

# 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](mailto: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...*

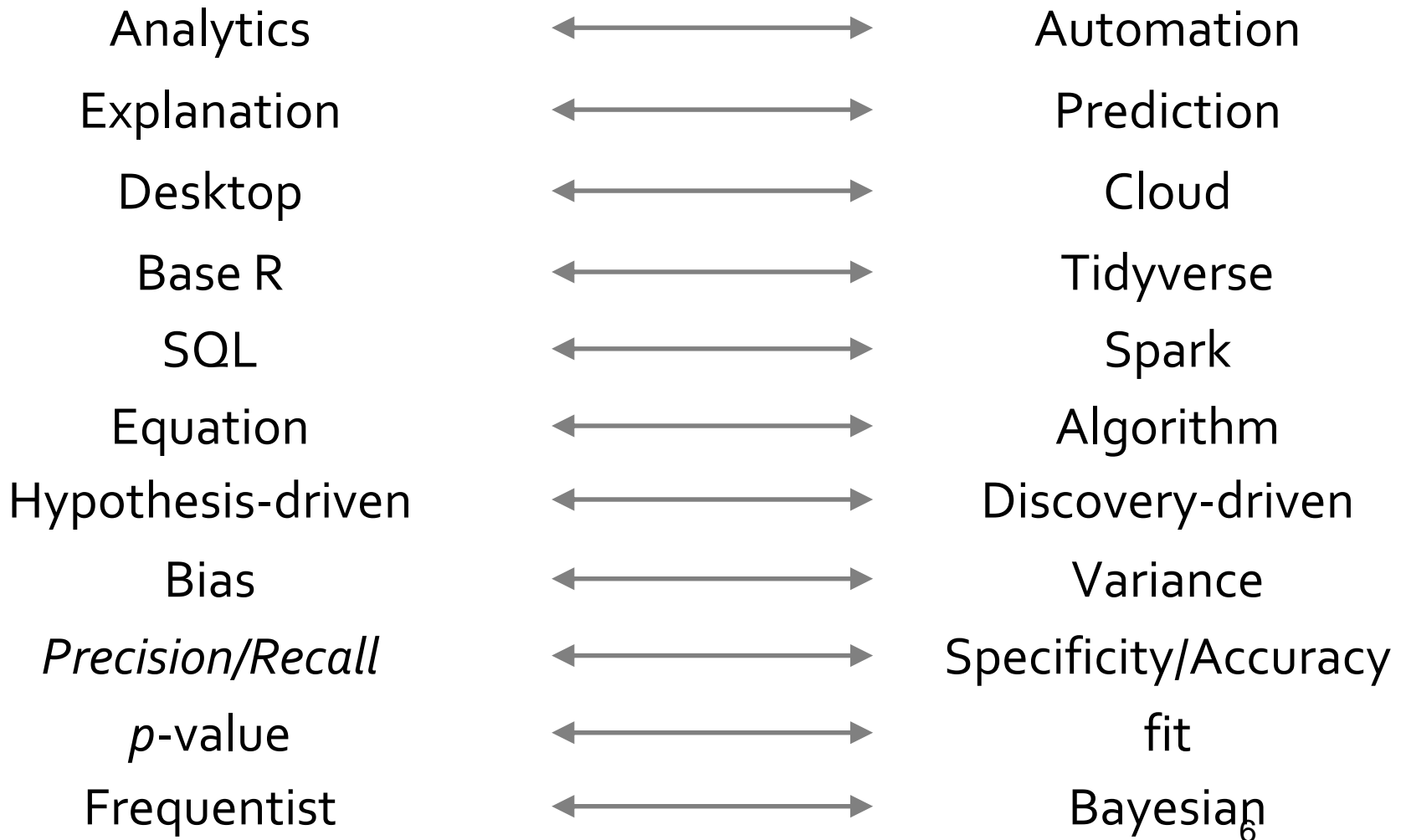
# Information Dynamics

- 
- *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?

