Bayesian Sense-Making in Data Science

Michael L. Thompson, Ph.D.
https://www.linkedin.com/in/mlthomps/

P&G data&modelingsciences

6th Annual BayesiaLab Conference in Chicago

Nov. 2, 2018

Outline

- **Prevalence of Bayesian Applications**
- **Whence Bayesian Analysis?**
  - The *Model Structure* Information Content Diagram
  - Motivation for Bayesian Sense-Making
- **Key Concepts of Bayesian Sense-Making**
  - Use Case: General Recommender/Advisor Systems
- **Future Implications**
  - References to Get Started

Prevalence of Bayesian Applications

Bayesian Analysis to Delight Consumers at P&G

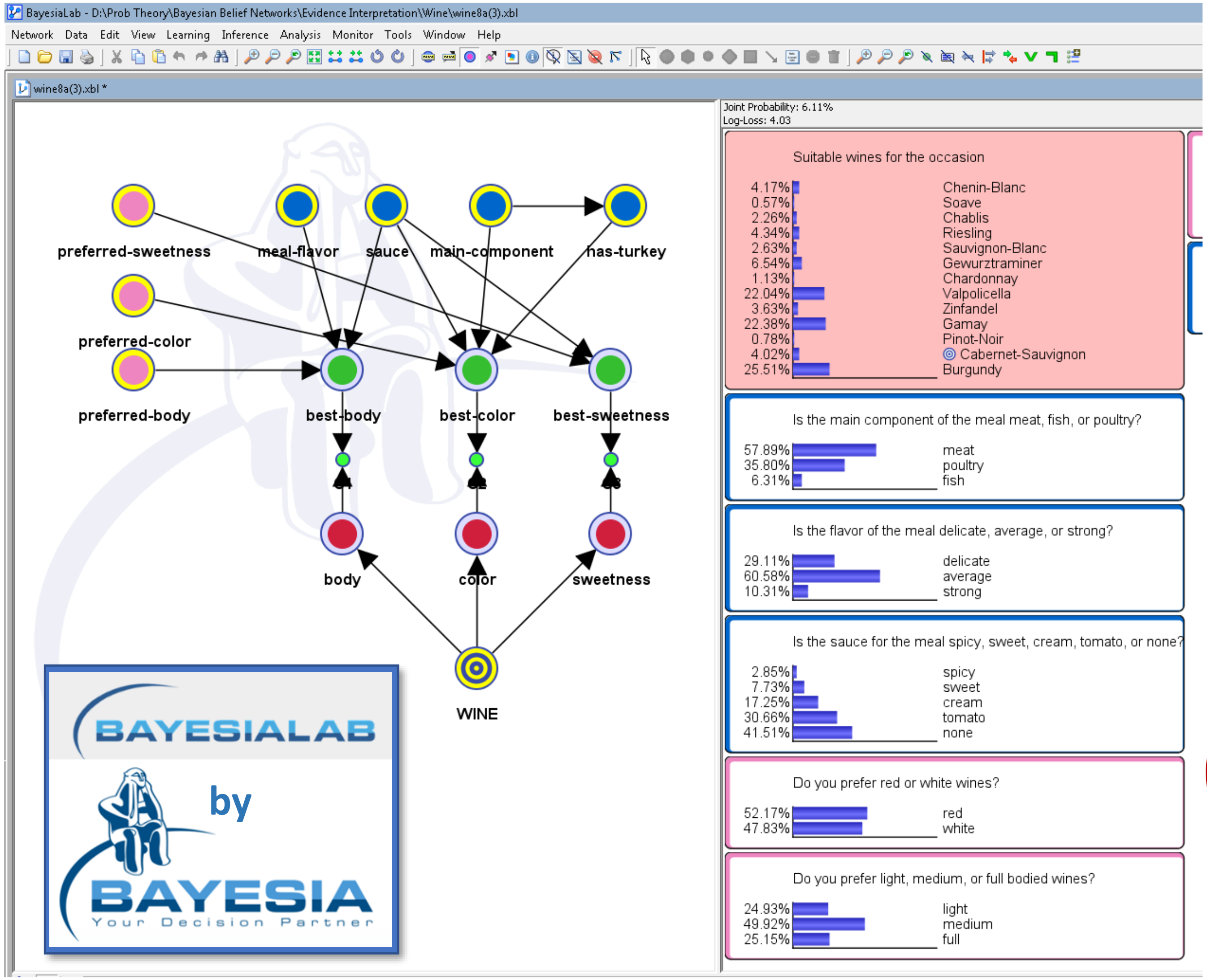
How Color Affects Decision

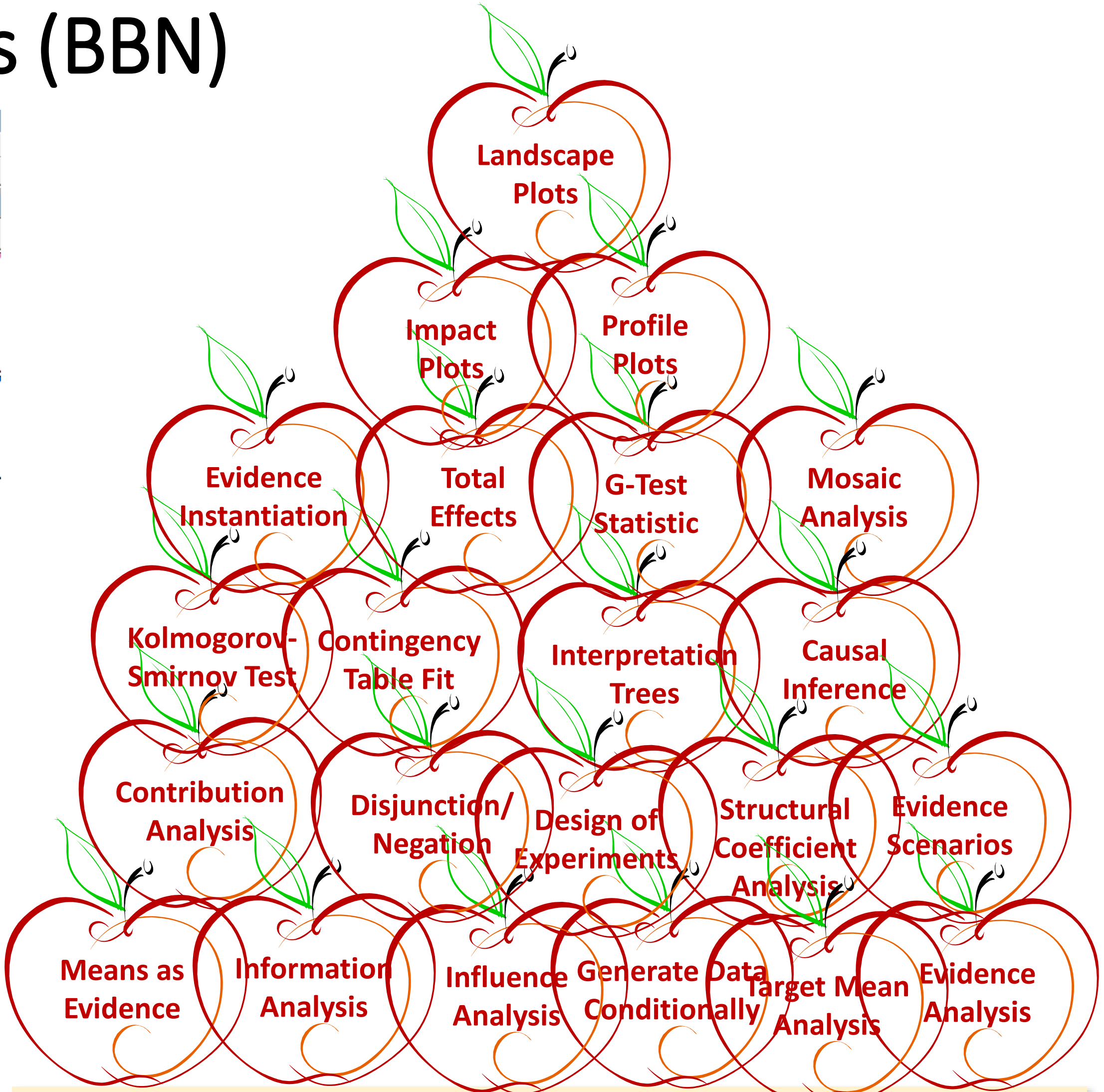
"…We have posed a … **basic question about consumer behavior, and the answer to this question is best captured by a multi-level, dichotomous, logistic regression model … using Bayesian inference by Markov Chain Monte Carlo simulation,….**"

Thompson, Michael L., et al, *P&G Core Technologies J.*, 2000

P&G/*Bayesia* Strategic Partnership

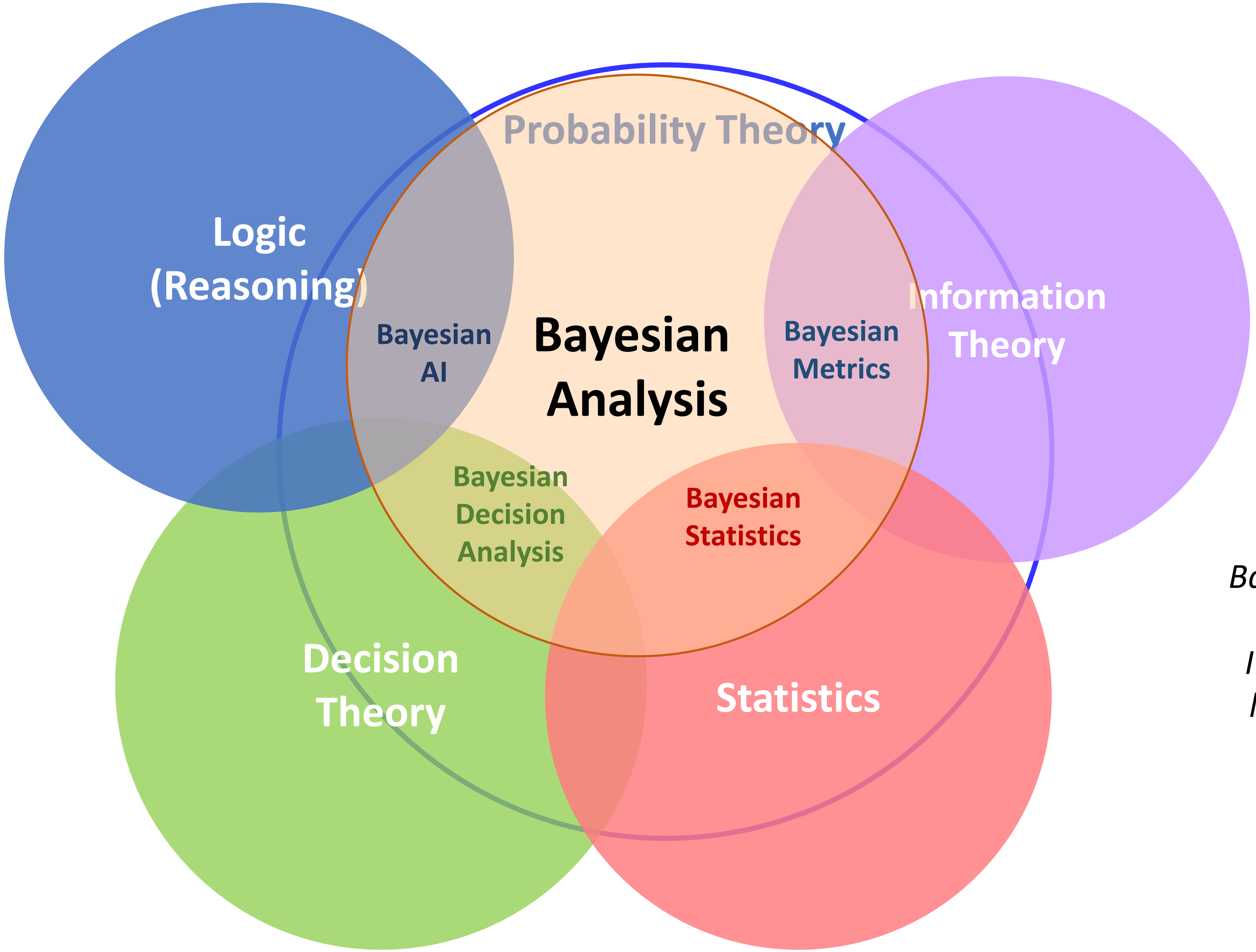
BayesiaLab for Bayesian Belief Networks (BBN)





Co-developed many *BayesiaLab* features

Whence Bayesian Analysis?



In short,
Bayesian Analysis is more than just adopting priors to model data.
It's reasoning about the world to learn and to drive decisions, i.e., Bayesian Sense-Making!

Model Structure: Sources of Information

Models are built by casting the information we have into mathematical functions.

$$V = \frac{4}{3}\pi r^3$$

- **K**nowledge representation
  - Domain knowledge (e.g., physical laws, theories of behavior, etc.)

$$y(x) = \theta_0 + \theta_1 x$$

- **M**athematical approximation
  - Simplifications and canonical functional forms (e.g., linear relationships, response surfaces, etc.)

