

It's All Fun and Games until Someone Annotates:

Video Games with a Purpose for
Linguistic Annotation.



David Jurgens

Stanford University
jurgens@stanford.edu



Roberto Navigli

Sapienza University of Rome
navigli@di.uniroma1.it



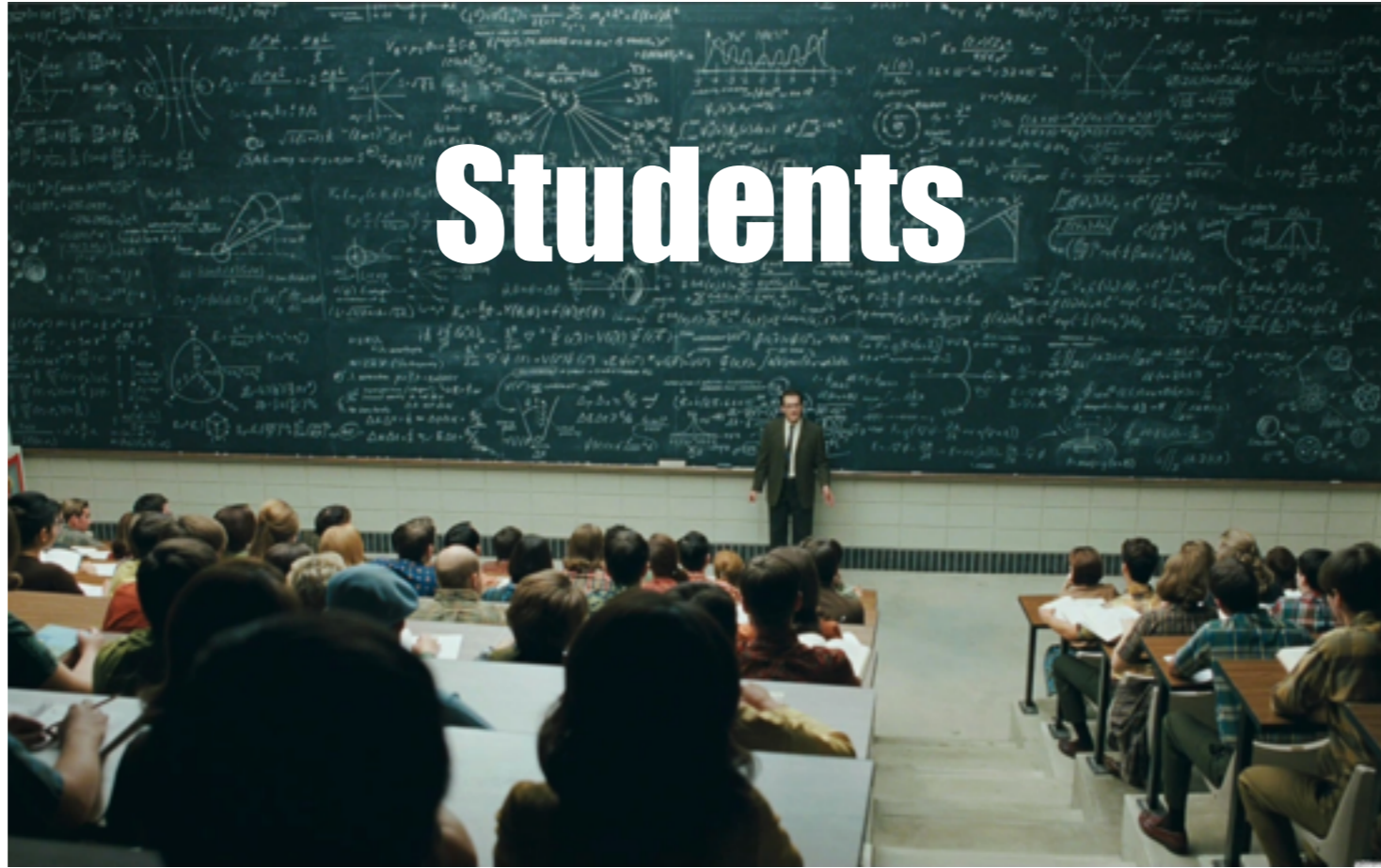
ERC Starting Grant
MultiJEDI No. 259234

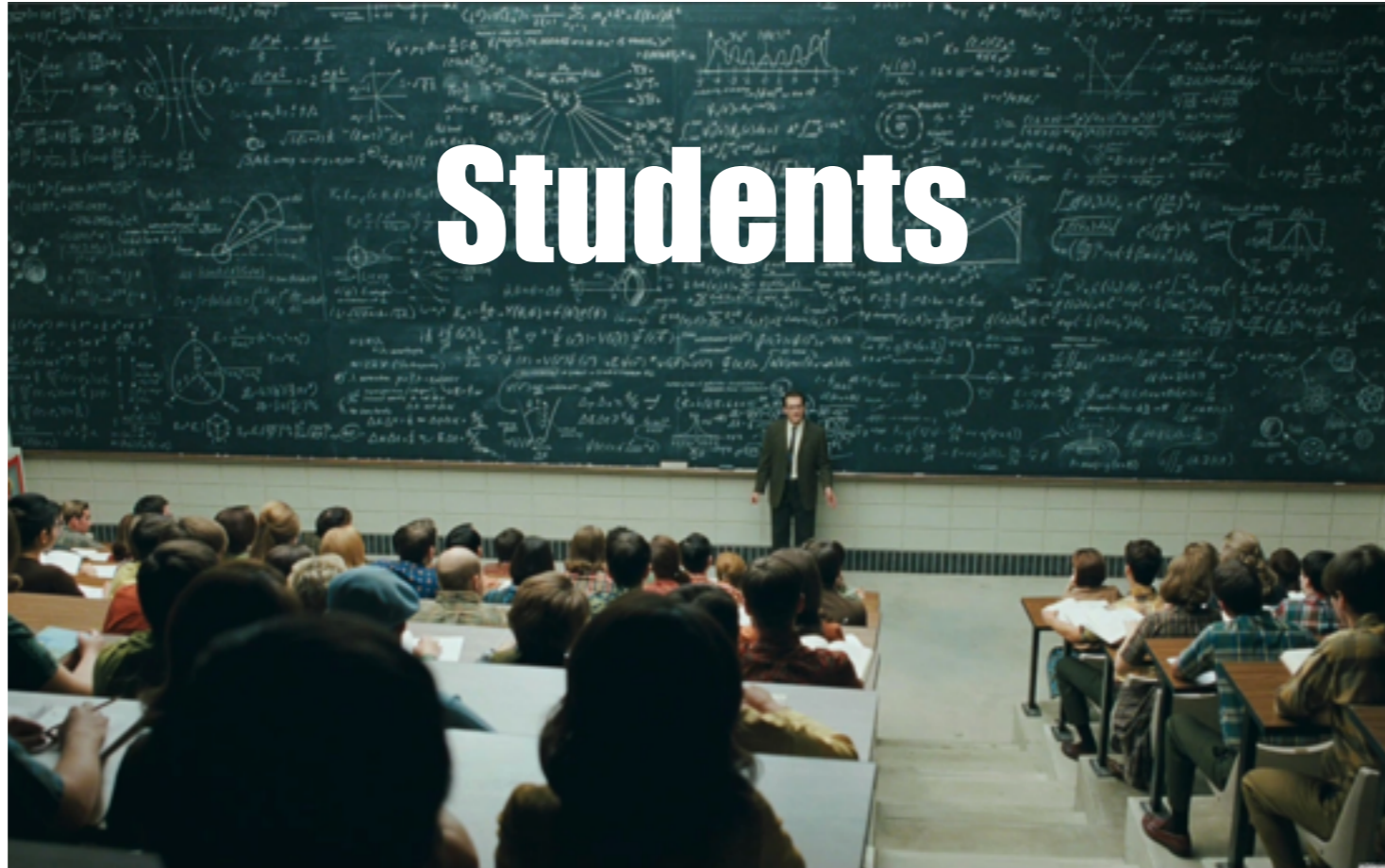
Annotating is no fun.



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





~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



The current state of NLP games

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

Game has started.
2 seconds remaining

Your score:
720

emergency

attack crash

Phrase Detectives (Poesio et al., 2013)

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.

Not mentioned before

This is a property

Done

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.

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.