# Persistent Tensions



# What is Data Science?

- Data Science is “the science of planning for, acquisition, management, analysis of, and inference from data.”
- Data science is case-based, hands-on, and interdisciplinary.
- Building upon experimentation, modeling, and computation, there are some that believe that data science *is, in fact, a new, type of scientific discovery*.
- Data scientists demonstrate mastery of skills and concepts, including many traditionally associated with the fields of Mathematics, Statistics, and Computer Science and. *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-Domain expertise. 7

# Math/Stat/CompSci/Subject Domain

- What can we leverage well for data science success?

<i>Evidence-Based...</i>	<i>...Language</i>	<i>...Approach</i>	<i>...Operational Scale</i>	<i>...Validity</i>
Practice	Equations, Theorems	Parametric tests, Regressions	Algorithms, Literate Programming	Various—Professional Standards
Theory	Probability, Matrices	General Linear Model	Computational Thinking	Various—Research-based
Discipline	Math	Statistics	Computer Science	Subject Matter (Domain)



## Practical Applications III: Deep (Machine) Learning

---

Mt. Sinai Hospital (NY) 2015 Research Program: “Deep Patient”

1. Tested on 700,000 patient records

*Able to predict disease far better than traditional methods*

2. Better than humans at predicting onset of schizophrenia

*Not even physicians can accurately predict that psychiatric disorder*

3. Algorithm was able to detect a pattern never before discovered

*Not only is pattern latent, so is its detection method (“black-box”)*

# Balance Breadth and Depth

- R
- Python
- Julia
- Visualization
- $t$ -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
- CranViews

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

- What can we leverage well for data science success?
- Math
- Stat
- CompSci
  - computational thinking

# Commercial Packages too

# Common Core (e.g., a graph in Grade 3)

ASA ID	award	Category	Title	LocalID	City
4-3550	1	GR10-12	Charting Chart Toppers	4-3550	Andover
MI4-194	2	GR10-12	Influenza 2018	MI4-194	
PA 4- 75	3	GR10-12	The Pen is Mightier than the Mouse	PA 4- 75	Glen Mills
CT10-12D13	HM	GR10-12	The Cost of Milk Production	CT10-12D13	Higganum
PA 2- 12	1	GR4-6	First Grade Shoelaces	PA 2- 12	Abington
2-3654	2	GR4-6	Screen Time	2-3654	Appleton
DC 2-3718	3	GR4-6	Sugary Soda Showdown	DC 2-3718	Columbia
PA 2- 61	HM	GR4-6	Duck...Duck...Rabbit?	PA 2- 61	Abington
OH 7-9 10	1	GR7-9	Pressure for Good Grades: Does It Exist?	OH 7-9 10	Lyndhurst
3-3712	2	GR7-9	Do High School Students Perform Acts of Kindness?	3-3712	West Nyack
DC 3-3729	3	GR7-9	Starbucks Consumption = mmm!	DC 3-3729	Williamsburg
MI3-121	HM	GR7-9	What's Up with Binge Watching?	MI3-121	
1-3635	1	GRK-3	Which Hand Rules	1-3635	Bar Harbor
OH K-3 9	2	GRK-3	Are You a Square or a Rectangle	OH K-3 9	Chagrin Falls
1-3552	3	GRK-3	Do You Know that Orangutans Are Endangered?	1-3552	Plano
DC 1-3939	HM	GRK-3	Speedcubing Statistics	DC 1-3939	Fairfax VA

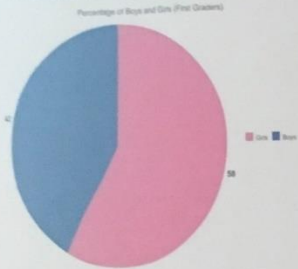
# "First-Grade Shoelaces" (Grades 4-6)

## Tie Your Shoelaces!!!

What is the likelihood of a first grader being able to tie their own shoelaces? What characteristics might affect the ability to tie shoelaces among first graders?

