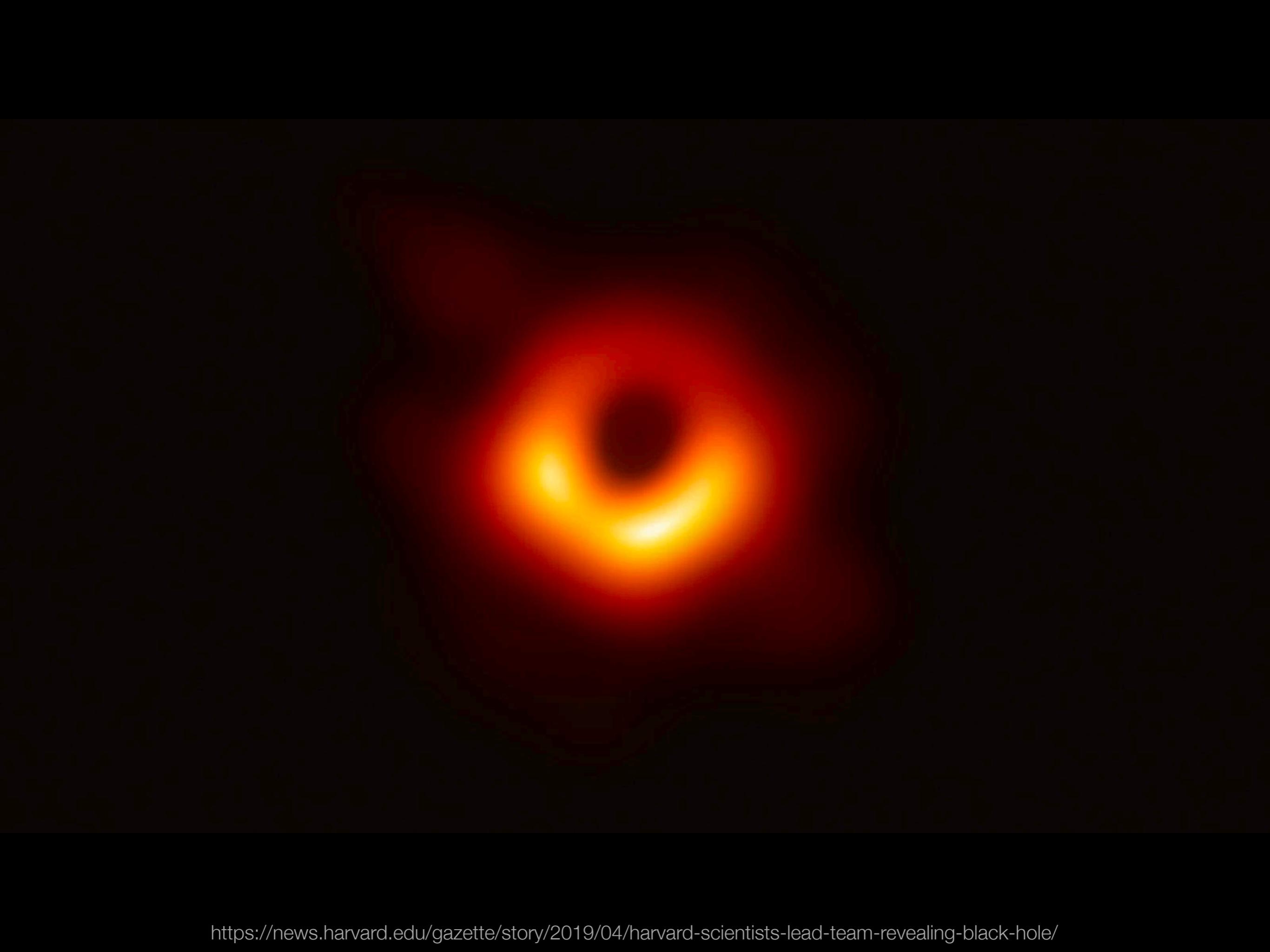




SI 630

Natural Language Processing: Algorithms and People

Lecture 13: Annotation and Crowdsourcing
April 15, 2020

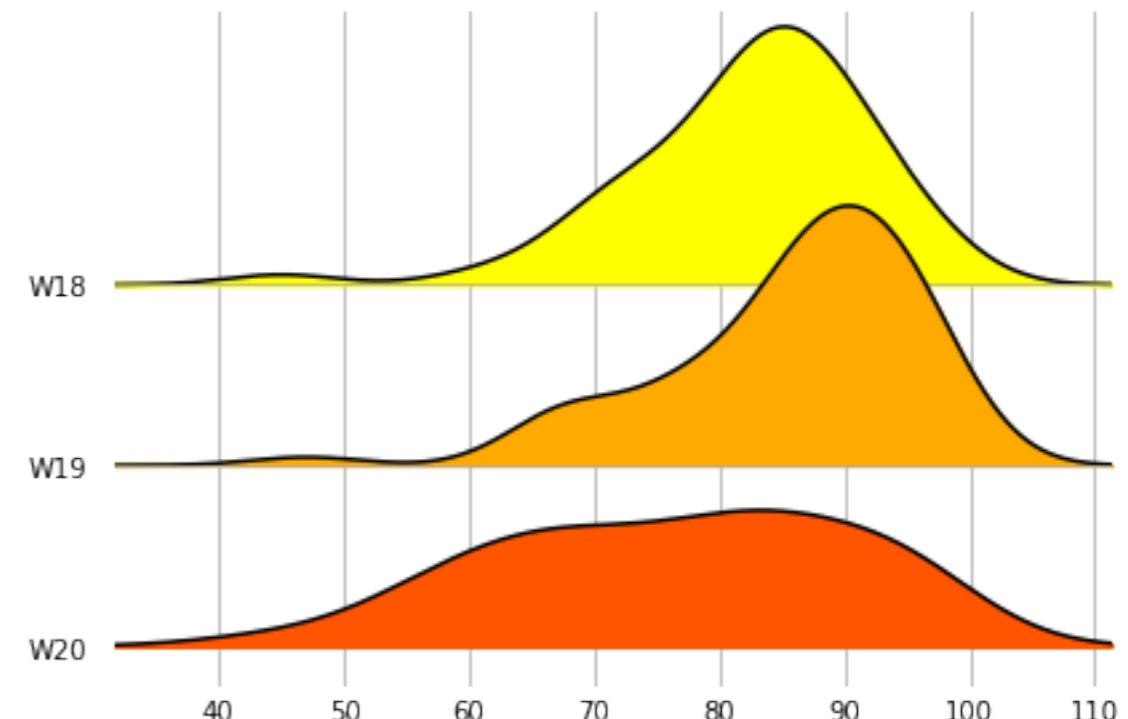
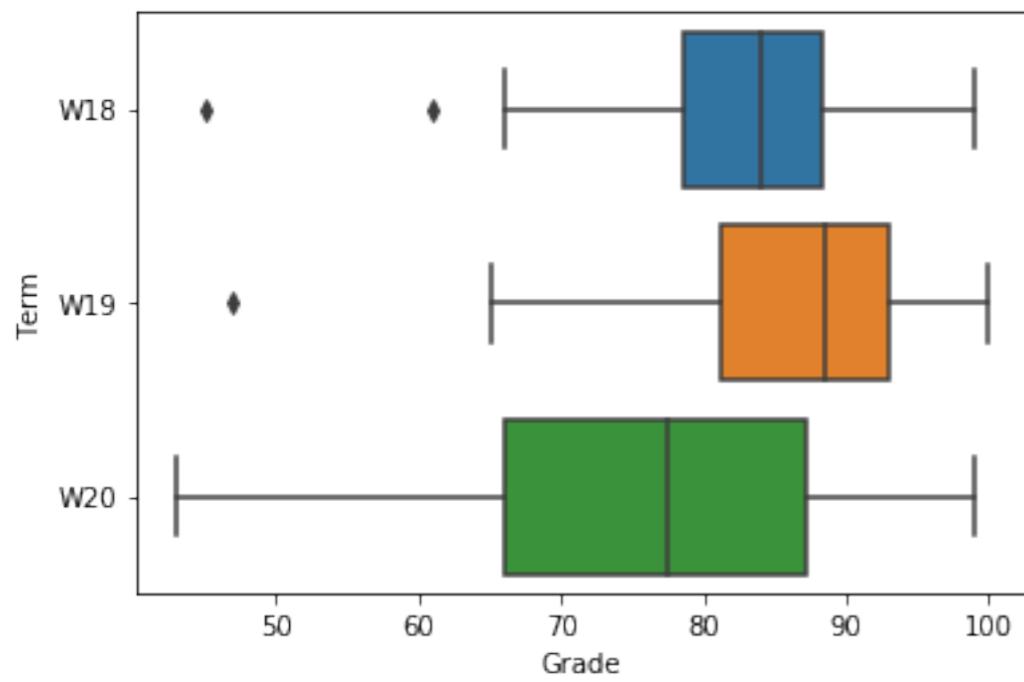


<https://news.harvard.edu/gazette/story/2019/04/harvard-scientists-lead-team-revealing-black-hole/>

Midterm Follow Up!

Overall Thoughts

- Scores lower this year than previous semesters :(



Common Mistakes

- Q1: none? 
- Q2: Proposed sentiment analysis systems when they were explicitly not asked for
- Q3: Described a Question Answering system without saying how it interacts
- Q4: Proposed using an LSTM on a 90 minute conversation, but this is too much input for modern deep learning systems (but modifications can be made so it works with an LSTM)

Q1a

1. Which of the following statements about the Skip-gram formulation of word2vec are correct (what you implemented in HW2)?
 1. When it comes to a small corpus, it often has better performance than tf-idf vectors on word similarity tasks
 2. It predicts the center word from the surrounding context words
 3. The final word vector for a word is the average or sum of the input vector v and output vector u corresponding to that word
 4. It makes use of global co-occurrence statistics
 5. None of the above
 6. All of the above

Q1b

1. Which of the following statements about dependency trees are correct?
 1. Each word is connected to exactly one dependent (i.e., each word has exactly one outgoing edge)
 2. A dependency tree is called a “projective” if all the words have a path to the ROOT
 3. A dependency tree may have internal nodes in the tree to represent syntactic elements not in the sentence
 4. None of the above
 5. All of the above

Q1c

1. Which of the following statements is true of neural language models?
 1. Neural feed-forward models with fixed-sized windows share weights across the window
 2. Neural feed-forward language models suffer from the sparsity problem, but count-based n-gram language models do not
 3. The number of parameters in an RNN language model grows with the number of time steps
 4. None of the above
 5. All of the above

Q1d

1. Which of the following statements is true when designing and testing NLP models
 1. Hyperparameters should be selected after repeated experiments measuring performance on the validation dataset
 2. Classification accuracy (as a metric) is usually uninformative for comparing performance when classes are heavily unbalanced in terms of frequency
 3. A performance improvement over the next-closest model with $p < 0.05$ means that the chance the models actually have the same performance is less than 5%
 4. None of the above
 5. All of the above

Q2

Explain the following in a 2-3 sentences each:

- The purpose of the language model for machine translation. If you want to translate from language A to language B, what data would you train this model on? Fluency
- The purpose of the translation model for machine translation. If you want to translate from language A to language B, what data would you train this model on? Translation quality
- In what circumstances would you want the texts for language B that are each used to train the language model and the translation model to look very different from each other? (i.e., the text written in B used to train the LM vs. the text written in B used to train the TM)

Q3

- Your friend is trying to analyze some 100K+ documents but after labeling 100 with their main topic, they give up because it will take forever. However, knowing you have taken 630, they ask you to build a custom topic model to analyze their data. When discussing the model with them, your friend mentions that they know specific words are in different topics and the main topic for the 100 documents they annotated (note that this is not the only topic for those documents!). However, they are not sure how many topics are in the corpus as a whole.

Given this prior knowledge,

- (1) describe in detail how would you change the Gibbs Sampling update procedure to build this new model,
- (2) in 2-3 sentences, what effect the prior knowledge will have on the output document-topic distribution when the model finishes.

LQ1: Human Trafficking Classifier

- Your task is to describe how you would create a classifier that (1) detects these types of pages and (2) is robust to the fact that the language on these pages changes regularly to avoid detection. Describe how you would design/adjust your classifier to match the focus on recall.

LQ2: Anything other than Sentiment Analysis

- [your goals are to] (1) better monetize a product (2) take actionable steps to improve a particular current project, or (3) decide what product to produce next.
- Your task is to design a system that takes in a massive volume of social media data about one company and produces insights (of your design) that satisfy at least one of those goals. You can safely assume that you are receiving millions of posts/comments/etc. from all of the major social media platforms and that you have existing metadata about the users from your company (e.g., demographics, location, past comment history), as you are a social media analytics company after all.

LQ3: Build a Chatbot for Kids

- Your task is to describe how you would approach designing such a chatbot. What techniques will you use, how will you determine what content is appropriate for children, and how will you adapt your responses to this particular audience.

LQ4: Summarize Long Interviews + Describe Ethics

- Your task is two-fold:
- First, describe a system design for this summarization. Note that you'll be summarizing a long document (~90 minutes!) so you need to be careful about input size and what all you report.
- Second, as a student of 630, you realize that 1,000 documents isn't necessarily a lot of data to train a summarization system (but it's still your job, so you have to do it). Describe what kind of technical *and* ethical issues you expect to arise in this situation, how you would test for these issues, and how you would design safeguards. Remember, therapist time is valuable so your solution can't be to have the therapist check everything.



Getting Data

Modern NLP is driven by
annotated data

Modern NLP is driven by annotated data

- Penn Treebank (1993; 1995; 1999); morphosyntactic annotations of WSJ

Modern NLP is driven by annotated data

- Penn Treebank (1993; 1995; 1999); morphosyntactic annotations of WSJ
- OntoNotes (2007–2013); syntax, predicate-argument structure, word sense, coreference

Modern NLP is driven by annotated data

- Penn Treebank (1993; 1995; 1999); morphosyntactic annotations of WSJ
- OntoNotes (2007–2013); syntax, predicate-argument structure, word sense, coreference
- FrameNet (1998–): frame-semantic lexica/annotations

Modern NLP is driven by annotated data

- Penn Treebank (1993; 1995; 1999); morphosyntactic annotations of WSJ
- OntoNotes (2007–2013); syntax, predicate-argument structure, word sense, coreference
- FrameNet (1998–); frame-semantic lexica/annotations
- MPQA (2005); opinion/sentiment

Modern NLP is driven by annotated data

- Penn Treebank (1993; 1995; 1999); morphosyntactic annotations of WSJ
- OntoNotes (2007–2013); syntax, predicate-argument structure, word sense, coreference
- FrameNet (1998–); frame-semantic lexica/annotations
- MPQA (2005); opinion/sentiment
- SQuAD (2016); annotated questions + spans of answers in Wikipedia

Modern NLP is driven by
annotated data

Modern NLP is driven by annotated data

- In most cases, the data we have is the product of **human judgments**.

Modern NLP is driven by annotated data

- In most cases, the data we have is the product of **human judgments**.
 - What's the correct part of speech tag?

Modern NLP is driven by annotated data

- In most cases, the data we have is the product of **human judgments**.
 - What's the correct part of speech tag?
 - Syntactic structure?

Modern NLP is driven by annotated data

- In most cases, the data we have is the product of **human judgments**.
 - What's the correct part of speech tag?
 - Syntactic structure?
 - Sentiment?

Ambiguity

“One morning I shot
an elephant in my pajamas”



Animal Crackers

Dogmatism

Fast and Horvitz (2016),
“Identifying Dogmatism in Social
Media: Signals and Models”

Given a comment, imagine you hold a well-informed, different opinion from the commenter in question. We'd like you to tell us how likely that commenter would be to engage you in a constructive conversation about your disagreement, where you each are able to explore the other's beliefs. The options are:

- (5):** It's unlikely you'll be able to engage in any substantive conversation. When you respectfully express your disagreement, they are likely to ignore you or insult you or otherwise lower the level of discourse.
- (4):** They are deeply rooted in their opinion, but you are able to exchange your views without the conversation degenerating too much.
- (3):** It's not likely you'll be able to change their mind, but you're easily able to talk and understand each other's point of view.
- (2):** They may have a clear opinion about the subject, but would likely be open to discussing alternative viewpoints.
- (1):** They are not set in their opinion, and it's possible you might change their mind. If the comment does not convey an opinion of any kind, you may also select this option.

Sarcasm

“In many respects you know they honor President Obama. ISIS is honoring President Obama! He is the founder of ISIS. He’s the founder of ISIS, O.K.! He’s the founder, he founded ISIS and I would say the co-founder would be crooked Hillary Clinton. Co-founder, crooked Hillary Clinton. And that’s what it’s about.”

Sarcasm

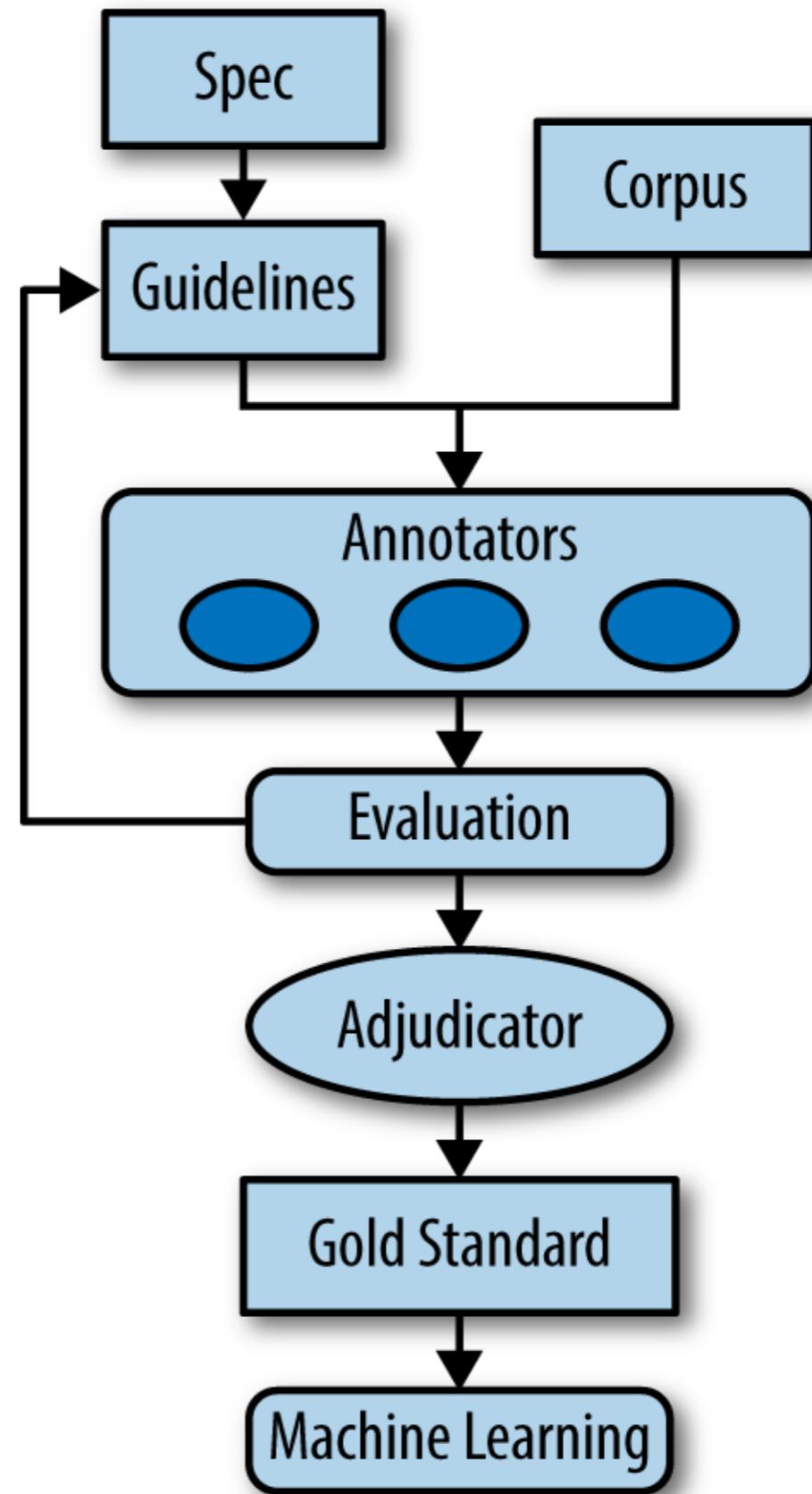
“In many respects you know they honor President Obama. ISIS is honoring President Obama! He is the founder of ISIS. He’s the founder of ISIS, O.K.! He’s the founder, he founded ISIS and I would say the co-founder would be crooked Hillary Clinton. Co-founder, crooked Hillary Clinton. And that’s what it’s about.”

The image shows a screenshot of a Twitter post from Donald J. Trump's account. The post includes his profile picture, his name with a blue verification checkmark, and his handle @realDonaldTrump. To the right is a blue "Follow" button. The tweet itself reads: "Ratings challenged @CNN reports so seriously that I call President Obama (and Clinton) "the founder" of ISIS, & MVP. THEY DON'T GET SARCASM?" Below the tweet is the timestamp "3:26 AM - Aug 12, 2016". At the bottom are engagement metrics: 9,730 replies, 7,787 retweets, and 23,837 likes. There is also a small "•" icon.

Fake News

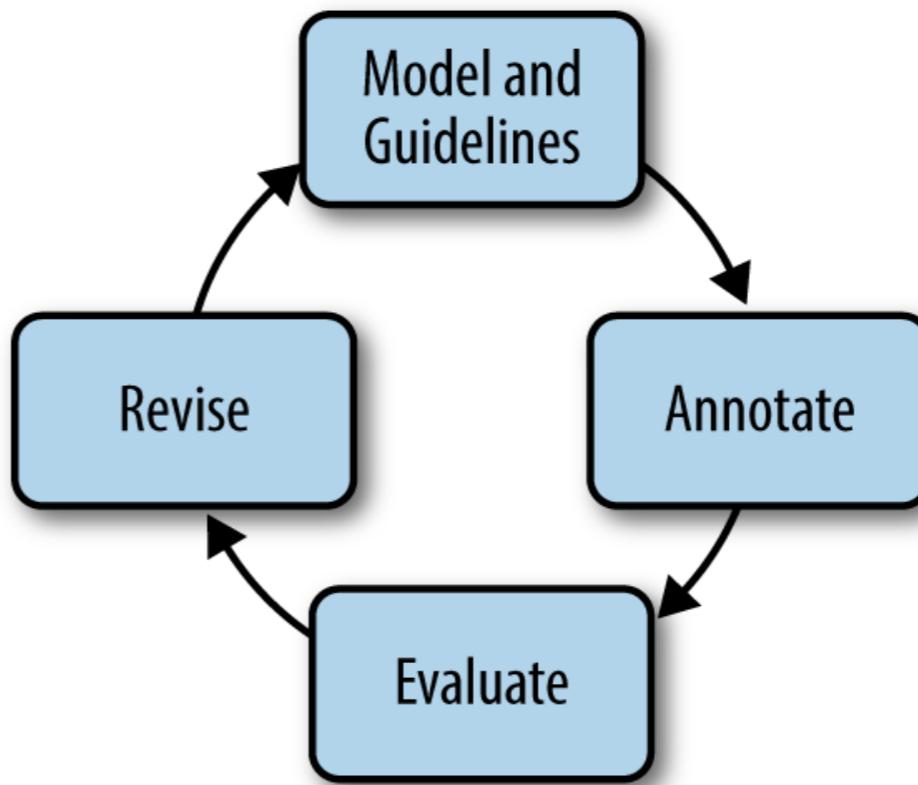
The word cloud is composed of numerous words in different sizes and colors, all related to the theme of fake news. The largest word, "Fake News Challenge", is in a large green font. Other prominent words include "Like" (large green), "Share" (large green), "fraud" (large green), "rumor" (large dark green), "hero" (large black), "vote" (large black), "trust" (large black), "truth" (large black), "secret" (large dark green), "election" (large dark green), "protest" (large dark green), "enemy" (large dark green), "foreign" (large dark green), "politics" (large dark green), "power" (large dark green), "photos" (large dark green), "fear" (large dark green), "risk" (large dark green), "speech" (large dark green), "terror" (large blue), "internet" (large blue), "war" (large blue), "share" (large green), "say" (large green), "mob" (medium green), "party" (medium green), "trigger" (medium green), "panic" (medium green), "leader" (medium green), "argue" (medium green), "democracy" (medium green), "unqualified" (medium green), "dictator" (medium green), "force" (medium green), "propaganda" (medium green), "site" (medium blue), "kill" (medium blue), "danger" (medium blue), "ugly" (medium blue), "crowd" (medium blue), "shout" (medium blue), "true" (medium blue), "cyber" (medium blue), "source" (medium blue), "gross" (medium blue), "negation" (small blue), "opponent" (small blue), "shouting" (small blue), "fighting" (small blue), "world" (small blue), and "law" (small blue).

Annotation pipeline



Pustejovsky and Stubbs (2012),
Natural Language Annotation for Machine Learning

Annotation pipeline



Pustejovsky and Stubbs (2012),
Natural Language Annotation for Machine Learning

Annotation Guidelines

- Our goal: given the constraints of our problem, how can we formalize our description of the annotation process to encourage multiple annotators to provide the same judgment?

Annotation Guidelines

- What is the goal of the project?
- What is each tag called and how is it used? (Be specific: provide examples, and discuss gray areas.)
- What parts of the text do you want annotated, and what should be left alone?
- How will the annotation be created? (For example, explain which tags or documents to annotate first, how to use the annotation tools, etc.)

Practicalities

- Annotation takes time, concentration (can't do it 8 hours a day)
- Annotators get better as they annotate (earlier annotations not as good as later ones)

Why not do it yourself?

- Expensive/time-consuming
- Multiple people provide a measure of consistency: is the task well enough defined?
- Low agreement = not enough training, guidelines not well enough defined, task is bad

Adjudication

- Adjudication is the process of deciding on a single annotation for a piece of text, using information about the **independent annotations**.
- Can be as time-consuming (or more so) as a primary annotation.
- Does not need to be identical with a primary annotation (both annotators can be wrong by chance)

Interannotator agreement



annotator B

annotator A

puppy	fried chicken	
puppy	6	3
fried chicken	2	5

observed agreement = 11/16 = 68.75%

Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

		annotator A	
		puppy	fried chicken
annotator B	puppy	7	4
	fried chicken	8	81

Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

		annotator A	
		puppy	fried chicken
annotator B	puppy	7	4
	fried chicken	8	81

Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$\kappa = \frac{0.88 - p_e}{1 - p_e}$$

		annotator A	
		puppy	fried chicken
annotator B	puppy	7	4
	fried chicken	8	81

Cohen's kappa

- Expected probability of agreement is how often we would expect two annotators to agree assuming **independent** annotations

$$p_e = P(A = \text{puppy}, B = \text{puppy}) + P(A = \text{chicken}, B = \text{chicken})$$

Cohen's kappa

- Expected probability of agreement is how often we would expect two annotators to agree assuming **independent** annotations

$$\begin{aligned} p_e &= P(A = \text{puppy}, B = \text{puppy}) + P(A = \text{chicken}, B = \text{chicken}) \\ &= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken}) \end{aligned}$$

Cohen's kappa

$$= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken})$$

$$P(A=\text{puppy}) \quad 15/100 = 0.15$$

$$P(B=\text{puppy}) \quad 11/100 = 0.11$$

$$P(A=\text{chicken}) \quad 85/100 = 0.85$$

$$P(B=\text{chicken}) \quad 89/100 = 0.89$$

$$= 0.15 \times 0.11 + 0.85 \times 0.89$$

$$= 0.773$$

annotator B

		annotator A	
		puppy	fried chicken
puppy	puppy	7	4
	fried chicken	8	81

Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$\kappa = \frac{0.88 - p_e}{1 - p_e}$$

$$\kappa = \frac{0.88 - 0.773}{1 - 0.773}$$

$$= 0.471$$

		annotator A	
		puppy	fried chicken
annotator B	puppy	7	4
	fried chicken	8	81

Cohen's kappa

- “Good” values are subject to interpretation, but rule of thumb:

0.80-1.00

Very good agreement

0.60-0.80

Good agreement

0.40-0.60

Moderate agreement

0.20-0.40

Fair agreement

< 0.20

Poor agreement

annotator A

annotator B

	puppy	fried chicken
puppy	0	0
fried chicken	0	100

annotator A

annotator B

	puppy	fried chicken
puppy	50	0
fried chicken	0	50

annotator A

annotator B

	puppy	fried chicken
puppy	0	50
fried chicken	50	0

Interannotator agreement

- Cohen's kappa can be used for any number of classes.
- Still requires **two** annotators who evaluate the same items.
- Fleiss' kappa generalizes to **multiple** annotators, each of whom may evaluate **different** items (e.g., crowdsourcing)

Fleiss' kappa

- Same fundamental idea of measuring the observed agreement compared to the agreement we would expect by chance.
- With $N > 2$, we calculate agreement among **pairs** of annotators

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Fleiss' kappa

Number of annotators who assign category j to item i n_{ij}

For item i with n annotations, how many annotators agree, among all $n(n-1)$ possible pairs

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^K n_{ij}(n_{ij} - 1)$$

Fleiss' kappa

For item i with n annotations, how many annotators agree, among all $n(n-1)$ possible pairs

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^K n_{ij}(n_{ij} - 1)$$

Annotator			
A	B	C	D
+	+	+	-

agreeing pairs
of annotators →

A-B
B-A
A-C
C-A
B-C
C-B

Label	n_{ij}
+	3
-	1

$$P_i = \frac{1}{4(3)}(3(2) + 1(0))$$

Fleiss' kappa

Average agreement among all items

$$P_o = \frac{1}{N} \sum_{i=1}^N P_i$$

Probability of category j

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}$$

Expected agreement by chance —
joint probability two raters pick the
same label is the product of their
independent probabilities of picking
that label

$$P_e = \sum_{j=1}^K p_j^2$$

Krippendorff's alpha

- Formulated as $\alpha = 1 - \frac{D_o}{D_e}$ where D_o and D_e denote the observed and expected disagreement, respectively
- Works with all kinds of annotations:
 - Numeric
 - Rankings (Ordinal)
 - Classifications (Nominal)
- The definitions of D_o and D_e change for each data

Annotator bias correction

- Dawid, A. P. and Skene, A. M. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm," Journal of the Royal Statistical Society, 28(1):20–28, 1979.
- Weibe et al. (1999), "Development and use of a gold-standard data set for subjectivity classifications," ACL (for sentiment)
- Carpenter (2010), "Multilevel Bayesian Models of Categorical Data Annotation"
- Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. EMNLP 2008
- Sheng et al. (2008), "Get another label? improving data quality and data mining using multiple, noisy labelers", KDD.
- Raykar et al. (2009), "Supervised learning from multiple experts: whom to trust when everyone lies a bit," ICML
- Hovy et al. (2013), "Learning Whom to Trust with MACE," NAACL

Annotator bias correction

		annotator label			
		positive	negative	mixed	unknown
truth	positive	0.95	0	0.03	0.02
	negative	0	0.80	0.10	0.10
	mixed	0.20	0.05	0.50	0.25
	unknown	0.15	0.10	0.10	0.70

$P(\text{label} \mid \text{truth})$

confusion matrix for a single annotator (David)

