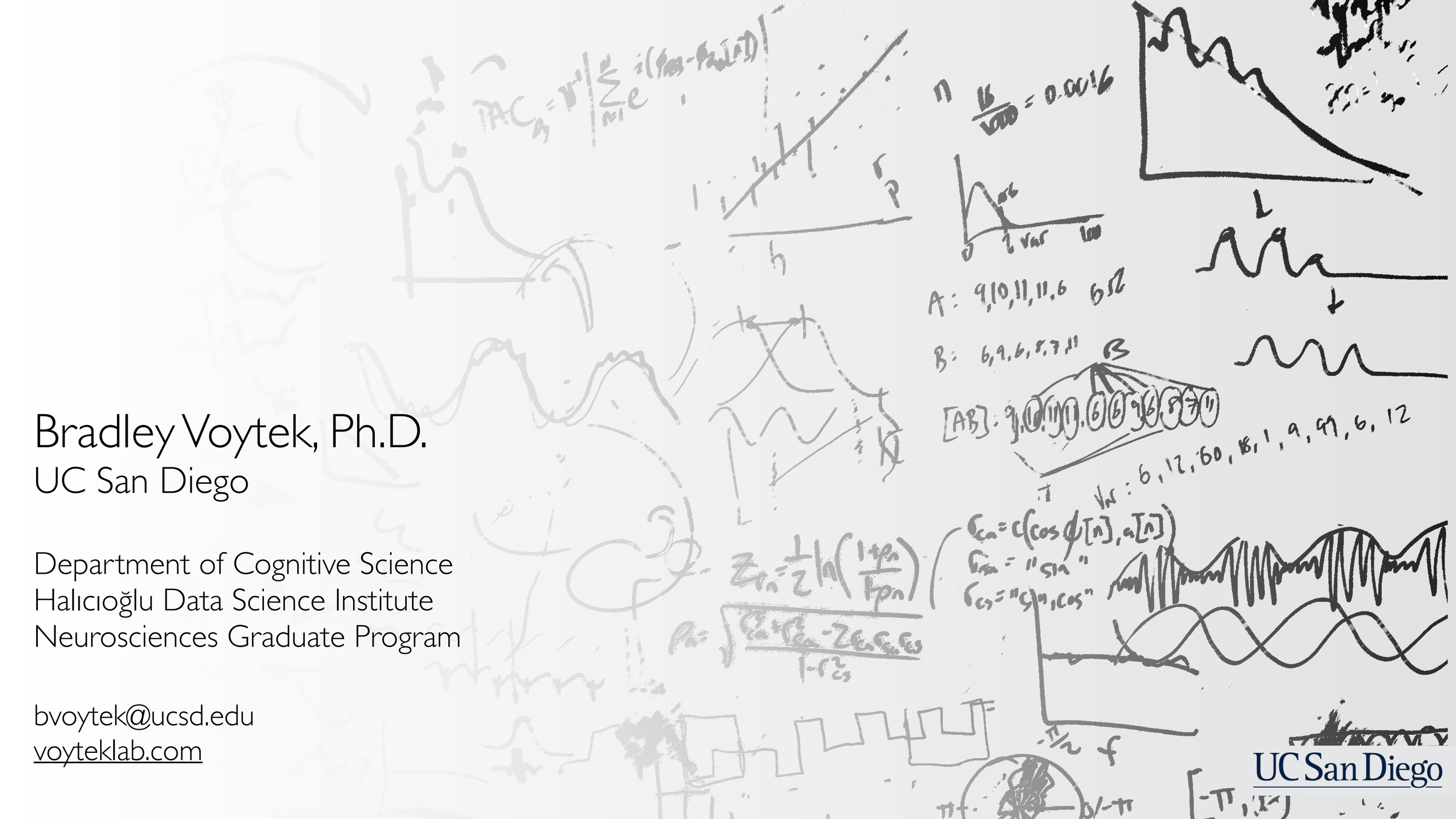
Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
voyteklab.com



UC San Diego

So what is Information?

- Data placed into a context that can be understood.
- Information allows us to make decisions.

Source: *Google for Education*

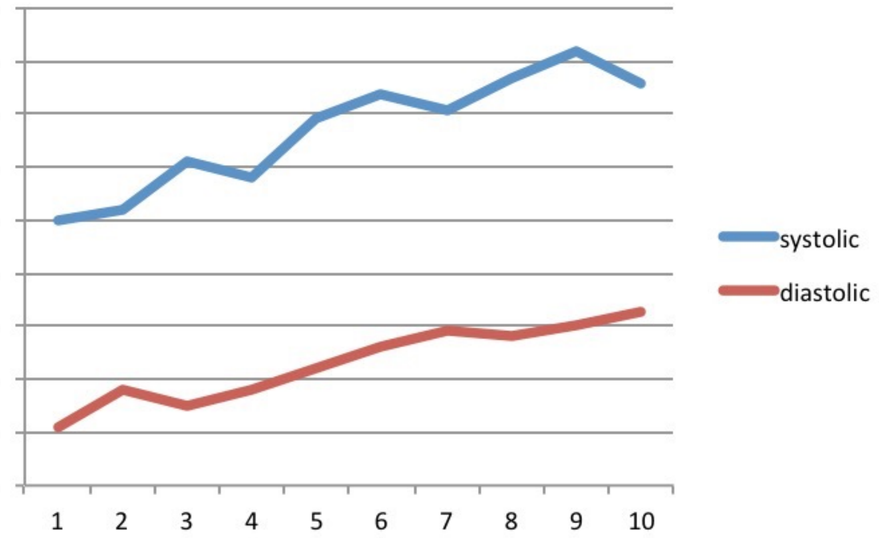
Is this Information?

