Introduction to Data Science

CSC 405/605

UNC GREENSBORO

Action Items

- Sign Up for Discord (See Syllabus on Canvas)
  - Complete form on announcement page
  - Form Groups on discord don’t wait
  - Otherwise random assignment
  - Email List of group (see lecture 01 notes)
- Will add GitHub IDs tomorrow morning to course repo
  - Can access course notes there
  - Homework release next wednesday
- Make a GitHub Repo
  - Follow instructions in syllabus (or see lecture 01 notes)
  - Sunday evening is the next time I’ll add GitHub IDs

Data Science

- What is Data Science?
  - Interdisciplinary field that uses tools, techniques, and science to make predictions or answers questions from data
- Involves:
  - Data curation
  - Data cleaning
  - Data Analysis
  - Fundamental Research
  - Machine learning
  - Deep Learning
  - Web Scraping
  - Statistics
  - Visualization
  - Information Privacy
  - …

Data Science

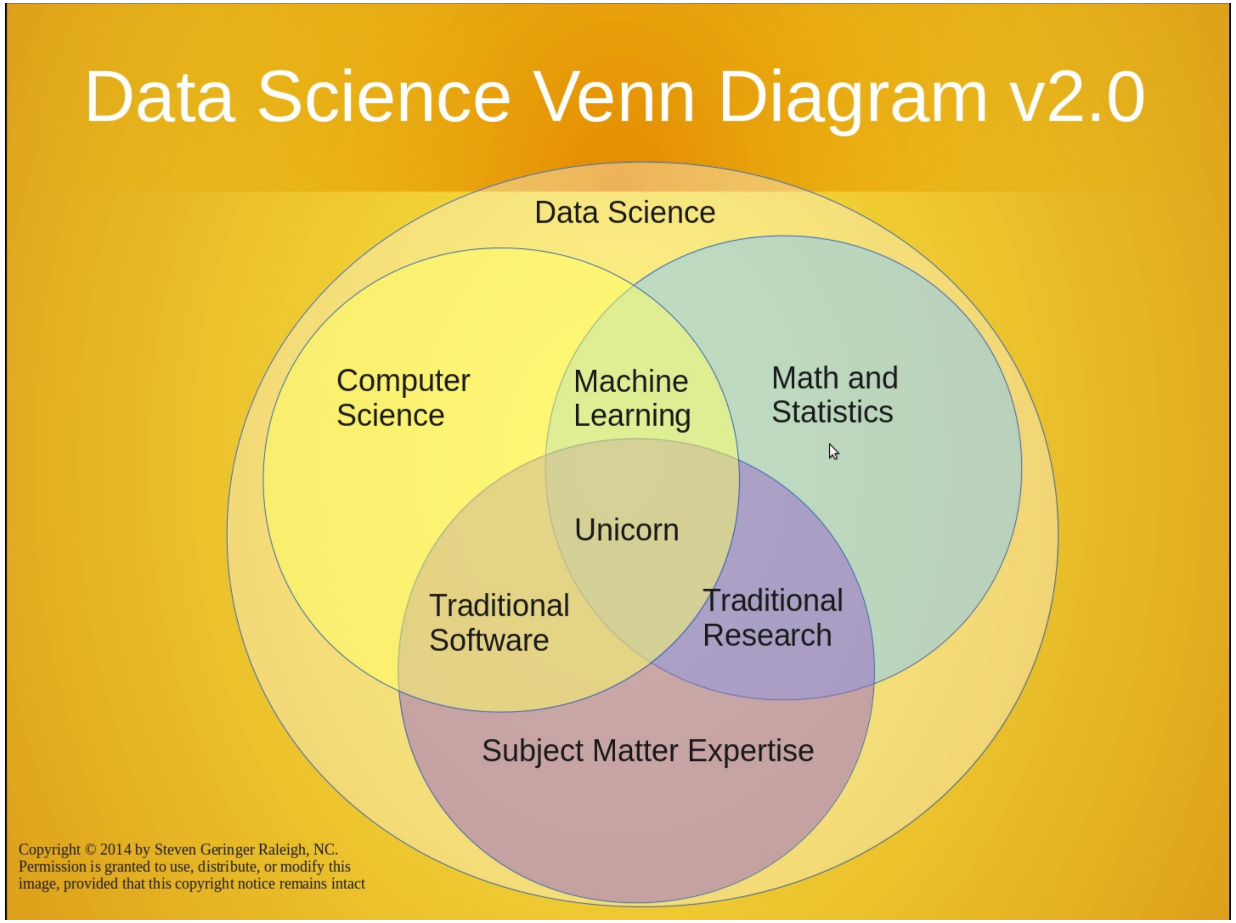
- Purpose of Data Science:
    - To find patterns
- Understanding patterns means understanding the world
    - Mechanic fixing a car
        - What is the problem?
        - Does it happen while stationary?
        - Does it happen when you accelerate? Brake?
        - ….
    - Narrows down problem based on observed patterns

