitioner-instructors we interviewed: F=female, M=male. For PhD†: P4 left a PhD program, and P10 is currently a PhD student. R1 means major research university. 'Students' is approximate number of students per class.

Neuroscience PhD
Pro Sports
Instructional Live Streams on Twitch

Psychology PhD
Medical School Professor

Business/Math BS
MOOC Instructor

Varied Prior Computing Experiences

“From a teaching perspective, I feel blessed that I didn’t study computer science. I’m self-taught, and I feel that makes it easier for me to empathize with my students and anticipate their problems.” - P17

Varied Motivations for Learning Data Science

“Most people I see have to learn to code in an absolute panic for their thesis.” - P7

Teaching Data-Analytic Programming

“Maybe ten percent of the people I teach are going to need to write their own R function.” - P14

Teaching Authentic Practices

“All of the courses are project based and all projects are done on GitHub. It helps them build a portfolio.” - P9

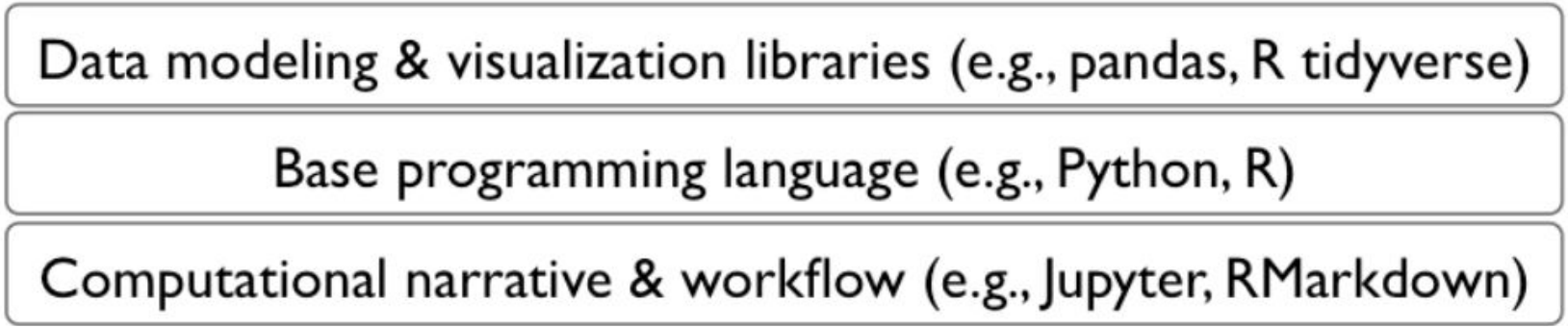
Delivering the Data Science Tech Stack

“How much do we really want to teach about system administration and .bashrc?” - P17

Data modeling & visualization libraries (e.g., pandas, R tidyverse)

Data modeling & visualization libraries (e.g., pandas, R tidyverse)

Base programming language (e.g., Python, R)

