

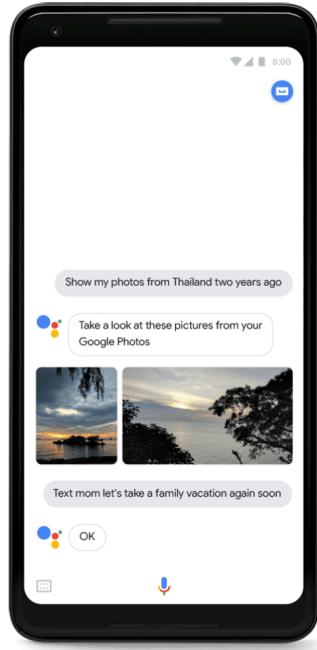
# Evaluating Conversational Agents

**João Sedoc**

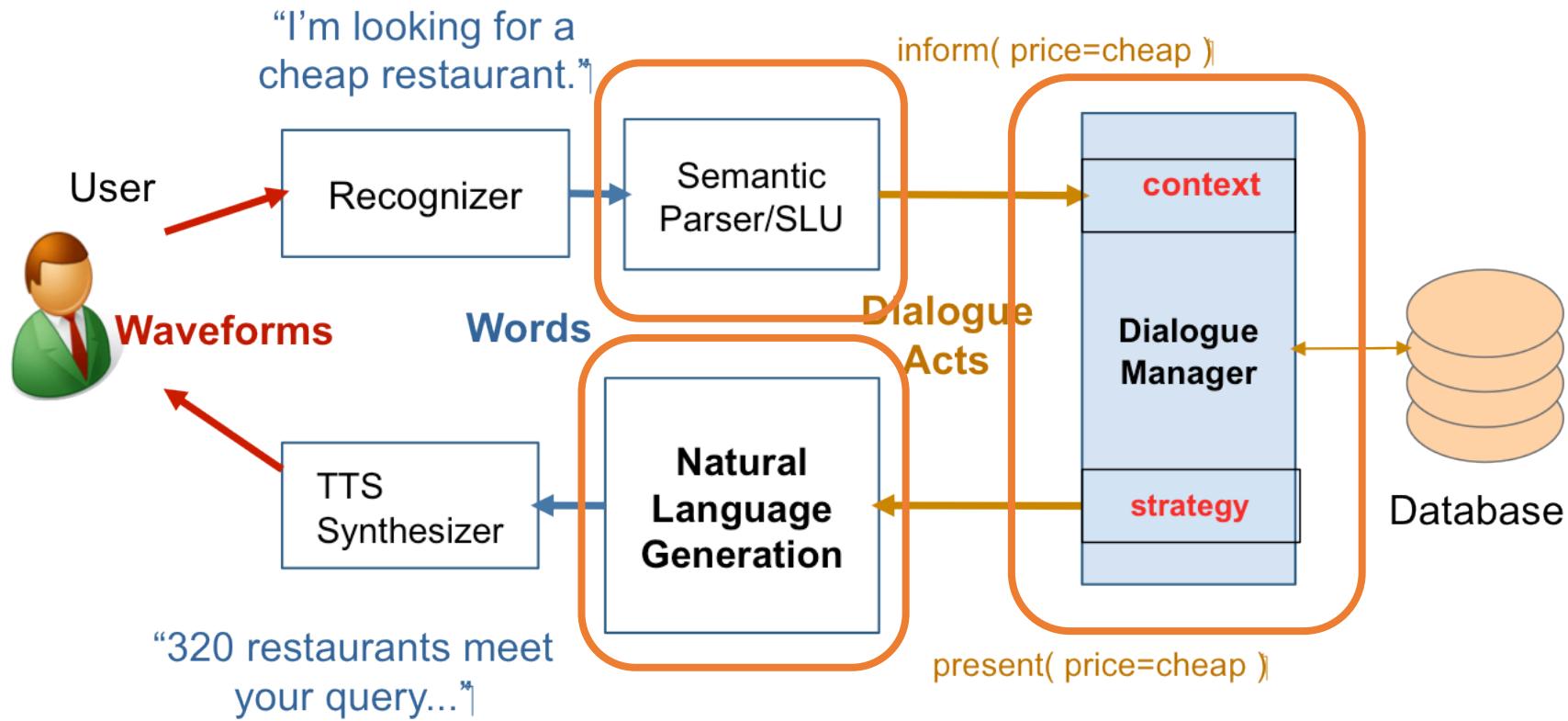
`jsedoc@jhu.edu`  
Johns Hopkins  
Computer Science

June 25<sup>th</sup>, 2019

# Chatbots are Ubiquitous: Personal Agents, Games, Education & Business



# Spoke Dialog System Architecture

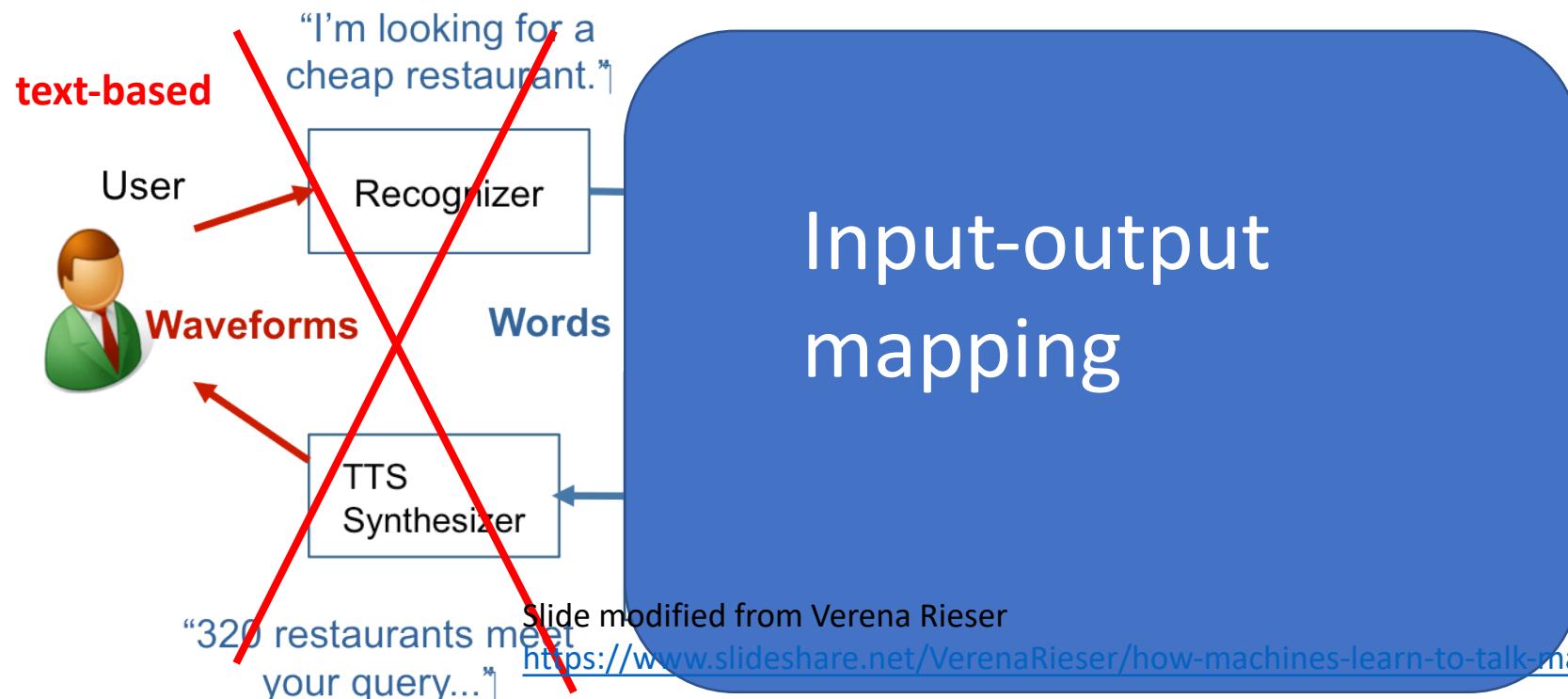


Slide modified from Verena Rieser

<https://www.slideshare.net/VerenaRieser/how-machines-learn-to-talk-machine-learning-for-conversational-ai>

