SEBASTIAN RASCHKA

GETTING STARTED WITH DATA SCIENCE

MICHIGAN STATE DATA SCIENCE

Speaker Series

September 27, 2016

Data Science?

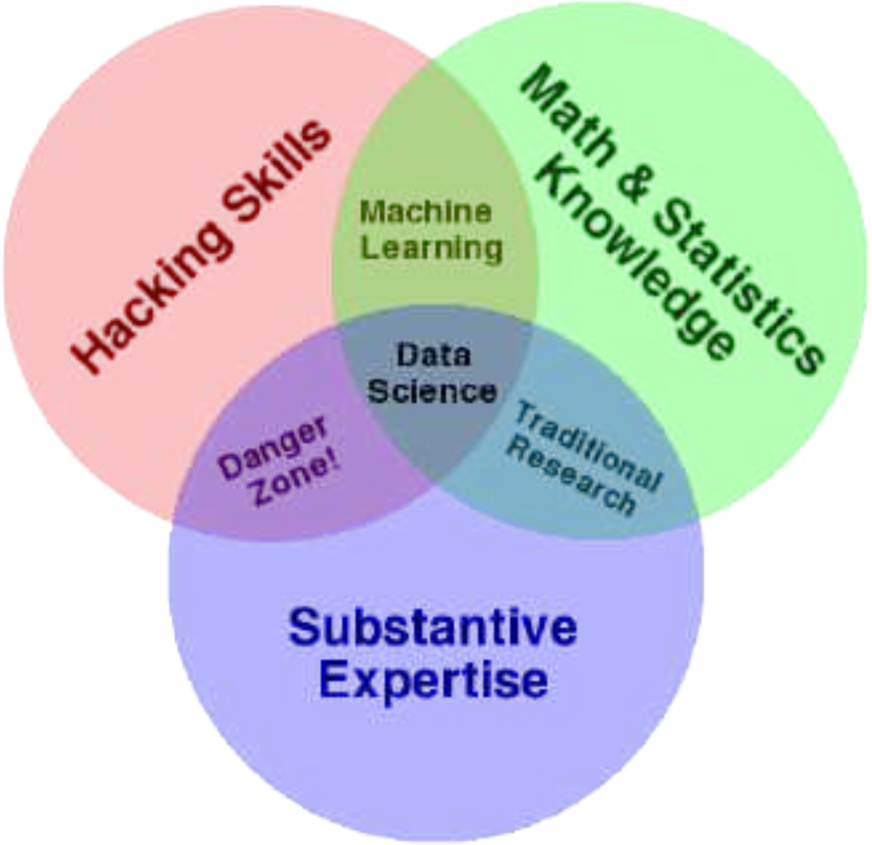
"Data Scientist" is a Data Analyst who lives in California.

– @nivertech

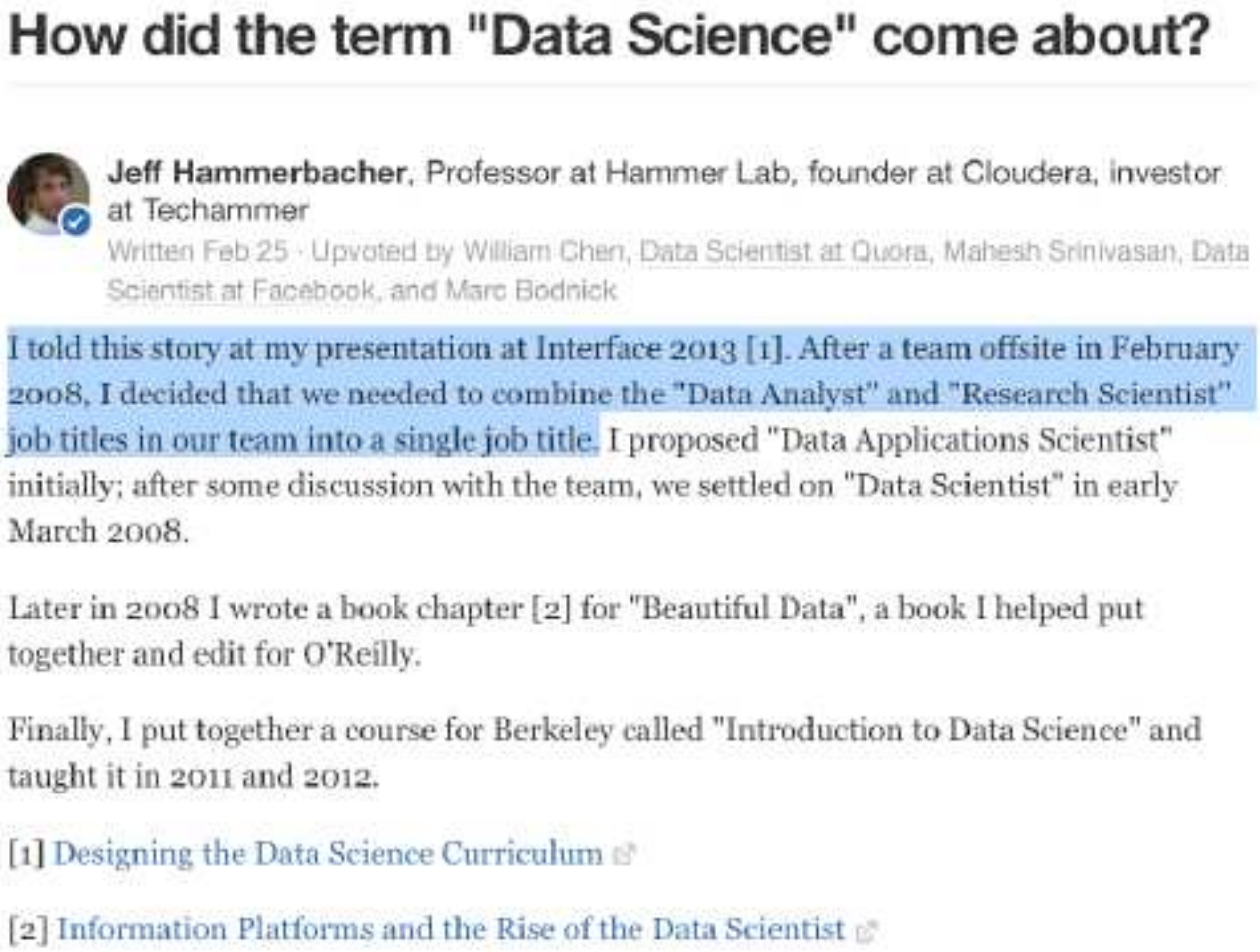
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

– Josh Wills (Cloudera)

DREW CONWAY'S CLASSIC



POPULARIZED
THE TERM
~2008



How did the term "Data Science" come about?

Jeff Hammerbacher, Professor at Hammer Lab, founder at Cloudera, investor at Techammer

Written Feb 25 · Upvoted by William Chen, Data Scientist at Quora, Mahesh Srinivasan, Data Scientist at Facebook, and Marc Bodnick

I told this story at my presentation at Interface 2013 [1]. After a team offsite in February 2008, I decided that we needed to combine the "Data Analyst" and "Research Scientist" job titles in our team into a single job title. I proposed "Data Applications Scientist" initially; after some discussion with the team, we settled on "Data Scientist" in early March 2008.

Later in 2008 I wrote a book chapter [2] for "Beautiful Data", a book I helped put together and edit for O'Reilly.

Finally, I put together a course for Berkeley called "Introduction to Data Science" and taught it in 2011 and 2012.

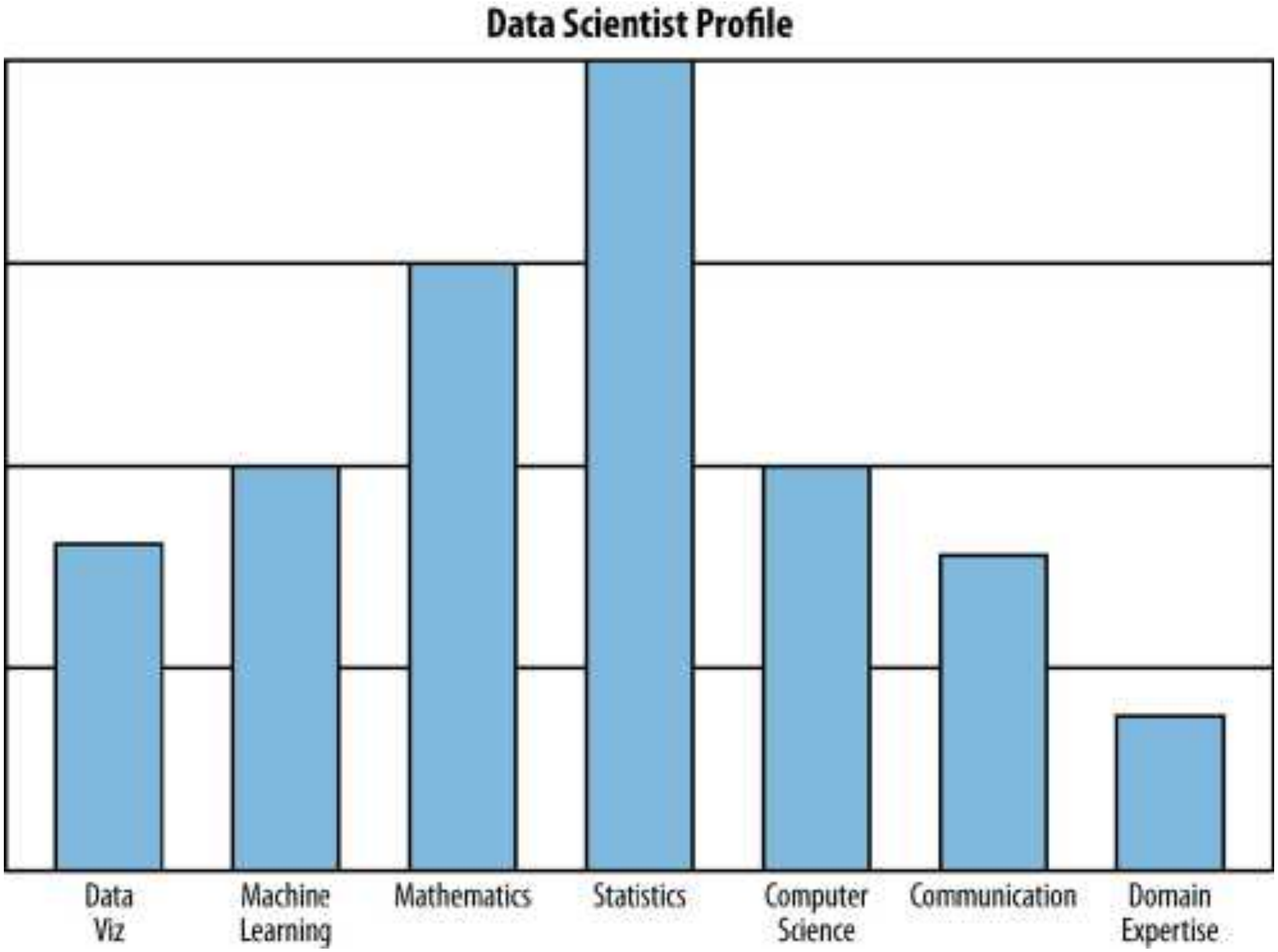
[1] Designing the Data Science Curriculum

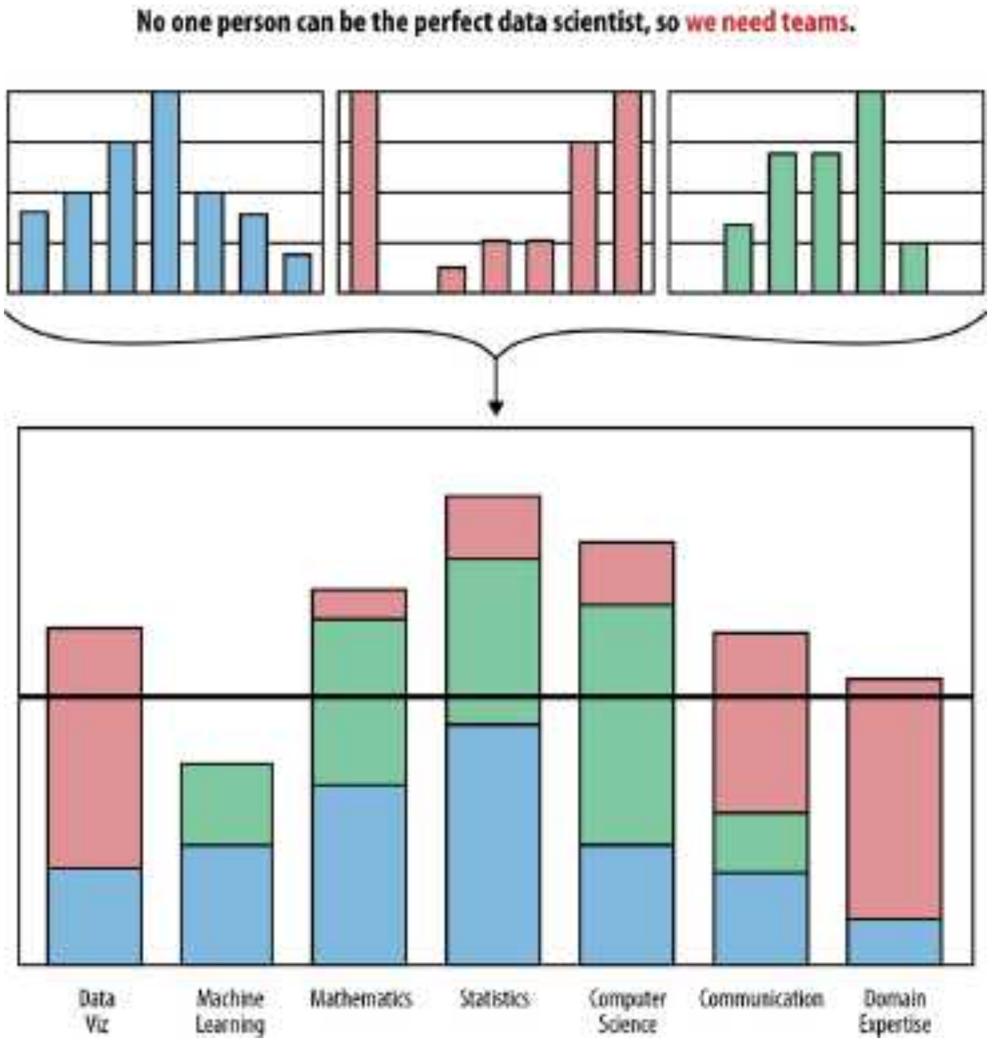
[2] Information Platforms and the Rise of the Data Scientist