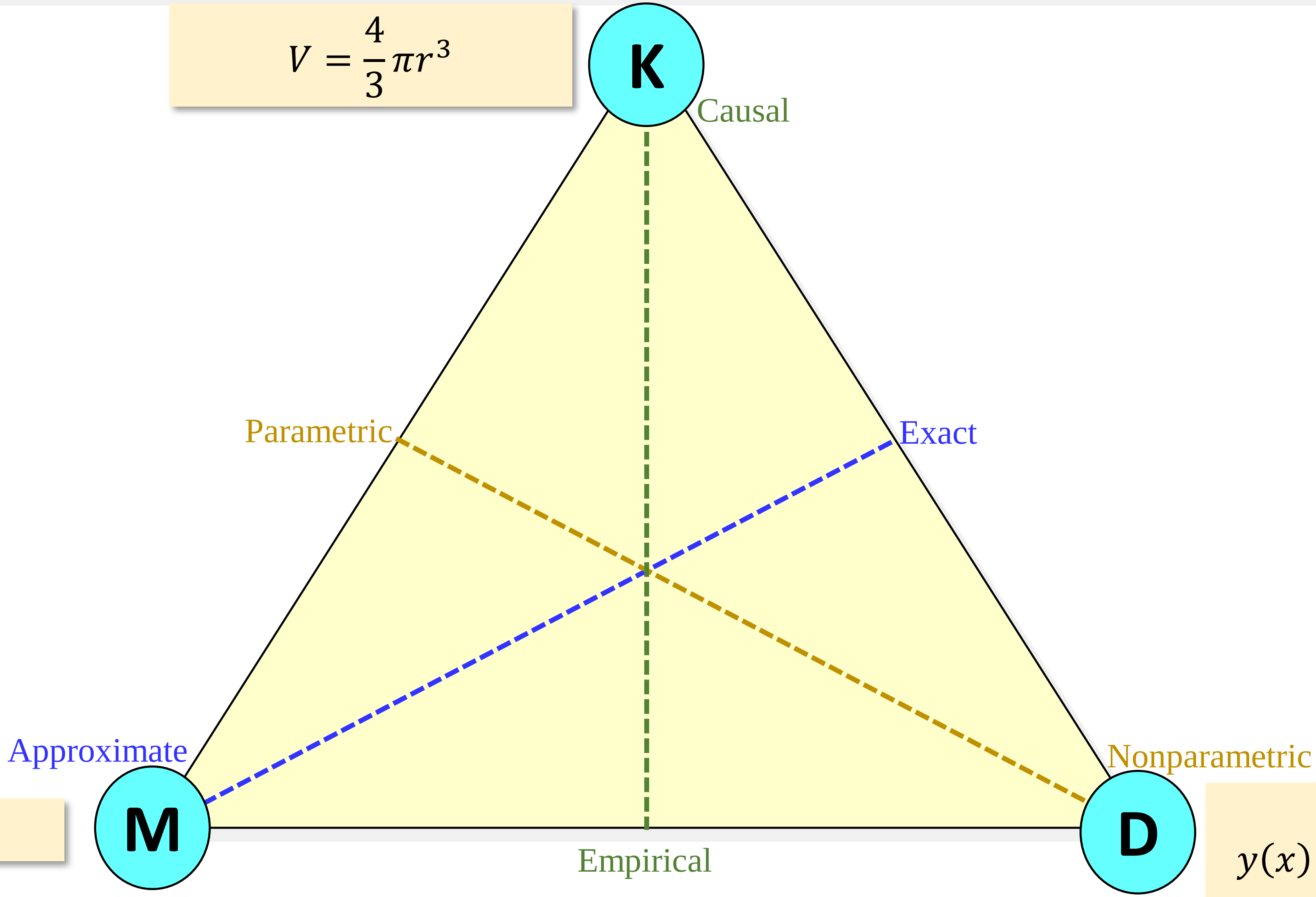
$$y(x) = \sum_{j=1}^{M} w_j f(x; \Theta)$$

- **D**ata considerations
  - Flexible combinations of basis functions that grow with the data (e.g., nonparametric density estimation, multivariate analyses, etc.)

Model Structure Information Content Diagram

A ternary mixture diagram of information sources that dictate model structure

$$V = \frac{4}{3}\pi r^3$$

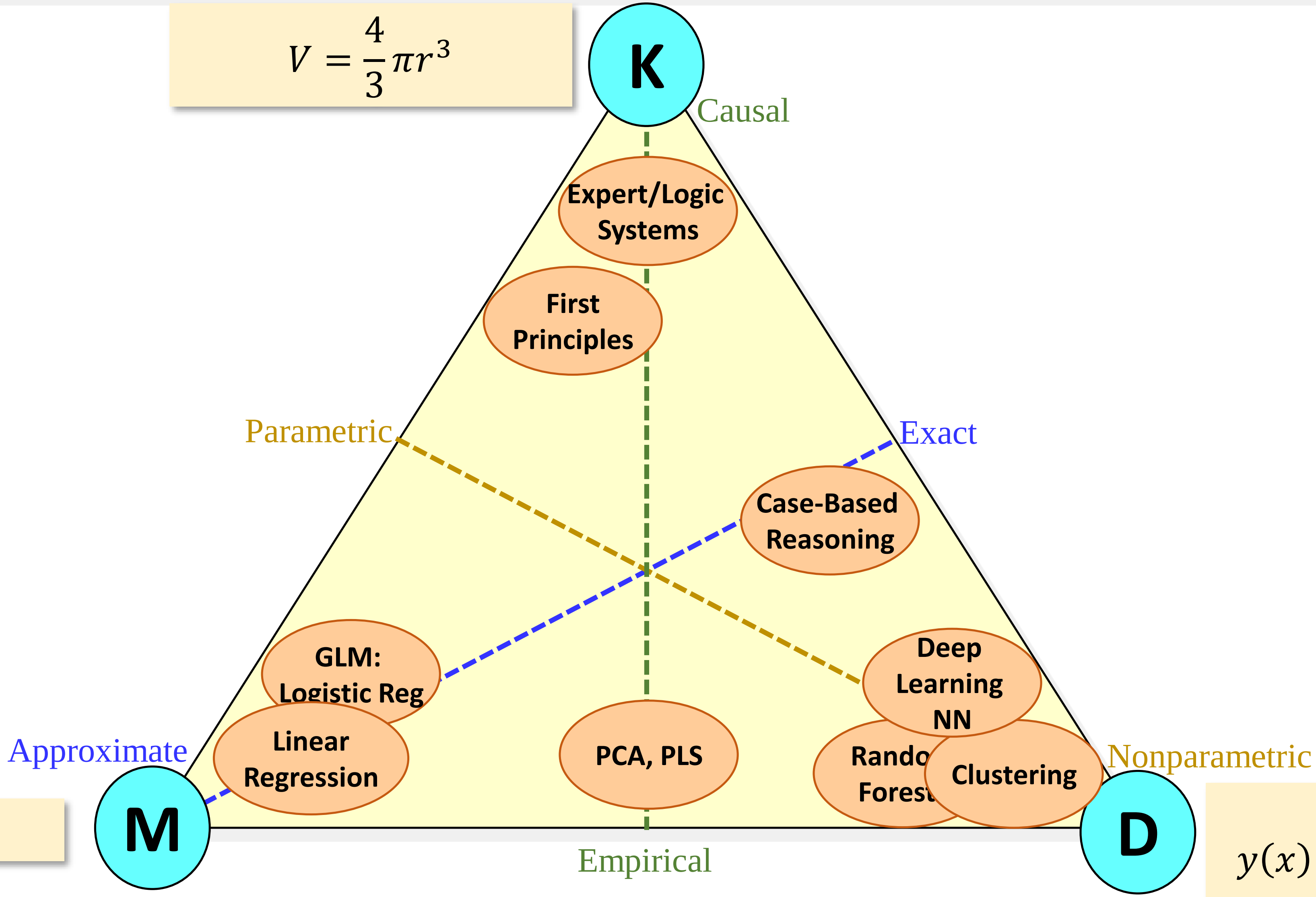


$$y(x) = \theta_0 + \theta_1 x$$

$$y(x) = \sum_{j=1}^{M} w_j f(x; \Theta)$$

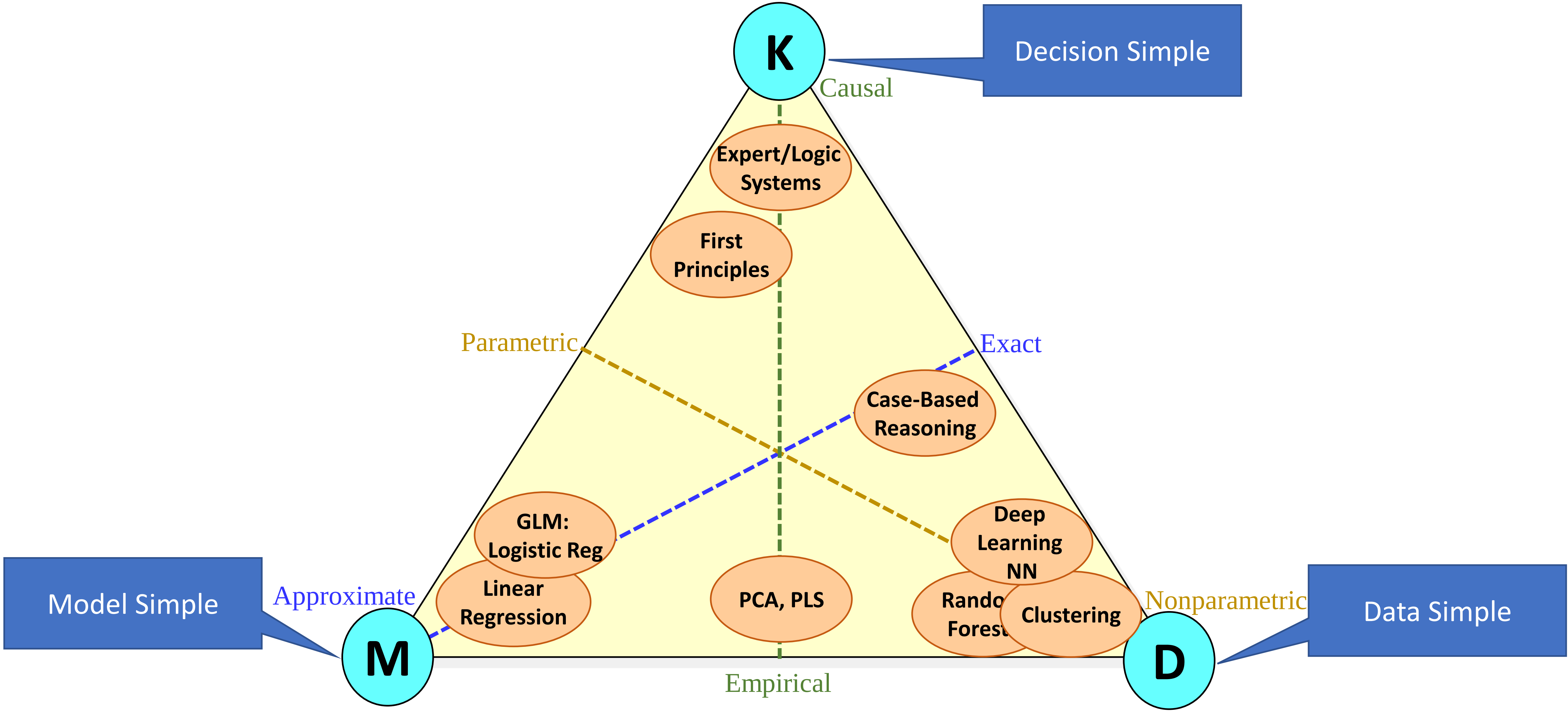
Common Modeling Paradigms

$$V = \frac{4}{3}\pi r^3$$



$$y(x) = \theta_0 + \theta_1 x$$

$$y(x) = \sum_{j=1}^{M} w_j f(x; \Theta)$$


Decision Simple
K
Causal
Expert/Logic Systems
First Principles
Parametric
Exact
Case-Based Reasoning
GLM: Logistic Reg
Linear Regression
PCA, PLS
Deep Learning NN
Rando
Forest
Clustering
Model Simple
Approximate
M
Nonparametric
D
Data Simple
Empirical

Motivation for Bayesian Analysis

Fusion of Data Complex, Model Complex, Decision Complex

"Simple"

- **Data**
  - Single source
  - Single variable types/distribution families
  - Tabular & Ample
    - Non-missing
    - Regular, exchangeable

Homogeneous

- **Model**
  - Observations linked to observations (Modeling the Data)
  - Empirical structure
    - Single-level
    - Acausal
  - Single hypothesis
  - Component-level estimation; Low-level integration

Data-to-Data

- **Decision**
  - Deterministic assumptions
  - Modal/point estimate solutions
  - Predictive inference (What will happen?)
  - Single objective, Static

