Machine Learning in Industry

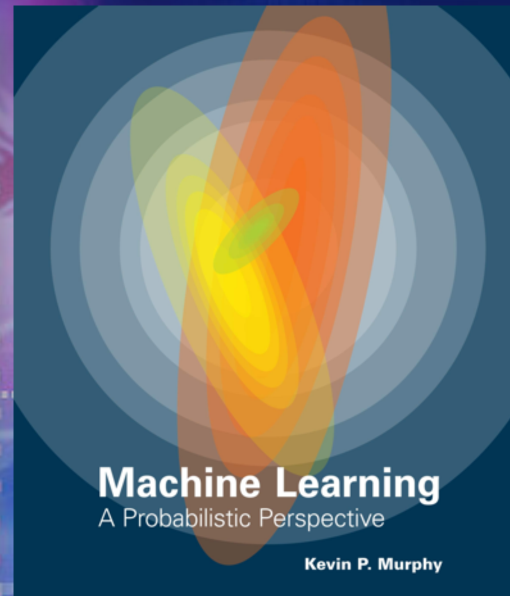
Ralf Herbrich
Amazon

Overview

- Theory
  - **Inference in Factor Graphs**
  - **Approximate Message Passing**
- Applications @ Microsoft
  - **TrueSkill: Gamer Rating and Matchmaking**
  - **TrueSkill Through Time: History of Chess**
  - **Click-Through Rate Prediction in Online Advertising**
  - **Matchbox: Recommendation Systems**
- Applications @ Amazon

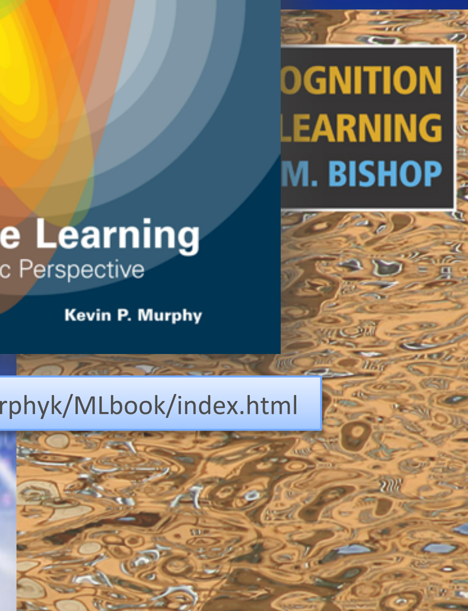
Background Material

Coursera



http://www.coursera.org

http://www.cs.ubc.ca/~murphyk/MLbook/index.html



http://www.cs.ucl.ac.uk/staff/d.barber/brml/

http://research.microsoft.com/en-us/um/people/cmbishop/PRML/index.htm

Overview

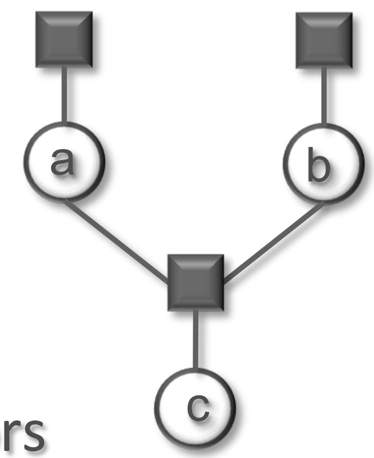
- **Theory**
  - Inference in Factor Graphs
  - Approximate Message Passing
- Applications
  - TrueSkill: Gamer Rating and Matchmaking
  - TrueSkill Through Time: History of Chess
  - Click-Through Rate Prediction in Online Advertising
  - Matchbox: Recommendation Systems
- Future Applications

Graphical Models

- **Definition:** Graphical representation of joint probability distribution
  - Nodes: ○ = Variables
  - Edges: Relationship between variables
- **Variables**:
  - Observed Variables: Data
  - Unobserved Variables: 'Causes' + Temporary/Latent
- **Key Questions**:
  - (Conditional) *Dependency*: $p(a, b|c) \stackrel{?}{=} p(a|c) \cdot p(b|c)$
  - *Inference*/Marginalisation: $p(a, b) = \sum_{c} p(a, b, c)$

Factor Graphs

- **Definition:** Graphical representation of product structure of a function (Wiberg, 1996)
  - Nodes: ■ = Factors ○ = Variables
  - Edges: Dependencies of factors on variables.



- **Semantic:**

$$p(\mathbf{x}) \;=\; \prod_f f\left(\mathbf{x}_{V(f)}\right)$$

  - Local variable dependency of factors

$$p(a,b,c) \;=\; f_1(a) \cdot f_2(b) \cdot f_3(a,b,c)$$

Factor Graphs and Bayes' Law

- Bayes' law

$$p(\mathbf{s}|y) \;\propto\; p(y|\mathbf{s}) \cdot p(\mathbf{s})$$

- Factorising prior

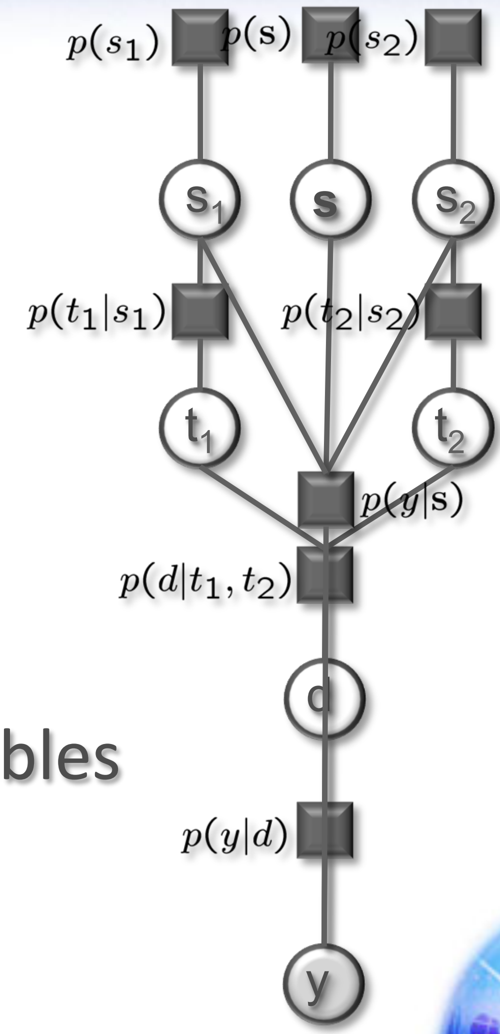
$$p(\mathbf{s}) \;=\; p(s_1) \cdot p(s_2)$$

- Factorising likelihood

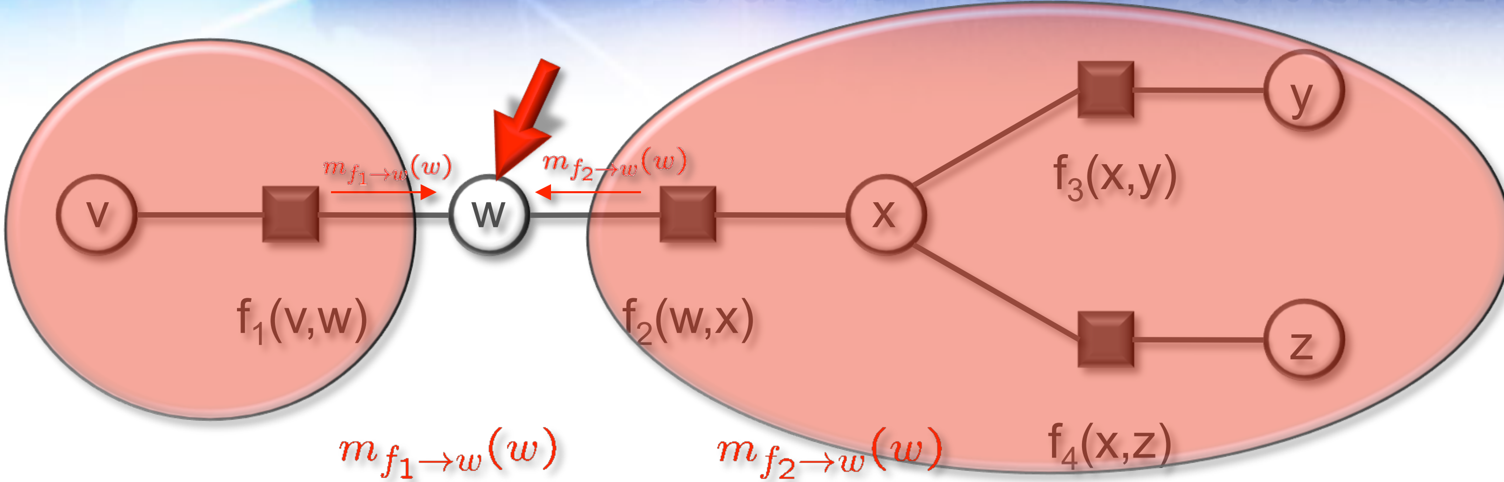
$$p(y, \mathbf{t}, d|\mathbf{s}) \;=\; \prod_i p(t_i|s_i) \cdot p(d|t_1, t_2) \cdot p(y|d)$$

- Inference: Sum out latent variables

$$p(y|\mathbf{s}) \;=\; \sum_{\mathbf{t}} \sum_{d} p(y, \mathbf{t}, d|\mathbf{s})$$



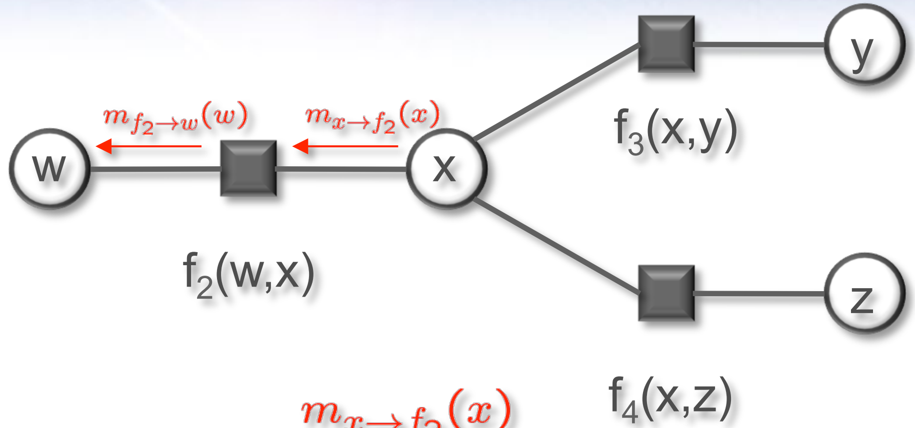
Factor Trees: Separation



$$p(w) = \overbrace{\left[\sum_{v} f_1(v,w)\right]}^{m_{f_1 \to w}(w)} \overbrace{\left[\sum_{x}\sum_{y}\sum_{z} f_2(w,x) f_3(x,y) f_4(x,z)\right]}^{m_{f_2 \to w}(w)}$$

Observation: Sum of products becomes product of sums of all messages from neighbouring factors to variable!

Messages: From Factors To Variables

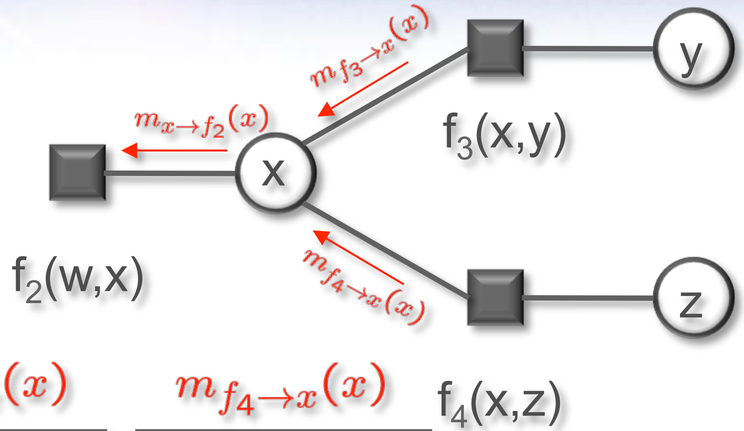


$$m_{f_2\to w}(w) = \sum_x \sum_y \sum_z f_2(w,x) f_3(x,y) f_4(x,z)$$

$$m_{f_2\to w}(w) = \sum_x f_2(w,x) \overbrace{\boxed{\left[\sum_y \sum_z f_3(x,y) f_4(x,z)\right]}}^{m_{x\to f_2}(x)}$$

Observation: Factors only need to sum out all their local variables!

Messages: From Variables To Factors



$$m_{x \to f_2}(x) = \overbrace{\left[\sum_y f_3(x,y)\right]}^{m_{f_3 \to x}(x)} \cdot \overbrace{\left[\sum_z f_4(x,z)\right]}^{m_{f_4 \to x}(x)}$$

Observation: Variables pass on the product of all incoming messages!

The Sum-Product Algorithm

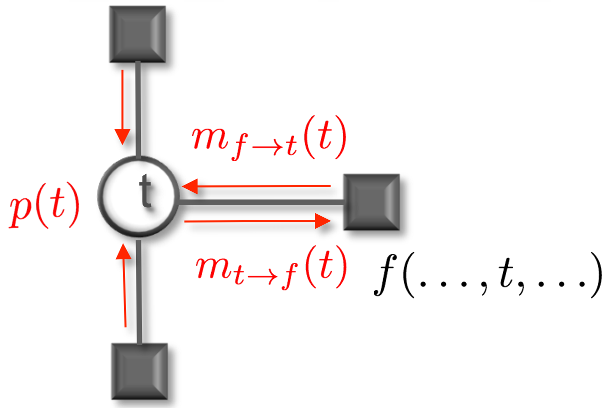
- Three update equations (Aji & McEliece, 1997)

$$
\begin{aligned}
p(t) &= \prod_{f \in F_t} m_{f \to t}(t) \\
m_{f \to t_1}(t_1) &= \sum_{t_2} \sum_{t_3} \cdots \sum_{t_n} f(t_1, t_2, t_3, \ldots) \prod_{i>1} m_{t_i \to f}(t_i) \\
m_{t \to f}(t) &= \prod_{f_j \in F_t \setminus \{f\}} m_{f_j \to t}(t)
\end{aligned}
$$

- Update equations can be directly derived from the distributive law.
- Calculate all marginals at the same time!
- Only need to pass messages twice along each edge!

Practical Considerations II

- **Redundant computations:**



$$p(t) = \prod_{f \in F_t} m_{f\to t}(t)$$

$$m_{t\to f}(t) = \prod_{f_j \in F_t \setminus \{f\}} m_{f_j \to t}(t)$$

$$\Rightarrow \quad p(t) = m_{t\to f}(t) \cdot m_{f\to t}(t)$$

- **Caching**: Only store $p(t)$ and $m_{f\to t}(t)$, then

$$m_{t\to f}(t) = \frac{p(t)}{m_{f\to t}(t)}$$

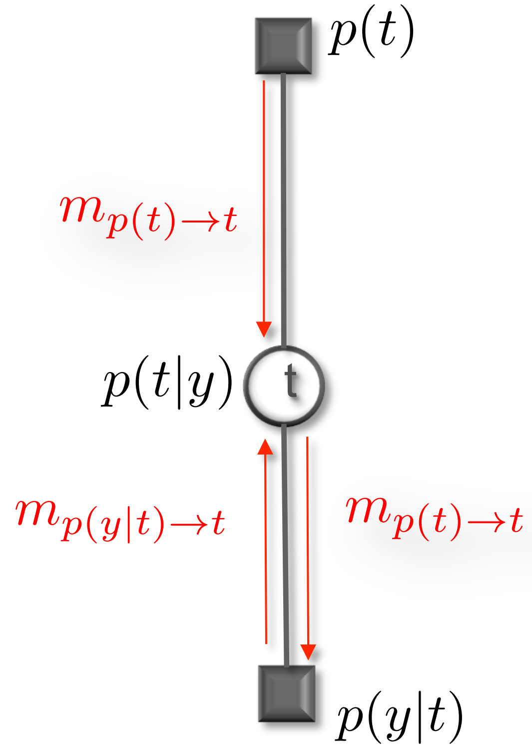
A Bayesian Interpretation

- **Recall Bayes' Law:**

$$p(t|y) \propto p(y|t) \cdot p(t)$$

- **Prior and Data Messages:**

$$\begin{aligned} p(t|y) &\propto m_{p(y|t)\to t} \cdot m_{p(t)\to t} \\ &\propto m_{p(y|t)\to t} \cdot m_{t\to p(y|t)} \end{aligned}$$



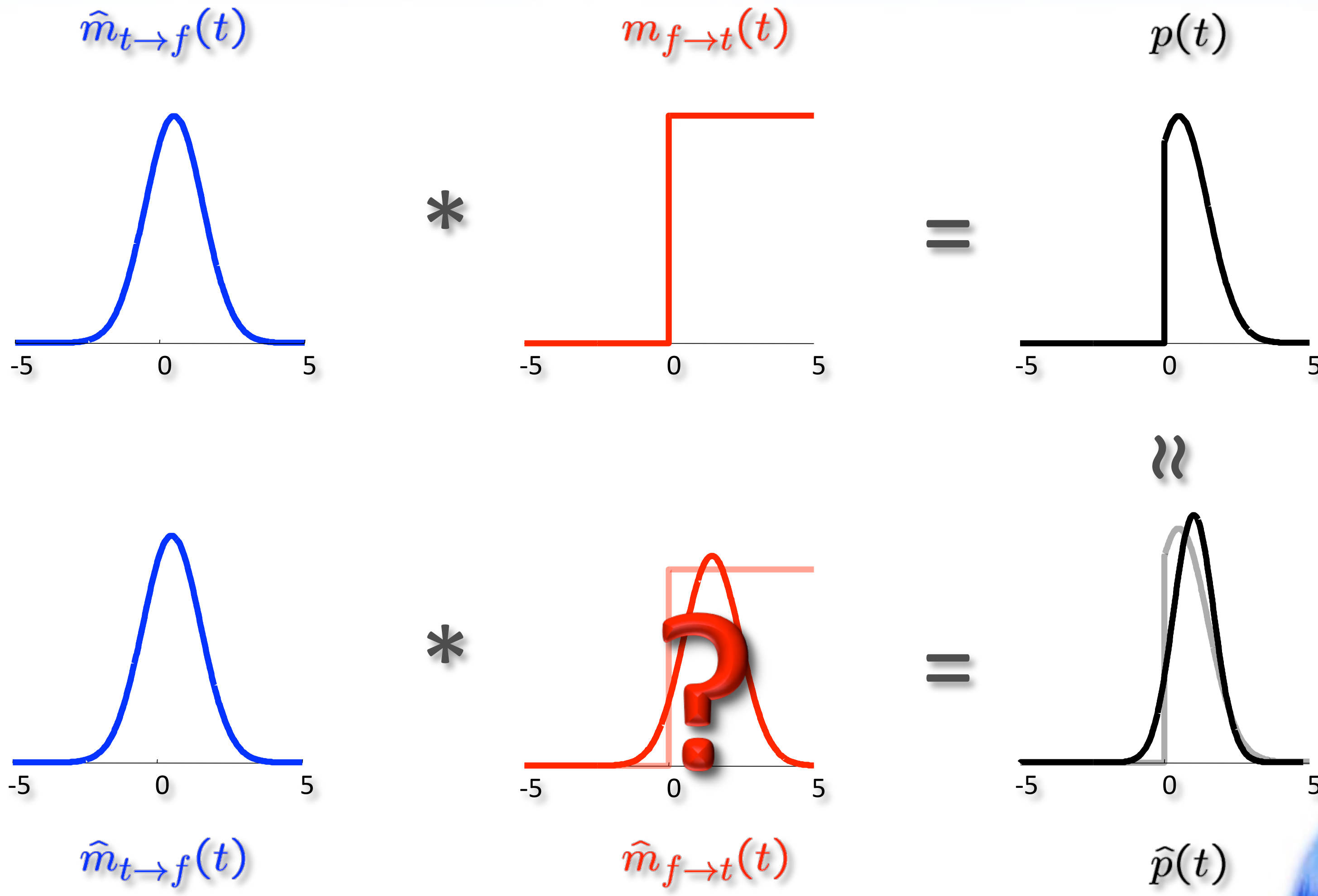
Message passing is separating the likelihood and prior into outgoing and incoming message!

Approximate Message Passing

- **Problem:** The exact messages from factors to variables may not be closed under products.
- **Solution:** Approximate *each* marginal as well as possible in using a divergence measure on beliefs.
- **General Idea:** Leave-one out approximation

$$
\begin{aligned}
\hat{p}(t) &= \text{argmin}_{\hat{p}}, D\left[\overset{p(t)}{\boxed{m_{f\to t}\cdot \hat{m}_{t\to f}}}, \hat{p}\right] \\
\hat{m}_{f\to t}(t) &= \frac{\hat{p}(t)}{\hat{m}_{t\to f}(t)}
\end{aligned}
$$

Approximate Message Passing



Divergence Measures

- **Kullback-Leibler Divergence:** Expected log-odd ratio between two distributions:

$$\mathsf{KL}(p,q) := \sum_t p(t) \log \left(\frac{p(t)}{q(t)} \right)$$

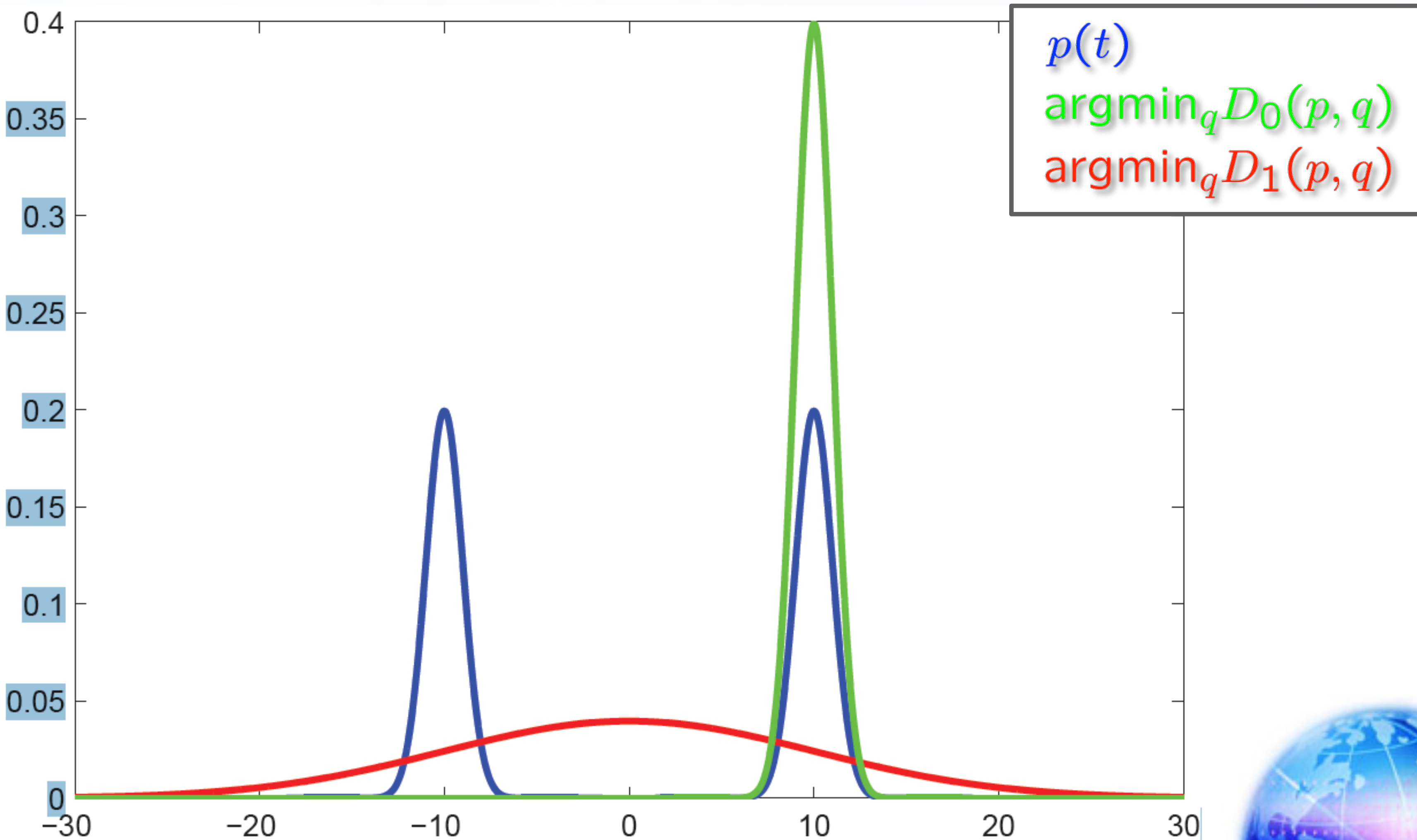
- **Minimizer for Exponential Families:** Matching the moments of the distribution $p(t)$!
- **General α-Divergence:**

$$D_\alpha(p,q) := \frac{1 - \sum_t \frac{p^{\alpha-1}(t)}{q^{\alpha-1}(t)}}{\alpha(1-\alpha)}$$

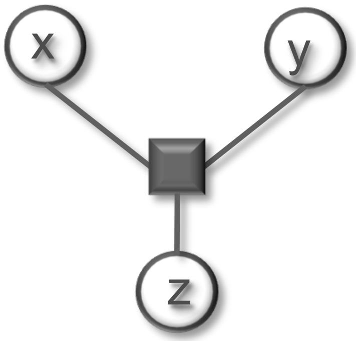
- **Special Cases:**

$$\begin{aligned} D_0(p,q) &= \mathsf{KL}(q,p) \\ D_1(p,q) &= \mathsf{KL}(p,q) \end{aligned}$$

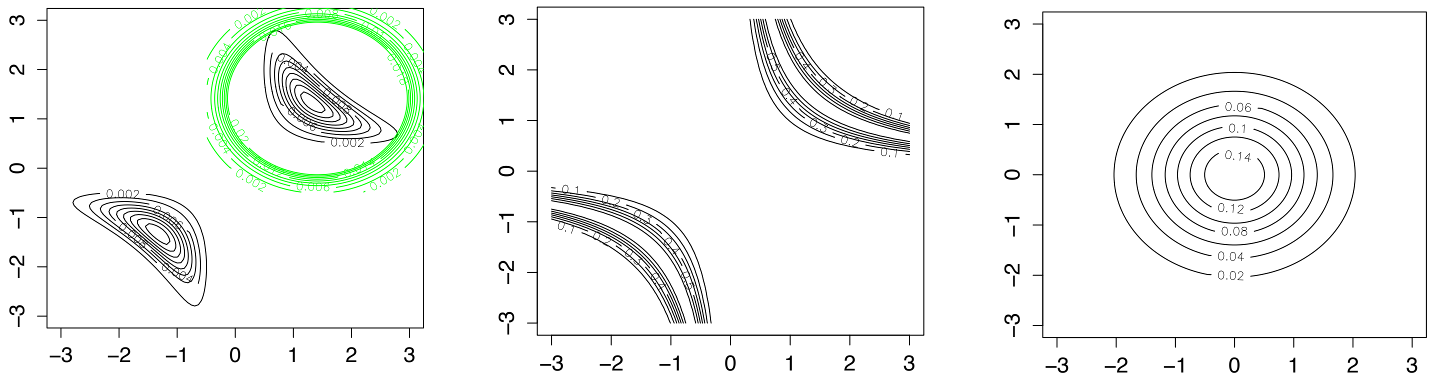
α-Divergence in Pictures



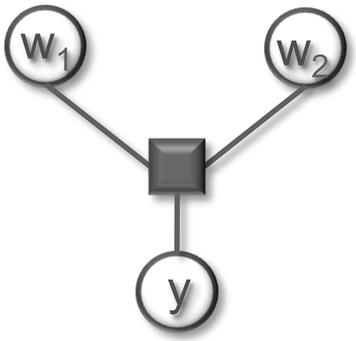
When to use which α-Divergence?



