Probability Theory for Data Science

Ronojoy Adhikari
The Institute of Mathematical Sciences

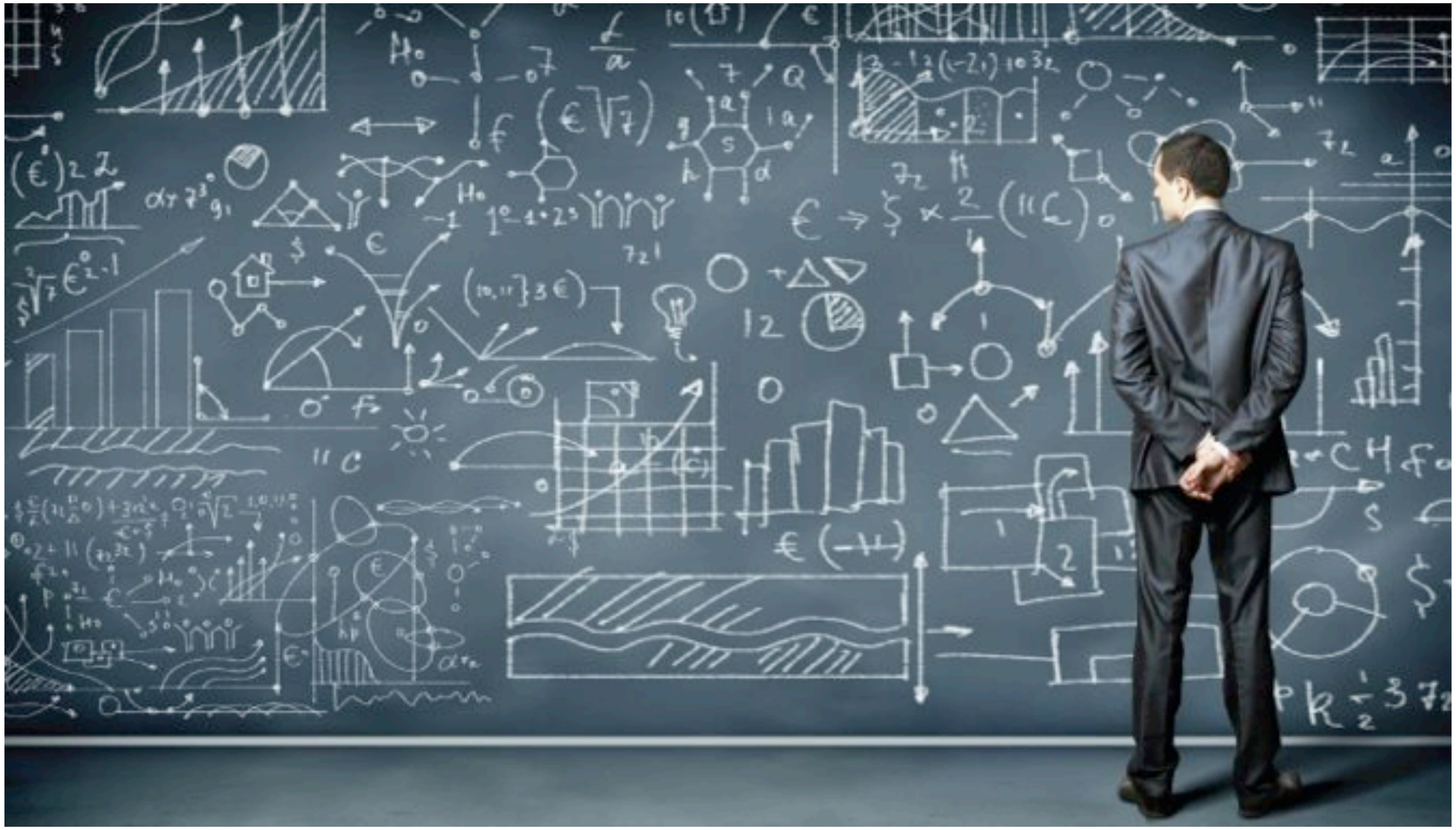
Resources

- pylearn : machine learning resources in Python (github.com/ronojoy/pylearn)
- slides on speakerdeck : (speakerdeck.com/ronojoy/data-science-theory)

Introduction and motivation

- Reasoning as the basis for a science of data
- Reasoning under certainty and under uncertainty
- Boolean logic and probability theory
- Rules of probability theory
- Assigning probabilities - indifference and maximum entropy
- Inference and learning
- Is this a fair coin ? Elementary example of reasoning under uncertainty

Lots of data - where is the science ?