[https://www.quora.com/How-did-the-term-Data-Science-come-about/answer/Jeff-Hammerbacher]

“Rachel’s data science profile, which she created to illustrate trying to visualize oneself as a data scientist;”

From: Cathy O'Neil & Rachel Schutt. “Doing Data Science

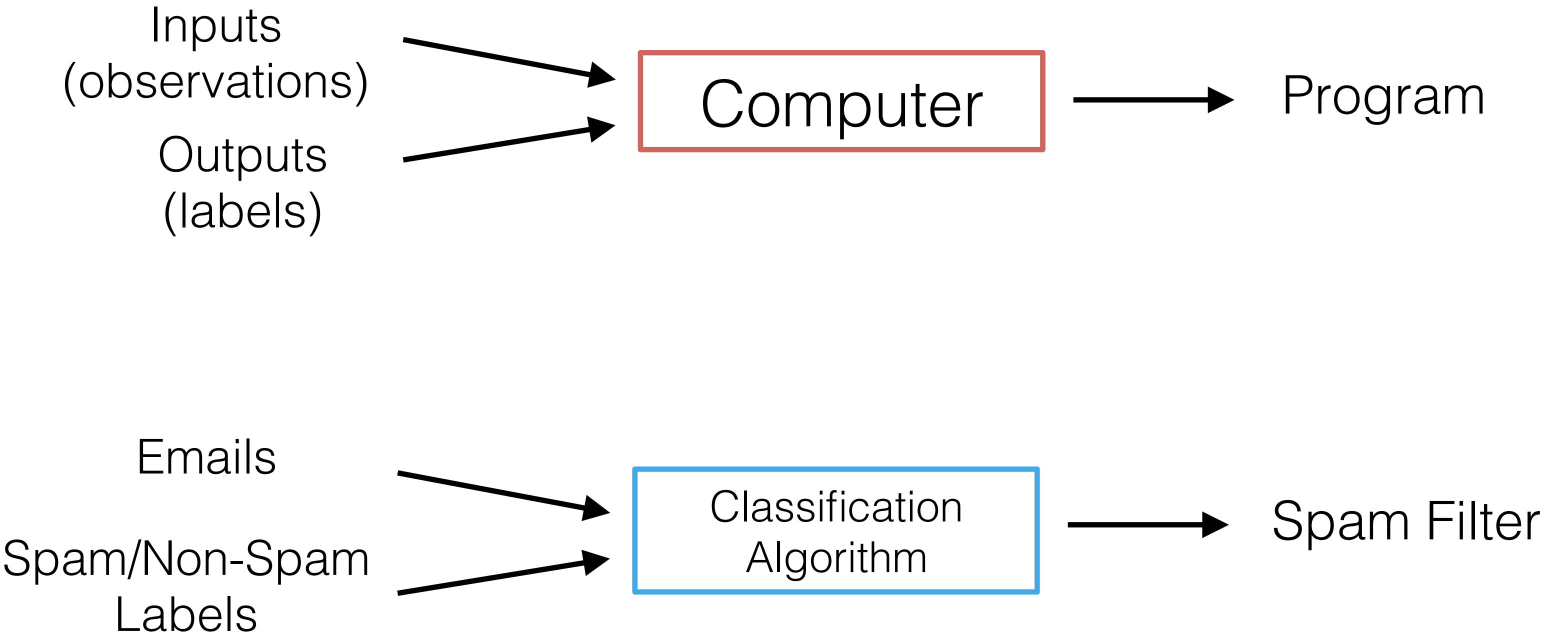




From: Cathy O'Neil & Rachel Schutt.
“Doing Data Science

Machine learning?

What is Machine Learning?



Inputs (observations)

Outputs (labels)

Computer

Program

Emails

Spam/Non-Spam Labels

Classification Algorithm

Spam Filter

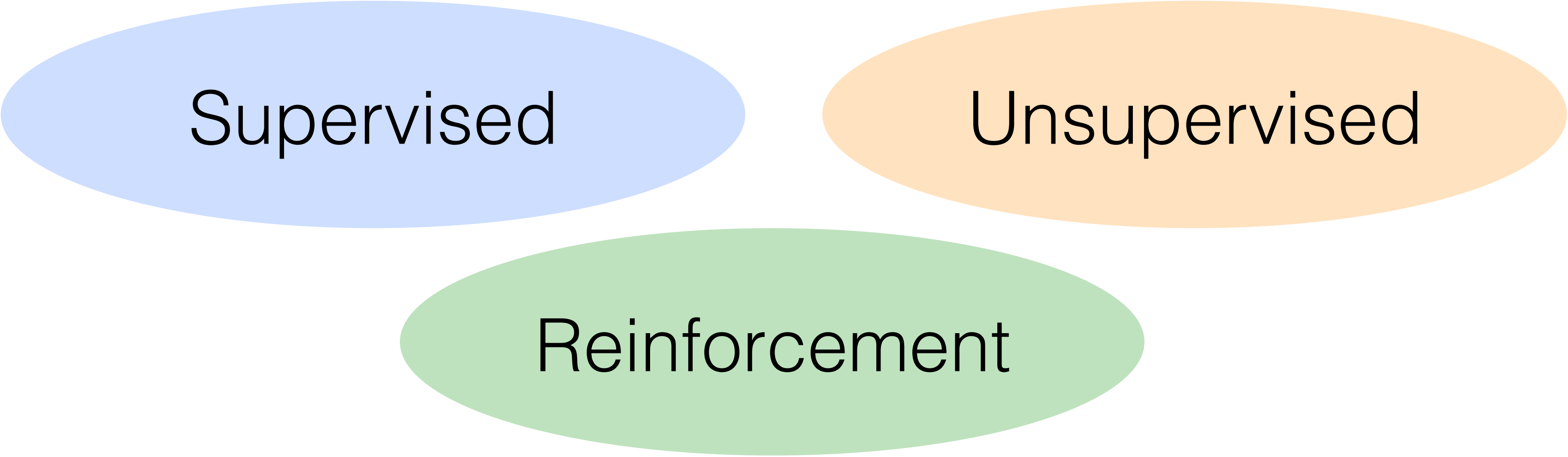
What can Machine Learning do for us?

https://flic.kr/p/5BLW6G [CC BY 2.0]

https://commons.wikimedia.org/wiki/
File:Google_self_driving_car_at_the_Googleplex.jpg

Photo by Michael Shick, CC BY-SA 4.0

3 Types of Learning

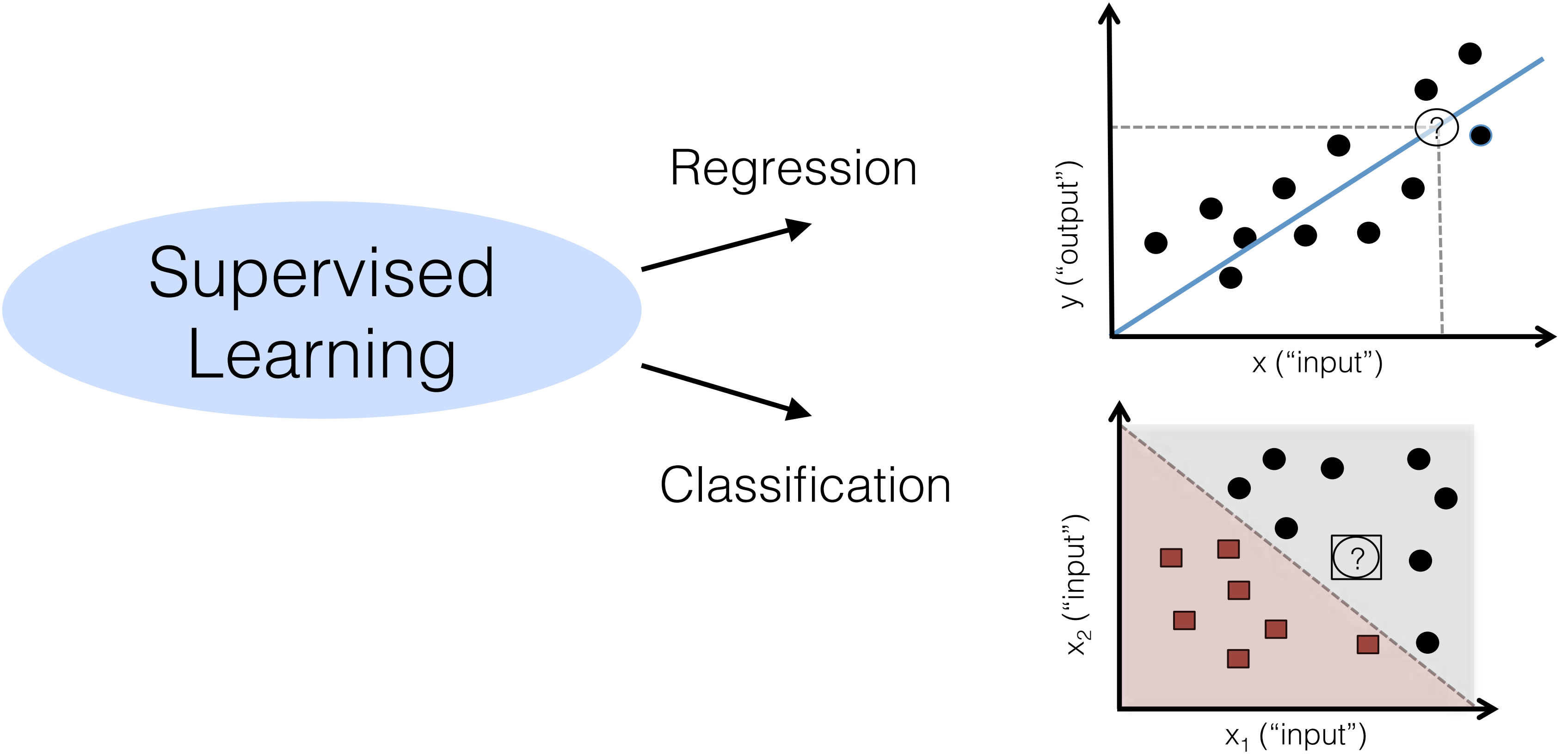


Supervised

Unsupervised

Reinforcement

Working with Labeled Data



Supervised Learning

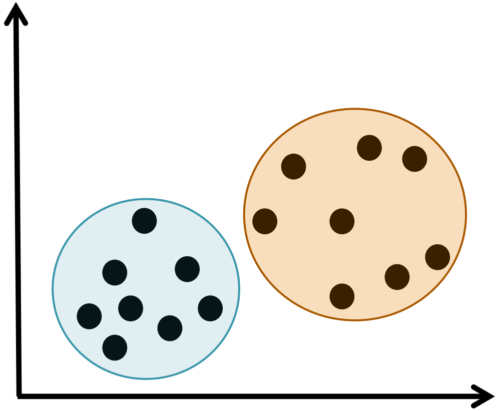
Regression

Classification

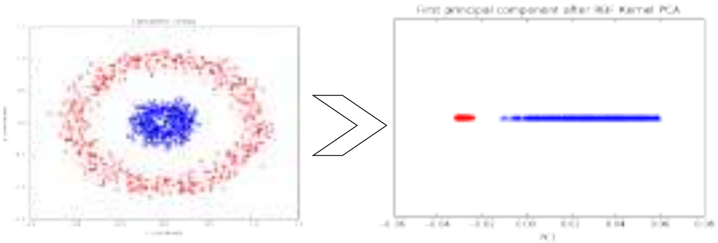
Working with Unlabeled Data

Unsupervised Learning

Clustering



Compression

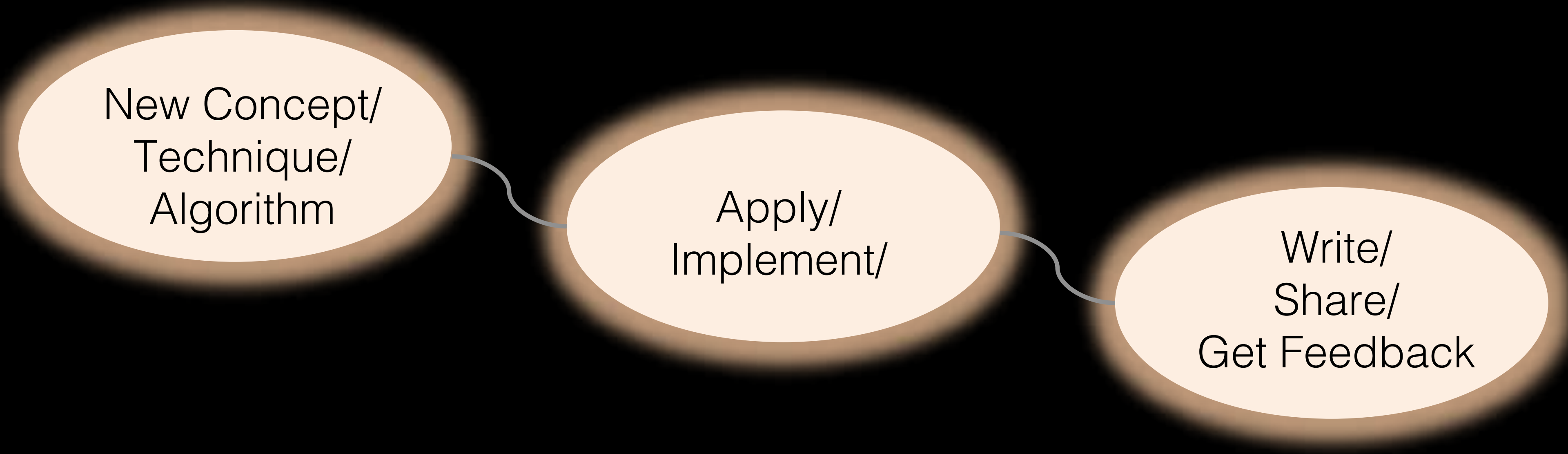


GETTING STARTED WITH DATA SCIENCE

☑ Reading (/ Classes)!

☑ Doing!

