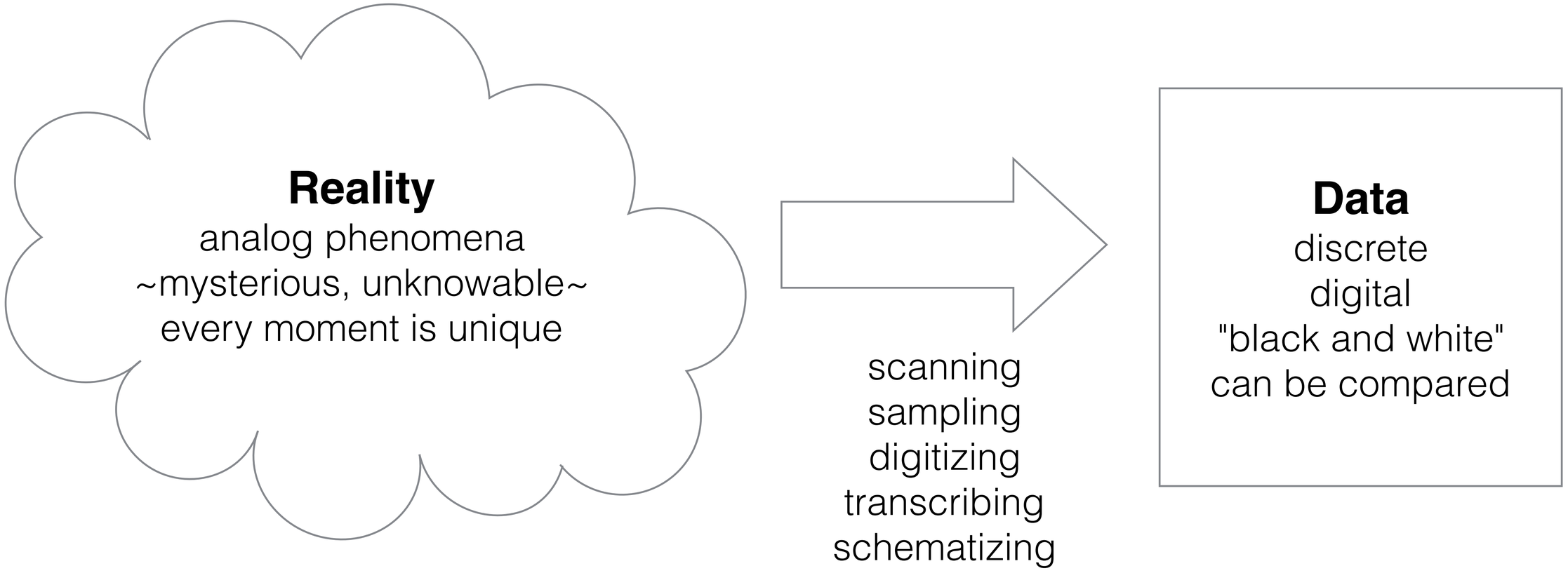
programming with data

Allison Parrish, Assistant Arts Professor at the Interactive Telecommunications Program and Interactive Media Arts Program of New York University

