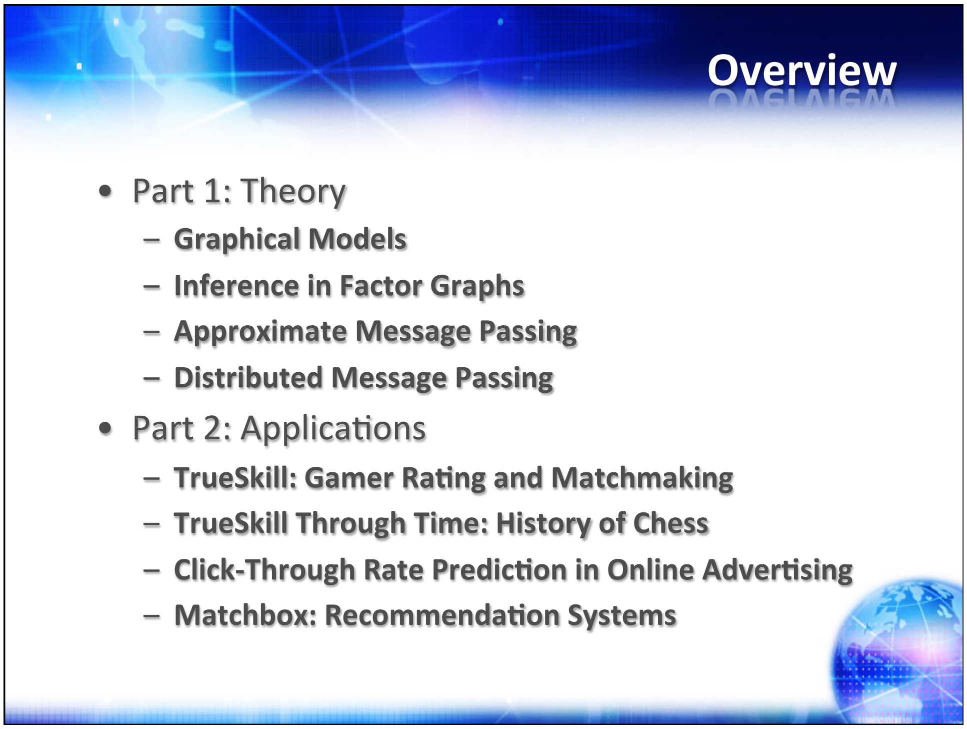
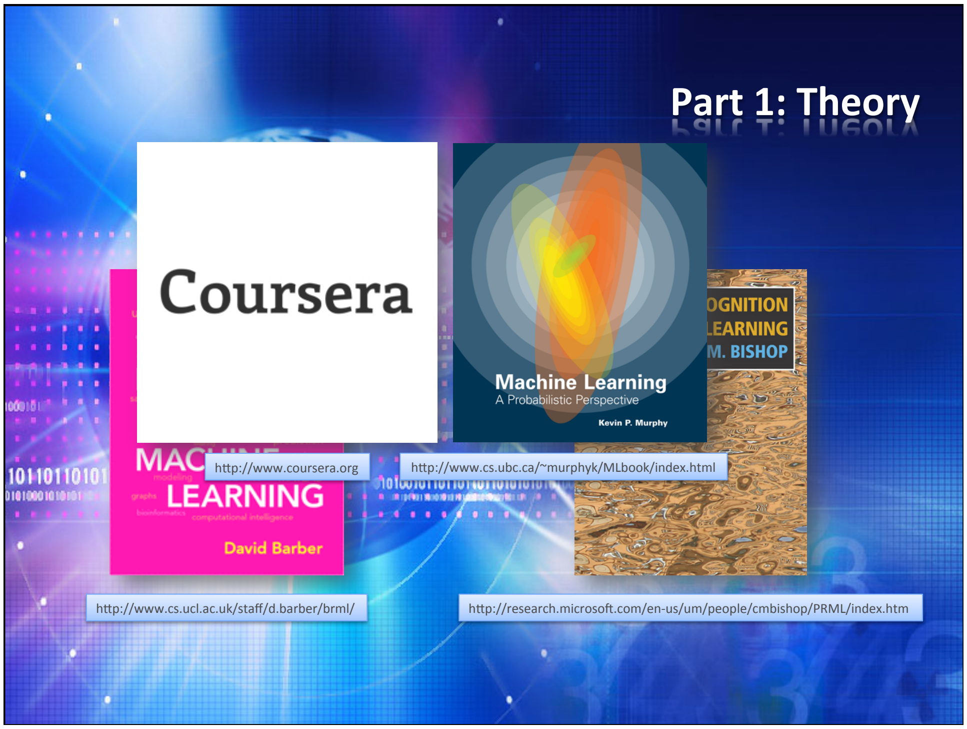
Distributed, Real-Time Bayesian Learning in Online Services

Ralf Herbrich
Amazon



Overview

- Part 1: Theory
  - Graphical Models
  - Inference in Factor Graphs
  - Approximate Message Passing
  - Distributed Message Passing
- Part 2: Applications
  - TrueSkill: Gamer Rating and Matchmaking
  - TrueSkill Through Time: History of Chess
  - Click-Through Rate Prediction in Online Advertising
  - Matchbox: Recommendation Systems



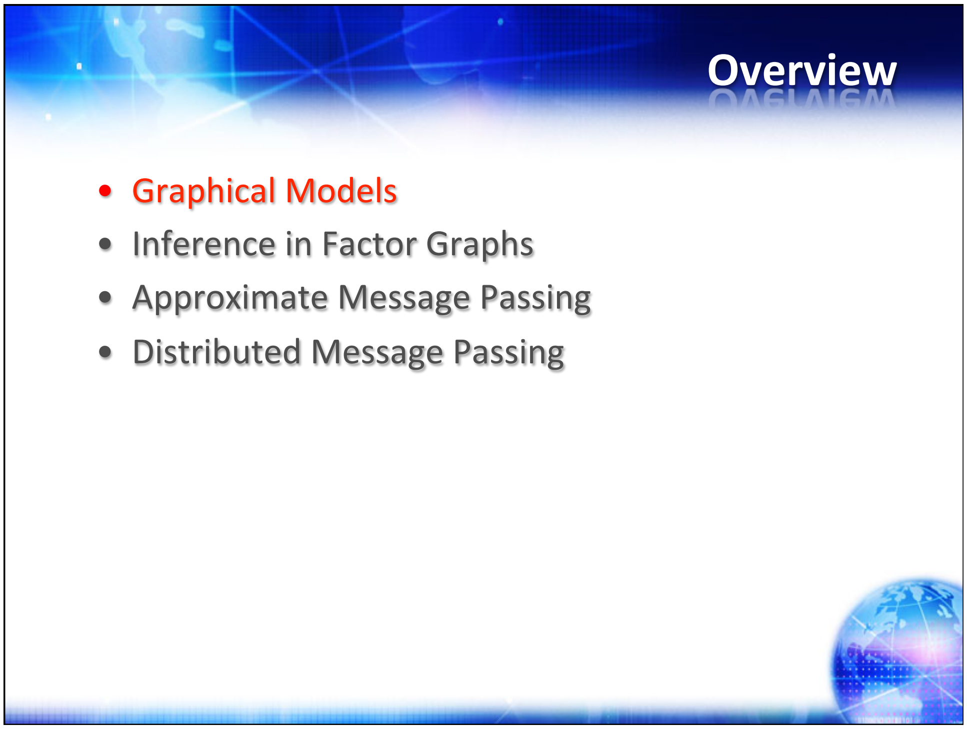
Part 1: Theory

http://www.coursera.org

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

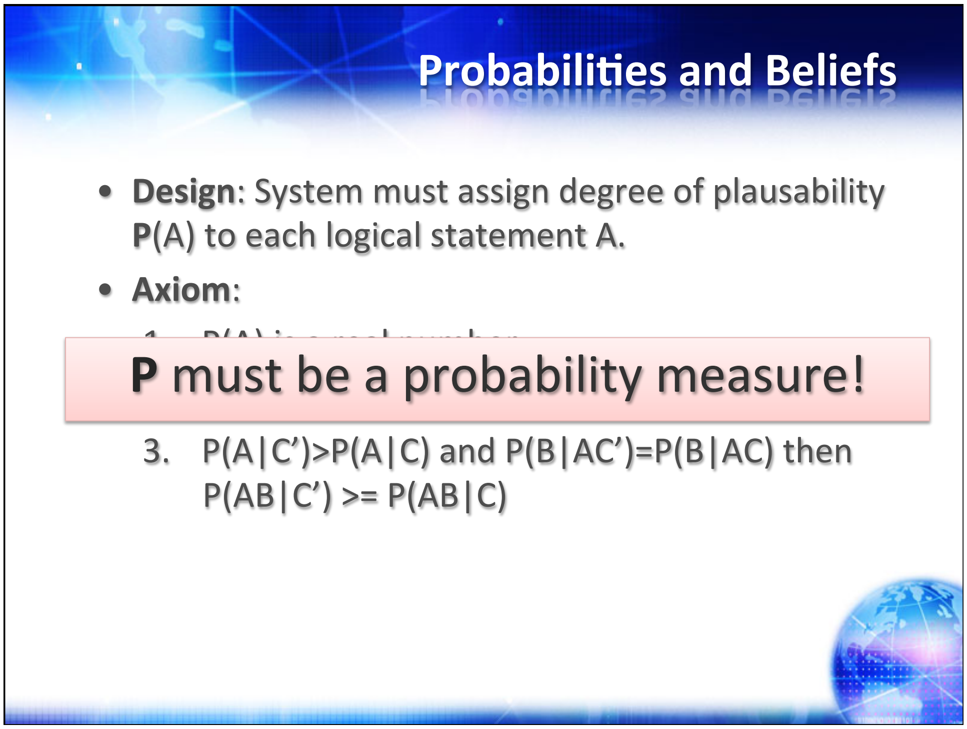
http://www.cs.ucl.ac.uk/staff/d.barber/brml/

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



Overview

- Graphical Models
- Inference in Factor Graphs
- Approximate Message Passing
- Distributed Message Passing

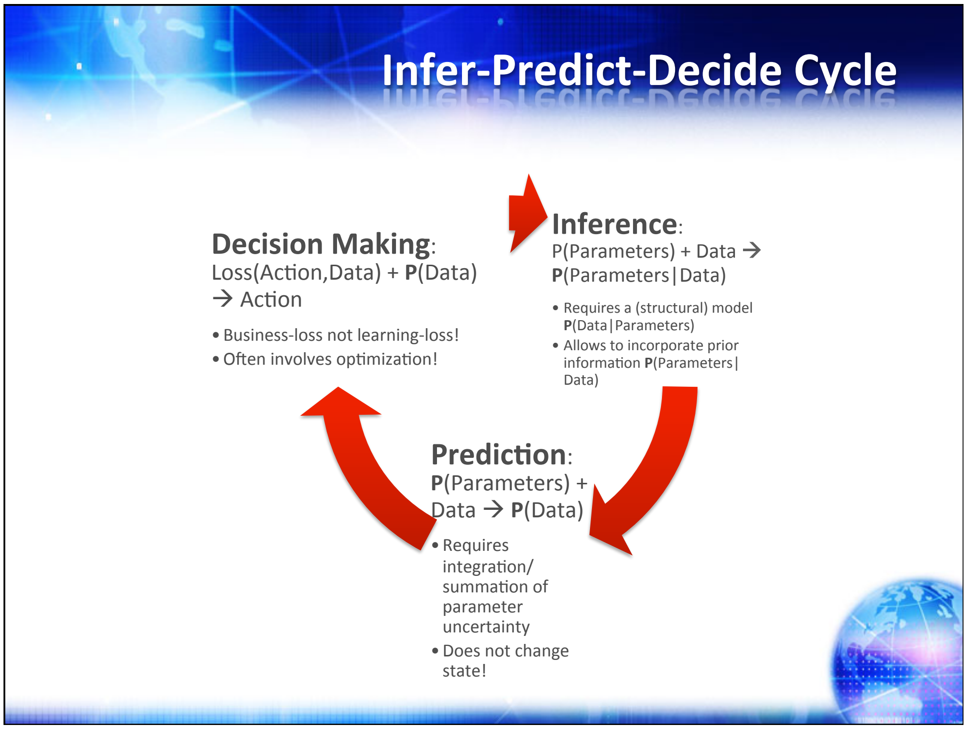


Probabilities and Beliefs

- **Design**: System must assign degree of plausability **P**(A) to each logical statement A.
- **Axiom**:
  1. P(A) is a real number

P must be a probability measure!

  3. P(A|C’)>P(A|C) and P(B|AC’)=P(B|AC) then P(AB|C’) >= P(AB|C)



Infer-Predict-Decide Cycle

Inference:
P(Parameters) + Data → **P**(Parameters|Data)

- Requires a (structural) model **P**(Data|Parameters)
- Allows to incorporate prior information **P**(Parameters|Data)

Prediction:
P(Parameters) + Data → **P**(Data)

- Requires integration/ summation of parameter uncertainty
- Does not change state!

Decision Making:
Loss(Action,Data) + **P**(Data) → Action

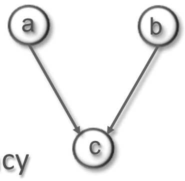
- Business-loss not learning-loss!
- Often involves optimization!

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*: $p(a,b|c) \overset{?}{=} p(a|c) \cdot p(b|c)$
  - *Inference*/Marginalisation: $p(a,b) = \sum_c p(a,b,c)$

Directed Models: Bayesian Networks

- **Definition:** Graphical representation of joint probability distribution (Pearl, 1988)
  - Nodes: ○ = Variables
  - Directed Edges: Conditional probability distribution
- **Semantic:**



$$p(\mathbf{x}) = \prod_i p\left(x_i | \mathbf{x}_{\text{parents}(i)}\right)$$

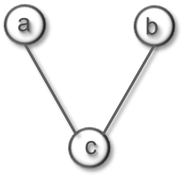
  - Ancestral relationship of dependency

$$p(a,b,c) = p(a) \cdot p(b) \cdot p(c|a,b)$$

Undirected Models: Markov Networks

- **Definition:** Graphical representation of joint probability distribution (Pearl, 1988)
  - Nodes: ○ = Variables
  - Edges: Dependency between variables
- **Semantic:**

$$p(\mathbf{x}) = \frac{1}{Z} \cdot \prod_{C} \phi(x_C) \qquad \phi \geq 0$$

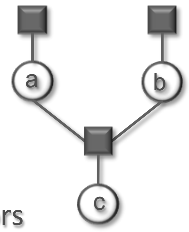


  - Local potentials over cliques

$$\begin{aligned} p(a,b,c) &= \frac{1}{Z} \cdot \phi_{ac}(a,c) \cdot \phi_{bc}(b,c) \\ Z &= \sum_a \sum_b \sum_c \phi_{ac}(a,c) \cdot \phi_{bc}(b,c) \end{aligned}$$

Factor Graphs

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

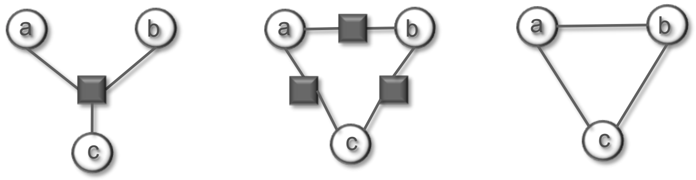


$$p(\mathbf{x}) = \prod_{f} f\left(\mathbf{x}_{V(f)}\right)$$

  - Local variable dependency of factors

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

Factor Graphs are Powerful!



$f_1(a, b, c)$ $\quad f_1(a, b) \cdot f_2(b, c) \cdot f_3(a, c)$ $\quad \phi(a, b, c)$

Undirected graphical models can hide the factorisation within a clique!

Factor Graphs and Bayes' Law

- Bayes' law

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

- Factorising prior

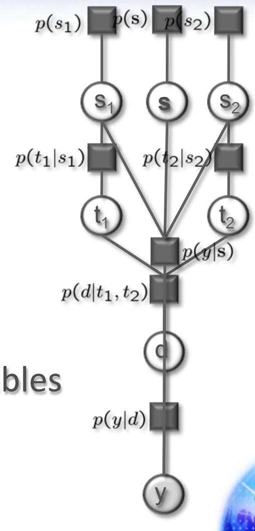
$$p(\mathbf{s}) \;=\; p(s_1) \cdot p(s_2)$$

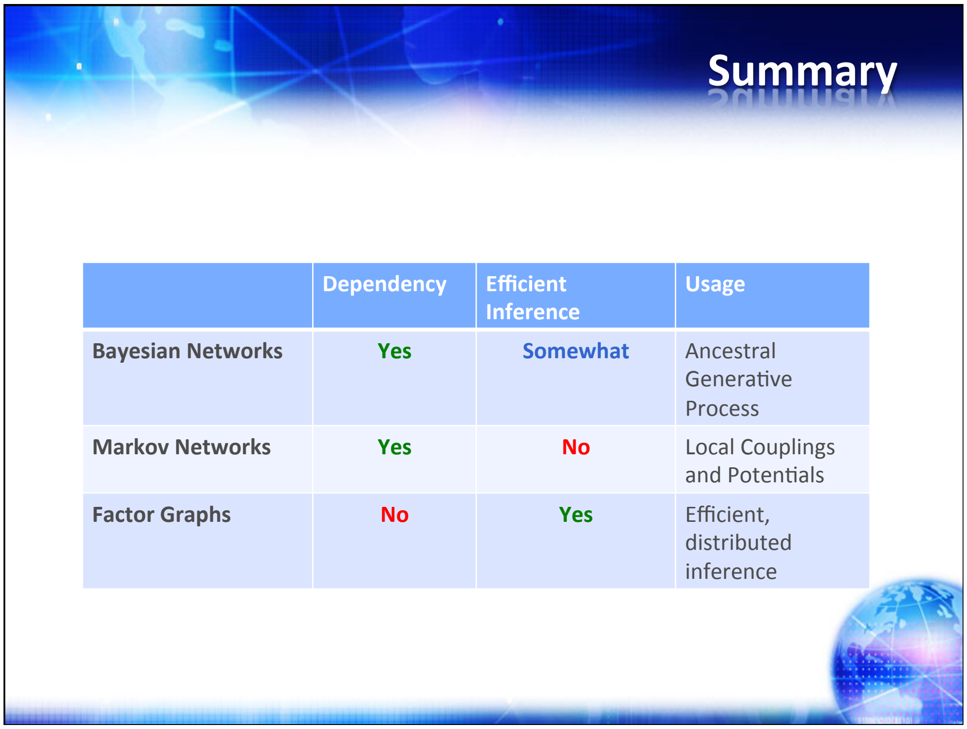
- Factorising likelihood

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

- Inference: Sum out latent variables

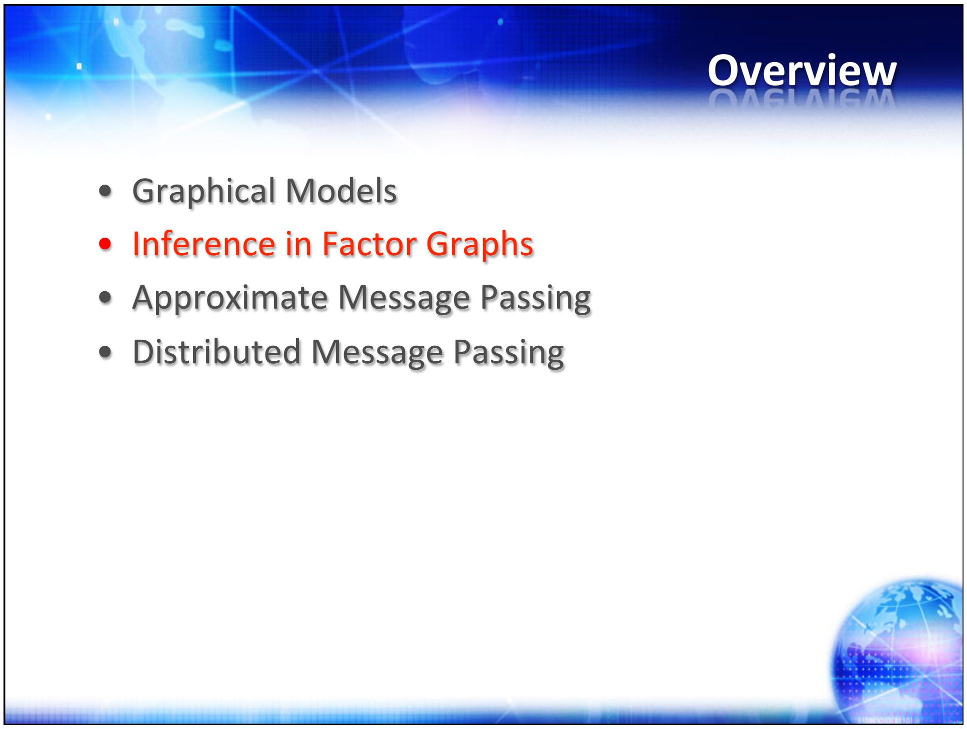
$$p(y|\mathbf{s}) \;=\; \sum_{\mathbf{t}} \sum_{d} p(y, \mathbf{t}, d|\mathbf{s})$$