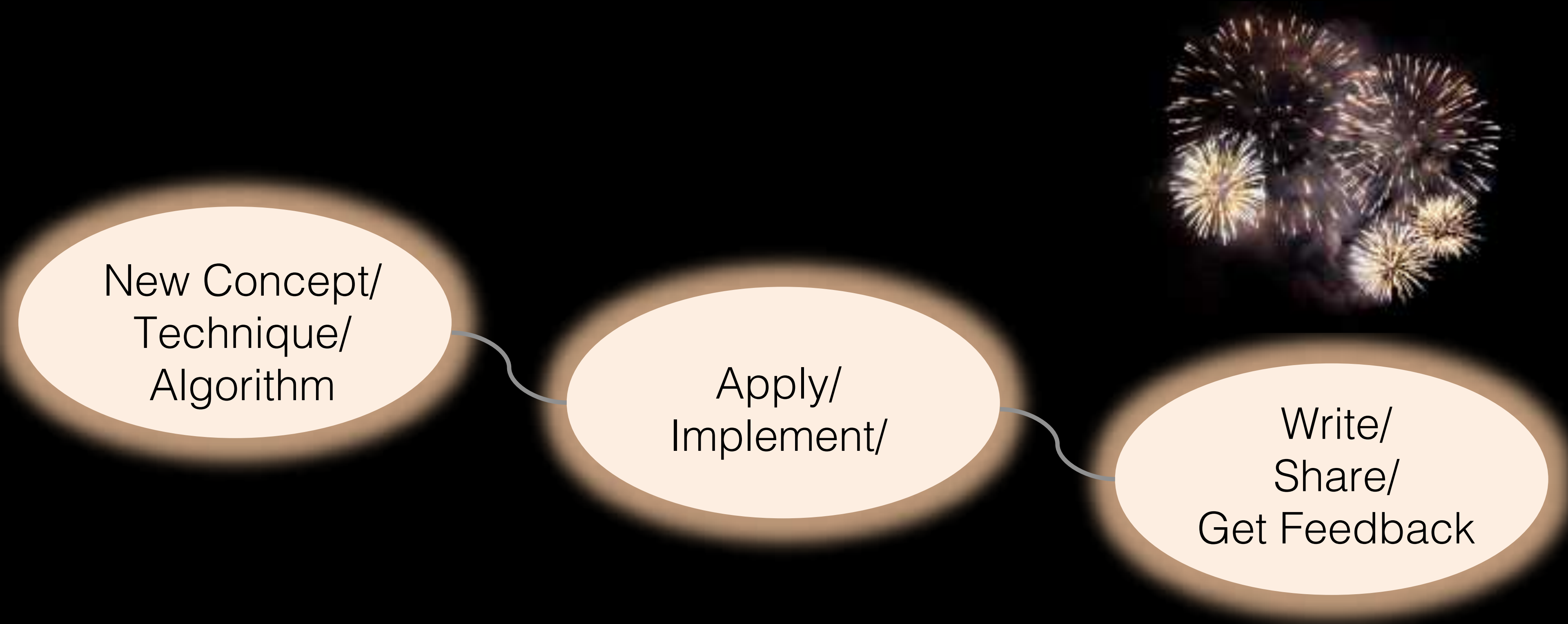
☑ Communicating!



New Concept/
Technique/
Algorithm

Apply/
Implement/

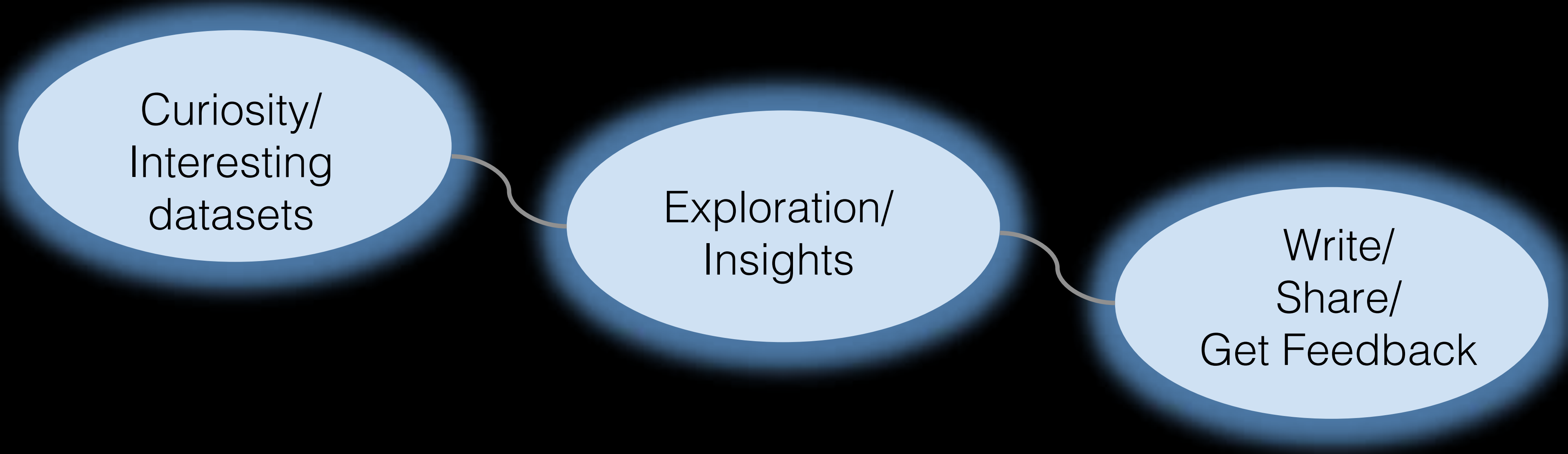
Write/
Share/
Get Feedback



New Concept/
Technique/
Algorithm

Apply/
Implement/

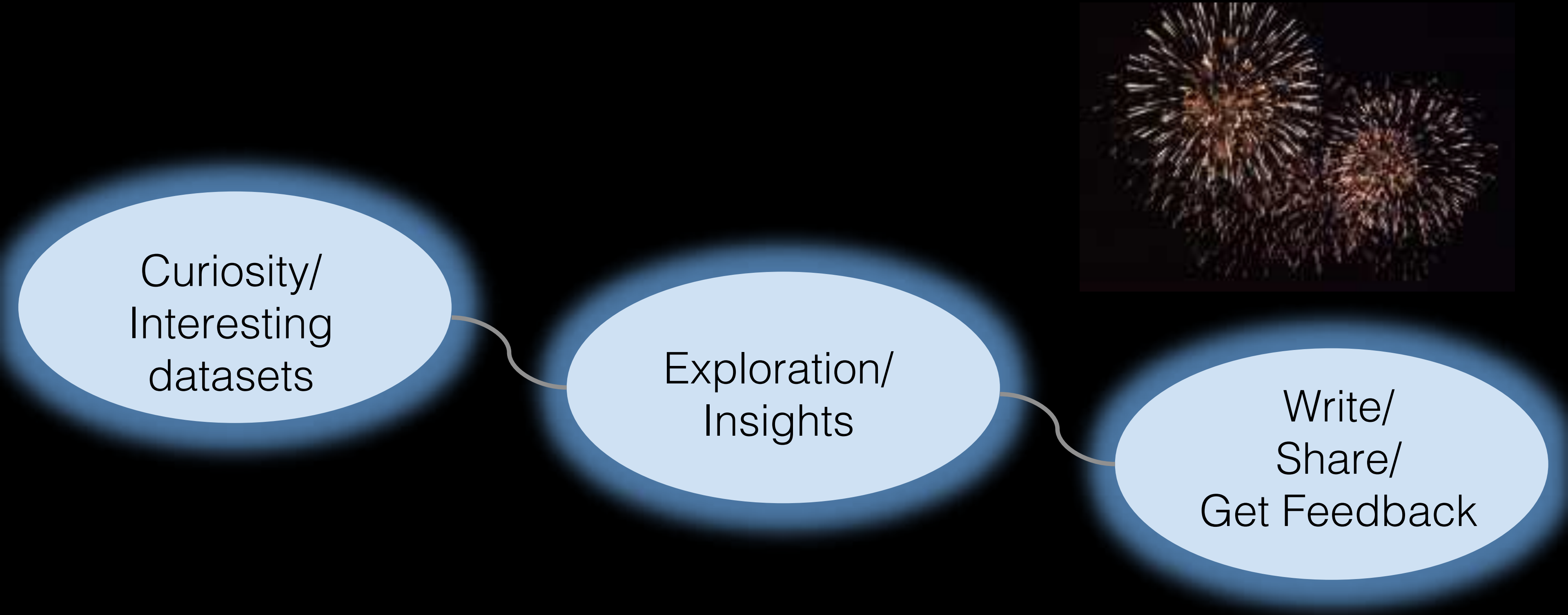
Write/
Share/
Get Feedback



Curiosity/
Interesting
datasets

Exploration/
Insights

Write/
Share/
Get Feedback

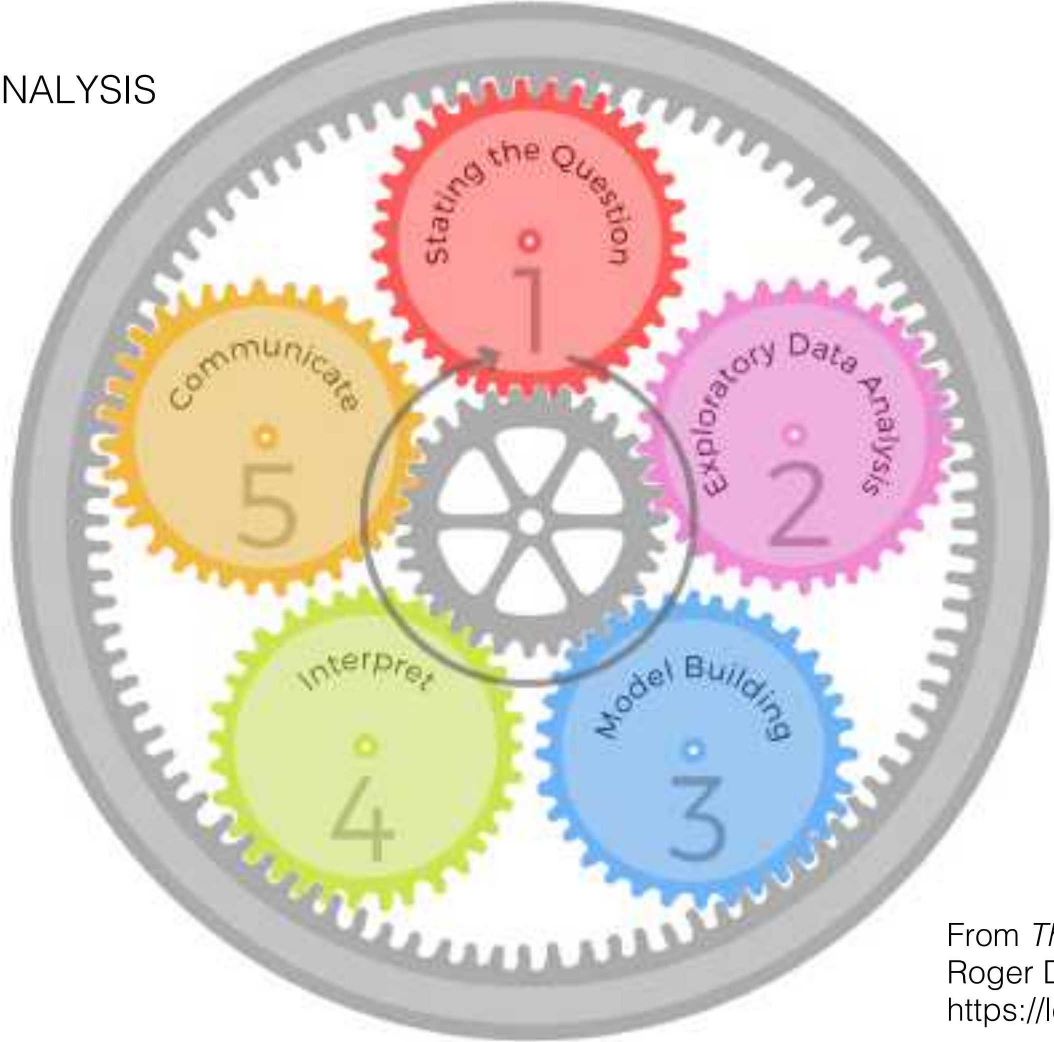


Curiosity/
Interesting
datasets

Exploration/
Insights

Write/
Share/
Get Feedback

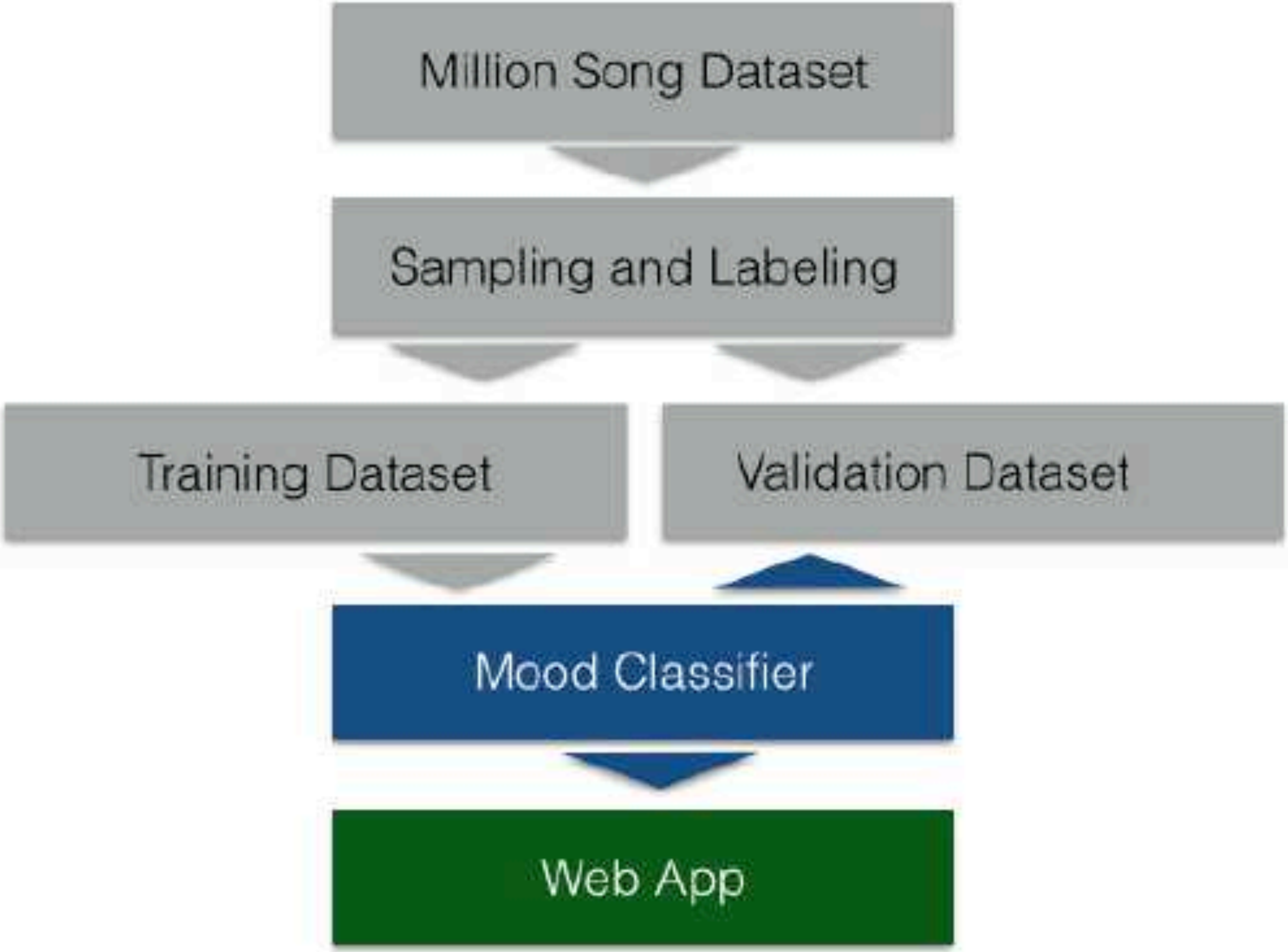
EPICYCLES OF ANALYSIS



From *The Art of Data Science* by
Roger D. Peng and Elizabeth Matsui
https://leanpub.com/artofdatascience

