

City University of New York (CUNY)

## CUNY Academic Works

---

Open Educational Resources

City College of New York

---

2018

### Intro to Data Science - Course Intro (Week One)

Grant Long  
*CUNY City College*

NYC Tech-in-Residence

[How does access to this work benefit you? Let us know!](#)

More information about this work at: [https://academicworks.cuny.edu/cc\\_oers/246](https://academicworks.cuny.edu/cc_oers/246)

Discover additional works at: <https://academicworks.cuny.edu>

---

This work is made publicly available by the City University of New York (CUNY).  
Contact: [AcademicWorks@cuny.edu](mailto:AcademicWorks@cuny.edu)

What is Data Science?



# Harvard Business Review



DATA

## Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

**W**hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early." Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see possibilities. He began

Is that helpful?



People you may know



**Soumith Chintala**  
Artificial Intelligence  
Research Engineer at  
New York University

Connect



**Jeonggon Mun**  
CEO at BlockBank  
Shangxuan Hou

Connect



**Michael Strain**  
Director of Economic Policy  
Studies, American Enterprise  
Federal Reserve Bank  
of New York

Connect

*Goldman started to test what would happen if you presented users with names of people they hadn't yet connected with but seemed likely to know—for example, people who had shared their tenures at schools and workplaces.*



*We're living through a golden age of behavioral research. It's amazing how much we can figure out about how people think now.*



## What's Even Creepier Than Target Guessing That You're Pregnant?

By Jordan Ellenberg

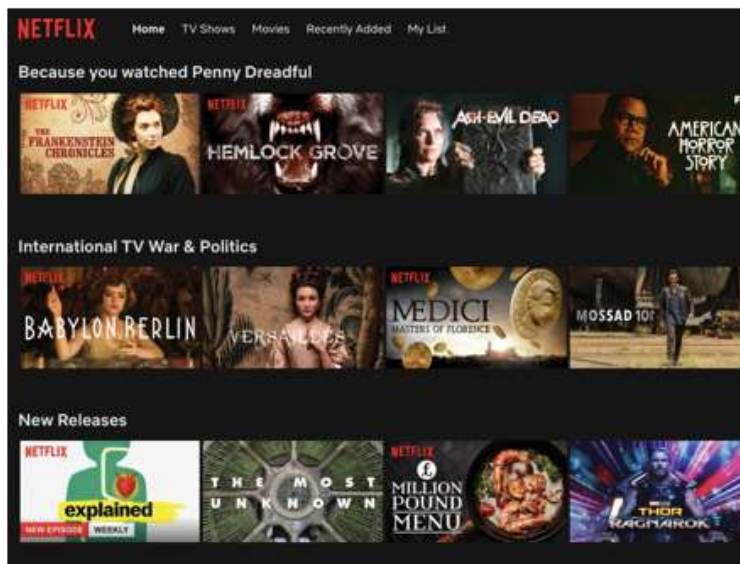


It can be spooky to contemplate living in a world where Google and Facebook and even Target know more about you than your parents do.

Photo illustration by James Emmesman. Photo Courtesy of Shutterstock.

*Here, the mistakes have real consequences. It's creepy and bad when Target intuits that you're pregnant. But it's even creepier and worse if you're not pregnant—or a terrorist, or a deadbeat dad—and an algorithm, doing its business in a closed and opaque box, decides that you are.*

# NETFLIX



*Hastings determined that a 10% improvement to the Cinematch algorithm would decrease customer churn and increase annual revenue by up to \$89 million. However, traditional options for improving the algorithm, such as hiring and training new employees, were time intensive and costly. Hastings decided to improve Netflix's software by crowdsourcing, and began planning the Netflix Prize, an open contest searching for a 10% improvement on Cinematch.*

