Computational Structures in Data Science

UC Berkeley EECS
Lecturer
Michael Ball

UC Berkeley EECS
Adj. Asst. Prof.
Dr. Gerald Friedland

Lecture #1: Welcome to CS88!

January 24, 2020

http://cs88.org

CS88 Team - Michael

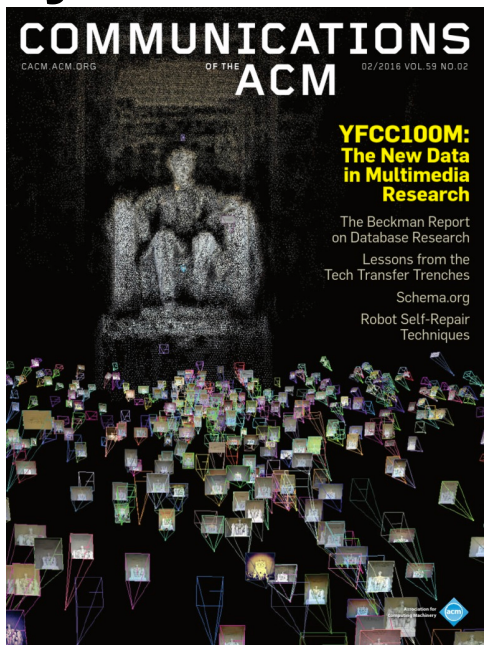
- **Michael Ball**
  - ball@Berkeley.edu – You're best off by using Piazza! ☺
  - 625 Soda Hall
  - http://michaelball.co – I don't update this much…
    - » It was great procrastination when I was a CS student.
  - Office hours: TBD @ 625 Soda
  - A few minutes after class
- **Things I do:**
  - Intro CS Research
    - » Tools, curriculum
  - Training TAs
  - Building Educational Software (Gradescope)
  - Tools for web accessibility

CS88 Team - Gerald

- **Dr. Gerald Friedland**
  - fractor@Berkeley.edu – **You're best off by using Piazza! ☺**
- **Projects you might want to check out:**

http://mmcommons.org

- Work with 100M images, 1M videos in your own Amazon instance.



http://www.teachingprivacy.org

Creating teaching materials informing about data over sharing.

<Teaching Privacy>

CS88 — Our Awesome Team!

Head Teaching Assistants

Alex Kassil

Email: alexkassil@berkeley.edu

Brian Mi

Email: bmi@berkeley.edu

Julia Yu

Email: juliayu@berkeley.edu

Teaching Assistants

Alec Kan

Email: alec.kan@berkeley.edu

Cameron Malloy

Email: cmalloy@berkeley.edu

Shreya Kannan

Email: shreyakannan@berkeley.edu

Sophia Qin

Email: sophia.qin@berkeley.edu

Srinath Goli

Email: srig@berkeley.edu

Vandana Ganesh

Email: vandanag@berkeley.edu

Welcome

- **We are all here to learn:**

```
Knowledge(end) – Knowledge(start)
```

Goals today

- Introduce you to
  - the field
  - the course
  - the team

- Answer your questions

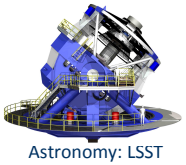
- Big Ideas:
  - Abstraction
  - Data Types

Logistics

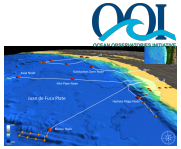
- Labs are Mon & Tues
  - Please see the signup form on Piazza
  - Go to any lab section next week
  - Attendance is tracked, but not required. (”Bonus” points.)
- There are 2 sections on CalCentral, it does not matter which you are enrolled in.
- 1-2pm lecture is webcast, 3pm is not.

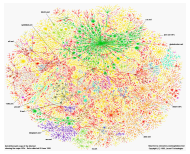
Data Science

Nearly every field of discovery is transitioning from "data poor" to "data rich"


Astronomy: LSST


Physics: LHC

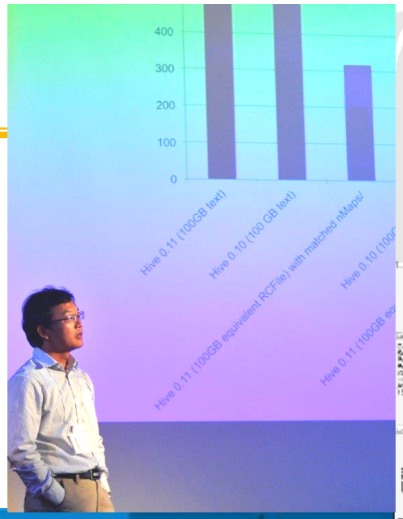

Oceanography: OOI

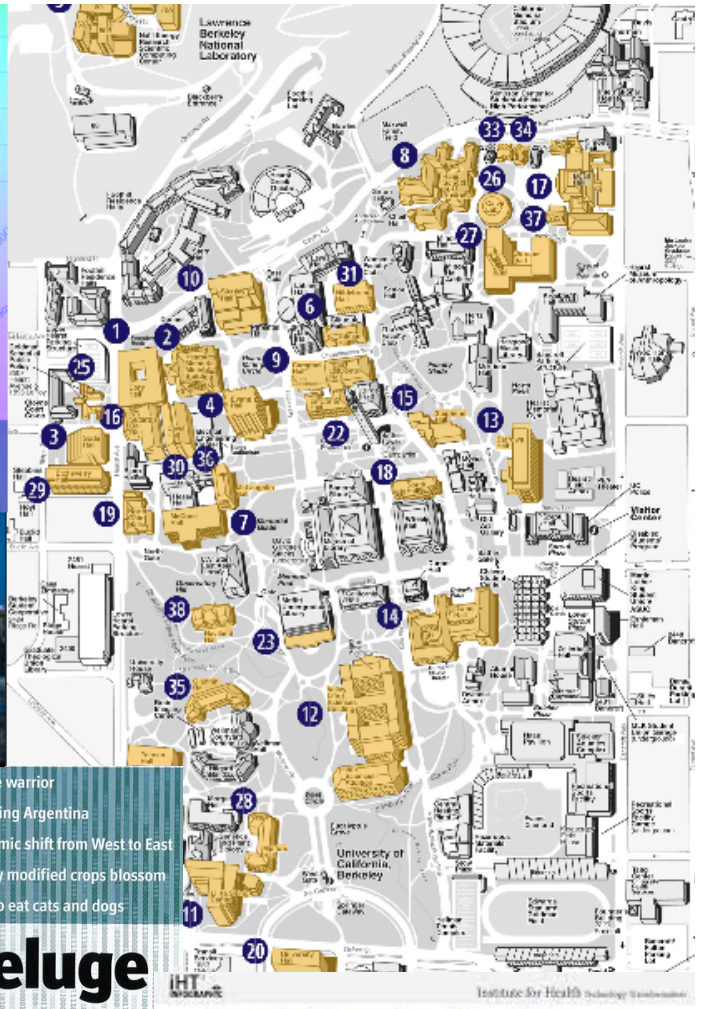

Sociology: The Web


Biology: Sequencing


Economics: POS terminals


Neuro



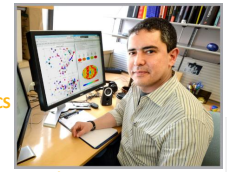


Berkeley UNIVERSITY OF CALIFORNIA

Data Science growing organically everywhere

WIRED Spark: Open Source Superstar Rewrites Future of Big Data
BY CADE METZ 06.19.13 6:30 AM