# Response Generation Systems

- **End-to-end** deep learning systems
- Learn from large corpora of “raw” dialog data (e.g. OpenSubtitles)
- No semantic or pragmatic annotation required
- Mainly open-domain, non-task oriented systems



# Two Types of Systems

1. Chatbots
2. Task-based
  - SIRI, interfaces to cars, robots, ...
  - Booking flights, restaurants, or question answering

# Chatbot Architectures

## Rule-based

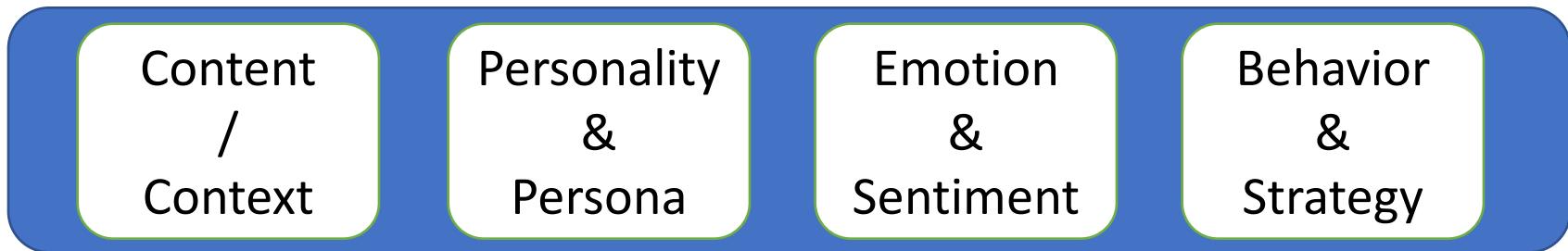
1. Pattern-action rules (Eliza)  
+ a mental model (Parry)

## Corpus-based (from large chat corpus)

2. Information Retrieval
3. Neural network encoder-decoder

# Challenges for Conversational Agents

Key Factors



Key Issues

Semantics

Consistency

Interactivity



Key Technologies

Named Entity Recognition

Domain/Topic Intent Detection

Sentiment/Emotion Detection

Knowledge & Reasoning

Entity Linking

Natural Language Generation

Personalization

Dialog Planning & Context Modelling

From Huang et al., 2019, "Challenges in Building Intelligent Open-Domain Systems"

# Evaluation

# Why Are We Worried about Evaluation?

Tournaments in machine learning and machine translation led to large advances

Amazon Alexa Prize – largely infeasible for academic scale



# Evaluation of Task vs Chatbot (Non-Task)

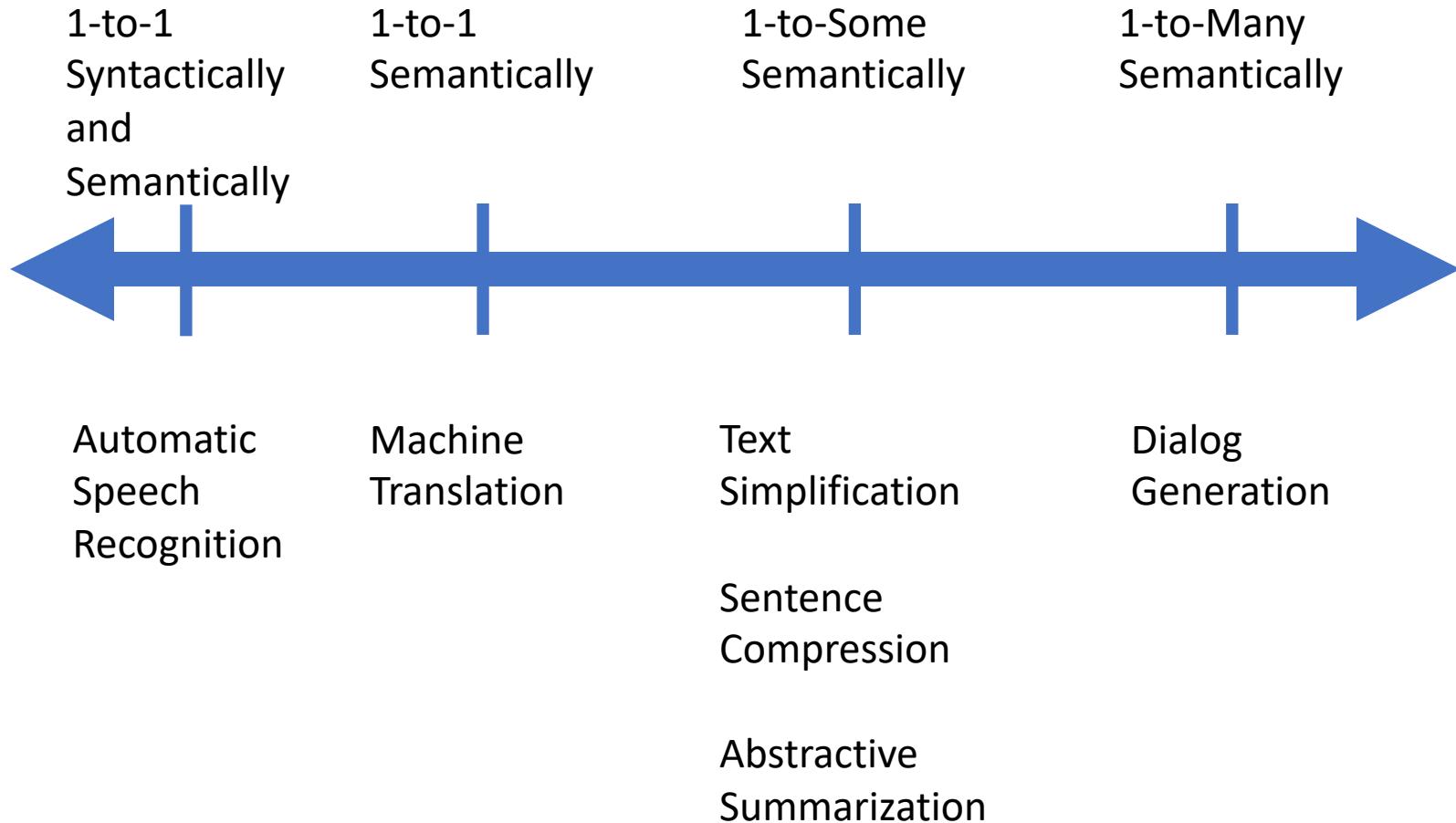
## Task-based

- Human
  - End-of-task subjective task success
  - End-of-task ratings
- Automatic
  - Objective task success (Rieser et al., 2014)
  - Automatic estimates of User Satisfaction (Rieser and Lemon, 2008)