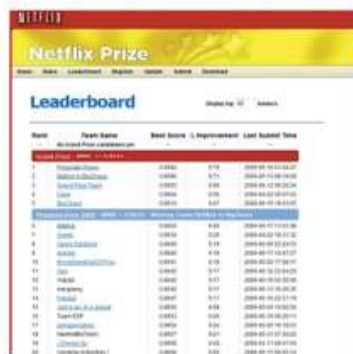


# NETFLIX

CASEY JOHNSTON, ARS TECHNICA BUSINESS 04.16.12 08:20 AM

## NETFLIX NEVER USED ITS \$1 MILLION ALGORITHM DUE TO ENGINEERING COSTS

*We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment. Also, our focus on improving Netflix personalization had shifted to the next level by then.*



Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	Golden Thread	0.8166	0.74	2006-09-10 14:33:27
2	Netflix Prize Team	0.8166	0.74	2006-09-12 16:24:24
3	Team	0.8166	0.74	2006-09-10 10:00:00
4	Netflix	0.8166	0.74	2006-09-10 10:00:00
5	Netflix	0.8166	0.74	2006-09-10 10:00:00
6	Netflix	0.8166	0.74	2006-09-10 10:00:00
7	Netflix	0.8166	0.74	2006-09-10 10:00:00
8	Netflix	0.8166	0.74	2006-09-10 10:00:00
9	Netflix	0.8166	0.74	2006-09-10 10:00:00
10	Netflix	0.8166	0.74	2006-09-10 10:00:00
11	Netflix	0.8166	0.74	2006-09-10 10:00:00
12	Netflix	0.8166	0.74	2006-09-10 10:00:00
13	Netflix	0.8166	0.74	2006-09-10 10:00:00
14	Netflix	0.8166	0.74	2006-09-10 10:00:00
15	Netflix	0.8166	0.74	2006-09-10 10:00:00
16	Netflix	0.8166	0.74	2006-09-10 10:00:00
17	Netflix	0.8166	0.74	2006-09-10 10:00:00
18	Netflix	0.8166	0.74	2006-09-10 10:00:00
19	Netflix	0.8166	0.74	2006-09-10 10:00:00
20	Netflix	0.8166	0.74	2006-09-10 10:00:00

Who are Data Scientists?



Lyft Engineering

[Follow](#)



[HOME](#) [ENGINEERING](#) [DATA SCIENCE](#) [SECURITY](#) [MOBILE](#)



Nicholas Chamandy

[Follow](#)

