Prof. Alex Rogers
Department of Computer Science
University of Oxford
11th October 2020

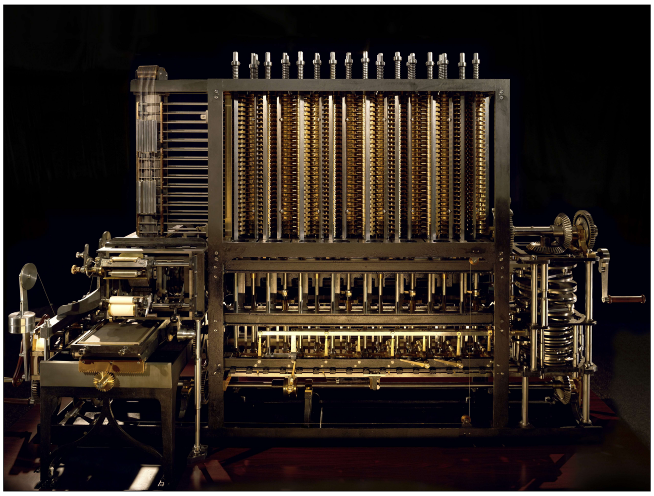
Probabilistic Machine Learning
Session 1

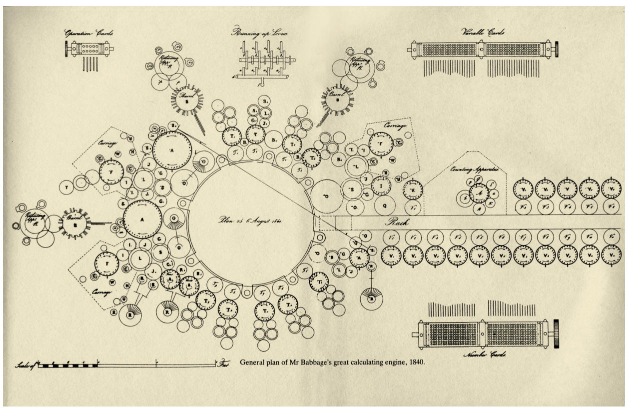
Contents

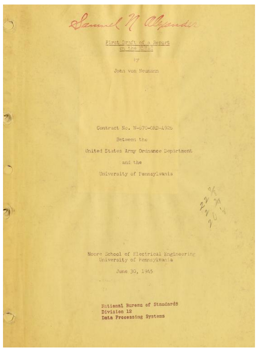
Introductions

University of Oxford

Computer Architecture







Samuel N. Alexander

First Draft of a Report on the EDVAC

by

John von Neumann

Contract No. W-670-ORD-4926

Between the

United States Army Ordnance Department

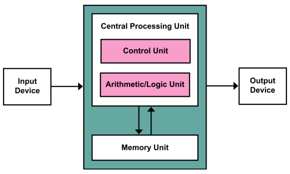
and the

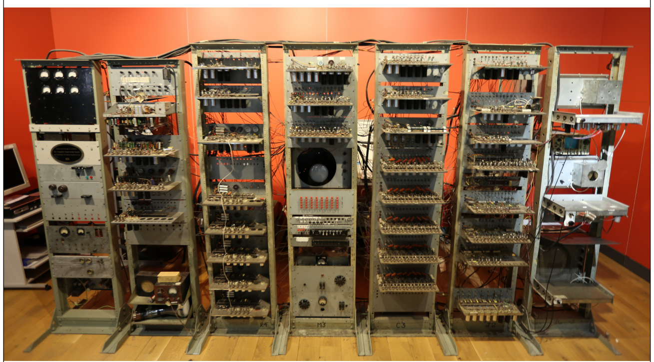
University of Pennsylvania