Data Science

- Purpose of Data Science:
  - To find patterns
- Understanding patterns means understanding the world
  - Scientist making a research breakthrough?
- Starts with identifying a pattern
- Data Science identifies patterns to make predictions and inferences on data

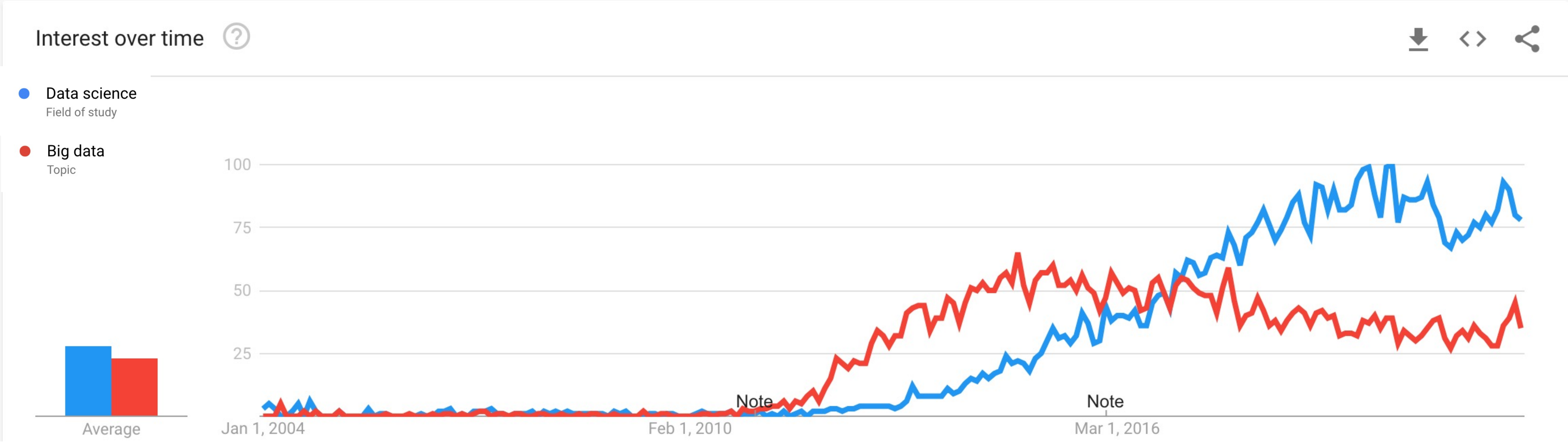
Data Science

- More an art than science
- Core
  - Subject Matter
  - Computer Science
  - Math and Statistics
- Different ratios of core areas used for different applications
  - Forecasting Demand of Sales
  - Classifying people in Images
  - Self Driving AI