A few years back …

```
Position  gameday              Name  Salary            GameInfo
478       GK        1            Adrian    4900  WHU@SUN 10:00AM ET
280        M        1      Wes Hoolahan    4400  LEI@NOR 10:00AM ET
309        D        1            Cedric    4200  SOU@CHE 12:30PM ET
480        M        1  Cheikhou Kouyate    5300  WHU@SUN 10:00AM ET
25         D        1      Jordan Amavi    4500  STK@AVL 10:00AM ET
142        M        1      Riyad Mahrez   10100  LEI@NOR 10:00AM ET
143        F        1       Jamie Vardy    9000  LEI@NOR 10:00AM ET
334        F        1        Mame Diouf    6400  STK@AVL 10:00AM ET
```



Million Song Dataset

Sampling and Labeling

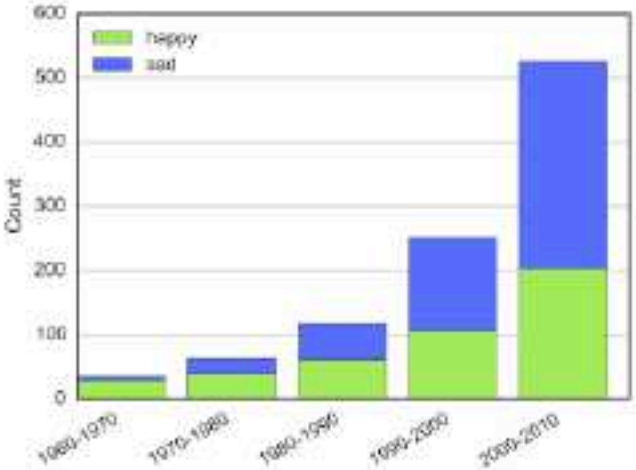
Training Dataset

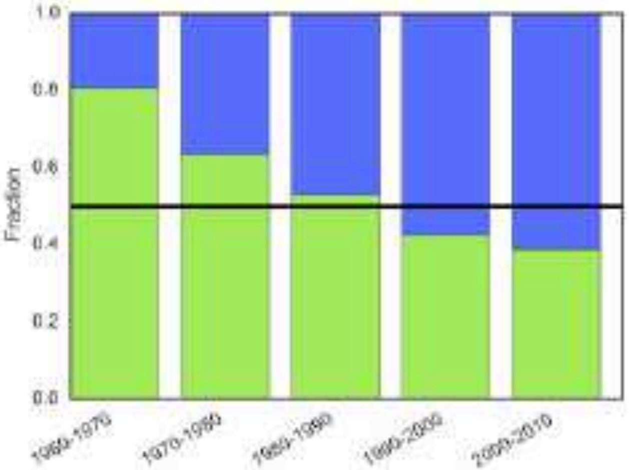
Validation Dataset

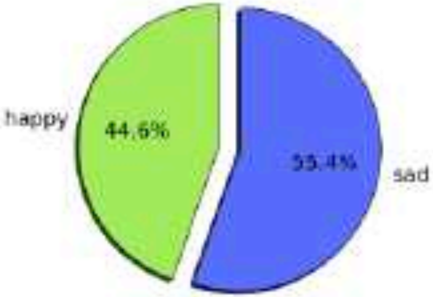
Mood Classifier

Web App

Why so sad?
The mood of music over the last 50 years



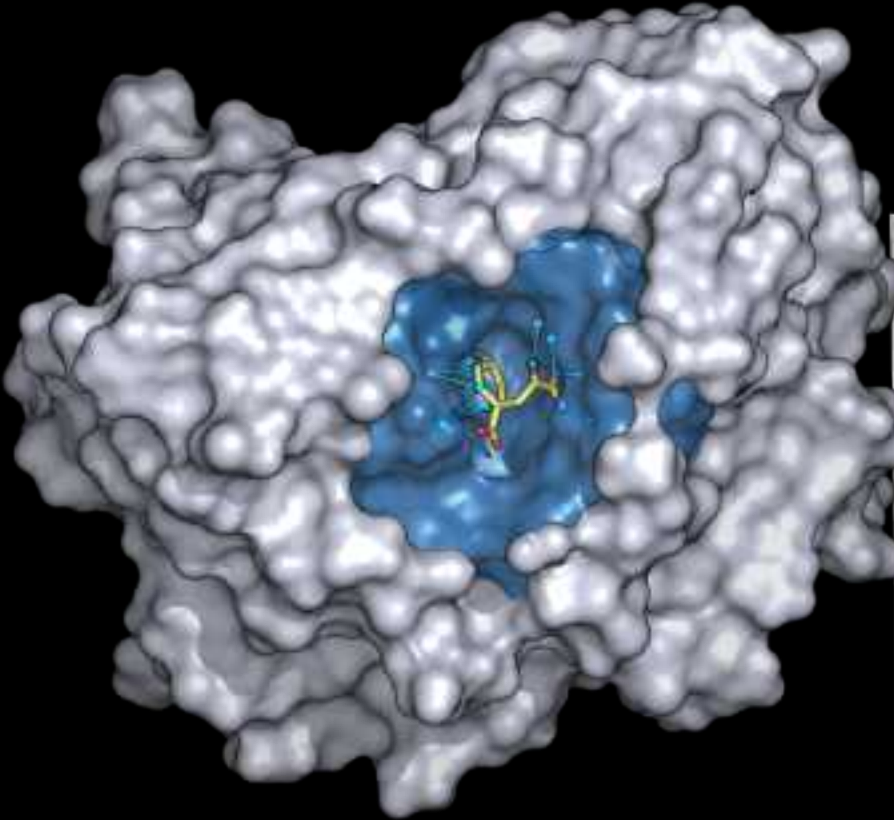


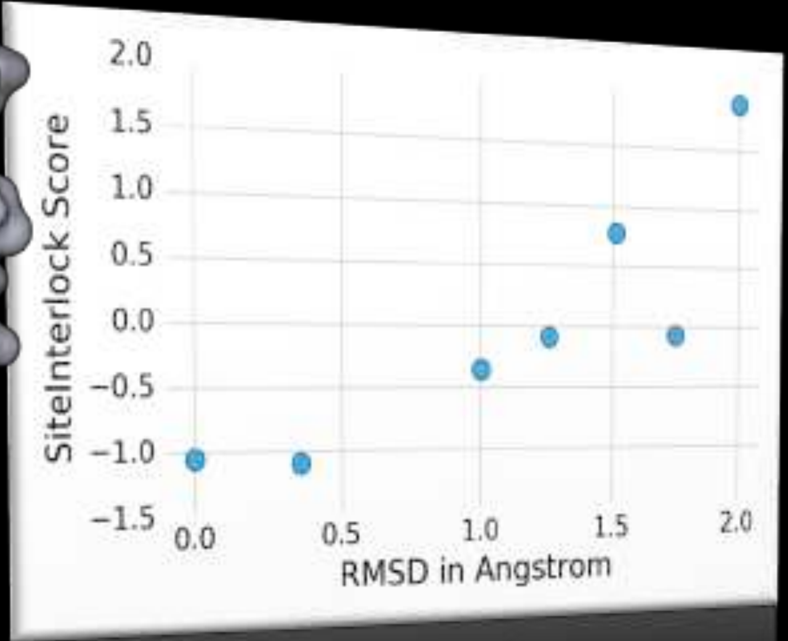


[based on the 1000-song training dataset]

https://github.com/rasbt/musicmood

Sebastian Raschka 2014




SiteInterlock Score
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
RMSD in Angstrom

Screenlamp: A Software Framework for Hypothesis-Driven Ligand Discovery Based on Virtual Screening and Machine Learning

Sebastian Raschka, Santosh Gunturu, Anne M. Scott, Mar Huertas, Weiming Li, and Leslie A. Kuhn
Michigan State University, East Lansing, MI 48824, U.S.A.
http://kuhnlab.bmb.msu.edu

Introduction

- The goal in *virtual screening*, the high-throughput computational evaluation of small molecules as potential protein activators or inhibitors, is to select a small set likely to show activity in experimental tests.
- The challenge is to identify features that distinguish a small number of active compounds (typically 10 or fewer) from 100,000s to millions of molecules being screened.
- We developed Screenlamp, a computational tool
  - to increase the computational efficiency and success rate in virtual screening
  - and to facilitate hypothesis-driven molecular selection and the analysis of structure-activity relationships using machine learning.

References

1. Allen, F. (2002). The Cambridge Structural Database: A quarter of a million crystal structures and rising. Acta Crystallographica Section B: Structural Science. [illegible]
2. Irwin, J. J., & Shoichet, B. K. (2005). ZINC – A free database of commercially available compounds for virtual screening. Journal of Chemical Information and Modeling, 45(1), 177-82. [illegible]
3. Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., & Stahl, M. T. (2010). Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. Journal of Chemical Information and Modeling, 50(4), 572-84. [illegible]
4. Hawkins, P. C. D., Skillman, A. G., & Nicholls, A. (2007). Comparison of shape-matching and docking as virtual screening tools. Journal of Medicinal Chemistry, 50(1), 74-82. [illegible]
5. Gatica, E. A., & Cavasotto, C. N. (2012). Ligand and decoy sets for docking to G protein-coupled receptors. Journal of Chemical Information and Modeling, 52(1), 1-6. [illegible]
6. Raschka, S., Scott, A. M., Liu, N., Gunturu, S., Huertas, M., Li, W., & Kuhn, L. A. [illegible] (manuscript in preparation).
7. Raschka, S. [illegible] (2016). [illegible]
8. [illegible]

Acknowledgements

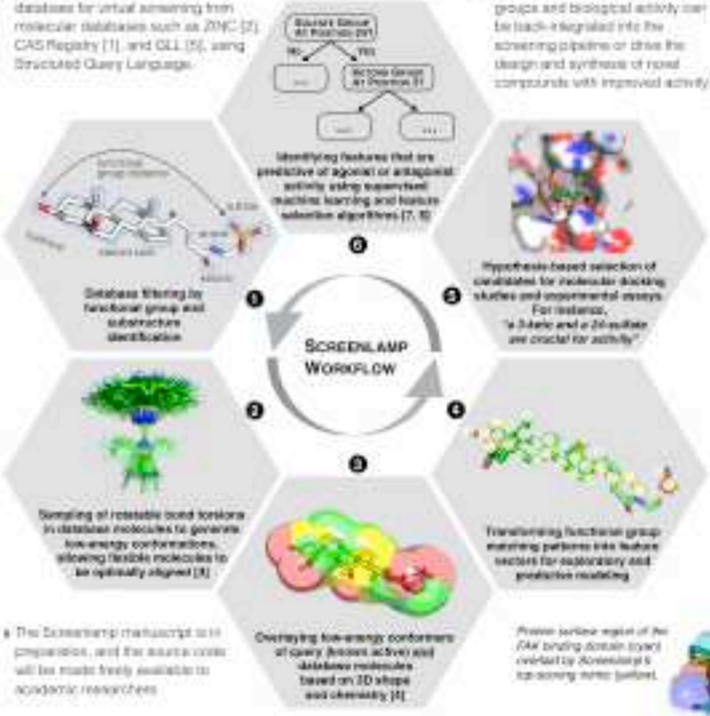
This research was supported by grants from the Great Lakes Fishery Commission. [illegible]

Concept and Methods

- Screenlamp is a virtual screening framework to identify structural, volumetric, and chemical mimics of a known query molecule interacting with the protein target of interest. Our framework allows scientists to incorporate hypotheses about the importance of certain functional groups, their spatial orientation to each other, and experimental data to facilitate the identification of biologically active molecules.

- Screenlamp curates a relational database for virtual screening from molecular databases such as ZINC [2], CAS Registry [1], and GLL [5], using Structured Query Language.

- The relationship between functional groups and biological activity can be back-integrated into the screening pipeline or drive the design and synthesis of novel compounds with improved activity.



1 Database filtering by functional group and substructure identification

2 Sampling of rotatable bond torsions in database molecules to generate low-energy conformations, allowing flexible molecules to be optimally aligned [3]

3 Overlaying low-energy conformers of query (known active) and database molecules based on 3D shape and chemistry [4]

4 Transforming functional group matching patterns into feature vectors for exploratory and predictive modeling

5 Hypothesis-based selection of candidates for molecular docking studies and experimental assays. For instance, *"a 3-keto and a 24-sulfate are crucial for activity"*

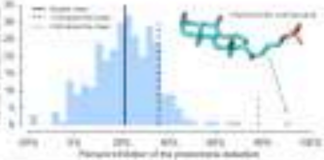
6 Identifying features that are predictive of agonist or antagonist activity using supervised machine learning and feature selection algorithms [7, 8]

- The Screenlamp manuscript is in preparation, and the source code will be made freely available to academic researchers.

