Resources, Capabilities, and Strategies for Data Science Learners

2018 CSUN DataJam
CSU Northridge
Friday, September 28, 2018

Wayne Smith, Ph.D.
Department of Management
CSU Northridge
ws@csun.edu

This presentation is available at (.odp, .pptx, .pdf):
https://github.com/wsphd/datajam2018

Overview

- *Thinking* as a Data Scientist
  - Positioning Yourself in the Contemporary Paradigm
- *Learning* as a Data Scientist
  - What the Best Autodidacts (self-learners) Know
- *Acting* as a Data Scientist
  - Earning Career Success

- Q&A
- Give away some goodies…

Motivating Question

Are you a *data science* learner?

…*quantitative reasoning*…

…*big data*…

…*statistical computing*…

…*applied math*…

…*business intelligence*…

…*predictive analytics*…

…*decision support modeling*…

…*artificial intelligence*…

What is Data Science?

Curriculum Guidelines for Undergraduate Programs in Data Science (September, 2016)

- Data Science is "the science of planning for, acquisition, management, analysis of, and inference from data."

- Students would demonstrate mastery of skills and concepts, including many traditionally associated with the fields of Statistics, Computer Science and Mathematics. *Data Science blends much of the pedagogical content from all three disciplines, but it is* ***neither the simple intersection, nor the superset of the three****.*

- There is a fourth area of demonstrated mastery too: subject-matter expertise.
- Building upon experimentation, modeling, and computation, there are some that believe that data science *is, in fact, a new, type of scientific discovery*.

- Case-based, hands-on, and interdisciplinary
- Additionally, some existing courses in statistics, math, and computer science, should be partly re-designed for use in a data science curriculum.

Information Dynamics

- *Wisdom*
  - Extraordinary Insight (Explanation) for Foresight (Prediction)
  - ***Restaurant*: How should our menu change in the future to best optimize nightly sales?**
- *Knowledge*
  - Combination of Explicit Information and Tacit Information
  - ***Restaurant*: What action led to the change in last night's sales?**
- *Information*
  - Meaningful Data
  - ***Restaurant*: How does last night's sales compare to that night the previous year? How does last night's sales compare to our goals?**
- *Data*
  - Raw, atomic, basic
  - ***Restaurant*: What were the total sales for last night?**

Analytics for Decision-making

- *Prescriptive* Analytics
  - What should we do?
  - ***HR Department*: What should we (the HR Department) do to meet or exceed the organization's hiring and retention goals for next year? What data/information/knowledge/wisdom should we provide to our hiring and technical managers to help? What are we missing?**
- *Predictive* Analytics
  - What is likely to happen?
  - ***HR Department:* How many new employees will our organization need next year? How will the mix change? What is our competition likely to do?**
- *Diagnostic* Analytics
  - Why did it happen?
  - ***HR Department:* Did our emphasis on recruiting from campus A (over campus B, etc.) matter? What do the managers of these entry-level employees think?**
- *Descriptive* Analytics
  - What happened?
  - ***HR Department:* How many entry-level professionals did we hire last year? How many of them are still with us now?**

Math/Stat/CompSci/Subject Domain w/ Theory and Practice

- What can we leverage well for data science success?

- Math
- Stat
- CompSci
  - computational thinking

Persistent Tensions

Analytics	↔	Automation
Explanation	↔	Prediction
Desktop	↔	Cloud
Base R	↔	Tidyverse
SQL	↔	Spark
Equation	↔	Algorithm
Hypothesis-driven	↔	Discovery-driven
Bias	↔	Variance
Precision/Recall	↔	Specificity/Accuracy
p-value	↔	fit
Frequentist	↔	Bayesian

Balance Breadth and Depth

- R
- Python
- Julia
- *t*-test for two independent groups, OLS linear regression, k-means clustering (or classification)
- LP/convex optimization, GIS/spatial analysis, network analysis, text analysis

- computational thinking

Commercial Packages too

Next-Generation Science Standards

There are *seven* “cross-cutting” skills to support *four* core subject domains: Physical Science, Life Science, Earth and Space Science, and Engineering.

