ECE 175B: Probabilistic Reasoning and Graphical Models

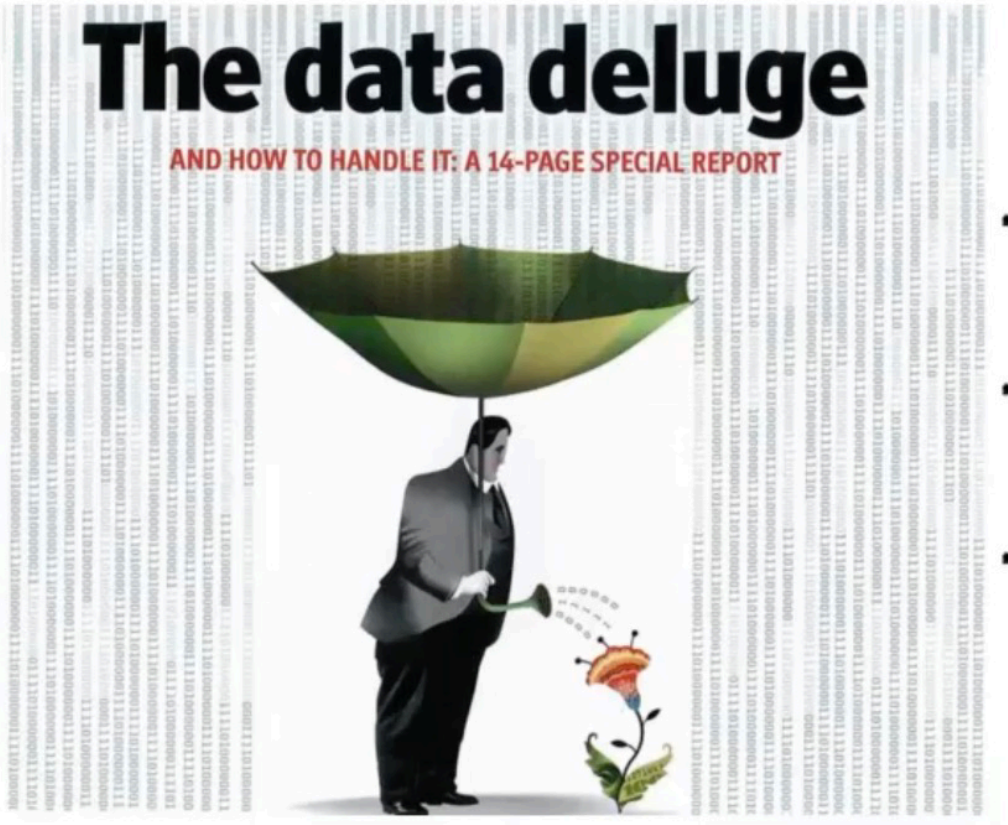
Florian Meyer

Scripps Institution of Oceanography
Electrical and Computer Engineering Department
University of California San Diego