Applications and Results

Project 1: Discovering pheromone antagonists for a G-protein coupled receptor

- Screenlamp screened more than 6 million commercially available compounds, identifying 311 for experimental assays testing 12 hypotheses. Based on *in vivo* experiments performed by our collaborators (Weiming Li lab, MSU), 11 of these compounds were found to block 45-100% of the pheromone detection in sea lamprey, an invasive species in the Great Lakes of North America.
- One compound, a non-toxic bile acid, was highly active, blocking 80% of sea lamprey pheromone detection in very low (10⁻⁷M) concentration and nullified the sea lamprey response to the mating pheromone in a natural stream [8].



Activity distribution of 311 Screenlamp-selected compounds from biological assays (3-4 replicas per experiment).

Project 2: Stimulating bone regeneration

- Screenlamp is also being used in collaboration with Kurt Hankenson's lab at MSU to develop mimics of Notch ligands to stimulate bone regrowth, funded by the Department of Defense.

Project 3: Blocking FAK interaction to block cell adhesion in cancer

- Cell-cell adhesion is an important step in cancer metastasis. In collaboration with Bai Zeng and Marc Basson (University of North Dakota), we are using Screenlamp to discover focal adhesion kinase (FAK) mimics that block cell adhesion.

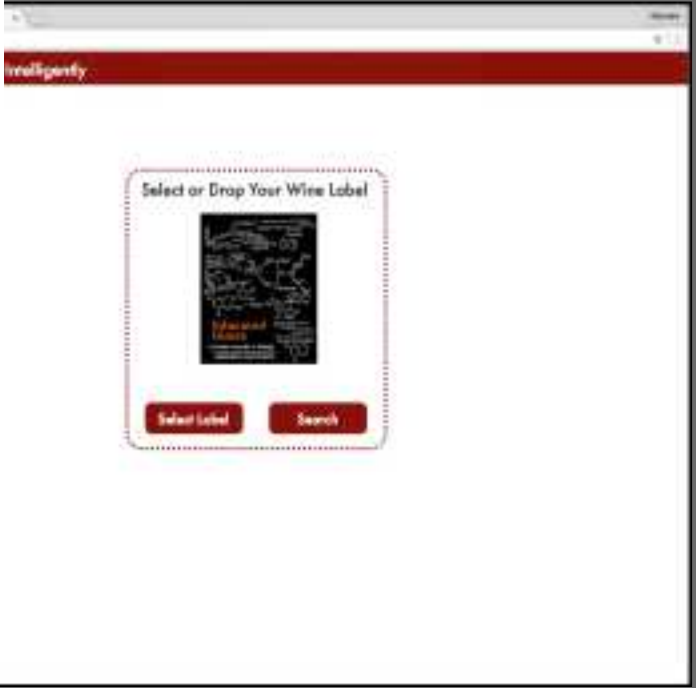
Protein surface region of the FAK binding domain (cyan) overlaid by Screenlamp's top-scoring mimic (yellow).

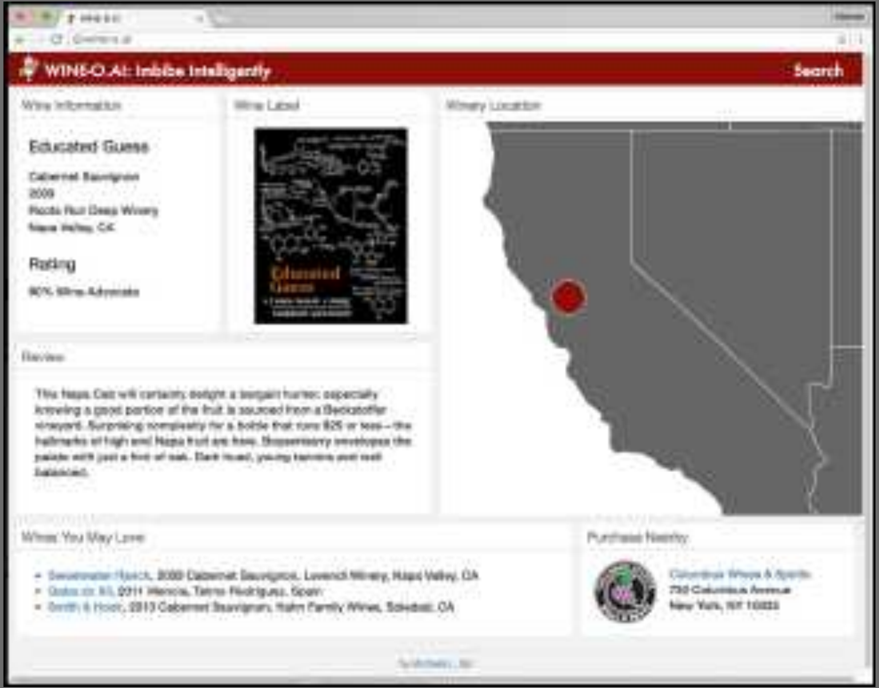
Some Interesting Project Ideas …

Michelle Gill
mlgill

Biophysicist in transition to data scientist. Part of the Summer 2016 @thisismetis Data Science Bootcamp cohort. My Git [illegible]

WINE-O.AI:
Computer Vision Assisted Wine Recommendations





Educated Guess

Cabernet Sauvignon
2009
Roots Run Deep Winery
Napa Valley, CA

Rating

90% Wine Advocate

Review

This Napa Cab will certainly delight a bargain hunter, especially knowing a good portion of the fruit is sourced from a Beckstoffer vineyard. Surprising complexity for a bottle that runs $25 or less—the hallmarks of high end Napa fruit are here. Boysenberry envelopes the palate with just a hint of oak. Dark fruit, young tannins and well balanced.

Wines You May Love

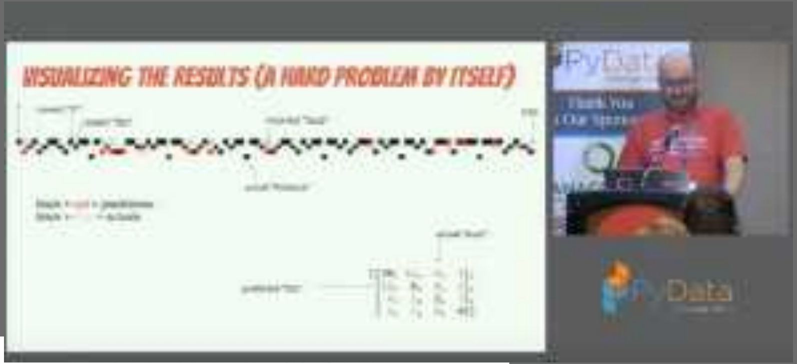
- Sweetwater Ranch, 2009 Cabernet Sauvignon, Lovenci Winery, Napa Valley, CA
- Gaba do Xil, 2011 Mencia, Telmo Rodriguez, Spain
- Smith & Hook, 2013 Cabernet Sauvignon, Hahn Family Wines, Soledad, CA

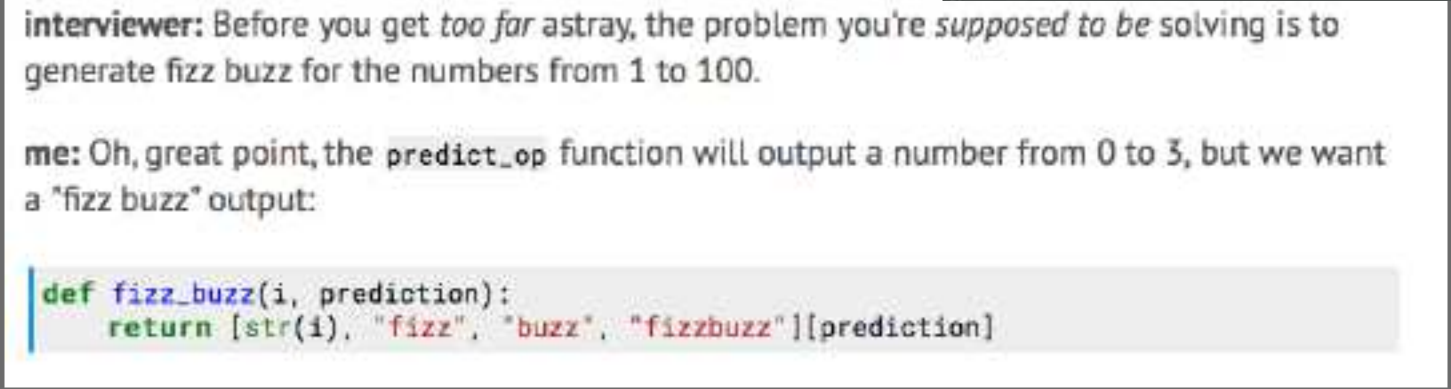
Purchase Nearby

Columbus Wines & Spirits
730 Columbus Avenue
New York, NY 10025

by Michelle L. Gill

Fizz Buzz in Tensorflow

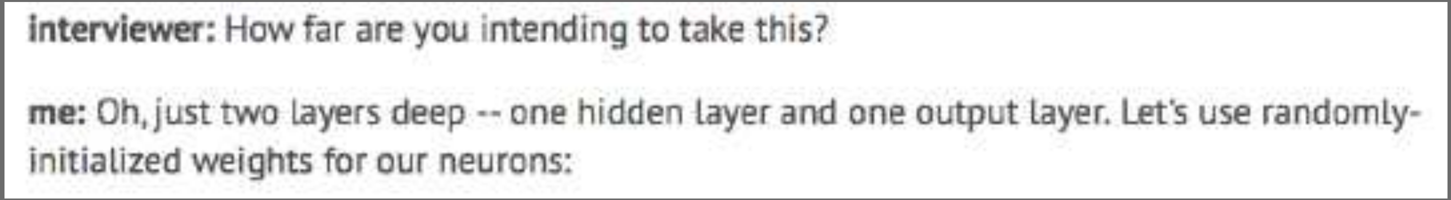




interviewer: Before you get *too far* astray, the problem you're *supposed to be* solving is to generate fizz buzz for the numbers from 1 to 100.

me: Oh, great point, the `predict_op` function will output a number from 0 to 3, but we want a "fizz buzz" output:

```
def fizz_buzz(i, prediction):
    return [str(i), "fizz", "buzz", "fizzbuzz"][prediction]
```



interviewer: How far are you intending to take this?

me: Oh, just two layers deep -- one hidden layer and one output layer. Let's use randomly-initialized weights for our neurons:

http://joelgrus.com/2016/05/23/fizz-buzz-in-tensorflow/

Pomegranate: fast and flexible probabilistic models in Python

http://pomegranate.readthedocs.io/en/latest/

WHY
PROJECTS?

What are the

TOOLS?

R

python™

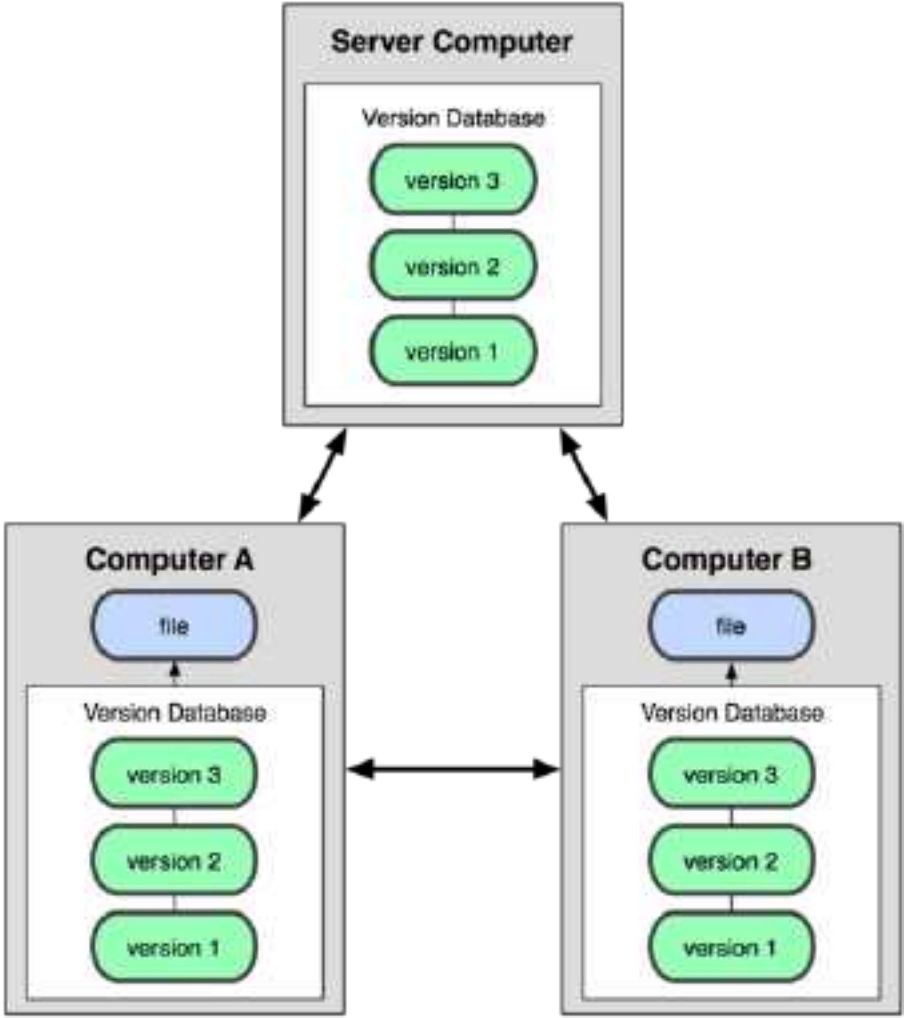
“R is a programming language developed by statisticians for statisticians; Python was developed by a computer scientist, and it can be used by programmers to apply statistical techniques.”

