



Computational Structures in Data Science



UC Berkeley EECS
Lecturer
Michael Ball

Lecture #1: Welcome to CS88!





Welcome

- We are all here to learn:
Knowledge (end) – Knowledge (start)



CS88 Team

Head Teaching Assistants



Alex Kassil

Email: alexkassil@berkeley.edu



Amir Shahait

Email: ashahait@berkeley.edu

Teaching Assistants



Alec Kan

Email: alec.kan@berkeley.edu



Brian Mi

Email: bmi@berkeley.edu



Julia Yu

Email: juliayu@berkeley.edu



Lyric Yu

Email: lyricyu@berkeley.edu



Sophia Qin

Email: sophia.qin@berkeley.edu



Srinath Goli

Email: srig@berkeley.edu



CS88 Team - me

- 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: Tues 5:00-7:00pm @ 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





Goals today

- Introduce you to
 - the field
 - the course
 - the team
- Answer your questions



- Big Ideas:
 - Abstraction
 - Data Type





Announcements

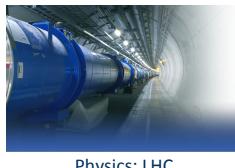
- **Labs are Fridays – we do take attendance**
- **Lab 1 is due tomorrow night**
- **HW 1 is due Thurs night**
- **CS Mentors Sections:**
 - <http://www.bit.ly/88csm1>

Data Science

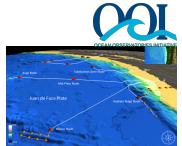
Nearly every field of discovery is transitioning from “data poor” to “data rich”



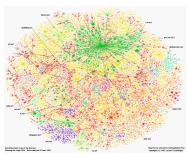
Astronomy: LSST



Physics: LHC



Oceanography: OOI



Sociology: The Web



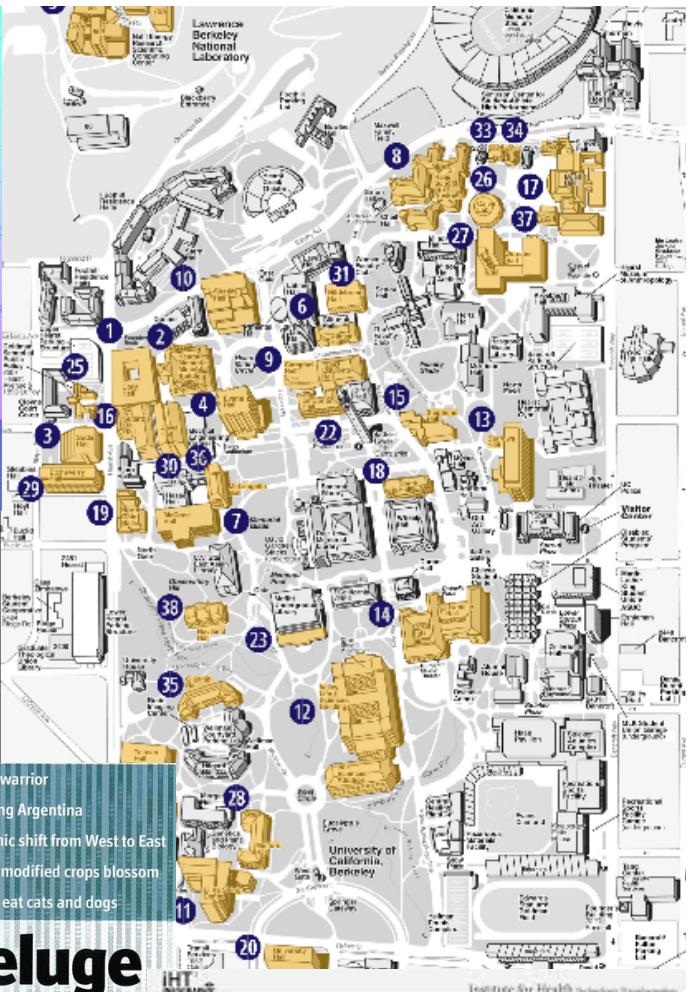
Biology: Sequencing



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 08.19.13 6:30 AM



Reconstructing the movies in your mind



Bin Yu, Statistics
Jack Gallant, Neuroscience



Richard Allen
Earth & Plan.
Science
Geospatial Lab

KBase
PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

Adam Arkin,
Bioengineering



Fernando Perez,
Brain Imaging Center
iPython tools and community



Charles Marshall
Rosie Gillespie
Integrative Biology
Digitized Museum

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



Emmanuel Saez, Economics

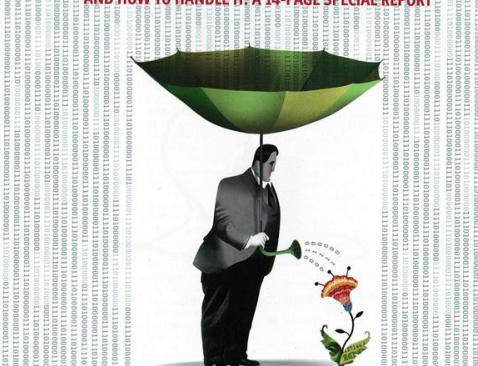
The Economist

FEBRUARY 27TH-MARCH 9TH 2010

Economist.com

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT



Analytics in Healthcare

Analytics: The Nervous System of IT-Enabled Healthcare

The healthcare industry is moving from volume-based reimbursement to value-based reimbursement that is designed to achieve higher quality, lower costs, and a better patient experience. To succeed, healthcare providers are forming accountable care organizations (ACOs) and collaborating through care delivery systems.



