AGENDA

1. Data Science Process

2. Distributions and Random Variables

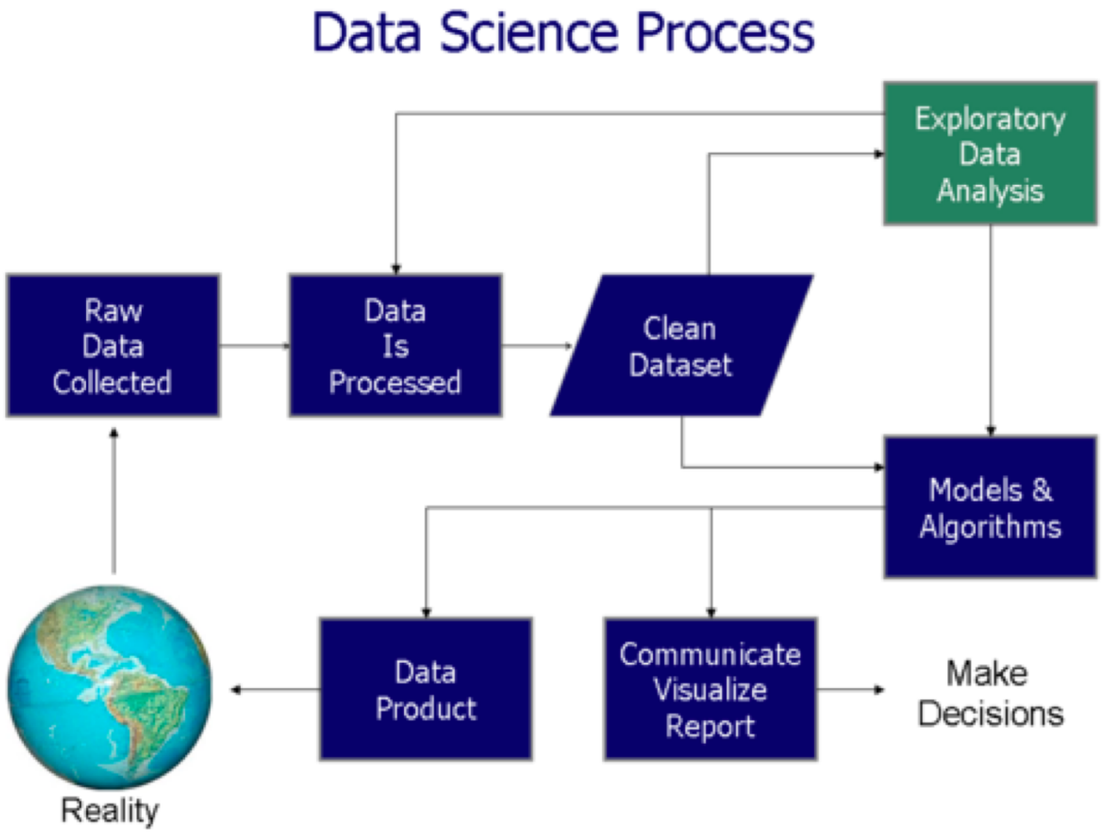
3. Discrete Distributions

DATA SCIENCE PROCESS

- The process of doing data science is usually not linear.