Science : observation - hypothesis - experiment - theory

What are we observing ?
What is our hypothesis ?
Can we experiment ?
Will there be a theory ?

The scientific method

The scientific method

The scientific method

“Now this is the peculiarity of scientific method, that when once it has become a habit of mind, that mind converts all facts whatsoever into science. The field of science is unlimited; its solid contents are endless, every group of natural phenomena, every phase of social life, every stage of past or present development is material for science. The unity of all science consists alone in its method, not in its material.

The man who classifies facts of any kind whatever, who sees their mutual relation and describes their sequence, is applying the scientific method and is a man of science. The facts may belong to the past history of mankind, to the social statistics of our great cities, to the atmosphere of the most distant stars, to the digestive organs of a worm, or to the life of a scarcely visible bacillus. It is not the facts themselves which form science, but the method in which they are dealt with.”

The scientific method

The

Grammar of Science

BY

KARL PEARSON, M.A., F.R.S.

PART I.—PHYSICAL

THIRD EDITION, REVISED AND ENLARGED

LONDON

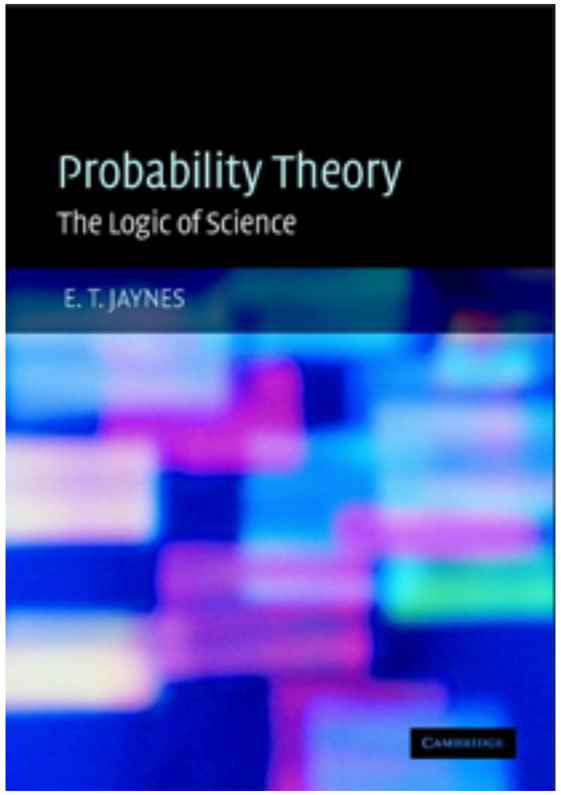
ADAM AND CHARLES BLACK

1911

"Now this is the peculiarity of scientific method, that when once it has become a habit of mind, that mind converts all facts whatsoever into science. The field of science is unlimited; its solid contents are endless, every group of natural phenomena, every phase of social life, every stage of past or present development is material for science. The unity of all science consists alone in its method, not in its material.

The man who classifies facts of any kind whatever, who sees their mutual relation and describes their sequence, is applying the scientific method and is a man of science. The facts may belong to the past history of mankind, to the social statistics of our great cities, to the atmosphere of the most distant stars, to the digestive organs of a worm, or to the life of a scarcely visible bacillus. It is not the facts themselves which form science, but the method in which they are dealt with."