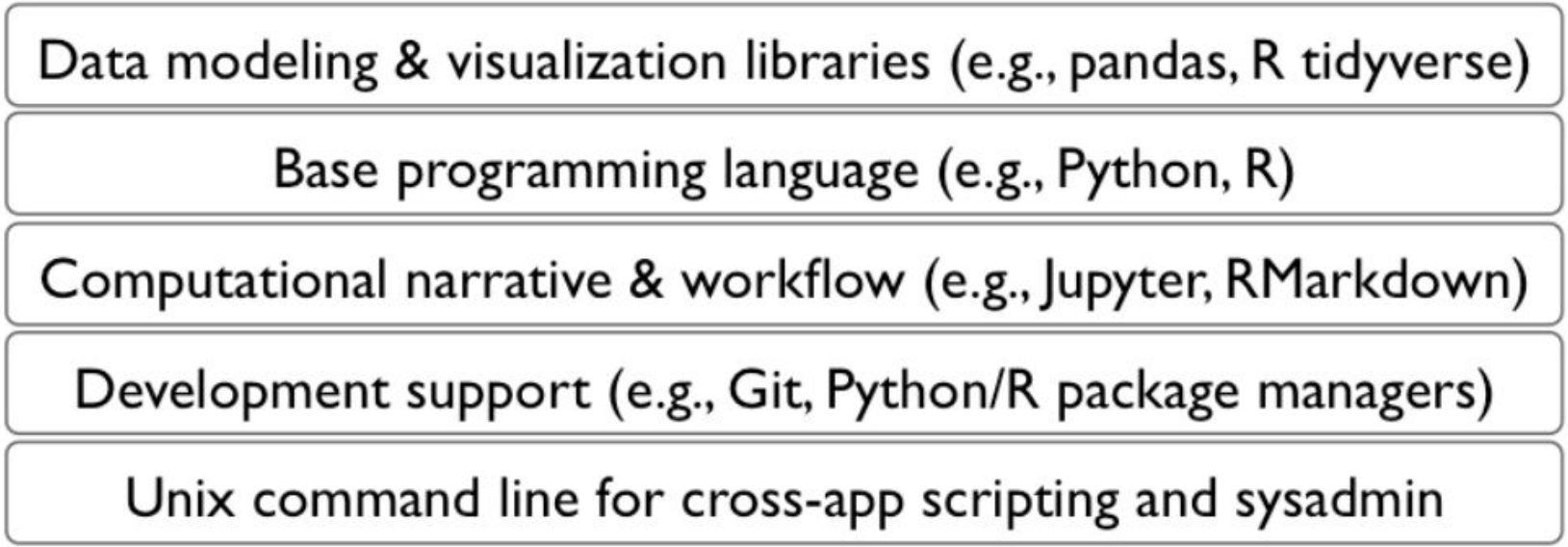

Data modeling & visualization libraries (e.g., pandas, R tidyverse)
Base programming language (e.g., Python, R)
Computational narrative & workflow (e.g., Jupyter, RMarkdown)

Data modeling & visualization libraries (e.g., pandas, R tidyverse)

Base programming language (e.g., Python, R)

Computational narrative & workflow (e.g., Jupyter, RMarkdown)

Development support (e.g., Git, Python/R package managers)


Data modeling & visualization libraries (e.g., pandas, R tidyverse)
Base programming language (e.g., Python, R)
Computational narrative & workflow (e.g., Jupyter, RMarkdown)
Development support (e.g., Git, Python/R package managers)
Unix command line for cross-app scripting and sysadmin

Data modeling & visualization libraries (e.g., pandas, R tidyverse)

Base programming language (e.g., Python, R)

Computational narrative & workflow (e.g., Jupyter, RMarkdown)

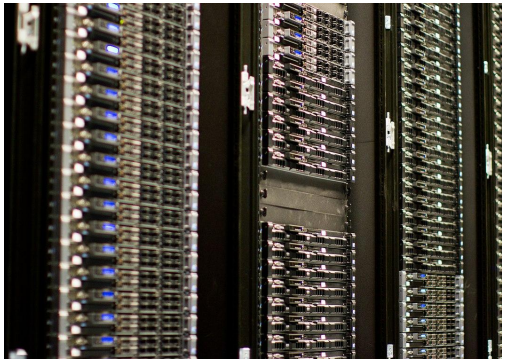
Development support (e.g., Git, Python/R package managers)

Unix command line for cross-app scripting and sysadmin

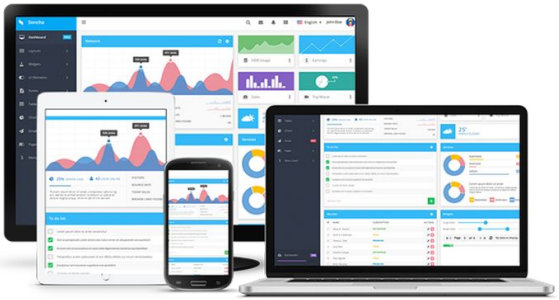
Reproducibility infrastructure (e.g., Docker, virtual machines)

Software Setup Solutions:

Desktop



Server



Web Application

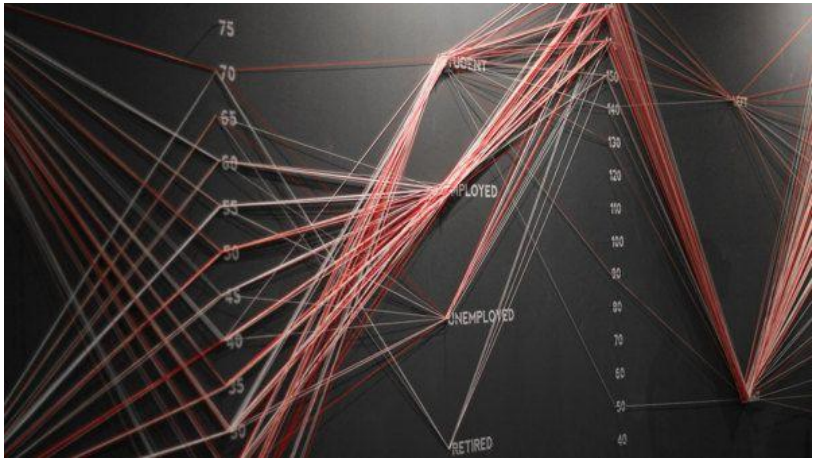
Finding and Curating Datasets

“It’s hard to find a dataset that exactly fits your problem. I’ve spent weeks looking for a dataset to teach with.” - P1

Coping with Uncertainty

“Everything is always on fire! How do we teach people to live with this reality?” - P4

How Can We Make Things Better?



Obtaining High-Quality Datasets for Teaching

New Data Science Learning Environments

How do we create a student-and-instructor-friendly data science learning environment?

Paper: seankross.com/chi-2019

Talk slides: seankross.com/chi-2019-talk

Let's Talk!

Twitter: @seankross

Email: seankross@ucsd.edu

UC San Diego
The Design Lab