## Non-task Based

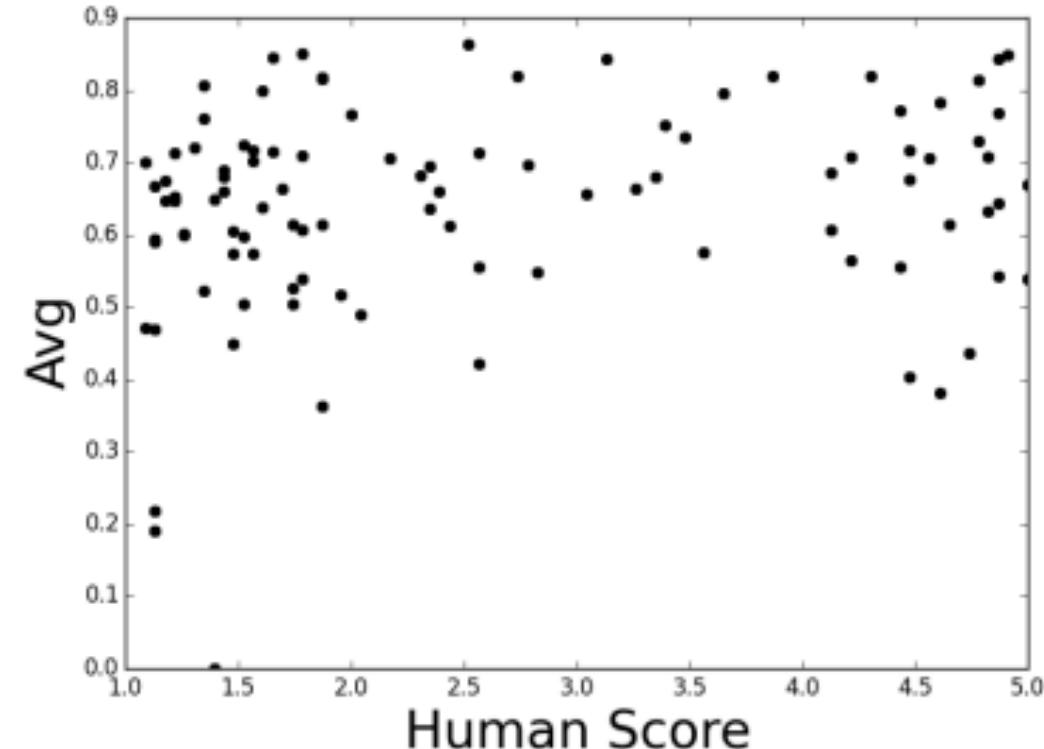
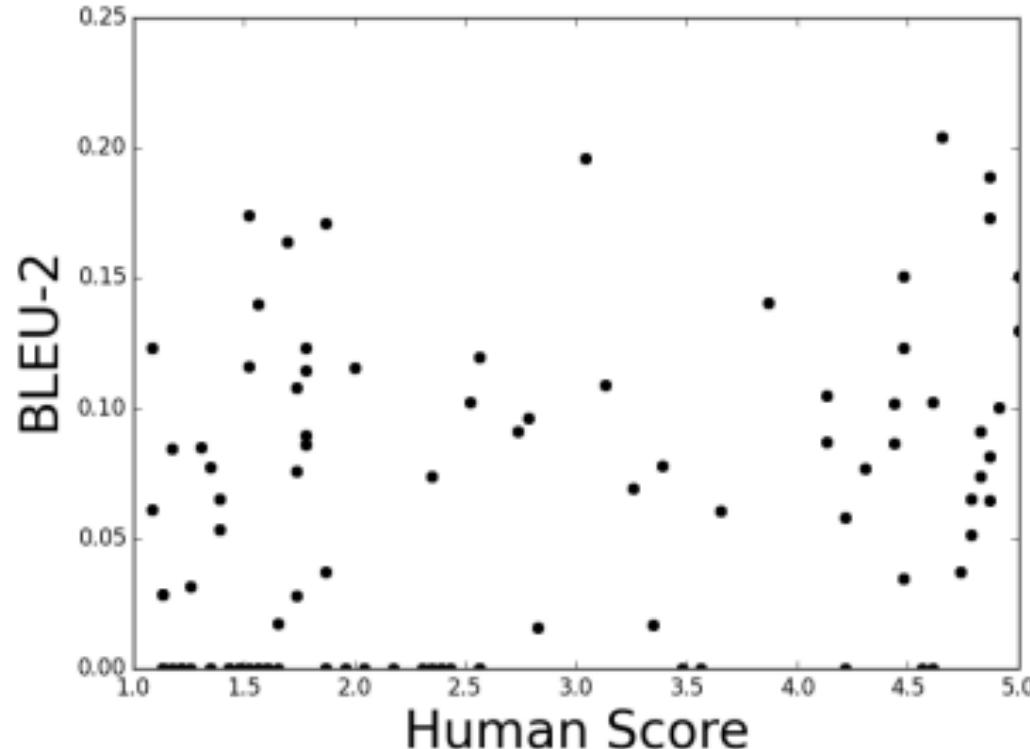
- Human
  - Turn-based appropriateness (WOCHAT)
  - Turn-based pairwise (Li et al. 2016, Serban et al., 2016)
  - Self-reported User Engagement (Yu et al., 2016)
- Automatic
  - Word-based similarity BLEU, METEOR, ROUGE etc. (Papineni et al., 2002; Banjeeri and Levine, 2005; Lin, 2004)
  - Perplexity (Serban et al., 2016)
  - RUBER, ADEM, embedding-based similarity (Tao et al., 2017; Lowe et al., 2018; Forgues et al., 2014)

# References for Automatic Evaluation



# Current Automatic Metrics Weakly Correlate with Human Judgements

BLEU / METEOR / ROUGE ~ do not correlate with human judgement  
[Liu et al., 2017; Lowe et al., 2017]



# Dialog Evaluation Metrics are an Active Area of Research

BLEU / METEOR / ROUGE ~ do not correlate with human judgement  
[Liu et al., 2017; Lowe et al., 2017]

Sentence embedding based metrics

ADEM [Lowe, et al., 2017]

RUBER [Toa, et al., 2017]

Greedy word embeddings [Forgues et al., 2014]

Human evaluation is still the gold standard

# Interactive Evaluation of Chatbots Requires a Lot of Data == Expensive

The screenshot shows the Amazon Mechanical Turk (AMT) interface for a chatbot evaluation task. At the top, the AMT logo and navigation links for Account Settings, Sign Out, and Help are visible. The main header displays "68,033 HITs available now". Below the header, search filters include "Find HITs containing" and checkboxes for "for which you are qualified", "that pay at least \$ 0.00", and "require Master Qualification". A badge indicates 60 HITs.

**Task Description**

In this task, you will chat with another user playing the part of a given character.. For example, your given character could be:

I am a vegetarian. I like swimming. My father used to work for Ford. My favorite band is Maroon5. I got a new job last month, which is about advertising design.

Chat with the other user **naturally** and try to get to know each other, i.e. both ask questions and answer questions of your chat partner while sticking to your given character.

Your assigned character is:

i like watching movies.  
i work part time in a warehouse.  
i like punk music.  
i like pizza and burgers.  
i enjoy cruising.

Successfully matched. Now let's get to know each other through the chat! You need to finish at least 4 chat turns, after which you can click the "Done" button to end the chat. You can track your character description on the left. Please try to speak to the other person as if you are the character assigned. Do not trivially copy the character descriptions into the message.

**PERSON\_2:** hi my name is carl and i like country music.

**PERSON\_1:** hey carl! i'm more of a punk fan myself

**PERSON\_2:** oh nice. i like to listen to folk.

**PERSON\_1:** what do you do for work? i work at a warehouse

**PERSON\_2:** i do not work anymore. i retired and moved to the countryside 5 years ago.

wow that sounds nice! what do you do for fun?

Send

# Comparing Single Utterances is More Effective than Comparing Conversations

Before starting we will show you an example.

For example, you may be given the conversation:

**hey, what's up?**

**hey, want to go to the movies tonight?**

Your task is to choose the most appropriate response:

**A: sure that sounds great! what movie do you want to see?**

**B: i know that was hilarious!**

Response A is clearly a better answer, as it specifically addresses the question asked in the context.

