EPBI 414

Unit 1
Course Orientation & Concepts

Introduction

- Instructor Background
- Course Info and Syllabus
- Data Concepts
- Basic Data Types
- Student Self-Assessment

Instructor Background

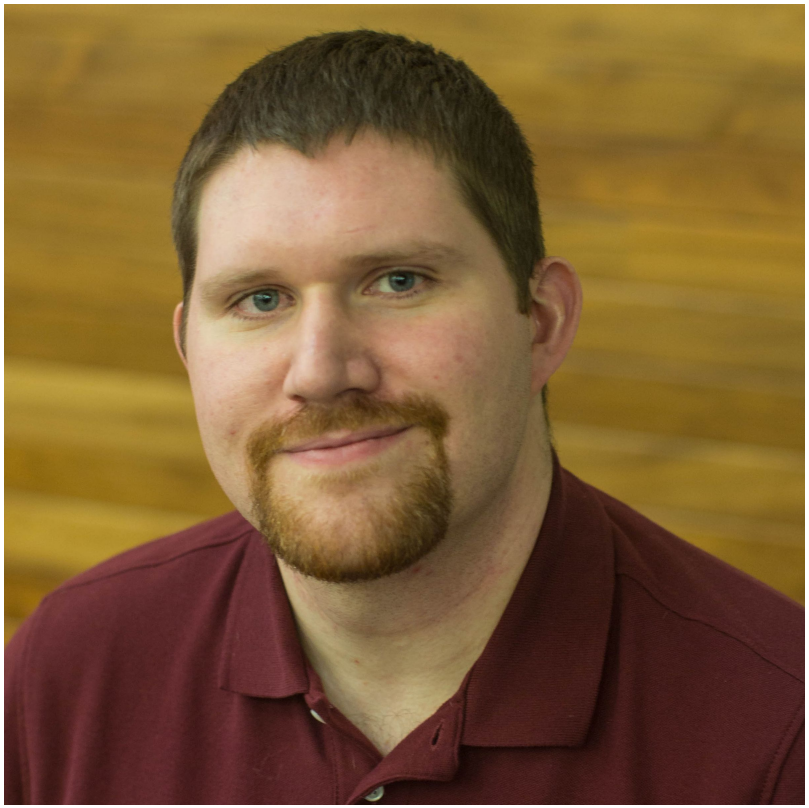
Thomas A. Rehman

- CWRU '07, '09
- PSU '11

Call me Tom

Living in Madison, WI

- Data scientist



Course Info and Syllabus

- Academic Honesty
- Email List & Course Website
- Blackboard
- Syllabus & Schedule
- Course Logistics

Academic Honesty

- Online course means unanticipated challenges
- When in doubt, *please ask* - don't risk your academic career
- Guidelines:
  - Homework: work together, submit your own work
  - Quizzes: open book, timed, ***on your own***

Email List

- Email is primary method of communication
- Will be using your CWRU email (xyz123@case.edu)
- Everyone will be enrolled on discussion list
  - Emails everyone, and preserves discussions
  - Also has a web interface if you would prefer

Course Website

- Contains slides, videos, and content
- Access limited to current students
- URL (for copy / paste):
  https://sites.google.com/a/case.edu/epbi414/

Blackboard

- Blackboard used for submissions and grades
- All due dates are given in Eastern time
- Please use email list for discussions and questions
- Check your access!
  - Course name may be wrong

Syllabus & Schedule

- Brief review of syllabus and schedule
- Please review it in depth on your own
- Inconsistencies? Tell me!
- URL (for copy / paste):
  https://sites.google.com/a/case.edu/epbi414/syllabus

Homework and Tests

- Weekly homework, due (generally) on Friday by 12:00PM
- Occasional open-book quizzes, done through Blackboard
- Proctored examinations - you find a proctor
  - Generally, students group together to do the exam

Course Logistics

- I live in Madison, Wisconsin
  - Time difference - Central vs. Eastern
- Class due dates / times always **Eastern**
- Assignments always submitted through Blackboard

More Logistics

- To reach me, the fastest and most reliable way will be my email: tar9@case.edu
- For course material questions (i.e. class questions), use the mailing list
  - If you have a question - someone else does too
- One-on-one office hours are available by request (Google Hangouts or phone)

Weekly Review Sessions

- Past experience shows that live review sessions help a lot
- During self-assessment, provide the best nights for you
- Attendance not required, but encouraged
- You come with questions!

