Data as a Resource

David Bamman
Info 202: Information Organization and Retrieval

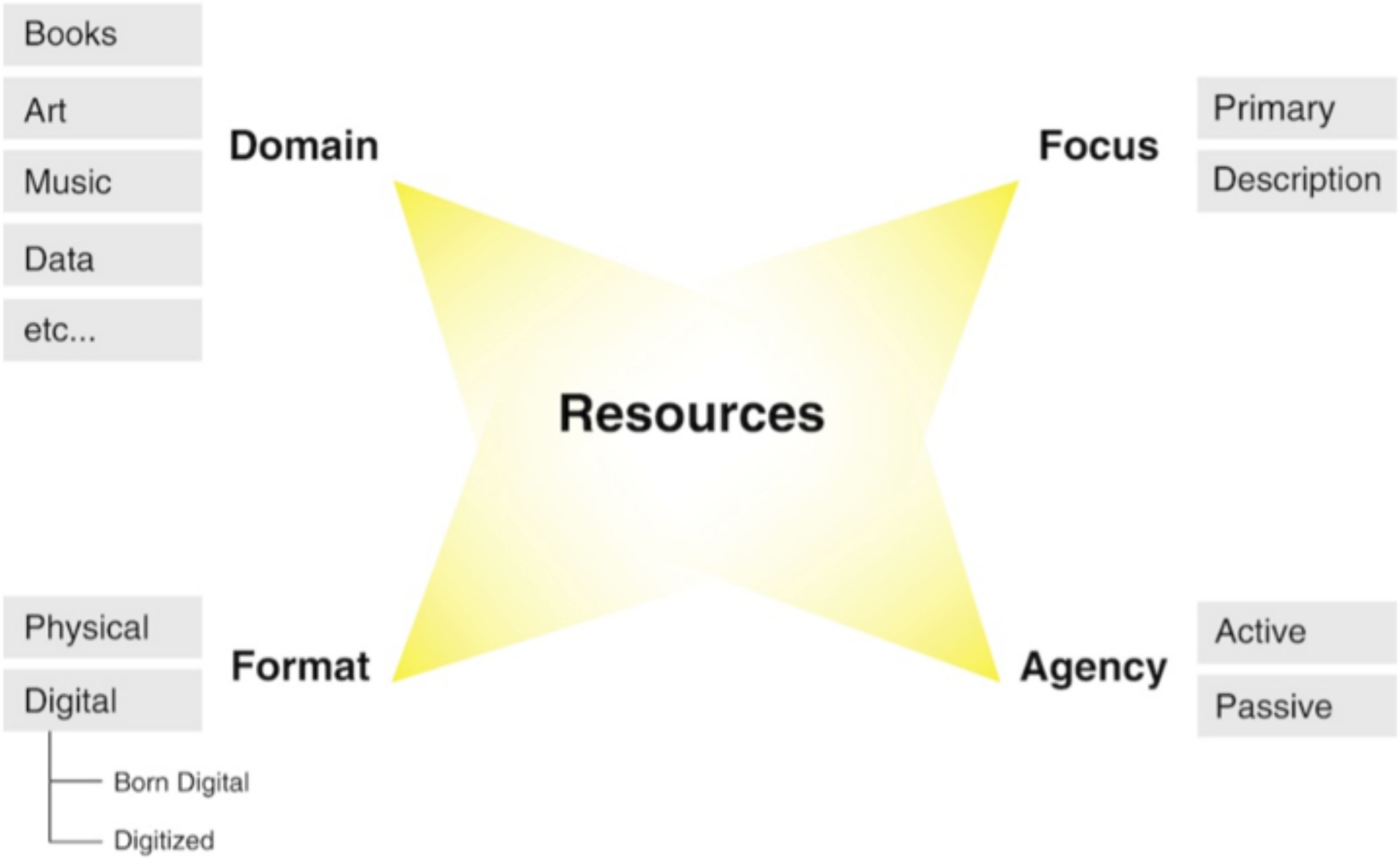
September 19, 2016

Data as resource

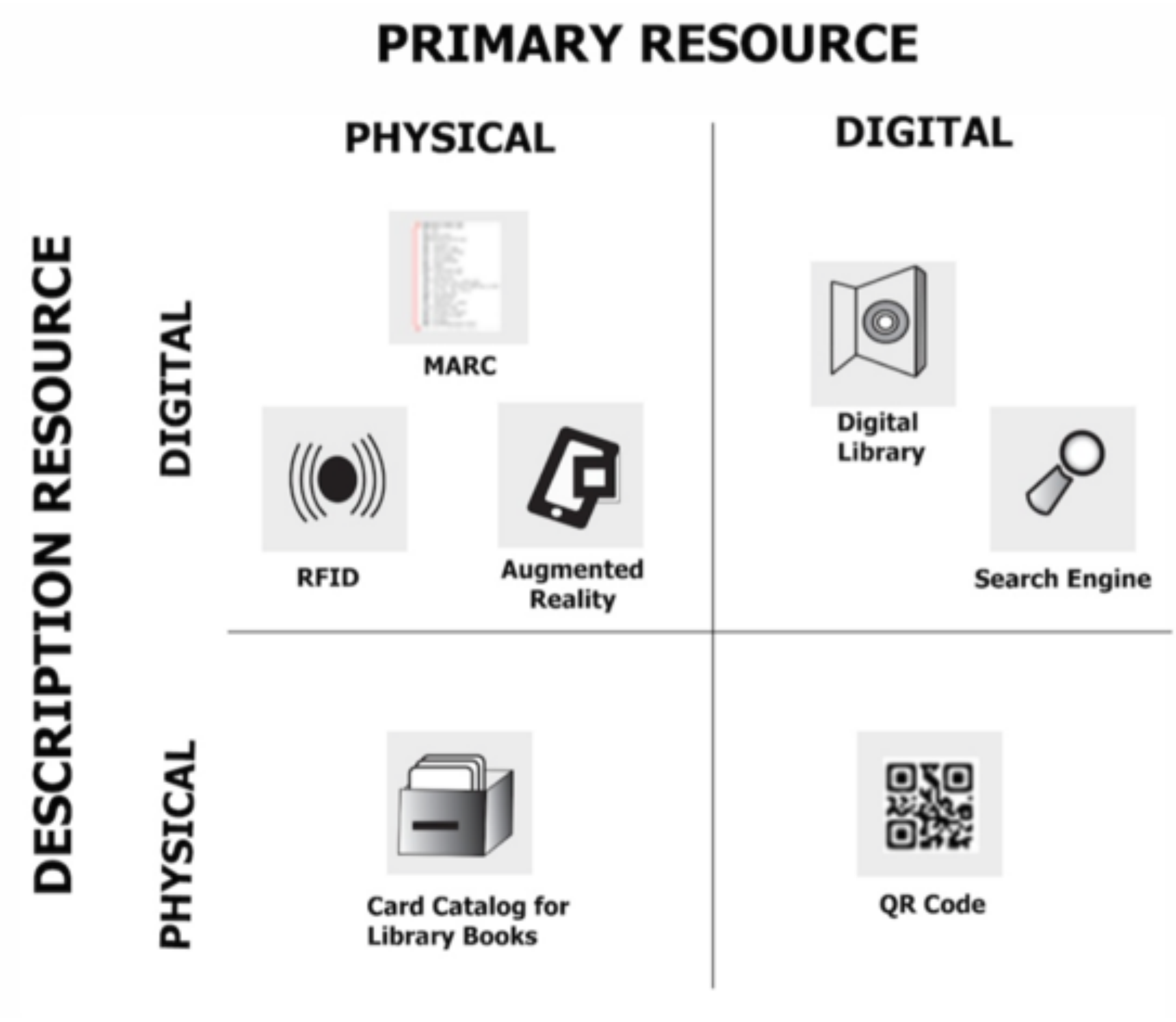
- Data as a resource to be organized
    - data analysis
    - sensor data streams
    - internet of things

“data” science

Data as resource



Format x Focus



Data as a resource

- In many analyses, the data you have is not what you actually care about; it’s the population they represent.
- What you organize is your knowledge about the population.

