Revamping Student Education with Real-World Data

Prof. Brian D. Davison

LEHIGH UNIVERSITY

Photo credit: Nikos Pappas, Yannis Voutsalas (CC BY 2.0)

Photo credit: Ecole Polytechnique Executive Education (CC BY 2.0)

Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J.

FROM THE OCTOBER 2012 ISSUE

Forbes

IBM Predicts Demand F

MAY 13, 2017

January 2016

Glassdoor Ranks the 25 Best Jobs in A

Of no surprise to us, "Data scientist" ranked #1 on Glassdoor's list of 25 b
on earnings potential, career opportunities, and potential career growth.
list of 25 jobs with the best work-life balance last October. (The Washing

OSIsoft. PI World SAN FRANCISCO 2018

50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? Find out how.

United States | 2017 | 8.1k Shares

5 Analytics Manager

4.6 / 5 Job Score

4.1 / 5 Job Satisfaction

$112,000 Median Base Salary

1,958 Job Openings

View Jobs

Universities are educating more and more data-savvy students

- MS degrees in data science…
- BS degrees…
- Every semester I teach a popular **Introduction to Data Science** course to **students from any discipline**

Image Credit: NASA/Ames/JPL-Caltech

- In all of these environments, students need to apply methods to data sets to learn, gain experience, and test intuitions

One machine learning training website says:

Why Do We Need Practice Datasets?

If you are interested in practicing applied machine learning, you need datasets on which to practice.

This problem can stop you dead.

- Which dataset should you use?
- Should you collect your own or use one off the shelf?
- Which one and why?

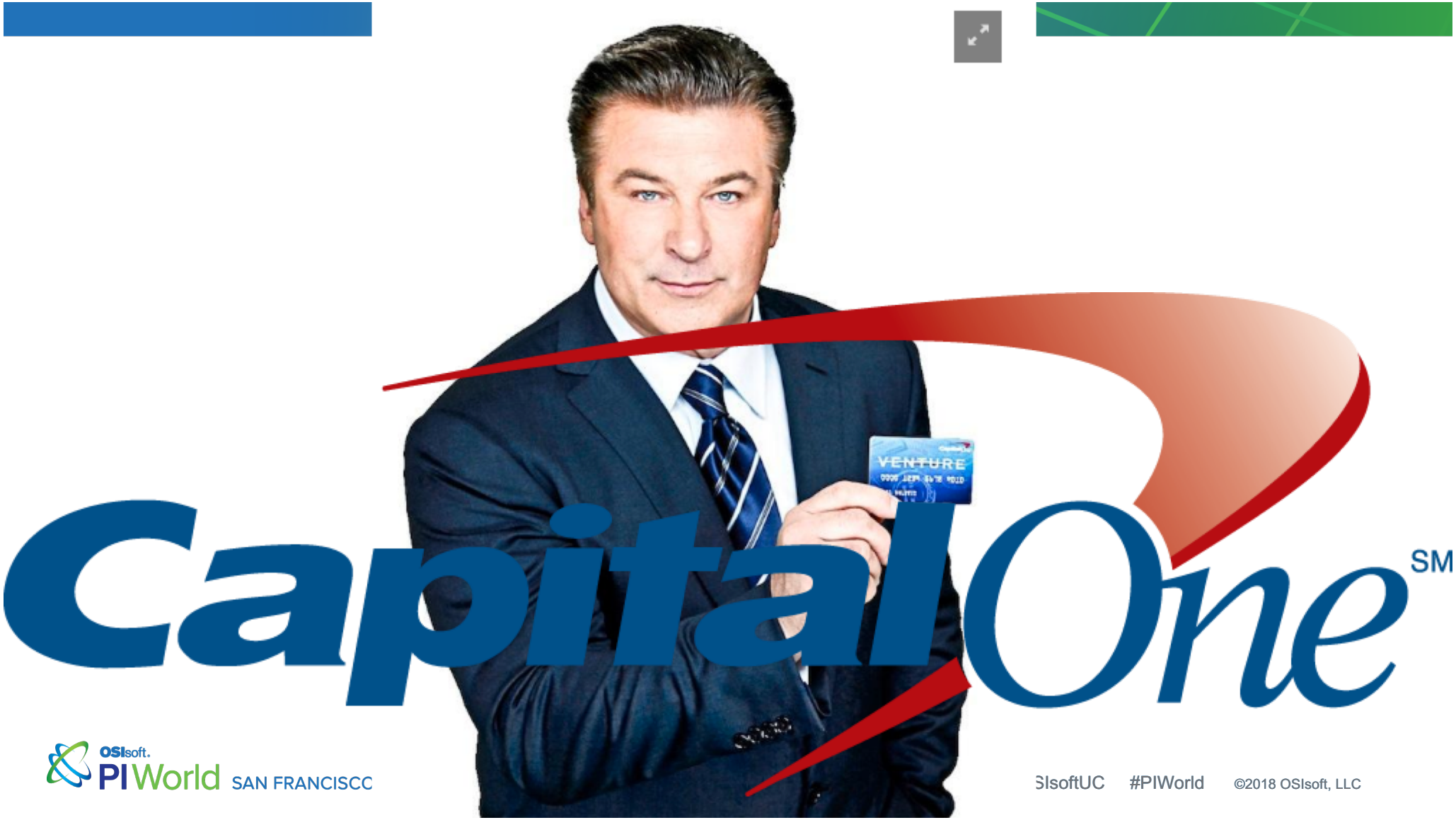
⋮

For beginners, you can get everything you need and more in terms of datasets to practice on from the UCI Machine Learning Repository.