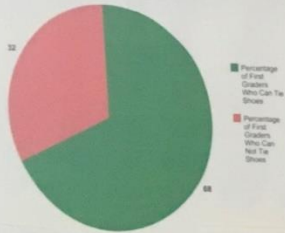
First a little information about the 1st graders: there were a total of 74 students surveyed and the percentage of boys and girls are as follows:

Percentage of Boys and Girls (First Graders)

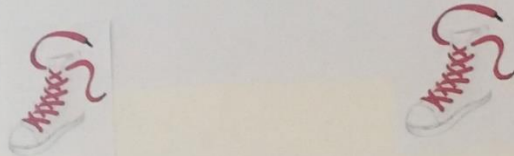


Percentage of First Graders Who Can Tie Their Shoes Vs. First Graders Who Cannot Tie Shoes

Percentage of First Graders Who Can Tie Shoes Vs. First Graders Who Cannot Tie Shoes

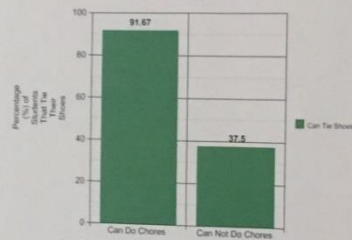


## FIRST GRADE SHOELACES.



Percentage of First Graders Who Can Tie Their Shoes Based on Whether or Not They do Chores

Percentage of First Graders Who Can Tie Their Shoes Based on Whether or Not They Do Chores



First Graders Distinguished Between Students Who Do Chores Vs. Students Who Do Not Do Chores

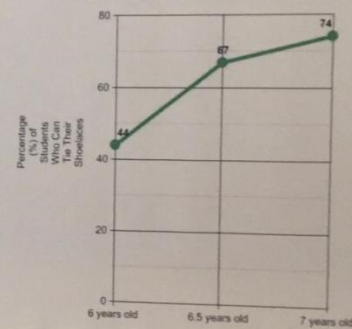
Percentage of First Graders Who Can Tie Their Shoes Based on Gender

Percentage of First Graders Who Can Tie Their Shoes Based on Gender



The Effect of Age on the Ability to Tie Shoelaces Among First Graders

The Effect of Age on the Ability to Tie Shoelaces Among First Graders



## Conclusion:

Overall, the likelihood of a first grader being able to tie their shoelaces is about 68%. If a first grader does their chores, they have a 92% chance of being able to tie their shoes! Girls are more likely to be able to tie their shoes than boys. Plus, as first graders grow older, they get more and more likely to be able to tie their shoes.





# "First-Grade Shoelaces" (Grades 4-6) (results)

## Conclusion:

Overall, the likelihood of a first grader being able to tie their shoelaces is about 68%. If a first grader does their chores, they have a 92% chance of being able to tie their shoes! Girls are more likely to be able to tie their shoes than boys. Plus, as first graders grow older, they get more and more likely to be able to tie their shoes.



# 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.

	<i>Cross-cutting</i> skill	<i>Data Science</i> 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”)

## How does Disney do it?

## Personas: Business

- Information Worker
  - Excel, Powerpoint
  - Prepared BI reports
  - Light Statistics
- Business Analyst
  - Excel, Powerpoint
  - COTS Reporting tool
  - Light Statistics
- Data Analyst
  - Excel, Powerpoint
  - COTS Reporting tool
  - SQL

