From intelligent data analysis to medical decision support

Péter Antal
antal@mit.bme.hu
http://bioinfo.mit.bme.hu/

Computational Biomedicine (ComBine) Workgroup
Intelligent Systems Group
Faculty of Electrical Engineering and Informatics (VIK)
Department of Measurement and Information Systems
Budapest University of Technology and Economics (BME)
Outline

• Factors behind the A.I./machine learning „hype”
• Probability theory
• Bayesian networks, decision networks
• Value of information, optimal decisions
• Exercise: construction of a decision support system
• ➔ next lecture: learning
Computational Biomedicine

COMBINE lab

Team

Bence Bolgár
András Gézsi
Gábor Hullám
András Millinghoffer
Péter Sárközy
Péter Antal

http://bioinfo.mit.bme.hu/
ComBineLab.hu: Themes

- Knowledge engineering
- Study design
- Genetic measurements
- Data engineering
- Data analysis
- Interpretation
- Decision support
ComBineLab.hu: tools

- **BayesEye: Bayesian, systems-based data analysis**
  - Bayesian model averaging over Bayesian network structures.

- **BayesCube: Probabilistic decision support**
  - Semantically enriched Bayesian and decision network models.

- **BysCyc/QSF (Bayesian Encyclopedia):**
  - Large-scale probabilistic inference

- **QDF: Kernel-based fusion methods for repositioning**
  - Multi-aspect rankings and multi-aspect metrics in drug discovery

- **Variant Meta Caller: precision NGS**
  - Next-generation sequencing pipelines

- **VB-MK-LMF: drug-target interaction prediction**
  - Variational Bayesian Multiple Kernel Logistic Matrix Factorization

- ... see Tools @ [http://bioinfo.mit.bme.hu/](http://bioinfo.mit.bme.hu/)
Hallmarks of a new AI era?

Mastering the game of Go with deep neural networks and tree search
David Silver\textsuperscript{1*}, Aja Huang\textsuperscript{1*}, Chris J. Maddison\textsuperscript{1}, Arthur Guez\textsuperscript{1}, Laurent Sifre\textsuperscript{1}, George van den Driessche\textsuperscript{1}, Julian Schrittwieser\textsuperscript{1}, Ioannis Antonoglou\textsuperscript{1}, Veda Panneershelvam\textsuperscript{1}, Marc Lanctot\textsuperscript{1}, Sander Dieleman\textsuperscript{1}, Dominik Grewe\textsuperscript{1}, John Nham\textsuperscript{2}, Nal Kalchbrenner\textsuperscript{1}, Ilya Sutskever\textsuperscript{2}, Timothy Lillicrap\textsuperscript{1}, Madeleine Leach\textsuperscript{1}, Koray Kavukcuoglu\textsuperscript{1}, Thore Graepel\textsuperscript{1} & Demis Hassabis\textsuperscript{1}

Human-level control through deep reinforcement learning
Volodymyr Mnih\textsuperscript{1*}, Koray Kavukcuoglu\textsuperscript{1*}, David Silver\textsuperscript{1*}, Andrei A. Rusu\textsuperscript{1}, Joel Veness\textsuperscript{1}, Marc G. Bellemare\textsuperscript{1}, Alex Graves\textsuperscript{1}, Martin Riedmiller\textsuperscript{1}, Andreas K. Fidjeland\textsuperscript{1}, Georg Ostrovski\textsuperscript{1}, Stig Petersen\textsuperscript{1}, Charles Beattie\textsuperscript{1}, Amir Sadik\textsuperscript{1}, Ioannis Antonoglou\textsuperscript{1}, Helen King\textsuperscript{1}, Dharshan Kumaran\textsuperscript{1}, Daan Wierstra\textsuperscript{1}, Shane Legg\textsuperscript{1} & Demis Hassabis\textsuperscript{1}
Medical decision support systems

Watson for Oncology – assessment and advice cycle
www.avanteoconsulting.com/machine-learning-accelerates-cancer-research-discovery-innovation/
Automated discovery systems


- (Gene prioritization...)
„Machine science”


D. R. Swanson et al.: *An interactive system for finding complementary literatures: a stimulus to scientific discovery*, Artificial Intelligence, 1997


“Soon, computers could generate many useful hypotheses with little help from humans.”
Factors behind the „A.I./learning hype”

• New theory?
  – Unified theory of AI?
  – A new machine learning approach?

• New hardware? (computing power..)
  – GPUs?
  – Quantum computers?

• New resources?
  – Data?
  – Knowledge?
  – Money?
  – Brains/Minds?
Computing power: Moore’s Law

1965, Gordon Moore, founder of Intel: „The number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years "... "for at least ten years"

Integration and parallelization wont bring us further. End of Moore’s law?
Definitions of „big data”


The 3xV: *volume, variety, and velocity* (2001).