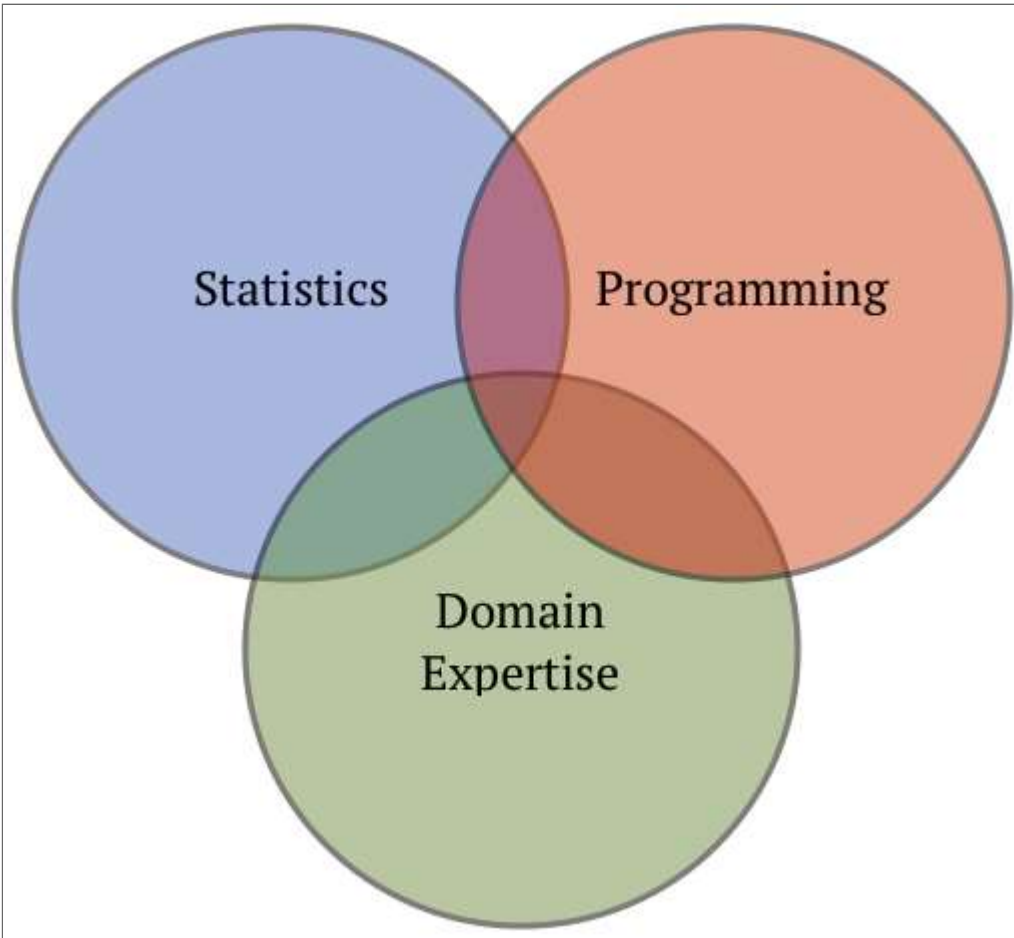
Scientific Director at Lyft

Apr 18 · 5 min read

## What's in a name?

The semantics of Science at Lyft

At Lyft, we're rebranding our Data Analyst function as *Data Scientist*, and our Data Scientist function as *Research Scientist*. In this short post, we describe the reasoning behind the change, which we believe will set Lyft up to make better decisions and build better products as we scale.



# About Me









# About Tech in Residence



**[ CUNY ]**  
**[ 2X TECH ]**



About this Course

**[grantmlong.com/teaching](http://grantmlong.com/teaching)**

## Official Course Objectives

1. Explain the key steps in a data science project.
2. Apply Python to load, clean, and process data sets.
3. Identify key elements of and patterns in a data set using computational analysis and statistical methods.
4. Explain and visualize empirical findings using with Python and other resources.
5. Explain fundamental principles of machine learning.
6. Apply predictive algorithms to a data set.
7. Work effectively in a team dedicated to analyzing data.

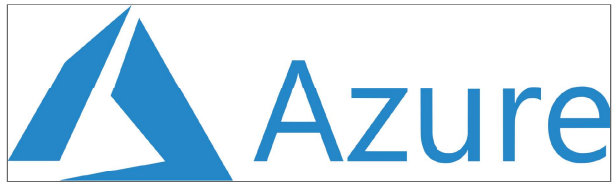
## Why Take this Course?

1. Careers in data are abundant, lucrative, and rewarding.
2. Learn how to detect BS.
3. Be a more informed person.

## Resources: Coding



## Resources: Notebooks





Resources: Class Communications

Course Page

**Blackboard<sup>®</sup>**

## How to Get Help



How to Get Help

**itds.ccny@gmail.com**

# How to Get Help

Classmates

## Grading

Project	40%
Assignments & Quizzes	30%
Midterm Exam	20%
Class Participation	10%

## Project

The bulk of the course grade will be a group project that will be due in December (exact date TBD). Students will be expected to work on the project during the second half of the class and will be required to present their progress throughout the course of the semester. Grades will be assigned on the basis of overall project quality, demonstration of core principles taught in the class, and individual contributions to the group's effort. More details on the project will be discussed in the second week of class.

## Assignments and Exams

- **Assignments.** This class includes short, frequent assignments to check comprehension. All assignments and quizzes will be graded on a 5-point scale. All quizzes will be announced in advance of class.
  - **No late assignments accepted.** Assignments not turned in by the set deadline will be scored as 0/5. Exceptions will be granted only as mandated by CUNY policy.
  - **Worst two assignments dropped,** includes missed assignments.
- **Exam.** A short midterm exam will be held in October and will focus on broad concepts the course has surveyed thus far. The format will mimic the style of questions frequently asked in interviews for data-related roles.