Berkeley
UNIVERSITY OF CALIFORNIA

UCB CS88 Fa19 L1

A National Challenge

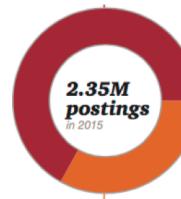
Increasingly US jobs require data science and analytics skills. Can we meet the demand? The current shortage of skills in the national job pool demonstrates that business-as-usual strategies won't satisfy the growing need. If we are to unlock the promise and potential of data and all the technologies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

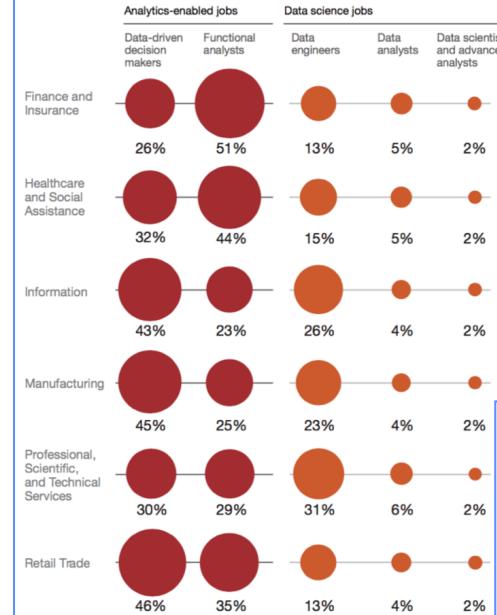
May 2011 | by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byrnes



April 2017

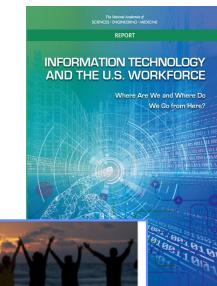
BHEF

Of 2.35 million job postings in the US.