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

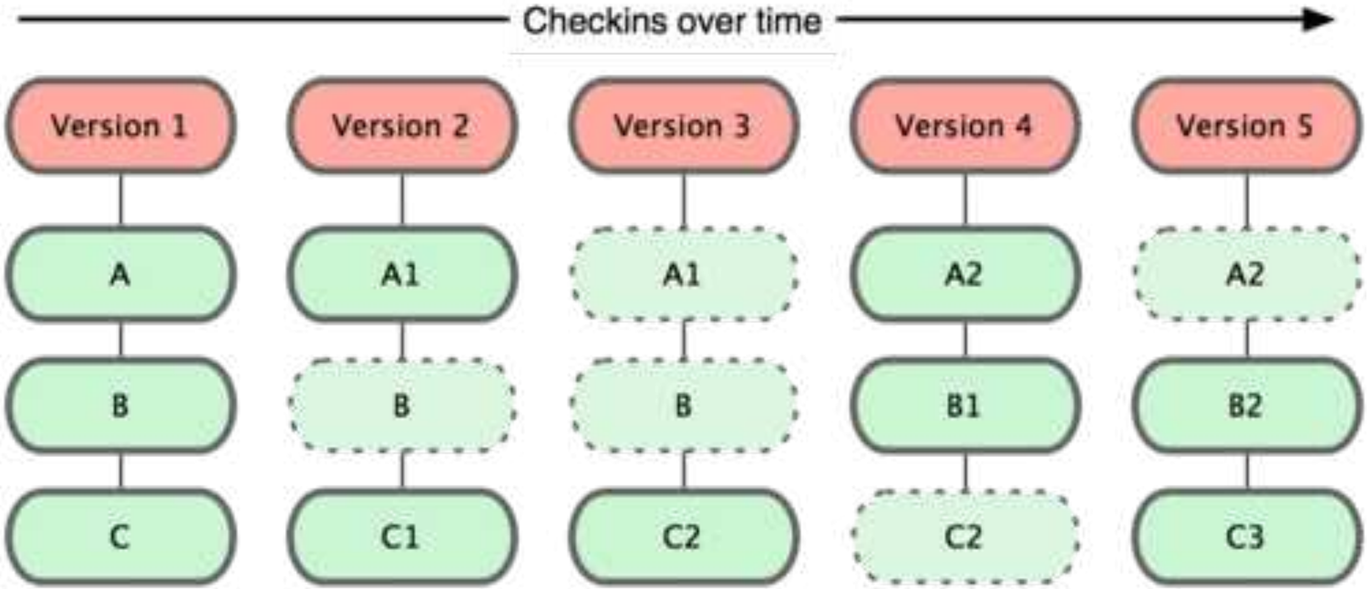
The image cannot be displayed. Your computer may not have enough memory to

GitHub

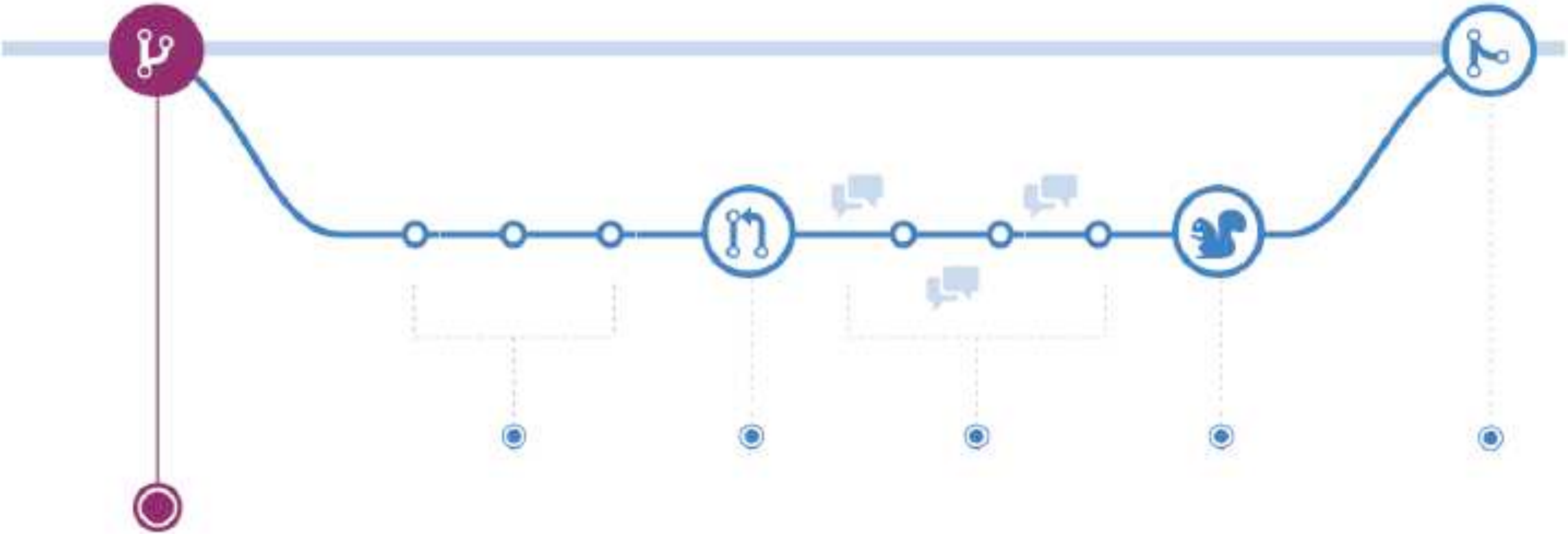


From: Scott Chacon. “Pro Git.”

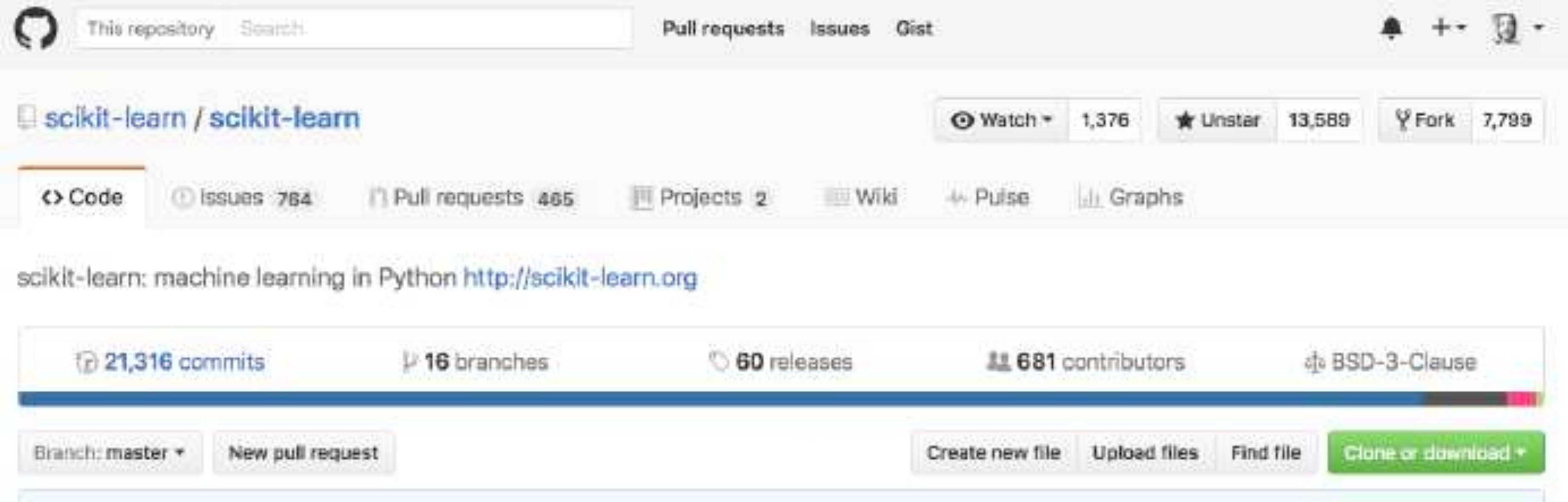
git



From: Scott Chacon. “Pro Git.”



from https://guides.github.com/introduction/flow/index.html

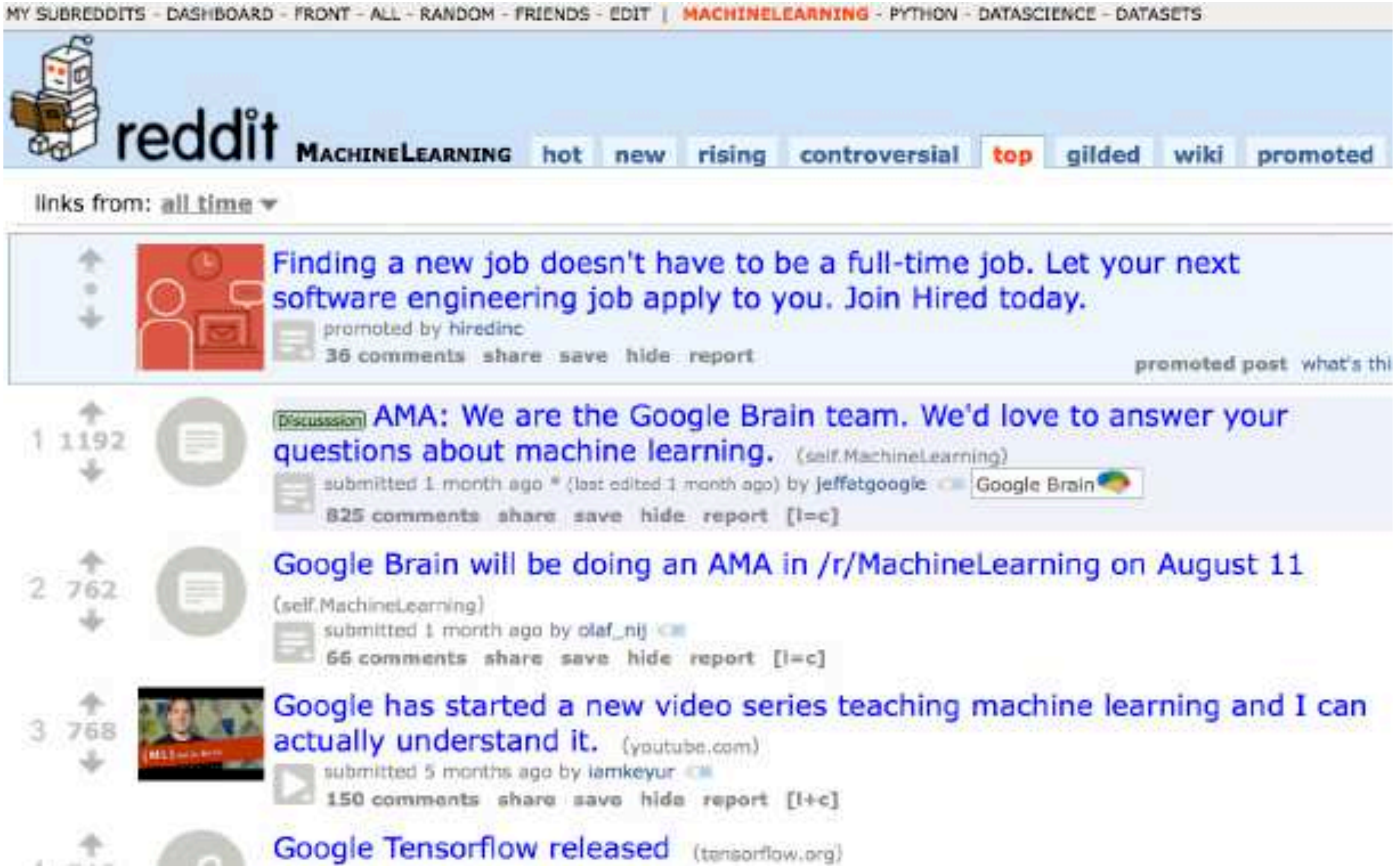


scikit-learn / scikit-learn

scikit-learn: machine learning in Python http://scikit-learn.org

21,316 commits 16 branches 60 releases 681 contributors BSD-3-Clause

Keeping Up to Date
&
Exchanging Ideas/Tips



MY SUBREDDITS - DASHBOARD - FRONT - ALL - RANDOM - FRIENDS - EDIT | MACHINELEARNING - PYTHON - DATASCIENCE - DATASETS
reddit
MachineLearning
hot
new
rising
controversial
top
gilded
wiki
promoted

links from: all time

Finding a new job doesn't have to be a full-time job. Let your next software engineering job apply to you. Join Hired today.

promoted by hiredinc

36 comments share save hide report

promoted post what's thi

1 1192 Discussion AMA: We are the Google Brain team. We'd love to answer your questions about machine learning. (self.MachineLearning)

submitted 1 month ago * (last edited 1 month ago) by jeffatgoogle Google Brain

825 comments share save hide report [l=c]

2 762 Google Brain will be doing an AMA in /r/MachineLearning on August 11 (self.MachineLearning)

submitted 1 month ago by olaf_nij

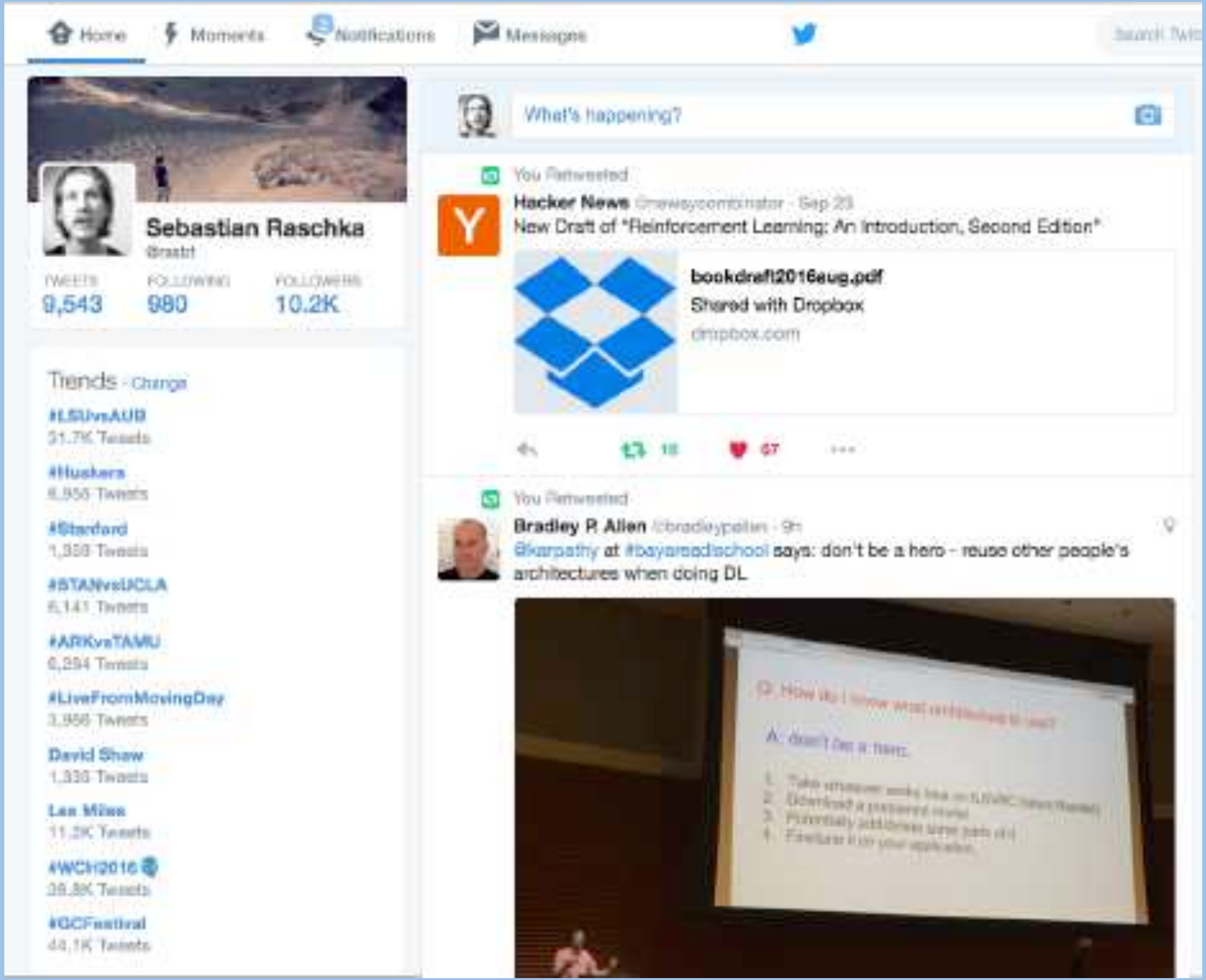
66 comments share save hide report [l=c]