The scientific method

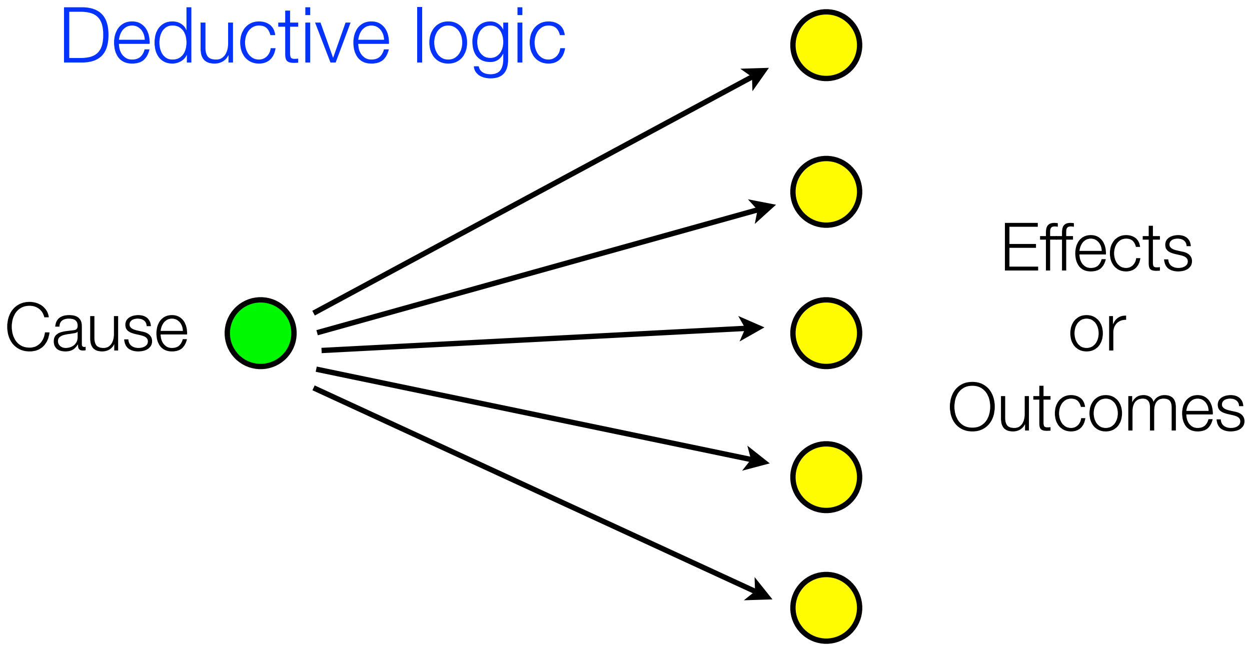


"Now this is the peculiarity of scientific method, that when once it has become a habit of mind, that mind converts all facts whatsoever into science. The field of science is unlimited; its solid contents are endless, every group of natural phenomena, every phase of social life, every stage of past or present development is material for science. The unity of all science consists alone in its method, not in its material.

The man who classifies facts of any kind whatever, who sees their mutual relation and describes their sequence, is applying the scientific method and is a man of science. The facts may belong to the past history of mankind, to the social statistics of our great cities, to the atmosphere of the most distant stars, to the digestive organs of a worm, or to the life of a scarcely visible bacillus. It is not the facts themselves which form science, but the method in which they are dealt with."