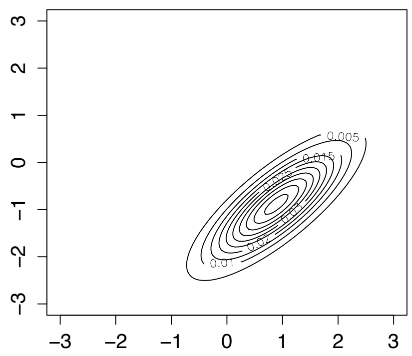
α=0 resolves multi-modality in the posterior at the expense of too much certainty!

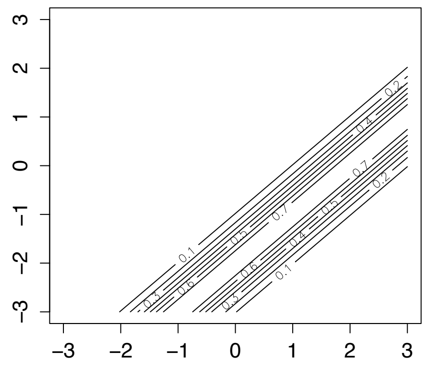


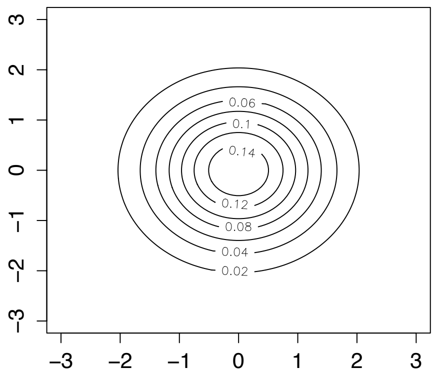
When to use which α-Divergence?



α=1 captures all uncertainty for uni-modal posterior distributions!

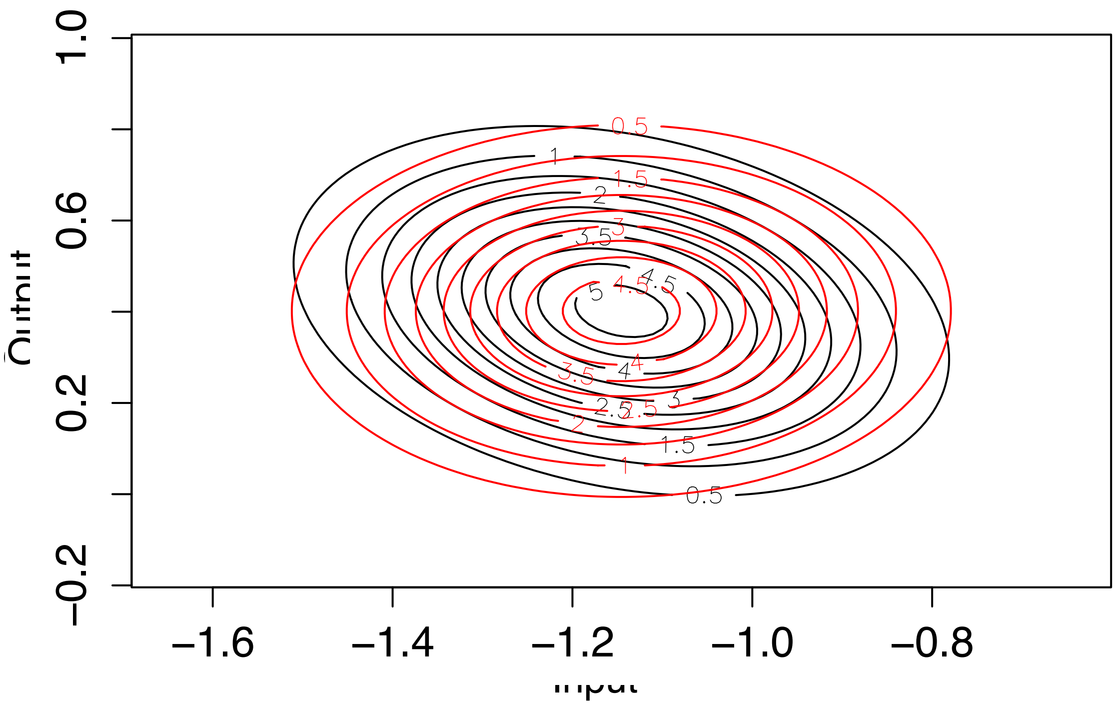






Sample (ctd)

$$p(\mathbf{w}|y, \mathbf{x}) = \mathcal{N}(y; [\sin(3t); \sin(t)\cos(6t)]^{\mathrm{T}}\mathbf{w}, 0.1) \cdot \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$$



Overview

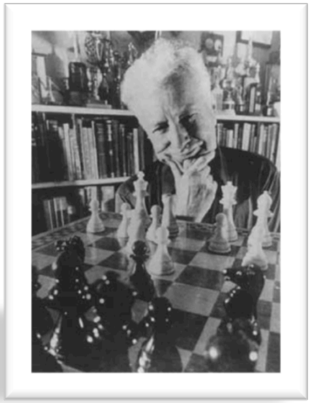
- Theory
  - Inference in Factor Graphs
  - Approximate Message Passing
- **Applications @ Microsoft**
  - TrueSkill: Gamer Rating and Matchmaking
  - TrueSkill Through Time: History of Chess
  - Click-Through Rate Prediction in Online Advertising
  - Matchbox: Recommendation Systems
- Applications @ Amazon

TrueSkill™

Joint work with Thore Graepel, Tom Minka & Phillip Trelford

Motivation

- Competition is central to our lives
  - Innate biological trait
  - Driving principle of many sports
- Chess Rating for fair competition
  - ELO: Developed in 1960 by Árpád Imre Élő
  - Matchmaking system for tournaments
- Challenges of online gaming
  - Learn from few match outcomes efficiently
  - Support multiple teams and multiple players per team

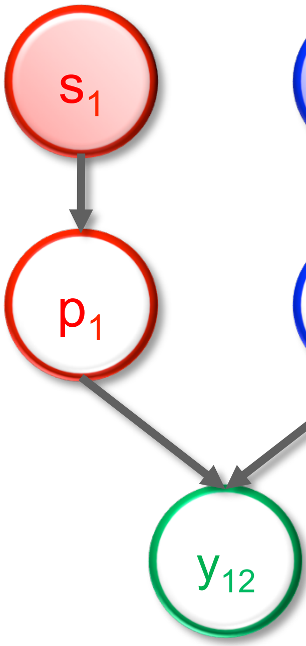


The Skill Rating Problem

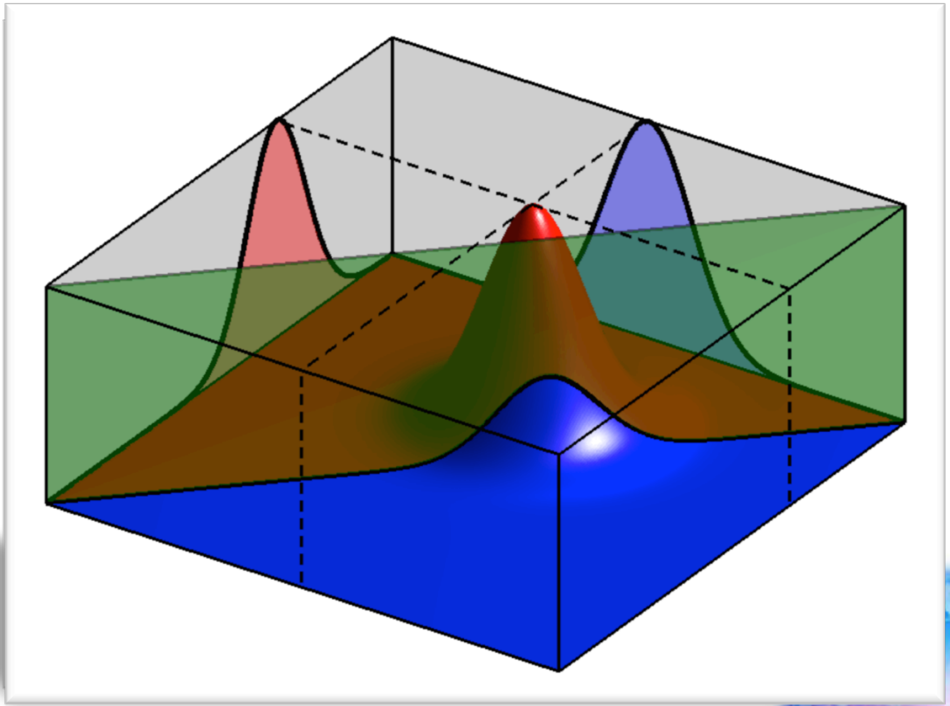
- **Given**:
  - Match outcomes: Orderings among k teams

Two Player Match Outcome Model

- Latent Gaussian performance model for fixed skills
- Possible outcomes: Player 1 wins over 2 (and vice versa)

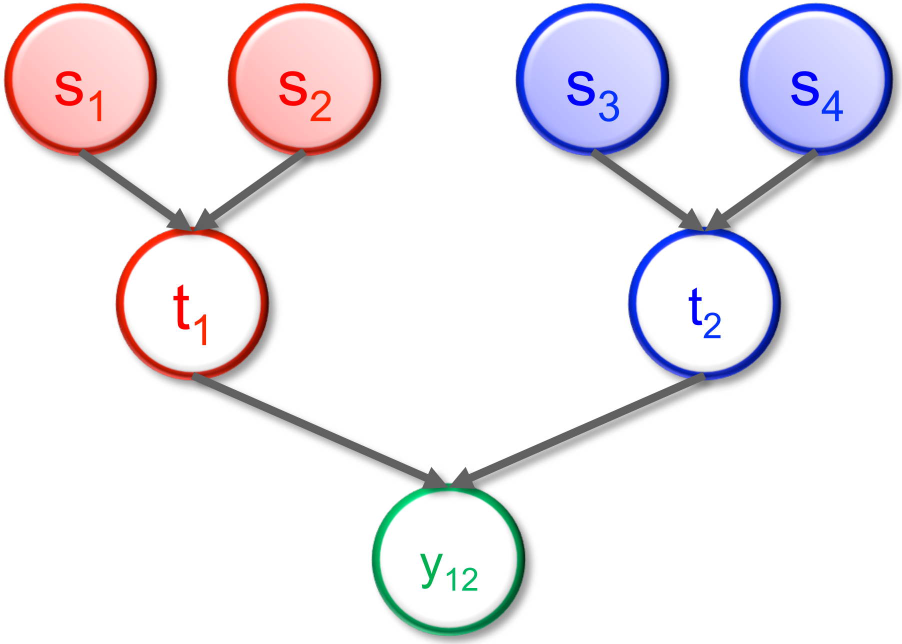


$$\mathbf{P}(y_{12} = (1,2) | p_1, p_2) = \mathbb{I}(p_1 > p_2)$$



Two Team Match Outcome Model

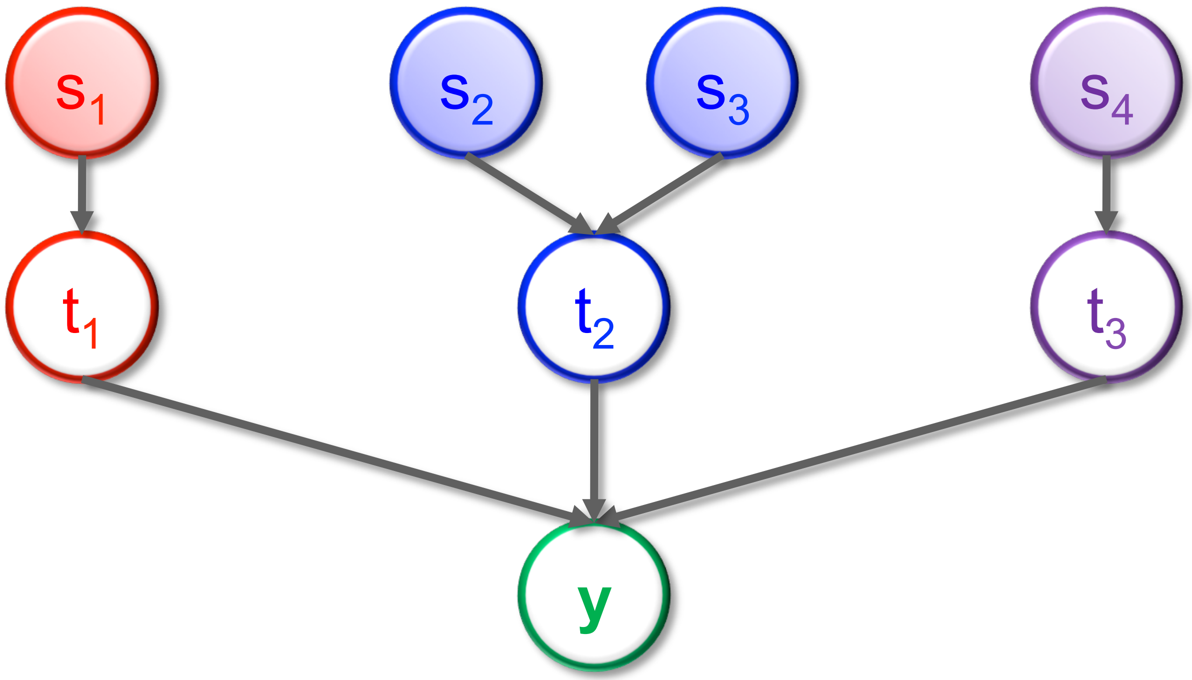
- Skill of a team is the sum of the skills of its members



$$\mathbf{P}(t_1 | s_1, s_2) = \mathcal{N}\left(t_1; s_1 + s_2, 2 \cdot \beta^2\right)$$

Multiple Team Match Outcome Model

- Possible outcomes: Permutations of the teams

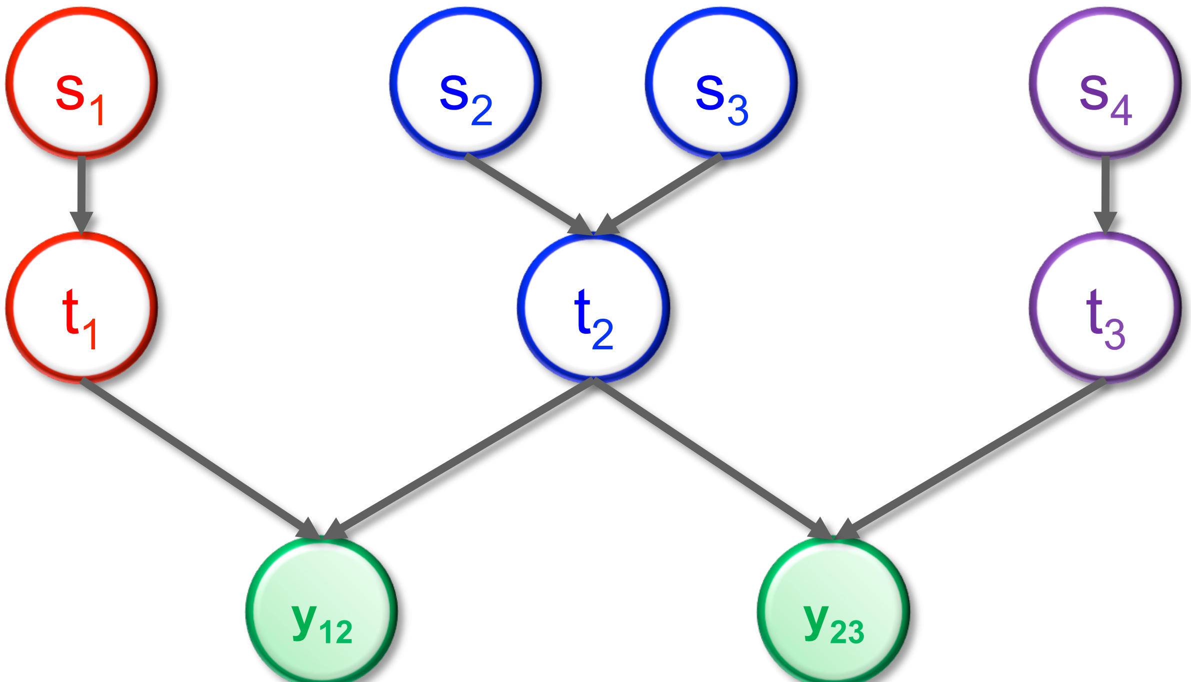


$$\mathbf{P}(\boldsymbol{y}|t_1, t_2, t_3) = \mathbb{I}(\boldsymbol{y} = (i, j, k)) \text{ where } t_i > t_j > t_k$$

Multiple Team Match Outcome Model

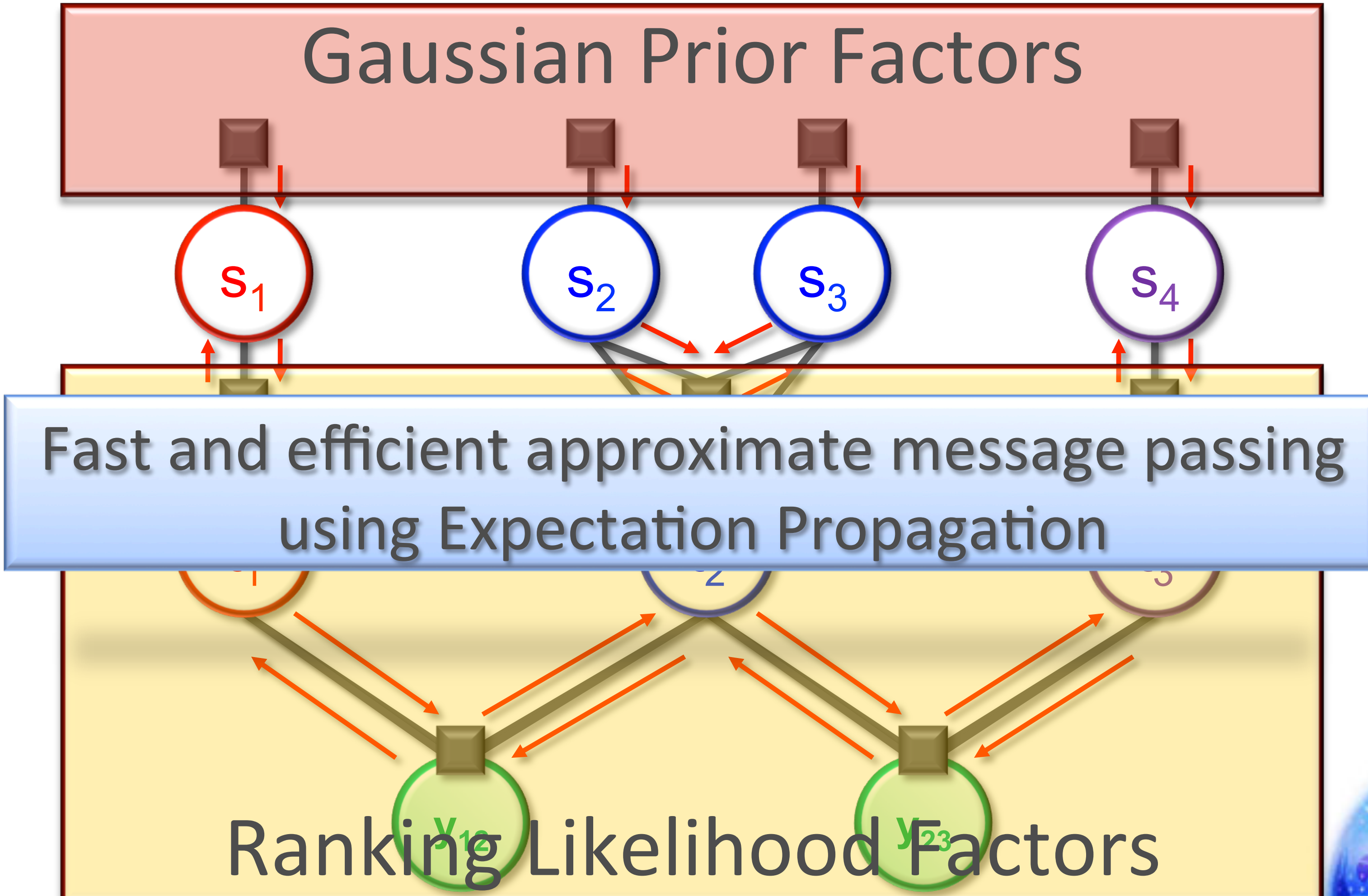
- But we are interested in the (Gaussian) posterior!

$$\mathbf{P}(s_i|\boldsymbol{y} = (1,2,3)) = \mathcal{N}(s_i; \mu_i, \sigma_i^2)$$



$y_{12} = (1,2)$ $y_{23} = (2,3)$

Efficient Approximate Inference

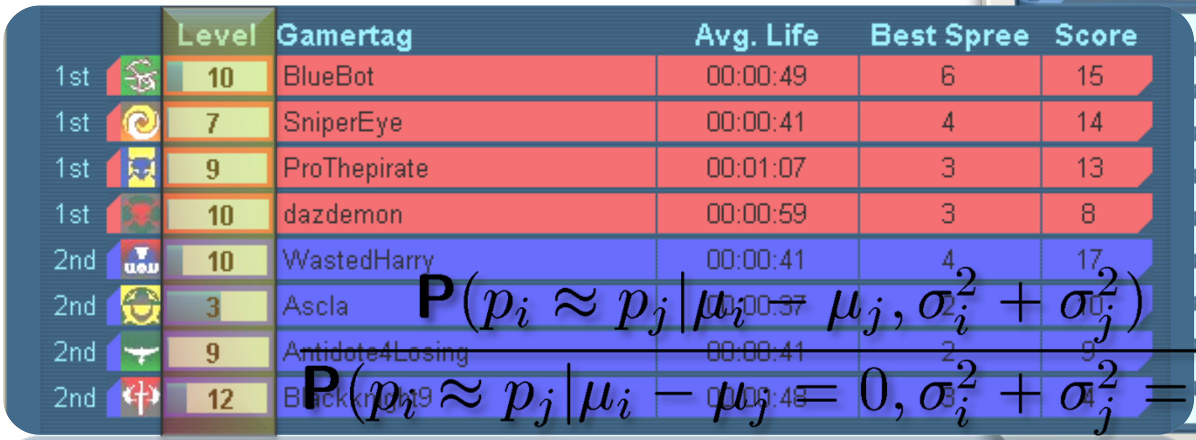


Applications to Online Gaming

- **Leaderboard**
  - Global ranking of all players

$$\mu_i - 3 \cdot \sigma_i$$

- **Matchmaking**
  - For gamers: Most uncertain outco

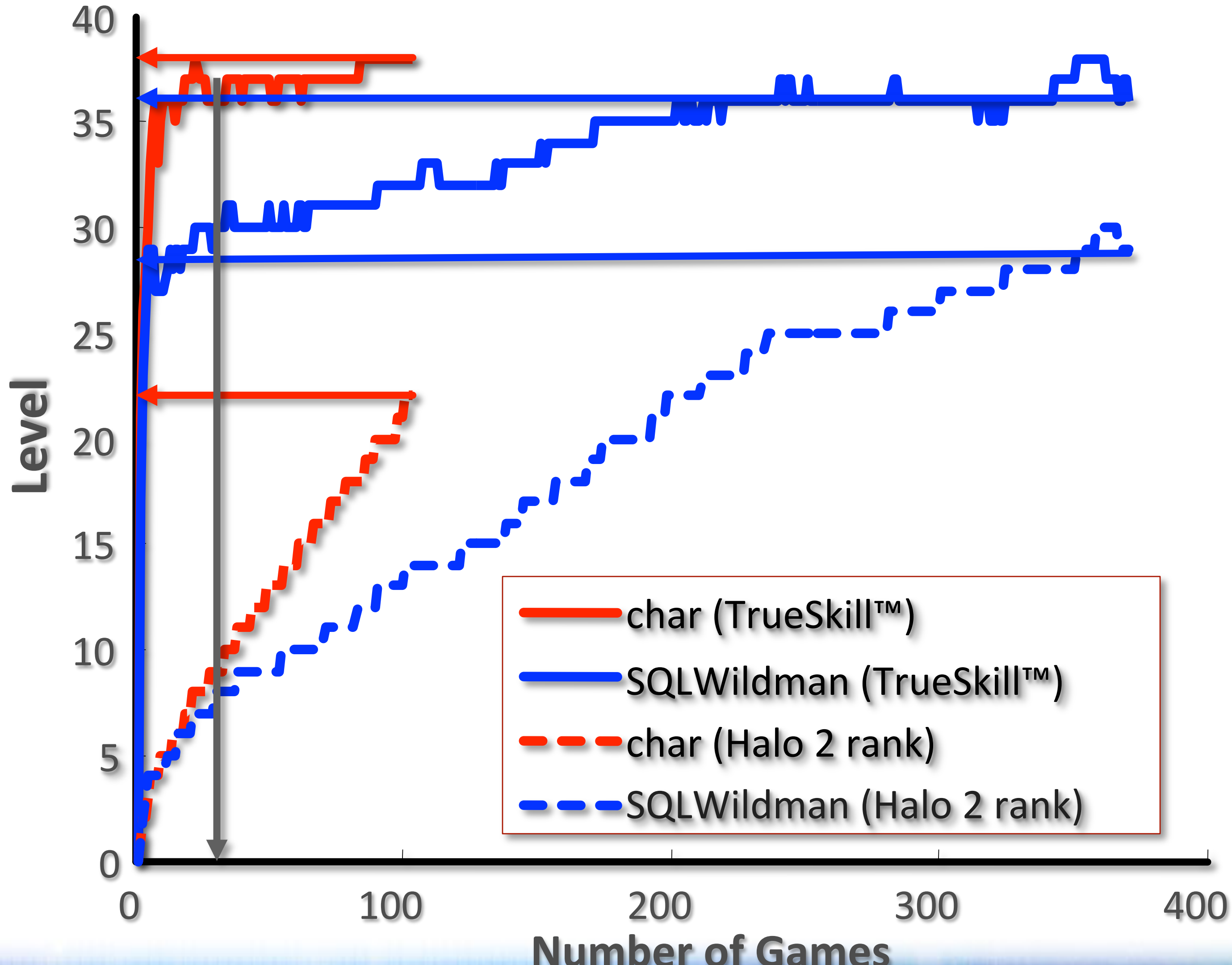


$$\frac{\mathbf{P}(p_i \approx p_j \,|\, \mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)}{\mathbf{P}(p_i \approx p_j \,|\, \mu_i - \mu_j = 0, \sigma_i^2 + \sigma_j^2 = 0)}$$