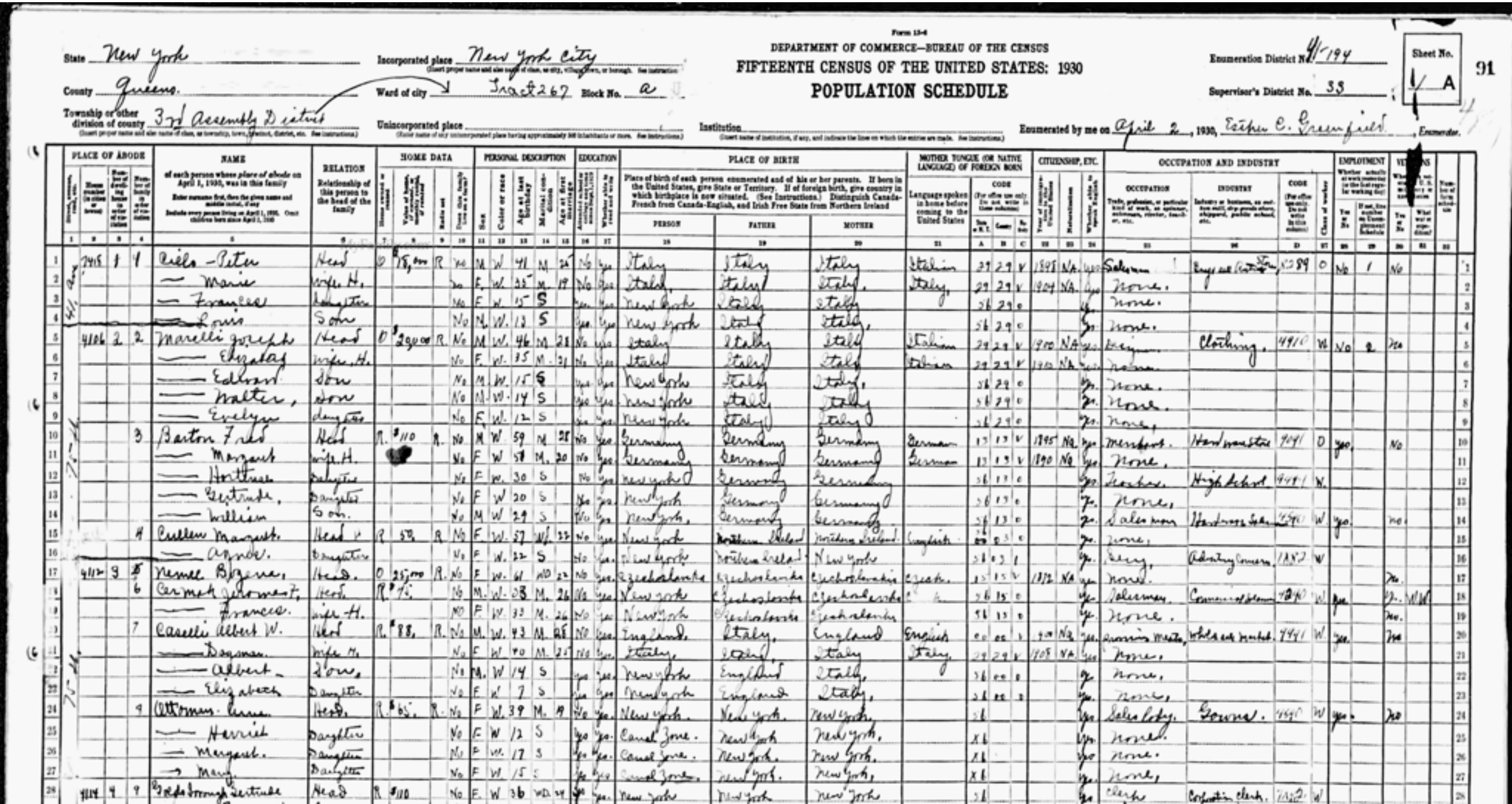
Identity

- What is the unit of analysis?
- Population vs. sample

Identity

Average age of Shamu?

Population



Form 15-6

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS

FIFTEENTH CENSUS OF THE UNITED STATES: 1930

POPULATION SCHEDULE

State New York

County Queens.

Township or other division of county 3rd Assembly District

Incorporated place New York City

Ward of city Tract 267 Block No. a

Unincorporated place

Institution

Enumeration District No. 41-194

Supervisor's District No. 33

Sheet No. 1 A

91

Enumerated by me on April 2, 1930, Esther E. Greenfield, Enumerator.

	PLACE OF ABODE				NAME	RELATION	HOME DATA				PERSONAL DESCRIPTION					EDUCATION		PLACE OF BIRTH			MOTHER TONGUE (OR NATIVE LANGUAGE) OF FOREIGN BORN	CODE			CITIZENSHIP, ETC.			OCCUPATION AND INDUSTRY				EMPLOYMENT		VETERANS		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18 PERSON	19 FATHER	20 MOTHER	21	A	B	C	22	23	24	25 OCCUPATION	26 INDUSTRY	D	27	28	29	30	31	32
1	41 Ave	4115	1	1	Aiello - Peter	Head	O	$18,000	R	No	M	W	41	M	25	No	Yes	Italy	Italy	Italy	Italian	29	29	V	1898	NA	Yes	Salesman	Bags and Auto Store	5389	O	No	1	No		
2					— Marie	Wife H.				No	F	W	35	M	19	No	Yes	Italy	Italy	Italy	Italy	29	29	V	1904	NA	Yes	None.								
3					— Frances	Daughter				No	F	W	15	S		Yes	Yes	New York	Italy	Italy		56	29	0			Yes	None.								
4					— Louis	Son				No	M	W	13	S		Yes	Yes	New York	Italy	Italy		56	29	0			Yes	None.								
5		4106	2	2	Marelli Joseph	Head	O	$29000	R	No	M	W	46	M	28	No	Yes	Italy	Italy	Italy	Italian	29	29	V	1900	NA	Yes	Designer	Clothing.	4910	W	No	2	No		
6					— Elizabeth	Wife H.				No	F	W	35	M	21	No	Yes	Italy	Italy	Italy	Italian	29	29	V	1902	NA	Yes	None.								
7					— Edward	Son				No	M	W	15	S		Yes	Yes	New York	Italy	Italy		56	29	0			Yes	None.								
8					— Walter,	Son				No	M	W	14	S		Yes	Yes	New York	Italy	Italy		56	29	0			Yes	None.								
9					— Evelyn	daughter				No	F	W	12	S		Yes	Yes	New York	Italy	Italy		56	29	0			Yes	None,								
10	75 St			3	Barton Fred	Head	R	$110	R	No	M	W	59	M	28	No	Yes	Germany	Germany	Germany	German	13	13	V	1895	Na	Yes	Merchant.	Hardware Store	9041	O	Yes,		No		
11					— Margaret	Wife H.				No	F	W	57	M	20	No	Yes	Germany	Germany	Germany	German	13	13	V	1890	Na	Yes	None.								
12					— Hortense	Daughter				No	F	W	30	S		No	Yes	New York	Germany	Germany		56	13	0			Yes	Teacher.	High School	9441	W.					
13					— Gertrude.	Daughter				No	F	W	20	S		No	Yes	New York	Germany	Germany		56	13	0			Yes	None,								
14					— William	Son.				No	M	W	29	S		No	Yes	New York.	Germany	Germany		56	13	0			Yes	Salesman	Hardware Sales	4540	W	Yes.		No		
15				4	Cullen Margaret.	Head V	R	$50	R	No	F	W	57	Wd	22	No	Yes	New York	Northern Ireland	Northern Ireland.	English	56	03	0			Yes	None,								
16					— Agnes.	Daughter				No	F	W	22	S		No	Yes	New York	Northern Ireland	New York		56	03	1			Yes	Secy,	Advertising Concern.	1882	W					
17		4112	9	5	Nemec Bozena,	Head.	O	$25,000	R.	No	F	W.	61	Wd	22	No	Yes	Czechoslovakia	Czechoslovakia	Czechoslovakia	Czech.	15	15	V	1892	NA	Yes	None.						No.		
18				6	Cermak Jerome F.	Head.	R	$75.		No	M.	W.	33	M.	26	No	Yes	New York	Czechoslovakia	Czechoslovakia		56	15	0			Yes	Salesman.	Commercial ...	4540	W.	Yes		Yes.	WW	
19					— Frances.	Wife H.				No	F	W.	33	M.	26	No	Yes	New York	Czechoslovakia	Czechoslovakia		56	15	0			Yes	None.						No.		
20				7	Caselli Albert W.	Head	R.	$88.	R.	No	M.	W.	43	M.	25	No	Yes	England.	Italy.	England	English	00	00	1	1910	Na	Yes	Receiving Master,	Wholesale market.	7441	W.	Yes.		No		
21					— Dagmar.	Wife H.				No	F	W	40	M.	25	No	Yes	Italy,	Italy,	Italy	Italy	29	29	V	1908	NA	Yes	None.								
22					— Albert.	Son,				No	M.	W	14	S		Yes	Yes	New York	England	Italy,		56	00	0			Yes	None,								
23					— Elizabeth	Daughter				No	F	W	7	S		Yes	Yes	New York	England	Italy,		56	00	2			Yes	None,								
24				8	Ottman Anna	Head.	R.	$65.	R.	No	F	W.	39	M.	19	No	Yes	New York.	New York.	New York.		56					Yes	Saleslady.	Gowns.	4540	W	Yes.		No		
25					— Harriet	Daughter				No	F	W	12	S		Yes	Yes	Canal Zone.	New York	New York,		X6					Yes	None.								
26					— Margaret.	Daughter				No	F	W.	17	S		Yes	Yes	Canal Zone.	New York.	New York.		X6					Yes	None.								
27					— Mary	Daughter				No	F	W	15	S		Yes	Yes	Canal Zone.	New York.	New York,		X6					Yes	None,								
28		4114	4	9	Goldsborough Gertrude	Head	R	$100		No	F.	W	36	Wd.	24	Yes	Yes	New York	New York	New York		56					Yes	Clerk	Corporation clerk.	7452	W					