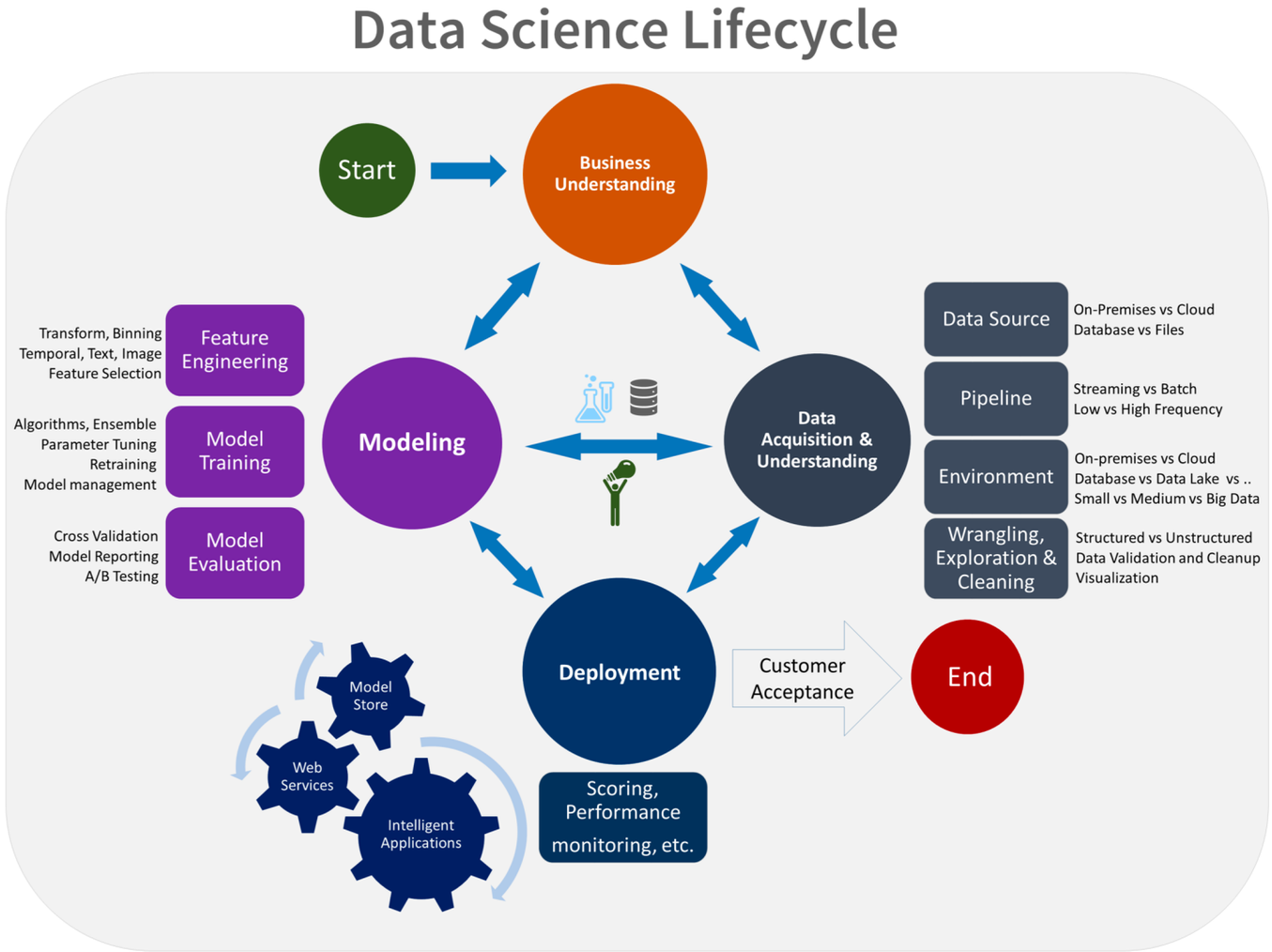
DATA SCIENCE PROCESS

- The process of doing data science is usually not linear.



