

# Revamping Student Education with Real-World Data

**Prof. Brian D. Davison**















Photo credit: Nikos Pappas, Yannis Voutsalas (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 For  
MAY 13, 2017

January 2016

Glassdoor Ranks the 25 Best Jobs in America

Of no surprise to us, "Data scientist" ranked #1 on Glassdoor's list of 25 best jobs in America based on earnings potential, career opportunities, and potential career growth. It also ranked #1 on our list of 25 jobs with the best work-life balance last October. (The Washington Post)

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



Image Credit: NASA/Ames/JPL-Caltech



UCI Machine Learning Repository  
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Google Custom Search Search

View ALL Data Sets

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 332 data sets as a service to the machine learning community. You may view all data sets through our searchable interface. Our old web site is still available, for those who prefer the old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians. We have also set up a mirror site for the Repository.

Supported By: In Collaboration With:

**Latest News:**

2013-04-04: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!

2010-03-01: Note from donor regarding Netflix data

2009-10-16: Two new data sets have been added.

2009-09-14: Several data sets have been added.

2008-07-23: Repository mirror has been set up.

2008-03-24: New data sets have been added!

2007-06-25: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

**Featured Data Set: Molecular Biology (Protein Secondary Structure)**

Task: Classification  
Data Type: Sequential  
# Instances: 128

From CMU connectionist bench repository; Classifies secondary structure of certain globular proteins

**Newest Data Sets:**

2015-08-04: [Mice Protein Expression](#)

2015-07-29: [Smartphone-Based Recognition of Human Activities and Postural Transitions](#)

2015-07-27: [Cuff-Less Blood Pressure Estimation](#)

2015-07-11: [Taxi Service Trajectory - Prediction Challenge, ECML PKDD 2015](#)

2015-07-05: [Folio](#)

2015-07-03: [Chronic Kidney Disease](#)

2015-06-06: [Machine Learning based ZAlpha Ltd. Stock Recommendations 2012-2014](#)

2015-05-31: [Online News Popularity](#)

2015-05-30: [Sentiment Labelled Sentences](#)

2015-05-25: [Forest type mapping](#)

2015-05-19: [Online Video Characteristics and Transcoding Time Dataset](#)

2015-05-04: [wiki4HE](#)

**Most Popular Data Sets (hits since 2007):**

759694: Iris

533604: Adult

443222: Wine

373775: Car Evaluation

352061: Breast Cancer Wisconsin (Diagnostic)

293690: Abalone

256377: Wine Quality

247141: Heart Disease

237122: Poker Hand

211967: [Human Activity Recognition Using Smartphones](#)

204720: Forest Fires

187997: Internet Advertisements

About || Citation Policy || Donation Policy || Contact || CML

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 of 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





# Capital One<sup>SM</sup>

 OSIsoft.  
**PIWorld** SAN FRANCISCO

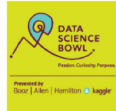
SIsoftUC #PIWorld ©2018 OSIsoft, LLC

NET

kaggle

Search

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

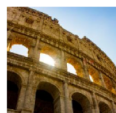


Google Landmark Retrieval Challenge

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

Sign In

New to Data  
Get started with  
our most popular  
for beginners, **Time Series**  
Learning from



omission  
diction file for  
g & a spot on  
rboard.

