Construction and Analysis of Skill Modeling Frameworks in Esports

MS Project Defense – Spring 2020

Alexander J. Bisberg
alex.bisberg@utah.edu
University of Utah

Advisor
Rogelio E. Cardona-Rivera
rogelio@cs.utah.edu

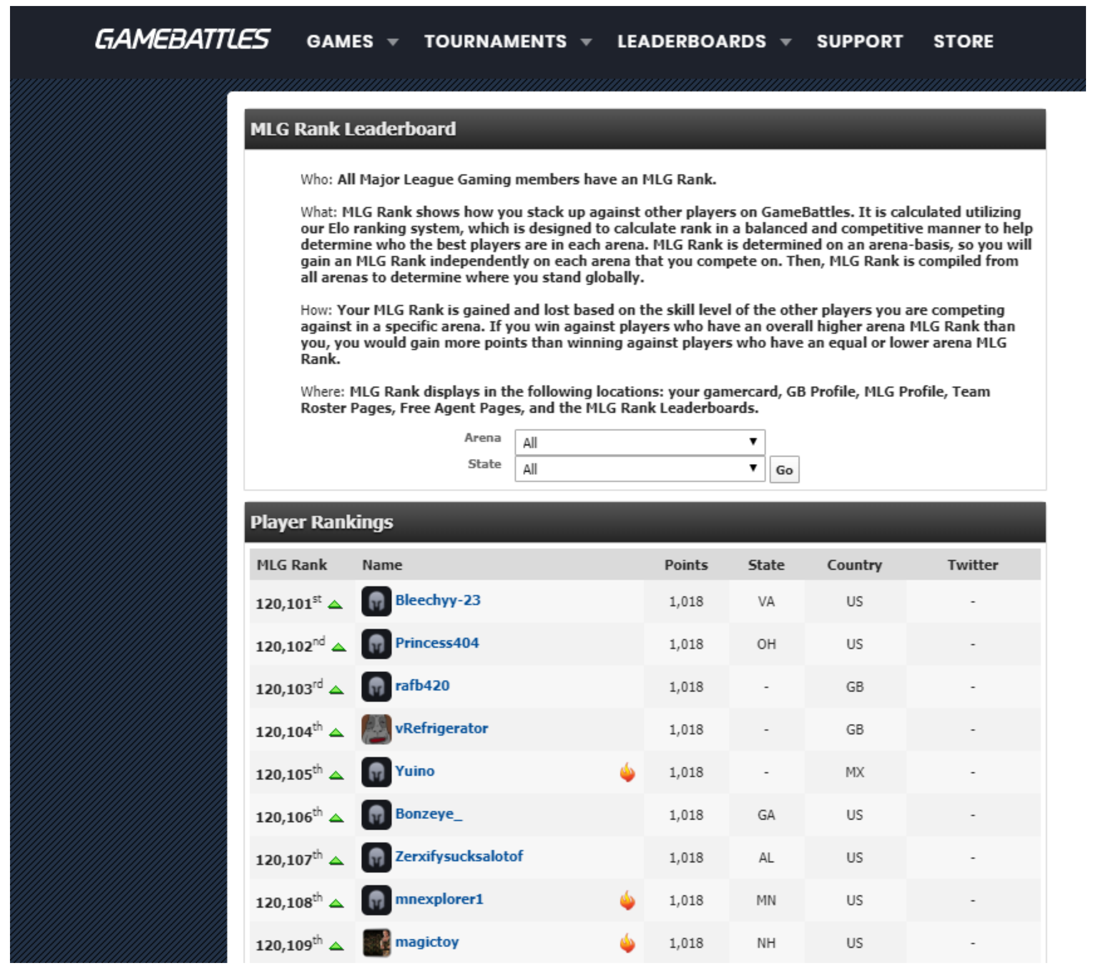
The Laboratory for Quantitative Experience Design

What are Esports and why do they matter?

Multiplayer competitive games played individually or in teams

What is skill and why does it matter?

The ability and capacity to execute activities that **overcome challenges** around ideas, things, or people



GAMEBATTLES GAMES TOURNAMENTS LEADERBOARDS SUPPORT STORE

MLG Rank Leaderboard

Who: All Major League Gaming members have an MLG Rank.

What: MLG Rank shows how you stack up against other players on GameBattles. It is calculated utilizing our Elo ranking system, which is designed to calculate rank in a balanced and competitive manner to help determine who the best players are in each arena. MLG Rank is determined on an arena-basis, so you will gain an MLG Rank independently on each arena that you compete on. Then, MLG Rank is compiled from all arenas to determine where you stand globally.

How: Your MLG Rank is gained and lost based on the skill level of the other players you are competing against in a specific arena. If you win against players who have an overall higher arena MLG Rank than you, you would gain more points than winning against players who have an equal or lower arena MLG Rank.

Where: MLG Rank displays in the following locations: your gamercard, GB Profile, MLG Profile, Team Roster Pages, Free Agent Pages, and the MLG Rank Leaderboards.

Arena: All
State: All
Go

Player Rankings

MLG Rank	Name	Points	State	Country	Twitter
120,101st	Bleechyy-23	1,018	VA	US	-
120,102nd	Princess404	1,018	OH	US	-
120,103rd	rafb420	1,018	-	GB	-
120,104th	vRefrigerator	1,018	-	GB	-
120,105th	Yuino	1,018	-	MX	-
120,106th	Bonzeye_	1,018	GA	US	-
120,107th	Zerxifysucksalotof	1,018	AL	US	-
120,108th	mnexplorer1	1,018	MN	US	-
120,109th	magictoy	1,018	NH	US	-

Ranking

Matchmaking

...

What is the problem?

Skill modeling frameworks have been developed for *traditional sports* in an *ad-hoc, unsystematic* way.

How do we measure skill?

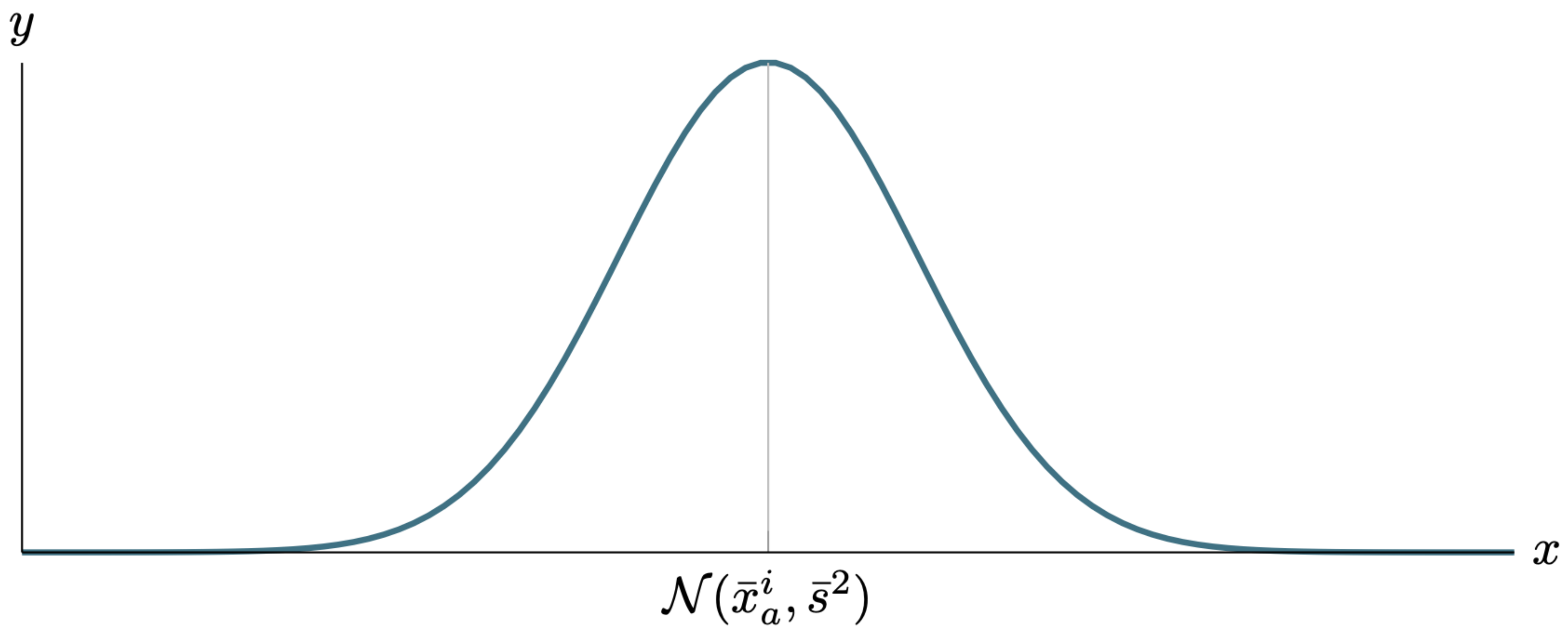
- Arpad Elo - Hungarian chess player and mathematician