The 8xV: Vast, Volumes of Vigorously, Verified, Vexingly Variable Verbose yet Valuable Visualized high Velocity Data (2013)

**Not „conventional” data:** „Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it (E.Dumbill: Making sense of big data, Big Data, vol.1, no.1, 2013)
Carlson’s Law for Biological Data

NATURE, Vol 464, April 2010

- Sequencing costs per mill. base
- Publicly available genetic data
  - x10 every 2-3 years
  - Data volumes and complexity that IT has never faced before…
Data: „Big” data in life sciences

Healthcare Industry is dealing with data overload

**Exogenous data**
(Behavior, Socio-economic, Environmental, …)
60% of determinants of health
*Volume, Variety, Velocity, Veracity*

1100 Terabytes
Generated per lifetime

**Genomics data**
30% of determinants of health
*Volume*

6 TB
Per lifetime

**Clinical data**
10% of determinants of health
*Variety*

0.4 TB
Per lifetime

Source: “The Relative Contribution of Multiple Determinants to Health Outcomes”, Lauren McGover et al., Health Affairs, 33, no.2 (2014)

*IBM Watson Health*

Wearable Unit

**Fabric Electrodes & Sensors**

ECG (Einthoven’s triangle) AVR
Thoracic respiration
EMG troads
Preordial leads
AVL
Abdominal respiration
AVF

Elbow articulation
Shoulder articulation
Biomedical omic data/big data

2010<: “Clinical phenotypic assay”/drugome: open clinical trials, adverse drug reaction DBs, adaptive licensing, Large/scale cohort studies (~100,000 samples)

Environment&life style

Phenome (disease, side effect)

Metabolome

Proteome

Transcriptome

Genome(s), epigenome, microbiome

Drugs

Moore’s law

Carlson’s law
UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Scientists, please ensure you read the background materials before registering. To our participants, we say thank you for supporting this important resource to improve health. Without you, none of the research featured on this website would be possible.


Large-scale cohorts in UK

UK Biobank:
- 1 million< adults
- aged 40-69,
- 2006-2036<
- genes x lifestyle x environment ➔ diseases
- open 2012-
Further national biobanks: FinnGen

- [https://www.finngen.fi/en](https://www.finngen.fi/en)
- 500k participants
- 2017-
- **Personalized medicine project**
- genome information (WGS) + digital health care data
- **The study is funded[!!!] by Business Finland and seven international pharmaceutical companies:** Abbvie, AstraZeneca, Biogen, Celgene, Genentech (a member of the Roche Group), Merck & Co., Inc., Kenilworth, NJ, USA and Pfizer.
Further health data

• FlatIron Health (acquired by Roche):
  – 7 major academic research centers
  – 280+ community oncology practices
  – top 15 therapeutic oncology companies
  – 2500 clinicians
  – 2.1 million active patient records
    • complete, electronic health records
    • +patient-reported data
Big health data streams

<table>
<thead>
<tr>
<th>New “Omics” Data Streams</th>
<th>Traditional Data Streams</th>
<th>Quantified Self Data Streams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome</td>
<td>Personal and Family Health History</td>
<td>Self-reported data: health, exercise, food, mood journals, etc.</td>
</tr>
<tr>
<td>- SNP mutations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Structural variation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Epigenetics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microbiome</td>
<td>Prescription History</td>
<td>Mobile Application Data</td>
</tr>
<tr>
<td>Transcriptome</td>
<td>Lab Tests: History and Current</td>
<td></td>
</tr>
<tr>
<td>Metabolome</td>
<td>Demographic Data</td>
<td>Quantified Self Device Data</td>
</tr>
<tr>
<td>Proteome</td>
<td>Standardized Instrument Response</td>
<td>Biosensor Data Objective Metrics</td>
</tr>
<tr>
<td>Diseasome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmentome</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend: Consumer-available

M.Swan: THE QUANTIFIED SELF: Fundamental Disruption in Big Data Science and Biological Discovery, Big data, Vol 1., No. 2., 2013
On the thresholds of data: health

- Local datasets: 1k ➔ 10k participants
- International datasets: 10k ➔ 100k
- National biobanks: <1 million
- International biobanks: x1 million
- Regular health records: 100 million (➔ Meta-analysis using summary statistics)

➔ Federated learning: separation of data and model
  1. Data is standardized (using ontologies)
  2. Stays at the institutes/individuals
  3. Model updates are communicated
  4. Using privacy-preserving techniques

Disease specific
Cross-sectional
Longitudinal
Patient-reported
Self-quantified

2010<
2010<
Number of biomedical publications

**Little Science, Big Science**, by Derek J. de Solla Price, 1963

*Fig. 2. Cumulative number of abstracts in various scientific fields, from the beginning of the abstract service to given date*

It will be noted that after an initial period of rapid expansion to a stable growth rate, the number of abstracts increases exponentially, doubling in approximately 15 years.
Semantic publishing: papers vs DBs/KBs

M. Gerstein, "E-publishing on the Web: Promises, pitfalls, and payoffs for bioinformatics," Bioinformatics, 1999
M. Gerstein: Blurring the boundaries between scientific 'papers' and biological databases, Nature, 2001
P. Bourne, "Will a biological database be different from a biological journal?," Plos Computational Biology, 2005
Biomedical databases by 2000
Knowledge: Linked open data

E-science, data-intensive science, the fourth paradigm

All Scientific Data Online

- Many disciplines overlap and use data from other sciences
- Internet can unify all literature and data
- Go from literature to computation to data back to literature
- Information at your fingertips for everyone-everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity
Factors behind the „hype” II.