The current state of NLP games

Wordrobe (Venhuizen et al., 2013)



Jinx (Seemakurty et al., 2010)



More similar to gamified tasks than taskified games

Phrase Detectives (Poesio et al., 2013)



Can we take a video game
and taskify it?

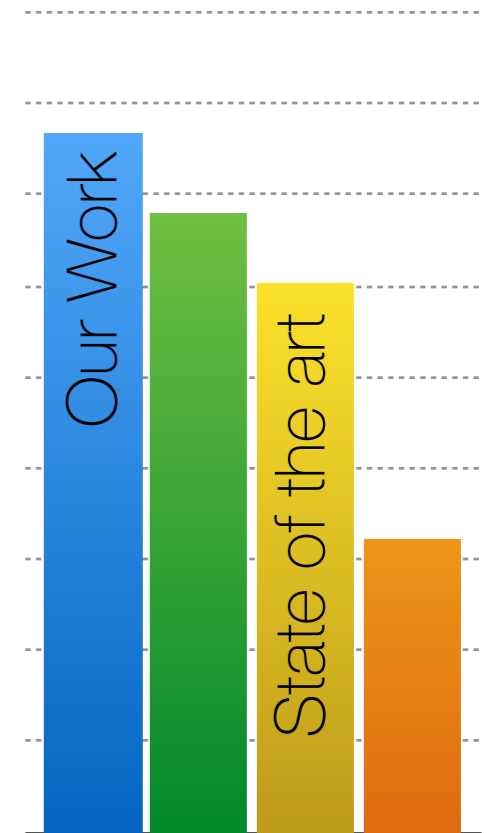
Contributions



Taskified Games



Design Methodologies



Expert-level Performance

Taskify a game with a popular design