AMP Lab
Ion Stoica, CS
Michael Franklin, CS


Fernando Perez,
Brain Imaging Center
iPython tools and community

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase


Adam Arkin,
Bioengineering


Charles Marshall
Rosie Gillespie
Integrative Biology
Digitized Museum

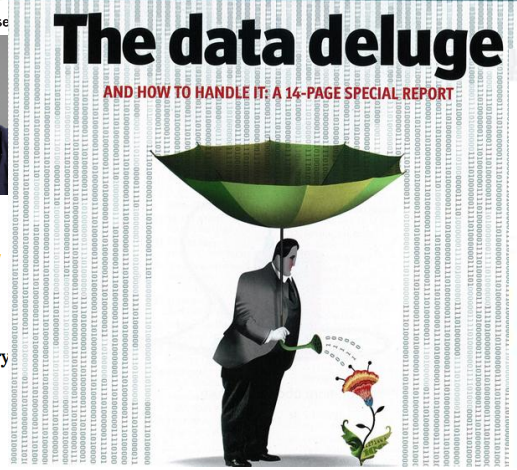
Reconstructing the movies in your mind

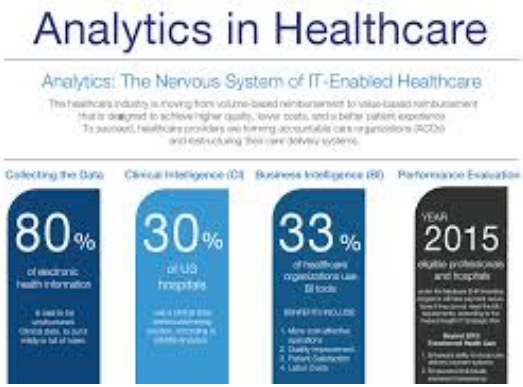

Bin Yu, Statistics
Jack Gallant, Neuroscience


Richard Allen
Earth& Plan.
Science
Geospatial Lab

The New York Times
Incomes Flat in Recovery but Not for the 1%
Feb 15, 2013

Emmanuel Saez, Economics





Berkeley UNIVERSITY OF CALIFORNIA

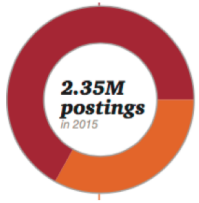
A National Challenge

Increasingly US jobs require data science and analytics skills. Can we meet the demand? The current shortage of skills in the national job pool demonstrates that business-as-usual strategies won't satisfy the growing need. If we are to unlock the promise and potential of data and all the technologies that depend on it, employers and educators will have to transform.

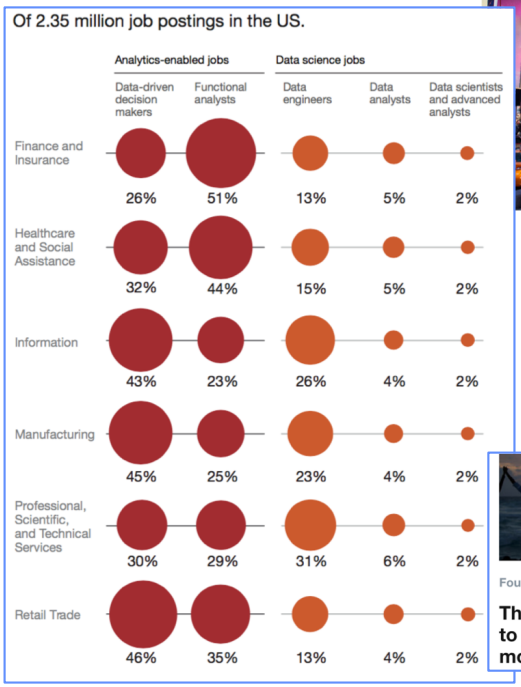
By 2021, **69% of employers expect** candidates with DSA skills to get preference for jobs in their organizations.
Only **23% of college** and university leaders say their graduates will have those skills.

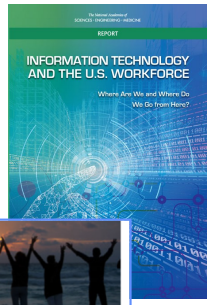
Report | McKinsey Global Institute
Big data: The next frontier for innovation, competition, and productivity
May 2011 | by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers

pwc.com/us/dsa-skills
Investing in America's data science and analytics talent
The case for action



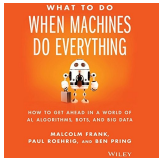
April 2017
BHEF





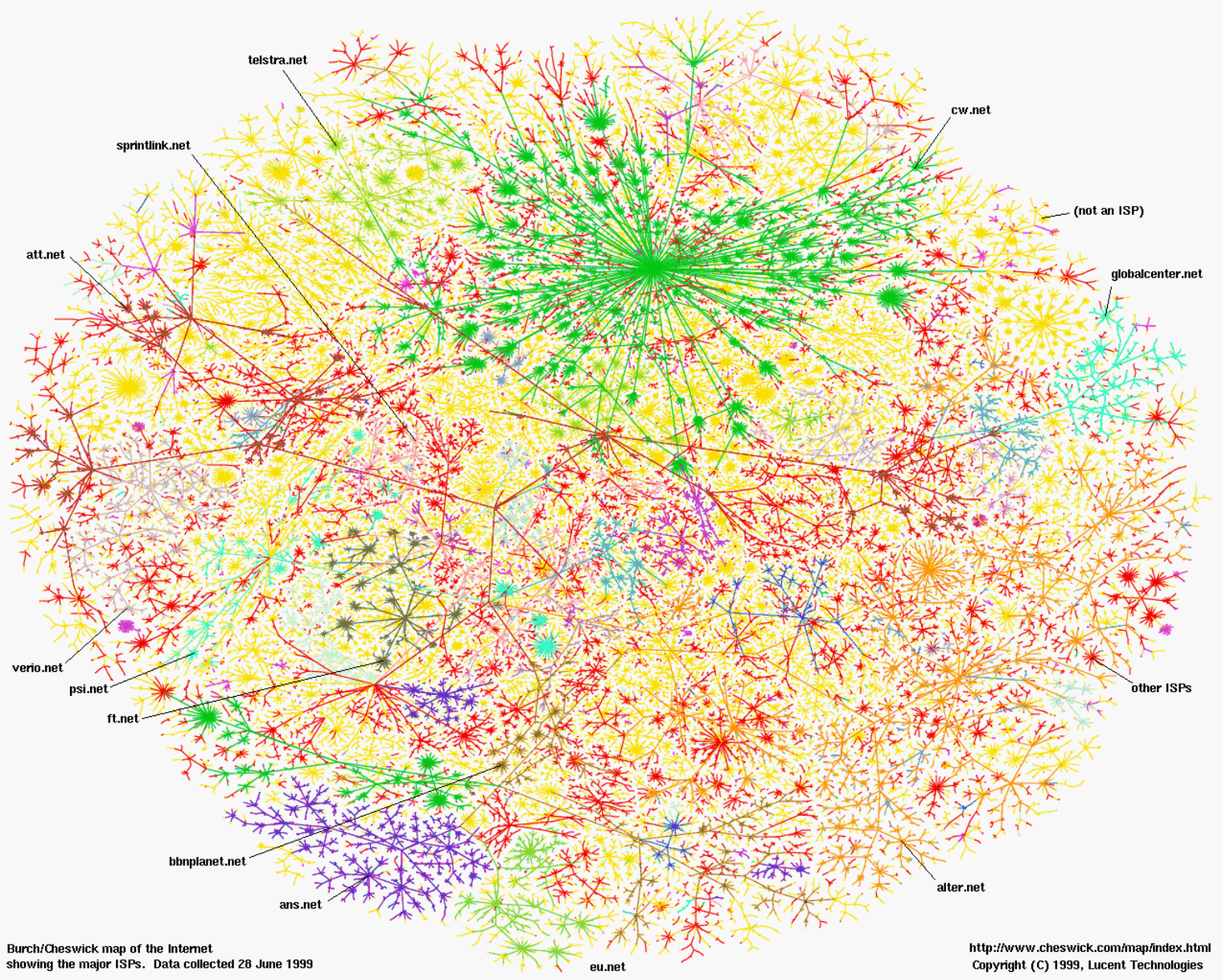
Fourth Industrial Revolution
The fourth sector is a chance to build a new economic model for the benefit of all