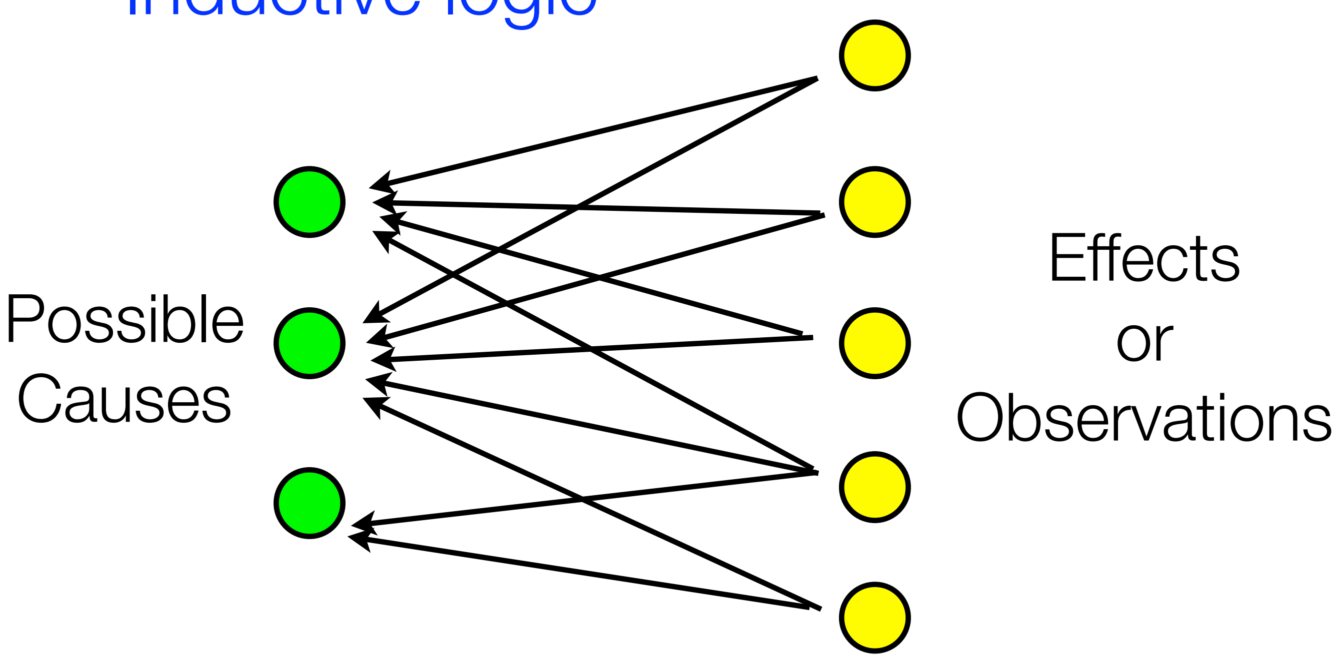
Logical reasoning

Deductive logic



Boolean algebra

Inductive logic



Bayesian probability

Boolean algebra

- Formalization of Aristotelian logic
- Propositions : are either TRUE or FALSE
- Operations : conjunction (AND), disjunction (OR), negation (NOT)
- Laws : algebraic identities between compound propositions
- Ex.1 : NOT(A AND B) = (NOT A) OR (NOT B)
- Ex. 2 : NOT(A OR B) = (NOT A) AND (NOT B)
- Rules for reasoning consistently with certain propositions.

Probability Theory

- Generalization of Boolean logic
- Propositions have a truth value p, with p = 0 (FALSE) and p = 1 (TRUE)
- Operations : conjunction (AND), disjunction (OR), negation (NOT)
- sum rule : P(A) + P(NOT A) = 1
- product rule : P(A AND B) = P(A|B)P(B) = P(B|A) P(A)
- => P(A OR B) = P(A) + P(B) - P(A AND B)
- independent => P(A|B) = P(A) ; mutually exclusive => P(A OR B) = P(A) + P(B)

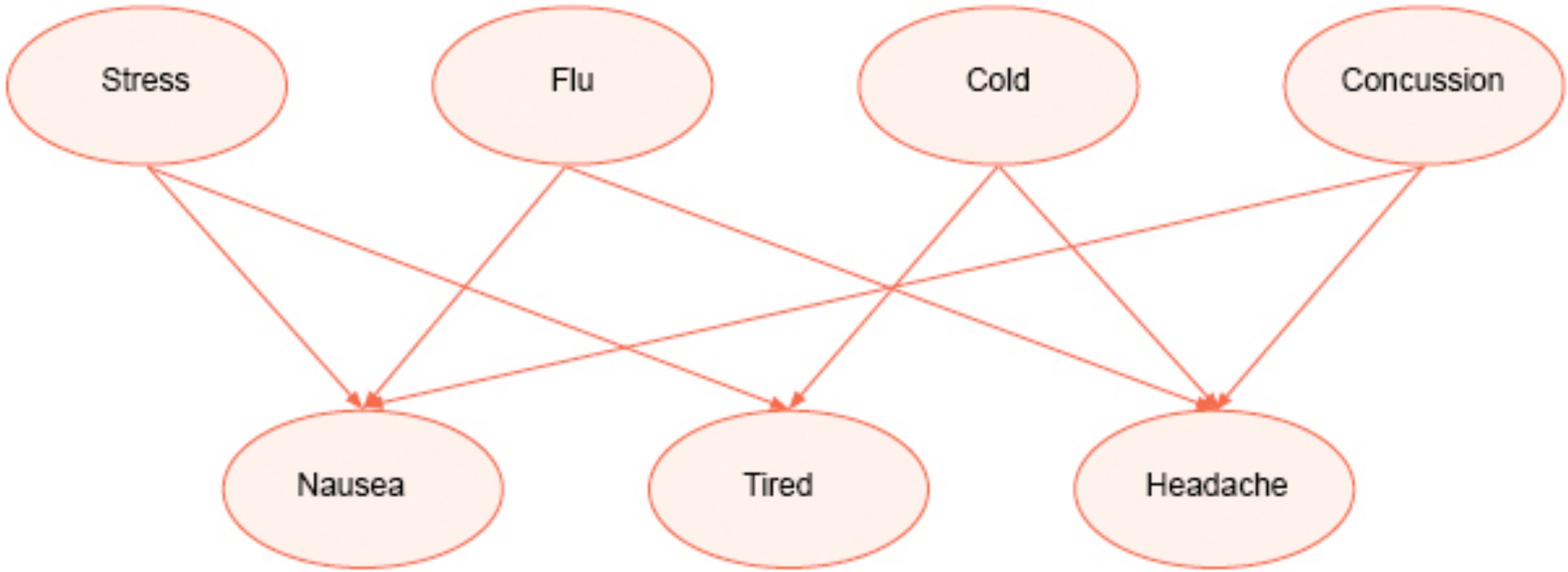
Assigning probabilities

- Probabilities are ALWAYS conditioned on information $P(A) = P(A \mid I)$
- Consider a set of propositions A1, A2, ... An, that are exhaustive and mutually exclusive. In the absence of any other information, the principle of indifference says that P(Ai) = 1/N (Laplace)
- When additional information is available, probabilities are assigned taking the additional information into account. The principle of maximum entropy says that P should be assigned by maximizing $\sum P_i \log P_i$, subject to the constraints that derive from the additional information.
- Maximum entropy reduces to indifference when there are no constraints.

