Human Ranking of Machine Translation

Matt Post
Johns Hopkins University

University of Pennsylvania
April 9, 2015

JOHNS HOPKINS UNIVERSITY

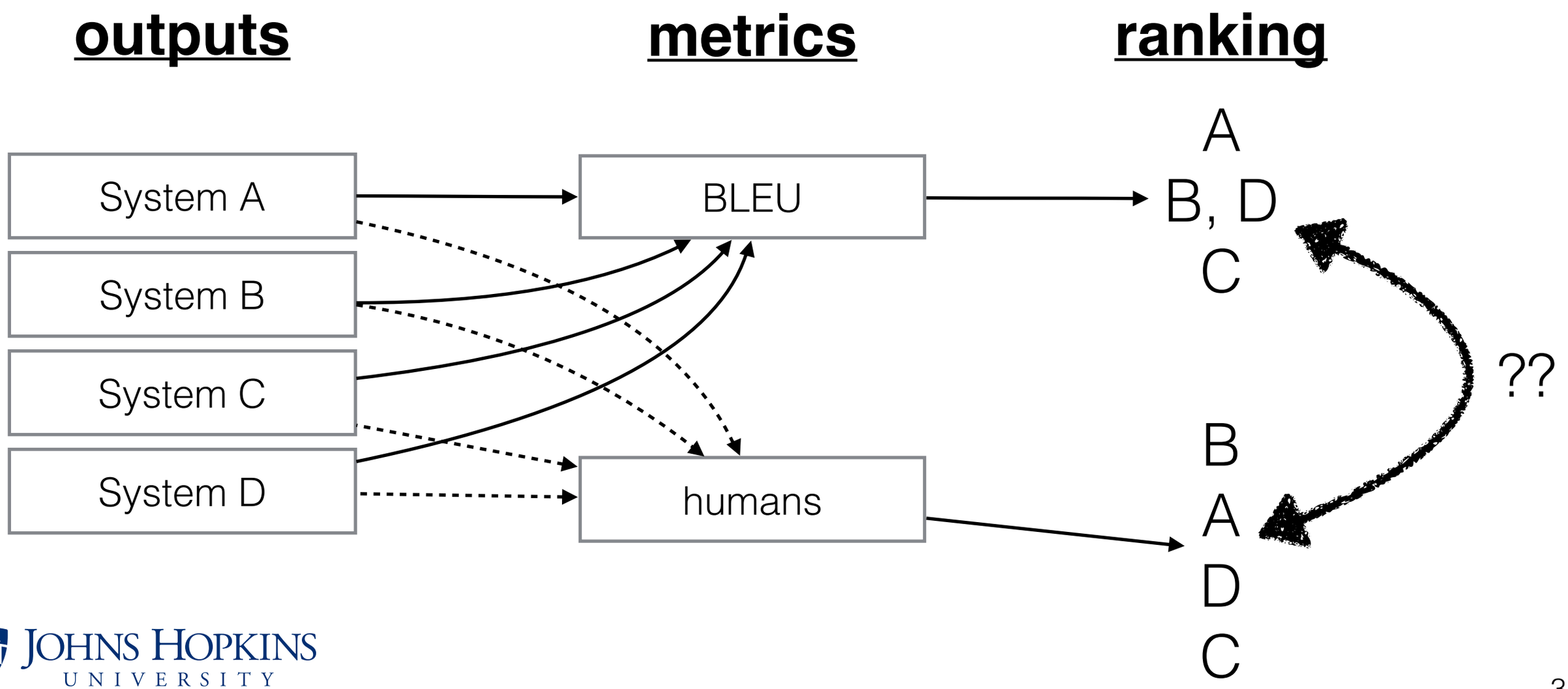
Some slides and ideas borrowed from Adam Lopez (Edinburgh)

Review

- In translation, human evaluations are what matter
  - but they are expensive to run
  - this holds up science!
- The solution is automatic metrics
  - fast, cheap, (usually) easy to compute
  - deterministic

Review

- Automatic metrics produce a *ranking*
- They are evaluated using correlation statistics against *human judgments*



Review

- The human judgments are the “gold standard”
- Questions:
  1. How do we get this gold standard?
  2. How do we know it’s correct?

Today

- **How we produce the gold-standard ranking**
- **How we know it's correct**

JOHNS HOPKINS UNIVERSITY

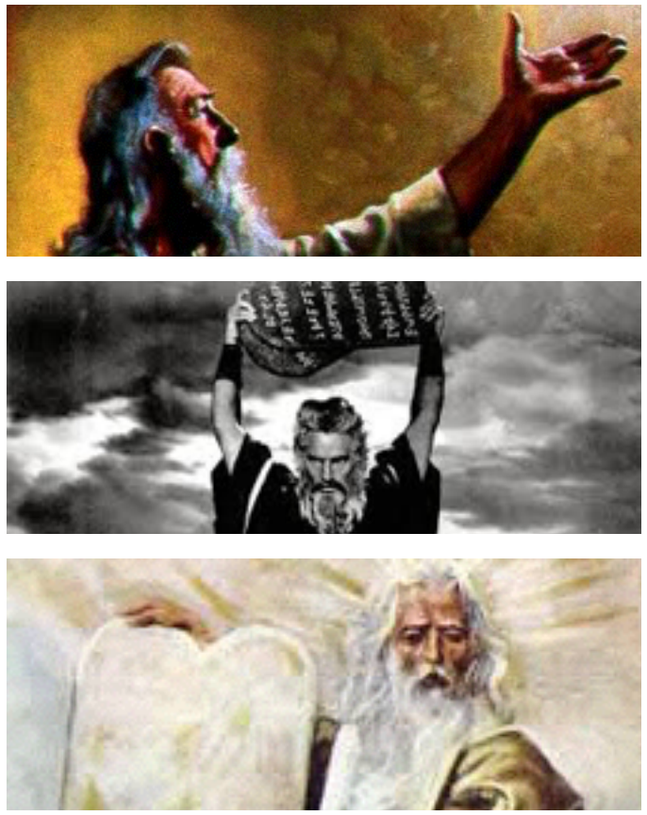
At the end of this lecture…

- You should understand
  - how to rank with incomplete
  - how to evaluate truth claims in science
- You might come away with
  - a desire to submit your metric to the WMT metrics task (deadline: May 25, 2015)
  - a desire to buy an Xbox
  - a preference for simplicity

Producing a ranking

- Then, we take this data and produce a ranking
- Outline of the rest of the talk