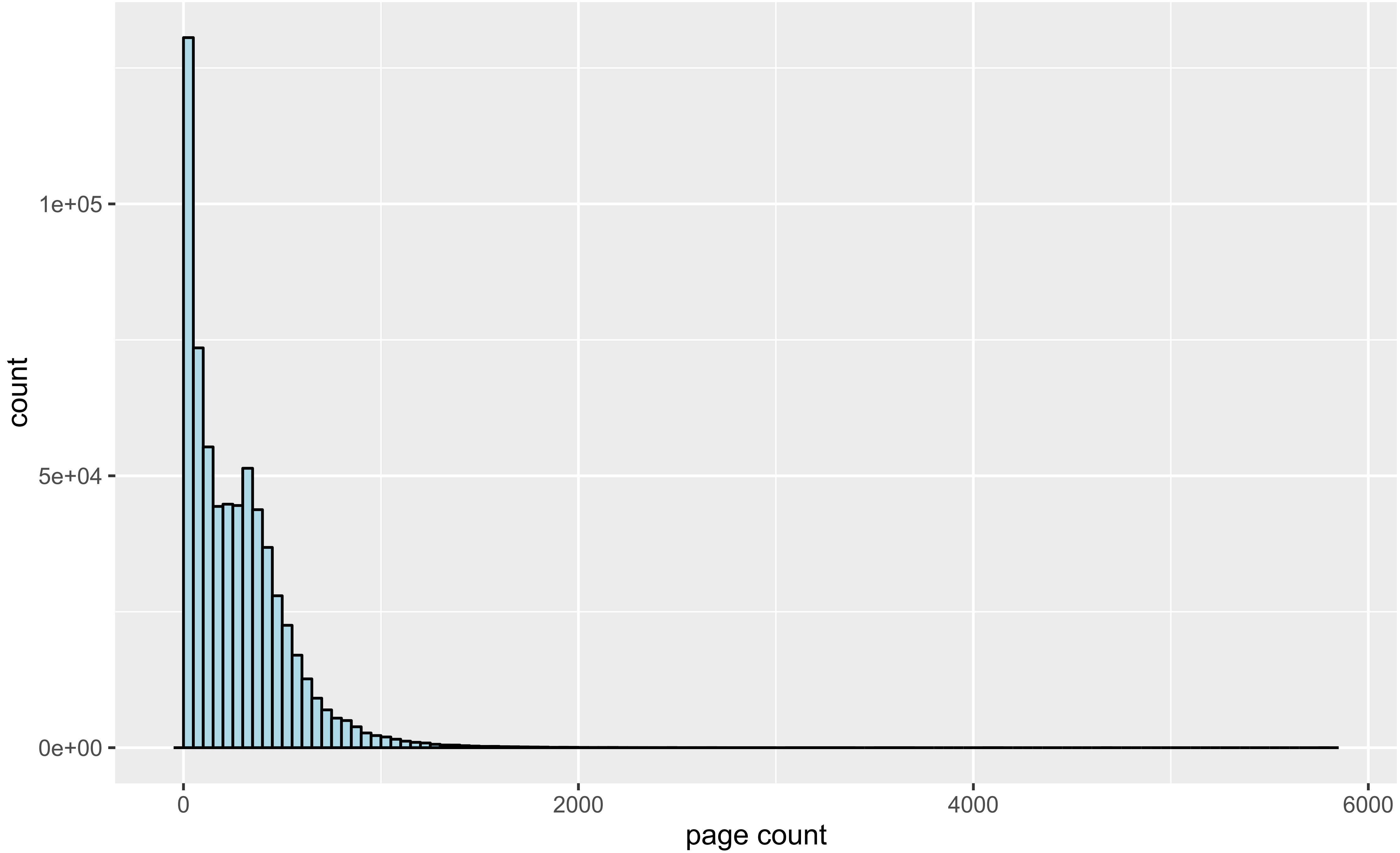
Population

population vs. hypothetical population

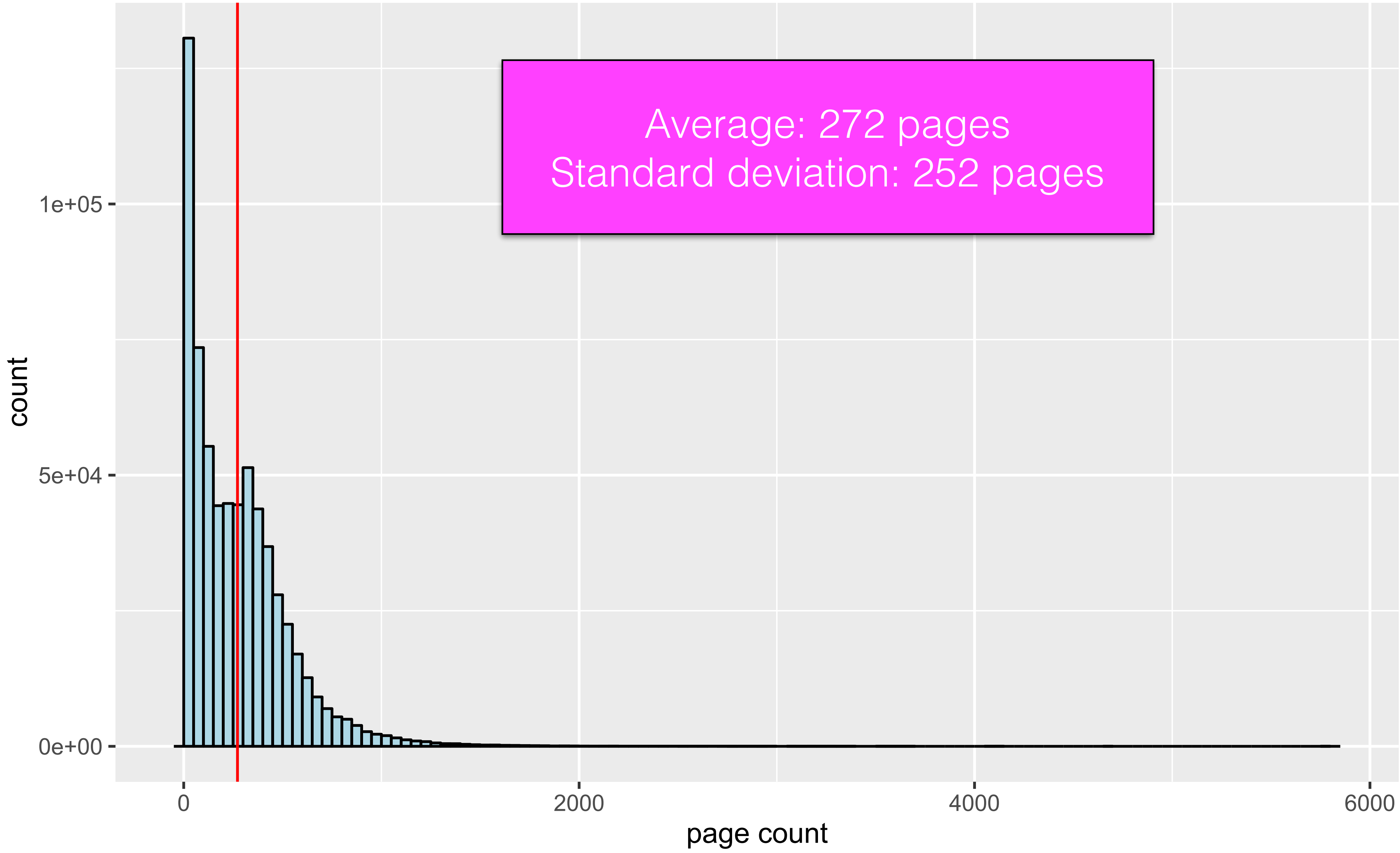
Sample

- You can’t always measure an entire population
  - expensive
  - can be destructive (measure how long light bulbs last)
  - impossible to measure (future users)

How Big is Enough?



How Big is Enough?