SCRIPPS INSTITUTION OF OCEANOGRAPHY

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Big Data



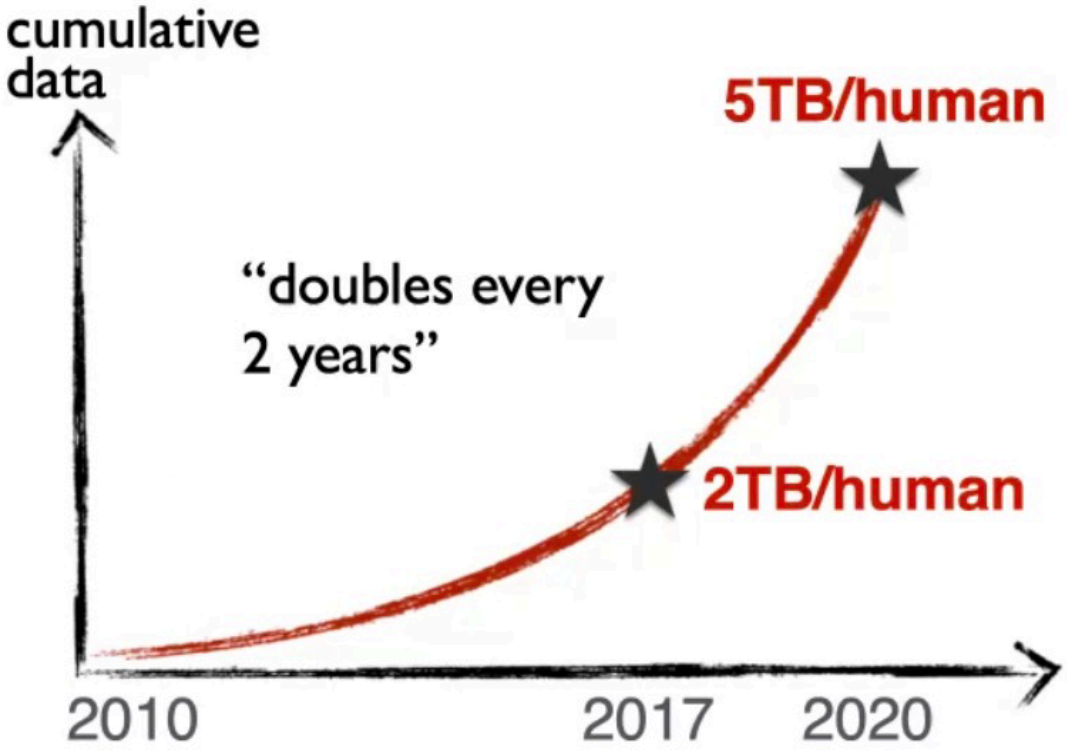
→ Volume

→ Velocity

→ Variety

The Economist
2010

Volume and Velocity



Source: IDC Digital Universe

Mathematical Formulation of a Graph

Graph: $\mathcal{G}=(\mathcal{V},\mathcal{E})$ such that $\mathcal{E}\subseteq[\mathcal{V}]^2$

Vertex (node) set $\mathcal{V}$: a collection of objects

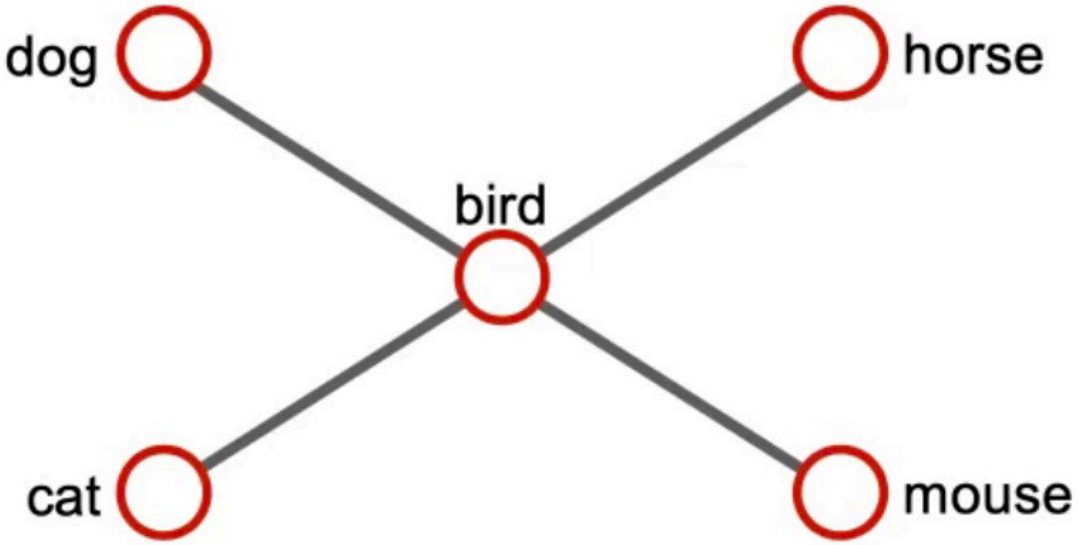
Edge set $\mathcal{E}$: relation between objects

Example:

