Lecture 6.1

# Introduction to Probabilistic models

Machine Learning

Ivan Smetannikov

09.04.2021

# Lecture plan

- Overview and Motivation
- Distributions
- Factors

# Lecture plan

- Overview and Motivation
- Distributions
- Factors

# What is machine learning?

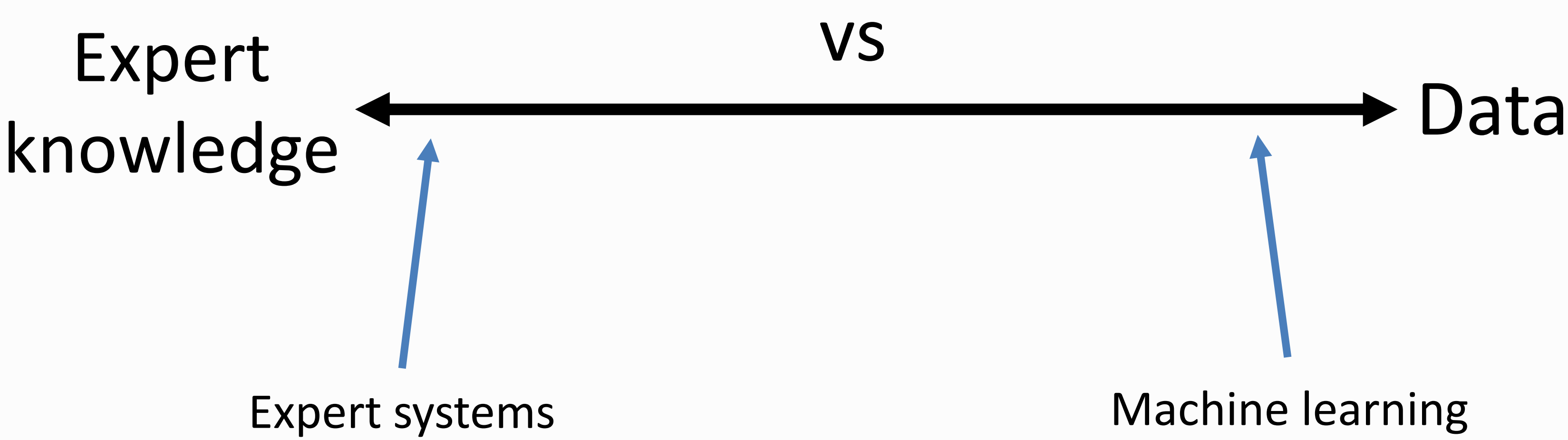
# What is machine learning?

Expert knowledge

vs

Data

# What is machine learning?



# Problem

An illness, which is spread among 1% of population. This illness test returns true answers in 95% of cases. Someone receives a positive result.

What is the probability, he actually suffers the illness?

# PGM: Motivation and Overview

predisposing
symptoms
test results
diseases
treatment outcomes

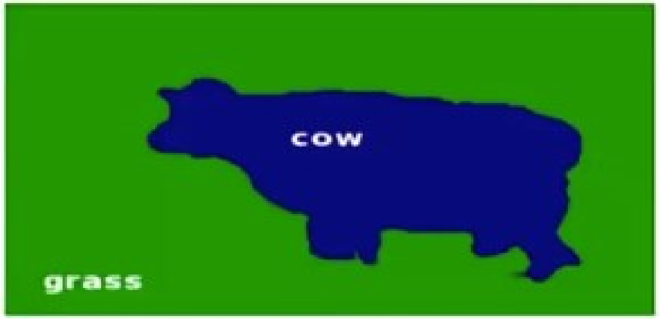
# PGM: Motivation and Overview

predisposing
symptoms
test results
diseases
treatment outcomes

millions of pixels or
thousands of superpixels

Each, needs to be labeled
{grass, sky, water, cow, horse, …}

# PGM: Motivation and Overview




predisposing
symptoms
test results
diseases
treatment outcomes

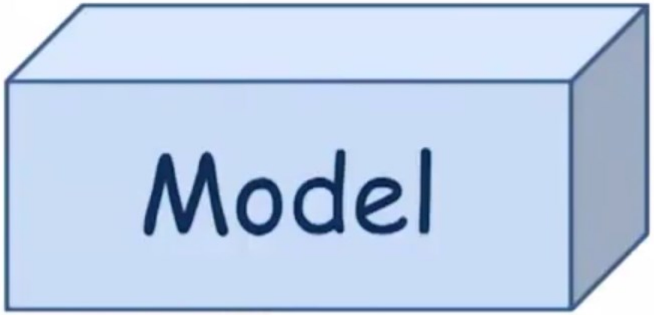
millions of pixels or
thousands of superpixels