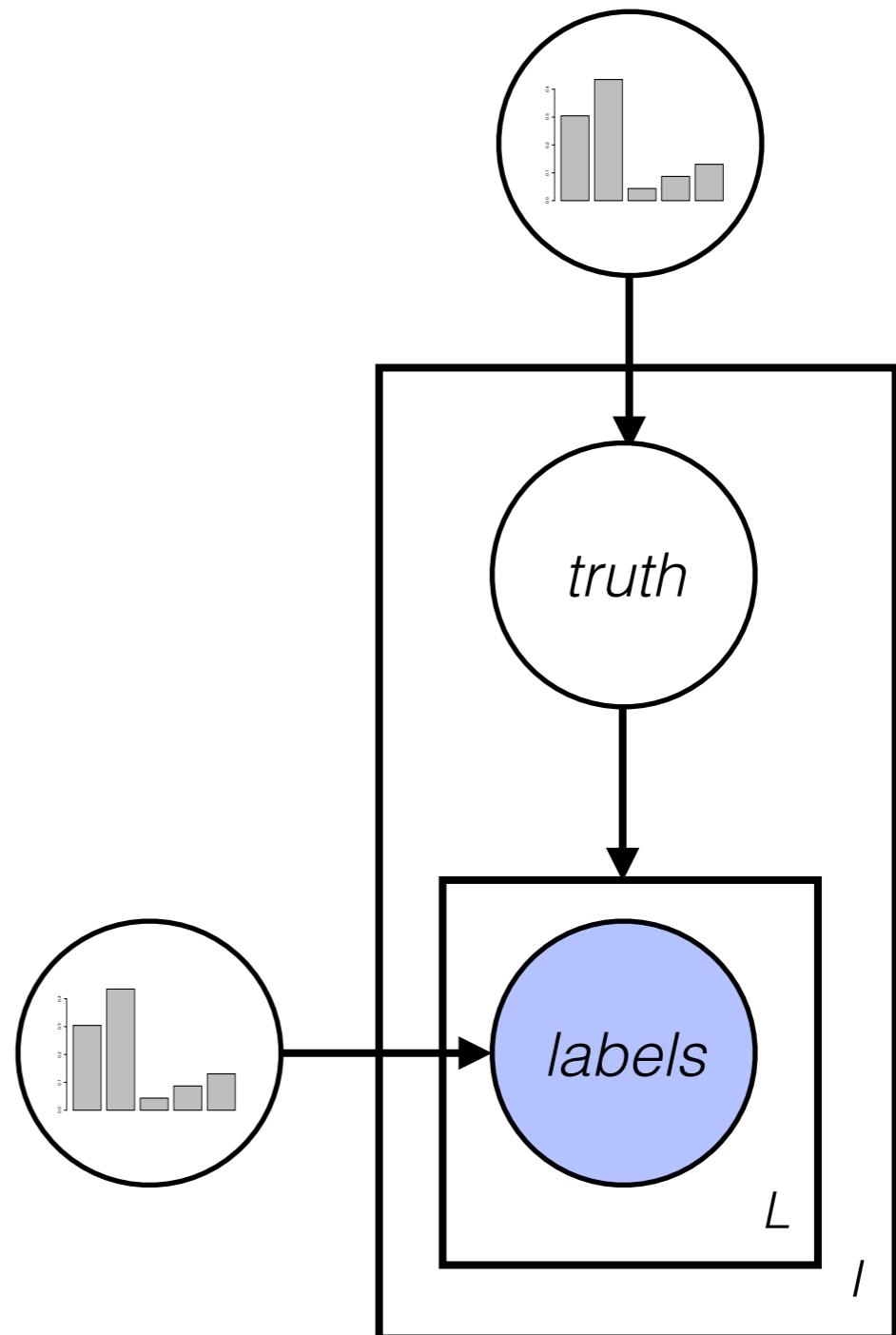
Annotator bias correction

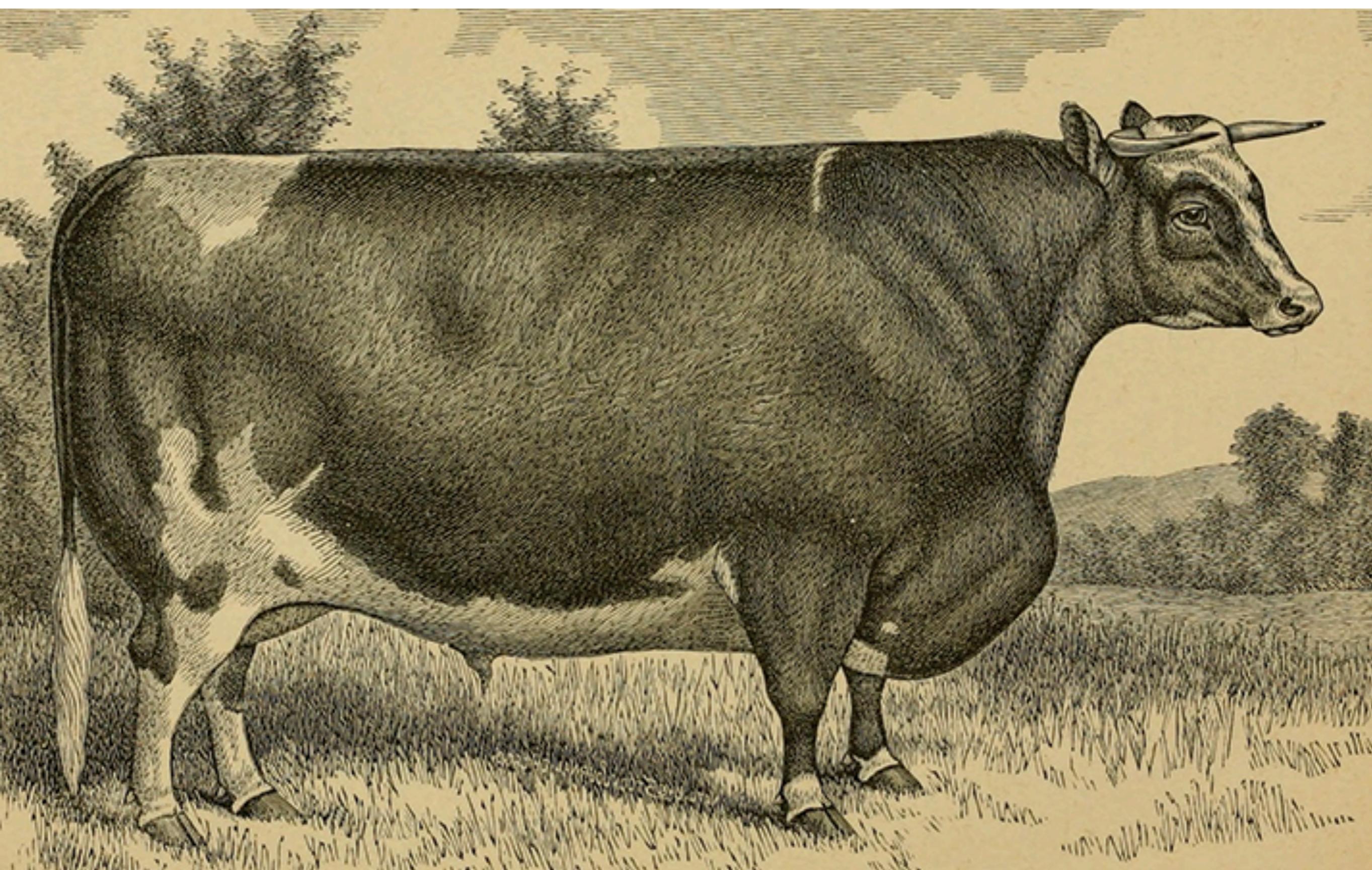
Annotator bias
correction

Dawid and Skene 1979

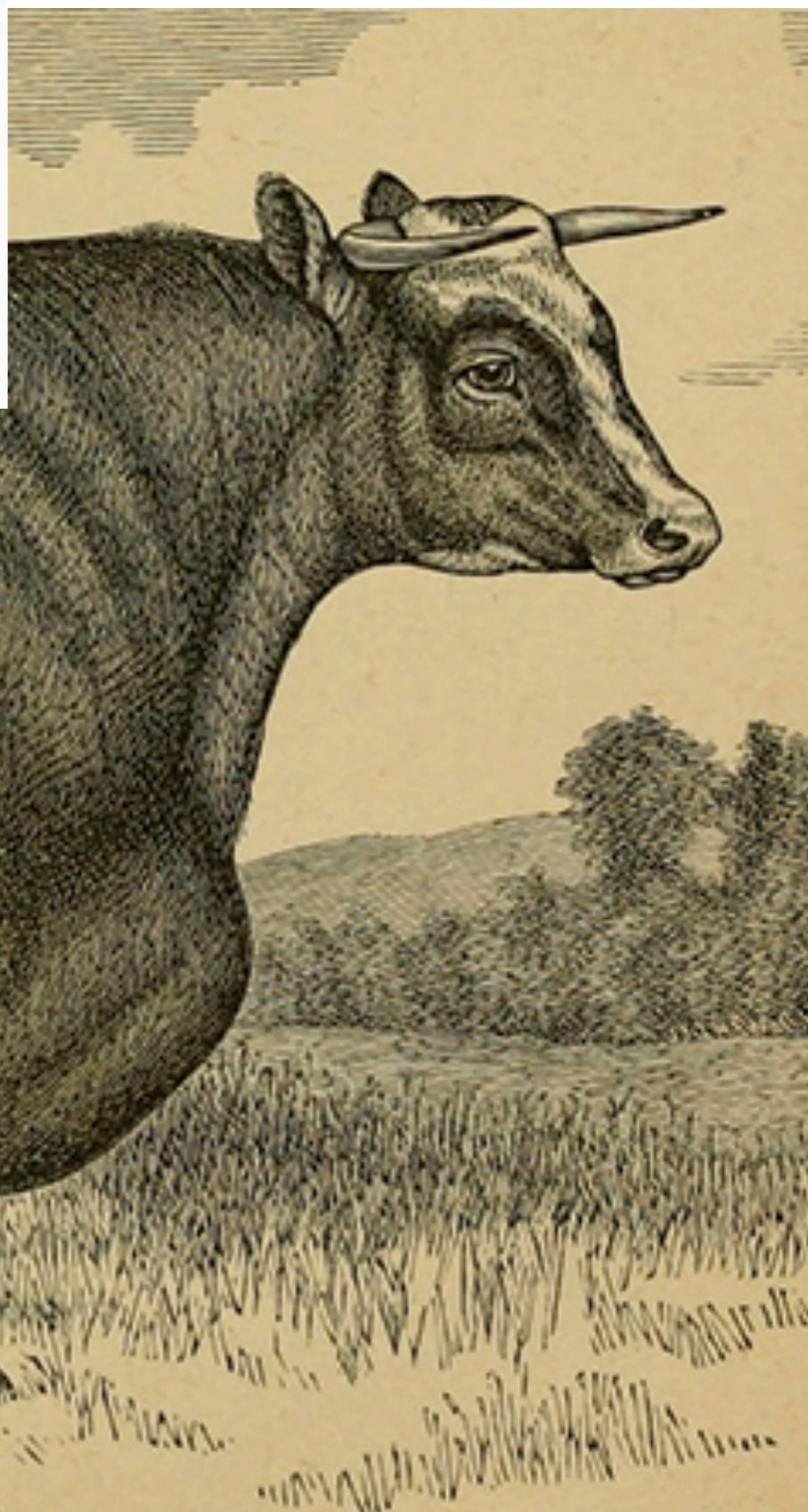
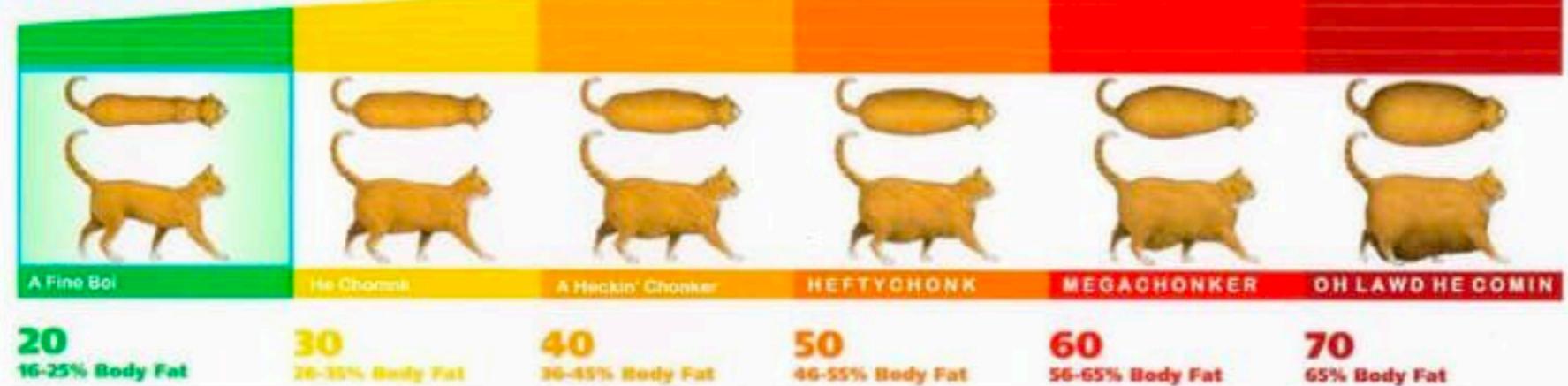
Basic idea: the true label is unobserved; what we observe are noisy judgments by annotators

annotator confusion matrix
 $P(\text{label} \mid \text{truth})$





CHONK Chart



Crowdsourcing

Crowdsourcing

Crowdsourcing

- Basic idea: Pay a small amount of money for random people to complete a small task
 - Can train annotators or require certain skills

Crowdsourcing

- Basic idea: Pay a small amount of money for random people to complete a small task
 - Can train annotators or require certain skills
- Many people work on crowdsourcing platforms, which enables massive parallel work

Crowdsourcing

- Basic idea: Pay a small amount of money for random people to complete a small task
 - Can train annotators or require certain skills
- Many people work on crowdsourcing platforms, which enables massive parallel work
- Amazon Mechanical Turk is by far the most common one used

[Home](#)[Create](#)[Manage](#)[Developer](#)[Help](#)[New Project](#)

New Batch with an Existing Project

Create HITs individually

Start a New Project

Categorization

[Data Collection](#)[Moderation of an Image](#)[Sentiment](#)[Survey](#)[Survey Link](#)[Tagging of an Image](#)[Transcription from A/V](#)[Transcription from an Image](#)[Writing](#)[Other](#)

Example of Categorization

Choose the best category for this image



- kitchen
- living
- bath
- bed
- outside

[View Instructions ↓](#)

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

You must ACCEPT the HIT before you can submit the results.

[Create Project ▾](#)

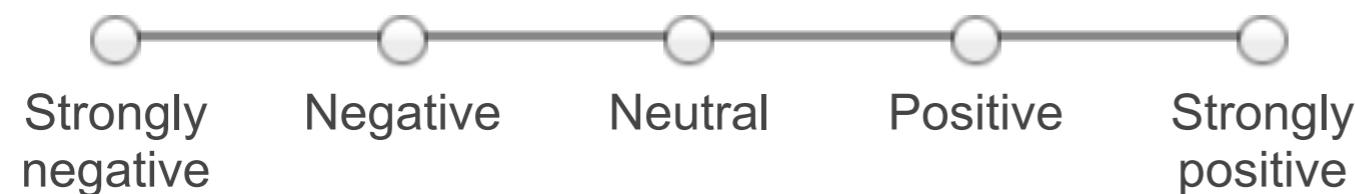
Sentiment

Pick the best sentiment based on the following criterion.

Strongly positive	Select this if the item embodies emotion that was extremely happy or excited toward the topic. For example, "Their customer service is the best that I've seen!!!!"
Positive	Select this if the item embodies emotion that was generally happy or satisfied, but the emotion wasn't extreme. For example, "Sure I'll shop there again."
Neutral	Select this if the item does not embody much of positive or negative emotion toward the topic. For example, "Yeah, I guess it's ok." or "Is their customer service open 24x7?"
Negative	Select this if the item embodies emotion that is perceived to be angry or upsetting toward the topic, but not to the extreme. For example, "I don't know if I'll shop there again because I don't trust them."
Strongly negative	Select this if the item embodies negative emotion toward the topic that can be perceived as extreme. For example, "These guys are terrific... NOTTTT!!!!!!" or "I will NEVER shop there again!!!"

Judge the sentiment expressed by the following item toward: Amazon

If you loved Firefly TV show, amazing Amazon price for entire series: about \$27 BlueRay & \$17 DVD.



Data Collection

Find the Website Address for this Restaurant

- For this restaurant below, enter the website address for the official website of the restaurant
- Include the full address, e.g. <http://www.thecheesecakefactory.com>
- Do not include URLs to city guides and listings like Citysearch.

Restaurant Name: **Olive Garden**

Address: **310 Strander Blvd Tukwila, WA 98188**

Phone Number: **(206) 241-4899**

Website Address:

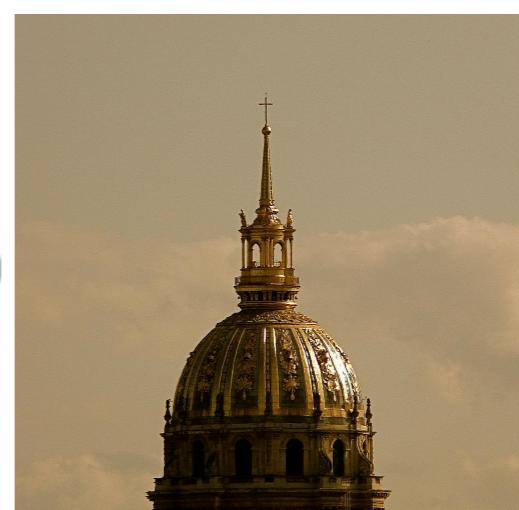
You must ACCEPT the HIT before you can submit the results.

Content Moderation

Select all images containing adult content

Guidelines for flagging an image as adult content. Flag the image if you consider any of the following to be true.

- Does the image contain nudity?
- Does the image portray hate or hate crimes?
- Does the image contain bloody violence?
- Does the image contain offensive gestures?



Surveys

1. What is your gender?

Male Female

2. What is your age?

3. Which of the following best describes your highest achieved education level?

Some High School

4. What is the total income of your household?

Less than \$12,500
\$12,500 - \$24,999
\$25,000 - \$37,499
\$37,500 - \$49,999

5. What is your favorite type of food?

Italian

Transcription

Transcribe the text contained in the image

- Look at the receipt and copy the number of items purchased.
- Provide the dollar amount for tax.
- Provide the dollar amount for the total sale.
- Do not use dollar signs (\$) but make sure you have two decimal points. (ie 4.35)

Image:



Number of items on receipt

Tax amount

Total spent on all items

Writing

Write a brief description of a website.

- Write short article summarizing what a website is about and their products and services.
- Click the link below to review the website and browse the products and services.
- Your submission must be at least 50 words long but no more than 100 words.
- No award will be given for submissions of less than 50 words.
- Your writing must be original and can not simply be a copy of part of the website.

Website name: **The Website Name Here**

Website link: **<http://www.linktowebsitewhere.com>**

Select a customizable template to start a new project

Survey

Survey Link

Survey

Vision

Image Classification

Bounding Box

Semantic Segmentation

Instance Segmentation

Polygon

Keypoint

Image Contains

Video Classification

Moderation of an Image

Image Tagging

Image Summarization

Language

Sentiment Analysis

Intent Detection

Collect Utterance

Emotion Detection

Semantic Similarity

Audio Transcription

Conversation Relevance

Document Classification

Translation Quality

Audio Naturalness

Other

What sentiment does this text convey?

Everything is wonderful!

Select an option

Positive	1
Negative	2
Neutral	3
N/A	4

Submit

Create Project

- Reward per assignment can't be blank

1 Enter Properties 2 Design Layout 3 Preview and Finish

Project Name: Sentiment Analysis This name is not displayed to Workers.

Describe your task to Workers

Title Sentiment analysis
Describe the task to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.

Description Sentiment analysis
Give more detail about this task. This gives Workers a bit more information before they decide to view your task.

Keywords sentiment, text
Provide keywords that will help Workers search for your tasks.

Setting up your task

Reward per assignment \$
This is how much a Worker will be paid for completing an assignment. Consider how long it will take a Worker to complete each assignment.

Number of assignments per task 3
How many unique Workers do you want to work on each task?

Time allotted per assignment 1 Hours
Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.

Task expires in 7 Days
Maximum time your task will be available to Workers on Mechanical Turk.

① Enter Properties

② Design Layout

③ Preview and Finish

```
15      header= What sentiment does this text convey?  
16    >  
17  
18    <classification-target>  
19      <!-- The text you want classified will be substituted for the "text" variable when you<br/>20          publish a batch with a CSV input file containing multiple text items -->  
21      ${text}  
22    </classification-target>  
23  
24    <!-- Use the short-instructions section for quick instructions that the Worker  
25        will see while working on the task. Including some basic examples of  
26        good and bad answers here can help get good results. You can include  
27        any HTML here. -->  
28    <short-instructions>  
29      Choose the primary sentiment that is expressed by the text.  
30    </short-instructions>  
31  
32    <!-- Use the full-instructions section for more detailed instructions that the  
33        Worker can open while working on the task. Including more detailed  
34        instructions and additional examples of good and bad answers here can  
35        help get good results. You can include any HTML here. -->  
36    <full-instructions header="Sentiment Analysis Instructions">  
37      <p><strong>Positive</strong> sentiment include: joy, excitement, delight</p>  
38      <p><strong>Negative</strong> sentiment include: anger, sarcasm, anxiety</p>  
39      <p><strong>Neutral</strong>: neither positive or negative, such as stating a fact</p>  
40      <p><strong>N/A</strong>: when the text cannot be understood</p>  
41      <p>When the sentiment is mixed, such as both joy and sadness, use your judgment to choose th  
42    </full-instructions>  
43  
44  </crowd-classifier>
```

① Enter Properties

② Design Layout

③ Preview and Finish

```
15     header= "What sentiment does this text convey?  
16     >  
17  
18     <classification-target>  
19         <!-- The text you want classified will be substituted for the "text" variable when you<br/>20             publish a batch with a CSV input file containing multiple text items -->  
21             ${text}  
22     </classification-target>  
23  
24     <!-- Use the short-instructions section for quick instructions that the Worker  
25         will see while working on the task. Including some basic examples of  
26         good and bad answers here can help get good results. You can include  
27         any HTML here. -->  
28     <short-instructions>  
29         Choose the primary sentiment that is expressed by the text.  
30     </short-instructions>  
31  
32     <!-- Use the full-instructions section for more detailed instructions that the  
33         Worker can open while working on the task. Including more detailed  
34         instructions and additional examples of good and bad answers here can  
35         help get good results. You can include any HTML here. -->  
36     <full-instructions header="Sentiment Analysis Instructions">  
37         <p><strong>Positive</strong> sentiment include: joy, excitement, delight</p>  
38         <p><strong>Negative</strong> sentiment include: anger, sarcasm, anxiety</p>  
39         <p><strong>Neutral</strong>: neither positive or negative, such as stating a fact</p>  
40         <p><strong>N/A</strong>: when the text cannot be understood</p>  
41         <p>When the sentiment is mixed, such as both joy and sadness, use your judgment to choose th  
42     </full-instructions>  
43  
44     </crowd-classifier>
```

Sentiment analysis

Requester: David Jurgens**Reward:** \$0.05 per task**Tasks available:** 0**Duration:** 1 Hours**Qualifications Required:** None**Previewing Answers Submitted by Workers**

This message is only visible to you and will not be shown to Workers.

You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Instructions[View full instructions](#)[View tool guide](#)

Choose the primary sentiment that is expressed by the text.

What sentiment does this text convey?

\${text}

Select an option

Positive	1
Negative	2
Neutral	3
N/A	4

Finish

[Results](#) [Workers](#) [Qualification Types](#)

Items Completed 40 of 1000

Negative 2%

Neutral 2%

Positive 0%

Incomplete 96%

[Download Results](#)[Add Time](#)[Cancel](#)**Sentiment Project****Answer Summary**[Results](#)[Cost](#)**Details**

Project Obama sentiment 9/13

Status in Progress

Question

Judge the sentiment expressed by the following item toward: President Obama

Created 09/08/13 13:10

Time elapsed 26 minutes

Est. completion 09/08/13 14:35

Expiration time 09/12/13 13:12

Worker time limit 1 hour

Effective hourly rate \$4.000

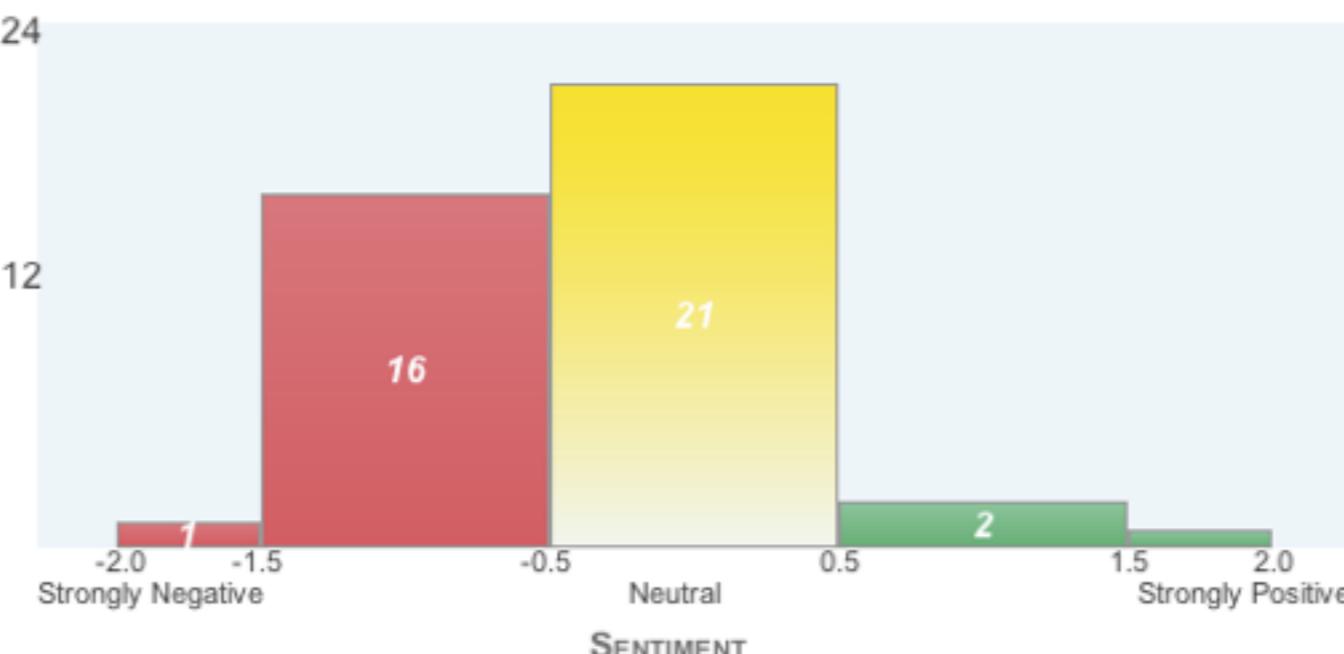
Number of Workers 5

Input file tweets.csv

40 of your 1000 items are done. We are waiting for Workers to complete the remaining 960 items.

The chart below shows you the distribution of the mean sentiment ratings ([How is it calculated?](#)) for the 1000 items in this batch. Click on each bar to view those items on the Results page.

You can also download the results in a .csv file.

**Summary for the batch**

This batch has a mean sentiment rating of **-0.41**. Of the 40 items, 2 items have a positive rating and 17 items have a negative rating.

Building your own crowdsourcing task (HIT)

Building your own crowdsourcing task (HIT)

- Set the parameters of your HIT

Building your own crowdsourcing task (HIT)

- Set the parameters of your HIT
- Optionally, specify requirements for which Turkers can complete your HIT

Building your own crowdsourcing task (HIT)

- Set the parameters of your HIT
- Optionally, specify requirements for which Turkers can complete your HIT
- Design an HTML template with \${variables}

Building your own crowdsourcing task (HIT)

- Set the parameters of your HIT
- Optionally, specify requirements for which Turkers can complete your HIT
- Design an HTML template with \${variables}
- Upload a CSV file to populate the variables

Building your own crowdsourcing task (HIT)

- Set the parameters of your HIT
- Optionally, specify requirements for which Turkers can complete your HIT
- Design an HTML template with \${variables}
- Upload a CSV file to populate the variables
- Pre-pay Amazon for the work

Building your own crowdsourcing task (HIT)

- Set the parameters of your HIT
- Optionally, specify requirements for which Turkers can complete your HIT
- Design an HTML template with \${variables}
- Upload a CSV file to populate the variables
- Pre-pay Amazon for the work
- Approve/reject work from Turkers

Building your own crowdsourcing task (HIT)

- Set the parameters of your HIT
- Optionally, specify requirements for which Turkers can complete your HIT
- Design an HTML template with \${variables}
- Upload a CSV file to populate the variables
- Pre-pay Amazon for the work
- Approve/reject work from Turkers
- Analyze results

HIT Parameters

- title
- description
- keywords
- reward amount
- max time allotted for work
- auto approval time

1 Enter Properties

2 Design Layout

3 Preview and Finish

Project Name: **Reading Comprehension Test**

This name is not displayed to Workers.

Describe your HIT to Workers

Title

Write questions and answers for a reading comprehension test

Describe the task to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.

Description

Read an article and then write 5 questions and 5 example answers that test a reader's comprehension

Give more detail about this task. This gives Workers a bit more information before they decide to view your HIT.

Keywords

reading, comprehension, writing, questions, answers, research

Provide keywords that will help Workers search for your HITs.

This project may contain potentially explicit or offensive content, for example, nudity. [\(See details\)](#)

Setting up your HIT

Reward per assignment

\$ 0.25

Tip: Consider how long it will take a Worker to complete each task. A 30 second task that pays \$0.05 is a \$6.00 hourly wage.

Number of assignments per HIT

10

How many unique Workers do you want to work on each HIT?

Time allotted per assignment

2

Hours

Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.

HIT expires in

14

Days

Maximum time your HIT will be available to Workers on Mechanical Turk.

Results are automatically approved in

7

Days

After this time, all unreviewed work is approved and Workers are paid.

The Purpose of Redundancy

The Purpose of Redundancy

- MTurk lets you set the number of assignments per HIT

The Purpose of Redundancy

- MTurk lets you set the number of assignments per HIT
- That gives you different (redundant) answers from different Turkers

The Purpose of Redundancy

- MTurk lets you set the number of assignments per HIT
- That gives you different (redundant) answers from different Turkers
- This lets you conduct surveys (num assignments = num respondents)

The Purpose of Redundancy

- MTurk lets you set the number of assignments per HIT
- That gives you different (redundant) answers from different Turkers
- This lets you conduct surveys (num assignments = num respondents)
- Also, lets you take votes and do tie-breaking, or do quality control

The Purpose of Redundancy

- MTurk lets you set the number of assignments per HIT
- That gives you different (redundant) answers from different Turkers
- This lets you conduct surveys (num assignments = num respondents)
- Also, lets you take votes and do tie-breaking, or do quality control
- Redundancy $\geq 10x$ incurs higher fees on MTurk

The art of designing a
good HIT

Crowdsourcing works for
tasks that are...

Crowdsourcing works for tasks that are...

- Natural and easy to explain to non-experts

Crowdsourcing works for tasks that are...

- Natural and easy to explain to non-experts
- Decomposable into simpler tasks that can be joined together

Crowdsourcing works for tasks that are...

- Natural and easy to explain to non-experts
- Decomposable into simpler tasks that can be joined together
- Parallelizable into small, quickly completed chunks

Crowdsourcing works for tasks that are...

- Natural and easy to explain to non-experts
- Decomposable into simpler tasks that can be joined together
- Parallelizable into small, quickly completed chunks
- Well-suited to quality control (some data has correct gold standard annotations)

Crowdsourcing works for
tasks that are...

Crowdsourcing works for tasks that are...

- Robust to some amount of noise/errors (the downstream task is training a statistical model)

Crowdsourcing works for tasks that are...

- Robust to some amount of noise/errors (the downstream task is training a statistical model)
- Balanced and each task contains the same amount of work

Crowdsourcing works for tasks that are...

- Robust to some amount of noise/errors (the downstream task is training a statistical model)
- Balanced and each task contains the same amount of work
 - Don't have tons of work in one assignment but not another

Crowdsourcing works for tasks that are...

- Robust to some amount of noise/errors (the downstream task is training a statistical model)
- Balanced and each task contains the same amount of work
 - Don't have tons of work in one assignment but not another
 - Don't ask Turkers to annotate something occurs in the data <<10% of the time

Guidelines for your own tasks

- Simple instructions are required
- If your task can't be expressed in one paragraph + bullets, then it may need to be broken into simpler sub-tasks
- Include examples, but remember workers are on the clock when they read instructions

Guidelines for your own tasks

Guidelines for your own tasks

- Quality control is paramount
 - Measuring redundancy doesn't work if people answer incorrectly in systematic ways
 - Embed gold standard data as controls
 - Some platforms like FigureEight will embed these controls for you!

Guidelines for your own tasks

- Quality control is paramount
 - Measuring redundancy doesn't work if people answer incorrectly in systematic ways
 - Embed gold standard data as controls
 - Some platforms like FigureEight will embed these controls for you!
- Qualification tests v. no qualification test
 - Reduce participation, but usually ensures higher quality

Guidelines for your own tasks

- Remember that MTurk has a reputation system
- Rejecting Turkers has consequences for them, blocking Turkers is likely to get them expelled from MTurk
- In one example, Chris Callison-Burch rejects all work performing near chance, accept all work doing ~80%, reject proportionally to performance on controls for in-between

Guidelines for your own tasks

- Reputation goes both ways
- If using MTurk, monitor your profile on Turker Nation and TurkOpticon
- Pay your workers generously
- Pay quickly, and be responsive to email questions
 - Easier if using platforms like FigureEight (formerly CrowdFlower) that auto-pay
 - Still need to address challenges to gold standard then!

Why do people participate on crowdsourcing platforms?

- How can we motivate people to participate?
- Even with a low barrier to entry (anyone with a computer can contribute) we still need to make a case **why** they should contribute.

Motivation: Pay

- Easiest way to recruit workers.
- Downside: provides incentive to cheat
- Problem might be exacerbated when the crowd workers are anonymous
- MTurk uses micropayments
- Online temping services provide higher wages:
LiveOps, ODesk, etc
- FigureEight tried non-monetary payments (virtual goods and currencies, SwagBucks)

Motivation Altruism

- People want to do good
- When Jim Gray went missing, volunteers searched 500k satellite images
- After the Haitian earthquake, diaspora translated 1000 messages per day



Motivation: Reputation

- Sometimes people will contribute in order to build their profile within a community
 - Example: stackoverflow

Users

reputation

new us

Type to find users:

[Gordon Linoff](#)

New York, United States

1,260

sql, mysql, sql-server

[Sotirios Delimanolis](#)**1,020**

java, spring, spring-mvc

[Arun P Johny](#)

Bangalore, India

960

jquery, javascript, html

[Jon Skeet](#)

Reading, United Kingdom

915

c#, java, .net

[Hans Passant](#)

Madison, WI

905

c#, .net, winforms

[Martijn Pieters](#)

Cambridge, United Kingdom

880

python, python-2.7, list

[alecx](#)

Russia

870

python, django, scrapy

[BalusC](#)

Willemstad, Curaçao

855

java, jsf, jsf-2

[dasblinkenlight](#)

United States

825

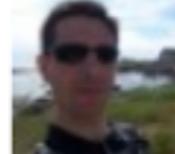
c#, c++, java

[falsetru](#)

Seoul, South Korea

800

python, regex, ruby

[VonC](#)

France

780

git, eclipse, java

[arshajii](#)

Boston, MA

777

java, python, string

Motivation: Enjoyment

- Games with a purpose is a strategy to try to make a task fun
- In the ESP game two players look at an image and try to guess what words the other is thinking
- In doing so they label images on the web



Luis Von Ahn
==
Tom Sawyer



Tom Sawyer (Whitewashing the Fence), 1936

Motivation: Implicit Work

- Make people do work for you as some part of another task

The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.

morning

morning overtook

Type the two words:



reCAPTCHA™
stop spam.
read books.

Motivation

- Pay
- Altruism
- Reputation
- Enjoyment
- Implicit Work
- Can you think of others?

Quality Control

- Even if people are motivated to participate, how do we know that they are doing work conscientiously?
- Can we trust them not to cheat or sabotage the system?
- Even if they are acting in good faith, how do we know that they're doing things right?

Quality Control: Reputation Check

- Mechanical Turk uses a reputation system
- When a Turker submits poor work, Requesters reject it
- The Turker's approval rate is displayed to all other Requesters

Quality Control: Agreement and Redundancy

- The ESP game uses the labels that two players independently agree on
- Similar technique is often used in MTurk, when each item is done independently
- Redundancy allows a voting on ambiguous answers / opinions
- It is also helpful for identifying workers who are consistently divergent

Quality Control: Gold Standard

- In MTurk we commonly mix in questions with a known answer alongside new questions
 - Some platforms like CrowdFlower include these automatically
- This is similar to agreement, but now we check agreement against experts or trusted workers
- For multiple choice questions, gold standard allows for automatic grading

Quality Control: Defensive Task Design

- Try to design tasks so that they are nearly as hard to cheat as they are to complete
- Example from Chris Callison-Burch:
 - For translation HITs, people frequently would paste text into Google Translate
 - Text is converted into images, then people had to transcribe it.

Quality Control: Economic Incentives

- When money is the motivator, it may be possible to use different incentive structures to illicit better results
- Pay people more when they reach a certain level of mastery, or when their output passes second pass reviews
- CastingWords uses bonuses to do this

[All HITs](#) | [HITs Available To You](#) | [HITs Assigned To You](#)Find **HITs** containing **castingwords**that pay at least \$ **0.00** for which you are qualified require Master Qualification**GO****HITs containing 'castingwords'**

1-10 of 47 Results

Sort by: Reward Amount (most first)  **GO!**[Show all details](#) | [Hide all details](#)1 [2](#) [3](#) [4](#) [5](#) > [Next](#) >> [Last](#)

Express Transcription: Frank King X-ray #159143 (195% premium) (avg rwd+bsn: \$7.60) [13:00 mmss]		Not Qualified to work on this HIT (Why?) View a HIT in this group	
Requester:	CastingWords	HIT Expiration Date:	Sep 11, 2013 (56 minutes 15 seconds) Reward: \$3.80
		Time Allotted:	4 hours 30 minutes HITs Available: 1
Express Transcription: Easy #159160 (avg rwd+bsn: \$6.3) [21:00 mmss]		Not Qualified to work on this HIT (Why?) View a HIT in this group	
Requester:	CastingWords	HIT Expiration Date:	Sep 11, 2013 (3 hours 58 minutes) Reward: \$3.15
		Time Allotted:	5 hours 30 minutes HITs Available: 1
Express Transcription: Frank King X-ray #159143 (195% premium) (avg rwd+bsn: \$4.68) [08:00 mmss]		Not Qualified to work on this HIT (Why?) View a HIT in this group	
Requester:	CastingWords	HIT Expiration Date:	Sep 11, 2013 (56 minutes 13 seconds) Reward: \$2.34
		Time Allotted:	4 hours HITs Available: 1
Express Transcription: Easy #159160 (avg rwd+bsn: \$3.9) [13:00 mmss]		Not Qualified to work on this HIT (Why?) View a HIT in this group	
Requester:	CastingWords	HIT Expiration Date:	Sep 11, 2013 (3 hours 58 minutes) Reward: \$1.95
		Time Allotted:	4 hours 30 minutes HITs Available: 1
Express Transcription: Easy #159160 (avg rwd+bsn: \$2.4) [08:00 mmss]		Not Qualified to work on this HIT (Why?) View a HIT in this group	
Requester:	CastingWords	HIT Expiration Date:	Sep 11, 2013 (3 hours 58 minutes) Reward: \$1.20
		Time Allotted:	4 hours HITs Available: 1
Express Transcription: Easy #159160 (avg rwd+bsn: \$1.5) [05:00 mmss]		Not Qualified to work on this HIT (Why?) View a HIT in this group	
Requester:	CastingWords	HIT Expiration Date:	Sep 11, 2013 (3 hours 58 minutes) Reward: \$0.75
		Time Allotted:	4 hours HITs Available: 1
Express Transcription: Frank King X-ray #159143 (195% premium) (avg rwd+bsn: \$1.16) [02:00 mmss]		Not Qualified to work on this HIT (Why?) View a HIT in this group	
Requester:	CastingWords	HIT Expiration Date:	Sep 11, 2013 (56 minutes 9 seconds) Reward: \$0.58
		Time Allotted:	3 hours 30 minutes HITs Available: 1



Annotation Quality

For any dataset, it is crucial to consider the following questions

- Is my data reliable enough for evaluation?
 - Is the annotation outcome reproducible?
- Is there good enough agreement on what is the right answer?
 - How do we measure inter-annotator agreement?

People label incorrectly for many reasons!

- Mistakes
- People get tired
- Adversarial annotators
- Subjective/hard annotation task
- Insufficient training







Annotator Agreement

- Lots of possibilities (to name few)
 - Percentage of times annotators agreed
 - Cohen's Kappa
 - Krippendorff's Alpha

Annotator Agreement

- Lots of possibilities (to name few)
 - Percentage of times annotators agreed
 - Cohen's Kappa
 - Krippendorff's Alpha

Why might percent agreement
be a bad measure?





A B





A B

muffin muffin





A B

muffin muffin



dog dog



A B

muffin muffin



dog dog

muffin muffin



A B

muffin muffin



dog dog

muffin muffin

muffin muffin



A B

muffin muffin



dog dog

muffin muffin

muffin muffin

muffin muffin



A B

muffin muffin



dog dog

muffin muffin

muffin muffin

muffin muffin

muffin muffin



A B

muffin muffin



dog dog

muffin muffin

muffin muffin

muffin muffin

muffin muffin

muffin dog



A B

muffin muffin



dog dog

muffin muffin

muffin muffin

muffin muffin

muffin muffin

muffin dog

muffin muffin



A

B

muffin muffin



dog dog

muffin muffin

muffin muffin

muffin muffin

muffin muffin

muffin dog

muffin muffin

A and B have 88% agreement!



A B C

muffin muffin muffin



dog dog muffin

muffin muffin muffin

muffin muffin muffin

A and B have 88% agreement!

muffin muffin muffin

muffin muffin muffin

muffin dog muffin

muffin muffin muffin



A B C

muffin muffin muffin



dog dog muffin

muffin muffin muffin

muffin muffin muffin

A and B have 88% agreement!

muffin muffin muffin

A and C have 88% agreement!

muffin muffin muffin

muffin dog muffin

muffin muffin muffin



A

B

C

muffin muffin muffin

dog dog muffin

muffin muffin muffin

muffin muffin muffin

A and B have 88% agreement!

A and C have 88% agreement!

How might we control for one class being more frequent?

muffin muffin muffin

muffin muffin muffin

muffin dog muffin

muffin muffin muffin

Controlling for Agreement by chance

	A: Dog	A: Muffin	
B: Dog	1	1	2
B: Muffin	0	6	6
	1	7	8

Controlling for Agreement by chance

Many NLP methods use
Cohen's κ :

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

		A: Dog	A: Muffin
		B: Dog	1
		B: Muffin	0
			2
			6
			8

Controlling for Agreement by chance

Many NLP methods use

Cohen's κ :

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

p_o = Observed percentage
agreement

		A: Dog	A: Muffin
		B: Dog	1
		B: Muffin	0
			1
			7
			8

Controlling for Agreement by chance

Many NLP methods use

Cohen's κ :

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

p_o = Observed percentage agreement

p_e = Expected percentage agreement if annotators were randomly choosing

		A: Dog	A: Muffin
		B: Dog	1
		B: Muffin	0
			2
			6
			8

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

		A: Dog	A: Muffin
B: Dog	1	1	2
B: Muffin	0	6	6
	1	7	8

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

Step 1: Calculate observed agreement:
• $(6+1)/8 = 0.875$

		A: Dog	A: Muffin
		1	1
B: Dog	1	1	2
	0	6	6
B: Muffin	1	7	8

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

Step 1: Calculate observed agreement:

- $(6+1)/8 = 0.875$

Step 2: Calculate prob. raters would randomly both say “dog”

- $(2/8) * (1/8) = 0.03125$

	A: Dog	A: Muffin	
B: Dog	1	1	2
B: Muffin	0	6	6
	1	7	8

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

	A: Dog	A: Muffin	
B: Dog	1	1	2
B: Muffin	0	6	6
	1	7	8

Step 1: Calculate observed agreement:

- $(6+1)/8 = 0.875$

Step 2: Calculate prob. raters would randomly both say “dog”

- $(2/8) * (1/8) = 0.03125$

Step 3: Calculate prob. raters would randomly both say “muffin”

- $(6/8) * (7/8) = 0.65625$

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

	A: Dog	A: Muffin	
B: Dog	1	1	2
B: Muffin	0	6	6
	1	7	8

Step 1: Calculate observed agreement:

- $(6+1)/8 = 0.875$

Step 2: Calculate prob. raters would randomly both say “dog”

- $(2/8) * (1/8) = 0.03125$

Step 3: Calculate prob. raters would randomly both say “muffin”

- $(6/8) * (7/8) = 0.65625$

Step 4: Add steps 2 and 3 to get overall prob. of random agreement

- $0.03 + 0.66 = 0.6875$

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

	A: Dog	A: Muffin	
B: Dog	1	1	2
B: Muffin	0	6	6
	1	7	8

Step 1: Calculate observed agreement:

- $(6+1)/8 = 0.875$

Step 2: Calculate prob. raters would randomly both say “dog”

- $(2/8) * (1/8) = 0.03125$

Step 3: Calculate prob. raters would randomly both say “muffin”

- $(6/8) * (7/8) = 0.65625$

Step 4: Add steps 2 and 3 to get overall prob. of random agreement

- $0.03 + 0.66 = 0.6875$

Step 5: Fill in formula:

- $(0.875 - 0.6875) / (1 - 0.6875) = 0.6$

What if annotators had picked at random?