Deterministic, Predictions

"Complex"

- **Data**
  - Multiple sources
  - Multiple variable types/distribution families
  - Ragged & Sparse
    - Missing
    - Multigranular aggregation

Heterogeneous

- **Model**
  - Latent spaces (Modeling the Domain)
  - Causal structure
    - Multi-level
    - Mechanisms
  - Mixture phenomena/Multi-Hypothesis
  - System-level integration

True-to-True

- **Decision**
  - Reasoning under uncertainty (UQ)
  - Risk analysis
  - Explanatory inference (Why did it happen?)
  - Multi-Objective, Dynamic updating

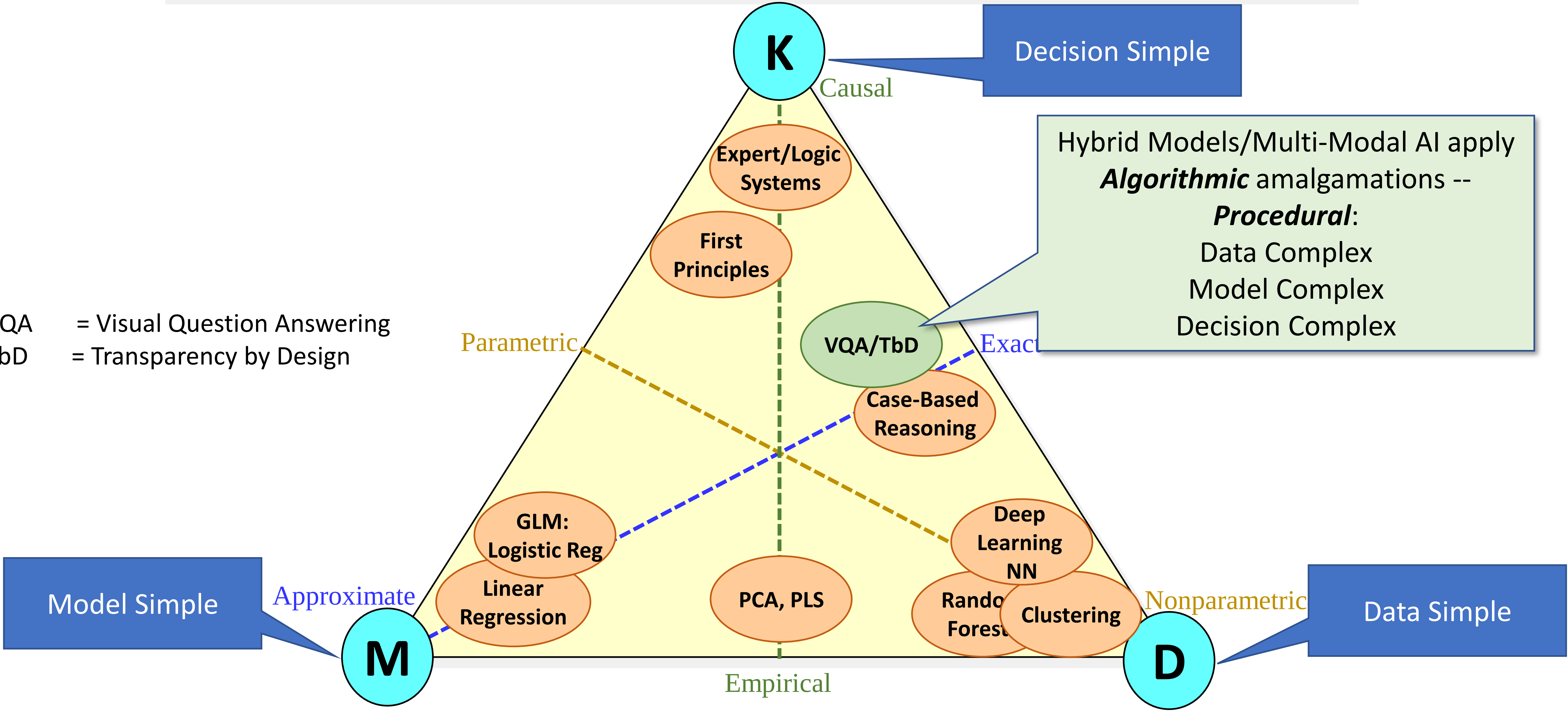
Probabilistic, Explanations

Hybrid Models

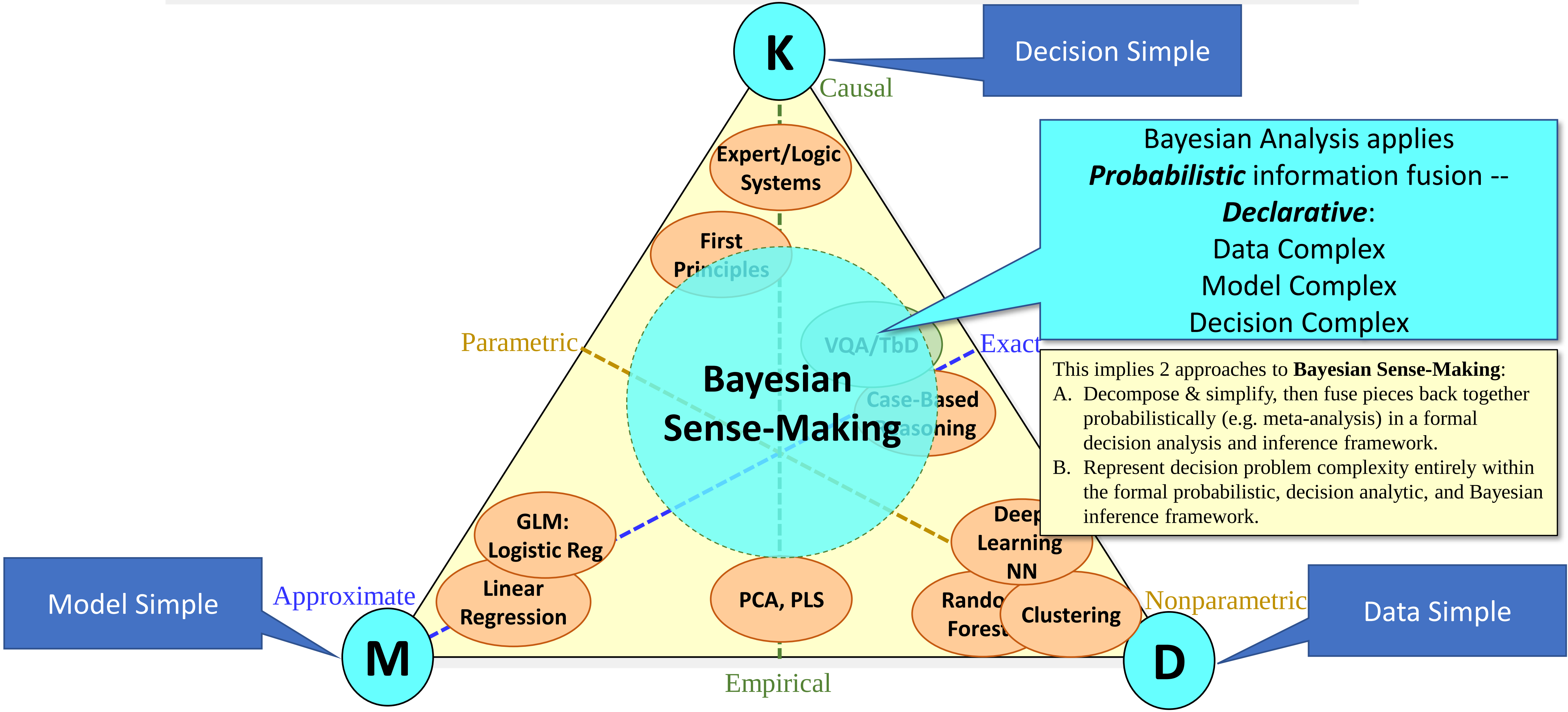
Combine Components to Make Sense in the Face of Real-World Complexity