Moore School of Electrical Engineering
University of Pennsylvania

June 30, 1945

National Bureau of Standards
Division 12
Data Processing Systems

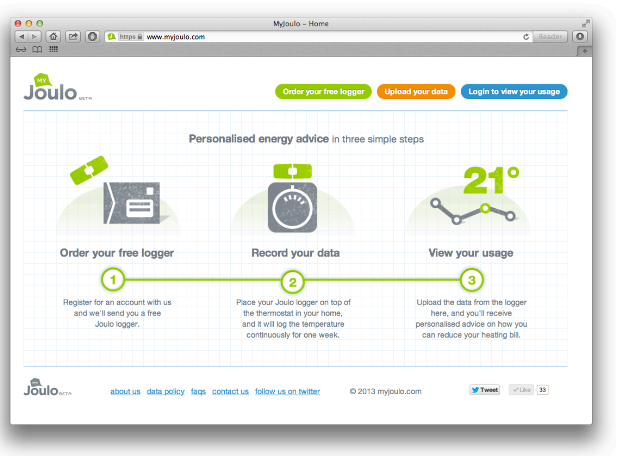


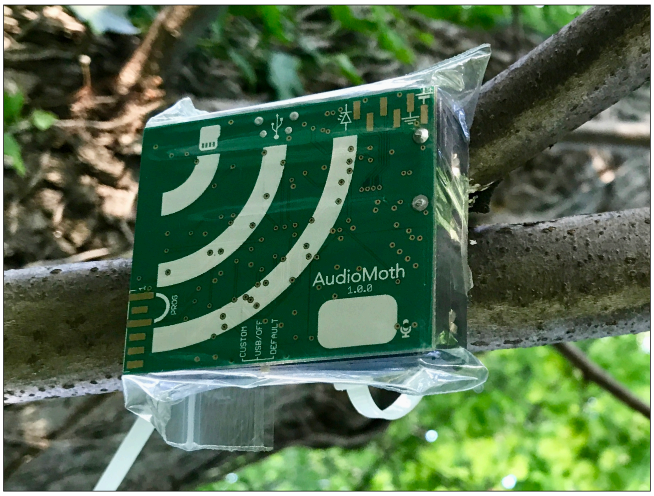


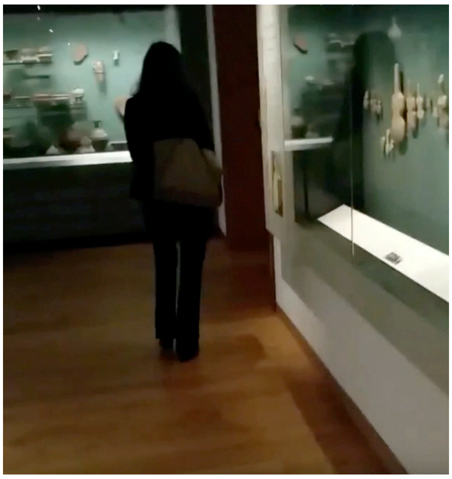
SPECTRE

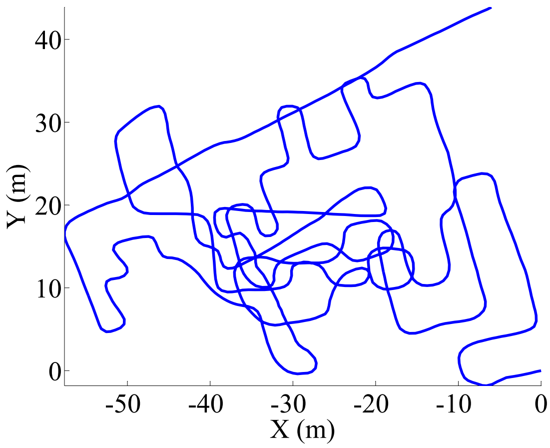
MELTDOWN

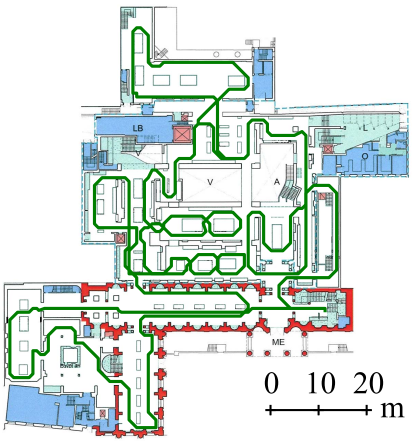
Cyber-Physical Systems









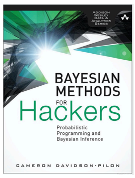


Probabilistic Machine Learning

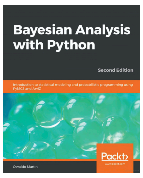
1. To understand the fundamentals of probability theory, Bayes rule and Bayesian inference.