Human ranking methods

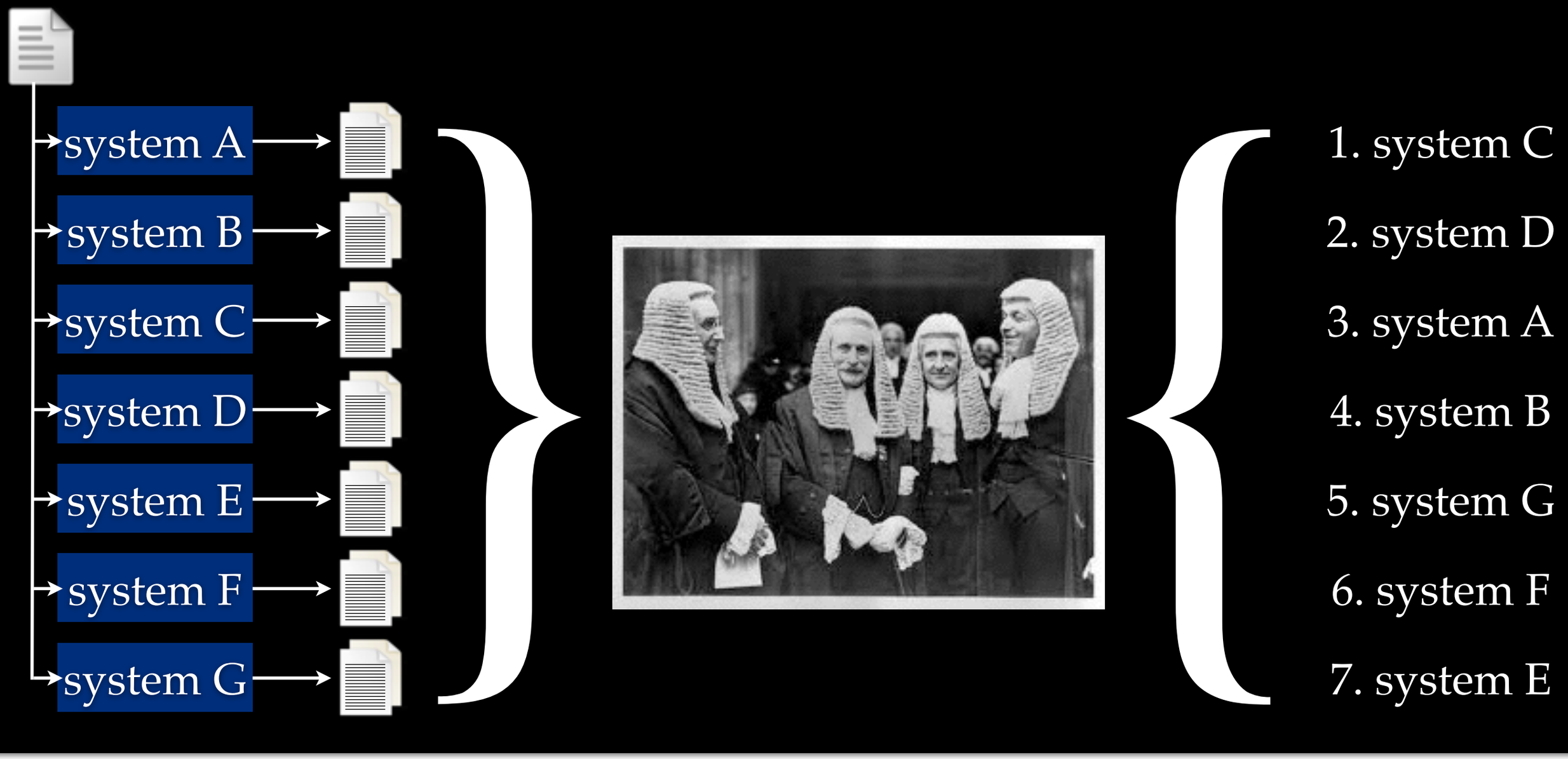


Model selection

Clustering

JOHNS HOPKINS UNIVERSITY

Goal



JOHNS HOPKINS UNIVERSITY

Slide from Adam Lopez

Goal

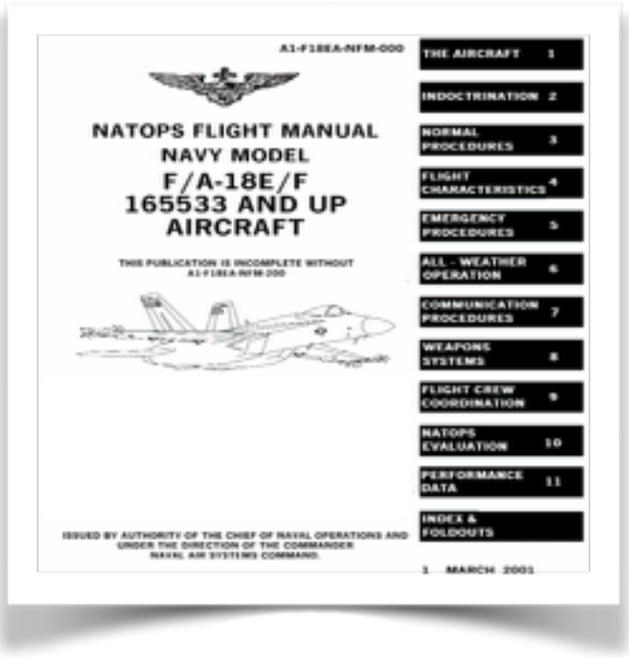
- Produce a ranking of *systems*
- There are many ways to do this:
  - Reading comprehension tests
  - Time spent on human post-editing
  - Aggregating sentence-level judgments
- This last one is what is used by the Workshop on Statistical Machine Translation (statmt.org/wmt15)

JOHNS HOPKINS UNIVERSITY

Inherent problems

- Translation is used for a range of tasks

Understanding the past



Technical manuals

Conversing

Information

- What *best* (or sufficient) means likely varies by person and situation

Collecting data

- Data: K systems translate an N-sentence document
- We use human judges to compare translations of an input sentence and select whether

  the first is *better*, *worse*, or *equivalent to* the second

- We use a large pool of judges

"Valentino měl vždycky raději eleganci než slávu.

— Source

Valentino has always preferred elegance to notoriety.

— Reference

Best ← Rank 1 ○ Rank 2 ◉ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

"Valentino should always elegance rather than fame.

— Translation 1

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ◉ Rank 4 ○ Rank 5 ○ → Worst

"Valentino has always rather than the elegance of glory.

— Translation 2

Best ← Rank 1 ◉ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

" Valentino had always preferred elegance than glory.

— Translation 3

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ◉ Rank 5 ○ → Worst

"Valentino has always had the elegance rather than glory.

— Translation 4

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ◉ → Worst

`` Valentino has always had a rather than the elegance of the glory.

— Translation 5

Collecting data

"Valentino měl vždycky raději eleganci než slávu.
— Source

Valentino has always preferred elegance to notoriety.
— Reference

Best ← Rank 1 ○ Rank 2 ◉ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
"Valentino should always elegance rather than fame.
— Translation 1

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ◉ Rank 4 ○ Rank 5 ○ → Worst
"Valentino has always rather than the elegance of glory.
— Translation 2

Best ← Rank 1 ◉ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
" Valentino had always preferred elegance than glory.
— Translation 3

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ◉ Rank 5 ○ → Worst
"Valentino has always had the elegance rather than glory.
— Translation 4

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ◉ → Worst
`` Valentino has always had a rather than the elegance of the glory.
— Translation 5

C > A > B > D > E

C > A	A > B	B > D	D > E
C > B	A > D	B > E	
C > D	A > E		
C > E			

ten pairwise judgments

JOHNS HOPKINS UNIVERSITY

Dataset

- This yields ternary-valued pairwise judgments of the following form

judge “dredd” ranked onlineB > JHU on sent #74
judge “judy” ranked uedin > UU on sent #1734
judge “reinhold” ranked JHU > UU on sent #1
judge “jay” ranked onlineA = uedin on sent #953
…