- Purpose: Wanted a system he could use to easily compute skill and win probability (by hand) between tournaments

A. Elo, *The Rating of Chess Players, Past and Present*. Ishi Press International, 1978.

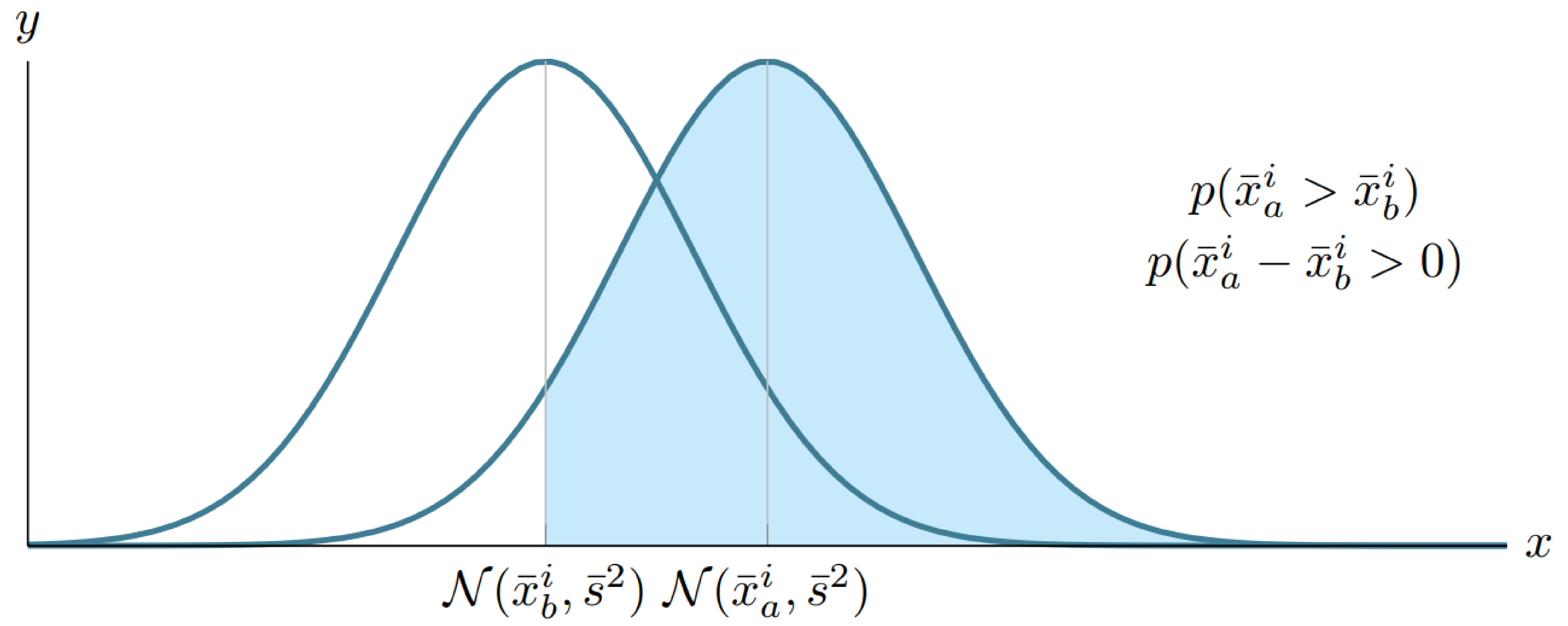
Elo skill representation

- Skill is represented by a Gaussian with some **mean** and **constant variance**