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

	A: Dog	A: Muffin	
B: Dog	45	45	90
B: Muffin	45	45	90
	90	90	180

What if annotators had picked at random?

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}, \quad p_o = 0.5 \quad p_e = 0.5$$

	A: Dog	A: Muffin	
B: Dog	45	45	90
B: Muffin	45	45	90
	90	90	180

What if annotators had picked at random?

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}, \quad p_o = 0.5 \quad p_e = 0.5$$

	A: Dog	A: Muffin	
B: Dog	45	45	90
B: Muffin	45	45	90
	90	90	180

$$\kappa = \frac{0.5 - 0.5}{1 - 0.5}$$

What if annotators had picked at random?

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}, \quad p_o = 0.5 \quad p_e = 0.5$$

	A: Dog	A: Muffin	
B: Dog	45	45	90
B: Muffin	45	45	90
	90	90	180

$$\kappa = \frac{0.5 - 0.5}{1 - 0.5}$$

$$\kappa = 0$$

What if we have ...

- More than one rater
- Ratings along a scale
- Ratings that are numeric
- Situations where not everyone rated each item

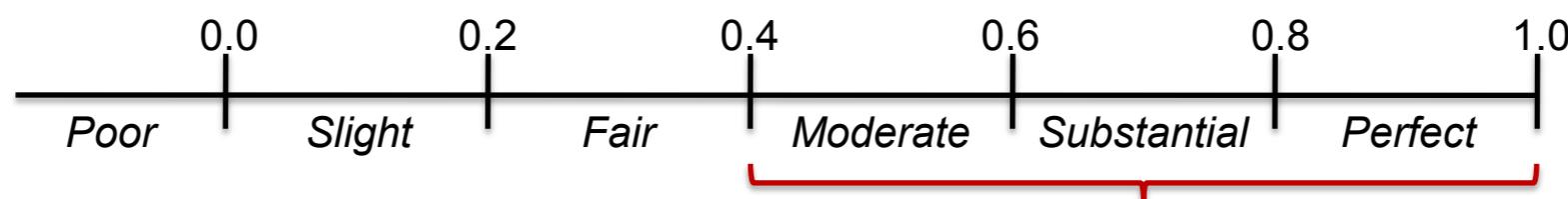
What if we have . . .

- More than one rater
- Ratings along a scale
- Ratings that are numeric
- Situations where not everyone rated each item

Krippendorff's α solves all of these

How much agreement do you need?

- Landis and Koch (1977)



- Krippendorff (1980), Carletta (1996)
 - $0.67 < K < 0.8$ “allowing tentative conclusions to be drawn”
 - above 0.8 “good reliability”
- Krippendorff (2004)
 - “even a cutoff point of 0.8 is a pretty low standard”
- Neuendorf (2002)
 - “reliability coefficients of 0.9 or greater would be acceptable to all, 0.8 or greater [...] in most situations”

How much agreement do you *really* need?

- Aim high in general
- Some tasks are *very hard* or *fundamentally subjective*
 - Microaggressions: $\kappa = 0.464$
 - Connotations: % Agreement = 0.52
 - Social factors of relationships: $\kappa = 0.59$
 - Word sense disambiguation: $\alpha \leq 0.3$

It's possible even to use these agreements to identify useful things

It's possible even to use these agreements to identify useful things

- Idea: identify annotators with low agreement, then downweight their judgments

It's possible even to use these agreements to identify useful things

- Idea: identify annotators with low agreement, then downweight their judgments
- Idea: identify items (e.g., sentences) that annotators all disagree on

It's possible even to use these agreements to identify useful things

- Idea: identify annotators with low agreement, then downweight their judgments
- Idea: identify items (e.g., sentences) that annotators all disagree on
- Idea: identify classes that have low disagreement

It's possible even to use these agreements to identify useful things

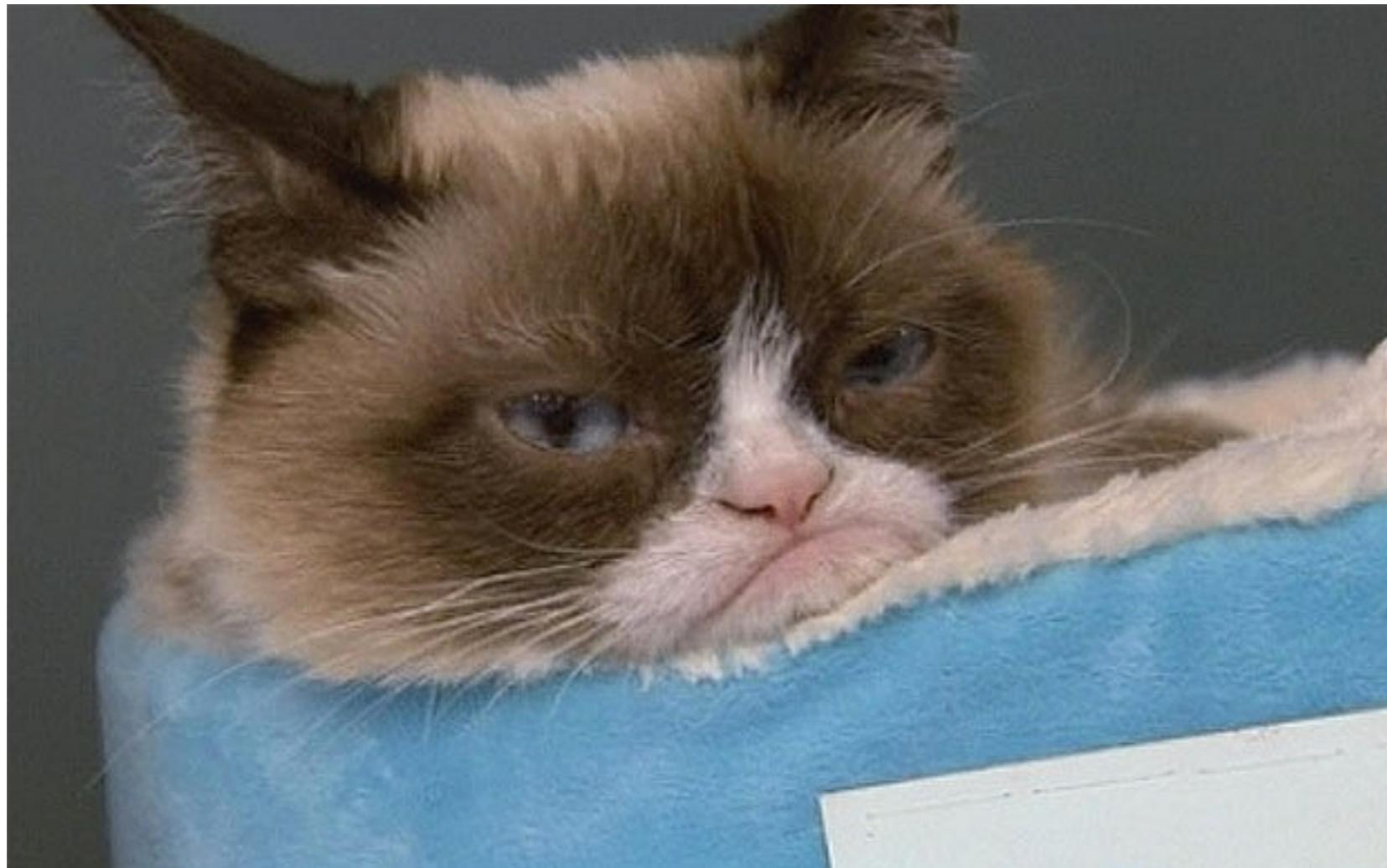
- Idea: identify annotators with low agreement, then downweight their judgments
- Idea: identify items (e.g., sentences) that annotators all disagree on
- Idea: identify classes that have low disagreement
- Many papers have been written on these ideas since the 70s some will be in the weekly readings





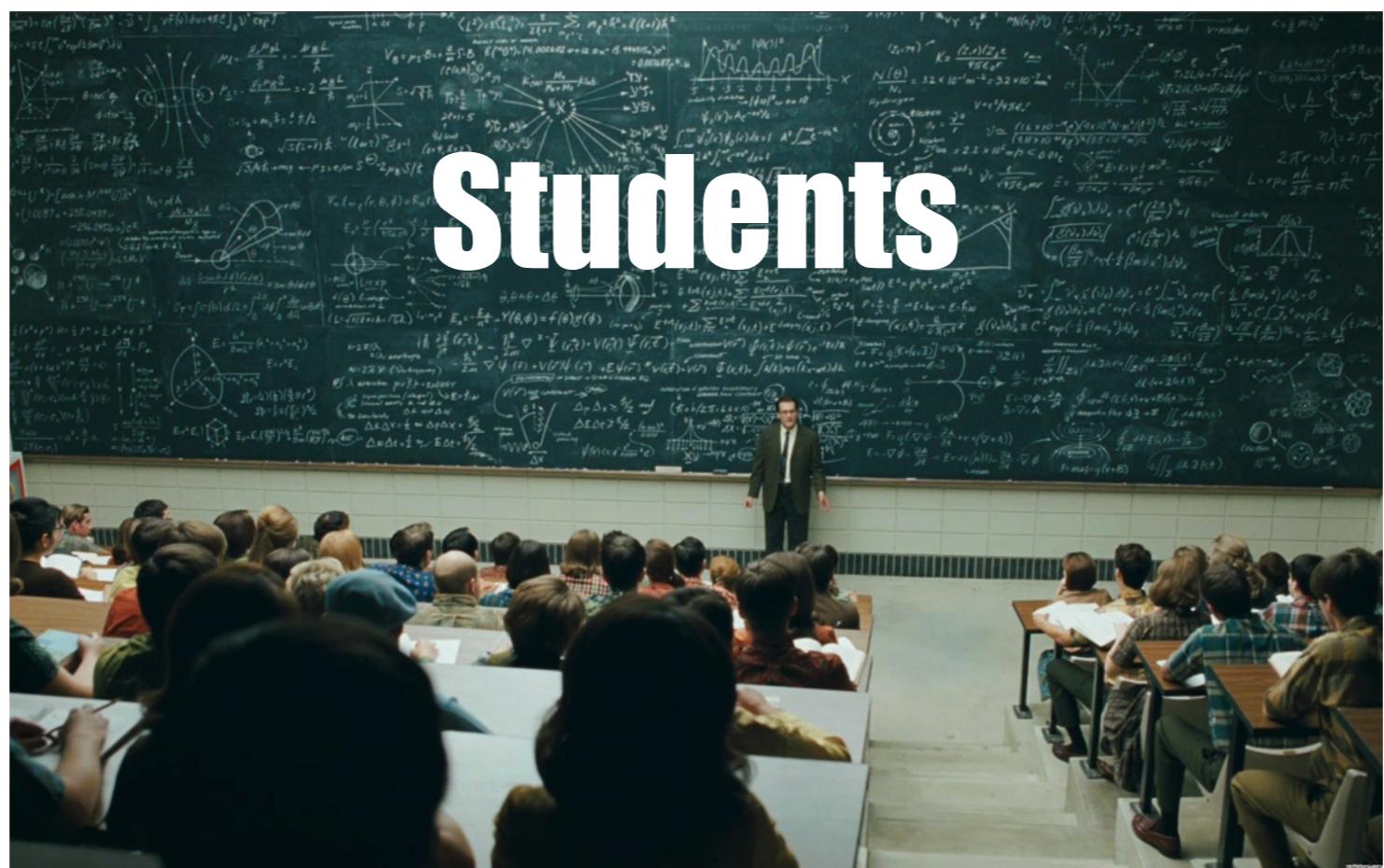
Games with a Purpose

Annotating is no fun.

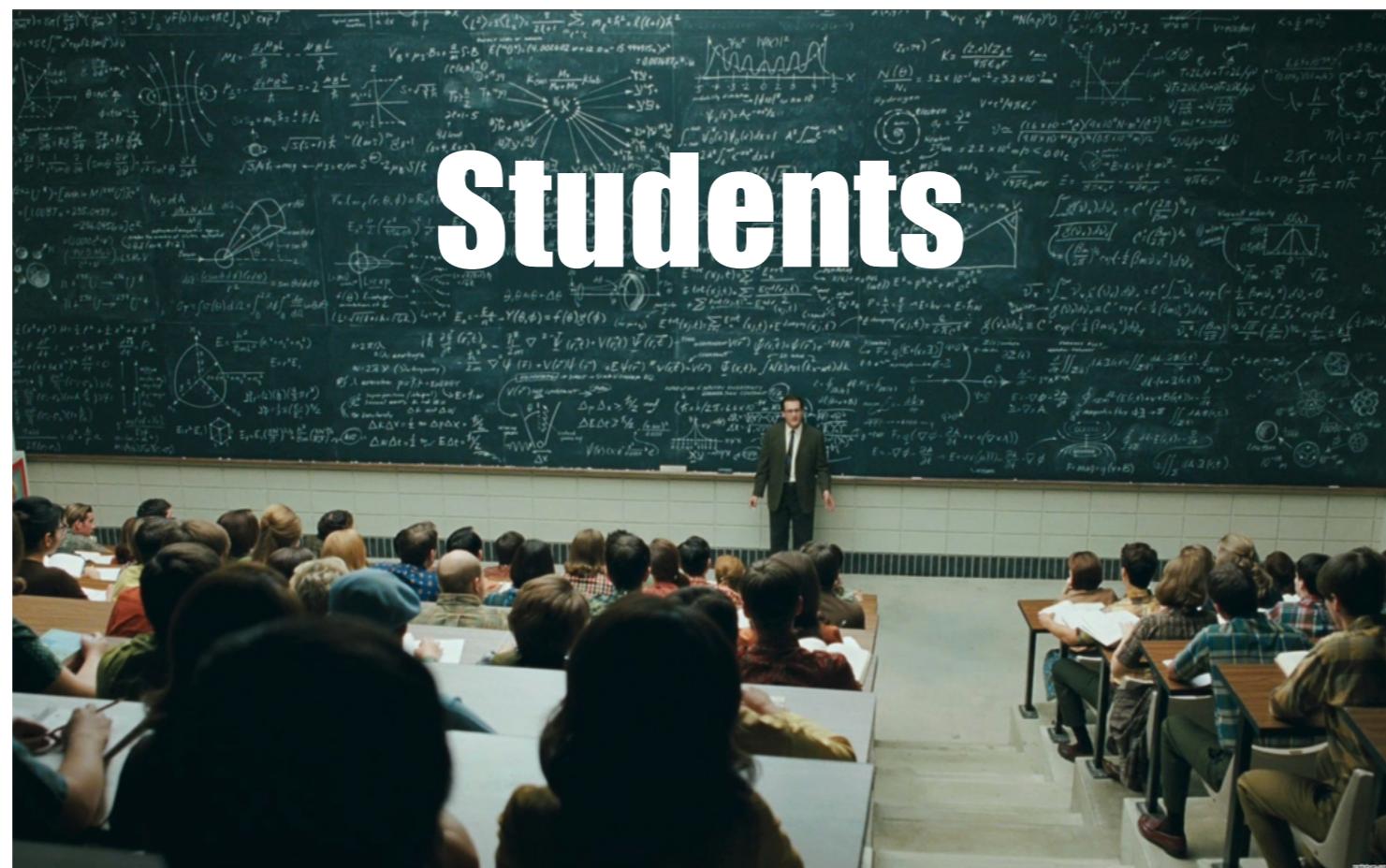


The sprinter won the race

- 1) be the winner in a contest or competition
- 2) win something through one's efforts
- 3) obtain advantages, such as points
- 4) attain success or reach a desired goal



Students



Students



~40K Turkers Active Concurrently × 1 Week
= 6.7M hours of possible MTurk time per week

~40K Turkers Active Concurrently × 1 Week
= 6.7M hours of possible MTurk time per week

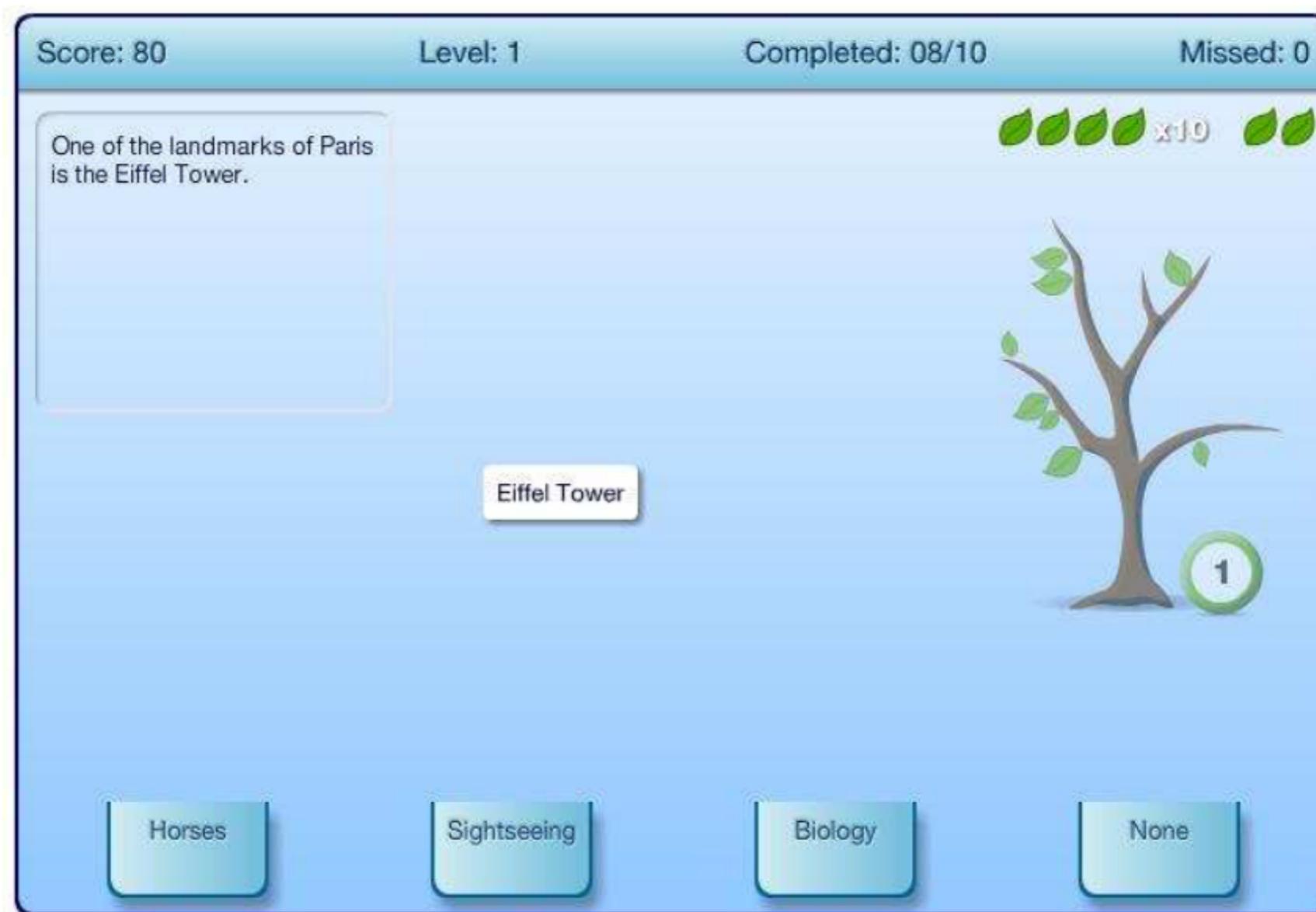
3,000,000,000 hours spent playing video games per week

~40K Turkers Active Concurrently × 1 Week
= 6.7M hours of possible MTurk time per week

3,000,000,000 hours spent playing video games per week



Serious games for NLP



Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012). Quality through flow and immersion: gamifying crowdsourced relevance assessments. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12 (pp. 871–880).

Serious games for NLP

Score: 80 Level: 1 Completed: 08/10 Missed: 0

One of the landmarks of Paris is the Eiffel Tower.

Table 5: Effective annotation cost.

	Conventional	Game-based
Cost per doc	\$0.06	\$0.0004
Whole corpus	\$192	\$1.28
Effective hourly rate	\$1.80	\$0.18

Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012). Quality through flow and immersion: gamifying crowdsourced relevance assessments. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12 (pp. 871–880).

Serious games for NLP

Score: 80 Level: 1 Completed: 08/10 Missed: 0

One of the landmarks of Paris is the Eiffel Tower.

Table 5: Effective annotation cost.

	Conventional	Game-based
Cost per doc	\$0.06	\$0.0004
Whole corpus	\$192	\$1.28
Effective hourly rate	\$1.80	\$0.18

Horses Sightseeing Biology None

Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012). Quality through flow and immersion: gamifying crowdsourced relevance assessments. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12 (pp. 871–880).

Serious games for NLP

Serious games for NLP



duoLingo

Serious games for NLP



duoLingo

Originally a game with a purpose: Teaching its users a foreign language while having them translate simple phrases in documents (the translation feature has since been removed)

Serious games for NLP

Serious games for NLP

- Why are serious games for NLP good?
 - Save money, avoid expensive annotations or crowdsourcing
 - Get high-quality results from experts
 - Make people happy!

Serious games for NLP

- Why are serious games for NLP good?
 - Save money, avoid expensive annotations or crowdsourcing
 - Get high-quality results from experts
 - Make people happy!



Serious games for NLP

- Why are serious games for NLP good?
 - Save money, avoid expensive annotations or crowdsourcing
 - Get high-quality results from experts
 - Make people happy!



It looks very
promising,
doesn't it?!

The state of NLP games in mid-2010s

Wordrobe (Venhuizen et al., 2013)

Senses Questions left until drawer is completed: 5

Authorities say the accident **occurred** Saturday, near the town of Veligonda in southern Andhra Pradesh state.

come to pass (synonyms: happen, hap, go on, pass off, pass, fall out, come about, take place)
 come to one's mind – suggest itself
 to be found to exist

Place your bet: low high

answer skip

Jinx (Seemakurty et al., 2010)

Jinx

Play now
Score remaining

Your score
720

Emergency
check crash

Phrase Detectives (Poesio et al., 2013)

USERPROFILE
jibjub
0 this week
0 decisions
0 agreements
0 extras
0 this month
0 all time
Level: Trainee
Your rating: 0%
CASE OPEN
11 tasks remaining
0 completed cases
[EDIT PROFILE | LOGOUT](#)
Like 16 people like this.

NAME THE CULPRIT
Has the phrase shown in orange been mentioned before in this text or is it a property? Use your mouse to select the closest phrase(s) if it has been mentioned before.

Banhammer (Wikipedia)
The term banhammer, is a satirical term for the banning or blocking of users of Internet forums or online games.

SEARCHCLUES
Phrases beginning with a, an or the can serve two different purposes.
1. As an object
They can be used to identify an object in the text, for example "The postman delivered a letter" or "Jane owns a laptop".
2. As a property
They can also be used to say something about an object. For example "Fred, the postman, delivered a letter" describes the object "Fred" as having the property of being "the postman".
If you think the phrase describes a property try to select the closest phrase it refers to.

Not mentioned before This is a property Done

Comment on this phrase
▶ Skip this one
▶ Skip - closest phrase can't be selected
▶ Skip - closest phrase is no longer visible
▶ Skip - error in the text

Games4NLP: State of the game

uComp Language Quiz

Games4NLP: State of the game

uComp Language Quiz

The screenshot shows the homepage of the Games4NLP language quiz. At the top, there's a navigation bar with icons for back, forward, search, and user profile. The URL in the address bar is `quiz.ucomp.eu`. Below the bar, the title "language quiz" is displayed with a checkmark icon. A sidebar on the left lists "Play", "Progress", "Leaderboard", and "About". On the right, there are three small square icons representing different games or features.

Term Sentiment (with a warning icon)

Does the following term express a **negative**, neutral or **positive** opinion?

perfektní

Available languages: EN, DE, FR, ES, RU, ZH, CS (English, German, French, Spanish, Russian, Chinese, Czech). There are also red, grey, green, and blue square buttons below the language list.

Game Level

Level 2 (highlighted in yellow) vs You (at 220 points). A green arrow with "+5" indicates an increase in level.

Top Scores February 2018

Rank	User	Score
1	zb1022705 mvrht	220
	Level: 2	

Bonus Points February 0

Invite your friends to play! You'll receive a bonus of 5% of their points after they have accepted your invitation.

Games4NLP: State of the game

uComp Language Quiz

The screenshot shows the homepage of the Games4NLP uComp Language Quiz. At the top, there is a navigation bar with icons for back, forward, search, and user profile. Below it, the title "language quiz" is displayed with a checkmark icon. A sub-navigation bar includes links for "Play", "Progress", "Leaderboard", and "About". On the right side of the header, there are three small square icons representing different language or resource logos.

The main content area features a section titled "Term Sentiment" with a warning icon. It asks, "Does the following term express a negative, neutral or positive opinion?" followed by the word "perfektní". Below this, there is a row of five colored squares (red, orange, grey, green, blue) with a plus sign on the right. To the right of the term section, there is a "Top Scores February 2018" table showing one entry:

Rank	User	Score
1	zb1022705 mvrht	220

Below the scores, there is a "Bonus Points February" section with a value of 0. To the right, there is a call-to-action for inviting friends with the text: "Invite your friends to play! You'll receive a bonus of 5% of their points after they have accepted your invitation.".

At the bottom left, there is a "Game Level" section showing a level of "2" with a "Level +5" button. To the right, there is a summary of the quiz's purpose: "uComp Language Quiz - A Game with a Purpose for Multilingual Language Resource Acquisition".

At the very bottom, two contact details are provided:

Arno Scharl
MODUL University Vienna
Department of New Media Technology
Am Kahlenberg 1, Vienna, Austria
scharl@modul.ac.at

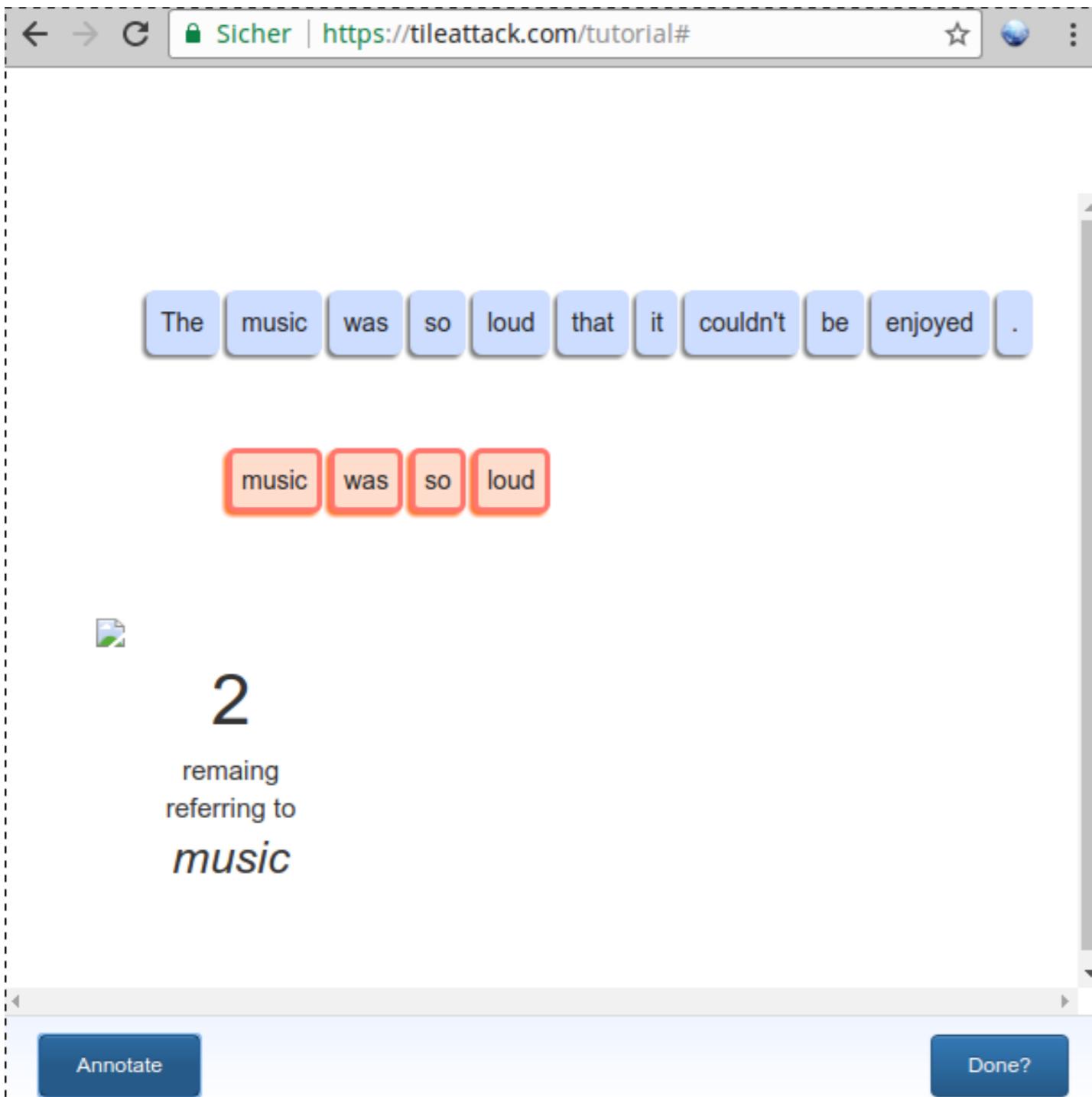
Michael Föls
Vienna Univ. of Economics & Business
RI for Computational Methods
Welthandelsplatz 1, Vienna, Austria
michael.foels@wu.ac.at

Games4NLP: State of the game

Tileattack

Games4NLP: State of the game

Tileattack



Games4NLP: State of the game

Tileattack

The screenshot shows a web browser window for 'Tileattack' at <https://tileattack.com/tutorial#>. The main content area displays a sequence of words in blue rounded rectangular boxes: 'The', 'music', 'was', 'so', 'loud', 'that', 'it', 'couldn't', 'be', 'enjoyed', followed by a period. Below this, four specific words ('music', 'was', 'so', 'loud') are highlighted with red rounded rectangular boxes. In the bottom left corner of the main area, there is a small icon of a document and the number '2'. To the right of the main content, a white box contains the text: 'Testing game mechanics in games with a purpose for NLP applications' in bold, followed by the names 'Chris Madge, Udo Kruschwitz, Jon Chamberlain, Richard Bartle, Massimo Poesio' in bold, and the text 'School of Computer Science and Electronic Engineering' and 'University Of Essex' below them. At the bottom of this box is an email address: '{cjmadv, udo, jchamb, rabartle, poesio}@essex.ac.uk'. At the very bottom of the browser window, there are two buttons: 'Annotate' on the left and 'Done?' on the right.

Sicher | <https://tileattack.com/tutorial#>

The music was so loud that it couldn't be enjoyed .

music was so loud

2

remaing referring to *music*

Testing game mechanics in games with a purpose for NLP applications

Chris Madge, Udo Kruschwitz, Jon Chamberlain, Richard Bartle, Massimo Poesio
School of Computer Science and Electronic Engineering
University Of Essex
{cjmadv, udo, jchamb, rabartle, poesio}@essex.ac.uk

Annotate Done?

Games4NLP: State of the game

Zombilingo

Games4NLP: State of the game

Zombilingo

The screenshot shows a web browser window for the Zombilingo game at <https://zombilingo.org/game/upl>. The page has a green header with the Zombilingo logo and navigation links for Accueil, Jouer, Forum, and FAQ. On the right, there's a user profile icon for 'zb1034181' with a dropdown arrow, and icons for a brain (0), a book (0), and a flask (0). The main content area has a green background. At the top, it says "Trouve les expressions multi-mots présentes dans la phrase". Below this is a progress bar consisting of a red bone shape with the text "90%" in the center. To the right of the progress bar is a question mark icon. A large text box contains the following French text: "L'affaire des diamants est une affaire politique révélée par Le Canard enchaîné le 10 octobre 1979 qui impliquait le président Valéry Giscard d'Estaing lorsqu'il était ministre des finances et le chef d'État de la République centrafricaine, Jean-Bedel Bokassa, dans les années 1970." At the bottom left is a dashed rectangular input field, and at the bottom right is a green button labeled "Valider 0".

Games4NLP: State of the game

Zombilingo

The screenshot shows the Zombilingo game interface. At the top, there's a browser header with a back button, forward button, refresh, home, and address bar showing the URL <https://zombilingo.org/game/upl>. To the right of the address bar are search, download, and other browser controls. Below the header, the Zombilingo logo is on the left, followed by menu links: Accueil, Jouer, Forum, and FAQ. On the right side of the header are user icons: a cartoon zombie head, the username 'zb1034181' with a dropdown arrow, a brain icon with '0', a green book icon with '0', and a flask icon with '0'. The main content area has a green background. In the center, the text 'Trouve les expressions multi-mots présentes dans la phrase' is displayed above a progress bar. The progress bar is a red bone shape with the number '90%' in the middle. To the right of the progress bar is a link 'Besoin d'aide?' next to a circular icon of an elderly man with glasses. Below the progress bar, a text box contains the following French text: 'L'affaire des diamants est une affaire politique révélée par Le Canard enchaîné le 10 octobre 1979 qui impliquait le président Valéry Giscard d'Estaing lorsqu'il était ministre des finances et le chef d'État de la République centrafricaine, Jean-Bedel Bokassa, dans les années 1970.' At the bottom right, there's a white box containing the text 'Who wants to play Zombie? A survey of the players on ZOMBILINGO' and contact information for three researchers.

**Who wants to play Zombie?
A survey of the players on ZOMBILINGO**

Karën Fort
Université Paris-Sorbonne / STIH
karen.fort@paris-sorbonne.fr

Bruno Guillaume
Inria Nancy Grand-Est/LORIA
bruno.guillaume@inria.fr

Nicolas Lefebvre
Inria Nancy Grand-Est/LORIA
nicolas.lefebvre@inria.fr

Games4NLP: State of the game

Word Sheriff

Games4NLP: State of the game

Word Sheriff

Play with others, make a guess, and climb the leaderboards!

The screenshot shows a game interface for 'Word Sheriff' on the 'Games4NLP' platform. The interface is set against a teal background.

Top Left: A blue circular icon labeled 'Me' with a brown cowboy hat containing a yellow star. Below it is a yellow progress bar with the text '0 (+ 0)'.

Top Right: An orange button labeled 'Sound Off' with a speaker icon.

Middle Top: A yellow box containing the text 'You are the narrator! 3 clues allowed'.

Middle Center: The message 'Round Finished! 0 out of 3 were correct' and 'Your word is:' followed by the word 'berlin' in a brown font.

Middle Bottom: Two buttons: 'Give a clue' (light gray) and 'Give clue' (blue).

Bottom Left: An orange circular icon labeled 'Player' with a yellow lightbulb icon. Below it is a yellow progress bar with the text '0 (+ 0)'.

Bottom Center: The heading 'Your current clues are:' followed by a list of three clues: 1. Germany, 2. Capital, 3. Bear.

Bottom Right: Three green circular icons labeled 'Player' with yellow lightbulb icons, each showing a yellow progress bar with '0 (+ 0)'.

Bottom Far Right: The heading 'The current guesses are:' followed by a list of six guesses: 1. macedonia, 2. soldier, 3. cartel, 4. yacht, 5. volcanic, 6. query.