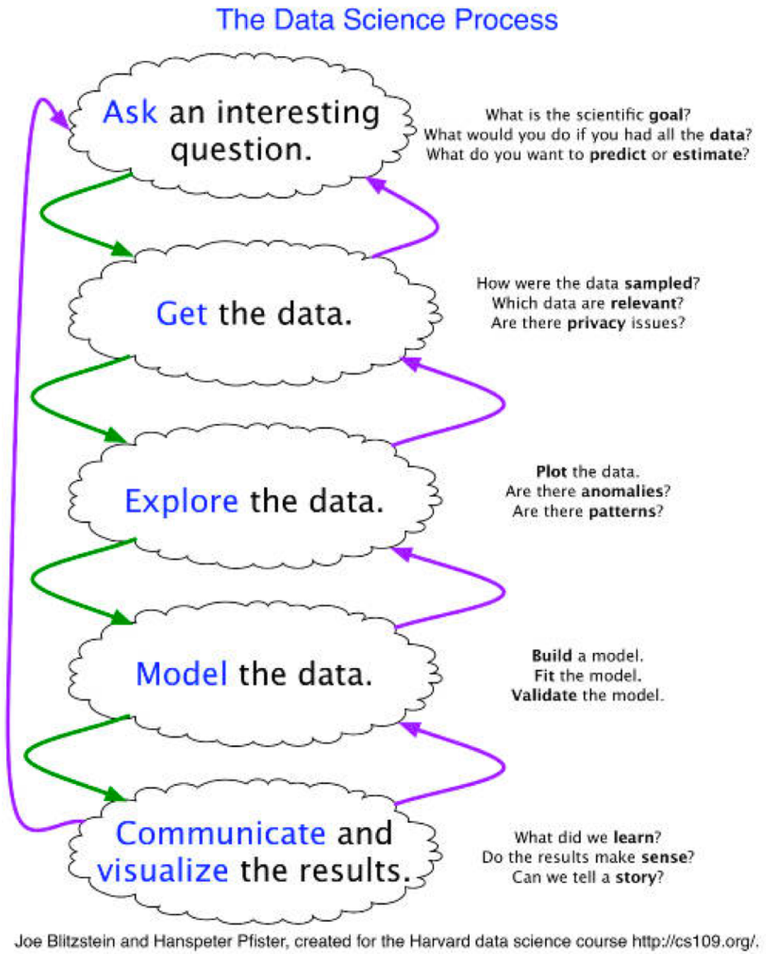
DATA SCIENCE PROCESS

- The process of doing data science is usually not linear.



DATA SCIENCE PROCESS

- The process of doing data science is usually not linear.



DATA SCIENCE PROCESS

- The process of doing data science is usually not linear.
- That doesn't mean that we won't try to make it as linear as possible!

