Introduction

Vibhav Gogate

THE UNIVERSITY OF TEXAS AT DALLAS
Erik Jonsson School of Engineering and Computer Science

The Course Title is a Misnomer

Contrived name: Statistical Methods in AI/ML

- A ploy to attract students

Real name: Probabilistic Graphical models (PGMs)

Why Study Probabilistic Graphical Models

- "PGMs have revolutionized AI and machine learning" – Eric Horvitz (Microsoft Research Head)
- Active Research area: UAI, NeurIPS, AAAI, ICML, ACL, EMNLP, ICCCV, CVPR and so on.
- Industrial Applications: Microsoft, IBM, Bio-tech, Yahoo!, Ebay and so on.

Data Scientist is the sexiest job of the 21st century

Back in the 1990s, computer engineer and Wall Street "quant" were the hot occupations in business. Today data scientists are the hires firms are competing to make. As companies wrestle with unprecedented volumes and types of information, demand for these experts has raced well ahead of supply. Indeed, Greylock Partners, the VC firm that backed Facebook and LinkedIn, is so worried about the shortage of data scientists that it has a recruiting team dedicated to channeling them to the businesses in its portfolio.
Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions. They find the story buried in the data and communicate it. And they dont just deliver reports: They get at the questions at the heart of problems and devise creative approaches to them. One data scientist who was studying a fraud problem, for example, realized it was analogous to a type of DNA sequencing problem. Bringing those disparate worlds together, he crafted a solution that dramatically reduced fraud losses.

Data Scientist is the sexiest job of the 21st century

- A good data scientist must have a good understanding of PGMs, machine learning, AI and statistics.
- Along with Data structures, Algorithms, Databases, Computer architecture, programming languages, systems, software engineering and theory of computation!
- Computer Scientist + AI + Stat + ML. All components are equally important.

Grading

Subject to Change! Tentative.

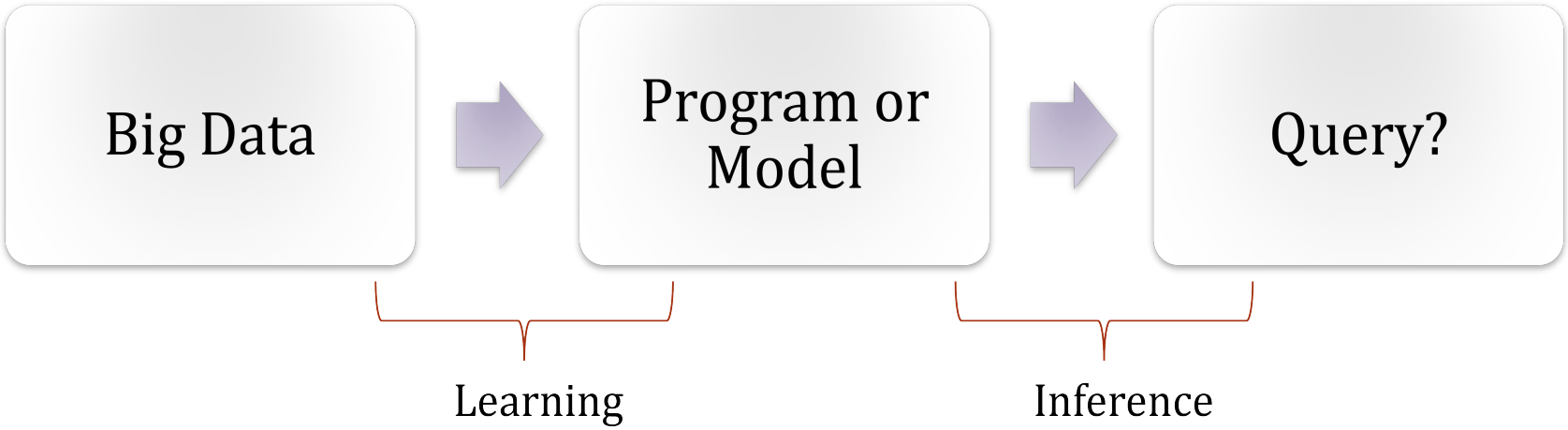
- 4-5 Homework assignments (60%)
  - Due in two weeks. More like mini-projects!
  - Programming involved
- Individual Project (20%)
- Take home Final Exam (20%). The final exam will be a mini-project. Due in two weeks.
- A (90+), B(80+) C (70+) D(60+) F (below 60).