Each, needs to be labeled
{grass, sky, water, cow, horse, …}

# Probabilistic Graphical Models

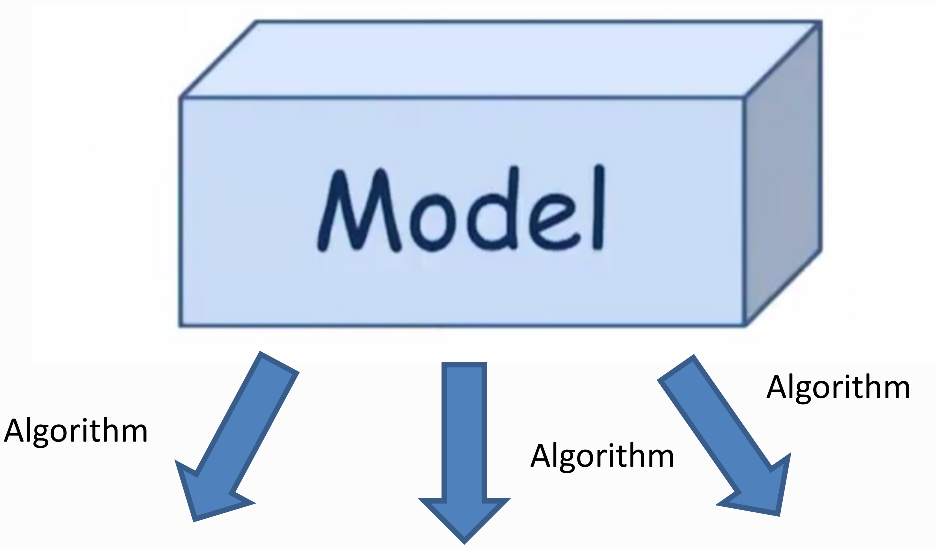
# Models

Declarative representation

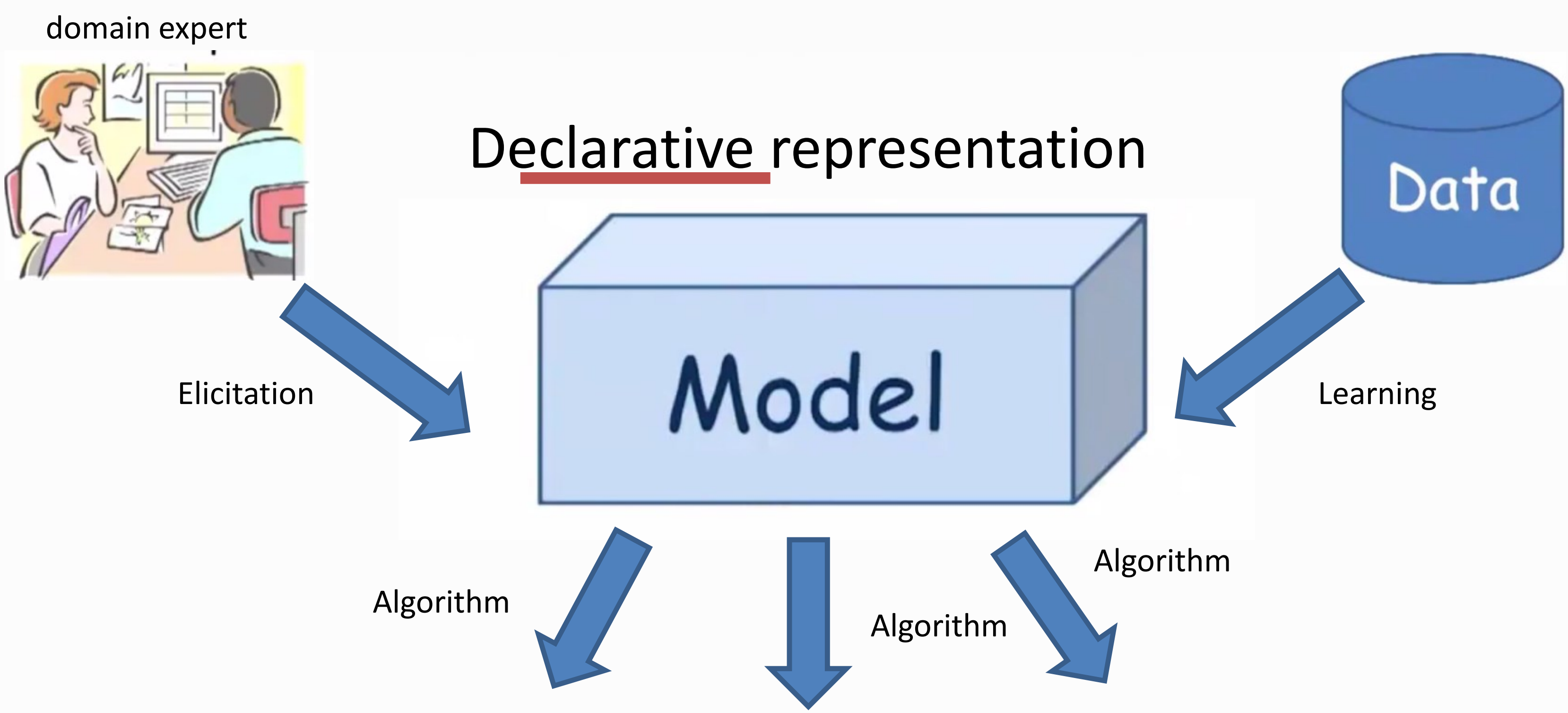


# Models

Declarative representation



# Models


domain expert
Declarative representation
Data
Elicitation
Model
Learning
Algorithm
Algorithm
Algorithm