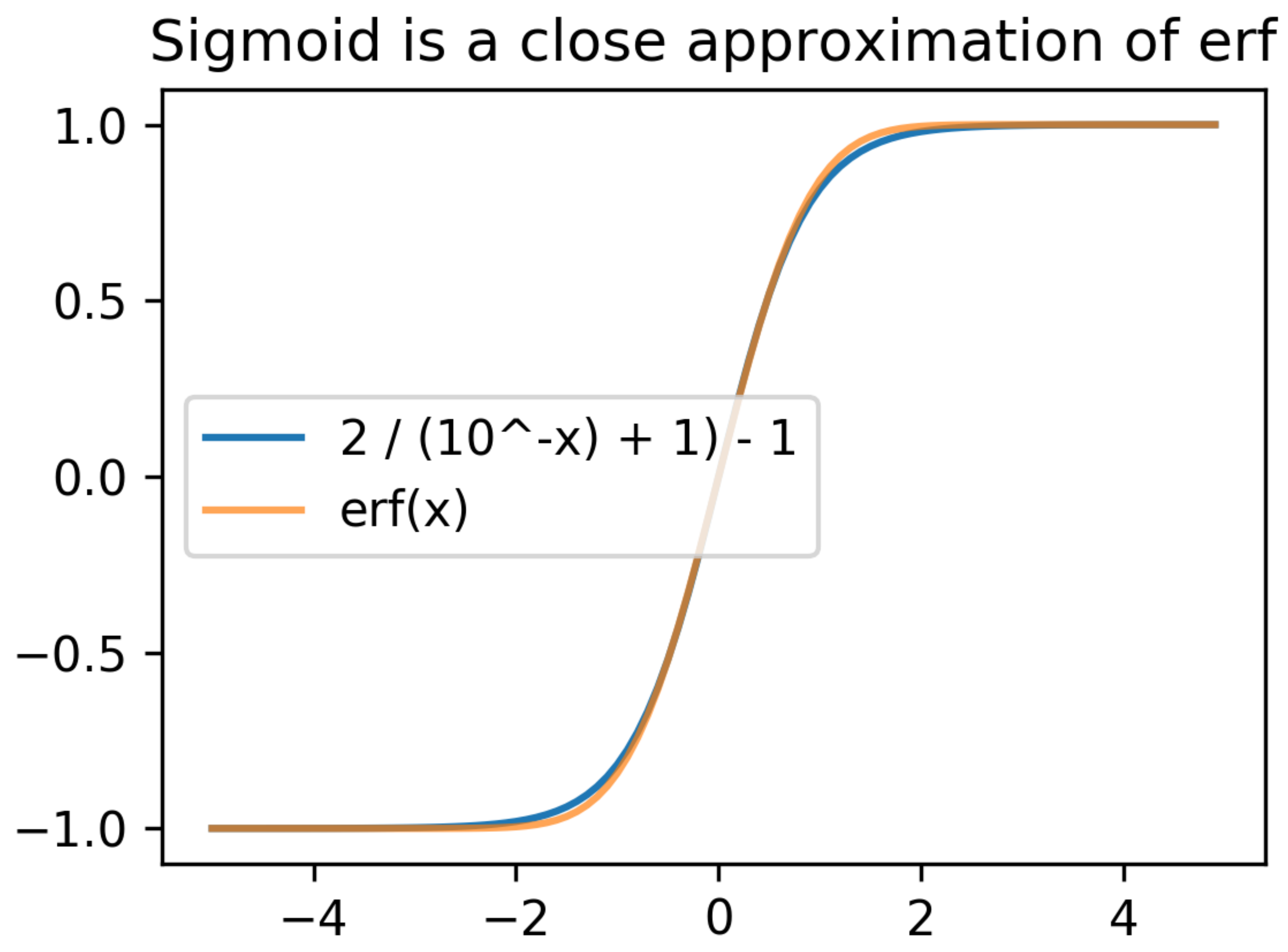
Elo skill difference

- The probability that one skill is greater than the other is the *difference of their distributions*
  - What does this entail?



Elo win probability

- A team's likelihood to win is quantified by the *difference in skill from their opponent*
- Calculate the probability by integrating (evaluate the CDF)
  - Can use approximation for ease of understanding



$$erf(\bar{x}_a^i - \bar{x}_b^i) \longrightarrow \Pr\left(W_a^i\right) = \frac{1}{1 + 10^{(\bar{x}_a^i - \bar{x}_b^i)/w90}}$$

Elo score update

- What happens after we observe the result of a game?
- We make an update to the score based on our expectation of the outcome, scaled by *K*

$$\underbrace{\bar{x}_a^{i+1}}_{\text{Skill at next game}} = \underbrace{\bar{x}_a^{i}}_{\text{Skill at last game}} + \overbrace{K}^{\text{Match importance}}(\underbrace{S_a^i}_{\text{Score of last game}} - \underbrace{\mathrm{E}[S]_a^i}_{\text{Expected score of last game}})$$

What is the problem?

Skill modeling frameworks have been developed for traditional sports in *an ad-hoc, unsystematic* way.

- **Even among Elo based models, there is not a unified understanding of how to build them**

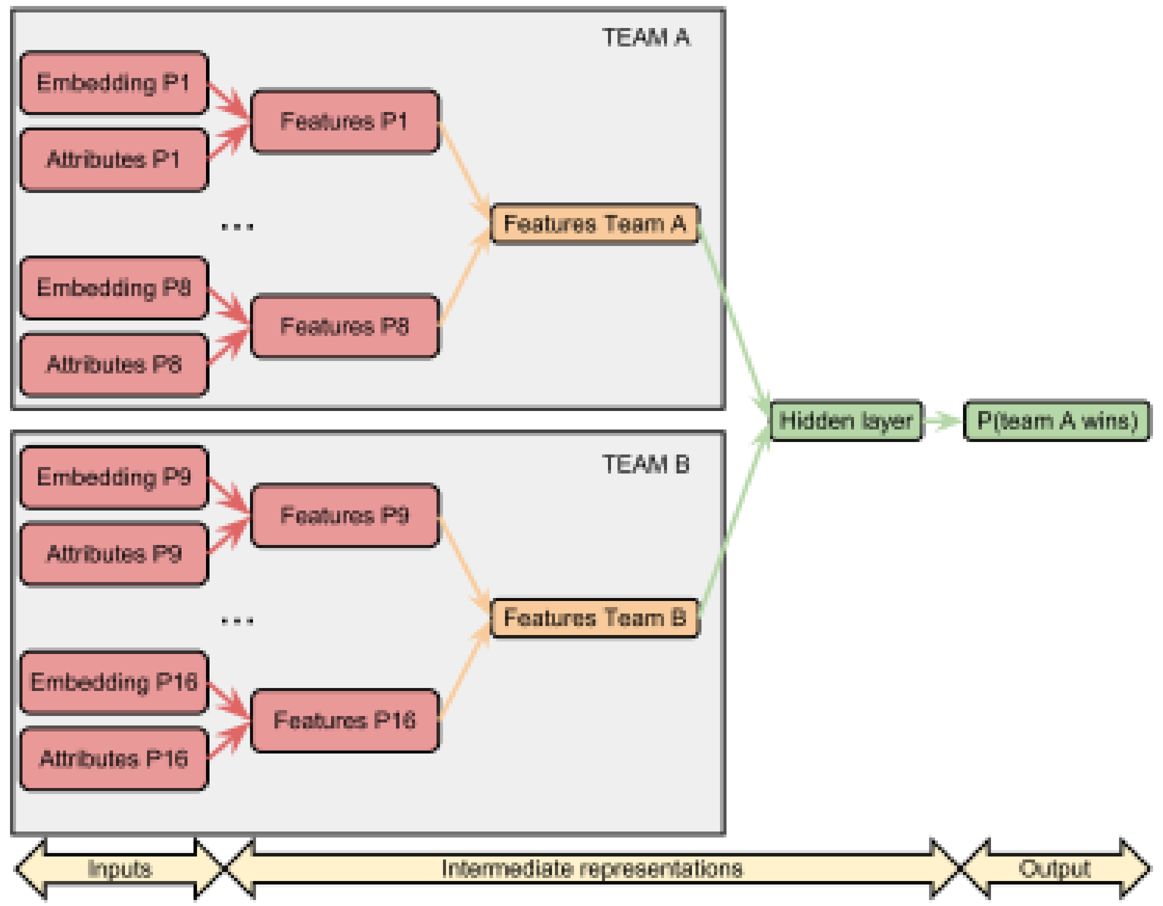
  S. Lacy "Implementing an Elo rating system for European Football," 2018.

