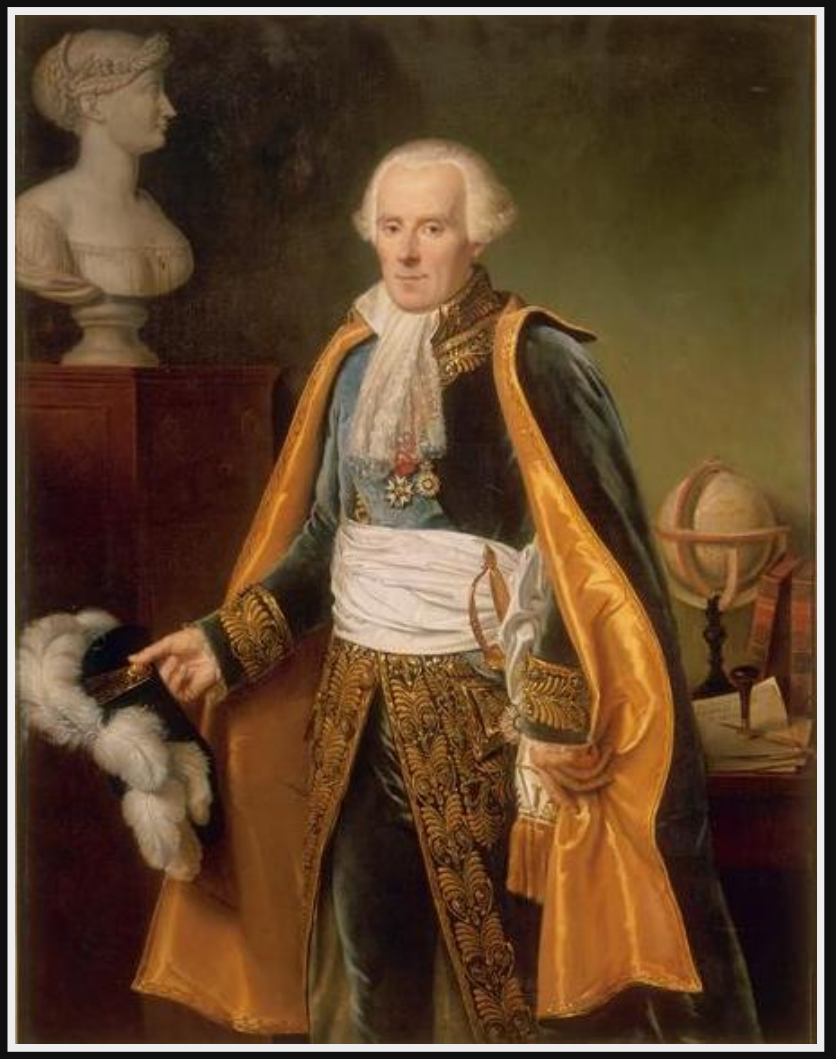
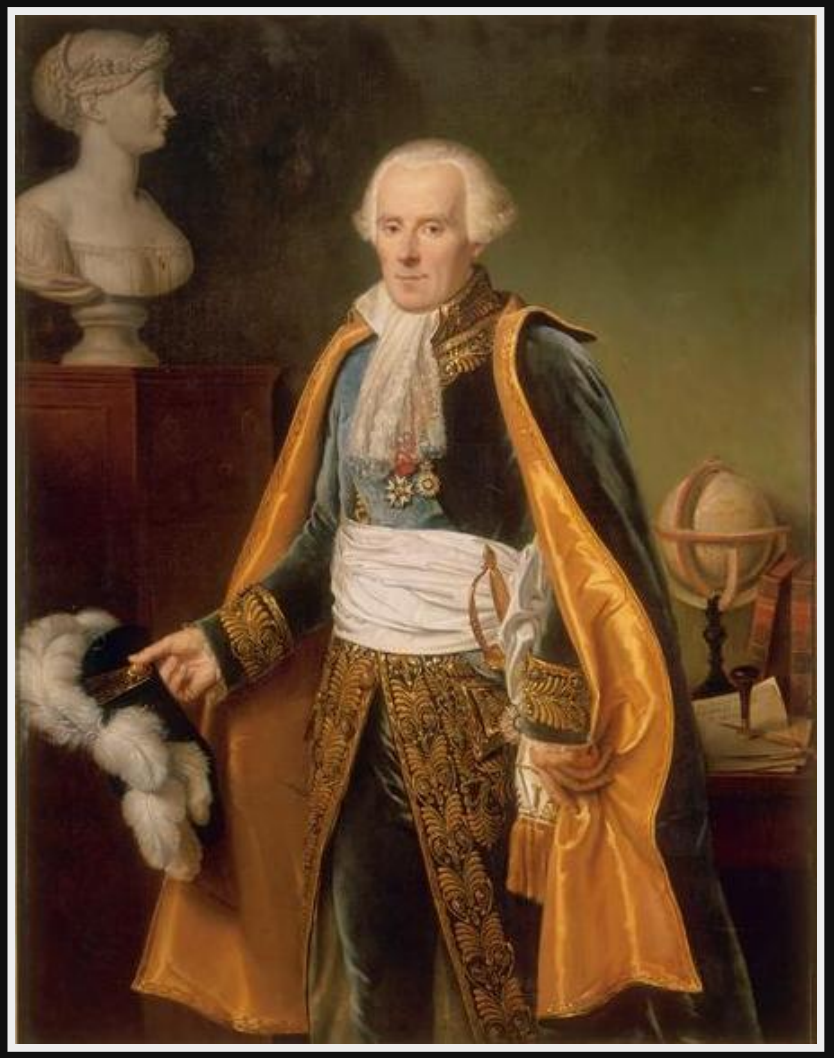
Foundations of Data Analysis