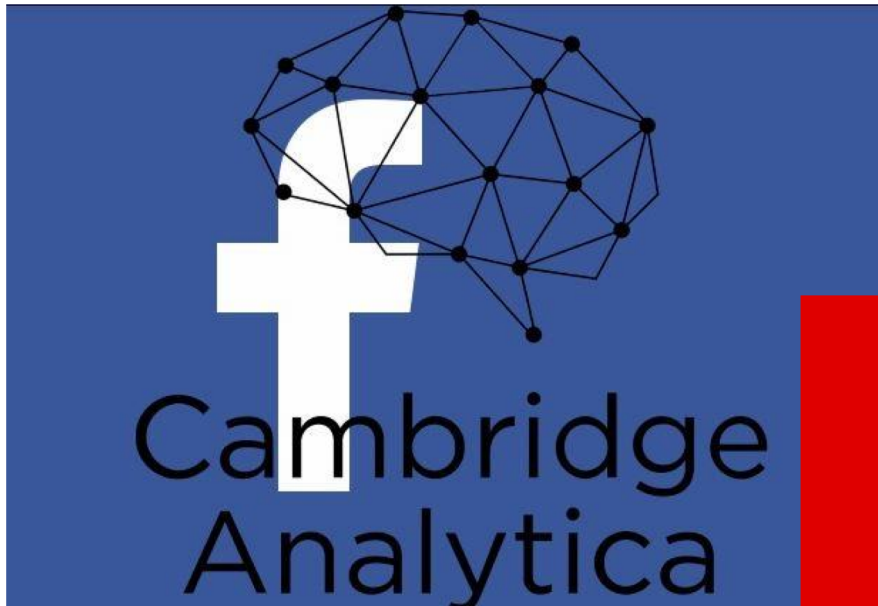
Sisoft, LLC



# 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 a non-disclosure agreement~~



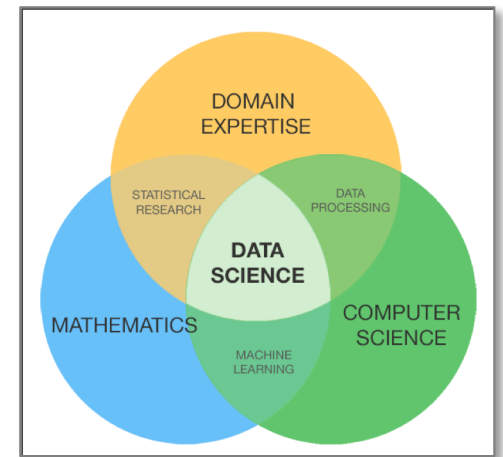
# If you really want to impact students...

- 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

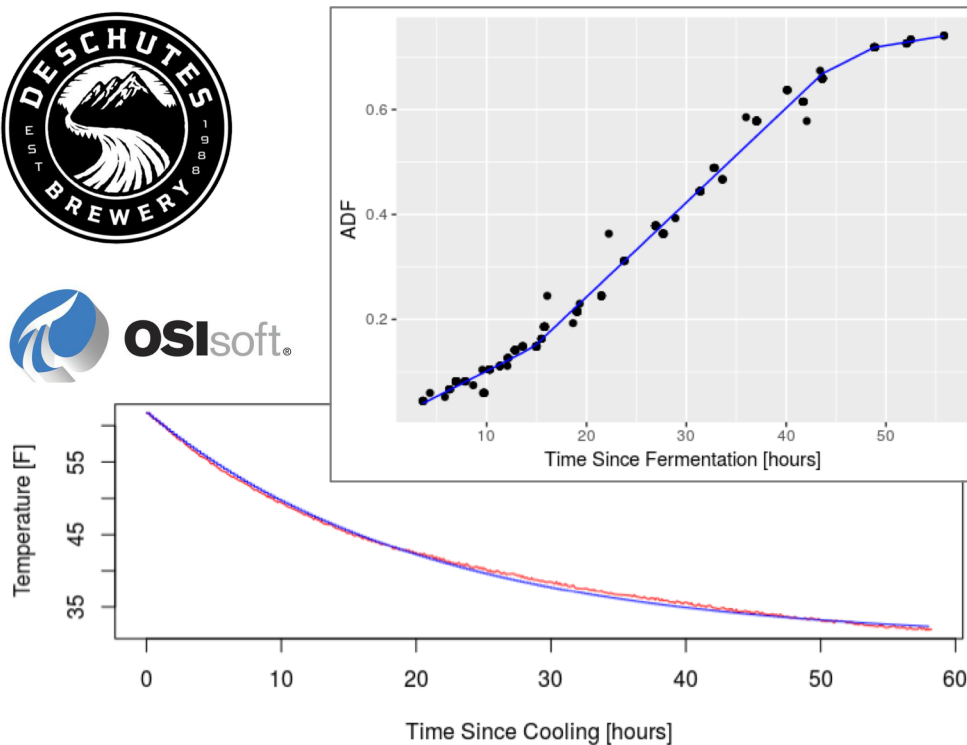


Image source: Wikipedia

*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 OSIssoft 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

Спасибо

Danke

Gracias

Thank You

감사합니다

ありがとう

Grazie

Obrigado