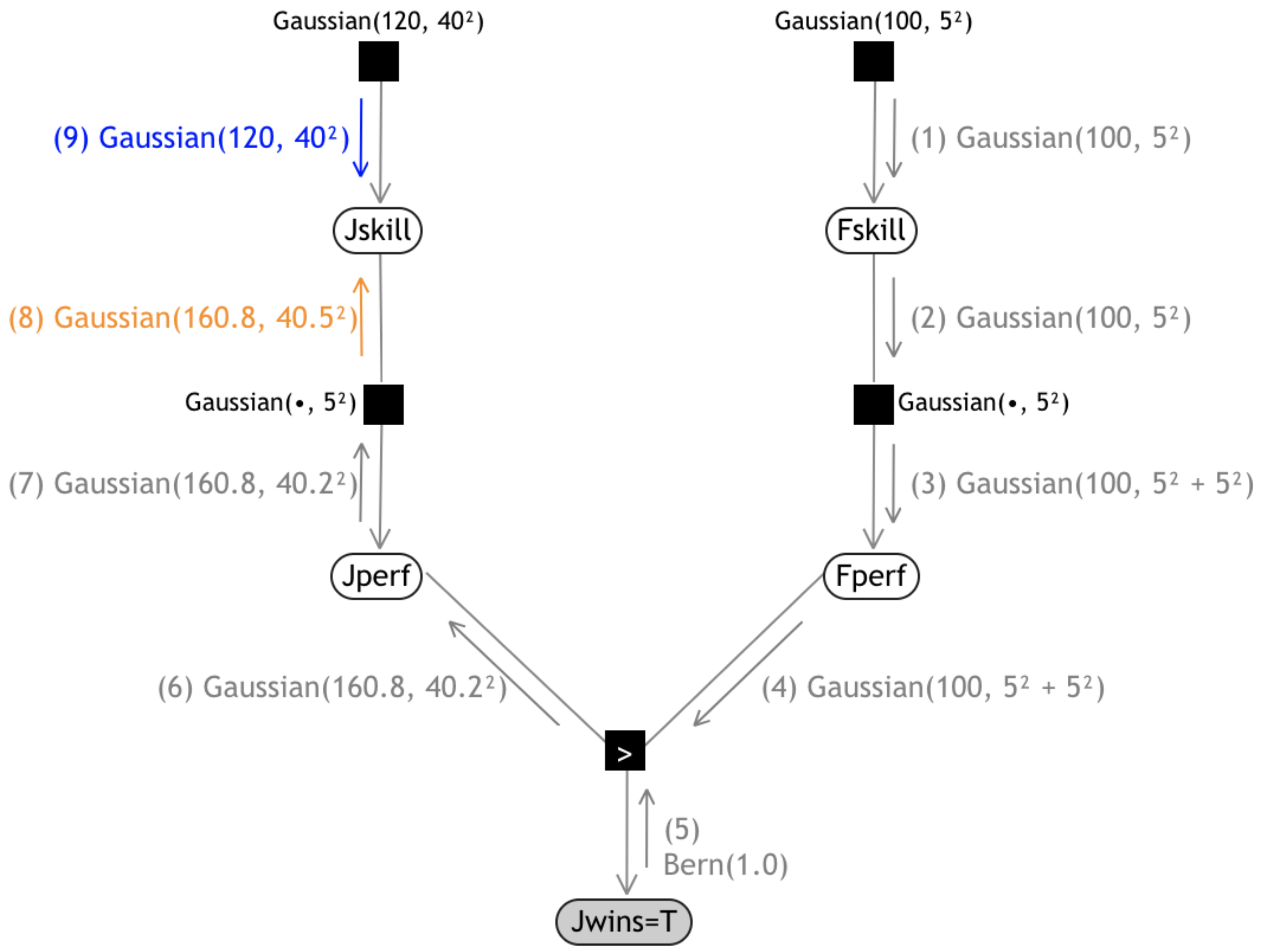
  J. Boice "How our MLB predictions Work," *FiveThirtyEight.* 2018. fivethirtyeight.com

  S. A. Kovalchik, "Searching for the GOAT of tennis win prediction," *J. Quant. Anal. Sport.*, vol. 12, no. 3, pp. 127–138, 2016.

- But people have tried to make other models…

What other skill models are out there?





Glicko

- Considers rating reliability
- Play frequency

Neural Nets

- Creates player profiles (embeddings)
- Takes into account ping, player experience

TrueSkill

- Probabilistic (graphical) modeling approach
- Models skill as distribution

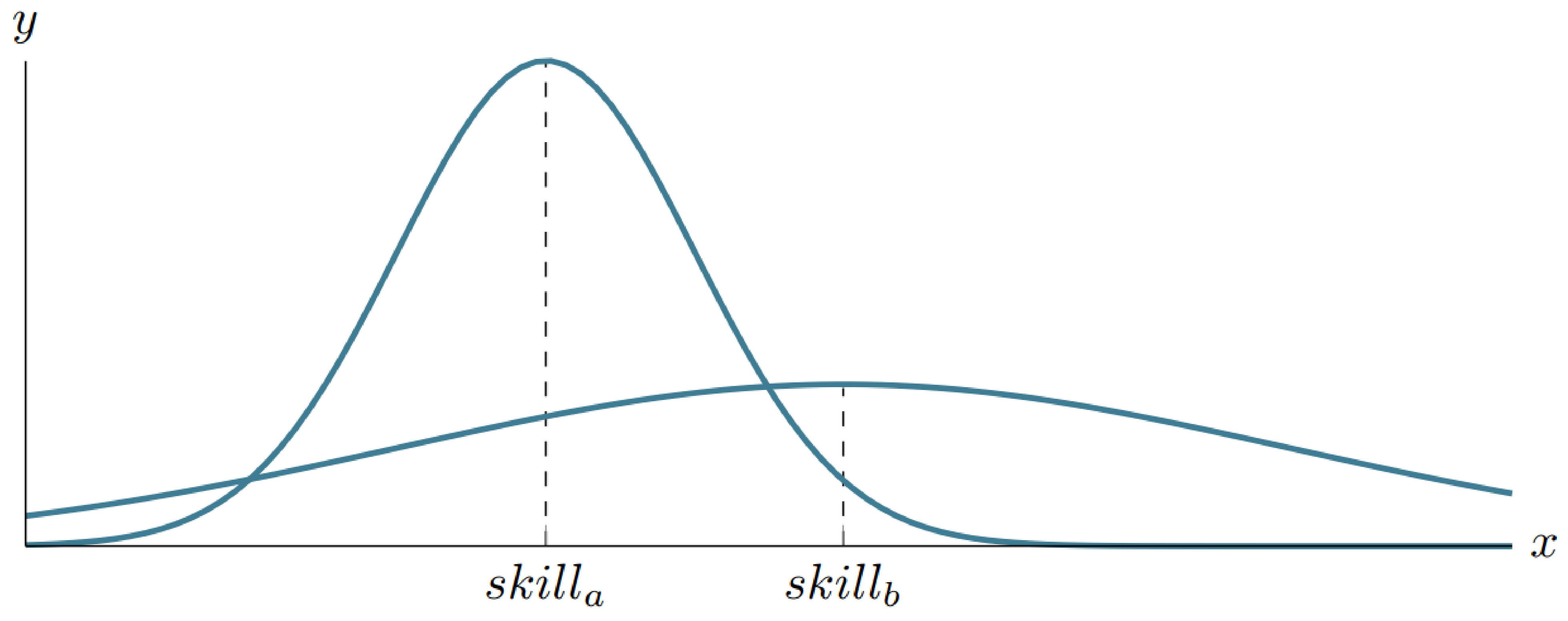
M. E. Glickman, “The Glicko system.” 1999.

O. Delalleau, E. Contal, E. Thibodeau-Laufer, R. C. Ferrari, Y. Bengio, and F. Zhang, “Beyond Skill Rating: Advanced Matchmaking in Ghost Recon Online,” in *IEEE Transactions on Computational Intelligence and AI in Games*, 2012.

P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel, “TrueSkill Through Time: Revisiting the History of Chess,” in *Advances in Neural Processing Systems (NIPS)*, 2007.