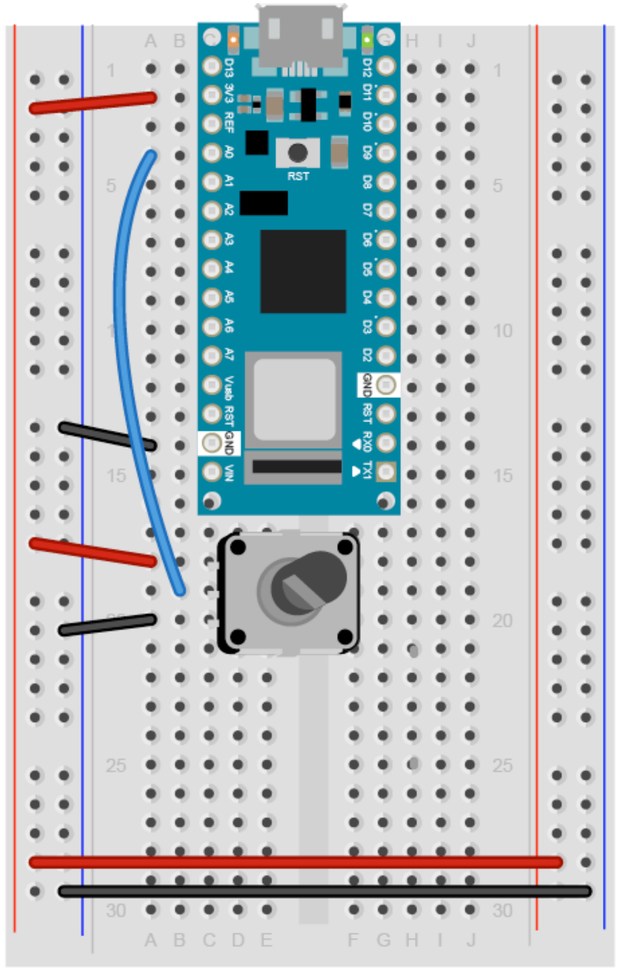
what even is data??

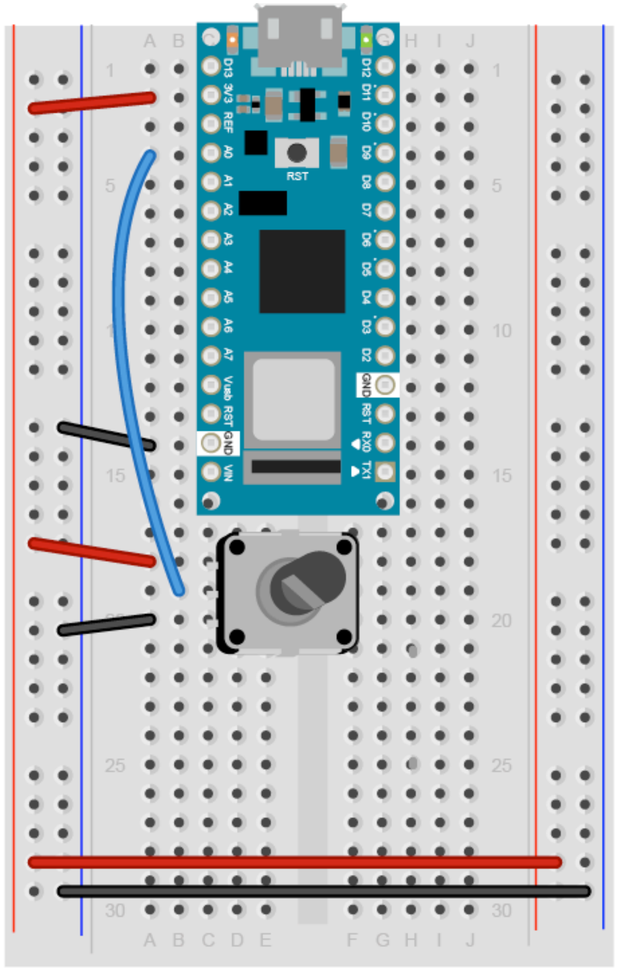
the etymology

- *data* is borrowed from Latin—it's the nominative plural of the past participle of the verb *dō* "I give"
    - ... from which we also get the English words *donate*, *donor*, *date*, *die* (as in the six-sided thing you roll), *mandatory*...
- a fair etymological translation in English might be "givens" or "gifts"
- which is weird because data is very much *taken* (in my opinion), rather than "given"
- "a gift doth blind the eyes of the wise" — Deuteronomy 16:19, KJV
- *data* is "technically" plural, and *datum* is technically the singular form and I might say *datum* sometimes but that's because i'm a virgo


Reality
analog phenomena
~mysterious, unknowable~
every moment is unique
scanning
sampling
digitizing
transcribing
schematizing
Data
discrete
digital
"black and white"
can be compared

a physical computing example


A B H I J
1
5
10
15
20
25
30
A B C D E F G H I J
D13
3V3
REF
A0
A1
A2
A3
A4
A5
A6
A7
Vusb
RST
GND
VIN
D12
D11
D10
D9
D8
D7
D6
D5
D4
D3
D2
GND
RST
RX0
TX1
RST