Temple Run



1 Billion Downloads

Fruit Ninja

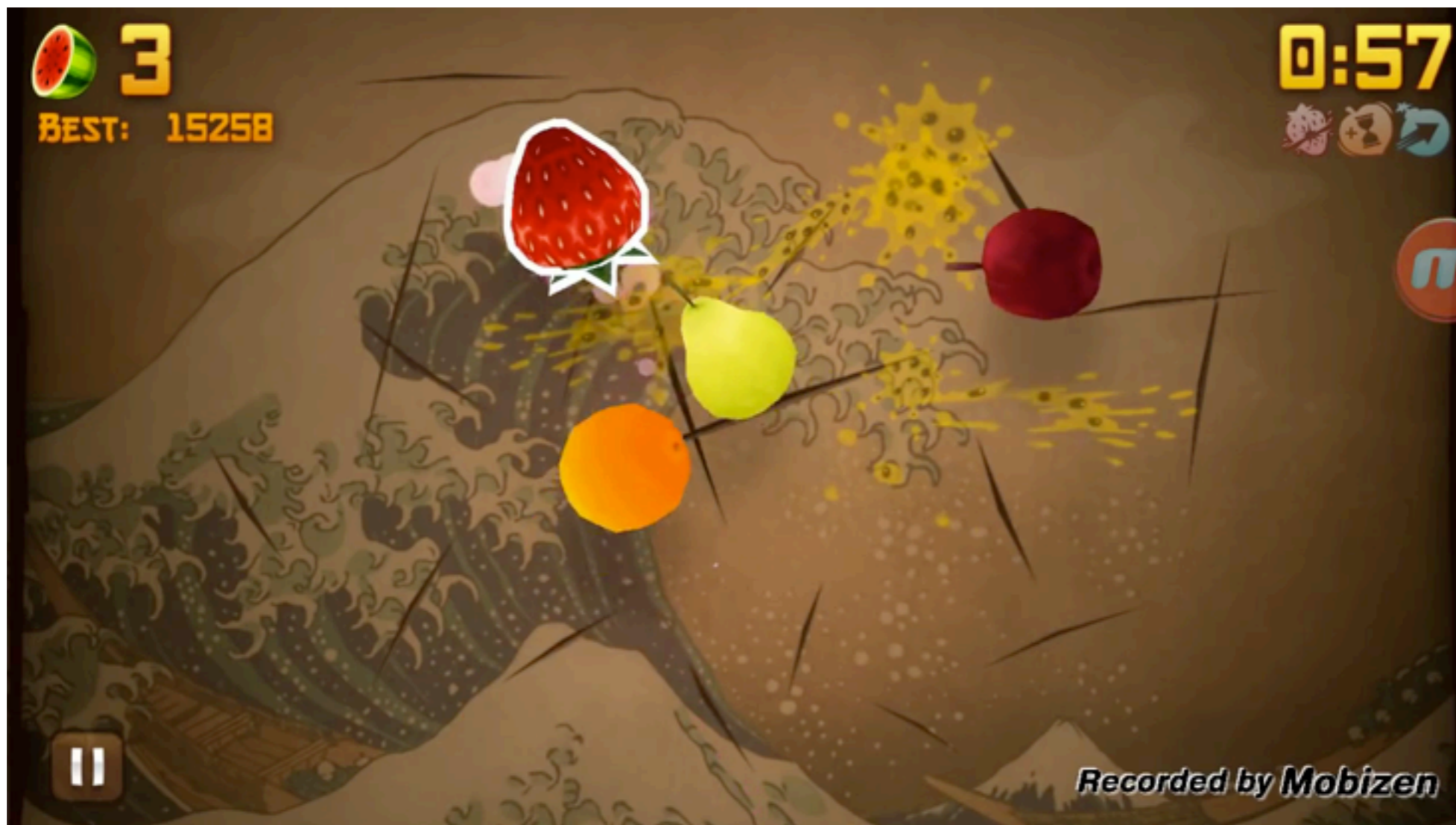


300 Million Downloads
1/3 of all iPhones

Can we adapt Fruit Ninja
for disambiguation?



Key mechanic:
Click on certain kinds of things




Player needs to avoid these

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



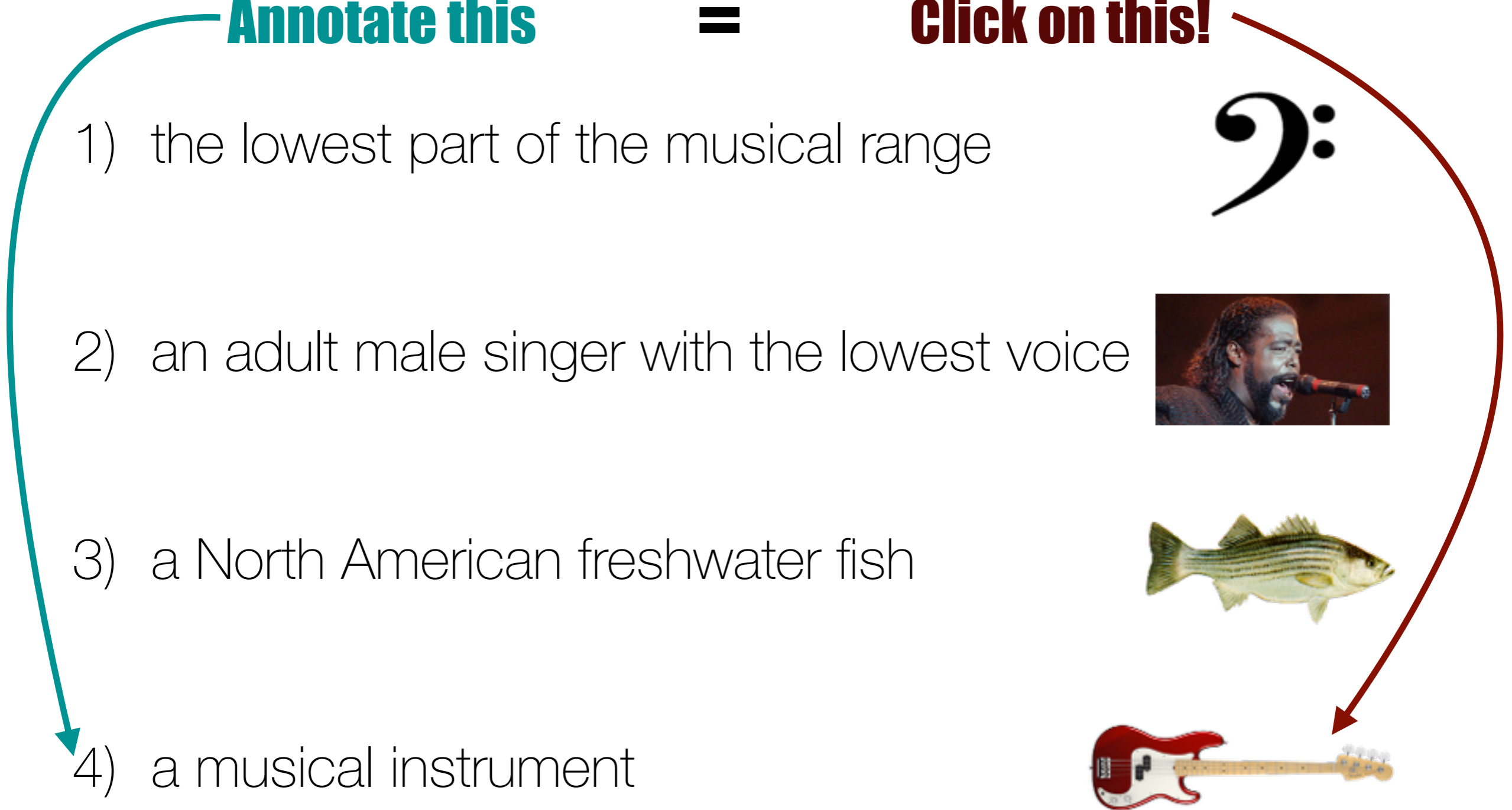
2) an adult male singer with the lowest voice