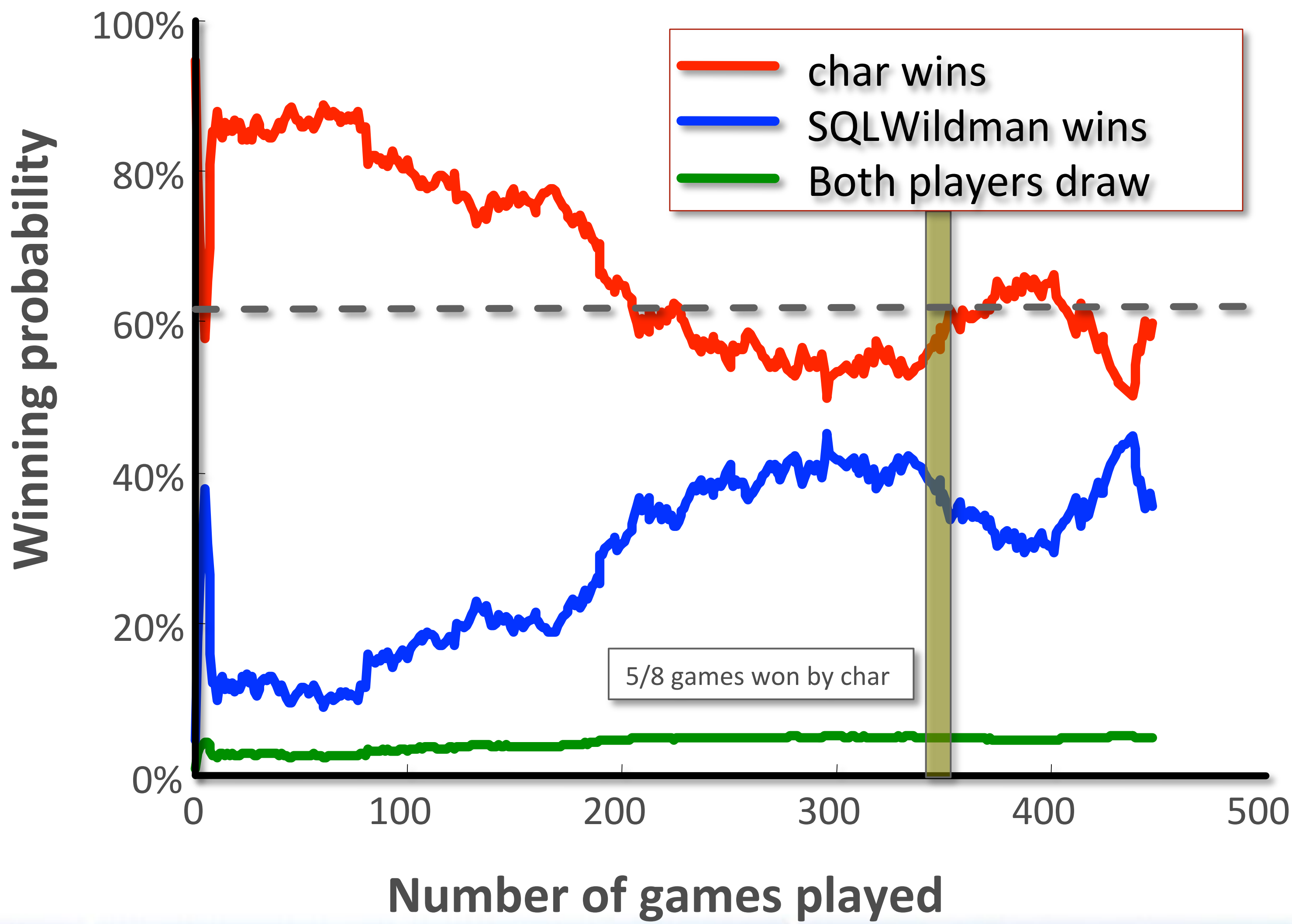
Experimental Setup

- **Data Set: Halo 2 Beta**
  - 3 game modes
    - Free-for-All
    - Two Teams
    - 1 vs. 1
  - > 60,000 match outcomes
  - ≈ 6,000 players
  - 6 weeks of game play
  - Publically available

Convergence Speed



Convergence Speed (ctd.)

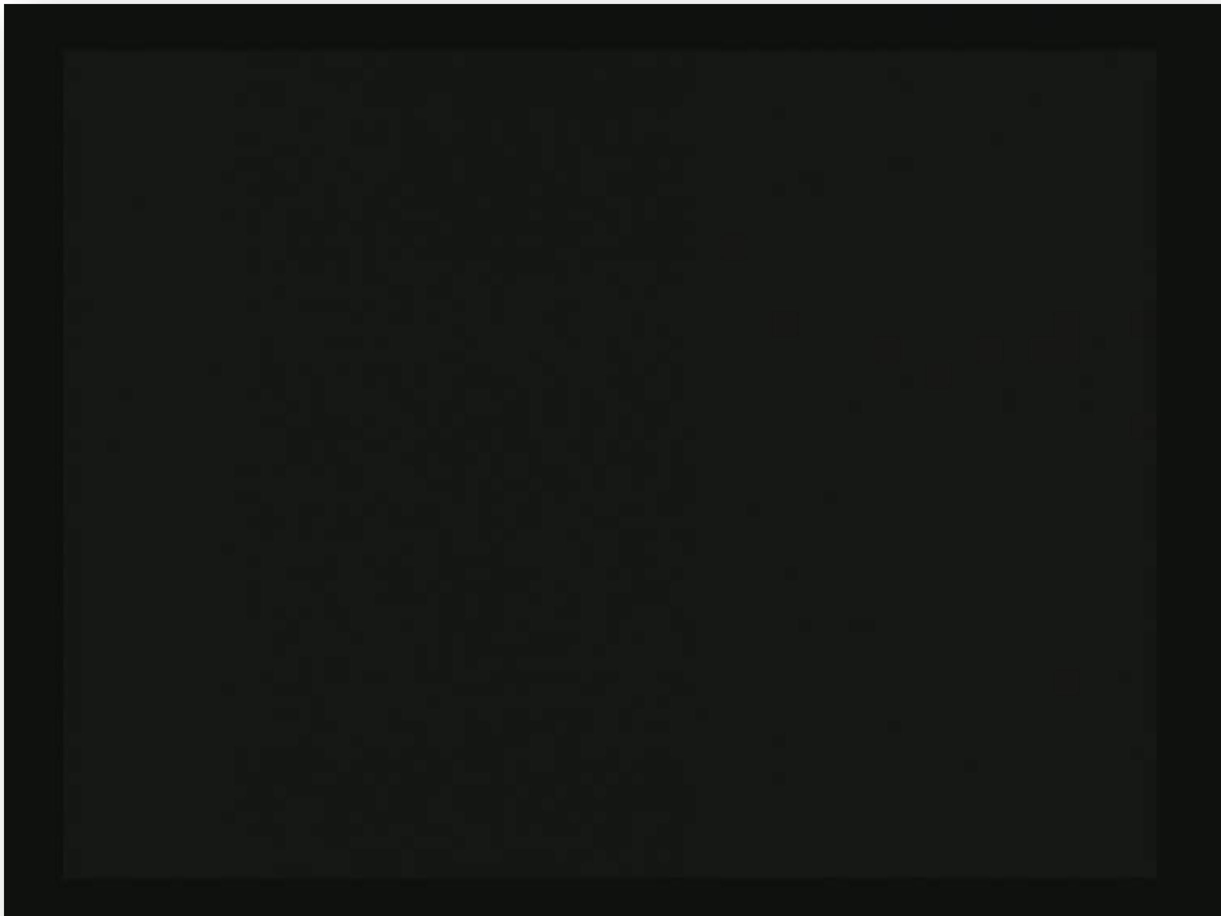


Xbox 360 & Halo 3

- **Xbox 360 Live**
  - Launched in September 2005
  - Every game uses TrueSkill™ to match players
  - > 10 million players
  - > 2 million matches per day
  - > 2 billion hours of gameplay

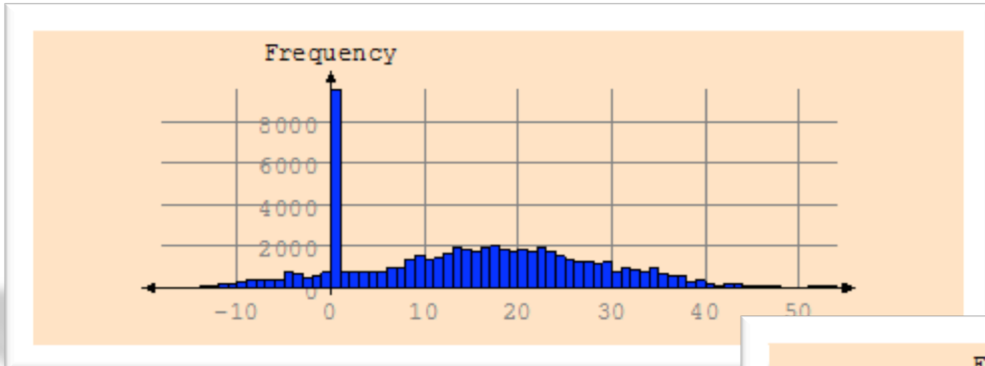
- **Halo 3**
  - Launched on 25th September 2007
  - Largest entertainment launch in history
  - > 200,000 player concurrently (peak: 1,000,000)

Halo 3 in Action



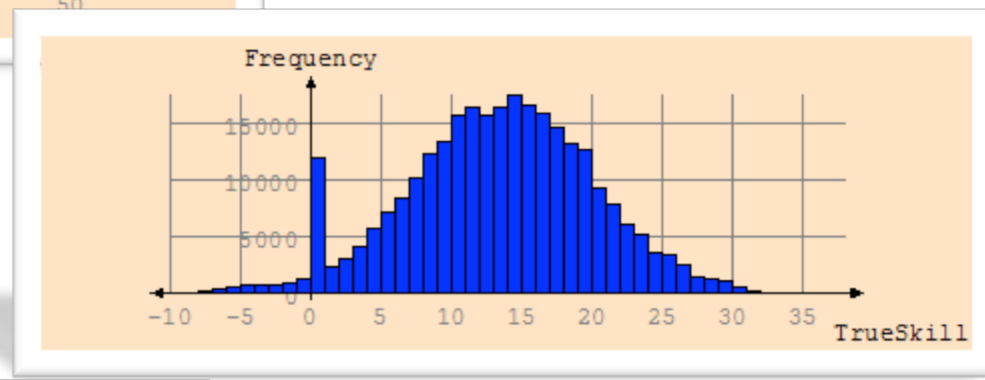
Halo 3 Public Beta Analysis

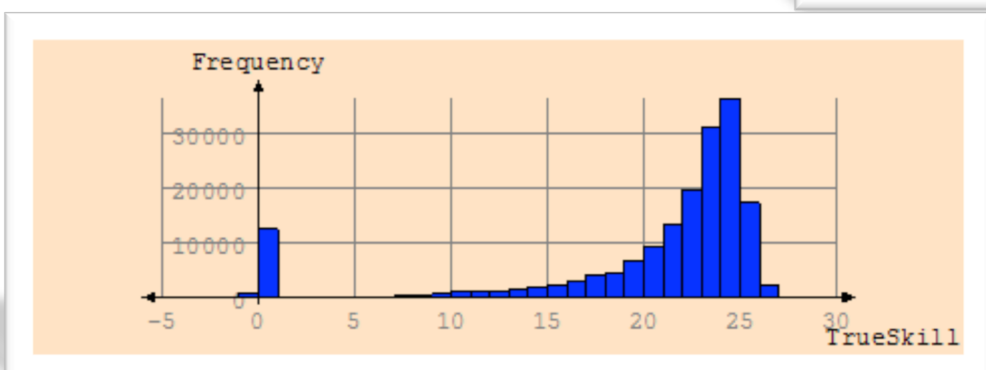
Skill Distributions of Online Games



Golf (18 holes): 60 levels

Car racing (3-4 laps): 40 levels

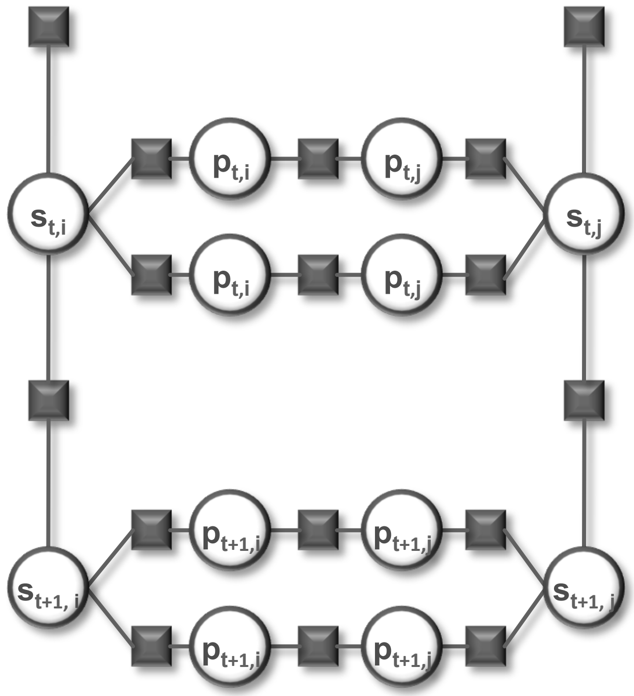




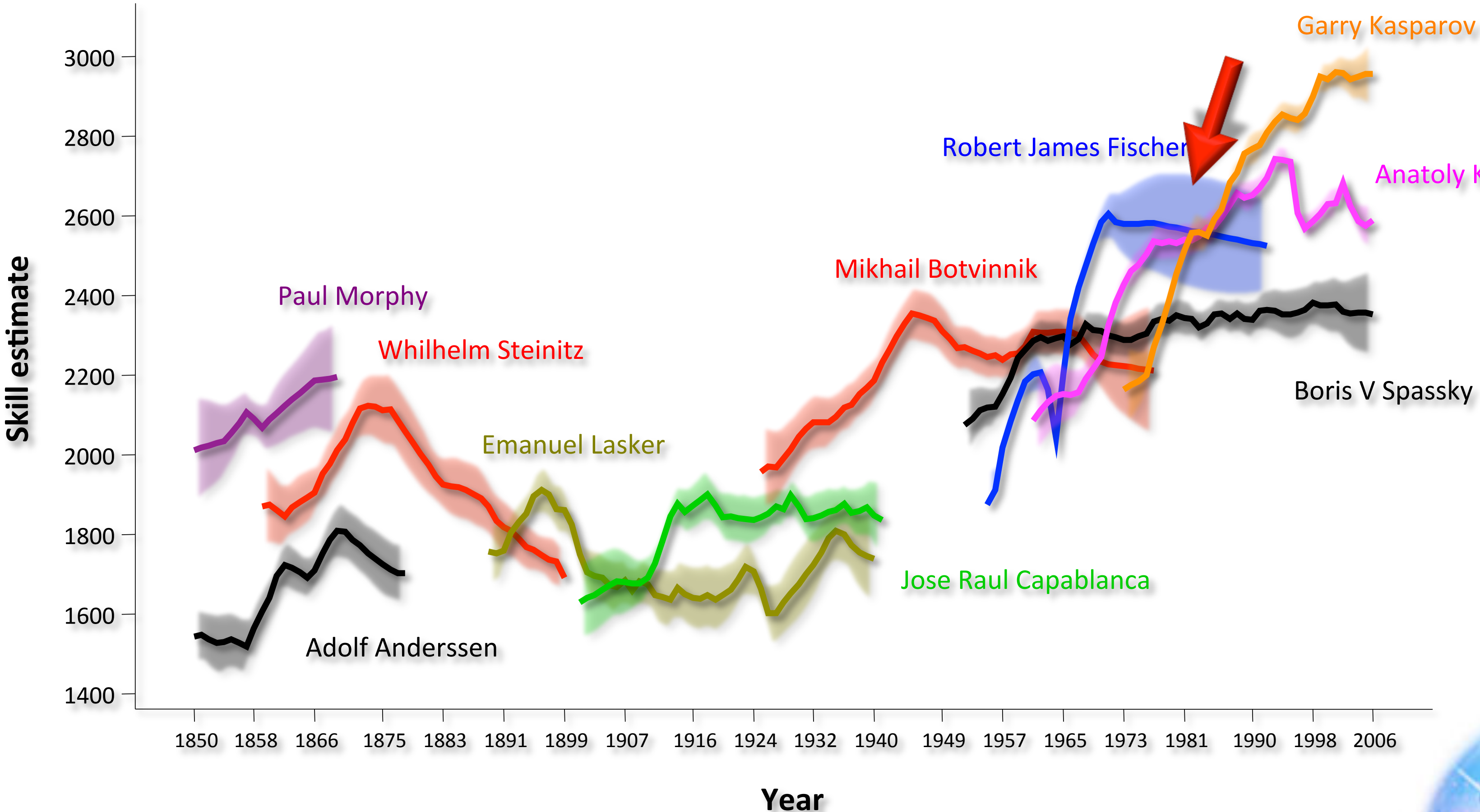
UNO (chance game): 10 levels

TrueSkill™ Through Time: Chess

- Model time-series of skills by smoothing across time
- History of Chess
  - 3.5M game outcomes (ChessBase)
  - 20 million variables (each of 200,000 players in each year of lifetime + latent variables)
  - 40 million factors



ChessBase Analysis: 1850 - 2006



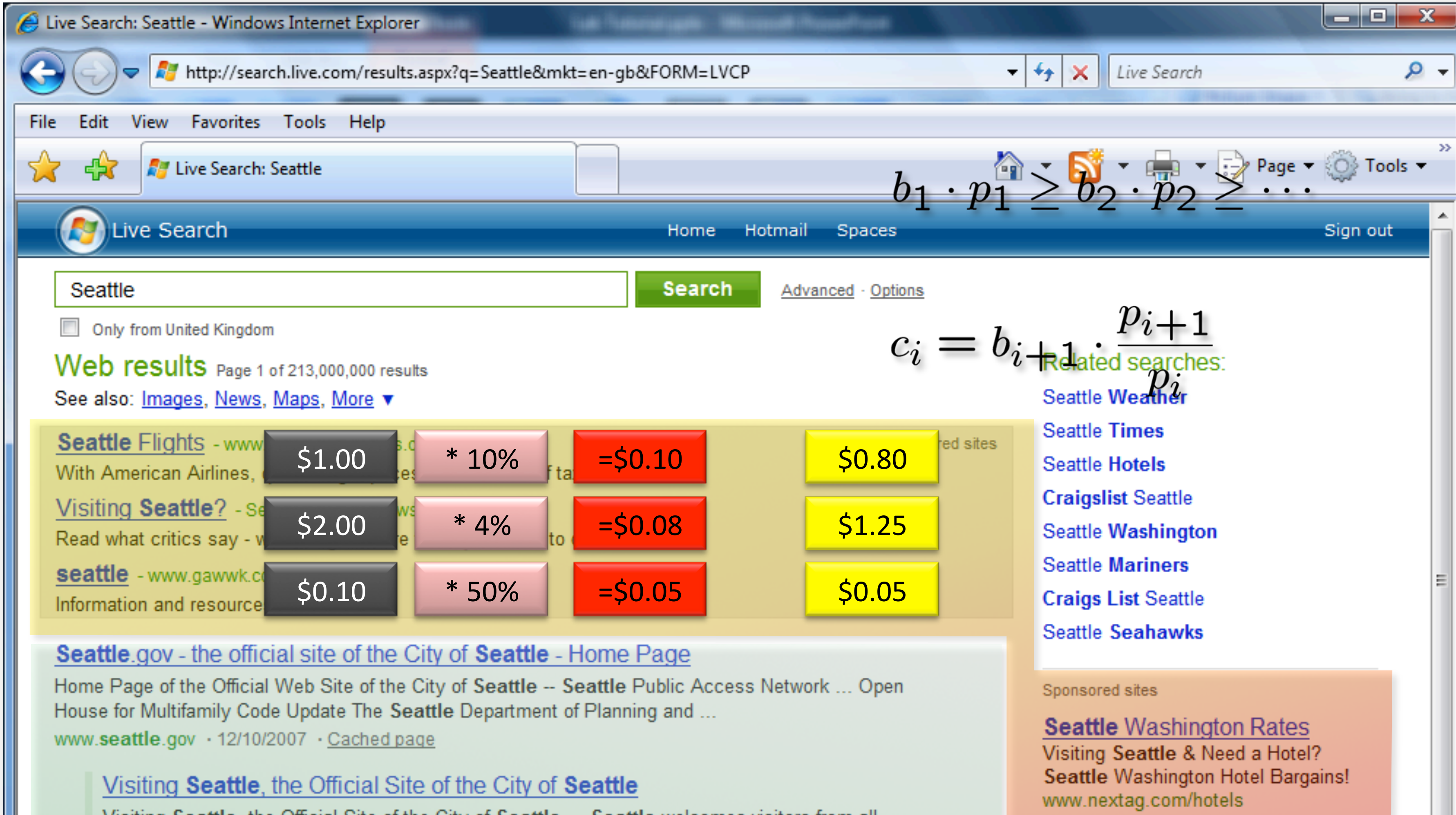
Online Advertising

Joint work with Thore Graepel, Joaquin Quiñonero Candela, Onno Zoeter, Tom Borchert , Phillip Trelford

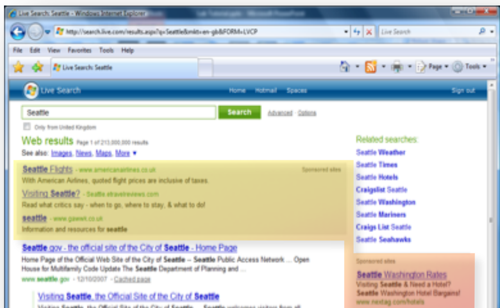
Why Predict Probability-of-Click?