Investing in America's data science and analytics talent

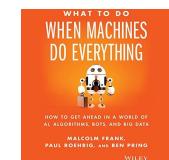
The case for action



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

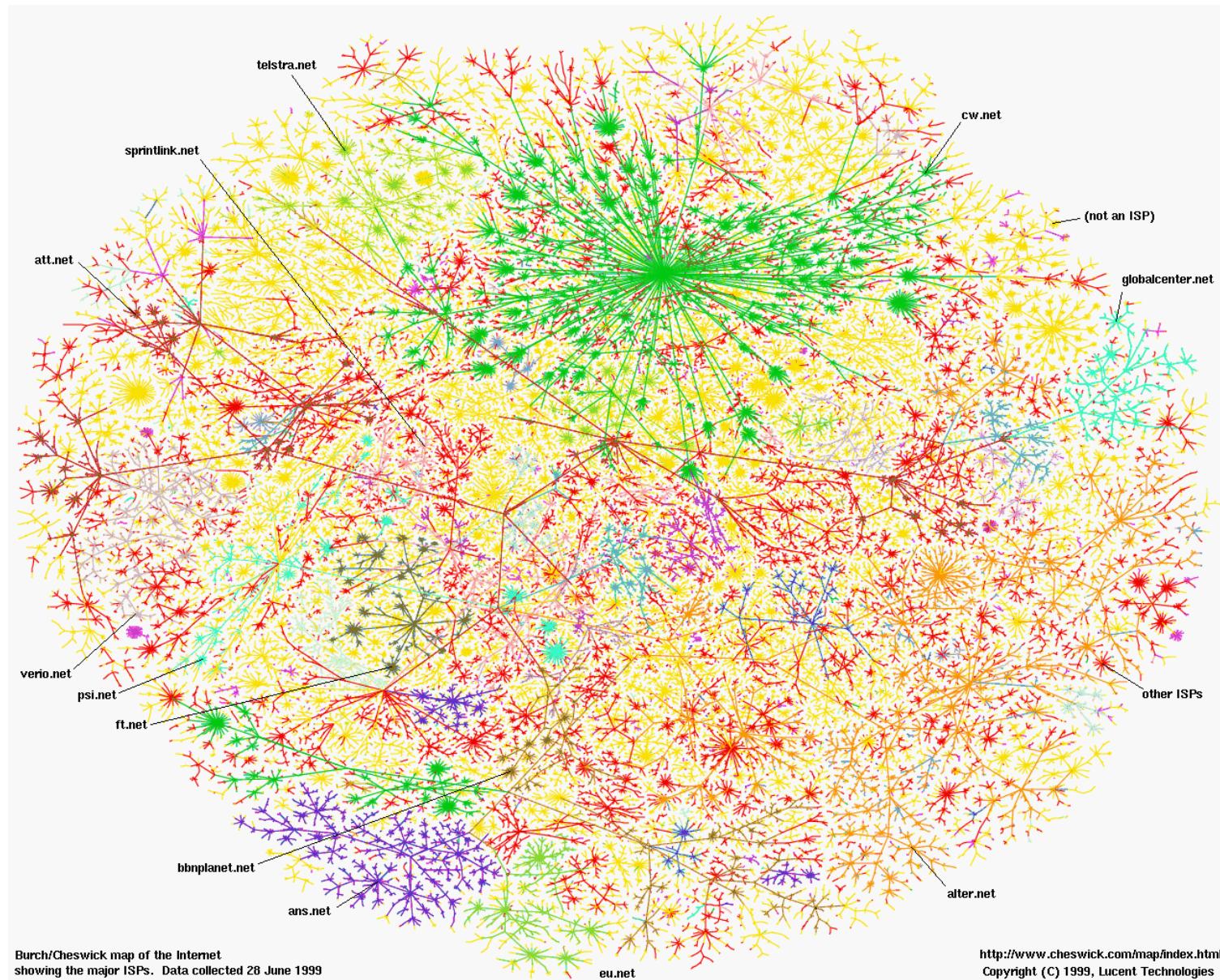
Augmenting Human Intelligence

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





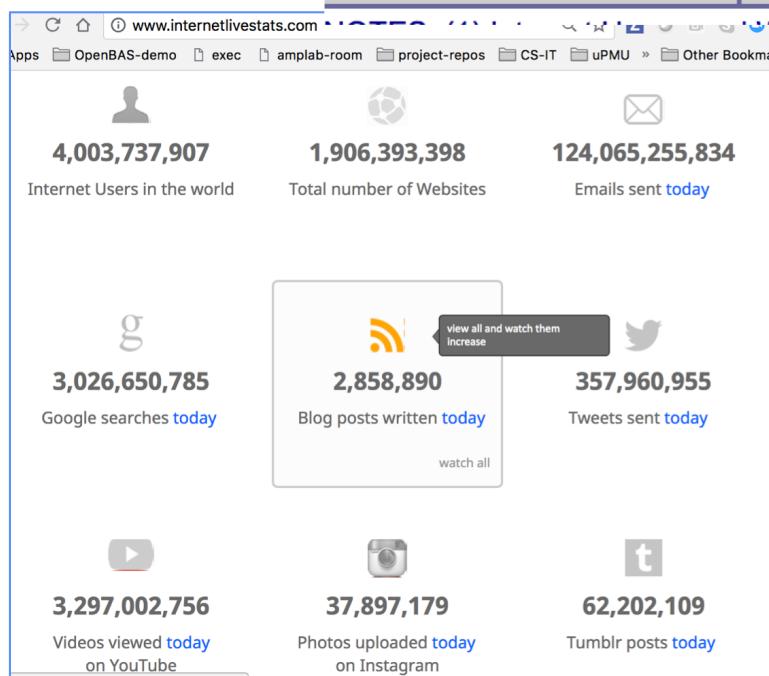
Greatest Artifact of Human Civilization ...





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

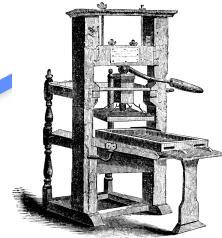
World Regions	Population (2018 Est.)	Population % of World	Internet Users 31 Dec 2017	Penetration Rate (% Pop.)	Growth 2000-2018
Africa	1,287,914,329	16.9 %	453,329,534	35.2 %	9,941 %
Asia	4,207,588,157	55.1 %	2,023,630,194	48.1 %	1,670 %
Europe	827,650,849	10.8 %	704,833,752	85.2 %	570 %
Latin America / 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

Age of Enlightenment



Industrial Revolution



Connected





A Connected World of Data

- The world's knowledge at our finger tips
- *Digitization* 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



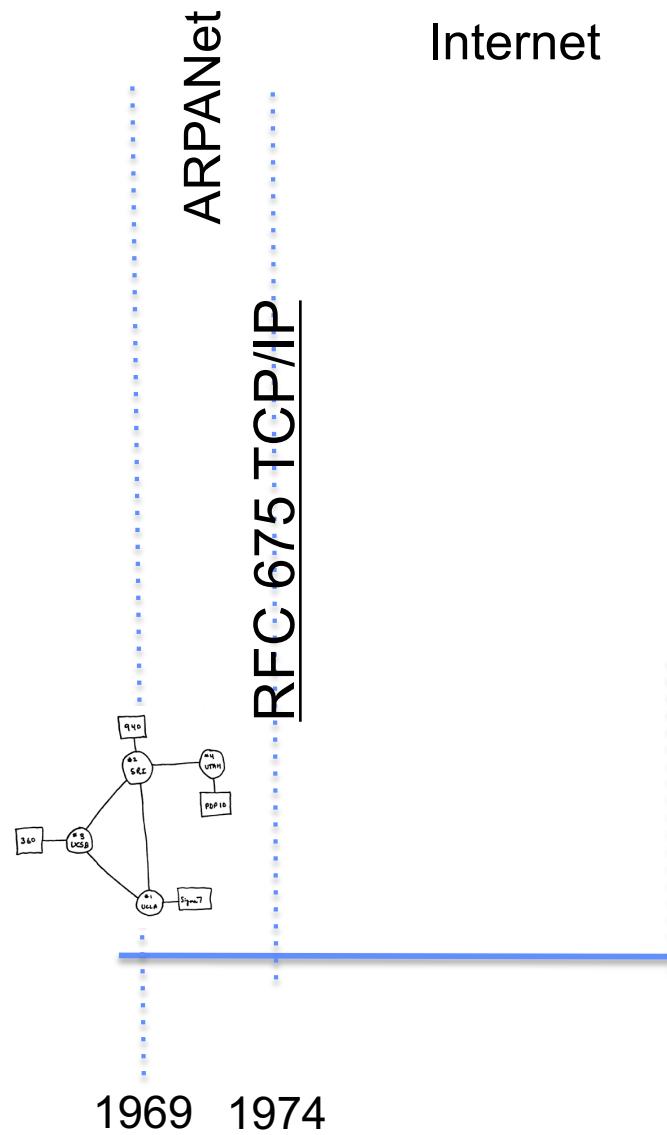
What if I could ... ?