• New theory:
  – Unified theory of AI: Probabilistic models
  – RE-new-ed machine learning approaches:
    • „Sequential” learning, „Deep” learning

• New hardware
  – GPUs: yes
  – Quantum computers: no

• New resources
  – Data & Knowledge: Linked Open Data
  – Society: open for „smart” solutions
Milestones and phases in AI

• ~1930: Zuse, Neumann, Turing..: „instruction is data“:
  – Laws of nature can be represented, „executed“/simulated with modifications, learnt
  – Knowledge analogously: representation, execution, adaptation and learning

• 1943 McCulloch & Pitts: Boolean circuit model of brain
• 1950 Turing's "Computing Machinery and Intelligence"
• 1956 Dartmouth meeting: the term "Artificial Intelligence”
• 1950s Early AI programs (e.g. Newell & Simon's Logic Theorist)
  • The physical symbol system hypothesis: search
• 1965 Robinson's complete algorithm for logical reasoning
• 1966—73 AI discovers computational complexity
  Neural network research almost disappears
• 1969—79 Early development of knowledge-based systems
  • The knowledge system hypothesis: knowledge is power
• 1986-- Neural networks return to popularity
• 1988-- Probabilistic expert systems
• 1995-- Emergence of machine learning
  • The „big data” hypothesis: let data speak
• 2005/2015-- Emergence of autonomous adaptive decision systems („robots”, agents)
  • The autonomy hypothesis??
Optimal decision: decision theory
probability theory+utility theory

- Decision situation:
  - Actions
  - Outcomes
  - Probabilities of outcomes
  - Utilities/losses of outcomes
  - Maximum Expected Utility Principle (MEU)
  - Best action is the one with maximum expected utility

\[
\begin{align*}
&\text{Action: } a_i \\
&\text{Outcomes: } o_j \\
&\text{Probability: } p(o_j \mid a_i) \\
&\text{Utility: } U(o_j \mid a_i) \\
&\text{Expected Utility: } EU(a_i) = \sum_j U(o_j \mid a_i) p(o_j \mid a_i) \\
&\text{Best action: } a^* = \arg \max_i EU(a_i)
\end{align*}
\]
Types of inference

- (Passive, observational) inference
  - $P(\text{Query} \mid \text{Observations, Observational data})$

- Interventionist inference
  - $P(\text{Query} \mid \text{Observations, Interventions})$

- Counterfactual inference
  - $P(\text{Query} \mid \text{Observations, Counterfactual conditionals})$

- Biomedical applications
  - Prevention
  - Screening
  - Diagnosis
  - Therapy selection
  - Therapy modification
  - Evaluation of therapeutic efficiency
Probabilistic graphical models: Bayesian Networks

- A directed acyclic graph (DAG)
- Nodes are random variables
- Edges represent direct dependence (causal relationship)
- Local models: \( P(X_i | \text{Pa}(X_i)) \)
- Offers three interpretations

\[
P(\text{Model} | \text{Data}) \propto P(\text{Data} | \text{Model})P(\text{Model})
\]

Thomas Bayes (c. 1702 – 1761)
Ovarian tumor diagnostics

International Ovarian Tumor Analysis (IOTA, Dirk Timmerman)

Interpretations of probability

- Sources of uncertainty
  - inherent uncertainty in the physical process;
  - inherent uncertainty at macroscopic level;
  - ignorance;
  - practical omissions;

- Interpretations of probabilities:
  - combinatoric;
  - physical propensities;
  - frequentist;
  - personal/subjectivist;
  - instrumentalist;

\[
\lim_{N \to \infty} \frac{N_A}{N} = \lim_{N \to \infty} \hat{p}_N(A) = p(A) \approx p(A \mid \xi)
\]
Uncertainty

• A. Einstein: „God does not play dice.”

• Einstein-Podolski-Rosen paradox / Bell Test

• S. Hawking: „Does god play dice?”
  http://www.hawking.org.uk/does-god-play-dice.html

• The BIG Bell Test (Nov30, 2016)
A chronology of uncertain inference

- [1713] Ars Conjectandi (The Art of Conjecture), Jacob Bernoulli
  - Subjectivist interpretation of probabilities
- [1718] The Doctrine of Chances, Abraham de Moivre
  - the first textbook on probability theory
  - Forward predictions
    - „given a specified number of white and black balls in an urn, what is the probability of drawing a black ball?”
    - his own death
- [1764, posthumous] Essay Towards Solving a Problem in the Doctrine of Chances, Thomas Bayes
  - Backward questions: „given that one or more balls has been drawn, what can be said about the number of white and black balls in the urn”
- [1812], Théorie analytique des probabilités, Pierre-Simon Laplace
  - General Bayes rule
- ...
- [1921]: Correlation and causation, S. Wright’s diagrams
- [1933]: A. Kolmogorov: Foundations of the Theory of Probability
Basic concepts of probability theory

- Joint distribution
- Conditional probability
- Bayes’ rule
- Chain rule
- Marginalization
- General inference
- Independence
  - Conditional independence
  - Contextual independence
Syntax

• **Atomic event**: A complete specification of the state of the world about which the agent is uncertain

  E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

  \[
  \begin{align*}
  Cavity = \text{false} \land Toothache = \text{false} \\
  Cavity = \text{false} \land Toothache = \text{true} \\
  Cavity = \text{true} \land Toothache = \text{false} \\
  Cavity = \text{true} \land Toothache = \text{true}
  \end{align*}
  \]

• Atomic events are mutually exclusive and exhaustive
Axioms of probability

• For any propositions $A$, $B$

  – $0 \leq P(A) \leq 1$
  – $P(true) = 1$ and $P(false) = 0$
  – $P(A \lor B)^{True}$
Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.