Material

- Reasoning with Probabilistic and Deterministic Graphical Models, 2nd Edition Exact Algorithms, by Rina Dechter.
- Modeling and Reasoning with Bayesian networks, by Adnan Darwiche.
- Probabilistic Graphical Models: Principles and Techniques, by Daphne Koller and Nir Friedman.

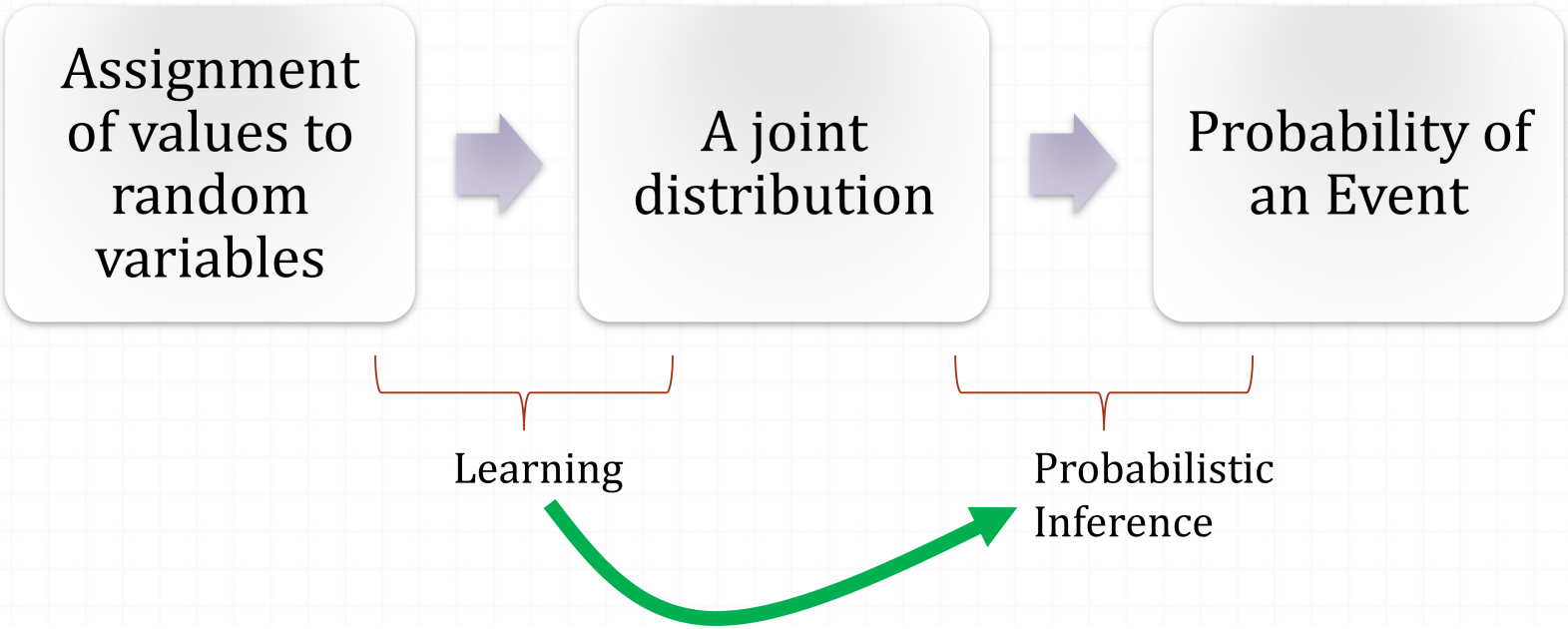
- Plus, class notes and papers!

