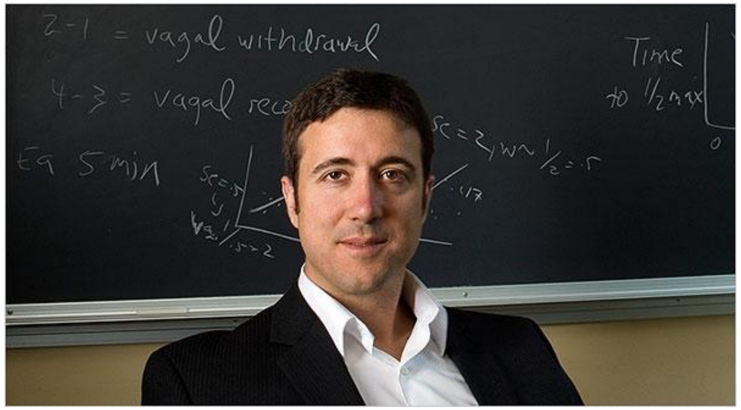
Lecture 02
Data

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Statistics

- "Statistics is about asking questions and using data to understand the answers." (Tor D. Wager)



- Professor, Psychology and Neuroscience, University of Colorado Boulder
- My PhD and postdoc advisor

Statistics

- “Statistics is about asking questions and using data to understand the answers.” (Tor D. Wager)
- “Statistics is the science of collecting, analyzing and interpreting data, or factual information.” (Martin A. Lindquist)

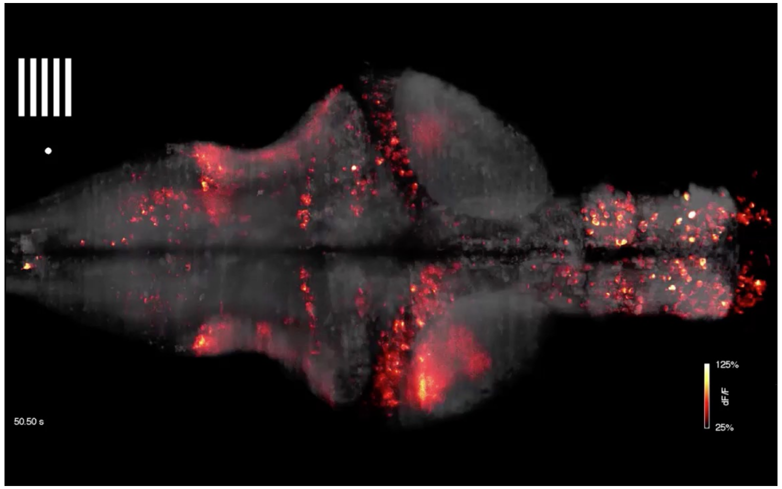
- Professor, Biostatistics, Johns Hopkins University
- Tor’s good friend and colleague, something like a co-advisor for me.

Statistics

- “Statistics is about asking questions and using data to understand the answers.” (Tor D. Wager)
- “Statistics is the science of collecting, analyzing and interpreting data, or factual information.” (Martin A. Lindquist)
- “Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.” (Textbook)
- Statistics is becoming more and more important. Why??

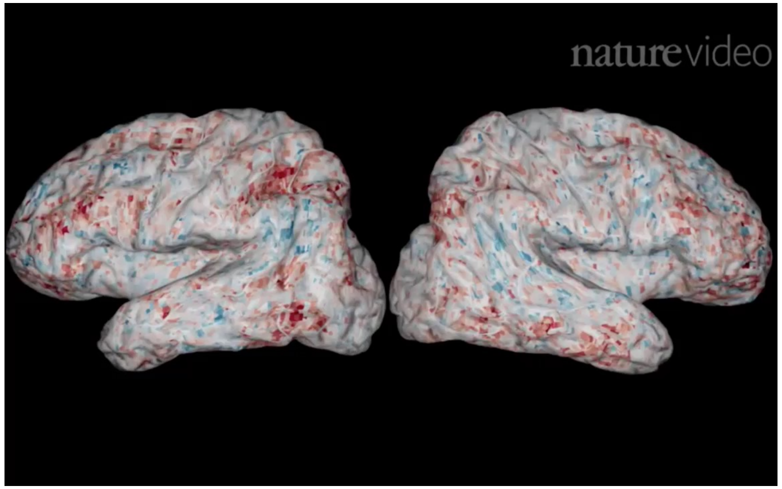
Big data era!