3 768 Google has started a new video series teaching machine learning and I can actually understand it. (youtube.com)

submitted 5 months ago by iamkeyur

150 comments share save hide report [l+c]

Google Tensorflow released (tensorflow.org)



Sebastian Raschka
@rasbt

TWEETS	FOLLOWING	FOLLOWERS
9,543	980	10.2K

Trends · Change

#LSUvsAUB
31.7K Tweets

#Huskers
8,956 Tweets

#Stanford
1,359 Tweets

#STANvsUCLA
6,141 Tweets

#ARKvsTAMU
6,294 Tweets

#LiveFromMovingDay
3,956 Tweets

David Shaw
1,336 Tweets

Les Miles
11.2K Tweets

#WCH2016
39.8K Tweets

#GCFestival
44.1K Tweets

You Retweeted

Hacker News @newsycombinator · Sep 23

New Draft of "Reinforcement Learning: An Introduction, Second Edition"

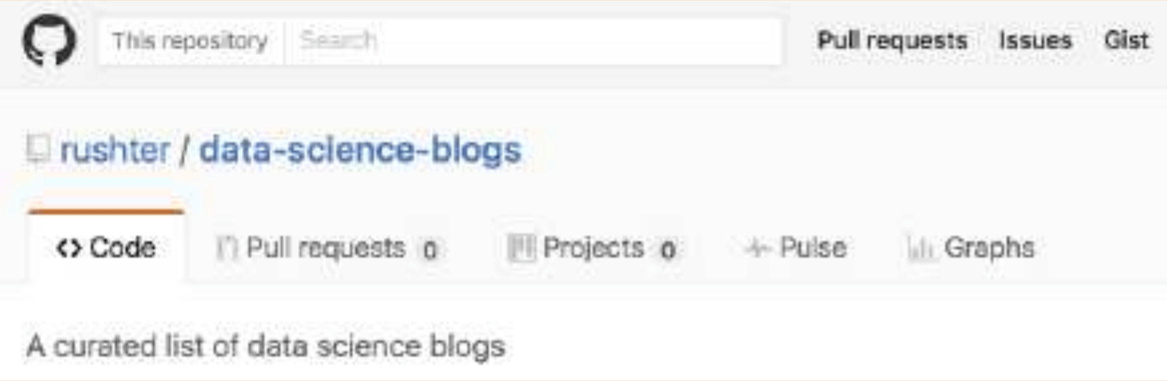
bookdraft2016aug.pdf
Shared with Dropbox
dropbox.com

18 67

You Retweeted

Bradley P. Allen @bradleypallen · 9h

@karpathy at #bayareadlschool says: don't be a hero - reuse other people's architectures when doing DL

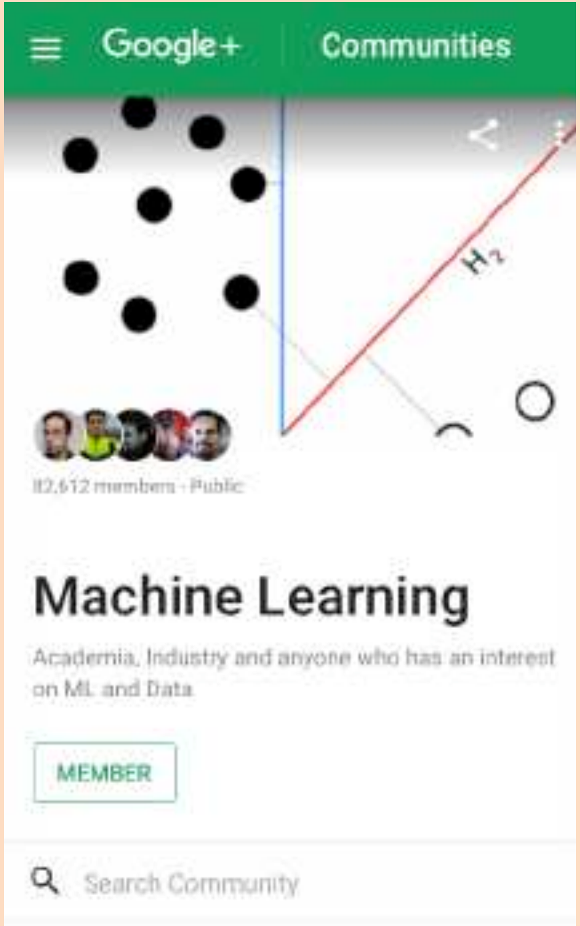


rushter / data-science-blogs

A curated list of data science blogs

https://github.com/rushter/data-science-blogs

https://news.ycombinator.com

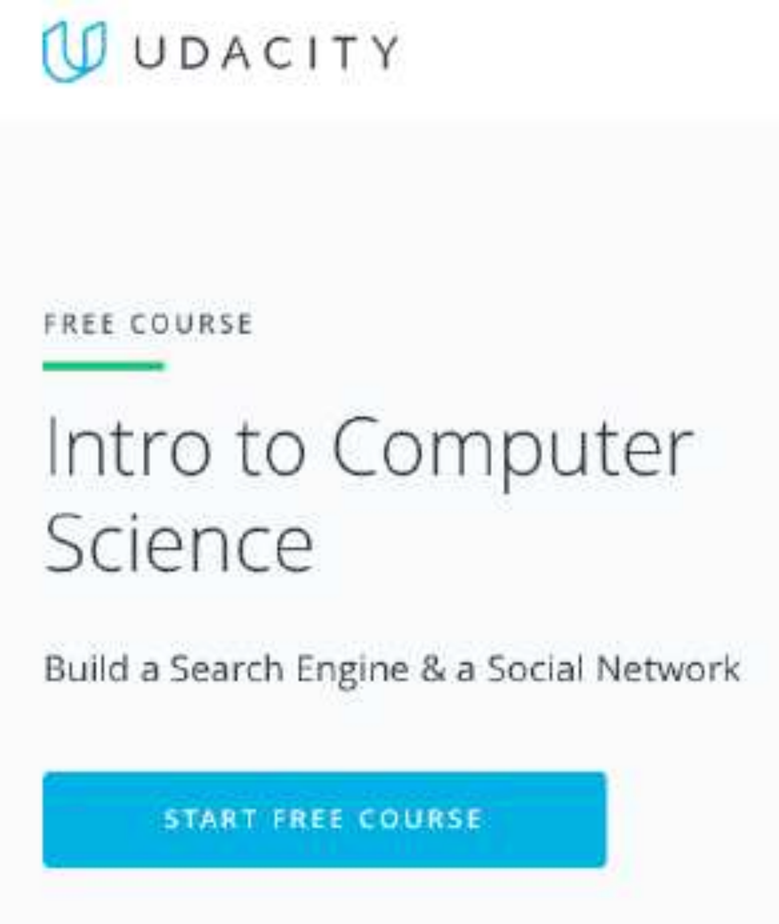


82,612 members · Public

Machine Learning

Academia, Industry and anyone who has an interest on ML and Data

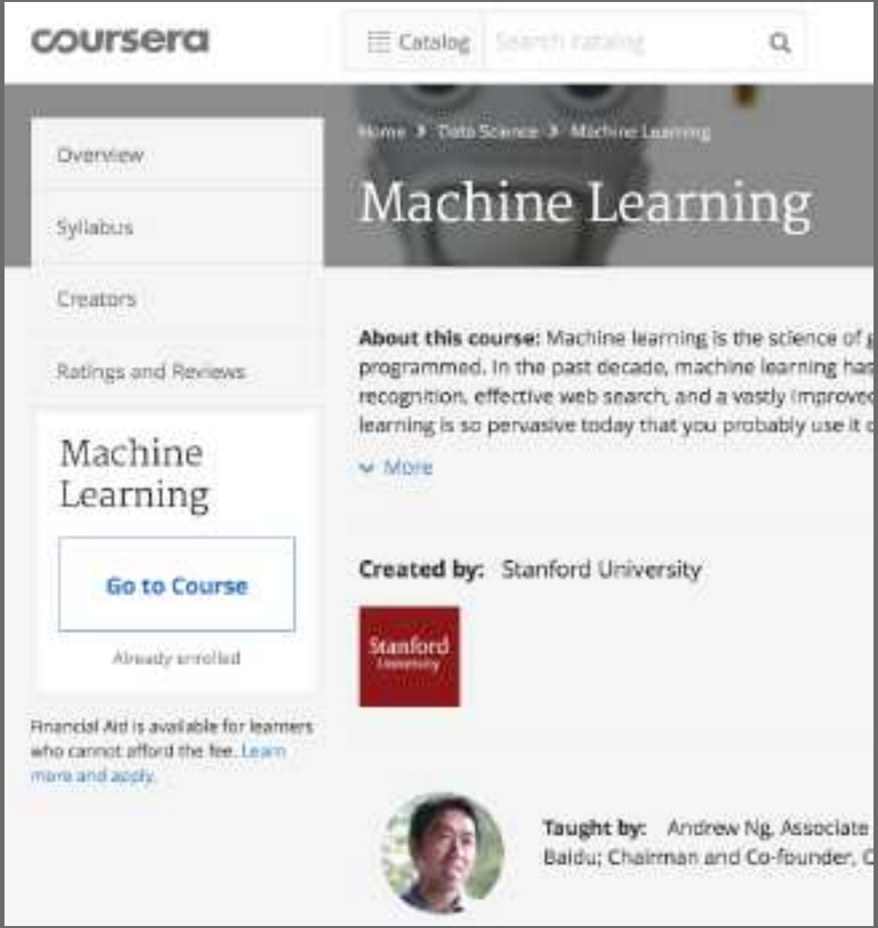
RESOURCES



FREE COURSE

Intro to Computer Science

Build a Search Engine & a Social Network



Machine Learning

About this course: Machine learning is the science of g
programmed. In the past decade, machine learning has
recognition, effective web search, and a vastly improved
learning is so pervasive today that you probably use it c

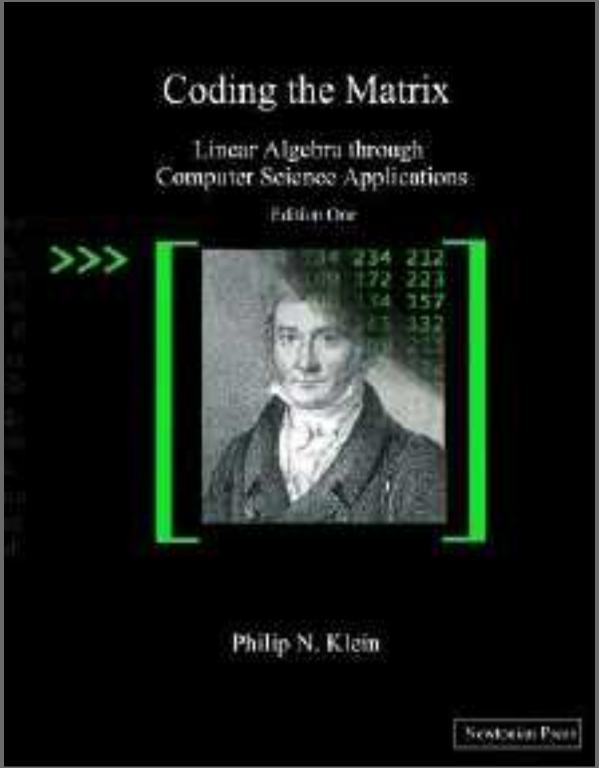
Machine Learning

Already enrolled

Financial Aid is available for learners who cannot afford the fee. Learn more and apply.