# Workflow (“How does Disney do it?”)

## Personas: Data Scientist

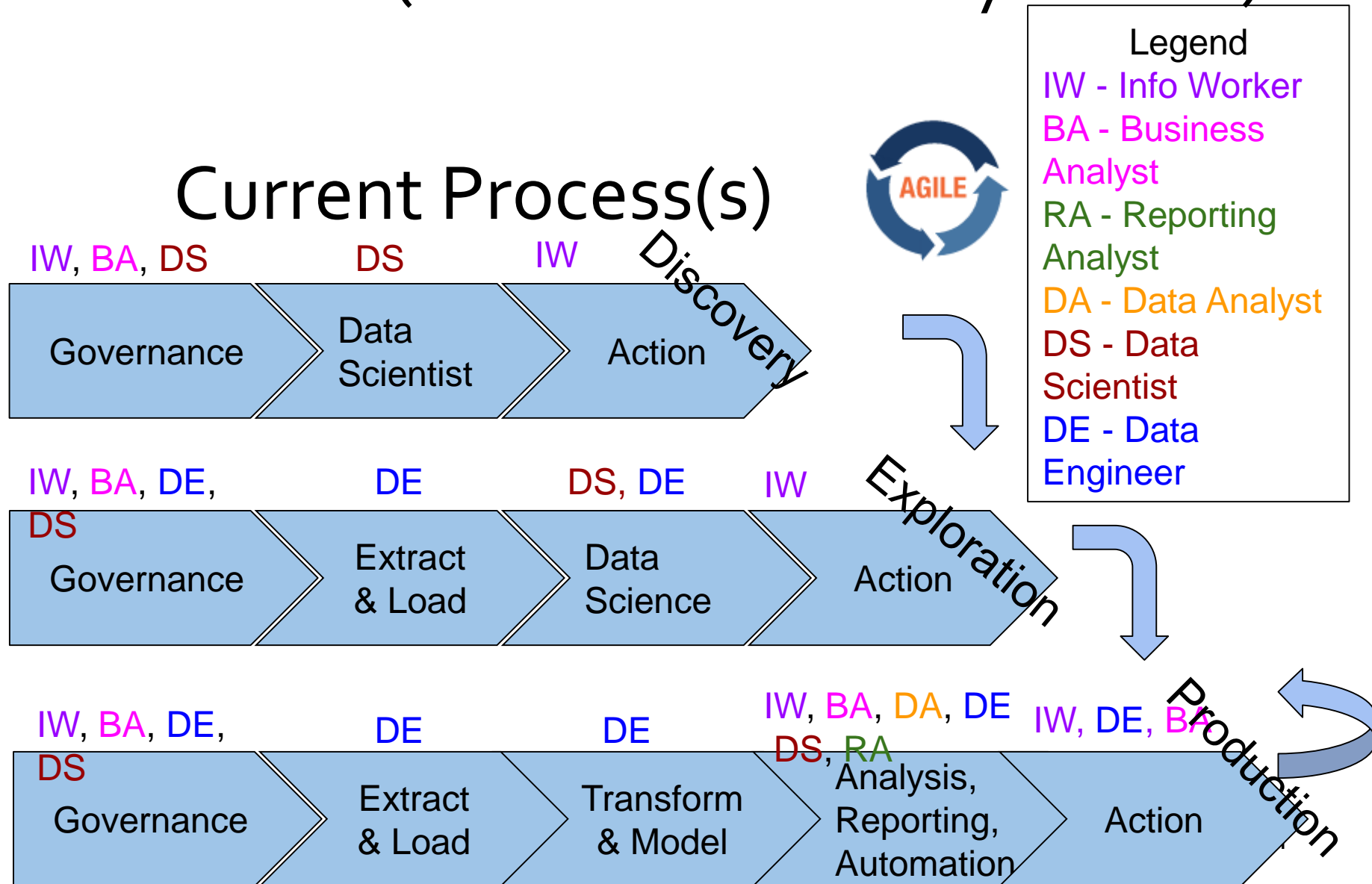
- Data Scientist
  - Required: Python & SQL; Nice to have: Java, Scala
  - Machine Learning, Statistics, Deep Learning
  - Data wrangling skills
  - Distributed systems & algorithms
  - Data Sampling, approximate aggregations, extrapolation
  - Scientific Method - Notebooks
  - Data communication, visualization
  - Cloud services, Linux CLI
  - Bonus: NLP, image recognition

# Workflow (“How does Disney do it?”)

## Personas: Technical

- Data Engineer
  - Distributed Systems, Stream Processing
  - Tools, Infrastructure, Frameworks, Services
  - Java, Scala, SQL, Python, R, Bash/Zsh
  - Linux, Git, DevOps, Cloud
  - Medium Stats
  - Medium ML/DL
  - Hadoop, Yarn, HDFS, ElasticSearch
- Reporting Programmer Analyst
  - COTS Reporting tool
  - SQL