DATA SCIENCE PROCESS

1. Define problem.

DATA SCIENCE PROCESS

1. Define problem.

2. Gather data.

DATA SCIENCE PROCESS

1. Define problem.

2. Gather data.

3. Explore data.

DATA SCIENCE PROCESS

1. Define problem.
2. Gather data.
3. Explore data.
4. Model with data.

DATA SCIENCE PROCESS

1. Define problem.
2. Gather data.
3. Explore data.
4. Model with data.
5. Evaluate model.

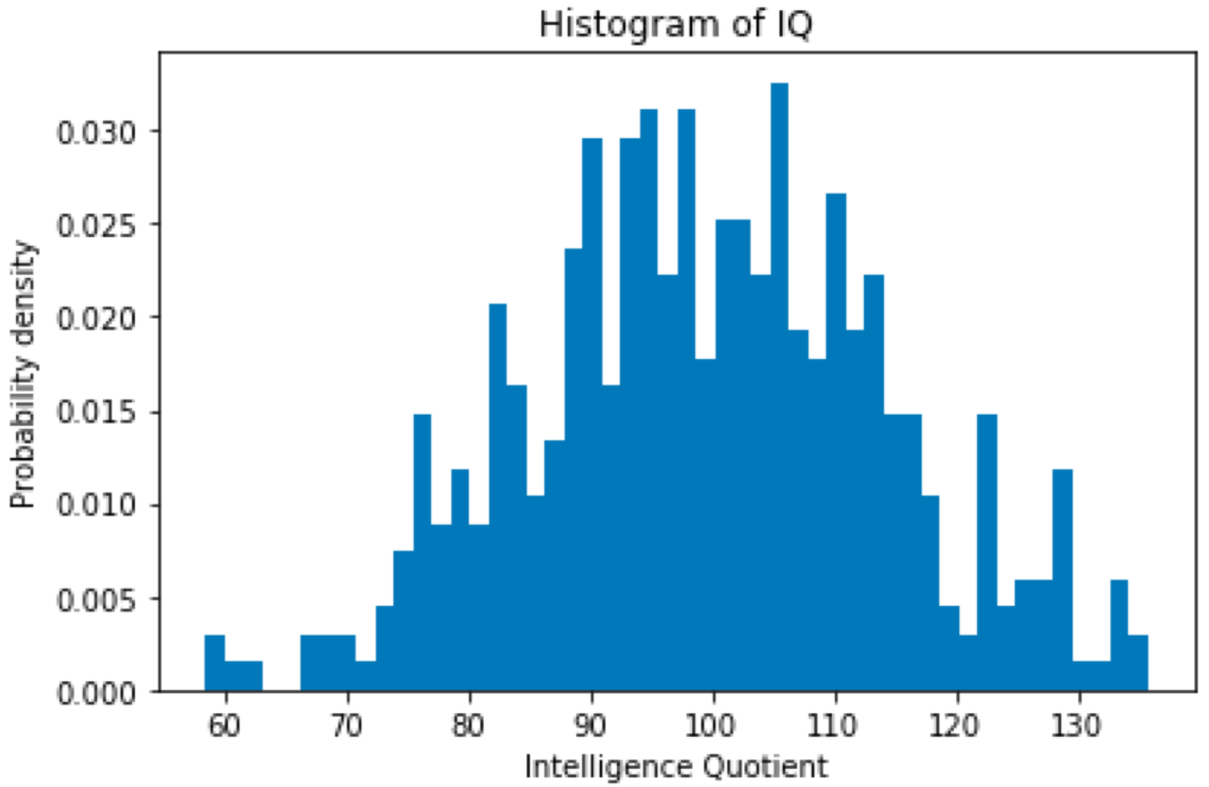
DATA SCIENCE PROCESS

1. Define problem.
2. Gather data.
3. Explore data.
4. Model with data.
5. Evaluate model.
6. Answer problem.

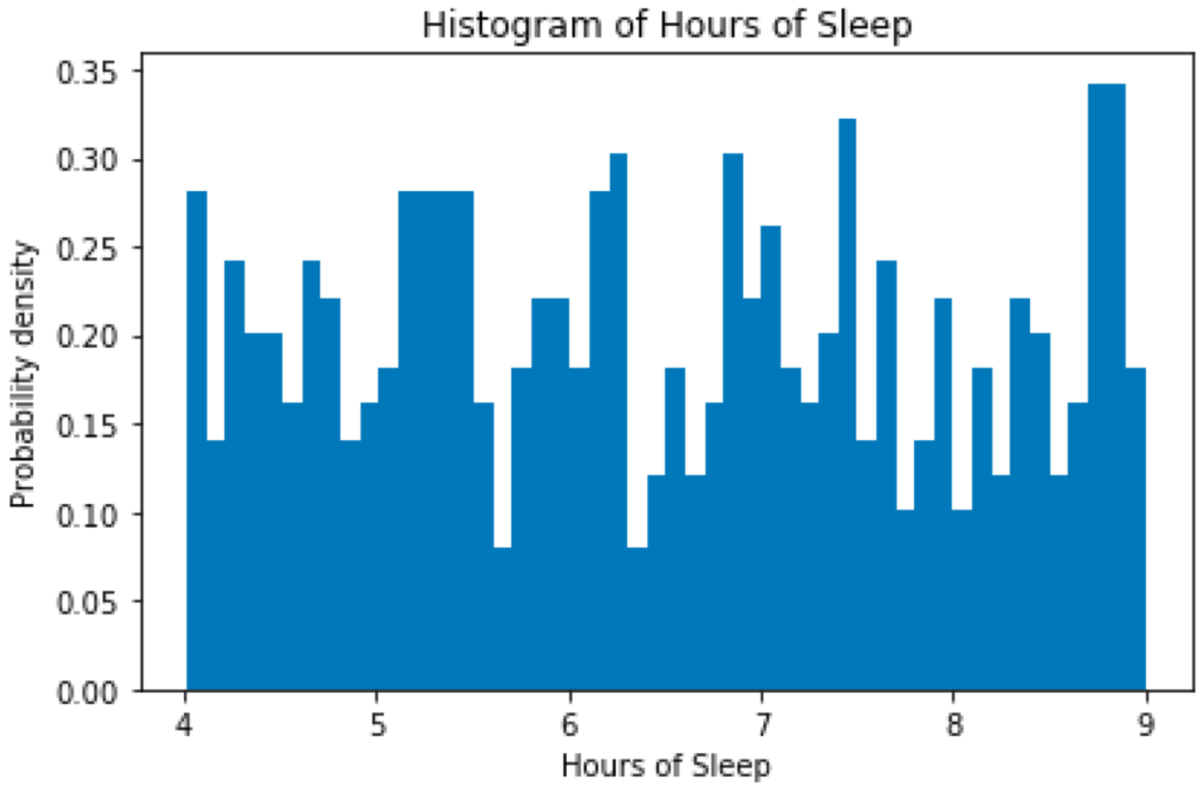
EXPLORING DATA

- Let's suppose you're studying three phenomena:
  - The intelligence quotient (IQ) of individuals.
  - The number of hours of sleep individuals get in a night.
  - The income of individuals.

