Lecture 1
Course Introduction

Dennis Sun
Stanford University
DATASCI / STATS 112

January 9, 2023

1. Background

2. Course Logistics

3. Course Overview

4. A Look Ahead

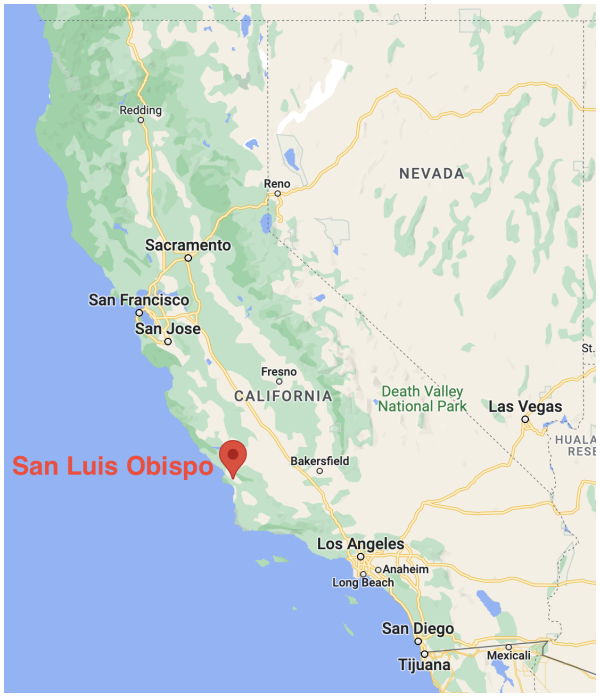
1 Background

2 Course Logistics

3 Course Overview

4 A Look Ahead

About Me



- I am a new professor at Stanford in the Statistics department.
- I was a Ph.D. student at Stanford from 2010-2015.
- I have been a professor at Cal Poly, San Luis Obispo.
- I am also concurrently a Data Scientist at Google.

About this Course

- DATASCI / STATS 112 is a new course at Stanford.
- The plan is to make this course a gateway for the B.S. in Data Science (the new name for the MCS major).
- For now, it counts as an elective for the B.S., as well as for the "Applied Quantitative Reasoning" WAYS requirement.
- (I will be talking to departments to get this course to count as a major elective, but no guarantees.)
- This course is based on a Data Science course I designed and taught at Cal Poly.

How is this Course Different?

- Unlike most data science or machine learning classes on campus, DATASCI 112 has no math or statistics prereqs.
- To do data science, you need to know how to program (in Python). But just a little. So CS 106A is a prereq.
- Don't mistake this class for a "baby" version of Data Science. You'll be prepared for most Data Scientist internships after taking this class.
- This class will hint at many mathematics and statistics connections. This will hopefully motivate you to take classes like MATH 51 and STATS 116, if you haven't already!

1 Background

2 Course Logistics

3 Course Overview

4 A Look Ahead

Course Website

Your one-stop shop for everything related to this course:
`https://dlsun.github.io/stats112`

Please bookmark this page.

(You don't even need to look on Canvas!)

Course Requirements

- Lectures on MWF will introduce the concepts. Exercises will be assigned after lectures on MW.
- Sections on TuTh are for presenting and discussing solutions to the exercises. Participation is required and counts toward your grade.
- There will be weekly homework.
- There will be two in-class midterms.
- There will be a final project. This project is an opportunity for you to explore a data set that is meaningful to you and start to build a portfolio.