3) a North American freshwater fish



4) a musical instrument



She plays the bass



She plays the bass



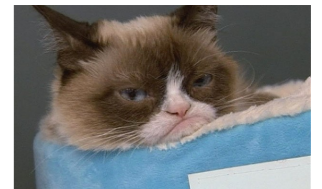
She plays the bass



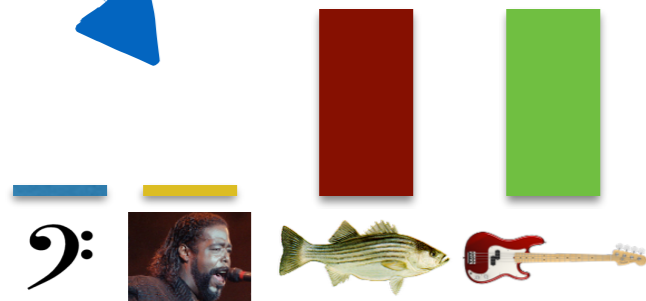
She plays the bass



Key Problem #1: This is boring

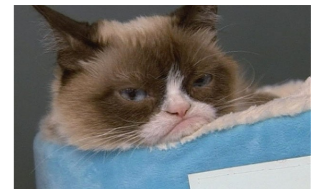


She plays the bass



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



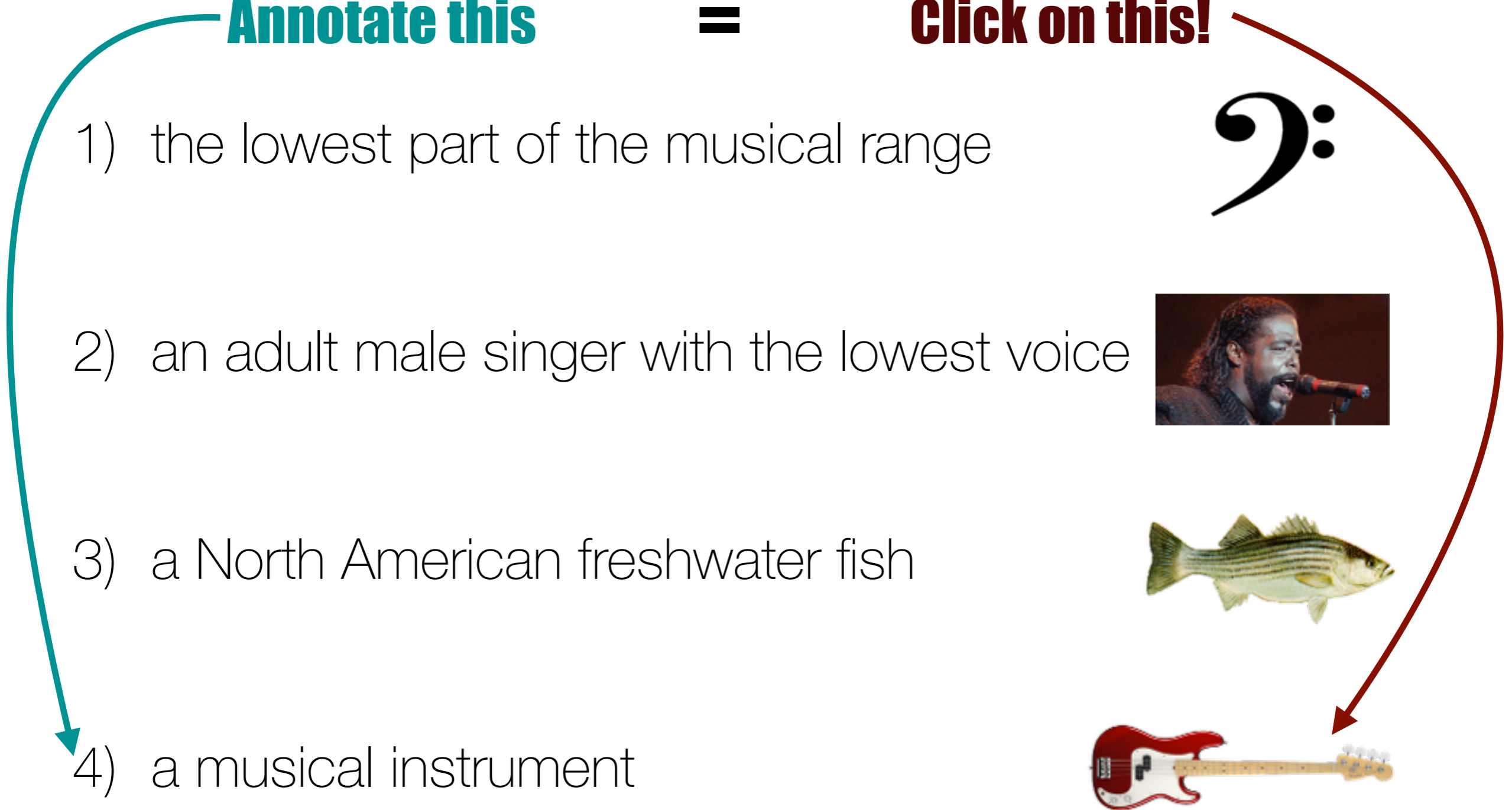
2) an adult male singer with the lowest voice