- **Boolean** random variables
  - e.g., *Cavity* (do I have a cavity?)

- **Discrete** random variables
  - e.g., *Weather* is one of <sunny, rainy, cloudy, snow>
  - Domain values must be exhaustive and mutually exclusive

- Elementary proposition constructed by assignment of a value to a random variable: e.g., *Weather* = sunny, *Cavity* = false
  - (abbreviated as \(\neg\text{cavity}\))

- Complex propositions formed from elementary propositions and standard logical connectives e.g., *Weather* = sunny \(\lor\) *Cavity* = false
Joint (probability) distribution

- **Prior or unconditional probabilities** of propositions
  - e.g., \( P(Cavity = \text{true}) = 0.1 \) and \( P(Weather = \text{sunny}) = 0.72 \) correspond to belief prior to arrival of any (new) evidence

- **Probability distribution** gives values for all possible assignments:
  - \( P(Weather) = <0.72, 0.1, 0.08, 0.1> \) (normalized, i.e., sums to 1)

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables
  - \( P(Weather, Cavity) = a 4 \times 2 \) matrix of values:

<table>
<thead>
<tr>
<th></th>
<th>sunny</th>
<th>rainy</th>
<th>cloudy</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Weather</strong> = sunny</td>
<td>0.144</td>
<td>0.02</td>
<td>0.016</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Cavity</strong> = true</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cavity</strong> = false</td>
<td>0.576</td>
<td>0.08</td>
<td>0.064</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Conditional probability

- **Conditional or posterior probabilities**
  - e.g., $P(\text{cavity} \mid \text{toothache}) = 0.8$
  - i.e., given that \text{toothache} is all I know

- (Notation for conditional distributions:
  - $P(\text{Cavity} \mid \text{Toothache}) = 2$-element vector of 2-element vectors)

- If we know more, e.g., \text{cavity} is also given, then we have
  - $P(\text{cavity} \mid \text{toothache}, \text{cavity}) = 1$

- New evidence may be irrelevant, allowing simplification, e.g.,
  - $P(\text{cavity} \mid \text{toothache}, \text{sunny}) = P(\text{cavity} \mid \text{toothache}) = 0.8$

- This kind of inference, sanctioned by domain knowledge, is crucial
Conditional probability

- Definition of conditional probability:
  \[ P(a | b) = \frac{P(a \land b)}{P(b)} \text{ if } P(b) > 0 \]

- **Product rule** gives an alternative formulation:
  \[ P(a \land b) = P(a | b) P(b) = P(b | a) P(a) \]

- A general version holds for whole distributions, e.g.,
  \[ P(Weather,Cavity) = P(Weather | Cavity) P(Cavity) \]
  (View as a set of 4 × 2 equations, not matrix mult.)
Bayes’ rule

An algebraic triviality

\[ p(X | Y) = \frac{p(Y | X) p(X)}{p(Y)} = \frac{p(Y | X) p(X)}{\sum_X p(Y | X) p(X)} \]

A scientific research paradigm

\[ p(Model | Data) \propto p(Data | Model) p(Model) \]

A practical method for inverting causal knowledge to diagnostic tool.

\[ p(Cause | Effect) \propto p(Effect | Cause) \times p(Cause) \]
Chain rule

- **Chain rule** is derived by successive application of product rule:

\[
P(X_1, ..., X_n) = P(X_1, ..., X_{n-1}) \cdot P(X_n \mid X_1, ..., X_{n-1})
\]

\[
= P(X_1, ..., X_{n-2}) \cdot P(X_{n-1} \mid X_1, ..., X_{n-2}) \cdot P(X_n \mid X_1, ..., X_{n-1})
\]

\[
= \ldots
\]

\[
= \pi \cdot P(X_i \mid X_1, ..., X_{i-1})
\]
Marginalization

• ~Summing out/averaging out

• Start with the joint probability distribution:

  - For any proposition $\phi$, sum the atomic events where it is true: $P(\phi) = \sum_{\omega: \omega \models \phi} P(\omega)$
Inference by enumeration

• Start with the joint probability distribution:

<table>
<thead>
<tr>
<th>cavity</th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>¬ catch</td>
<td>.016</td>
<td>.064</td>
</tr>
<tr>
<td>catch</td>
<td>.072</td>
<td>.008</td>
</tr>
<tr>
<td>¬ catch</td>
<td>.144</td>
<td>.576</td>
</tr>
</tbody>
</table>

• Can also compute conditional probabilities:

\[
P(\neg cavity \mid toothache) = \frac{P(\neg cavity \land toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4\]
Normalization

- Denominator can be viewed as a normalization constant $\alpha$

$$P(Cavity \mid toothache) = \alpha, \ P(Cavity, toothache) = \alpha, \ [P(Cavity, toothache, catch) + P(Cavity, toothache, \neg catch)] = \alpha, \ [<0.108,0.016> + <0.012,0.064>] = \alpha, \ <0.12,0.08> = <0.6,0.4>$$

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables
Any question about observable events in the domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the query variables \( Y \) given specific values \( e \) for the evidence variables \( E \)

Let the hidden variables be \( H = X - Y - E \)

Then the required summation of joint entries is done by summing out the hidden variables:

\[
P(Y | E = e) = \alpha P(Y, E = e) = \alpha \sum_h P(Y, E = e, H = h)
\]

- The terms in the summation are joint entries because \( Y, E \) and \( H \) together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity \( O(d^n) \) where \( d \) is the largest arity
  2. Space complexity \( O(d^n) \) to store the joint distribution
  3. How to find the numbers for \( O(d^n) \) entries?
Independence, Conditional independence

$I_p(X;Y|Z)$ or $(X \perp Y|Z)_p$ denotes that $X$ is independent of $Y$ given $Z$ defined as follows

for all $x,y$ and $z$ with $P(z)>0$: $P(x;y|z) = P(x|z)P(y|z)$

(Almost) alternatively, $I_p(X;Y|Z)$ iff

$P(X|Z,Y) = P(X|Z)$ for all $z,y$ with $P(z,y)>0$.