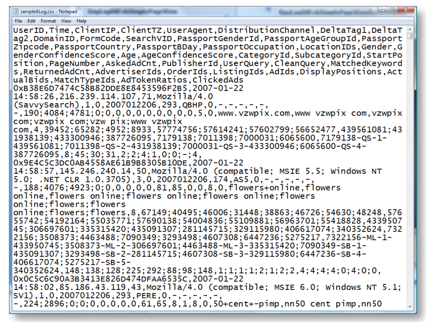
$$b_1 \cdot p_1 \geq b_2 \cdot p_2 \geq \cdots$$

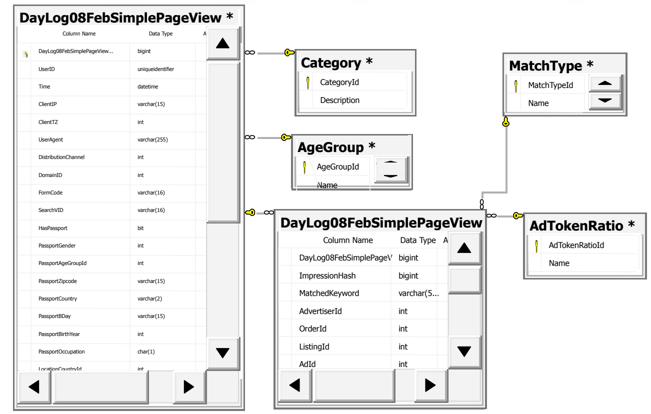
$$c_i = b_{i+1} \cdot \frac{p_{i+1}}{p_i}$$



The Flow of Information



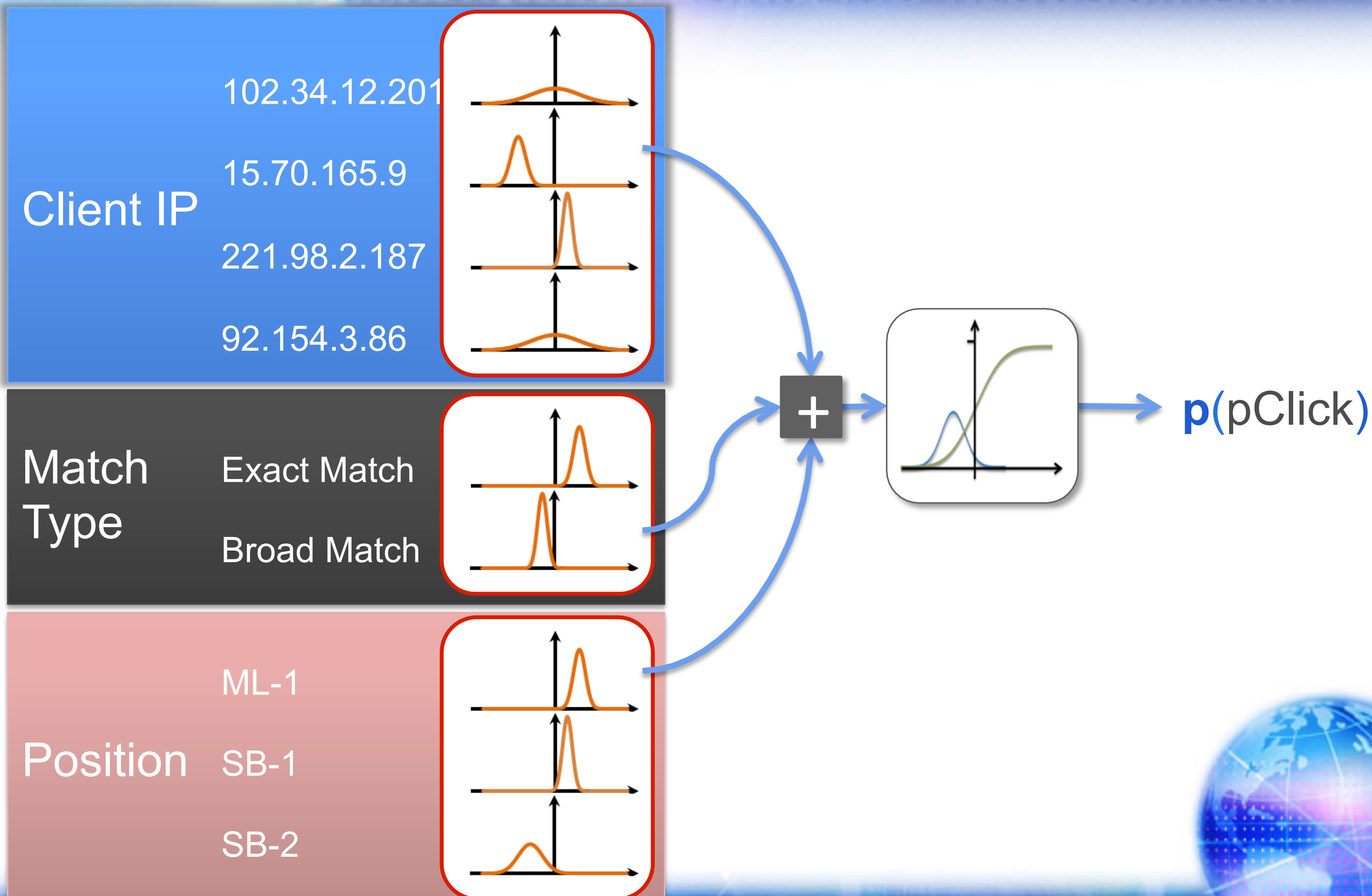




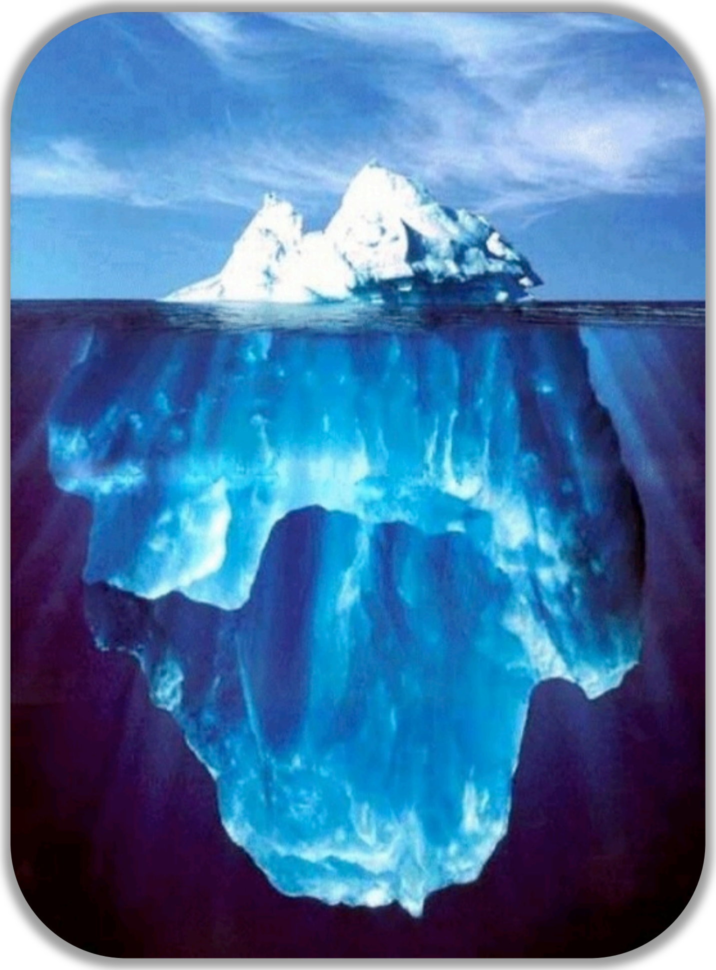
User interaction → Raw Logs → Structured Data

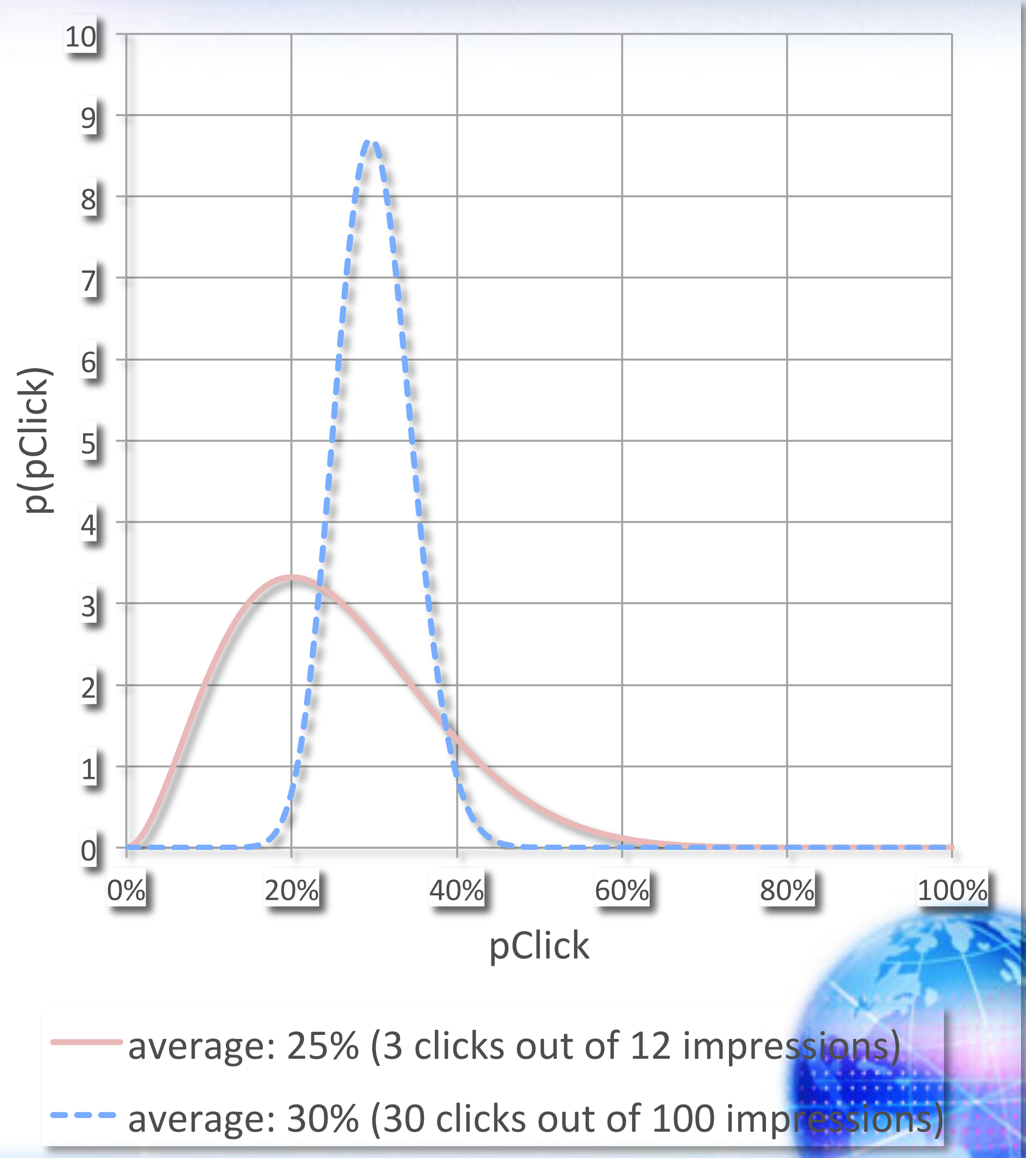
- Why structured data?
  - Data validation and cleaning
  - Principled feature transformations

Uncertainty: Bayesian Probabilities

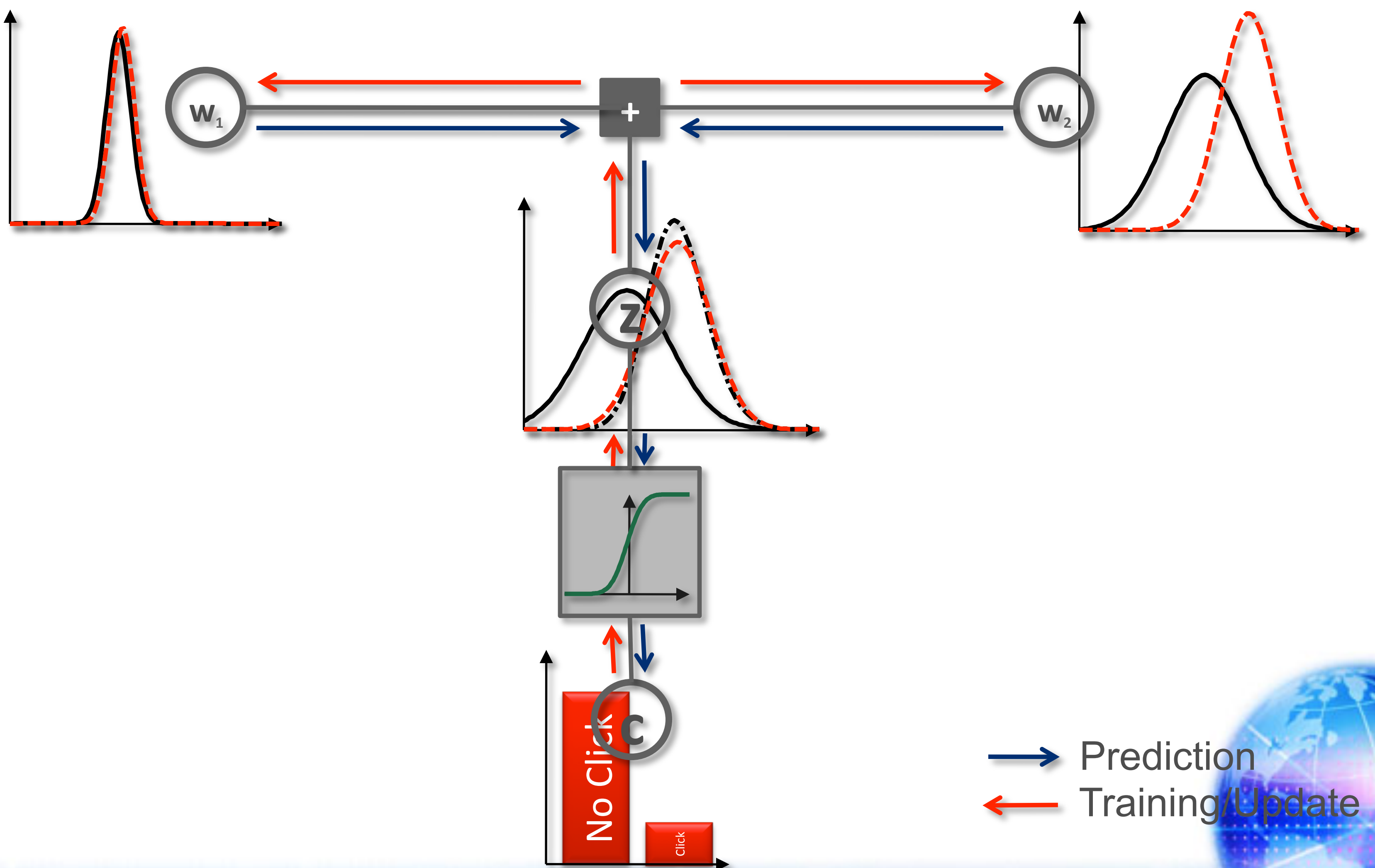


Principled Exploration





Training Algorithm in Action

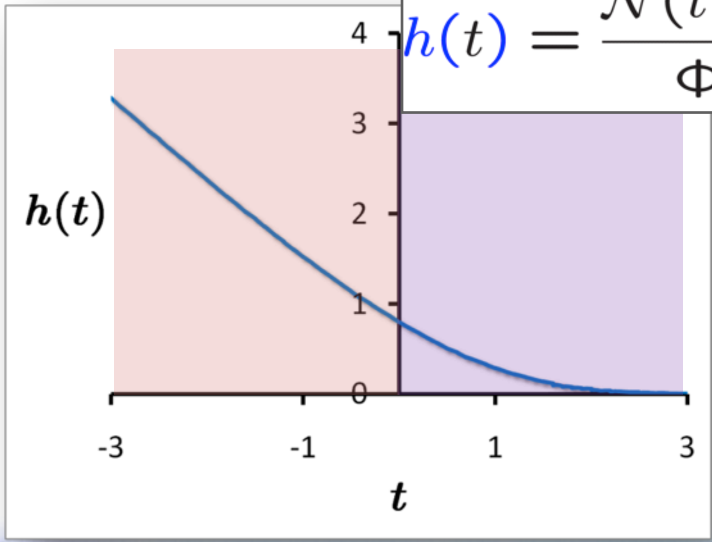


Inference: An Optimization View

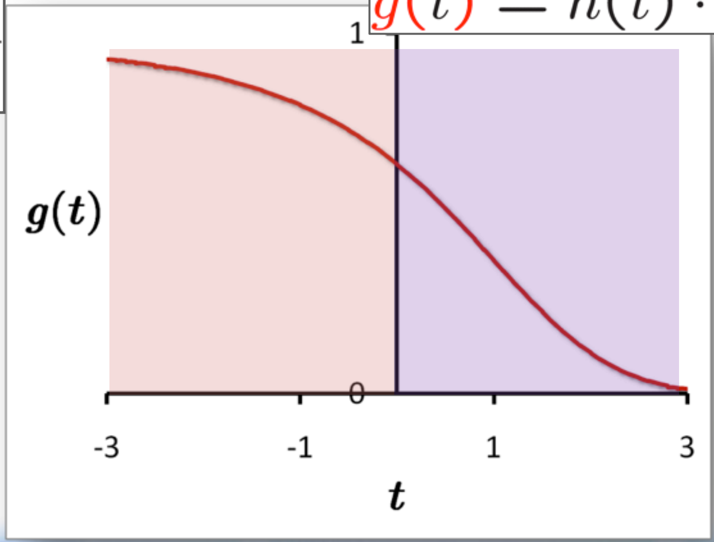
$$\mu_i \leftarrow \mu_i + \frac{\sigma_i^2}{s} \cdot h\left[\frac{\sum_{j=1}^d \mu_j}{s}\right] \quad \sigma_i^2 \leftarrow \sigma_i^2 \left(1 - \frac{\sigma_i^2}{s^2} \cdot g\left[\frac{\sum_{j=1}^d \mu_j}{s}\right]\right)$$

$$s^2 = \beta^2 + \sum_{j=1}^d \sigma_j^2$$

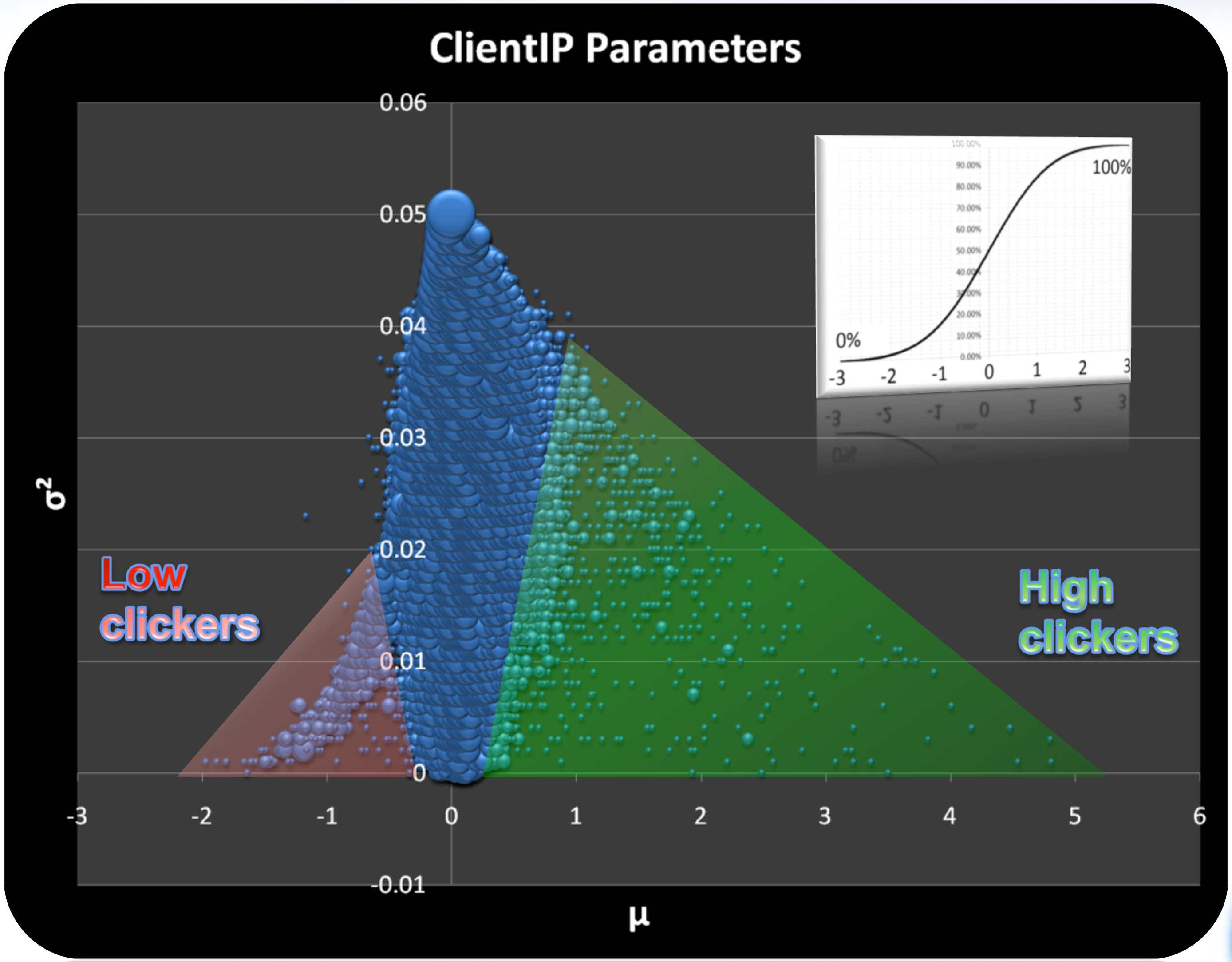
$$h(t) = \frac{\mathcal{N}(t;\, 0, 1)}{\Phi(t)}$$