- 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 Connected



09/9/19



3.0 B 11/15

3,293,151,639

Internet Users in the world

g

2,652,887,737

Google searches **today**

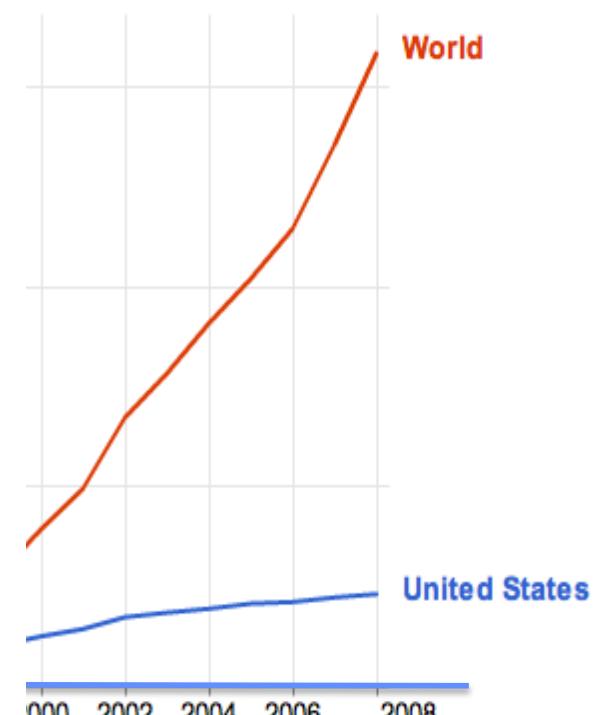


5,835,884,253

Videos viewed **today**
on YouTube

UCB CS88 Fa19 L1

fo »



Internet Indicators - Last updated December 21, 2010

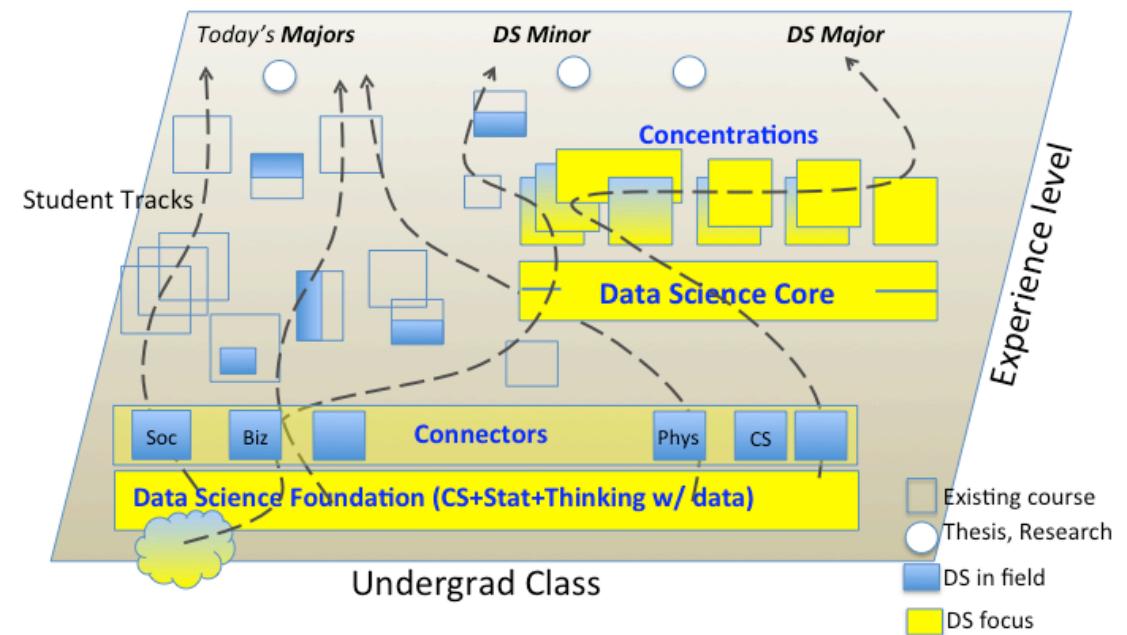
2010

14



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”