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



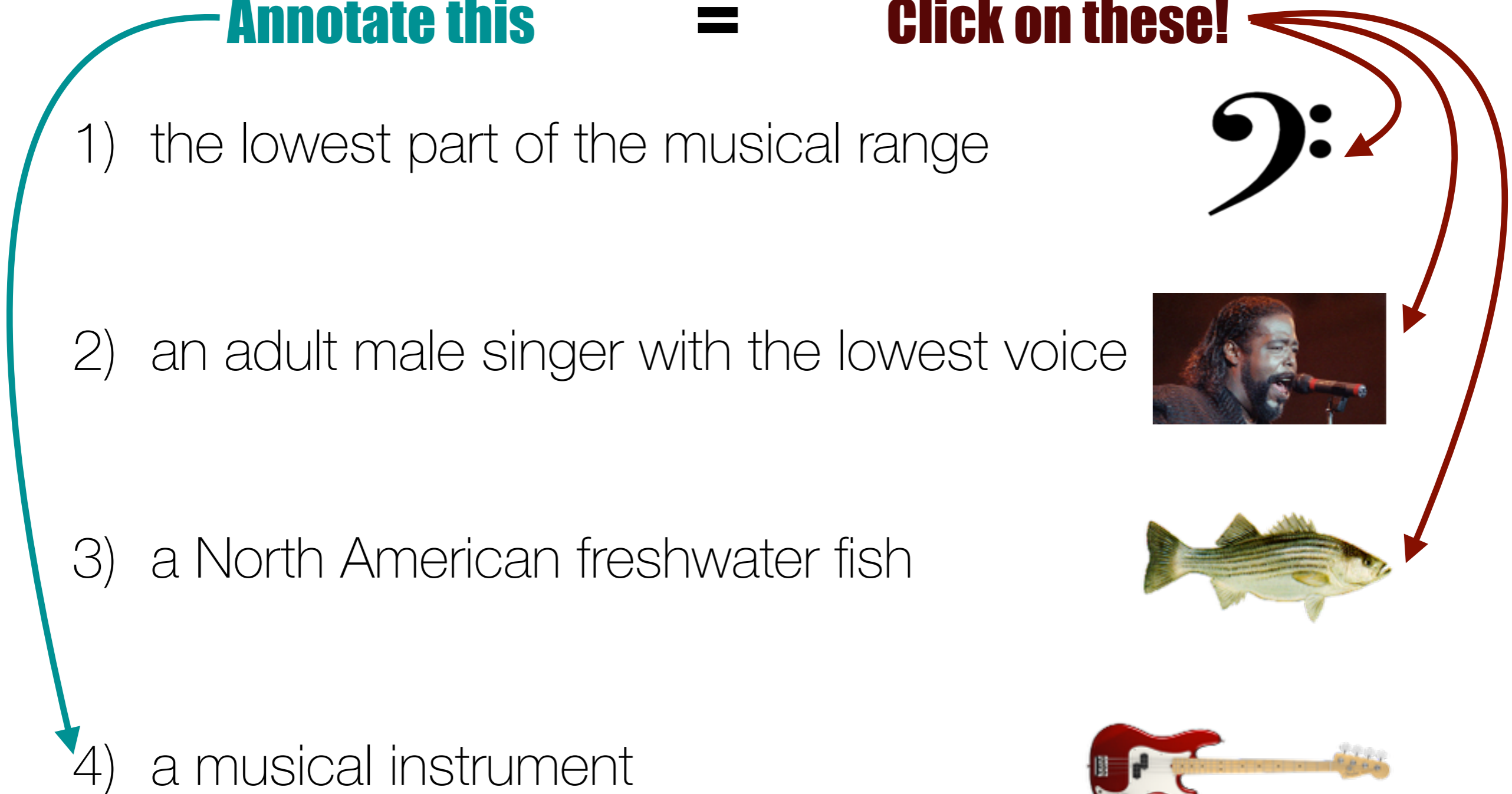
2) an adult male singer with the lowest voice



3) a North American freshwater fish



4) a musical instrument



She plays the bass



She plays the bass



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



IM  GENET

 **BabelNet**

A very large multilingual encyclopedic dictionary and semantic network

Current image-sense libraries



IMAGENET

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



000

Start

Player Setup

▶ Instructions

Game Options