# Uncertainty

- Partial knowledge of state of the world

# Uncertainty

- Partial knowledge of state of the world
- Noisy observations

# Uncertainty

- Partial knowledge of state of the world
- Noisy observations
- Phenomena not covered by our model

# Uncertainty

- Partial knowledge of state of the world
- Noisy observations
- Phenomena not covered by our model
- Inherent stochasticity

# Probability Theory

- Declarative representation with clear semantics

# Probability Theory

- Declarative representation with clear semantics
- Powerful reasoning patterns

Conditioning
Decision making

# Probability Theory

- Declarative representation with clear semantics
- Powerful reasoning patterns

**Conditioning**
**Decision making**

- Established learning methods

# Complex Systems

predisposing
symptoms
test results
diseases
treatment outcomes

class labels for
thousands of superpixels

# Complex Systems

predisposing
symptoms
test results
diseases
treatment outcomes

class labels for
thousands of superpixels

Random variables $\boxed{X_1, \ldots, X}_n$

# Complex Systems

predisposing
symptoms
test results
diseases
treatment outcomes

class labels for
thousands of superpixels

Random variables $\boxed{X_1, \ldots, X}_n$

$\boxed{\text{Joint}}$ distribution $P(\underline{X_1, \ldots, X_n})$

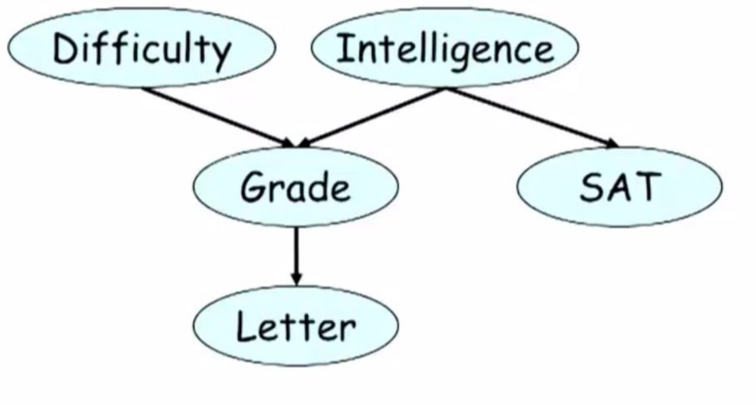
Binary valued distibrution over $2^n$ possible states