Augmenting Human Intelligence



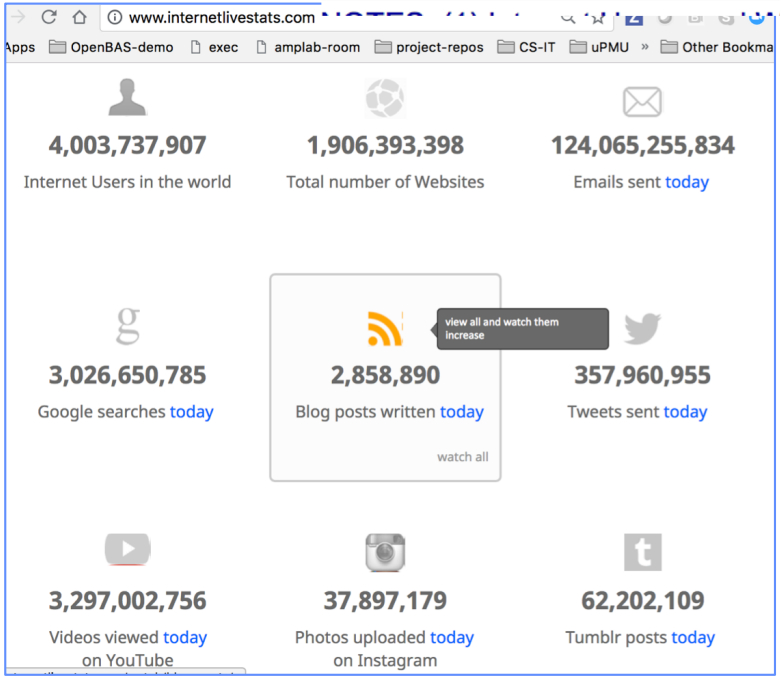
WORLD ECONOMIC FORUM
The Fourth Industrial Revolution: what it means, how to respond

Greatest Artifact of Human Civilization …

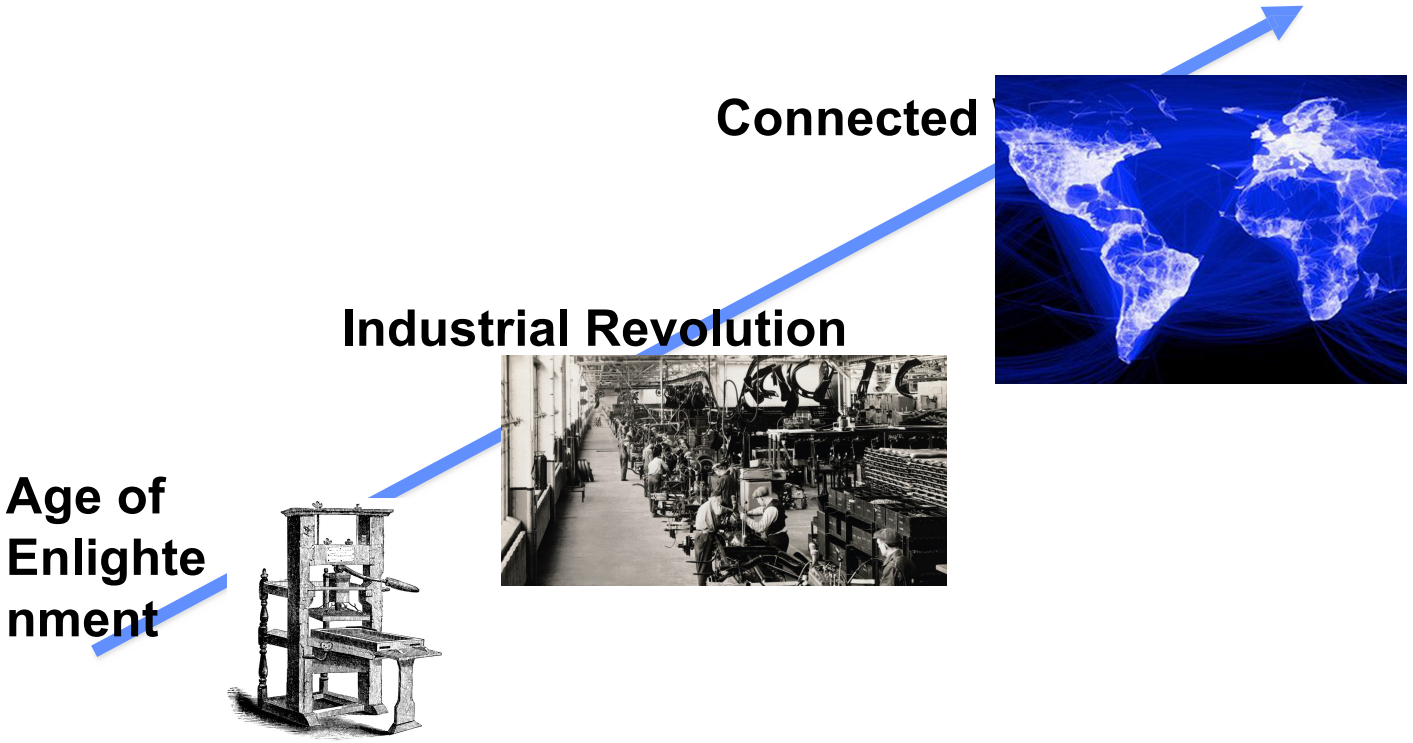


WORLD INTERNET USAGE AND POPULATION STATISTICS
DEC 31, 2017 - Update

World Regions	Population (2018 Est.)	Population % of World	Internet Users 31 Dec 2017	Penetration Rate (% Pop.)	Growth 2000-2018
Africa	1,287,914,329	16.9 %	453,329,534	35.2 %	9,941 %
Asia	4,207,588,157	55.1 %	2,023,630,194	48.1 %	1,670 %
Europe	827,650,849	10.8 %	704,833,752	85.2 %	570 %
Latin America / Caribbean	652,047,996	8.5 %	437,001,277	67.0 %	2,318 %
Middle East	254,438,981	3.3 %	164,037,259	64.5 %	4,893 %
North America	363,844,662	4.8 %	345,660,847	95.0 %	219 %
Oceania / Australia	41,273,454	0.6 %	28,439,277	68.9 %	273 %
WORLD TOTAL	7,634,758,428	100.0 %	4,156,932,140	54.4 %	1,052 %