Summary

	Dependency	Efficient Inference	Usage
Bayesian Networks	Yes	Somewhat	Ancestral Generative Process
Markov Networks	Yes	No	Local Couplings and Potentials
Factor Graphs	No	Yes	Efficient, distributed inference



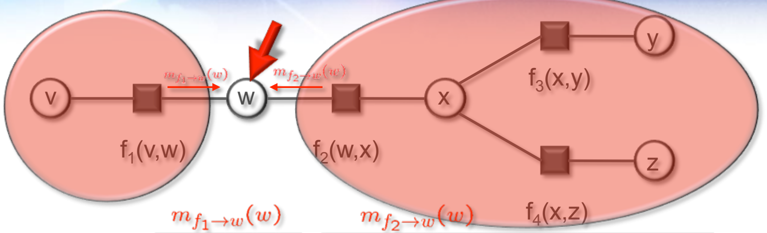
Overview

- Graphical Models
- Inference in Factor Graphs
- Approximate Message Passing
- Distributed Message Passing

Factor Graphs and Factor Trees

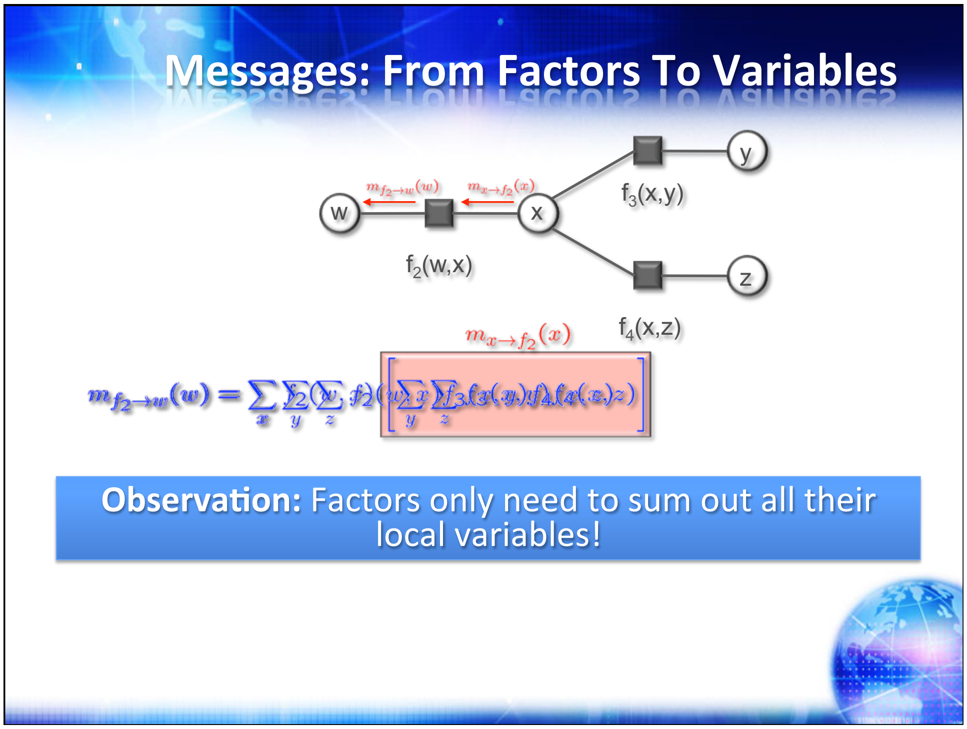
- **Factor Graphs:** Arbitrary functions
  - Bayesian Networks
  - Markov Networks
- **Factor Trees**: Functions where the variable indices never decrease from left to right
- **Factor Graph ➔ Factor Tree**:
  1. Pick an arbitrary node
  2. Build the spanning tree

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!



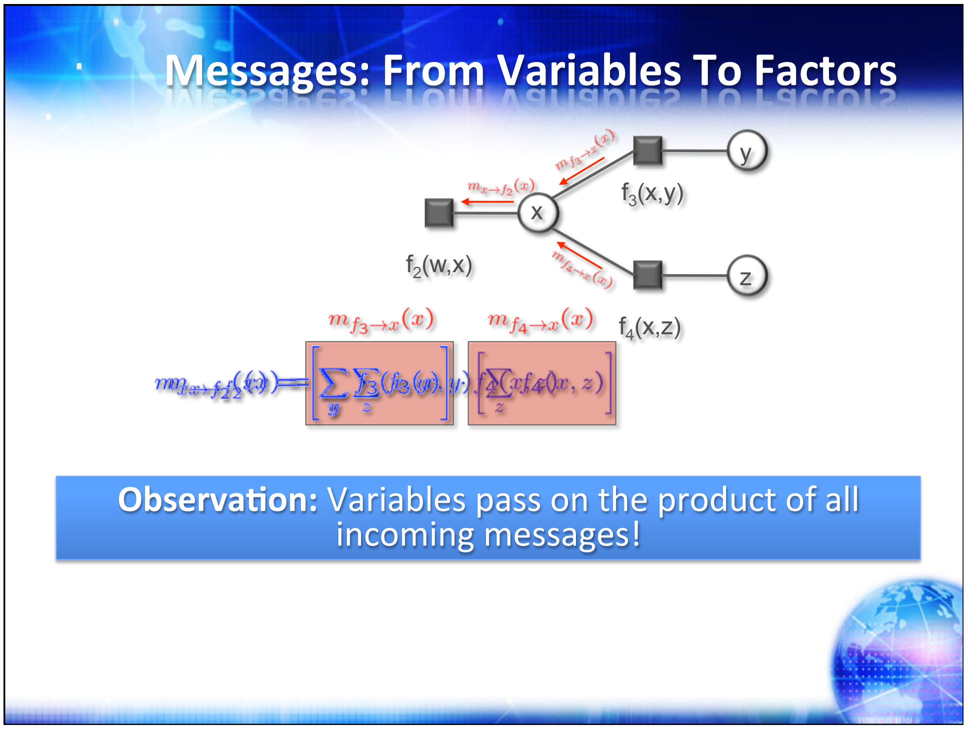
Messages: From Factors To Variables

$m_{x \to f_2}(x)$

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

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

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



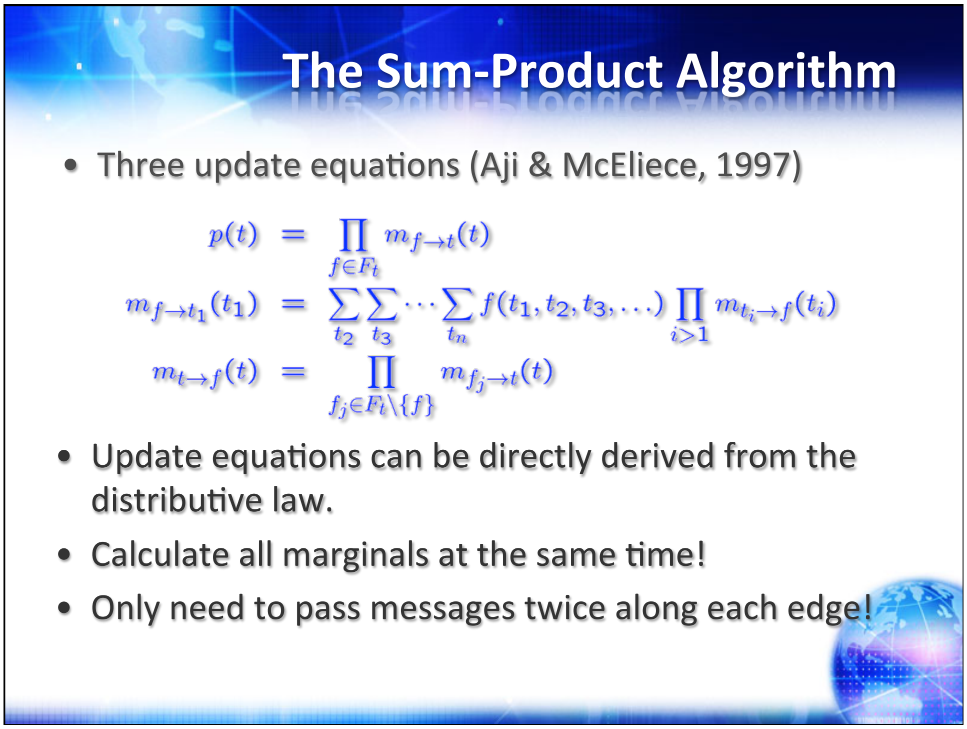
Messages: From Variables To Factors

$m_{f_3 \to x}(x)$ $m_{f_4 \to x}(x)$

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

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

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

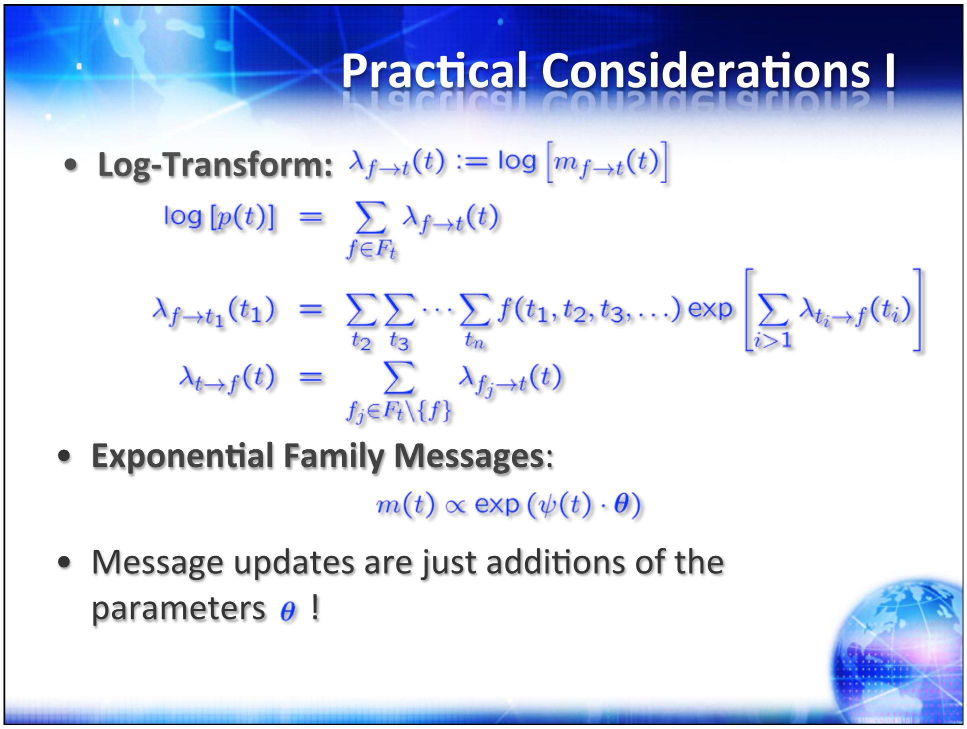


The Sum-Product Algorithm

- 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 I

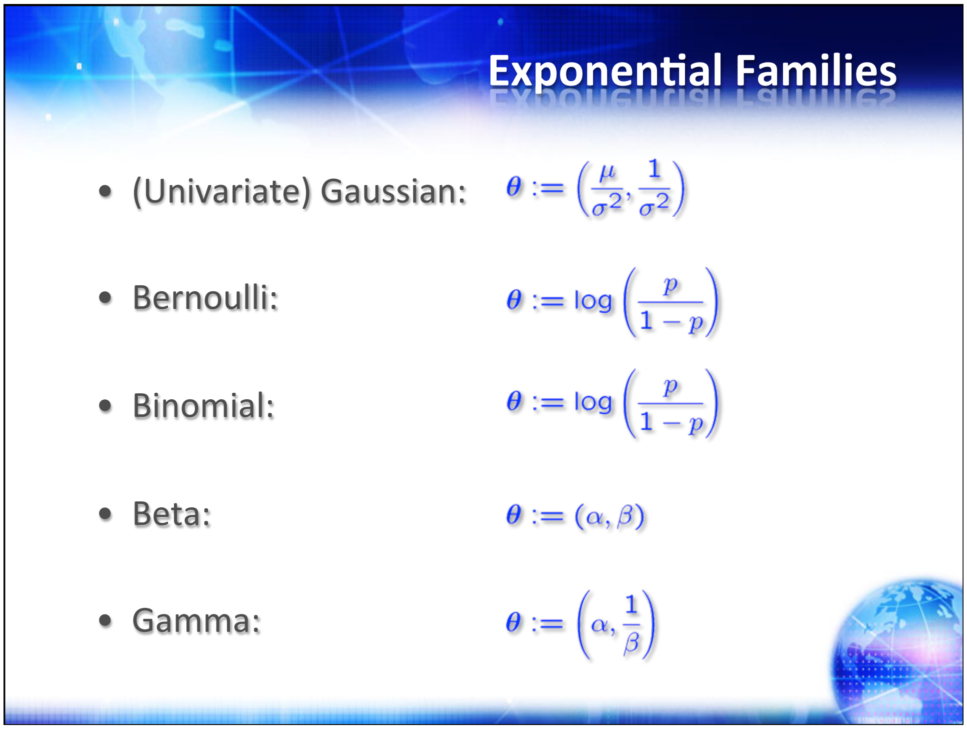
- **Log-Transform:** $\lambda_{f \to t}(t) := \log \left[m_{f \to t}(t) \right]$

$$
\begin{aligned}
\log \left[p(t) \right] &= \sum_{f \in F_t} \lambda_{f \to t}(t) \\
\lambda_{f \to t_1}(t_1) &= \sum_{t_2} \sum_{t_3} \cdots \sum_{t_n} f(t_1, t_2, t_3, \ldots) \exp \left[\sum_{i>1} \lambda_{t_i \to f}(t_i) \right] \\
\lambda_{t \to f}(t) &= \sum_{f_j \in F_t \setminus \{f\}} \lambda_{f_j \to t}(t)
\end{aligned}
$$

- **Exponential Family Messages:**

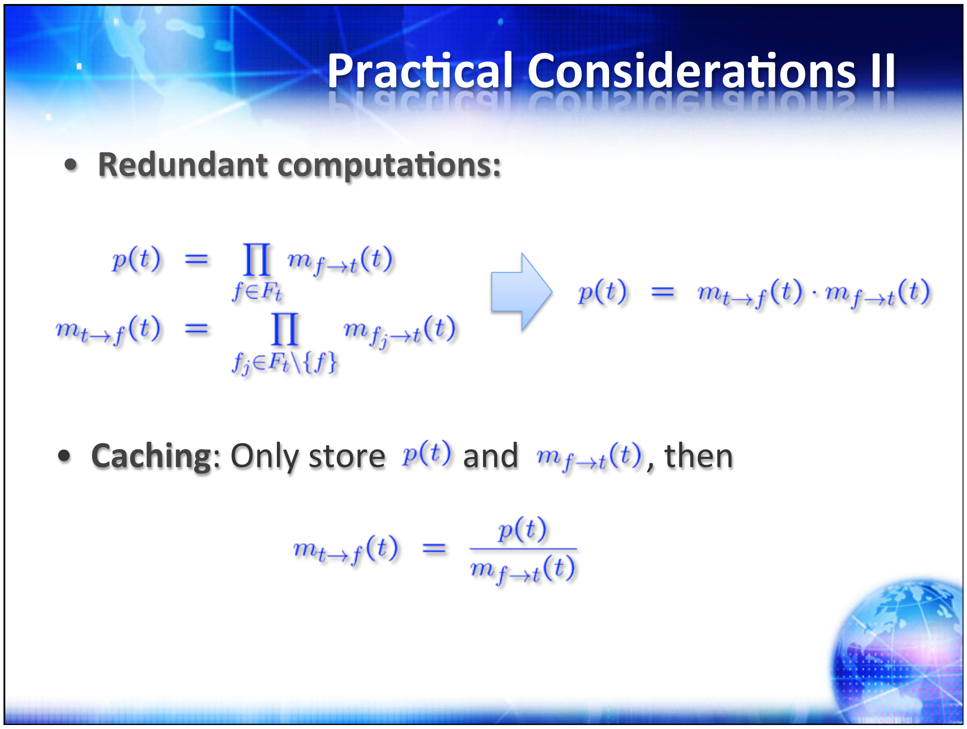
$$
m(t) \propto \exp \left(\psi(t) \cdot \boldsymbol{\theta} \right)
$$

- Message updates are just additions of the parameters $\boldsymbol{\theta}$!



Exponential Families

- (Univariate) Gaussian: $\boldsymbol{\theta} := \left(\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}\right)$
- Bernoulli: $\boldsymbol{\theta} := \log\left(\frac{p}{1-p}\right)$
- Binomial: $\boldsymbol{\theta} := \log\left(\frac{p}{1-p}\right)$
- Beta: $\boldsymbol{\theta} := (\alpha, \beta)$
- Gamma: $\boldsymbol{\theta} := \left(\alpha, \frac{1}{\beta}\right)$



Practical Considerations II

- **Redundant computations:**

$$\begin{aligned} p(t) &= \prod_{f \in F_t} m_{f \to t}(t) \\ m_{t \to f}(t) &= \prod_{f_j \in F_t \setminus \{f\}} m_{f_j \to t}(t) \end{aligned} \quad \Rightarrow \quad p(t) = m_{t \to f}(t) \cdot m_{f \to t}(t)$$

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

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

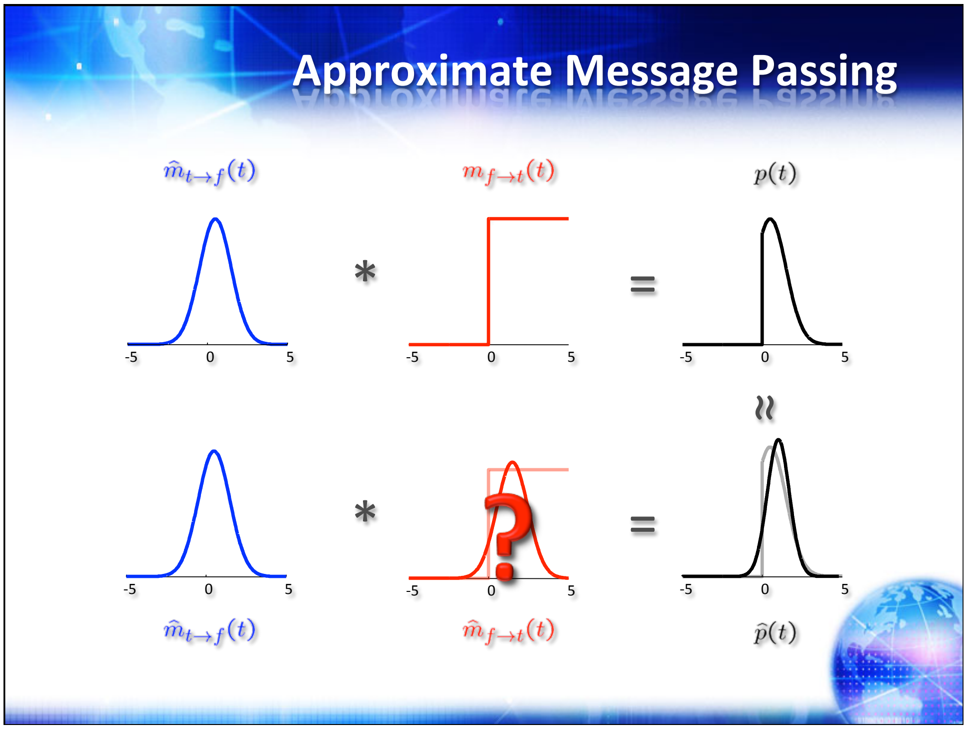
Overview

- Graphical Models
- Inference in Factor Graphs
- Approximate Message Passing
- Distributed Message Passing