Other notations: $D_p(X;Y|Z) = \gamma I_p(X;Y|Z)$

Direct dependence: $D_p(X;Y|V/{X,Y})$
Context-specific independence

Decision tree: Each internal node represents a (univariate) test, the leaf nodes contain the conditional probabilities given the values along the path.

Decision graph: If conditions are equivalent, then subtrees can be merged. E.g. If (Bleeding=absent, Onset=late) ~ (Bleeding=weak, Regularity=irreg)
The independence model of a distribution

The independence map (model) \( M \) of a distribution \( P \) is the set of the valid independence triplets:

\[
M_P = \{ I_{P,1}(X_1;Y_1|Z_1), \ldots, I_{P,K}(X_K;Y_K|Z_K) \}
\]

If \( P(X,Y,Z) \) is a Markov chain, then

\[
M_P = \{ D(X;Y), D(Y;Z), I(X;Z|Y) \}
\]

Normally/almost always: \( D(X;Z) \)
Exceptionally: \( I(X;Z) \)
The semi-graphoid axioms

1. Symmetry: The observational probabilistic conditional independence is symmetric.
   \[
   I_p(X; Y|Z) \iff I_p(Y; X|Z)
   \]

2. Decomposition: Any part of an irrelevant information is irrelevant.
   \[
   I_p(X; Y \cup W|Z) \Rightarrow I_p(X; Y|Z) \text{ and } I_p(X; W|Z)
   \]

3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.
   \[
   I_p(X; Y \cup W|Z) \Rightarrow I_p(X; Y|Z \cup W)
   \]

4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.
   \[
   I_p(X; Y|Z) \text{ and } I_p(X; W|Z \cup Y) \Rightarrow I_p(X; Y \cup W|Z)
   \]

Semi-graphoids (SG): Symmetry, Decomposition, Weak Union, Contraction (holds in all probability distribution). SG is sound, but incomplete inference.
Measures of dependence

- Information theoretic based dependence
  - Entropy: $H(X)$
  - Conditional entropy: $H(X|Y)$
  - Kullback-Leibler divergence ($KL(p||q)$)
    - Not distance (asymmetric, triangle inequality)
    - Always positive
  - Mutual information: $MI(X;Y)$, $MI(X;Y|Z)$
    - $MI(X;Y)=H(X)-H(X|Y)$
    - $MI(X;Y)=KL(p(X,Y)||p(X)p(Y))$
Naive Bayesian network

Assumptions:

1. Two types of nodes: a cause and effects.
2. Effects are conditionally independent of each other given their cause.

Variables (nodes)

- Flu: present/absent
- FeverAbove38C: present/absent
- Coughing: present/absent

Model

\[
\begin{align*}
P(\text{Fever} = \text{present}) &= 0.6 \\
P(\text{Fever} = \text{absent}) &= 1 - 0.6 \\
P(\text{Fever} = \text{present} | \text{Flu} = \text{absent}) &= 0.01 \\
P(\text{Fever} = \text{absent} | \text{Flu} = \text{absent}) &= 1 - 0.01 \\
\end{align*}
\]

\[
\begin{align*}
P(\text{Coughing} = \text{present} | \text{Flu} = \text{present}) &= 0.3 \\
P(\text{Coughing} = \text{absent} | \text{Flu} = \text{present}) &= 1 - 0.7 \\
P(\text{Coughing} = \text{present} | \text{Flu} = \text{absent}) &= 0.02 \\
P(\text{Coughing} = \text{absent} | \text{Flu} = \text{absent}) &= 1 - 0.02 \\
\end{align*}
\]
Naive Bayesian network (NBN)

Decomposition of the joint:
\[
P(Y, X_1, \ldots, X_n) = P(Y) \prod_i P(X_i | Y, X_1, \ldots, X_{i-1}) \quad \text{// by the chain rule}
\]
\[
= P(Y) \prod_i P(X_i | Y) \quad \text{// by the N-BN assumption}
\]

2n+1 parameters!