## Texts and Materials

- **Required Text:** *Data Science from Scratch*, Joel Grus. 2nd Edition, April 2015 (O'Reilly). Available **online**.
- **Additional required readings and videos** will be made available to students in advance of each week's assignments. All will be available online at no cost.
- In addition to the required materials, students may find the following resources helpful in supplementing course materials:
  - **Recommended Text:** *Python for Data Analysis*, Wes McKinney. 2nd Edition, October 2017 (O'Reilly). Available **online**.
  - **Recommended Text:** *Elements of Statistical Learning*, Trevor Hastie, Robert Tibshirani and Jerome Friedman. 2nd Edition, 2009 (Springer). Available free online **here**.

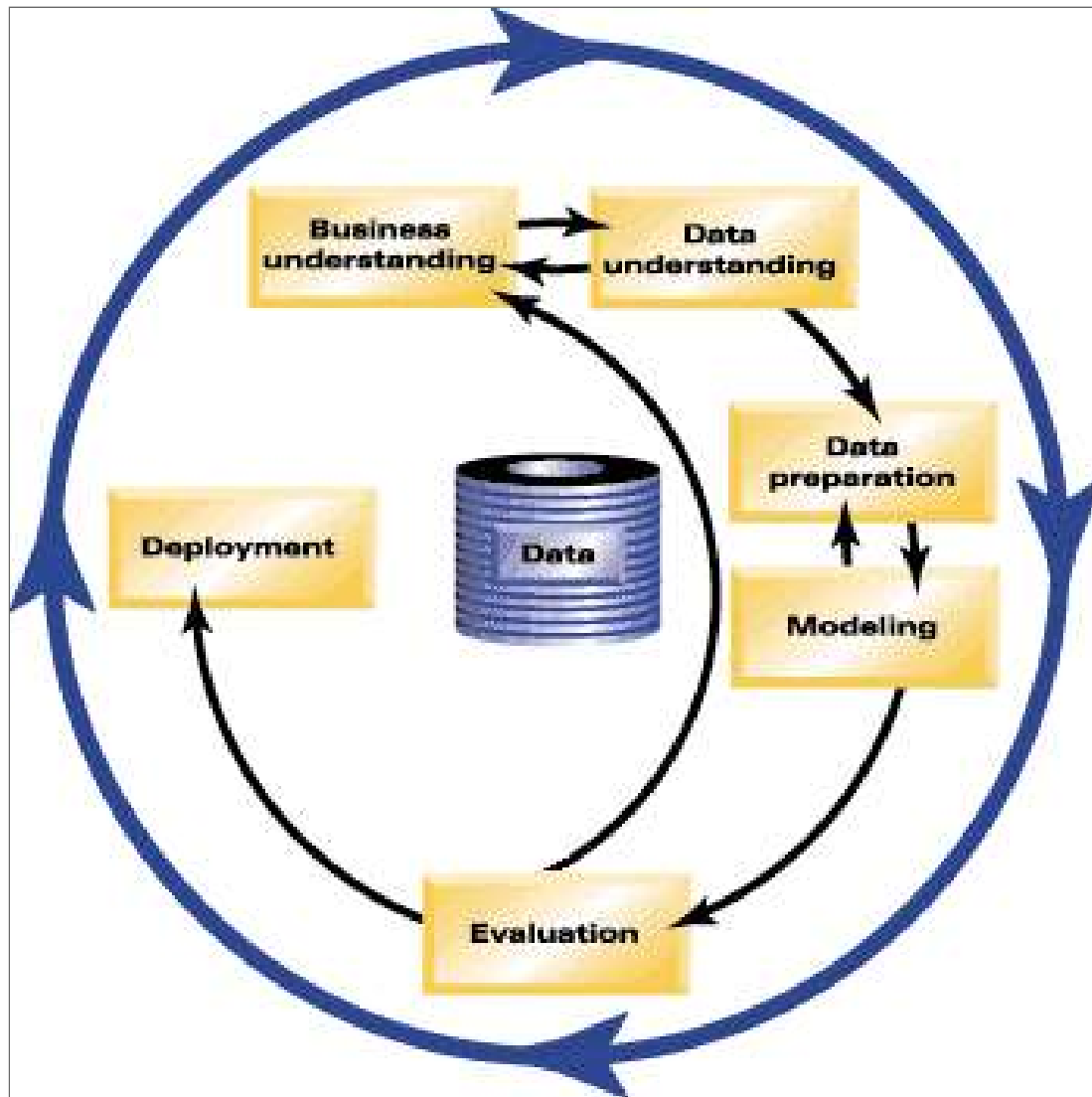


## Cheating

*Academic dishonesty is prohibited in The City University of New York. Penalties for academic dishonesty include academic sanctions, such as failing or otherwise reduced grades, and/or disciplinary sanctions, including suspension or expulsion.*

**CUNY Policy on Academic Integrity.**

# Data Science in Practice



# Today's Data

# H1-B Visa Data



**UNITED STATES DEPARTMENT OF LABOR**  
Employment & Training Administration

A to Z | Site Map | FAQs | Forms | About DOL | Contact Us | Español

Enter Search Term

[ETA Home](#)

[Find Job & Career Info](#)

[Business & Industry](#)

[Workforce Professionals](#)

[Grants & Contracts](#)

[TAA Program](#)

[Foreign Labor Certification](#)

[Performance & Results](#)

[Regions & States](#)

[ETA Home](#) > [Foreign Labor Certification](#) > [OFLC Performance Data](#)

**Foreign Labor Certification**  
*Helping U.S. employers fill jobs while protecting U.S. and foreign workers*

**OFLC Performance Data**

In carrying out its responsibility for the processing of labor certification and labor attestation applications, the Office of Foreign Labor Certification (OFLC) generates program data that is essential both for internal assessment of program effectiveness and for providing the Department's external stakeholders with useful information about the immigration programs administered by OFLC. In line with the Department's commitment to the Open Government initiative and specific regulatory disclosure requirements, this page includes