A	B	C
1	110	71
2	112	78
3	121	75
4	118	78
5	129	82
6	134	86
7	131	89
8	137	88
9	142	90
10	136	93

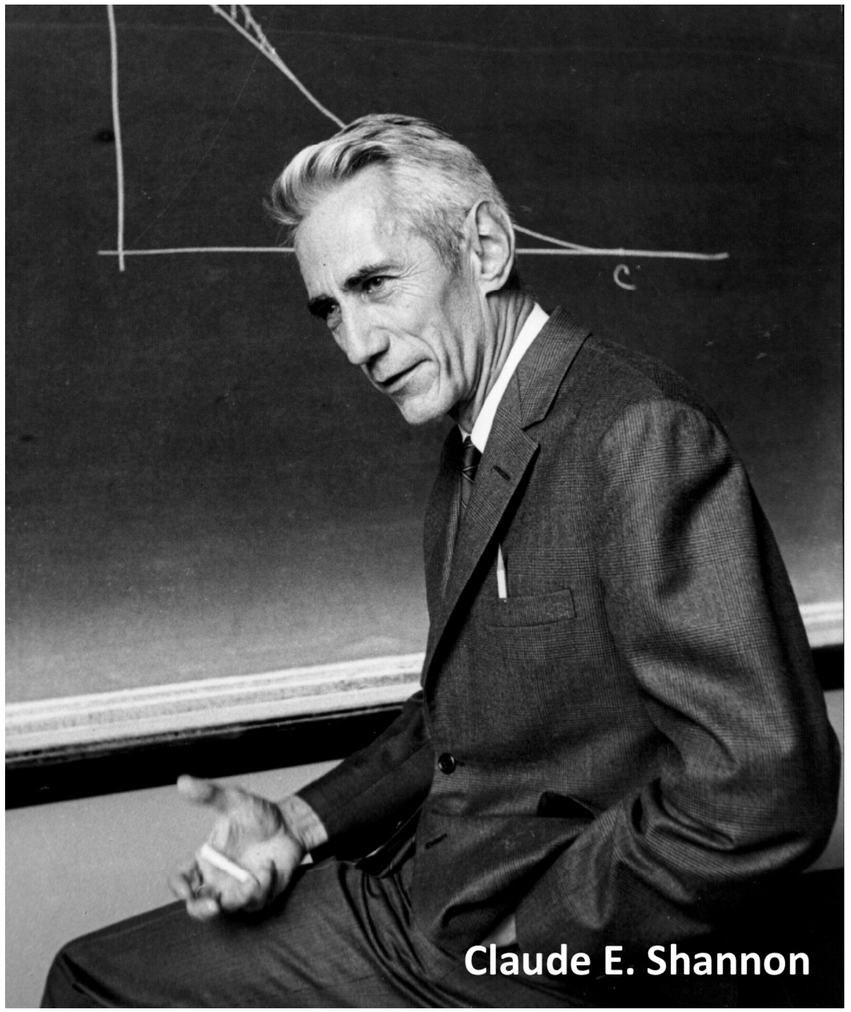
Is *this* Information?

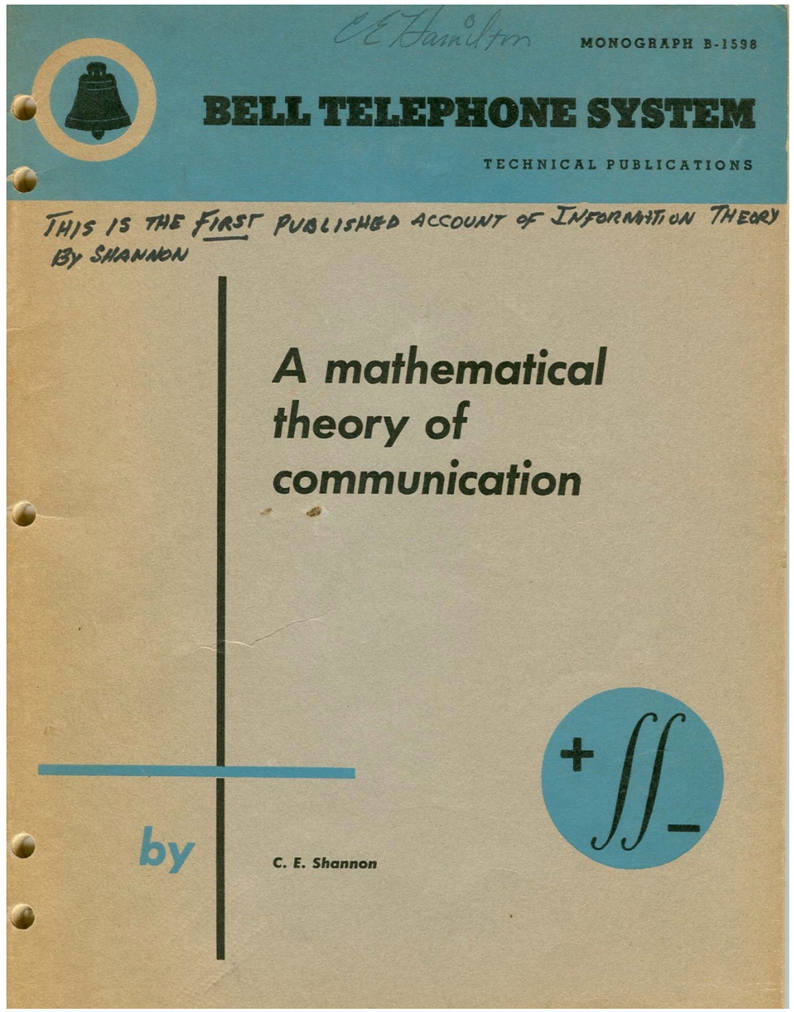
A	B	C
1	110	71
2	112	78
3	121	75
4	118	78
5	129	82
6	134	86
7	131	89
8	137	88
9	142	90
10	136	93

time	systolic	diastolic
1	110	71
2	112	78
3	121	75
4	118	78
5	129	82
6	134	86
7	131	89
8	137	88
9	142	90
10	136	93