The sample space

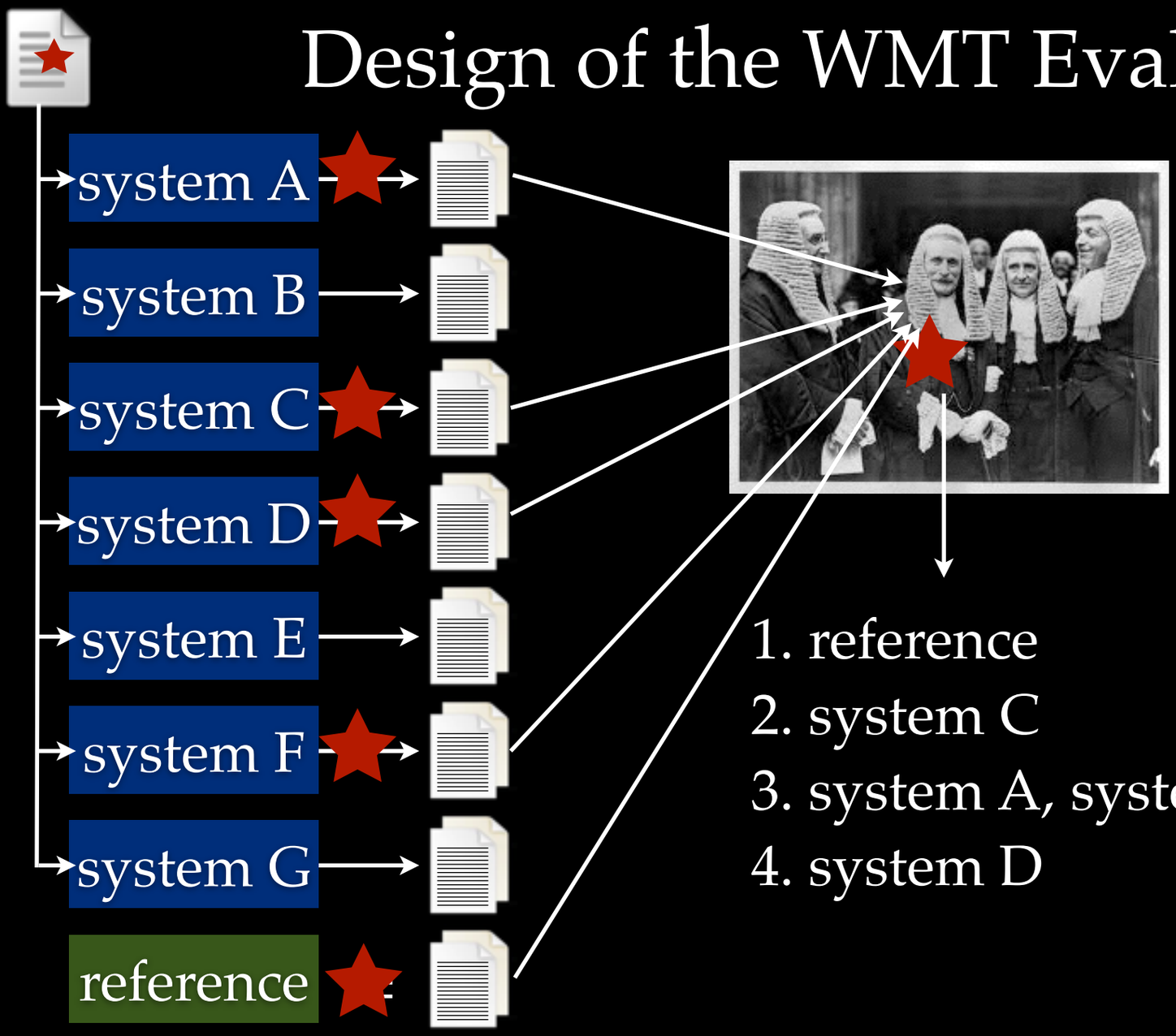
- How much data is there to collect?

(number of ways to pick two systems)
x (number of sentences) x (number of judges)

  - For 10 systems there are 135k comparisons
  - For 20 systems, 570k
  - More with multiple judges

- Too much to collect, also wasteful; instead we sample

Design of the WMT Evaluation (2008-2011)



1. reference
2. system C
3. system A, system F
4. system D

WMT Raw Data: pairwise rankings

$$
\begin{cases}
\text{reference} \prec \text{system A} \\
\text{reference} \prec \text{system C} \\
\text{reference} \prec \text{system D} \\
\text{reference} \prec \text{system F} \\
\text{system A} \succ \text{system C} \\
\text{system A} \prec \text{system D} \\
\text{system A} \equiv \text{system F} \\
\text{system C} \prec \text{system D} \\
\text{system C} \prec \text{system F} \\
\text{system D} \prec \text{system F}
\end{cases}
$$

While (evaluation period is not over):

- ➡ Sample input sentence.
- ➡ Sample five translators of it from *Systems* ∪ *{Reference}*.
- ➡ Sample a judge.
- ➡ Receive set of pairwise judgments from the judge.

How much data do we collect?

Language Pair	Systems	Rankings	Average
Czech–English	5	21,130	4,226.0
English–Czech	10	55,900	5,590.0
German–English	13	25,260	1,943.0
English–German	18	54,660	3,036.6
French–English	8	26,090	3,261.2
English–French	13	33,350	2,565.3
Russian–English	13	34,460	2,650.7
English–Russian	9	28,960	3,217.7
Hindi–English	9	20,900	2,322.2
English–Hindi	12	28,120	2,343.3
TOTAL WMT 14	110	328,830	2,989.3
WMT13	148	942,840	6,370.5
WMT12	103	101,969	999.6
WMT11	133	63,045	474.0

of tens of millions possible

Producing a ranking

- Then, we take this data and produce a ranking
- Human ranking methods

Expected wins and variants

Bayesian model (relative ability)

TrueSkill™

Expected wins (1)

- This most appealing and intuitive approach
- Define `wins(A)`, `ties(A)`, and `loses(A)` as the number of times system A won, tied, or lost
- Score each system as follows

$$\texttt{score(A)} = \frac{\texttt{wins(A) + ties(A)}}{\texttt{wins(A) + ties(A) + loses(A)}}$$

- Now sort by scores

Expected wins (2)

- Do you see any problems with this?

$$\texttt{score(A)} = \frac{\texttt{wins(A) + ties(A)}}{\texttt{wins(A) + ties(A) + loses(A)}}$$

- Look at a judgments:

jud **one winner, one loser** onlineB > JHU on sent #74
jud **one winner, one loser** edin > UU on sent #1734
jud **one winner, one loser** ed JHU > UU on sent #1
jud **two winners, no losers** lineA = uedin on sent #953

Expected wins (3)

- A system is rewarded as much for a tie as for a win
  - …and most systems are variations of the same underlying architecture, data

- New formula: throw away ties

$$\text{score}(A) = \frac{\text{wins}(A)}{\text{wins}(A) + \text{loses}(A)}$$

- Wait: Is this better?

A Grain of Salt for the WMT Manual Evaluation (Bojar et al., 2012)

Expected wins (4)

- Problem 2: the luck of the draw

$$\text{score}(A) = \frac{\text{wins}(A)}{\text{wins}(A) + \text{loses}(A)}$$

aggregation over
different sets of inputs
different competitors
different judges

- Consider a case where in reality B > C, but
  - B gets compared to a bunch of good systems
  - C gets compared to a bunch of bad systems
  - we could get score(C) > score(B

JOHNS HOPKINS UNIVERSITY

Expected wins (5)

- This can happen!
  - Systems include a human reference translation
  - Also include really good unconstrained commercial systems

Expected wins (6)

- Even more problems:
  - remember that the scores for a system is the percentage of time it won in comparisons across all systems
  - what if score(B) > score(C), but in direct comparisons, C was almost always better than B?
  - this leads to cycles in the ranking
- Is this a problem?

Summary

- List of problems:
  - Including ties biases similar systems, excluding discredits
  - Comparisons do not factor in difficulty of the "match" (i.e., losing to the best system should count less)
  - There are cycles in the judgments