Era of Transformation



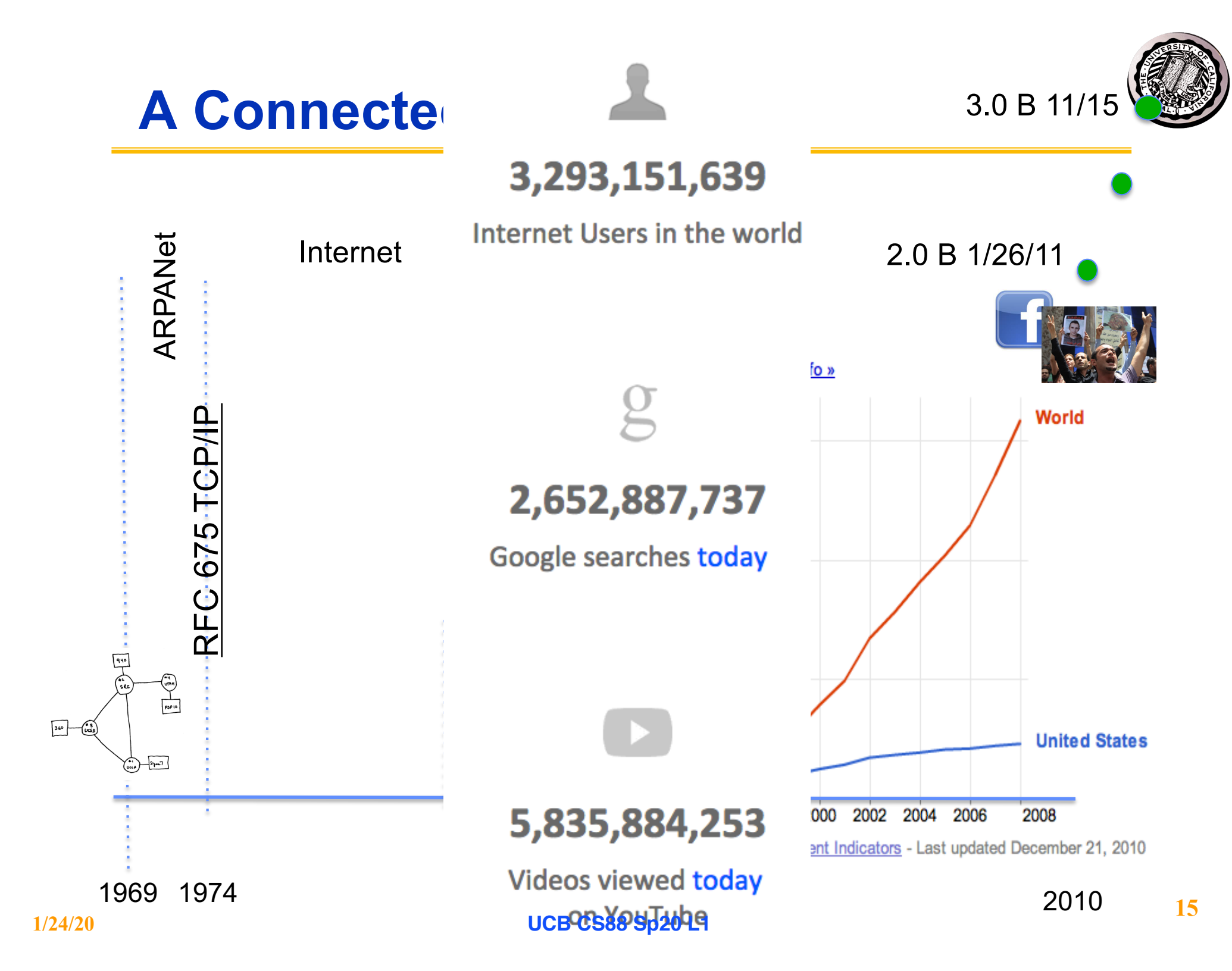
A Connected World of Data

- The world's knowledge at our finger tips
- *Digitialization* of life, industry and society
- Intimately connected to billions of us, globally
- Explosion of observational instruments
  - Genomics, Microscopy, Astronomical, …
- Vast Computational power to do analytics
- Synthetic design exploration thru simulation
- Machine reading of everything
- Statistical machine learning algorithms to "discover" structure

What if I could … ?

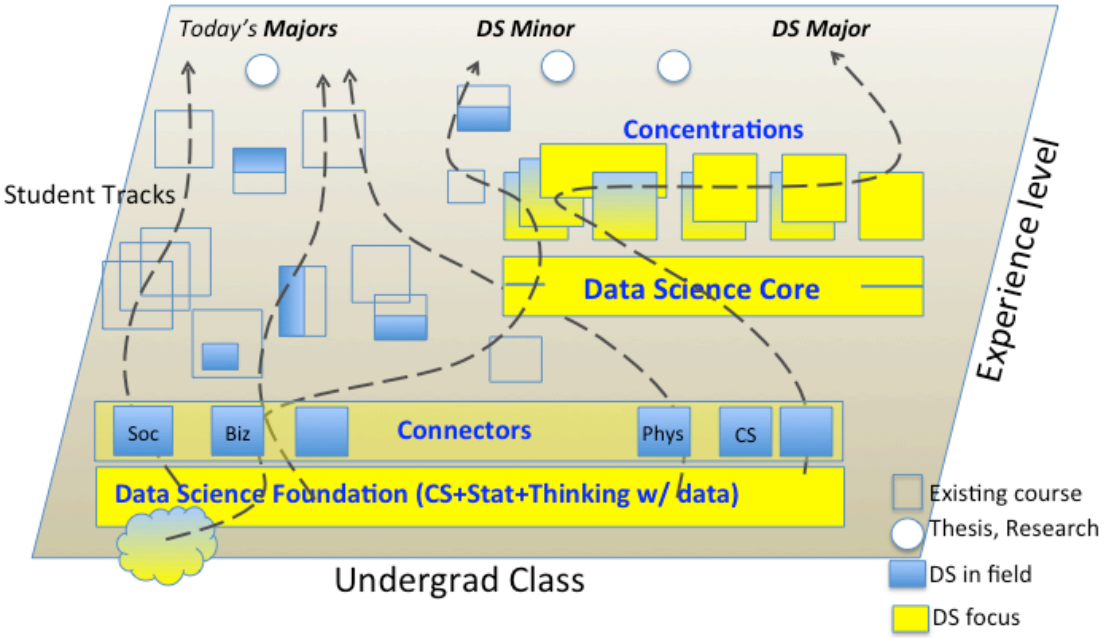
- See the world's digital footprints?
- Read everything that's ever been written?
- Take it all in and dive down anywhere as far as the science can take me?
- Learn the physical/chemical/biological /sociological/neurological… models from the data?
- Explore billions of designs and pick the one I want?
- … ?

