Lecture 1
Introduction to Data Science

Dennis Sun (modified by Jonathan Taylor)
Stanford University

DATASCI 112

March 30, 2026

1. Background
2. Course Logistics
3. Introduction to Data
4. A Look Ahead

1 Background

2 Course Logistics

3 Introduction to Data

4 A Look Ahead

About this Course

- **DATASCI 112** is the gateway course for the B.A. and the B.S. in Data Science.
- This course is designed for freshmen and sophomores who are exploring Data Science as a major, but everyone is welcome!

How is this Course Different?

- Unlike most data science or machine learning classes on campus, **DATASCI 112** has no math or statistics prereqs.
- To begin doing data science, you need to know how to program (a bit). So **CS 106A** is a prereq.
- But you don't need a lot of math. We will rely on geometric intuition in this class.
- But there are many mathematics connections, which I will hint at and which I hope will encourage you to take classes like **MATH 51** (linear algebra) and **STATS 117** (probability).

Don't mistake this class for a "baby" version of Data Science. You'll be prepared for most "Data Scientist" and "Data Analyst" internships after taking this class.

1 Background

2 Course Logistics

3 Introduction to Data

4 A Look Ahead

Course Website

Your one-stop shop for everything related to this course today is
`https://datasci112.stanford.edu`

Please bookmark this page.

(P.S. There is no Canvas for this course!)

Course Requirements

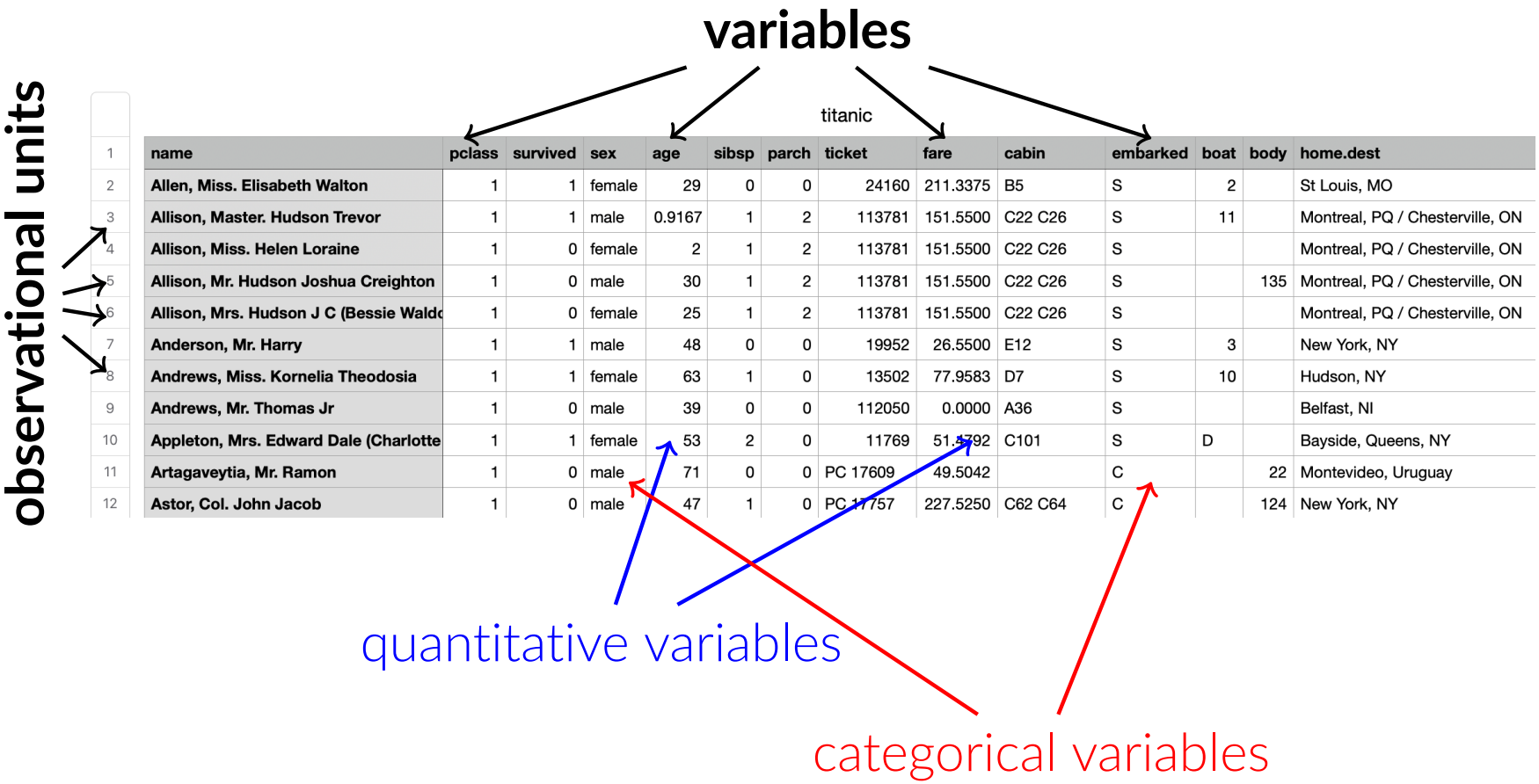
- Lectures on MWF will *introduce* the concepts.
- Sections on TuTh will *reinforce* the concepts.
  - Before each section, exercises will be posted on the course website.
  - In section, you will present and discuss solutions to these exercises.
  - Participation is required and counts $\mathbf{15}\%$ toward your grade.