TrueSkill skill representation

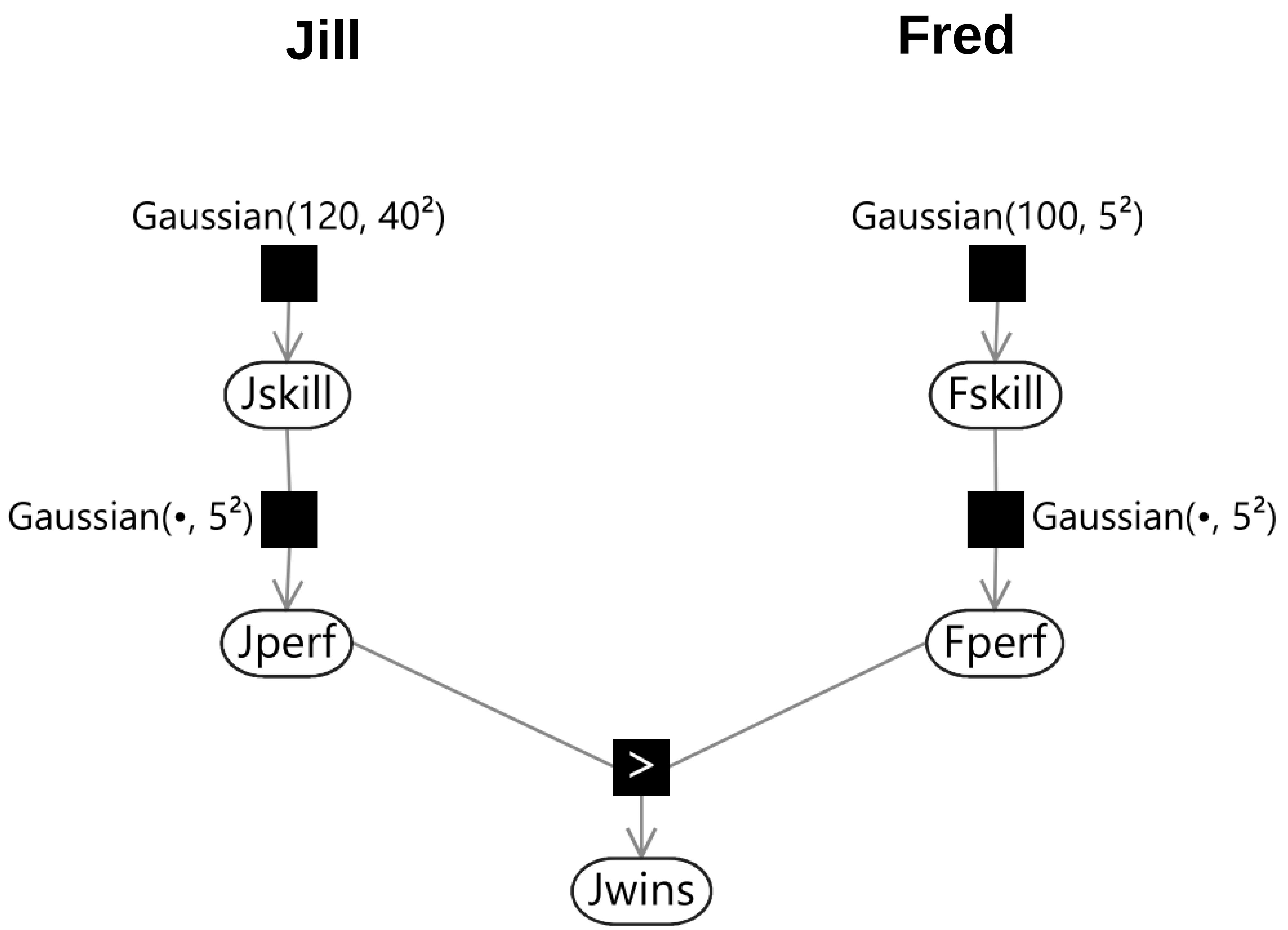
- Skill is represented by a Gaussian with some **mean and variance**
  - Key difference – variance is now a model parameter!
- Goal: calculate the **posterior skill distribution** after a game
- Purpose: matchmaking



How do we represent a game?

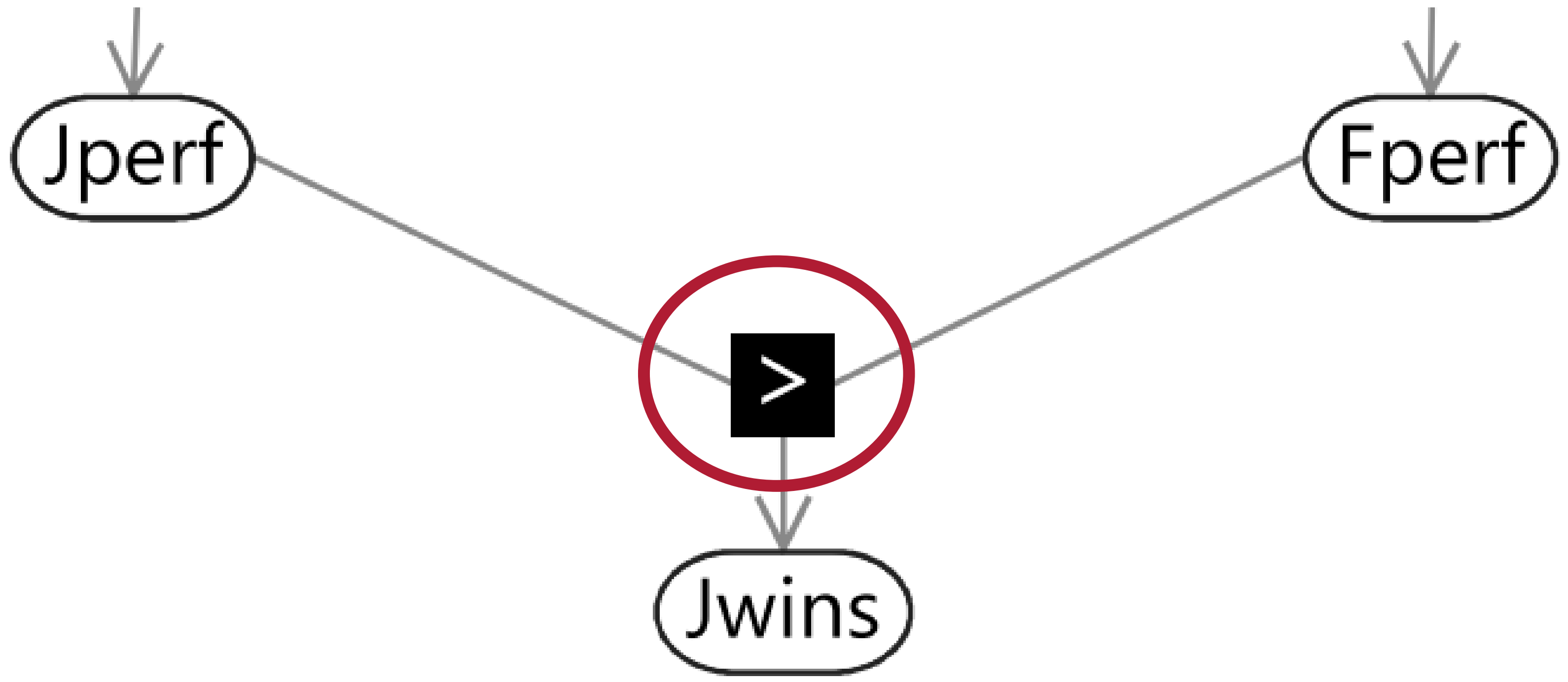
- Factor Graphs
  - Circle – variable node
  - Square – factor node

J. Winn and C. M. Bishoph, *Model-Based Machine Learning*. Microsoft, 2013.



How do we represent a game?