program information organized in three main categories: 1. OFLC's annual reports, providing a cumulative overview of program information and data; 2. Selected Statistics by Program providing cumulative quarterly excerpts of program information by major immigration program to provide snapshot views of the OFLC programs; and 3. Cumulative quarterly and annual releases of program disclosure data to assist with external research and program evaluation.

Annual Report Performance Data

Annual Performance Reports

Selected Statistics by Program

Disclosure Data

This page allows the public to access the latest quarterly and annual disclosure data in easily accessible formats for the purpose of performing in-depth longitudinal research and analysis. OFLC case disclosure data is available for download by the federal fiscal year cycle covering the October 1 through September 30 period (all disclosure data sets are saved in the Microsoft Excel (.xls) file format). Each data set is cumulative, containing unique records identified by the applicable OFLC case number, and any noticeable typographical or other data anomalies may be due to internal data entry or other external customer errors in completing the application form.

Select data fields for each case record are extracted from foreign labor certification application tables within OFLC case management systems based on the most recent date a determination decision was issued. Please refer the File Structure document for each field title and respective field description.

Please refer to our [Disclosure Data User Guide](#) that provides an example of how, using Microsoft Excel, the data can be filtered or sorted to provide relevant information specific to your needs.



# Homework

- No class next Monday, September 4.
- **First assignment (due Tuesday, September 4 at 11:59pm):**
  - Email **itds.ccny@gmail.com** with:
    1. Two original insights *that we did not discuss in class* from our H1B data dive.
    2. How you prefer to be addressed in class (name, pronouns).
    3. The email you prefer to correspond in with the class.
    4. Your GitHub handle. (Sign up for one if you do not already have it, a free account is fine.)
    5. The top **three** things you hope to get out of this class. (No wrong answers)