2. To understand how to use PyMC3, a powerful probabilistic programming framework based on Python, to build sophisticated probabilistic models.

3. To develop experience applying these models to solve real-world machine learning and data analysis tasks.



Probabilistic Programming and Bayesian Methods for Hackers
Cameron Davidson-Pilon
https://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/



Bayesian Analysis with Python
Osvaldo Martin

GitHub

https://github.com/AlexRogersCS/probabilistic_machine_learning

Python 3.8, Numpy, Matplotlib, Jupyter

https://www.anaconda.com/products/individual

PyMC3

https://docs.pymc.io

Google Colab Notebooks

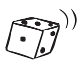
https://colab.research.google.com/

Motivating Examples

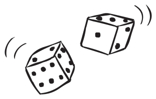
Probability Theory

Single and Joint Probabilities

Probability theory typically describes random variables, $P(X)$, and the probability that these random variables take on actual values.



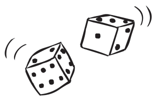
$$P(D = 6) = 1/6$$



$$P(D_1 = 6, D_2 = 6) = 1/36$$

Independent Probabilities

When events are **independent** their **joint** probability is given by the product of their individual probabilities.

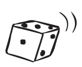


$$P(D_1 = 6, D_2 = 6) = 1/36$$

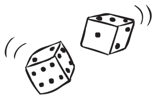
$$P(D_1 = 6, D_2 = 6) = P(D_1 = 6) \times P(D_2 = 6)$$

Conditional Probabilities

Conditional probabilities describe the probability of any event **given** the occurrence of another event.



$$P(D = 6 | D \text{ is even}) = 1/3$$



$$P(D_1 = 6 | D_1 + D_2 > 8) = \ ?$$

The Rules of Probabilities

There are just two fundamental rules of probability theory.

Sum rule $$P(X) = \sum_Y P(X,Y)$$

Product rule $$P(X,Y) = P(Y|X) \times P(X)$$

Plus the observation that probabilities must sum to 1.

Frequentist probability is the interpretation of probability that defines an event's probability as the limit of its relative frequency in a large number of trials.

Bayesian probability is the interpretation of the probability that holds that it can be defined as the degree to which a person (or community) believes that a proposition is true.

Rev. Thomas Bayes (1702 - 1761)

- Framework for plausible reasoning developed over the nineteenth century
- Developed by three physicists:
  - Harold Jeffreys (1891 – 1989)
  - Richard Cox (1898 - 1991)
  - Edwin Jaynes (1922 – 1998)
- Growth of computational power since 2000 has seen this become one of the dominant model for much of machine learning.
- Recent development of probabilistic programming make using Bayesian reasoning ever more accessible.

Cox's axioms describe desirable properties of a framework for reasoning about uncertainty

1. Degrees of plausibility are represented by real numbers.
2. Qualitative correspondence with common sense:
   - Increasing the plausibility of a statement, decreases the plausibility of its negative.
3. Consistency:
   - If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.
   - It must be possible to incorporate all the evidence relevant to a question into the reasoning process.
   - All equivalent states of knowledge are represented by equivalent plausibility assignments.

Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Theorem

$$\underbrace{P(H|D)}_{\textbf{Posterior}} = \frac{\overbrace{P(D|H)}^{\textbf{Likelihood}}\overbrace{P(H)}^{\textbf{Prior}}}{\underbrace{P(D)}_{\textbf{Evidence}}}$$

Example 1



- Suppose that Bob can decide to go to work by one of three modes of transportation, car, bus, or train.
  - Because of high traffic, if he decides to go by car, there is a 50% chance he will be late.
  - If he goes by bus, which has special reserved lanes but is sometimes overcrowded, the probability of being late is only 20%.
  - The train is almost never late, with a probability of only 1%, but is more expensive than the bus.