# Graphical Models

$X_1, \ldots, X_n$ — nodes

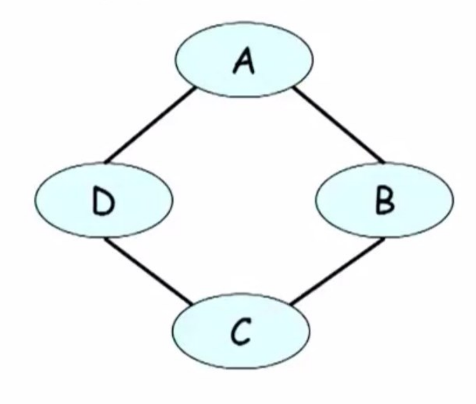
## Bayesian networks

Directed graph



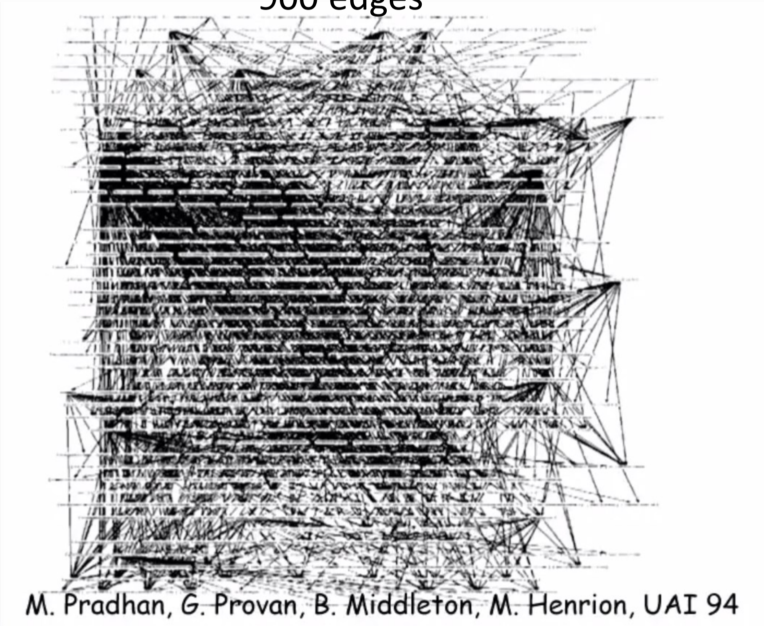

## Markov networks

Undirected graph



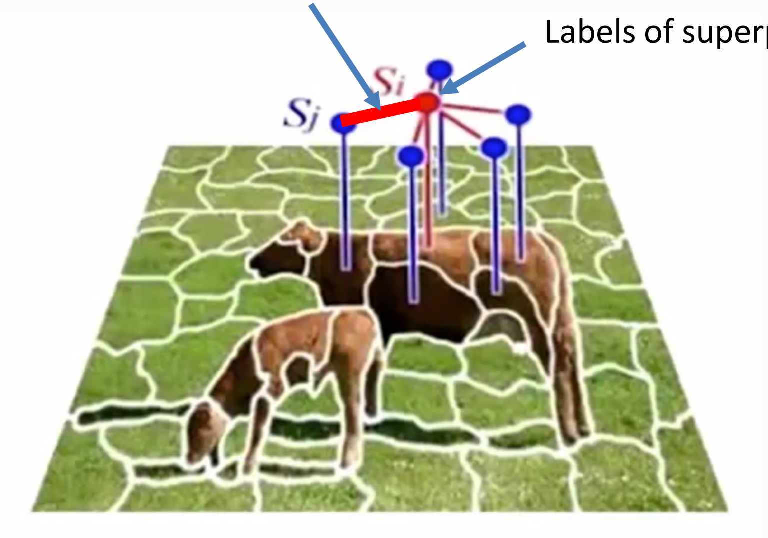
# Graphical Models

CPCS diagnosis
~480 nodes
~900 edges




Probabilistic relationships

Labels of superpixels



# Graphical Representation

- Intuitive and compact data structure
- Efficient reasoning using general–purpose algorithms
- Sparse parameterization
  - feasible elicitation ← by hand
  - learning from data ← automatically

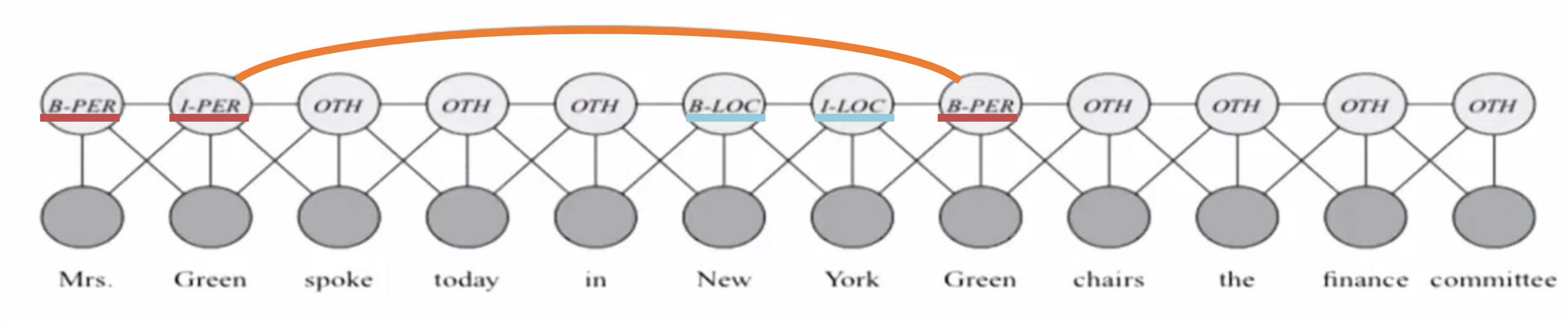
# Many Applications

- Medial diagnosis
- Fault diagnosis
- Natural language processing
- Traffic analysis
- Social network models
- Message decoding
- Computer vision
  - Image segmentation
  - 3D reconstruction
  - Holistic scene analysis
- Speech recognition
- Robot localization and mapping

# Textual Information Extracion

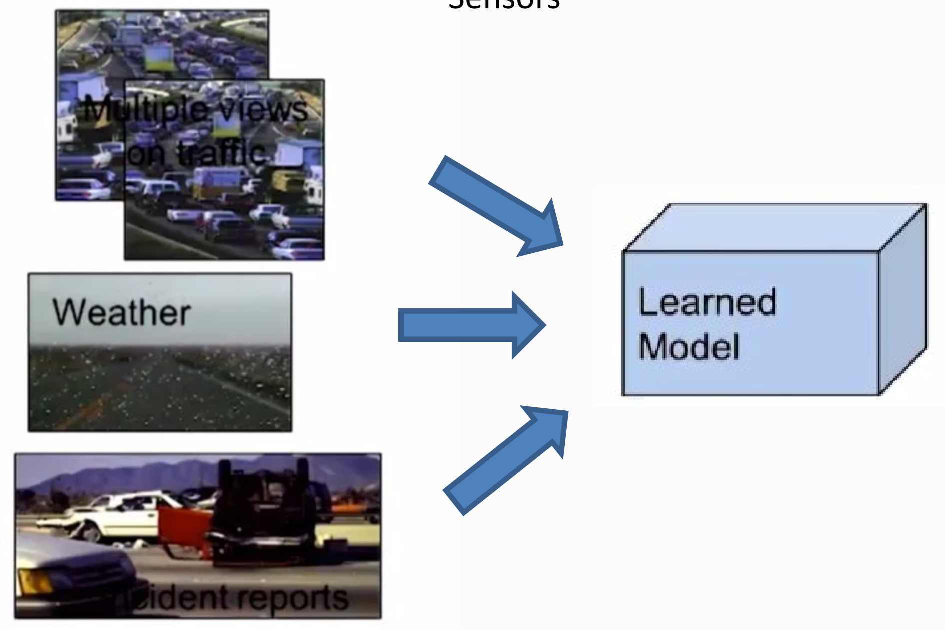
Mrs. Green spoke today in New York. Green chairs the finance committee.