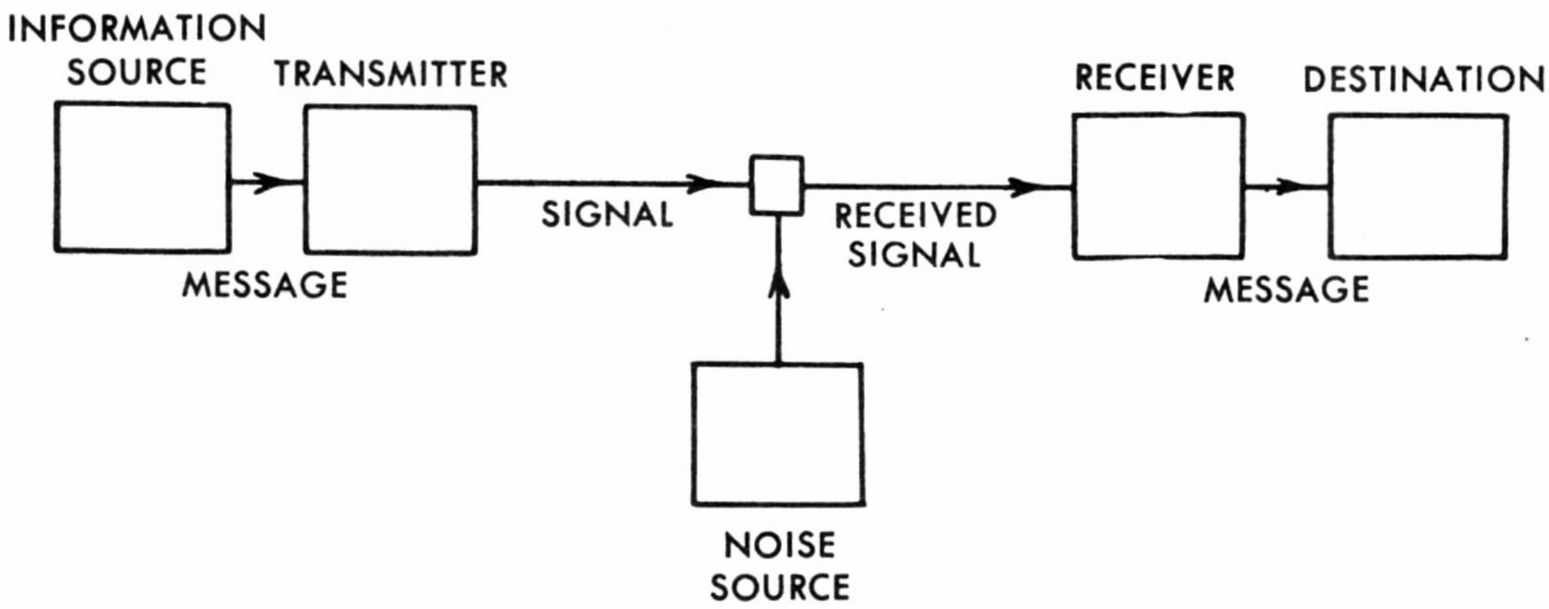


Information Theory





Information Theory



Information Theory

How can we achieve perfect communication over an imperfect, noisy communication channel?

- Use more reliable components
- Stabilize the environment
- Reduce noise sources

These are all costly and/or outside our ability!

Source: *Kumar, Bryn Mawr College*

Information Theory

Information content of an outcome x:

$$h(x) = log_2(1/p(x))$$

Information Theory

Flipping a (fair) coin once will give us either heads or tails, each with $p = 0.5$

$$h(x) = log_2(1/p(x))$$

A single flip of a coin gives us:

$log_2(1/0.5) = 1$ bit of information (whether it comes up heads or tails)

UC San Diego

Source: *Kumar, Bryn Mawr College*

Information Theory

We could write a sequence of 25 flips as, for example:

hthhtthhhhthttthhhhthtt

or, using 1 for *h* and 0 for *t*, the 25 bits:

1011001011101000101110100

Information Theory

We thus get the nice fact that *n* flips of a fair coin gives us *n* bits of information, and takes *n* binary digits to encode.

Source: *Kumar, Bryn Mawr College*

COGS 9
Introduction to Data Science

Data and information

UC San Diego

Today's Learning Objective

Identify variable types within a dataset

So are we Data?

Imagine a future society with teleportation technology. Instead of having to spend all day traveling to get from Orlando to L.A., you can now step into a teleporter booth, hit a glowing green button and be more or less instantly transported anywhere on the planet. Here's how it works: the machine scans your full atomic structure, stores the pattern, then beams it to another teleporter, where a matter-assembler puts you back together again from a stock pile of atoms. You have used this machine many times with no qualms whatsoever. Now, imagine that one day you step in the booth, press the green button, but nothing happens. You are then told in a polite, robotic voice, "I'm sorry traveler, but something went wrong. Although we successfully scanned your body and reassembled you in L.A., the disintegration process failed. Would you please press the purple button in order to finish the disintegration process?"