VQA = Visual Question Answering

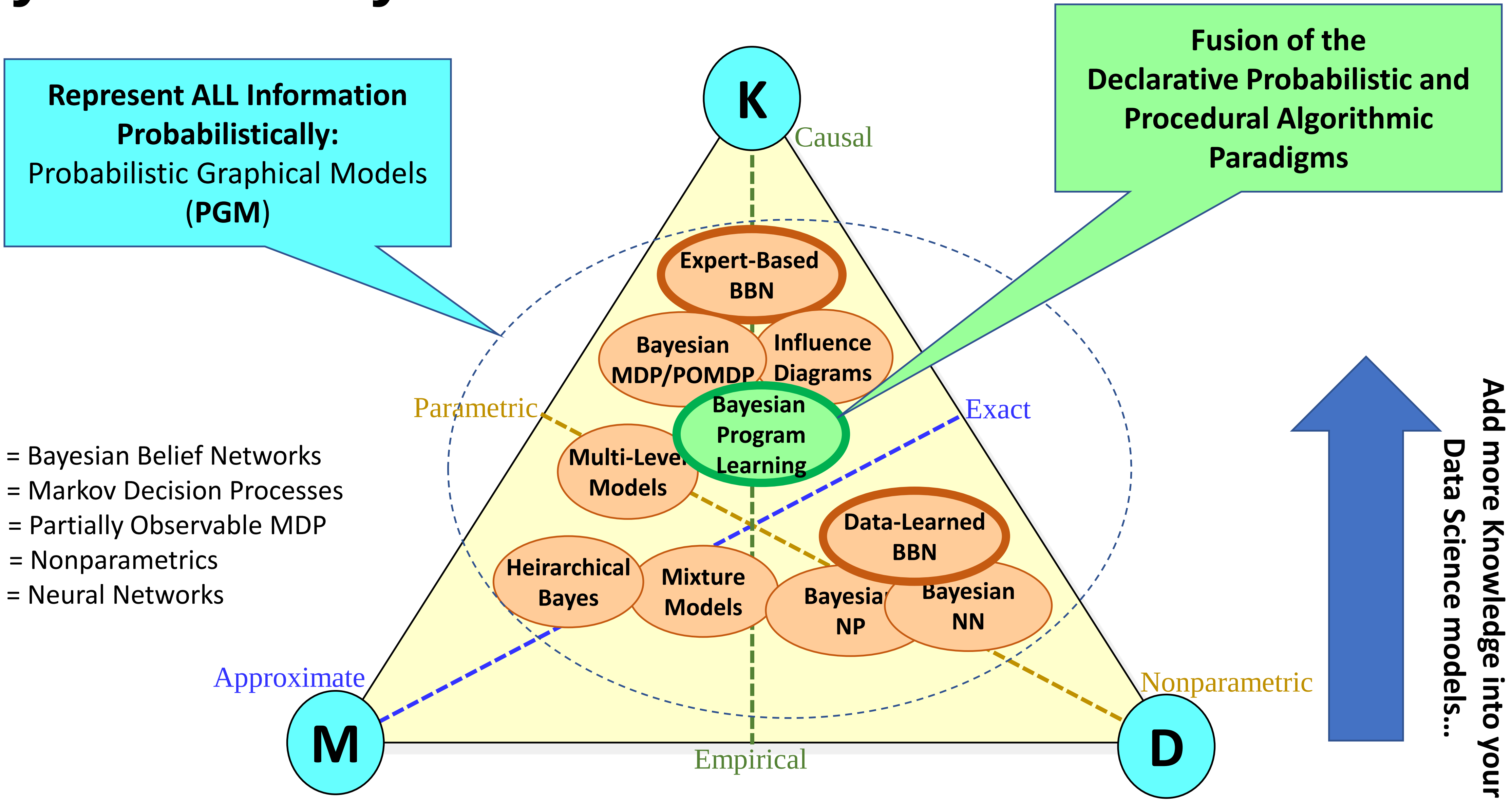
TbD = Transparency by Design



Bayesian Analysis Formalizes Sense-Making



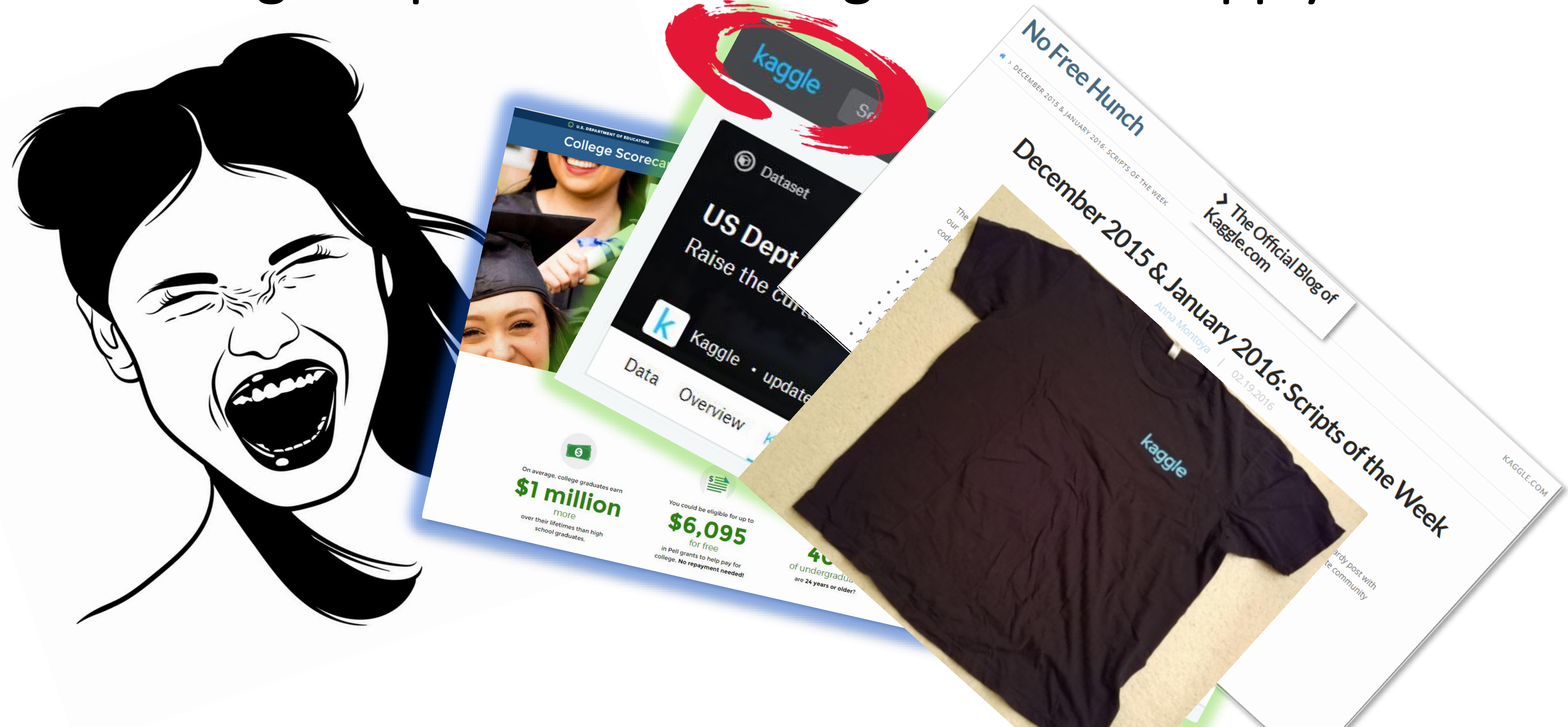
Bayesian Analysis



BBN = Bayesian Belief Networks
MDP = Markov Decision Processes
POMDP = Partially Observable MDP
NP = Nonparametric
NN = Neural Networks

Use Case: Background

A daughter picks which colleges to visit & apply to...



Bayesian Sense-Making: Key Concepts

I. Domain-Relevant Model Structure
- Generative Probabilistic Graphical Models (PGM)
- Latent Spaces, Mixture & Multi-Level Models Causal Structural Models

II. Information Theoretic Principles
- Informative but Least Committal Probability Distributions
- Probability-Based Metrics for Association, Goodness, and Discrepancy

III. Bayesian Inference within Probabilistic Programming Languages
- Model-Based Machine Learning
- Declarative Probabilistic Programs

IV. Explanatory and Causal Inference
- Most Relevant Explanations
- Simulation & Implications of Interventions &Counterfactuals

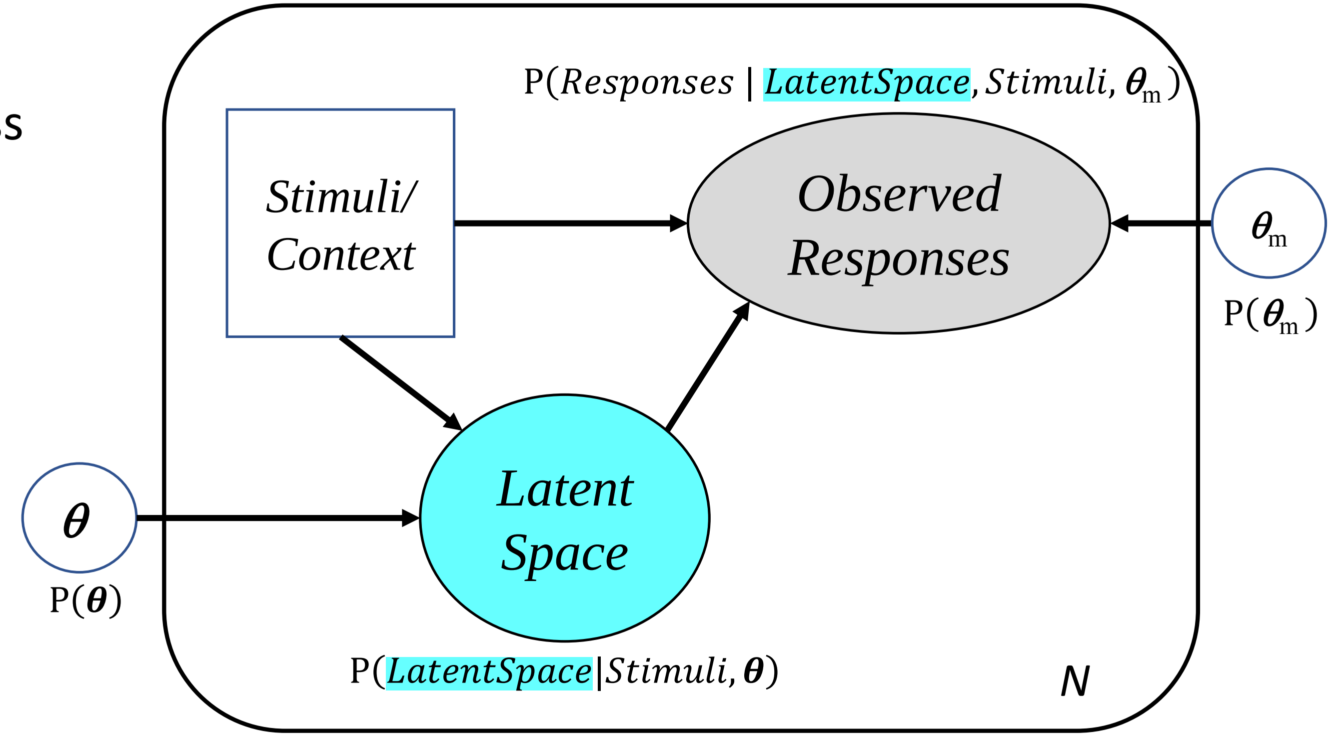
V. Risk Analysis and Decision Analysis
- Uncertainty Quantification (UQ) & Optimization – Maximum Expected Utility
- Value of Information

VI. Optimal Learning
- Sequential & Adaptive Design of Experiments: Optimal Exploration & Exploitation

I. Domain-Relevant Structure

Generative Probabilistic Graphical Models (PGM)

- Declarative specification of data generation process
  - Exploit ***Conditional Independence***
  - ***Probabilistic Programming Languages***
  - ***Model-Based Machine Learning***
- Explicitly represent Latent Spaces
  - **Model the System, *NOT* the Data**



The Latent Space is the link that fuses together observations from many different contexts.

$$P(Responses, LatentSpace, \Theta \mid Stimuli)$$

$$= P(\Theta) \prod^{N} P(Responses \mid LatentSpace, Stimuli, \Theta)\, P(LatentSpace \mid Stimuli, \Theta)$$

$$\rightarrow \prod_{Contexts} P(Responses = D \mid LatentSpace = Z, Stimuli) \prod_j P\big(LatentSpace = Z_j \mid Par(Z_j), Stimuli\big)$$

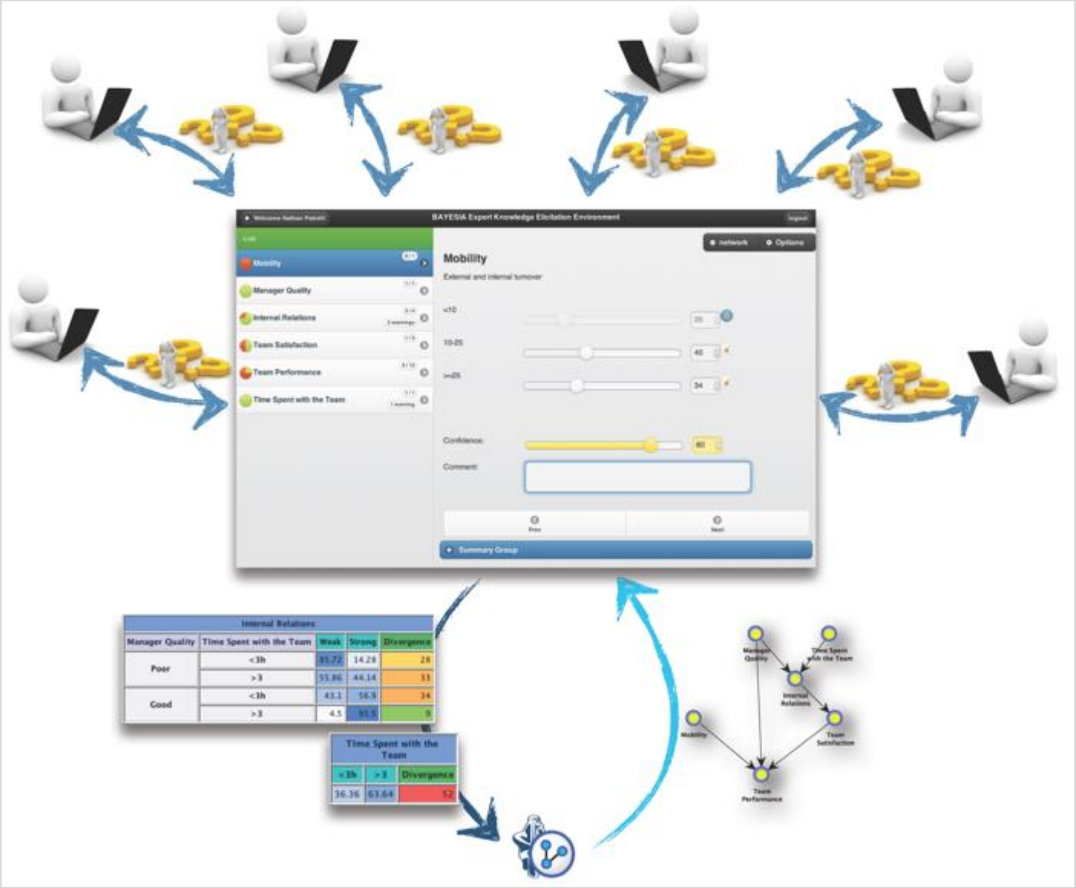
Measurement Models over multiple Contexts — Causal Structural Model

Knowledge Elicitation

Capturing and representing domain knowledge

BEKEE, *Bayesia* Expert Knowledge Elicitation Environment

Seminar: Knowledge Elicitation & Reasoning with Bayesian Networks (video)



Source: Bayesia S.A.S

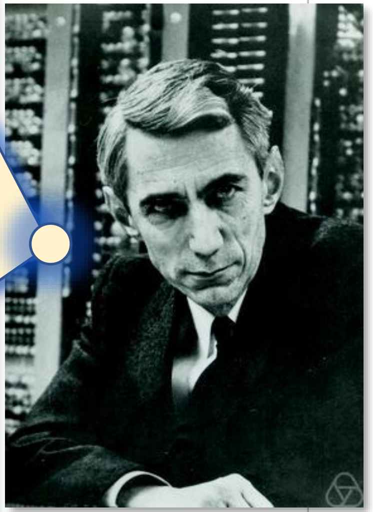
II. Information Theoretic Concepts

Basis for prior distributions & discrepancy/association metrics

- Basics
  - Surprisal, $S(x) = \log(1/P(x))$
  - Entropy , $H(x) = \Sigma_x P(x)S(x)$
  - Information, $I(x|y) = H(x)-H(x|y)$

"My greatest concern was what to call it. **I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.'** When I discussed it with John von Neumann, he had a better idea. **Von Neumann told me, 'You should call it entropy, for two reasons.** In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. **In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage."**

Claude E. Shannon,
Scientific American, 1971, v225, p180



- **Distribution Derivations**
  - **MaxEnt**: Given moments, quantiles, and/or bounds, derive the probability distribution that satisfies these constraints while admitting no other information.

Fitness Function Metrics

- **MDL(p(x,D,Θ))**: Measure of information content of a probabilistic model p(x,D,**Θ**)
- **KLD(p||q)**: Measure of discrepancy between a probability distribution p and a reference distribution q.