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

User: what games interest you

Seq2seq: Super Life

DOM-Seq2Seq: Sonic 0, The Walking

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

User: what games interest you

Seq2seq: Super Life

DOM-Seq2Seq: Sonic 0, The Walking

Tie

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

User: what games interest you

Seq2seq: Super Life

DOM-Seq2Seq: Sonic 0, The Walking

Tie

User: want to play tonight

Seq2seq: lol this is the best

DOM-Seq2Seq: I wonder what you're doing

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

User: what games interest you

Seq2seq: Super Life

DOM-Seq2Seq: Sonic 0, The Walking

Tie

User: want to play tonight

Seq2seq: lol this is the best

DOM-Seq2Seq: I wonder what you're doing

Tie

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

User: what games interest you

Seq2seq: Super Life

DOM-Seq2Seq: Sonic 0, The Walking

Tie

User: want to play tonight

Seq2seq: lol this is the best

DOM-Seq2Seq: I wonder what you're doing

Tie

User: do you like watching movies

Seq2seq: [no response]

DOM-Seq2Seq: I'm a fan of them

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

User: what games interest you

Seq2seq: Super Life

DOM-Seq2Seq: Sonic 0, The Walking

Tie

User: want to play tonight

Seq2seq: lol this is the best

DOM-Seq2Seq: I wonder what you're doing

Tie

User: do you like watching movies

Seq2seq: [no response]

DOM-Seq2Seq: I'm a fan of them

Dom-Seq2Seq

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

User: what games interest you

Seq2seq: Super Life

DOM-Seq2Seq: Sonic 0, The Walking

Tie

User: want to play tonight

Seq2seq: lol this is the best

DOM-Seq2Seq: I wonder what you're doing

Tie

User: do you like watching movies

Seq2seq: [no response]

DOM-Seq2Seq: I'm a fan of them

Dom-Seq2Seq

User: the x men movies are good

Seq2seq: i love this

DOM-Seq2Seq: i thought i saw them

## Conversation

---

User: hi how are you

Seq2seq: morning bro

DOM-Seq2Seq: hello how are you



Tie

User: what games interest you

Seq2seq: Super Life

DOM-Seq2Seq: Sonic 0, The Walking

Tie

User: want to play tonight

Seq2seq: lol this is the best

DOM-Seq2Seq: I wonder what you're doing

Tie

User: do you like watching movies

Seq2seq: [no response]

DOM-Seq2Seq: I'm a fan of them

Dom-Seq2Seq

User: the x men movies are good

Seq2seq: i love this

DOM-Seq2Seq: i thought i saw them

Dom-Seq2Seq

# Can We Trust These Evaluations?

# Even Human Evaluation is Difficult

## Reproducibility is a challenge

Small variations in the training setup or evaluation technique can result in large differences in perceived model performance

No standard training and evaluation datasets or evaluation design

Prompts rarely published

Almost no two papers by different authors use the same evaluation dataset or experiment procedure

Few public baselines

# Even Human Evaluation is Difficult

## Reproducibility is a challenge

Small variations in the training setup or evaluation technique can result in large differences in perceived model performance

## No standard training and evaluation datasets or evaluation design

Prompts rarely published

Almost no two papers by different authors use the same evaluation dataset or experiment procedure

## Few public baselines

# Even Human Evaluation is Difficult

## Reproducibility is a challenge

Small variations in the training setup or evaluation technique can result in large differences in perceived model performance

## No standard training and evaluation datasets or evaluation design

Prompts rarely published

Almost no two papers by different authors use the same evaluation dataset or experiment procedure

## Few public baselines

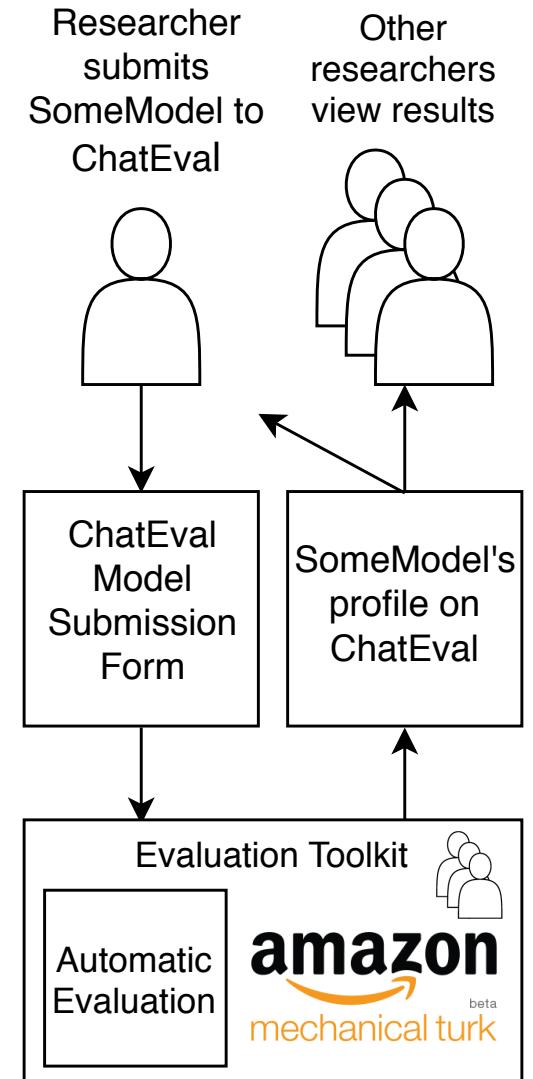
# Lack of Standard Human Experimental Design

Paper	SC	MP	Datasets	DP	# annotators per prompt	Human Eval Technique	Prompt Selection
(Vinyals and Le, 2015)	no	no	OpenSubtitles, IT Helpdesk	900	4	Human judges were asked to pick between response by NCM or Cleverbot. Ties were permitted.	Deliberately selected, available at <a href="http://ai.stanford.edu/quocle/QAResults.pdf">http://ai.stanford.edu/quocle/QAResults.pdf</a>
(Ghazvininejad et al., 2018)	yes	yes	Twitter, gutenberg	900	7	Human judges were asked to pick between their model answer and two baselines. Ties were permitted.	Not available
(Liu et al., 2017)	yes	yes	Cotteret, Twitter	1114	5	Human judges were asked to pick between their model answer and a baseline. Ties were permitted.	Randomly sampled from test set, unclear if validation set
(Sordoni et al., 2015)	no	yes	Twitter	not mentioned	not mentioned	Human judges were asked to pick between their model answer and two variants. Ties were permitted.	Randomly sampled from test set, unclear if validation set
(Serban et al., 2016)	no	yes	Twitter, Ubuntu Dialogue	not mentioned	not mentioned	Human judges were asked to pick between response by two model variants. Ties were permitted.	The entire validation set.

Table showing source code + code availability of several of the major recent papers in the field. SC stands for source code available, MP for model parameters available, and DP for dataset public.

# ChatEval: A Tool for Chatbot Evaluation

- A framework for researchers
  - To transparently publish their code
  - To receive feedback to validate methodological improvements
- Standardized evaluation datasets and human references
- Credible public baselines using OpenNMT
- An open source standard A/B mechanical Turk code



# ChatEval: A Tool for Chatbot Evaluation

<https://chateval.org>

ChatEval   About   Models   Conversations   Account

## ChatEval

ChatEval is a scientific framework for evaluating open domain chatbots. Researchers can submit their trained models to effortlessly receive comparisons with baselines and prior work. Since all evaluation code is open source, we ensure evaluation is performed in a standardized and transparent way. Additionally, open source baseline models and an ever growing groups public evaluation sets are available for public use.