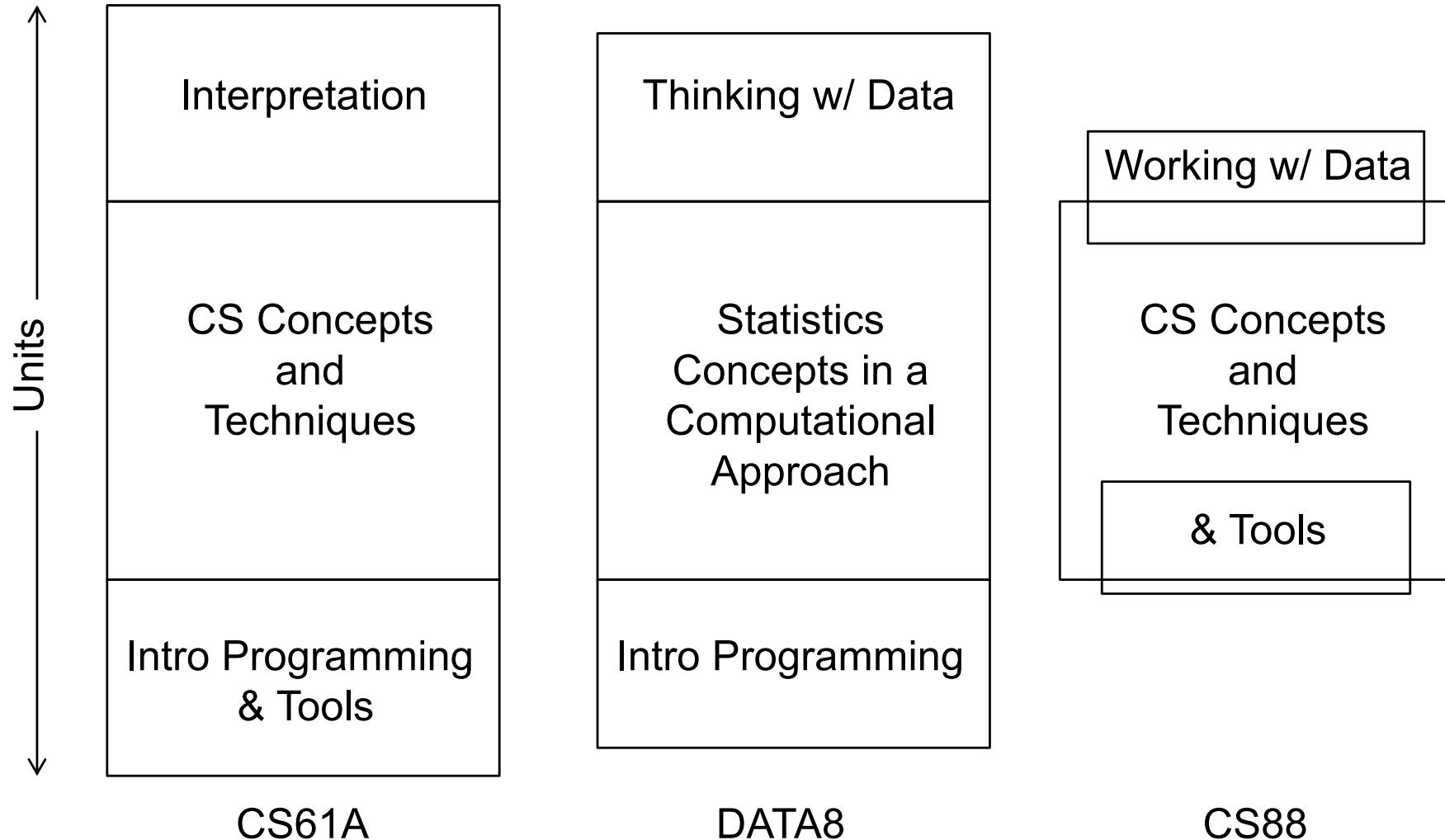
CS88 – Computational Structures in Data Science