Association Metrics

- **I(X,Y)**: Mutual Information is the KLD of the true joint probability distribution $P(X,Y)$ from the joint under independence $P(X)P(Y)$

III. Bayesian Inference in Probabilistic Programming Languages

Declarative Probabilistic Specification Distinct from Inference Algorithms

Model-Based Machine Learning

E.g., Microsoft's C. Bishop ([PDF])

Environment:

BayesiaLab by Bayesia: solely

Bayesian Belief Networks & Influence Diagrams

Probabilistic Programming Languages:

(see https://github.com/topics/bayesian-inference)

***Stan* (esp. R)**, PyMC3 (Python)

Google: TensorFlow Probability;

Uber AI: Pyro

Microsoft: Infer.NET

- Natively encode probability distributions
- Syntax for conditioning upon evidence
- Make available a variety of inference algorithms for any model: e.g. Hamiltonian Monte *Carlo-No-U-Turn* Sampling (HMC-NUTS); Automatic Differentiation Variational Inference (ADVI); and robust optimizers

Stan

```
// Bayesian Plackett
data {
  int N;
  int M;
  int<lower=0> K[N];
  int Kmax;
  int D;
  int Nsource;
  int Ncountry;
  int w[N,Kmax];
  row_vector[D] x[M]
  int country[M];
  int source[N];
  // HYPERPARAMETER
  vector[D] mu0;
  cov_matrix[D] Sigm
}
parameters {
  vector[D] beta;
  vector[M] etaraw;
  vector[Ncountry] b
  vector[Nsource] bs
  real<lower=0> Smag
strengths)
}
transformed paramete
  vector[M] eta;
  vector[Ncountry] b
  vector[Nsource] bs
  vector[M] strength
  real etamean;
  real<lower=0> etas
  etamean = mean(eta
  etastdv =   sd(etaraw);
  eta     = (etaraw    - etamean)/etastdv;
  bCntry = (bCntryraw - etamean)/etastdv;
  bSrc    = (bSrcraw   - etamean)/etastdv;
  for (j in 1:M) {
    strength[j] = Smag * ( eta[j] + x[j] * beta + bCntry[country[j]] );
  }
}
```

```
model {
  vector[M] pinstitution; // conditional probability of
  vector[M] vi;           // strength of each institutio
  // Priors for the coefficients/random effects.
  beta        ~ multi_normal(mu0,Sigma);
  Smag        ~ exponential(1);
  etaraw      ~ normal(0,1);
  bCntryraw ~ normal(0,1);
  bSrcraw     ~ normal(0,1);
  // Compute conditional probabilities of ranks for each
  for ( i in 1:N ) { // for each of N competitions/ranki
    // Strength of each institution adjusted for source.
    vi = exp( strength + Smag * bSrc[source[i]]  ); // (
    for ( r in 1:K[i] ) { // for each position, i.e. ran
      for ( j in 1:M ) {
        pinstitution[j] = 0.0;
      }
      for ( rj in r:K[i] ) { // for institutions with ra
        pinstitution[w[i,rj]] = vi[w[i,rj]];
      }
      pinstitution = pinstitution / sum(pinstitution);
      // Likelihood based on the ordering data: Sample f
      w[i,r] ~ categorical(pinstitution);
    }
  }
}
```


Source: "Bayesian Plackett-Luce Rankings Model", M.L.Thompson, Kaggle.com kernel, 2016, Apache 2.0 license

IV. Explanatory and Causal Inference

Deriving Insights & Reliable Policies by Explaining Why

- Most Relevant (Representative) Explanations: **Generalized Bayes Factor, GBF(H;E)**
  - *Which hypothesis, H, best explains given evidence, E?*
- Implications of Interventions and Decision Policies: Causal inference

- $\mathrm{GBF}(H;E) = \dfrac{\mathrm{P}(Evidence{=}E \mid Hypothesis{=}H)}{\mathrm{P}(Evidence{=}E \mid Hypothesis{\neq}H)}$

  $= \dfrac{\mathrm{Odds}(Hypothesis{=}H \mid Evidence{=}E)}{\mathrm{Odds}(Hypothesis{=}H)}$

- Weight of Evidence, $\mathrm{WE}(H;E) \triangleq \log \dfrac{\mathrm{P}(E \mid {=}H)}{\mathrm{P}(E \mid {\neq}H)}$

where

$$\mathrm{Odds}(X = x) \equiv \frac{\mathrm{P}(X{=}x)}{\mathrm{P}(X{\neq}x)} = \frac{\mathrm{P}(X{=}x)}{1-\mathrm{P}(X{=}x)}$$

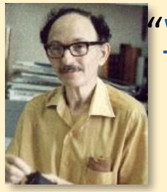
Yuan, C., et al., Most relevant explanations in Bayesian networks, *J. AI Research*, 2011
Good, I.J., Weight of Evidence: A Brief Survey, Bayesian Statistics 2, 1985

IV. Explanatory and Causal Inference

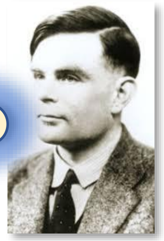
Deriving Insights & Reliable Policies by Explaining Why

"**It is therefore natural to call it *the factor in favour of H provided by E* and this was the name given to it by A.M. Turing in a vital cryptanalytic application in WWII in 1941.** He did not mention Bayes's theorem, with which it is of course closely related, because he always liked to work out everything for himself. **When I said to him that the concept was essentially an application of Bayes's theorem he said 'I suppose so'.**

… **Thus *weight of evidence* is equal to the logarithm of the Bayes factor.**"



"Weight of Evidence: A Brief Survey",
Good, I.J., *Bayesian Statistics 2;*
Bernardo, et al. (eds), 1985



- $\mathrm{GBF}(H;E) = \dfrac{\mathrm{P}(Evidence{=}E \mid Hypothesis{=}H)}{\mathrm{P}(Evidence{=}E \mid Hypothesis{\neq}H)} = \dfrac{\mathrm{Odds}(Hypothesis{=}H \mid Evidence{=}E)}{\mathrm{Odds}(Hypothesis{=}H)}$

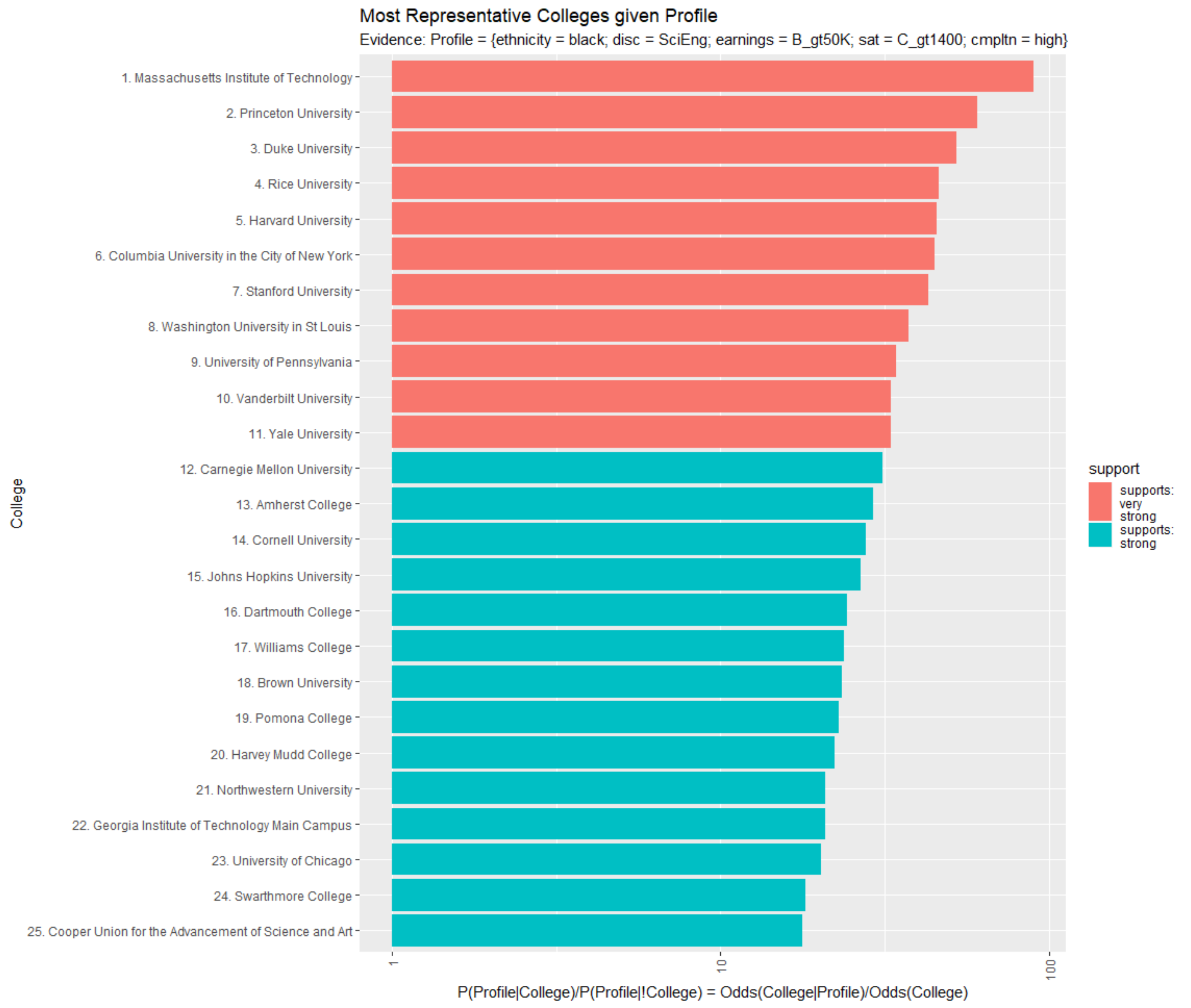
- Weight of Evidence, $\mathrm{WE}(H;E) \triangleq \log \dfrac{\mathrm{P}(E \mid =H)}{\mathrm{P}(E \mid \neq H)}$

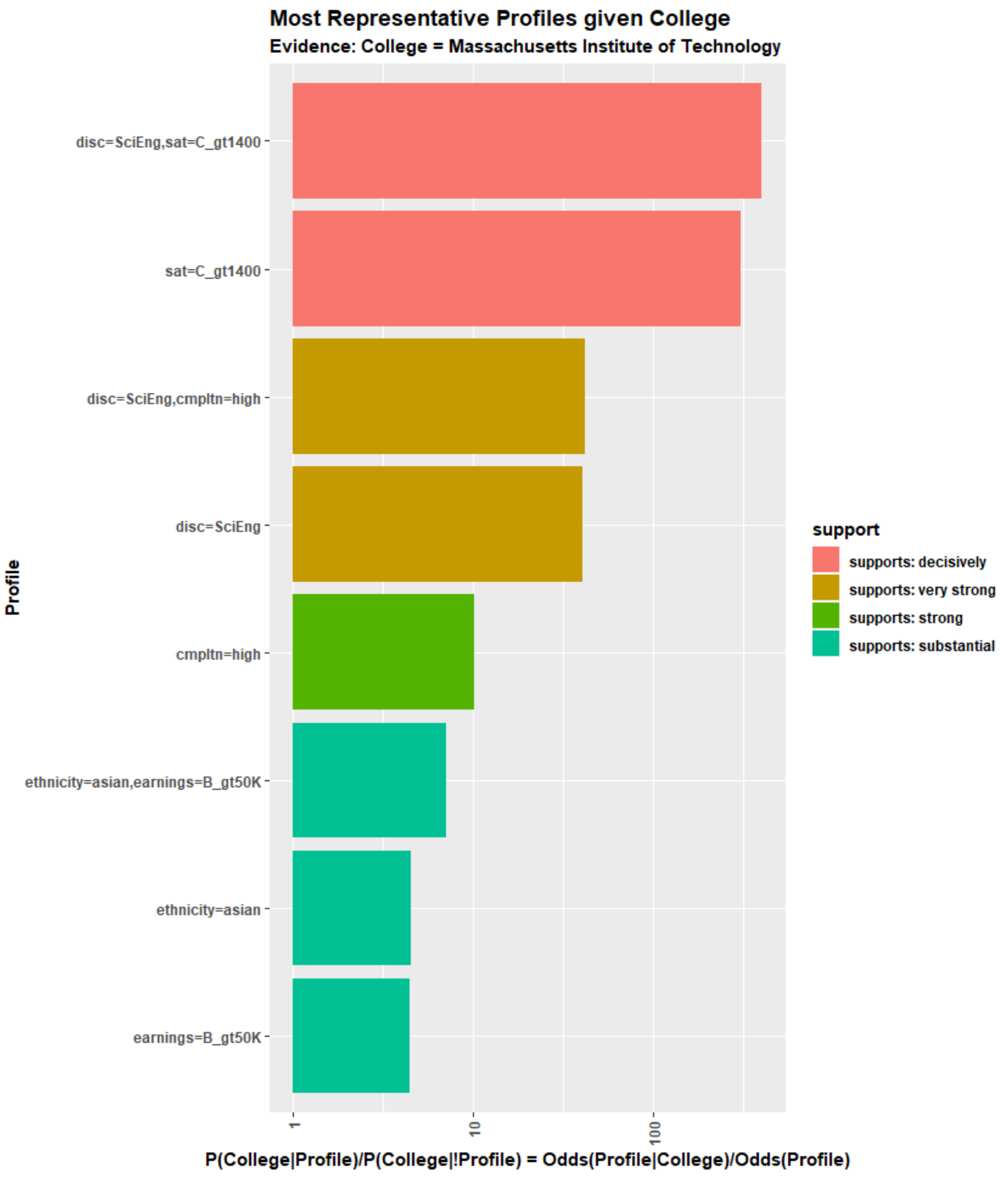
"**…**the terminology of Bayes factors and weights of evidence has more intuitive appeal [than log-likelihood ratio]. This intuitive appeal persists in the general case when the weight of evidence is not the logarithm of a likelihood ratio.
I conjecture that juries, detectives, doctors, and perhaps most educated citizens, will eventually express their judgments in these intuitive terms.", *ibid*