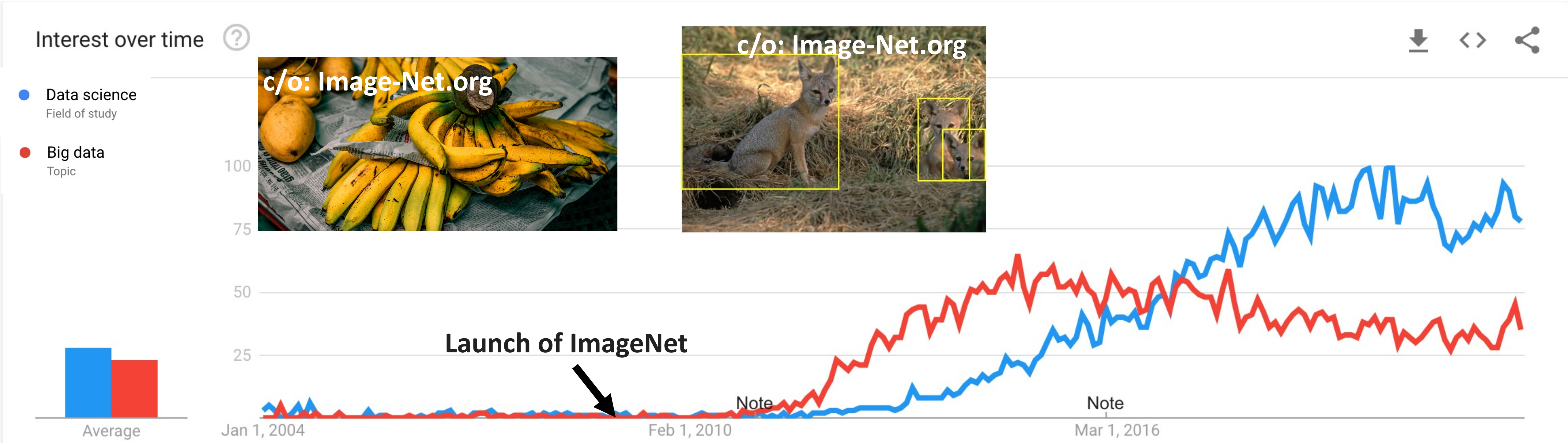
Data Science

- Why is Data Science exciting? – Google Trends



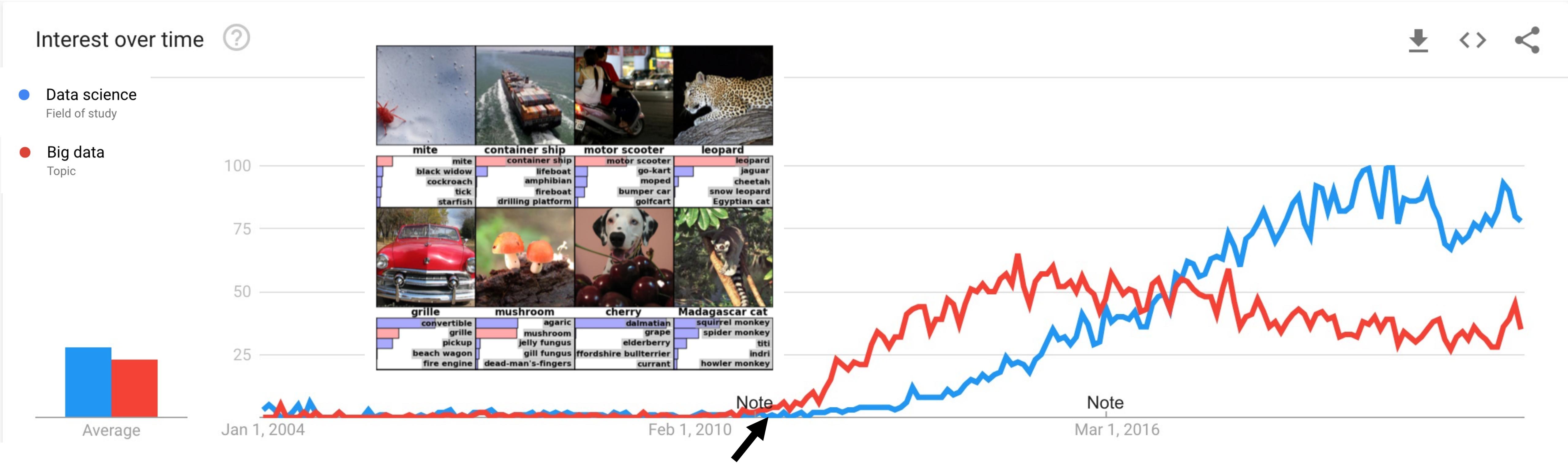
Data Science

- Why is Data Science exciting? – Google Trends



Data Science

- Why is Data Science exciting? – Google Trends

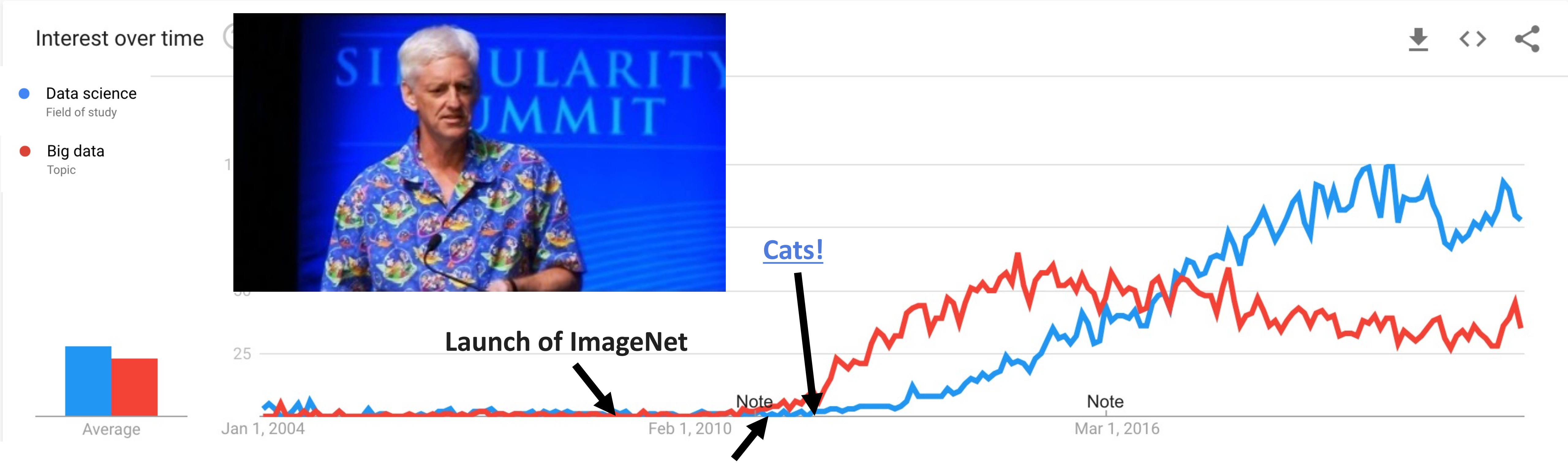


Creation of AlexNet

Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.