Light-sheet microscopy: a larval zebrafish



One-hour scan yields 1TB data.

Functional Magnetic Resonance Imaging



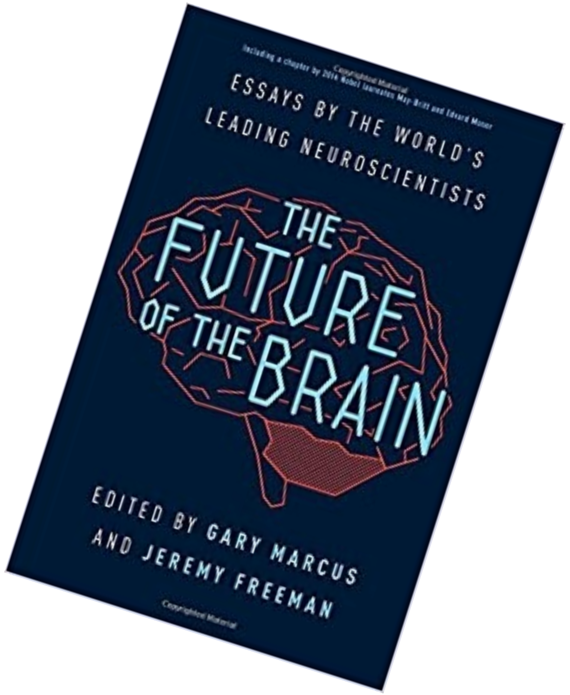
fMRI: whole-brain scan in 460ms with 2 mm^3 resolution

Videos from Freeman et al., 2014, *Nat Methods;* Huth et al., 2016, *Nature*

Big data era! What's the challenge?

Jeremy Freeman

Previously, Janelia Farm Research Campus
Currently, Chan-Zuckerberg Science Initiative



Including a chapter by 2014 Nobel laureates May-Britt and Edvard Moser
ESSAYS BY THE WORLD'S LEADING NEUROSCIENTISTS
THE FUTURE OF THE BRAIN
EDITED BY GARY MARCUS AND JEREMY FREEMAN

New technologies are now, finally, allowing us to probe the activity of thousands of neurons simultaneously while animals perform rich, ethologically relevant behaviors…

The first challenge to converting this newfound torrent of neural measurement into fundamental scientific meaning is to ask how to "make sense of the data."

Statistics!!!

Statistics

- “Statistics is about asking questions and using data to understand the answers.” (Tor D. Wager)
- “Statistics is the science of collecting, analyzing and interpreting data, or factual information.” (Martin A. Lindquist)
- “Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.” (Textbook)
- Statistics is a scientific way of searching for “meaning” out of data. (Choong-Wan Woo)

Statistics

- The problem is that people see only what they want to see!

  and statistics can serve as a lens to filter data!

From Martin's stat class:

- Three types of lies: Lies, damn lies and statistics.
- Always approach statistics with a critical eye.

From Tor's stat class:

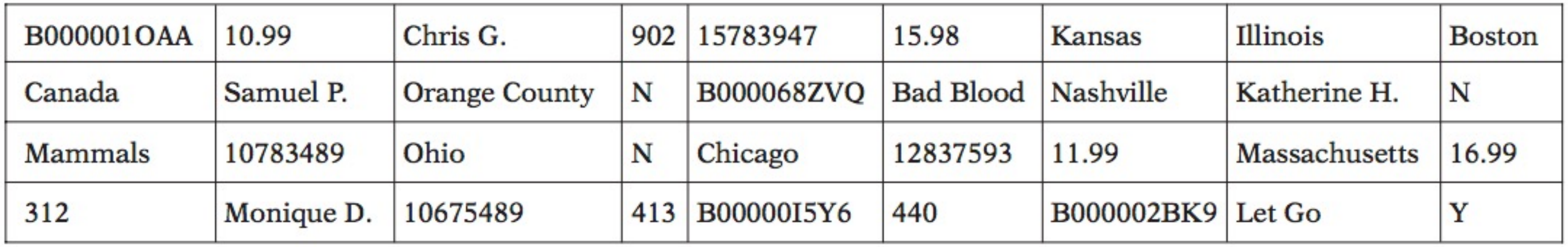
- "You can say anything with statistics"
- Statistics don't lie, but people can lie with statistics