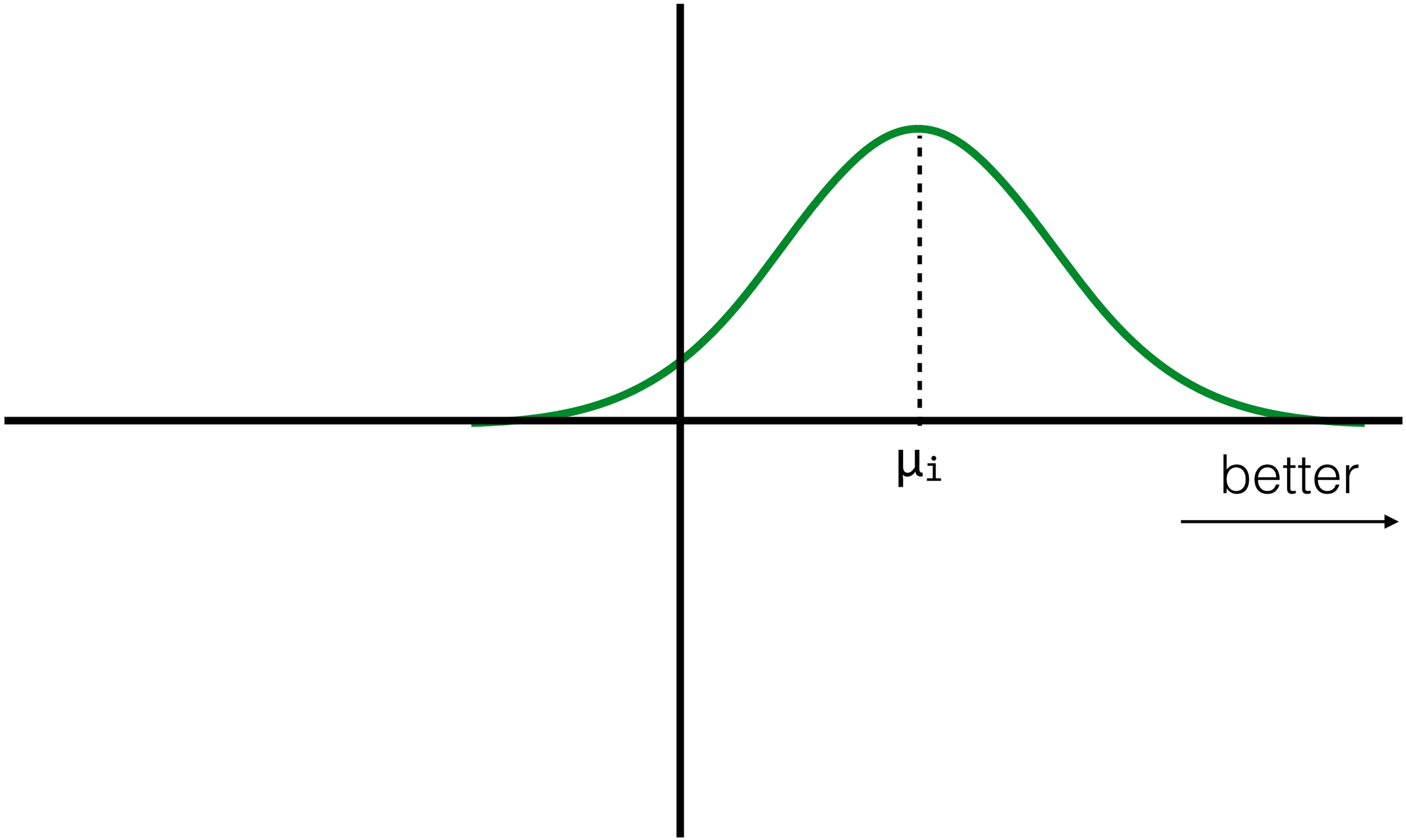
- We made intuitive changes, but how do we know whether they're correct?

Relative ability model

Models of Translation Competitions (Hopkins & May, 2013)

- In Expected Wins, we estimate a probability of each system winning a competition
- We now move to a setup that models the *relative ability* of a system
  - Assume each system S_i has an inherent ability, μ_j
  - Its translations are then represented by draws from a Gaussian distribution centered at μ_j

Relative ability



Relative ability

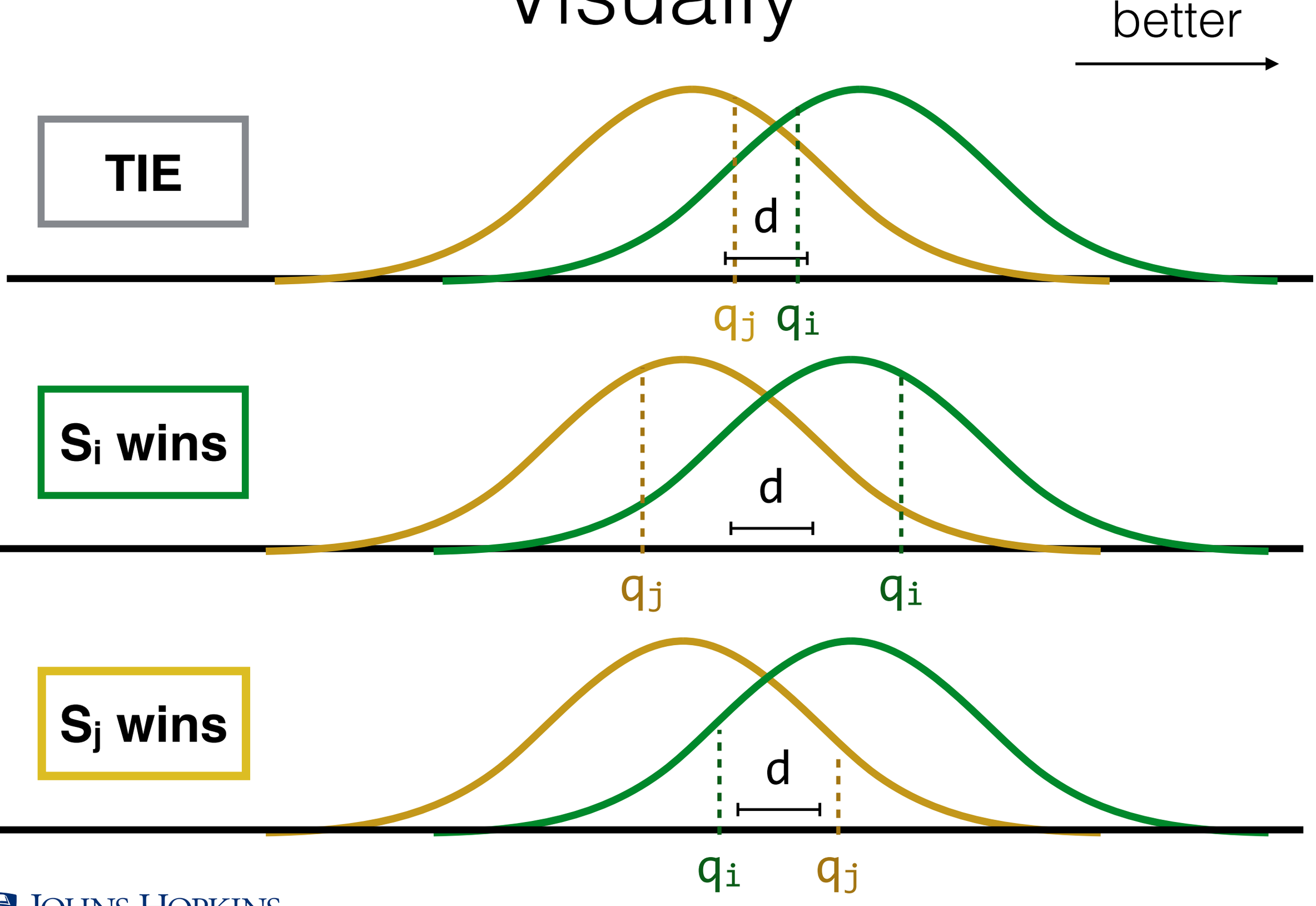
- A “competition” proceeds as follows:
  - Choose two systems, S_i and S_j, from the set {S}
  - Sample a “translation” from their distributions

$$q_i \sim N(S_i; \mu_i, \sigma 2)$$

$$q_j \sim N(S_j; \mu_j, \sigma 2)$$

  - Compare their values to determine who won
    - Define d as a “decision radius”
    - Record a tie if $|q_i - q_j| < d$
    - Else record a win or loss

Visually



Observations

- We can compute exact probabilities for all these events (difference of Gaussians)

- On average, a system with a higher “ability” will have higher draws, and will win

- Systems with close μs will tie more often

Learning the model

- If we knew the system means, we could rank them
- We assume the data was generated by the process above; we need to infer values for hidden params:
  - System means {μ}
  - Sampled translation qualities {q}
- We'll use Gibbs sampling
  - Uses simple random steps to learn a complicated joint distributions
  - Converges under certain conditions

Gibbs sampling

judge “dredd” ranked onlineB > JHU on sent #74
judge “judy” ranked uedin > UU on sent #1734
judge “reinhold” ranked JHU > UU on sent #1
judge “jay” ranked onlineA = uedin on sent #953

- Represent data as tuples `(Si, Sj, π, qi, qj)`

(onlineB, JHU, >, ?, ?)
(uedin, UU, >, ?, ?)
(JHU, UU, >, ?, ?)
(onlineA, uedin, =, ?, ?)