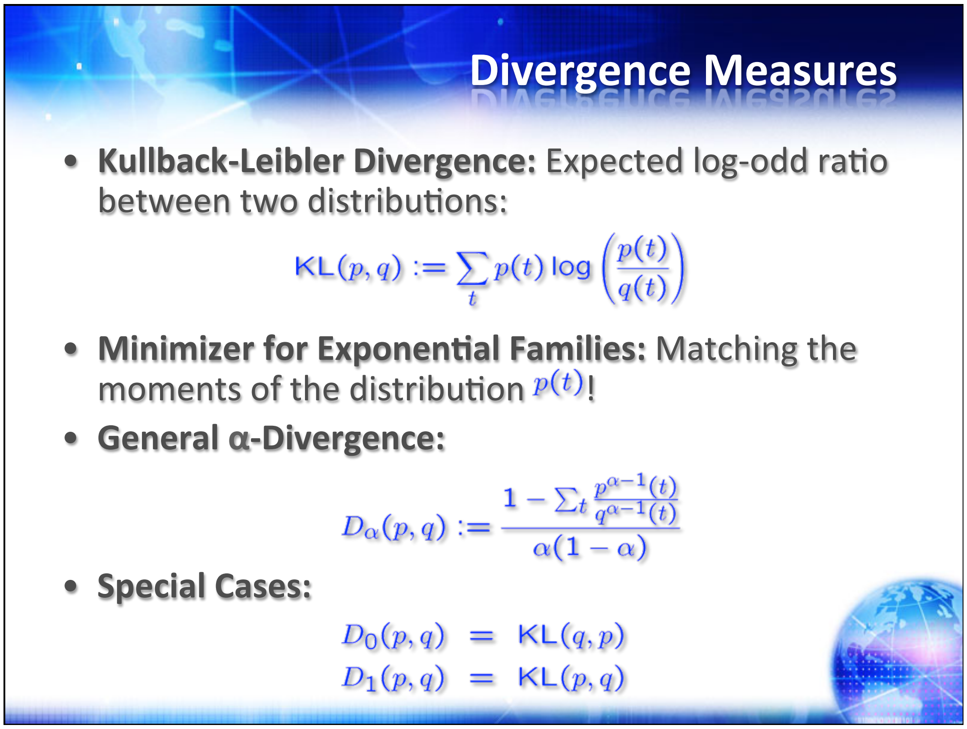
Approximate Message Passing

- **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

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



Approximate Message Passing



Divergence Measures

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

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

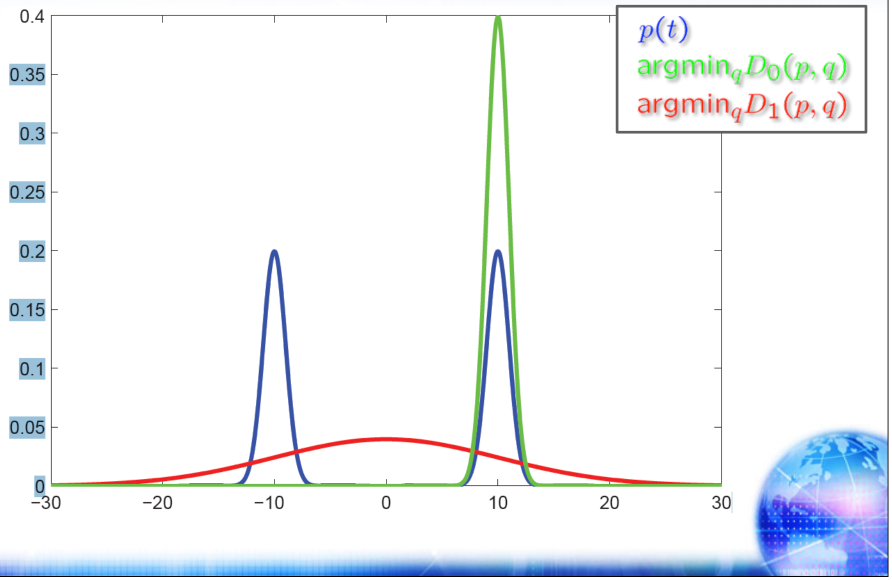
- **Minimizer for Exponential Families:** Matching the moments of the distribution $p(t)$!
- **General α-Divergence:**

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

- **Special Cases:**

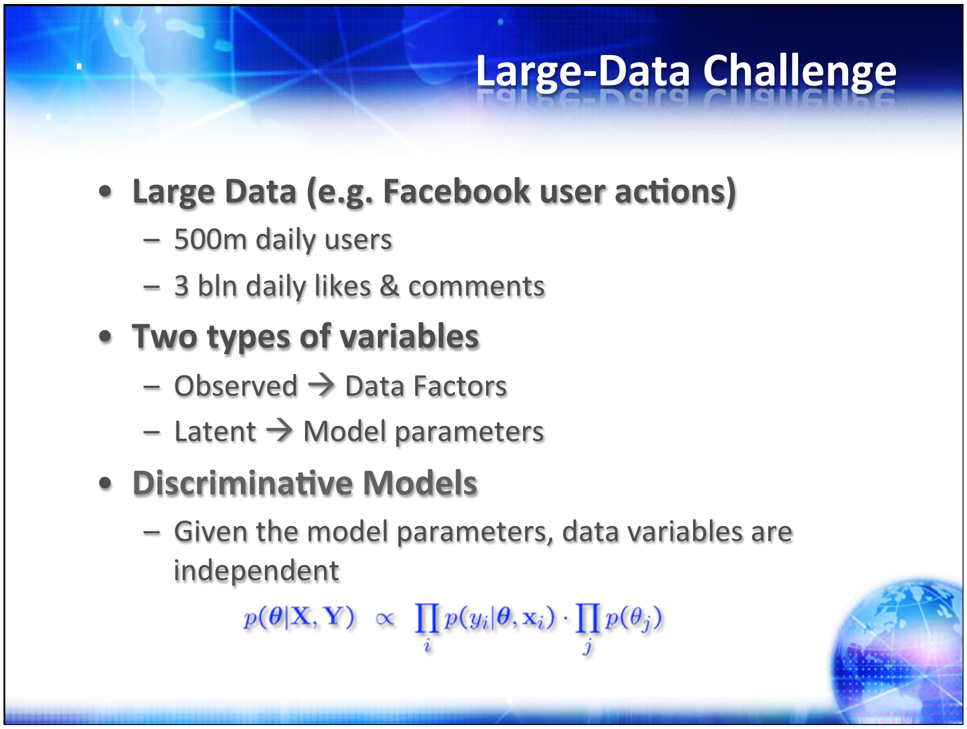
$$\begin{aligned} D_0(p,q) &= \mathsf{KL}(q,p) \\ D_1(p,q) &= \mathsf{KL}(p,q) \end{aligned}$$

α-Divergence in Pictures



Overview

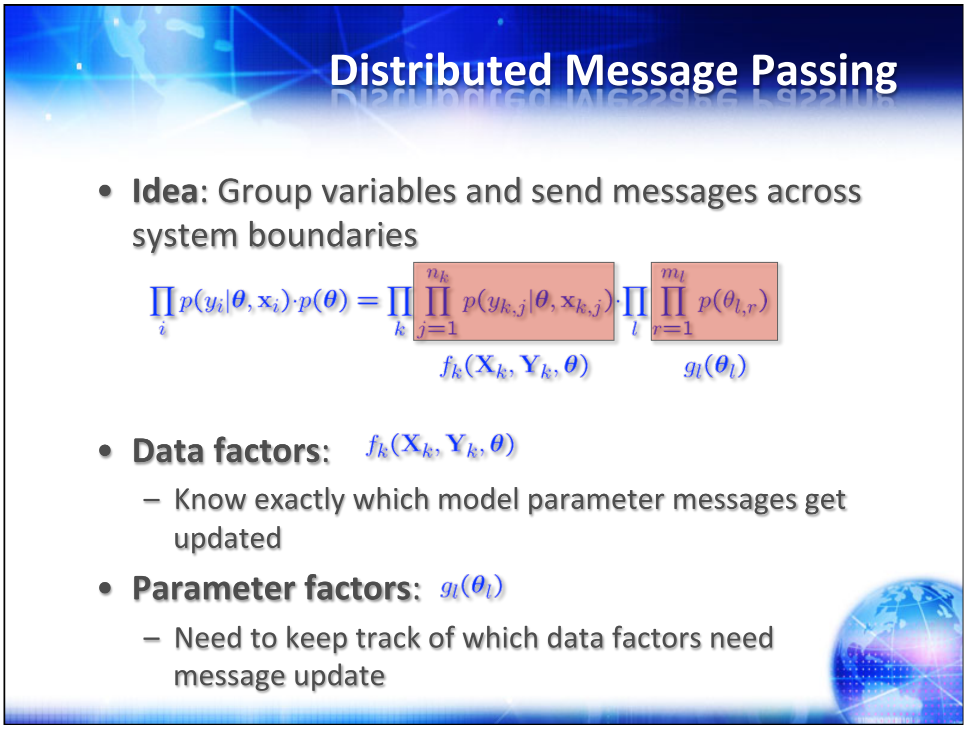
- Graphical Models
- Inference in Factor Graphs
- Approximate Message Passing
- Distributed Message Passing



Large-Data Challenge

- **Large Data (e.g. Facebook user actions)**
  - 500m daily users
  - 3 bln daily likes & comments
- **Two types of variables**
  - Observed → Data Factors
  - Latent → Model parameters
- **Discriminative Models**
  - Given the model parameters, data variables are independent

$$p(\boldsymbol{\theta}|\mathbf{X},\mathbf{Y}) \propto \prod_i p(y_i|\boldsymbol{\theta},\mathbf{x}_i)\cdot\prod_j p(\theta_j)$$



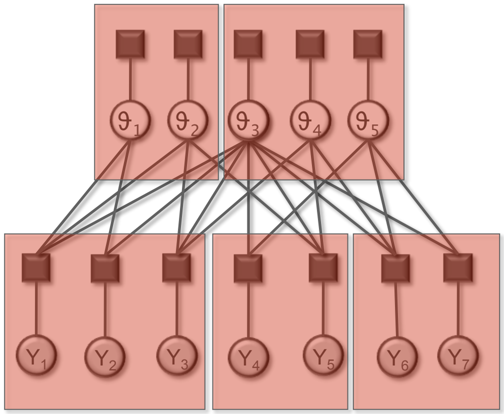
Distributed Message Passing

- **Idea**: Group variables and send messages across system boundaries

$$\prod_i p(y_i|\boldsymbol{\theta},\mathbf{x}_i)\cdot p(\boldsymbol{\theta}) = \prod_k \underbrace{\boxed{\prod_{j=1}^{n_k} p(y_{k,j}|\boldsymbol{\theta},\mathbf{x}_{k,j})}}_{f_k(\mathbf{X}_k,\mathbf{Y}_k,\boldsymbol{\theta})}\cdot\prod_l \underbrace{\boxed{\prod_{r=1}^{m_l} p(\theta_{l,r})}}_{g_l(\boldsymbol{\theta}_l)}$$

- **Data factors**: $f_k(\mathbf{X}_k,\mathbf{Y}_k,\boldsymbol{\theta})$
  - Know exactly which model parameter messages get updated
- **Parameter factors**: $g_l(\boldsymbol{\theta}_l)$
  - Need to keep track of which data factors need message update

Distributed Conditional Models



Belief Store
("Memory")

Message Passing
("Communicate")

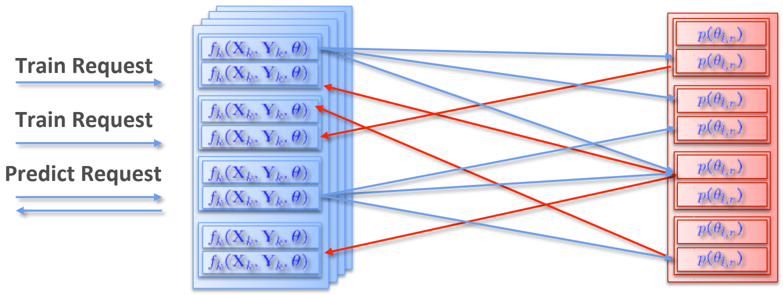
Data Messages
("Compute")

A Systems Service View

Compute

Communicate

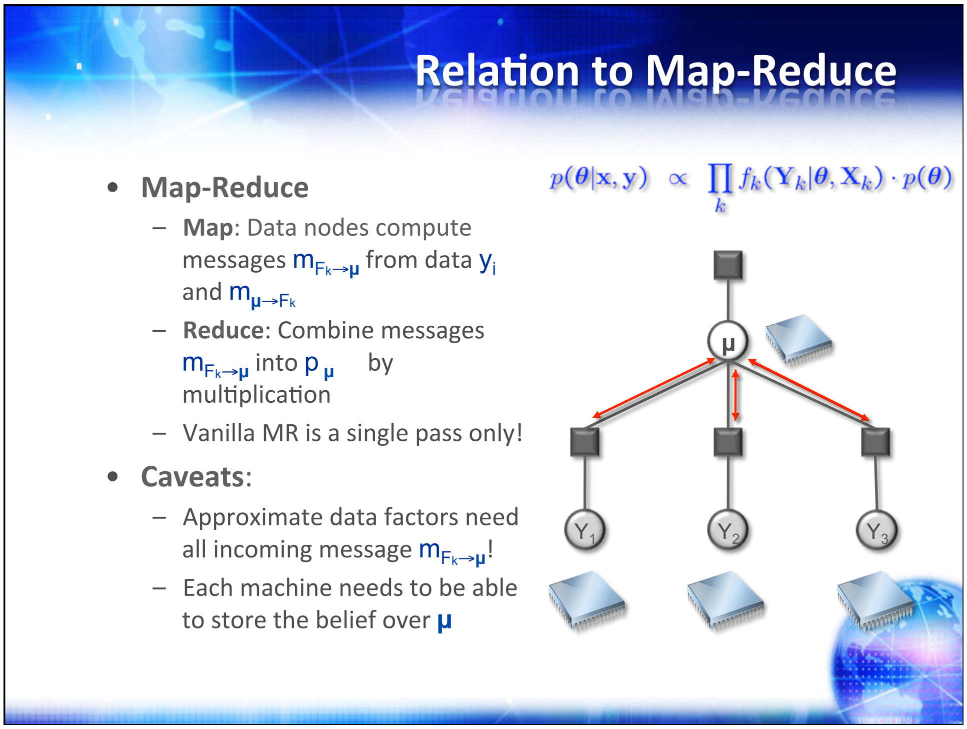
Store



Train Request

Train Request

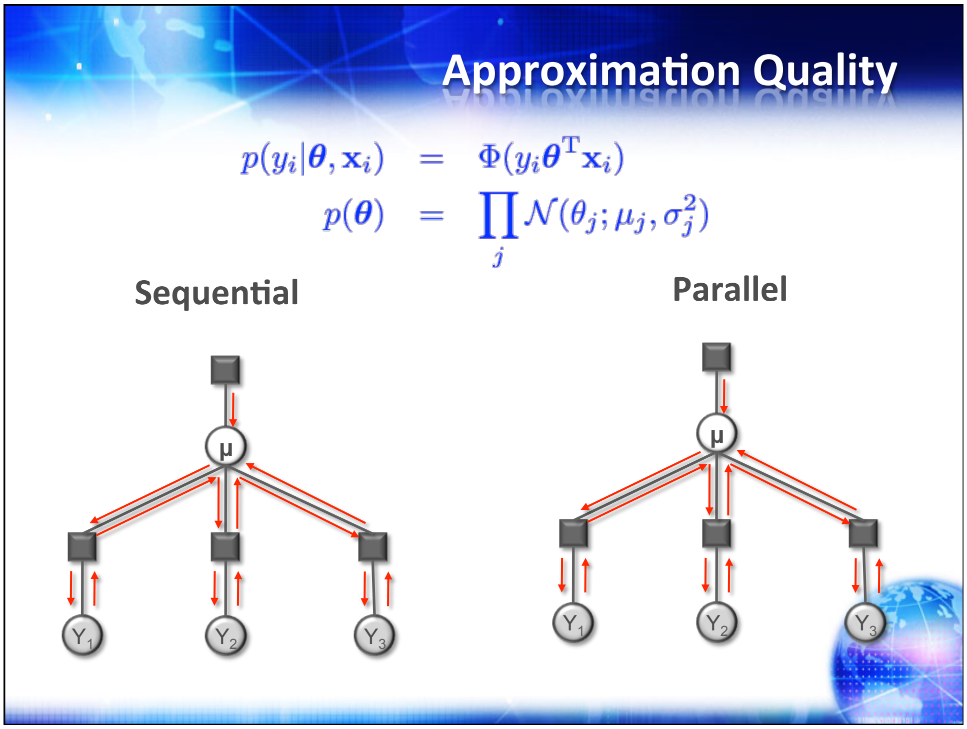
Predict Request



Relation to Map-Reduce

- **Map-Reduce**
  - **Map**: Data nodes compute messages $m_{F_k\to\mu}$ from data y_i and $m_{\mu\to F_k}$
  - **Reduce**: Combine messages $m_{F_k\to\mu}$ into p_{μ} by multiplication
  - Vanilla MR is a single pass only!
- **Caveats**:
  - Approximate data factors need all incoming message $m_{F_k\to\mu}$!
  - Each machine needs to be able to store the belief over $\boldsymbol{\mu}$

$$p(\boldsymbol{\theta}|\mathbf{x},\mathbf{y}) \propto \prod_k f_k(\mathbf{Y}_k|\boldsymbol{\theta},\mathbf{X}_k)\cdot p(\boldsymbol{\theta})$$

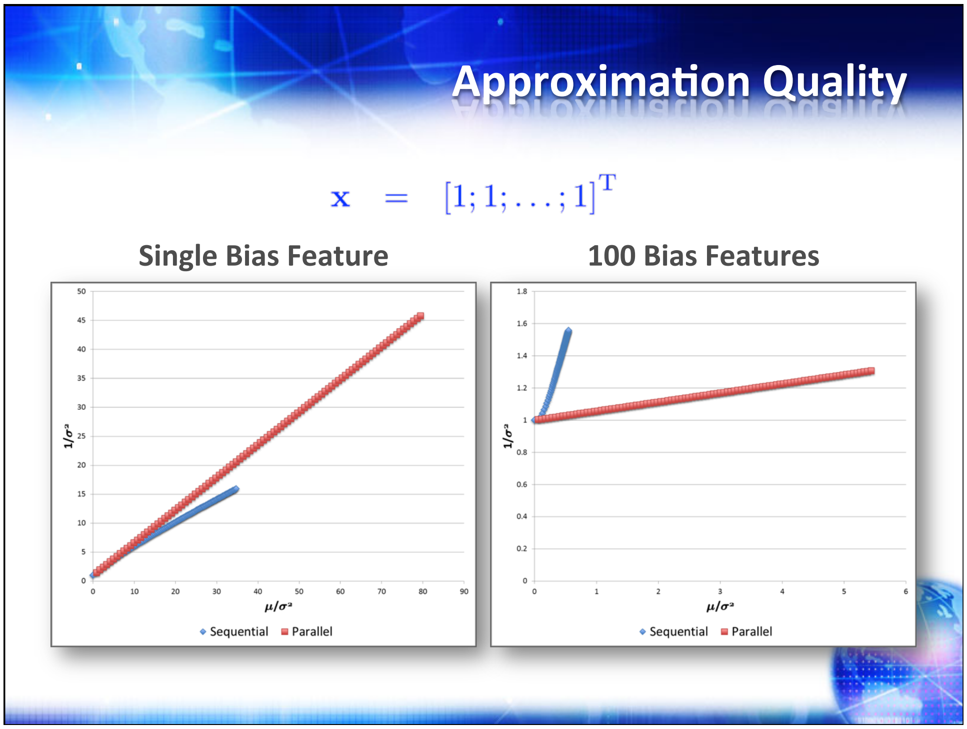


Approximation Quality

$$\begin{aligned} p(y_i|\boldsymbol{\theta},\mathbf{x}_i) &= \Phi(y_i\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i) \\ p(\boldsymbol{\theta}) &= \prod_j \mathcal{N}(\theta_j;\mu_j,\sigma_j^2) \end{aligned}$$

Sequential

Parallel

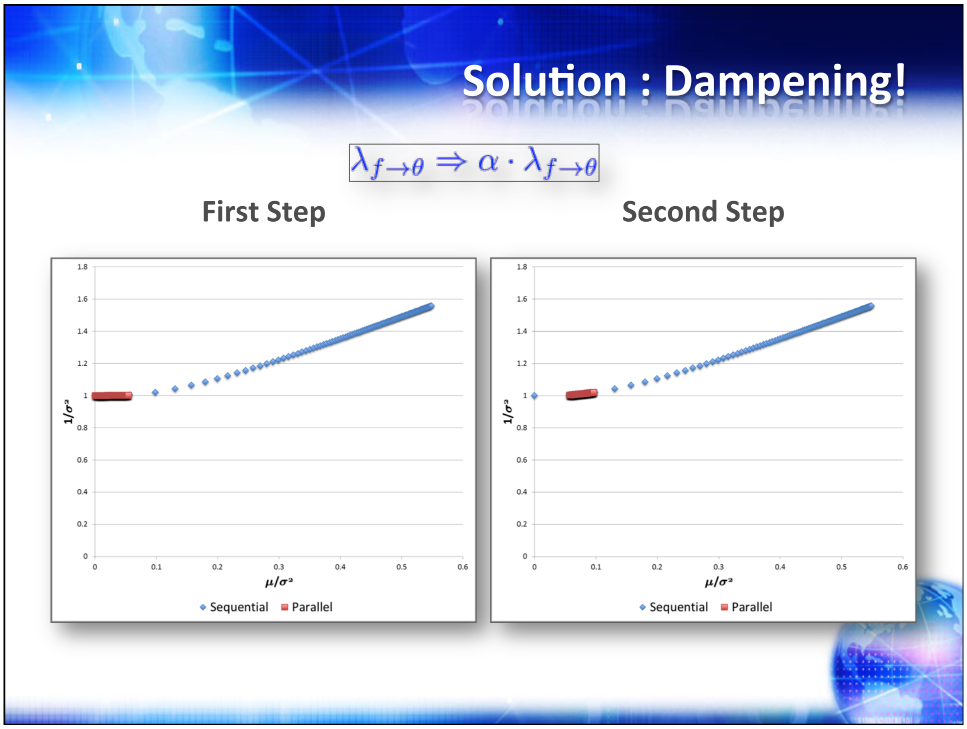


Approximation Quality

$$\mathbf{x} = [1;1;\ldots;1]^{\mathrm{T}}$$

Single Bias Feature

100 Bias Features



Solution : Dampening!