Yuan, C., et al., Most relevant explanations in Bayesian networks, *J. AI Research*, 2011
Good, I.J., Weight of Evidence: A Brief Survey, Bayesian Statistics 2, 1985

Most Relevant Explanations

Ranking Colleges as hypotheses given Student Profiles as evidence, & vice versa

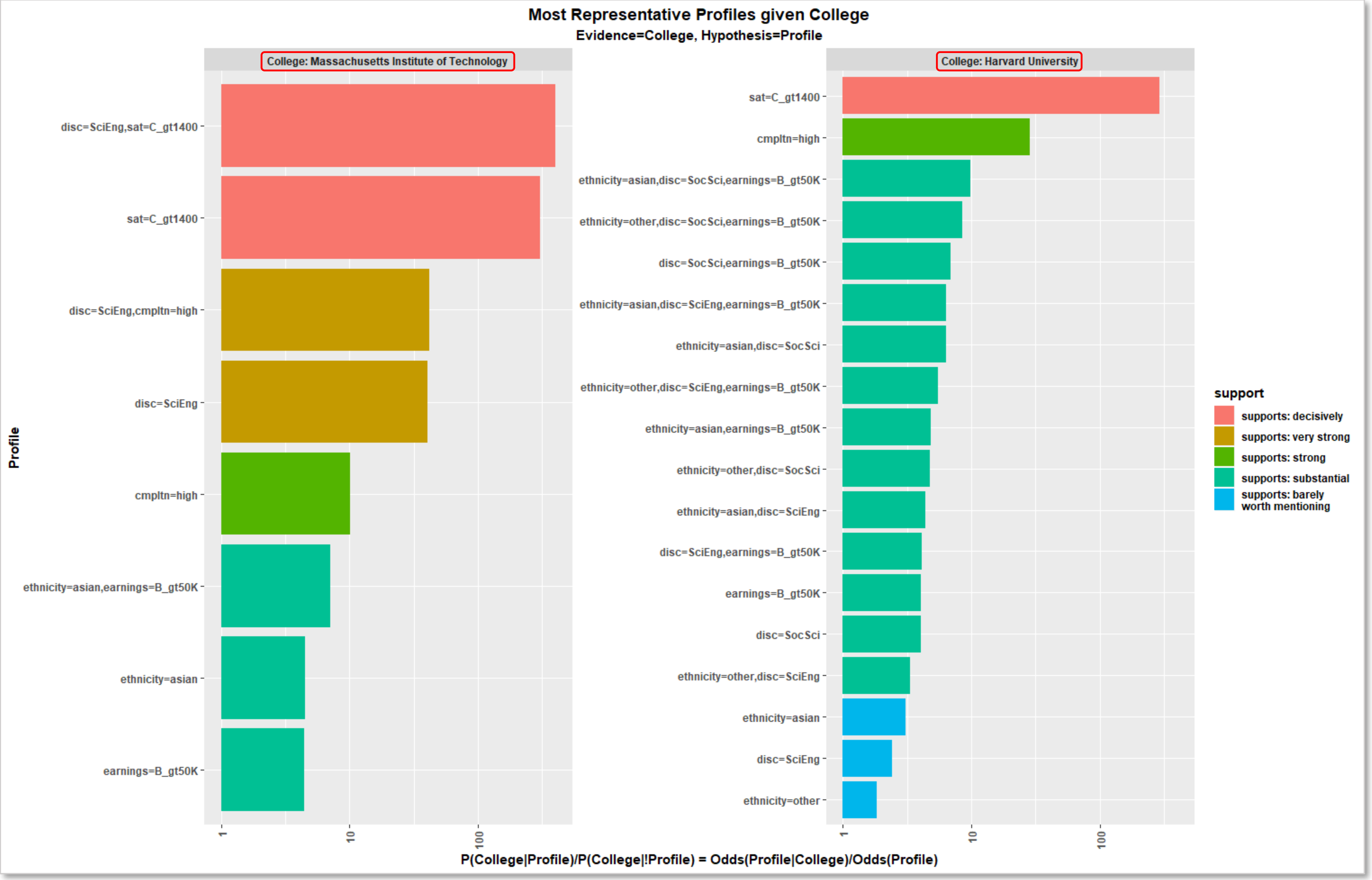




Caveat: Naïve Bayes model

Most Relevant Explanations

Contrasting Colleges as evidence scenarios



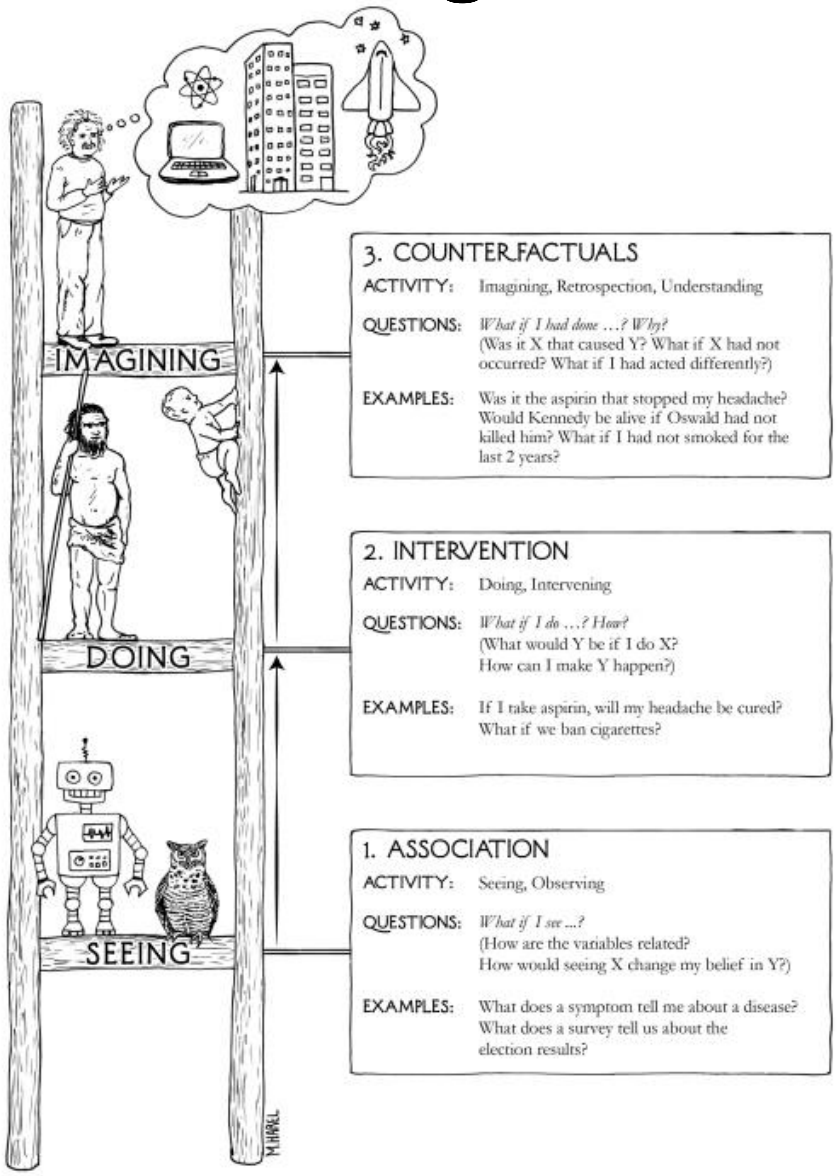
Caveat: Naïve Bayes model

The Book of Why: The New Science of Cause and Effect – Pearl and Mackenzie

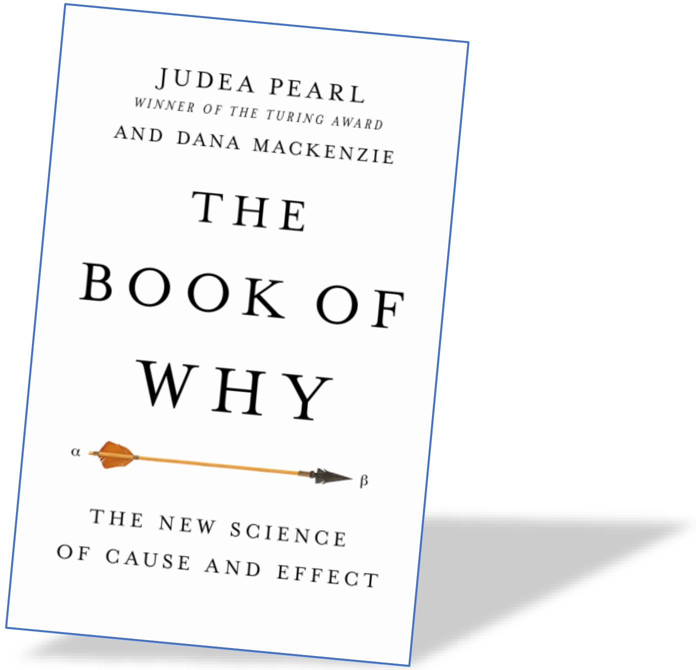
Causal Inference: Climbing Pearl's Causal Ladder

- Motivates imposing a ***Causal Structural Model of the Latent Space***
- Predictive ***Simulations***: Implications of Interventions/Decision Policies
- Pearl, J. and Mackenzie, D., *[The Book of Why: The New Science of Cause and Effect](#)*, 2018
  - Downloadable [Chapter 1](#)

- **DAGitty** [http://www.dagitty.net/] ...
  - "... is a browser-based environment for creating, editing, and analyzing causal models (also known as directed acyclic graphs or causal Bayesian networks). **The focus is on the use of causal diagrams for minimizing bias in empirical studies** in epidemiology and other disciplines."
  - Developed & maintained by [Johannes Textor](#) ([Tumor Immunology Lab](#) and [Institute for Computing and Information Sciences](#), [Radboud University Nijmegen](#))
  - Textor, J., et al., "[Robust causal inference using directed acyclic graphs: the R package 'dagitty'](#)", *Intl. J. Epidemiology*, 45, 6, 1 Dec. 2016, 1887–1894

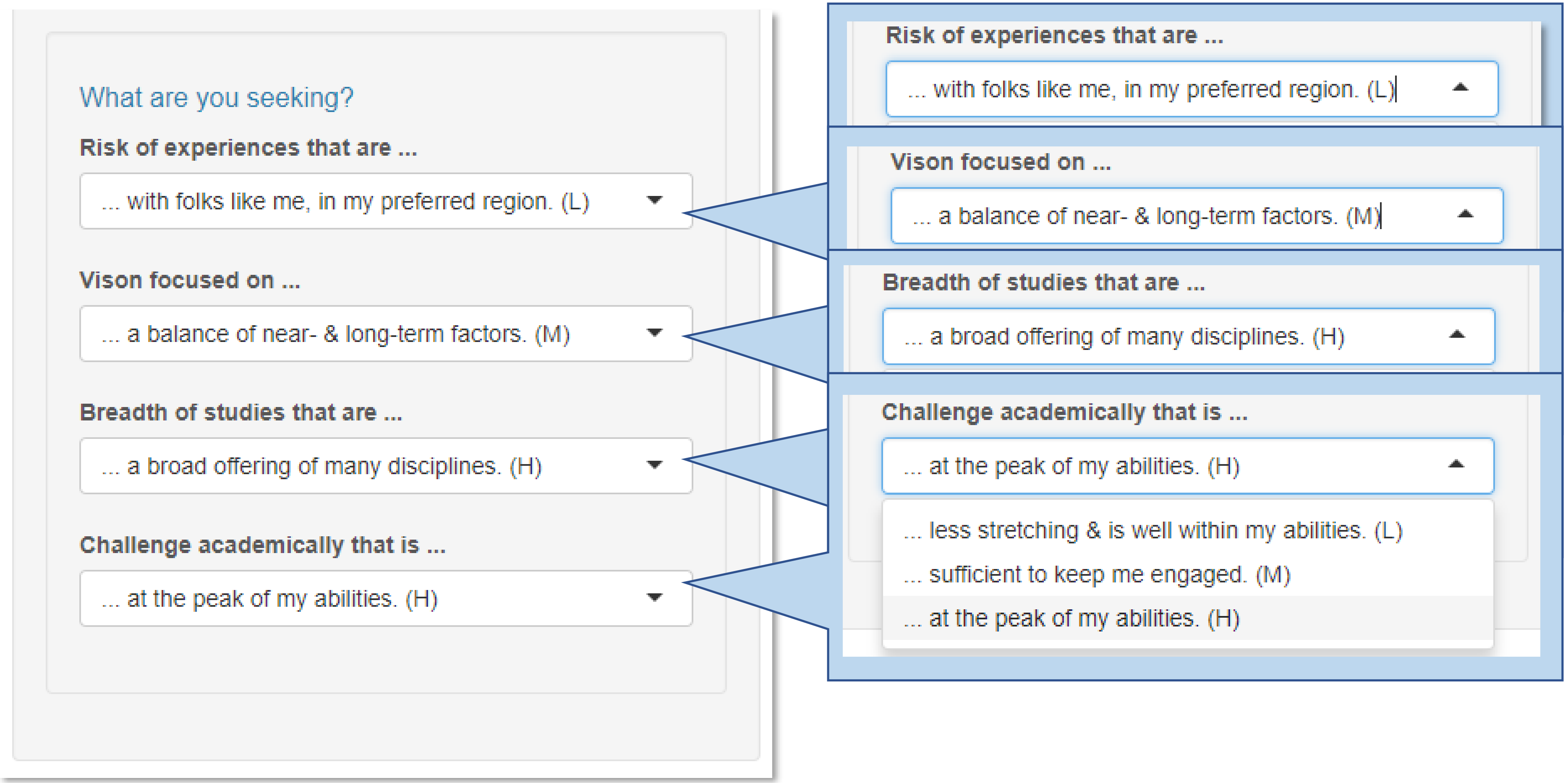


Leads to plausible reasoning about a person's underlying motives …. Hence, we go beyond measured data and into latent constructs.