$\mathcal{V}=\{\text{dog, cat, bird, horse, mouse}\}$

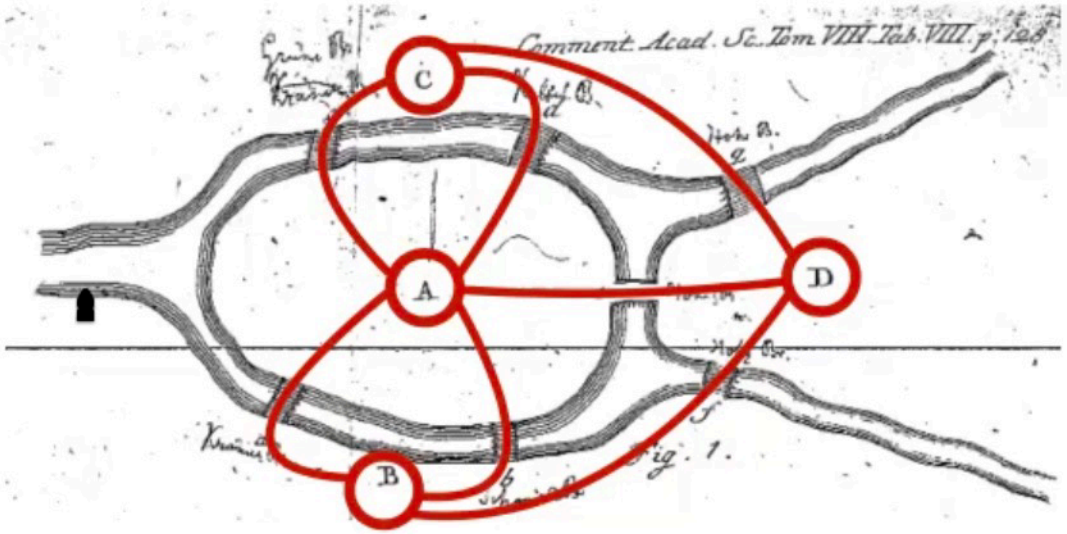
$\mathcal{E}=\{\{\text{dog, bird}\},\{\text{cat, bird}\},\{\text{bird, horse}\},\{\text{bird, mouse}\}\}$

Graph Representation

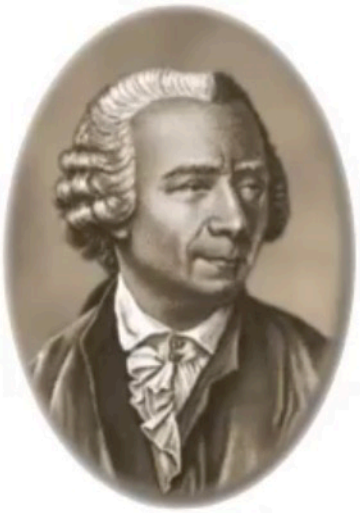


it's all about structure

History of Graphs



Seven Bridges of Königsberg



Leonhard
EULER

How to Use Graphs

- **data model** - captures abstract structure
- computational vehicle for **efficient algorithm**
- natural tool for **distributed** implementations

Tools and Applications

- **Graphical models** combine
  - graph theory: intuitive appeal
  - probability theory: interface between model and data
  - linear algebra & optimization: efficient algorithms
- Allow to deal with **uncertainty** and **complexity**
- **Application areas:** machine learning, information theory, communications, statistics, image processing, bioinformatics, statistical mechanics, social networks, sensor networks,
- **Unifying framework:** factor analysis, mixture models, hidden Markov models, Kalman filters, error correcting codes, Ising models, etc.

Example 1

- Consider binary (yes, no) **variables** relating to patient's health:

  asia, smoking, tuberculosis, cancer, bronchitis, xray, dyspnea

  Can you specify their joint distribution?

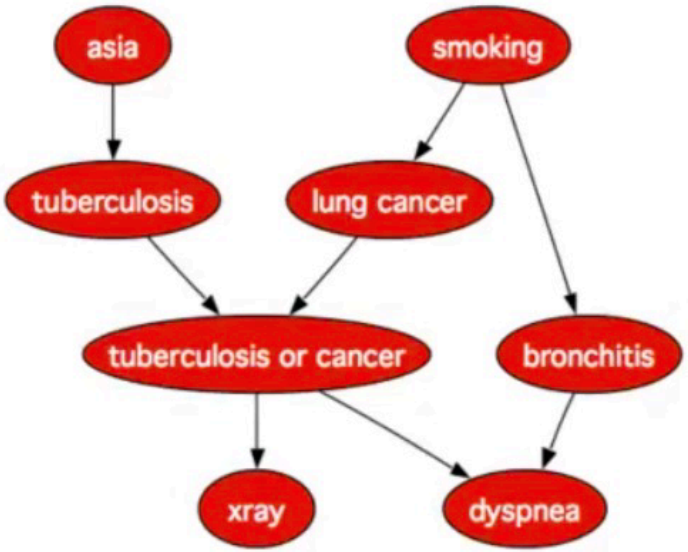
- Assume we know the **joint distribution**

  $p(\text{asia}, \text{smoking}, \text{tuberculosis}, \text{cancer}, \text{bronchitis}, \text{xray}, \text{dyspnea})$

  Can you efficiently calculate $p(\text{asia}|\text{dyspnea}=\text{no})$?