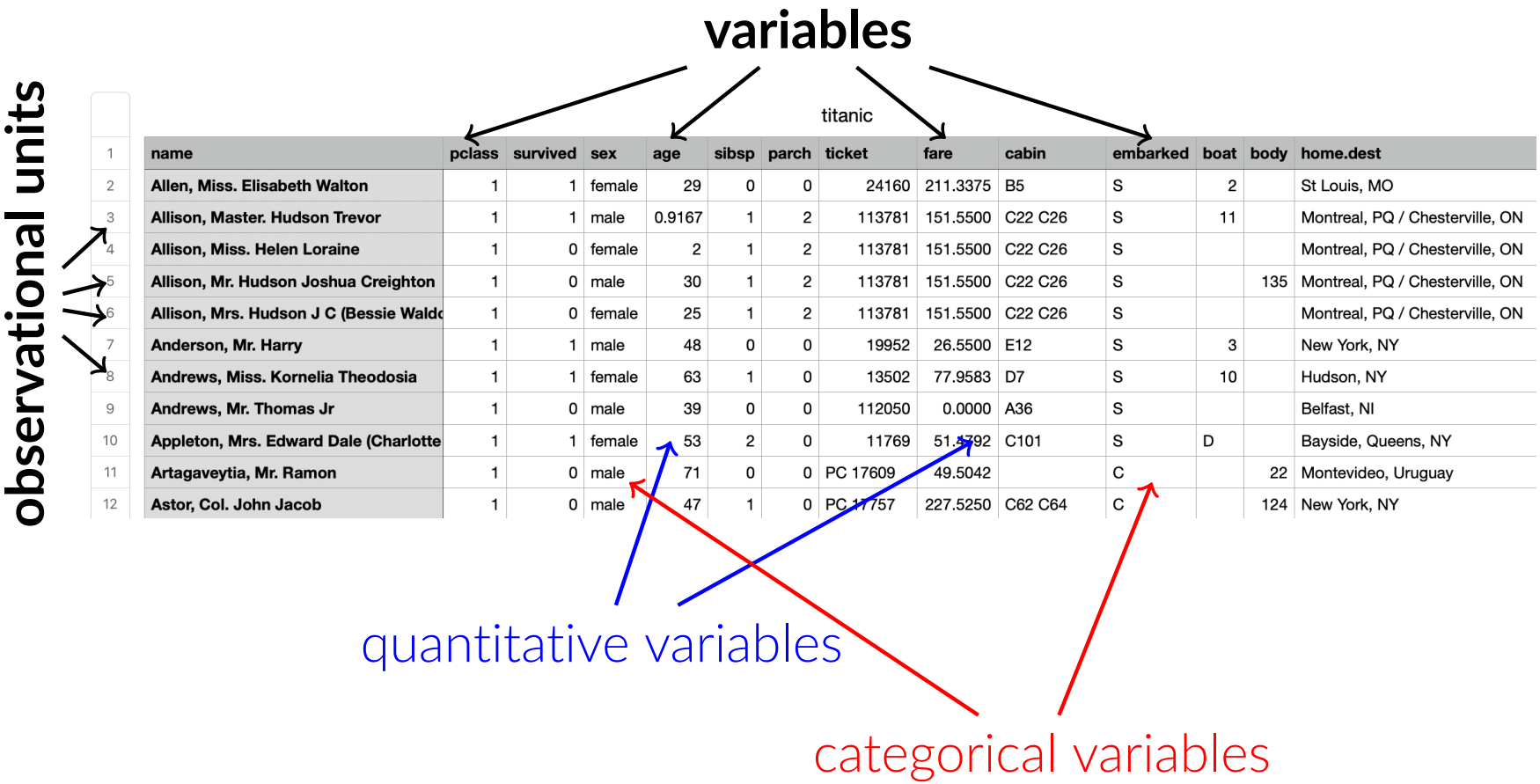
1 Background

2 Course Logistics

3 Course Overview

4 A Look Ahead

What Does Data Look Like?



titanic

	name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
1	Allen, Miss. Elisabeth Walton	1	1	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
2	Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville, ON
3	Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville, ON
4	Allison, Mr. Hudson Joshua Creighton	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville, ON
5	Allison, Mrs. Hudson J C (Bessie Waldo	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville, ON
6	Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
7	Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
8	Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
9	Appleton, Mrs. Edward Dale (Charlotte	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
10	Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
11	Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY

This kind of data is called **tabular data**.

Course Outline

DATASCI / STATS 112 is divided into roughly three units:

1. tabular data (analyzing and visualizing)
2. other shapes of data: text, images, maps
3. machine learning

How to Represent Data on Disk?