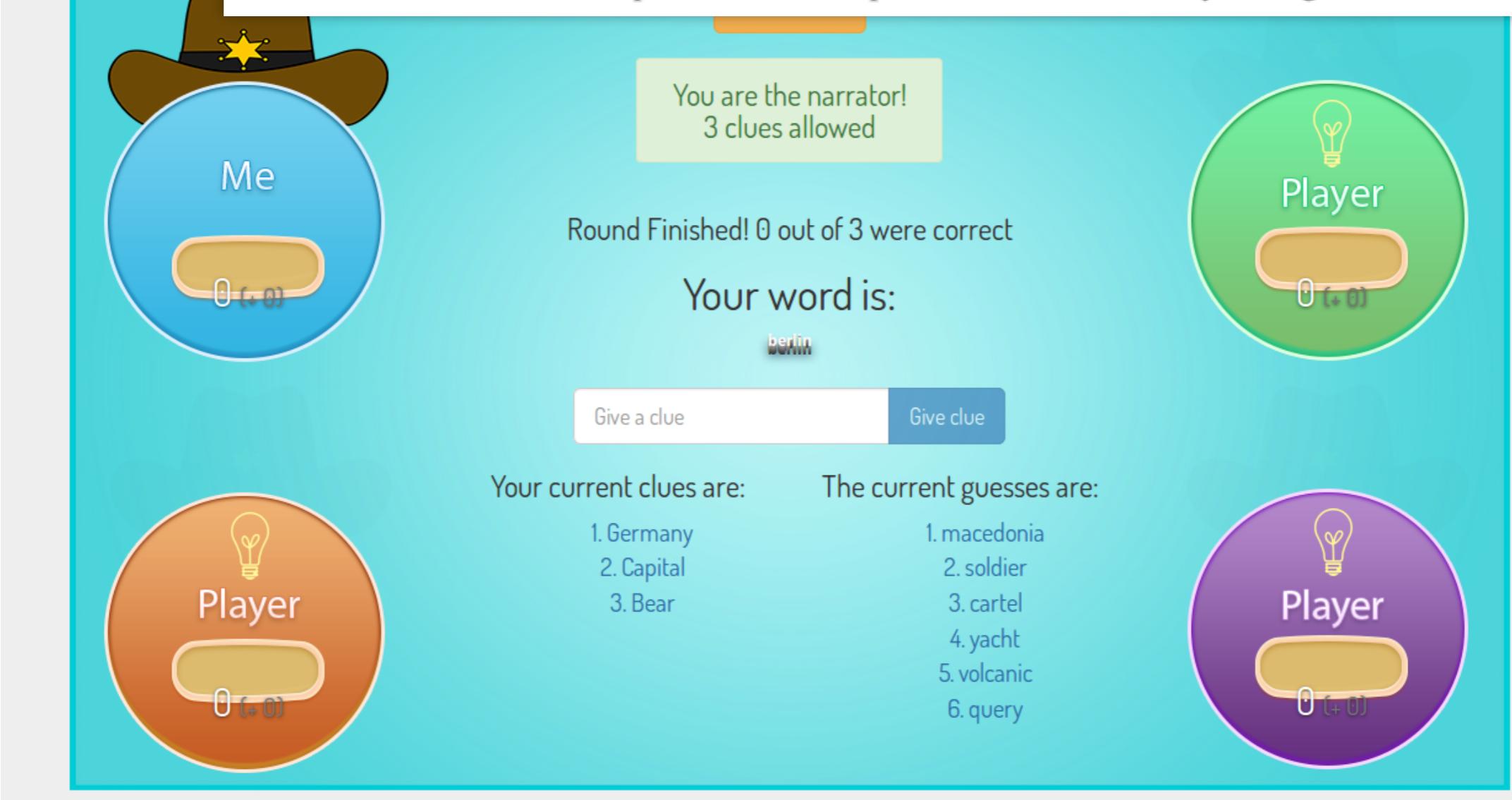
Games4NLP: State of the game

Defining Words with Words: Beyond the Distributional Hypothesis

Iuliana-Elena Parasca* Andreas Lukas Rauter* Jack Roper* Aleksandar Rusinov*
Guillaume Bouchard Sebastian Riedel Pontus Stenetorp

{iuliana.parasca, andreas.rauter, jack.roper, aleksandar.rusinov}.13@ucl.ac.uk
{g.bouchard, s.riedel, p.stenetorp}@cs.ucl.ac.uk

Department of Computer Science, University College London



Games4NLP: State of the game

Player's perspective

Games4NLP: State of the game

Player's perspective

- It may sound harsh but these are not games but **annotation tools in disguise**

Games4NLP: State of the game

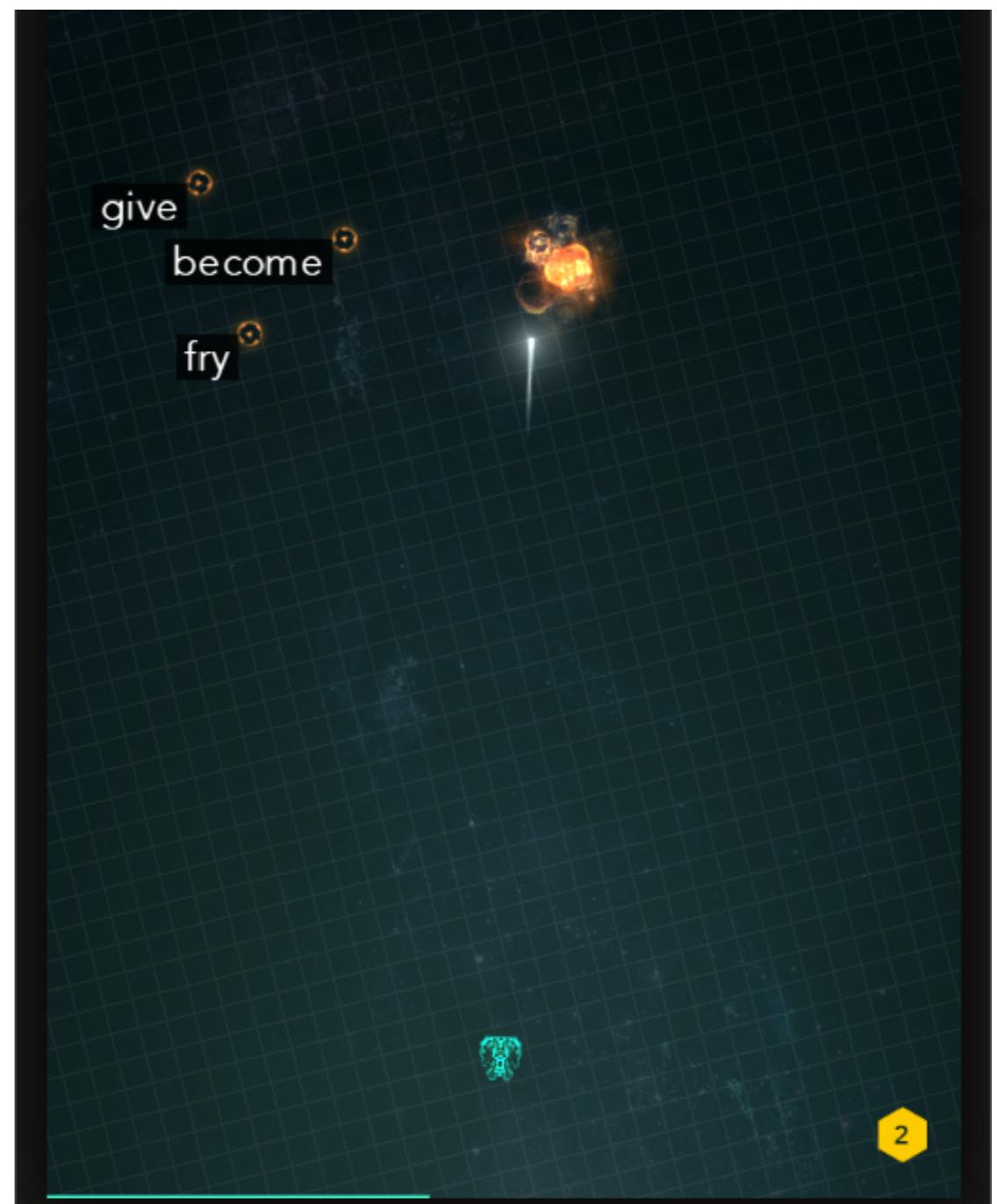
Player's perspective

- It may sound harsh but these are not games but **annotation tools in disguise**
- Key components of serious games (Garris et al., 2002)
 - Interest
 - Enjoyment
 - Task involvement
 - Confidence

Games4NLP: State of the game

Player's perspective

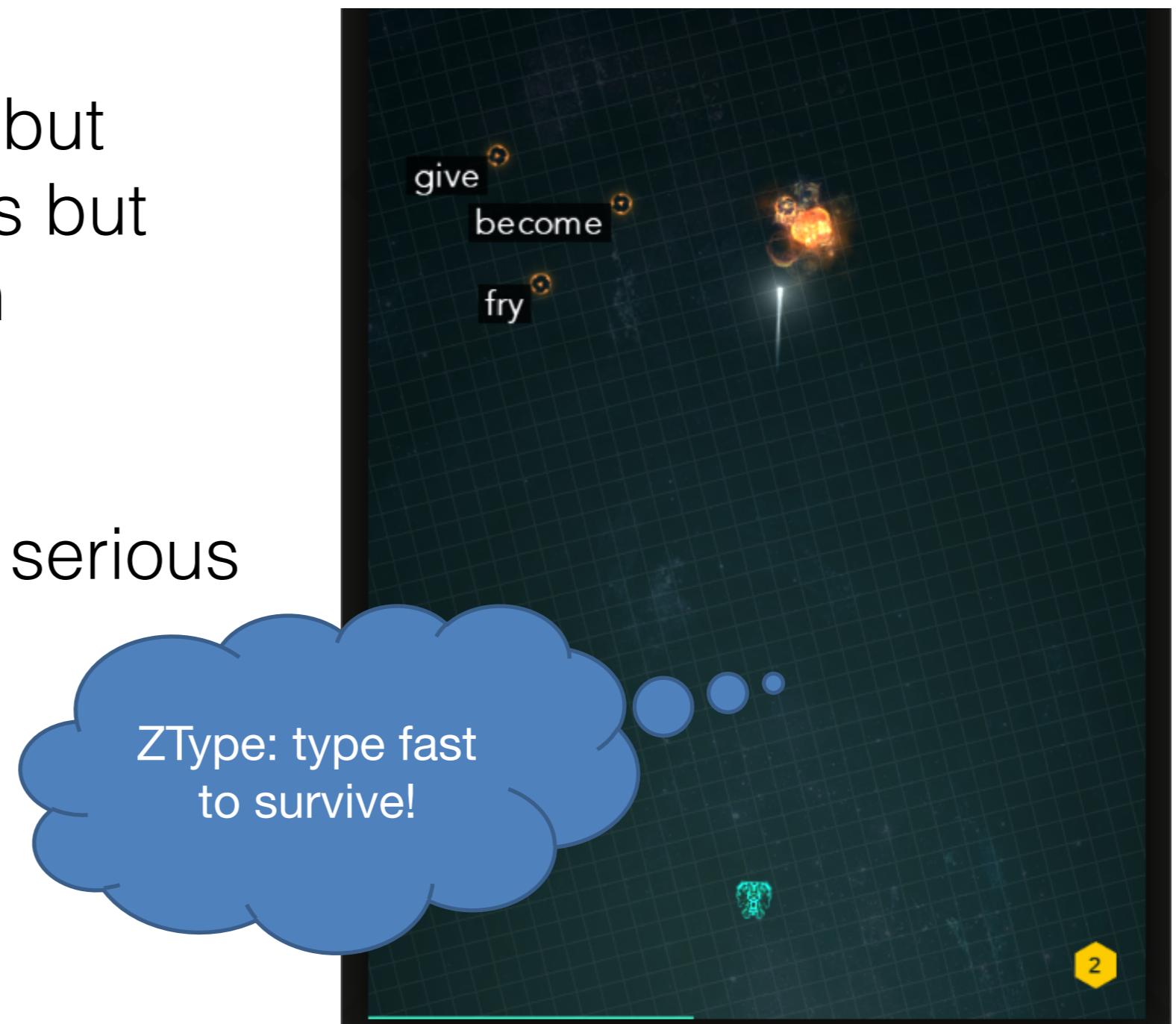
- It may sound harsh but these are not games but **annotation tools in disguise**
- Key components of serious games (Garris et al., 2002)
 - Interest
 - Enjoyment
 - Task involvement
 - Confidence



Games4NLP: State of the game

Player's perspective

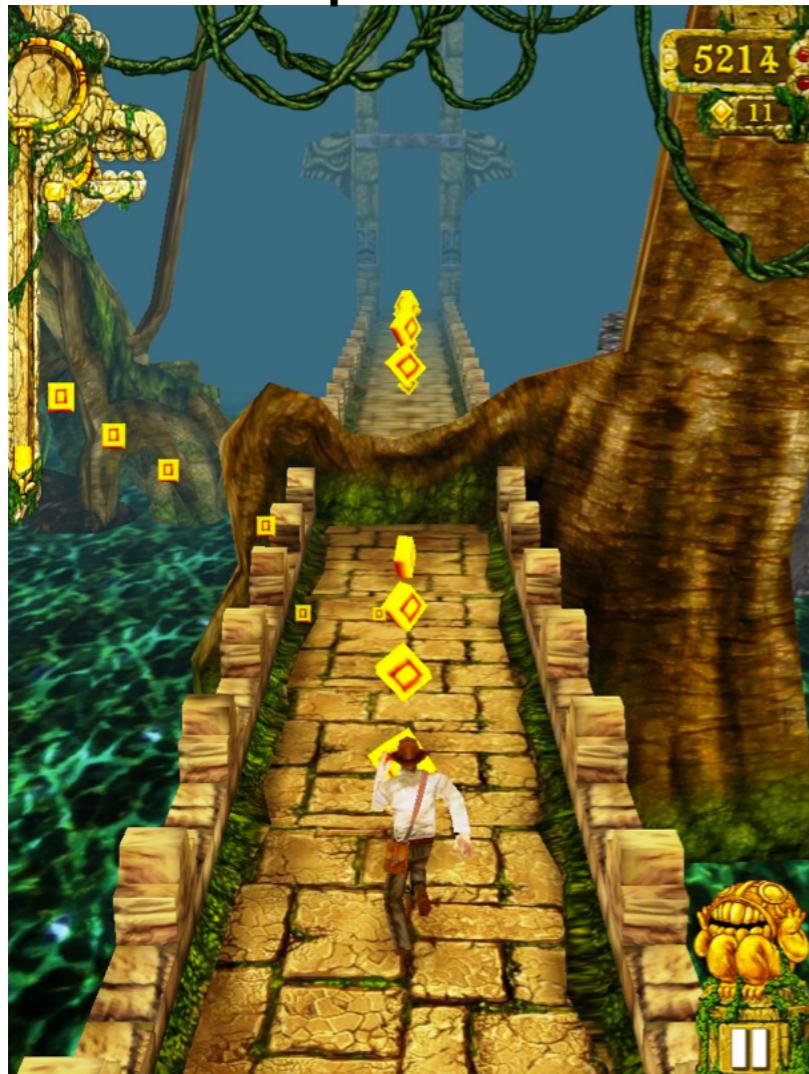
- It may sound harsh but these are not games but **annotation tools in disguise**
- Key components of serious games (Garris et al., 2002)
 - Interest
 - Enjoyment
 - Task involvement
 - Confidence



Can we take a video game
and **taskify** it?

Idea: Taskify a popular game with a simple game mechanic

Temple Run



1 Billion Downloads

Fruit Ninja



300 Million Downloads
1/3 of all iPhones

Can we adapt Fruit Ninja for disambiguation?



Key mechanic: Click on certain kinds of things



Player needs to avoid these

Key mechanic: Click on certain kinds of things



Player needs to avoid these

She plays the bass

- 1) the lowest part of the musical range
- 2) an adult male singer with the lowest voice
- 3) a North American freshwater fish
- 4) a musical instrument

She plays the bass

Annotate this

- 1) the lowest part of the musical range
- 2) an adult male singer with the lowest voice
- 3) a North American freshwater fish
- 4) a musical instrument

She plays the bass

Annotate this

- 1) the lowest part of the musical range



- 2) an adult male singer with the lowest voice



- 3) a North American freshwater fish



- 4) a musical instrument



She plays the bass

Annotate this

=

Click on this!

- 1) the lowest part of the musical range



- 2) an adult male singer with the lowest voice



- 3) a North American freshwater fish



- 4) a musical instrument



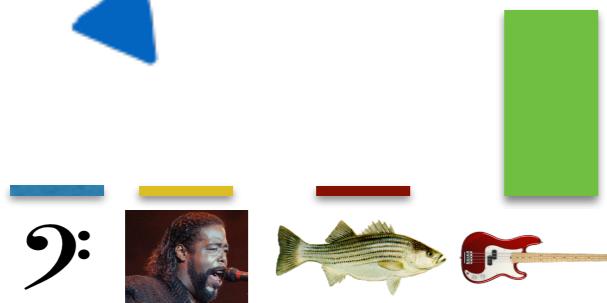
She plays the bass



She plays the bass



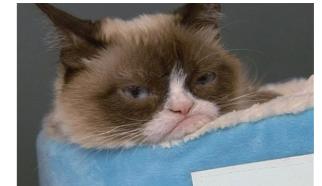
She plays the bass



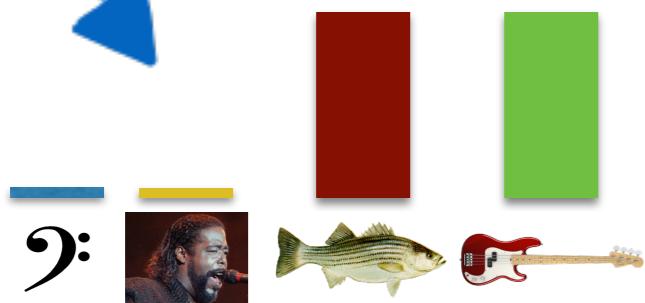
She plays the bass



Key Problem #1: This is boring



She plays the bass



Key Problem #1: This is boring
Key Problem #2: Game mistakes radically change results



She plays the bass

Annotate this

=

Click on this!

- 1) the lowest part of the musical range



- 2) an adult male singer with the lowest voice



- 3) a North American freshwater fish



- 4) a musical instrument



She plays the bass

Annotate this

=

Click on these!

- 1) the lowest part of the musical range



- 2) an adult male singer with the lowest voice



- 3) a North American freshwater fish



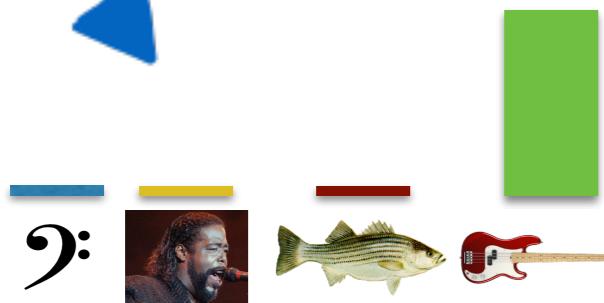
- 4) a musical instrument



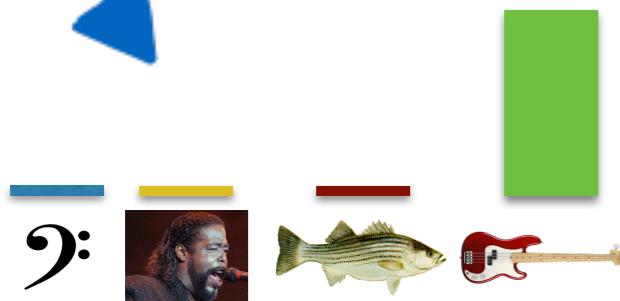
She plays the bass



She plays the bass



She plays the bass



**The most common gameplay mistake
has no effect on the annotation**

Where do we get the images?

- 1) the lowest part of the musical range



- 2) an adult male singer with the lowest voice



- 3) a North American freshwater fish



- 4) a musical instrument

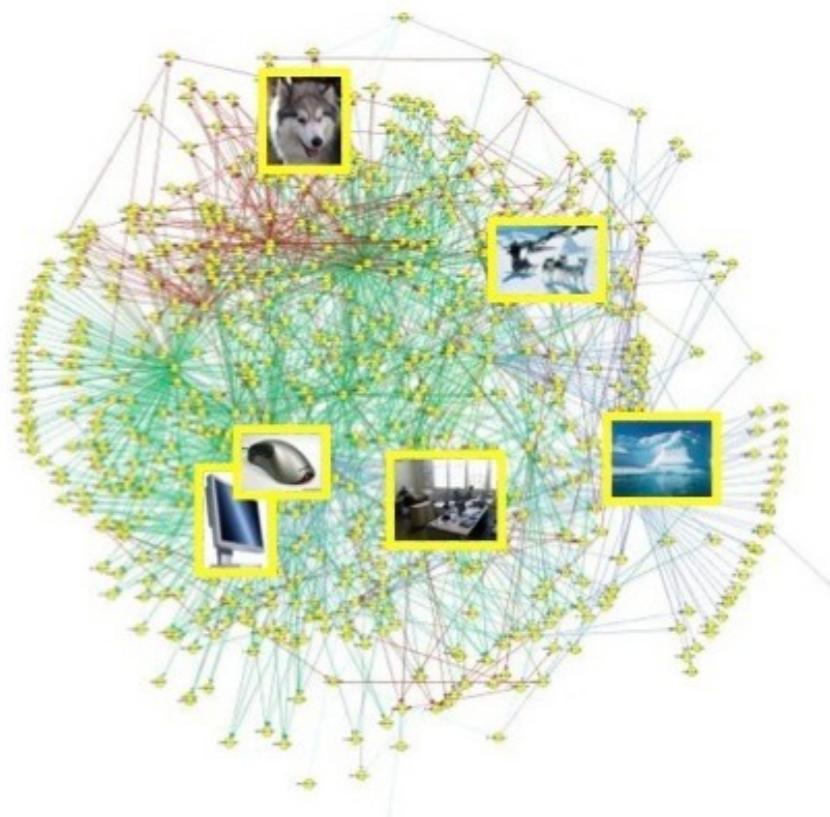


Current image-sense libraries



A very large multilingual encyclopedic dictionary and semantic network

Current image-sense libraries



IM^{GENET}

No abstract nouns,
no verbs

BabelNet

A very large multilingual encyclopedic dictionary and semantic network

Few verbs,
relatively few pictures

Build a game in order to
create resources for another
game!



Puzzle Racer



0 0 0

Start

Player Setup

► Instructions

Game Options

View Leaderboard

Puzzle Racer



0 0 0

Start

Player Setup

► Instructions

Game Options

View Leaderboard

Real game features!

Unlockable Racers



Lots of Power-ups



Enemies!



Leaderboards

All-Time Top Puzzle Racers:		
Player	Lv.	Total Score
the knight who ...	2	519324
jim	2	325224
N Chompskinator	2	245013
david	2	241552
charnimanik	1	532313

► Done (new) Click here for full list!



love#n#1: a strong positive emotion of affection



cat#n#1: a feline mammal



...

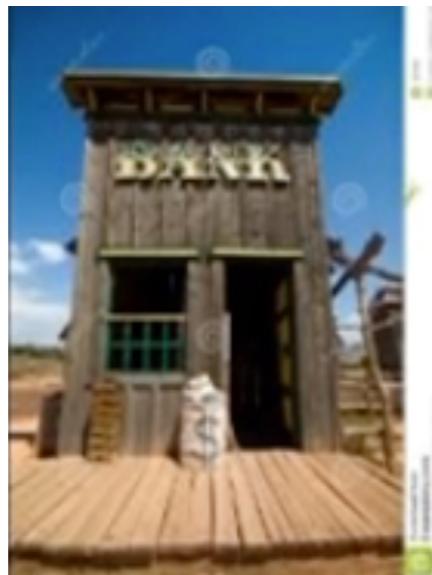
Task Question:

Which of these pictures best shows the following definition:
“a building in which the business of banking transacted”



Task Question:

Which of these pictures best shows the following definition:
“a building in which the business of banking transacted”



>



>



Taskify by making bad pictures in-game obstacles



Players must identify
obstacles and dodge them



Taskify by making bad pictures in-game obstacles



Players must identify
obstacles and dodge them



How do we get rid of the text to make it a game?

Which of these pictures best shows
“a building in which the business of banking
transacted”



0

Time:

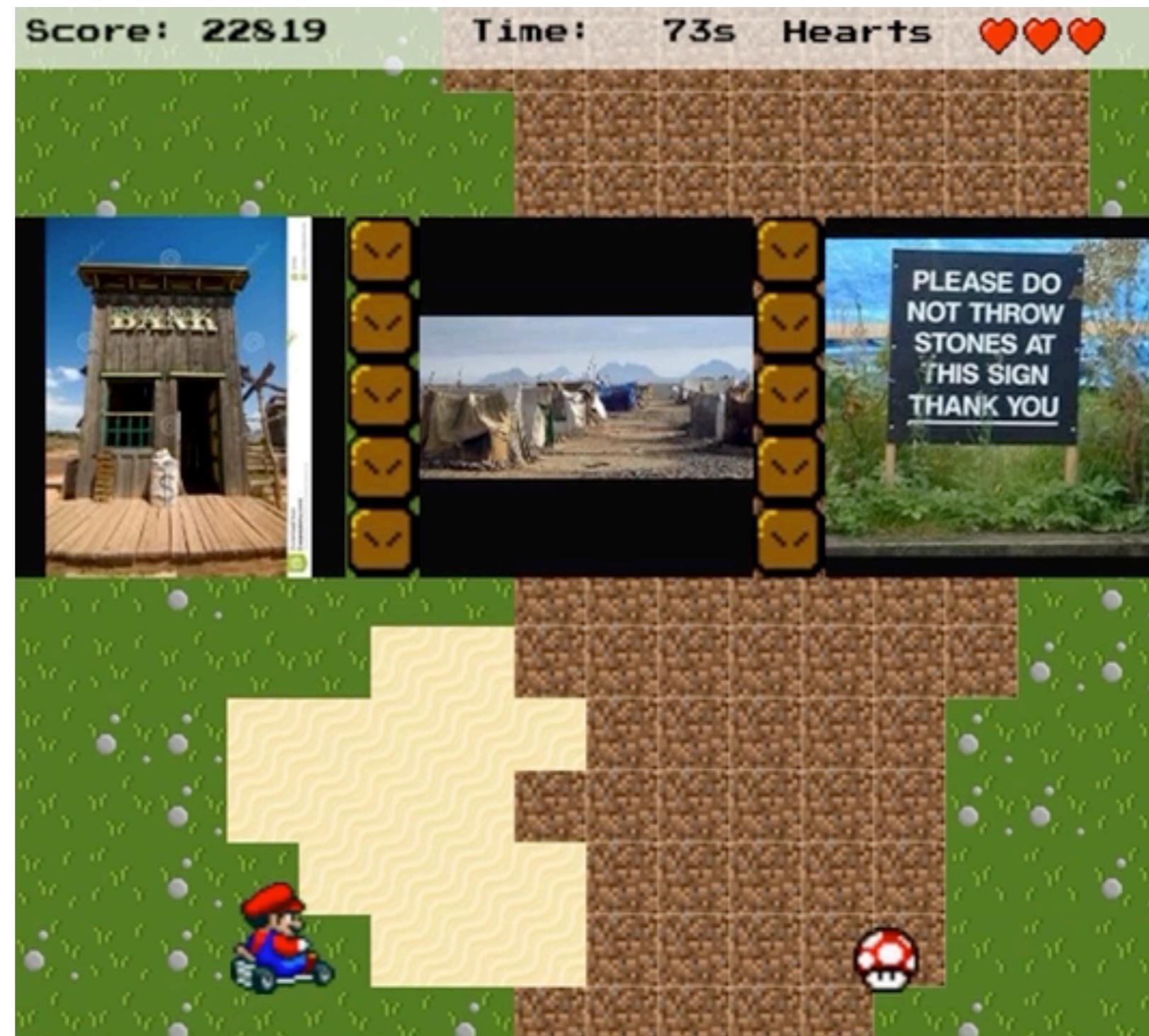
Hearts

This race's puzzle clues:

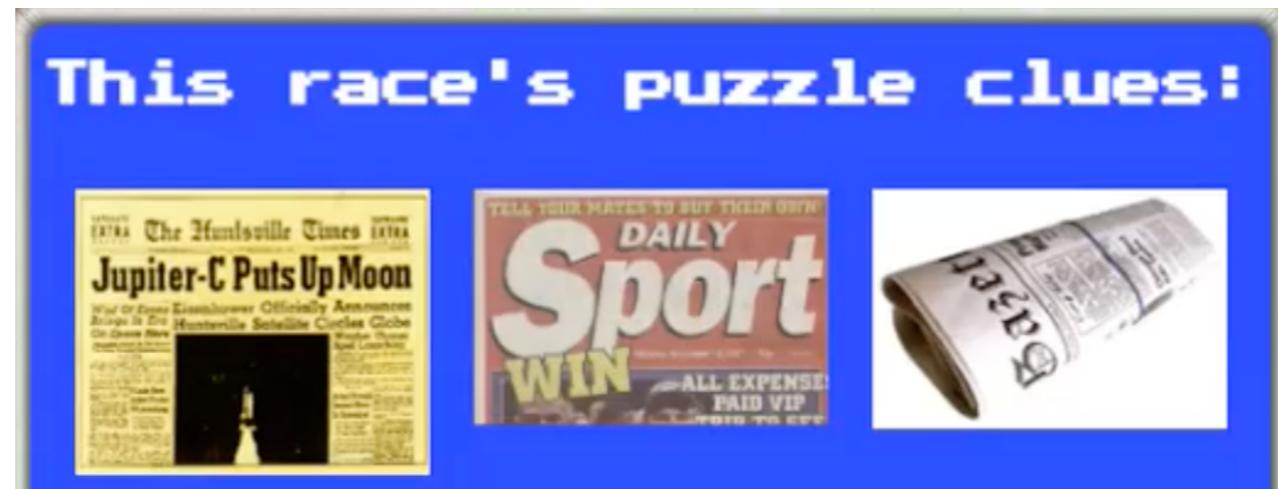


**Find the idea in common and
guide your racer over similar
pictures to stay alive!**

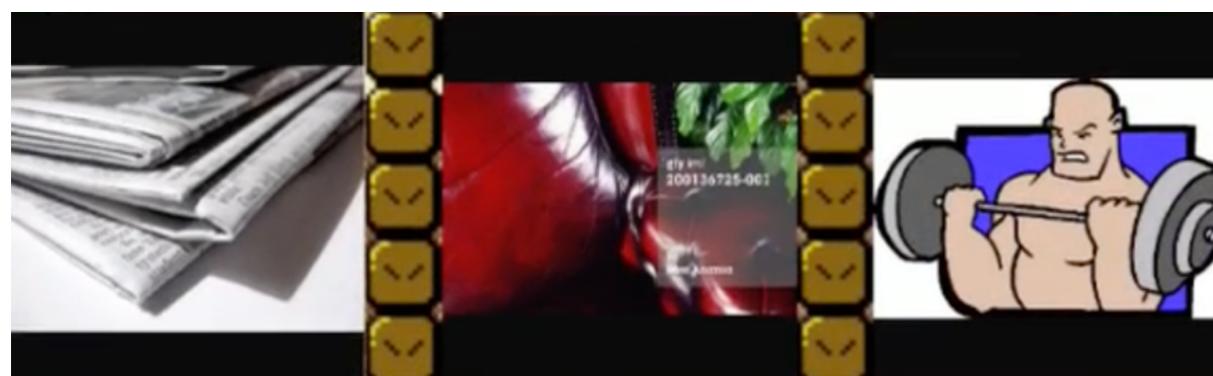
► Let's race!







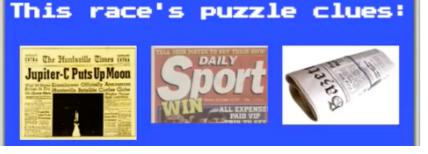
Players are shown
two types of puzzle gates



Golden Gate



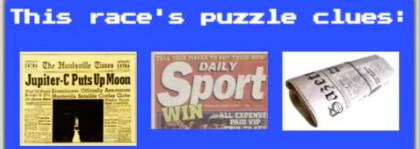
Mystery Gate





Accuracy: 100%

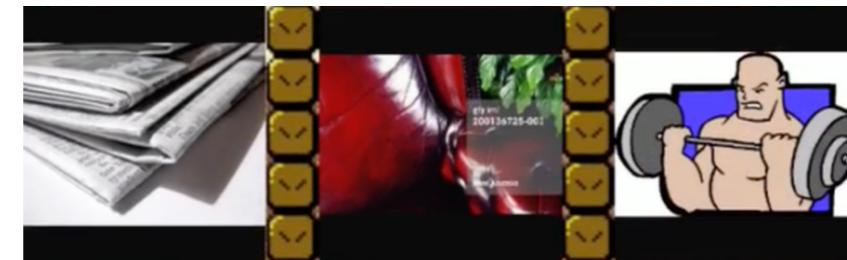
First three
gates are
always
golden



Accuracy: 66%

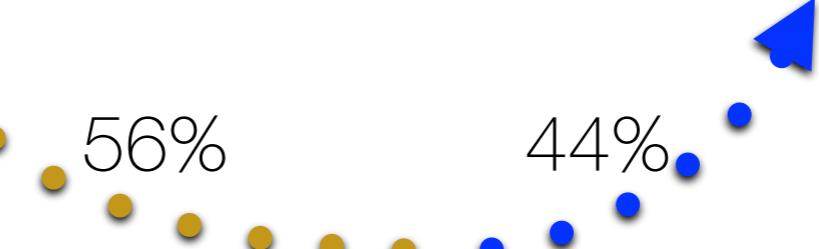
Accuracy: 100%

Accuracy: 100%

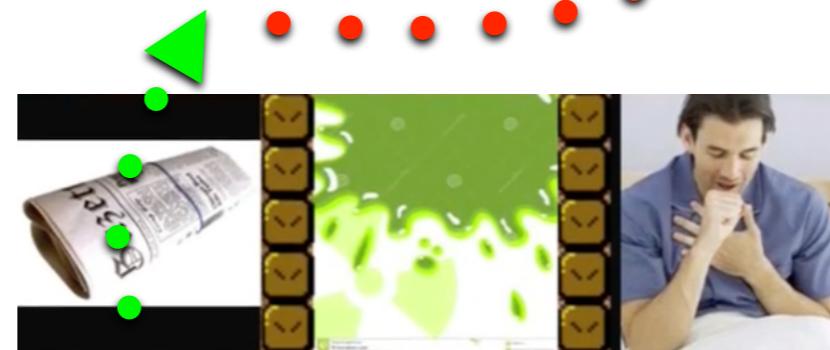


Show a mystery gate with probability =
 $0.66 \times \text{Accuracy}$

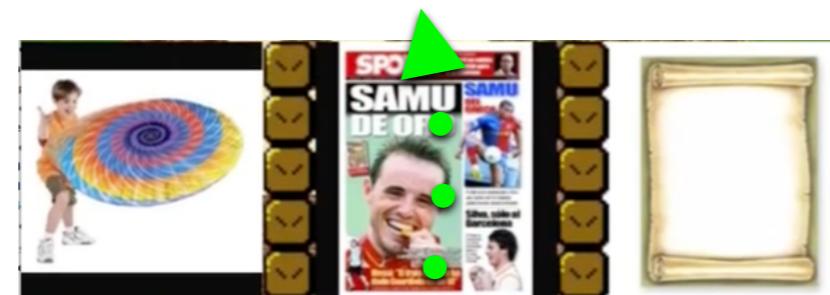
First three gates are
always
golden



Accuracy: 66%



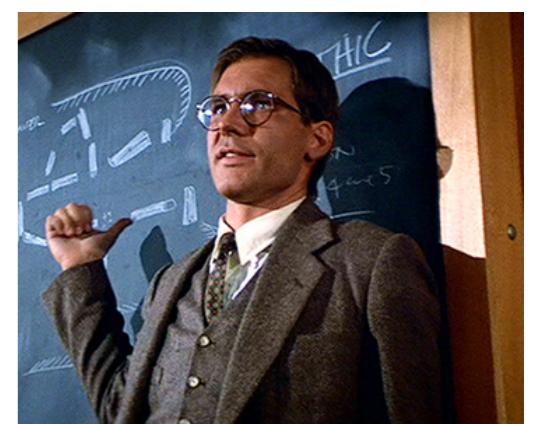
Accuracy: 100%



Accuracy: 100%



Does it work?



Game Setup

- Picked 23 nouns, verbs, and adjectives
 - 4-10 senses each; 132 senses total
- Start with ~10 gold images per sense and 16.6K unlabeled images total
- Recruited students to play, with offer of gift cards for top positions in leaderboard after two weeks (\$70 total)

Gameplay Results

- 126 people played at least one game
- 7,199 races over two weeks
- 20,254 ratings across all images
 - 231 – 329 ratings per sense
- 83% accuracy at **Golden Gates**

How does PuzzleRacer
compare in quality with
Crowdsourcing?

Recreate the Puzzle Racer annotation task on CrowdFlower



Given the three example images in the instructions, which of the following images most resembles underlying idea?

Recreate the Puzzle Racer annotation task on CrowdFlower



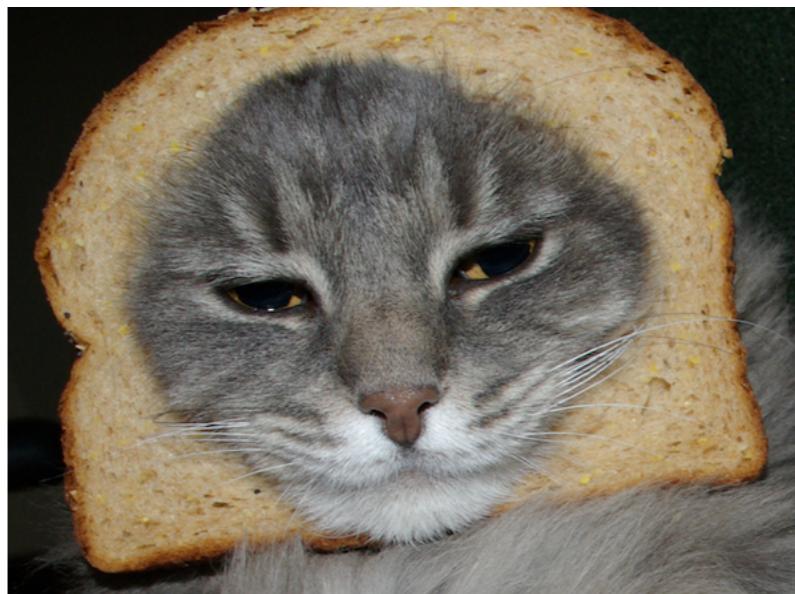
Given the three example images in the instructions, which of the following images most resembles underlying idea?

One of these questions is from a **Golden Gate**, the others are from **Mystery Gates**



Evaluate by comparing top-ranked images

cat (n): a feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats



which is better?

left

equal

right

About equal in quality...



7%

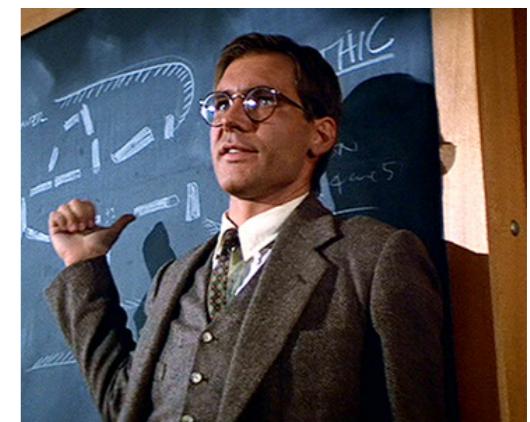


79%

14%

... but Puzzle Racer was 27% the price!*

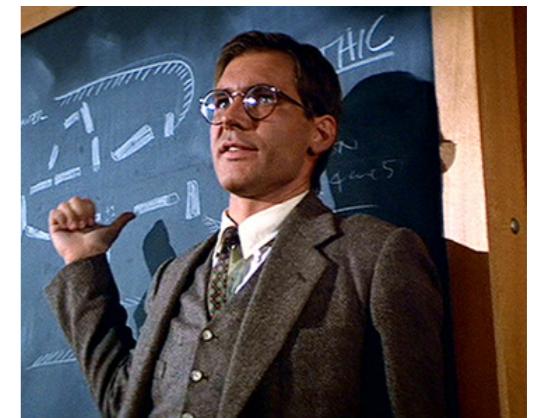
Does it work?



No statistically-significant difference in quality between Puzzle Racer-created and Expert-selected images



=



argument (n): a contentious speech act; a dispute where there is strong disagreement



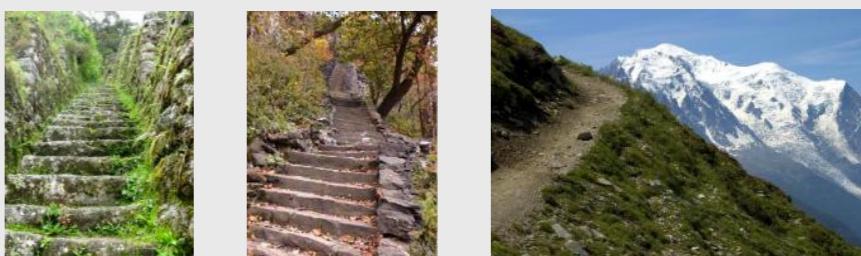
atmosphere (n): the weather or climate at some place



important (a): of great significance or value



climb (v): go upward with gradual or continuous progress



smell (v): smell bad



Now back to disambiguation!