A Connecte


3.0 B 11/15
3,293,151,639
Internet Users in the world
2.0 B 1/26/11
2,652,887,737
Google searches today
5,835,884,253
Videos viewed today
on YouTube
Internet
ARPANet
RFC 675 TCP/IP
World
United States
000 2002 2004 2006 2008
ent Indicators - Last updated December 21, 2010
1969 1974
2010

Data 8 – Foundations of Data Science

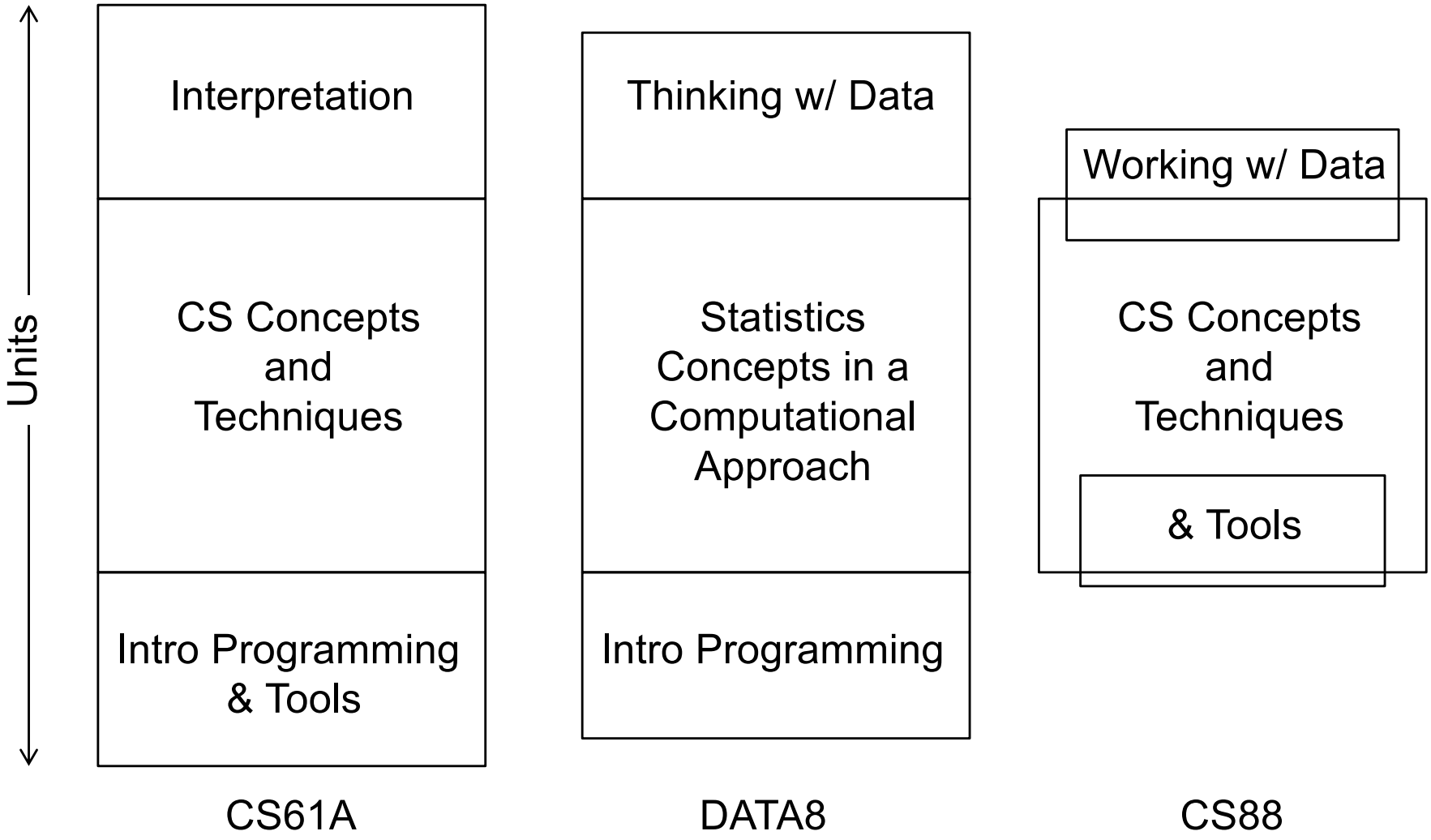
- **Computational Thinking + Inferential Thinking in the context of working with real world data**
- **Introduce you to several computational concepts in a simple data-centered setting**
  - **Authoring computational documents**
  - **Tables**
  - **Within Python3 and "SciPy"**



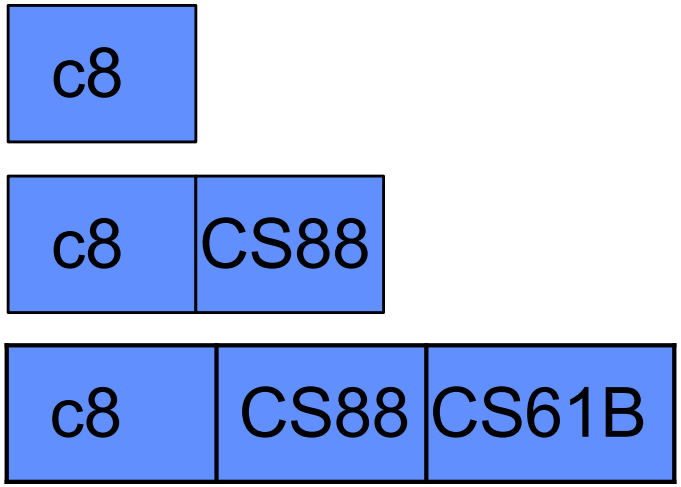
CS88 – Computational Structures in Data Science

- **Deeper understanding of the computing concepts introduced in C8**
  - Hands-on experience => Foundational Concept
  - How would you create what you use in c8 ?
- **Extend your understanding of the structure of computation**
  - What is involved in interpreting the code you write ?
  - Deeper CS Concepts: Recursion, Objects, Classes, Higher-order Functions, Declarative programming, …
  - Managing complexity in creating larger software systems through composition
- **Create complete (and fun) applications**
- **In a data-centric approach**

How does CS88 relate to CS61A ?