View Leaderboard

Real game features!

Unlockable Racers



Lots of Power-ups



Enemies!



Leaderboards





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



cat#n#1: a feline mammal



...

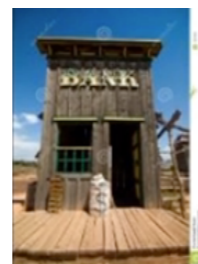
Task Question:

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



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”



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!


Time: Hearts


This race's puzzle clues:



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

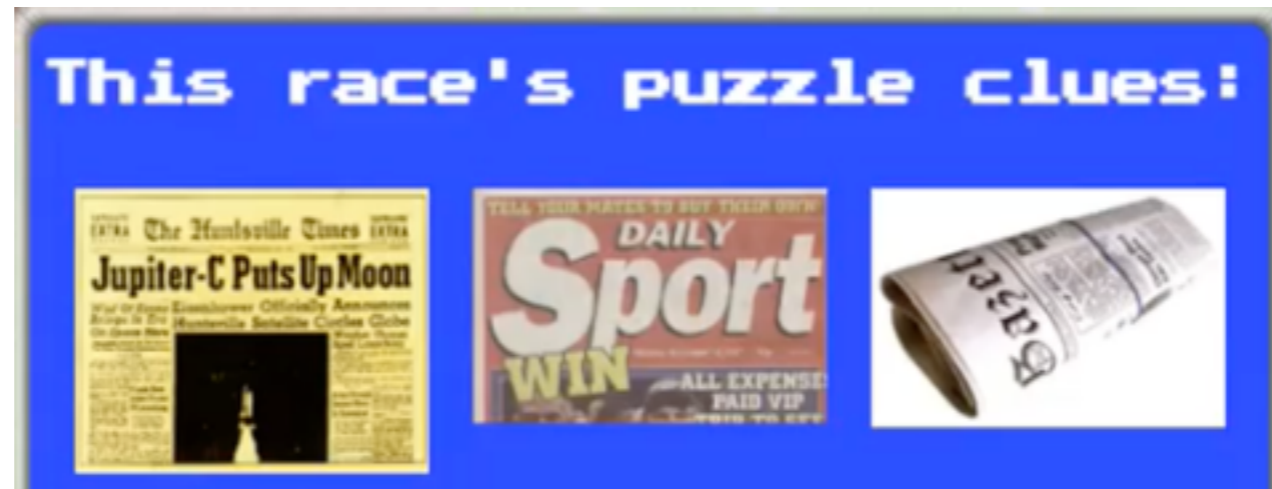
▶ Let's race!