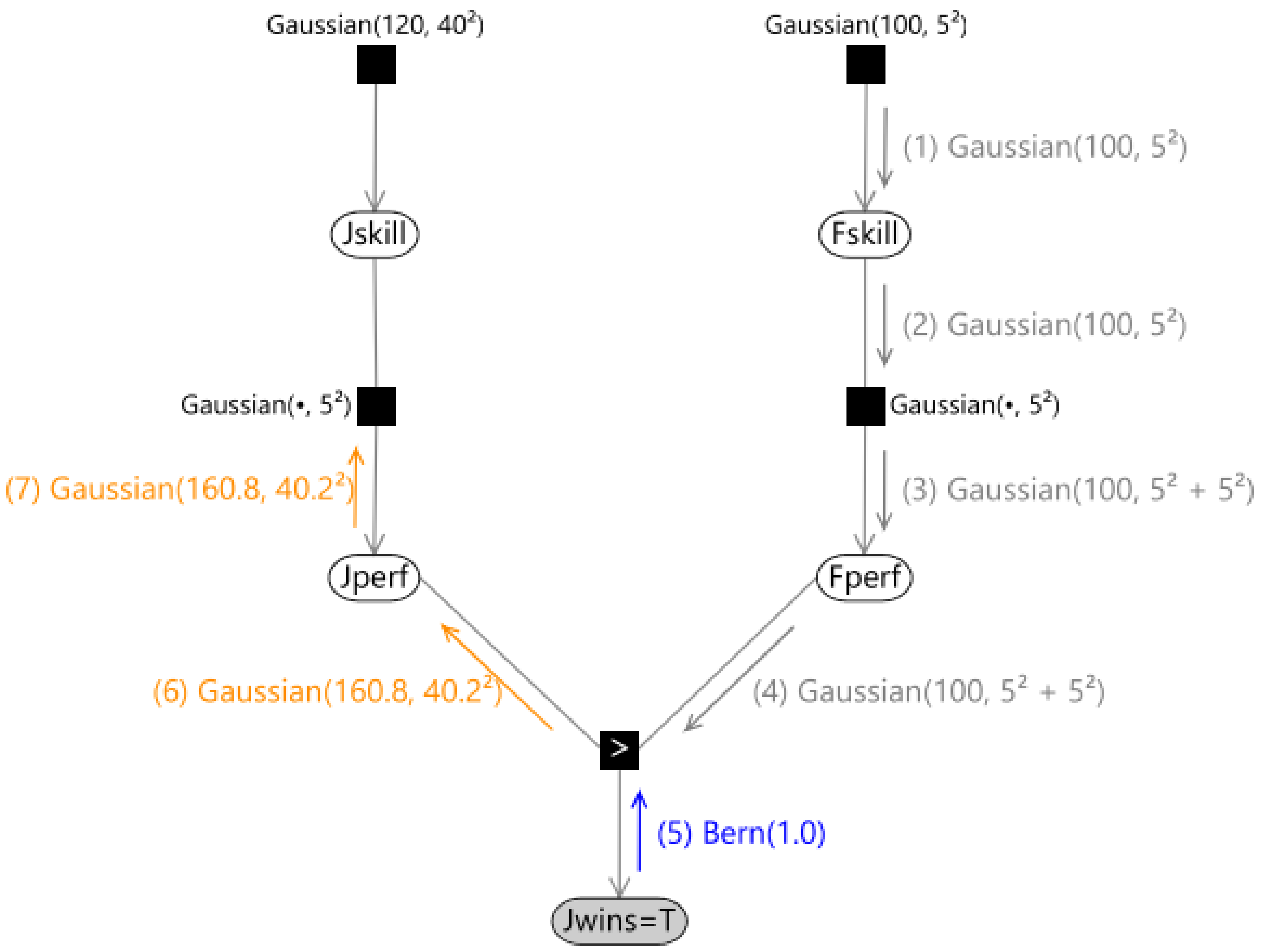
- Probabilistically, what is winning or losing?
  - One score is greater than the other
  - Similar to Elo!
  - *p(Jperf > Fperf)* – this is now a distribution



J. Winn and C. M. Bishoph, *Model-Based Machine Learning*. Microsoft, 2013.

How is the skill updated?

- Skill is updated with inference
  - Arrows pass the distributions from nodes to factors
  - Tom Minka’s expectation propagation



T Minka, “Expectation Propagation for Approximate Bayesian Inference” 2013.

What is the problem?

Skill modeling frameworks have been developed for *traditional sports* in an *ad-hoc, unsystematic* way.

- Even among Elo based models, there is not a unified understanding of how to build them
- **Now that there are many models in the ecosystem, how do we choose which to use?**

Who has compared skill models?

GOAT of Tennis Ratings
Kovalchik et al.

- Compared conventional models, Elo, and BCM
- Accuracy, calibration and log loss

S. A. Kovalchik, “Searching for the GOAT of tennis win prediction,” *J. Quant. Anal. Sport.*, vol. 12, no. 3, pp. 127–138, 2016.

Ranking rankings
Barrow et al.

- Compared conventional win percent, RPI, page rank, Elo
- Most rankings have similar predictive power

D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting, “Ranking rankings: An empirical comparison of the predictive power of sports ranking methods,” *J. Quant. Anal. Sport.*, vol. 9, no. 2, pp. 187–202, 2013

Comparison of Rating Systems
Glickman et al.

- Compared Elo and Glicko
- Very comparable in log loss and misclassification rate

M. E. Glickman, J. Hennessy, and A. Bent, “A comparison of rating systems for competitive women’s beach volleyball,” *Stat. Appl.*, vol. 30, no. 2, pp. 233–254, 2018.

My Claim

Skill models and skill modeling comparisons are done in an ad-hoc, unsystematic way.

I have developed a way to (1) systematically build Elo models and (2) systematically analyze skill models.

What I did

- (1) Systematically build Elo models
  - SCOPE – **S**elective **C**ross-validation **O**ver **P**arameters for **E**lo
- (2) Systematically compare skill models
  - FRAGEM-S – **FR**amework for **A**nalysis of **G**ame and **E**sports **M**odeling - **S**kill
  - Example: SCOPE vs. TrueSkill
- Future Work
  - FRAGEM-R (Roles) and more

What is SCOPE and why is it different?

- Selective Cross-validation Over Parameters for Elo
  - Problem addressed: Inconsistent, ad-hoc Elo models
- SCOPE model parameters
  1. Score initialization
  2. K baseline and updates
  3. Margin of victory
  4. Change in skill over time

Parameter	Range
Base K	1 – 50
MoV	4 functions
K Scale	0.1 – 0.9
Cutoff	1600-1750
w90	100-500
Regression	0.1 – 0.3

1: Score Initialization

- How do we know what a team's starting skill is?
  - Unsolved problem by Elo
- **SCOPE**: Use data from a pervious season to inform initialization

2: *K* baseline and updates

- Should *K* be the same for all games? All teams?
  - Some games are more important
- **SCOPE**: More certain about highly skilled teams
  - Decrease *K* above a certain point

J. Boice "How our MLB predictions Work," *FiveThirtyEight.* 2018. fivethirtyeight.com

3: Margin of victory

- **SCOPE**: Margin of victory (MoV) captures meaningful data that can impact our perception of skill
    - Other modelers have had success with this idea
- How much should we scale based on MoV?
    - Linear? Exponential?

S. Lacy “Implementing an Elo rating system for European Football,” 2018.

4: Change in skill over time

- **SCOPE**: Teams regress to the mean over time
- The framework is flexible enough to continue adding assumptions

How do we choose these parameters?

- Cross-validation
  - Common technique - grid search over hyper parameters
  - Could use other ways, i.e. optimization
- Time series data
  - Day-forward chaining

How do we measure model performance?

- Accuracy
  - Correct predictions are when we predict a team with over 50% chance to win actually wins

$$correct = \begin{cases} 1 & \text{if } \Pr(W) > 0.5 \text{ and } S = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$accuracy = \frac{\sum^{n} correct}{n}$$

- Log loss
  - Penalizes confident incorrect predictions

$$log - loss = \frac{-\sum^{n}[S_a \log \Pr(W_a) + S_b \log \Pr(W_b)]}{n}$$

- Calibration
  - Rewards confident correct predictions

$$calibration = \frac{\sum^{n} \max[\Pr(W_a), \Pr(W_b)]}{\sum^{n} \operatorname{argmax}_{S}[\Pr(W_a), \Pr(W_b)]}$$

How does SCOPE work on real data?