Diagnostic inference:
\[
P(Y | x_{i1}, \ldots, x_{ik}) = \frac{P(Y) \prod_j P(x_{ij} | Y)}{P(x_{i1}, \ldots, x_{ik})}
\]

If Y is binary, then the odds
\[
P(Y=1 | x_{i1}, \ldots, x_{ik}) / P(Y=0 | x_{i1}, \ldots, x_{ik}) = \frac{P(Y=1)/P(Y=0)}{\prod_j P(x_{ij} | Y=1) / P(x_{ij} | Y=0)}
\]

![Diagram showing Flu, Fever, and Coughing nodes with arrows connecting them.]

\[
p(Flu = \text{present} \mid Fever = \text{absent}, Coughing = \text{present})
\]
\[
\propto p(Flu = \text{present}) p(Fever = \text{absent} \mid Flu = \text{present}) p(Coughing = \text{present} \mid Flu = \text{present})
\]
Bayesian networks: three facets

3. Concise representation of joint distributions

\[ P(M, O, D, S, T) = P(M)P(O|M)P(D|O,M)P(S|D)P(T|S,M) \]

2. Graphical representation of (in)dependencies

\[ M_P = \{ I_{P,1}(X_1; Y_1|Z_1), \ldots \} \]

1. Causal model
Bayesian networks

• A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

• Syntax:
  – a set of nodes, one per variable
  – a directed, acyclic graph (link ≈ "directly influences")
  – a conditional distribution for each node given its parents:  
    \( P(X_i \mid \text{Parents}(X_i)) \)

• In the simplest case, conditional distribution represented as a **conditional probability table** (CPT) giving the distribution over \( X_i \) for each combination of parent values
Example

• I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

• Variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls

• Network topology reflects "causal" knowledge:
  – A burglar can set the alarm off
  – An earthquake can set the alarm off
  – The alarm can cause Mary to call
  – The alarm can cause John to call
Compactness

• A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values.

• Each row requires one number $p$ for $X_i = true$ (the number for $X_i = false$ is just $1-p$).

• If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers.

• I.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution.

• For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$).
Noisy-OR

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)
2) Independent failure probability $q_i$ for each cause alone

\[ P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i \]

<table>
<thead>
<tr>
<th>Cold</th>
<th>Flu</th>
<th>Malaria</th>
<th>$P(\text{Fever})$</th>
<th>$P(\neg \text{Fever})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.98</td>
<td>0.02 = 0.2 × 0.1</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.94</td>
<td>0.06 = 0.6 × 0.1</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.88</td>
<td>0.12 = 0.6 × 0.2</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.988</td>
<td>0.012 = 0.6 × 0.2 × 0.1</td>
</tr>
</tbody>
</table>

Number of parameters **linear** in number of parents
Constructing Bayesian networks

1. Choose an ordering of variables $X_1, ..., X_n$

2. For $i = 1$ to $n$
   - add $X_i$ to the network
   - select parents from $X_1, ..., X_{i-1}$ such that
     \[
     P(X_i \mid \text{Parents}(X_i)) = P(X_i \mid X_1, ..., X_{i-1})
     \]

This choice of parents guarantees:

\[
P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, ..., X_{i-1}) \quad \text{//(chain rule)}
\]

\[
= \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i)) \quad \text{//(by construction)}
\]
Semantics

The full joint distribution is defined as the product of the local conditional distributions:

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{Parents}(X_i)) \]

e.g., \( P(j \land m \land a \land \neg b \land \neg e) \)

\[ = P(j | a) \cdot P(m | a) \cdot P(a | \neg b, \neg e) \cdot P(\neg b) \cdot P(\neg e) \]
Inferring independencies from structure: d-separation

$I_G(X;Y|Z)$ denotes that $X$ is d-separated (directed separated) from $Y$ by $Z$ in directed graph $G$.
Markov blanket (and boundary)

Each node is conditionally independent of all others given its
Markov blanket: parents + children + children’s parents
The building block of causality: v-structure (arrow of time)

\[
\begin{align*}
\text{transitive } M &\neq \text{intransitive } M \\
M_p &= \{D(X;Z), D(Z;Y), D(X,Y), I(X;Y|Z)\} \\
M_p &= \{D(X;Z), D(Y;Z), I(X;Y), D(X;Y|Z)\}
\end{align*}
\]

Often: present knowledge renders future states conditionally independent. (confounding)

Ever(?) : present knowledge renders past states conditionally independent. (backward/atemporal confounding)
Observational equivalence of causal models

**Definition 11** Two DAGs $G_1, G_2$ are observationally equivalent, if they imply the same set of independence relations (i.e. $(X \perp Y | Z)_{G_1} \iff (X \perp Y | Z)_{G_2}$).

The implied equivalence classes may contain $n!$ number of DAGs (e.g. all the full networks representing no independencies) or just 1.

**Theorem 2** Two DAGs $G_1, G_2$ are observationally equivalent, iff they have the same skeleton (i.e. the same edges without directions) and the same set of v-structures (i.e. two converging arrows without an arrow between their tails).

**Definition 12** The essential graph representing observationally equivalent DAGs is a partially oriented DAG (PDAG), that represents the identically oriented edges called compelled edges of the observationally equivalent DAGs (i.e. in the equivalence class), such a way that in the common skeleton only the compelled edges are directed (the others are undirected representing inconclusiveness).
Compelled edges and PDAG

(“can we interpret edges as causal relations?” ➞ compelled edges)
The Causal Markov Condition

• A DAG is called a *causal structure* over a set of variables, if each node represents a variable and edges direct influences. A *causal model* is a causal structure extended with local probabilistic models.

• A causal structure $G$ and distribution $P$ satisfies the Causal Markov Condition, if $P$ obeys the local Markov condition w.r.t. $G$.

• The distribution $P$ is said to stable (or faithful), if there exists a DAG called *perfect map* exactly representing its (in)dependencies (i.e. $I_G(X;Y|Z) \iff I_P(X;Y|Z) \forall X,Y,Z \subseteq V$).