Average: 272 pages
Standard deviation: 252 pages

Standard Deviation

A measure of the dispersion of a set of data points

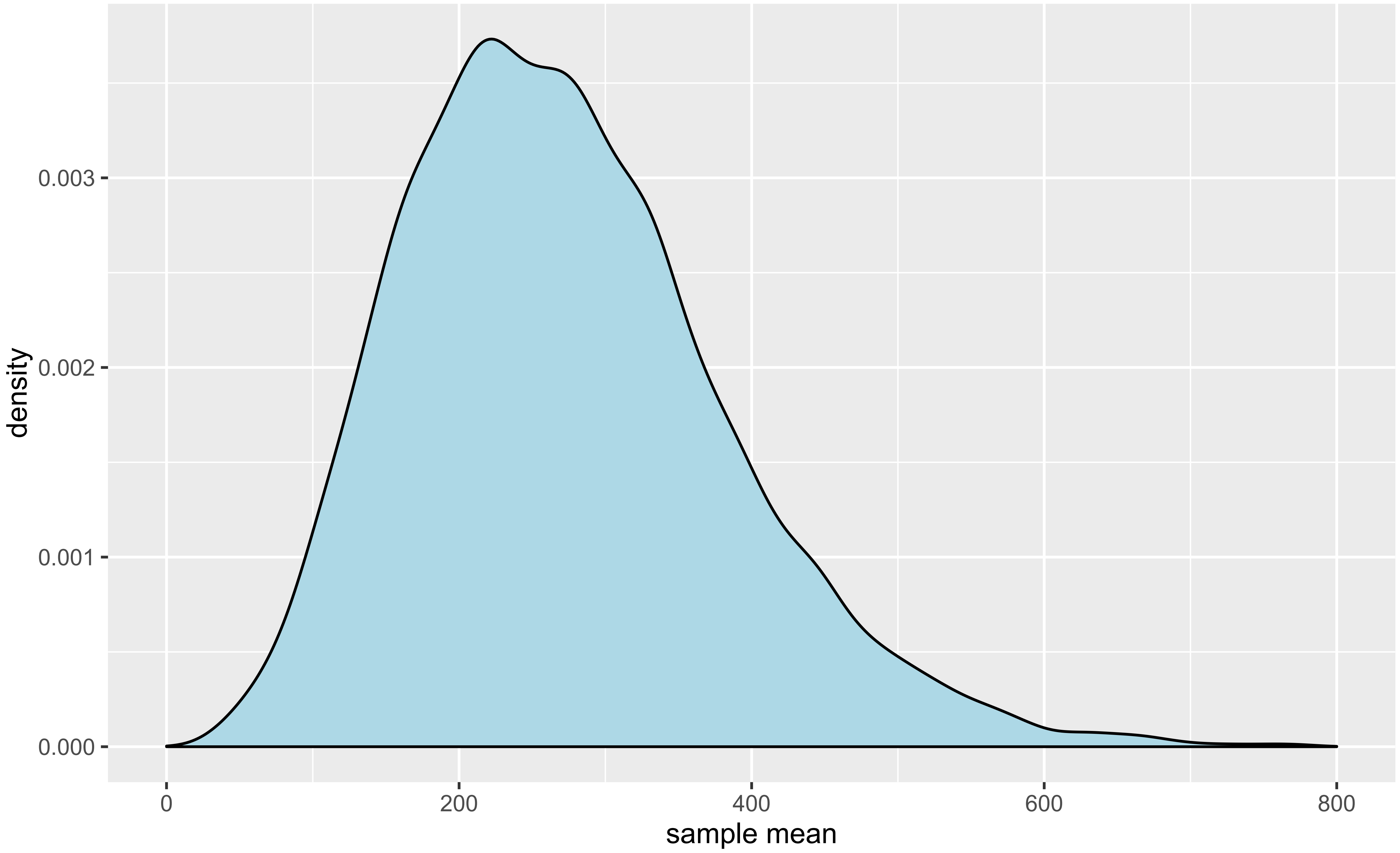
x_i	$(x_i-272)^2$
3	72361
89	33489
273	1
501	52441
494	49284

Average: 41515.2

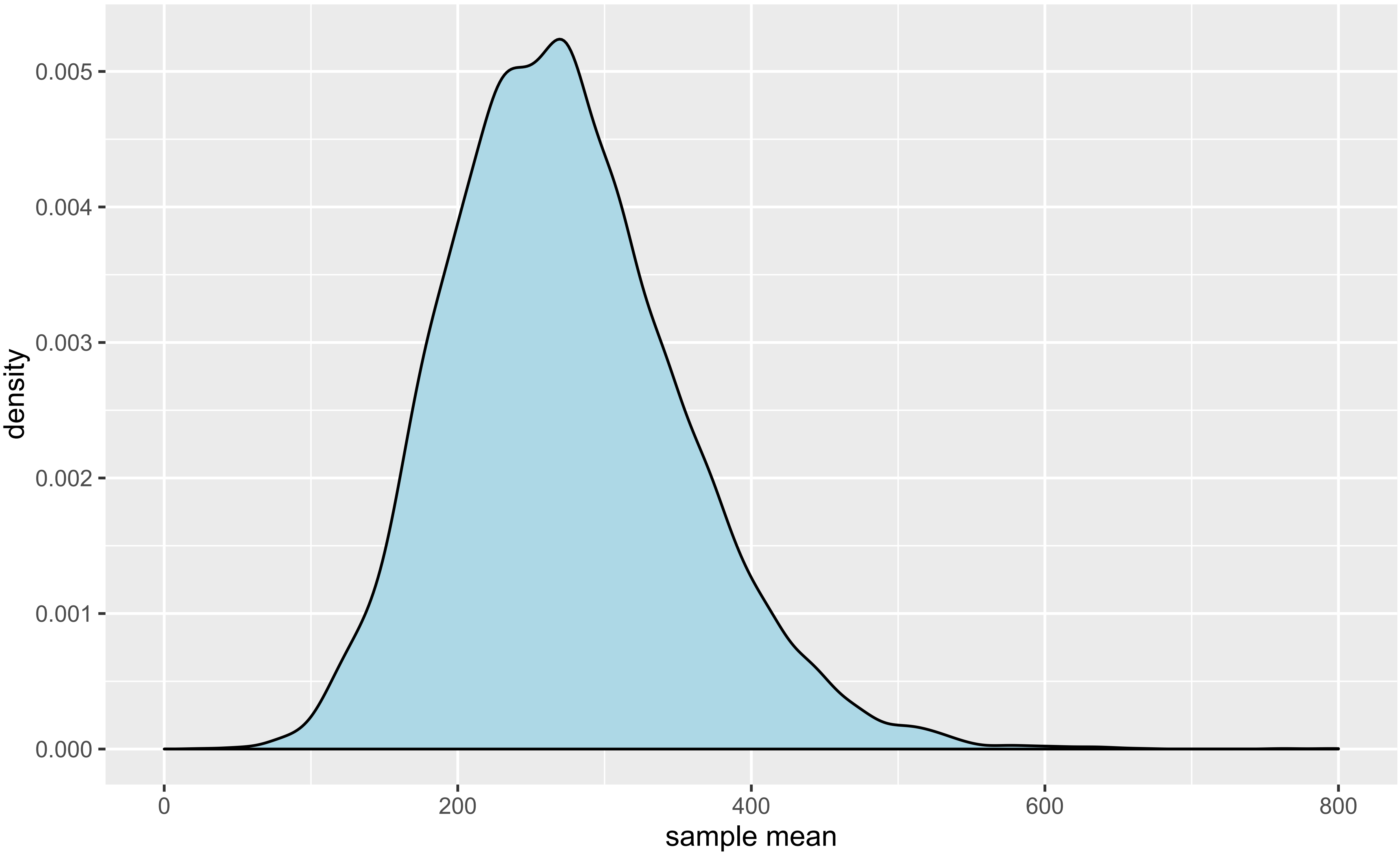
Variance is the average squared distance from the mean (41515.2)
Standard deviation = square root of the variance (203.8)

From the true population, let's take samples of size n (= 5) and measure the average of those samples
to see how much they vary