Opportunities for students

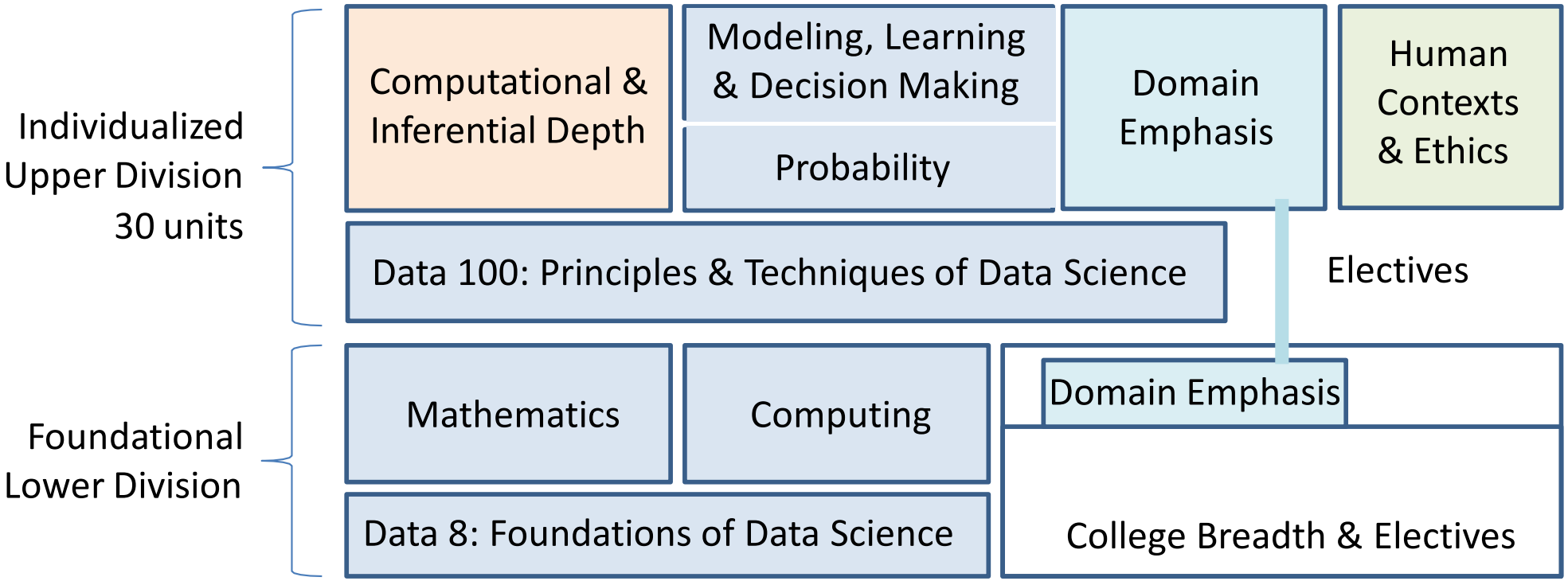


CS minor

***

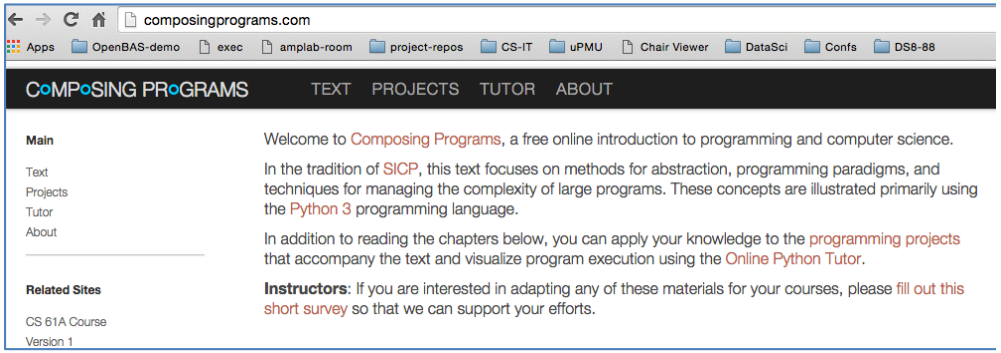
CS major

The Data Science Major



Course Structure

- 2 hours lecture, 2 hours lab, ~5 hours HW+study
- Lecture introduces concepts (quickly!), answers why questions.
- Lab provides concrete detail hands-on
- Homework (12) cements your understanding
  - Out Tuesdays, Due Next Thurs (~9 days)
- Projects (2) put your understanding to work in building complete applications
  - Maps
  - Ants vs Some Bees
  - Maybe: open-ended final?

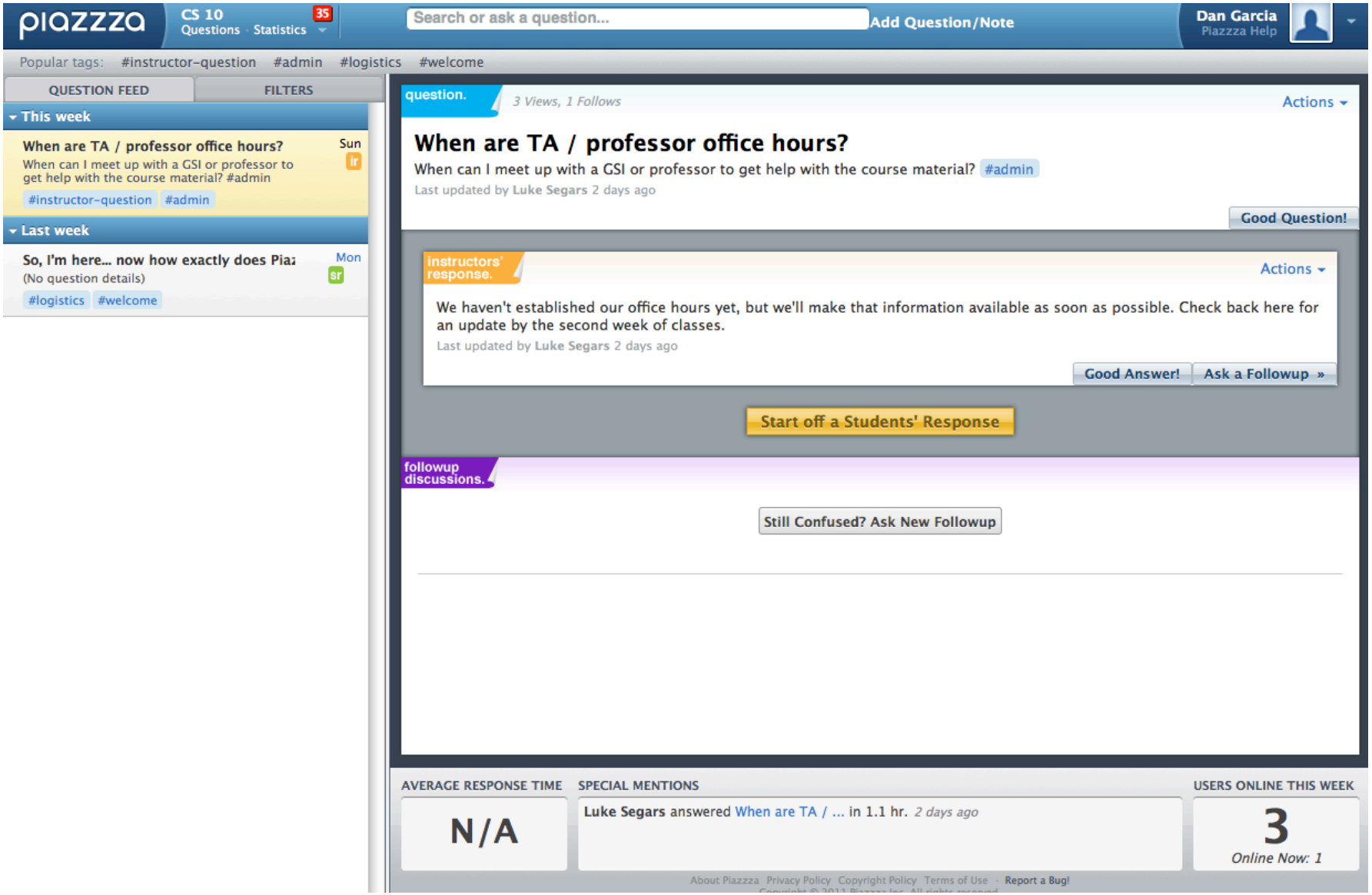


- Readings: http://composingprograms.com
  - Same as CS61a

Course Culture

- Learning
- Community
- Respect
- Collaboration
- Peer Instruction

Piazza for {ask,answer}ing questions



Where will we work?

- Your laptop
  - Using an editor and a terminal
- cs88.org
- datahub.berkeley.edu
  - "Jupyter Notebooks". An industry standard tool.
  - Not as often, but an option

iClicker Check In

- Are you enrolled in Data 8?