person location person organization

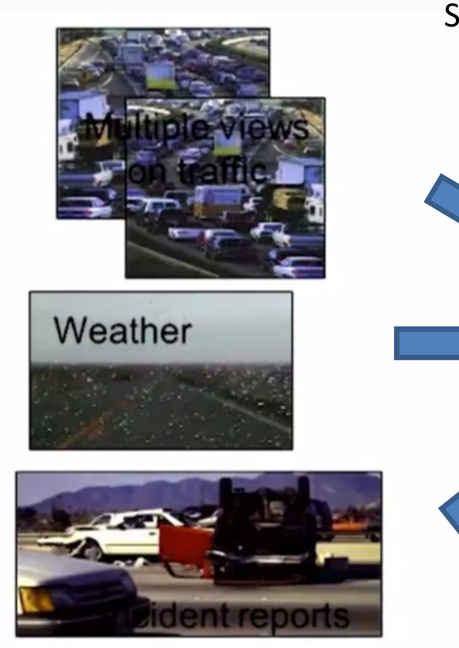


# Multi-Sensor Integration

Sensors

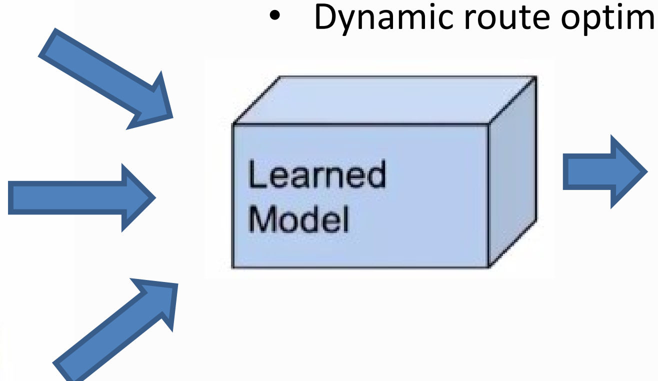


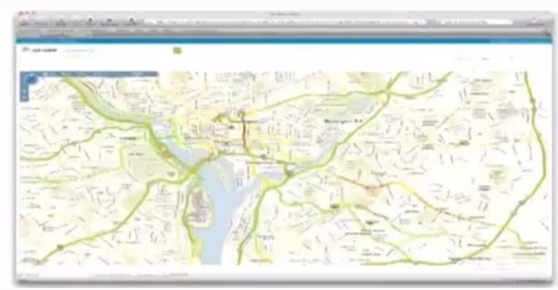
# Multi-Sensor Integration




Sensors

- Trained on historical data
- Learn to predict current and **future** road speed, including on **unmeasured** roads
- Dynamic route optimization



- I95 corridor experiment: accurate to ±5 MPH in 85% of cases
- Fielded in 72 cities

# Multi-Sensor Integration

- Representation
  - Directed and undirected
  - Temporal and plate models
- Inference
  - Exact and approximate
  - Decision making
- Learning
  - Parameters and structure
  - With and without complete data

# Lecture plan

- Overview and Motivation
- Distributions
- Factors

# Joint Distribution

- Intelligence (I)

  $i^0$ (low), $i^1$ (high)
- Difficulty (D)

  $d^0$ (easy), $d^1$ (hard)
- Grade (G)

  $g^1$ (A), $g^2$ (B), $g^3$ (C)

# Joint Distribution

- Intelligence (I) $\leftarrow$ 2
  $i^0$ (low), $i^1$ (high)
- Difficulty (D) $\leftarrow$ 2
  $d^0$ (easy), $d^1$ (hard)
- Grade (G) $\leftarrow$ 3
  $g^1$ (A), $g^2$ (B), $g^3$ (C)

# Joint Distribution

- Intelligence (I) ← 2

  $i^0$ (low), $i^1$ (high)
- Difficulty (D) ← 2

  $d^0$ (easy), $d^1$ (hard)
- Grade (G) ← 3

  $g^1$ (A), $g^2$ (B), $g^3$ (C)

**Parameters: 2x2x3=12**

| I | D | G | P(I,D,G) |
|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 |
| $i^0$ | $d^0$ | $g^2$ | 0.168 |
| $i^0$ | $d^0$ | $g^3$ | 0.126 |
| $i^0$ | $d^1$ | $g^1$ | 0.009 |
| $i^0$ | $d^1$ | $g^2$ | 0.045 |
| $i^0$ | $d^1$ | $g^3$ | 0.126 |
| $i^1$ | $d^0$ | $g^1$ | 0.252 |
| $i^1$ | $d^0$ | $g^2$ | 0.0224 |
| $i^1$ | $d^0$ | $g^3$ | 0.0056 |
| $i^1$ | $d^1$ | $g^1$ | 0.06 |
| $i^1$ | $d^1$ | $g^2$ | 0.036 |
| $i^1$ | $d^1$ | $g^3$ | 0.024 |