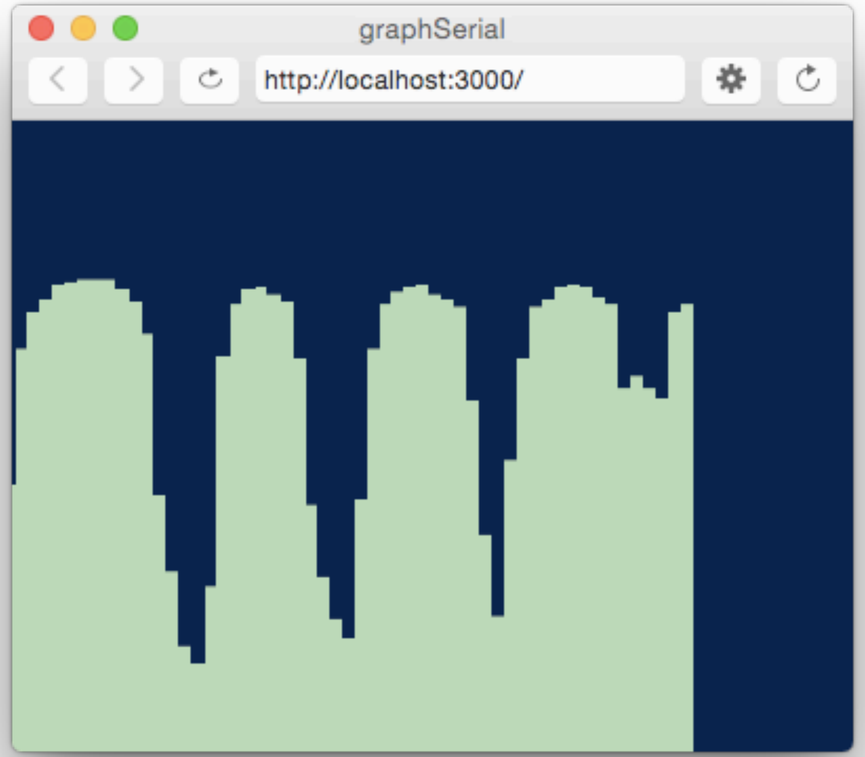


```
void setup() {
  Serial.begin(9600); // initialize serial
communications
}

void loop() {
  // read the input pin:
  int potentiometer = analogRead(A0);
  // remap the pot value to fit in 1 byte:
  int mappedPot = map(potentiometer, 0, 1023, 0, 255);
  // print it out the serial port:
  Serial.write(mappedPot);
  // slight delay to stabilize the ADC:
  delay(1);
}
```

how a sensor becomes data

- "sampling": readings from a sensor taken at discrete intervals
- always results in discrete values
- incomparable moments in time suddenly become comparable (i.e., we can say that two samples are "the same")
- this process (sensor sampled at discrete steps and turned into discrete values) is (more or less) how *all* data comes into existence (audio data, temperature data, air quality data, demographic data...)
- you can increase the resolution, but you will always still be collecting discrete values



thought experiment: how many cats are there in New York City?


Ellis Island
Ellis Island National
Museum of Immigration
Exhibits on US immigration history
Ellis Island Cafe
Fort Gibson
NEW JERSEY
NEW YORK
NEW YORK
NEW JERSEY

when you count, you decide what "counts"

language as data

I mean, police are always messing with me! I'm 45 years old now. I get stopped all the time.

'Is this your car?'

What do you mean, *'Is this my car?'* Of course it's my car.

But now there's videotape.

There's no middle ground now. You've got to choose your side, you see. You are either part of the problem or you are part of the solution.