Lecture 1: Introduction





Pierre-Simon Laplace (1749–1827)

Births in Paris

1745 - 1770

Births in Paris

1745 - 1770

251,527 Boys

Births in Paris

1745 - 1770

251,527 Boys

241,945 Girls

Are males born at a higher rate than females?

251,527 Boys

241,945 Girls

Some possible statistics

251,527 Boys

241,945 Girls

Some possible statistics

251,527 Boys 241,945 Girls

- **Difference:** +9,582 Boys

Some possible statistics

251,527 Boys 241,945 Girls

- **Difference:** +9,582 Boys
- **Ratio:** 104 Boys to 100 Girls

Some possible statistics

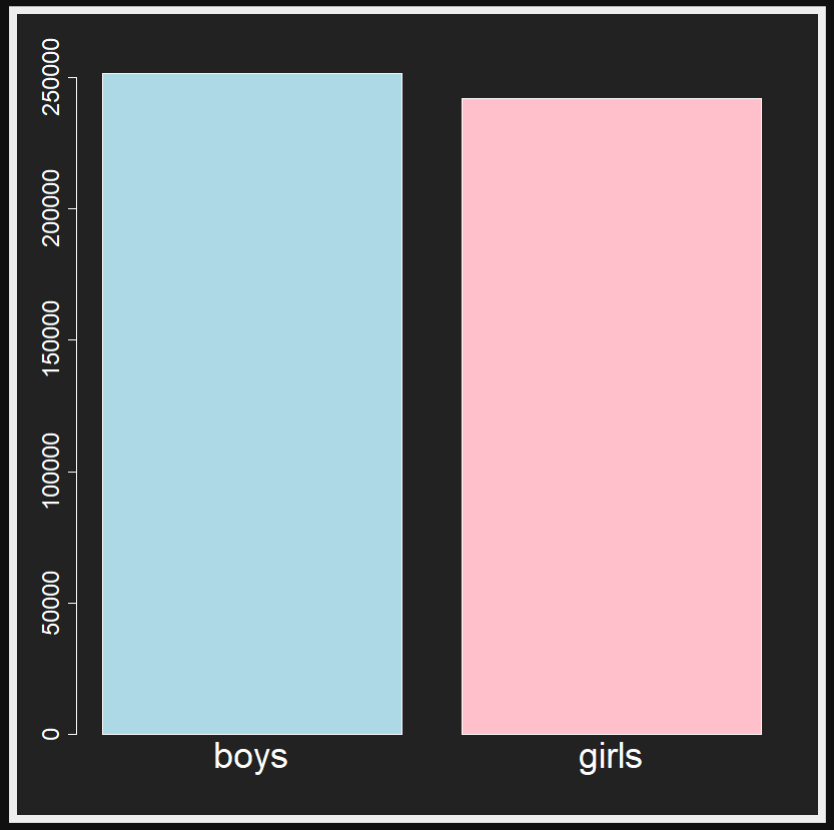
251,527 Boys 241,945 Girls

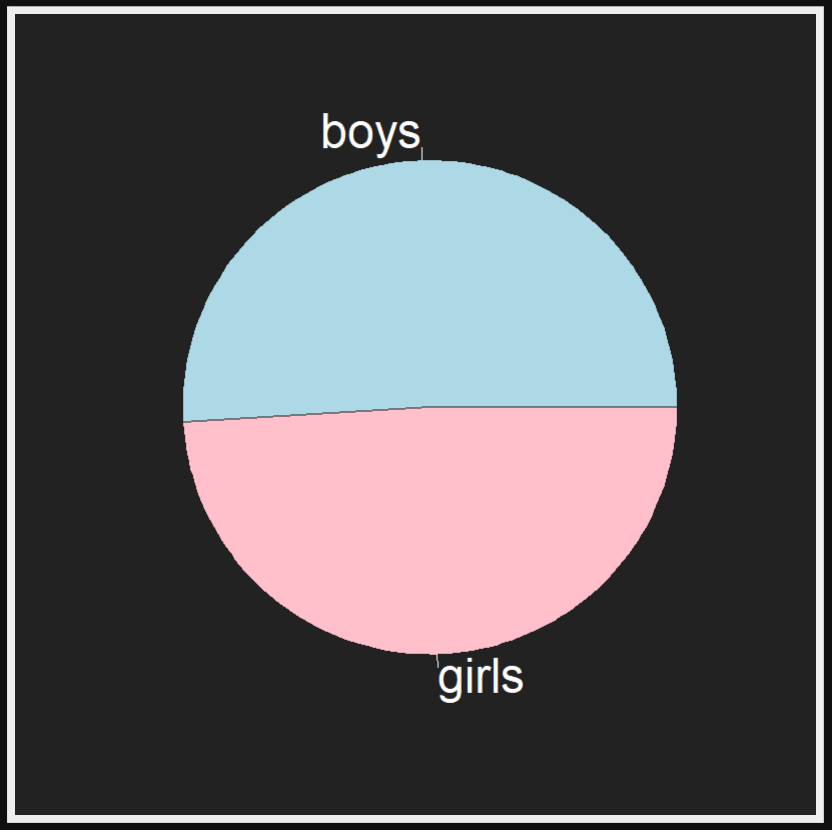
- **Difference:** +9,582 Boys
- **Ratio:** 104 Boys to 100 Girls
- **Proportion:** 50.97% Boys

Some possible visualizations

251,527 Boys

241,945 Girls





How did Laplace solve this?

How did Laplace solve this?

Conditional Probability that:

rate of boys, θ, is greater than girls,

given

observed data

Answer?

Answer?

$$P(\theta > 0.5 \mid \text{data}) = 1 - \epsilon,$$

where $\epsilon \approx\approx 1 \times 10^{-42}$.

What is probability?

What is probability?

Definition: *Probability* is the study of the mathematical rules that govern random events.

But what is randomness?

But what is randomness?

Informally, a *random event* is an event where we do not know the outcome without observing it.

But what is randomness?

Informally, a *random event* is an event where we do not know the outcome without observing it.

Probability tells us what we can say about such events, given our assumptions about the possible outcomes.

What is statistics?

What is statistics?

Definition: *Statistics* is the application of probability to the collection, analysis, and description of random data.

Statistics is used to:

Statistics is used to:

- **Design** experiments

Statistics is used to:

- **Design** experiments
- **Summarize** data

Statistics is used to:

- **Design** experiments
- **Summarize** data
- **Make conclusions** about the world

Statistics is used to:

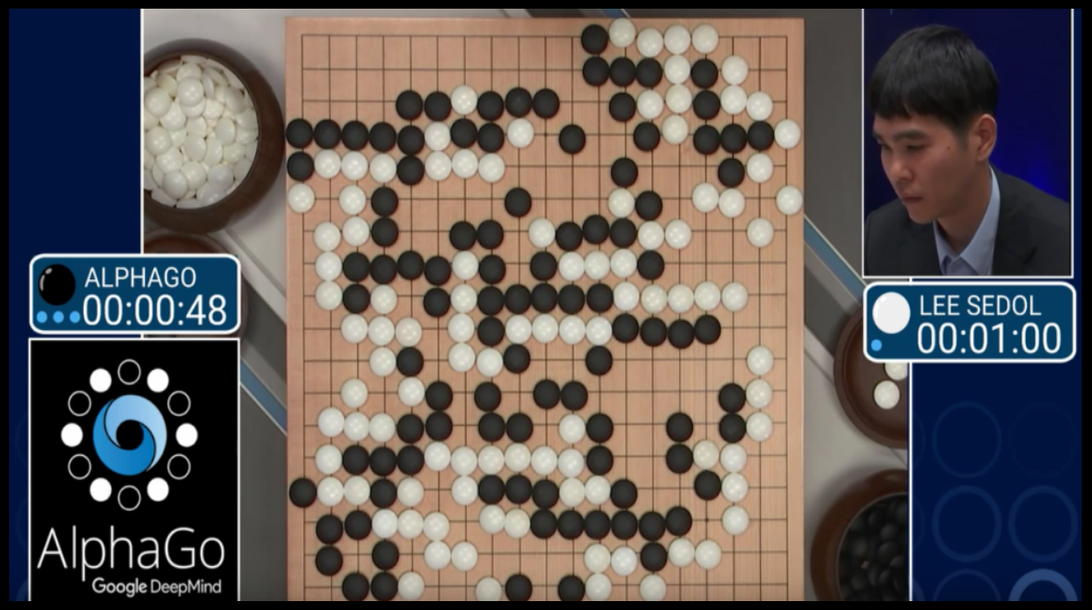
- **Design** experiments
- **Summarize** data
- **Make conclusions** about the world
- **Explore** complex data

What is machine learning?

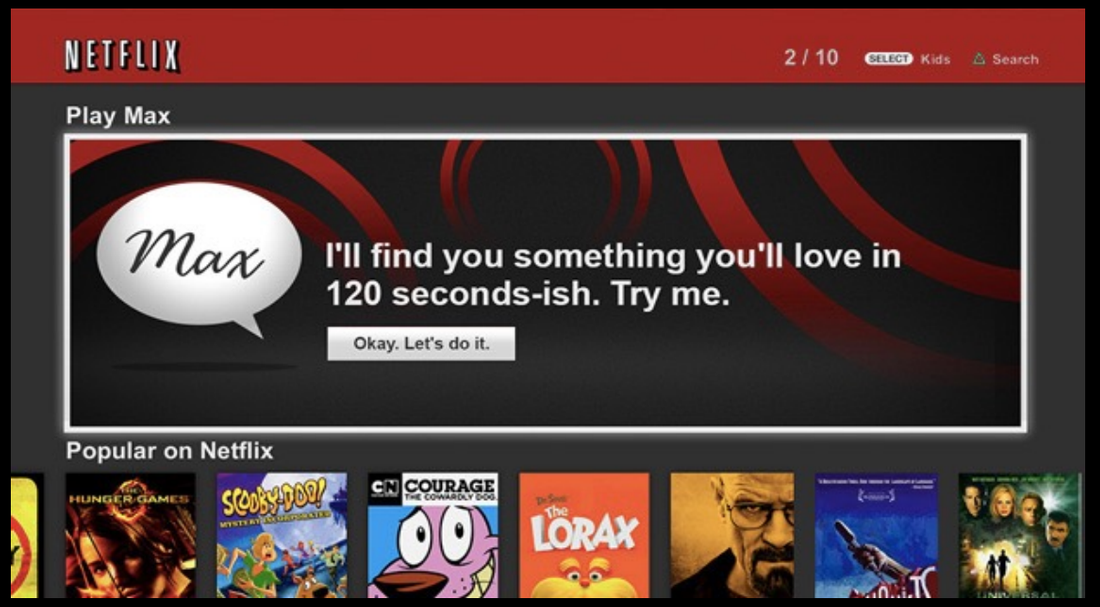
What is machine learning?

Definition: *Machine Learning* builds statistical models of data in order to recognize complex patterns and to make decisions based on these observations.

Machine Learning is Everywhere?