Example 1 cont.

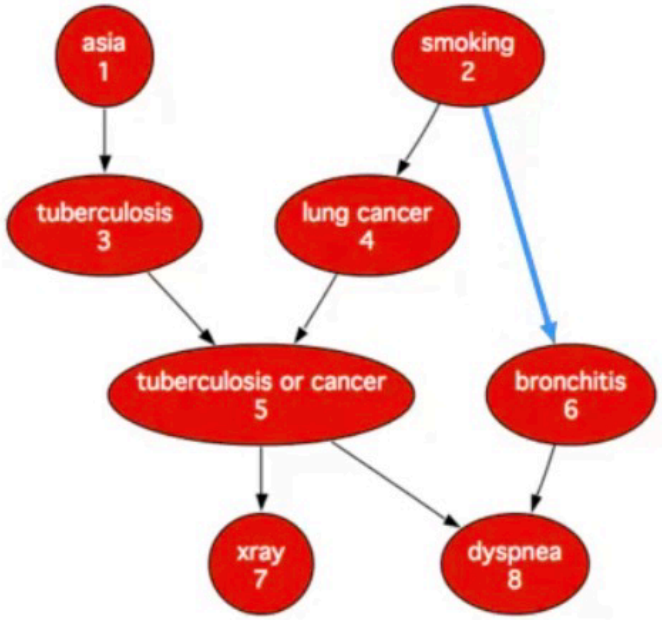
Bayesian network: directed acyclic graph (DAG)



- stay in asia increases risk of tuberculosis
- smoking can cause lung cancer and bronchitis
- tuberculosis, cancer, and bronchitis may result in dyspnea
- tuberculosis and cancer may lead to positive xray result

- 8 nodes (vertices): represent relevant variables
- 8 directed links (edges): represent interactions (causal influence)
- strength of interactions to be quantified via probabilities

Example 1 cont.



graph adjacency matrix

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 2 \\ 1 \\ 1 \\ 2 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

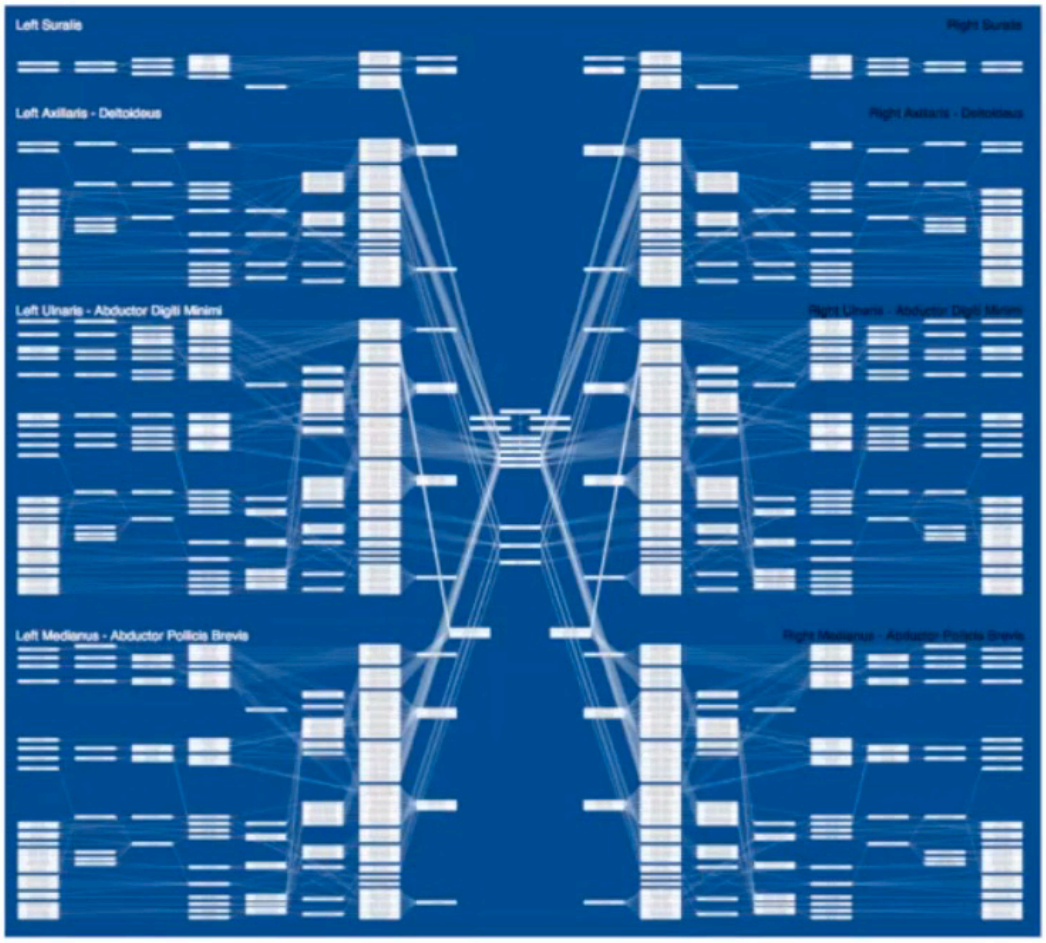
row sum: out-degrees (#children)