- Data
  - Call of Duty World League
- Using different assessment metrics changes our model parameters
- The most accurate model is somewhere in between

Team	Player	Series id	Match id	Won	Kills
Evil Geniuses	Freddy	1123	321	1	32

…

⋮

Parameters			Metrics		
Cutoff	K Scale	w90	Accuracy	Calibration	Log Loss
1650	0.10	200	$\boxed{\mathbf{.684} \pm .11}$	$1.01 \pm .15$	$.374 \pm .046$
1650	0.10	200	$.684 \pm .11$	$\mathbf{1.01} \pm .15$	$.374 \pm .046$
1650	0.75	100	$.662 \pm .12$	$1.17 \pm .17$	$\mathbf{.253} \pm .048$

A. J. Bisberg and R. E. Cardona-Rivera, "SCOPE : Selective Cross-validation Over Parameters for Elo," in *AIIDE*, 2019.

Takeaway

- Using SCOPE to build an understandable Elo model to represent team skill in esports produces accurate, easily understandable results
    - Comparable accuracy to TrueSkill

How does it compare to other models?

The skill model comparison ecosystem is fragmented for sports and *non-existent* for esports.

The use case of the model significantly affects model selection and evaluation

Considerations for Skill Modeling

- **Model performance metrics**
- Initialization
- Primary application
- Integrating external data
- Data representation
- Explicit player performance
- Team modeling

A. J. Bisberg, K. N. McKay-Bishop and R. E. Cardona-Rivera. "A Comparative Framework and Analysis of Skill Modeling in Esports," Submitted to IEEE Conference on Games, 2020

Model performance metrics

- Win prediction
  - Accuracy
  - Calibration
  - Log loss
- **Convergence**
  - Important for matchmaking
  - Measure with relative squared error

$$\mathrm{RSE} = \frac{\sum_{j=1}^{n} \left(P_j - T_j\right)^2}{\sum_{j=1}^{n} \left(T_j - \overline{T}_j\right)^2}$$

Considerations for Skill Modeling

- Model performance metrics
- **Initialization**
- Primary application
- Integrating external data
- Data representation
- Explicit player performance
- Team modeling

Do you have historical data?

Considerations for Skill Modeling

- Model performance metrics
- Initialization
- **Primary application**
- Integrating external data
- Data representation
- Explicit player performance
- Team modeling

Is this being used for matchmaking, win prediction or something else?

Considerations for Skill Modeling

- Model performance metrics
- Initialization
- Primary application
- **Integrating external data**
- Data representation
- Explicit player performance
- Team modeling

Do you care about more than just win/loss data?

Considerations for Skill Modeling

- Model performance metrics
- Initialization
- Primary application
- Integrating external data
- **Data representation**
- Explicit player performance
- Team modeling

Can you describe all of your data as a distribution?

Considerations for Skill Modeling

- Model performance metrics
- Initialization
- Primary application
- Integrating external data
- Data representation
- **Explicit player performance**
- Team modeling

Do you care about day-of performance?

Considerations for Skill Modeling

- Model performance metrics
- Initialization
- Primary application
- Integrating external data
- Data representation
- Explicit player performance
- **Team modeling**

Do you want to model your team as a collection of individual players?

Experimental Setup

- Data
  - Call of Duty World League
- Train for each of the 4 metrics
  - Accuracy
  - Calibration
  - Log Loss
  - Convergence

