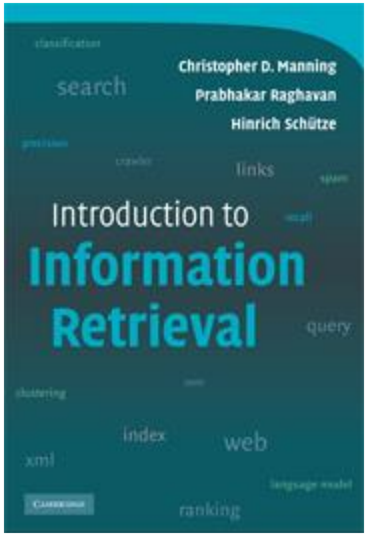
Information Retrieval and Organisation

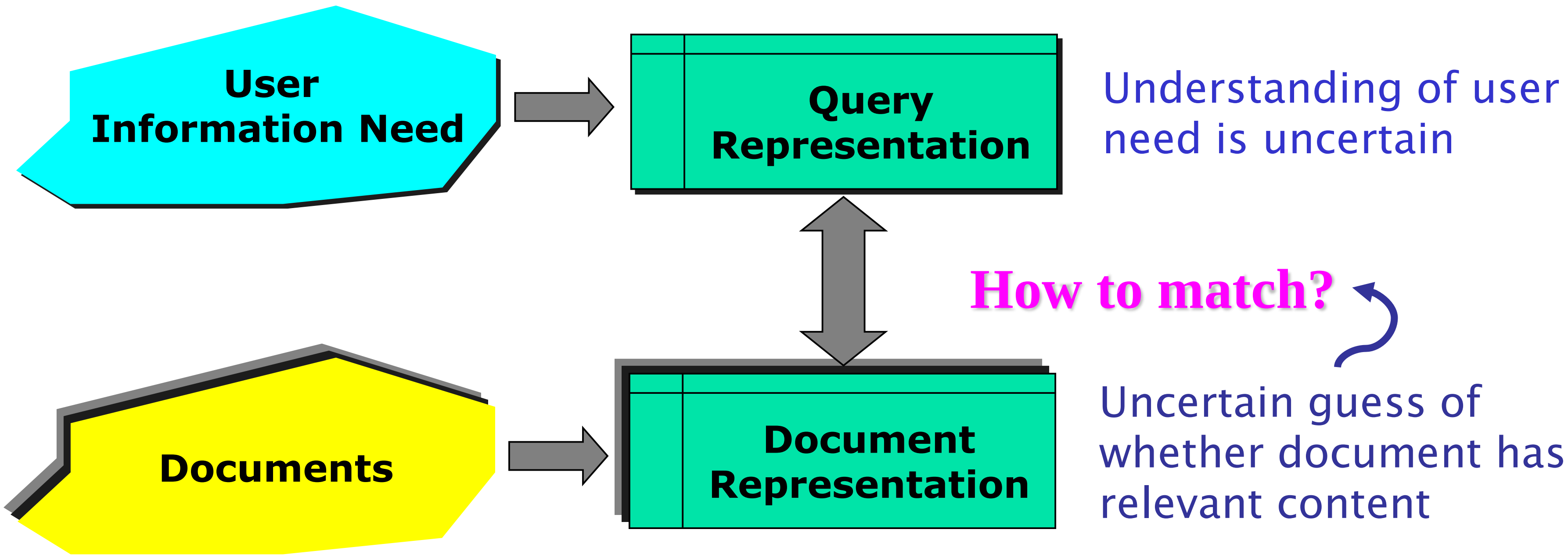


Chapter 11

Probabilistic Information Retrieval

Dell Zhang
Birkbeck, University of London

Why Probabilities in IR?



In IR systems, matching between each document and query is attempted in a semantically imprecise space of index terms.

Probabilities provide a principled foundation for uncertain reasoning.
Can we use probabilities to quantify our uncertainties?

Why Probabilities in IR?

- Problems with vector space model
  - Ad-hoc term weighting schemes
  - Ad-hoc basis vectors
  - Ad-hoc similarity measurement
- We need something more principled!

Probability theory is nothing but common sense reduced to calculation.