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 2 & 1 & 1 & 2 \end{pmatrix}$$

column sum: in-degrees (#parents)

- {asia, smoking}: in-degree zero; {xray, dyspnea}: out-degree zero
- **no directed cycle:** moving along arrows cannot get us back to any starting node

Example 2: MUNIN

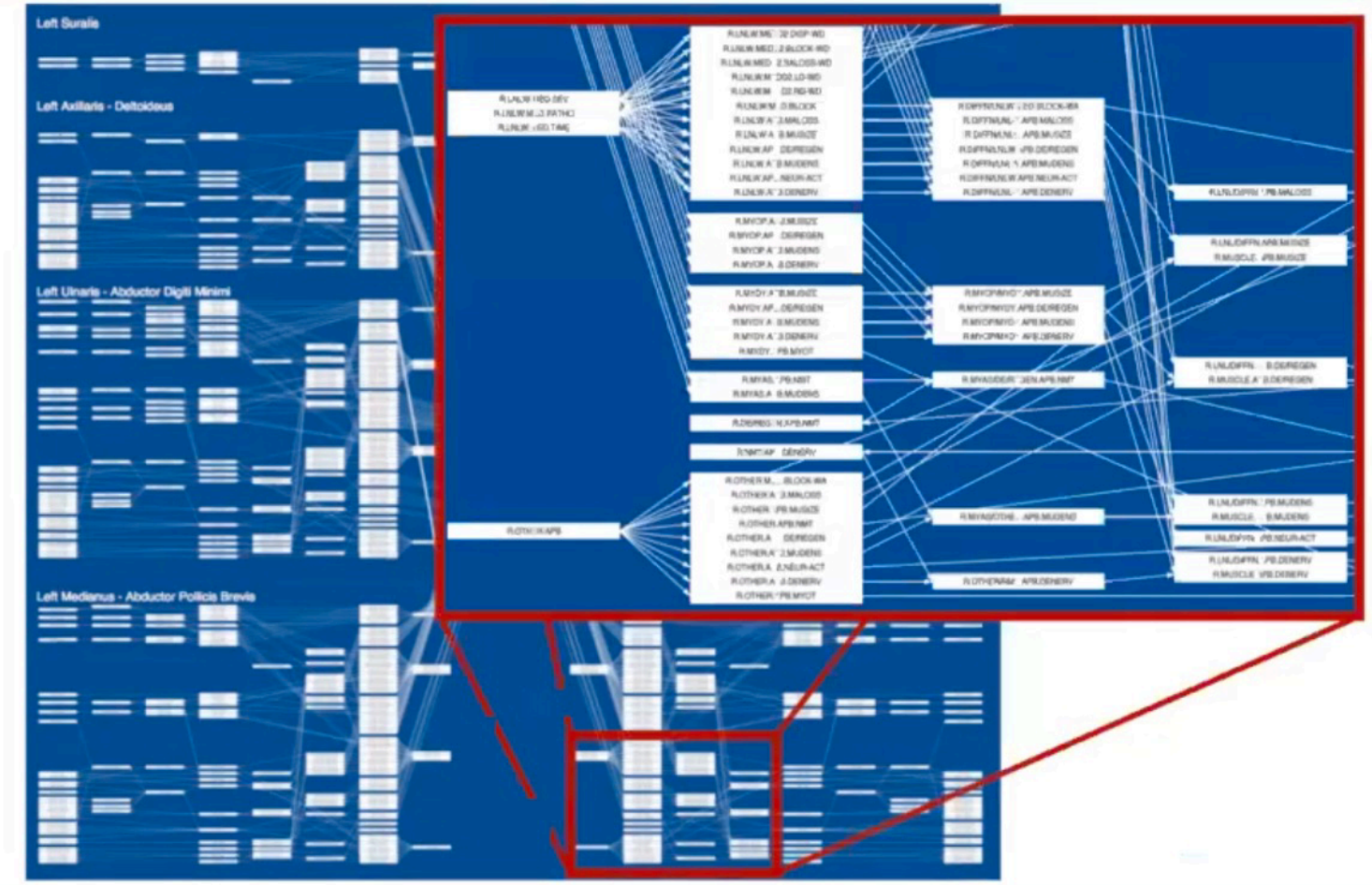


MUscle and **N**erve
Inference **N**etwork

Expert system for EMG-based diagnosis of neuromuscular disorder