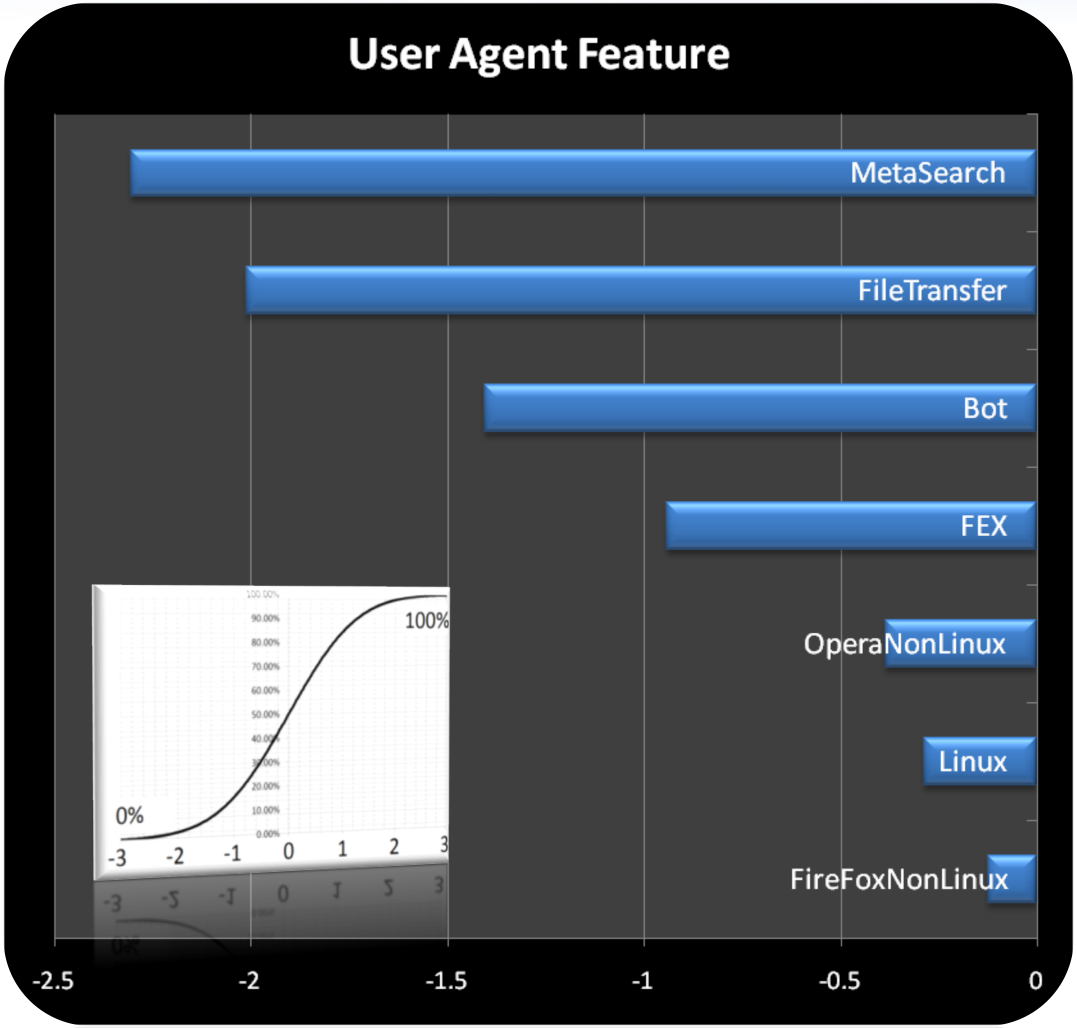
$$g(t) = h(t) \cdot [h(t) + t]$$



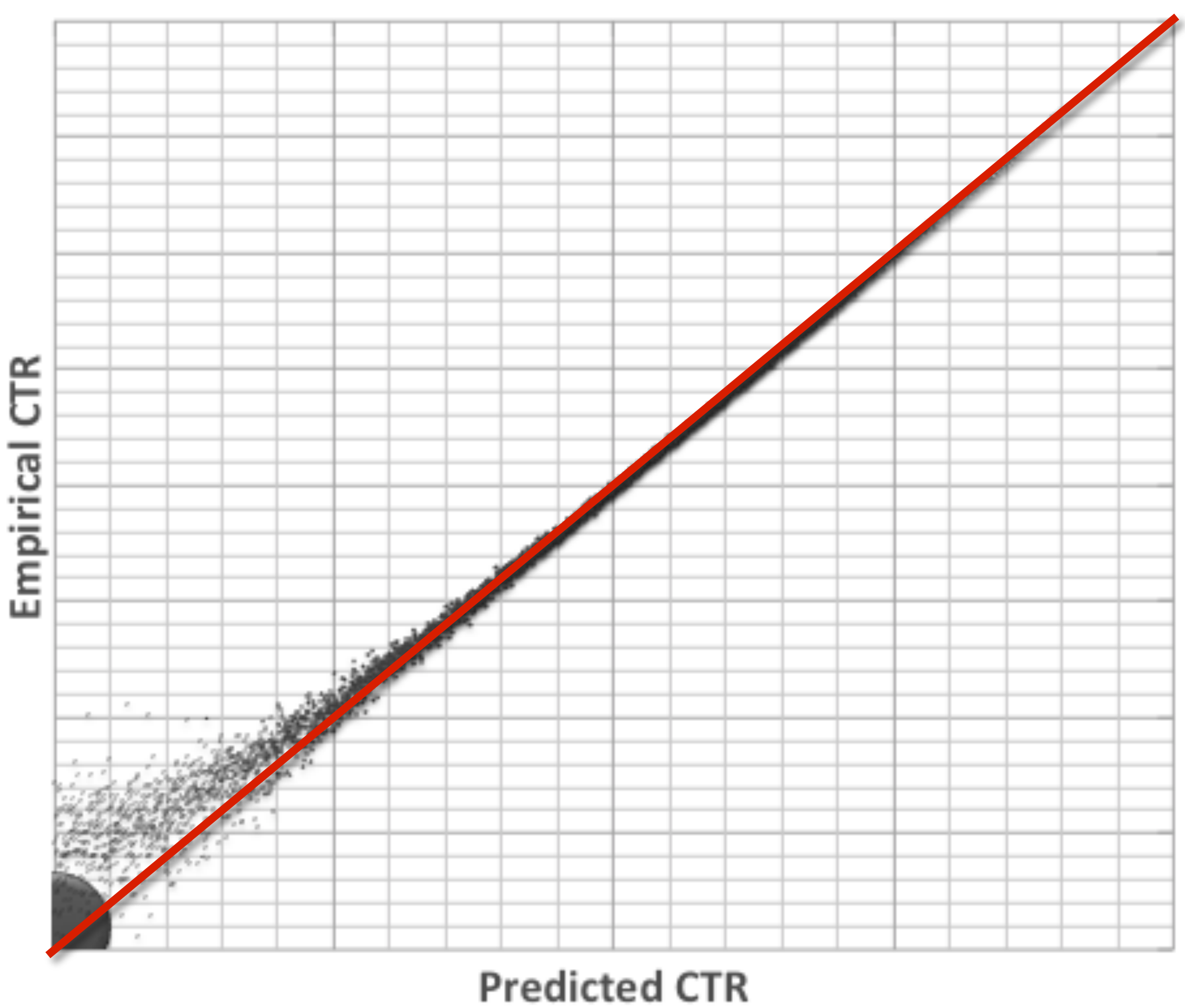
Client IP: Mean & Variance



UserAgent: Mean Posterior Effects

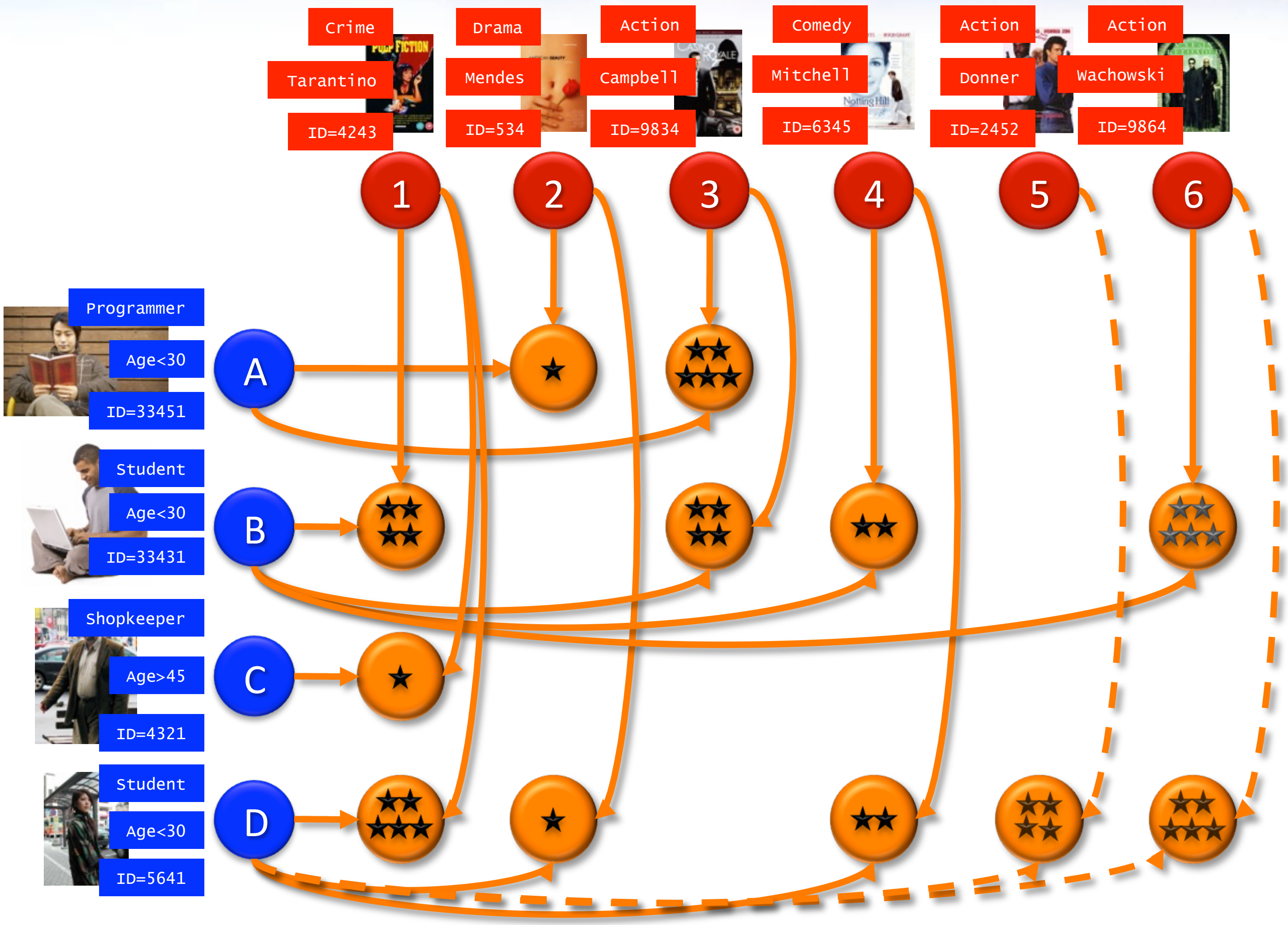


Accuracy



MatchBox

Joint work with Thore Graepel, Joaquin Quiñonero Candela, David Stern, Ulrich Paquet



Crime
Tarantino
ID=4243
Drama
Mendes
ID=534
Action
Campbell
ID=9834
Comedy
Mitchell
ID=6345
Action
Donner
ID=2452
Action
Wachowski
ID=9864
1
2
3
4
5
6
Programmer
Age<30
ID=33451
A
Student
Age<30
ID=33431
B
Shopkeeper
Age>45
ID=4321
C
Student
Age<30
ID=5641
D