High-level view of Machine Learning



Basic Tenet of probability theory: The world consists of random variables and the joint distribution over them represents complete knowledge!

Probabilistic Machine Learning



The Underlying Themes

Representation, Inference and Learning

- Representation: A modeling language. How to represent the joint distribution compactly?
  - Joint distribution grows exponentially! We don't have exponential space.
- Inference: Given a query and a model, find an answer to the query
- Learning: Given data, find the best possible model

The Underlying Themes

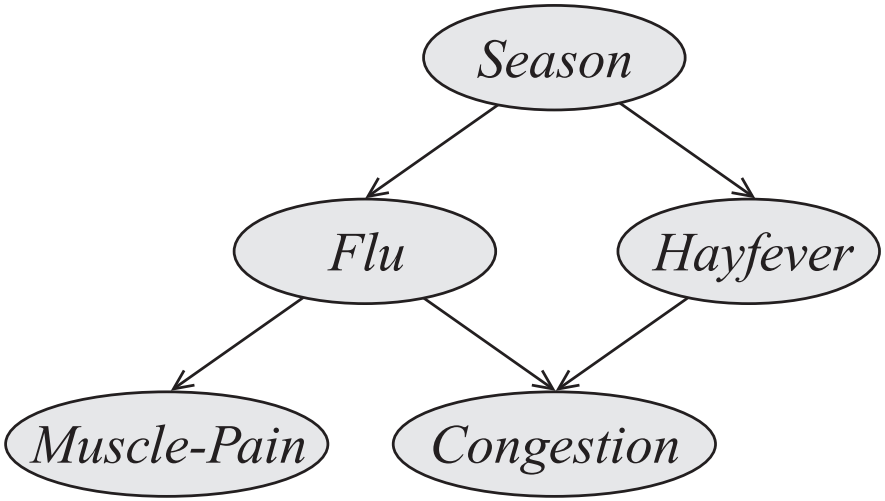
Declarative Knowledge Representation

- Most popular and widely used approach in AI and ML!
- Separate knowledge from reasoning.
- Model encodes how a system works, separate from algorithms that one can apply to it.
- Facilitates development of general-purpose algorithms.

The Underlying Themes

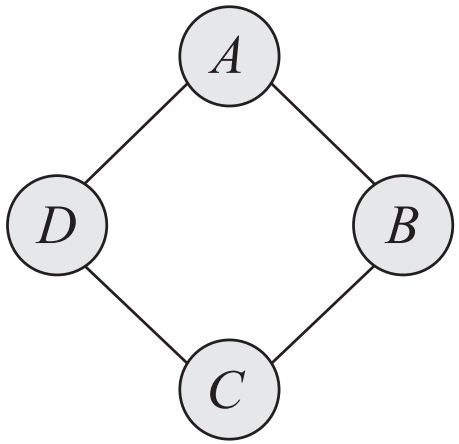
- Focus on complex, uncertain domains.
- Automated Reasoning
- Representing Uncertainty
  - Probability theory
- Tackling Complexity
  - Graph (Graphical models)
  - Propositional Logic
  - First-order logic (Relational models)
- Role of Statistics
  - Undergraduate class: describe the data succinctly.
  - This class: Inferring and learning graphical and relational models from data

Probabilistic Graphical Models



Factorization:

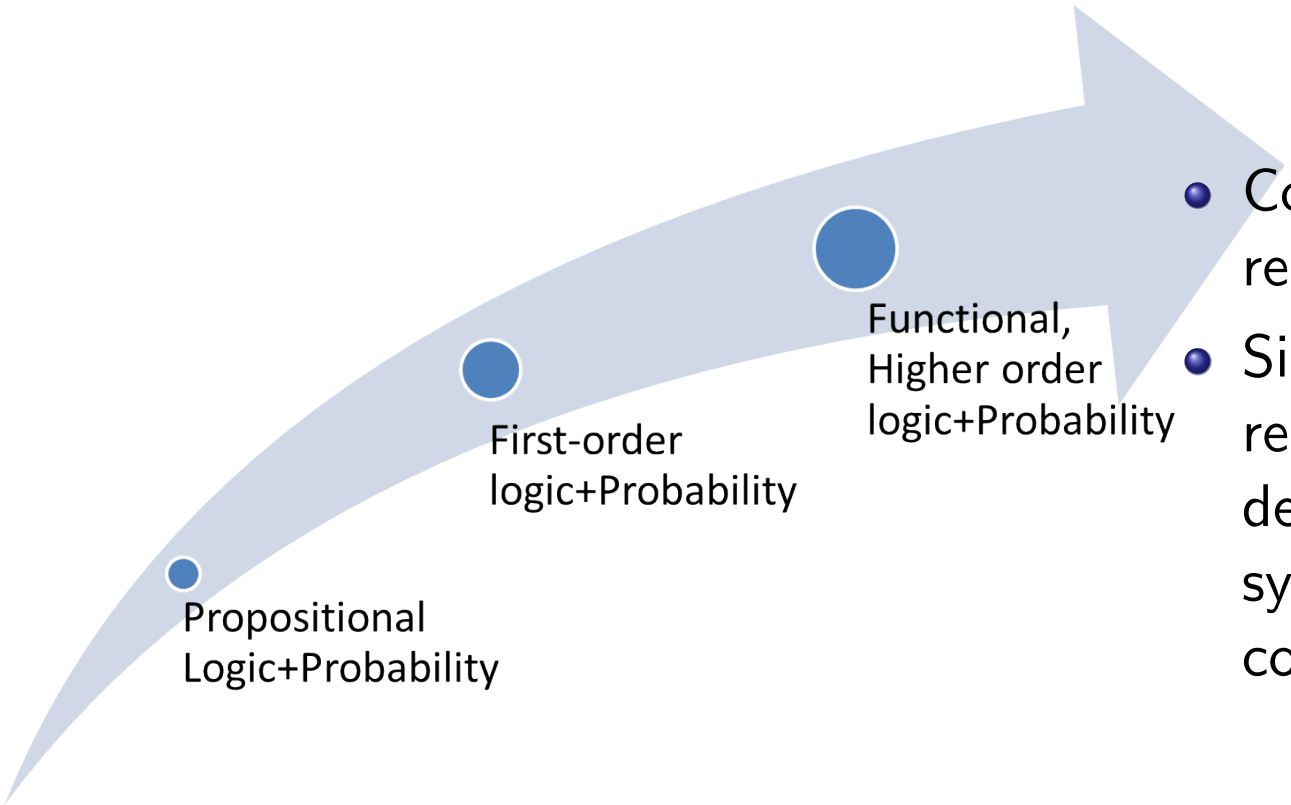
$P(S, F, H, C, M) =$
$P(S)P(F|S)$
$P(H|S)P(C|F, H)P(M|F)$



Factorization:

$P(S, F, H, C, M) =$
$\frac{1}{Z}\phi_1(A, B)$
$\phi_2(A, D)\phi_3(B, C)\phi_4(C, D)$

Where are we as a research community?



- Complexity of reasoning increases!
- Size of the representation decreases! (But then syntax is more complex)

Applications

- Computer Vision
- Natural language processing
- Speech processing
- Information Retrieval
- Databases and Data management
- Fraud Detection (e.g., at Experian, Ebay)

Applications (Continued)

- Ubiquitous computing
- Robotics
- Bio-informatics
- Chem-informatics
- Transportation science
- Human-computer interaction
- Network and Systems
- Computation Neuroscience

Course Objectives

[Internet figure]

By the time, we are at the end of this course

- You will put probability distributions on everything
- The hammer will be a graphical or a relational model and reasoning techniques for them
- You will beat every possible "application nail" with it

Course Diagram