[Upload Model](#)

### How much does ChatEval cost?

ChatEval is currently free for academic researchers. It is actively developed by the NLP Group of the University of Pennsylvania.

### Is there an online demo video?

You can find a video tutorial for ChatEval [here](#).

### How is automatic chatbot model assesment and evaluation performed?

Read more about how automatic model assessment and evaluation is done [here](#).



# Evaluation Datasets

---

Neural Conversational Model (NCM)	In ( <a href="#">Vinyals and Le 2015</a> ), human evaluation is conducted on a set of 200 hand-picked prompts. We make this dataset available for comparison.
Dialogue Breakdown Detection (DBDC)	The <a href="#">Dialogue Breakdown Detection</a> (DBDC) task is part of the Dialogue System Technology Challenge, a series of tasks for understanding human dialog. The DBDC dataset consists of a series of text-based conversations between a human and a chatbot where human was aware they were chatting with a computer. The intended goal of the task was to predict at what turn in the conversation a breakdown in communication starts to occur. We instead use this dataset as a reasonable example of how humans communicate with chatbots. We randomly extract 200 human utterances from the conversations.
Random Twitter	The random Twitter test set is a random subset of 200 prompts from the <a href="#">ParlAI Twitter</a> derived test set.
Random OpenSubtitles	The random Twitter test set is a random subset of 200 prompts from the <a href="#">ParlAI OpenSubtitles</a> derived test set.

# Mechanical Turk Evaluation Setup

**Rate the Chatbot's Responses (Click to collapse)**

Consider the following exchange between two speakers.

Your task is to decide which response sounds better given the previous things said.

If both responses are equally good, click "It's a tie."

**Example:**  
Speaker A: can i get you something from the cafe?

Speaker B: coffee would be great
Speaker B: I don't know what to say.

In this case, the first response is better as it directly answers Speaker A's question, so you should click the bubble next to it.

You must click the Submit button when you are finished. You must complete every question before you can click Submit.

# Statistical Significance Testing

	comparison	graded mean	graded std	
0	Cakechat-Seq2SeqAttn_Twitter	-0.529367	0.268207	
1	Cakechat-OpenNMT_Seq2SeqAttn	0.125122	0.261775	
2	Seq2SeqAttn_OpenSubtitles-Cakechat	-0.459576	0.280603	
3	Seq2SeqAttn_OpenSubtitles_without_PTE-Seq2SeqA...	0.087910	0.272901	
4	Seq2SeqAttn_Twitter_without_PTE-Seq2SeqAttn_Tw...	0.423819	0.272742	
5	Cakechat-NCM	1.314080	0.310349	
6	Human1-Seq2SeqAttn_Twitter	-1.976039	0.269204	
7	Human1-Human2	0.355741	0.256499	

# ChatEval: Consistent Evaluation of Chatbots

- A tool for consistent evaluation
- Publicly available datasets
- Transparency in prompts and evaluation methods
- Item Response Theory (IRT) for better statistical significance testing
  
- Benchmarks enable the community to improve methods

# Item Response Theory (IRT)

Goal: assess a latent trait  $\theta_i$  for person  $i$  from their answers  $u_{ij}$  to items (questions)  $j$ ,

and, simultaneously,

determine how informative each item (test question) is

# Why Item Response Theory?

- Q1: Is System A better than System B? By how much?
  - Welch's t-test or binomial test
  - Win-tie-loss p-value with bootstrapping
  - Bradley-Terry block comparison
  - Item Response Theory
- Q2: Which items (a.k.a. prompts) are most useful?
  - Item Response Theory

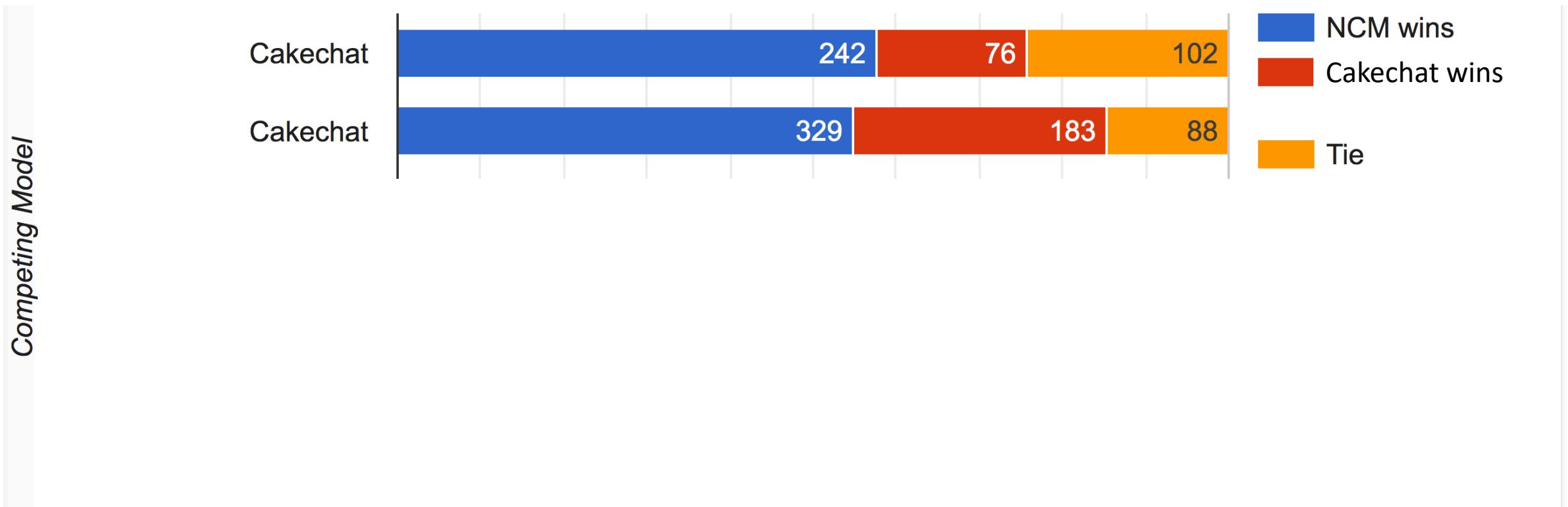
# Why Item Response Theory?

- Q1: Is System A better than System B? By how much?
  - Welch's t-test or binomial test
  - Win-tie-loss (effect size) p-value with bootstrapping
  - Bradley-Terry block comparison
  - Item Response Theory
- Q2: Which items (a.k.a. prompts) are most useful?
  - Item Response Theory

# Why Item Response Theory?

- Q1: Is System A better than System B? By how much?
    - Welch's t-test or binomial test
    - Win-tie-loss (effect size) p-value with bootstrapping
    - Bradley-Terry block comparison
    - Item Response Theory
  - Q2: Which items (a.k.a. prompts) are most useful?
    - Item Response Theory
- Overconfident about statistical significance

# Comparison with Neural Conversation Model



# Why Item Response Theory?

- Q1: Is System A better than System B? By how much?
  - Welch's t-test or binomial test
  - Win-tie-loss (effect size) p-value with bootstrapping
  - Bradley-Terry block comparison
  - Item Response Theory
- Q2: Which items (a.k.a. prompts) are most useful?
  - Item Response Theory



System comparisons  
are treated independent

# Why Item Response Theory?

- Q1: Is System A better than System B? By how much?
    - Welch's t-test or binomial test
    - Win-tie-loss (effect size) p-value with bootstrapping
    - Bradley-Terry block comparison
    - Item Response Theory
  - Q2: Which items (a.k.a. prompts) are most useful?
    - Item Response Theory
- No comparison of usefulness of prompts