- **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

c8

c8 CS88

c8 CS88 CS61B

CS minor

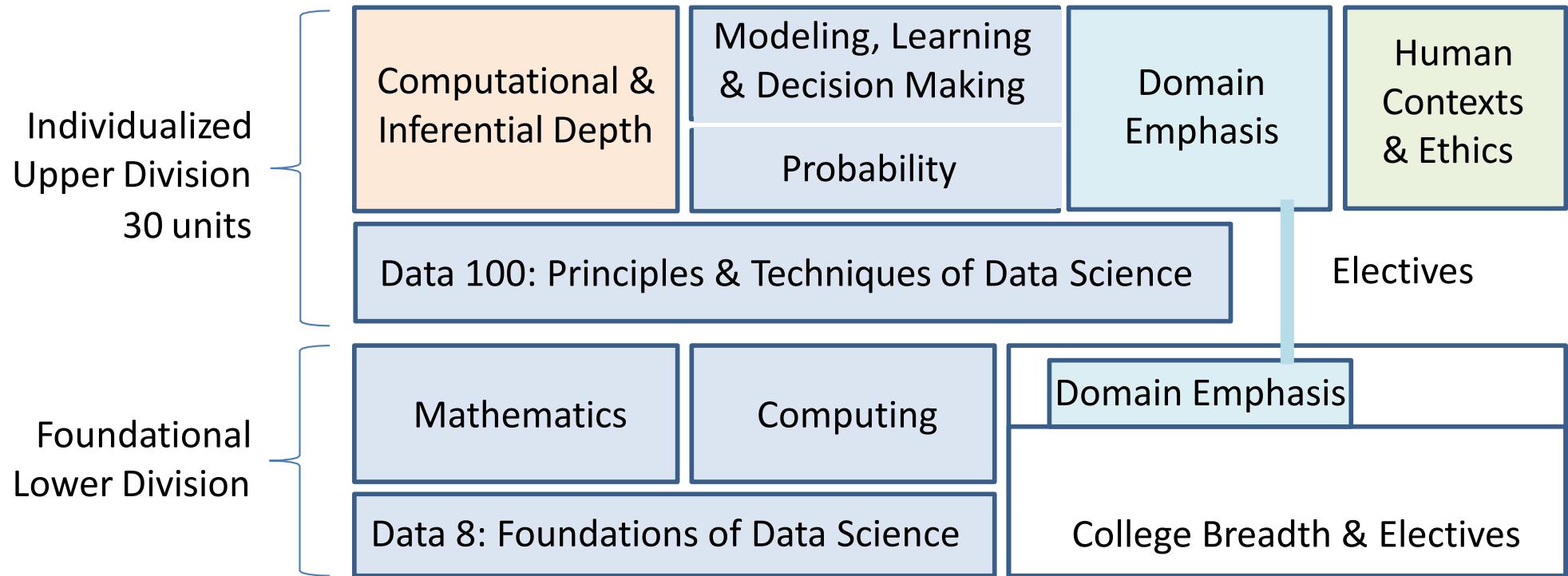
CS major

c8 cs61a

cs61a



The Data Science Major





Course Structure

- **1 Lecture + 1 Lab/Discussion on Friday (!!!)**
- **Lecture introduces concepts (quickly!), answers why questions.**
- **Lab provides concrete detail hands-on**
- **Homework (10) cements your understanding**
 - Out Tuesdays, Due Next Thurs (~9 days)
- **Projects (2) put your understanding to work in building complete applications**
 - Maps
 - Probably: Ants vs Some Bees

A screenshot of a web browser displaying the homepage of composingprograms.com. The page has a header with navigation links for Apps, OpenBAS-demo, exec, amplab-room, project-repos, CS-IT, uPMU, Chair Viewer, DataSci, Confs, and DS8-88. Below the header is a main menu with 'COMPOSING PROGRAMS' on the left and 'TEXT', 'PROJECTS', 'TUTOR', and 'ABOUT' on the right. On the left side, there's a sidebar with links for Main, Text, Projects, Tutor, and About. The main content area welcomes users to 'Composing Programs' and describes it as a free online introduction to programming and computer science, focusing on abstraction, programming paradigms, and managing complexity using Python 3. It also mentions the 'Online Python Tutor'. At the bottom, there's a note for 'Instructors' about adapting materials for courses and a link to a 'short survey'.

- **Readings:** <http://composingprograms.com>
 - Same as cs61a



Course Culture

- Learning
- Community
- Respect
- Collaboration
- Peer Instruction





Piazza for {ask,answer}ing questions

Screenshot of the Piazza platform interface.

Header: piazzza CS 10 Questions - Statistics 35 | Search or ask a question... Add Question/Note | Dan Garcia Piazza Help

Left Sidebar (QUESTION FEED):

- This week:**
 - When are TA / professor office hours? (Sun) [Ir]
When can I meet up with a GSI or professor to get help with the course material? #admin
#instructor-question #admin
- Last week:**
 - So, I'm here... now how exactly does Pia (Mon)
(No question details)
#logistics #welcome

Question Detail View:

Question: When are TA / professor office hours?
When can I meet up with a GSI or professor to get help with the course material? #admin
Last updated by Luke Segars 2 days ago

Instructors' response:
We haven't established our office hours yet, but we'll make that information available as soon as possible. Check back here for an update by the second week of classes.
Last updated by Luke Segars 2 days ago

Actions: Good Question!

Followup: Still Confused? Ask New Followup

Bottom Metrics:

- AVERAGE RESPONSE TIME: N/A
- SPECIAL MENTIONS: Luke Segars answered When are TA / ... in 1.1 hr. 2 days ago
- USERS ONLINE THIS WEEK: 3 Online Now: 1

About Piazza | Privacy Policy | Copyright Policy | Terms of Use | Report a Bug!
Copyright © 2013 Piazza Inc. All rights reserved.



Where will we work?