What is data?

- From Merriam-Webster:
  - "facts or information used usually to calculate, analyze, or plan something"
  - "information that is produced or stored by a computer"
- Appropriate for a statistical programming class

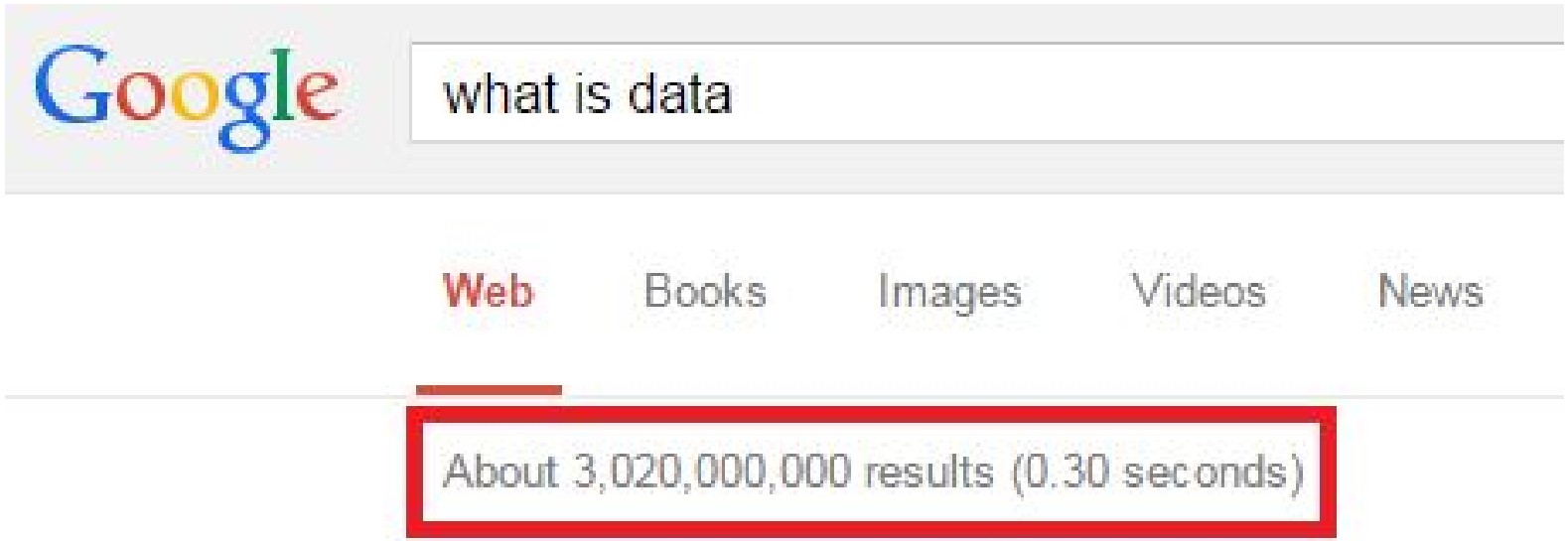
The misconception

When you ask what data is, you usually end up with something like this[1]:

```
> clinic_values
  month_year ob.hgba1c bf.hgba1c ej.hgba1c rv.hgba1c ob.bmi bf.bmi ej.bmi rv.bmi
1   Jan 2015       6.2       7.3       8.1       6.6   22.2   31.0   44.9   28.8
2   Feb 2015       6.4       7.5       7.5       6.9   22.6   33.0   41.1   28.6
3   Mar 2015       6.3       7.4       7.3       7.2   23.5   34.0   39.9   28.1
4   Apr 2015       6.1       7.3       6.9       7.4   23.3   33.0   39.3   26.6
5   May 2015       5.9       7.2       6.7       7.7   23.0   32.3   37.0   25.9
6   Jun 2015       5.8       7.3       6.5       7.9   22.9   31.1   35.5   25.8
> |
```

But what is data, really?