A statistician's manifesto

(From T. Hastie, via J. McAuliffe, via Jordan Boyd-Graber)

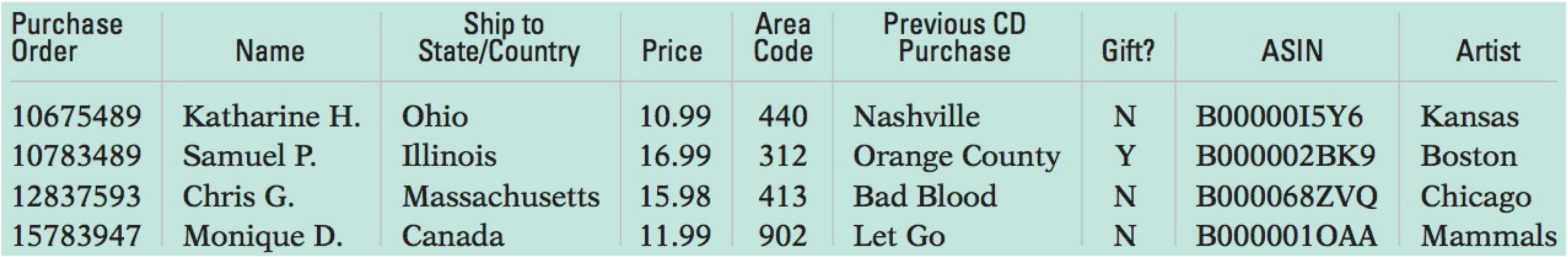
- **Understand the ideas behind the statistical methods**, so you know how to use them, when to use them, when not to use them.
- Complicated methods build on simple methods. **Understand simple methods first**.
- The results of a method are of little use without **an assessment of how well or poorly it is doing**.

Data



B0000010AA	10.99	Chris G.	902	15783947	15.98	Kansas	Illinois	Boston
Canada	Samuel P.	Orange County	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.	N
Mammals	10783489	Ohio	N	Chicago	12837593	11.99	Massachusetts	16.99
312	Monique D.	10675489	413	B00000I5Y6	440	B000002BK9	Let Go	Y

- What is the problem?
  - No context: *"We cannot make sense out of this table"*



Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B0000010AA	Mammals

- It is important to understand the context of your data set: " 5W1H"
  - WHO, WHAT, WHEN, WHERE, WHY and HOW.

Data

- It is important to understand the context of your data set: ”5W1H”
  - **WHO, WHAT, WHEN, WHERE, WHY and HOW.**

- We want to know who was measured, what was measured, how and where the data were collected and when and why the study was performed.

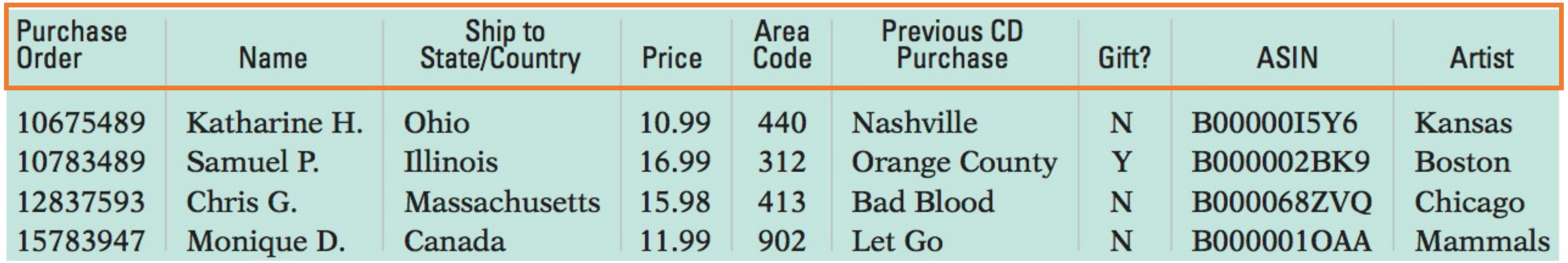
- Always keep track of the unit of measurement.