Matchbox With Metadata

User Metadata

Item Metadata

User $\mathbf{s} = \mathbf{U}\mathbf{x}$

Item $\mathbf{t} = \mathbf{V}\mathbf{y}$

Rating potential ~ $\mathcal{N}(\mathbf{s}^\top \mathbf{t}, \beta^2)$

Recommender System: MatchBox

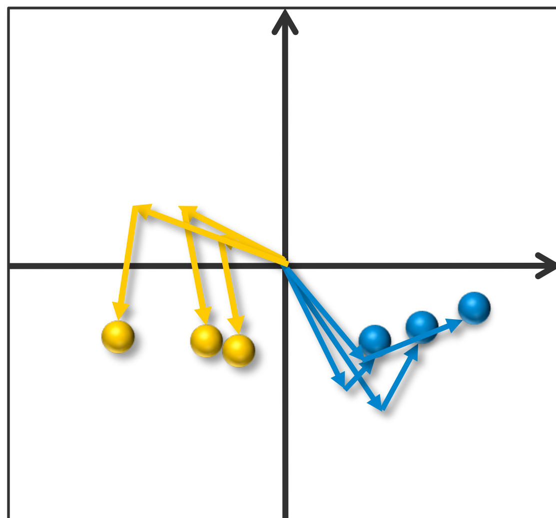
User

- mark
- ralf
- tao
- sheryl

Gender

- Male
- Female

User dislikes Movie



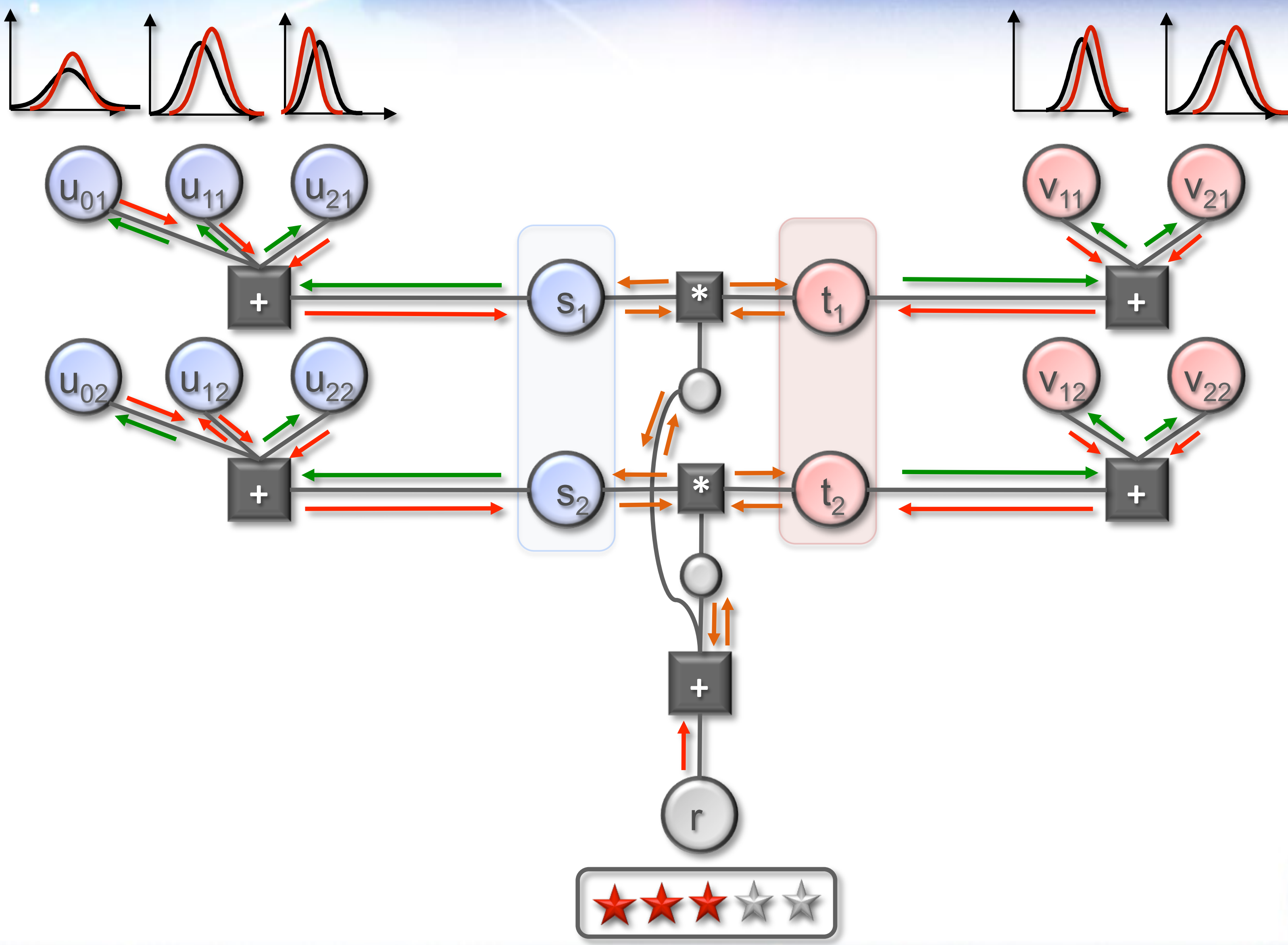
Movie

- Social Network
- Heat
- The Rock
- The Godfather

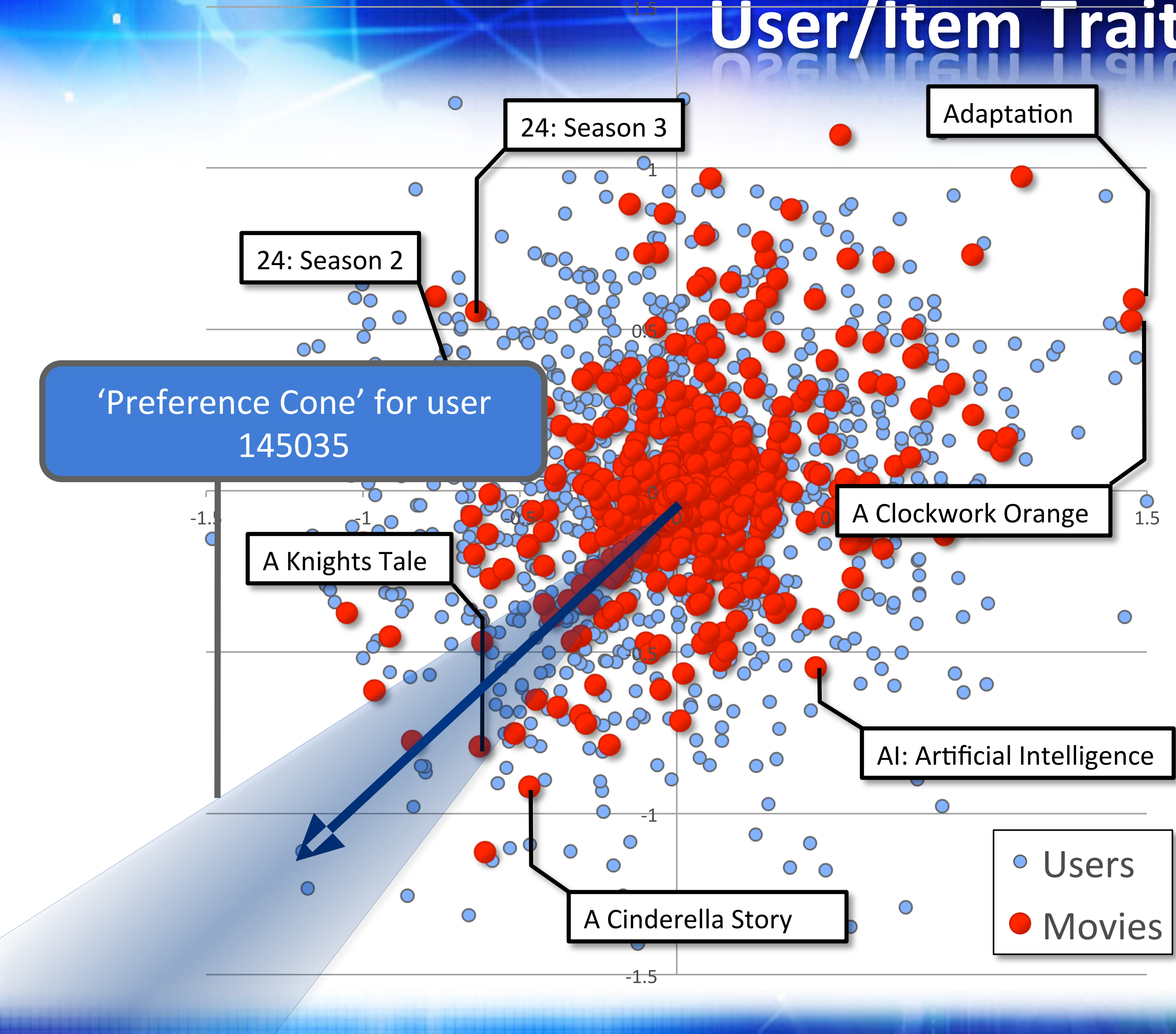
Director

- R. Scott
- C. Eastwood
- Q. Tarantino
- R. Howard

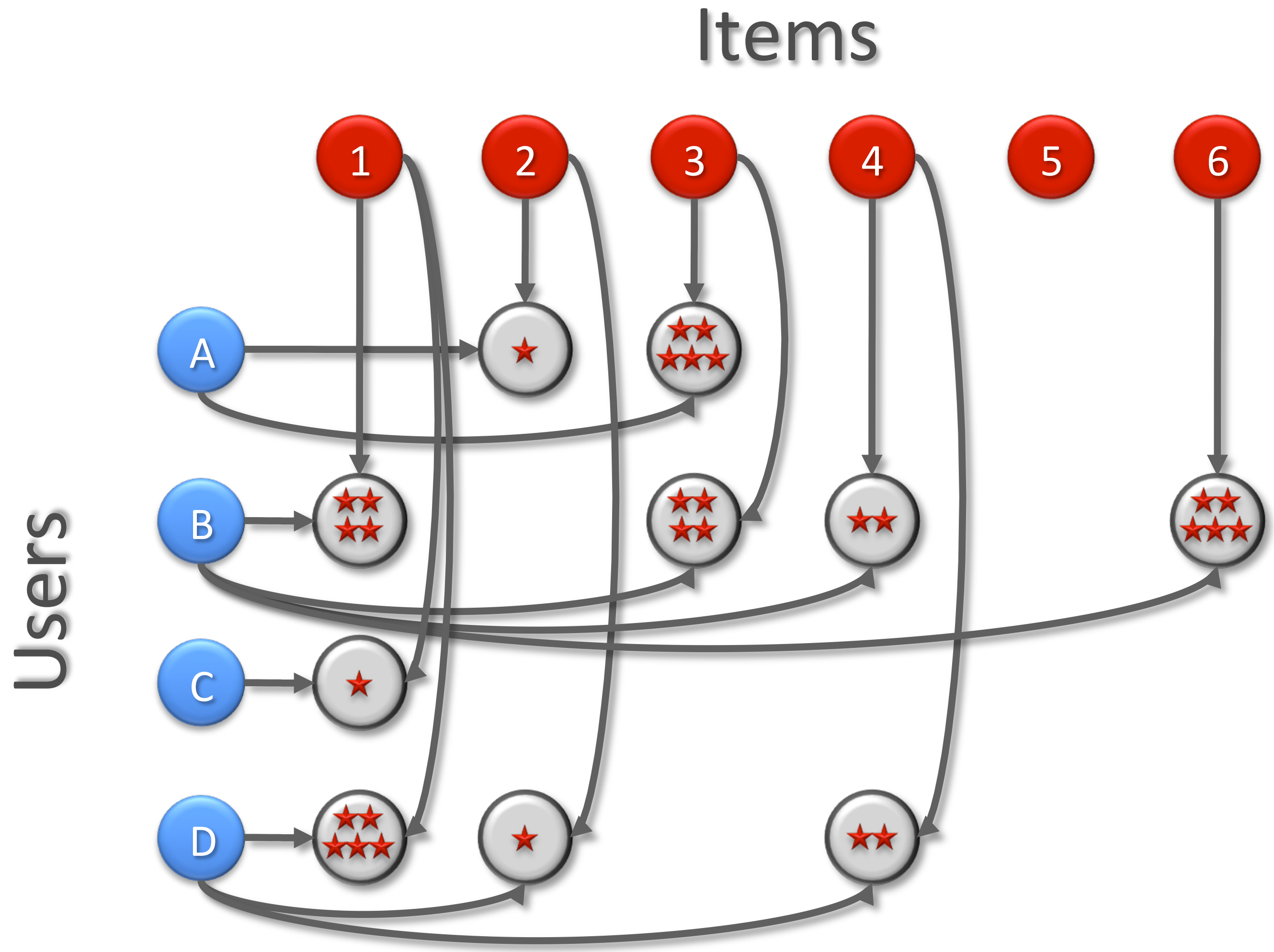
Message Passing For Matchbox



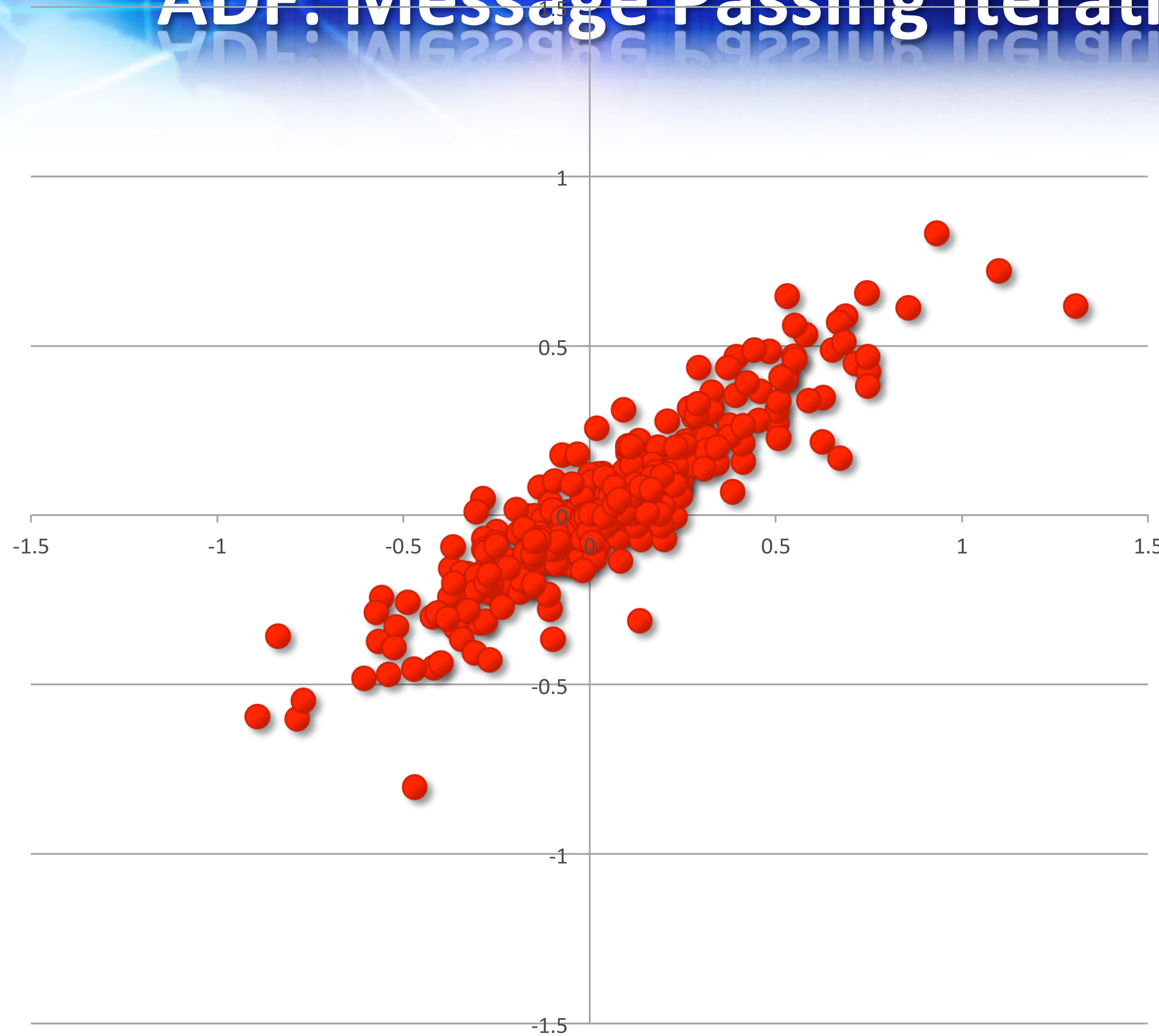
User/Item Trait Space



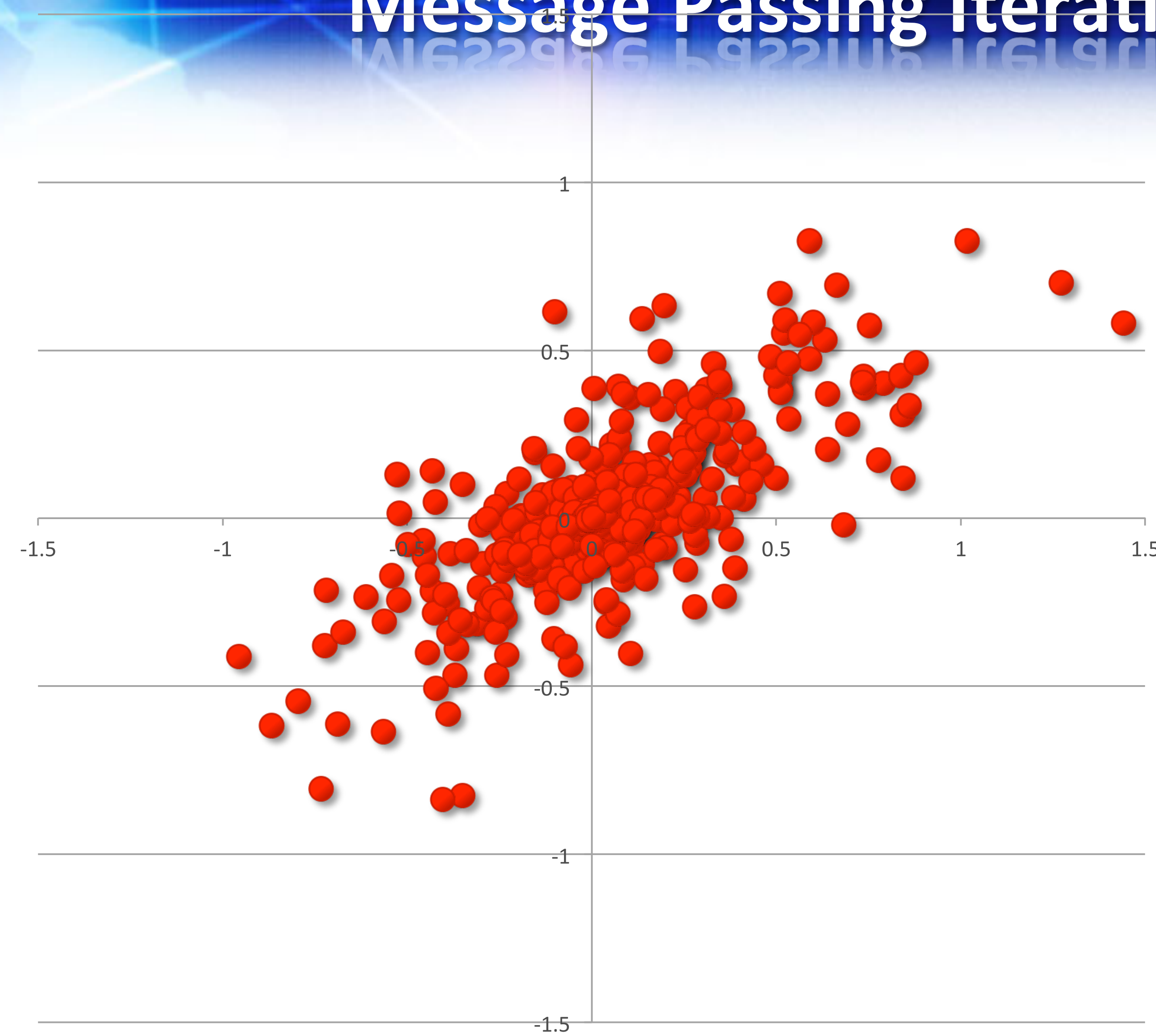
Incremental Training with ADF



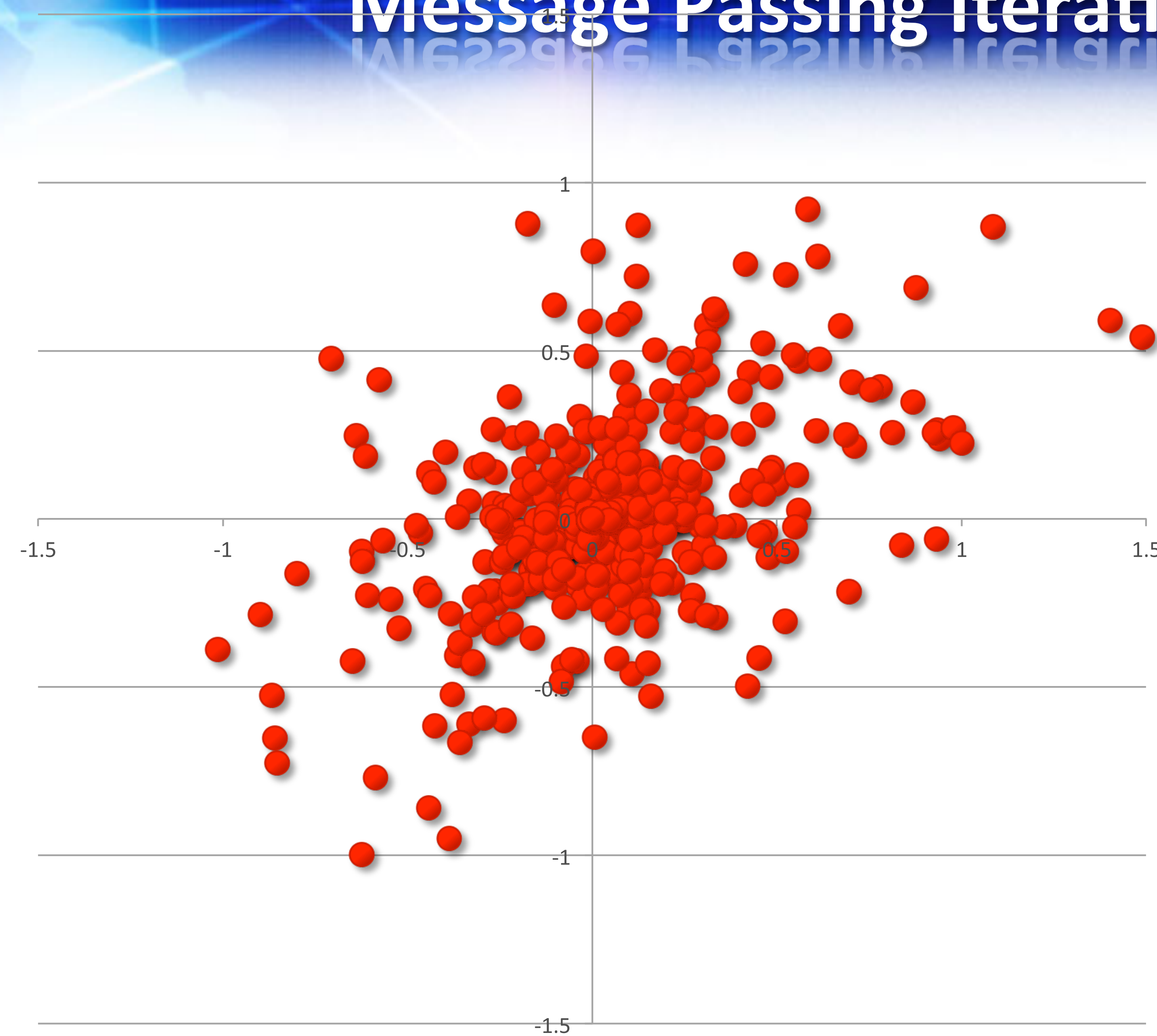
ADF: Message Passing Iteration 1



Message Passing Iteration 2



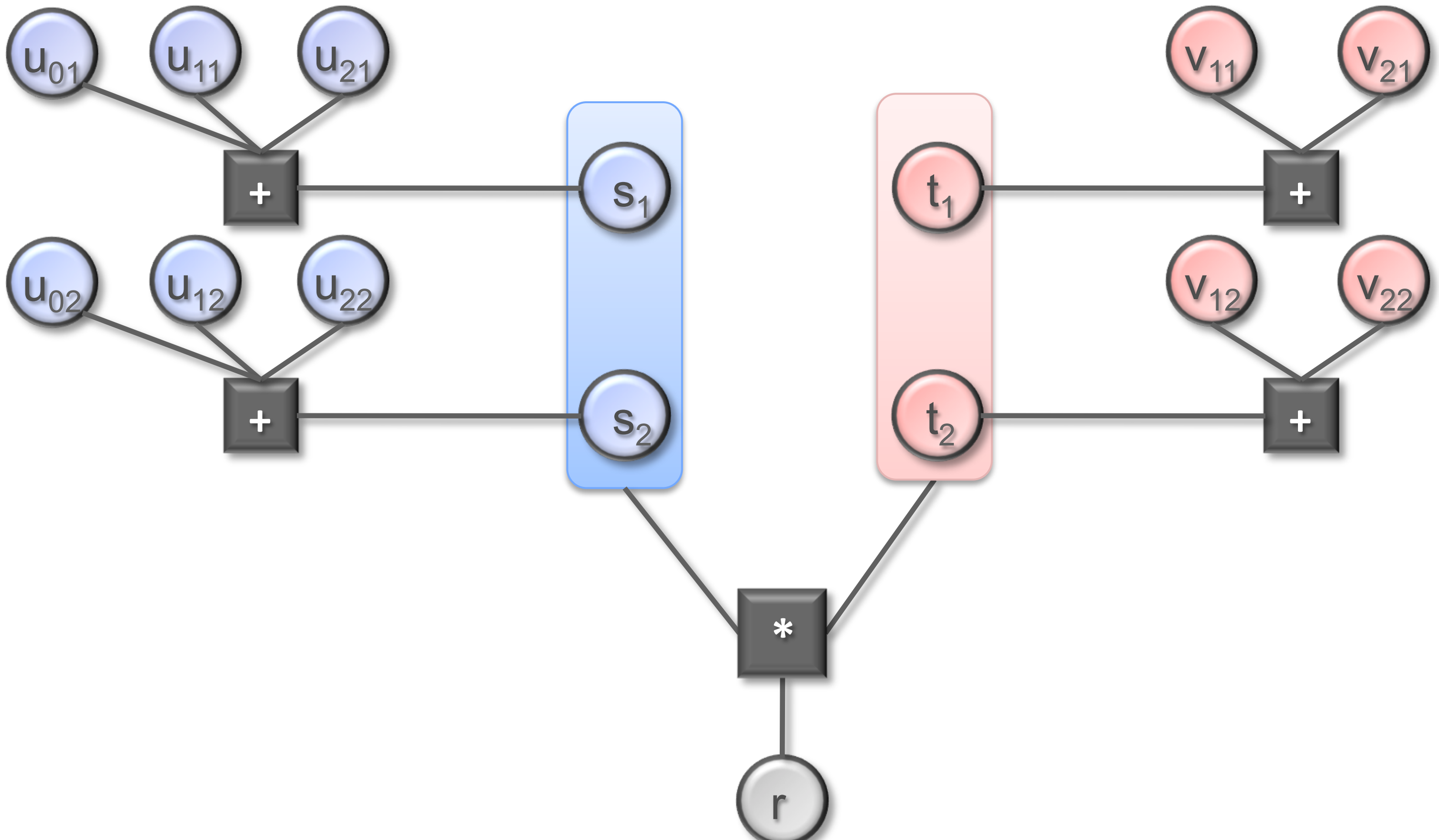
Message Passing Iteration 3



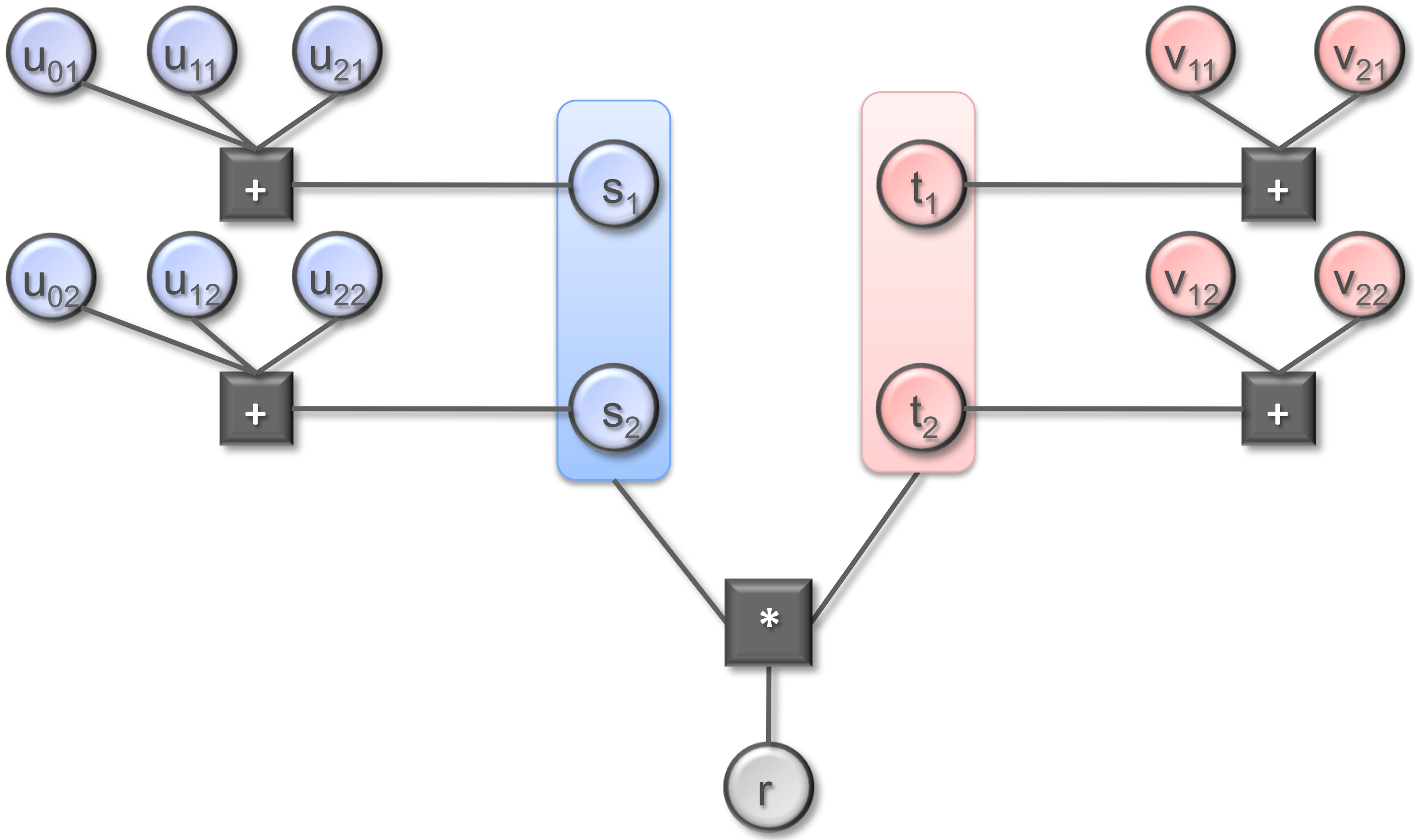
Message Passing Iteration 4

feedback models

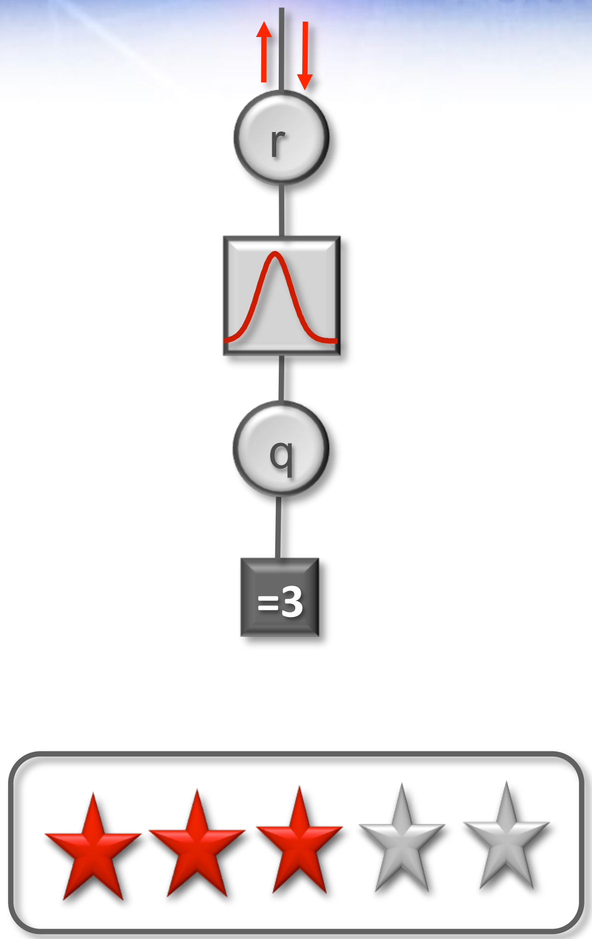
Feedback Models



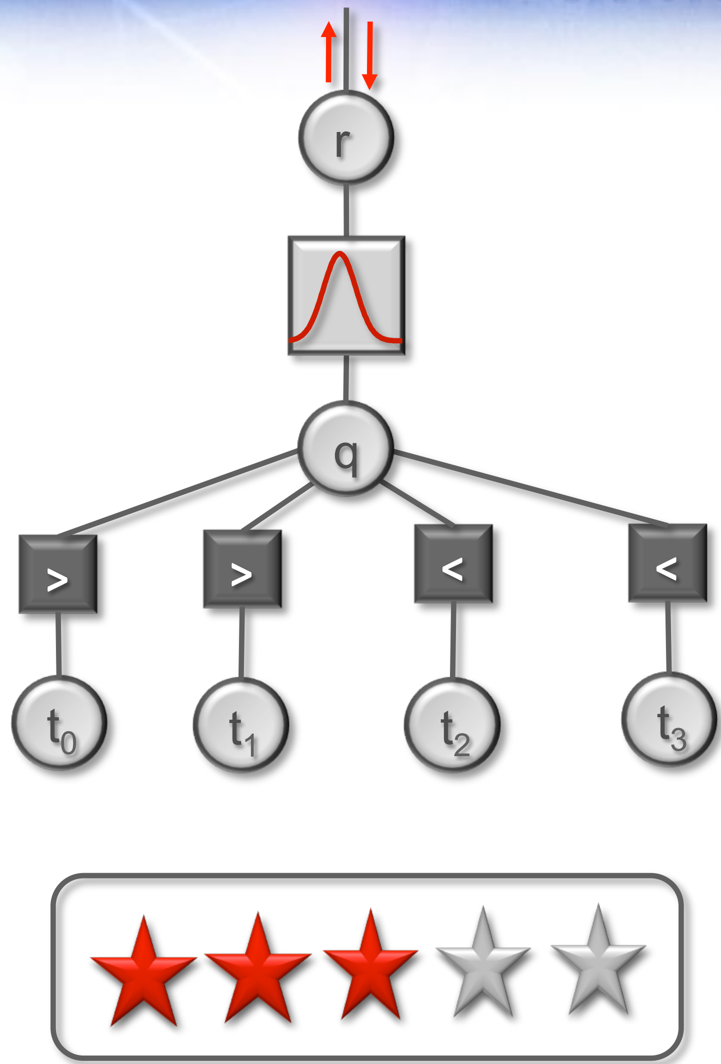
Feedback Models



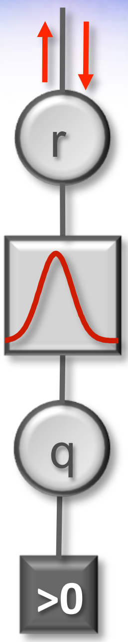
Feedback Models



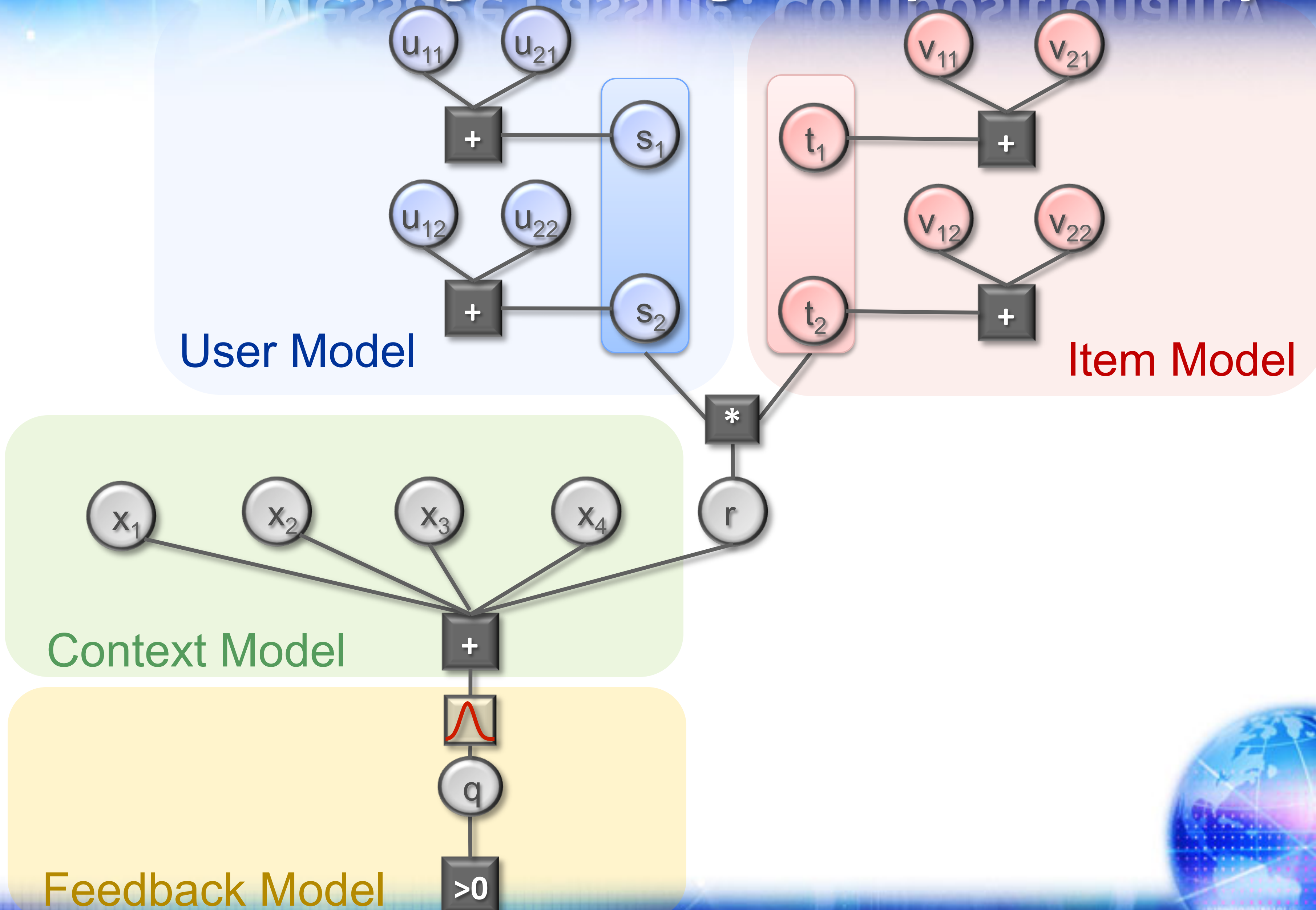
Feedback Models



Feedback Models



Message Passing: Compositionality



Learning to Play Go

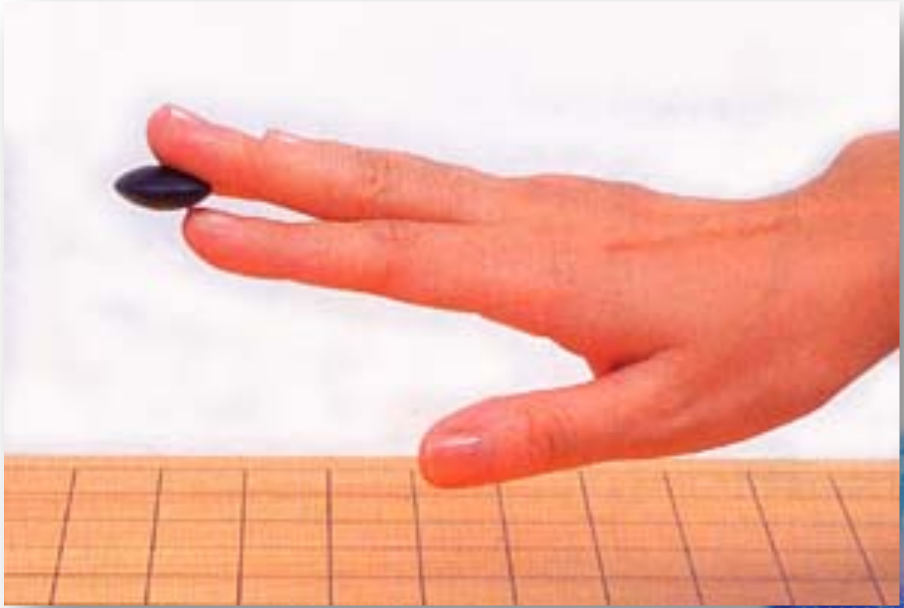
Joint work with Thore Graepel & David Stern

Uncertainty in Go

- Go is game of perfect information.
- Complexity of game tree +
  limited computer speed → uncertainty.
- 味 'aji' = 'taste'.
- Our Approach:
  Represent uncertainty using probabilities.

Machine Learning

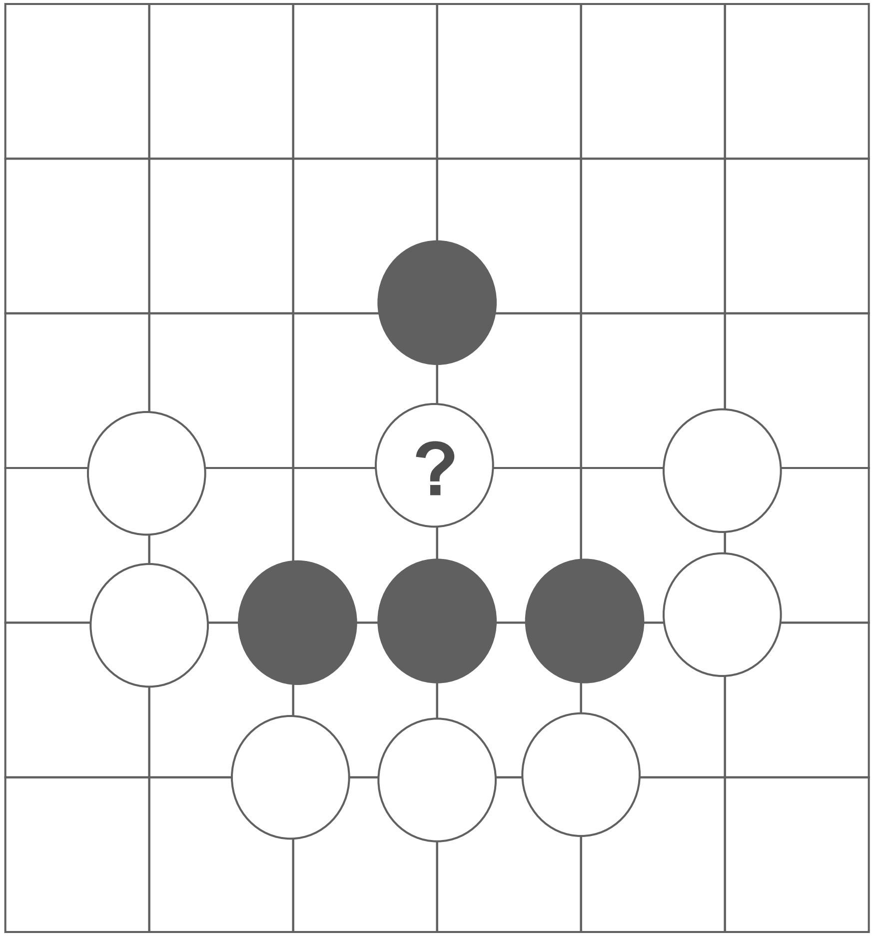
- Automatic knowledge Acquisition.
- Principled management of uncertainty.
- Applications to Go:
  - **Move Prediction.**
  - **Tactical Search.**
  - **Territory Prediction.**
  - **Monte Carlo Go.**



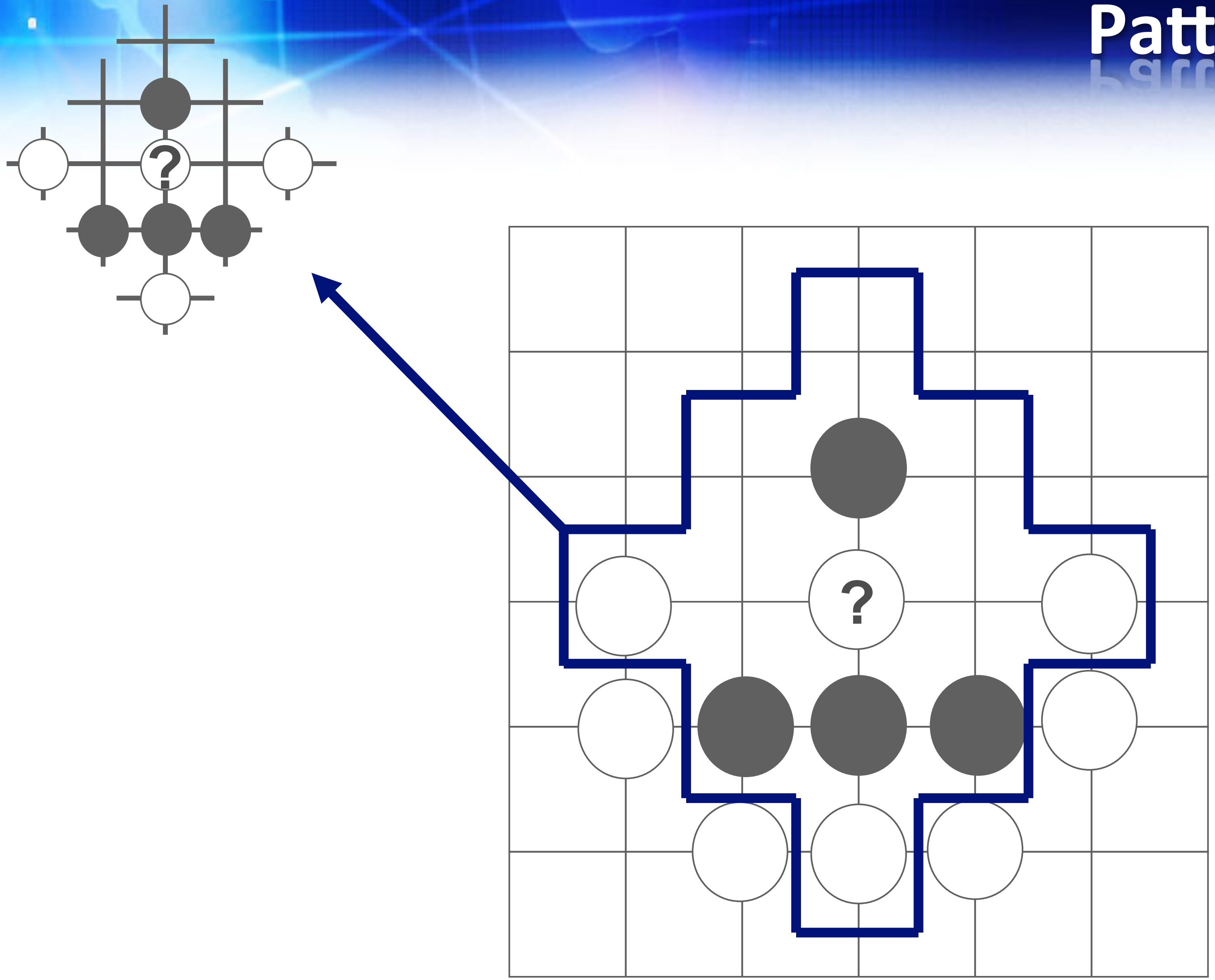
Move Prediction

- Learning from Expert Game Records
- Move associated with a set of patterns.
  - **Exact arrangement of stones.**
  - **Centred on proposed move.**
- Sequence of nested templates.
- Inspired by work by David Stoutamire and Frank de Groot