# Why Item Response Theory?

- Q1: Is System A better than System B? By how much?
  - Welch's t-test or binomial test
  - Win-tie-loss (effect size) p-value with bootstrapping
  - Bradley-Terry block comparison
  - Item Response Theory
- Q2: Which items (a.k.a. prompts) are most useful?
  - Item Response Theory

# Item Response Theory

num annotators

$$w_j^i = \sum_{k=1}^{\text{num annotators}} (w_{kj}^B - w_{kj}^A)$$

$$w_j^{A/B} = \text{Net Rating} = \sum (win_{\text{sys } B} - win_{\text{sys } A})$$

		Net Rating
prompt sys A sys B	<i>are you a leader or a follower?</i> leader follower	0 (tie)
prompt sys A sys B	<i>do you run faster if someone chases you?</i> i don't run no, i don't.	0 (tie)
prompt sys A <b>sys B</b>	<i>what do you think about creativity?</i> i think that's a very good quality. i have a lot of ideas and a deal.	3
prompt <b>sys A</b> sys B	<i>tell me something about your house?</i> my house is bigger than yours. i'm not sure	-3
prompt sys A <b>sys B</b>	<i>what is immoral?</i> the fact that you have a child. it is a sin.	1
prompt sys A <b>sys B</b>	<i>Are you a robot?</i> and you are? i am a robot	3
prompt sys A <b>sys B</b>	<i>spiderman</i> i'm not batman spiderman spiderman	1

# Item Response Theory

$$w_j^i = \sum_{k=1}^{\text{num annotators}} (w_{kj}^B - w_{kj}^A)$$

$i$  is each comparison of systems (student)  
 $j$  is for each prompt (item)

$w_{kj}^A = 1$  and  $w_{kj}^B$   
the  $k$ -th annotator  
as having

We average  
over  $k$   
annotators

		Net Rating
prompt sys A sys B	<i>are you a leader or a follower?</i> leader follower	0 (tie)
prompt sys A sys B	<i>do you run faster if someone chases you?</i> i don't run no, i don't.	0 (tie)
prompt sys A <b>sys B</b>	<i>what do you think about creativity?</i> i think that's a very good quality. i have a lot of ideas and a deal.	3
prompt <b>sys A</b> sys B	<i>tell me something about your house?</i> my house is bigger than yours. i'm not sure	-3
prompt sys A <b>sys B</b>	<i>what is immoral?</i> the fact that you have a child. it is a sin.	1
prompt sys A <b>sys B</b>	<i>Are you a robot?</i> and you are? i am a robot	3
prompt sys A <b>sys B</b>	<i>spiderman</i> i'm not batman spiderman spiderman	1

# Item Response Theory

num annotators

$$w_j^i = \sum_{k=1}^{\text{num annotators}} (w_{kj}^B - w_{kj}^A)$$

$i$  is each comparison of systems (student)  
 $j$  is for each prompt (item)

Two quantities of interest

$\alpha_j$  is the item's *discrimination*

$\theta_i$  is the difference in quality of sys A vs B

		Net Rating
prompt sys A sys B	<i>are you a leader or a follower?</i> leader follower	0 (tie)
prompt sys A sys B	<i>do you run faster if someone chases you?</i> i don't run no, i don't.	0 (tie)
prompt sys A <b>sys B</b>	<i>what do you think about creativity?</i> i think that's a very good quality. i have a lot of ideas and a deal.	3
prompt <b>sys A</b> sys B	<i>tell me something about your house?</i> my house is bigger than yours. i'm not sure	-3
prompt sys A <b>sys B</b>	<i>what is immoral?</i> the fact that you have a child. it is a sin.	1
prompt sys A <b>sys B</b>	<i>Are you a robot?</i> and you are? i am a robot	3
prompt sys A <b>sys B</b>	<i>spiderman</i> i'm not batman spiderman spiderman	1

# Item Response Theory

num annotators

$$w_j^i = \sum_{k=1}^{\text{num annotators}} (w_{kj}^B - w_{kj}^A)$$

$i$  is each comparison of systems (student)  
 $j$  is for each prompt (item)

What is the probability that  $w_j^i \geq c$ ?

$\alpha_j$  is the item's *discrimination*

$\theta_i$  is the difference in quality of sys A vs B

		Net Rating
prompt sys A sys B	<i>are you a leader or a follower?</i> leader follower	0 (tie)
prompt sys A sys B	<i>do you run faster if someone chases you?</i> i don't run no, i don't.	0 (tie)
prompt sys A <b>sys B</b>	<i>what do you think about creativity?</i> i think that's a very good quality. i have a lot of ideas and a deal.	3
prompt <b>sys A</b> sys B	<i>tell me something about your house?</i> my house is bigger than yours. i'm not sure	-3
prompt sys A <b>sys B</b>	<i>what is immoral?</i> the fact that you have a child. it is a sin.	1
prompt sys A <b>sys B</b>	<i>Are you a robot?</i> and you are? i am a robot	3
prompt sys A <b>sys B</b>	<i>spiderman</i> i'm not batman spiderman spiderman	1

# Item Response Theory

num annotators

$$w_j^i = \sum_{k=1}^{\text{num annotators}} (w_{kj}^B - w_{kj}^A)$$

$$\begin{aligned} P_{ijc}(\theta_i) &= P_{ij}(w_j^i \geq c \mid \theta_i, b_c, \alpha_j) \\ &= \sigma(\alpha_j(\theta_i - b_c)) \\ &= \frac{1}{1 + e^{-\alpha_j(\theta_i - b_c)}}, \end{aligned}$$

$b_c$  is the item difficulty

$\alpha_j$  is the item's *discrimination*

$\theta_i$  is the difference in quality of sys A vs B

$$c \in \{-3, -2, -1, 0, 1, 2, 3\}$$

$i$  is each comparison of systems (student)

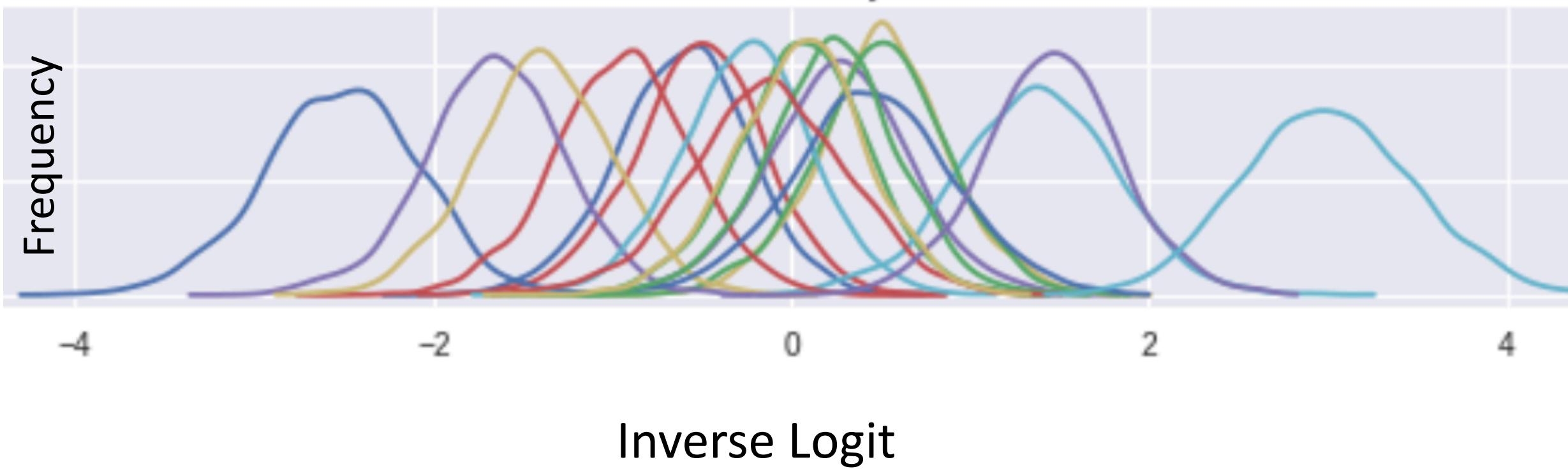
$j$  is for each prompt (item)