	Cross-cutting skill	*Data Science* analogue
1	Patterns	Observations, Variables, and Pattern Matching
2	Cause and Effect	Causality
3	Scale, Proportion, and Quantity	Measure Theory
4	Systems and System Models	Model Building
5	Energy and Matter	Systems Dynamics, Moments, and Entropy
6	Structure and Function	Hierarchical, Structural, and Latent Variable Analysis
7	Stability and Change	Variation, Complexity, and Interactions

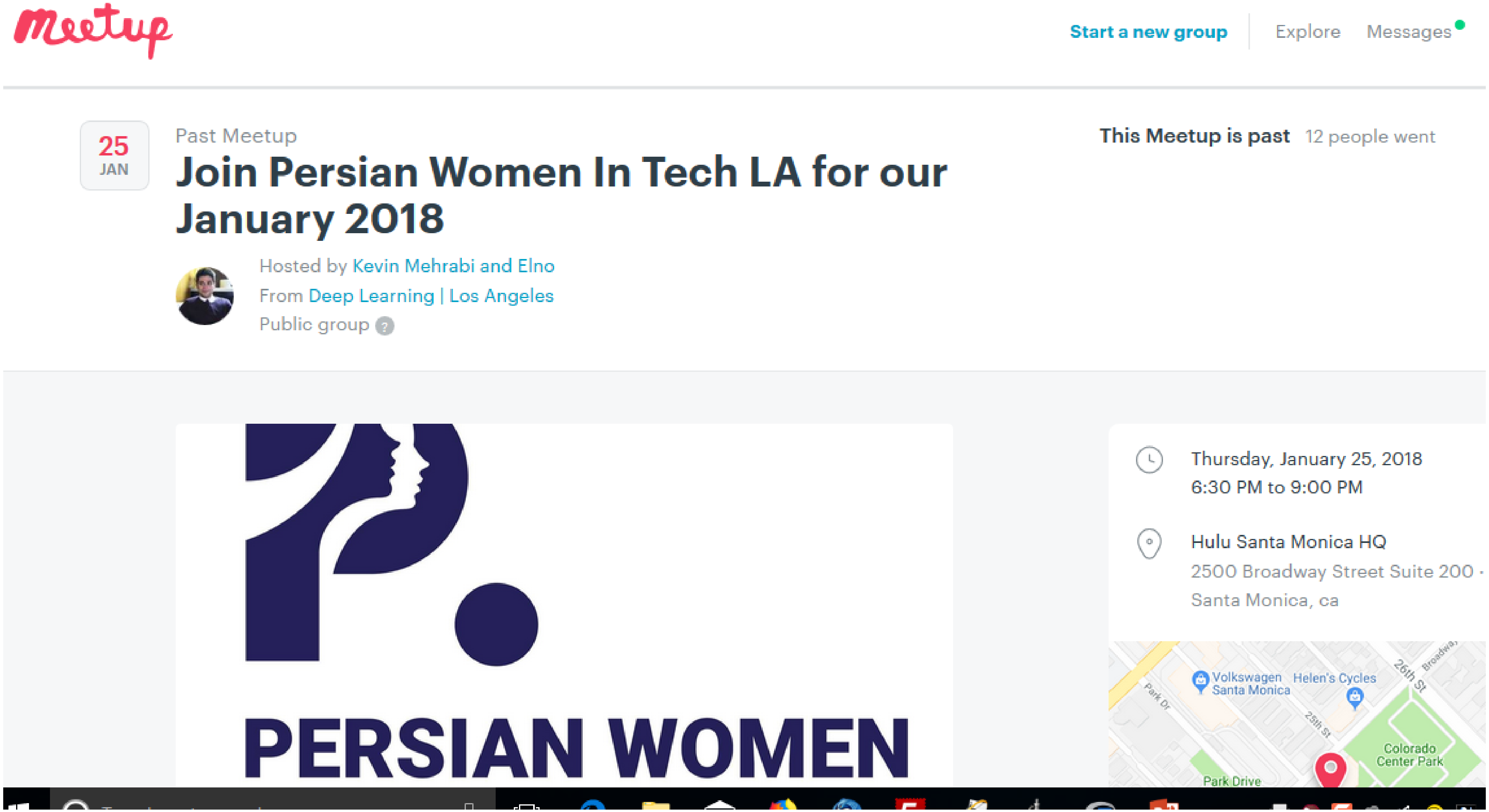
Reproducibility (“academic” perspective)

Workflow (“professional” perspective)

Twitter, Podcasts, Textbooks

- Twitter
  - #rstats
- Podcasts
  - https://www.analyticsvidhya.com/blog/2018/01/10-data-science-machine-learning-ai-podcasts-must-listen/

Meetups and Big Data Day LA, etc.



25
JAN

Past Meetup

Join Persian Women In Tech LA for our January 2018

Hosted by Kevin Mehrabi and Elno
From Deep Learning | Los Angeles
Public group ?