# Joint Distribution

- Intelligence (I) ← **2**

  $i^0$ (low), $i^1$ (high)
- Difficulty (D) ← **2**

  $d^0$ (easy), $d^1$ (hard)
- Grade (G) ← **3**

  $g^1$ (A), $g^2$ (B), $g^3$ (C)

| I | D | G | P(I,D,G) |
|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 |
| $i^0$ | $d^0$ | $g^2$ | 0.168 |
| $i^0$ | $d^0$ | $g^3$ | 0.126 |
| $i^0$ | $d^1$ | $g^1$ | 0.009 |
| $i^0$ | $d^1$ | $g^2$ | 0.045 |
| $i^0$ | $d^1$ | $g^3$ | 0.126 |
| $i^1$ | $d^0$ | $g^1$ | 0.252 |
| $i^1$ | $d^0$ | $g^2$ | 0.0224 |
| $i^1$ | $d^0$ | $g^3$ | 0.0056 |
| $i^1$ | $d^1$ | $g^1$ | 0.06 |
| $i^1$ | $d^1$ | $g^2$ | 0.036 |
| $i^1$ | $d^1$ | $g^3$ | 0.024 |

**sum=1**

**Parameters: 2x2x3=12**

**Independent parameters: 11**

# Conditioning

condition on $g^1$

| I | D | G | P(I,D,G) |
|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 |
| $i^0$ | $d^0$ | $g^2$ | 0.168 |
| $i^0$ | $d^0$ | $g^3$ | 0.126 |
| $i^0$ | $d^1$ | $g^1$ | 0.009 |
| $i^0$ | $d^1$ | $g^2$ | 0.045 |
| $i^0$ | $d^1$ | $g^3$ | 0.126 |
| $i^1$ | $d^0$ | $g^1$ | 0.252 |
| $i^1$ | $d^0$ | $g^2$ | 0.0224 |
| $i^1$ | $d^0$ | $g^3$ | 0.0056 |
| $i^1$ | $d^1$ | $g^1$ | 0.06 |
| $i^1$ | $d^1$ | $g^2$ | 0.036 |
| $i^1$ | $d^1$ | $g^3$ | 0.024 |

# Conditioning

condition on $g^1$

| I | D | G | P(I,D,G) |
|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 |
| ~~$i^0$~~ | ~~$d^0$~~ | ~~$g^2$~~ | ~~0.168~~ |
| ~~$i^0$~~ | ~~$d^0$~~ | ~~$g^3$~~ | ~~0.126~~ |
| $i^0$ | $d^1$ | $g^1$ | 0.009 |
| ~~$i^0$~~ | ~~$d^1$~~ | ~~$g^2$~~ | ~~0.045~~ |
| ~~$i^0$~~ | ~~$d^1$~~ | ~~$g^3$~~ | ~~0.126~~ |
| $i^1$ | $d^0$ | $g^1$ | 0.252 |
| ~~$i^1$~~ | ~~$d^0$~~ | ~~$g^2$~~ | ~~0.0224~~ |
| ~~$i^1$~~ | ~~$d^0$~~ | ~~$g^3$~~ | ~~0.0056~~ |
| $i^1$ | $d^1$ | $g^1$ | 0.06 |
| ~~$i^1$~~ | ~~$d^1$~~ | ~~$g^2$~~ | ~~0.036~~ |
| ~~$i^1$~~ | ~~$d^1$~~ | ~~$g^3$~~ | ~~0.024~~ |

# Conditioning: Reduction

condition on $g^1$

| I | D | G | P(I,D,G) |
|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 |
| | | | |
| | | | |
| $i^0$ | $d^1$ | $g^1$ | 0.009 |
| | | | |
| | | | |
| $i^1$ | $d^0$ | $g^1$ | 0.252 |
| | | | |
| | | | |
| $i^1$ | $d^1$ | $g^1$ | 0.06 |
| | | | |
| | | | |

# Conditioning: Renormalization

| I | D | G | P(I,D,G) |
|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 |
| $i^0$ | $d^1$ | $g^1$ | 0.009 |
| $i^1$ | $d^0$ | $g^1$ | 0.252 |
| $i^1$ | $d^1$ | $g^1$ | 0.06 |

**sum=0.447**

P(I,D, $g^1$)

unnormalized measure

# Conditioning: Renormalization

| I | D | G | P(I,D, $g^1$) | |
|---|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 | /0.447 |
| $i^0$ | $d^1$ | $g^1$ | 0.009 | /0.447 |
| $i^1$ | $d^0$ | $g^1$ | 0.252 | /0.447 |
| $i^1$ | $d^1$ | $g^1$ | 0.06 | /0.447 |

sum=0.447

$\Rightarrow$

| I | D | P(I,D\|$g^1$) |
|---|---|---|
| $i^0$ | $d^0$ | 0.282 |
| $i^0$ | $d^1$ | 0.02 |
| $i^1$ | $d^0$ | 0.564 |
| $i^1$ | $d^1$ | 0.134 |

P(I,D, $g^1$)

unnormalized

measure

# Conditioning: Marginalization

**Marginalize I**

| I | D | P(I,D) |
|---|---|---|
| $i^0$ | $d^0$ | 0.282 |
| $i^0$ | $d^1$ | 0.02 |
| $i^1$ | $d^0$ | 0.564 |
| $i^1$ | $d^1$ | 0.134 |