Source: *Gary Williams*, Washington University

What is data?

- Information that is not yet interpreted.
- Anything we can see, hear, smell, touch, taste is "data".
- Data can be transmitted from one place to another and stored.

Source: *Google for Education*

What is data?

- In simpler terms:
  - **data**: anything that can be stored on a computer

What is data?

- In simpler terms:
  - **data**: anything that can be stored on a computer

Note: This doesn't mean it *has* to be on a computer to be data. But, it *could* be stored on a computer.

measurements of your height over time, text messages you've sent, Facebook posts your friends have posted, websites you visit, things you buy with a credit card, pictures of your car on speed cameras, audio files downloaded from Spotify, text in emails you've sent to your professors, pictures of your pet, information you fill out in profiles for your school, job, or community organizations, information provided to your physician, list of all the items in your closet, number of clicks on a website's advertisement, list of the sizes of all the shoes ever sold in the world, number of items purchased from a website, information collected in chemistry lab, information entered onto Yelp about your favorite restaurant, photos from your family's vacation, twitter follows, your grades, butterflies in the park

If you can enter it into a spreadsheet, take a picture of it, write about it, make a video of it, or record it on audio - then it is probably data.

Some data

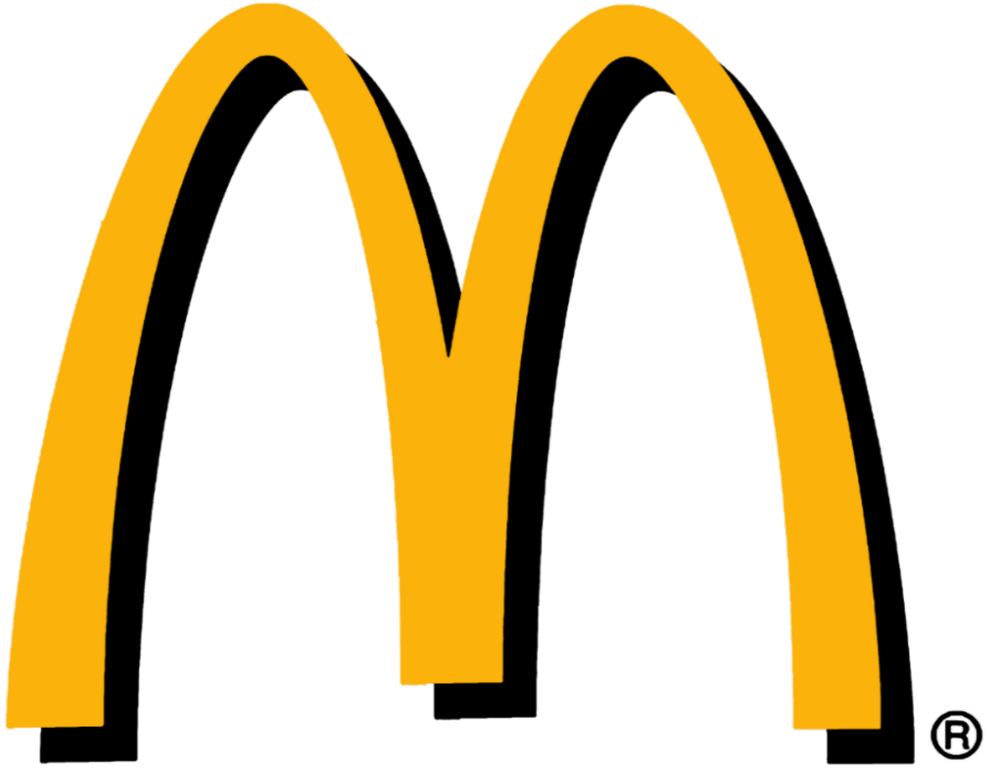
What's the largest group of people who can understand this?

Some data

This?