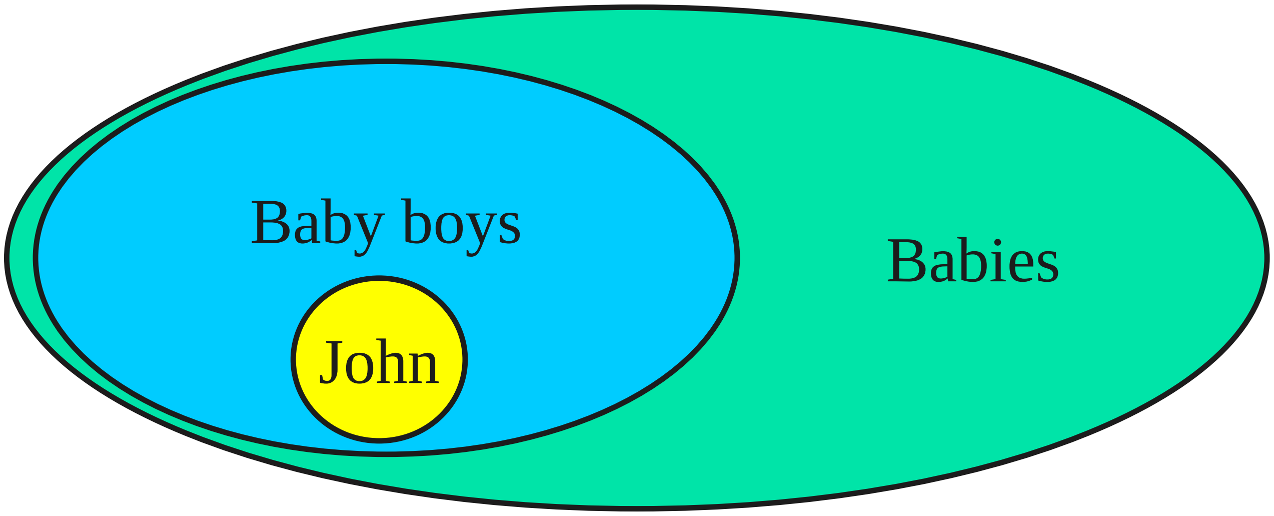
--- Pierre-Simon Laplace

The Bean Machine

- Wikipedia
  - http://en.wikipedia.org/wiki/Bean_machine
- Demonstration:
  - http://www.youtube.com/watch?v=9xUBhhM4vbM
- Simulation:
  - http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html

Probability

- $P(A)$ means **probability** that A is true
  - P(baby is a boy) ≈ 0.5 (% of total that are boys)
  - P(baby is named John) ≈ 0.001 (% of total named John)

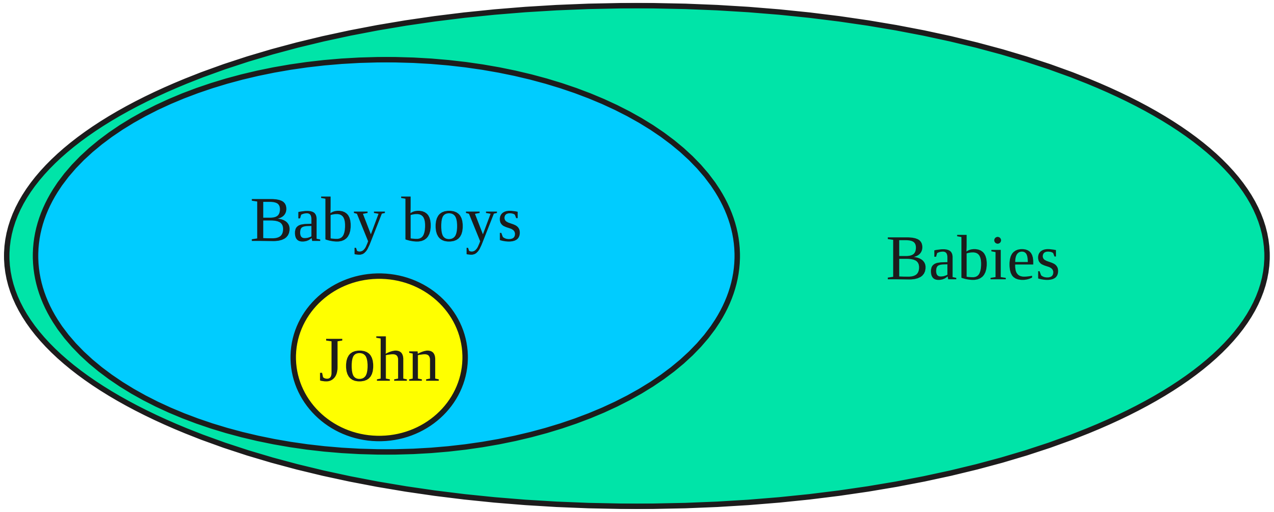


Odds

- The **odds** of an event A:

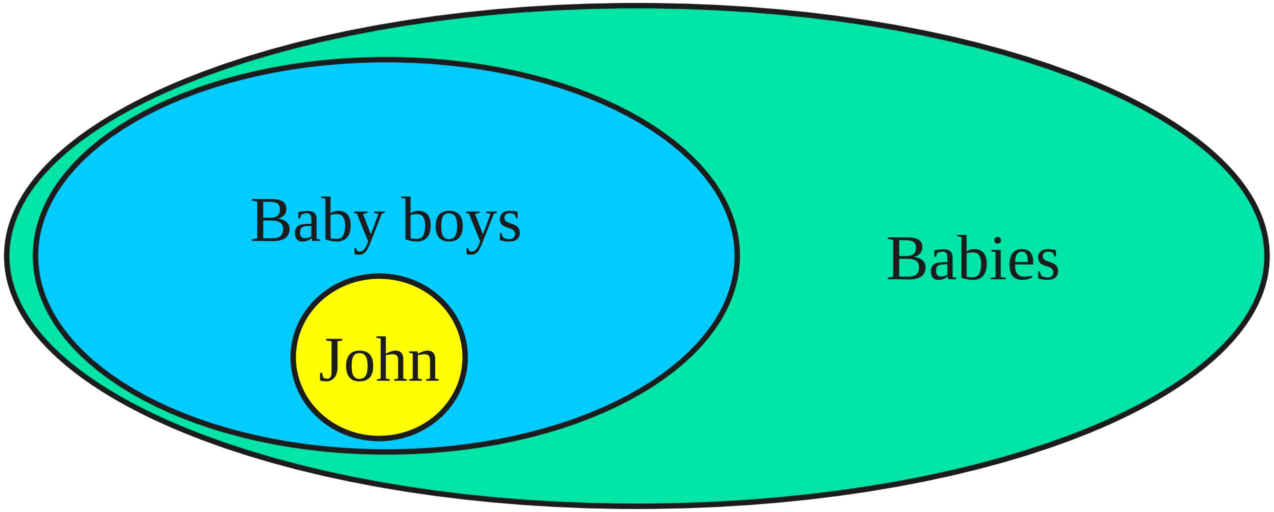
$$O(A) = \frac{P(A)}{P(\overline{A})} = \frac{P(A)}{1 - P(A)}.$$

  - O(baby is a boy) = 0.5 / 0.5 = 1
  - O(baby is named John) = 0.001 / 0.999 = 1/999



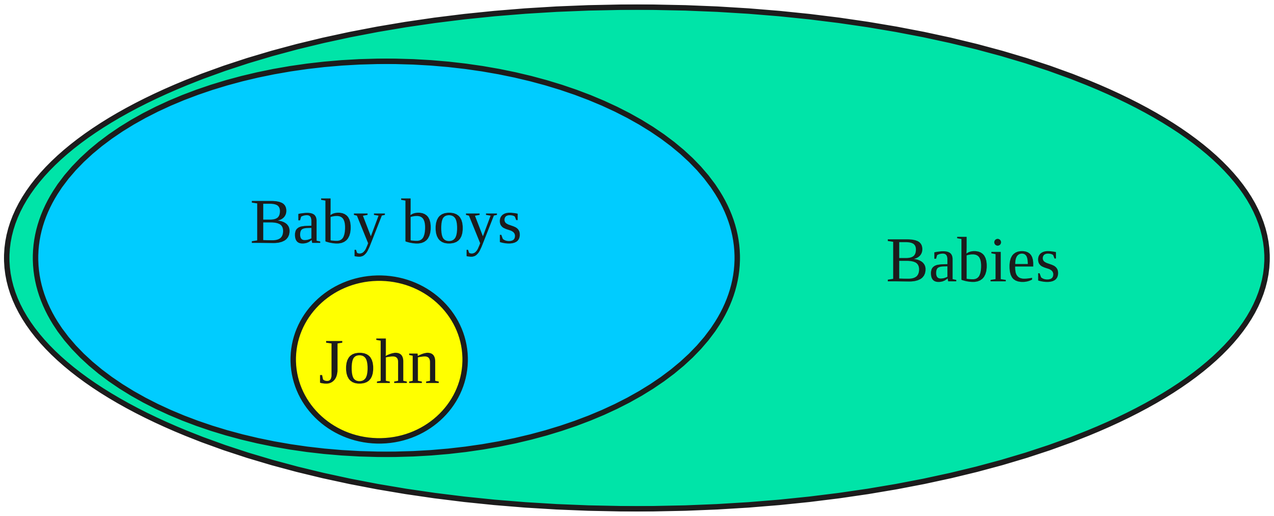
Joint Probability

- $P(A,B)$ means probability that A and B are both true
  - P(baby is named John, baby is a boy)