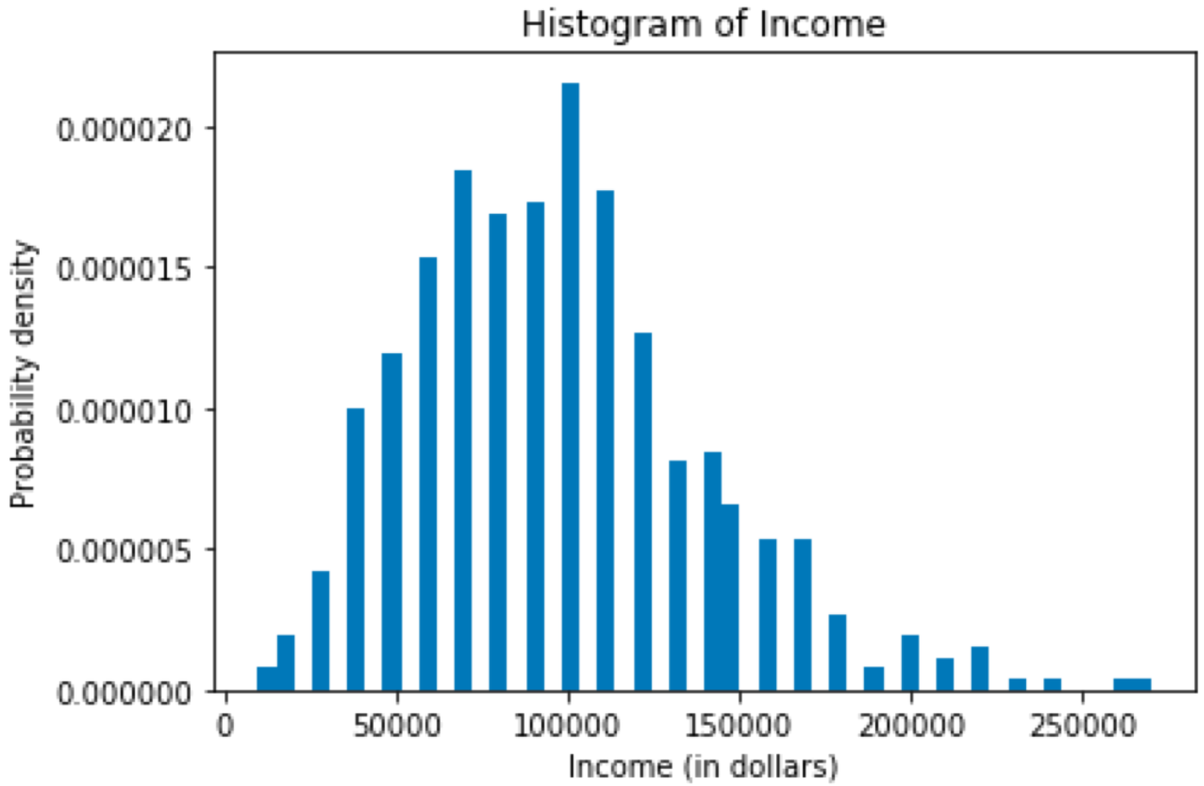
EXPLORING DATA – HISTOGRAM 1



EXPLORING DATA – HISTOGRAM 2



EXPLORING DATA – HISTOGRAM 3



EXPLORING DATA

- We just saw three **distributions**.
    - A **distribution** is the set of all values of a variable and how frequently we observe each of those values.

EXPLORING DATA

- We just saw three **distributions**.
  - A **distribution** is the set of all values of a variable and how frequently we observe each of those values.

- Whether we're describing our own data or trying to communicate it to someone else, looking at the distribution of one variable is usually a really good place to start.
  - However, if we want to summarize our distribution, we usually want to focus on three aspects.
  - Even though these histograms were of completely different information and each look very different, what did we consistently describe in each histogram?

OTHER TERMINOLOGY

- An **experiment** is an infinitely-repeatable procedure with a well-defined set of outcomes.

OTHER TERMINOLOGY

- An **experiment** is an infinitely-repeatable procedure with a well-defined set of outcomes.

- The **sample space** for an experiment is the set of all possible outcomes of an experiment.

OTHER TERMINOLOGY

- An **experiment** is an infinitely-repeatable procedure with a well-defined set of outcomes.

- The **sample space** for an experiment is the set of all possible outcomes of an experiment.

- A **random variable** is any function that maps our sample space to the real number line.

OTHER TERMINOLOGY

- An **experiment** is an infinitely-repeatable procedure with a well-defined set of outcomes.
- Example: Flip a coin twice.
- The **sample space** for an experiment is the set of all possible outcomes of an experiment.
- $S =$
- A **random variable** is any function that maps our sample space to the real number line.
- Example:

OTHER TERMINOLOGY

- An **experiment** is an infinitely-repeatable procedure with a well-defined set of outcomes.
- Example: Time slept.
- The **sample space** for an experiment is the set of all possible outcomes of an experiment.
- $S =$
- A **random variable** is any function that maps our sample space to the real number line.
- Example:

DISCRETE VS. CONTINUOUS

- When we flipped two coins, our sample space was **discrete**.
    - I can count up the number of items in the sample space.

DISCRETE VS. CONTINUOUS

- When we flipped two coins, our sample space was **discrete**.
    - I can count up the number of items in the sample space.

- When we recorded the time slept, our sample space was **continuous**.
    - I cannot count up the number of items in the sample space.

DISCRETE VS. CONTINUOUS CHECK

- Are each of the following discrete or continuous?

    1. The number of shoppers who come into my store.

    2. The probability that an individual votes in the upcoming election.

    3. The weight of a shipping container at the Port of Los Angeles.

DISCRETE VS. CONTINUOUS CHECK

- Are each of the following discrete or continuous?

1. The number of shoppers who come into my store.
   - Answer: Discrete

2. The probability that an individual votes in the upcoming election.
   - Answer: Continuous

3. The weight of a shipping container at the Port of Los Angeles.
   - Answer: Continuous

DISTRIBUTIONS

- Thinking forward for the rest of the class, it'll sometimes be convenient for us to make assumptions about how data are distributed.

- There are distributions that are common enough that they have a name.

DISTRIBUTIONS

- Thinking forward for the rest of the class, it'll sometimes be convenient for us to make assumptions about how data are distributed.

- There are distributions that are common enough that they have a name.

    - I might assume the number of people who log onto my website follows a Poisson distribution.

    - I build a model predicting how long my commute is and I might assume that my errors (how far my predictions are from the truth) follow a Normal distribution.