Data table

- Variables are the characteristics observed/recorded about each observation.

Variables: →
(column)

Observations →
(row) →

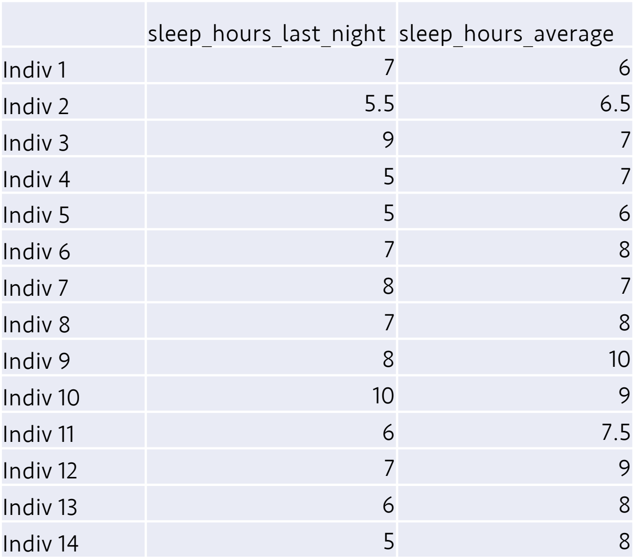


Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammals

- Let's make a data table using the data I collected last year:

Observations (row) →

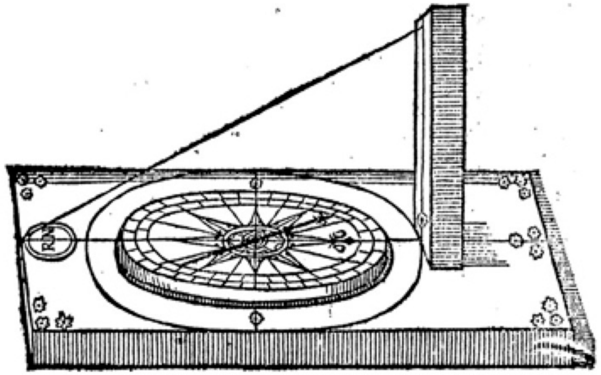
Variables
(column)



	sleep_hours_last_night	sleep_hours_average
Indiv 1	7	6
Indiv 2	5.5	6.5
Indiv 3	9	7
Indiv 4	5	7
Indiv 5	5	6
Indiv 6	7	8
Indiv 7	8	7
Indiv 8	7	8
Indiv 9	8	10
Indiv 10	10	9
Indiv 11	6	7.5
Indiv 12	7	9
Indiv 13	6	8
Indiv 14	5	8

Data table

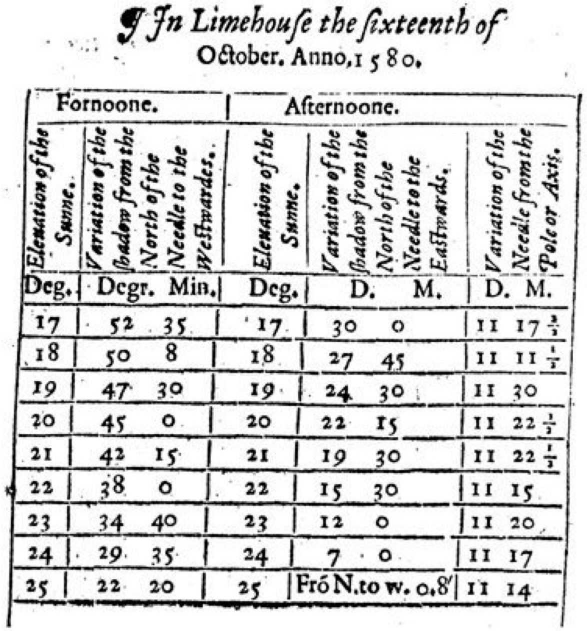
- It’s an old tradition.