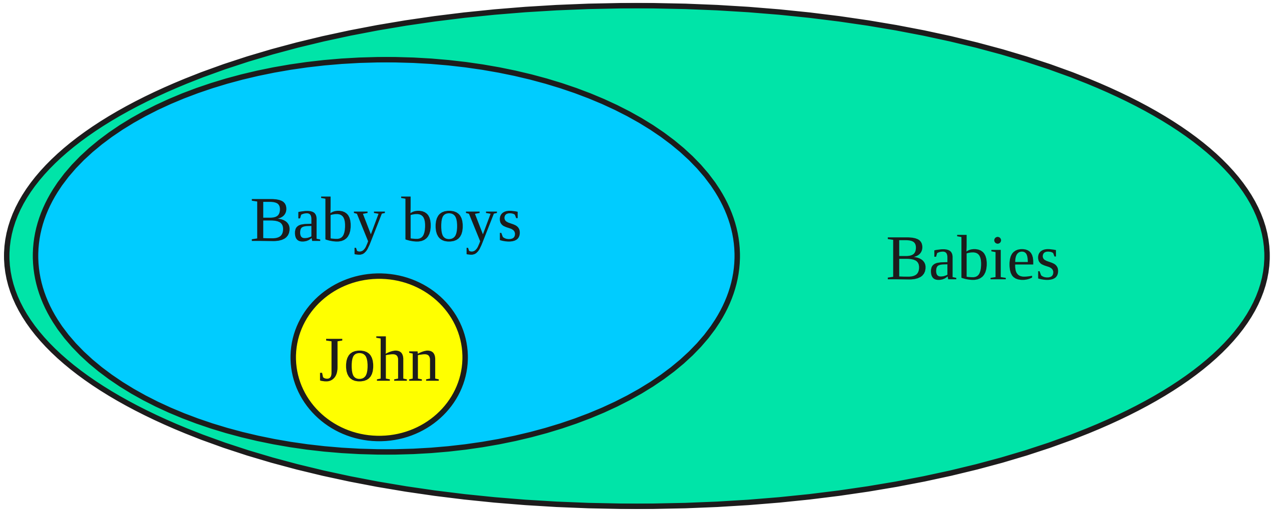
Conditional Probability

- $P(A|B)$ means probability that A is true when we already know B is true
  - P(baby is named John | baby is a boy) ≈ 0.002
  - P(baby is a boy | baby is named John) ≈ 1



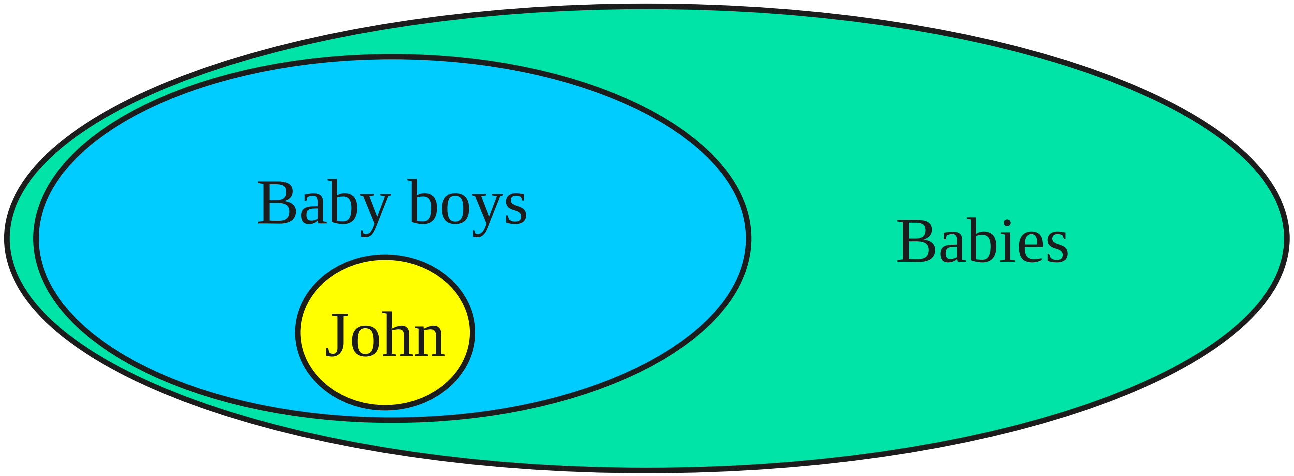
Basic Rules of Probability

- Chain Rule: $P(A, B) = P(A|B)P(B)$
  - P(named John, boy)
    = P(named John | boy) × P(boy)
    = 0.002 * 0.5 = 0.001