• CMC: *sufficiency* of $G$ (there is no extra, acausal edge)

• Faithfulness/stability: *necessity* of $G$ (there are no extra, parametric independency)
Interventional inference in causal Bayesian networks

• (Passive, observational) inference
  – $P(\text{Query} \mid \text{Observations})$

• **Interventionist inference**
  – $P(\text{Query} \mid \text{Observations, Interventions})$

• Counterfactual inference
  – $P(\text{Query} \mid \text{Observations, Counterfactual conditionals})$
Interventions and graph surgery

If $G$ is a causal model, then compute $p(Y|\text{do}(X=x))$ by

1. deleting the incoming edges to $X$
2. setting $X=x$
3. performing standard Bayesian network inference.
A deterministic concept of causation

- H. Simon
  - \( X_i = f_i(X_1, \ldots, X_{i-1}) \) for \( i = 1 \ldots n \)
  - In the linear case the system of equations indicates a natural causal ordering (flow of time?)

The probabilistic conceptualization is its generalization:

\[
P(X_i | X_1, \ldots, X_{i-1}) \sim X_i = f_i(X_1, \ldots, X_{i-1})
\]
Ovarian tumor diagnostics

International Ovarian Tumor Analysis (IOTA, Dirk Timmerman)

Decision networks
Inference tasks

Simple queries: compute posterior marginal $P(X_i|E=e)$
   e.g., $P(\text{NoGas}|\text{Gauge=empty, Lights=on, Starts=false})$

Conjunctive queries: $P(X_i, X_j|E=e) = P(X_i|E=e)P(X_j|X_i, E=e)$

Optimal decisions: decision networks include utility information;
   probabilistic inference required for $P(\text{outcome}|\text{action, evidence})$

Value of information: which evidence to seek next?

Sensitivity analysis: which probability values are most critical?

Explanation: why do I need a new starter motor?

Causal inference: what is the effect of an intervention?

Counterfactual inference: what would have been the effect of a hypothetical/imagery past intervention&observation?
Inference by enumeration

Let $X$ be all the variables. Typically, we want the posterior joint distribution of the query variables $Y$ given specific values $e$ for the evidence variables $E$.

Let the hidden variables be $H = X - Y - E$.

Then the required summation of joint entries is done by summing out the hidden variables:

$$P(Y|E = e) = \alpha P(Y, E = e) = \alpha \sum_h P(Y, E = e, H = h)$$

The terms in the summation are joint entries!

Obvious problems:
1) Worst-case time complexity $O(d^n)$ where $d$ is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???
Complexity of exact inference

Singly connected networks (or polytrees):
- any two nodes are connected by at most one (undirected) path
- time and space cost of exact inference $O(d^k n)$

Multiply connected networks:
- can reduce 3SAT to exact inference: $0 < p(\text{AND})? \Rightarrow \text{NP-hard}$
- equivalent to counting 3SAT models $\Rightarrow \#P$-complete
Következtetés többszörösen összekötött hálókban

Összevonós eljárások:
átalakítják a hálót egy valószínűségek szempontjából ekvivalens (de más topológiájú) fa gráffá, a nem megfelelő csomópontokat összevonva.

Sztochasztikus szimulációs eljárások:
a tárgytartomány nagyon nagy számú konkrét modelljét generálják le, ami konzisztens a valószínűségi háló által definiált eloszlással. Ez alapján az egzakt eredmények közelítését adják.
Sensitivity of the inference

\[ P(\text{Pathology}=\text{malignant}|E=e) \]
Decision theory = probability theory + utility theory

- Decision situation:
  - Actions
  - Outcomes
  - Probabilities of outcomes
  - Utilities/losses of outcomes
    - QALY, micromort
  - Maximum Expected Utility Principle (MEU)
    - Best action is the one with maximum expected utility

\[
EU(a_i) = \sum_j U(o_j | a_i) p(o_j | a_i)
\]

\[
a^* = \arg \max_i EU(a_i)
\]
Preferences

An agent chooses among prizes \((A, B, \text{ etc.})\) and lotteries, i.e., situations with uncertain prizes

Lottery \(L = [p, A; (1 - p), B]\)

Notation:

\(A \succ B\)  \(A\) preferred to \(B\)
\(A \sim B\)  indifference between \(A\) and \(B\)
\(A \preceq B\)  \(B\) not preferred to \(A\)
Rational preferences

Idea: preferences of a rational agent must obey constraints.

Rational preferences $\Rightarrow$

behavior describable as maximization of expected utility

Constraints:

Orderability

$(A \succ B) \lor (B \succ A) \lor (A \sim B)$

Transitivity

$(A \succ B) \land (B \succ C) \Rightarrow (A \succ C)$

Continuity

$A \succ B \succ C \Rightarrow \exists p \ [p, A; 1 - p, C] \sim B$

Substitutability

$A \sim B \Rightarrow [p, A; 1 - p, C] \sim [p, B; 1 - p, C]$  

Monotonicity

$A \succ B \Rightarrow (p \geq q \iff [p, A; 1 - p, B] \succeq [q, A; 1 - q, B])$
An irrational preference

Violating the constraints leads to self-evident irrationality

For example: an agent with intransitive preferences can be induced to give away all its money

If \( B \succ C \), then an agent who has \( C \) would pay (say) 1 cent to get \( B \)