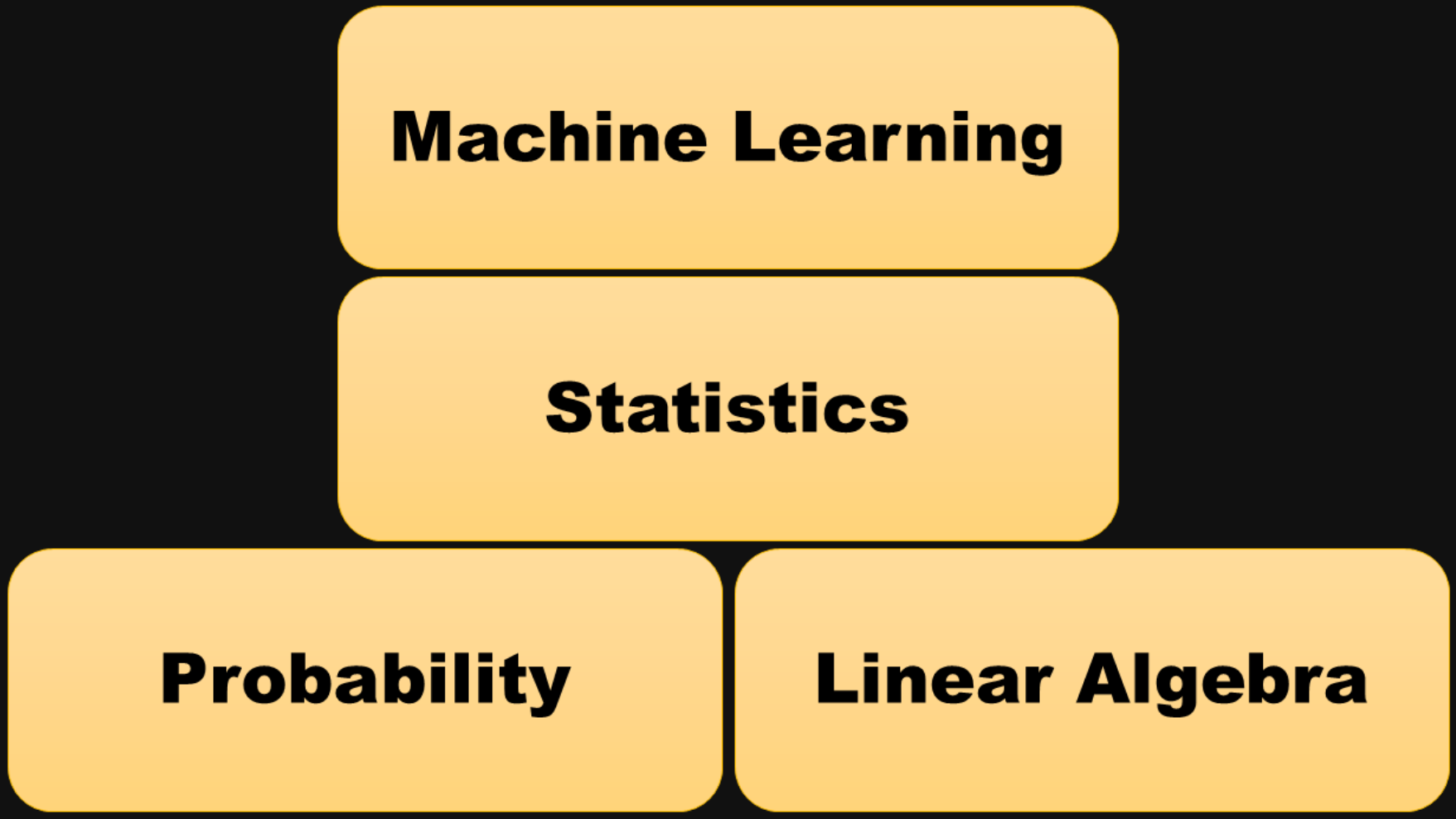
Chess player



Recommendation system

Assisted driving

Cancer diagnosis


Machine Learning
Statistics
Probability
Linear Algebra

Levels of data analysis expertise

Levels of data analysis expertise

0: What is data?

Levels of data analysis expertise

0: What is data?
1: I know how to run data analysis software

Levels of data analysis expertise

0: What is data?
1: I know how to run data analysis software
2: I understand the math behind the analysis

Levels of data analysis expertise

0: What is data?
1: I know how to run data analysis software
2: I understand the math behind the analysis
3: I'm able to invent new data analysis methods

Why should you know the mathematical foundations?

When machine learning goes wrong

from Goodfellow et al. ICLR 2015

When machine learning goes wrong

Panda (57.7% confidence)

from Goodfellow et al. ICLR 2015

When machine learning goes wrong

Panda (57.7% confidence)

Gibbon (99.3% confidence)

from Goodfellow et al. ICLR 2015