Some data

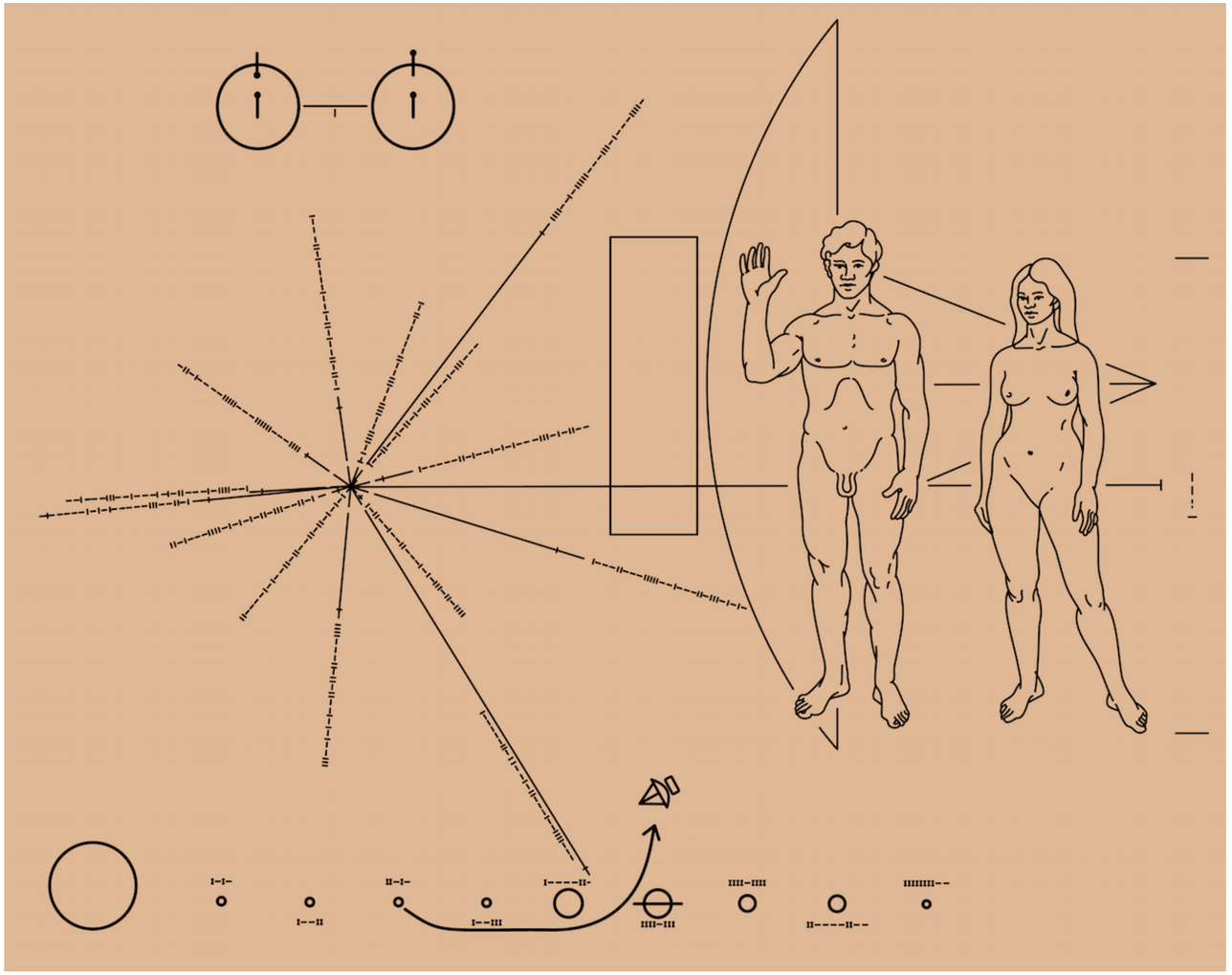
This?



Some data

This?

Are data universal?



Types of data

Quantitative

- Integers (whole numbers, i.e. 10)
- Continuous: Float or Numeric (200.78)

Qualitative

- String or Character (“Hello World”)
- Categorical or Discrete (limited number of options) - (*i*.e., minor, adult)

Types of data

We categorize **data storage** by how **structured** it is

- Semi-structured*
  - spreadsheets (tabular data)
  - JSON & XML
- Structured
  - relational databases (SQL)
- Unstructured
  - everything else: video, audio, images, websites, apps, text, etc.

*Data Scientists work with semi-structured data most frequently, but have to be familiar with and comfortable in all types

JSON

JSON: key-value pairs

nested/hierarchical data

ENCODE SOME INFORMATION!

TELEPHONE GAME

play is training for the unexpected

Information Theory

Shannon information is:

- *Uncertainty*: The number of possible messages.
- *Surprise*: Likelihood of a message.
- *Difficulty*: How do you transmit the message?

Uncertainty

The number of possible messages

- b_ink

Uncertainty

The number of possible messages

- b/ink?

Uncertainty

The number of possible messages

- b*r*ink?

Uncertainty

The number of possible messages

- boink?

Uncertainty

The number of possible messages

- b4ink?

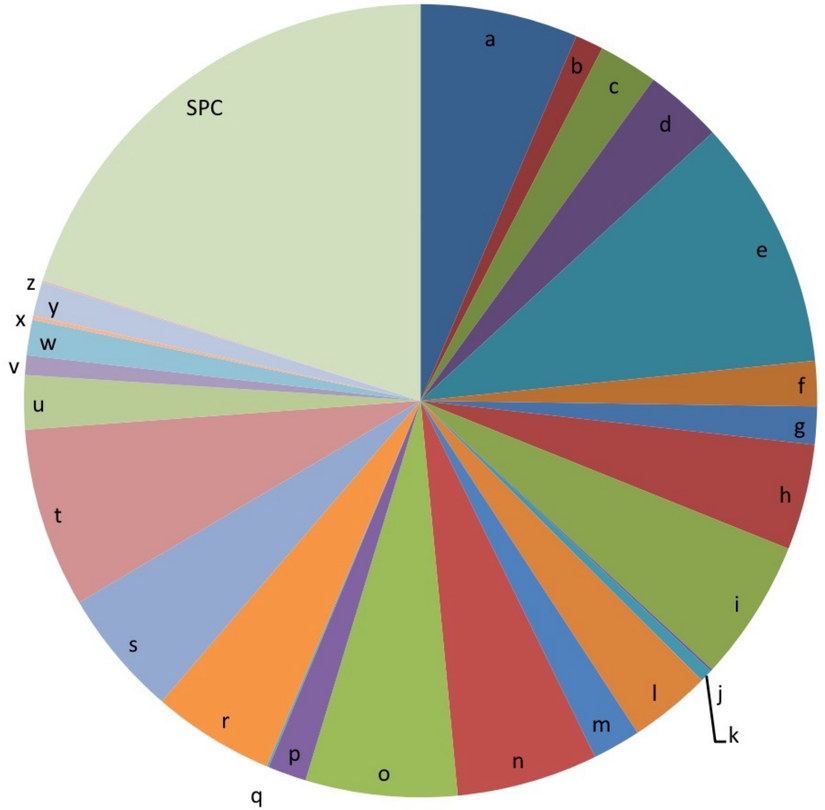
Uncertainty

The number of possible messages

- oran_e

English letter frequency

a	0.06428
b	0.01147
c	0.02413
d	0.03188
e	0.10210
f	0.01842
g	0.01543
h	0.04313
i	0.05767
j	0.00082
k	0.00514
l	0.03338
m	0.01959
n	0.05761
o	0.06179
p	0.01571
q	0.00084
r	0.04973
s	0.05199
t	0.07327
u	0.02201
v	0.00800
w	0.01439
x	0.00162
y	0.01387
z	0.00077
SPC	0.20096