# Conditioning: Marginalization

**Marginalize I**

| I | D | P(I,D) |
|---|---|---|
| $i^0$ | $d^0$ | 0.282 |
| $i^0$ | $d^1$ | 0.02 |
| $i^1$ | $d^0$ | 0.564 |
| $i^1$ | $d^1$ | 0.134 |

| D | P(D) |
|---|---|
| $d^0$ | 0.846 |
| $d^1$ | 0.154 |

# Lecture plan

- Overview and Motivation
- Distributions
- Factors

# Factors

- A factor $\phi(X_1, \ldots, X_k)$

  $\phi: Val(X_1, \ldots, X_k) \rightarrow R$

- Scope = $\{X_1, \ldots, X_k\}$

# Joint distribution

**P(I,D,G)**

| I | D | G | P(I,D,G) |
|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 |
| $i^0$ | $d^0$ | $g^2$ | 0.168 |
| $i^0$ | $d^0$ | $g^3$ | 0.126 |
| $i^0$ | $d^1$ | $g^1$ | 0.009 |
| $i^0$ | $d^1$ | $g^2$ | 0.045 |
| $i^0$ | $d^1$ | $g^3$ | 0.126 |
| $i^1$ | $d^0$ | $g^1$ | 0.252 |
| $i^1$ | $d^0$ | $g^2$ | 0.0224 |
| $i^1$ | $d^0$ | $g^3$ | 0.0056 |
| $i^1$ | $d^1$ | $g^1$ | 0.06 |
| $i^1$ | $d^1$ | $g^2$ | 0.036 |
| $i^1$ | $d^1$ | $g^3$ | 0.024 |

# Unnormalized measure P(I,D, $g^1$)

**P(I,D, $g^1$)**

| I | D | G | P(I,D,G) |
|---|---|---|---|
| $i^0$ | $d^0$ | $g^1$ | 0.126 |
| $i^0$ | $d^1$ | $g^1$ | 0.009 |
| $i^1$ | $d^0$ | $g^1$ | 0.252 |
| $i^1$ | $d^1$ | $g^1$ | 0.06 |

Scope = {I,D}

# Conditional Probability Distribution (CPD)

**P(G|I,D)**

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

# Conditional Probability Distribution (CPD)

**P(G|I,D)**

| | $g^1$ | $g^2$ | $g^3$ | |
|---|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 | |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 | |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 | |
| context → $i^1, d^1$ | 0.5 | 0.3 | 0.2 | sum = 1 |
| | A | B | C | |

# General factors

| A | B | $\phi$ |
|---|---|---|
| $a^0$ | $b^0$ | 30 |
| $a^0$ | $b^1$ | 5 |
| $a^1$ | $b^0$ | 1 |
| $a^1$ | $b^1$ | 10 |

Scope = {A,B}

# Factor product

$\phi_1$

| | | |
|---|---|---|
| $a^1$ | $b^1$ | 0.5 |
| $a^1$ | $b^2$ | 0.8 |
| $a^2$ | $b^1$ | 0.1 |
| $a^2$ | $b^2$ | 0 |
| $a^3$ | $b^1$ | 0.3 |
| $a^3$ | $b^2$ | 0.9 |

Scope = {A,B}

$\phi_2$

| | | |
|---|---|---|
| $b^1$ | $c^1$ | 0.5 |
| $b^1$ | $c^2$ | 0.7 |
| $b^2$ | $c^1$ | 0.1 |
| $b^2$ | $c^2$ | 0.2 |

Scope = {B,C}

| | | | |
|---|---|---|---|
| $a^1$ | $b^1$ | $c^1$ | 0.5x0.5=0.25 |
| $a^1$ | $b^1$ | $c^2$ | 0.5x0.7=0.35 |
| $a^1$ | $b^2$ | $c^1$ | 0.8x0.1=0.08 |
| $a^1$ | $b^2$ | $c^2$ | 0.8x0.2=0.16 |
| $a^2$ | $b^1$ | $c^1$ | 0.1x0.5=0.05 |
| $a^2$ | $b^1$ | $c^2$ | 0.1x0.7=0.07 |
| $a^2$ | $b^2$ | $c^1$ | 0x0.1=0 |
| $a^2$ | $b^2$ | $c^2$ | 0x0.2=0 |
| $a^3$ | $b^1$ | $c^1$ | 0.3x0.5=0.15 |
| $a^3$ | $b^1$ | $c^2$ | 0.3x0.7=0.21 |
| $a^3$ | $b^2$ | $c^1$ | 0.9x0.1=0.09 |
| $a^3$ | $b^2$ | $c^2$ | 0.9x0.2=0.18 |

# Factor product

$\phi_1$

| | | |
|---|---|---|
| $a^1$ | $b^1$ | 0.5 |
| $a^1$ | $b^2$ | 0.8 |
| $a^2$ | $b^1$ | 0.1 |
| $a^2$ | $b^2$ | 0 |
| $a^3$ | $b^1$ | 0.3 |
| $a^3$ | $b^2$ | 0.9 |

Scope = {A,B}

$\phi_2$

| | | |
|---|---|---|
| $b^1$ | $c^1$ | 0.5 |
| $b^1$ | $c^2$ | 0.7 |
| $b^2$ | $c^1$ | 0.1 |
| $b^2$ | $c^2$ | 0.2 |