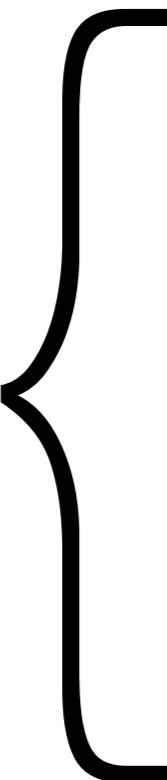


Disambiguate by clicking on
pictures for the wrong senses



She plays the bass

Show one
picture for
each of the
 n senses



She plays the bass

Show one picture for each of the n senses



$$3x^2 - 2xy + c$$

Diagram illustrating the terms of the equation:

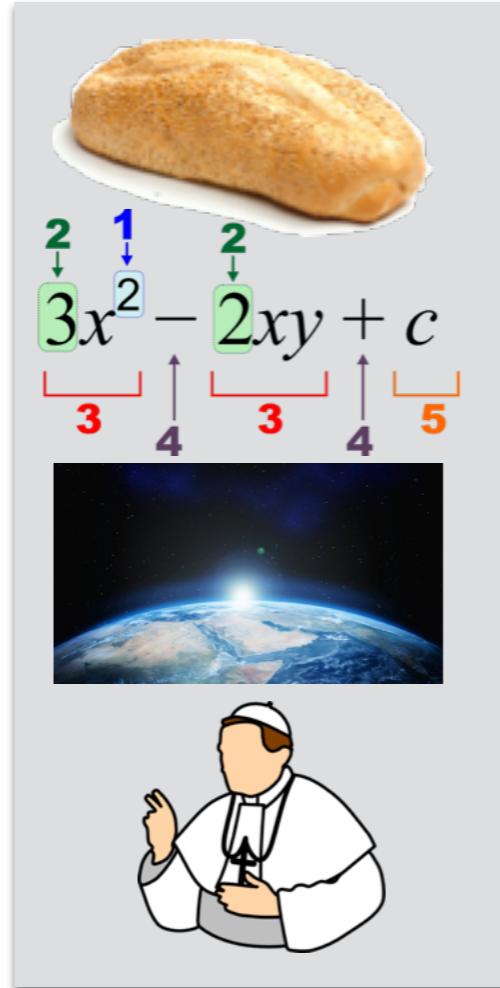
- $3x^2$ is highlighted with a green box and a green arrow pointing to the term.
- 2 is highlighted with a green box and a green arrow pointing to the coefficient.
- 1 is highlighted with a blue box and a blue arrow pointing to the exponent of y .
- 3 is highlighted with a red bracket below the term.
- 4 is highlighted with a red bracket below the term.
- 3 is highlighted with a red bracket below the term.
- 4 is highlighted with a red bracket below the term.
- 5 is highlighted with an orange bracket below the constant term.



Include n pictures from random senses

She plays the bass

Show one picture for each of the n senses



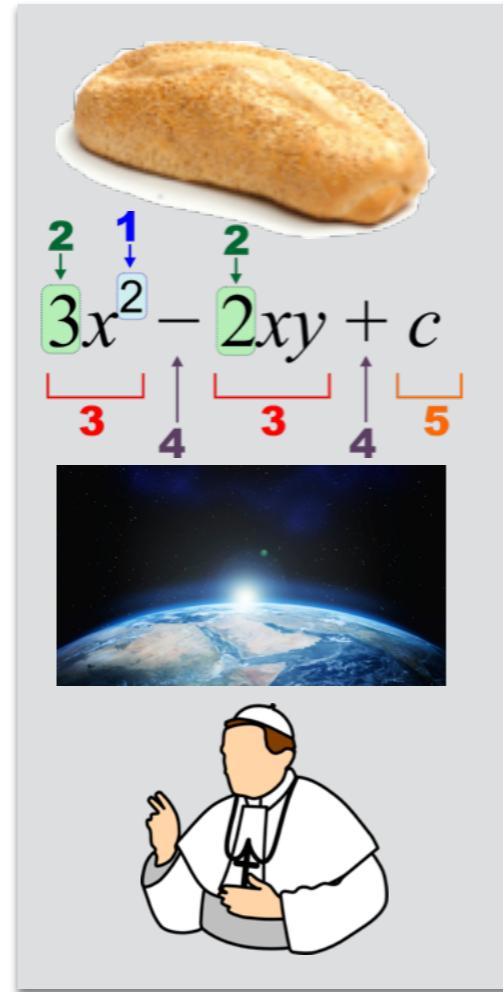
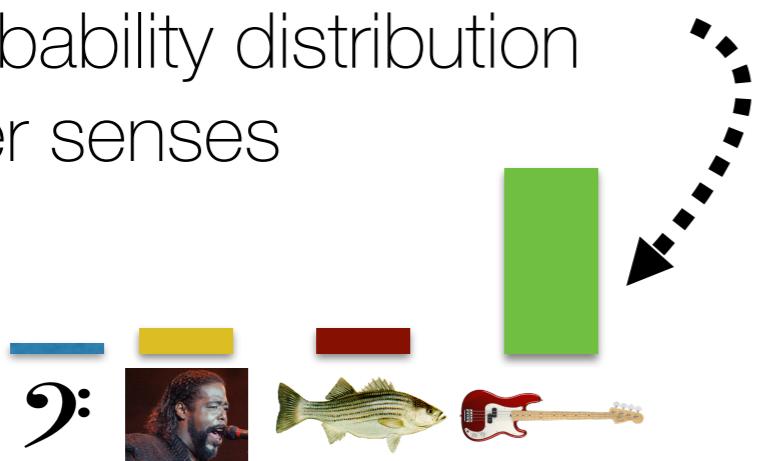
Include n pictures from random senses



Monitor player's ability by them destroying unrelated images from random senses

She plays the bass

Show one picture for each of the n senses



Include n pictures from random senses

Each game produces a probability distribution over senses

Monitor player's ability by them destroying unrelated images from random senses

Disambiguate by clicking on pictures for the wrong senses

Objective

The following sentence contains a clue to your survival. Look at **wins** and think of pictures that remind you of its meaning.

It is the one exercise that drastically influences the definition of the thighs at the hipline - that mark of the champion that sets him apart from all other bodybuilders - a criterion of muscle " drama " that is unforgettable to judges and audiences alike - the facet of muscular development that **wins** prizes .

When you click below, pictures will be thrown on screen. Your job is destroy every picture that **does not** remind you of **wins** in the sentence above. When in doubt, blow it up!

Let me blow stuff up.

Disambiguate by clicking on pictures for the wrong senses

Objective

The following sentence contains a clue to your survival. Look at **wins** and think of pictures that remind you of its meaning.

It is the one exercise that drastically influences the definition of the thighs at the hipline - that mark of the champion that sets him apart from all other bodybuilders - a criterion of muscle " drama " that is unforgettable to judges and audiences alike - the facet of muscular development that **wins** prizes .

When you click below, pictures will be thrown on screen. Your job is destroy every picture that **does not** remind you of **wins** in the sentence above. When in doubt, blow it up!

Let me blow stuff up.

Does it work?

Direct comparison with Wordrobe, a WSD game

 Senses Questions left until drawer is completed: 5 

Authorities say the accident **occurred** Saturday, near the town of Veligonda in southern Andhra Pradesh state.

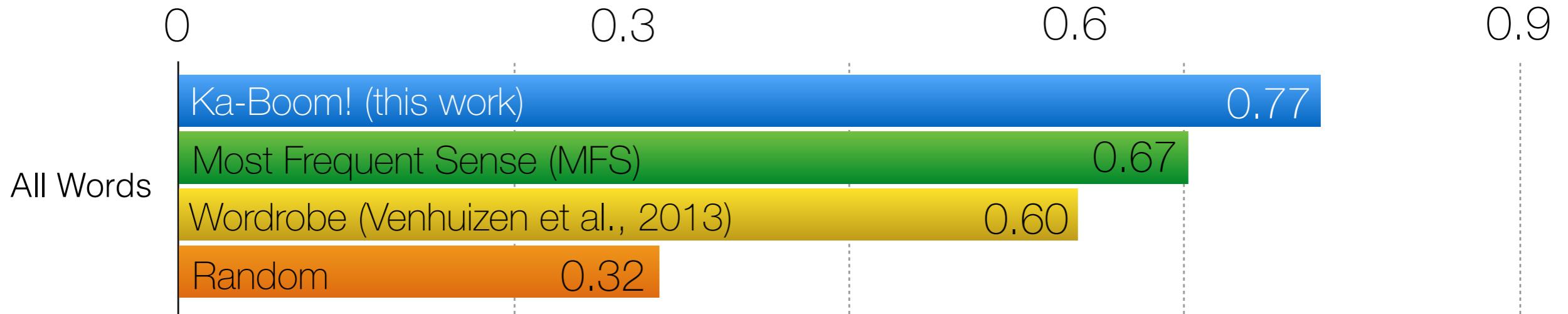
- come to pass (synonyms: happen, hap, go on, pass off, pass, fall out, come about, take place)
- come to one's mind – suggest itself
- to be found to exist

Place your bet: low  high

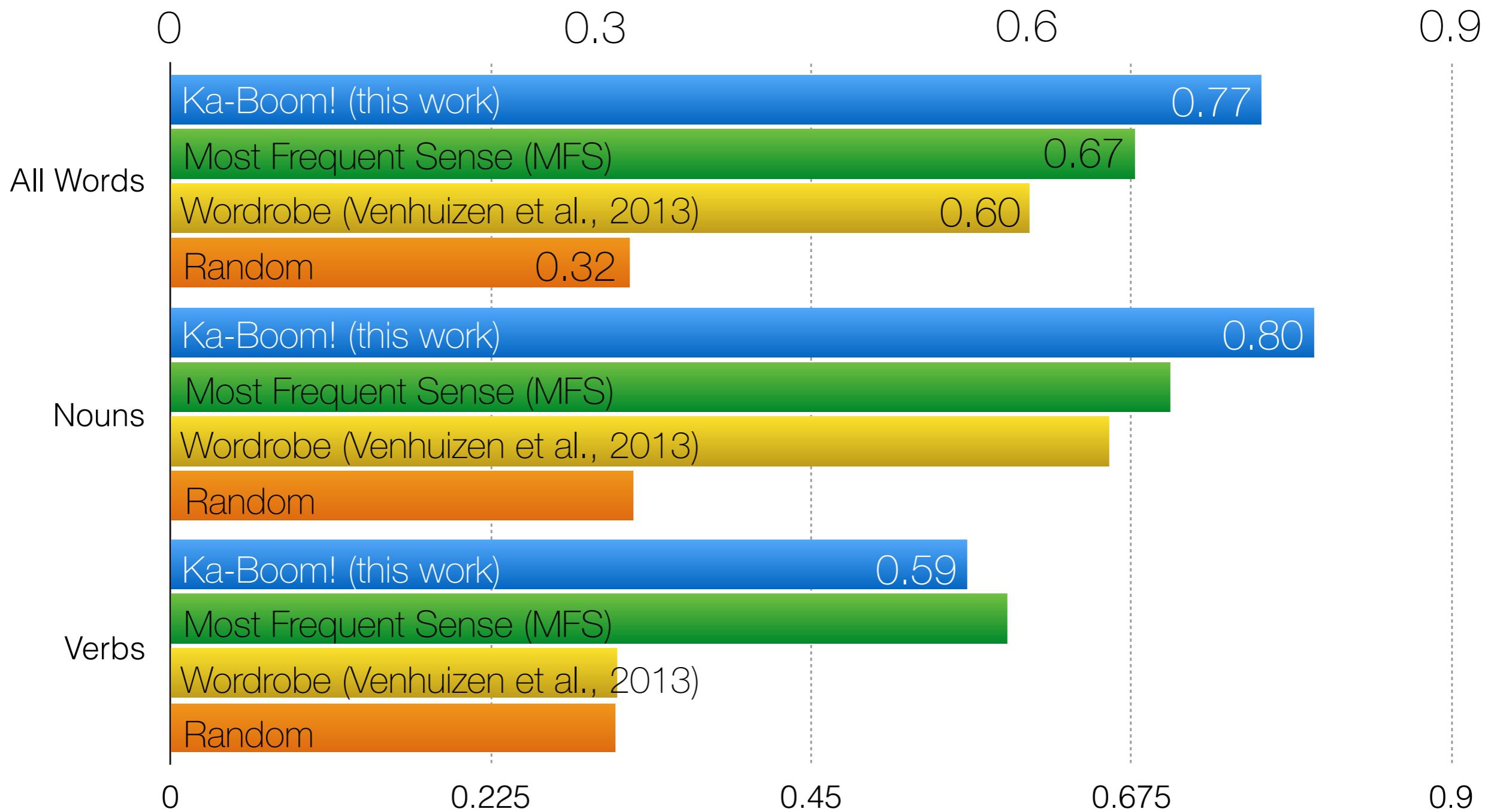
answer skip

Tested on 111 sentences total for
74 nouns and 16 verbs (3.4 senses on average)

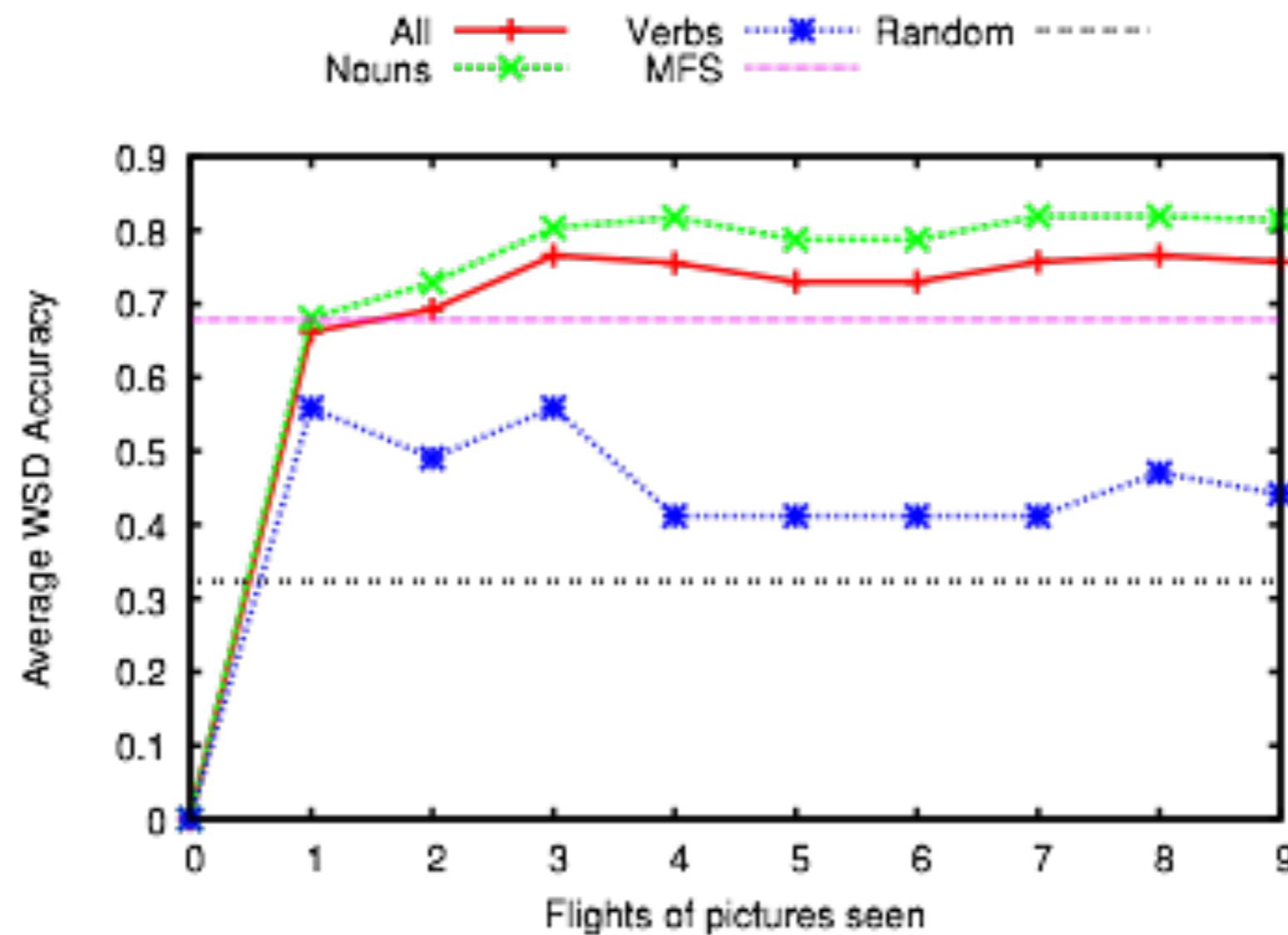
Disambiguation Accuracy



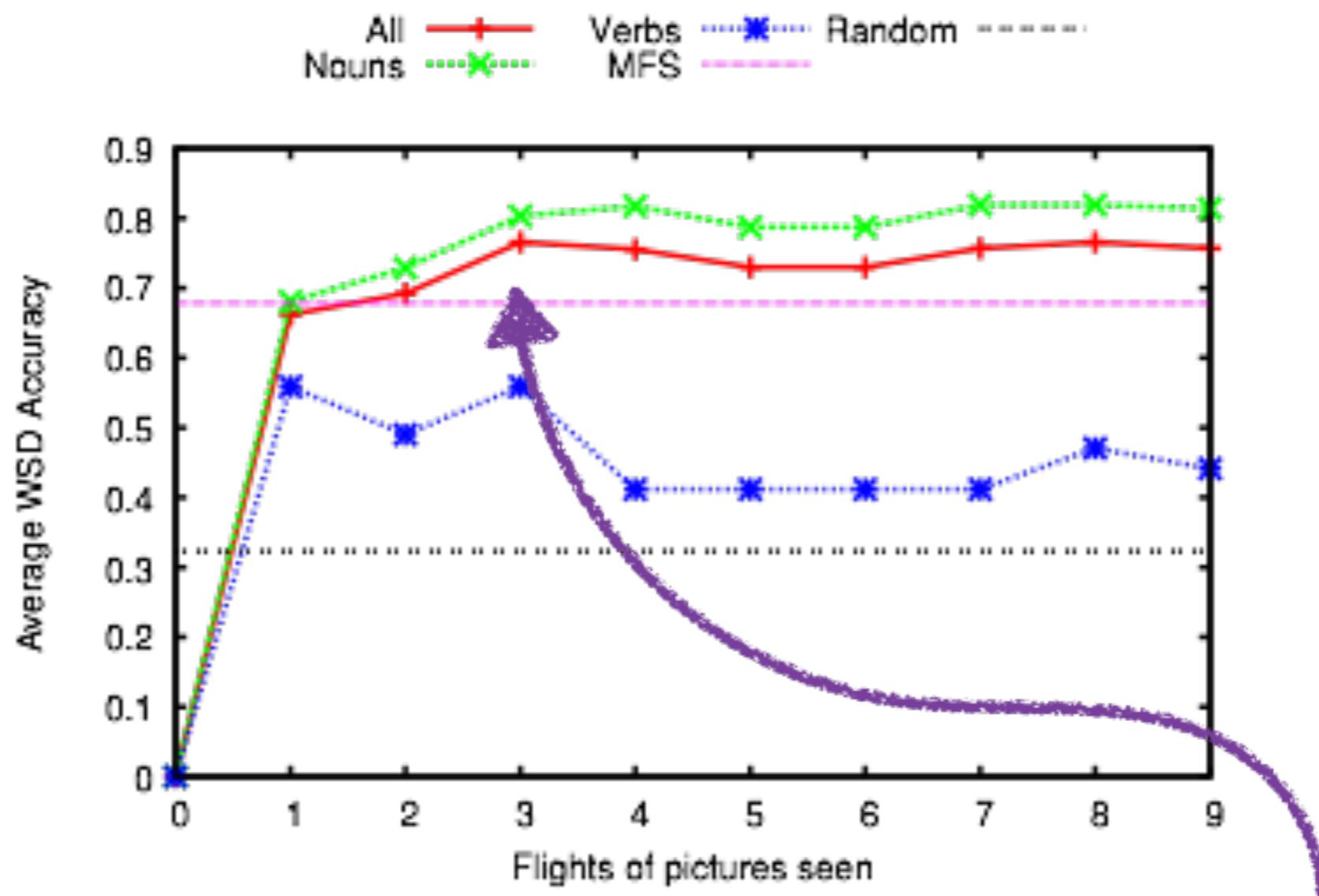
Disambiguation Accuracy



How long did players take to converge on the right sense?



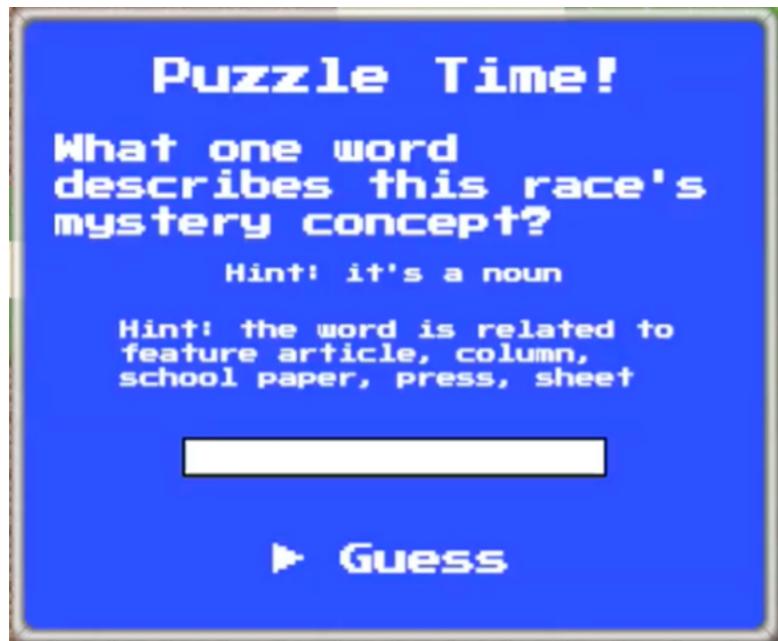
How long did players take to converge on the right sense?



Three flights takes under a minute, which is equivalent to the annotation speed of experts

(Krishnamurthy and Nicholls, 2000)

What went right?



Game elements proved fun and addicting (for us too)



Identified reusable patterns for taskifying games

What could have gone better?

- Game development is hard if you have no experience
 - 2 Months for Puzzle Racer vs. 1 week for Ka-boom!
- Still needed manual annotation to bootstrap the games
- Games were slower than crowdsourcing
 - But only because we didn't have a ready pool of players

Games4NLP: State of the game

Games4NLP: State of the game

- So far, it looks that successful games rely on intuitions, visual stimuli, and quick associations
 - WSD “Fruit Ninja” – possible to **visualize words, fast** reactions
 - Puzzle Racer – making sense of **words** during a **race**
 - ZType – do not think, **just type**

Games4NLP: State of the game

- So far, it looks that successful games rely on intuitions, visual stimuli, and quick associations
 - WSD “Fruit Ninja” – possible to **visualize words, fast** reactions
 - Puzzle Racer – making sense of **words** during a **race**
 - ZType – do not think, **just type**

Games4NLP: State of the game

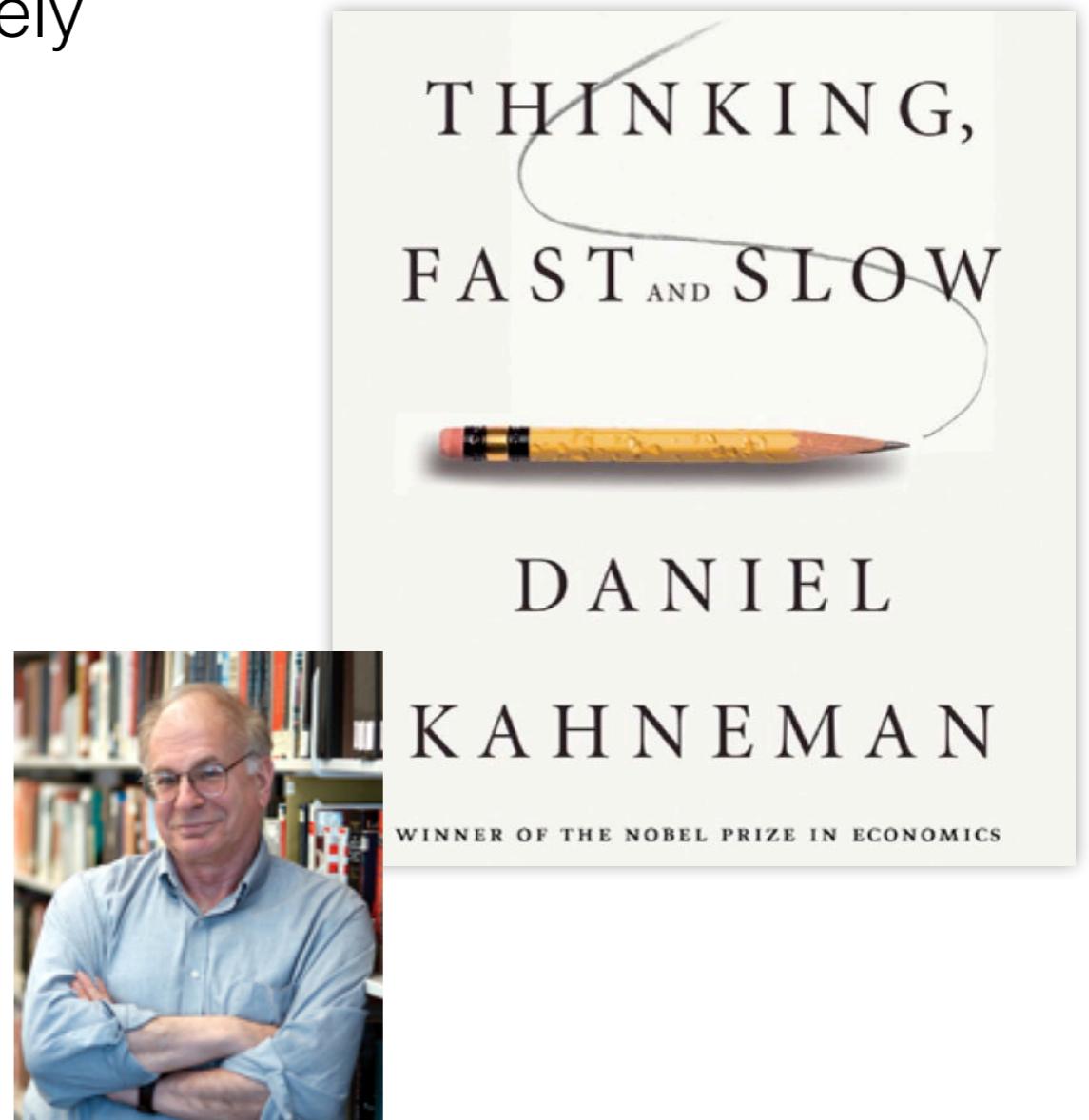
- So far, it looks that successful games rely on intuitions, visual stimuli, and quick associations
 - WSD “Fruit Ninja” – possible to **visualize words, fast** reactions
 - Puzzle Racer – making sense of **words** during a **race**
 - ZType – do not think, **just type**
- Sounds pretty much like “fast” thinking
 - as opposite to “slow” deliberative thinking

Games4NLP: State of the game

- So far, it looks that successful games rely on intuitions, visual stimuli, and quick associations
 - WSD “Fruit Ninja” – possible to **visualize words, fast** reactions
 - Puzzle Racer – making sense of **words** during a **race**
 - ZType – do not think, **just type**
- Sounds pretty much like “fast” thinking
 - as opposite to “slow” deliberative thinking

Games4NLP: State of the game

- So far, it looks that successful games rely on intuitions, visual stimuli, and quick associations
 - WSD “Fruit Ninja” – possible to **visualize words, fast** reactions
 - Puzzle Racer – making sense of **words** during a **race**
 - ZType – do not think, **just type**
- Sounds pretty much like “fast” thinking
 - as opposite to “slow” deliberative thinking
- But some NLP tasks actually need some “deep” thinking – **now what?**



Computational Argumentation

on the NLP landscape

Argumentation:

Verbal, social, and rational activity aimed at **convincing** a reasonable critic of the acceptability of a **standpoint** by putting forward a constellation of one or more propositions to justify this standpoint (van Eeemer et al., 2014)

Computational Argumentation

on the NLP landscape

UNDERSTANDING ONLINE STAR RATINGS:



<https://xkcd.com/1098/>

Argumentation:

Verbal, social, and rational activity aimed at **convincing** a reasonable critic of the acceptability of a **standpoint** by putting forward a constellation of one or more propositions to justify this standpoint (van Eemeren et al., 2014)

Computational Argumentation

on the NLP landscape

UNDERSTANDING ONLINE STAR RATINGS:

★★★★★ [HAS ONLY ONE REVIEW]

★★★★★ EXCELLENT

★★★★★ OK

★★★★☆]

★★★★☆

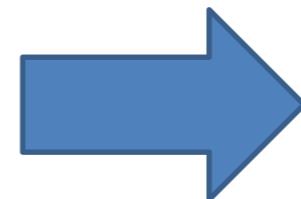
★★★★☆

★★★★☆ CRAP

★★★★☆

★★★★☆

★★★★☆



<https://xkcd.com/1098/>

Argumentation:

Verbal, social, and rational activity aimed at **convincing** a reasonable critic of the acceptability of a **standpoint** by putting forward a constellation of one or more propositions to justify this standpoint (van Eeemer et al., 2014)

Computational Argumentation

on the NLP landscape

UNDERSTANDING ONLINE STAR RATINGS:

★★★★★ [HAS ONLY ONE REVIEW]

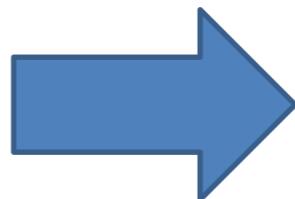
 EXCELLENT

OK

7

5 of 5

Page 5



ARE YOU COMING TO BED?

I CAN'T. THIS
IS IMPORTANT.

WHAT?

SOMEONE IS WRONG
ON THE INTERNET.



<https://xkcd.com/1098/>

<http://xkcd.com/386/>

Argumentation:

Verbal, social, and rational activity aimed at **convincing** a reasonable critic of the acceptability of a **standpoint** by putting forward a constellation of one or more propositions to justify this standpoint (van Eemeren et al., 2014)

(Computational) Argumentation: Argument

Introduction

- An argument is a **claim**, supported by **reasons**, intended to persuade

(Computational) Argumentation: Argument

Introduction

Claim

Physical education
should be mandatory in
schools

- An argument is a **claim**, supported by **reasons**, intended to persuade

(Computational) Argumentation: Argument

Introduction

Claim

Physical education
should be mandatory in
schools

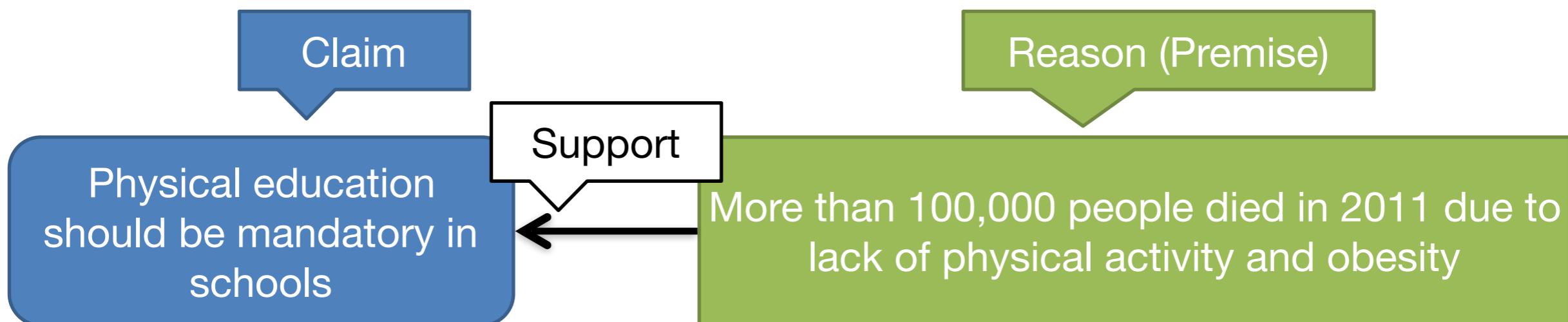
Reason (Premise)

More than 100,000 people died in 2011 due to
lack of physical activity and obesity

- An argument is a **claim**, supported by **reasons**, intended to persuade

(Computational) Argumentation: Argument

Introduction



- An argument is a **claim**, supported by **reasons**, intended to persuade

(Computational) Argumentation: Structures

Introduction

- Rebuttals: **attack** instead of support

(Computational) Argumentation: Structures

Introduction

Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet

attack instead of support

(Computational) Argumentation: Structures

Introduction

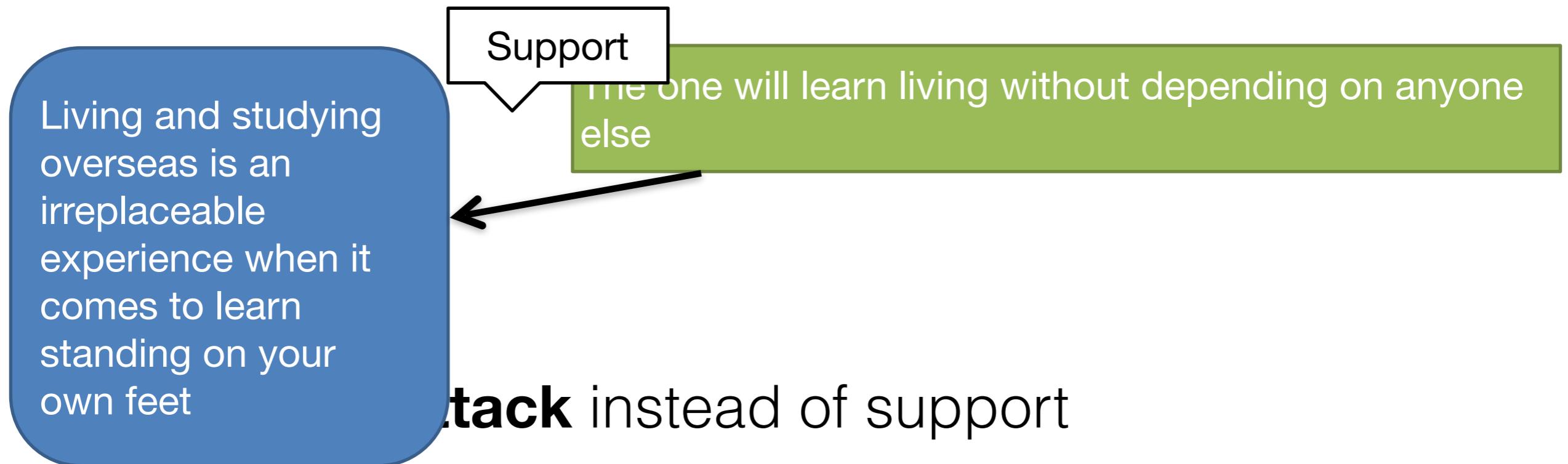
Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet

The one will learn living without depending on anyone else

Attack instead of support

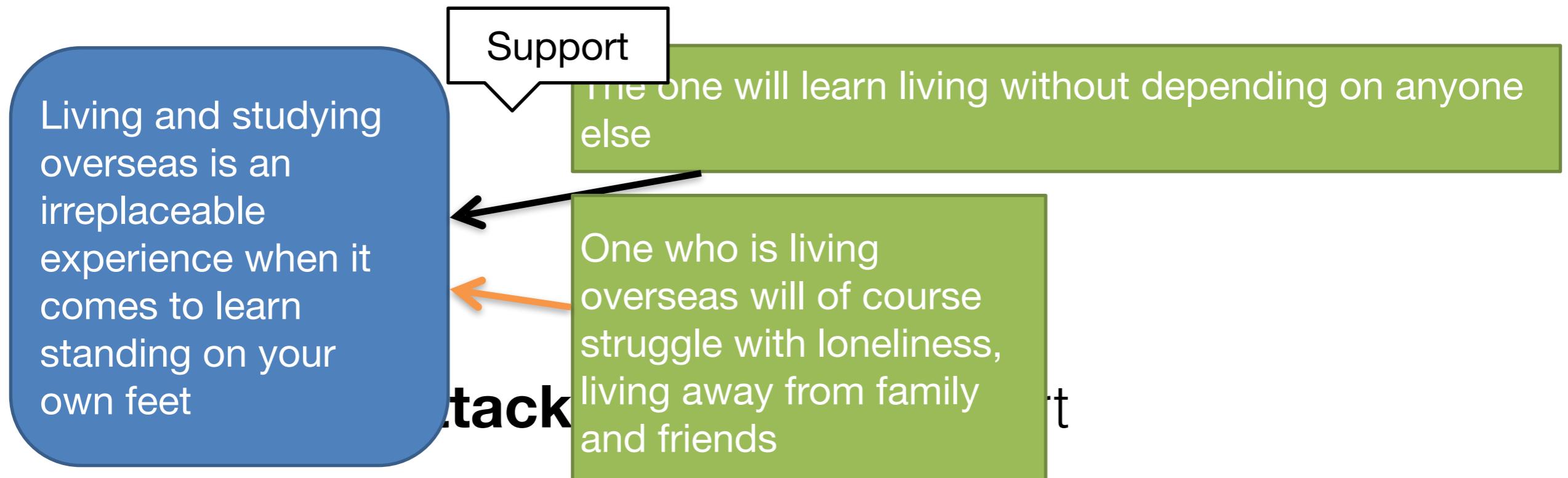
(Computational) Argumentation: Structures