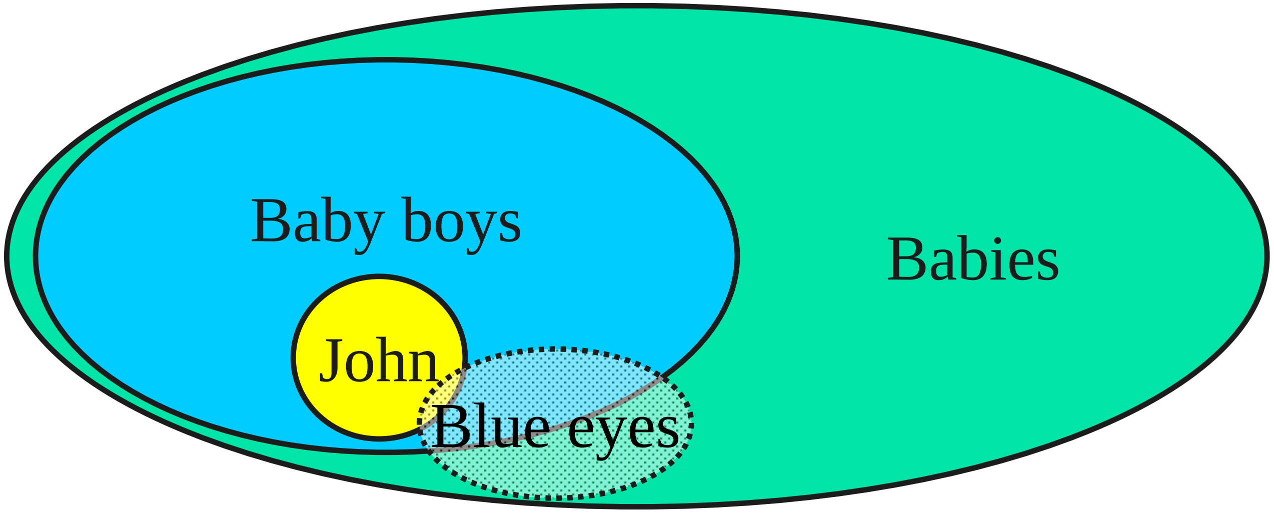
Basic Rules of Probability

- Partition Rule: $P(B) = P(A, B) + P(\overline{A}, B).$
  - P(boy)
    =P(named John, boy) + P(*not* named John, boy)
    = 0.001 + 0.499 = 0.5



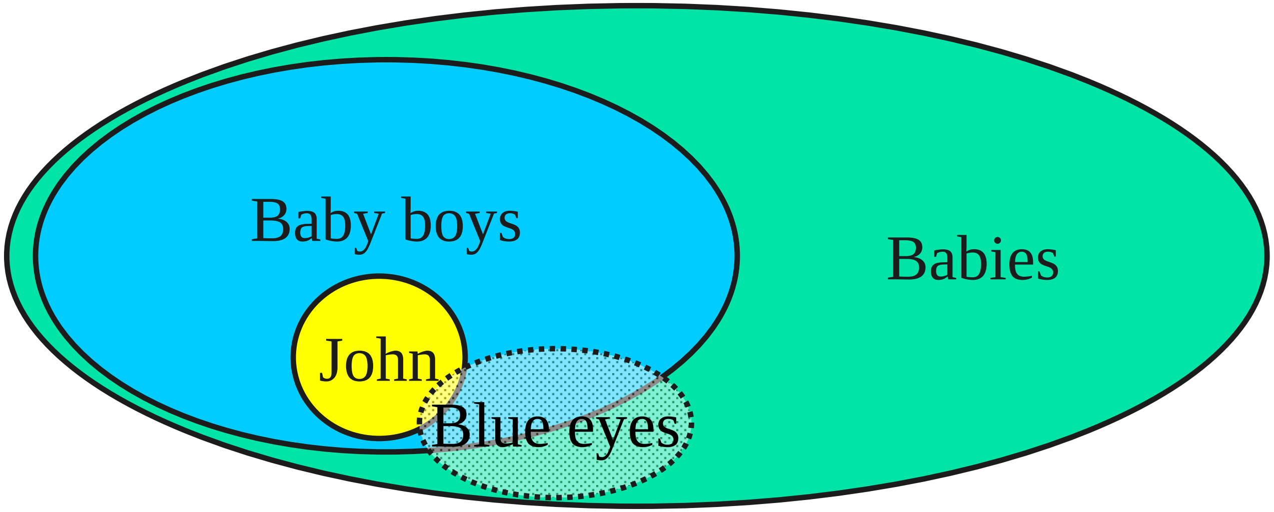
Independence

- $P(A,B) = P(A)P(B)$: A and B are **independent**
  - $P(\text{blue eyes, boy}) = P(\text{blue eyes}) \times P(\text{boy})$



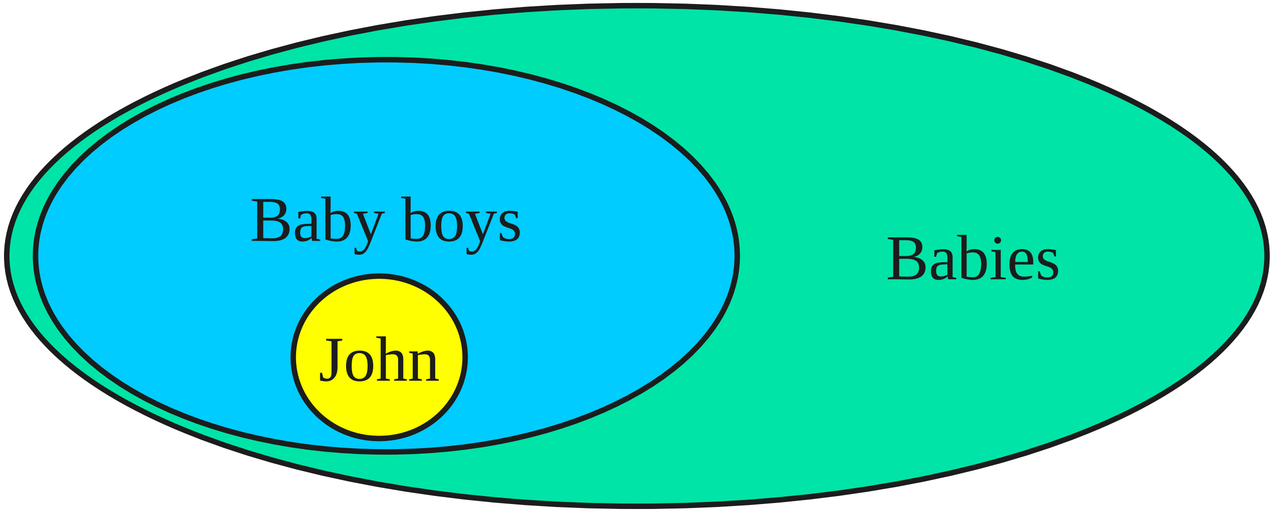
Conditional Independence

- $P(A,B|C) = P(A|C)P(B|C)$: A and B are **conditionally independent** given C
  - P(named John, blue eyes | boy) = P(named John | boy) × P(blue eyes | boy)