This Meetup is past 12 people went

Thursday, January 25, 2018
6:30 PM to 9:00 PM

Hulu Santa Monica HQ
2500 Broadway Street Suite 200 ·
Santa Monica, ca

CSUN Cloud Services? (e.g., IBM)

- Free-tier
  - IBM Watson Cloud (you want the “no time restrictions” option)
  - https://console.bluemix.net/registration/free/

  - Amazon Web Services (you want the “non-expiring” offer)
  - https://aws.amazon.com/free/

  - Google Cloud (you want the “always free” option)
  - https://cloud.google.com/free/

  - Microsoft Azure (you want the “start free” option)
  - https://azure.microsoft.com/en-us/free/

  - CSUN too (https://www.csun.edu/it/ibm-cloud-services-csun)

Canvas API Example

```
# do once
install.packages( “devtools” )
library( devtools )
install_github( “daranzolin/rcanvas” )
library( rcanvas )
set_canvas_token( “your-40-character-token-from-Account-Settings-here” )
set_canvas_domain( “https://canvas.csun.edu” )

# get course items
get_user_items( course_id = 12345, item = “assignments” )
get_user_items( course_id = 12345, item = “missing_submissions” )

# get course analytics
get_user_items( course_id = 12345, item = “activity” )

# upload a file (or files)
upload_course_file( course_id = 12345, file_name = “testfile.docx” )
```

Recent blog on LAPD arrests/crimes

Jobs/indeed.com, etc.

Conferences w/ videos (and convert to iTunes via VLC?)

A New Language for conversations

- Student peers
- Professional contacts

- Most important—Professors
  - What kinds of *research questions* have you worked on?
  - What kinds of *data* have you used?
  - What kinds of *analytical methods* have you used?
  - What kinds of *software tools* have you used?
  - What are you hoping to learn to do in the near future?

Complexity

What do real world data science needs look like?

Data Science needs: Current Examples from the City of LA

- Elected Officials
  - Affordable Housing Risk Scoring and Covent Risk Scoring
  - Downtown Transportation Analysis
  - Street Pavement Prioritization and Early Warning System
  - Property Values and Affordable and Low Income Housing
  - LAPD Recruitment Performance Dashboard
  - CAP tracking enhancements, dashboards, and integration with other Personnel systems
  - Attrition prediction tool
  - Homelessness Services Matcher

Data Science needs: Current Examples from the City of LA

- Information Technology Agency
  - ServiceNow Analysis and Dashboard
- Office of Finance
  - Call center operational improvements
  - Bill Collections
  - Revenue Forecasting
- Department of Transportation
  - Projecting Parking Demand
- Department of Cultural Affairs
  - Cultural Events Analytics, Neighborhood Arts Profile, and Cultural Desert Discovery

City of LA: Specific Examples

- **Downtown Transportation Analysis**
  - Analysis of bicyclists and pedestrian use on Spring and Main both before and after Spring and Main Forward project. This will build on existing work from CSULA and LADOT. The Downtown configuration analysis for Project Downtown streets could show an ideal mix of various improvements and interventions. We hope to be able to project throughput for various streets downtown in different configurations.

City of LA: Specific Examples

- **LAPD Recruitment Performance Dashboard**
  - LAPD Personnel recruiters' main metric for success for recruiting candidates is the number of tests administered. However, there is limited visibility into which recruiters are testing the highest proportion of successful candidates, what strategies are most viable, or which geographic areas and events yield the best results. In preparation for anticipated surges in retirement, smarter recruiting is essential. A paradigm shift that is outcome-oriented will lead to greater accountability and flexibility as LAPD and Personnel strive to meet hiring goals. For this project, we wish to provide recruiters with new metrics of success.

Counterpoint

- Hernandez, D., and Greenwald T. (August 11, 2018), “IBM Has a Dilemma”, *Wall Street Journal*.
- Muller, J. (2018), *The Tyranny of Metrics*, Princeton University Press.
- O’Neill, C. (2017), *Weapons of Math Destruction: How Big Data Increases Inequity and Threatens Democracy*, Broadway Books.
- Pearl, J. (2018), *The Book of Why: The New Science of Cause and Effect*, Basic Books.
- Tenner, E. (2018), *The Efficiency Paradox: What Big Data Can’t Do*, Alfred A. Knopf.

Bibliography/References

- https://www.stat.berkeley.edu/~nolan/Papers/Data.Science.Guidelines.16.9.25.pdf
- Davenport, T. (2009), "Make Better Decisions", *Harvard Business Review*, Nov. 87(11), p. 117-123.

fin

- Again: Are you a data science learner?

- Questions?

- Goodies