Data Science

- Why is Data Science exciting? – Google Trends



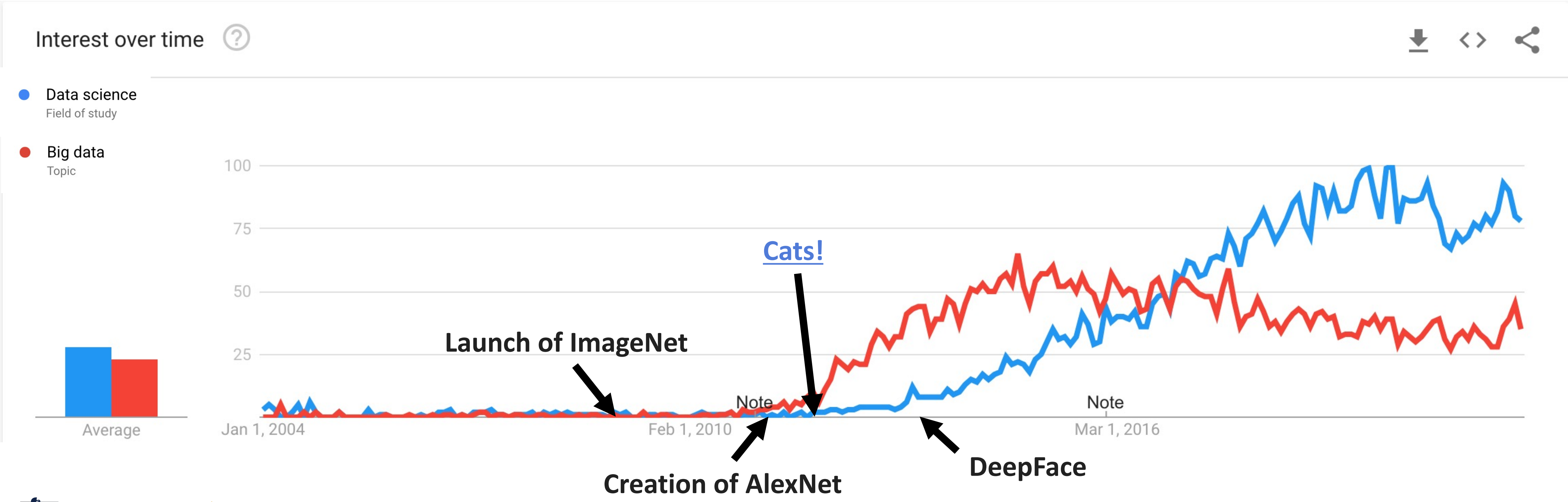
Launch of ImageNet

Cats!

Creation of AlexNet

Data Science

- Why is Data Science exciting? – Google Trends



Cats!

Launch of ImageNet

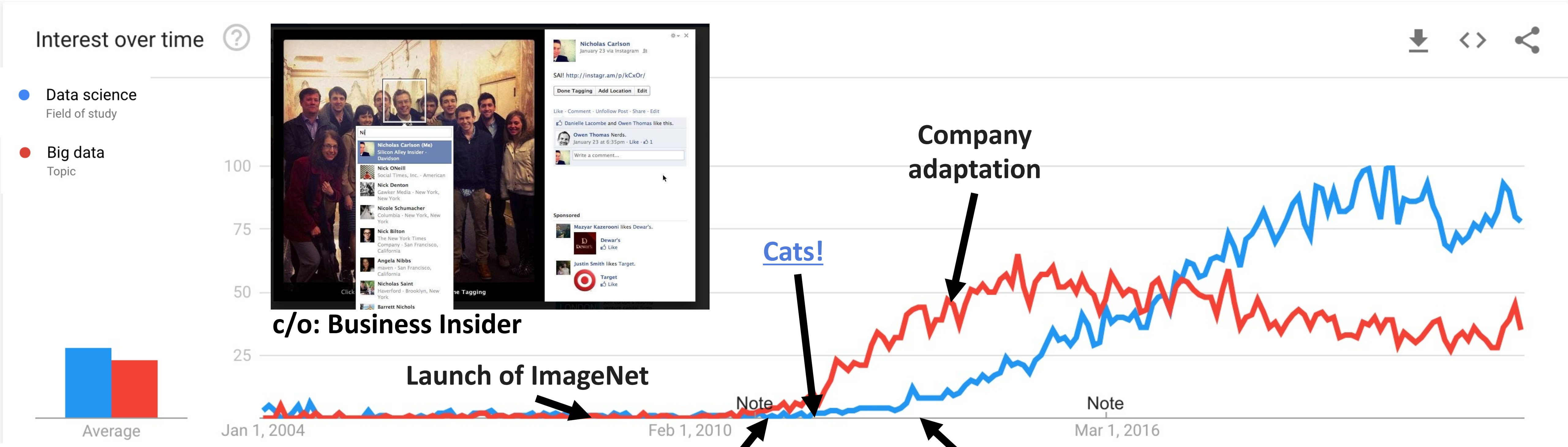
Creation of AlexNet

DeepFace

Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220.