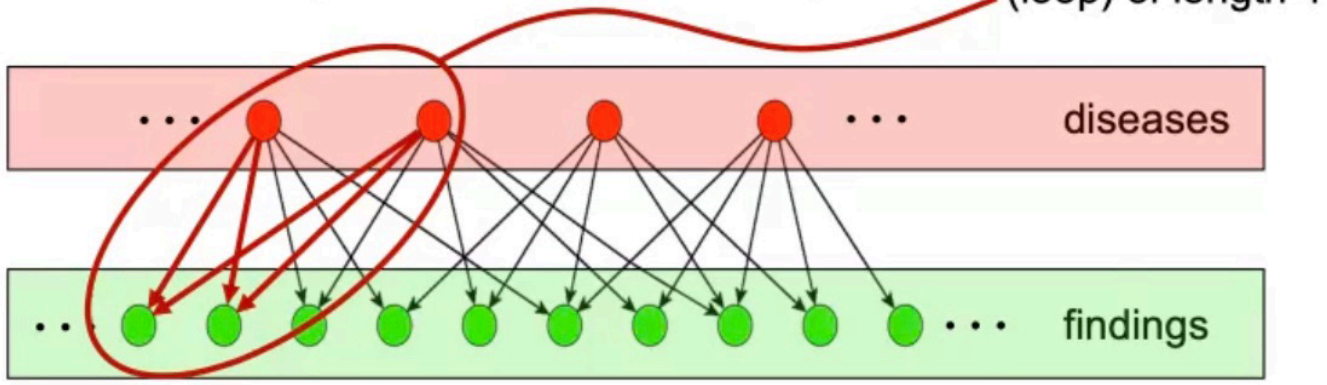
> 1000 nodes

Example 2: MUNIN



Example 3: Quick Medical Reference

- **Quick Medical Reference (QMR)**
  - bipartite (two-colorable) graph
  - variables represent diseases and findings

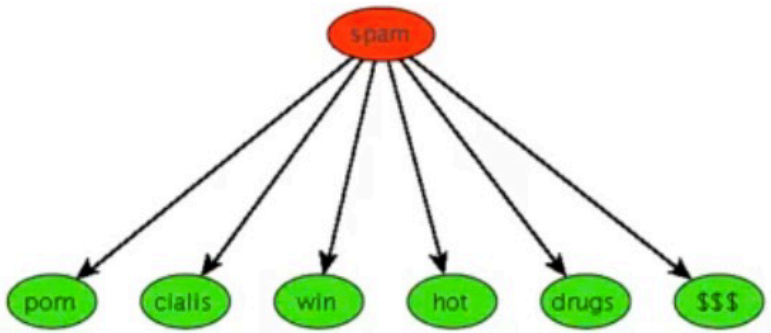


  - Univ. Pittsburgh QMR Database
    - ~40000 edges
    - ~400 diseases (average outdegree ~100)
    - ~4000 findings (average indegree ~10)