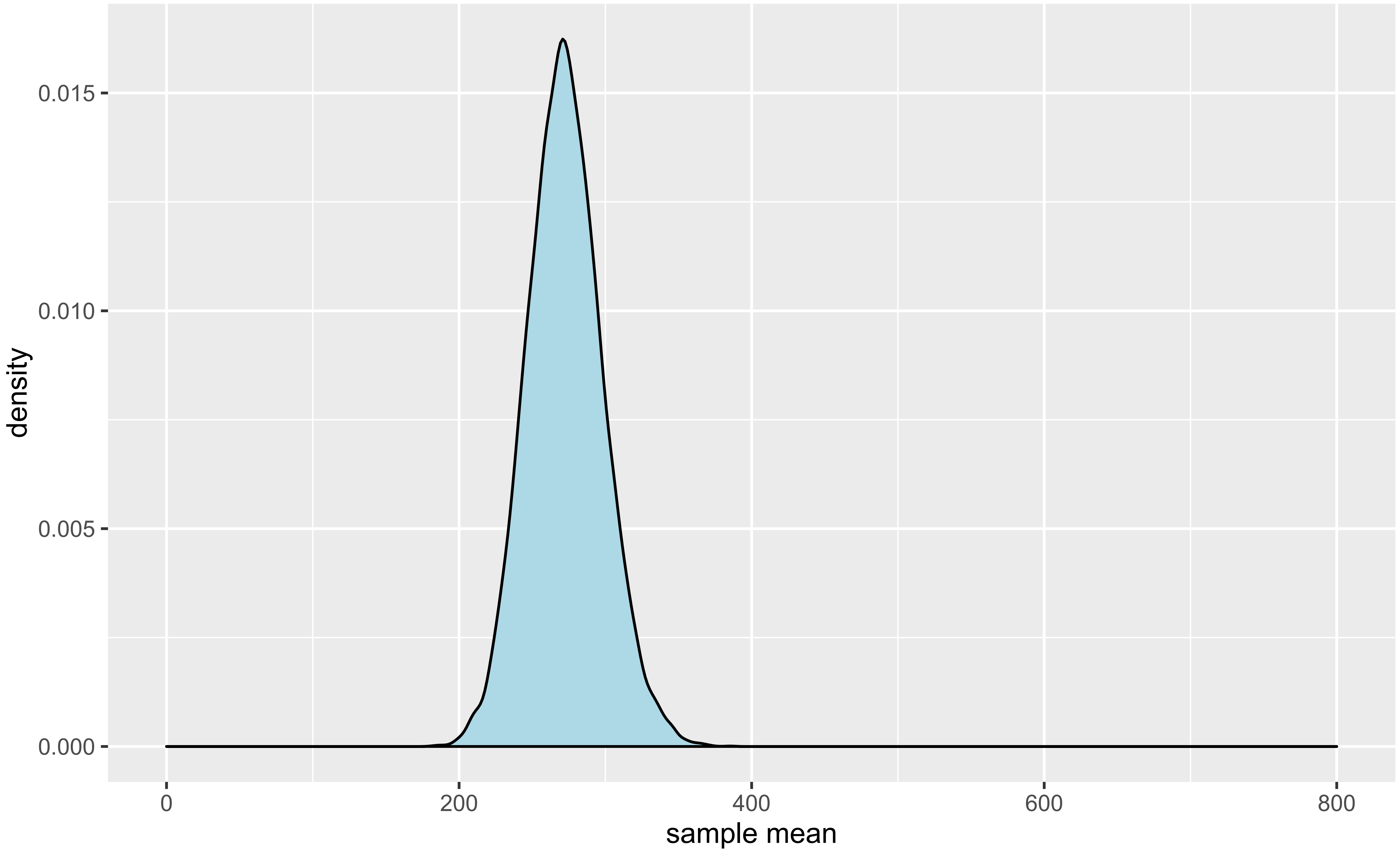
					Average
72	6	78	192	326	134.8
458	314	12	336	20	228.0
44	134	64	28	934	240.8
216	544	296	278	215	309.8
21	206	234	1024	330	363.0



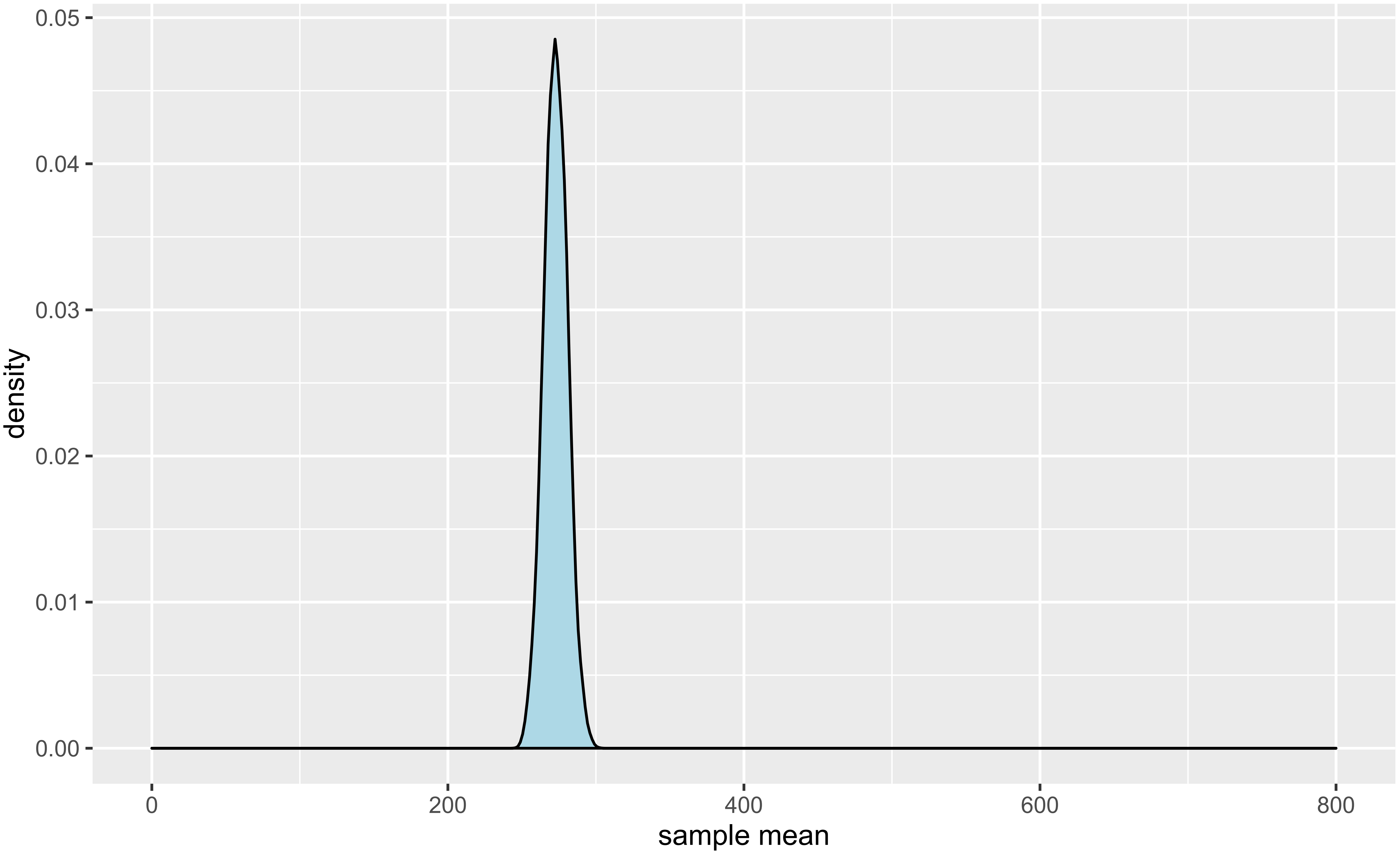
n=5, standard deviation of samples = 111



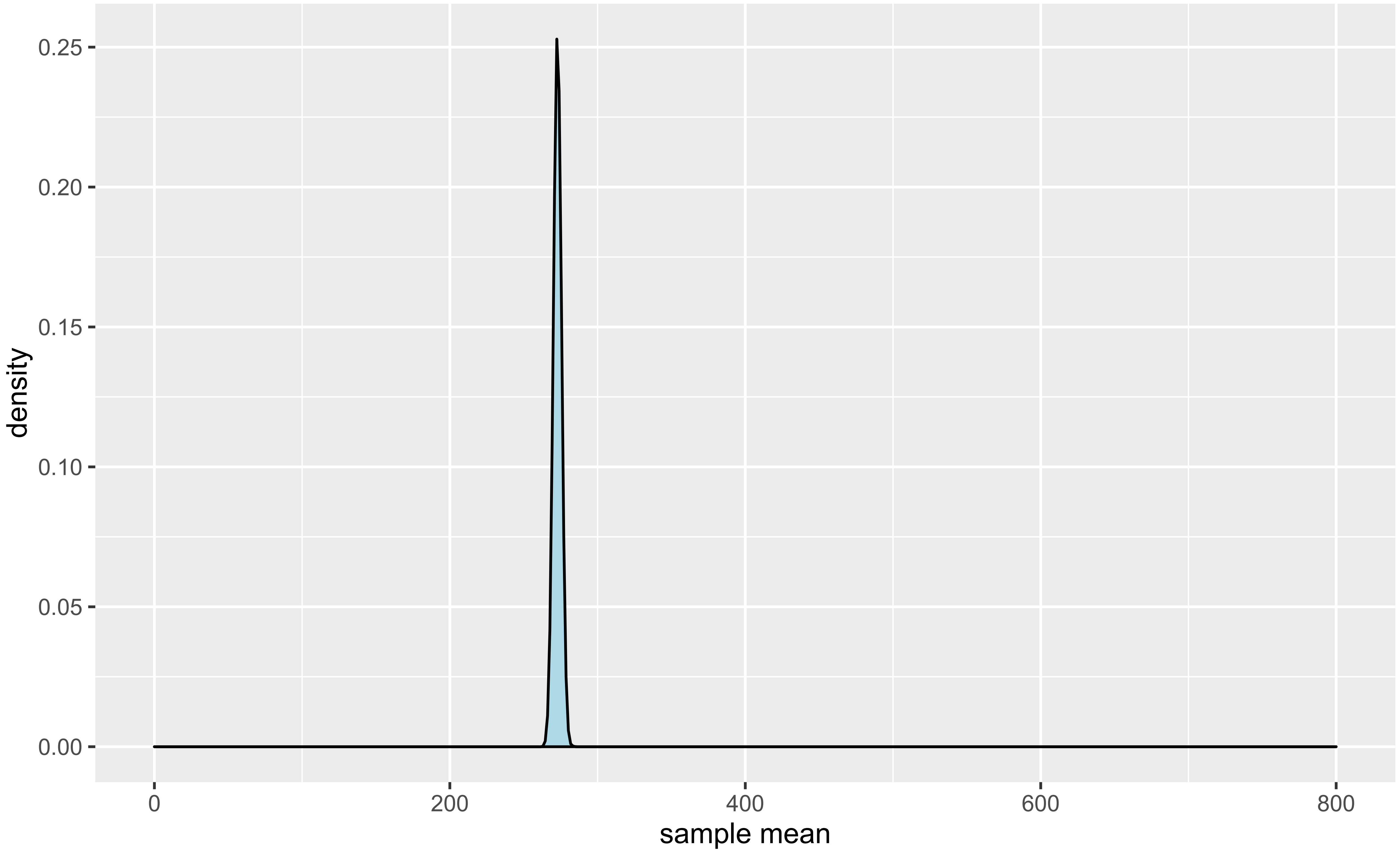
n=10, standard deviation of samples = 79