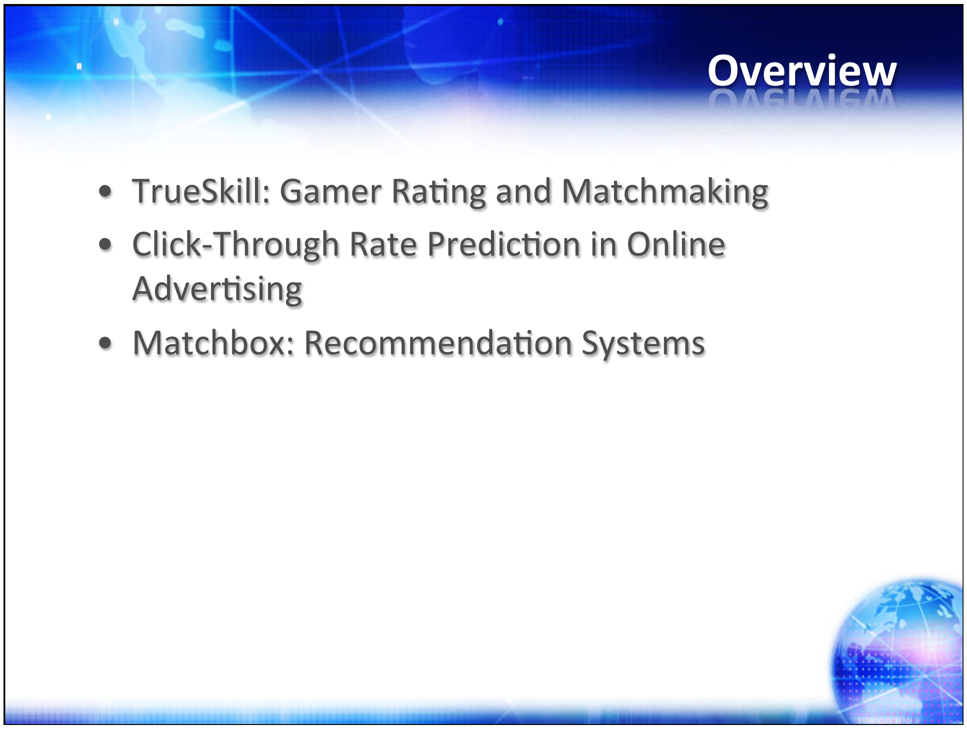
$$\lambda_{f\to\theta} \Rightarrow \alpha \cdot \lambda_{f\to\theta}$$

First Step

Second Step

Break!

Part 2: Applications



Overview

- TrueSkill: Gamer Rating and Matchmaking
- Click-Through Rate Prediction in Online Advertising
- Matchbox: Recommendation Systems

TrueSkill™

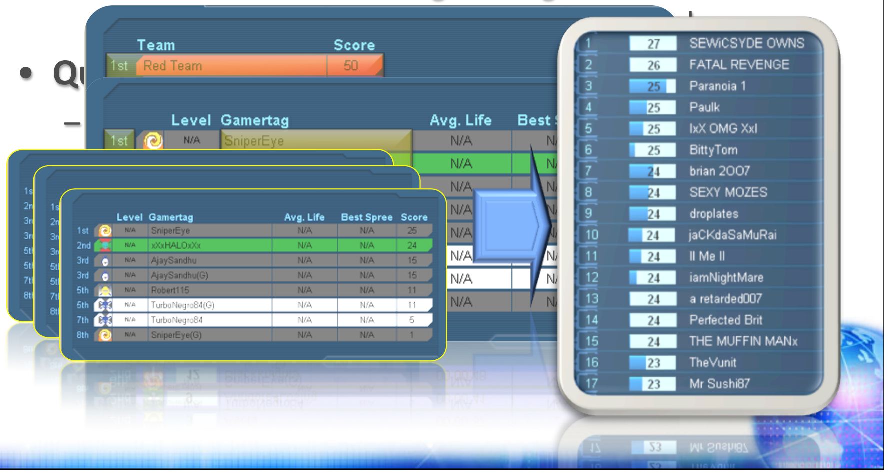
Joint work with Thore Graepel, Tom Minka & Phillip Trelford

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

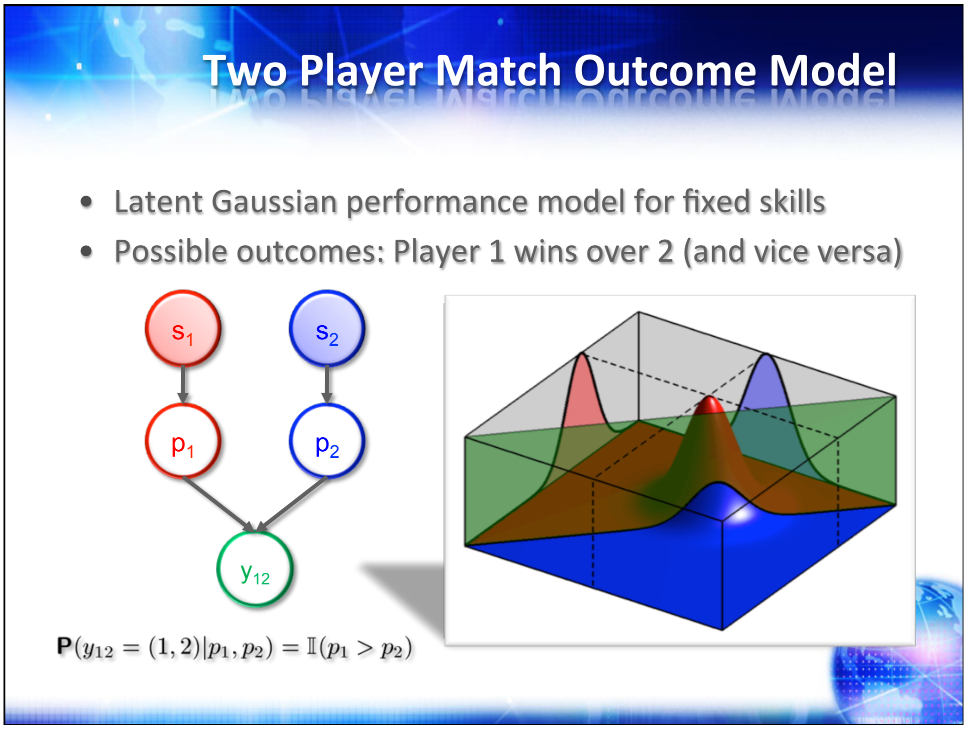
- **Given**:
  - Match outcomes: Orderings among k teams
- **Qu**



Team	Score
1st Red Team	50

	Level	Gamertag	Avg. Life	Best Spree	Score
1st	N/A	SniperEye	N/A	N/A	25
2nd	N/A	xXxHALOxXx	N/A	N/A	24
3rd	N/A	AjaySandhu	N/A	N/A	15
3rd	N/A	AjaySandhu(G)	N/A	N/A	15
5th	N/A	Robert115	N/A	N/A	11
5th	N/A	TurboNegro84(G)	N/A	N/A	11
7th	N/A	TurboNegro84	N/A	N/A	5
8th	N/A	SniperEye(G)	N/A	N/A	1

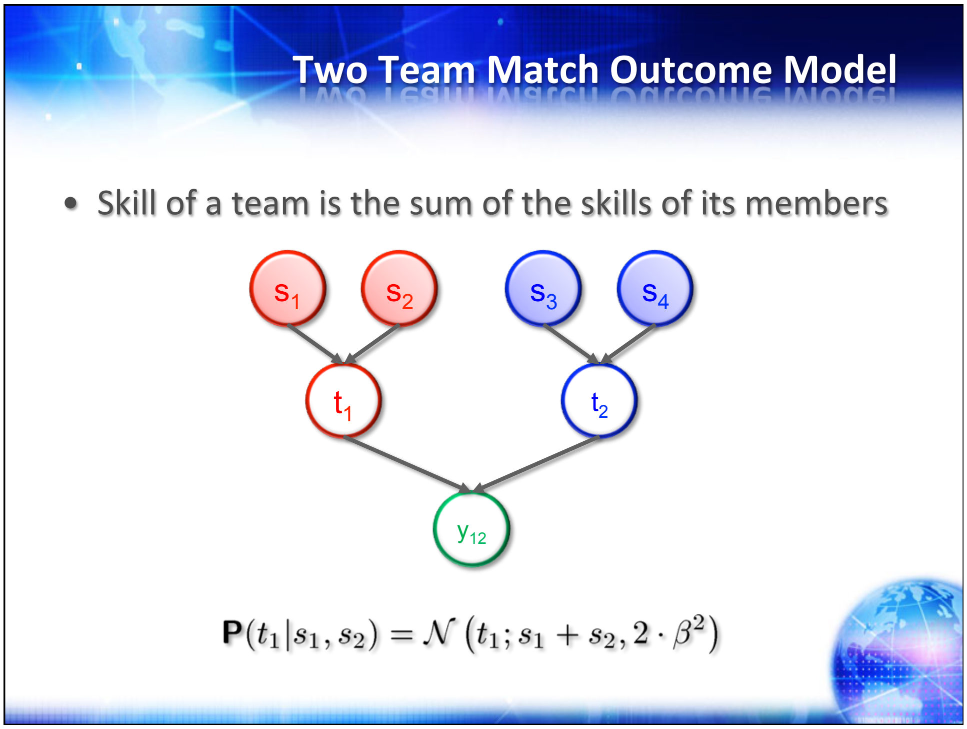
1	27	SEWiCSYDE OWNS
2	26	FATAL REVENGE
3	25	Paranoia 1
4	25	Paulk
5	25	IxX OMG XxI
6	25	BittyTom
7	24	brian 2OO7
8	24	SEXY MOZES
9	24	droplates
10	24	jaCKdaSaMuRai
11	24	II Me II
12	24	iamNightMare
13	24	a retarded007
14	24	Perfected Brit
15	24	THE MUFFIN MANx
16	23	TheVunit
17	23	Mr Sushi87



Two Player Match Outcome Model

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

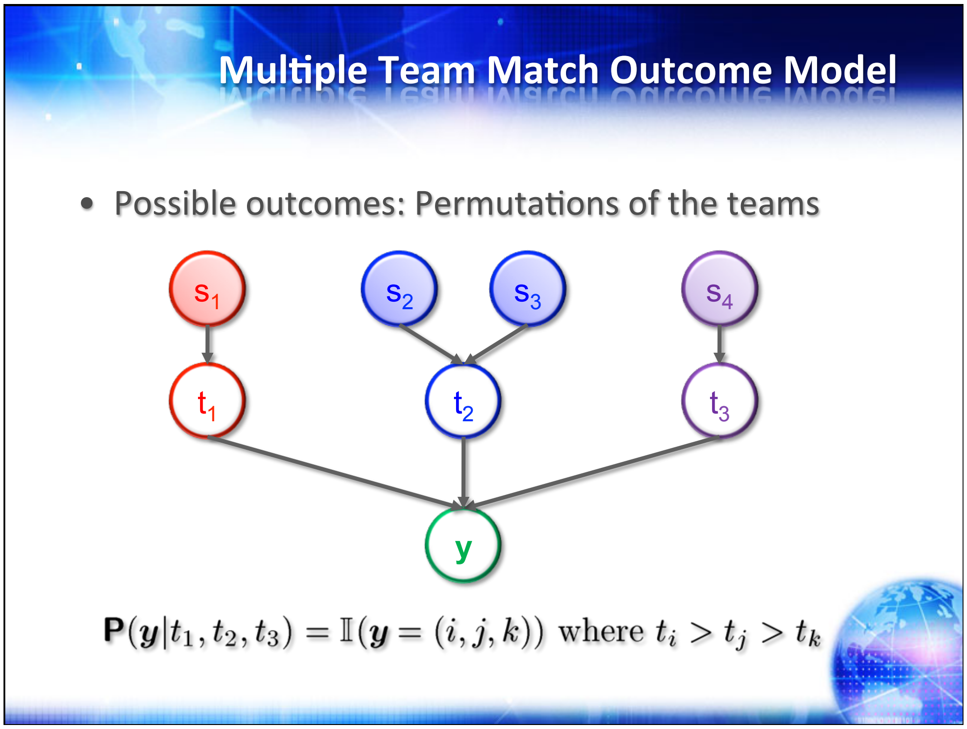
$$\mathbf{P}(y_{12} = (1,2)|p_1, p_2) = \mathbb{I}(p_1 > p_2)$$



Two Team Match Outcome Model

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

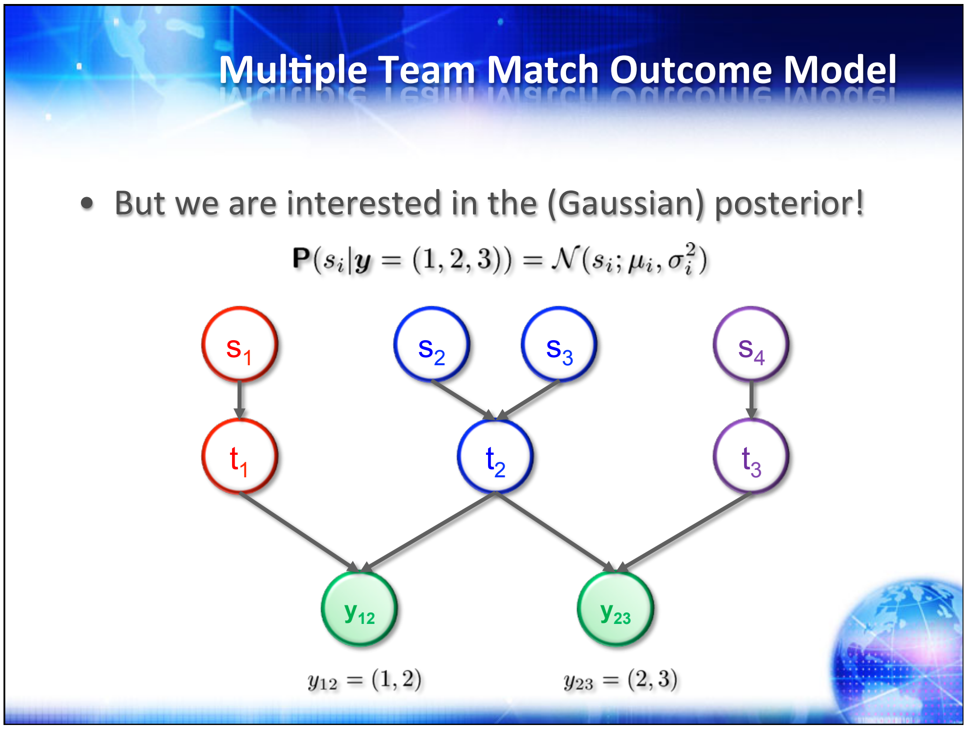
$$\mathbf{P}(t_1|s_1, s_2) = \mathcal{N}\left(t_1; s_1 + s_2, 2 \cdot \beta^2\right)$$



Multiple Team Match Outcome Model

- Possible outcomes: Permutations of the teams

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



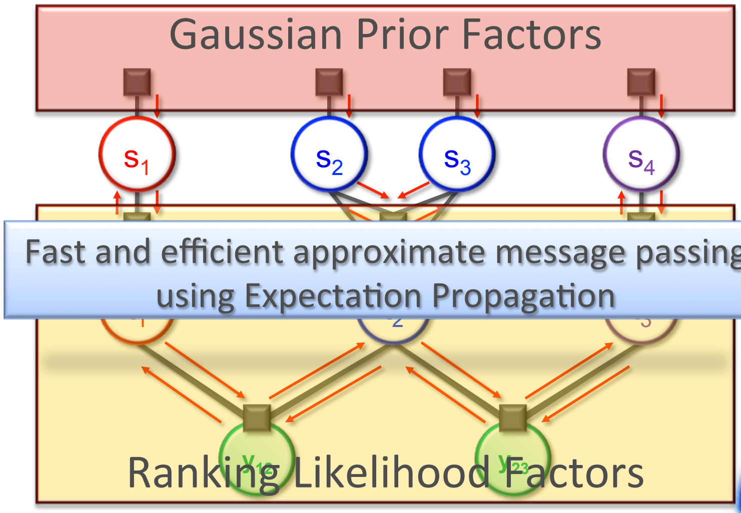
Multiple Team Match Outcome Model

- 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



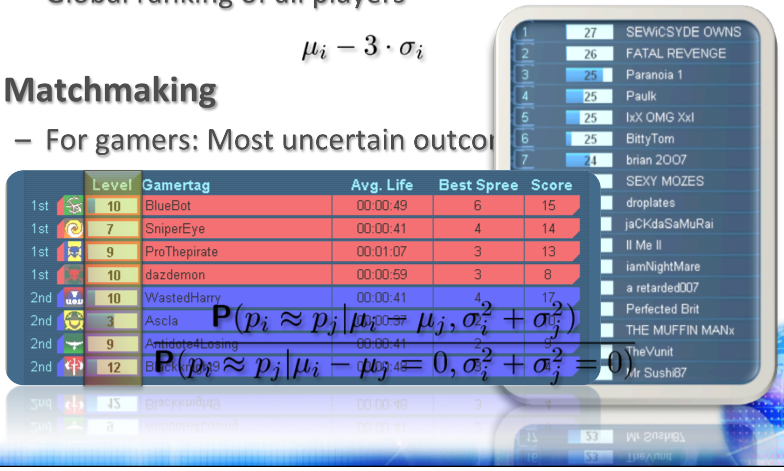
Fast and efficient approximate message passing using Expectation Propagation

Applications to Online Gaming

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

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

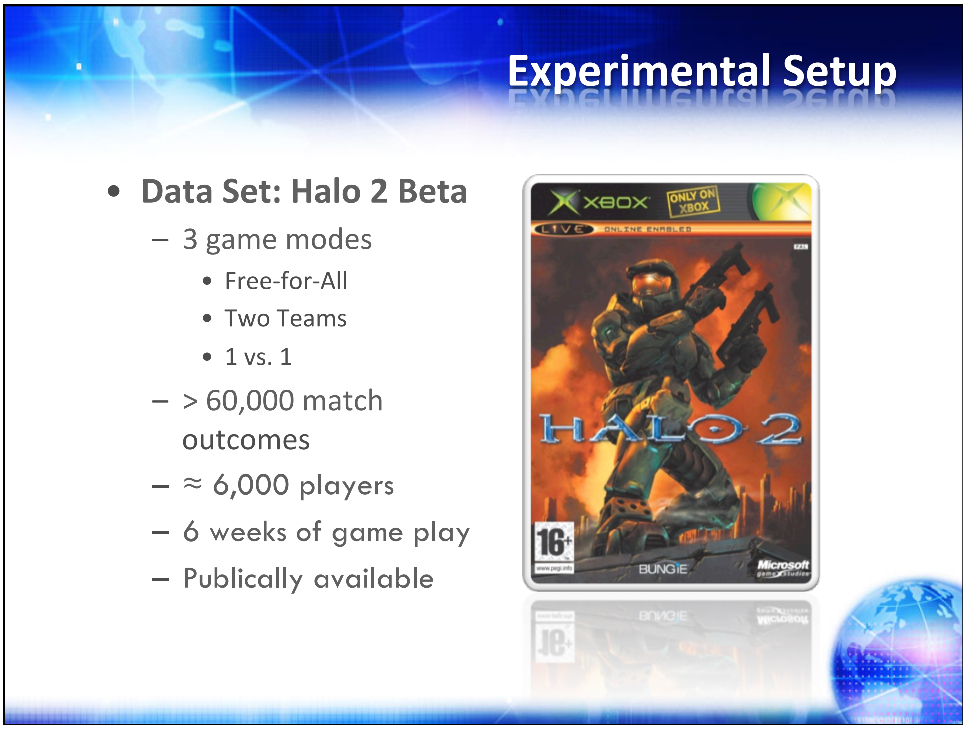
- **Matchmaking**
  - For gamers: Most uncertain outco