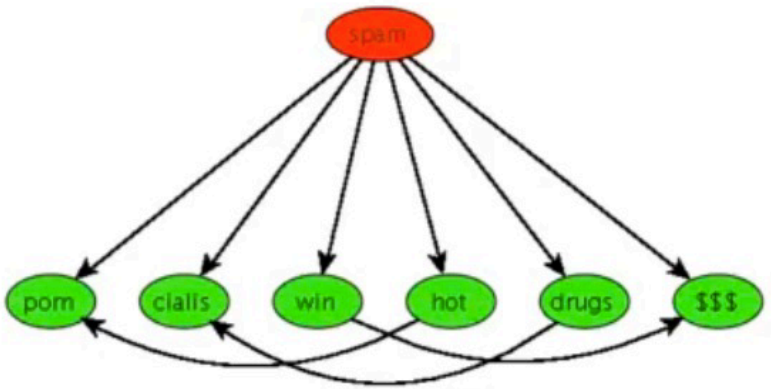
Example 4: Spam Filtering

Bayesian spam filters

- essentially text classifiers
- one root, many leaves representing words (or word fragments)

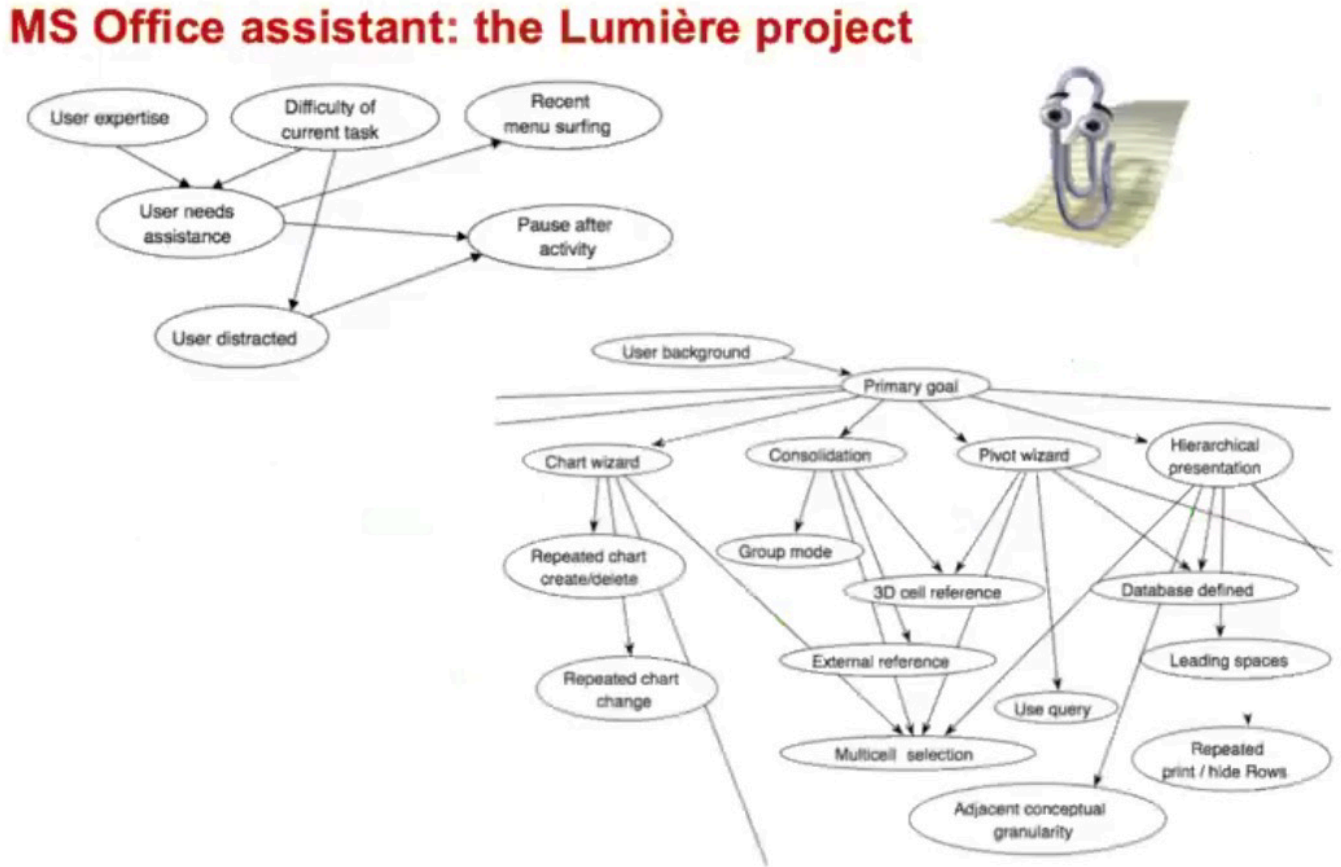


„naive Bayes“



Bayesian filter with dependencies

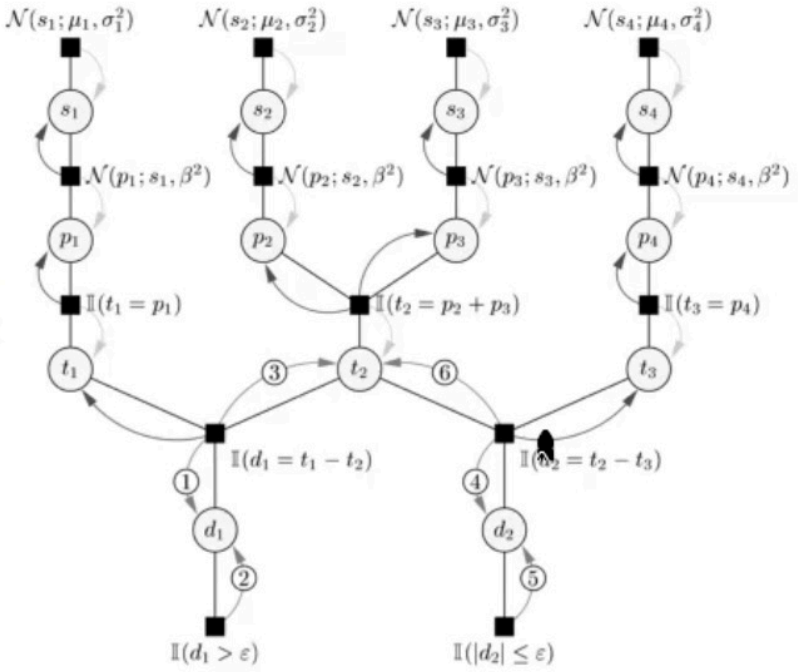
Example 5: MS Office Assistant



Example 6: Gaming

Microsoft's TrueSkill™: player skill rating

"planet-scale" application of Bayesian inference



Example 7: Autonomous Navigation

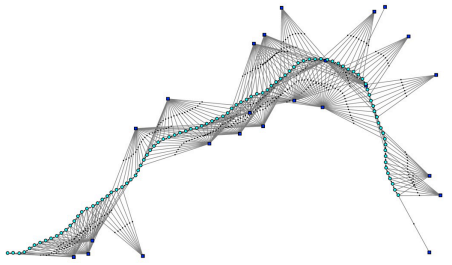
- **Data:** Stream of measurements from heterogeneous sensors




- **Goal:** Infer navigation information, objects in the environment, the floorplan, …

- Probabilistic reasoning and graphical models make it possible to consistently combine (``fuse'') data across time and sensors and provides explainable solutions

- Graphical representations are useful for
  - modeling large-scale problems
  - perform scalable inference



Course Overview

- **Online Platform:** Lecture notes, homework assignments, further handouts, and links to Zoom meetings will be posted on the *Canvas calendar.*
- **Teaching Assistant:** Parthasarathi Khirwadkar (pkhirwad@ucsd.edu)
- **Class Schedule:**
  - Lectures are Tuesdays and Thursdays 12:30PM – 1:50PM in WLH 2205
  - Discussion sessions are Mondays 12:00 – 12:50 in PETER 102
- **Office Hours:** Office hours are every *Friday at 3 PM via Zoom or in-person by appointment*
- **Grades:** Homework 30%, mid-term exam 30%, and final exam 40%

Homework

- **Homework assignments** and **their solution** will be posted approximately every 2-3 weeks on Canvas and will be due one week later. A scan or a high resolution photos of the solution must be uploaded to Canvas
- Collaborations are encouraged but the developed solutions you hand in should reflect your own understanding of the course material
- Homework is **graded using an "A for effort" scheme**. Individual homework problems are not corrected. Points are assigned proportionally to the percentage of work done
- Because homework is not graded for correctness, students must read the solutions, to determine if they performed the homework correctly or not

Exams

- The **mid-term and final exams** are graded in the traditional manner
- Both exams are closed book and notes
- Only nonprogrammable calculators can be used; all other electronic storage and communication devices are banned
- Cheating will result in penalties

Bibliography

- *Bayesian Reasoning & Machine Learning,* David Barber, Cambridge Univ. Press, 2012.

Lecture notes slides, and additional references will be posted on the Canvas website of the course.

Questions?