n=100, standard deviation of samples = 25



n=1000, standard deviation of samples = 8



n=10000, standard deviation of samples = 2.5

Standard error

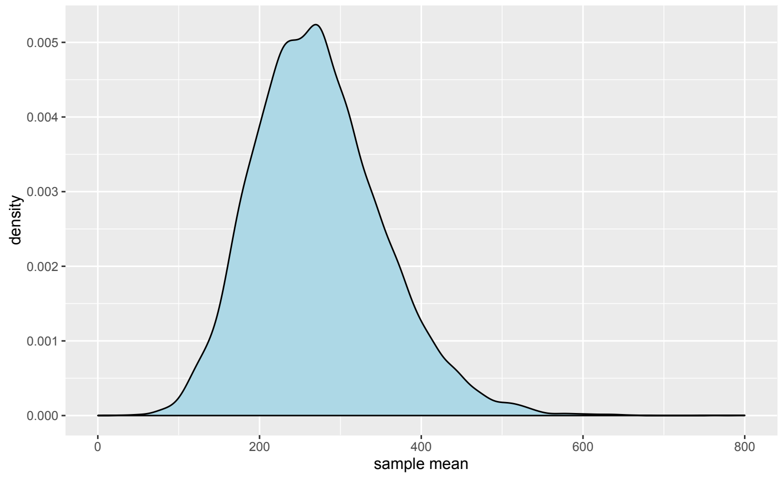
- The standard deviation of the sample is known as the standard error.

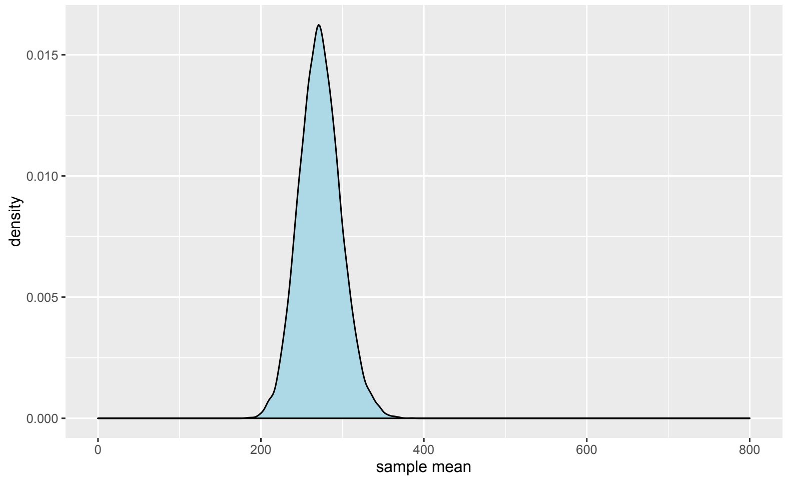
$$se = \frac{\sigma}{\sqrt{n}}$$

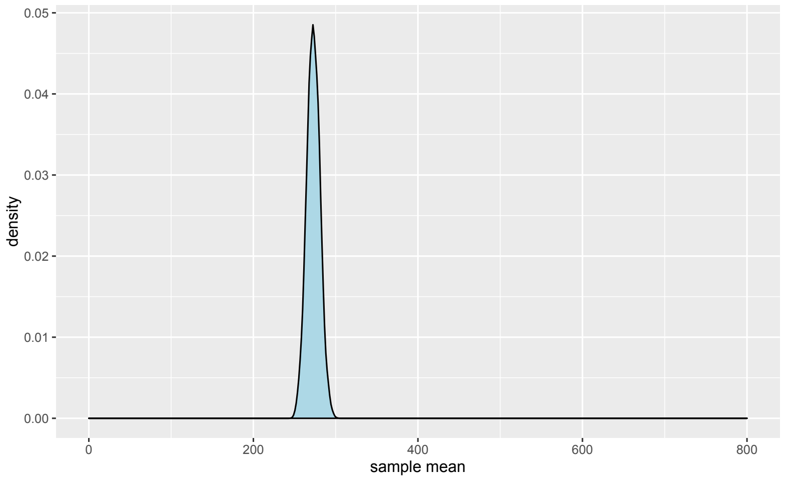
Margin of error

Under the assumption the sampling distribution is normally distributed (via the CLT), the standard error gives you confidence intervals for your measurement

90%	measurement	±1.65 x se
95%	measurement	±1.96 x se
99%	measurement	±2.58 x se







Bias in sampling

- Your knowledge about the data you really care about it is too uncertain.
    - Small samples
- Your knowledge about the data you have is not the same as the data you really care about.
    - Selection bias
    - Response bias

Sampling bias

- Non-random process by which data points are selected to be in the sample and others are not
- Canvassers conducting in-person interviews on voting preferences, avoiding:
  - all houses with pit bulls chained out front

⇒ only “nice” looking houses are canvassed

Watt et al., “Populations and Samples,” p. 52

Sampling bias

Lascaux cave painting

(Non-)Response bias

- Non-random process by which data points participate in the sample and others are not
- Survey companies about their organizational culture, only take measurements from those that let their employees respond

Watt et al., “Populations and Samples,” p. 53

(Non-)Response bias

1936 poll of presidential election (Alf Landon v. FDR)

NEW COLORED MAP OF POLAND IN THIS ISSUE
Showing the Territorial Changes Wrought by the War
The Literary Digest
(Trade Reg. U.S. Pat. Off.)
New York FUNK & WAGNALLS COMPANY London
PUBLIC OPINION New York combined with The LITERARY DIGEST
Vol. 68, No. 8. Whole No. 1609
FEBRUARY 19, 1921
Price 10 Cents

Example

- How many children are in your family (including you)?

Organizing data for analysis

- What considerations guide our choice for selecting data? What's the granularity of our unit of analysis?
    - Predict opening box office for a movie

Organizing data for analysis

- What considerations guide our choice for selecting data? What's the granularity of our unit of analysis?
    - Automatically identifying plagiarism