	1	27	SEWiCSYDE OWNS
	2	26	FATAL REVENGE
	3	25	Paranoia 1
	4	25	Paulk
	5	25	IxX OMG XxI
	6	25	BittyTom
	7	24	brian 2OO7
			SEXY MOZES
			droplates
			jaCKdaSaMuRai
			II Me II
			iamNightMare
			a retarded007
			Perfected Brit
			THE MUFFIN MANx
			TheVunit
			Mr Sushi87

	Level	Gamertag	Avg. Life	Best Spree	Score
1st	10	BlueBot	00:00:49	6	15
1st	7	SniperEye	00:00:41	4	14
1st	9	ProThepirate	00:01:07	3	13
1st	10	dazdemon	00:00:59	3	8
2nd	10	WastedHarry	00:00:41	4	17
2nd	3	Ascla	00:00:37	2	10
2nd	9	Antidote4Losing	00:00:41	2	9
2nd	12	BlackKnight9	00:00:48	3	4

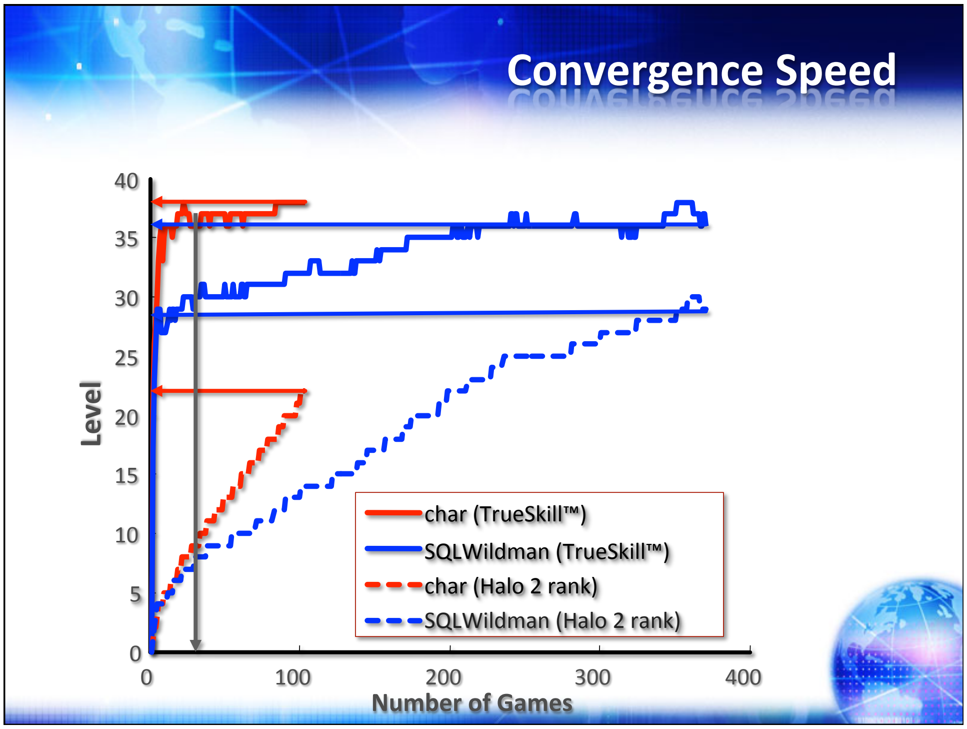
$$\frac{\mathbf{P}(p_i \approx p_j | \mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)}{\mathbf{P}(p_i \approx p_j | \mu_i - \mu_j = 0, \sigma_i^2 + \sigma_j^2 = 0)}$$



Experimental Setup

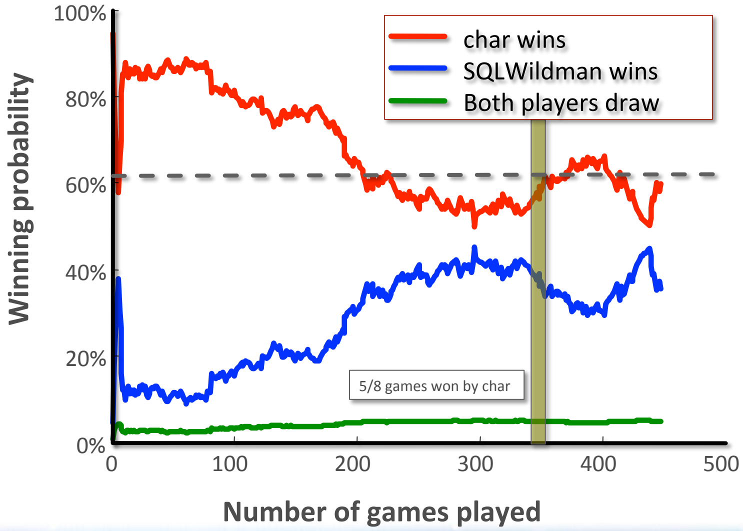
- **Data Set: Halo 2 Beta**
  - 3 game modes
    - Free-for-All
    - Two Teams
    - 1 vs. 1
  - > 60,000 match outcomes
  - ≈ 6,000 players
  - 6 weeks of game play
  - Publically available

XBOX
ONLY ON XBOX
LIVE ONLINE ENABLED
HALO 2
16+
BUNGIE
Microsoft game studios



Convergence Speed

Convergence Speed (ctd.)

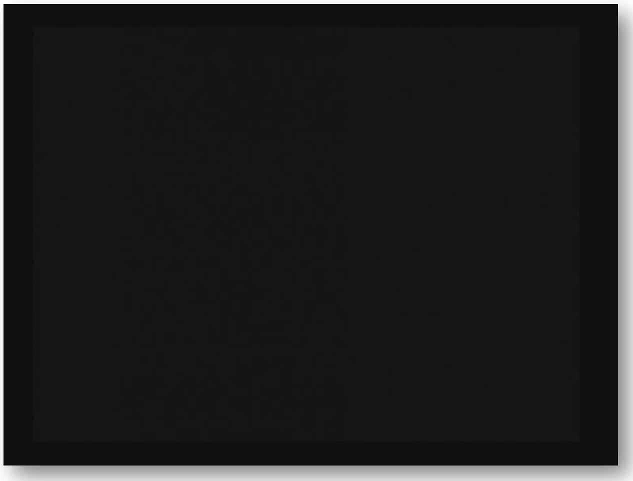


Xbox 360 & Halo 3

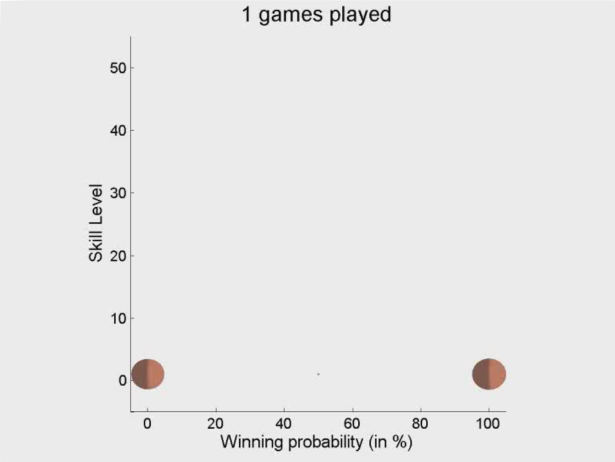
- **Xbox 360 Live**
  - Launched in September 2005
  - Every game uses TrueSkill™ to match players
  - > 10 million players
  - > 2 million matches per day
  - > 2 billion hours of gameplay

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

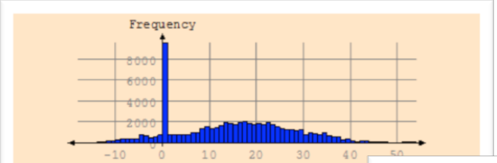
Halo 3 in Action



Halo 3 Public Beta Analysis

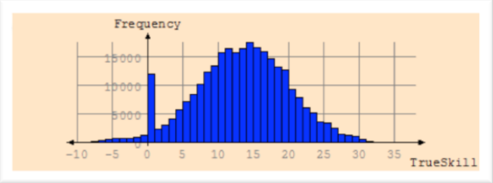


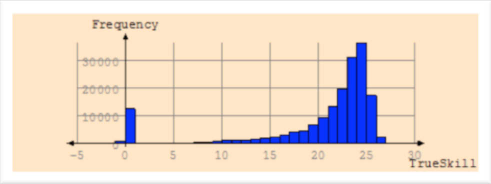
Skill Distributions of Online Games



Golf (18 holes): 60 levels

Car racing (3-4 laps): 40 levels

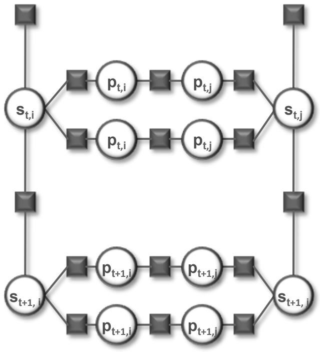




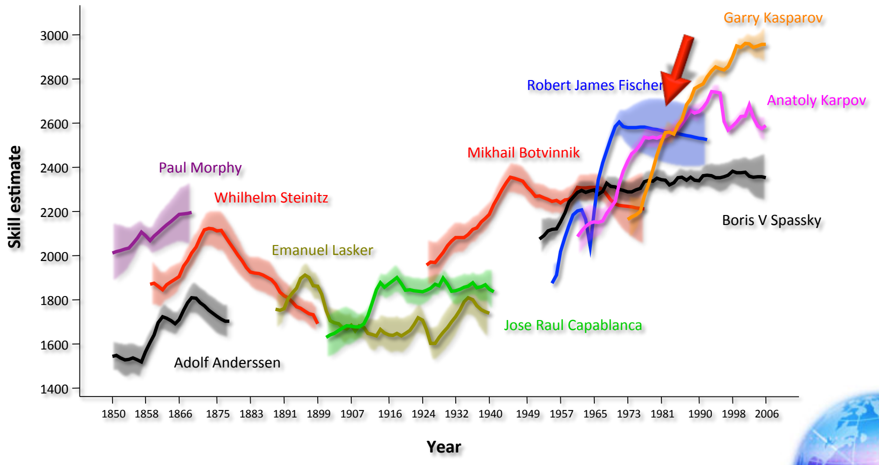
UNO (chance game): 10 levels

TrueSkill^{TM} 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



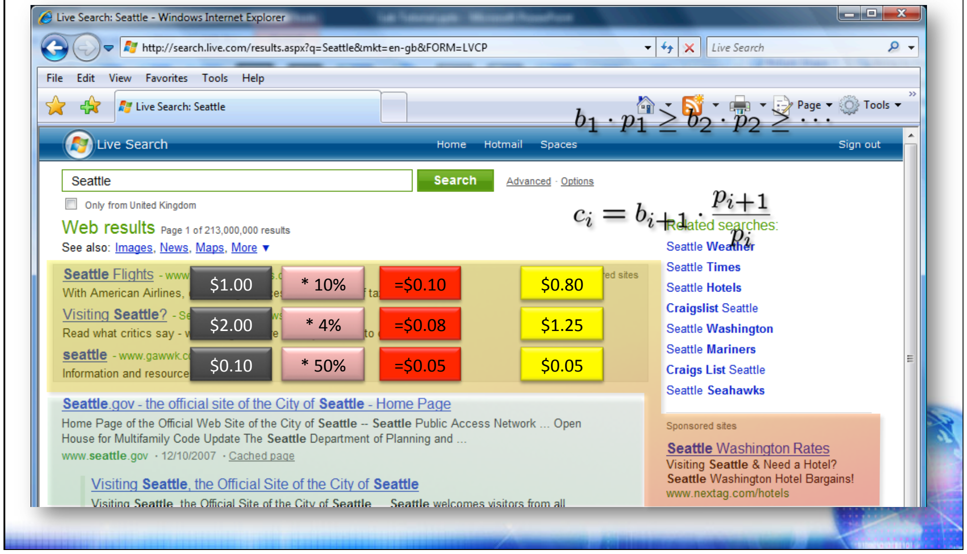
ChessBase Analysis: 1850 - 2006



Online Advertising

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

Why Predict Probability-of-Click?



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

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

□ Only from United Kingdom

Web results Page 1 of 213,000,000 results

See also: Images, News, Maps, More ▼

Seattle Flights - www

With American Airlines,

Visiting Seattle? - Se

Read what critics say - v

seattle - www.gawwk.c

Information and resource

$1.00	* 10%	=$0.10	$0.80
$2.00	* 4%	=$0.08	$1.25
$0.10	* 50%	=$0.05	$0.05

Seattle.gov - the official site of the City of Seattle - Home Page

Home Page of the Official Web Site of the City of Seattle -- Seattle Public Access Network ... Open House for Multifamily Code Update The Seattle Department of Planning and ...

www.seattle.gov · 12/10/2007 · Cached page

Visiting Seattle, the Official Site of the City of Seattle

Visiting Seattle, the Official Site of the City of Seattle ... Seattle welcomes visitors from all

Related searches:

- Seattle Weather
- Seattle Times
- Seattle Hotels
- Craigslist Seattle
- Seattle Washington
- Seattle Mariners
- Craigs List Seattle
- Seattle Seahawks

Sponsored sites

Seattle Washington Rates

Visiting Seattle & Need a Hotel?

Seattle Washington Hotel Bargains!

www.nextag.com/hotels

The Scale of Things

- Several weeks of data in training:

  7,000,000,000 impressions

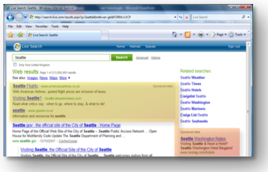
- 2 weeks of CPU time during training:

  2 wks × 7 days × 86,400 sec/day =

  1,209,600 seconds

- Learning algorithm speed requirement:
  - 5,787 impression updates / sec
  - 172.8 μs per impression update

The Flow of Information



Seattle

Web results Page 1 of 213,000,000 results

See also: Images, News, Maps, More

Sponsored sites

Seattle Flights - www.americanairlines.co.uk
With American Airlines, quoted flight prices are inclusive of taxes.

Visiting Seattle? - Seattle.etravelreviews.com
Read what critics say - when to go, where to stay, & what to do!

seattle - www.gawwk.co.uk
Information and resources for seattle

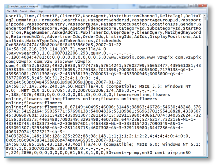
Seattle.gov - the official site of the City of Seattle - Home Page
Home Page of the Official Web Site of the City of Seattle – Seattle Public Access Network ... Open House for Multifamily Code Update The Seattle Department of Planning and ...
www.seattle.gov - 12/10/2007 - Cached page

Visiting Seattle, the Official Site of the City of Seattle
Visiting Seattle, the Official Site of the City of Seattle ... Seattle welcomes visitors from all

Related searches:
Seattle Weather
Seattle Times
Seattle Hotels
Craigslist Seattle
Seattle Washington
Seattle Mariners
Craigs List Seattle
Seattle Seahawks

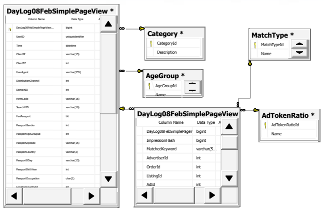
Sponsored sites
Seattle Washington Rates
Visiting Seattle & Need a Hotel? Seattle Washington Hotel Bargains!
www.nextag.com/hotels

User interaction



```
UserID,Time,ClientIP,ClientTZ,UserAgent,DistributionChannel,DeltaTag1,DeltaT
ag2,DomainID,FormCode,SearchVID,PassportGenderId,PassportAgeGroupId,Passport
Zipcode,PassportCountry,PassportBDay,PassportOccupation,LocationIDs,Gender,G
enderConfidenceScore,Age,AgeConfidenceScore,CategoryId,SubcategoryId,StartPo
sition,PageNumber,AskedAdCnt,PublisherId,UserQuery,CleanQuery,MatchedKeyword
s,ReturnedAdCnt,AdvertiserIds,OrderIds,ListingIds,AdIds,DisplayPositions,Act
ualBids,MatchTypeIds,AdTokenRatios,ClickedAds
0xB38E6D7474C58B82DDE8E8453596F2B5,2007-01-22
14:58:26,216.239.114.107,71,Mozilla/4.0
(SavvySearch),1,0,2007012206,293,QBHP,0,-,-,-,-,-,
-,190;4084;4781;0;0,0,0,0,0,0,0,0,5,0,www.vzwpix.com,www vzwpix com,vzwpix
com;vzwpix com;vzw pix;www vzwpix
com,4,39452;65282;4952;8933,57774756;57614241;57602799;56652477,439561081;43
1938139;433300946;387726095,7179138;7011398;7000031;6065600,7179138-QS-1-
439561081;7011398-QS-2-431938139;7000031-QS-3-433300946;6065600-QS-4-
387726095,8;45;30;31,2;2;4;1,0;0;-;4,
0x9E4C5C3DC0AB4558AE61B9B8305B10DE,2007-01-22
14:58:57,145.246.240.14,50,Mozilla/4.0 (compatible; MSIE 5.5; Windows NT
5.0; .NET CLR 1.0.3705),3,0,2007012206,174,AS5,0,-,-,-,-,-,
-,188;4076;4923;0;0,0,0,0,0,81,85,0,0,8,0,flowers+online,flowers
online,flowers online;flowers online;flowers online;flowers
online;flowers;flowers
online;flowers;flowers,8,67149;40495;46006;31448;38863;46726;54630;48248,576
55742;54192164;55035771;57690138;54004836;55109881;56963701;55418828,4339507
45;306697601;335315420;435091307;281145715;329115980;406617074;340352624,732
2156;3508373;4463488;7090349;3293498;4607308;6447236;5275217,7322156-ML-1-
433950745;3508373-ML-2-306697601;4463488-ML-3-335315420;7090349-SB-1-
435091307;3293498-SB-2-281145715;4607308-SB-3-329115980;6447236-SB-4-
406617074;5275217-SB-5-
340352624,148;138;128;225;292;88;98;148,1;1;1;1;2;1;2;2,4;4;4;4;0;4;0;0,
0x0C5C6C90A3B3413E826D474DFAA6535C,2007-01-22
14:58:02,85.186.43.119,43,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1;
SV1),1,0,2007012206,293,PERE,0,-,-,-,-,-,
-,224;2896;0;0;0,0,0,0,0,0,61,65,8,1,8,0,50+cent+-pimp,nn50 cent pimp,nn50
```

Raw Logs



Column Name	Data Type
DayLog08FebSimplePageView...	bigint
UserID	uniqueidentifier
Time	datetime
ClientIP	varchar(15)
ClientTZ	int
UserAgent	varchar(255)
DistributionChannel	int
DomainID	int
FormCode	varchar(16)
SearchVID	varchar(16)
HasPassport	bit
PassportGender	int
PassportAgeGroupId	int
PassportZipcode	varchar(15)
PassportCountry	varchar(2)
PassportBDay	varchar(15)
PassportBirthYear	int
PassportOccupation	char(1)
LocationCountryId	int

Category: CategoryId, Description

AgeGroup: AgeGroupId, Name

MatchType: MatchTypeId, Name

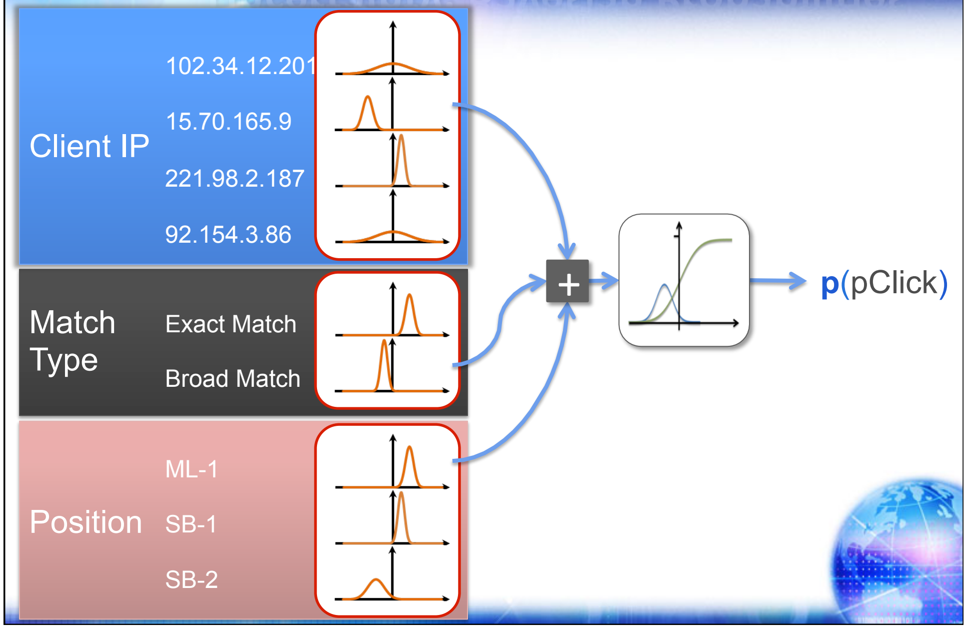
Column Name	Data Type
DayLog08FebSimplePageV	bigint
ImpressionHash	bigint
MatchedKeyword	varchar(5...
AdvertiserId	int
OrderId	int
ListingId	int
AdId	int