Norman, 1581, *“A Discours of the Variation of the Cumpas, or Magneticall Needle.”*

Variables: → (column)

Observations → (row)



¶ Jn Limehouſe the ſixteenth of Octobe r. Anno,1580.

Fornoone.			Afternoone.				
Elevation of the Sunne.	Variation of the ſhadow from the North of the Needle to the Weſtwardes.		Elevation of the Sunne.	Variation of the ſhadow from the North of the Needle to the Eaſtwards.		Variation of the Needle from the Pole or Axis.	
Deg.	Degr.	Min.	Deg.	D.	M.	D.	M.
17	52	35	17	30	0	11	17 $\frac{2}{3}$
18	50	8	18	27	45	11	11 $\frac{1}{3}$
19	47	30	19	24	30	11	30
20	45	0	20	22	15	11	22 $\frac{2}{3}$
21	42	15	21	19	30	11	22 $\frac{1}{3}$
22	38	0	22	15	30	11	15
23	34	40	23	12	0	11	20
24	29	35	24	7	0	11	17
25	22	20	25	Fró N.to w. 0,8'		11	14

From Stigler, *“The Seven Pillars of Statistical Wisdom”*

Categorical and Quantitative variables

- A categorical variable names groups or categories into which an individual case might fall.
  - E.g., Gender, hair color, car model.

- A quantitative variable contains numerical values that are measured in units.
  - E.g., Height, age, miles per gallon for a car.

Identifiers and Ordinal variables

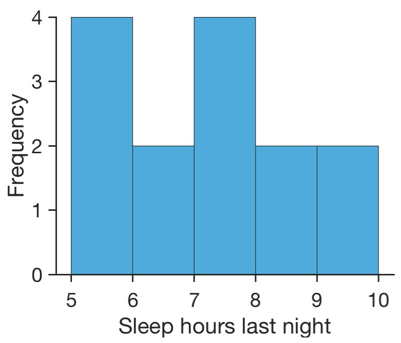
- An identifier is a unique value that each individual receives.
  - E.g., student ID, customer number
  - Personally identifiable information: any data that could potentially identify a specific individual.
    - E.g., social security number
    - For human data, we are usually required to use de-identified data. Identifiers are needed to link one dataset to other datasets. These are called relational database.

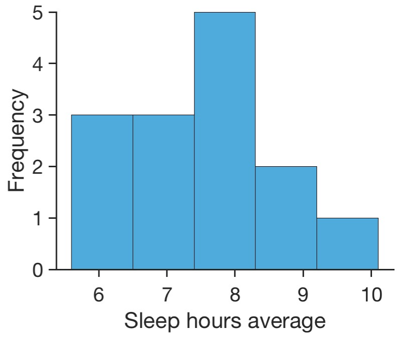
- An ordinal variable report order without natural units
  - E.g., How valuable do you think this course will be to you?
    - 1 = Worthless, 2 = Slightly, 3 = Middling, 4 = Reasonably, 5 = Invaluable

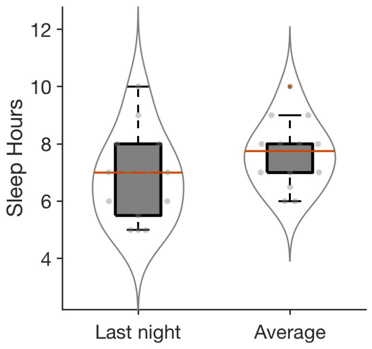
Distribution

- In statistics, variables typically have distributions.
- The distribution of a variable tells us what values it takes and how often it takes these values.

	sleep_hours_last_night	sleep_hours_average
Indiv 1	7	6
Indiv 2	5.5	6.5
Indiv 3	9	7
Indiv 4	5	7
Indiv 5	5	6
Indiv 6	7	8
Indiv 7	8	7
Indiv 8	7	8
Indiv 9	8	10
Indiv 10	10	9
Indiv 11	6	7.5
Indiv 12	7	9
Indiv 13	6	8
Indiv 14	5	8







Key Points

- Chapter 1
  - What is statistics?

- Chapter 2
  - Who, what, how, where, when, why
  - Data layout: Cases and variables
  - Categorical and Quantitative variables
  - Identifier, ordinal variables