Dataset	Data Point	Attribute	Class
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6*
WDBC	569	30	2
CMC	1473	9	3
Yeast	1484	8	10
Optical Digit	5620	64	10
Statlog	6435	36	6*
Thyroid	7200	21	3
Magic Gamma	19020	10	2

Why share real-world data with students?

- Improve student education!
  - Modern real-world problems are great motivators for students
  - Help them understand complex and dirty real-world data
- Educate students about your data science problems
  - Industry needs are often different from academic research
  - They will know what you care about, and can jump in when hired
- Increase visibility to students
  - Your brand becomes part of otherwise inaccessible conversations
- Indirectly educate students about your organization and needs
  - The best data science requires an understanding of the domain



kaggle

Sea

Sign In

18 Active Competitions

2018 Data Science Bowl
Find the nuclei in divergent images to advance medical discovery
Featured · 4 days to go · biology
$100,000
3,634 teams

TalkingData AdTracking Fraud Detection Challenge
Can you detect fraudulent click traffic for mobile app ads?
Featured · 25 days to go ·
$25,000
2,670 teams

CVPR 2018 WAD Video Segmentation Challenge
Can you segment each objects within image frames captured by vehicles?
Research · 2 months to go ·
$2,500
15 teams

iMaterialist Challenge (Fashion) at FGVC5
Image classification of fashion products.
Research · 2 months to go ·
$2,500
28 teams

iMaterialist Challenge (Furniture) at FGVC5
Image Classification of Furniture & Home Goods.
Research · 2 months to go ·
$2,500
177 teams

Google Landmark Retrieval Challenge
Given an image, can you find all of the same landmarks in a dataset?
Research · a month to go · image data
$2,500
113 teams

Google Landmark Recognition Challenge
Label famous (and not-so-famous) landmarks in images
Research · a month to go · image data
$2,500
231 teams

New to Data
Get started wit
our most popul
for beginners, Ti
Learning fro

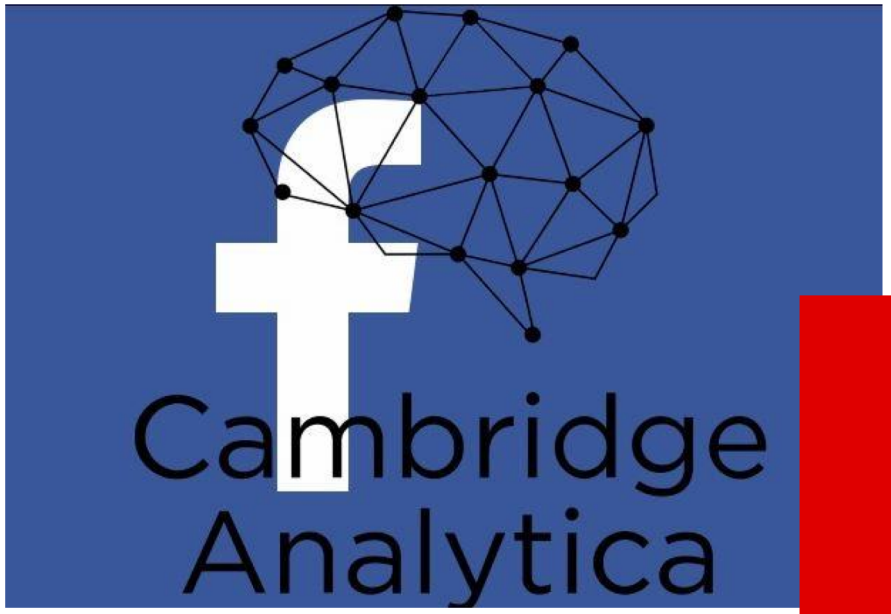
omission
diction file for
g & a spot on
rboard.

SIsoft, LLC

Why not share data?

- Your organization already has all the answers and staff it needs
- There is already lots of data out there
  - True! And it is growing! But process data is still pretty rare.
- Your competitor might learn something from your data
  - Not all data needs to expose intellectual property (e.g., pump runs)
- Concerns about privacy

Privacy concerns are real



How to share data?