JM: to get a conviction-
I mean police always messing with me.
I'm forty-five years old now right?
I get stopped all the time ([bæs]) says "Is this <u>your</u> car?"
What you mean is this my car? Of course it's my car.
This happened to me when I was sixteen years old walking the streets of Los Angeles.
This is crazy but now okay, it's videotaped,
everybody can see it,=
<EP>: =Uh huh.=
JM: =
like he said, and then to have - (0.6)
I mean that's serious m(h)an.=
EP: =It is.=
JM: =So you're ([yə]) going to ([gʌnə]) have - (.)
it's like I told you you got to -
there's no middle ground now.
You got to ([gɑrə]) choose your side you see.
You go either here you're either part of the problem or [you're part of the solution.]
EP: [That's right. That's right.]
JM: And it's going to ([gɔ̃:]) be I really feel it's going ([gɔ̃):])

data is information that "can be processed by a computer"

"Data... is information made *tractable*, to borrow a term from computer science. 'What distinguishes data from other forms of information is that it can be **processed by a computer, or by computer-like operations**,' as Lauren [Klein] has written in an essay coauthored with information studies scholar Miriam Posner. And to enable those operations, which range from counting to sorting and from modeling to visualizing, **the data must be placed into some kind of [computational] category...** like Boolean (a type of data with only two values, like true or false), integer (a type of number with no decimal points, like 237 or –1), or string (a sequence of letters or words, like 'this')."

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. The MIT Press, 2020. (my emphasis)

Table 46–U.S. honey production, imports, exports, stocks, and average price, by calendar year, since 1986										
Year 1/	Honey-producing colonies 2/	Yield per colony	Production	Imports	Exports	Stocks 3/	nplied domestic use	Average price per pound 4/	Value of production	U.S. resident population (as of July 1)
	Thousands	Pounds	Million pounds	Million pounds	Million pounds	Million pounds	Million pounds	Cents	Thousand dollars	Millions
1986	3,205	63	200	120	9	28	310	51	102,404	240.133
1987	3,316	71	235	58	12	34	275	50	118,424	242.289
1988	3,370	66	223	56	14	33	266	50	111,379	244.499
1989	3,528	51	181	77	10	33	247	50	89,956	246.819
1990	3,220	62	199	77	12	31	265	54	106,688	249.464
1991	3,211	69	221	92	10	36	298	56	122,830	252.153
1992	3,045	73	222	115	10	39	323	55	121,922	255.030
1993	2,875	80	231	134	9	49	346	54	124,280	257.783
1994	2,783	78	218	123	8	60	322	53	115,203	260.327
1995	2,655	80	211	89	9	42	308	69	144,585	262.803
1996	2,581	77	200	151	10	47	335	89	177,166	265.229
1997	2,631	75	197	167	9	71	332	75	147,795	267.784
1998	2,637	84	221	132	11	81	332	66	144,445	270.248
1999	2,652	77	203	183	11	79	377	60	122,044	272.691
2000	2,622	84	220	199	10	85	402	60	131,511	282.162
2001	2,550	73	186	145	7	65	344	70	130,980	284.969
2002	2,574	67	172	203	7	39	393	133	227,870	287.625
2003	2,599	70	182	200	7	41	374	139	252,051	290.108
2004	2,554	72	183	179	8	61	334	109	199,641	292.805
2005	2,409	72	175	233	8	62	399	92	160,994	295.517
2006	2,394	65	155	278	7	60	428	101	155,685	298.380
2007	2,443	61	148	233	8	53	381	108	159,763	301.231
2008	2,342	70	164	231	10	51	387	142	232,744	304.094
2009	2,498	59	146	210	10	38	361	147	215,671	306.772
2010	2,692	66	176	251	10	45	411	162	285,692	309.322
2011	2,491	60	148	288	12	37	433	177	261,850	311.557
2012	2,539	56	142	311	12	32	446	199	283,454	313.831
2013	2,641	57	150	338	12	38	469	214	320,037	315.994
2014	2,741	65	178	365	11	41	530	217	386,933	318.301
2015	2,661	59	157	386	11	42	531	208	325,946	320.635
2016	2,780	58	162	367	11	41	519	212	343,962	322.941