Introduction



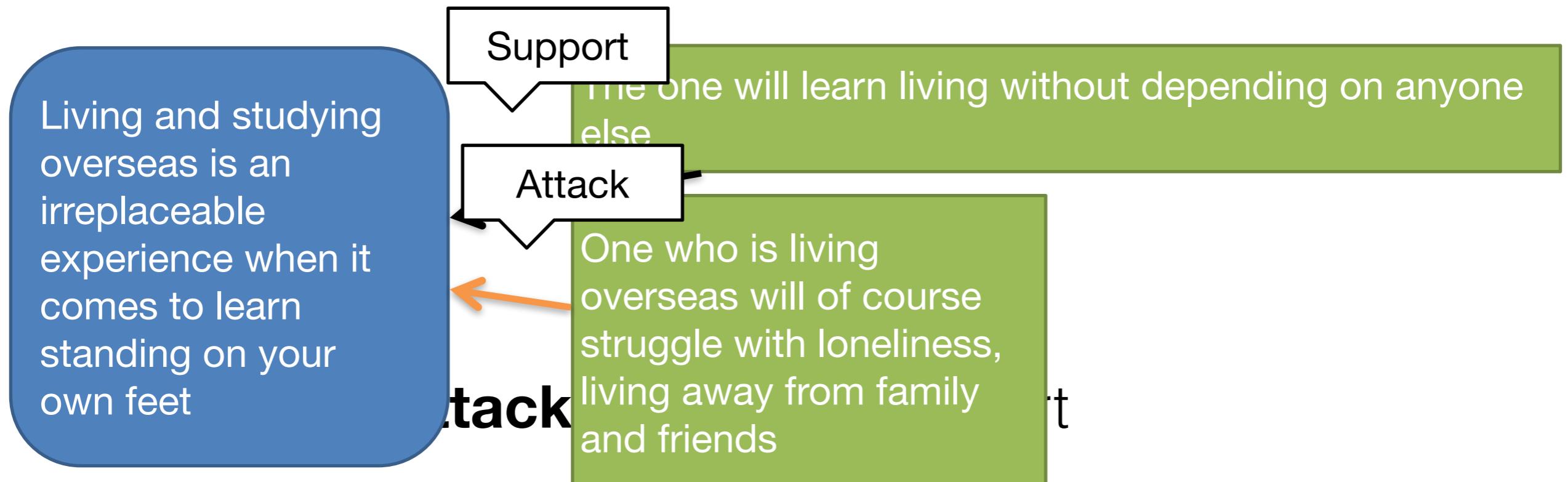
(Computational) Argumentation: Structures

Introduction



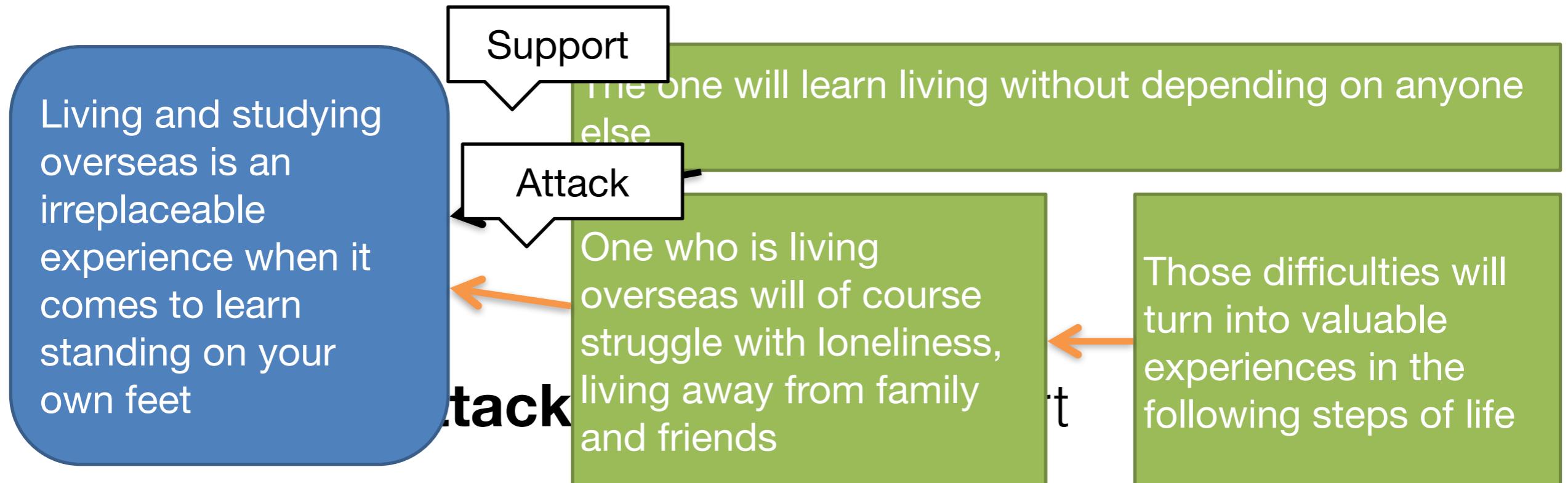
(Computational) Argumentation: Structures

Introduction



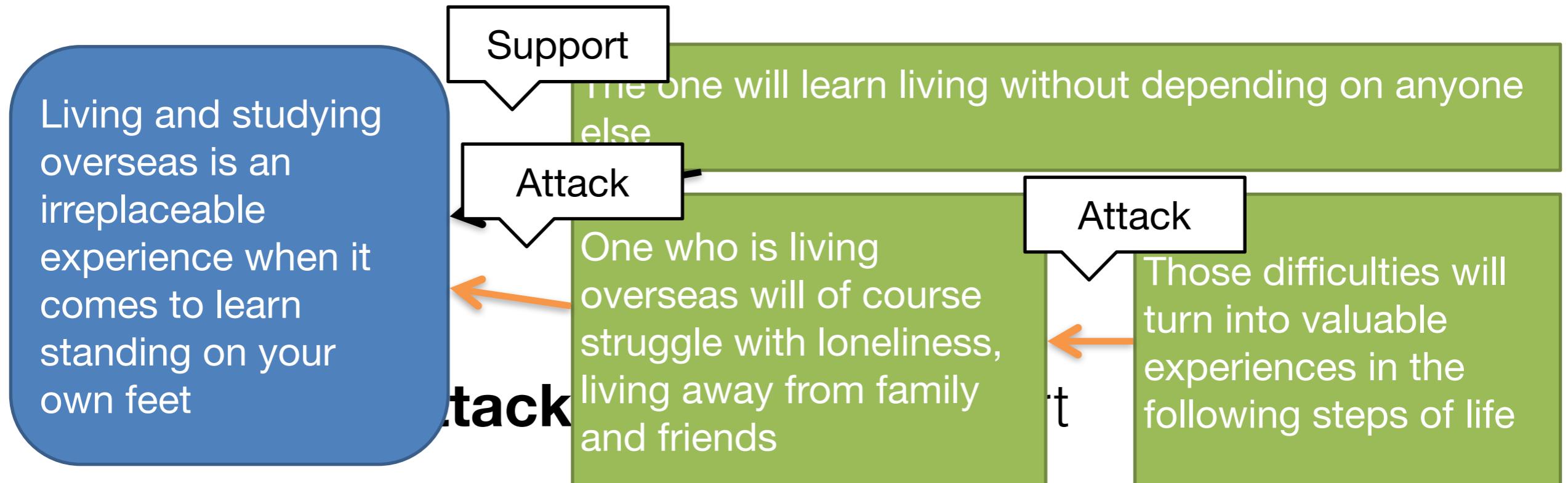
(Computational) Argumentation: Structures

Introduction



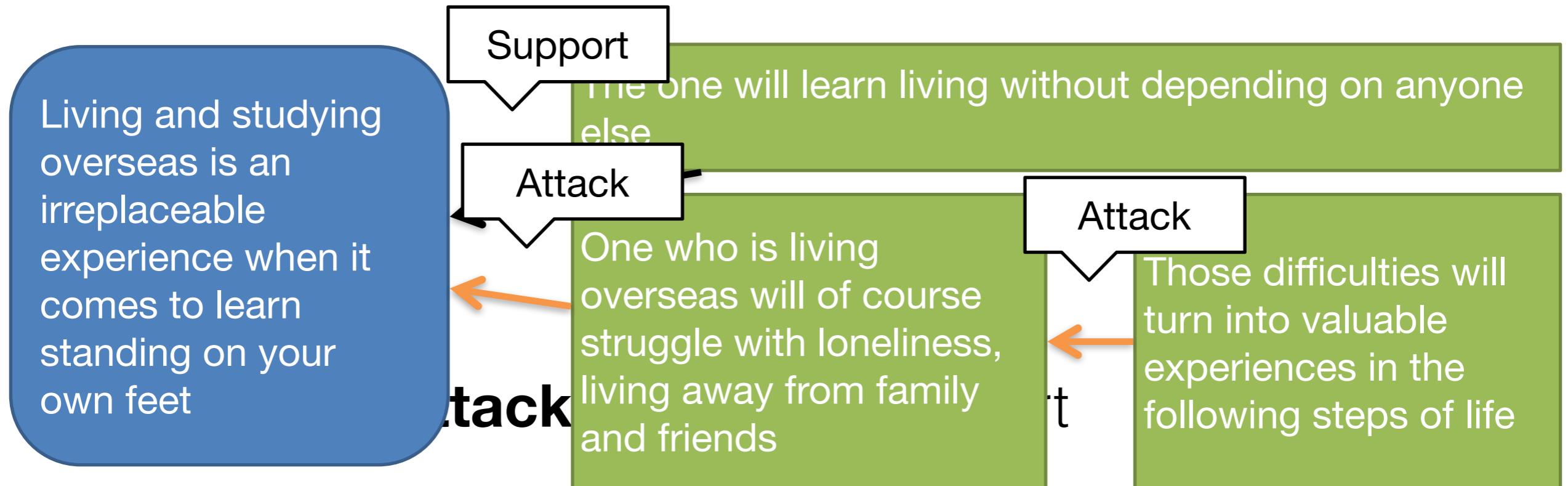
(Computational) Argumentation: Structures

Introduction



(Computational) Argumentation: Structures

Introduction



[...] Second, living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet. One who is living overseas will of course struggle with loneliness, living away from family and friends but those difficulties will turn into valuable experiences in the following steps of life. Moreover, the one will learn living without depending on anyone else. [...]

Computational Argumentation

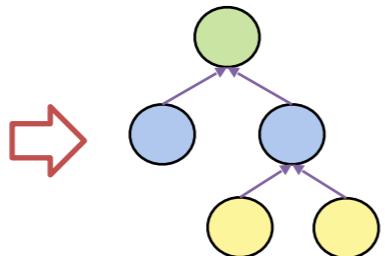
Variety of tasks, problems, and perspectives

Computational Argumentation

Variety of tasks, problems, and perspectives

Finding, “mining”, analyzing arguments and their structure

↓
Lorem ipsum dolor sit amet, consetetur
sadipscing elitr, sed diam nonumy
eirmod tempor invidunt ut labore et
dolore magna aliquyam erat, sed diam
voluptua. At vero eos et accusam et justo
duo dolores et ea rebum. Stet clita kasd
gubergren, no sea takimata sanctus est
Lorem ipsum dolor sit amet. Lorem
ipsum dolor sit amet, consetetur
sadipscing elitr, sed diam nonumy
eirmod tempor invidunt ut labore et
dolore magna aliquyam erat, sed diam
voluptua. At vero eos et accusam et justo
duo dolores et ea rebum.

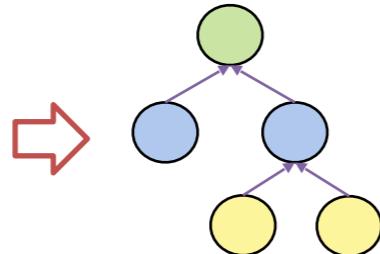


Computational Argumentation

Variety of tasks, problems, and perspectives

Finding, “mining”, analyzing arguments and their structure

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. **L**orem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum.



Assessing qualitative properties

Consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusamus et justus duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusamus et justus duo dolores et ea rebum.

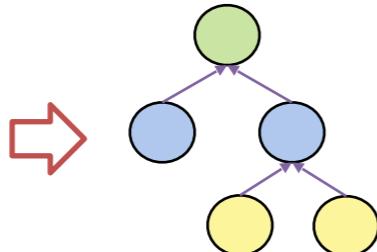


Computational Argumentation

Variety of tasks, problems, and perspectives

Finding, “mining”, analyzing arguments and their structure

↓
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum.



Assessing qualitative properties

↓
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum.



Explaining reasoning

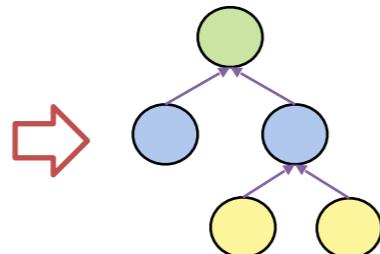


Computational Argumentation

Variety of tasks, problems, and perspectives

Finding, “mining”, analyzing arguments and their structure

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum.



Assessing qualitative properties

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum.



Explaining reasoning



Understanding social context



Computational Argumentation

From the serious game viewpoint

Computational Argumentation

From the serious game viewpoint

- Much beyond understanding a single word in context

Computational Argumentation

From the serious game viewpoint

- Much beyond understanding a single word in context
- Requires deliberative thinking

Computational Argumentation

From the serious game viewpoint

- Much beyond understanding a single word in context
- Requires deliberative thinking
- Even experts are bad at it (such as “find a claim”) (Paglieri, 2017)

Computational Argumentation

From the serious game viewpoint

- Much beyond understanding a single word in context
- Requires deliberative thinking
- Even experts are bad at it (such as “find a claim”) (Paglieri, 2017)

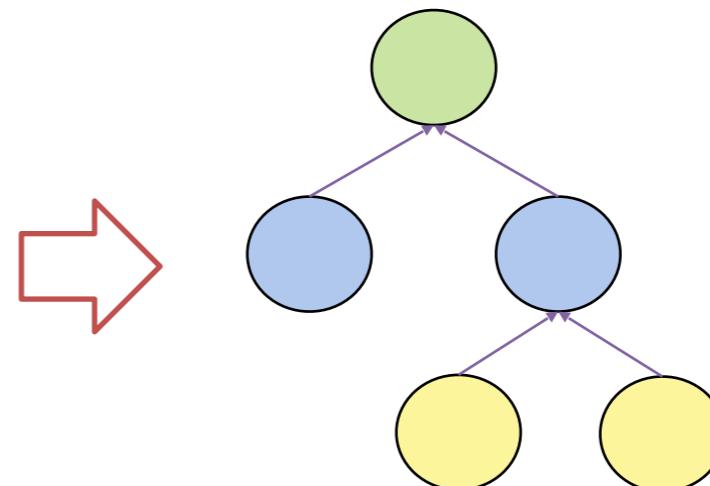


Argotario

Computational Argumentation meets Serious Games

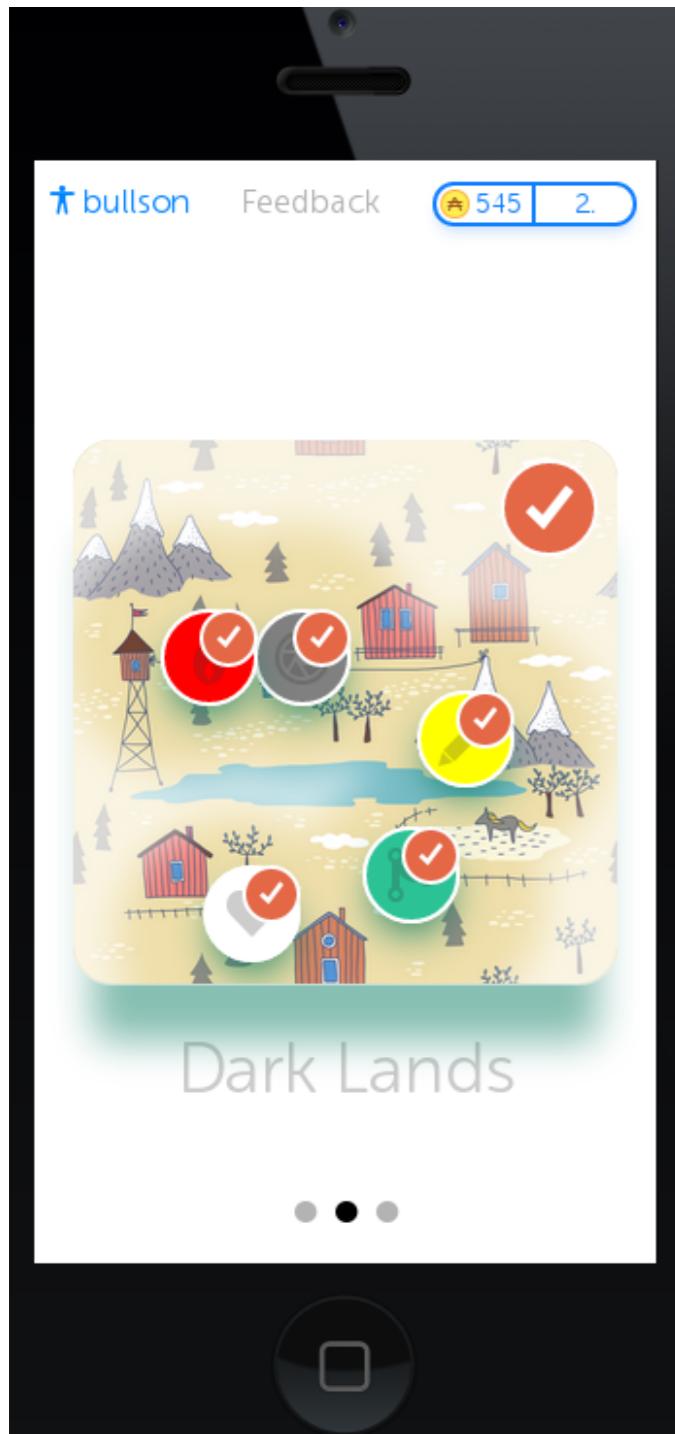
- Conceived with the following gaming idea
 - Find “premises” and “claims” in arguments
 - Identify the correct stance of an argument
 - Compose an argument
 - ...and earn points, get better in arguing!

Lorem ipsum dolor sit amet, consetetur
sadipscing elitr, sed diam nonumy
eirmod tempor invidunt ut labore et
dolore magna aliquyam erat, sed diam
voluptua. At vero eos et accusam et justo
duo dolores et ea rebum. Stet clita kasd
gubergren, no sea takimata sanctus est
Lorem ipsum dolor sit amet. Lorem
ipsum dolor sit amet, consetetur
sadipscing elitr, sed diam nonumy
eirmod tempor invidunt ut labore et
dolore magna aliquyam erat, sed diam
voluptua. At vero eos et accusam et justo
duo dolores et ea rebum.



Argotario

Design and usability have high priority



Built with modern full-stack technologies



Argotario (version 1)

Game rounds – easy tasks

Argotario (version 1)

Game rounds – easy tasks

Quit



Skip

Pro or Contra?

Only a part of the following argument is visible.
Do you think this is a Pro or a Contra argument?

"Are humans to blame for certain animal extinctions?"

Look at all the wonderful species that do not exist anymore because of the egoistic behavior of the human race.

Pro Contra



How many species are we losing?
On WWF.PANDA.ORG

[Tap here to read article](#)



Go On

Argotario (version 1)

Game rounds – easy tasks

- Guessing stance from incomplete arguments
- Finding “claim” or “reason”

Quit

450

Skip

Quit

420

Skip

Pro or Contra?

Only a part of the following argument is visible.
Do you think this is a Pro or a Contra argument?

“Are humans to blame for certain animal extinctions?”

Look at all the wonderful species that do not exist anymore because of the egoistic behavior of the human race.

Pro Contra

W How many species are we losing?
On WWF.PANDA.ORG

Tap here to read article

Can you pick the reason?

Read the following argument and tap on what you think is the reason of this argument!

“Are humans to blame for certain animal extinctions?”

Of course they are! Look at all the wonderful species that do not exist anymore because of the egoistic behavior of the human race.

W

How many species are we losing?
On WWF.PANDA.ORG

Tap here to read article

Go On



Go On

Argotario (version 1)

Game rounds – hard task

Quit  405 Skip

Read the following statement (Health) and compose a new argument!

"Should marijuana be legalized for individual use due to health condition?"

C It's 2015: Is weed legal in your state?
On CNN

Tap here to read article

Reorder Remove Components Help

Pro   Contra

Your Claim

/

Your Reason

/

Add another Reason



Go On

- Writing a new argument

Argotario (version 1)

Game rounds – hard task

Quit  405 Skip

Read the following statement (Health) and compose a new argument!

"Should marijuana be legalized for individual use due to health condition?"

 It's 2015: Is weed legal in your state?
On CNN
[Tap here to read article](#)

Reorder Remove Components Help

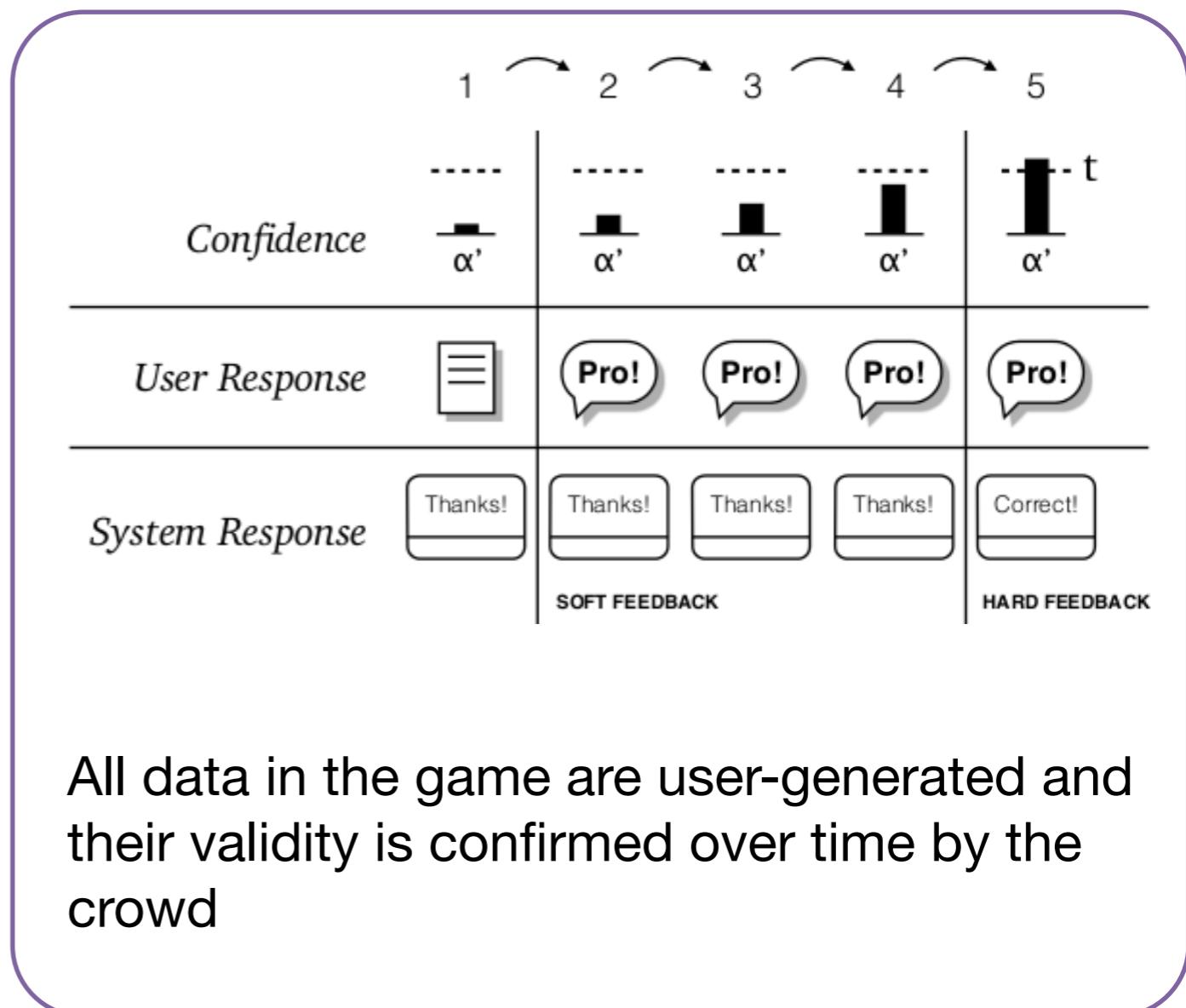
Pro  Contra 

Your Claim

Your Reason

Add another Reason

Go On



- Writing a new argument

Argotario (version 1)

How well we did?

Argotario (version 1)

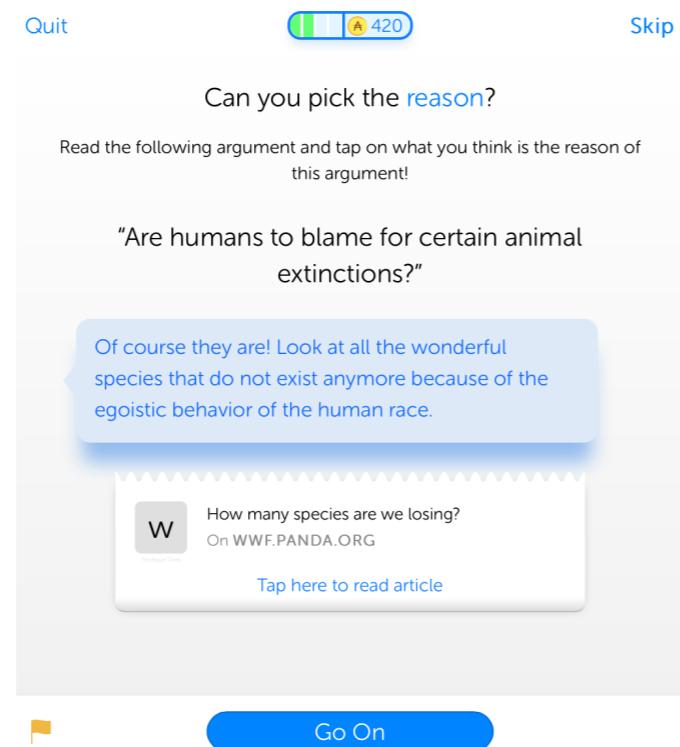
How well we did?

- The game was not really successful, despite its incentives (such as overall score rank)

Argotario (version 1)

How well we did?

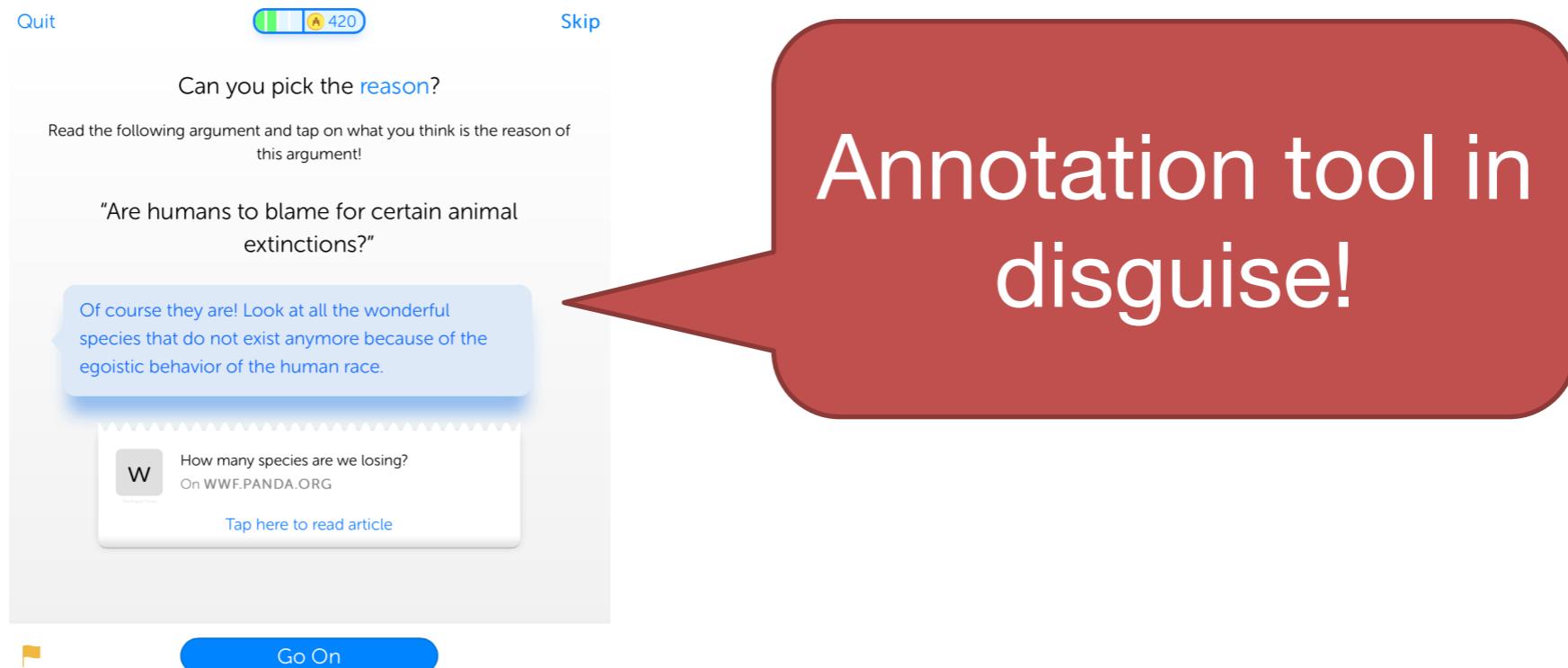
- The game was not really successful, despite its incentives (such as overall score rank)



Argotario (version 1)

How well we did?

- The game was not really successful, despite its incentives (such as overall score rank)



Argotario (version 1)

How well we did?

- The game was not really successful, despite its incentives (such as overall score rank)



- Reasons? We picked tasks which **people are usually bad at!**

Rethinking Argotario

“Pivoting”

Rethinking Argotario

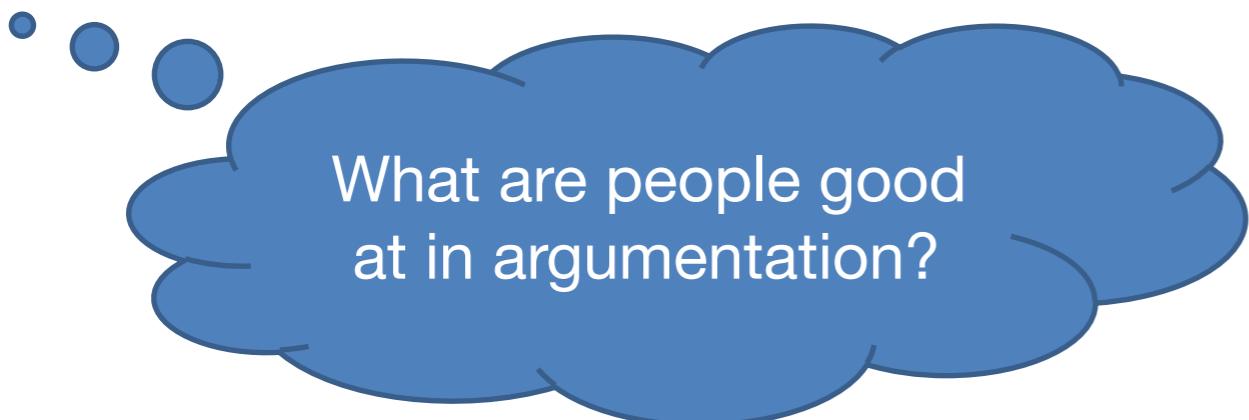
“Pivoting”

- Ask the right question

Rethinking Argotario

“Pivoting”

- Ask the right question



Rethinking Argotario

“Pivoting”

- Ask the right question



What are people good at in argumentation?



Spotting bad arguments!

Rethinking Argotario

“Pivoting”

- Ask the right question



What are people good at in argumentation?



Spotting bad arguments!

- We knew it even before reading the book by Mercier and Sperber (2017)

Rethinking Argotario

“Pivoting”

- Ask the right question



What are people good at in argumentation?



Spotting bad arguments!

- We knew it even before reading the book by Mercier and Sperber (2017)



But are “bad arguments” of any use?

Rethinking Argotario

“Pivoting”

- Ask the right question



What are people good at in argumentation?



Spotting bad arguments!

- We knew it even before reading the book by Mercier and Sperber (2017)



But are “bad arguments” of any use?



Have you heard of “fallacies”?

Rethinking Argotario

Introducing fallacies

- Fallacy is a prototypically bad argument, such as

Rethinking Argotario

Introducing fallacies

- Fallacy is a prototypically bad argument, such as



Ad Hominem

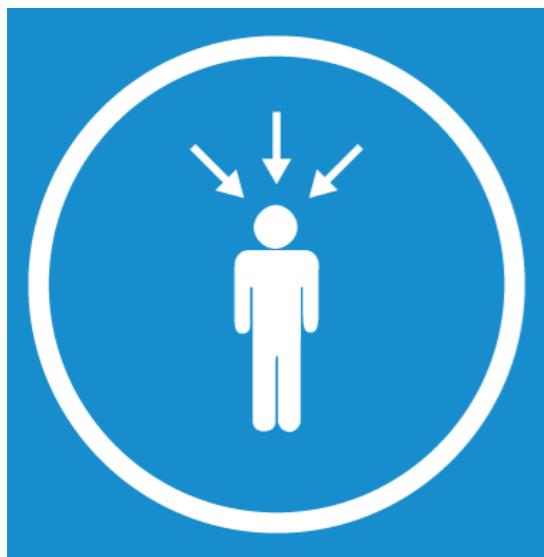
“Yeah, and you are a guy who loves war, that’s it. You like it when people die.”

Topic: Should the fight versus the Islamic State include military operations?

Rethinking Argotario

Introducing fallacies

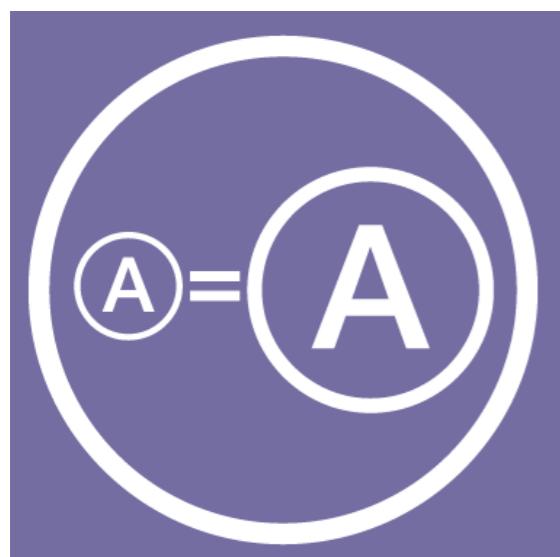
- Fallacy is a prototypically bad argument, such as



Ad Hominem

“Yeah, and you are a guy who loves war, that’s it. You like it when people die.”

Topic: Should the fight versus the Islamic State include military operations?



Hasty generalization

“Yes, Facebook is censoring racist comments against refugees. It works quite well. All media should be censored.”

Topic: Is it effective to censor parts of the media?

Rethinking Argotario

Introducing fallacies



Appeal to emotions

“Yes, all the polar-bears are dying, and we are next!”

Topic: Is global warming really an issue?

Rethinking Argotario

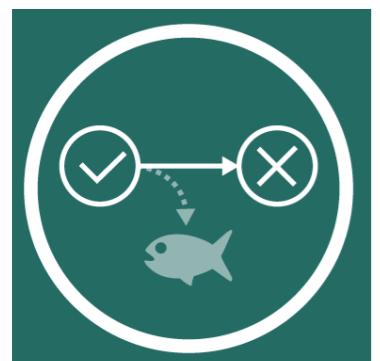
Introducing fallacies



Appeal to emotions

“Yes, all the polar-bears are dying, and we are next!”

Topic: Is global warming really an issue?



Red herring

“I am a hunter. Animals need to die in order to keep balance in the forest.”

Topic: Should we allow animal testing for medical purposes?

Rethinking Argotario

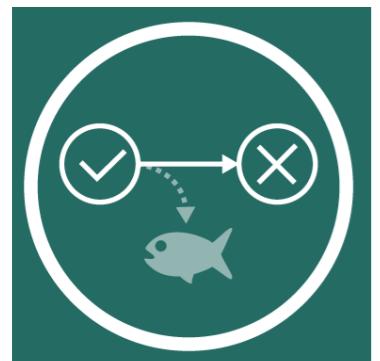
Introducing fallacies



Appeal to emotions

“Yes, all the polar-bears are dying, and we are next!”

Topic: Is global warming really an issue?



Red herring

“I am a hunter. Animals need to die in order to keep balance in the forest.”

Topic: Should we allow animal testing for medical purposes?



Irrelevant authority

“Yes, my husband has the same opinion”

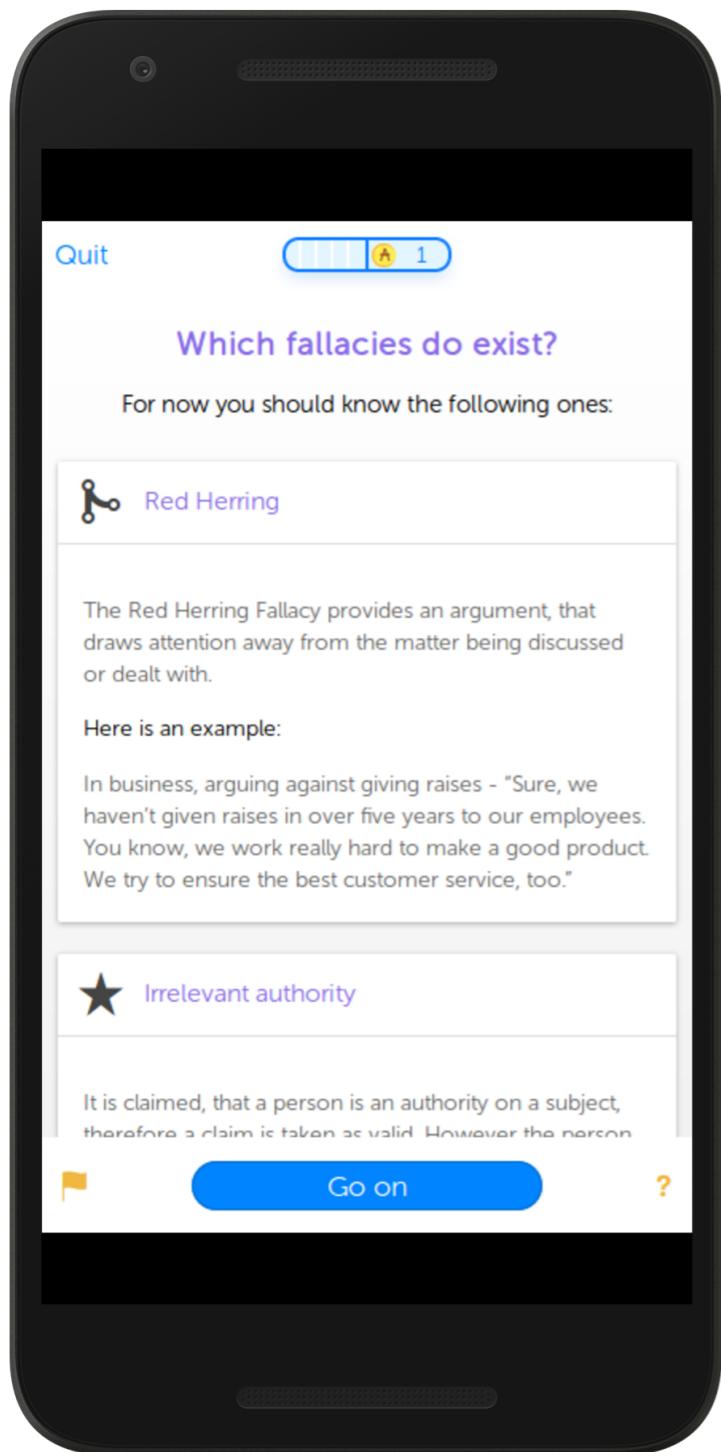
Topic: Is television an effective tool in building the minds of children?