Bayes' Rule

- $P(A|B) = P(B|A)\,P(A) \,/\, P(B)$
  - P(named John | boy)
    $= P$(boy | named John) $\times$ P(named John) / P(boy)



Bayes' Rule

likelihood   prior probability

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \left[\frac{P(B|A)}{\sum_{X\in\{A,\overline{A}\}} P(B|X)P(X)}\right] P(A).$$

posterior probability

Example: British Weather

- You are about to set off into town to do some shopping.
- You will only be out for an hour or so, but rain has been forecasted, so what are you going to do?
- You know the forecasts are pretty good-around 80% accurate, in fact.
- So the chances that you will need an umbrella are 80%, right?

Example: British Weather

- Wrong, they're actually more like 30%.
- On the hourly timescales relevant to shopping trips, Britain's base-rate of rain is about 10%. That is, there is only a 1 in 10 chance of rain falling in any particular hour, and thus a 9 in 10 chance of rain not falling. And this has a significant impact on how much trust we can put in even an 80% reliable forecast.

Example: British Weather

- Contingency Table

	Rain falls	No rain falls	Total
Forecast of rain	8	18	26
Forecast of no rain	2	72	74
Total	10	90	100

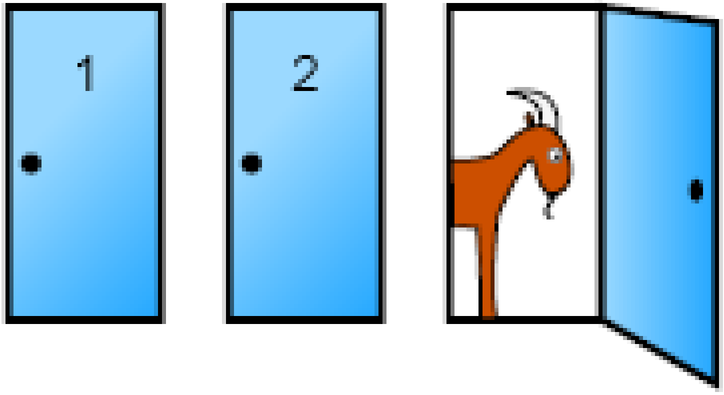
Robert Matthews, How right can you be?, New Scientist, 08 March 1997.

Example: British Weather

- In the language of probabilities
  - we can estimate P(rain-falls | forecast-of-rain)
  - by using Bayes' rule to combine the knowledge of P(rain-falls) and P(forecast-of-rain | rain-falls)!

Example: Monty Hall Problem

- YouTube videos on this problem
  - Cartoon
  - “21”
  - Numb3rs
- Wikipedia article on this problem



September 10, 1990

Mr. Lawrence A. Denenberg
Harvard University Center for
Research in Computing Technology
Aiken Computation Laboratory, Room 102
Harvard University
Cambridge, MA 02138

Dear Larry:

In sending you my okay for the use of "The Monty Hall Paradox," I should like to ask you a question. You mention that in part (a), the player should switch doors even without additional compensation -- indeed the player should be willing to pay Monty up to $21,845 for the privilege of switching.

Now, I am not well versed in algorithms; but as I see it, it wouldn't make any difference after the player has selected Door A, and having been shown Door C - why should he then attempt to switch to Door B? The major prize could only be in one of the three doors. He has made his selection of

Example: Discriminatory Drugs

- What is the probability that Drug I will treat a man successfully?
- Is the success of Drug I independent from gender?

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Example: Discriminatory Drugs

- Which drug is more effective for women?
- Which drug is more effective for men?
- Which drug is more effective overall?
  - Simpson's Paradox

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

"There are three kinds of lies: lies, damned lies, and statistics."

Example: Biased Coins

- Consider you have three coins
  - C_1, C_2, C_3
- Alex picked up one of the coins and flipped it 6 times.
- You didn't see which coin he picked out, but you observed the results of flipping coins
  - `THTHTT`
- How to guess which coin Alex chose?

Example: Biased Coins

- You experimented with the three coins, say 6 times
  - C_1: `HHHTHH`
  - C_2: `TTHTHH`
  - C_3: `THTTTH`
- Given the observation
  - O: `THTHTT`
- Which coin do you think Alex chose?