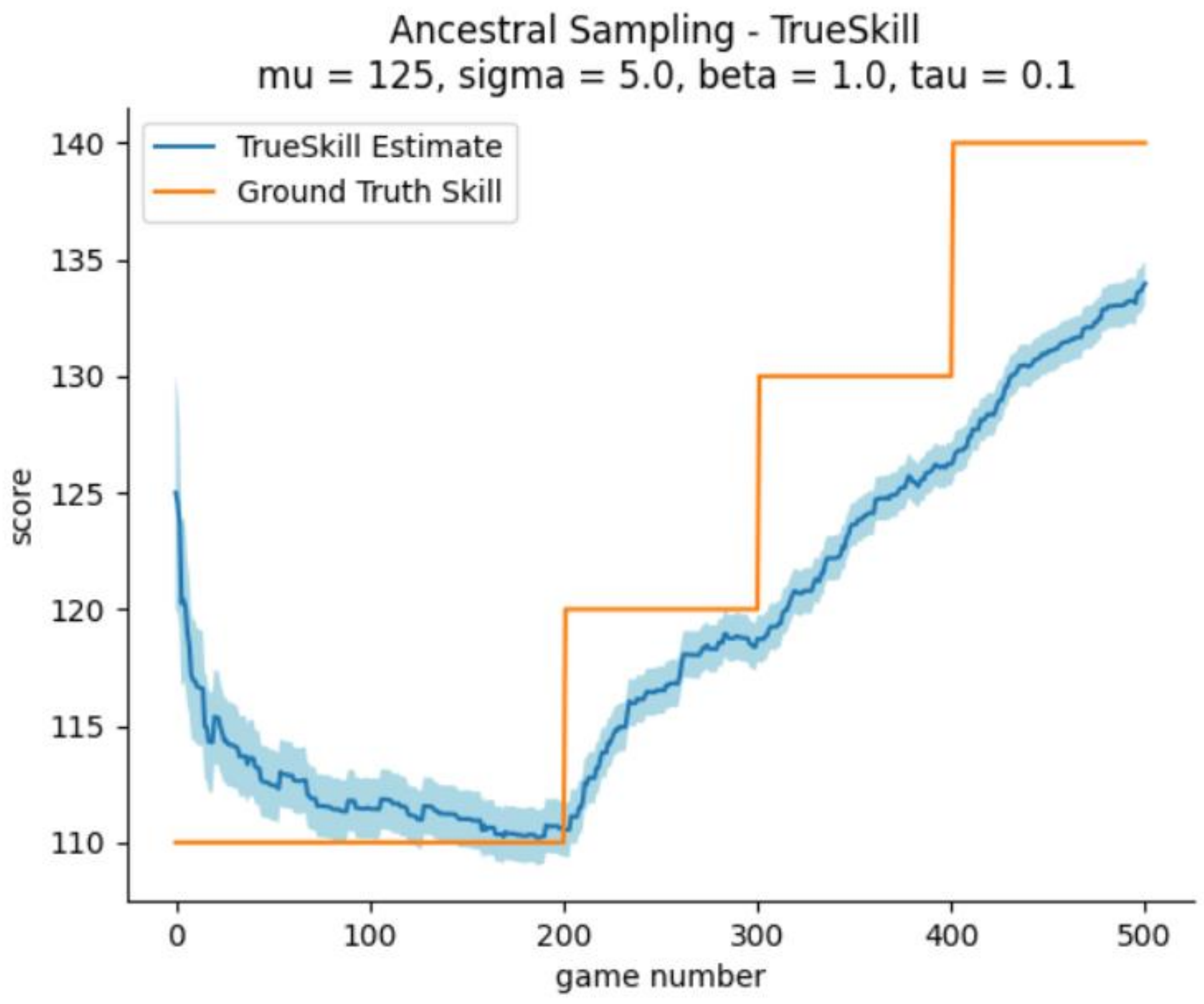
SCOPE often works very well

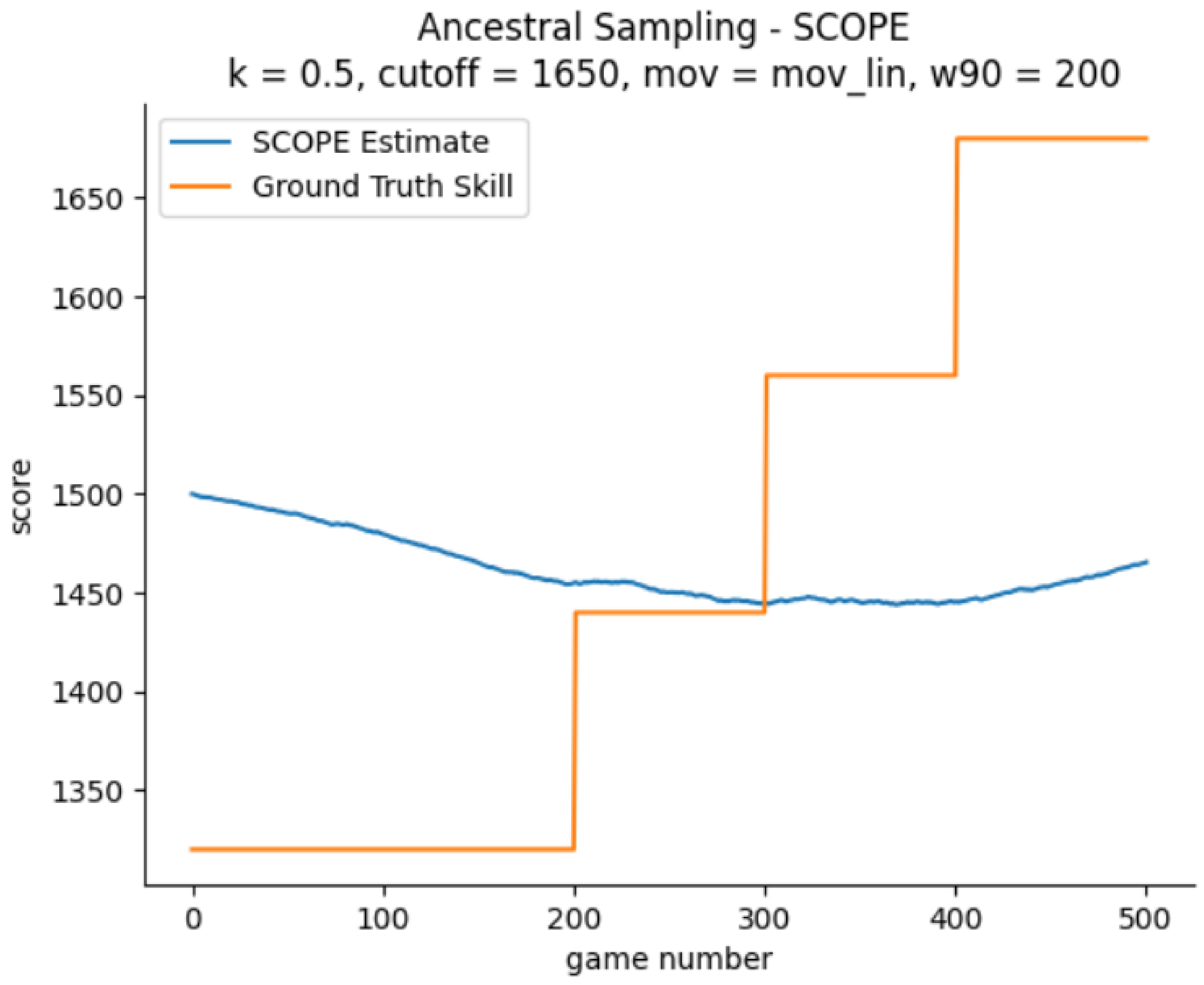
Best performing models highlighted in **bold**

Metric	Model			
	SCOPE	TS_Team	TS_Player	TS_MaxPlayer
Accuracy	$\mathbf{.684 \pm .11}$ †	$.646 \pm .064$	$.670 \pm .092$	$.670 \pm .097$
Calibration	$\mathbf{1.01 \pm .15}$ †	$1.08 \pm .099$	$.986 \pm .047$	$.900 \pm .074$
Log Loss	$\mathbf{.094 \pm .046}$ †	$.231 \pm .037$	$.232 \pm .037$	$.356 \pm .037$
RSE	$1.18 \pm .020$	$\mathbf{.309 \pm .047}$	-	-

Goes against common wisdom that TrueSkill is always better

Until it doesn’t





TrueSkill has better convergence properties

Discussion

- TrueSkill and SCOPE are used interchangeably when they shouldn't be
- This may seem obvious but…
  - SCOPE is designed for win prediction, and better at win prediction
  - TrueSkill is designed for matchmaking, ad better at matchmaking

Future Work

- How do roles impact player skill?
- Can we compare roles between esports?
  - FRAGEM-R
- Can we generalize skill, roles and shared information framework to other domains?

Recap and wrap up

Skill models and skill modeling comparisons are done in an *ad-hoc, unsystematic* way.

I have developed a way to
(1) systematically build Elo models

- **Using SCOPE**

(2) systematically analyze skill models.

- **Using FRAGEM-S**

Through experimentation, I have shown TrueSkill and SCOPE should not be used interchangeably

Thanks! Questions?

- Up next for me
  - Summer Internship at Activision
  - PhD in the fall

Alexander J. Bisberg
alex.bisberg@utah.edu
University of Utah

Advisor
Rogelio E. Cardona-Rivera
rogelio@cs.utah.edu

The Laboratory for Quantitative Experience Design

Bibliography

[1] A. Elo, *The Rating of Chess Players, Past and Present*. Ishi Press International, 1978.
[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[3] M. E. Glickman, "The Glicko system." 1999.
[4] R. Hebrich, T. Minka, and T. Graepel, "TrueSkill: A Bayesian Skill Rating System," in *Neural Information Processing Systems*, 2006.
[5] P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel, "TrueSkill Through Time: Revisiting the History of Chess," in *Advances in Neural Processing Systems (NIPS)*, 2007.
[6] T. Minka, M. Research, R. Cleven, and Y. Zaykov, "TrueSkill 2: An improved Bayesian skill rating system," 2018.
[7] O. Delalleau, E. Contal, E. Thibodeau-Laufer, R. C. Ferrari, Y. Bengio, and F. Zhang, "Beyond Skill Rating: Advanced Matchmaking in Ghost Recon Online," in *IEEE Transactions on Computational Intelligence and AI in Games*, 2012.
[8] D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting, "Ranking rankings: An empirical comparison of the predictive power of sports ranking methods," *J. Quant. Anal. Sport.*, vol. 9, no. 2, pp. 187–202, 2013
[9] S. A. Kovalchik, "Searching for the GOAT of tennis win prediction," *J. Quant. Anal. Sport.*, vol. 12, no. 3, pp. 127–138, 2016
[10] M. E. Glickman, J. Hennessy, and A. Bent, "A comparison of rating systems for competitive women's beach volleyball," *Stat. Appl.*, vol. 30, no. 2, pp. 233–254, 2018.
[11] J. Winn and C. M. Bishoph, *Model-Based Machine Learning*. Microsoft, 2013.
[12] A. J. Bisberg and R. E. Cardona-Rivera, "SCOPE : Selective Cross-validation Over Parameters for Elo," in *AIIDE*, 2019.
[13] A. J. Bisberg, K. N. McKay-Bishop and R. E. Cardona-Rivera. "A Comparative Framework and Analysis of Skill Modeling in Esports," Submitted to IEEE Conference on Games, 2020