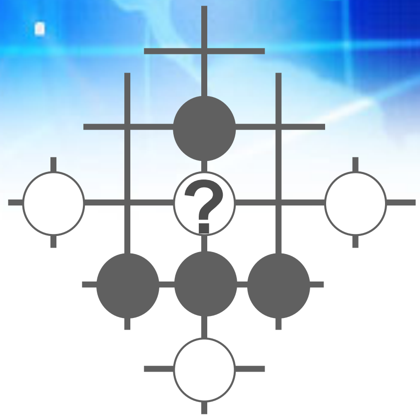
Patterns

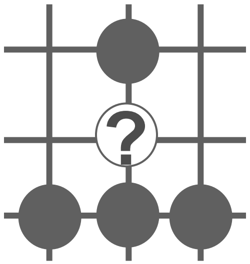


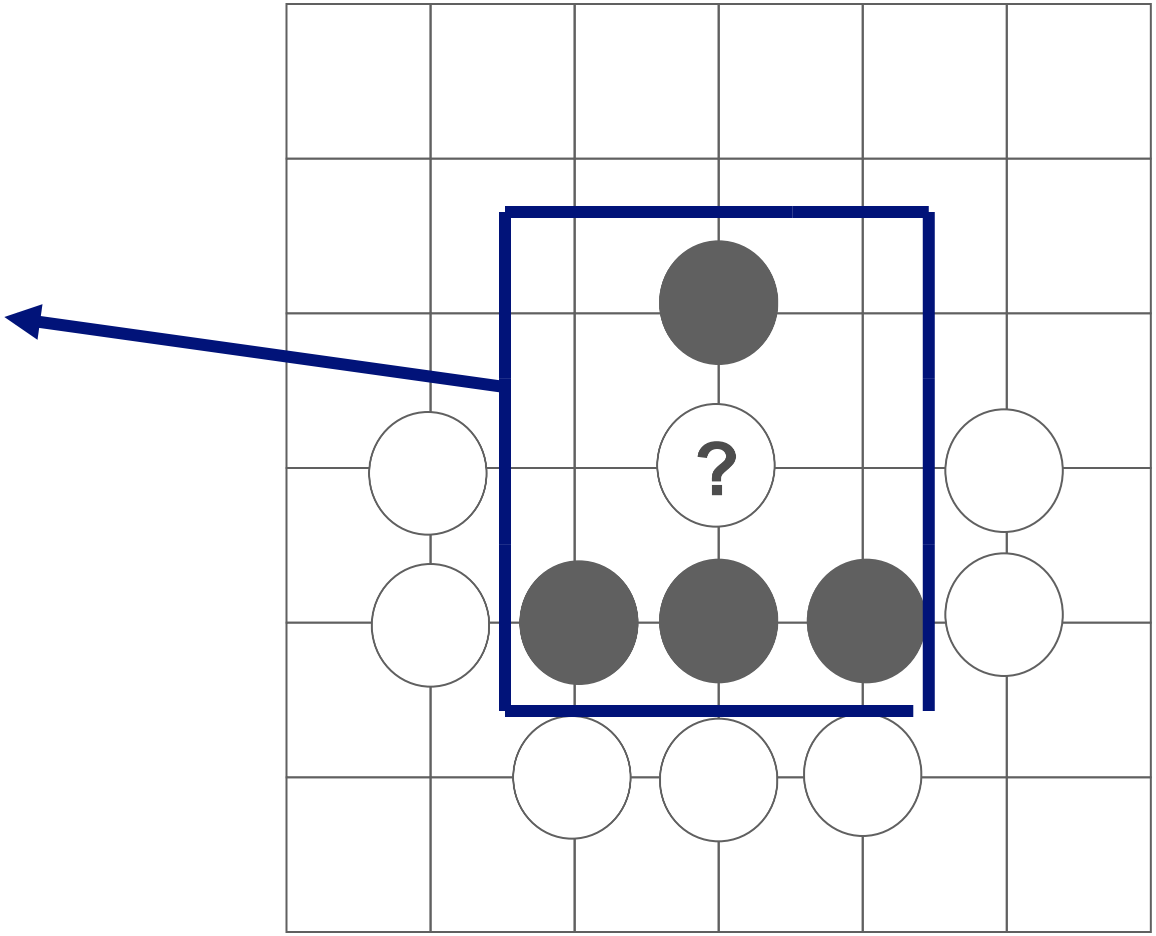
Patterns



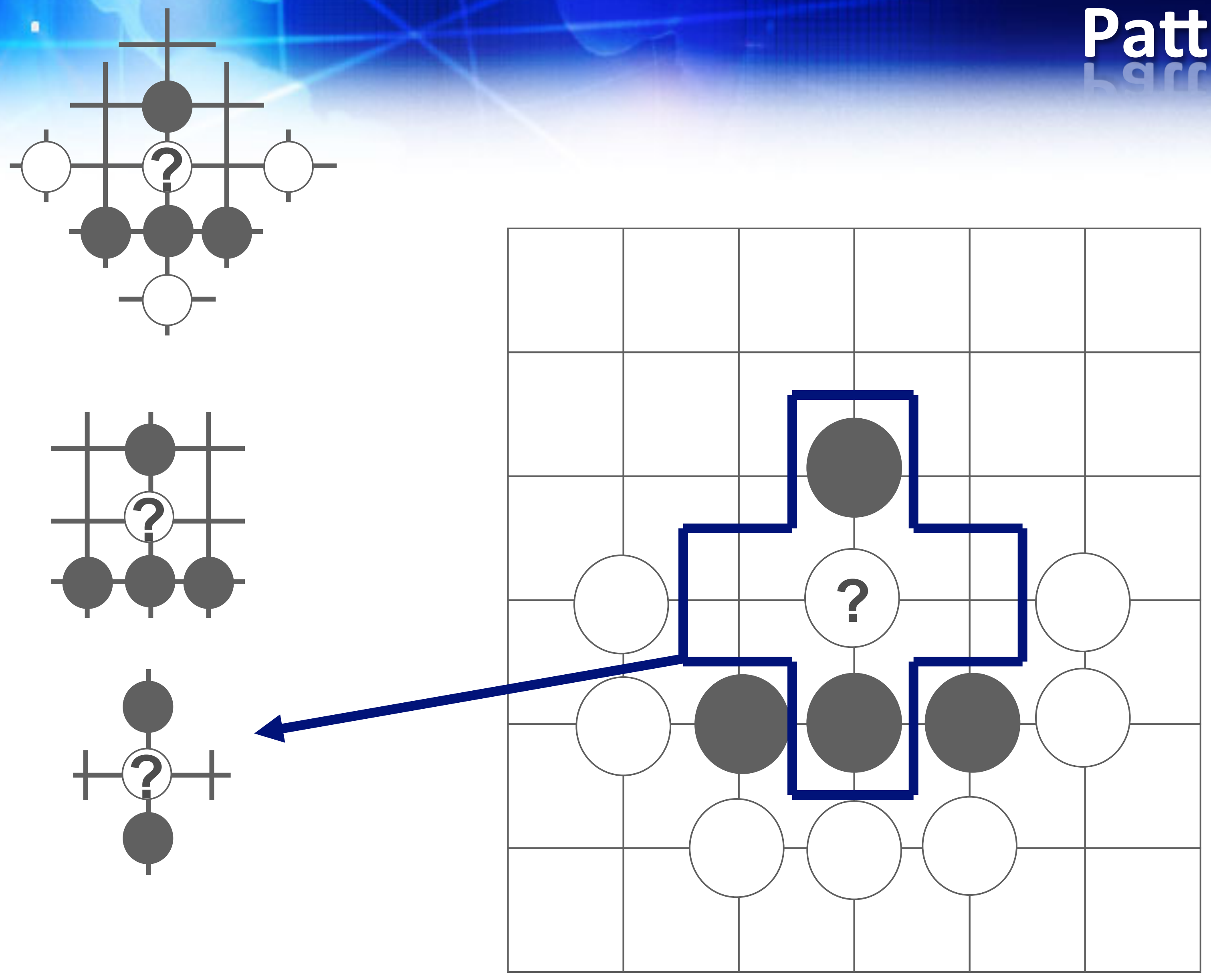
Patterns



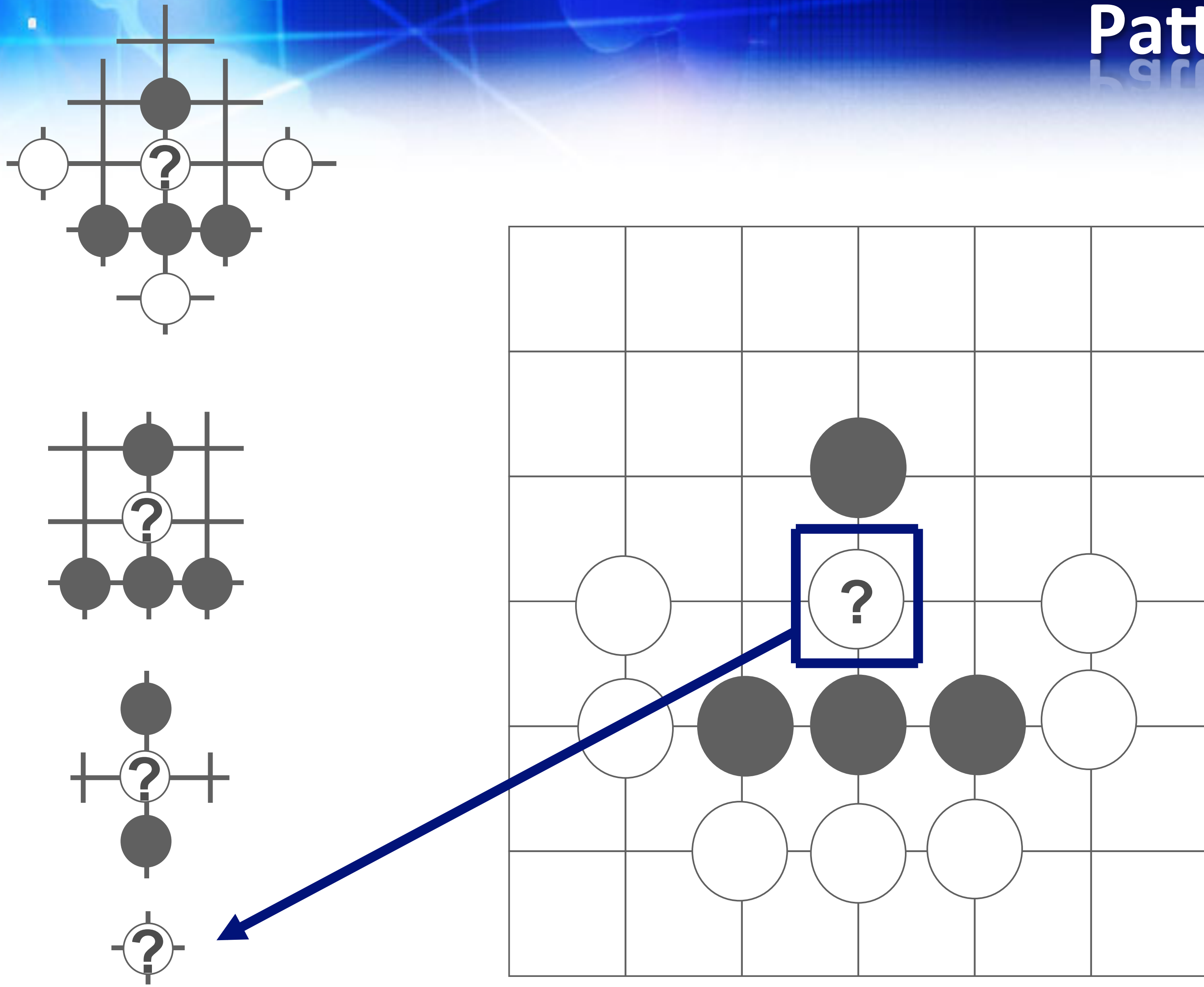




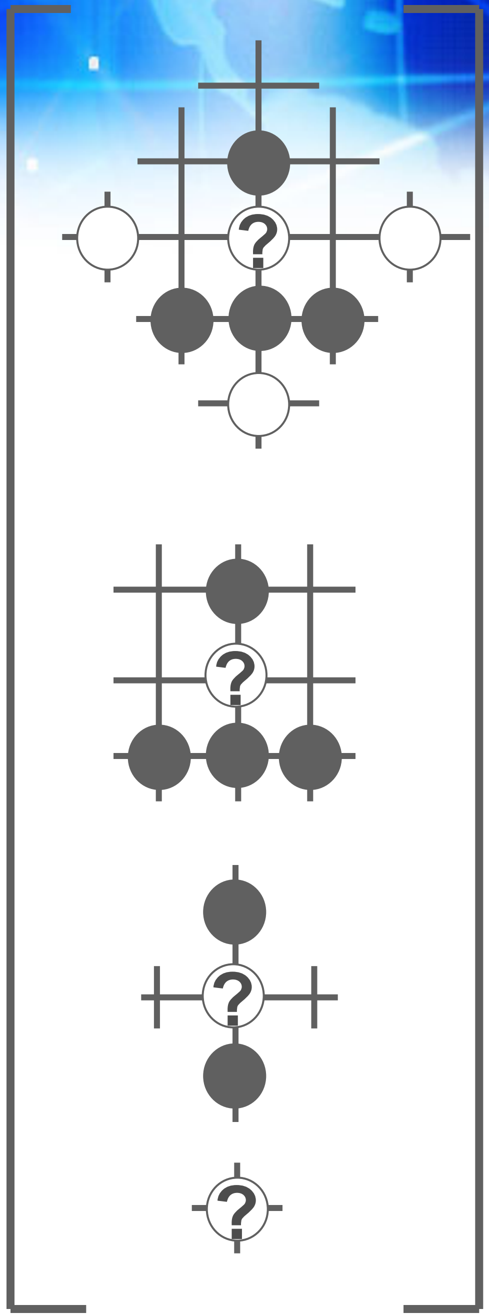
Patterns

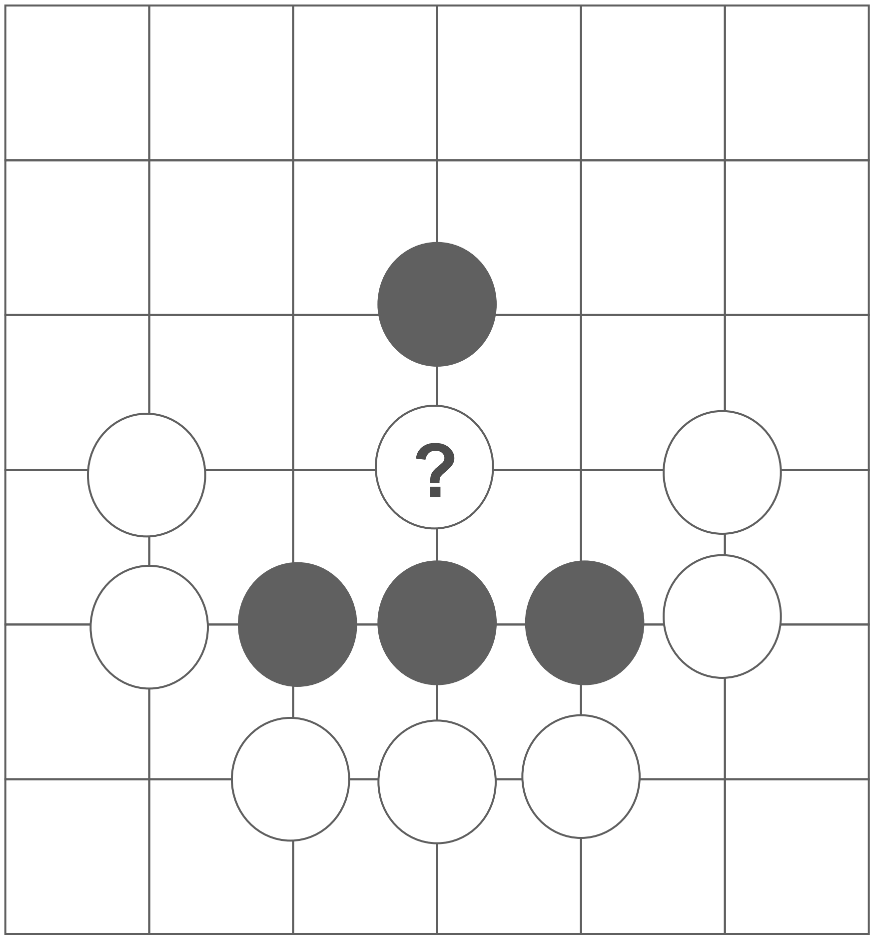


Patterns

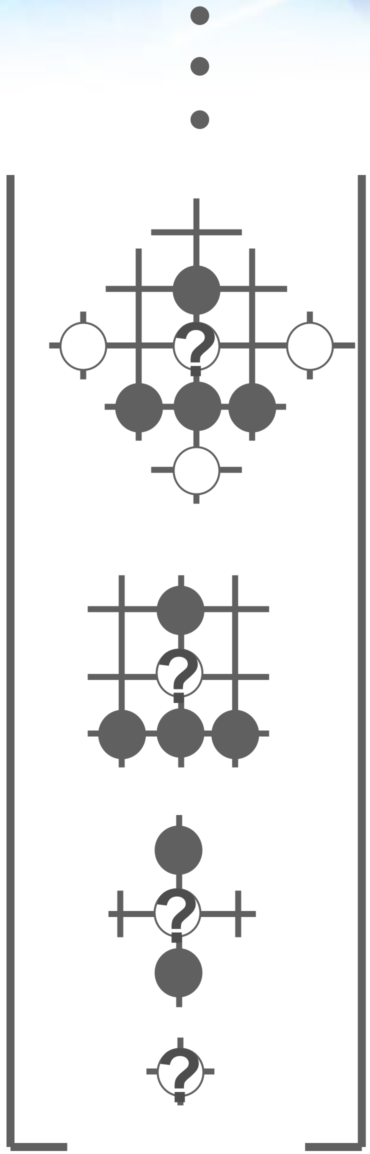


Patterns





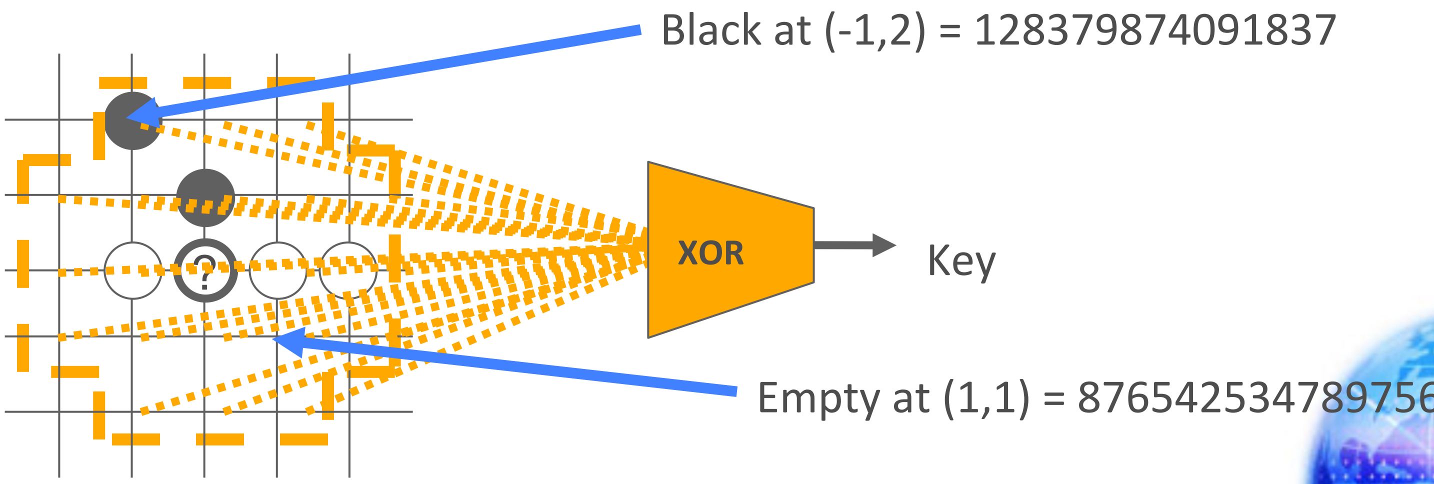
Patterns



- 13 Pattern Sizes
  - **Smallest is vertex only.**
  - **Biggest is full board.**

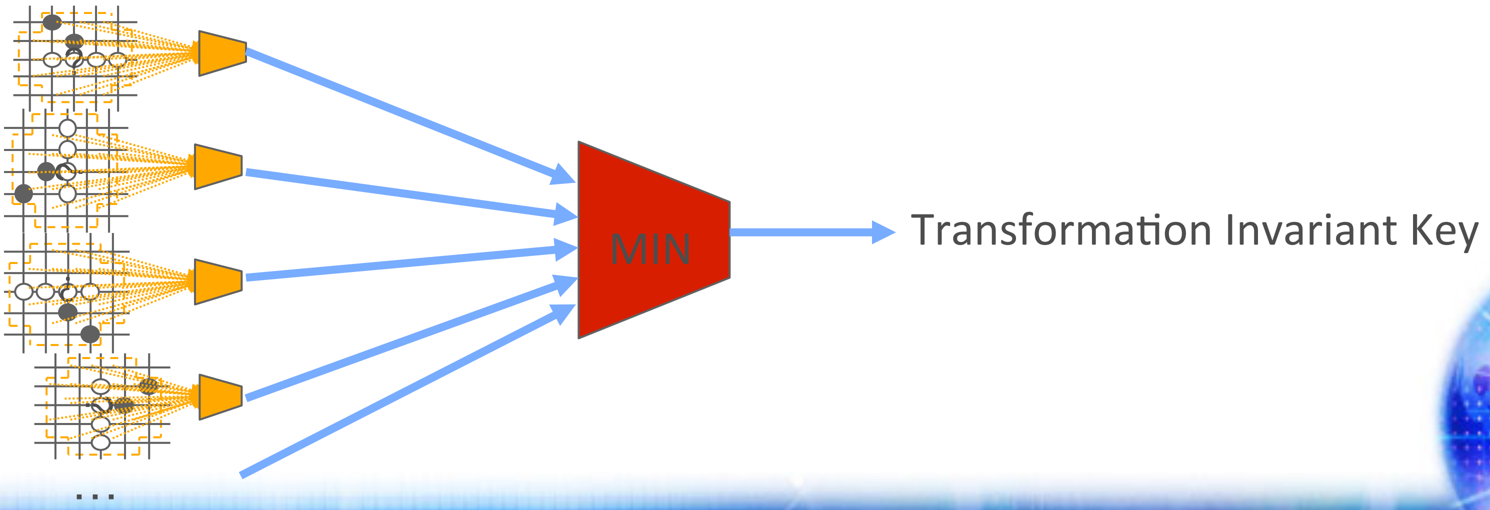
Pattern Matching

- **Goal**: Pattern information stored in hash table.
- **Idea**: 64 bit random numbers for each template vertex: One for each of {black, white, empty, off}.
- Combine with XOR (Zobrist, 1970).



Pattern Hash Key

- **Goal**: Pattern information stored in hash table.
- **Idea**: 64 bit random numbers for each template vertex: One for each of {black, white, empty, off}.
- Combine with XOR (Zobrist, 1970).



Harvesting

- **Data Size**: 180,000 games × 250 moves × 13 pattern sizes…

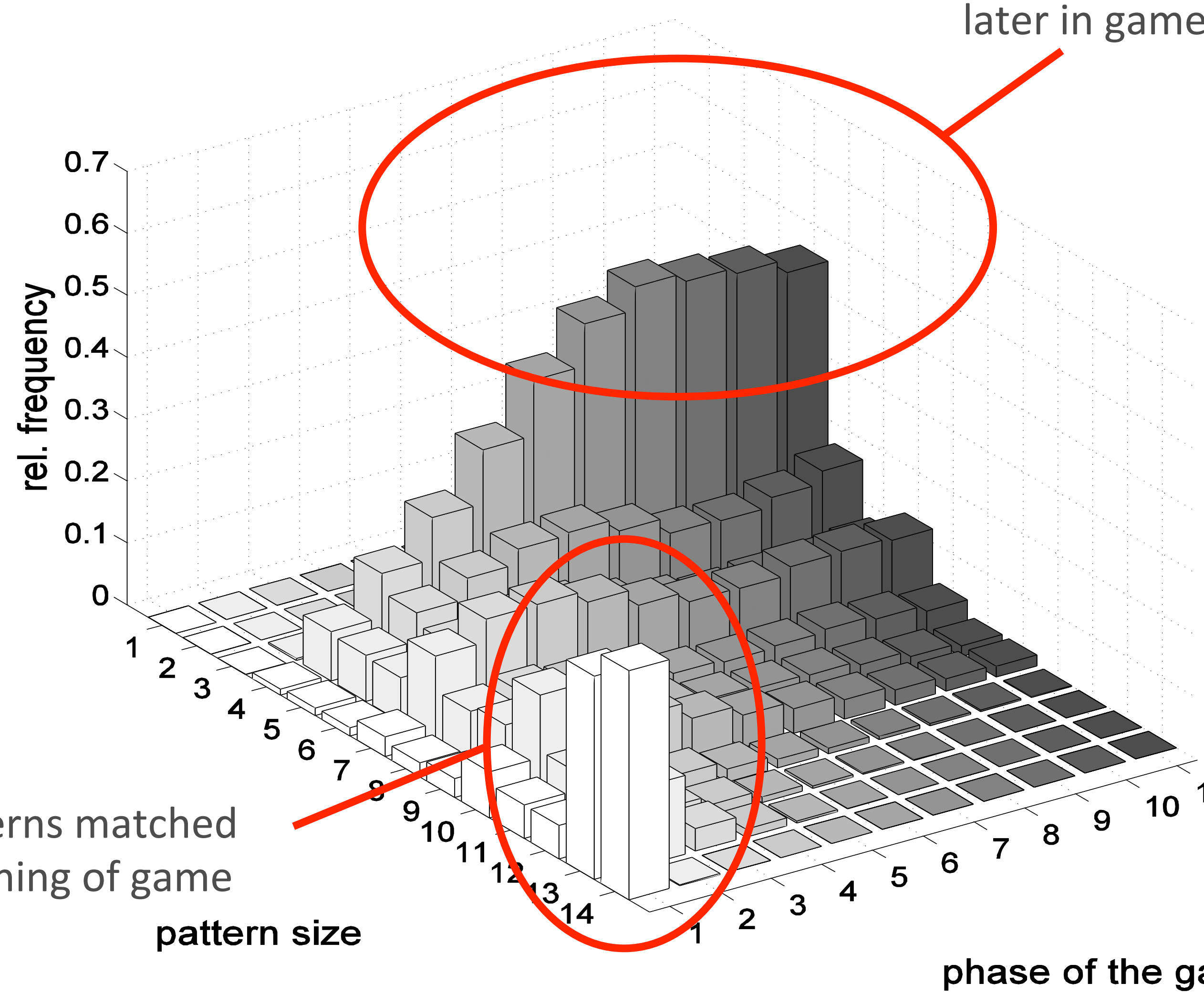
  …gives **600 million potential patterns**

- **Problem**: Need to limit number stored.