Sensor Data Streams

- Sensors for the analysis of human behavior

- Instrument a person

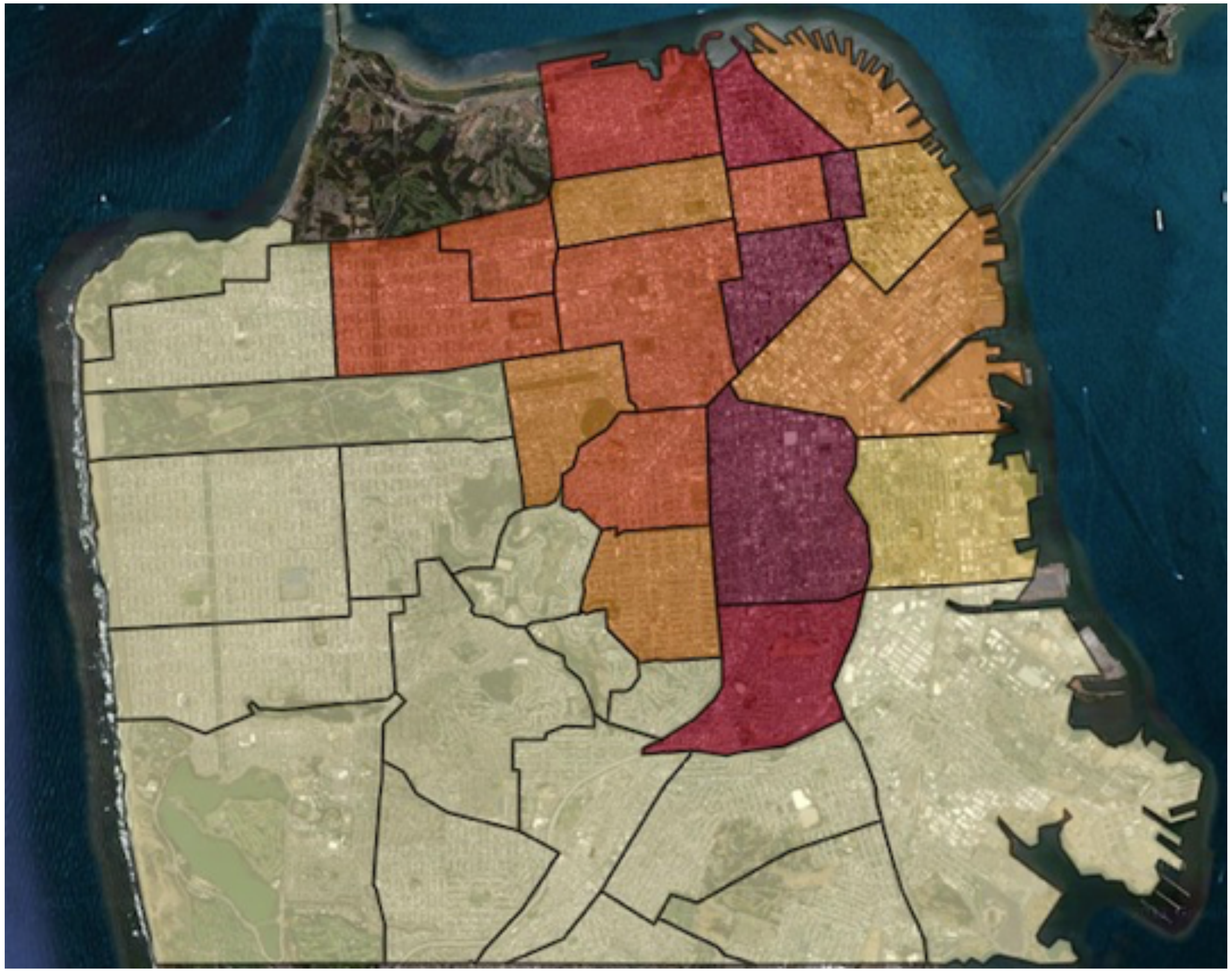
Sensor Data Streams

- Sensors for the analysis of human behavior

- Instrument a person

Sensor Data Streams

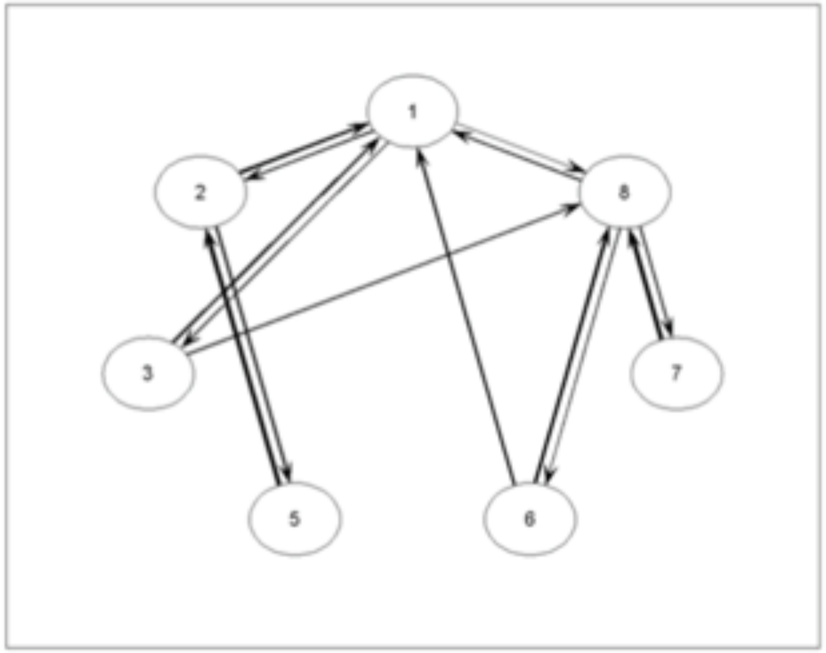
UBER

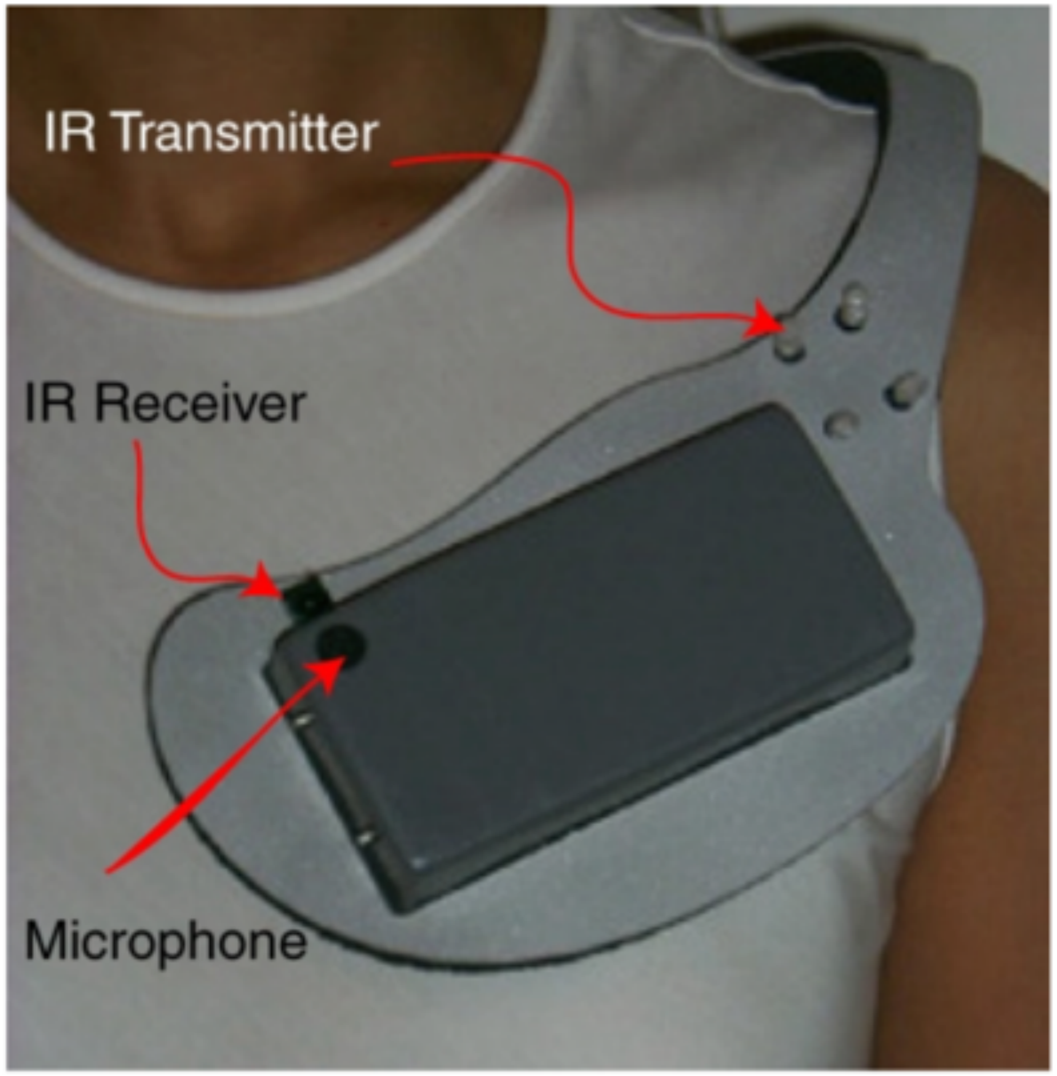


http://sanfrancisco.cbslocal.com/2014/11/18/uber-crunches-user-data-to-determine-where-the-most-one-night-stands-come-from/

Sensor Data Streams

- Instrument groups





Choudhury and Peatland, “The Sociometer”

Sensor Data Streams

- Instrument a space

Aware Home Research Initiative
http://www.awarehome.gatech.edu

Sensor Data Streams

- Instrument a space
- Deb Roy (MIT)

https://www.fastcompany.com/1733627/mit-scientist-captures-90000-hours-video-his-sons-first-words-graphs-it

Sensor Data Streams

- Where do people travel in a day?
- Who do they communicate with?
- What tools do they use during the day?
- What routines define a “typical” day?
- How healthy are their behaviors?