https://www.penn.museum/collections/object/521677 Object number CBS3293

	i.	ii.	iii.	iv.	v.	vi.	vii.	viii.	ix.	x.	xi.
1	**[Bull**	**Fully-grown cow**	**Two-year old heifer**	**One-year old heifer**	**Heifer calf**	**Bull calf**	**Total**	**Ghee**	**Herdsman**	**Mayor**	**"Kassite"]**
2	[…]	[…]	[…]	[…]	[…]	[…]	[19]5	60 silas	[…]	[…]	[…]
3	[…]	[…]	[…]	[…]	[…]	[…]	287	190 silas	Lultamar-[…]	[…]	[…]
4	[…]	[…]	[…]	[…]	[…]	[…]	33	20 silas	Apil-Nergal	[…]	[…]
5	[…]	[…]	[…]	[…]	[…]	[…]	13	7.5 silas	Izkur-Šamaš, son of Appāju	[…]	[…]
6	[…]	[…]	[…]	[…]	[…]	[…]	3	2.5 silas	Amurru-aḫa-iddina, son of Kurû	[…]	[…]
7	[…]	[…]	[…]	[…]	[…]	[x]+1	24	15 silas	Amurru-nādin-aḫḫē, in place of Muštē[…]	[…]	[…]
8	[…]	[…]	[…]	[…]	[…]	[x]+2	27	15 silas	Rabâ-ša-Nergal	In[?][...]	[…]

Huang, Ami. *State, Province, and Temple in Kassite Nippur: A Case Study of the Livestock Economy of the Ereš.Dingir Priestesses*. 2020. University of Chicago.

data is mundane

what is data even good for??

- at its best, data is a way to discover things about the world that aren't obvious at first glance

- think about: climate change, COVID statistics, keeping a journal...

- data can be a powerful way to hold power to account (we'll look at examples of this) — so much so that organizations often preemptively release data about their activities, or are required to do so by law

- ... but data can also contribute to the maintenance of the status quo. data interpreted as "given" is often understood to be what is "normal" (we'll look at a *ton* of examples of this)

goals of data analysis

- (cribbed from Spector, Alfred Z., et al. *Data Science in Context: Foundations, Challenges, Opportunities*. Cambridge University Press, 2022)
- "exploring" data in order to produce "insights" that are derived from the data: hypotheses, "aha" moments, identification of correlations that suggest causal relationships
- to produce "conclusions" from the data: predictions, recommendations, classification systems, etc.
- as artists and designers, using data might be part of our research process, or it might be the subject of the work itself

exploratory data analysis

- a type of data analysis that is focused on insights, summaries, and proposing hypotheses, rather than conclusions
- basically: take some data and figure out what's going on in there
- techniques include:
  - counting and finding the most common by group
  - descriptive statistics (mean, median, mode, percentiles, standard deviation)
  - visualizations: scatterplots, box plots, histograms, etc.
  - clustering
  - correlations

why python?

- it's ubiquitous in data science
- the tools for data analysis are mature and well-documented
- powerful and fast (when you're doing it right)
- surprisingly portable!
- I just think it's neat

drawbacks of python

- there are many different versions and distributions of python
- installing libraries can be a hassle
- lots of tutorials out there, but some of them aren't really that good :(
- stuck on your computer (difficult to share your work on the web)
- it's not the fastest or most memory-efficient language in the world

goals of the class

- by the end of the class, you should be able to:
  - locate or create data sets of interest
  - parse and clean data sets in various formats (CSV, JSON, Excel, SQL...)
  - perform exploratory data analysis on that data (with Python and Pandas)
  - communicate insights from those explorations (sharing data, work, and outcomes)
  - understand the nature of data and how it comes to be, and apply a critical eye toward how data is used for good and bad

what we're not going to do in this class

- like, actual statistics
    - we'll discover insights, correlations, and curiosities but we won't claim to find causes
- pretty data visualizations (this isn't a visual design class!)
- language and text stuff (at least not in depth) (take RWET if you want to learn more about this!)
- maps :(
- but the goal is to open the door to this kind of stuff if you want to learn it on your own or in another class!
- or help you clean up data sets in python so you can export friendlier data to other tools