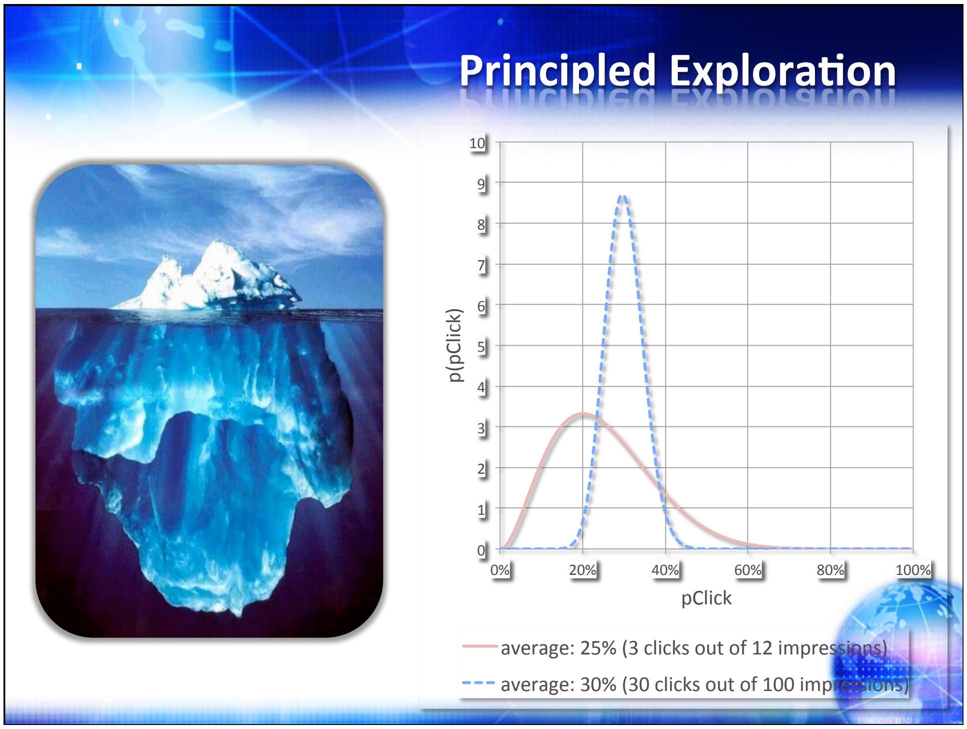
AdTokenRatio: AdTokenRatioId, Name

Structured Data

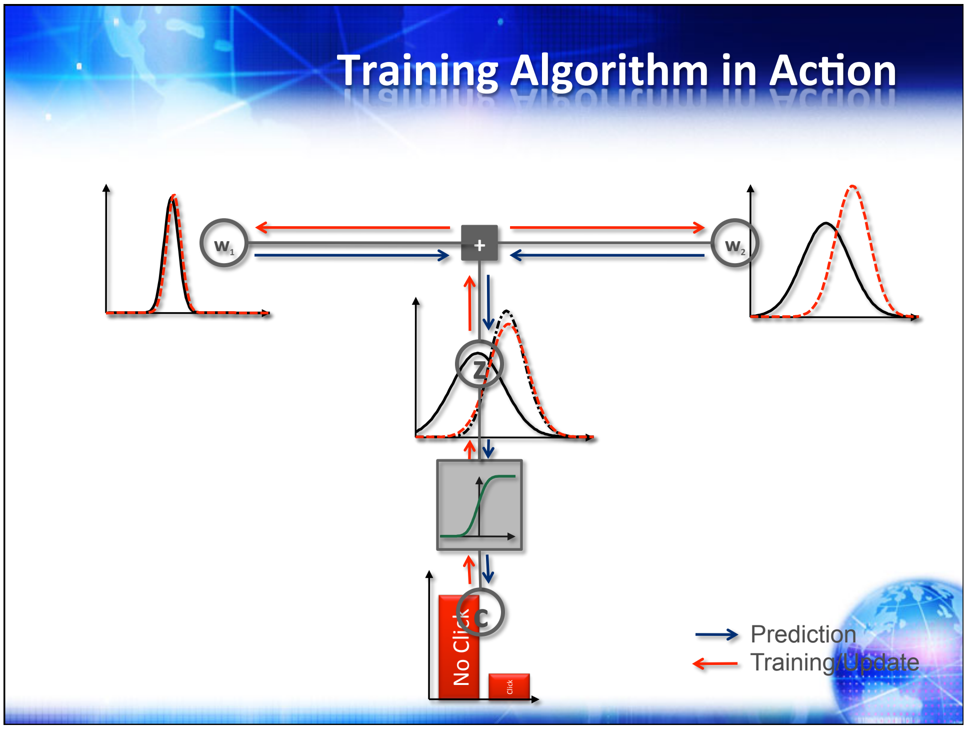
- Why structured data?
  - Data validation and cleaning
  - Principled feature transformations

Uncertainty: Bayesian Probabilities





Principled Exploration

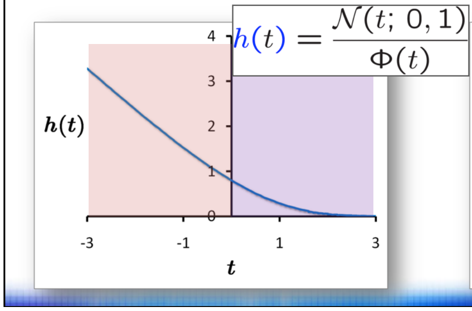


Training Algorithm in Action

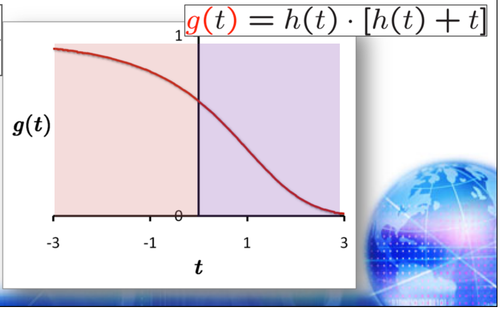
Inference: An Optimization View

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

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

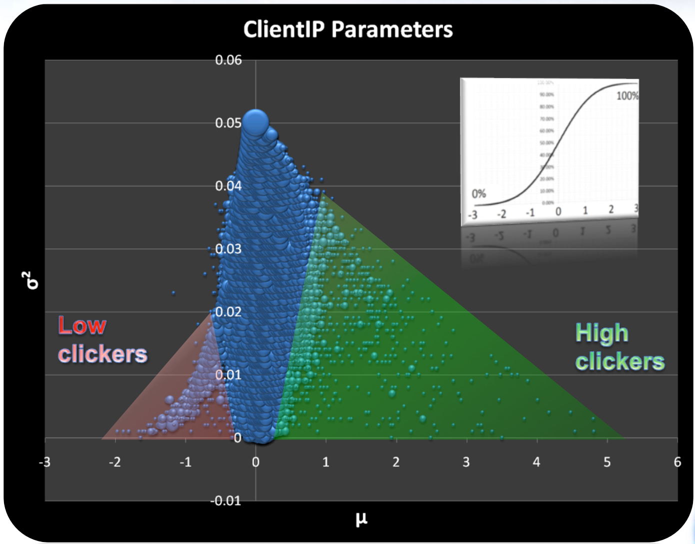


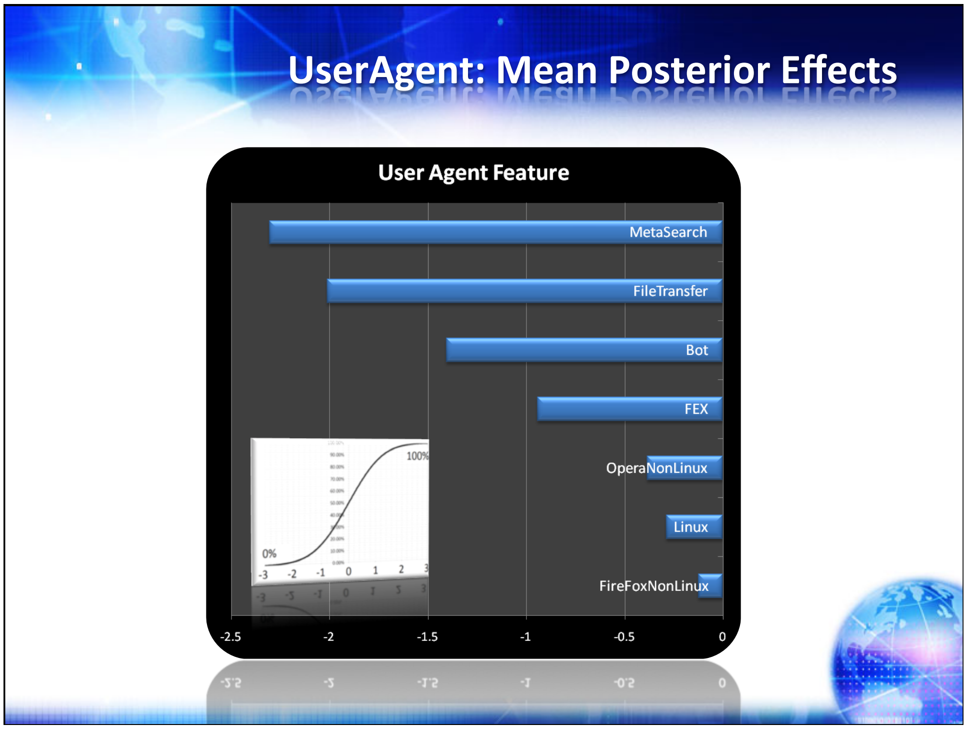
$$h(t) = \frac{\mathcal{N}(t;0,1)}{\Phi(t)}$$



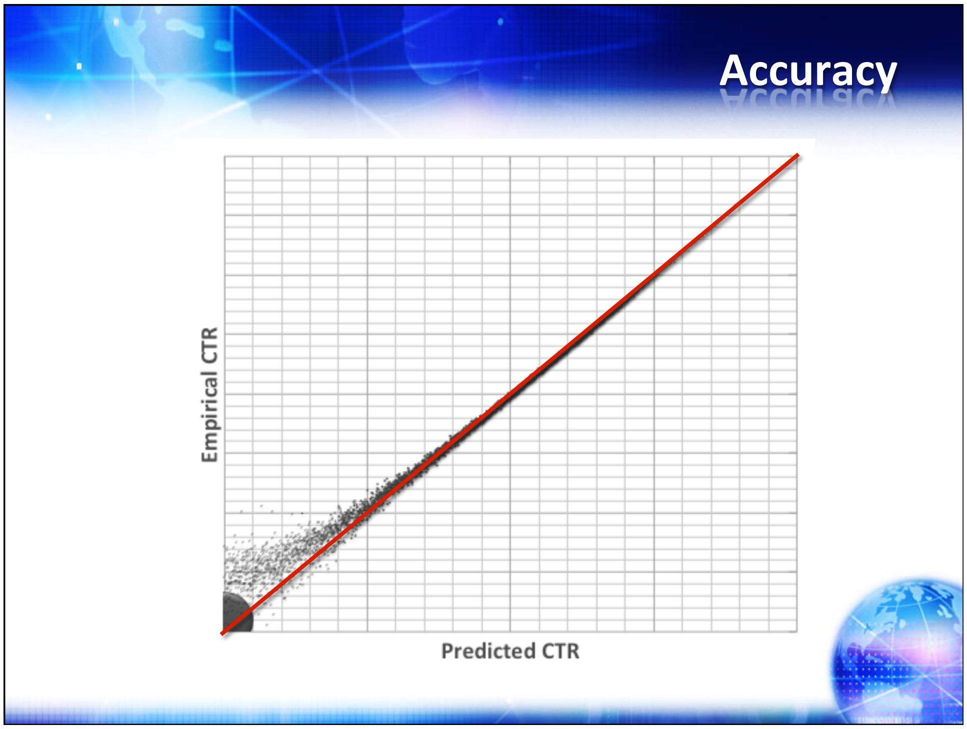
$$g(t) = h(t) \cdot [h(t) + t]$$

Client IP: Mean & Variance





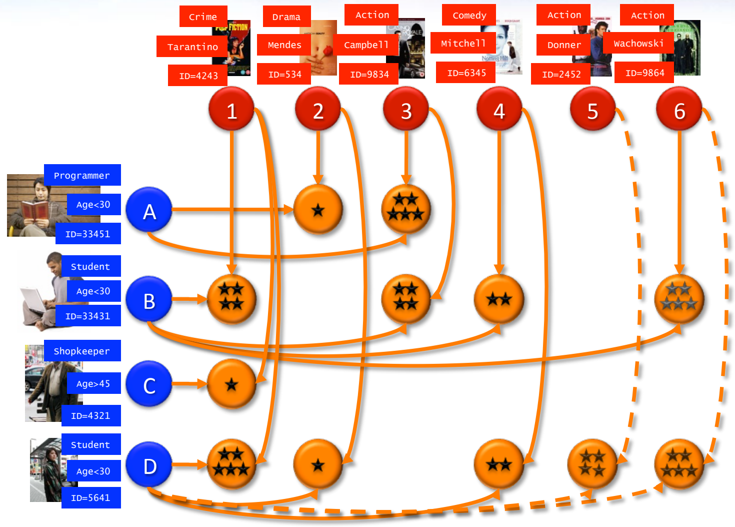
UserAgent: Mean Posterior Effects

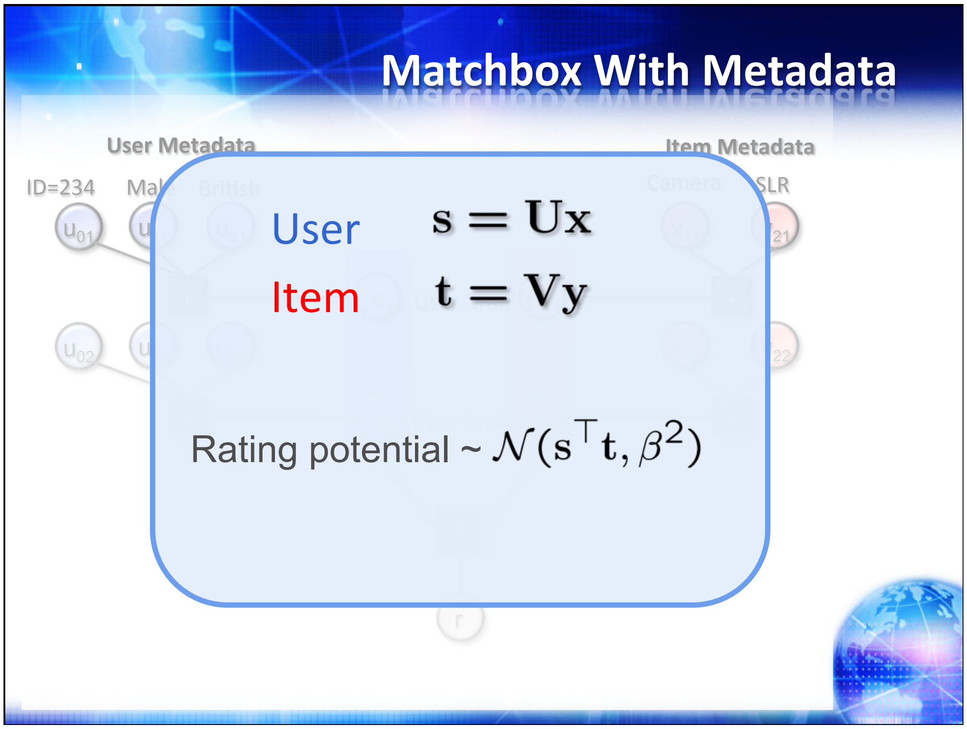


Accuracy

MatchBox

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



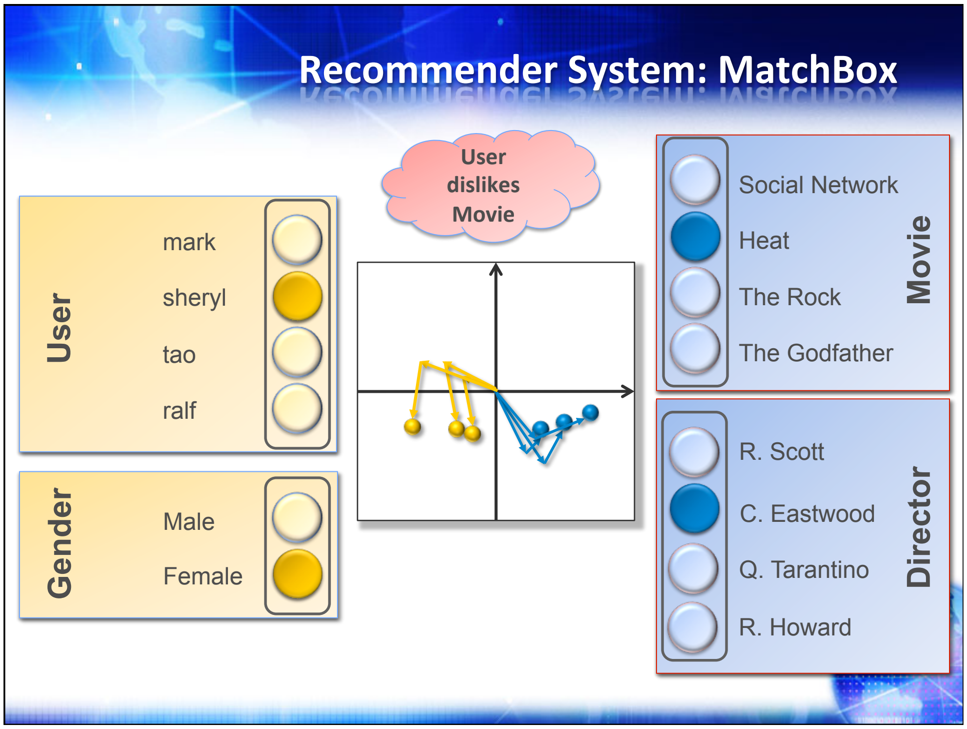


Matchbox With Metadata

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 dislikes Movie

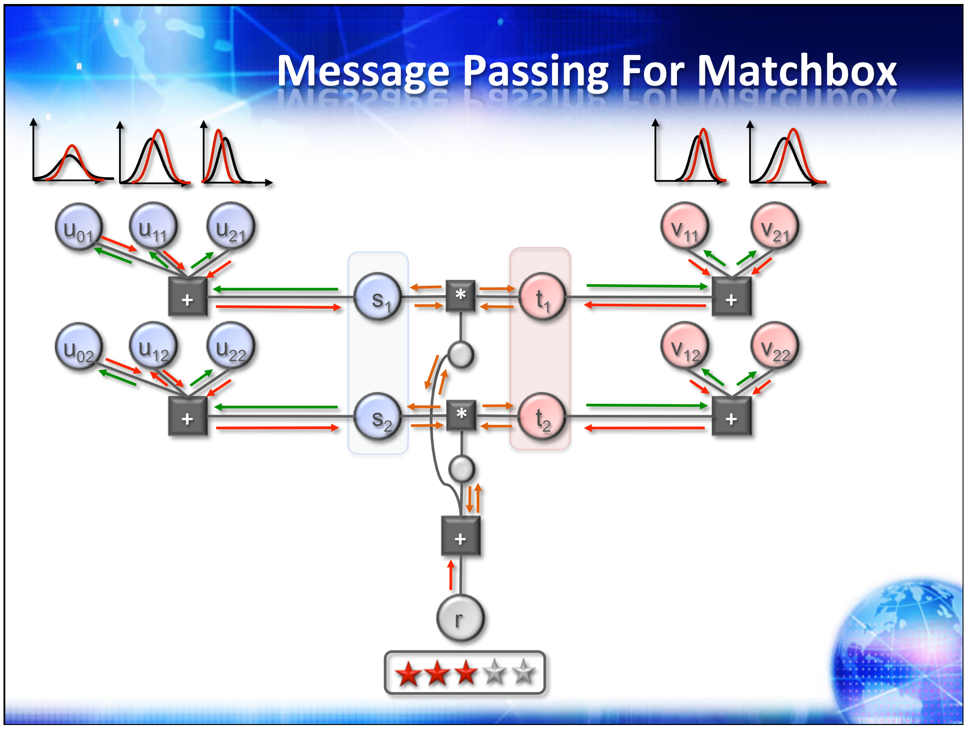
User: mark, sheryl, tao, ralf

Gender: Male, Female

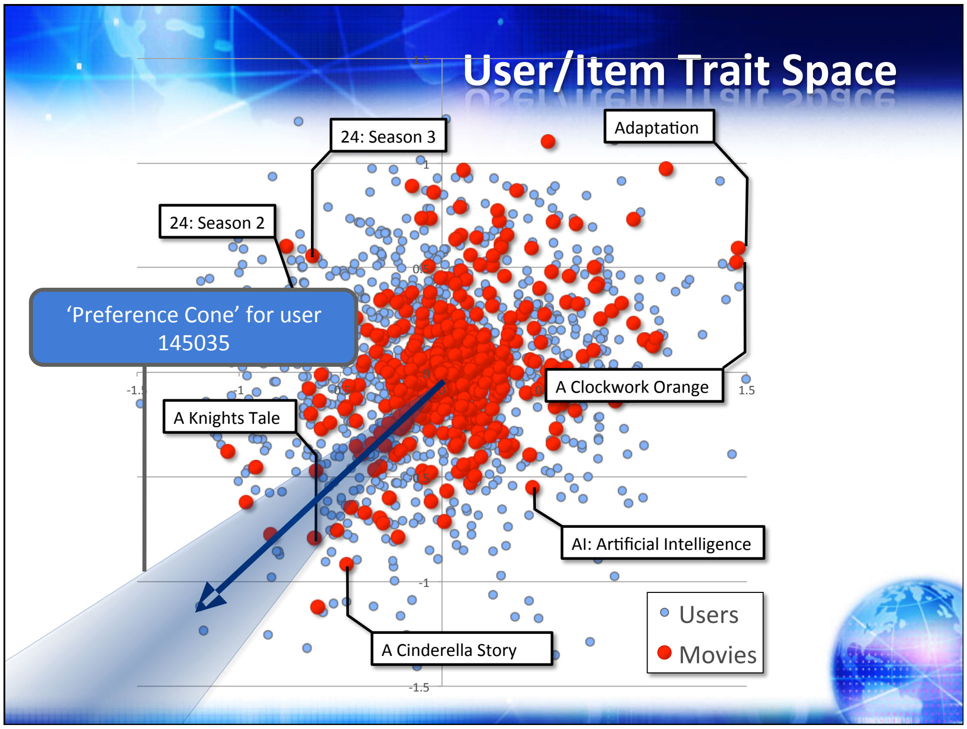
Movie: Social Network, Heat, The Rock, The Godfather