- Your laptop
 - Using an editor and a terminal
- cs88.org
- Datahub.berkeley.edu
 - Not as often, but an option



iClicker Check In

- How has lab gone so far?
- A. Labs have gone fantastic!
- B. Labs have gone alright...
- C. Labs have gone very well...
- D. I haven't been to lab yet.



iClicker Check In

- Are you enrolled in Data 8?
- A. I took it 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”**
 - 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



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, Unsupervised
Learning



Henri Matisse “Naked Blue IV”



Experiment

Standard Time Zones of the World

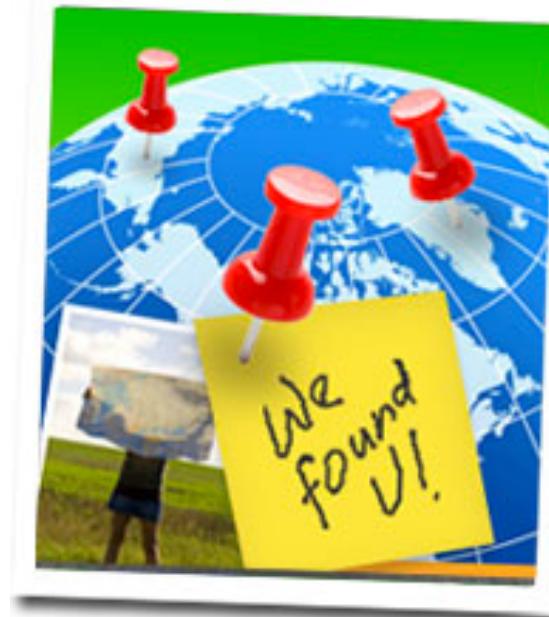




Where are you from?

Possible Answers:

- Planet Earth
- Europe
- California
- The Bay Area
- San Mateo
- 1947 Center Street,
Berkeley, CA
- $37.8693^\circ \text{ N}, 122.2696^\circ \text{ W}$



All correct but different levels of abstraction!



Abstraction gone wrong!



I Can Stalk U

Raising awareness about inadvertent information sharing

Home How Why About Us Contact Us

What are people *really* saying in their tweets?

 [denisluque](#): I am currently nearby <http://maps.google.com/?q=-23.6193333333,-46.5506666667>
1 minute ago · [Map Location](#) · [View Tweet](#) · [View Picture](#) · [Reply to denisluque](#)

 [nikosofficiel](#): I am currently nearby <http://maps.google.com/?q=48.8699833333,2.3282833333>
5 minutes ago · [Map Location](#) · [View Tweet](#) · [View Picture](#) · [Reply to nikosofficiel](#)

 [dilmanarede](#): I am currently nearby <http://maps.google.com/?q=-15.7878333333,-47.8291666667>
7 minutes ago · [Map Location](#) · [View Tweet](#) · [View Picture](#) · [Reply to dilmanarede](#)

 [downtownvan](#): I am currently nearby <http://maps.google.com/?q=49.2833333333,-123.119833333>
10 minutes ago · [Map Location](#) · [View Tweet](#) · [View Picture](#) · [Reply to downtownvan](#)

 [MommaGooseBC](#): I am currently nearby 15745 Weaver Lake Rd
Maple Grove MN

Links

- Mayhemic Labs
- PaulDotCom
- SANS ISC
- Electronic Frontier Foundation
- Center for Democracy & Technology

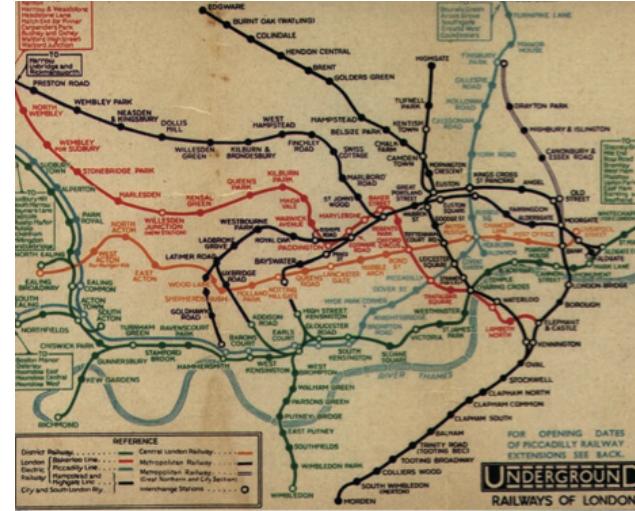
How did you find me?

Did you know that a lot of smart phones encode the location of where pictures are taken? Anyone who has a copy can access this information.



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)