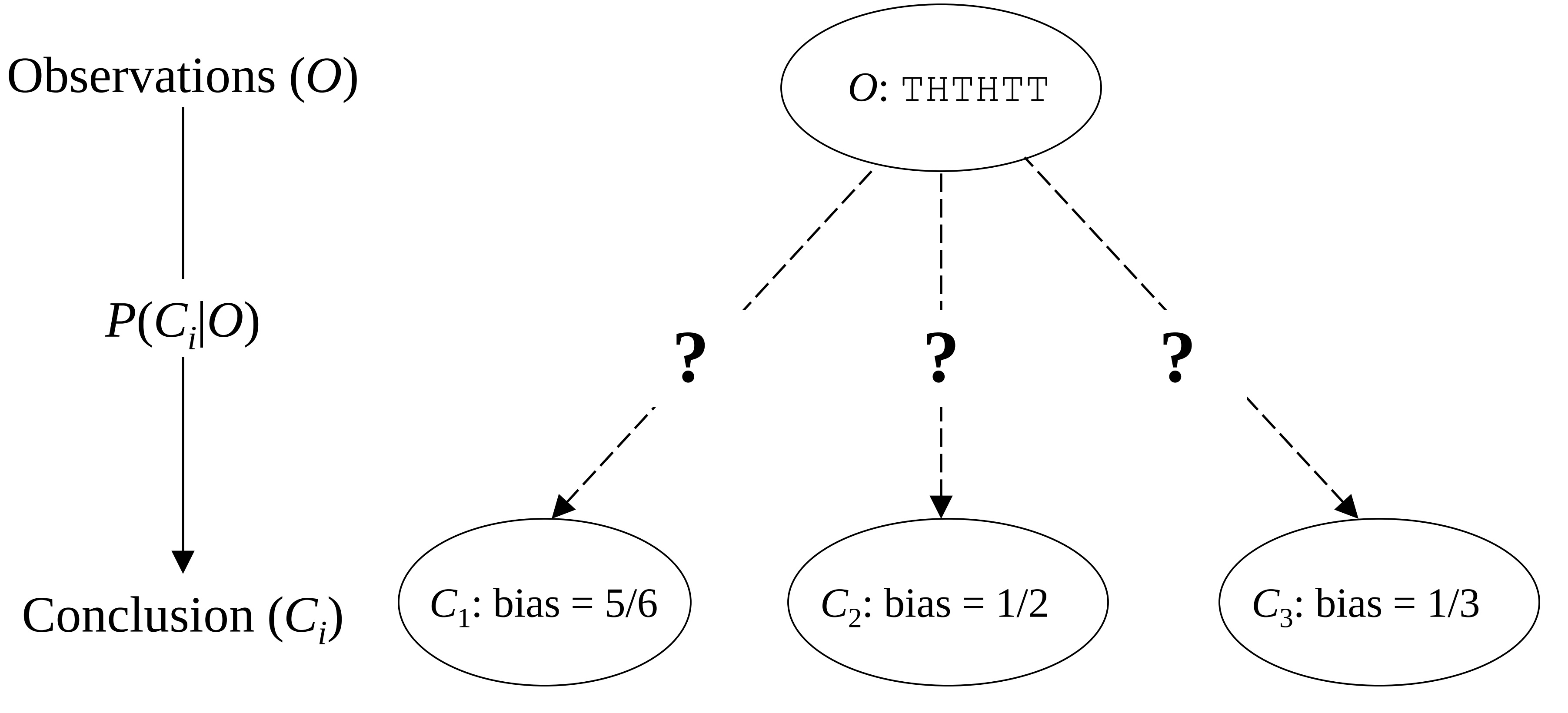
Example: Biased Coins

- A principled approach
  - Compare the posterior probability $P(C_i|O)$
  - It is not obvious how the posterior probability $P(C_i|O)$ can be computed directly
  - It is easy to compute the prior probability $P(C_i)$ and the likelihood $P(O|C_i)$.
- Build a model for each coin
  - C_1: `HHHTHH` → bias $P(\mathtt{H}|C_1) = 5/6$
  - C_2: `TTHTHH` → bias $P(\mathtt{H}|C_2) = 1/2$
  - C_3: `THTTTH` → bias $P(\mathtt{H}|C_3) = 1/3$