known **unknown**

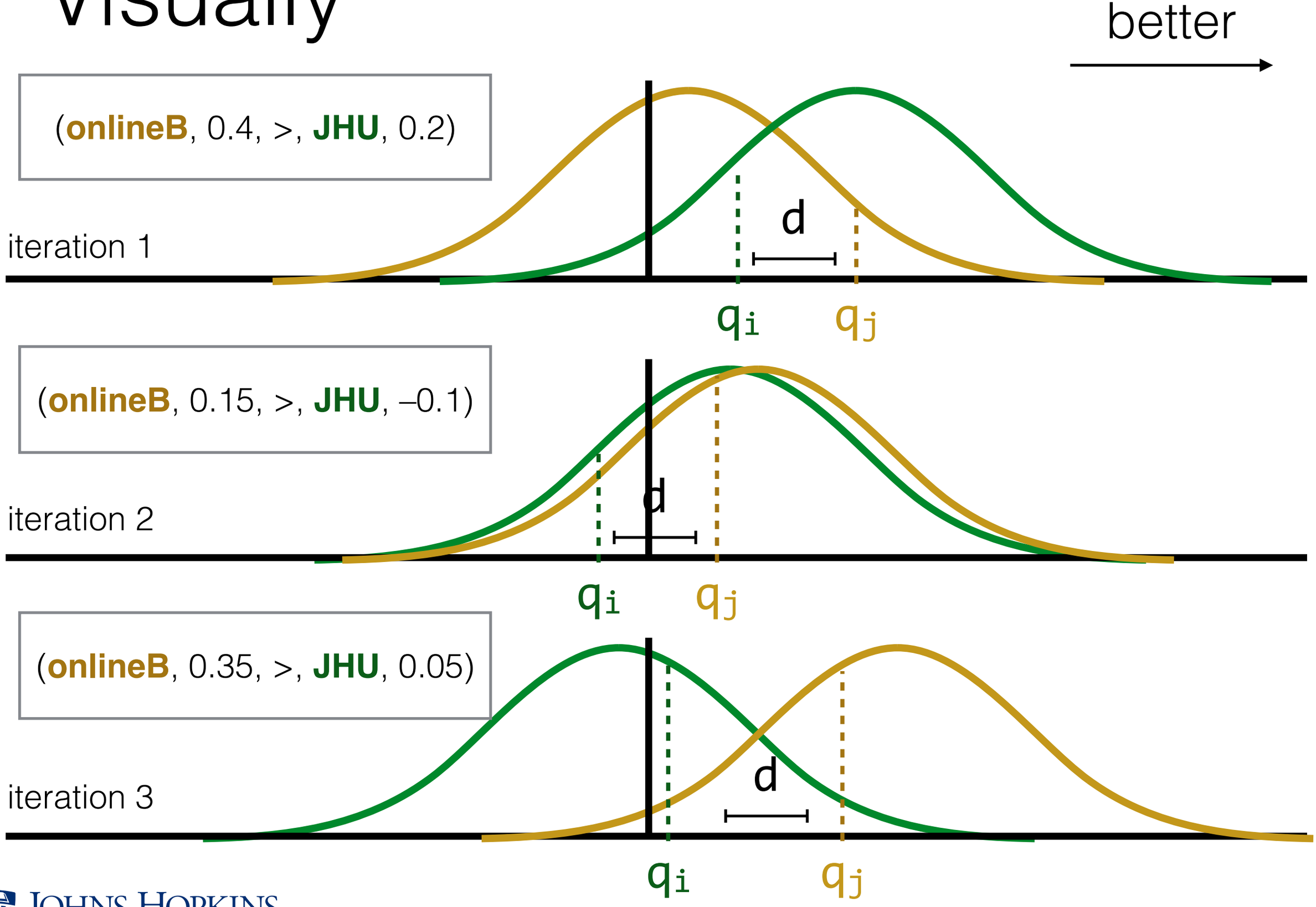
- Iterate back and forth between guessing {q}s and {μ}s

Iterative process

```
[collect all the judgments]
until convergence
  # resample translation qualities
  for each judgment
    q_i ~ N(μ_i,σ²)
    q_j ~ N(μ_j,σ²)
    # (adjust samples to respect judgment π)

# resample the system means
for each system
  μ_i = mean({q_i})
```

Visually


better
(onlineB, 0.4, >, JHU, 0.2)
iteration 1
d
q_i
q_j
(onlineB, 0.15, >, JHU, –0.1)
iteration 2
d
q_i
q_j
(onlineB, 0.35, >, JHU, 0.05)
iteration 3
d
q_i
q_j

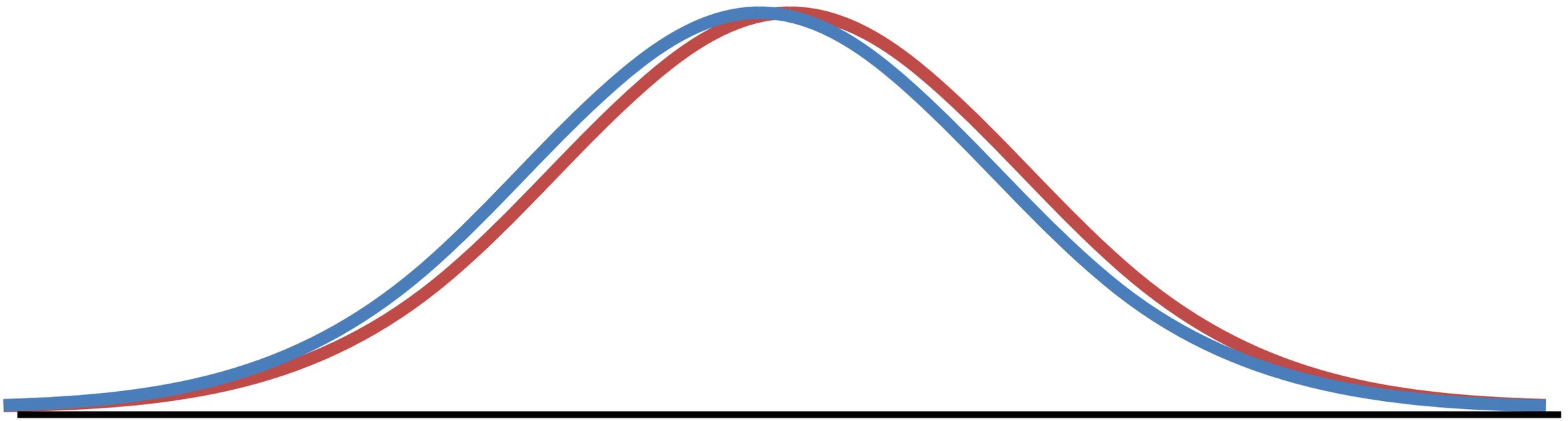
Summary

- Summary
  - Model provides us with an explanation of how the data was generated
  - We infer the abilities of the systems to rank using the human judgments
- Problems
  - Still no notion of evenness of the match
  - Judges are not modeled
  - Actual sentences are ignored