Argotario

Learn, recognize, and write fallacies

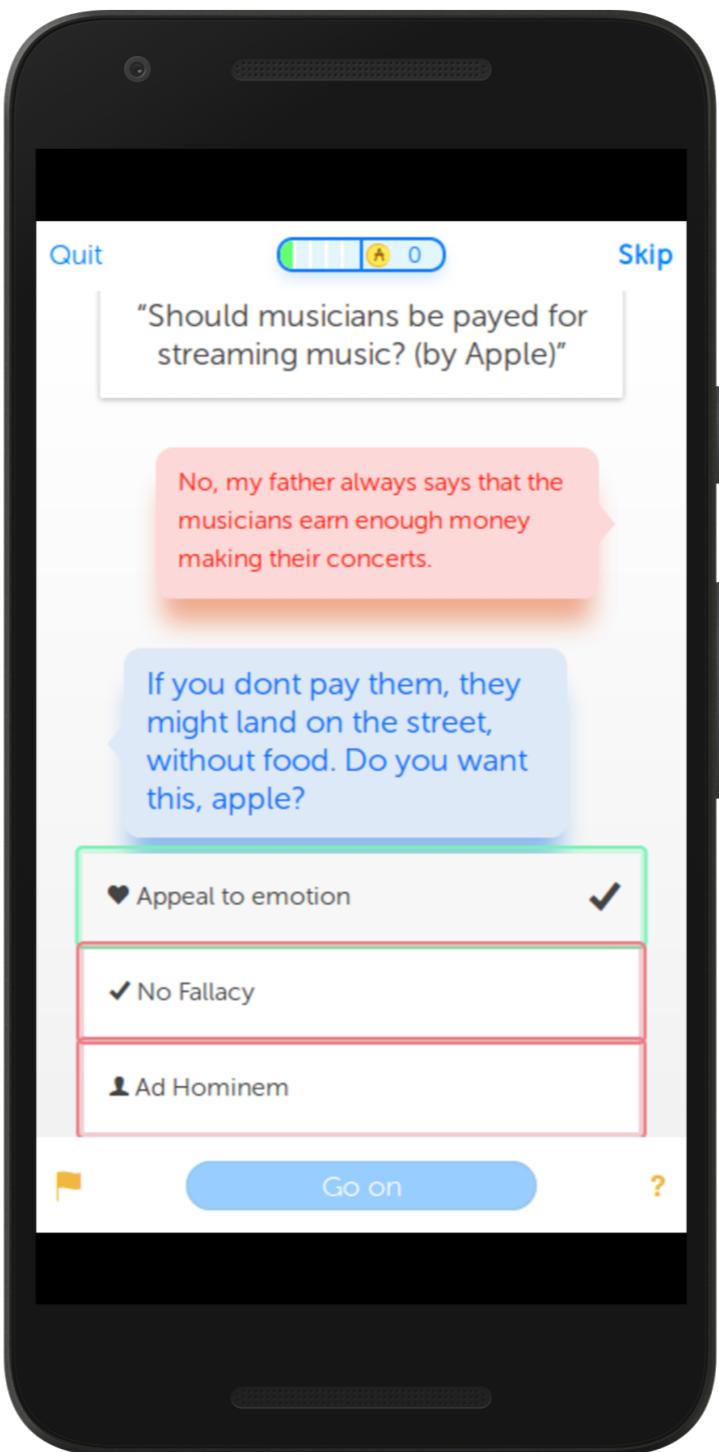
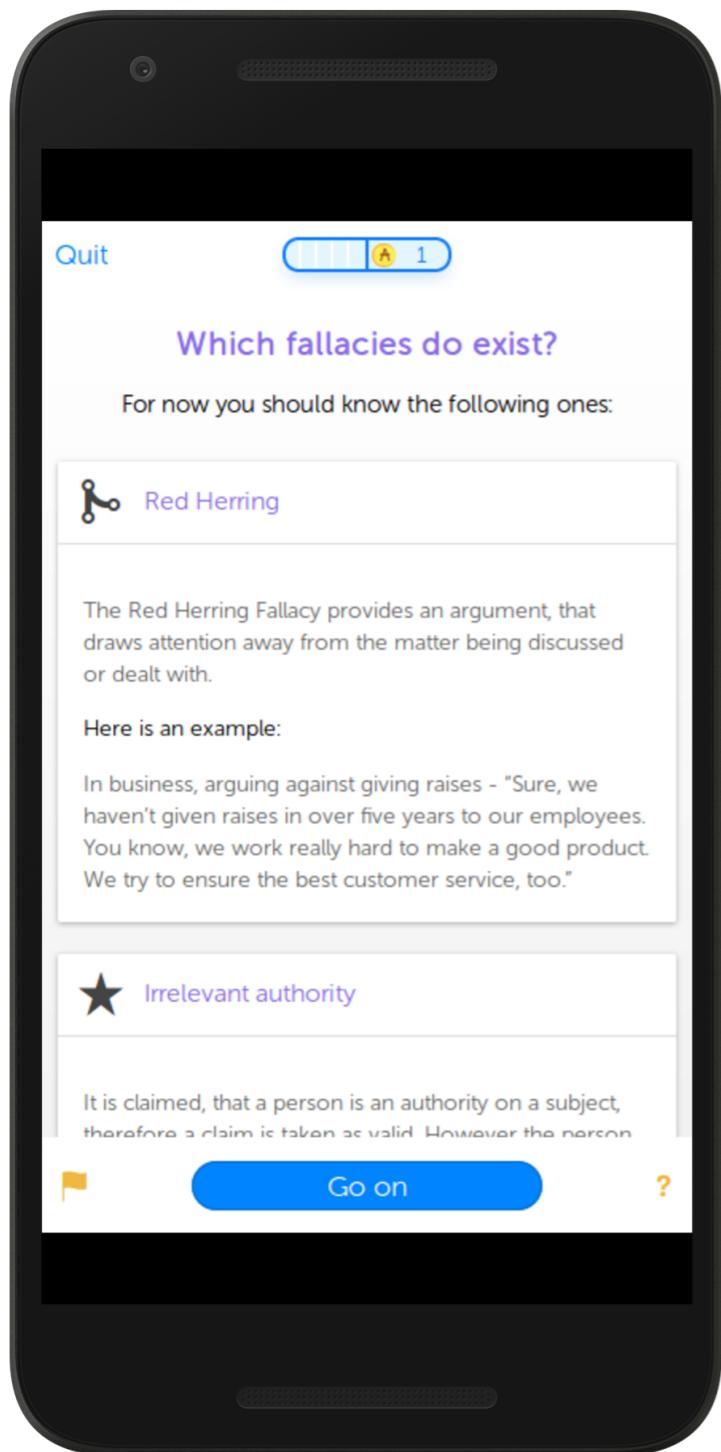
Argotario

Learn, recognize, and write fallacies



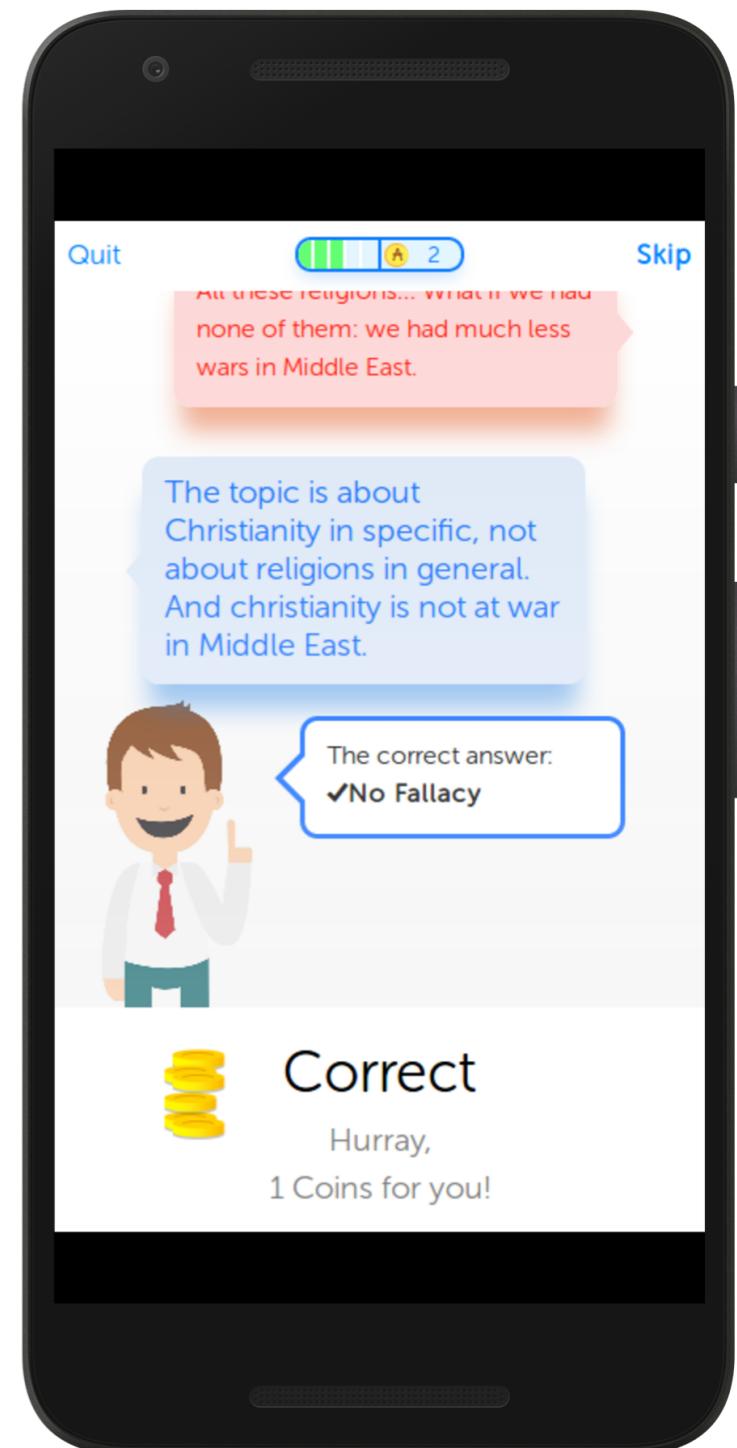
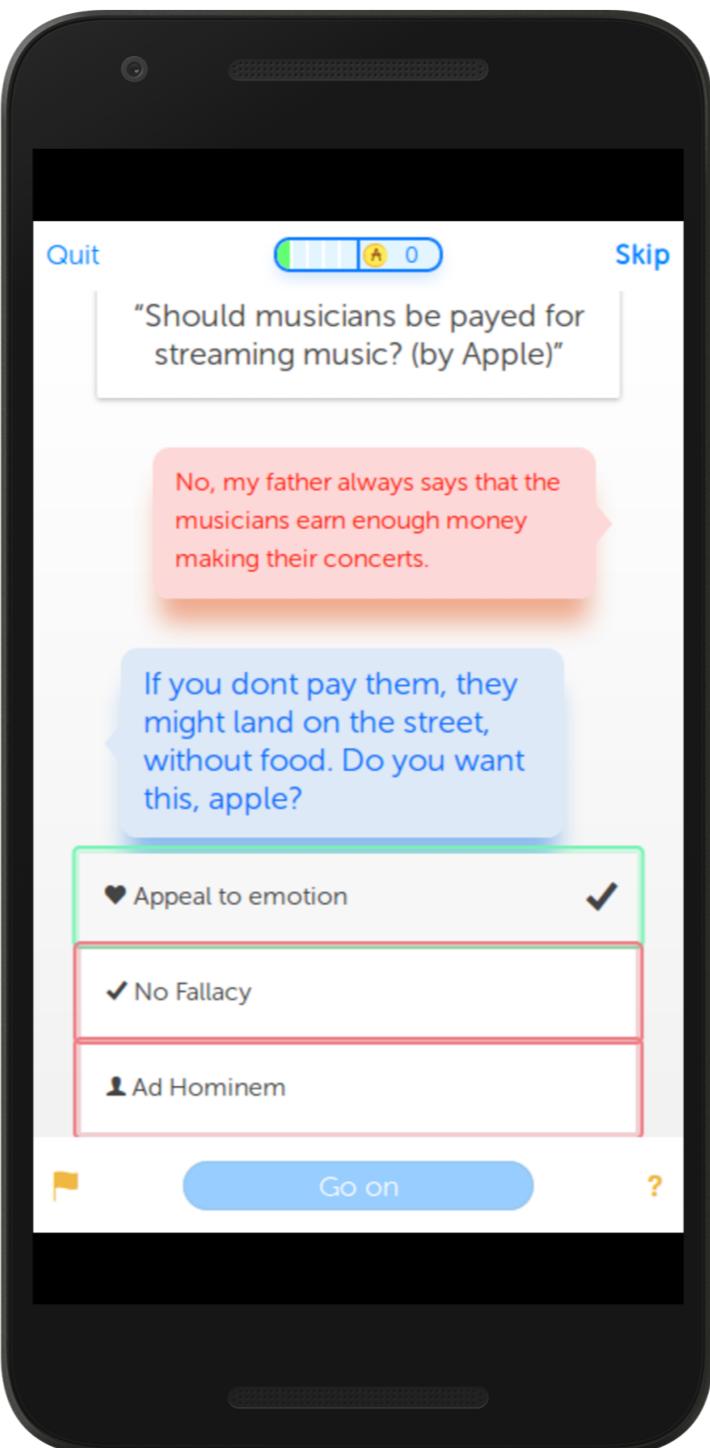
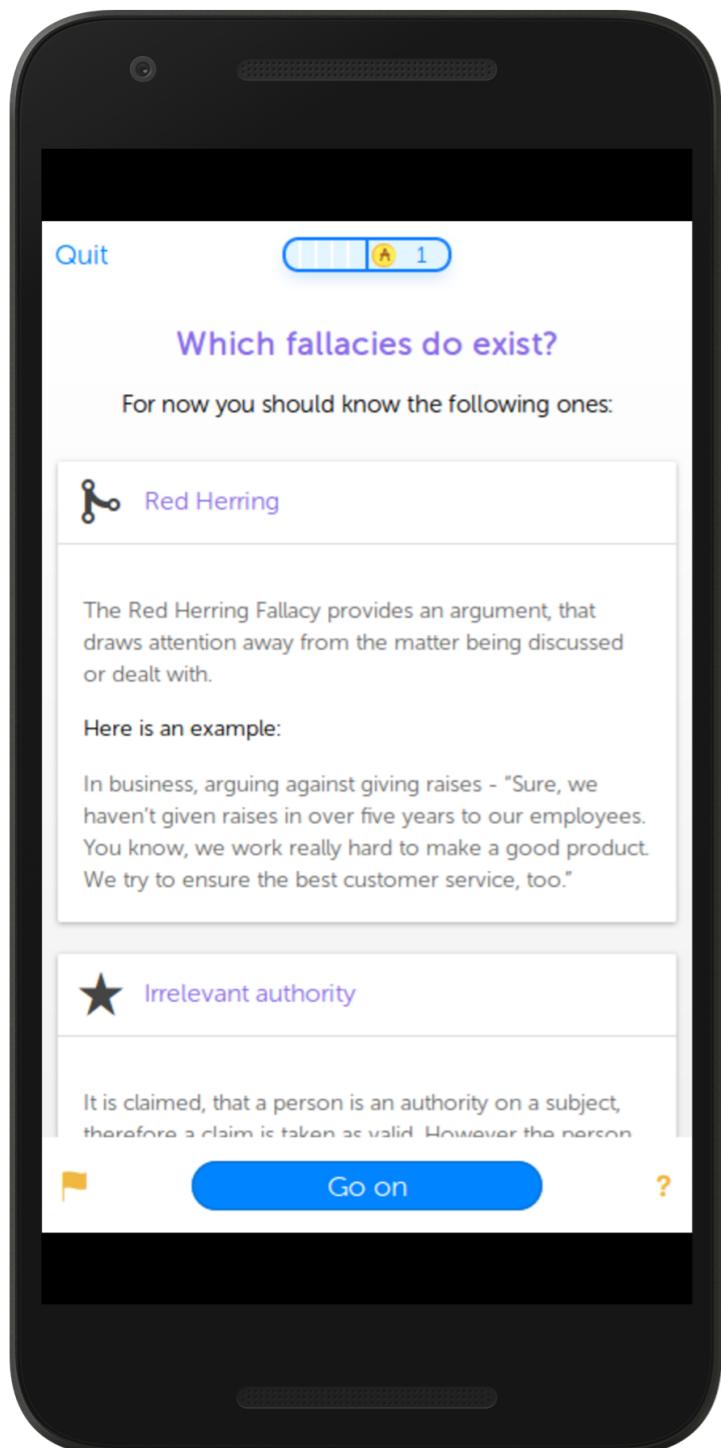
Argotario

Learn, recognize, and write fallacies



Argotario

Learn, recognize, and write fallacies

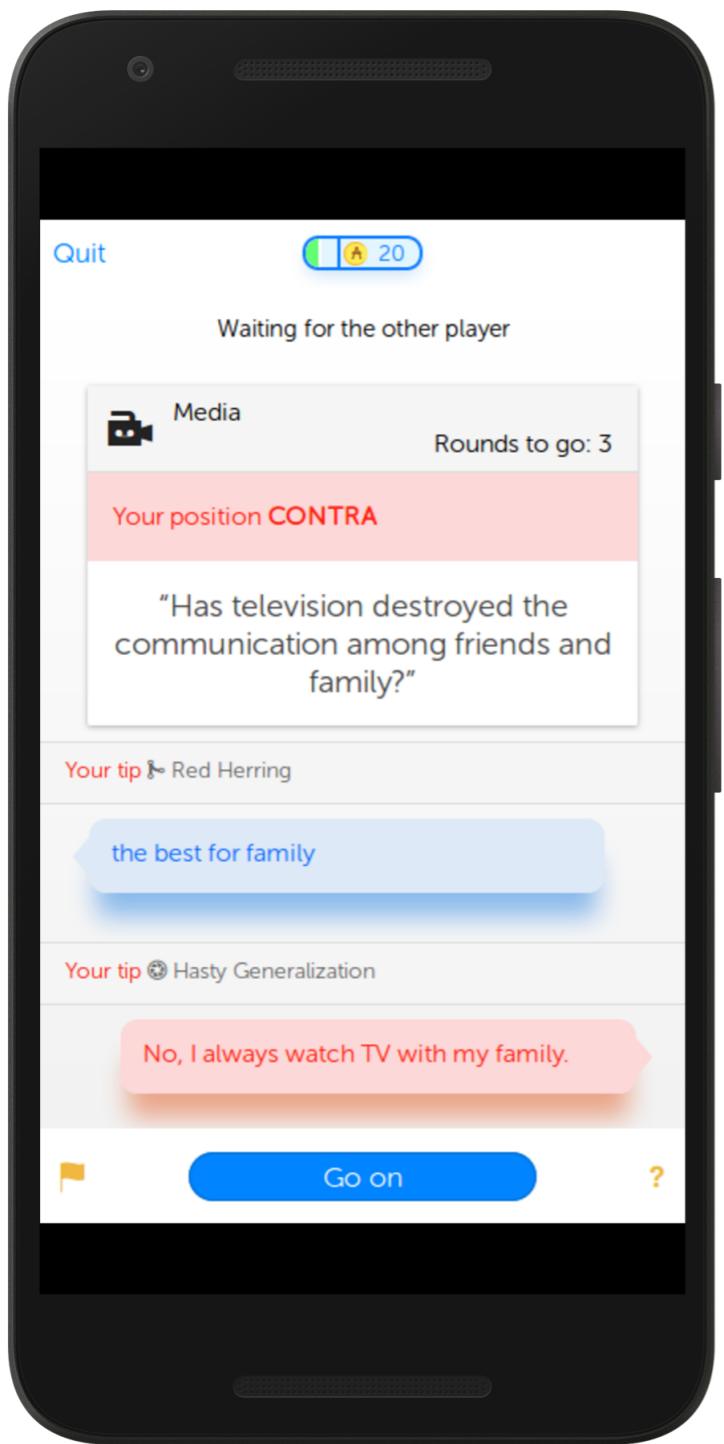


Argotario

Player versus player rounds

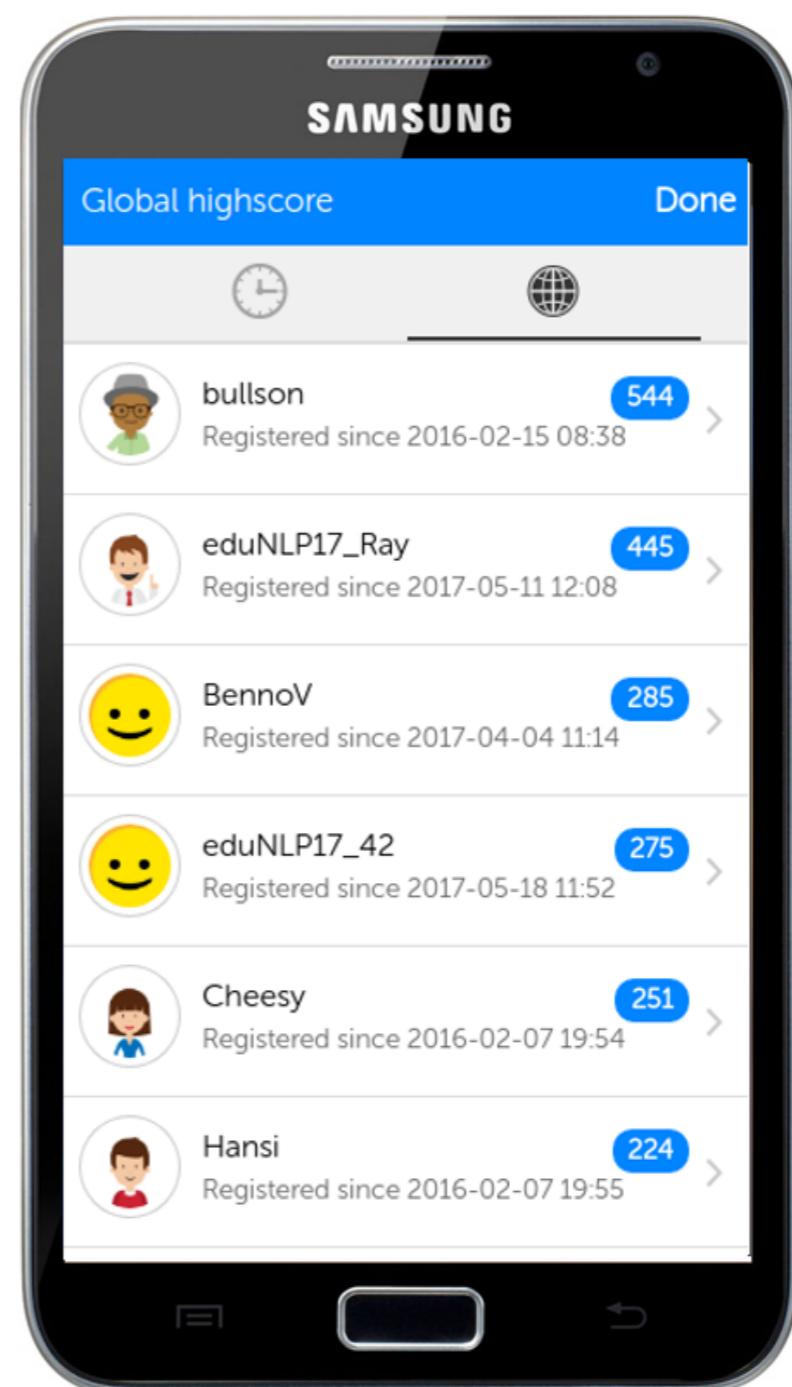
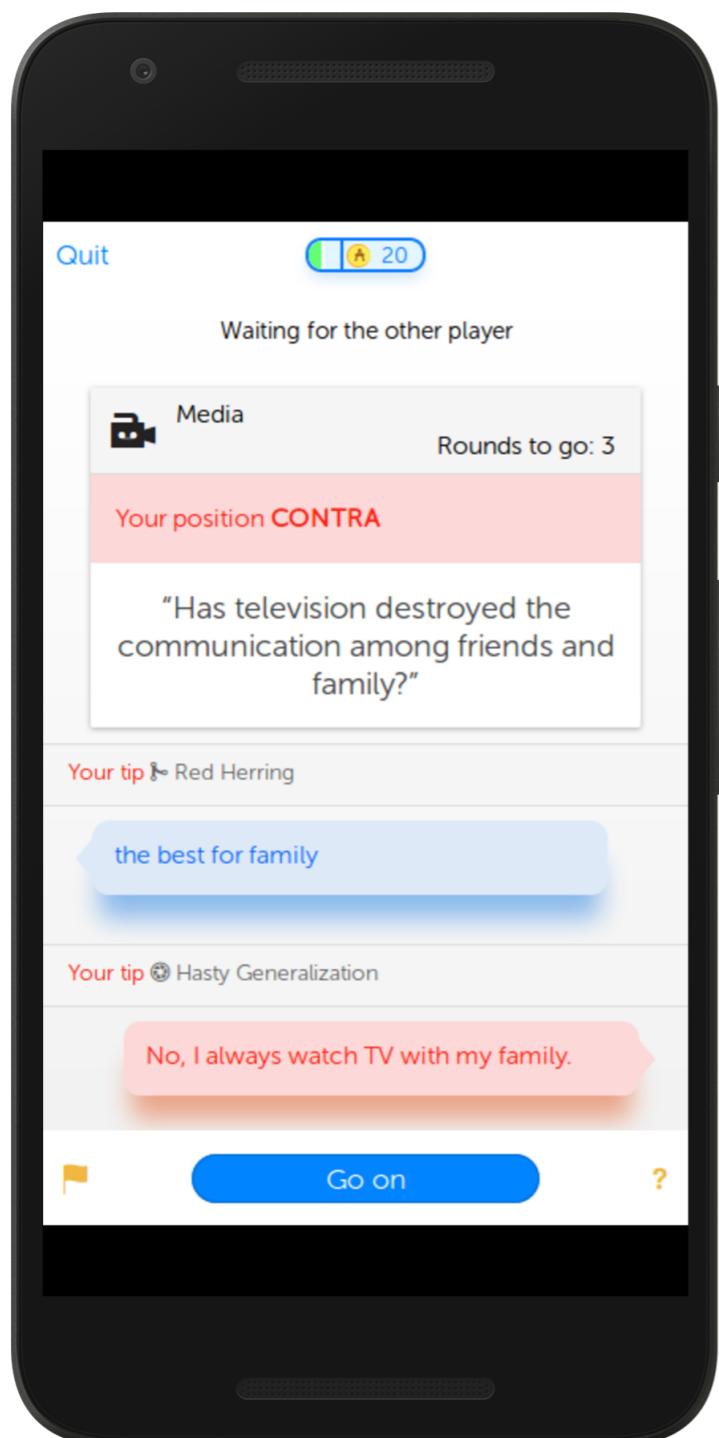
Argotario

Player versus player rounds



Argotario

Player versus player rounds



What have we learned?

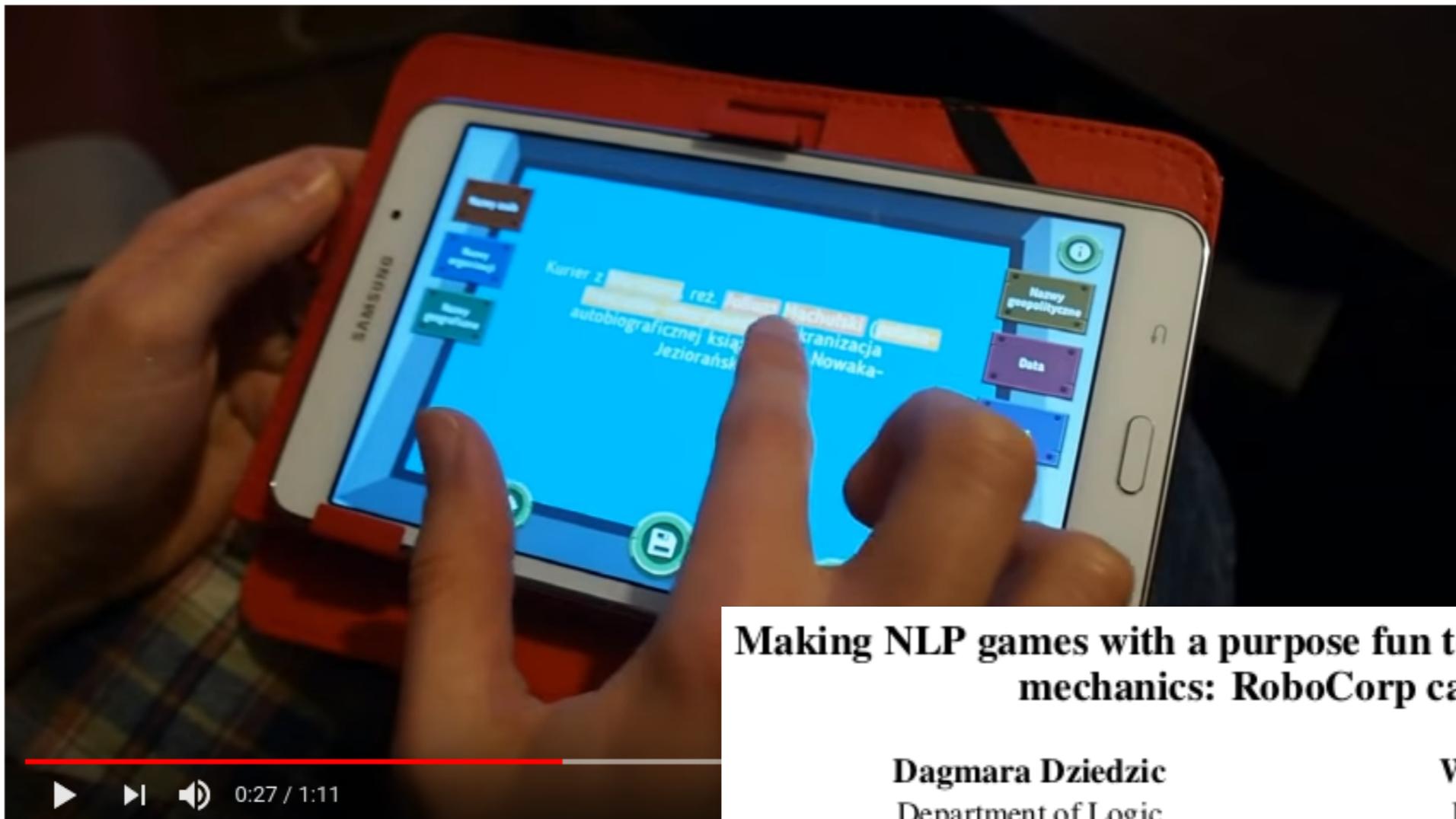
- Gamifying NLP tasks is hard but possible if done wisely
- “Taskifying” games is also hard because games are visual and NLP is mainly text



- Is there a better way?

Go where the users are

Mobile platforms, not computer screens



Making NLP games with a purpose fun to play using Free to Play mechanics: RoboCorp case study

Dagmara Dziedzic

Department of Logic
and Cognitive Science

Adam Mickiewicz University

Wieniawskiego 1, Poznań, Poland

dagmara.dziedzic@amu.edu.pl wojciech.wlodarczyk@amu.edu.pl

Wojciech Włodarczyk

Faculty of Mathematics
and Computer Science

Adam Mickiewicz University

Umultowska 87, Poznań, Poland

RoboCorp - Gra, która dzięki mądrości graczy rozwiąza

Leverage natural human potential

People are good at cooperating!



Leverage natural human potential

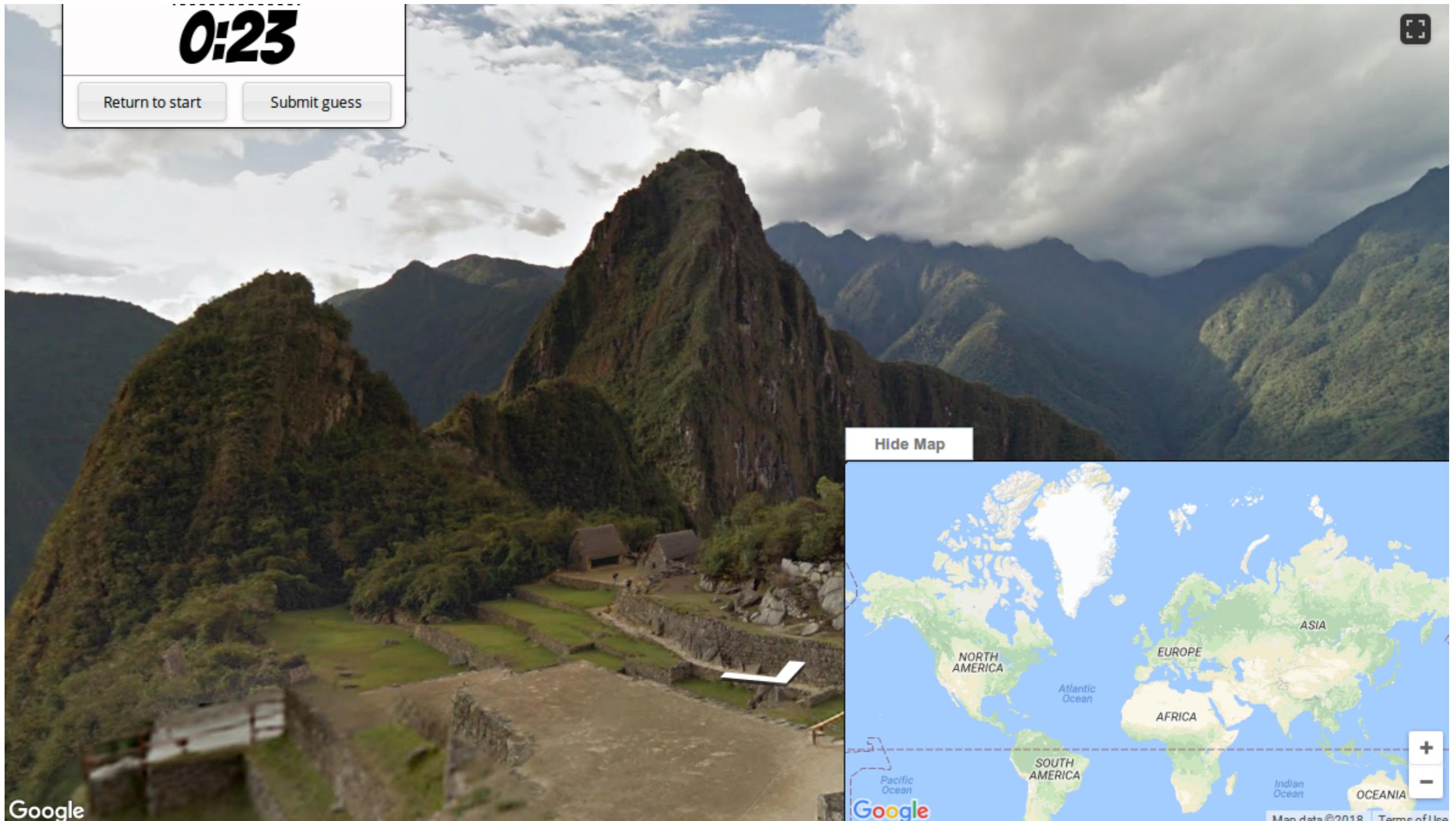
People are good at cooperating!



Niculae, V., & Danescu-Niculescu-Mizil, C. (2016). Conversational Markers of Constructive Discussions. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 568–578). San Diego, CA, USA: Association for Computational Linguistics.

Leverage natural human potential

People are good at cooperating!



Leverage natural human potential

People are good at cooperating!

3:26

I'm ready

Player Status

fritz AnotherTestPerson

Nobody made a guess yet. Click the map in the lower-right corner if you think you know the answer.
[Normal game.](#) **No spies in this game.** [\[?\]](#)

fritz: maybe maljsia?