Director: R. Scott, C. Eastwood, Q. Tarantino, R. Howard

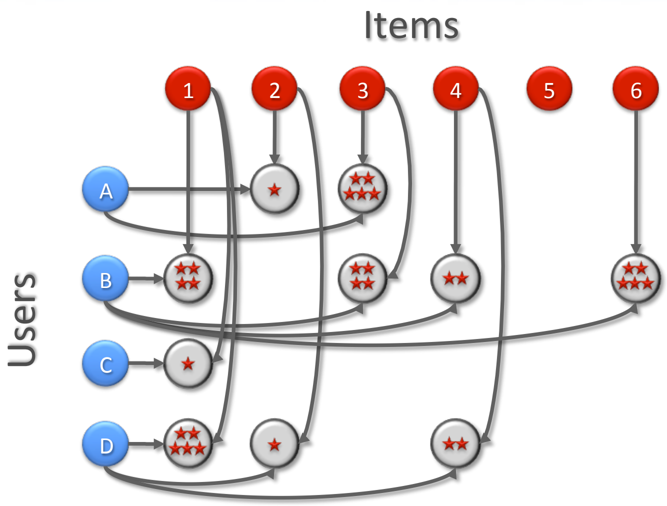
Message Passing For Matchbox



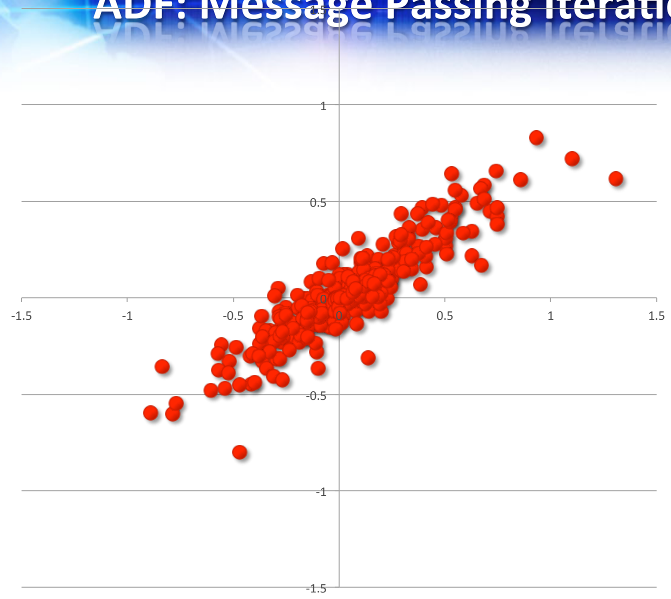
User/Item Trait Space

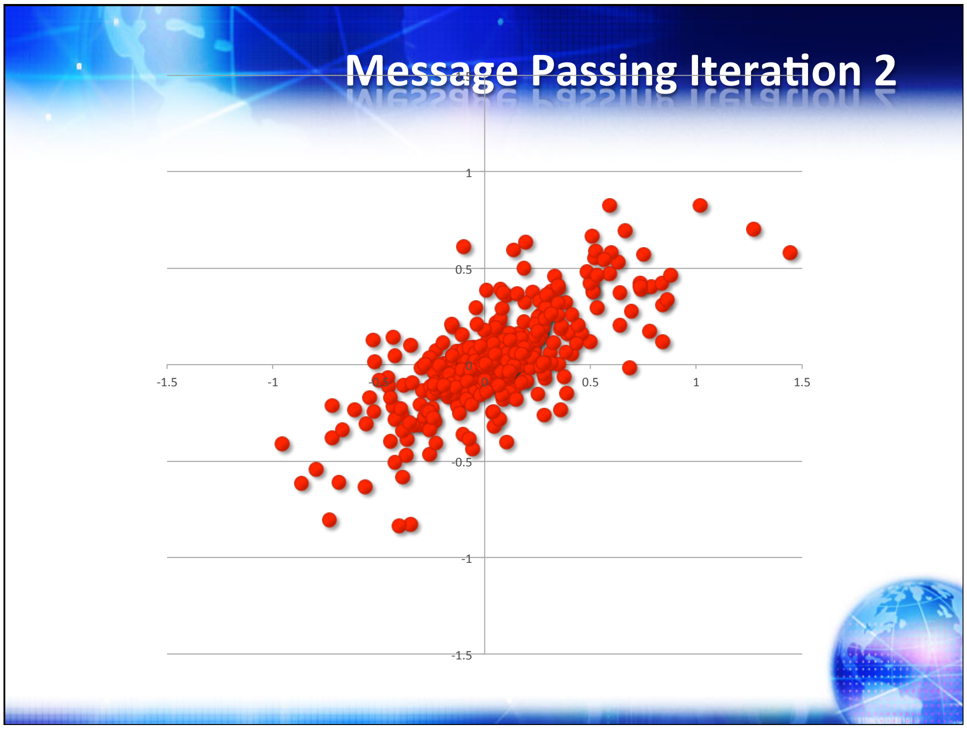


Incremental Training with ADF

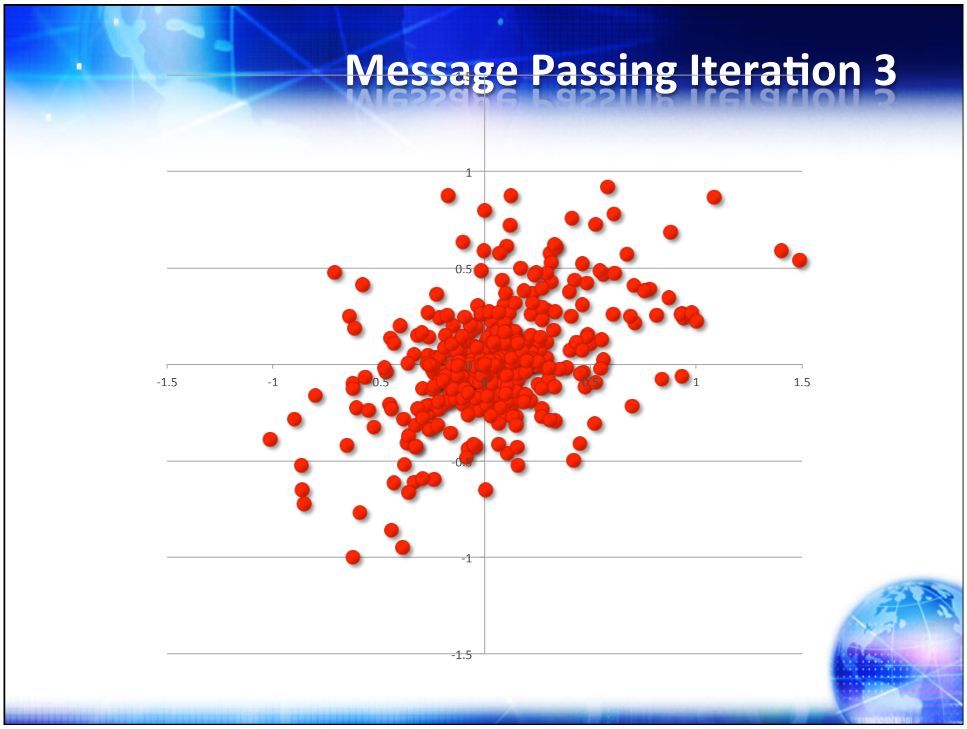


ADF: Message Passing Iteration 1



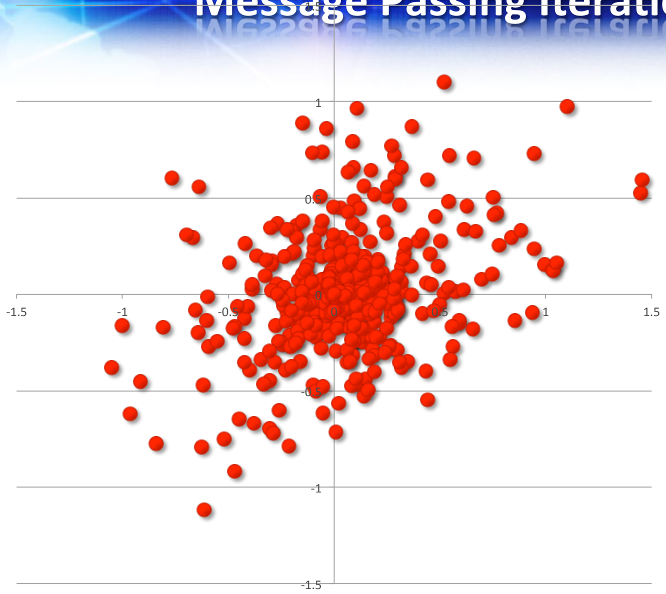


Message Passing Iteration 2



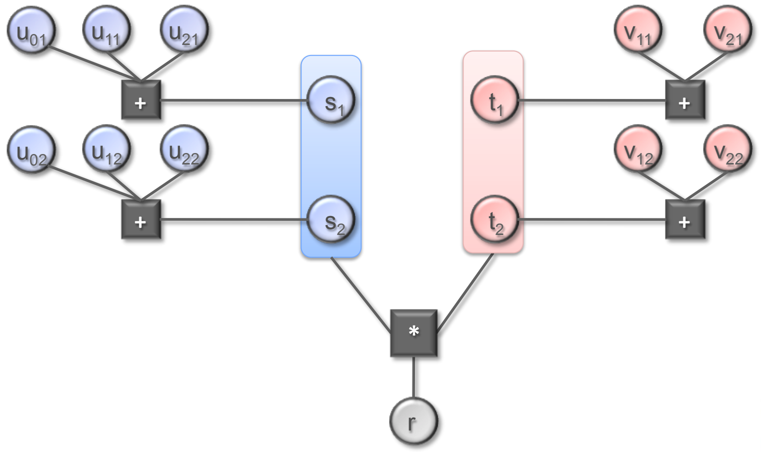
Message Passing Iteration 3

Message Passing Iteration 4

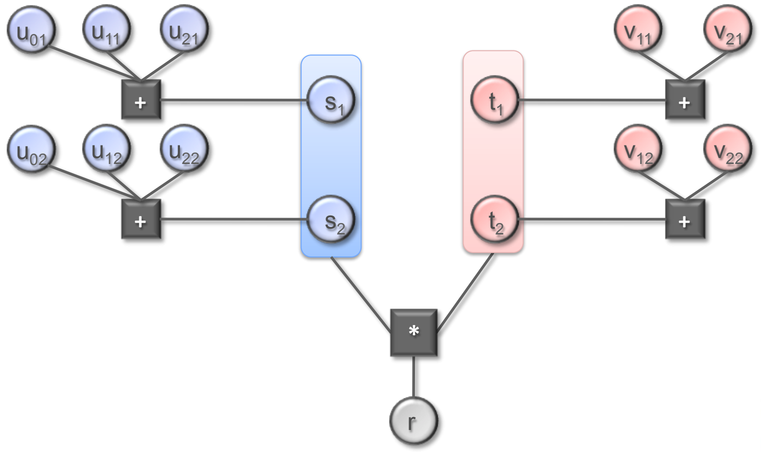


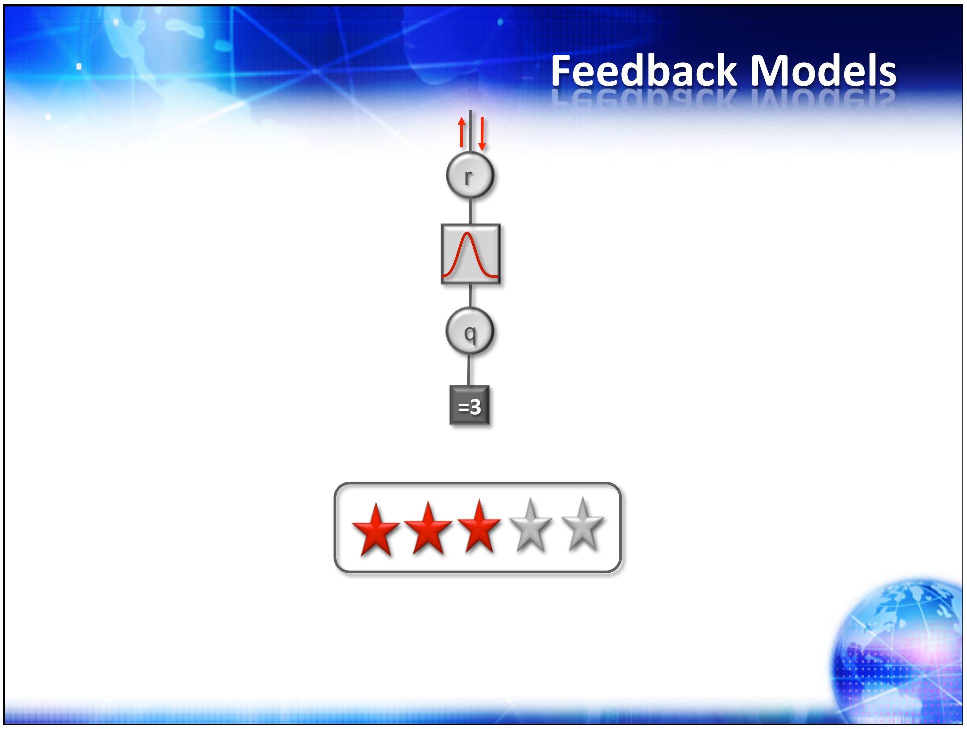
feedback models

Feedback Models

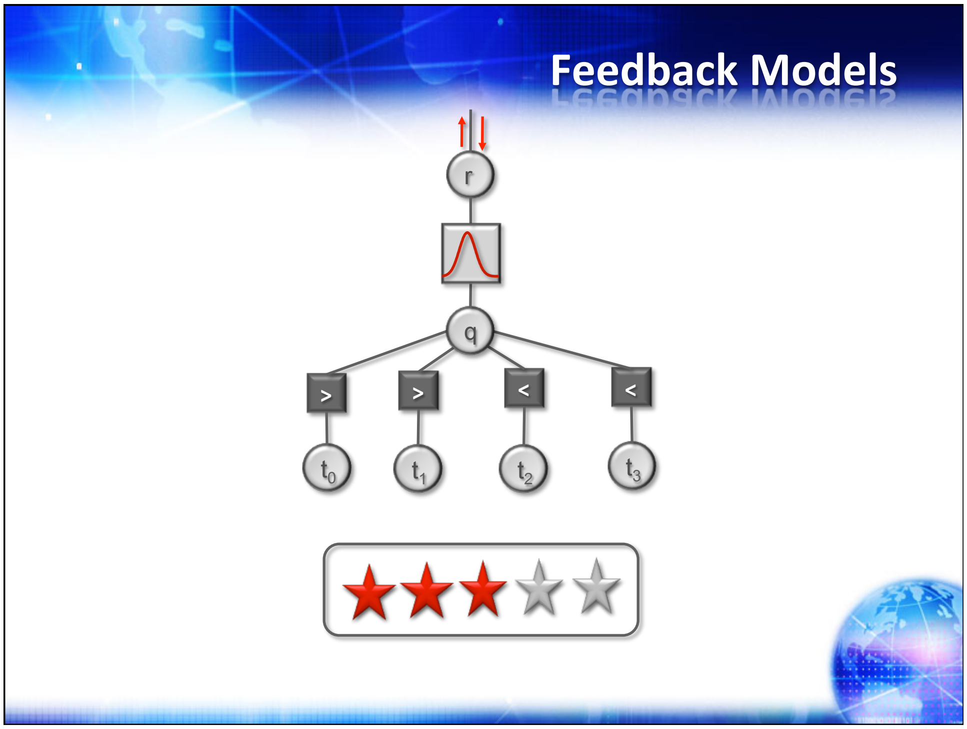


Feedback Models

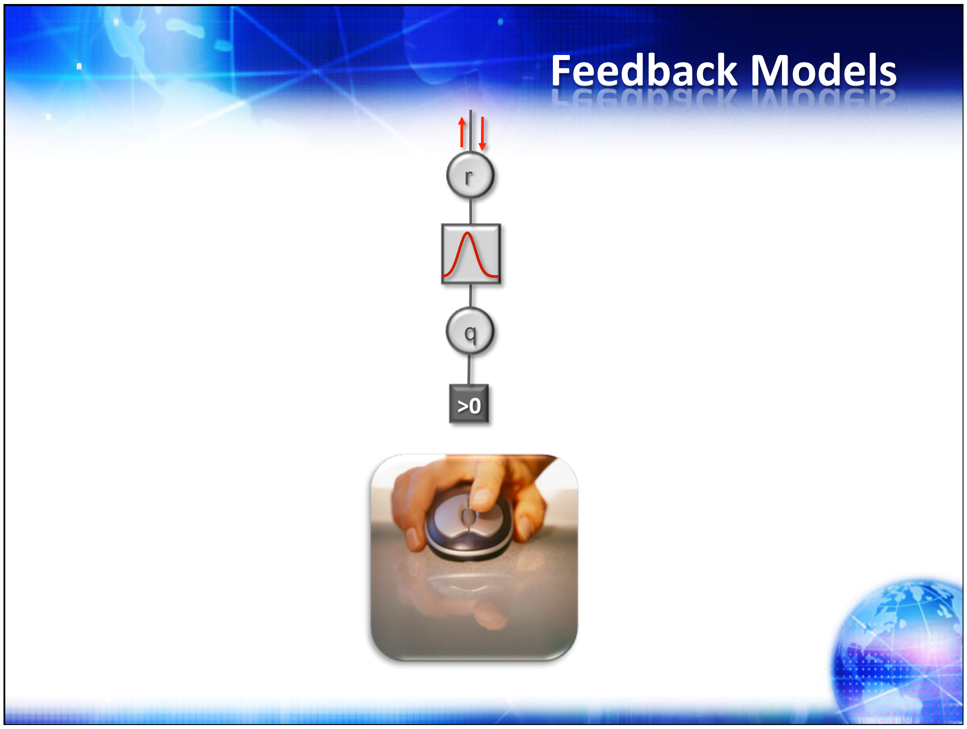




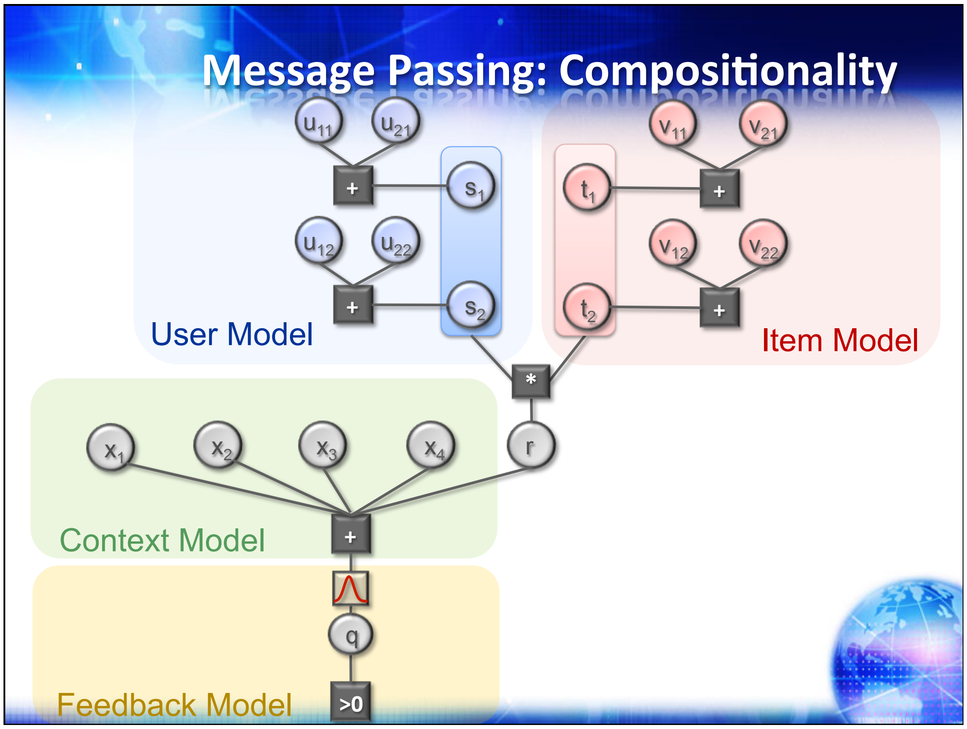
Feedback Models



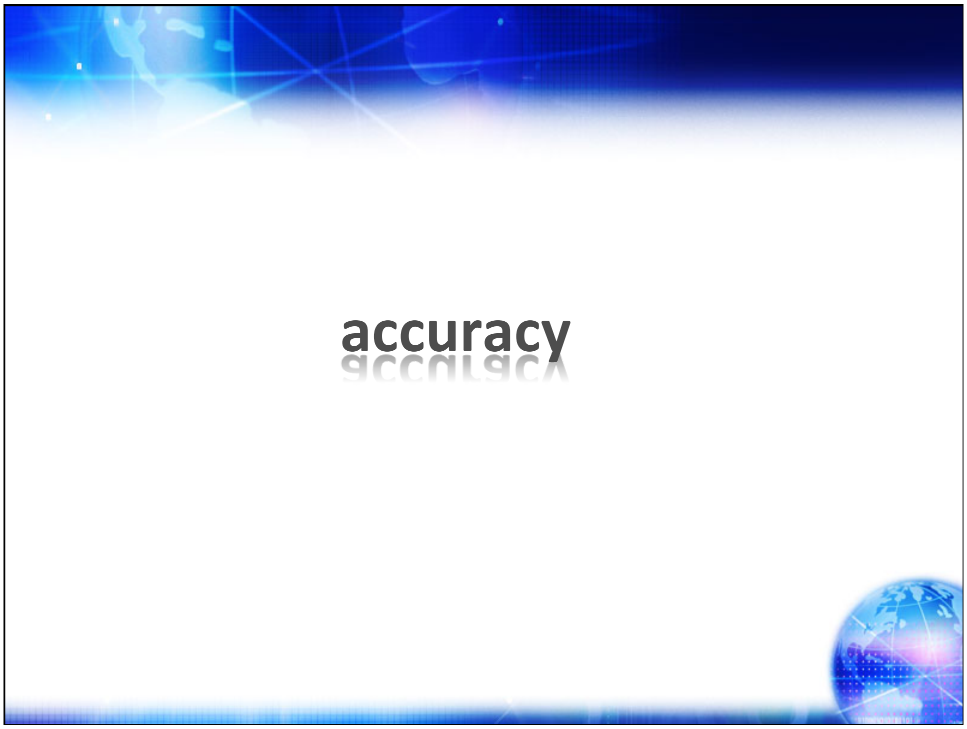
Feedback Models



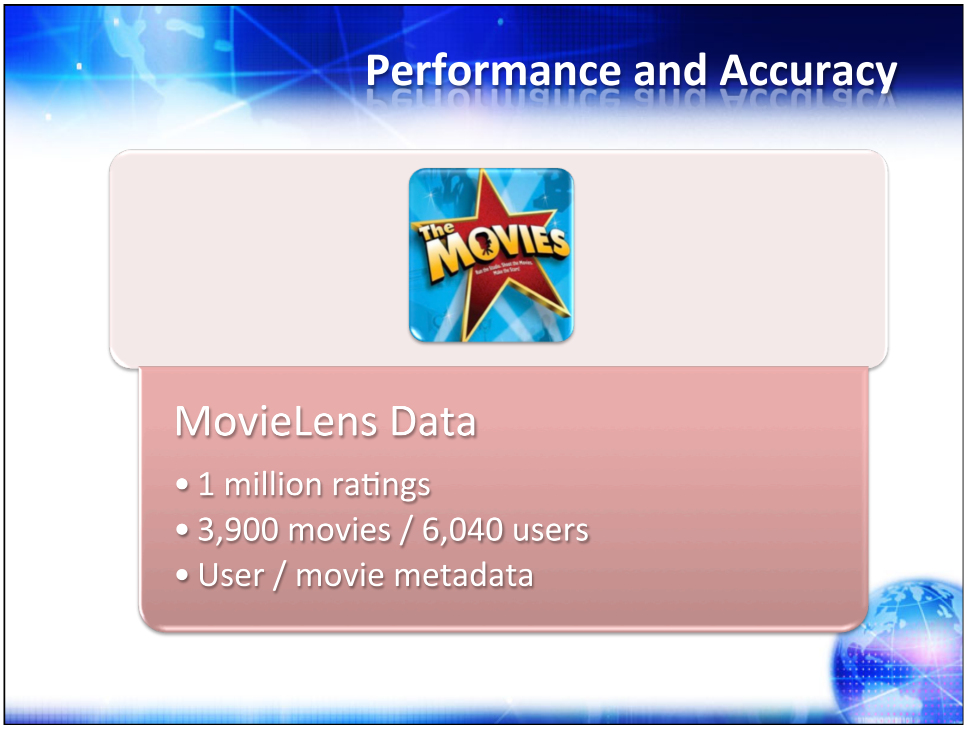
Feedback Models



Message Passing: Compositionality



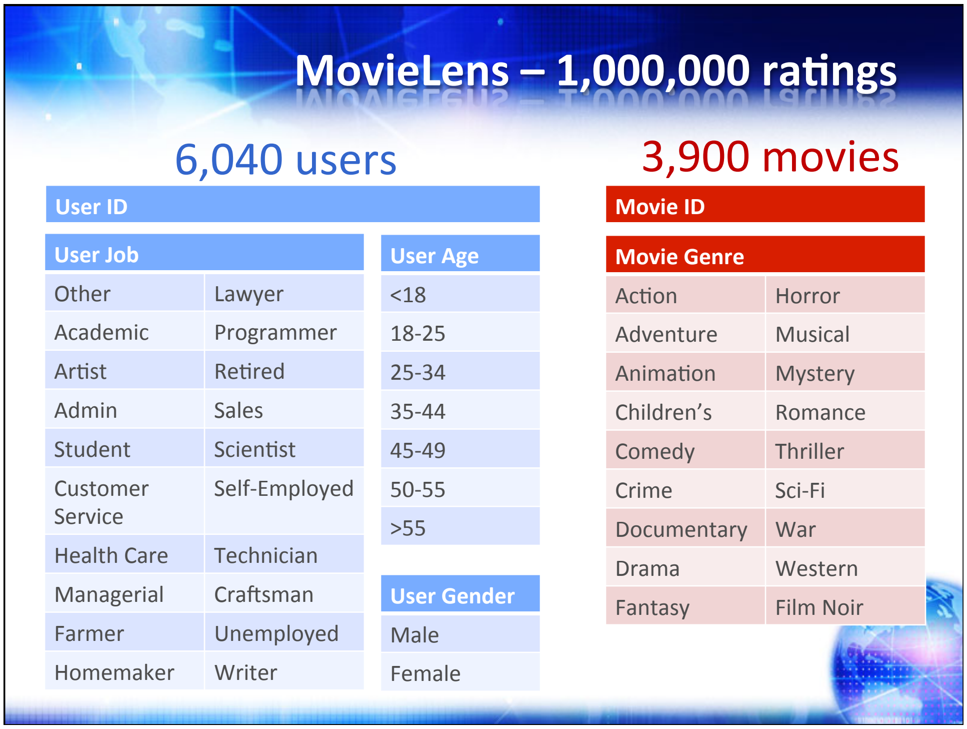
accuracy



Performance and Accuracy

MovieLens Data

- 1 million ratings
- 3,900 movies / 6,040 users
- User / movie metadata



MovieLens – 1,000,000 ratings

6,040 users

User ID

User Job	
Other	Lawyer
Academic	Programmer
Artist	Retired
Admin	Sales
Student	Scientist
Customer Service	Self-Employed
Health Care	Technician
Managerial	Craftsman
Farmer	Unemployed