Bayes theorem

- $P(A \text{ and } B) = P(A|B)\,P(B) = P(B|A)P(A)$
- $P(A|B) = P(B|A)\,P(A) / P(B)$
- Looks trivial but is extremely deep!
- P(disease | symptom) = want to know this.
- P(symptom | disease) = can estimate this (even empirically!)
- P(disease | symptom) = P(symptom | disease) P (disease) / P(symptom)

Bayesian networks



P(disease | symptoms) = diagnosis

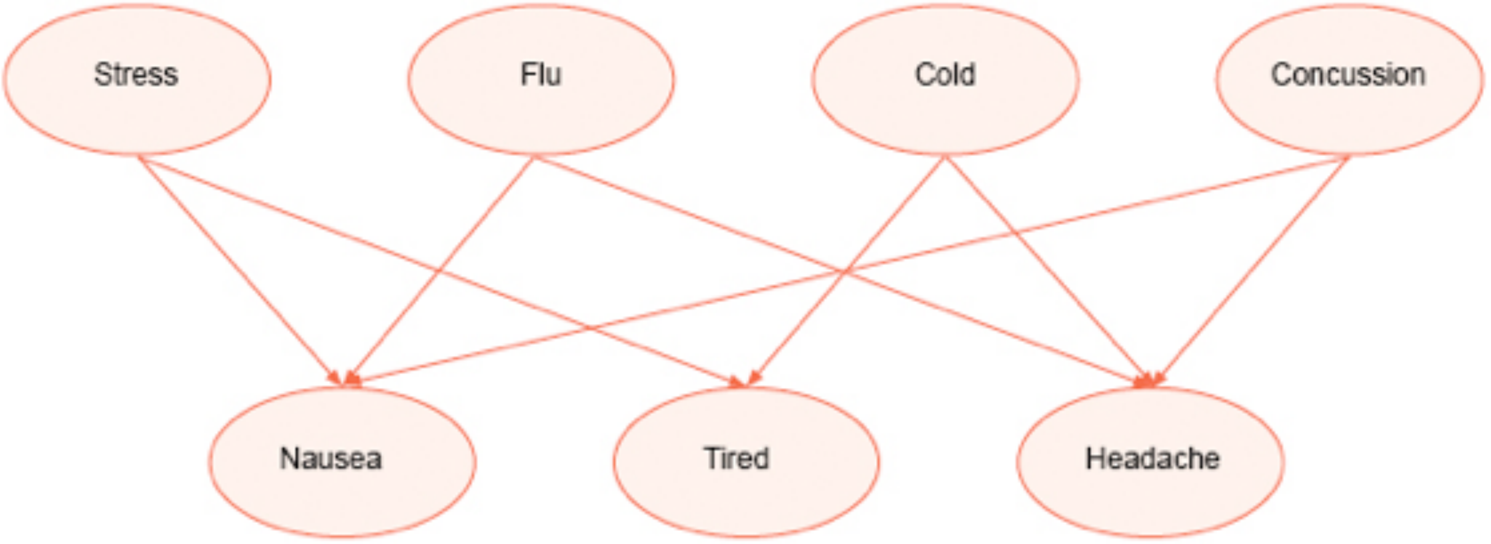
RILACS

Representation

Inference

Learning

Actions



Is this a fair coin ?

$$P(n_1|\theta, N) = \frac{N!}{n_1!(N-n_1)!}\theta^{n_1}(1-\theta)^{N-n_1}$$

P(D|H) - likelihood

$$P(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

P(H) = prior $\langle\theta\rangle = \frac{a}{a+b}$

$$P(\theta|n_1, N) \sim \theta^{n_1+a-1}(1-\theta)^{n-n_1+b-1}$$

P(H|D) = posterior

https://github.com/ronojoy/pylearn/blob/master/scripts/ex1-coin-tossing.py stuff we will use in the example

python™

NumPy

IP[y]: IPython

SciPy

matplotlib

SymPy

Mayavi

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Cython

SM StatsModels
Statistics in Python

scikits learn
machine learning in Python

scikits-image
image processing in python

PyTables

NetworkX