Reasons for other players' guesses:

AnotherTestPerson: Asian'taknow

AnotherTestPerson: it's in asia

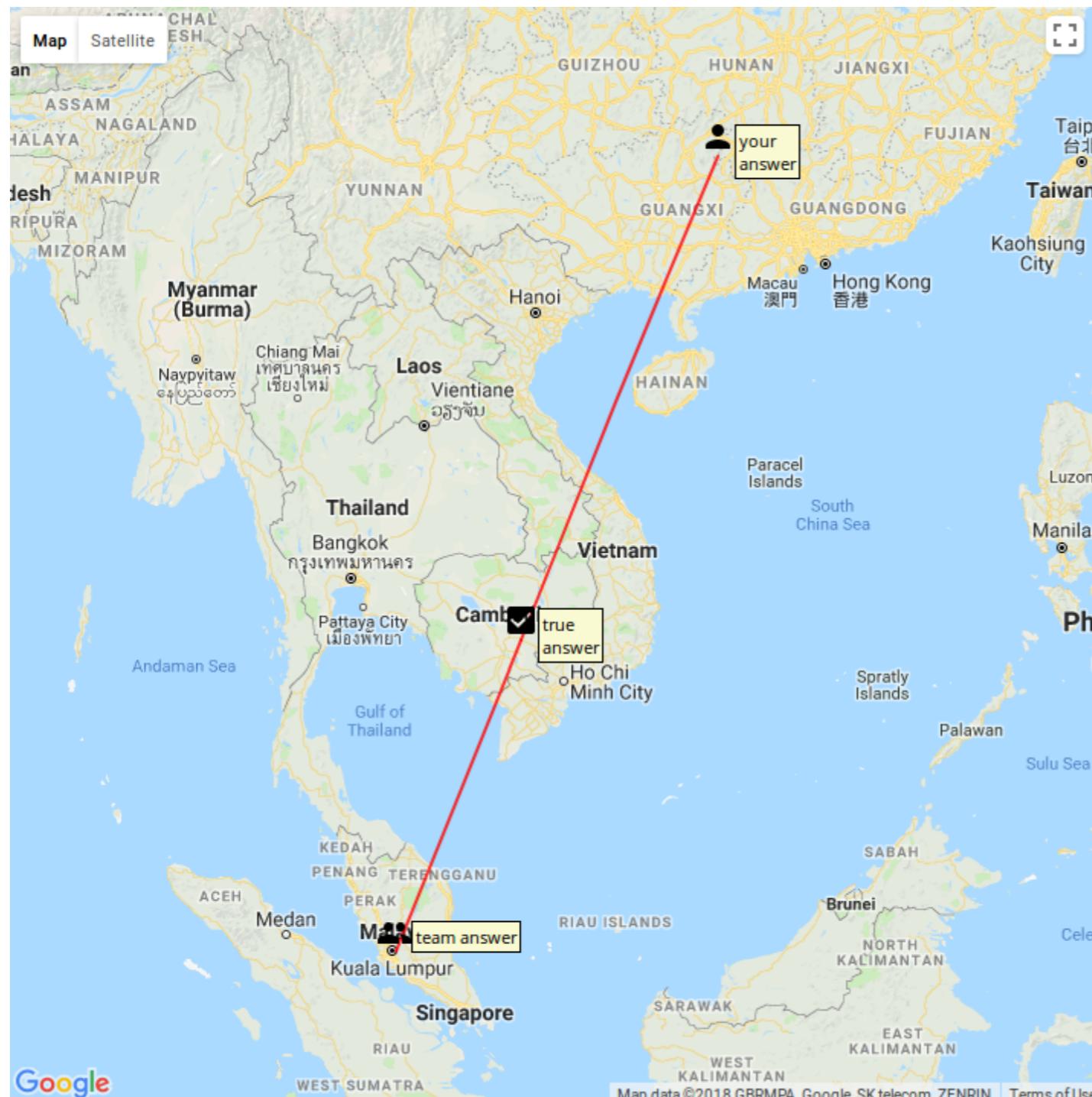
AnotherTestPerson: Could be Thailand?

fritz: Yes, would have been my second guess.

Chat with your team

Leverage natural human potential

People are good at cooperating!



YOUR SCORE

Show my teammates' guesses

The actual location is marked with .

Your team's answer was 1065.46 km away. This means you were better than 36% of the teams who played this map.

Your answer was 1579.87 km away. This means you were better than 43% of the people who played this map.

Congratulations! You won a site for finishing your first game!

Chat

|AnotherTestPerson has joined the chat.

Type your message

I want to play more!

Don't lose your progress! Connect with:

[Facebook](#) [Google](#) [Twitter](#)

Was it fun? Consider sharing StreetCrowd on , , or !

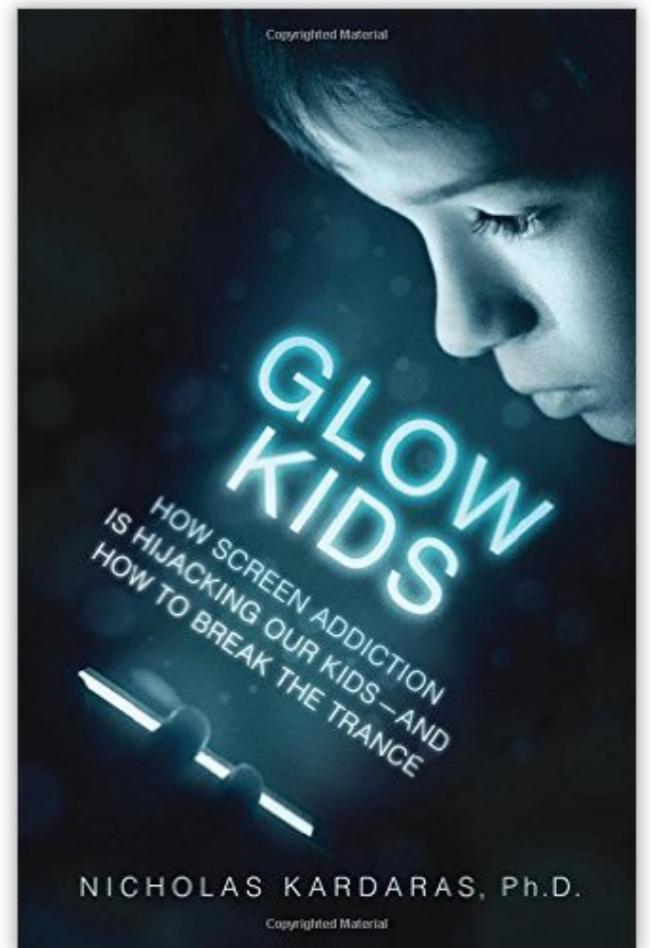
Leverage human desire for learning

Intrinsic motivation works in the long run

Leverage human desire for learning

Intrinsic motivation works in the long run

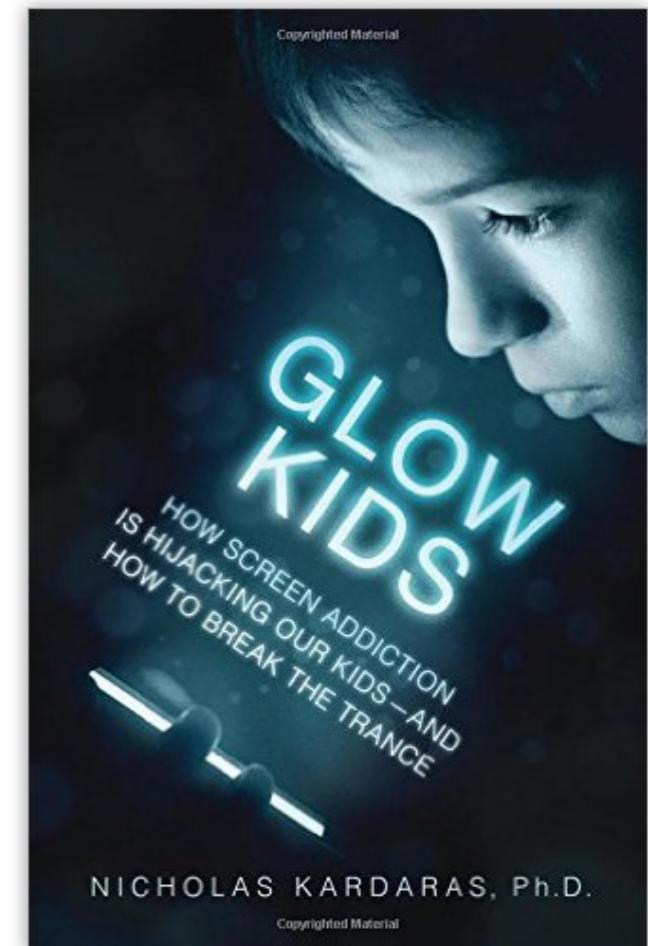
Educational games with extrinsic motivation (leaderboard, scores) get abandoned as soon as the motivation gets boring.



Leverage human desire for learning

Intrinsic motivation works in the long run

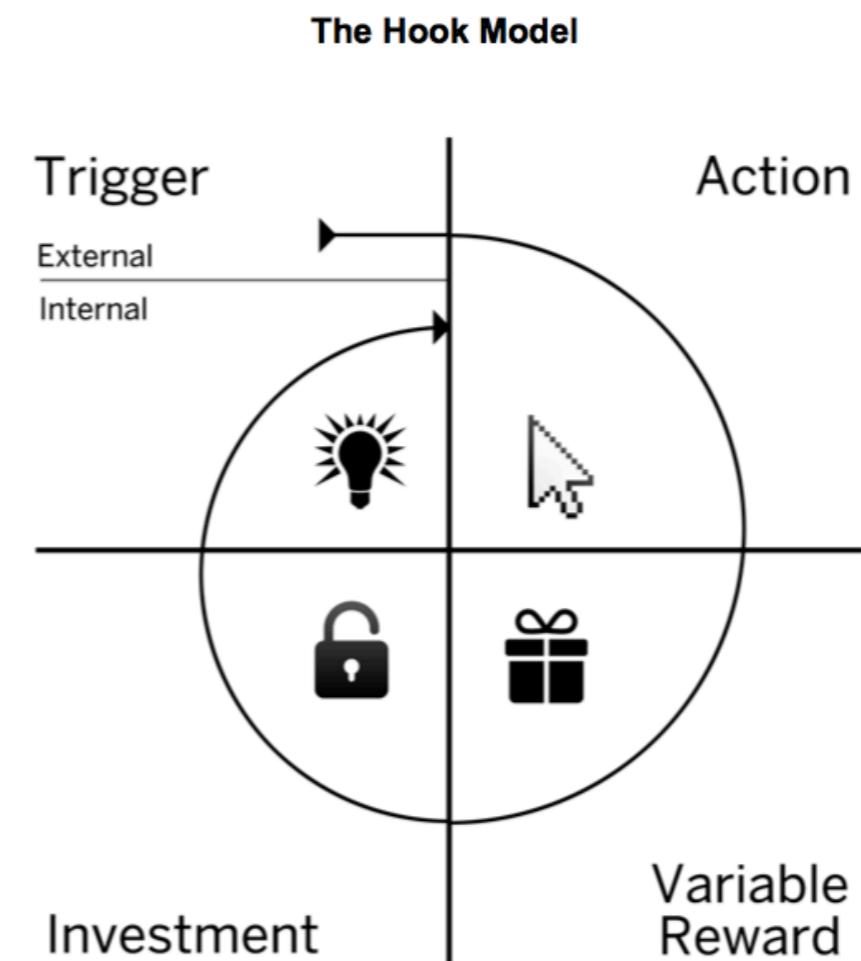
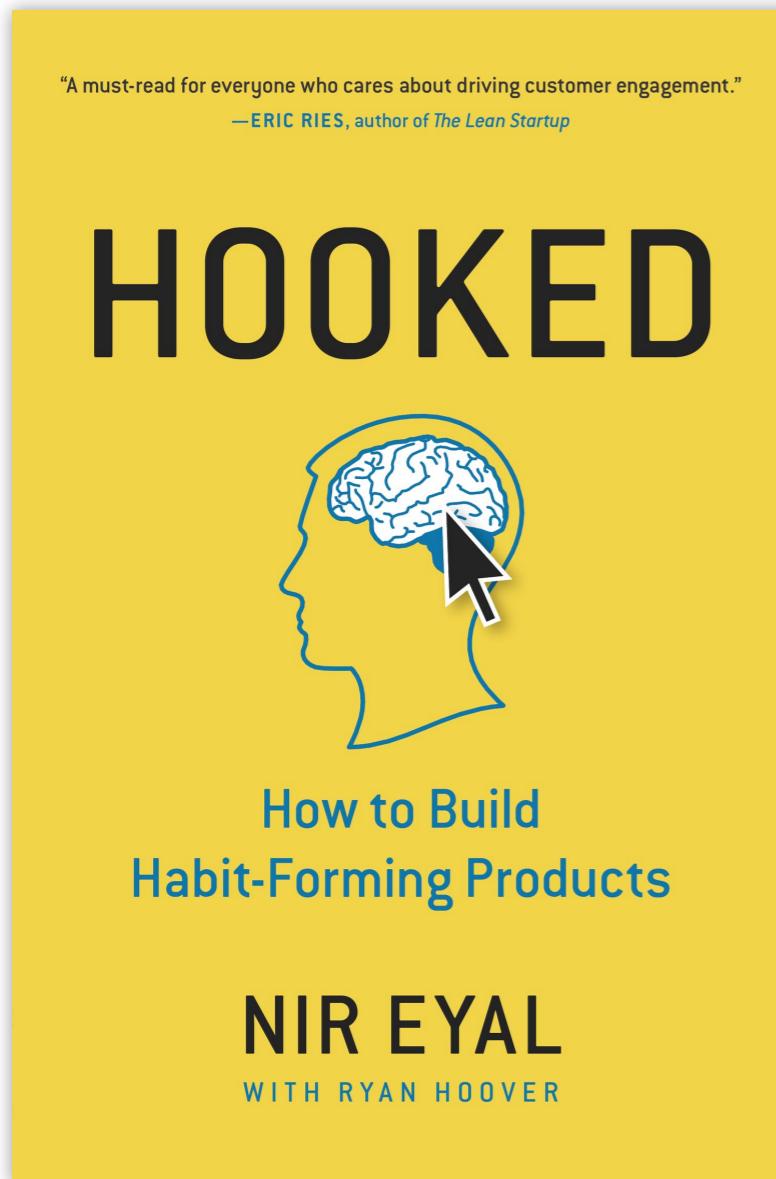
Educational games with extrinsic motivation (leaderboard, scores) get abandoned as soon as the motivation gets boring.



Duolingo uses gaming techniques but the intrinsic motivation is usually strong!

Leverage users' behavior (in a good way!)

Cognitive social psychology has many answers already



Overall Recommendations

Overall Recommendations

- Anything is better than nothing

Overall Recommendations

- Anything is better than nothing
- Give details on your study (who annotates and how?)

Overall Recommendations

- Anything is better than nothing
- Give details on your study (who annotates and how?)
- Use intensive training or professionals annotators

Overall Recommendations

- Anything is better than nothing
- Give details on your study (who annotates and how?)
- Use intensive training or professionals annotators
- Your classifier performance has an upper bound of your agreement

Overall Recommendations

- Anything is better than nothing
- Give details on your study (who annotates and how?)
- Use intensive training or professionals annotators
- Your classifier performance has an upper bound of your agreement
- Report also the agreement table/contingency matrix rather than only the obtained agreement

Overall Recommendations

- Anything is better than nothing
- Give details on your study (who annotates and how?)
- Use intensive training or professionals annotators
- Your classifier performance has an upper bound of your agreement
- Report also the agreement table/contingency matrix rather than only the obtained agreement
- Annotate with as many raters as possible, since it reduces the difference between the measures

Overall Recommendations

- Anything is better than nothing
- Give details on your study (who annotates and how?)
- Use intensive training or professionals annotators
- Your classifier performance has an upper bound of your agreement
- Report also the agreement table/contingency matrix rather than only the obtained agreement
- Annotate with as many raters as possible, since it reduces the difference between the measures
- Use Krippendorff's α

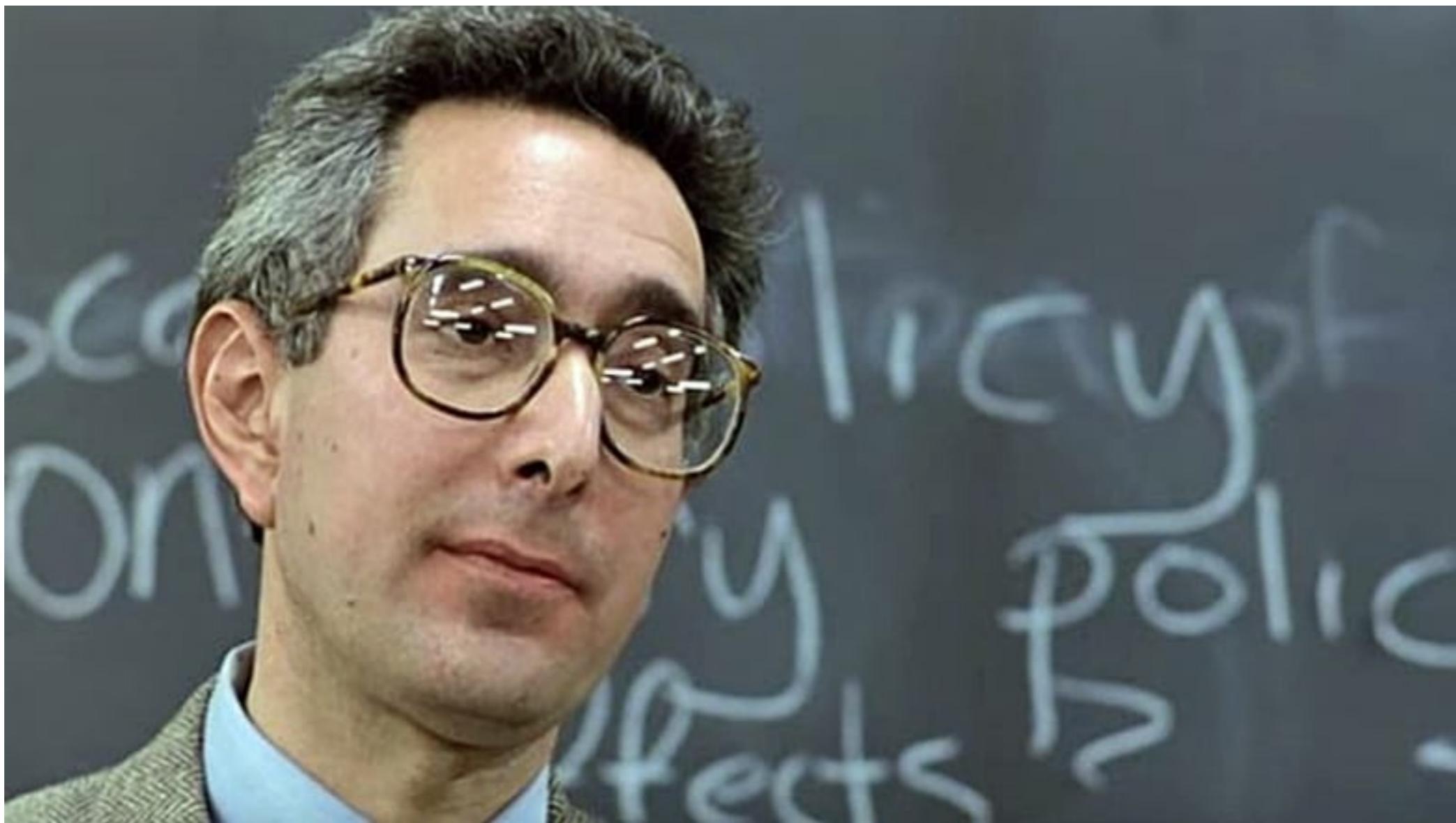
Overall Recommendations

- Anything is better than nothing
- Give details on your study (who annotates and how?)
- Use intensive training or professionals annotators
- Your classifier performance has an upper bound of your agreement
- Report also the agreement table/contingency matrix rather than only the obtained agreement
- Annotate with as many raters as possible, since it reduces the difference between the measures
- Use Krippendorff's α
- Be careful with weighted measures as they are hard to interpret

Overall Recommendations

- Anything is better than nothing
- Give details on your study (who annotates and how?)
- Use intensive training or professionals annotators
- Your classifier performance has an upper bound of your agreement
- Report also the agreement table/contingency matrix rather than only the obtained agreement
- Annotate with as many raters as possible, since it reduces the difference between the measures
- Use Krippendorff's α
- Be careful with weighted measures as they are hard to interpret
- Agreement should be above 0.8 to ensure data reliability (but depends on the case)

Student Projects (?)

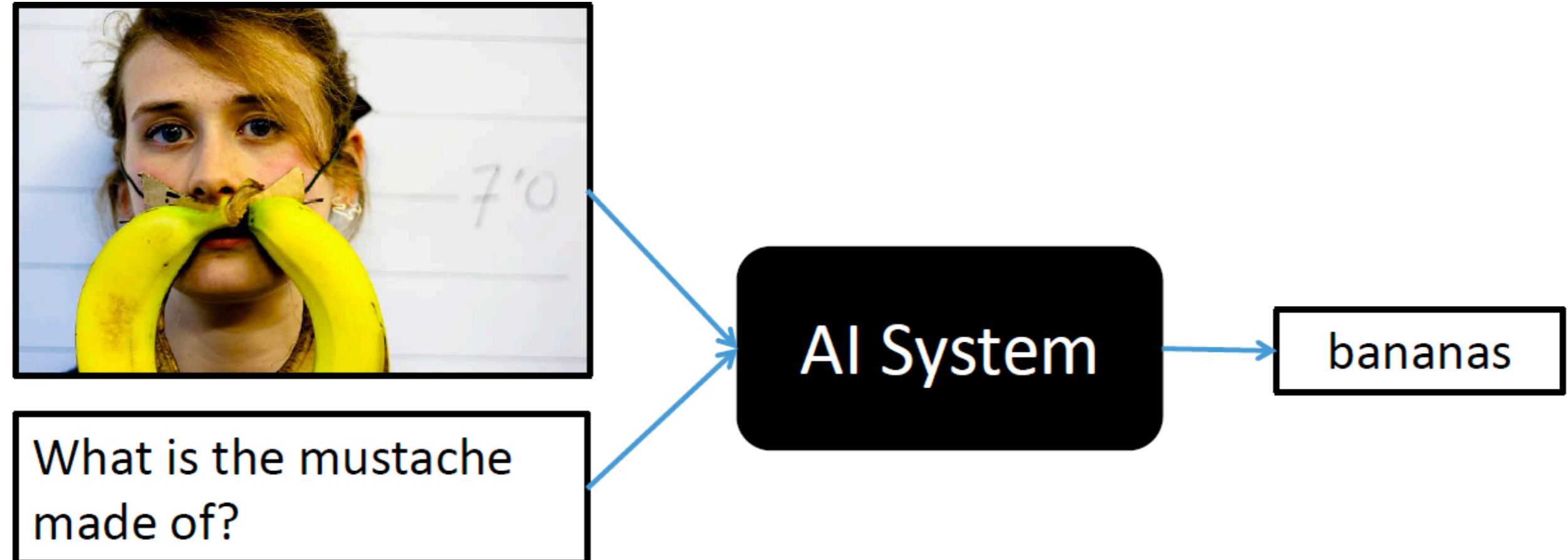






NLP is **too big** to cover in one class!

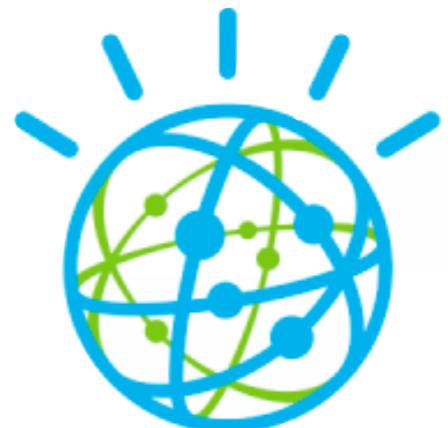
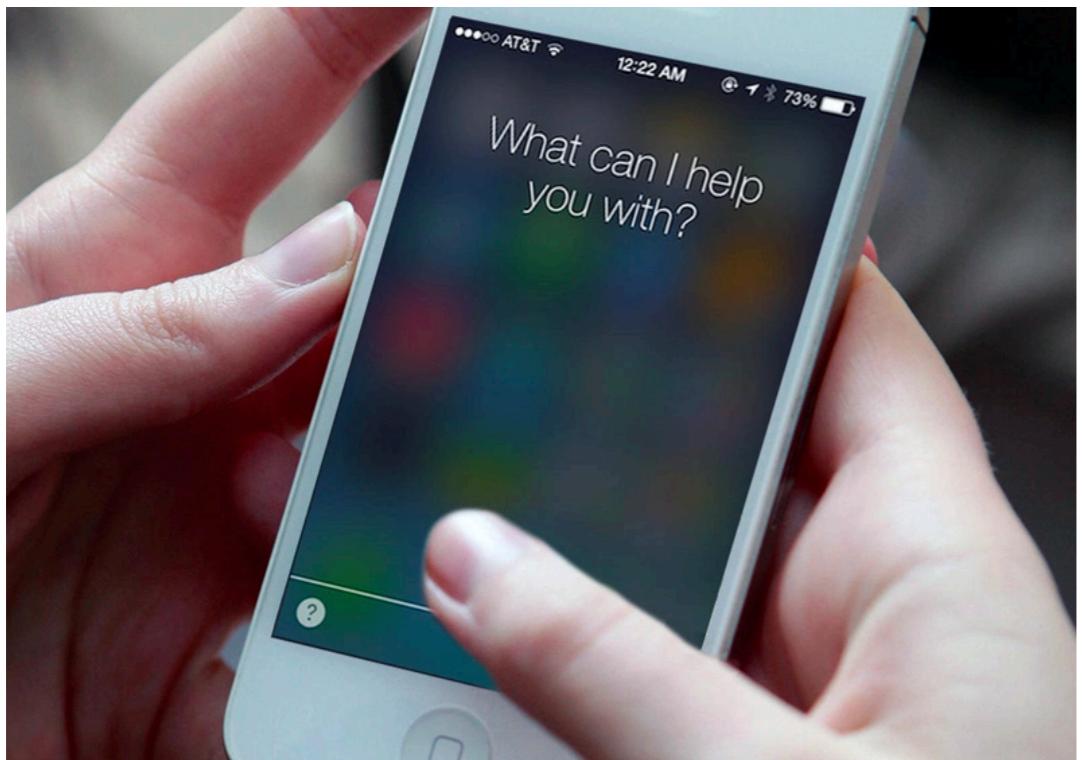
Multimodal NLP



Multilingual NLP



Speech



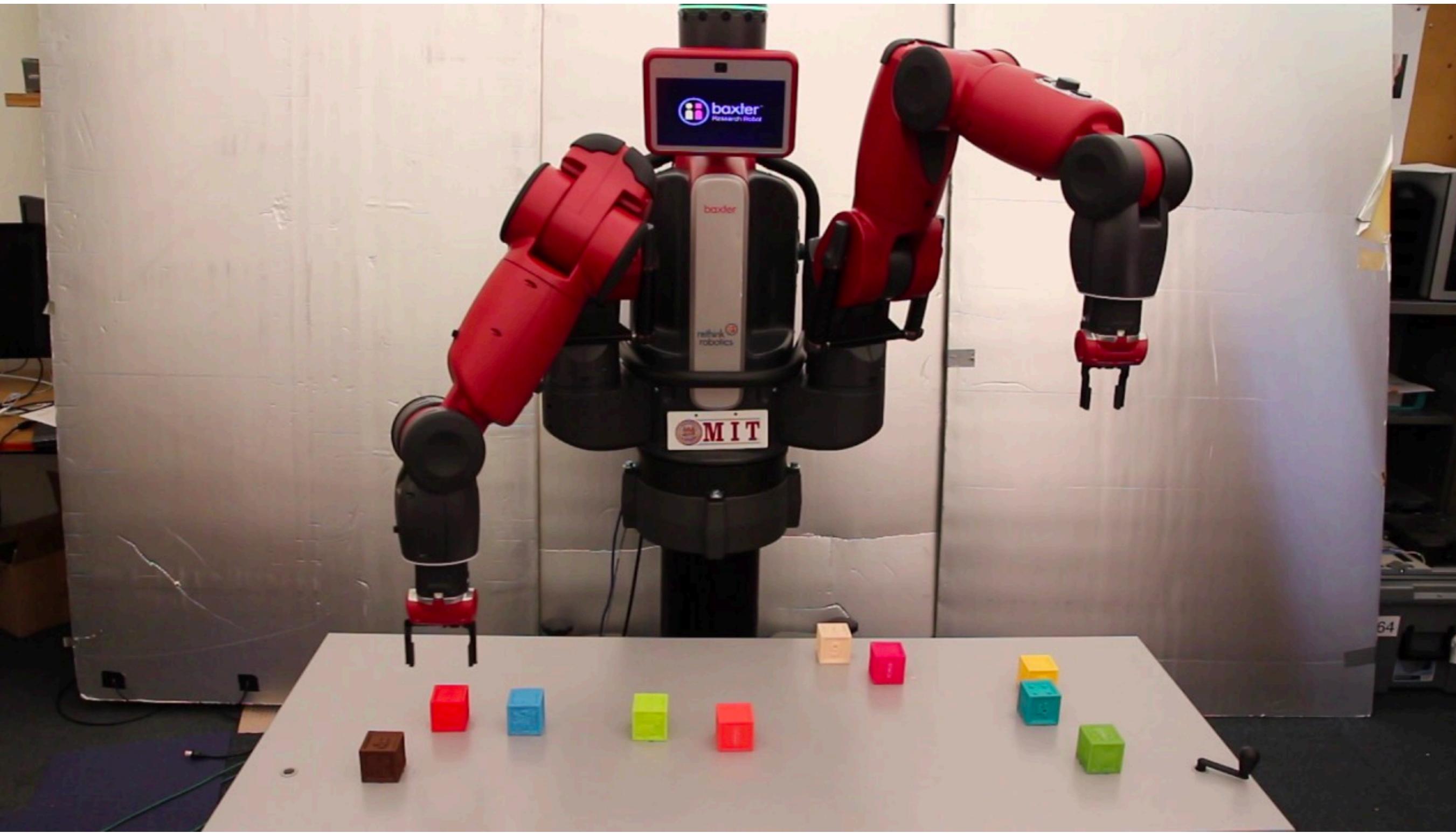
IBM **Watson**



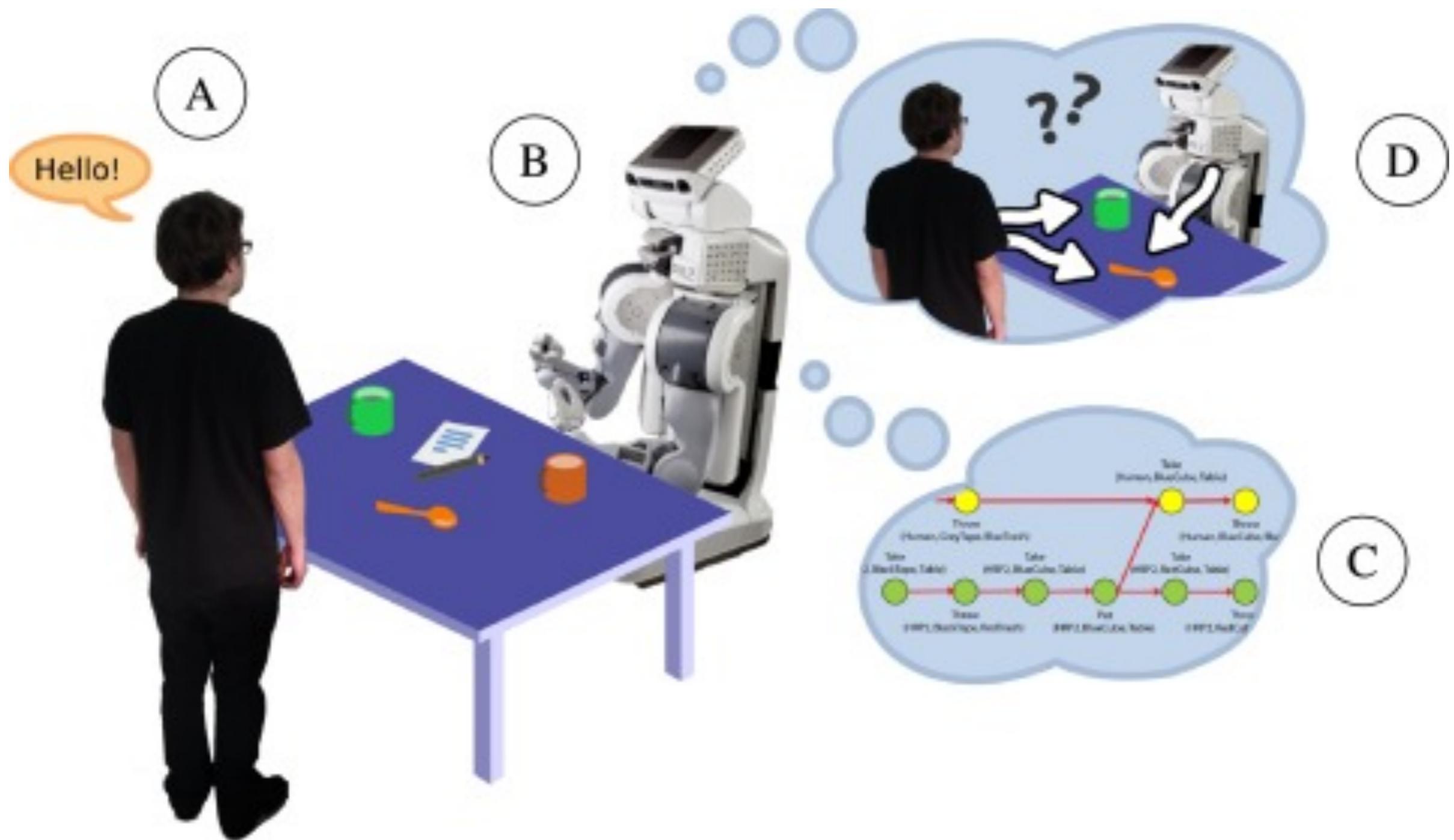
Language Learning



Language Grounding



Robotics



Psycholinguistics

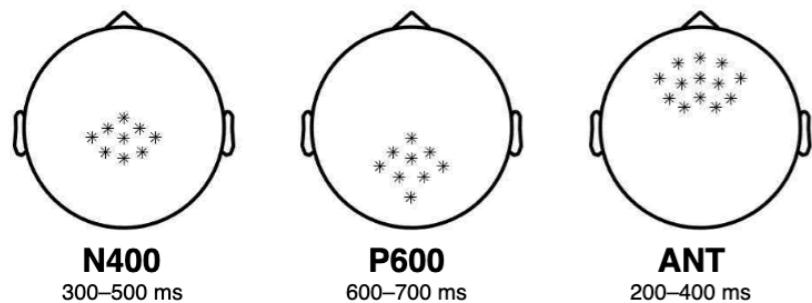
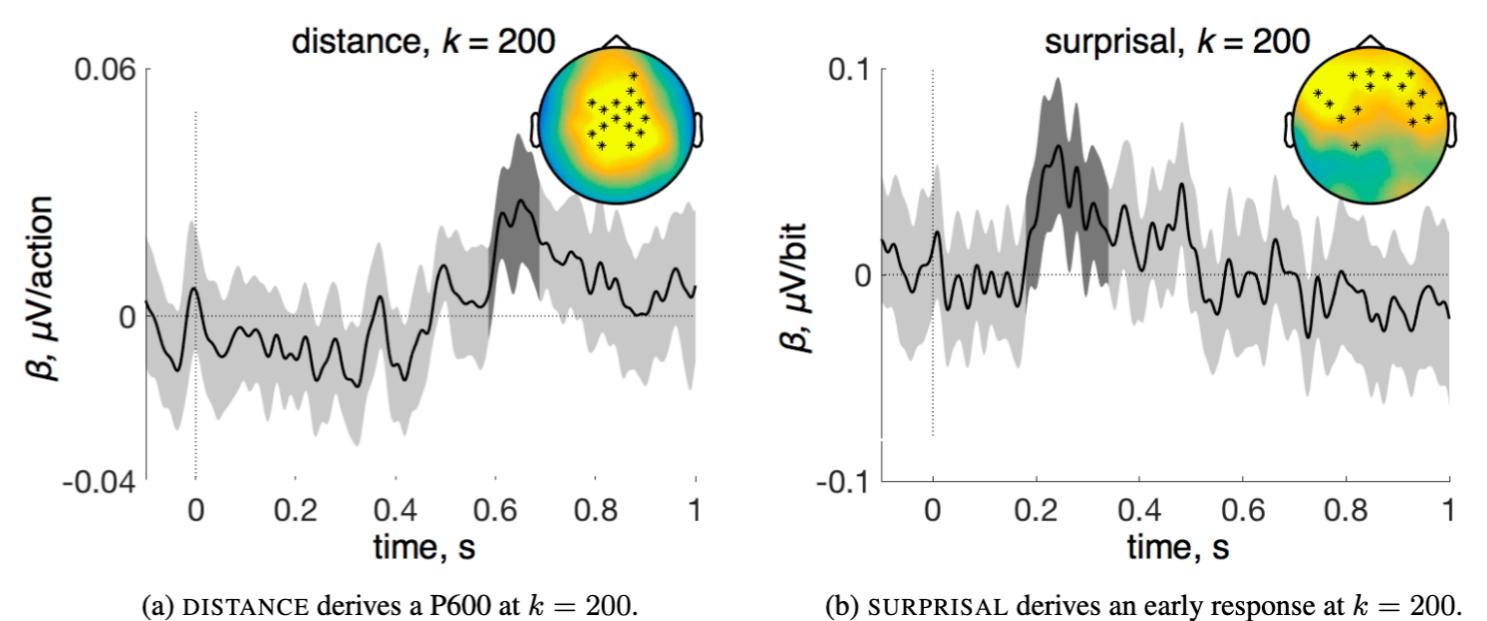


Figure 4: Regions of interest. The first region on the left, named “N400”, comprises central-posterior electrodes during a time window 300–500 ms post-onset. The middle region, “P600” includes posterior electrodes 600–700 ms post-onset. The rightmost region “ANT” consists of just anterior electrodes 200–400 ms post-onset.



Finding syntax in human encephalography with beam search
John Hale, Chris Dyer, Adhiguna Kuncoro, Jonathan Brennan

Language Generation



Document Analysis

Tasks

These are some of the tasks included in SciWING and their performance metrics

Task	Dataset	SciWING model	SciWING	Previous Best
Logical Structure Recovery	SectLabel	BiLSTM + Elmo Embeddings	73.2 (Macro F-score)	-
Header Normalisation	SectLabel	Bag of Words Elmo	93.52 (Macro F-Score)	-
Citation String Parsing	Neural Parscit	Bi-LSTM-CRF + GloVe + Elmo + Char-LSTM	88.44 (Macro F-Score)	90.45 Prasad et al (not comparable)
Citation Intent Classification	SciCite	Bi-LSTM + Elmo	82.16 (Fscore)	82.6 Cohan et al (without multi-task learning)
Biomedical NER - BC5CDR (Upcoming)	-	-	-	-
I2b2 NER (Upcoming)	-	-	-	-

Example: <https://github.com/abhinavkashyap/sciwing>

Story Understanding and Script Learning



Adversarial Learning for Text

(Idea shown using images)





It's done.



It's done.

Wrapping up

Reminders

- Project Update actually due today
- HW5 *nominally* due today
- Final Project Report due in two weeks
- Blog post due in two weeks

We're always trying to improve
and need your feedback

We're always trying to improve
and need your feedback

We're always trying to improve and need your feedback

- Course Evaluations are out!
 - Thoughts on homework/exam/projects
 - What content excited/bored you
 - How we can tweak earlier courses to better prepare you for 630
 - Tips for future students

Thanks for a great class!