- **Idea**: Keep patterns played more than n times.
- **Bloom filter**: Approximate test for set membership with minimal memory footprint.

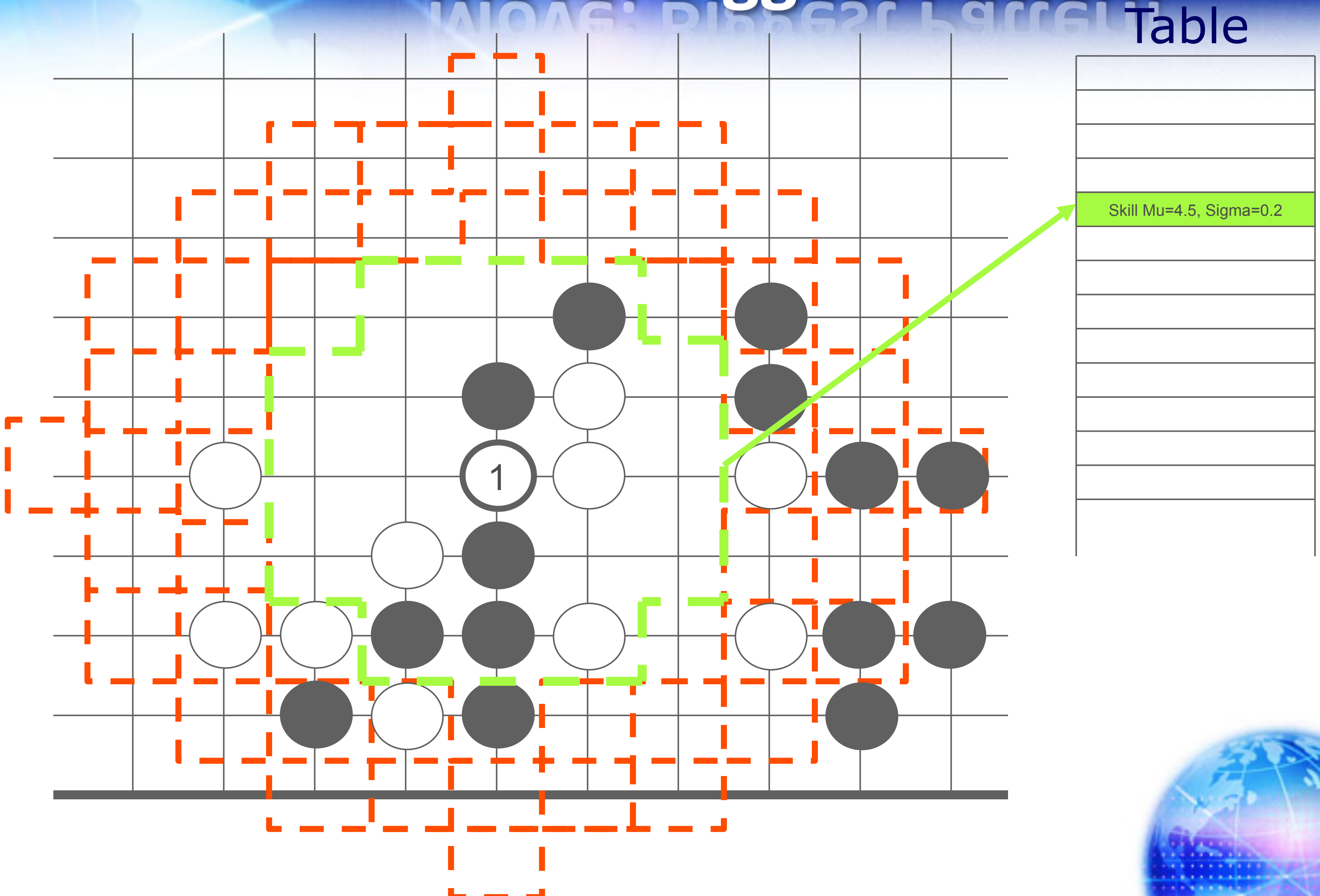
Relative Frequencies of Pattern Sizes

Smaller patterns matched later in game.

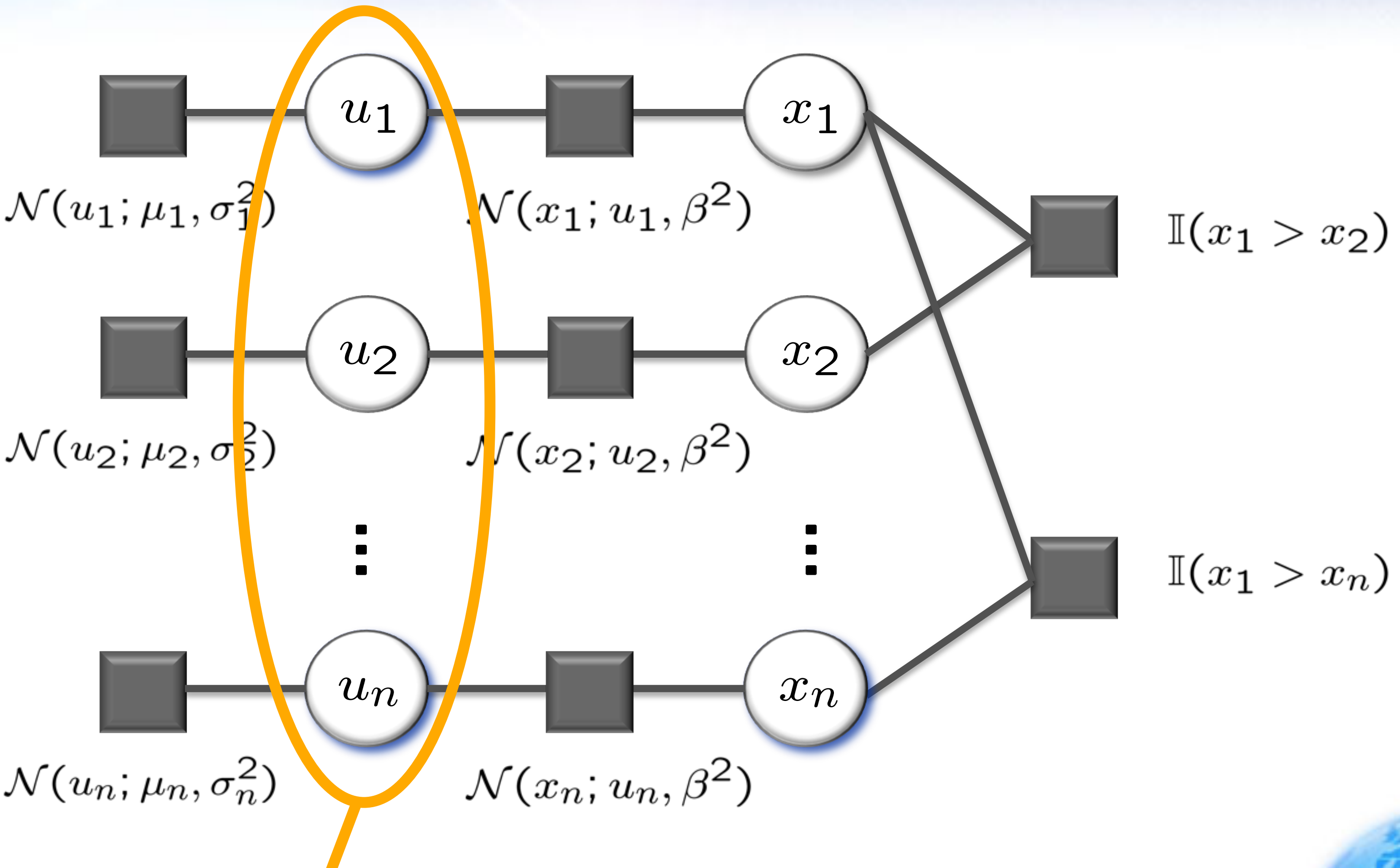


Big patterns matched at beginning of game

Move: Biggest Pattern

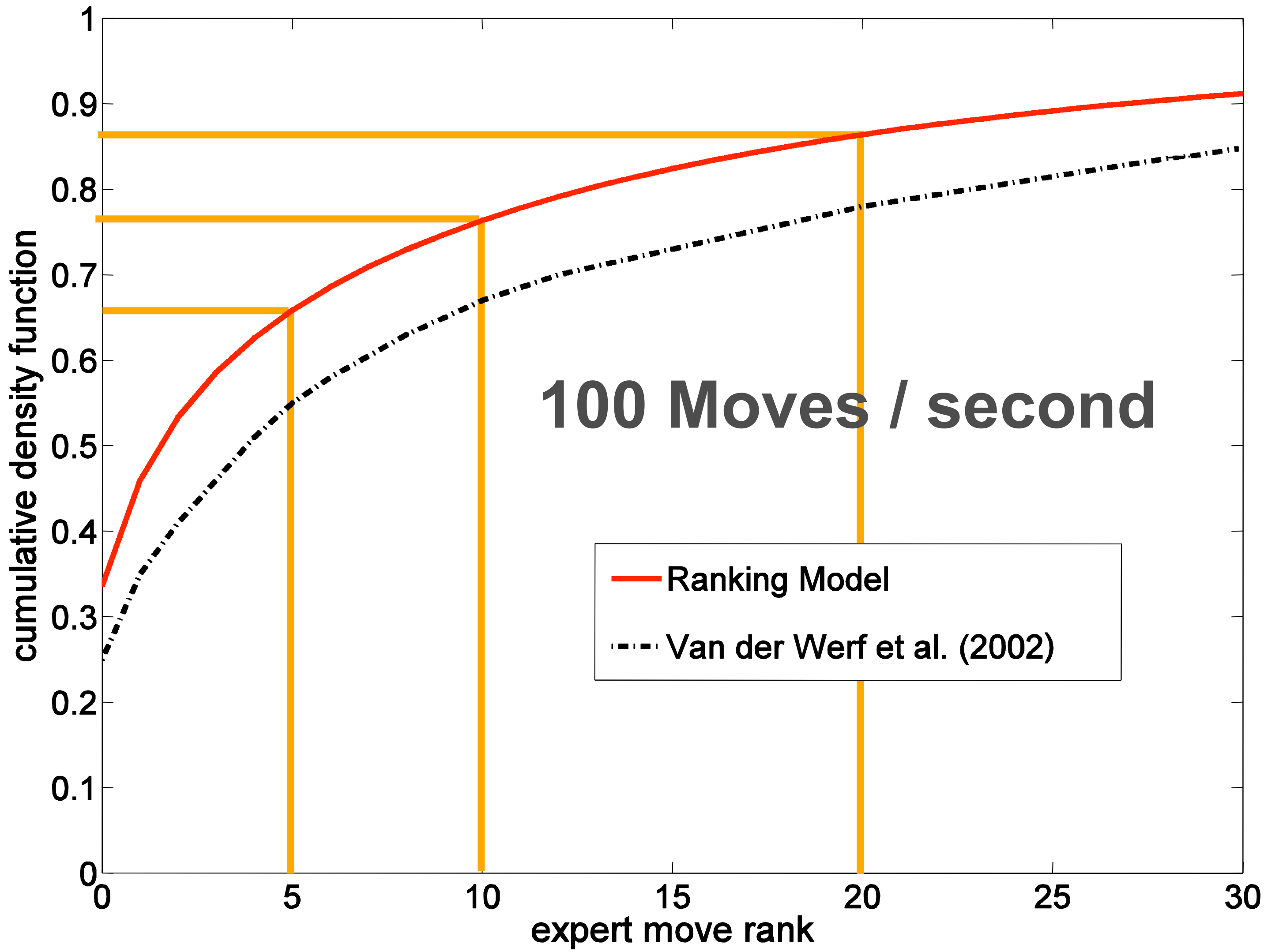


Bayesian Ranking Model

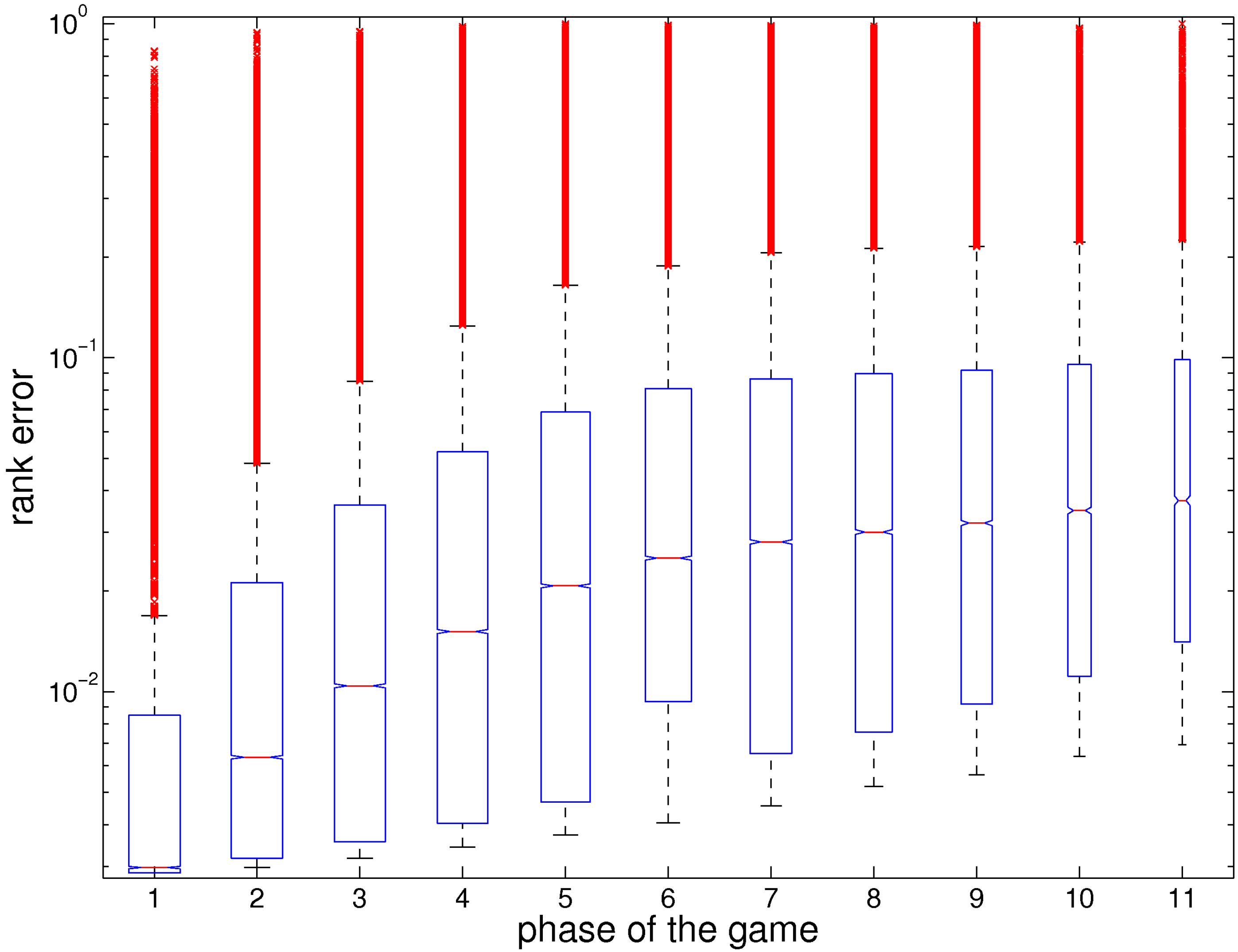


$$p(\mathbf{u}|\text{move, position}) = \int p(\mathbf{u}, \mathbf{x}|\text{move, position}) d\mathbf{x}$$

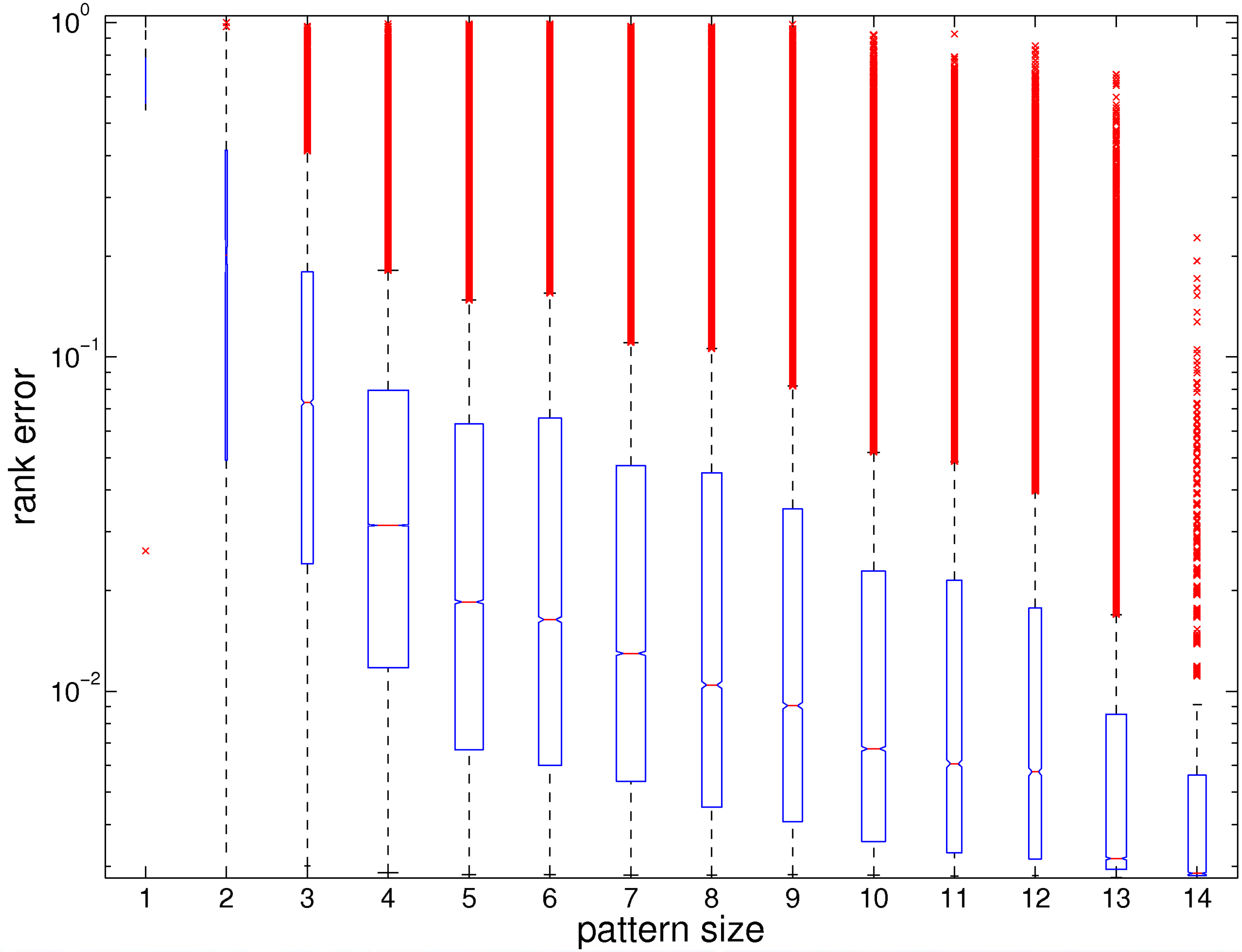
Move Prediction Performance



Rank Error vs Game Phase



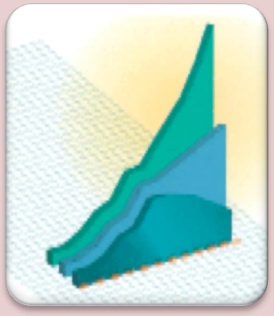
Rank Error vs Pattern Size

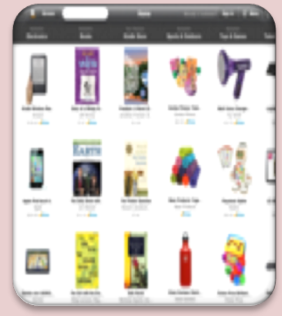


Overview

- Theory
  - Inference in Factor Graphs
  - Approximate Message Passing
- Applications @ Microsoft
  - TrueSkill: Gamer Rating and Matchmaking
  - TrueSkill Through Time: History of Chess
  - Click-Through Rate Prediction in Online Advertising
  - Matchbox: Recommendation Systems
- **Applications @ Amazon**

ML Opportunities @ Amazon





Retail

- Demand Forecasting
- Vendor Lead Time Prediction
- Pricing
- Packaging
- Substitute Prediction

Customers

- Product Recommendation
- Product Search
- Visual Search
- Product Ads
- Shopping Advice
- Customer Problem Detection

Seller

- Fraud Detection
- Predictive Help
- Seller Search & Crawling

Catalog

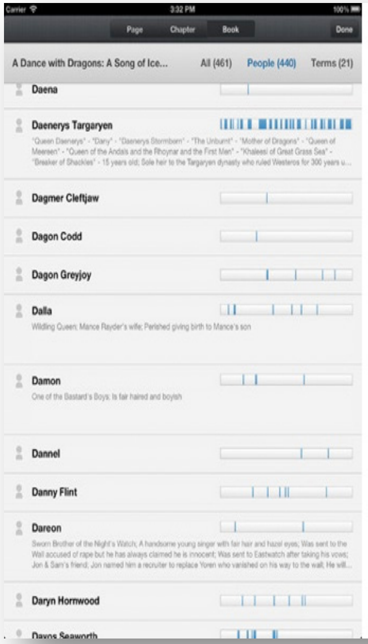
- Browse-Node Classification
- Meta-data validation
- Review Analysis

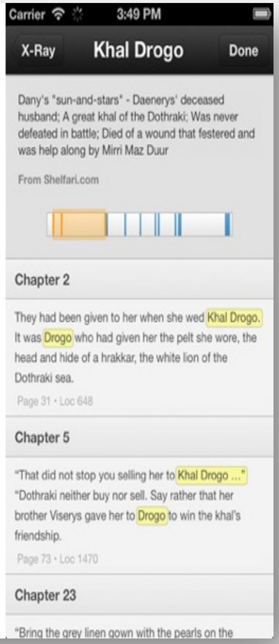
Digital

- Named-Entity Extraction
- XRay
- Plagiarism Detection

XRay


Emma's Kindle
X-Ray
IMDb
Elizabeth Banks
Effie Trinket
Jennifer Lawrence
Katniss Everdeen
See Full Cast
10s
The Hunger Games
00:12
HD
02:04


Page
Chapter
Book
Done
A Dance with Dragons: A Song of Ice...
All (461)
People (440)
Terms (21)
Daena
Daenerys Targaryen
Dagmer Cleftjaw
Dagon Codd
Dagon Greyjoy
Dalla
Damon
Dannel
Danny Flint
Dareon
Daryn Hornwood
Davos Seaworth


Carrier
3:49 PM
X-Ray
Khal Drogo
Done
Dany's "sun-and-stars" - Daenerys' deceased husband; A great khal of the Dothraki; Was never defeated in battle; Died of a wound that festered and was help along by Mirri Maz Duur
From Shelfari.com
Chapter 2
They had been given to her when she wed Khal Drogo. It was Drogo who had given her the pelt she wore, the head and hide of a hrakkar, the white lion of the Dothraki sea.
Page 31 • Loc 648
Chapter 5
"That did not stop you selling her to Khal Drogo ..."
"Dothraki neither buy nor sell. Say rather that her brother Viserys gave her to Drogo to win the khal's friendship.
Page 73 • Loc 1470
Chapter 23
"Bring the grey linen gown with the pearls on the

Machine Translation

This product page was automatically translated.
Was this translation helpful? Yes or No

1432 Girali (Grasping toy) - Selecta Wooden Toys/Selecta Spielzeug

by Selecta Spielzeug

3 customer reviews

Price: £12.03 & **FREE Delivery** in the UK. Details

Only 7 left in stock.

Sold by Alle-Spielwaren and Fulfilled by Amazon. Gift-wrap available.

Want it delivered to Germany by tomorrow, 18 March? Order within 5 hrs 41 mins and choose **One-Day Delivery to Germany** at checkout. Details

18 new from £7.11

- 10 cm / 4 in.
- This classic series of grasping toys has been perfected by Selecta for over 30 years.

› See more product details

Machine Translation: Deep Dive

$$p(English \mid Chinese) = \frac{p(English) \times p(Chinese \mid English)}{p(Chinese)}$$

$$\propto \underbrace{p(English)}_{\text{Language Model}} \times \underbrace{p(Chinese \mid English)}_{\text{Translation Model}}$$

- **Language Model:** What are good English sentences?
- **Translation Model:** What English sentences account well for a given Chinese sentence?

Translation Model

Although north wind howl , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

$$p(English \mid Chinese) = ?$$

Thanks!