# Workflow ("How does Disney do it?")



# Twitter, Podcasts, Tutorials

- Twitter
  - #rstats
- Podcasts
  - <https://www.analyticsvidhya.com/blog/2018/01/10-data-science-machine-learning-ai-podcasts-must-listen/>
- Tutorials
  - <http://tutorials.iq.harvard.edu/R/Rintro/Rintro.html>
- More references:
  - <https://smithw.org/datajam>

6  
SEP

Thursday, September 6, 2018, 7:00 PM

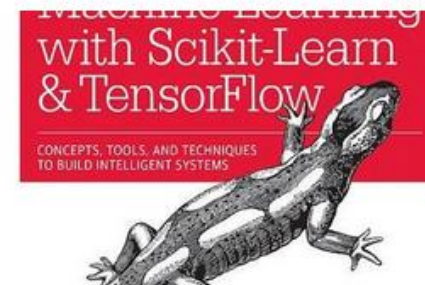
## Book Club: Hands-On Machine Learning with Scikit-Learn and TensorFlow

Hosted by [Cassie Borish](#)

Join our biweekly meetup as we work through Hands-On Machine Learning with Scikit-Learn & TensorFlow (<https://amzn.to/2jpUoiN>). Jupyter notebooks accompanying the text can be found here: <https://github.com/ageron/handson-ml> \*\*\*\*\* Tonight we are discussing and working through exercises from Chapter 8: Dimensionality Reduction. \*\*\*\*\* Haven't done the previous readings? That's ok! Summaries of previous chapters are posted here:



8 going

[Attend](#)**Zappbuddy Technologies**

25000 Avenue Stanford · Santa Clarita, CA



7 comments

### What we're about

This meetup is designed to create a meeting point in Santa Clarita for those

### Upcoming Meetups

[See all](#)

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



## PERSIAN WOMEN



Thursday, January 25, 2018

6:30 PM to 9:00 PM



Hulu Santa Monica HQ

2500 Broadway Street Suite 200 ·  
Santa Monica, ca





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

# Career Homework

- TechFest (CSUN Career Center)
  - Marketed mostly to Eng/CompSci, but a data science learner goes...
  - Also, if a display sign doesn't list "data science" (or whatever) on the sign, *you ask about it...*
- Indeed.com
  - "data science intern", "data science entry level"
  - E.g., 25 miles from CSUN
  - Do your homework from at least 20 job descriptions
    - What skills do I have? What skills don't I have but I can get? What skills don't I have that I don't even know what the skill is?

# Cloud Services

- 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>)
- Amazon certificate (<https://laedc.org/2018/08/09/amazon-los-angeles-colleges-cloud/>)
  - SMC (<http://www.smc.edu/NewsRoom/Pages/Cloud-Computing-Certificate.aspx>)
- But keep in mind threats to both *reproducibility* and *workflow*