Uncertainty

"___ clearly ___ that the needs of the many ___ the ___ of the few."

Redundancy

"___ clearly ___ that the needs of the many ___ the ___ of the few."

"Logic clearly dictates that ___ needs of ___ many outweigh ___ needs of ___ few."

Uncertainty

"I signed aboard ___ ship ___ practice medicine, not ___ have my atoms scattered back ___ forth across space."

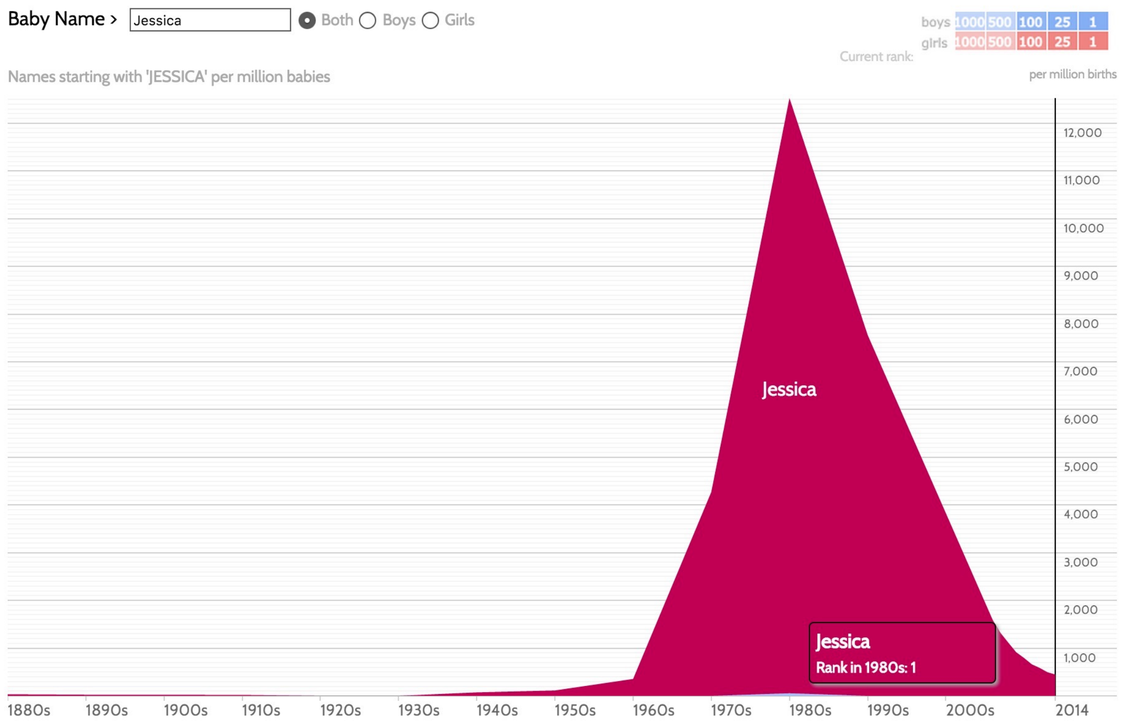
Language and context

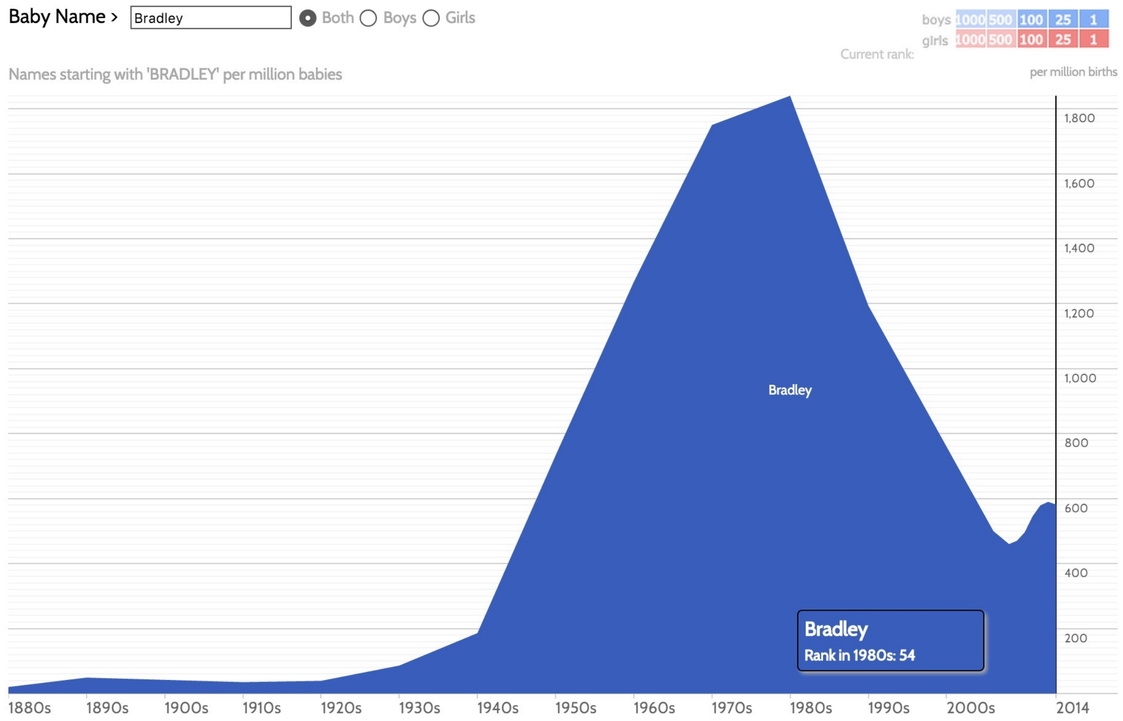
Definite and indefinite articles (corresponding to *the*, *a*, *an* in English) do not exist in the Russian language. The sense conveyed by such articles can be determined in Russian by context.