- A. I took if Fall 2018 or earlier
- B. I took it Spring 2019
- C. I'm taking it right now
- D. I am trying to enroll in Data 8
- E. I am not taking Data 8

Pro-student Grading Policies

- EPA
  - Rewards good behavior
  - Effort
    - » E.g., Office hours, doing every single lab, hw, reading Piazza pages
  - Participation
    - » E.g., Raising hand in lec or discussion, asking questions on Piazza
  - Altruism
    - » E.g., helping other students in lab, answering questions on Piazza
- You have 3 "Slip Days"
  - Homework and Projects
  - You use them to extend due date, 1 slip day for 1 day extension
  - You can use them one at a time or all at once or in any combination
  - They follow you around when you pair up (you are counted individually)
    - » E.g., A has 2, B has 0. Project is late by 1 day. A uses 1, B is 1 day late

Pro-student Grading Policies

- Turning it work will never harm you! Even if it's late or you can't get it completely correct.
- Labs are based primarily on effort.
- If you're struggling, come talk to us!
  - Especially your TAs. They have been there too.

Abstraction

- **Detail removal**

  "The act of leaving out of consideration one or more properties of a complex object so as to attend to others."

- **Generalization**

  "The process of formulating general concepts by abstracting common properties of instances"

- **Technical terms: Compression, Quantization, Clustering, Unsupervized Learning**

Henri Matisse *"Naked Blue IV"*

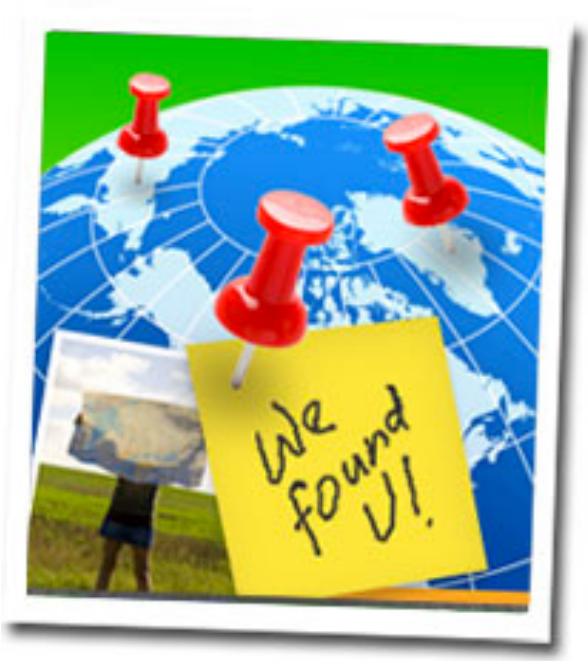
Experiment


Standard Time Zones of the World
WHERE ARE YOU FROM?
CANADA
U.S.
UNITED STATES
MEXICO
BRAZIL
ARGENTINA
CHILE
PERU
BOLIVIA
PARAGUAY
COLOMBIA
VENEZUELA
ECUADOR
RUSSIA
CHINA
INDIA
MONGOLIA
KAZAKHSTAN
JAPAN
PHILIPPINES
INDONESIA
AUSTRALIA
NEW ZEALAND
ALGERIA
LIBYA
EGYPT
SUDAN
NIGERIA
ETHIOPIA
ZAIRE
SOUTH AFRICA
MADAGASCAR
SAUDI ARABIA
IRAN
IRAQ
TURKEY
FRANCE
SPAIN
GERMANY
POLAND
UKRAINE
SWEDEN
NORWAY
FINLAND
ICELAND
North Atlantic Ocean
South Atlantic Ocean
North Pacific Ocean
South Pacific Ocean
Indian Ocean
International Date Line
Coordinated Universal Time (U.T.C.) formerly Greenwich Mean Time
C.A.R. -Central African Republic
F.Y.R.O.M. -The Former Yugoslav Republic of Macedonia

Where are you from?

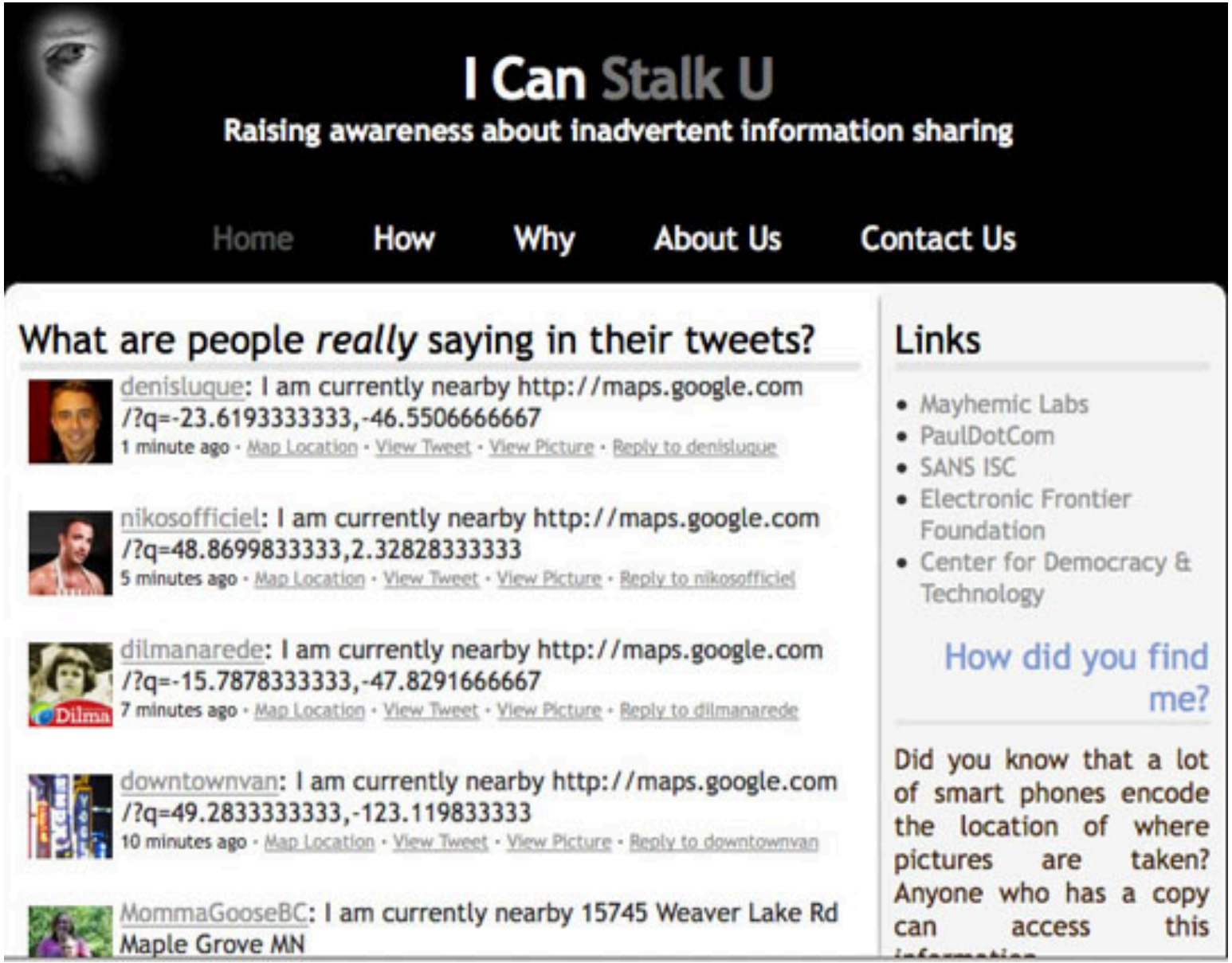
Possible Answers:

- **Planet Earth**
- **Europe**
- **California**
- **The Bay Area**
- **San Mateo**
- **1947 Center Street, Berkeley, CA**
- **37.8693° N, 122.2696° W**



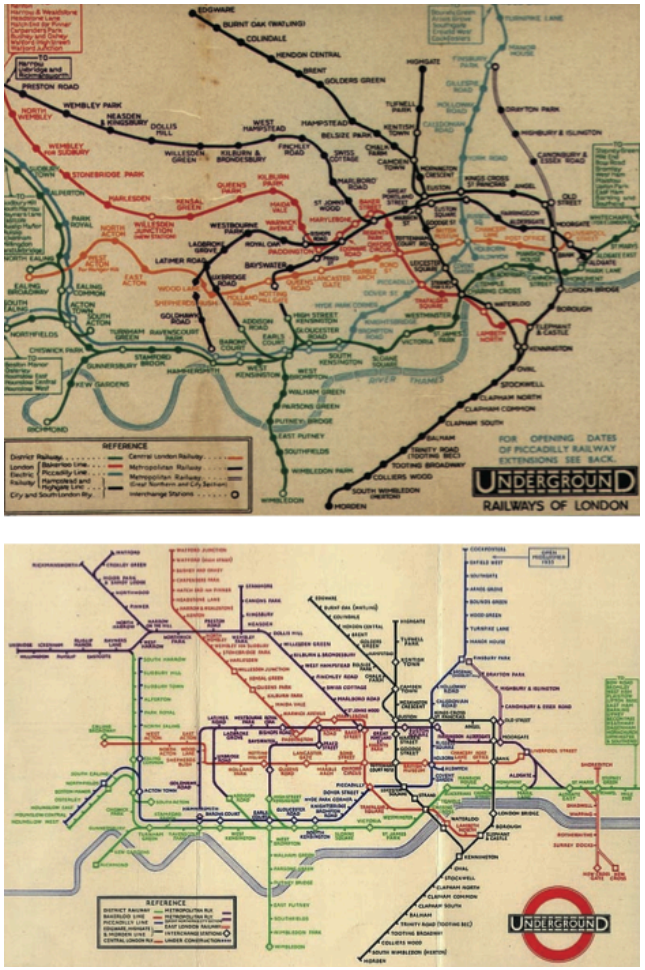
All correct but different levels of abstraction!

Abstraction gone wrong!



Detail Removal (in Data Science)

- You'll want to look at only the interesting data, leave out the details, zoom in/out…
- Abstraction is the idea that you focus on the essence, the cleanest way to map the messy real world to one you can build
- Experts are often brought in to know what to remove and what to keep!



The London Underground 1928 Map & the 1933 map by Harry Beck.

The Power of Abstraction, Everywhere!

- **Examples:**
  - **Functions (e.g., sin x)**
  - **Hiring contractors**
  - **Application Programming Interfaces (APIs)**
  - **Technology (e.g., cars)**
- **Amazing things are built when these layer**
  - **And the abstraction layers are getting deeper by the day!**

We only need to worry about the interface, or specification, or contract NOT how (or by whom) it's built

Above the abstraction line

Abstraction Barrier (Interface)
(the interface, or specification, or contract)

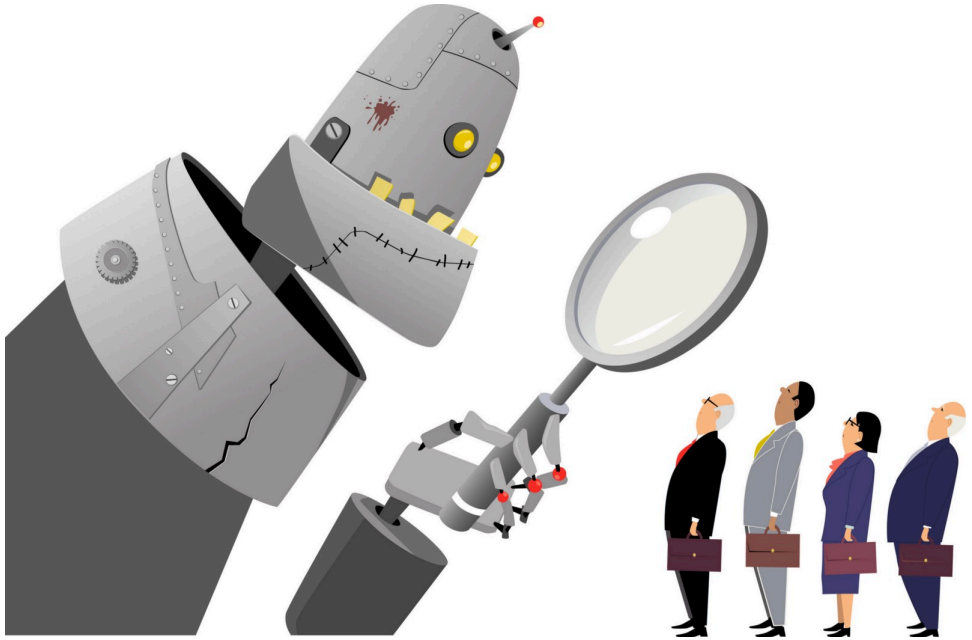
Below the abstraction line

This is where / how / when / by whom it is actually built, which is done according to the interface, specification, or contract.

Abstraction: Pitfalls

- **Abstraction is not universal without loss of information (mathematically provable). This means, in the end, the complexity can only be "moved around"**

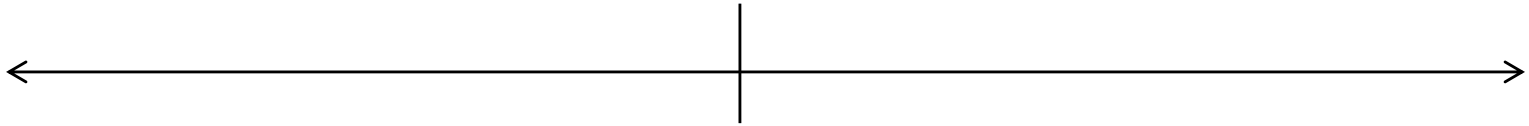
- **Abstraction makes us forget how things actually work and can therefore hide bias. Example: AI and hiring decisions.**



- **Abstraction makes things special and that creates dependencies. Dependencies grow longer and longer over time and can become unmanageable.**

Abstraction in CS: Data Type

- **What's this?**



Real (or ideal) world

42

Computer representation

Data Types and Operations

- **Set of elements**
  - with some internal representation
  - E.g. Integers, Floats, Booleans, Strings, …
- **Set of operations on elements of the type**
  - e.g. +, *, -, /, %, //, **
  - ==, <, >, <=, >=
- **Properties**
  - Commutative, Associative, …

- **Expressions are valid well-defined sets of operations on elements that produce a value of a type**

Lab and HW next week

- **Lab will get you setup with Python locally**
  - **You'll use your own computer**
  - **Learn about text files and the terminal**

Thoughts for the Wandering Mind

A binary digit (bit) is a symbol from {0,1}.

- **How many things can you represent with N bits?**

- **How many things can you represent with 1 digit (0-9)?**
- **2 digits? 6 digits?**