Data Science

- Why is Data Science exciting? – Google Trends



Nicholas Carlson
January 23 via Instagram

SAI! http://instagr.am/p/kCxOr/

Danielle Lacombe and Owen Thomas like this.

Owen Thomas Nerds.
January 23 at 6:35pm · Like · 1

Nicholas Carlson (Me)
Silicon Alley Insider · Davidson

Nick ONeill
Social Times, Inc. · American

Nick Denton
Gawker Media · New York, New York

Nicole Schumacher
Columbia · New York, New York

Nick Bilton
The New York Times Company · San Francisco, California

Angela Nibbs
maven · San Francisco, California

Nicholas Saint
Haverford · Brooklyn, New York

Barrett Nichols

Mazyar Kazerooni likes Dewar's.
Dewar's

Justin Smith likes Target.
Target

c/o: Business Insider

Company adaptation

Cats!

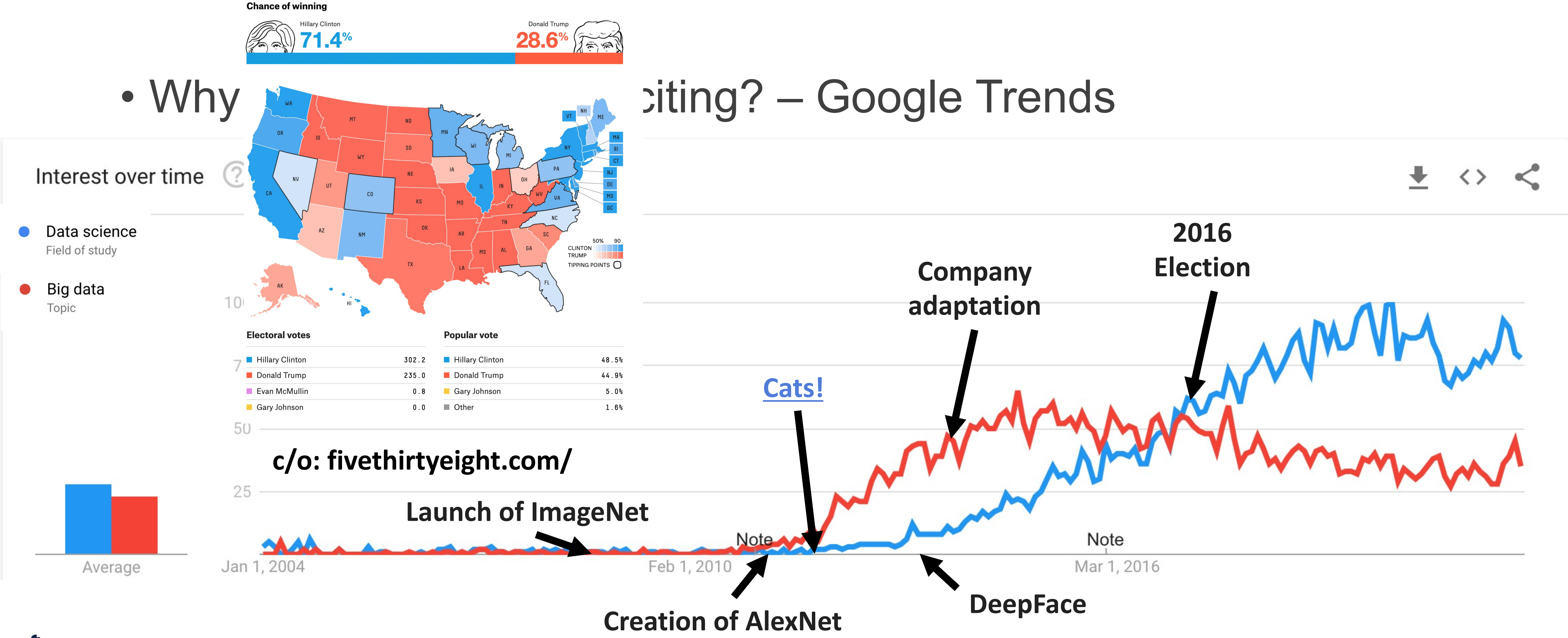
Launch of ImageNet

Creation of AlexNet

DeepFace

Data Science

- Why citing? – Google Trends



Electoral votes

🟦 Hillary Clinton	302.2
🟥 Donald Trump	235.0
🟪 Evan McMullin	0.8
🟨 Gary Johnson	0.0

Popular vote

🟦 Hillary Clinton	48.5%
🟥 Donald Trump	44.9%
🟨 Gary Johnson	5.0%
⬜ Other	1.6%

c/o: fivethirtyeight.com/

2016 Election

Company adaptation

Cats!