Created by: Stanford University

Taught by: Andrew Ng, Associate
Baidu; Chairman and Co-founder, C



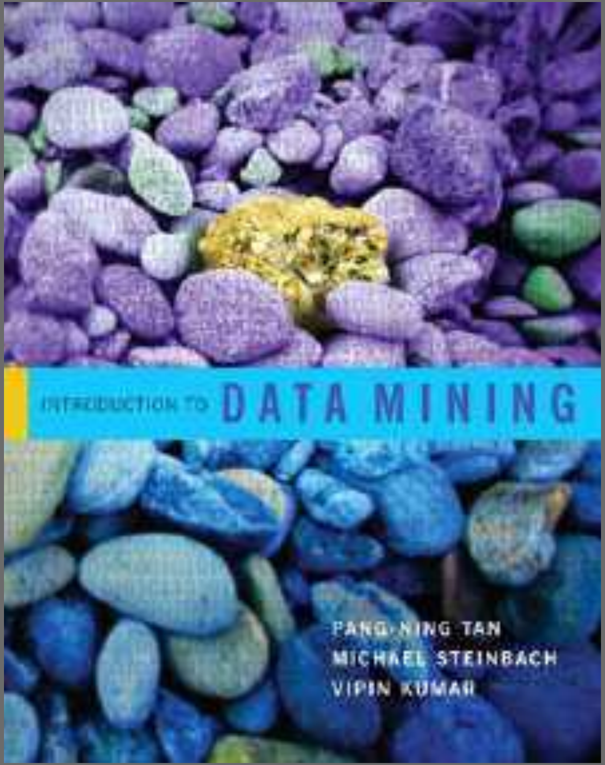
Coding the Matrix

Linear Algebra through
Computer Science Applications

Edition One

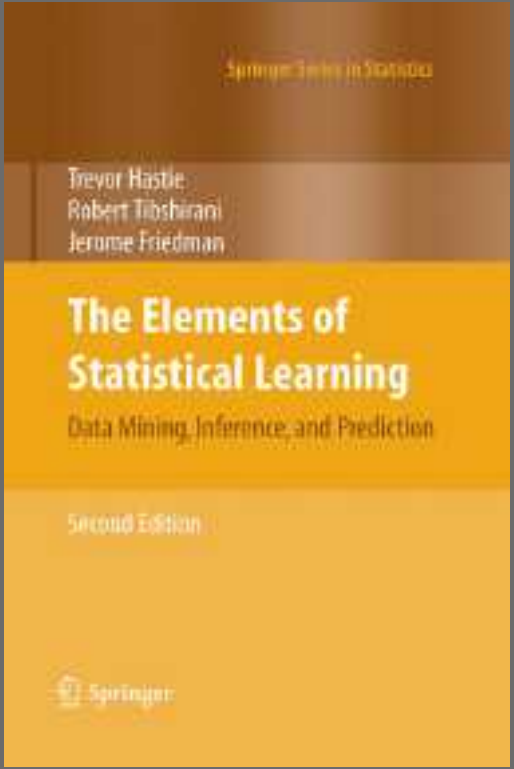
Philip N. Klein

Newtonian Press



INTRODUCTION TO DATA MINING

PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR



Springer Series in Statistics

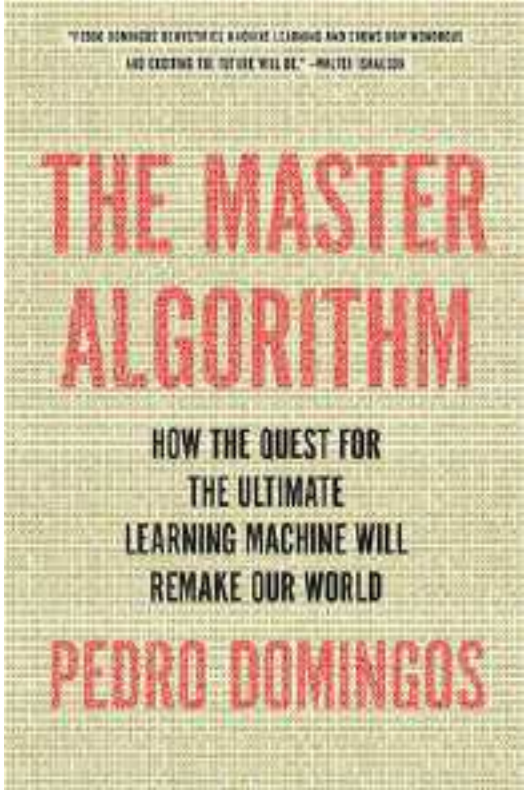
Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

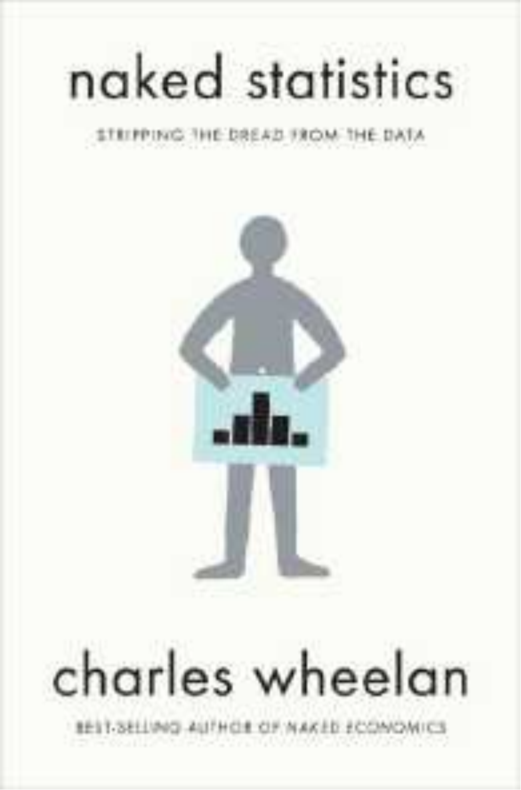
Data Mining, Inference, and Prediction

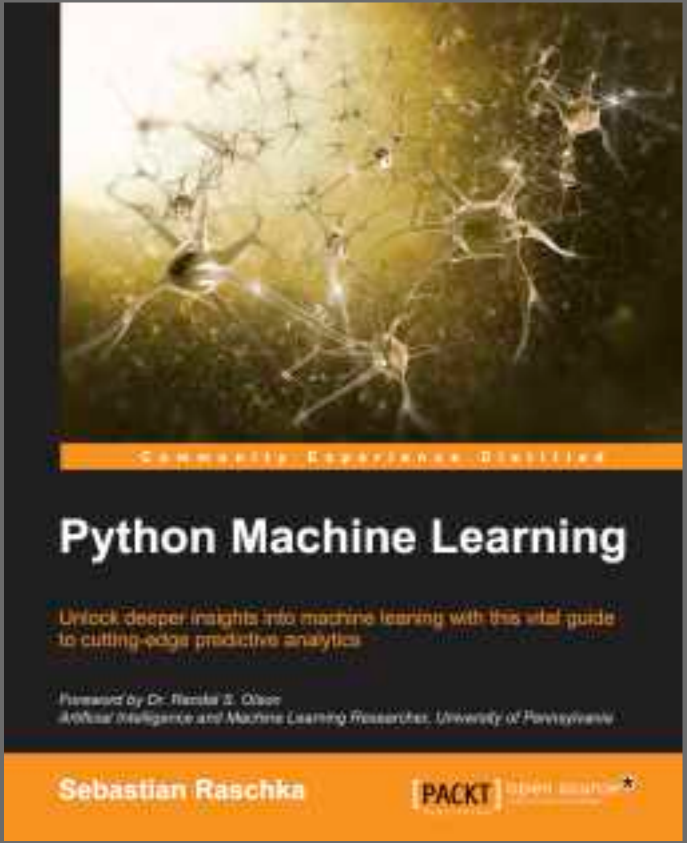
Second Edition

Springer


"PEDRO DOMINGOS DEMYSTIFIES MACHINE LEARNING AND SHOWS HOW WONDROUS AND EXCITING THE FUTURE WILL BE." —WALTER ISAACSON
THE MASTER ALGORITHM
HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD
PEDRO DOMINGOS

DATA SCIENTISTS AT WORK
SEBASTIAN GUTIERREZ
FOREWORD BY PETER NORVIG (GOOGLE)


naked statistics
STRIPPING THE DREAD FROM THE DATA
charles wheelan
BEST-SELLING AUTHOR OF NAKED ECONOMICS



Community Experience Distilled

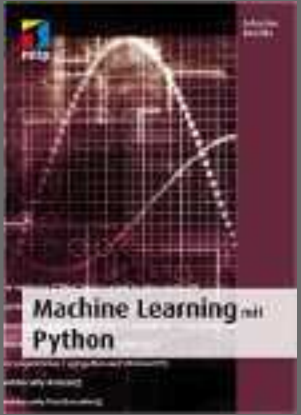
Python Machine Learning

Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics

Foreword by Dr. Randal S. Olson
Artificial Intelligence and Machine Learning Researcher, University of Pennsylvania

Sebastian Raschka

PACKT open source



Machine Learning mit Python

Python
Machine Learning

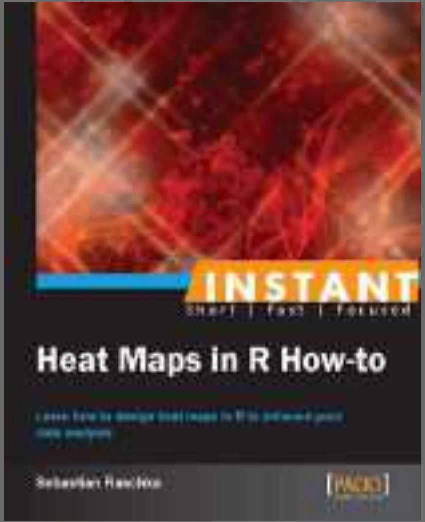
Packt

Python: Deeper Insights into Machine Learning

LEARNING PATH

Packt

Python
機械学習プログラミング



INSTANT
Short | Fast | Focused

Heat Maps in R How-to

Sebastian Raschka

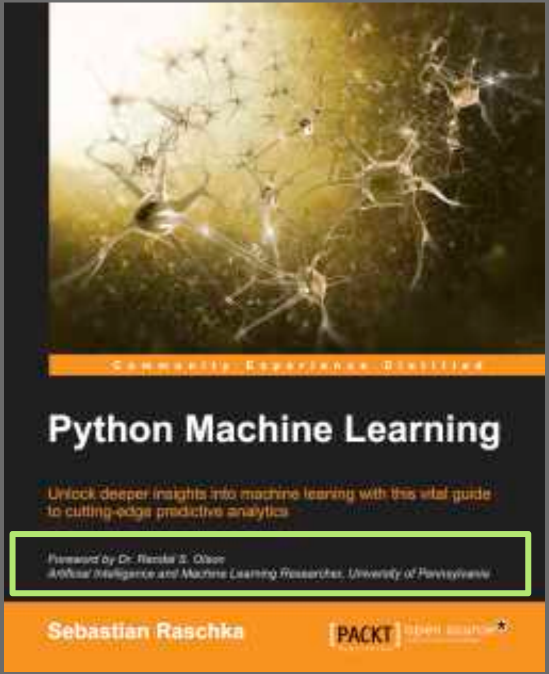
PACKT

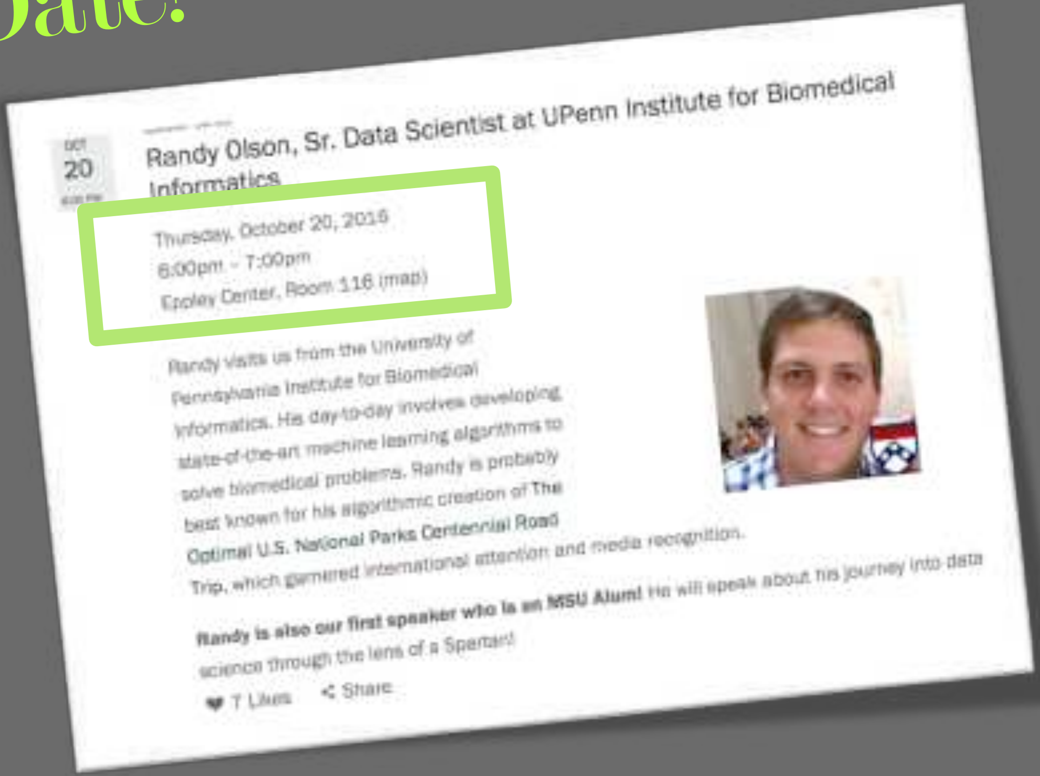
SEBASTIAN RASCHKA

Model Evaluation and Selection in Machine Learning

A Practical Guide with Applications in Python

Save the Date!


Community Experience Distilled
Python Machine Learning
Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics
Foreword by Dr. Randal S. Olson
Artificial Intelligence and Machine Learning Researcher, University of Pennsylvania
Sebastian Raschka
PACKT open source



OCT
20

Randy Olson, Sr. Data Scientist at UPenn Institute for Biomedical Informatics

Thursday, October 20, 2016
6:00pm - 7:00pm
Epply Center, Room 116 (map)

Randy visits us from the University of Pennsylvania Institute for Biomedical Informatics. His day-to-day involves developing state-of-the-art machine learning algorithms to solve biomedical problems. Randy is probably best known for his algorithmic creation of The Optimal U.S. National Parks Centennial Road Trip, which garnered international attention and media recognition.

Randy is also our first speaker who is an MSU Alum! He will speak about his journey into data science through the lens of a Spartan!

❤ 7 Likes   < Share

But the most important thing is to keep on learning. Not just for a few months, but for years. Every Saturday, you will have a choice between staying at home and reading research papers/implementing algorithms, vs. watching TV. If you spend all Saturday working, there probably won't be any short-term reward, and your current boss won't even know or say "nice work." Also, after that Saturday of hard work, you're not actually that much better at machine learning. But here's the secret: If you do this not just for one weekend, but instead study consistently for a year, then you will become very good.

There's a lot of demand today for ML people; once you get a job in ML, your learning will only accelerate further.

Andrew Ng, Chief Scientist at Baidu;
Chairman/Co-Founder of Coursera; Stanford faculty

https://www.quora.com/How-should-you-start-a-career-in-Machine-Learning/answer/Andrew-Ng https://github.com/rasbt

http://sebastianraschka.com

mail@sebastianraschka.com

@rasbt