Voida et al. (2014), “Sensor Data Streams”, p. 293

Sensor Data Streams

- Formulate research question
- Get/build sensors
- Determine how frequently to collection samples
- Install sensors
- Store data
- Sense making

Voida et al. (2014), “Sensor Data Streams”, p. 301

Data as resource

- What information do we actually get from sensors?

Granularity

→

- Heartbeat
- Sleep patterns
- Health

→

- Location
- Most common transit routes
- Health

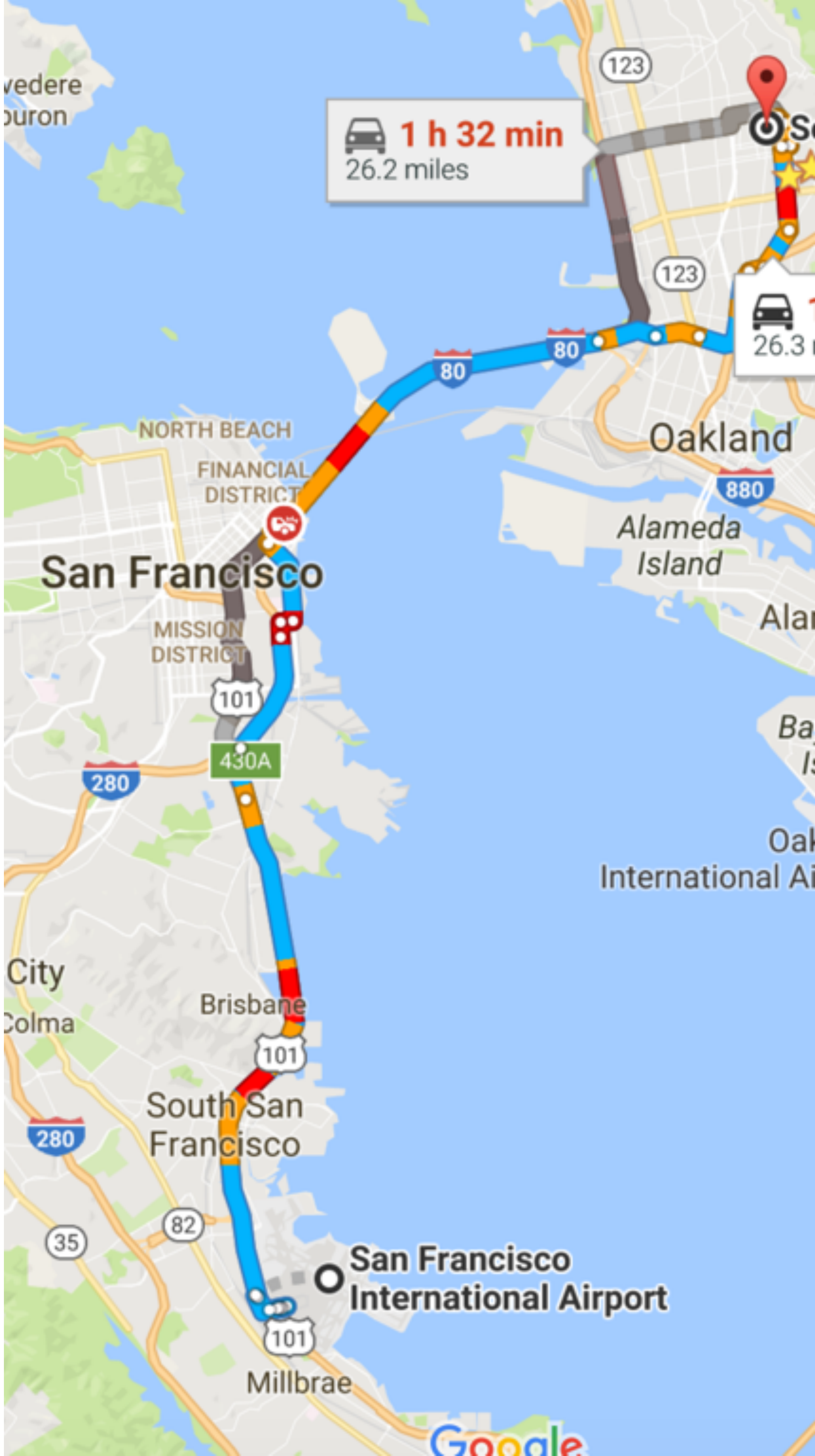
Sensor data in an organizing system

- what is being organized?
- why is it being organized?
- how much is it being organized?
- when is it being organized?
- how (or by whom) is it being organized?
- where is it being organized?

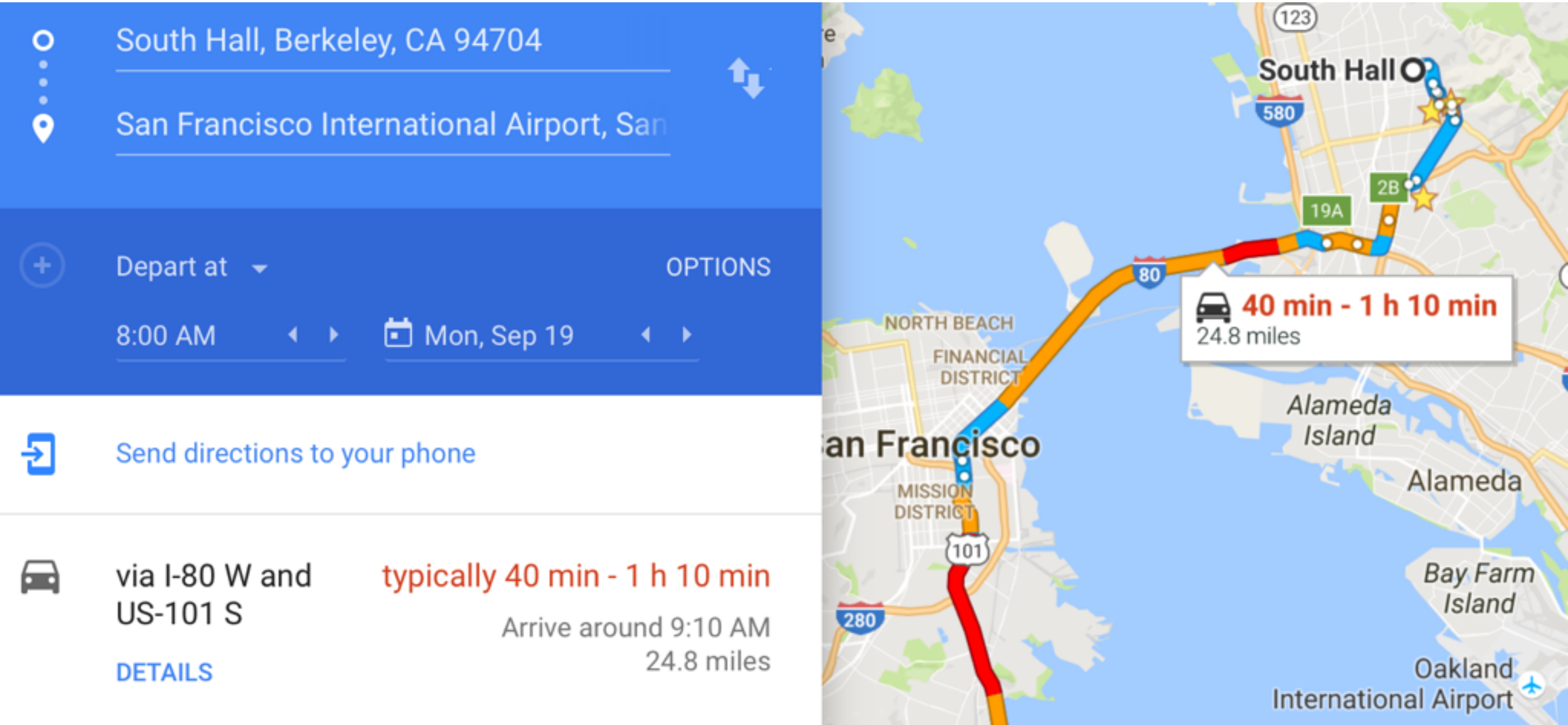
Granularity

Google Maps

- Low-level sensor data
- Route + timing
- Prediction of ETA



Granularity



South Hall, Berkeley, CA 94704

San Francisco International Airport, San

Depart at

8:00 AM

Mon, Sep 19

via I-80 W and US-101 S

typically 40 min - 1 h 10 min

Arrive around 9:10 AM

24.8 miles

40 min - 1 h 10 min

24.8 miles

Sensor data in an organizing system

- what is being organized?
- why is it being organized?
- how much is it being organized?
- when is it being organized?
- how (or by whom) is it being organized?
- where is it being organized?

Midterm

- Answer 4 out of 6 questions
- Covers material through today
- You have 90 minutes (contiguous)
- Due Friday 9/23