Scope = {B,C}

| | | | |
|---|---|---|---|
| $a^1$ | $b^1$ | $c^1$ | 0.5x0.5=0.25 |
| $a^1$ | $b^1$ | $c^2$ | 0.5x0.7=0.35 |
| $a^1$ | $b^2$ | $c^1$ | 0.8x0.1=0.08 |
| $a^1$ | $b^2$ | $c^2$ | 0.8x0.2=0.16 |
| $a^2$ | $b^1$ | $c^1$ | 0.1x0.5=0.05 |
| $a^2$ | $b^1$ | $c^2$ | 0.1x0.7=0.07 |
| $a^2$ | $b^2$ | $c^1$ | 0x0.1=0 |
| $a^2$ | $b^2$ | $c^2$ | 0x0.2=0 |
| $a^3$ | $b^1$ | $c^1$ | 0.3x0.5=0.15 |
| $a^3$ | $b^1$ | $c^2$ | 0.3x0.7=0.21 |
| $a^3$ | $b^2$ | $c^1$ | 0.9x0.1=0.09 |
| $a^3$ | $b^2$ | $c^2$ | 0.9x0.2=0.18 |

Scope = {A,B,C}

# Factor Marginalization

**Scope = {A,B,C}**

| | | | |
|---|---|---|---|
| $a^1$ | $b^1$ | $c^1$ | 0.25 |
| $a^1$ | $b^1$ | $c^2$ | 0.35 |
| $a^1$ | $b^2$ | $c^1$ | 0.08 |
| $a^1$ | $b^2$ | $c^2$ | 0.16 |
| $a^2$ | $b^1$ | $c^1$ | 0.05 |
| $a^2$ | $b^1$ | $c^2$ | 0.07 |
| $a^2$ | $b^2$ | $c^1$ | 0 |
| $a^2$ | $b^2$ | $c^2$ | 0 |
| $a^3$ | $b^1$ | $c^1$ | 0.15 |
| $a^3$ | $b^1$ | $c^2$ | 0.21 |
| $a^3$ | $b^2$ | $c^1$ | 0.09 |
| $a^3$ | $b^2$ | $c^2$ | 0.18 |

Marginalize B

**Scope = {A,C}**

| | | |
|---|---|---|
| $a^1$ | $c^1$ | 0.33 |
| $a^1$ | $c^2$ | 0.51 |
| $a^2$ | $c^1$ | 0.05 |
| $a^2$ | $c^2$ | 0.07 |
| $a^3$ | $c^1$ | 0.24 |
| $a^3$ | $c^2$ | 0.39 |

**=0.25+0.08**

# Factor reduction

| | | | |
|---|---|---|---|
| $a^1$ | $b^1$ | $c^1$ | 0.25 |
| $a^1$ | $b^1$ | $c^2$ | 0.35 |
| $a^1$ | $b^2$ | $c^1$ | 0.08 |
| $a^1$ | $b^2$ | $c^2$ | 0.16 |
| $a^2$ | $b^1$ | $c^1$ | 0.05 |
| $a^2$ | $b^1$ | $c^2$ | 0.07 |
| $a^2$ | $b^2$ | $c^1$ | 0 |
| $a^2$ | $b^2$ | $c^2$ | 0 |
| $a^3$ | $b^1$ | $c^1$ | 0.15 |
| $a^3$ | $b^1$ | $c^2$ | 0.21 |
| $a^3$ | $b^2$ | $c^1$ | 0.09 |
| $a^3$ | $b^2$ | $c^2$ | 0.18 |

**Reduce to the context $c^1$**

# Factor reduction

| | | | |
|---|---|---|---|
| $a^1$ | $b^1$ | $c^1$ | 0.25 |
| $a^1$ | $b^1$ | $c^2$ | 0.35 |
| $a^1$ | $b^2$ | $c^1$ | 0.08 |
| $a^1$ | $b^2$ | $c^2$ | 0.16 |
| $a^2$ | $b^1$ | $c^1$ | 0.05 |
| $a^2$ | $b^1$ | $c^2$ | 0.07 |
| $a^2$ | $b^2$ | $c^1$ | 0 |
| $a^2$ | $b^2$ | $c^2$ | 0 |
| $a^3$ | $b^1$ | $c^1$ | 0.15 |
| $a^3$ | $b^1$ | $c^2$ | 0.21 |
| $a^3$ | $b^2$ | $c^1$ | 0.09 |
| $a^3$ | $b^2$ | $c^2$ | 0.18 |

**Reduce to the context $c^1$**

| | | | |
|---|---|---|---|
| $a^1$ | $b^1$ | $c^1$ | 0.25 |
| $a^1$ | $b^2$ | $c^1$ | 0.08 |
| $a^2$ | $b^1$ | $c^1$ | 0.05 |
| $a^2$ | $b^2$ | $c^1$ | 0 |
| $a^3$ | $b^1$ | $c^1$ | 0.15 |
| $a^3$ | $b^2$ | $c^1$ | 0.09 |

**Scope = {A,B}**

# Why factors?

- Fundamental building block for defining distributions in high-dimensional spaces
- Set of basic operations for manipulating these probability distributions