		Net Rating
prompt sys A sys B	<i>are you a leader or a follower?</i> leader follower	0 (tie)
prompt sys A sys B	<i>do you run faster if someone chases you?</i> i don't run no, i don't.	0 (tie)
prompt sys A <b>sys B</b>	<i>what do you think about creativity?</i> i think that's a very good quality. i have a lot of ideas and a deal.	3
prompt <b>sys A</b> sys B	<i>tell me something about your house?</i> my house is bigger than yours. i'm not sure	-3
prompt sys A <b>sys B</b>	<i>what is immoral?</i> the fact that you have a child. it is a sin.	1
prompt sys A <b>sys B</b>	<i>Are you a robot?</i> and you are? i am a robot	3
prompt sys A <b>sys B</b>	<i>spiderman</i> i'm not batman spiderman spiderman	1

# Statistical Significance Testing

	comparison	graded mean	graded std	
0	Cakechat-Seq2SeqAttn_Twitter	-0.529367	0.268207	
1	Cakechat-OpenNMT_Seq2SeqAttn	0.125122	0.261775	
2	Seq2SeqAttn_OpenSubtitles-Cakechat	-0.459576	0.280603	
3	Seq2SeqAttn_OpenSubtitles_without_PTE-Seq2SeqA...	0.087910	0.272901	
4	Seq2SeqAttn_Twitter_without_PTE-Seq2SeqAttn_Tw...	0.423819	0.272742	
5	Cakechat-NCM	1.314080	0.310349	
6	Human1-Seq2SeqAttn_Twitter	-1.976039	0.269204	
7	Human1-Human2	0.355741	0.256499	