Example 1



- Suppose that Bob is late one day, and his boss wishes to estimate the probability that he drove to work that day by car.
  - Since he does not know which mode of transportation Bob usually uses, he gives a prior probability of 1 of 3 to each of the three possibilities.
- What is the boss' estimate of the probability that Bob drove to work?

Example 1

$$P(\text{car}) = \frac{1}{3} \qquad P(\text{late}|\text{car}) = \frac{1}{2}$$

$$P(\text{bus}) = \frac{1}{3} \qquad P(\text{late}|\text{bus}) = \frac{1}{5}$$

$$P(\text{train}) = \frac{1}{3} \qquad P(\text{late}|\text{train}) = \frac{1}{100}$$

Example 1

$$P(\text{car}|\text{late}) = \frac{P(\text{late}|\text{car})P(\text{car})}{P(\text{late})}$$

$$= \frac{P(\text{late}|\text{car})P(\text{car})}{P(\text{late}|\text{car})P(\text{car}) + P(\text{late}|\text{bus})P(\text{bus}) + P(\text{late}|\text{train})P(\text{train})}$$

$$= \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3} + \frac{1}{5}\times\frac{1}{3} + \frac{1}{100}\times\frac{1}{3}}$$

$$\approx 0.704$$

Example 1 (cont)

- Bob's co-worker Alice know Bob's travel preferences much better than his boss.
  - She knows that Bob never takes the bus, and 9 out of 10 times he will take the train to work.
- What is the Alice's estimate of the probability that Bob drove to work?

Example 1 (cont)

$$P(\text{car}) = \frac{1}{10} \qquad P(\text{late}|\text{car}) = \frac{1}{2}$$

$$P(\text{bus}) = 0 \qquad P(\text{late}|\text{bus}) = \frac{1}{5}$$

$$P(\text{train}) = \frac{9}{10} \qquad P(\text{late}|\text{train}) = \frac{1}{100}$$

Example 1

$$\begin{aligned}
P(\text{car}|\text{late}) &= \frac{P(\text{late}|\text{car})P(\text{car})}{P(\text{late})} \\
&= \frac{P(\text{late}|\text{car})P(\text{car})}{P(\text{late}|\text{car})P(\text{car}) + P(\text{late}|\text{train})P(\text{train})} \\
&= \frac{\frac{1}{2}\times\frac{1}{10}}{\frac{1}{2}\times\frac{1}{10}+\frac{1}{100}\times\frac{9}{10}} \\
&\approx 0.848
\end{aligned}$$

Example 2

- A clinical test, designed to diagnose a specific illness, comes out positive for a certain patient.
- We are told that
  - The test is 79 percent reliable (that is, it misses 21 percent of actual cases)
  - On average, this illness affects 1 percent of the population in the same age group as the patient
  - The test has a false positive rate of 10 percent.
- Taking this into account and assuming you know nothing about the patient's symptoms or signs, what is the probability that this patient actually has the illness?

Example 2

$$P(\mathrm{D}) = 0.01$$
$$P(\mathrm{T}|\mathrm{D}) = 0.79$$
$$P(\mathrm{T}|\neg\mathrm{D}) = 0.1$$

Example 2

$$\begin{aligned} P(\mathrm{D}|\mathrm{T}) &= \frac{P(\mathrm{T}|\mathrm{D})P(\mathrm{D})}{P(\mathrm{T})} \\ &= \frac{P(\mathrm{T}|\mathrm{D})P(\mathrm{D})}{P(\mathrm{T}|\mathrm{D})P(\mathrm{D}) + P(\mathrm{T}|\neg\mathrm{D})P(\neg\mathrm{D})} \\ &= \frac{0.79\times 0.01}{0.79\times 0.01 + 0.1\times 0.99} \\ &\approx 0.074 \end{aligned}$$

Next Time

Bayesian Inference and Markov Chain Monte Carlo