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ
Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ
Allison, Mr. Hudson Joshua Creighton	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ
Allison, Mrs. Hudson J C (Bessie Waldo	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ
Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlotte	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Qu
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo,
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY

```
name,pclass,survived,sex,age,sibsp,parch,ticket,fare,cabin,embarked,boat,body,home.dest
"Allen, Miss. Elisabeth Walton",1,1,female,29,0,0,24160,211.3375,B5,S,2,,"St Louis, MO"
"Allison, Master. Hudson Trevor",1,1,male,0.9167,1,2,113781,151.5500,C22 C26,S,11,,"Mont
"Allison, Miss. Helen Loraine",1,0,female,2,1,2,113781,151.5500,C22 C26,S,,,"Montreal, P
"Allison, Mr. Hudson Joshua Creighton",1,0,male,30,1,2,113781,151.5500,C22 C26,S,,135,"M
"Allison, Mrs. Hudson J C (Bessie Waldo Daniels)",1,0,female,25,1,2,113781,151.5500,C22
"Anderson, Mr. Harry",1,1,male,48,0,0,19952,26.5500,E12,S,3,,"New York, NY"
"Andrews, Miss. Kornelia Theodosia",1,1,female,63,1,0,13502,77.9583,D7,S,10,,"Hudson, NY
"Andrews, Mr. Thomas Jr",1,0,male,39,0,0,112050,0.0000,A36,S,,,"Belfast, NI"
"Appleton, Mrs. Edward Dale (Charlotte Lamson)",1,1,female,53,2,0,11769,51.4792,C101,S,D
"Artagaveytia, Mr. Ramon",1,0,male,71,0,0,PC 17609,49.5042,,C,,22,"Montevideo, Uruguay"
"Astor, Col. John Jacob",1,0,male,47,1,0,PC 17757,227.5250,C62 C64,C,,124,"New York, NY"
```

Comma-Separated Values (CSV) Format

How to Represent Data in Python?

```
[2] # Read in the Titanic data set using the Pandas `read_csv` function.
    df_titanic = pd.read_csv("https://dlsun.github.io/stats112/data/titanic.csv")

    # To look at the data, we make `df_titanic` the last line of the cell so that
    # the output is printed.
    df_titanic
```

	name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	Allen, Miss. Elisabeth Walton	1	1	female	29.0000	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	Allison, Miss. Helen Loraine	1	0	female	2.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	Allison, Mr. Hudson Joshua Creighton	1	0	male	30.0000	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	1	0	female	25.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1304	Zabour, Miss. Hileni	3	0	female	14.5000	1	0	2665	14.4542	NaN	C	NaN	328.0	NaN
1305	Zabour, Miss. Thamine	3	0	female	NaN	1	0	2665	14.4542	NaN	C	NaN	NaN	NaN
1306	Zakarian, Mr. Mapriededer	3	0	male	26.5000	0	0	2656	7.2250	NaN	C	NaN	304.0	NaN
1307	Zakarian, Mr. Ortin	3	0	male	27.0000	0	0	2670	7.2250	NaN	C	NaN	NaN	NaN
1308	Zimmerman, Mr. Leo	3	0	male	29.0000	0	0	315082	7.8750	NaN	S	NaN	NaN	NaN

1309 rows × 14 columns

All code in this class will be written in **notebooks**.

We will use a free service called **Colab**, which stores these notebooks on Google Drive.

1 Background

2 Course Logistics

3 Course Overview

4 A Look Ahead

Sections This Week

You should already been enrolled in one of 5 sections.

- Check your enrollment for the room and time.
- Your TA is listed on the course website.
- Tomorrow's section is *optional*. Take a look at the materials, and see if you would benefit from a walkthrough.
  (You may attend any section, not just the one you are enrolled in.)
- Starting Thursday, you are expected to attend the section you are enrolled in. (If you have a conflict, please e-mail your TA.)
- Bring your laptops to section so that you can follow along!

Office Hours

- Check the website for the latest office hours.
- I will have office hours today from 3 - 4 PM in Sequoia 124. Please drop by if you have questions about the course!

I am excited about this class, and I hope you are too!

If you have any friends whom you think would enjoy DATASCI 112, please let them know about it. It's not too late to enroll!

See you on Wednesday!