Latent Motivations of Students

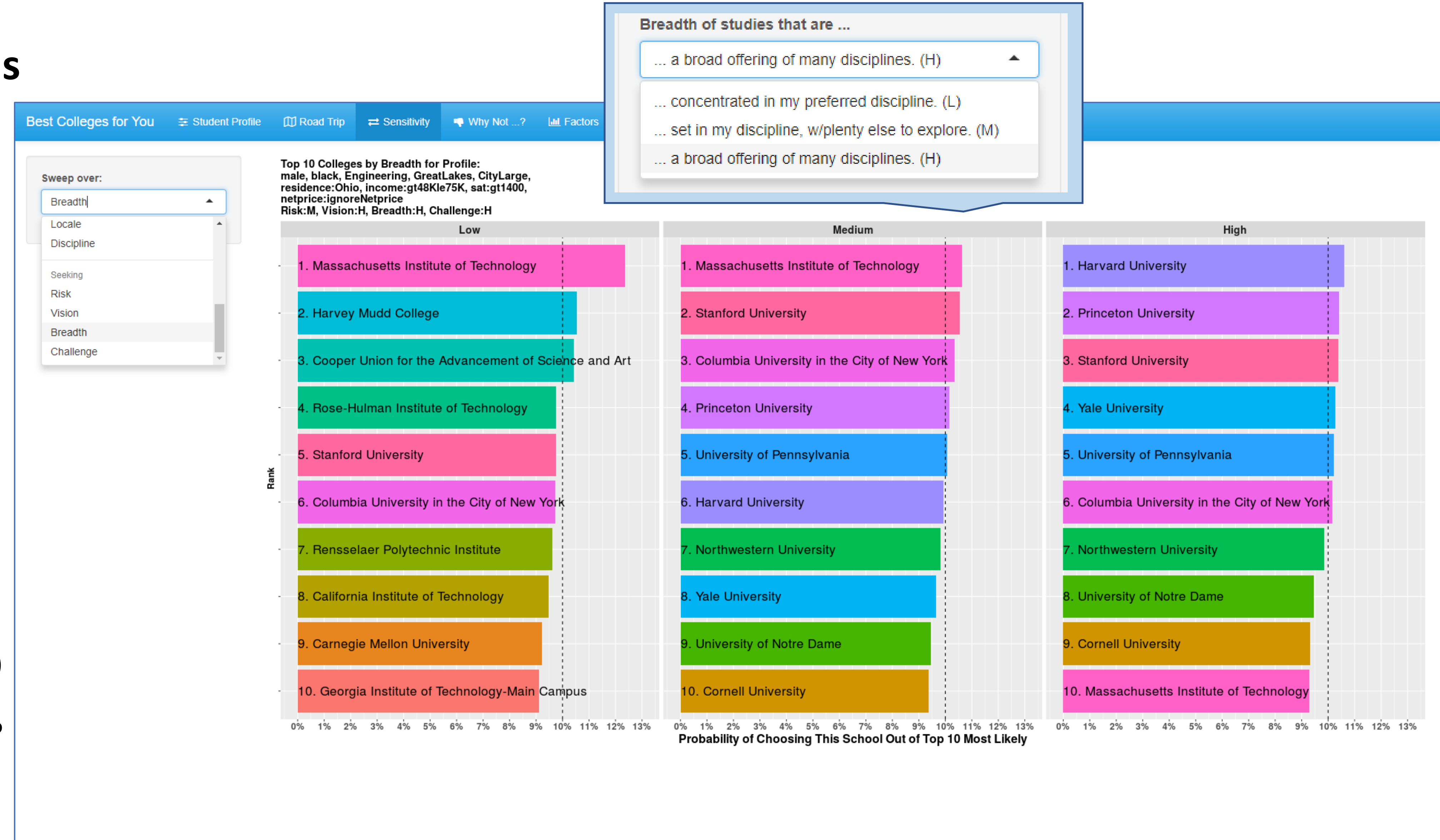
Manifest in behavioral theories & data and expressed attitudes



V. Risk Analysis & Decision Analysis

Quantifying the Uncertainty, Risk & Value of Decisions and Policies

- **Sensitivity Analysis**
- Uncertainty Quantification
- Optimization: Maximum Expected Utility
  - Influence Diagrams
- Learning Optimal Policies
  - Bayesian Reinforcement Learning
    - Markov Decision Processes (MDP)
    - Partially-Observable MDP (POMDP)

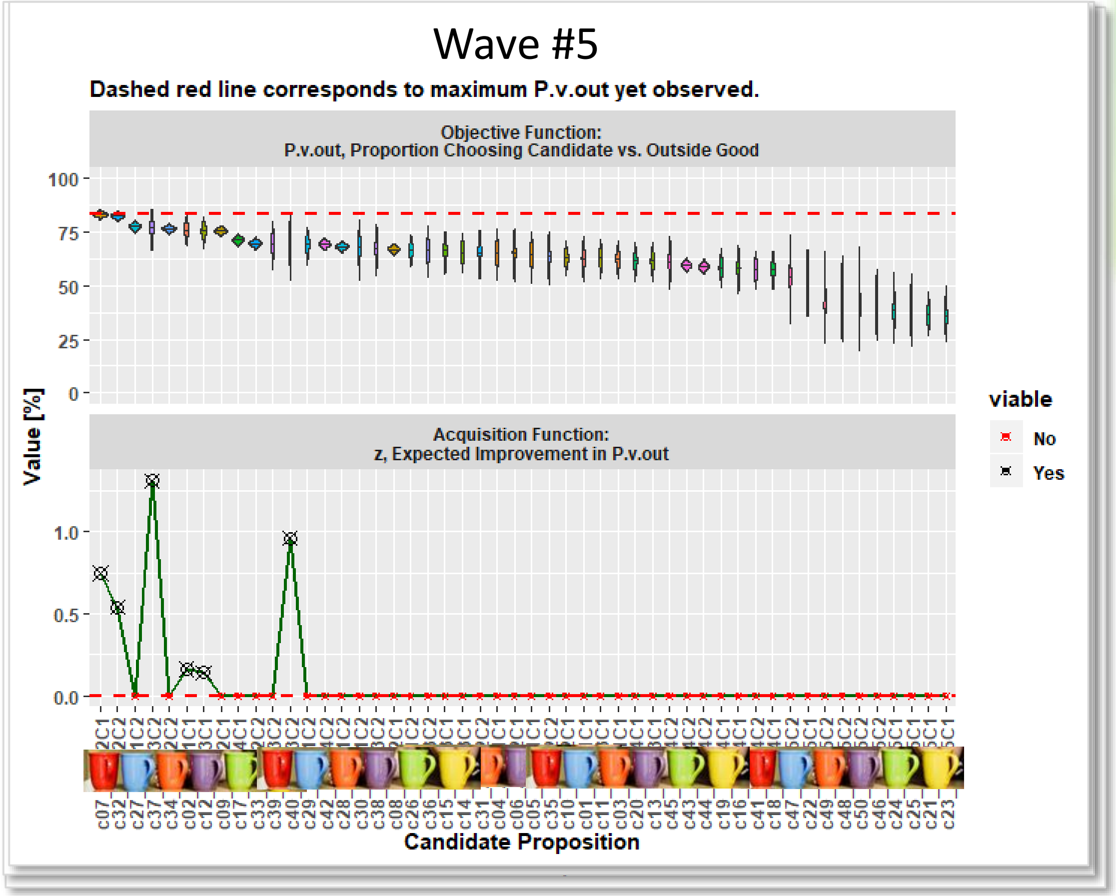


VI. Optimal Learning
Trading Off Exploration & Exploitation

- **Bayesian Optimization** for Adaptive/Sequential Experimental Design and Active Learning
  - **Maximum Expected Improvement** to rank order new stimuli

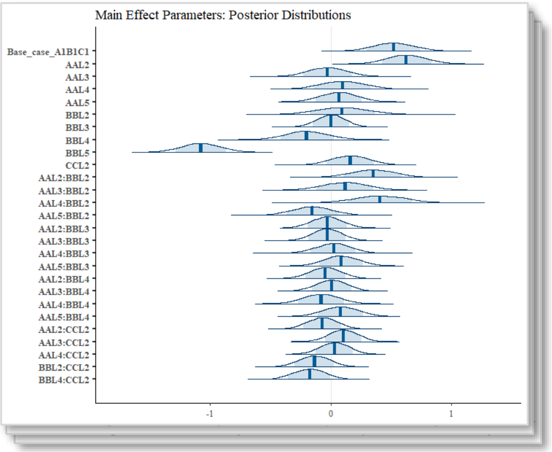
Joo, Mingyu,
Thompson, Michael L.,
Allenby, Greg M.,
Optimal Product Design by Sequential Experiments in High Dimensions,
Management Science (INFORMS),
Oct. 8, 2018

1. Evaluate & Pick Stimuli



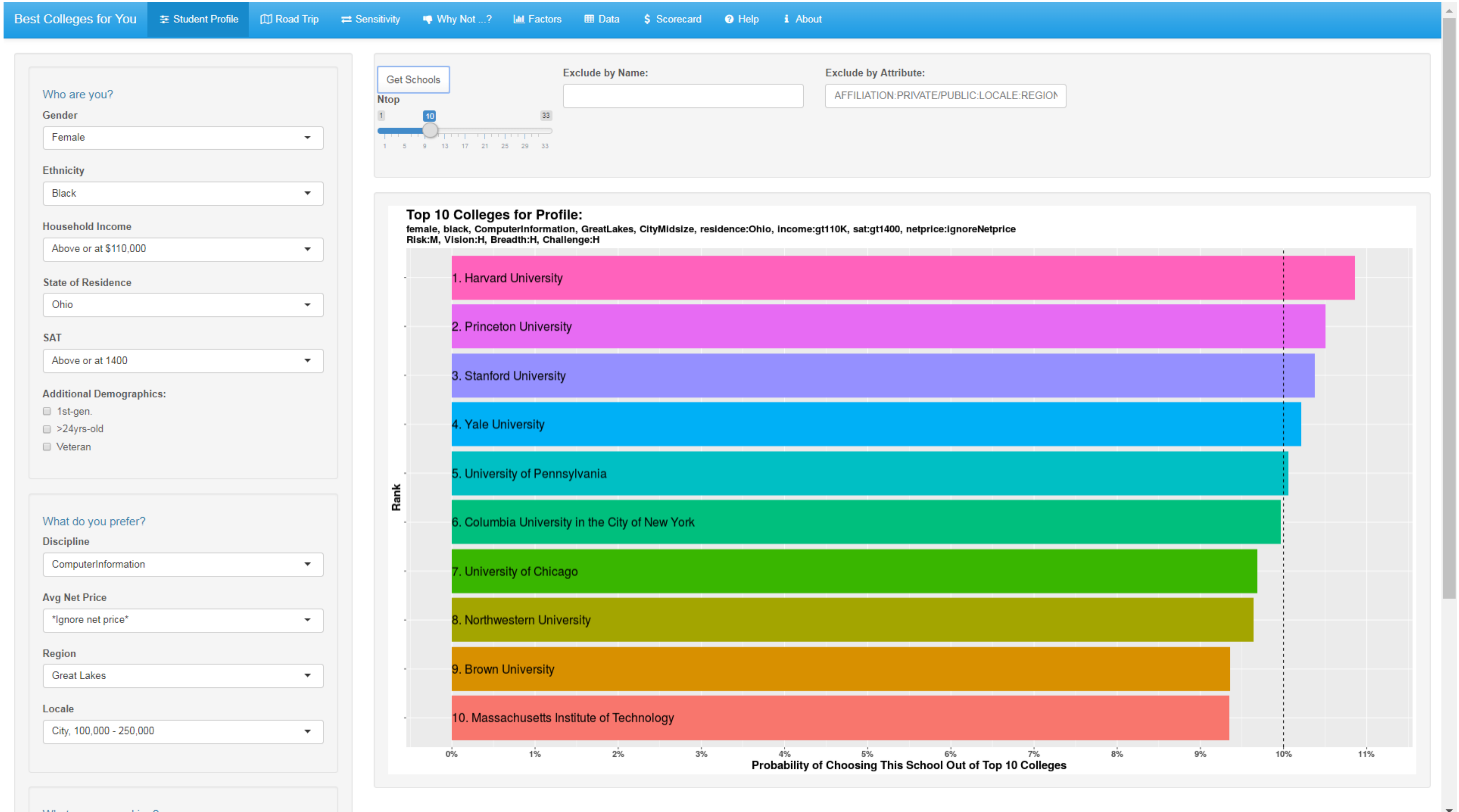
2. Perform Experiment

3. Update Model



And so, the “*Best Colleges for You*” App was born!