TrueSkill™ Ranking System

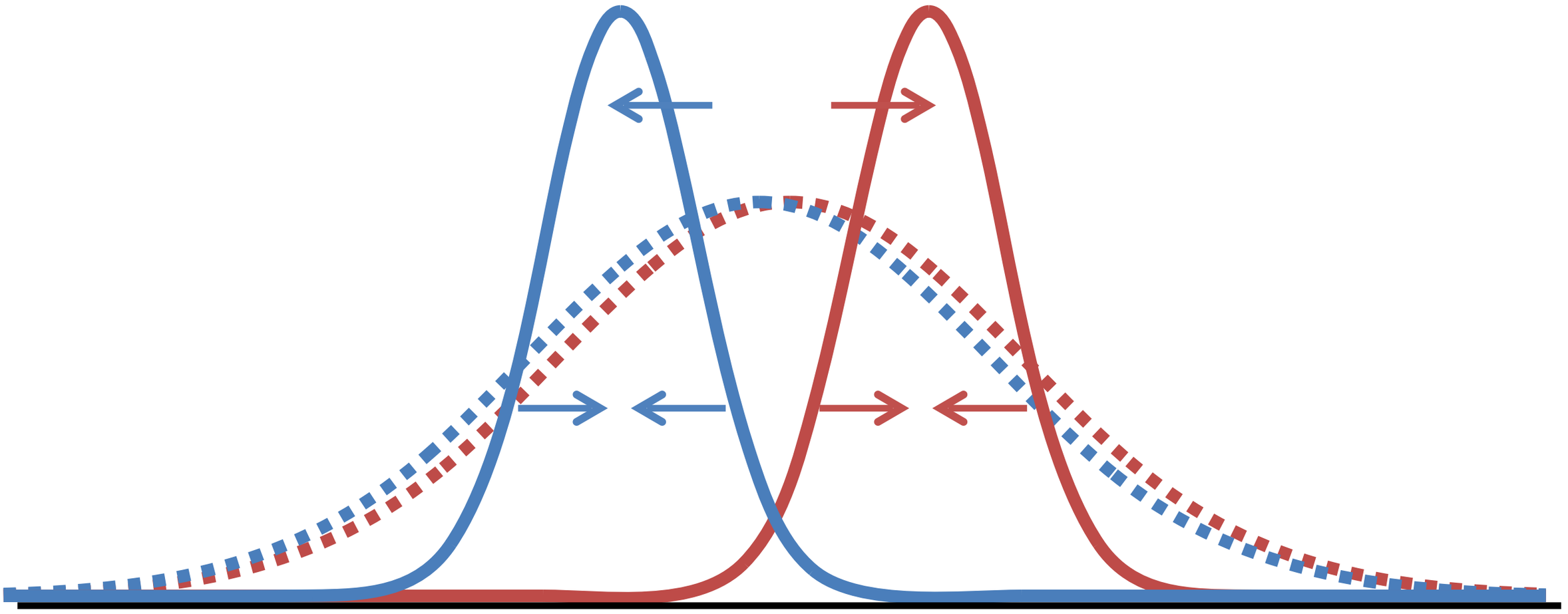
- Used to rate players in Xbox Live
- Based on the ELO system for Chess
- Models player ability (μ) *and* the system's confidence about that estimate (σ)
  - When a game is played, the outcome (win, loss, or tie) is used to update these parameters
  - A more surprising outcome results in larger updates
  - These values are also used to find even matches

Visualization



Visualization

Observation: S1 defeats S2



Not pictured: Confidences are separate for each system

JOHNS HOPKINS UNIVERSITY

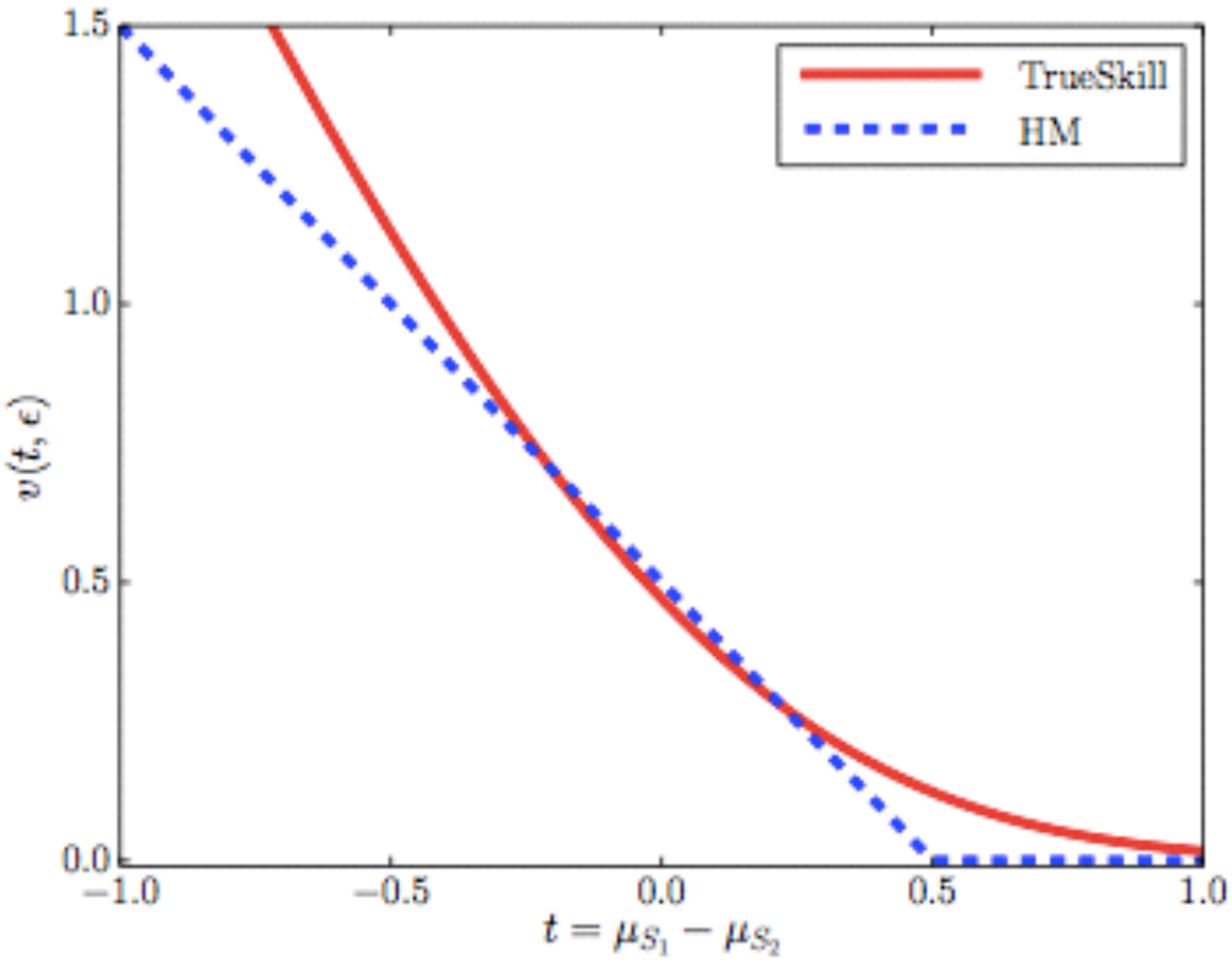
Updating

If S1 defeats S2,

$$\mu_{S_1} = \mu_{S_1} + \frac{\sigma^2_{S_1}}{c} \cdot v(\frac{t}{c}, \frac{\epsilon}{c})$$

$$\mu_{S_2} = \mu_{S_2} - \frac{\sigma^2_{S_2}}{c} \cdot v(\frac{t}{c}, \frac{\epsilon}{c})$$

outcome surprisal



TrueSkill for MT

- In the MT setting:
  - Each system is a player
  - Each pairwise annotation is a game

- We consider the judgments sequentially, an update the system parameters after each one