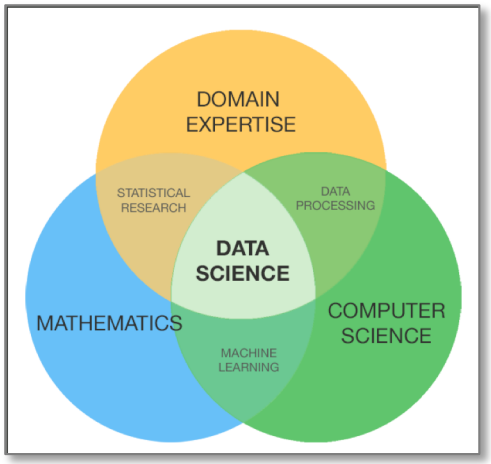
- Publicly through competitions
- Publicly via your website
- Publicly via data repositories or network
- Privately with a usage license
  - E.g., only educational or non-profit use
  - Perhaps with a data broker to enforce
- ~~Privately with an non-disclosure agreement~~

If you really want to impact students…

- Don’t just share data on the web (although that's a start)!
- Provide background into your industry, your organizational goals, and the problem domain
- Provide support for the data
  - Answer questions about the data and what kinds of analysis is valued
- Connect with the faculty
- Visit classes – either in person or via teleconference
- Sponsor (existing) hackathons that might use your data

Understanding the domain is essential

- How was the data generated or collected?
  - What do the data mean (i.e., how to interpret)?
  - What kinds of errors are likely to be present?
    - What are the sources of noise in the data sources or labels?
    - What does missing data mean?
  - What level of precision is important?
- What aspects of the data are already known (or believed) to be predictive of the desired target?
- What kinds of solutions are valuable to the organization?



A peek into my Thursday morning presentation:

Introduction to Time-Series Analysis with PI System and R

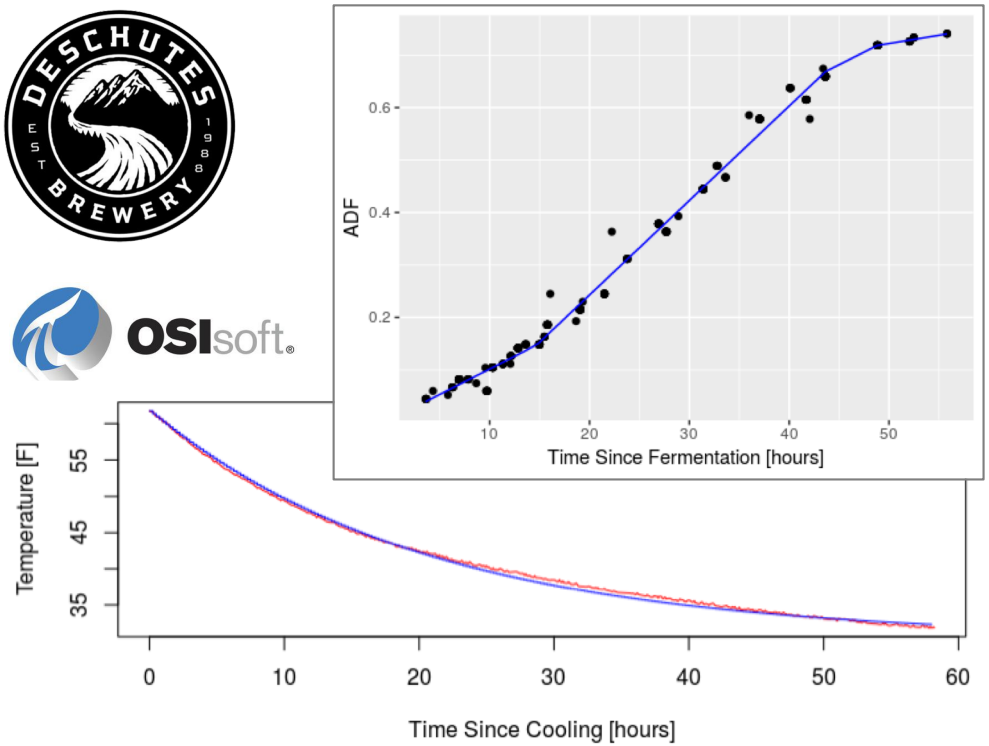


Image source: Wikipedia

Fermentation data was shared through and obtained from the OSIsoft Academic Community Service.

Sharing Data – Why and How?

Lehigh University (Bethlehem, PA) is considered one of the hidden ivies. Mission statement: *To advance learning through the integration of teaching, research, and service to others.*

CHALLENGE

Should organizations share data to help educate students?

- Concerns about exposing IP
- Concerns about privacy
- Already lots of data being shared

ADVANTAGES

Sharing data can impact students and benefit your organization

- Modern problems motivate students
- Students learn about issues and processes that you care about
- New venues to expose your brand

HOW

Many options for sharing are available

- PI data can be shared via OSIsoft Academic Community Service
- Share other data via data brokers, repositories, and web sites

Contact Information

Brian D. Davison

davison@cse.lehigh.edu

Associate Prof., Lehigh University

Feel free to share your data with me! ☺

Lehigh University's Linderman Library

Merci

谢谢

Danke

Спасибо

Gracias

Thank You

감사합니다

ありがとう

Grazie

Obrigado