Language and context

Articles are found in many Indo-European languages, Semitic languages (only the definite article), and Polynesian languages, but are formally absent from many of the world's major languages, such as Chinese, Indonesian, Japanese, Hindi, Punjabi, Urdu, the majority of Slavic and Baltic languages (including Russian), Yoruba, and the Bantu languages.

Extracting information from data

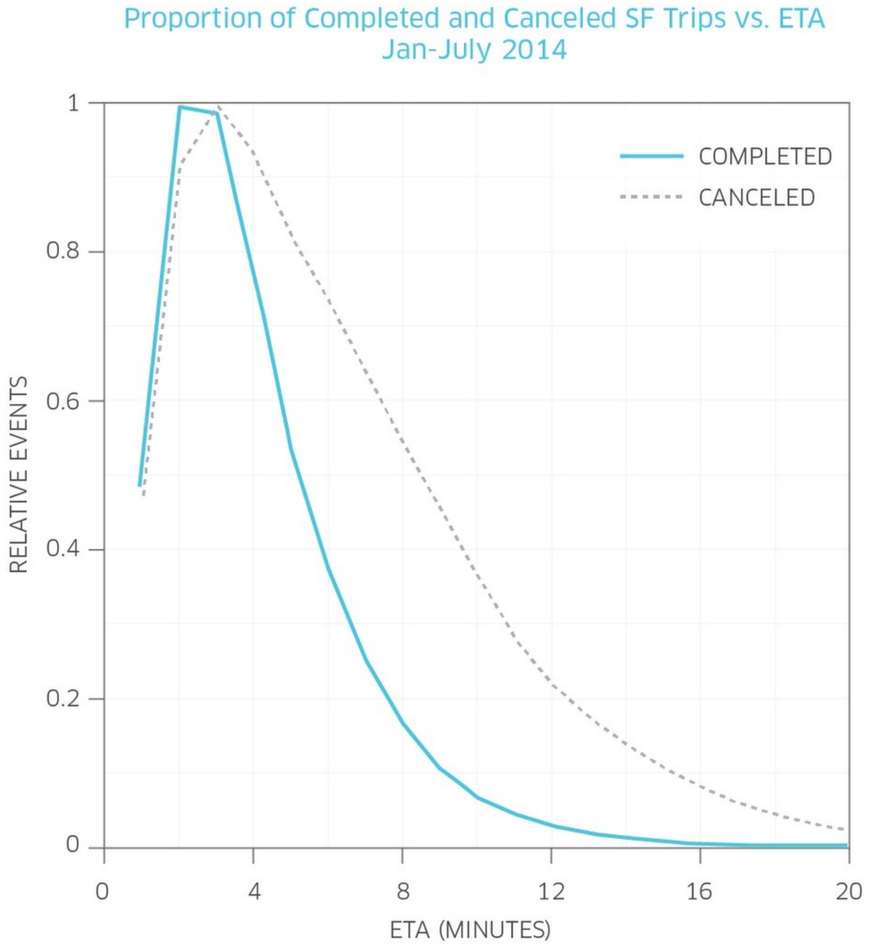




Source: babynamewizard.com

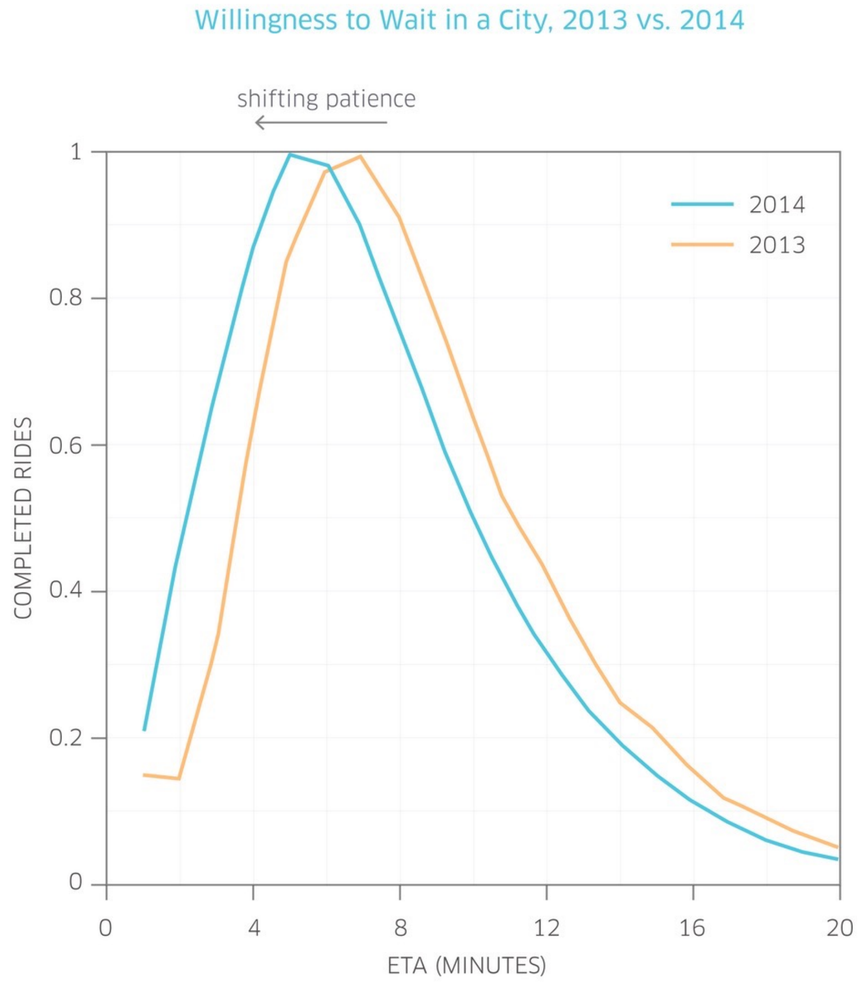
UC San Diego

Extracting information from data



Source: Voytek, *Uber, Inc.*

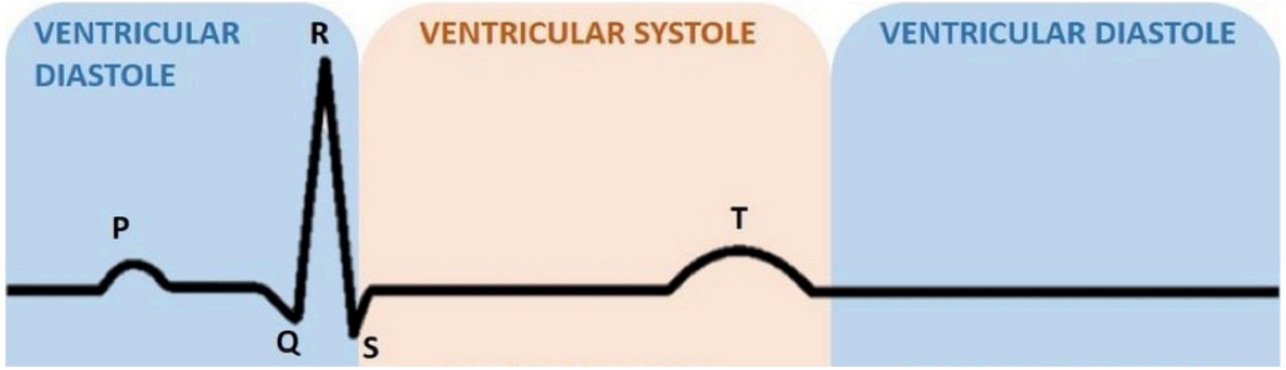
Extracting information from data



Source: Voytek, *Uber, Inc.*