Homemaker	Writer

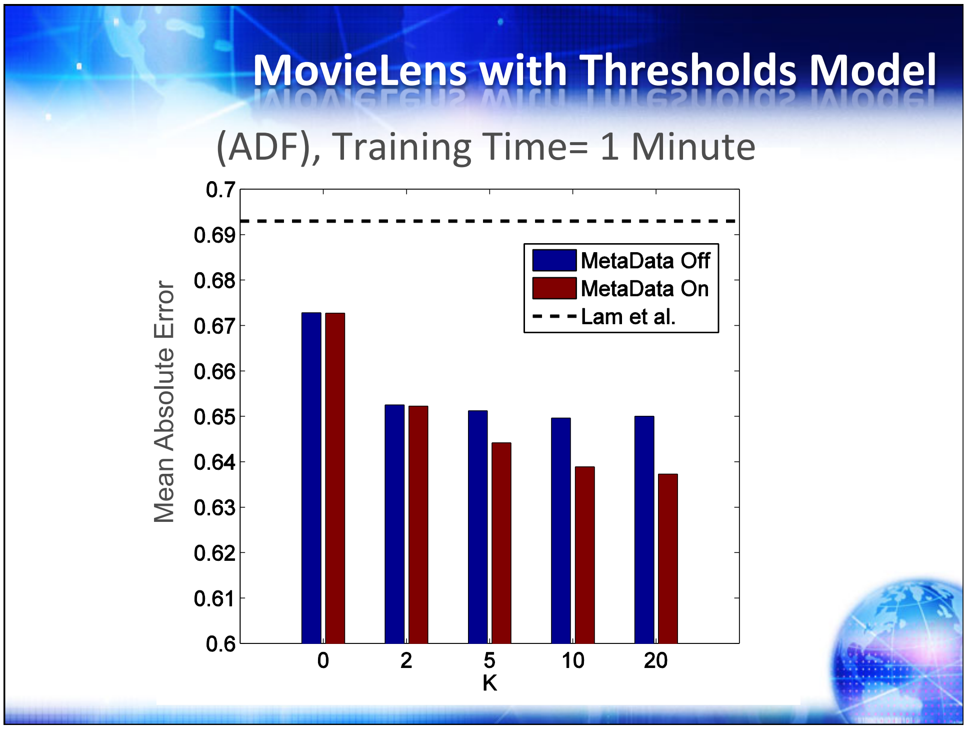
User Age
<18
18-25
25-34
35-44
45-49
50-55
>55

User Gender
Male
Female

3,900 movies

Movie ID

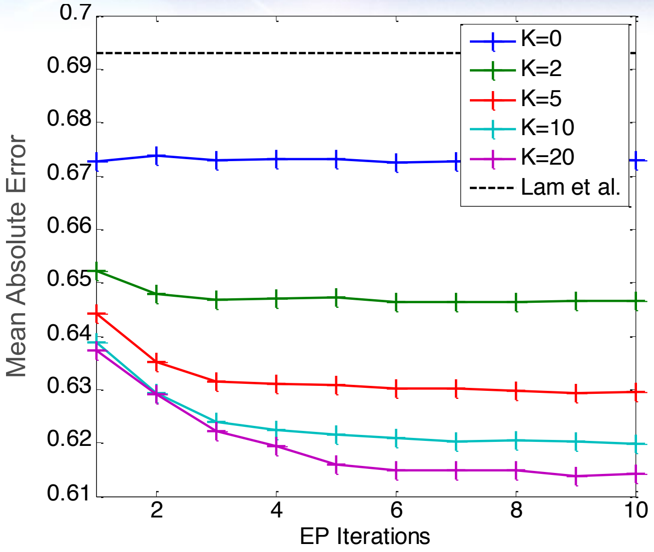
Movie Genre	
Action	Horror
Adventure	Musical
Animation	Mystery
Children’s	Romance
Comedy	Thriller
Crime	Sci-Fi
Documentary	War
Drama	Western
Fantasy	Film Noir



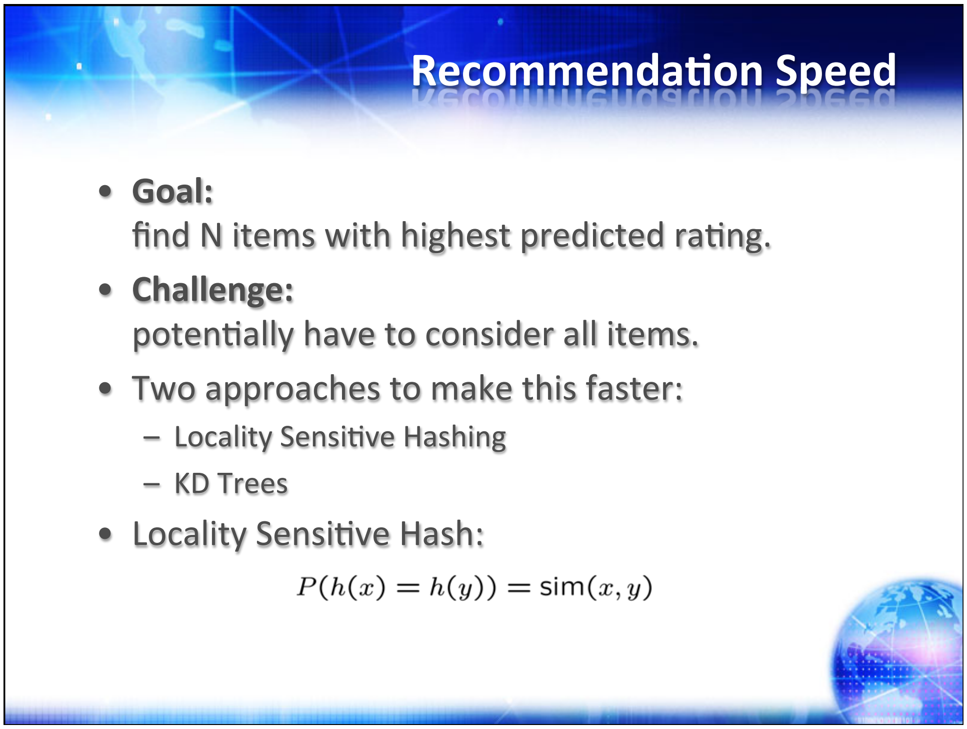
MovieLens with Thresholds Model

(ADF), Training Time= 1 Minute

MovieLens Error with Thresholds



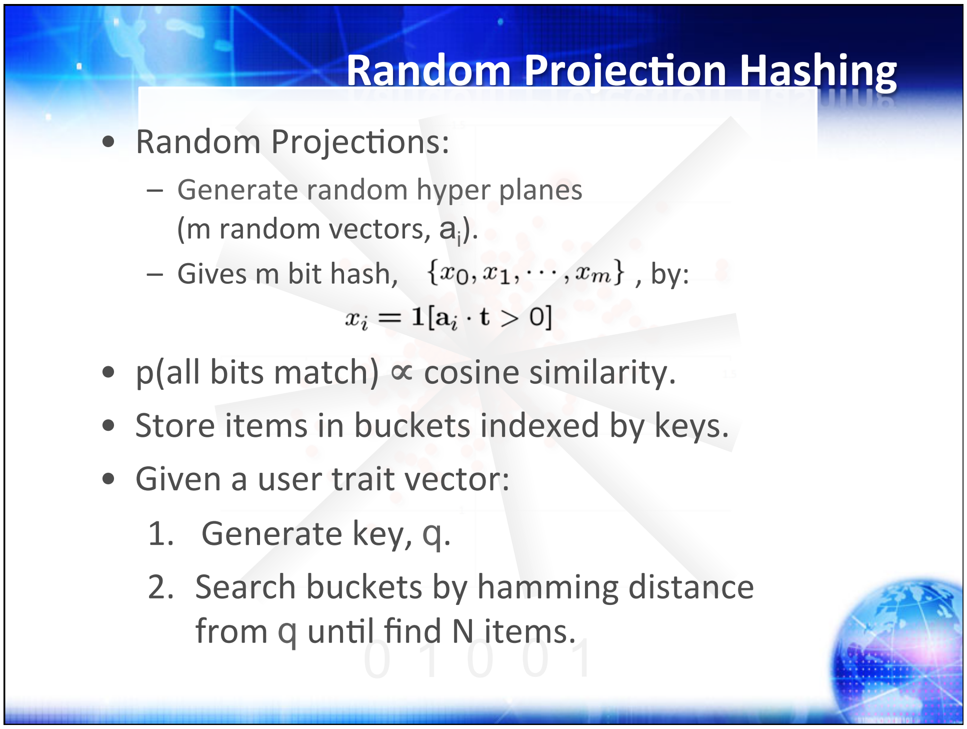
Recommendation Speed



Recommendation Speed

- **Goal:**
  find N items with highest predicted rating.
- **Challenge:**
  potentially have to consider all items.
- Two approaches to make this faster:
  - Locality Sensitive Hashing
  - KD Trees
- Locality Sensitive Hash:

$$P(h(x) = h(y)) = \text{sim}(x, y)$$

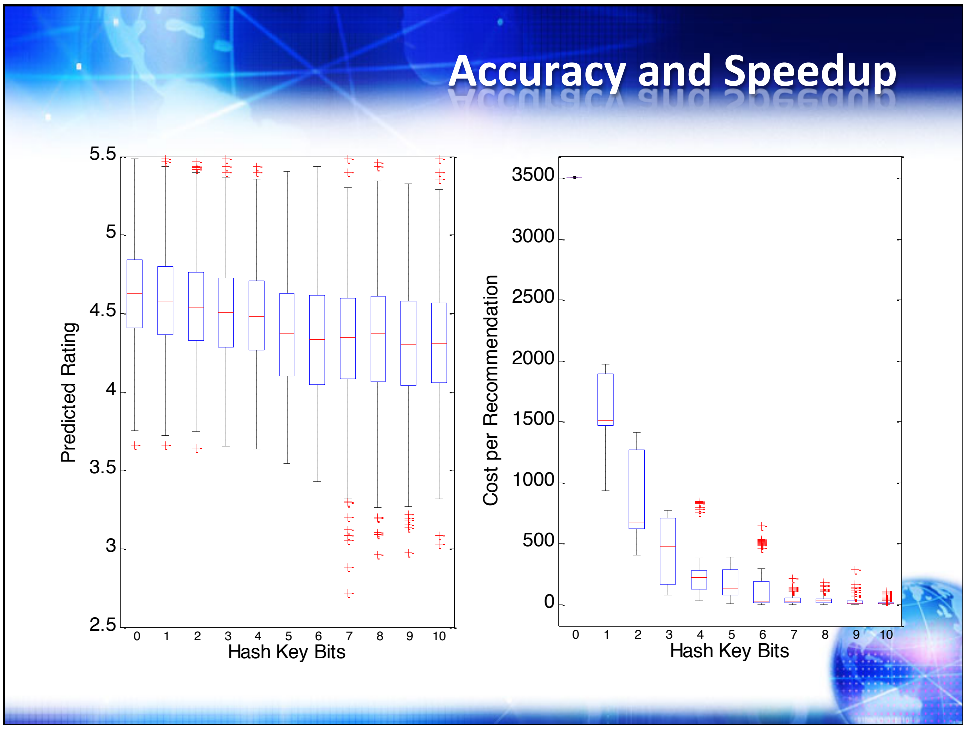


Random Projection Hashing

- Random Projections:
  - Generate random hyper planes (m random vectors, a_i).
  - Gives m bit hash, $\{x_0, x_1, \cdots, x_m\}$, by:

$$x_i = \mathbf{1}[\mathbf{a}_i \cdot \mathbf{t} > 0]$$

- p(all bits match) ∝ cosine similarity.
- Store items in buckets indexed by keys.
- Given a user trait vector:
  1. Generate key, q.
  2. Search buckets by hamming distance from q until find N items.



Accuracy and Speedup

Thanks!