Course Requirements

- There will be 5 labs, each consisting of two parts, counting **15**% toward your grade.
  - One part will be assigned every two lectures and due one week later at 8 AM.
  - No extensions will be granted under any circumstances, so please plan ahead.
  - There will be an optional Lab 6 due during Week 10 that will replace your lowest lab part.
- There will be two 50-minute in-class midterms, counting **35**% toward your grade.
- Instead of a final exam, there will be a final project, counting **35**% toward your grade.
  - This is an opportunity for you to explore a data set that is meaningful to you and start a portfolio.
  - There will be a poster session during the scheduled final exam time.

More information about grading on the course website.

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

Data like this, that can be stored in a spreadsheet, is called **tabular data**.

How is Tabular Data Represented 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,,"Montreal, PQ / Chester
"Allison, Miss. Helen Loraine",1,0,female,2,1,2,113781,151.5500,C22 C26,S,,,"Montreal, PQ / Chesterville,
"Allison, Mr. Hudson Joshua Creighton",1,0,male,30,1,2,113781,151.5500,C22 C26,S,,135,"Montreal, PQ / Ches
"Allison, Mrs. Hudson J C (Bessie Waldo Daniels)",1,0,female,25,1,2,113781,151.5500,C22 C26,S,,,"Montreal
"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,,"Bayside, Queens
"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 is Tabular Data Represented in Python?

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, N
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, N

`DataFrame`

Let's interact with this data using Python in a **notebook**.

All of our code will be written in Colab notebooks like this one.

Review: Categorical Variables

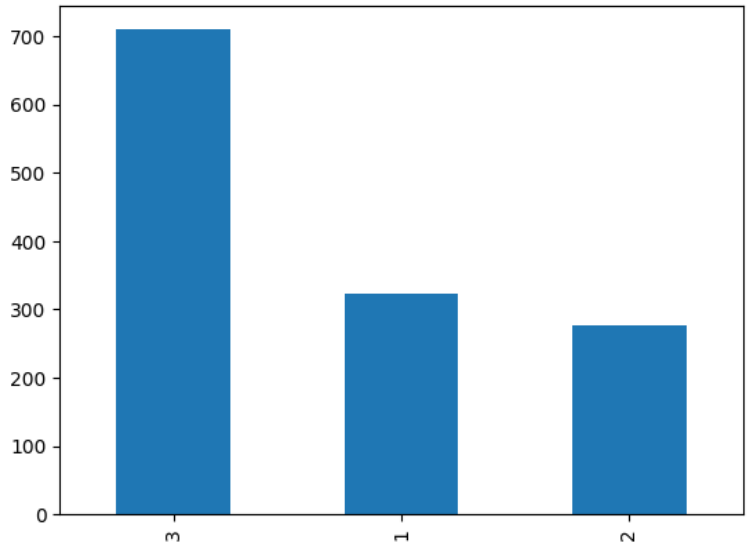
To *summarize* a categorical variable, we report the **counts** of each possible category.

```
df["pclass"].value_counts()
```

```
3    709
1    323
2    277
Name: pclass, dtype: int64
```

To *visualize* a categorical variable, we make a **bar plot**.

```
df["pclass"].value_counts().plot.bar()
```



Hmm...why are the classes out of order?

Review: Categorical Variables

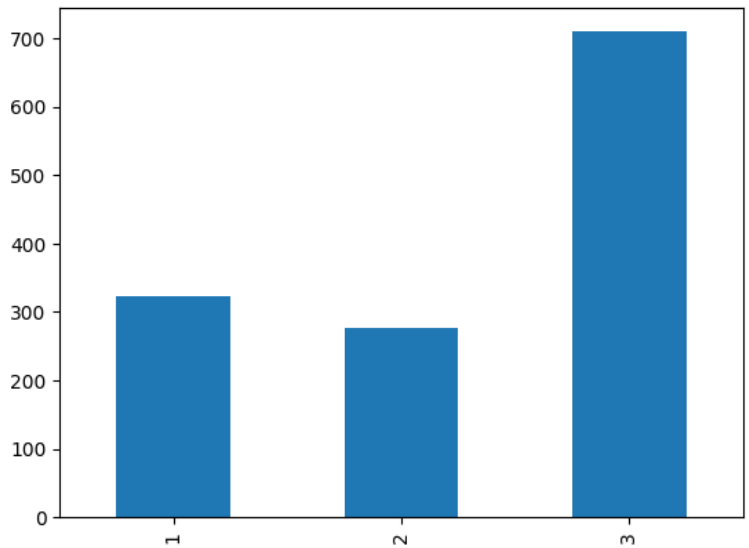
To *summarize* a categorical variable, we report the **counts** of each possible category.

```
df["pclass"].value_counts()
```

```
3    709
1    323
2    277
Name: pclass, dtype: int64
```

To *visualize* a categorical variable, we make a **bar plot**.

```
df["pclass"].value_counts().sort_index().plot.bar()
```



Notice that we can chain methods, one after the other.

1 Background

2 Course Logistics

3 Introduction to Data

4 A Look Ahead

The Three Parts of DATASCI 112

To paraphrase a famous historical figure,
"**DATASCI 112** est omnis divisa in partes tres..."

1. summarizing and visualizing tabular data
2. other shapes of data: textual, hierarchical, geospatial
3. machine learning

Sections This Week

You should already been enrolled in one of 6 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 Colab, 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 and CA.)
- Bring your laptops to section so that you can follow along!

- Office hours start tomorrow.
- The course website will be updated with the times and locations.
- Come see me after class if you have any questions about enrolling in this course.

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

See you on Wednesday!