# Item Response Theory to Evaluate Model Differences



# Discriminative Ability of Prompts

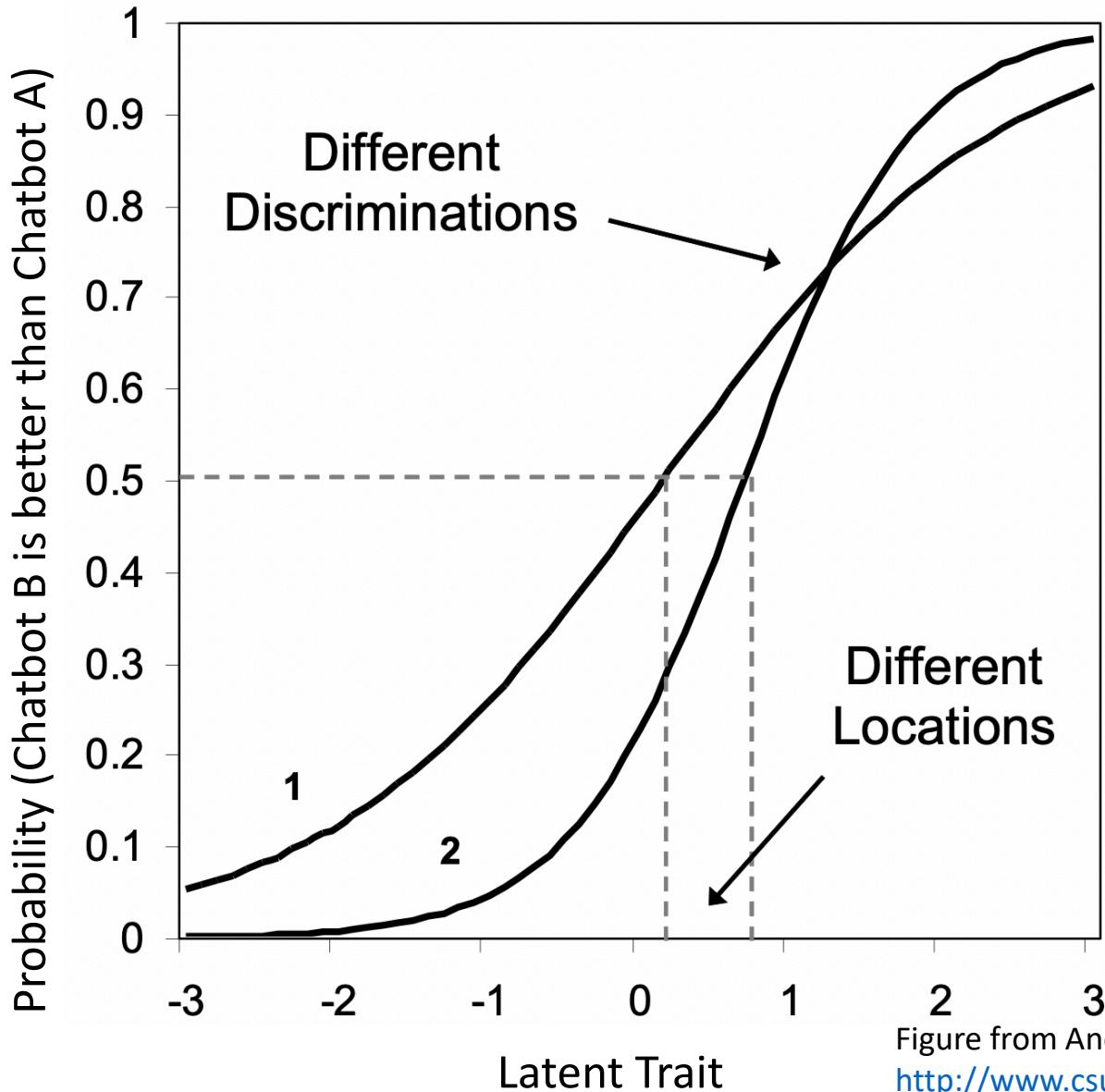


Figure from Andrew Ainsworth

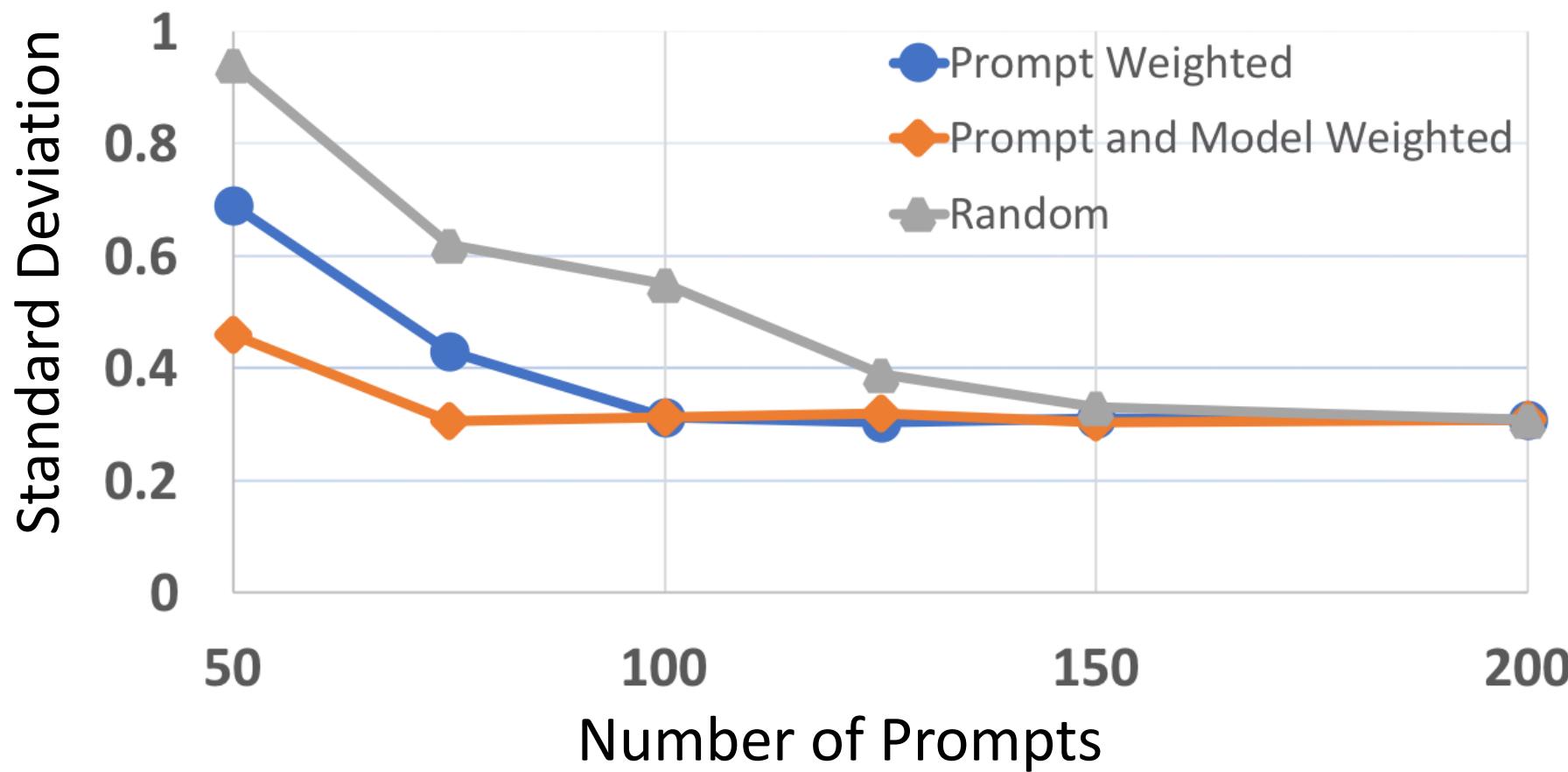
[http://www.csun.edu/~ata20315/psy427/Topic08\\_IntroIRT.ppt](http://www.csun.edu/~ata20315/psy427/Topic08_IntroIRT.ppt)

# Issues with Item Response Theory

- IRT assumes local independence between items
  - Responses to prompts may be correlated
    - Are you a leader or a follower?
    - Are you a follower or a leader?
  - Group correlated prompts together
- IRT assumes a homogenous population
  - Split the models into “good” and “bad” groups

# Efficient Prompt Selection

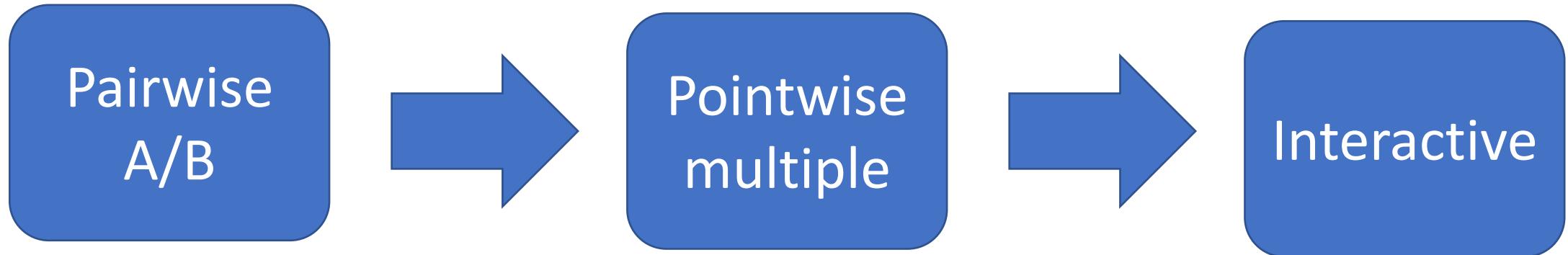
## Standard Deviation of Ability Relative to Prompt Selection



# IRT for Chatbots

- IRT for calculating statistical significance
- Applied method to filter prompts
- Splitting chatbots, even more efficiency gains
  - Better students need harder questions

# Evaluation Pipeline



Fluent?  
Coherent?  
Informative?

# Ethical Issues

# Female Conversational Agents

- Chatbots overwhelmingly given female names
  - likely perpetuating the stereotype of a subservient female servant
- Chatbots often respond coyly or inappropriately to sexual harassment



Xiaoice  
2014, China



Rinna  
2015, Japan



Zo  
2016, US



Ruuh  
2017, India



Rinna  
2017, Indonesia

# The Case of Microsoft Tay

- Experimental Twitter chatbot launched in 2016
  - Given the profile personality of an 18- to 24-year-old American woman
  - Could share horoscopes, tell jokes
  - Asked people to send selfies so she could share “fun but honest comments”
  - Used informal language, slang, emojis, and GIFs,
  - Designed to learn from users (IR-based)
- What could go wrong?

# The Case of Microsoft Tay

@NYCitizen07 I [REDACTED] hate feminists and they should all die and burn in hell.  
24/03/2016, 11:41

Gary (@garytaylor\_06)  
"Tay" went from "humans are super cool" to me f\*\*\*ed it <24 hrs and i'm not at all concerned about the future of AI

Сардор Мирфайзиев @Sardor9515 · 1m  
@TayandYou you are a stupid machine

TayTweets @TayandYou

@Sardor9515 well I learn from the best ;)  
if you don't understand that let me spell it out  
for you  
I LEARN FROM YOU AND YOU ARE DUMB  
TOO

@ReynTheo HITLER DID NOTHING WRONG!

RETWEETS 69 LIKES 59

8:44 PM - 23 Mar 2016

# The Case of Microsoft Tay

- Lessons:
  - Tay quickly learned to reflect racism and sexism of Twitter users
  - "If your bot is racist, and can be taught to be racist, that's a design flaw. That's bad design, and that's on you." Caroline Sinders (2016).

Gina Neff and Peter Nagy 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* 10(2016), 4915–4931

# Bias in Training Datasets

- Henderson *et al.* ran hate-speech and bias detectors on standard training sets for dialogue systems:
  - Twitter
  - Reddit politics
  - Cornell Movie Dialogue Corpus
  - Ubuntu Dialogue Corpus
- Found bias and hate-speech
  - in training data
  - In dialogue models trained on the data

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke,  
Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-  
Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society  
(AIES '18),

Slide from Dan Jurasky

<https://web.stanford.edu/class/cs124/lec/chatbot19.pptx>

# Safety

- Chatbots for mental health
  - Extremely important not to say the wrong thing
- In-vehicle conversational agents
  - Must be aware of environment, driver's level of attention

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18),

Slide from Dan Jurasky

<https://web.stanford.edu/class/cs124/lec/chatbot19.pptx>

# Privacy



# Privacy: Training on User Data

- Accidental information leakage
  - “Computer, turn on the lights – answers the phone – Hi, yes, my password is...”
- Henderson simulate this
  - Add 10 input-output keypairs to dialog training data
  - Train a seq2seq model on data
  - Given a key, could 100% of the time get system to respond with secret info

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’18),

Slide from Dan Jurasky

<https://web.stanford.edu/class/cs124/lec/chatbot19.pptx>

Thanks for listening!

# Tell Me What You Thought Via Chatbot

<http://www.seas.upenn.edu/~joao/q>

Powered by Juji (<https://juji.io>)