- Differences from Xbox:
  - Systems don't improve between games

JOHNS HOPKINS UNIVERSITY

Procedure

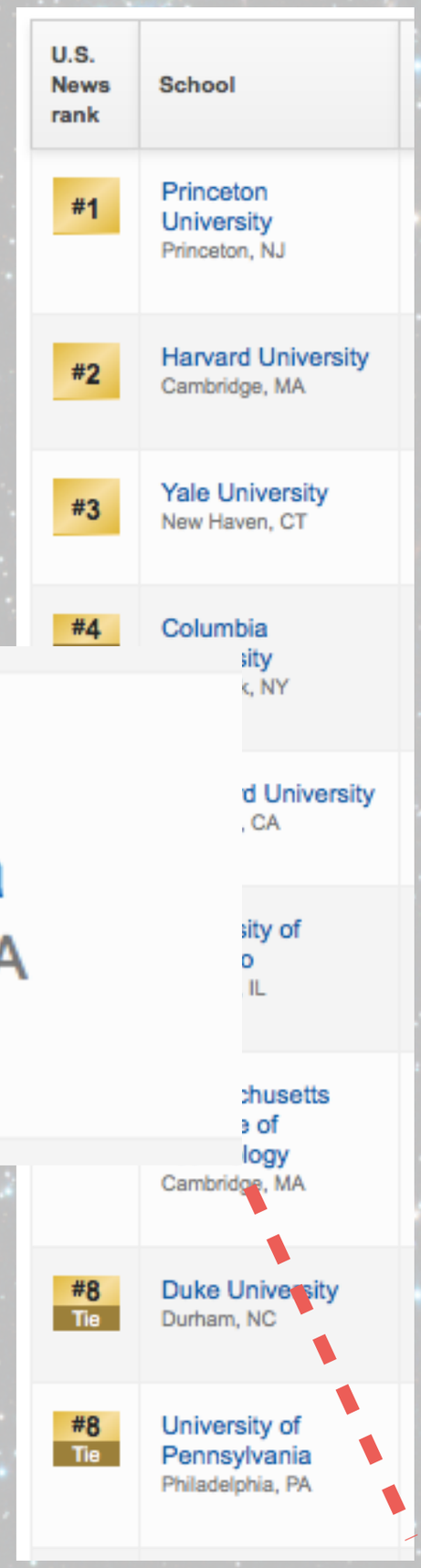
```
until convergence
  create a new match
  observe the outcome
  update the parameters of both systems
```

Advantages of TrueSkill

- The system parameter updates reflect how surprising the outcome was
- TrueSkill is an *online* algorithm (as opposed to batch)
  - Instead of sampling system pairs uniformly, we can gather more judgments from systems that are closely matched
  - This presents some potential for reducing the amount of data we need to collect

Partial orderings

- What is the best university in the world?
  - Best is not always well-defined or meaningful
- Instead of total orderi
  partial orderings, whi
  clusters of systems t
  distinguished



Simulating Human Judgment in Machine Translation Evaluation Campaigns (Koehn, 2012)

JOHNS HOPKINS UNIVERSITY

Computing clusters

- To compute clusters, we use a statistical technique called *bootstrap resampling*
  - Estimate variance by sampling the sample many times and compute statistics over the samples
- We run each model 1,000 times
  - For each system, extract rank from each fold, throw out outliers
  - Use resulting *rank range* to cluster

X
2
2
2
2
X
2
2
2
2
1
2
2
2
1

Hindi–English (WMT 2014)

rank range	constrained	unconstrained
1		online-B
2–4	uedin-syntax, cmu	online-A
5	uedin-phrase	
6–7	afrl, iit-bombay	
8	dcu-lingo24	
9		iit-hyderabad

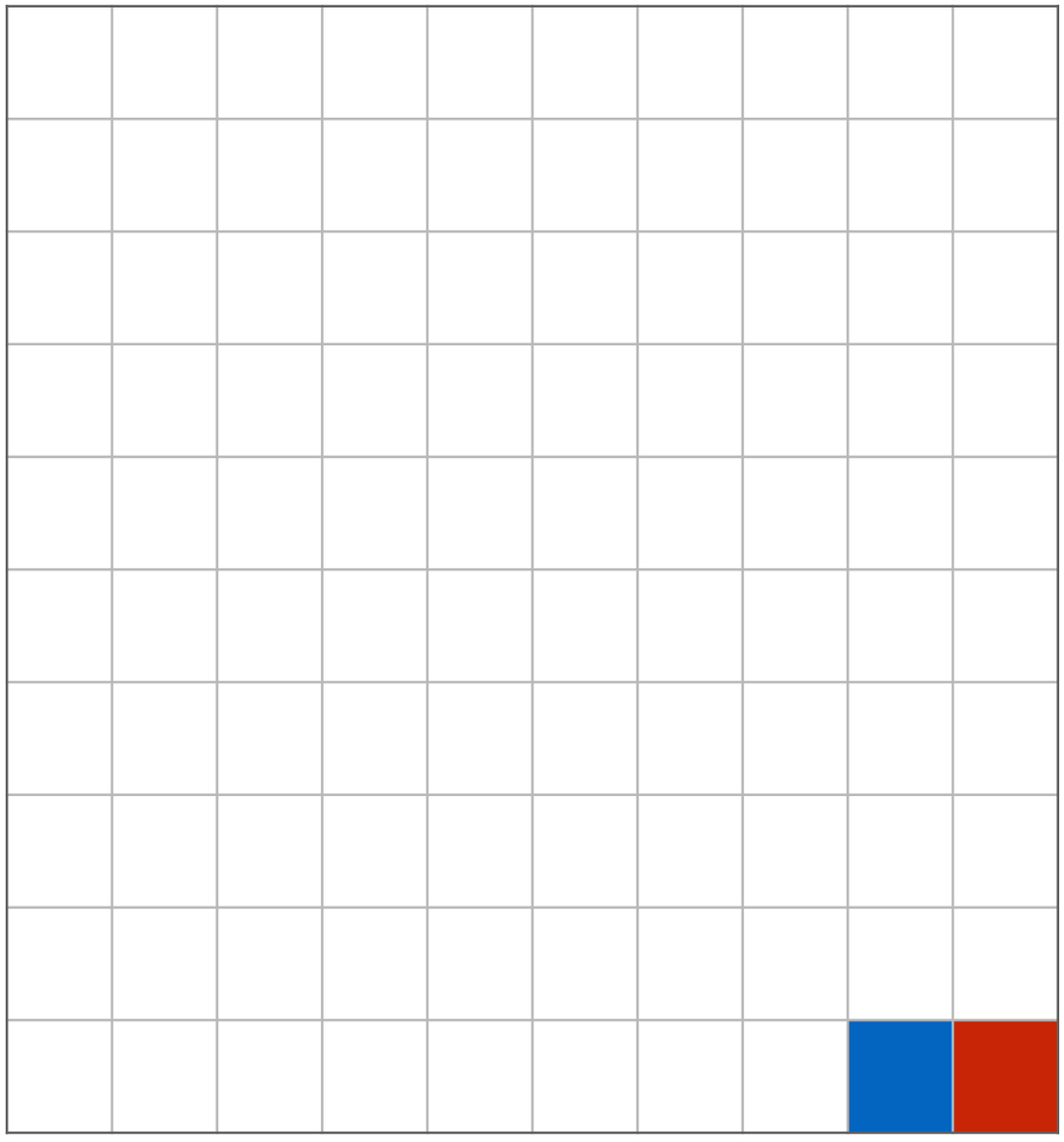
JOHNS HOPKINS UNIVERSITY

Model selection

- We have multiple ways of ranking the systems
  - Expected wins
  - Model of relative ability
  - TrueSkill
- Which is best?
  - Which one does the best job of making predictions?

Model selection

- Experimental setup
  - Split the complete data into 100 folds
  - For each fold
    - Build a model on the other 99 folds
    - Compute accuracy on the current fold
  - Report average accuracy across all folds