Launch of ImageNet

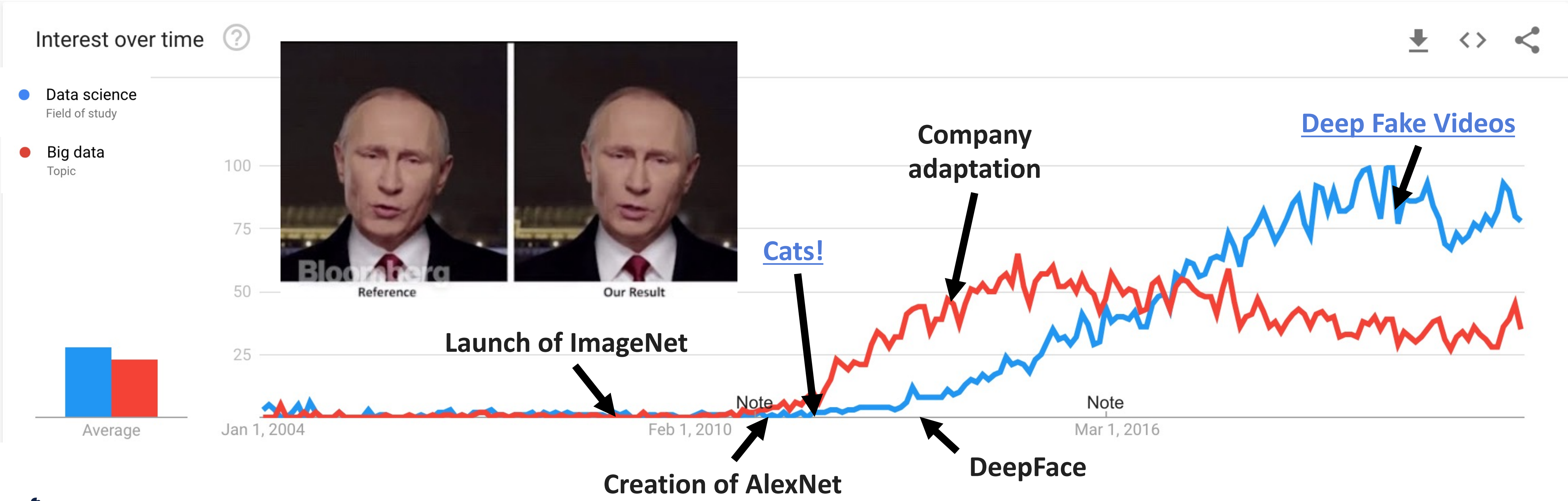
Creation of AlexNet

DeepFace

UNC GREENSBORO

Data Science

- Why is Data Science exciting? – Google Trends



Launch of ImageNet

Creation of AlexNet

Cats!

Company adaptation

DeepFace

Deep Fake Videos

Data Security with Self-learning Neural Networks Models

- American Express Example [Spotify] [Apple Podcasts]
- During Covid Spending Patterns changed
- Banned all spending that differs from past?
  - Great way to lose customers!
- Instead neural networks continuously learned in real time to help mitigate risk for customers and the company

Smart Cities

- Adopting complex technology to improve cities
  - Most of the technology uses sensors
  - Sensors collect data
  - **Insights derived from data**

Data Science

- Exciting Times!
  - Plethora of data
  - Many problems to solve
  - We have the computation means to solve them!
- Many Ethical Issues to tackle behind Data Science
  - Recommender Systems in Social Media
- An increasing demand for Data Scientists

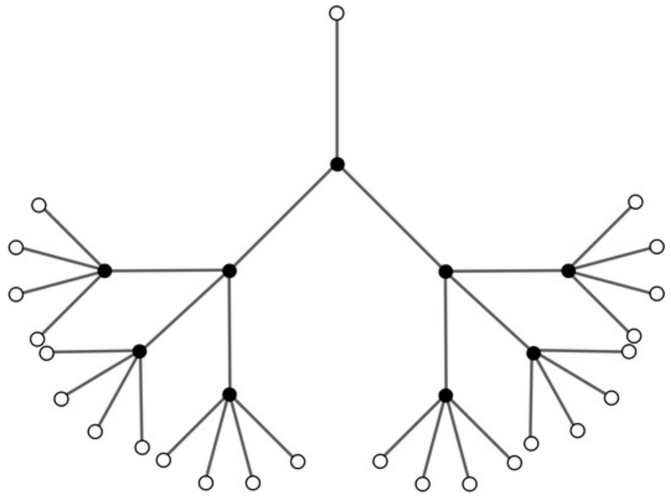
Data Scientist

- What is a Data Scientist Job?
    - Laundry List of Skills
        - Statistical modeling
        - Deep learning
        - Visualizations
        - Communicating effectively
        - …
    - Job role 1: Close to a statisticians
    - Job role 2: Masters Degree in Computer Science
    - Average Salary for a Data Scientist (according to Glassdoor)
        - Greensboro: $108,512
        - Charlotte: $114,918
        - Raleigh: $112,458
        - Generally above $100,000