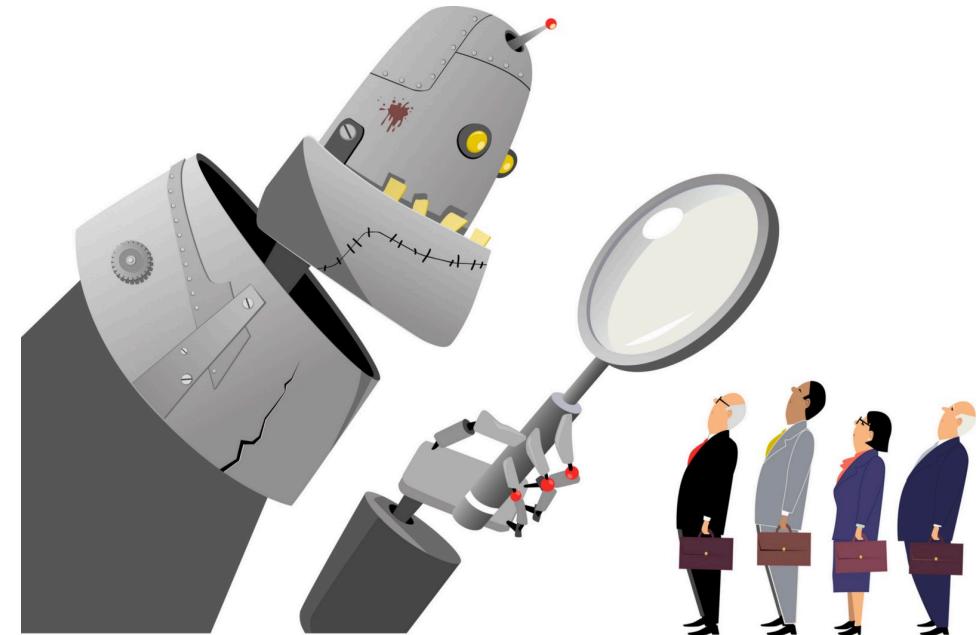
Below the abstraction line

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



Abstraction: Pitfalls

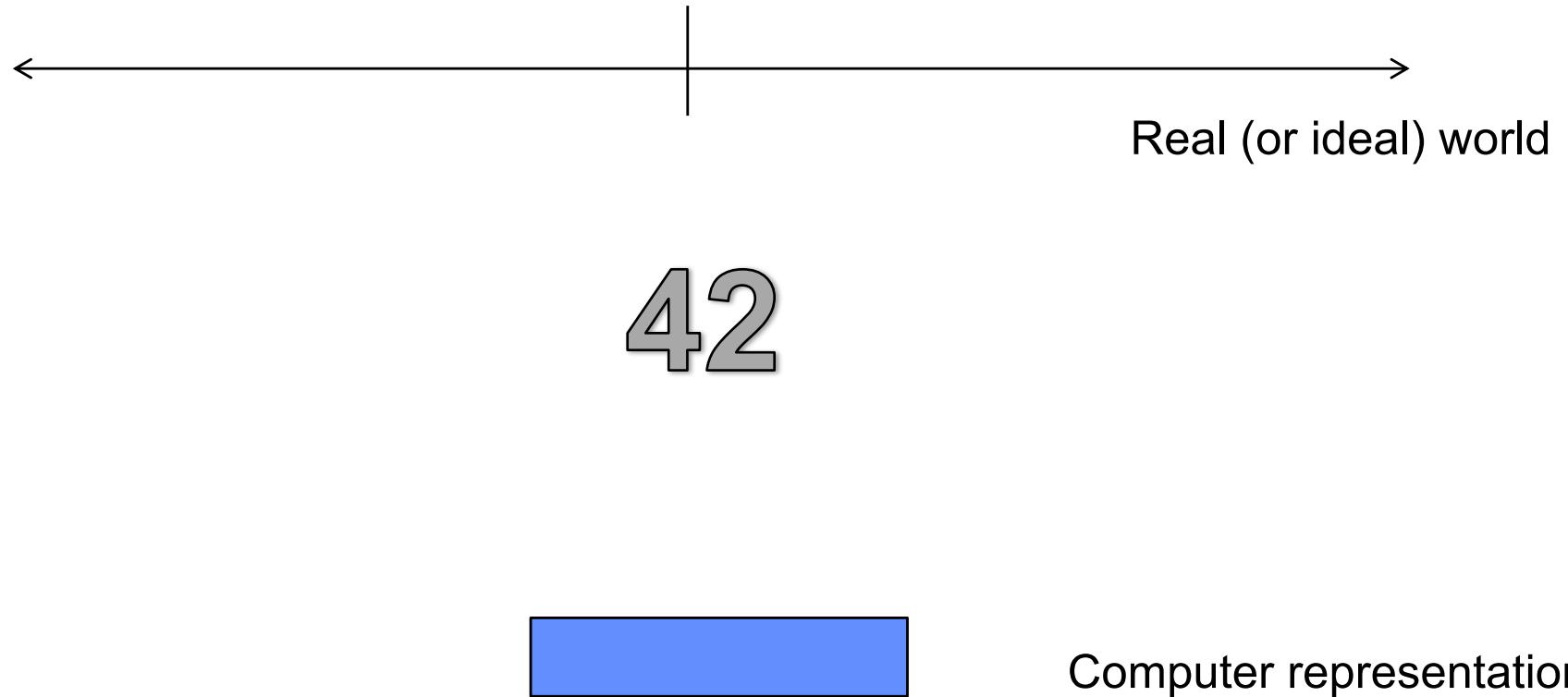
- Abstraction is not universal without loss of information (mathematically provable). This means, in the end, the complexity can only be “moved around”
- Abstraction makes us forget how things actually work and can therefore hide bias. Example: AI and hiring decisions.
- Abstraction makes things special and that creates dependencies. Dependencies grow longer and longer over time and can become unmanageable.





Abstraction in CS: Data Type

- What's this?





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 this week

- Lab will get you more practice with functions in Python
 - Make sure you've done Lab 1
- HW will give practice and explain subtleties of types, operators, and expressions
 - In a program development environment
- What's the difference between '==' and '=' ?



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?