If \( A \succ B \), then an agent who has \( B \) would pay (say) 1 cent to get \( A \)

If \( C \succ A \), then an agent who has \( A \) would pay (say) 1 cent to get \( C \)
Maximizing expected utility

**Theorem** (Ramsey, 1931; von Neumann and Morgenstern, 1944): Given preferences satisfying the constraints there exists a real-valued function $U$ such that

\[
U(A) \geq U(B) \iff A \succeq B
\]

\[
U([p_1, S_1; \ldots ; p_n, S_n]) = \sum_i p_i U(S_i)
\]

**MEU principle:**
Choose the action that maximizes expected utility

**Note:** an agent can be entirely rational (consistent with MEU) without ever representing or manipulating utilities and probabilities

E.g., a lookup table for perfect tic-tactoe
Utilities

Utilities map states to real numbers. Which numbers?

Standard approach to assessment of human utilities:
- compare a given state $A$ to a standard lottery $L_p$ that has
  - “best possible prize” $u_T$ with probability $p$
  - “worst possible catastrophe” $u_\bot$ with probability $(1 - p)$
- adjust lottery probability $p$ until $A \sim L_p$

pay $30 \sim L

\begin{tikzpicture}
  \def\r{1.5}
  \node (l) {L} ;
  \node (l1) [below left=\r cm and 0.5 cm of l] {0.999999} ;
  \node (l2) [below right=\r cm and 0.5 cm of l] {0.000001} ;
  \node (r) [below=2.5 \r cm of l] {continue as before} ;
  \node (r1) [below right=\r cm and 0.5 cm of r] {instant death} ;
  \draw[->] (l) -- (l1) ;
  \draw[->] (l) -- (l2) ;
  \draw[->] (l1) -- (r) ;
  \draw[->] (l2) -- (r1) ;
\end{tikzpicture}
Value of information

Current evidence \( E \), current best action \( \alpha \)
Possible action outcomes \( S_i \), potential new evidence \( E_j \)

\[
EU(\alpha|E) = \max_a \sum_i U(S_i) \ P(S_i|E, a)
\]

Suppose we knew \( E_j = e_{jk} \), then we would choose \( \alpha_{e_{jk}} \) s.t.

\[
EU(\alpha_{e_{jk}}|E, E_j = e_{jk}) = \max_a \sum_i U(S_i) \ P(S_i|E, a, E_j = e_{jk})
\]

\( E_j \) is a random variable whose value is currently unknown
⇒ must compute expected gain over all possible values:

\[
VPI_E(E_j) = \left( \sum_k P(E_j = e_{jk}|E) \ EU(\alpha_{e_{jk}}|E, E_j = e_{jk}) \right) - EU(\alpha|E)
\]

\( VPI = \text{value of perfect information} \)
Properties of VPI

Nonnegative—in expectation, not post hoc

$$\forall \ j, \ E \ VPI_E(E_j) \geq 0$$

Nonadditive—consider, e.g., obtaining $E_j$ twice

$$VPI_E(E_j, E_k) \neq VPI_E(E_j) + VPI_E(E_k)$$

Order-independent

$$VPI_E(E_j, E_k) = VPI_E(E_j) + VPI_{E,E_j}(E_k) = VPI_E(E_k) + VPI_{E,E_k}(E_j)$$

Note: when more than one piece of evidence can be gathered, maximizing VPI for each to select one is not always optimal

$\Rightarrow$ evidence-gathering becomes a **sequential** decision problem
Optimal binary decision

If the outcome $y$ and the prediction $\hat{y}$ are binary, the loss is defined by a binary cost matrix $C_{\hat{y}|y}$. The minimal loss decision is defined by

$$
\arg\min_{\hat{y}} C_{\hat{y}|0} P(Y = 0|x) + C_{\hat{y}|1} P(Y = 1|x),
$$

(8)

In case of $C_{0|0} = C_{1|1} = 0$, the prediction $\hat{y} = 1$ is optimal if

$$
\tau = \frac{C_{1|0}}{C_{1|0} + C_{0|1}} \leq P(Y = 1|x)
$$

(9)

where $\tau \in [0, 1]$ is the optimal decision threshold.
Exercise

– Select a domain, select candidate variables (5-10), and sketch the structure of the Bayesian network model.
– Consult it.
– Quantify the Bayesian networks.
– Evaluate it with global inference and „information sensitivity of inference” analysis.
– Generate a data set from your model.
– Learn a model from your data.
– Compare the structural and parametric differences between the two models.
– Extend your Bayesian network into a decision network.
– Investigate the value of further information.

• Optional tasks:
  – Analyse estimation biases.
  – Investigate the effect of model uncertainty and sample size on learning: vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning.
Subtask: test a decision network

• Investigate the value of further information as follows:
  • select values for some “evidence” variables (E=e),
  • using BayesCube calculate the current expected loss/utility EU(D|e),
  • select a variable “I” as potential “further” information,
  • using BayesCube calculate the conditional probabilities of potential further observations (i.e. the conditional probabilities of potential values of this “further information” variable, p(I=i|E=e)),
  • using BayesCube calculate the expected losses/utilities corresponding to these potential further observations EU(D|e,i),
  • calculate the (expected) value of (perfect) information corresponding to this variable “I”, Σi p(i|e)*EU(D|e,i)- EU(D|e).