https://thompsonml.shinyapps.io/BestCollegeApp/



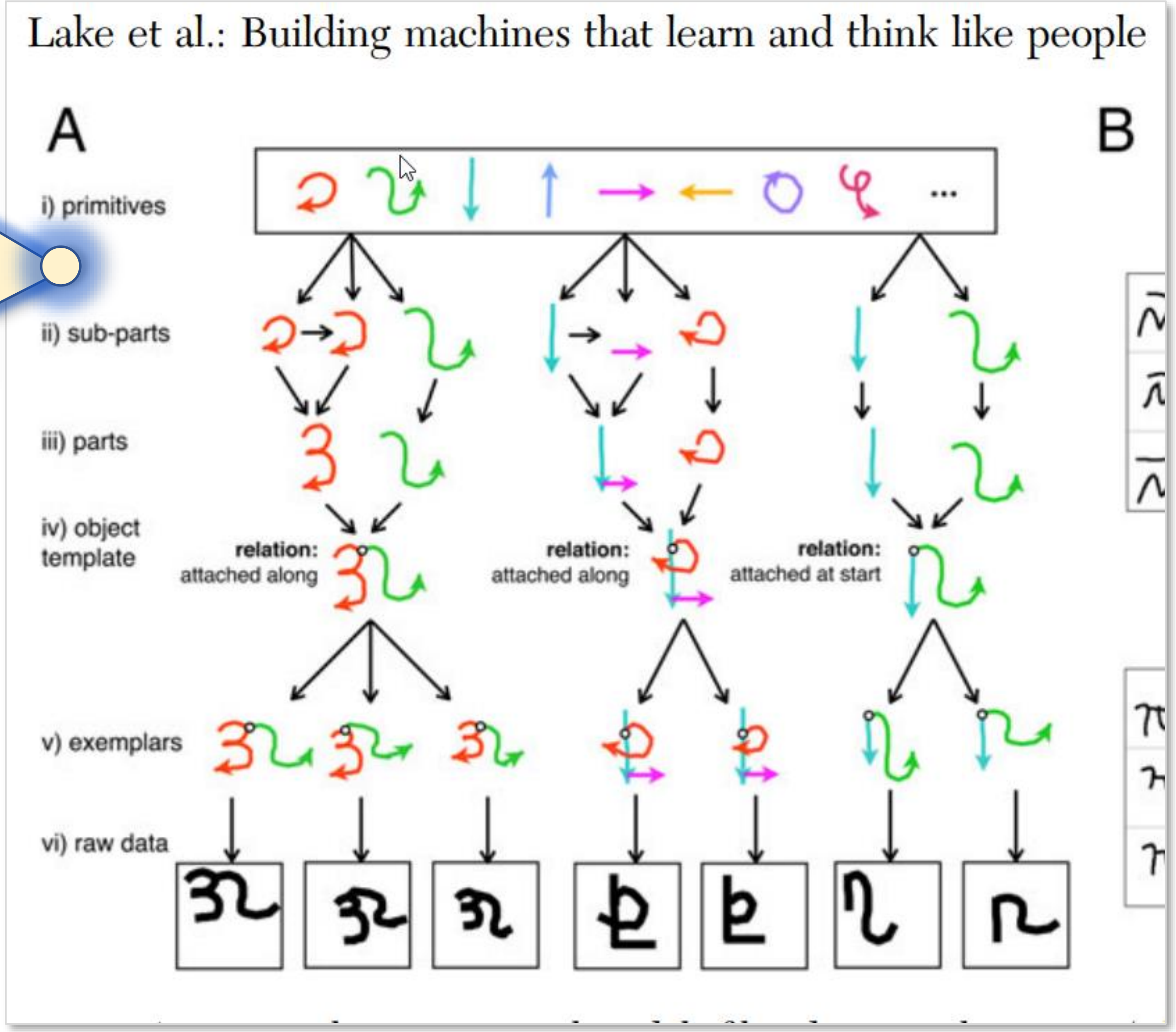
Future Implications

Sense-Making Systems

- **Compositionality**
- **Causality**
- **Learning-to-Learn**

"This trend is an avenue of potential integration of deep learning models with probabilistic models and probabilistic programming: **Training neural networks to help perform probabilistic inference in a generative model or a probabilistic program.**"
"Building machines that learn and think like people",
Lake, Brenden, et al. *Behavioral & Brain Sciences*, 40, E253. 2017

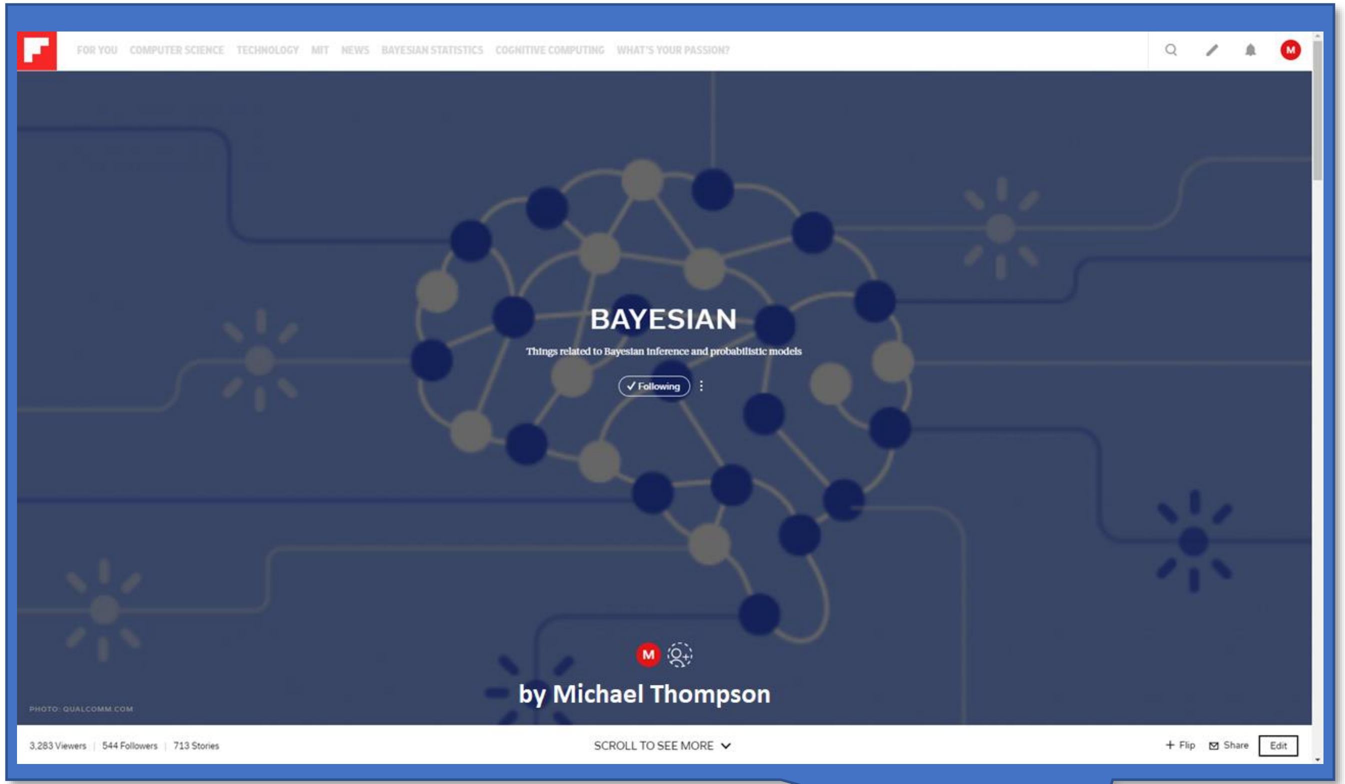
- **One-Shot Learning & General AI**
  - PGM over programs & complex schema (Brenden Lake, NYU; Josh Tenenbaum, MIT)
- **Explanatory AI**
  - DLNN as sensory apparatus fused with PGM answering "Why?" for diagnostic/advisory systems
- **Federated Learning**
  - "Computation at the Edges", e.g., Mobile Phone Deep Learning with Bayesian multi-level models



Future Implications

Organized to Innovate

- **The Procter & Gamble Company**
- **MIT Quest for Intelligence**
- **Notable Business Models**
  - Gamalon (customer intelligence)
  - Zighra (AI-powered continuous authentication): Decentralized AI through Bayesian Learning



Stay informed: Flipboard magazine "Bayesian"

References to Get Started

I. Basics

- Kurt, Will, "Count Bayesie" blog series, "A Guide to Bayesian Statistics", May 2, 2016

II. Domain-Relevant Model Structure

- Tenenbaum, Joshua, et al., How to Grow a Mind: Statistics, Structure, and Abstraction (PDF), *Science*, 2011; lecture video (must-see!)
- Koller, Daphne & Friedman, Nir, Probabilistic Graphical Models, 2009, downloadable excerpt Introduction (must-read!)
- Murphy, Kevin P., Introduction to Graphical Models, 2001
- Daly, Ronan, et al., Learning Bayesian Networks: Approaches and Issues, *The Knowledge Engineering Review*, 2011
- Van Horn, Kevin S., Constructing a Logic of Plausible Inference: A Guide to Cox's Theorem, *Intl. J. Approximate Reasoning*, 2003

III. Information Theoretic Principles

- Jaynes, E.T., The Relation of Bayesian and Maximum Entropy Methods, *Maximum-Entropy and Bayesian Methods in Sci. and Eng.*, 1988
- MacKay, David J.C., *Information Theory, Inference, and Learning Algorithms*, 2003; esp. Ch. 2 "Probability, Entropy, and Inference" and Ch. 3 "More About Inference".

IV. Bayesian Inference within Probabilistic Programming Languages

- Bishop, Christopher, Model-Based Machine Learning, *Phil. Trans. Roy. Soc. A*, 2013
- Davidson-Pilon, Cameron, Probabilistic Programming and Bayesian Methods for Hackers (PyMC3), 2015
- Conrady, Stefan & Jouffe, Lionel (Bayesia S.A.S), Bayesian Networks and BayesiaLab: A Practical Introduction for Researchers, 2015
  - Ch. 8, Probabilistic Structural Equation Models with Bayesian Networks for Key Drivers Analysis and Product Optimization, 2015
- Stan Development Team, Modeling Language User's Guide and Reference Manual, Version 2.17.0, 2018; esp. Section III. Example Models

V. Explanatory and Causal Inference

- Yuan, Changhe, et al., Most relevant explanations in Bayesian networks, *J. Artificial Intelligence Research*, 2011
- Pacer, Michael, et al., Evaluating computational models of explanation using human judgment, *Proc. 29th Conf. on Uncertainty in AI (UAI2013)*, 2013
- Rydall, Michael D. and Bramson, Aaron L., *Inference and Intervention: Causal Models for Business Analysis*, 2013
- Pearl, Judea and Mackenzie, Dana, *The Book of Why: The New Science of Cause & Effect*, 2018; downloadable excerpts Introduction, Chapter 1, Chapter 2

VI. Risk Analysis and Decision Analysis

- Fenton, Norman and Neil, Martin, Managing Risk in the Real World: Applications of Bayesian Networks, 2007
- Barry Matthew and Horvitz, Eric, Vista Goes Online: Decision Analytic Systems for Real-Time Decision-Making in Mission Control, 1994

VII. Optimal Learning

- Shahrari, Bobak, et al., Taking the Human Out of the Loop: A Review of Bayesian Optimization, *Proc. IEEE*, 2016
- Joo, Mingyu, et al., Optimal Product Design by Sequential Experiments in High Dimensions, *Management Science (INFORMS)*, Oct. 8, 2018