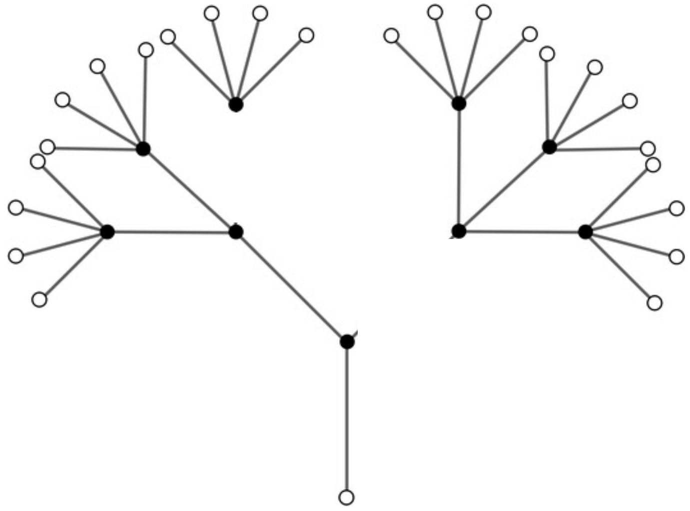
Data Scientist

- Becoming a Data Scientist?
  - Continuous Learning
    - Tools
    - Methods
    - Techniques
  - Allow the data to speak for itself
  - Creativity and Great Problem Solving Skills

Traditional Approach

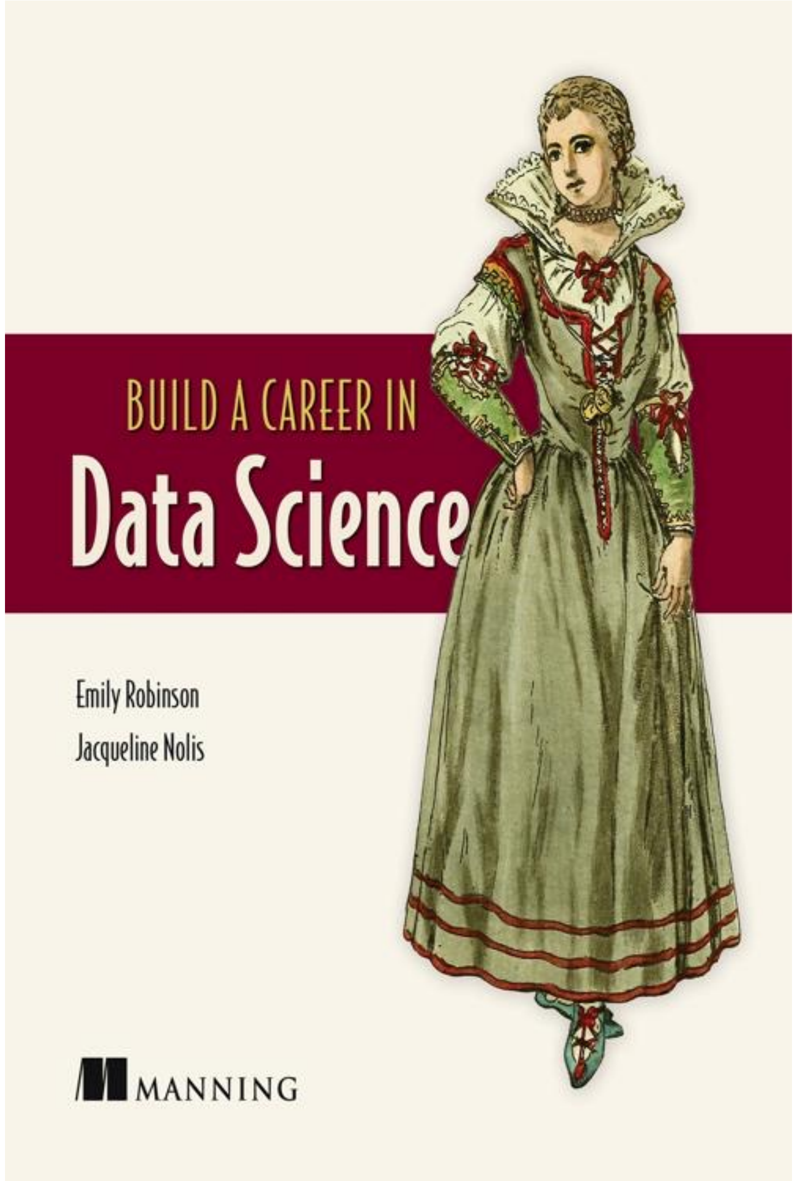


New Approach



Data Scientist - Industry

- Great Career Advice!
  - Create a Data Science Portfolio
  - Different companies types
  - How to identify good jobs through the job posting
  - Data Science Job Applications
  - Interviewing
  - Negotiating your salary
  - Navigating work in your job throughout different stages of your employment