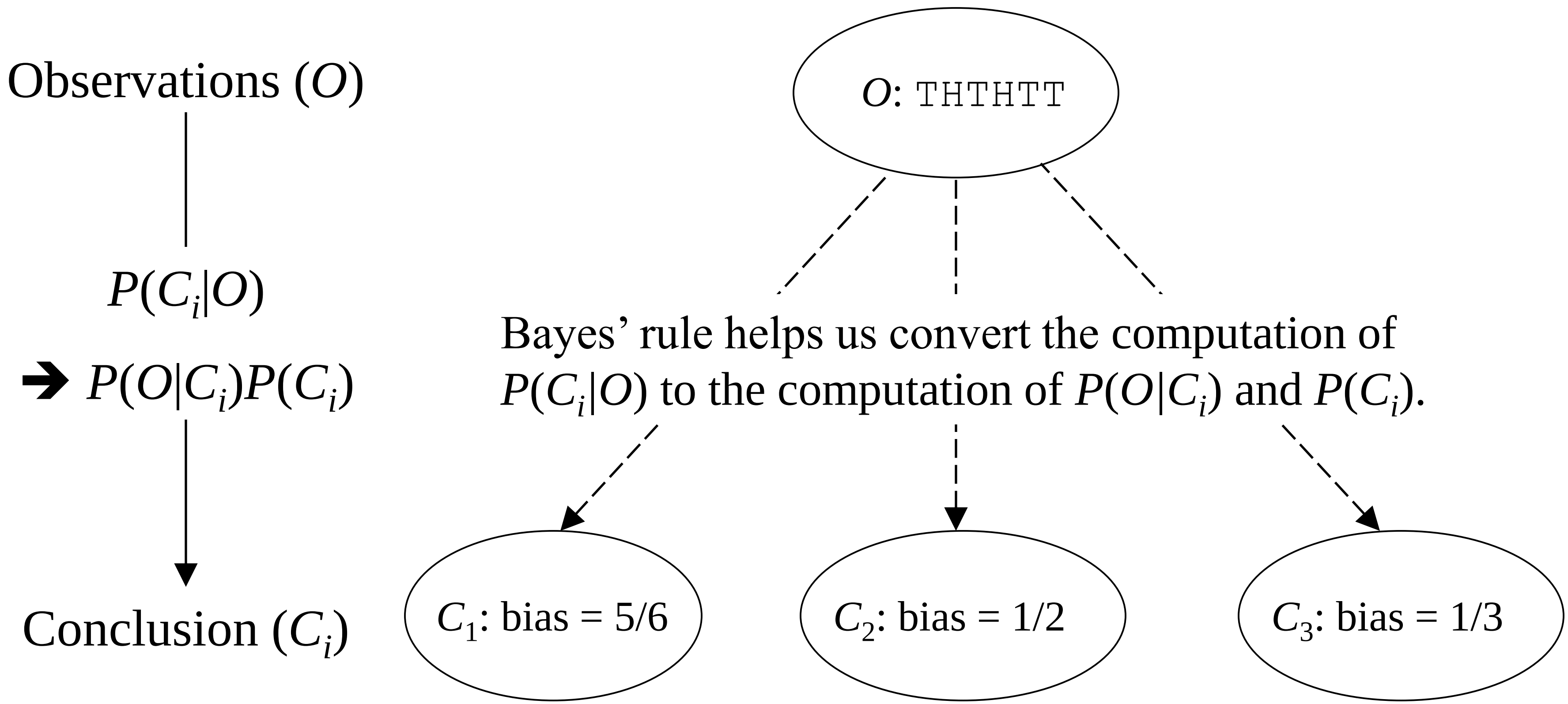
Example: Biased Coins

- Prior probability
  - $P(C_1) = P(C_2) = P(C_3) = 1/3$
- Likelihood
  - $P(O|C_1) = P(\mathtt{THTHTT}|C_1)$
    $= P(\mathtt{T}|C_1)P(\mathtt{H}|C_1)P(\mathtt{T}|C_1)P(\mathtt{H}|C_1)P(\mathtt{T}|C_1)P(\mathtt{T}|C_1)$
    $= P(\mathtt{H}|C_1)^2 P(\mathtt{T}|C_1)^4 = (5/6)^2 * (1/6)^4 \approx 0.0005$
  - $P(O|C_2) \approx 0.0156$
  - $P(O|C_3) \approx 0.0219$
- Posterior probability
  - Which coin has the largest posterior probability $P(C_i|O)$?

Example: Biased Coins



Example: Biased Coins



Probability Ranking Principle

- The document ranking method is the core of an IR system
  - We have a collection of documents. The user issues a query. A list of documents needs to be returned.
  - In what order do we present documents to the user? We want the “best” document to be first, second best second, etc….

Probability Ranking Principle

"If a reference retrieval system's response to each request is a ranking of the documents in the collection **in order of decreasing <u>probability of relevance</u>** to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data."

van Rijsbergen (1979:113-114)

$$P(R = 1|d, q)$$

Probability Ranking Principle

- **Theorem**. The PRP is optimal, in the sense that it minimizes the expected loss (also known as the Bayes risk) under 1/0 loss.
  - Provable if all probabilities are known correctly.

Appraisal

- Probabilistic methods are one of the ***oldest*** but also one of the currently ***hottest*** topics in IR.
  - Traditionally: neat ideas, but they've never won on performance.
  - It may be different now. For example, the Okapi BM25 term weighting formulas have been very successful, especially in TREC evaluations.

Okapi BM25

Retrieval Status Value

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d / L_{\mathrm{ave}})) + \mathrm{tf}_{td}}$$

IDF(t)

The document length of d

The average document length for the collection

The parameters k_1, b should ideally be tuned on a validation set.
The good values in practice are $1.2 \leq k_1 \leq 2$; $b = 0.75$.

Well-Known UK Researchers

Karen Sparck Jones

Stephen Robertson

Keith van Rijsbergen