# 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" )

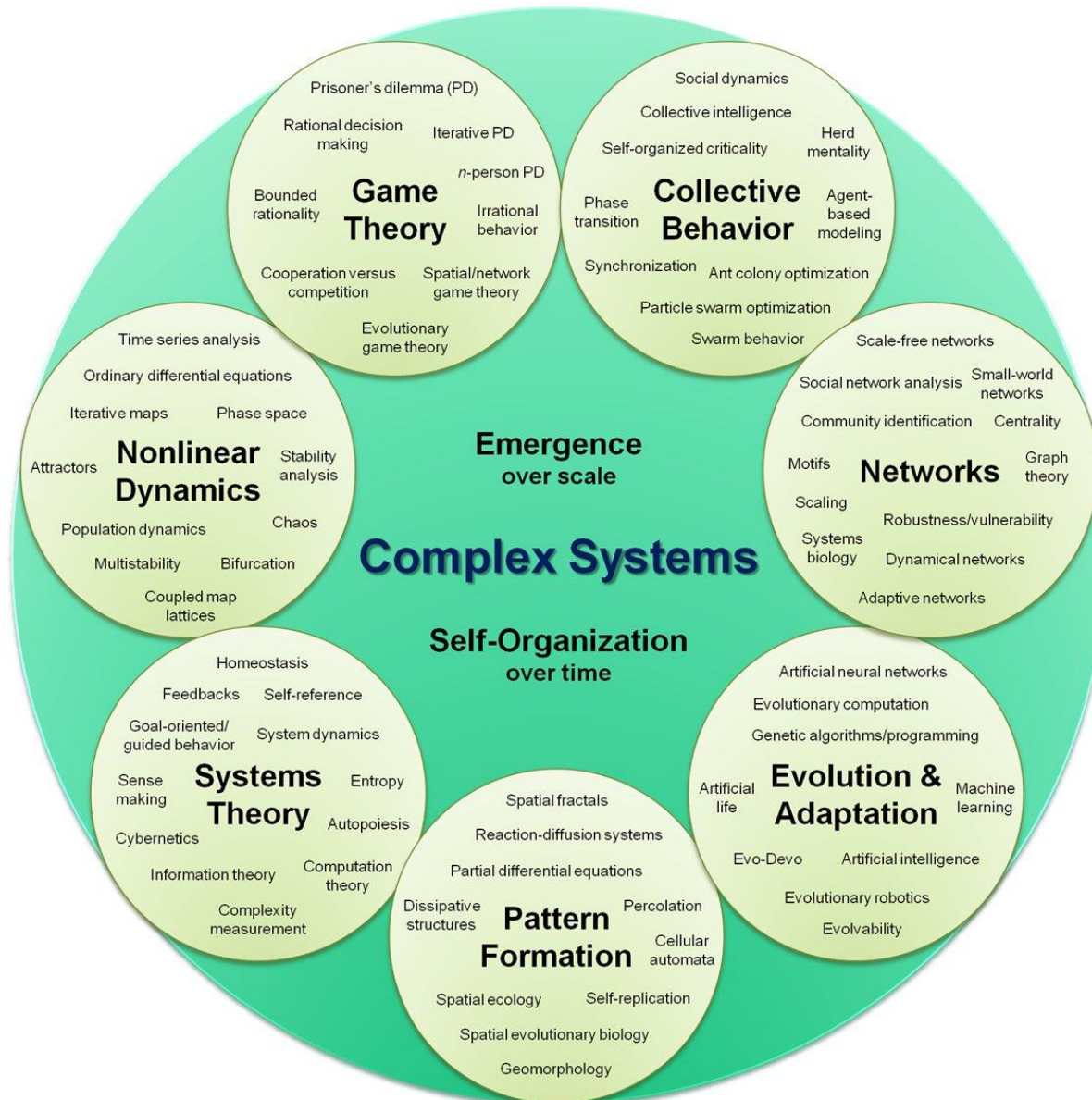
# get course items
get_user_items( course_id = 12345, item = "assignments" )
get_user_items( course_id = 12345, item = "missing_submissions" )

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

# upload a file
upload_course_file( course_id = 12345, file_name = "testfile.docx" )
```

# Recent blog on LAPD arrests/crimes

# Macro-level (Complexity)



# *Micro-level*

## (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 would you like to learn in the near future?
  - How do you learn new things related to using data?

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 Covert 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.

# References

- Automation vs. Analytics
  - Davenport, T. (2009), "Make Better Decisions", *Harvard Business Review*, Nov. 87(11), p. 117-123.
- Curriculum Guidelines for Undergraduate Programs in Data Science
  - <https://www.stat.berkeley.edu/~nolan/Papers/Data.Science.Guidelines.16.9.25.pdf>
- Complex Systems chart
  - By Hiroki Sayama, D.Sc. - Created by Hiroki Sayama, D.Sc., Collective Dynamics of Complex Systems (CoCo) Research Group at Binghamton University, State University of New York, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=12191267>

*fin*

- Again: Are you a data science learner?
- Questions?
- Goodies