A million and one definitions (+/- billions)[2]



A working definition

- Data is information from the world, organized and interpreted through explicit **epistemic models**
  - From the Greek: *episteme*, or knowledge
  - Some models:
    - classification (putting things into groups)
    - quantification (counting the number of things)
    - network (mapping relationships between things)

Really? Ancient Greek?

- Before you work with data, you need to understand what it represents
- At the core: data is how we represent the chaotic world in an orderly and understandable way
- Chaos surrounds us, and thus data comes from all around us!

Why frameworks matter

- Data has structure and shape
  - A key challenge of a statistical programmer
- The structure is informed by what is being measured - what framework is being applied
  - For instance, measuring things in time, versus at a point
- One key role you will likely have: getting data into the computer

Data...comes from people[3]

Data...comes from the stars[4]

Data...comes from everything[5]

Our train station photo

- What data might interest us?
  - Things we can see: heights
  - Things we kind of see: sex and gender
  - Things we don't see: religious beliefs
  - Things maybe they don't see: disease status

Data in Computers

- As far as a computer cares, all data is numbers - computers only work on numbers ("computer" comes from "compute")
- BUT - this is not a computer engineering class
- Our concern is how we abstract data into the computer in a practical sense

Basic Data Types

- Divide into two general types of data
- Numbers, or *numeric* variables
- Characters, or *character variables*
- These are building blocks with which most data can be built

Assigning types

- Assigning a type is about how we represent information in the computer

- The right type makes the representation better
  - Makes what is represented clear
  - Makes working with data easier
  - Makes analysis more logical by limiting options

Types and the CRF

- We'll discuss the ***Case Report Form***, or CRF, throughout the course
- A key tool for collecting data during clinical research trials
- Proper design of a CRF means thinking about ***structure*** and ***type***

CRF as data framework

- The CRF is a great way to conceptualize data types
  - It is a designed mechanism for turning chaos into data points
- Poor design of CRFs means poorly captured data
- Good CRFs mean your life is much easier!

Some simple types

- The name of a subject, Alan
  - Could just put Alan down
  - But could make numeric: 1,12,1,14
- Alan's height, 132cm
  - If we write 132cm, is that numeric, or character?
  - Would "one hundred thirty-two centimeters" be better?

Example - all characters

```
name    height
----    ------
Alan    one hundred thirty-two centimeters
Steve   one hundred eighty-eight centimeters
Joan    one hundred twelve centimeters
Mike    seventy-five inches
```

Still all characters

```
name    height
----    ------
Alan    132cm
Steve   188cm
Joan    112cm
Mike    75in
```

Character and numeric

name	height
Alan	132
Steve	188
Joan	112
Mike	75

Better character & numeric

name	height	units
Alan	132	cm
Steve	188	cm
Joan	112	cm
Mike	75	in

Using numeric categories

name	height	units
Alan	132	1
Steve	188	1
Joan	112	1
Mike	75	2

Types are important

- There are good reasons to use character and numeric types
- Every language does things a little differently
- Core concept: think how data should be represented
- Is it better as a word, or a number?

Student Self-Assessment

- Your first homework: complete the student self-assessment
- Be honest - you get credit simply for responding
- Helps me to know everyone's knowledge level

Attributions

1. Image taken by author from author's example data, found in: https://github.com/TARehman/shiny_demo/blob/master/server.R

2. Image taken by author from Google search, 8/12/2015

3. Image is public domain and acquired from: https://commons.wikimedia.org/wiki/File:Taipei_MRT_Crowds.jpg

4. Image is public domain and acquired from: https://commons.wikimedia.org/wiki/File:Constellation_Fornax,_EXtreme_Deep_Field.jpg

5. Image is public domain and acquired from: https://commons.wikimedia.org/wiki/Banana#/media/File:Banana_chips.JPG