Nolis, J., & Robinson, E. (2020). *Build a Career in Data Science* ([edition missing]). Manning Publications. Retrieved from https://www.perlego.com/book/1469326/build-a-career-in-data-science-pdf (Original work published 2020)

Data Scientist – Company Types

- MTC (Massive Tech Companies)
  - Google’s, Meta, Netflix, Apple, etc.
  - Hundreds or thousands of data related employees
  - High Salary
  - Data Infrastructure exists and is well documented
  - May build models for POC and hand off to a software engineer for implementation
  - Bureaucracy
    - Approval for new technology
    - Conferences
    - Freedom in Approach

Data Scientist – Company Types

- The Established Retailer
  - Payless, Best Buy, Bed Bath & Beyond, etc.
- Slower company to adopt new technology
  - See sales drop because a newer company has disrupted their business (Amazon)
- Newly formed data science team built to provide stakeholders (Executives, directors, managers) with more information and insights to improve the company

Data Scientist – Company Types

- The late-stage, successful tech start up
    - Lyft, Twitter, and Airbnb
- Data Science recognized on a company level
- Data Engineers to support your work
    - Data pipelines become slow or break, data engineers will fix them
- Agile, Fast pace environment
    - Projects may change rapidly

Data Scientist – Company Types

- Government Contractor
  - Boeing, Lockheed Martin, …
- Slow w.r.t data science
- Engineering divisions collecting data but struggle on how it can be used in existing processes
- Pace of work is slow
  - Greater chance of work life balance
- Use Older Technology

Data Scientist – Company Types

Criteria	Massive Tech Companies	Established Retailer	Late Stage Start-Up	Government Contractor
Bureaucracy	A lot	Little	None	A lot
Tech Stack	Complex	Old	Infancy	Ancient
Freedom	Little	A lot	A lot	None
Salary	Amazing	Decent	Poor	Decent
Job Security	Great	Decent	Poor	Great
Chances to Learn	A lot	Some	A lot	Few

Data Scientist

- How does the course help?
  - New methods and techniques
    - Storage
    - Analysis
    - Machine Learning
    - Visualization
  - Change the old way of thinking
  - Creative programming
- Unfortunately cannot tech you everything ☺
  - Will get you started on the path to becoming a data scientist

Datasets

- **Academic Datasets**
  - UC Irvine Machine Learning Repository
  - (http://archive.ics.uci.edu/ml/)
  - Stanford Large Network Dataset Collection
  - (http://snap.stanford.edu/data/)
  - Inter-university Consortium for Political and Social Research
  - (http://www.icpsr.umich.edu/)
  - Pittsburgh Science of Learning Center's DataShop
  - (https://pslcdatashop.web.cmu.edu/)
  - Academic Torrents (http://academictorrents.com/)
- **Private Companies**
  - Data.World (https://data.world/)
  - Quandl Financial Data (https://www.quandl.com/)
  - Amazon Web Services Public Data Sets (http://aws.amazon.com/datasets/)
  - Kaggle (http://www.kaggle.com/)
  - Nytimes (http://developer.nytimes.com/docs)

Datasets

- **Gov. and NGO’s**
  - Data.gov (https://www.data.gov/)
  - NYC Open Data (https://nycopendata.socrata.com/)
  - DC Open Data Catalog (http://data.dc.gov/)
  - OpenDataDC (http://www.opendatadc.org/)
  - DataLA (https://data.lacity.org/)
  - Project Open Data Dashboard (http://data.civicagency.org/))
  - data.gov.uk (http://data.gov.uk/)
  - US Census Bureau (http://www.census.gov/)
  - World Bank Open Data (http://data.worldbank.org/)
  - Humanitarian Data Exchange (http://docs.hdx.rwlabs.org/)
  - Sunlight Foundation (http://sunlightfoundation.com/api/)
  - ProPublica Data Store (https://projects.propublica.org/data-store/)

Datasets

- **Other resources**
  - 20 Big Data Sources (http://www.smartdatacollective.com/bernardmarr/235366/ big-data-20-free-big-data-sources-everyone-should-know)
  - Center for Data Innovation (http://www.datainnovation.org/category/publications/data- set-blog/)
  - Data Science Central (http://www.datasciencecentral.com/)
  - Python API's (http://www.pythonforbeginners.com/api/list- of-python-apis)
  - PyCoders Weekly (http://pycoders.com/)