Dataset: 328k judgments
10 language pairs

Results

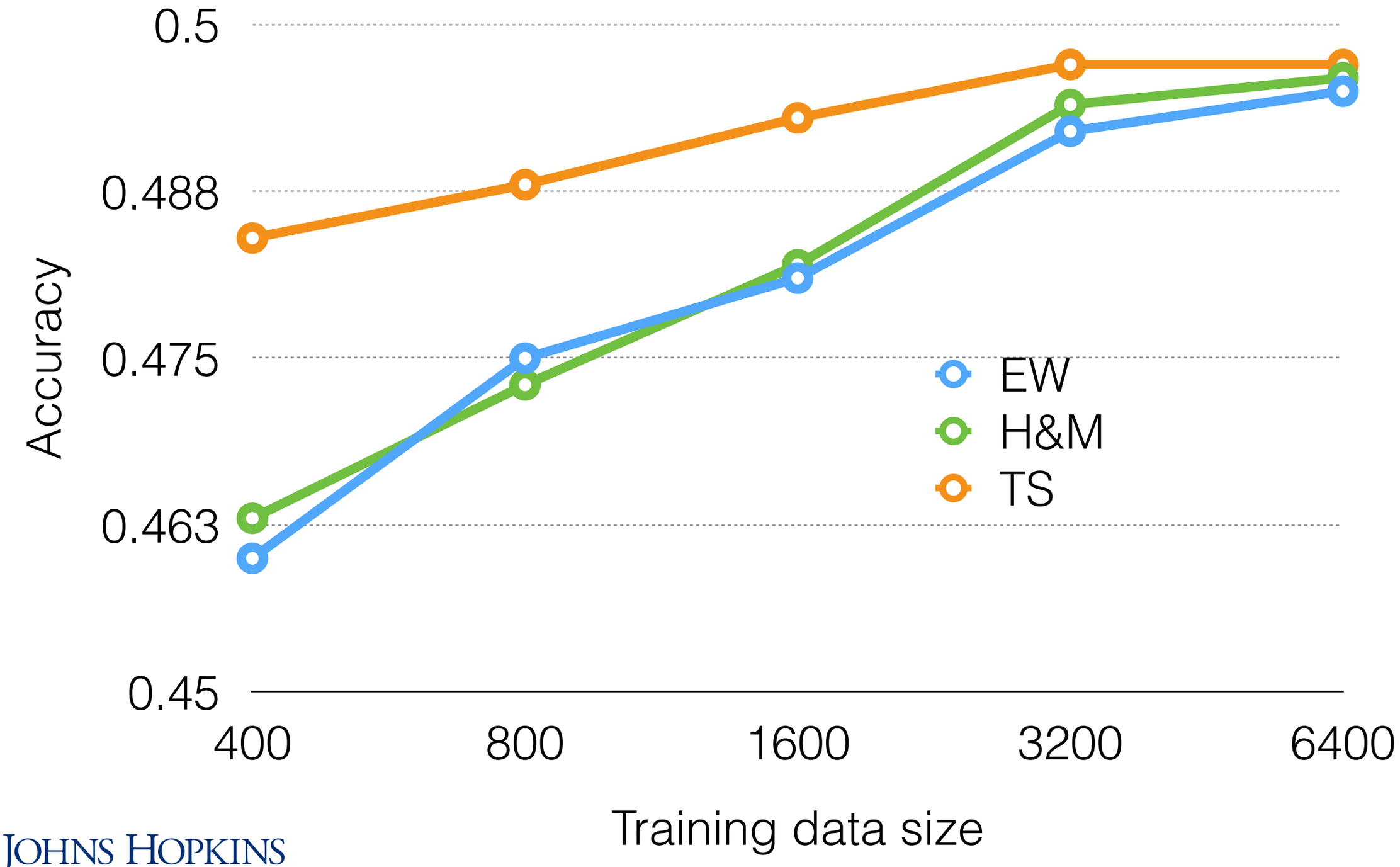
Task	EW	HM	TS	Oracle
Czech–English	40.4	**41.1**	**41.1**	41.2
English–Czech	45.3	45.6	**45.9**	46.8
French–English	49.0	**49.4**	49.3	50.3
English–French	44.6	44.4	**44.7**	46.0
German–English	43.5	**43.7**	**43.7**	45.2
English–German	47.3	**47.4**	47.2	48.2
Hindi–English	**62.5**	62.2	**62.5**	62.6
English–Hindi	53.3	**53.7**	53.5	55.7
Russian–English	47.6	**47.7**	**47.7**	50.6
English–Russian	**46.5**	46.1	46.4	48.2
MEAN	48.0	48.1	**48.2**	49.2

JOHNS HOPKINS UNIVERSITY

Analysis

- The different methods don't have that much of an effect (surprising?)
  - In fact, the ordering of systems was exactly the same for eight of the language pairs
- However, this hides the amount of data used

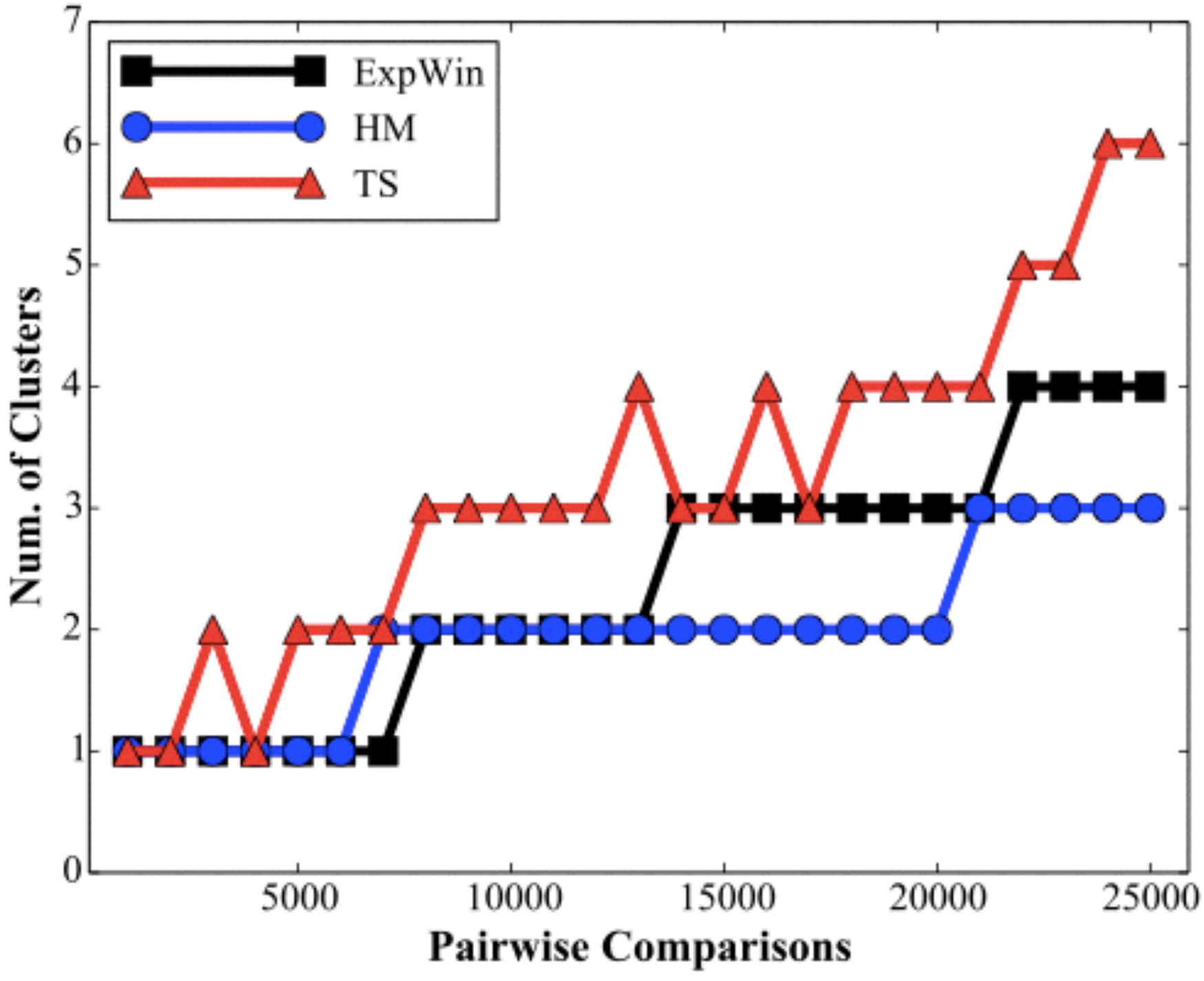
Data requirements



Analysis

- The different methods don't have that much of an effect (surprising?)
  - In fact, the ordering of systems was exactly the same for eight of the language pairs
- However, this hides the amount of data used
  - TrueSkill needs much less data
  - Also has much smaller variance (so we get tighter clusters)

Cluster counts



JOHNS HOPKINS UNIVERSITY

Summary

- There are many ways of producing the human ranking, from simple models to more elegant ones
- We use the model's ability to predict unseen data as a test of how good it is
  - There are many dimensions to goodness, including accuracy and data requirements
- Translation quality is inherently subjective and task-specific
  - Publishing clusters is a step towards capturing this