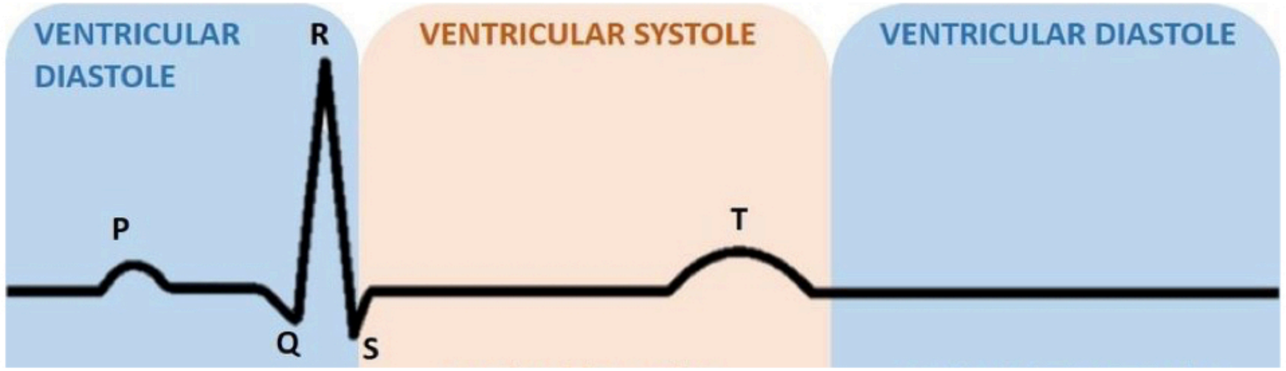

Dr. Muriel Landau

Extracting information from data



P Wave: Atrial depolarization (atria contract).
Q Wave: Initial ventricular depolarization (inter-ventricular septum).
R Wave: Main ventricular depolarization (ventricles contract).
S Wave: Final ventricular depolarization.
T Wave: Ventricular repolarization (ventricles recover).

Extracting information from data



P Wave: Atrial depolarization (atria contract).
Q Wave: Initial ventricular depolarization (inter-ventricular septum).
R Wave: Main ventricular depolarization (ventricles contract).
S Wave: Final ventricular depolarization.
T Wave: Ventricular repolarization (ventricles recover).