Score: 22819 Time: 73s Hearts 

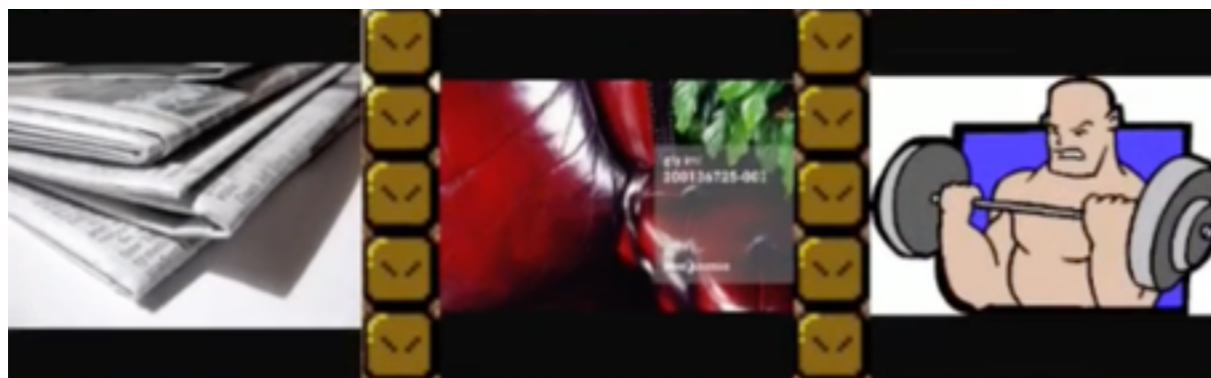


PLEASE DO NOT THROW STONES AT THIS SIGN THANK YOU





Players are shown
two types of puzzle gates



Golden Gate

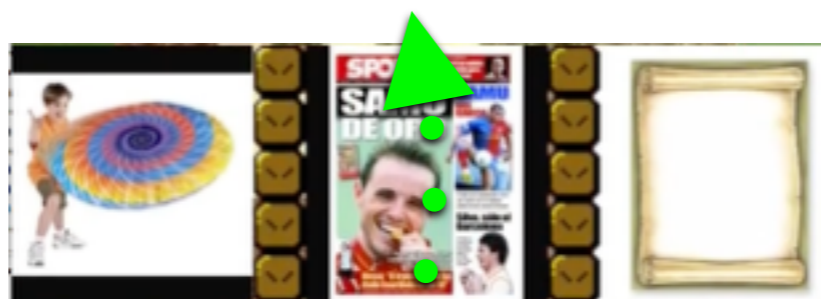



Mystery Gate

This race's puzzle clues:

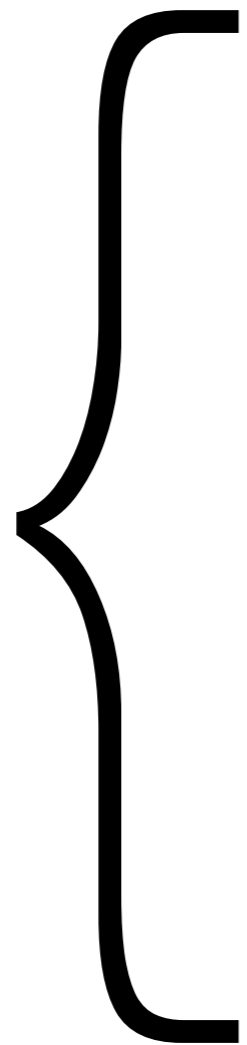


This race's puzzle clues:

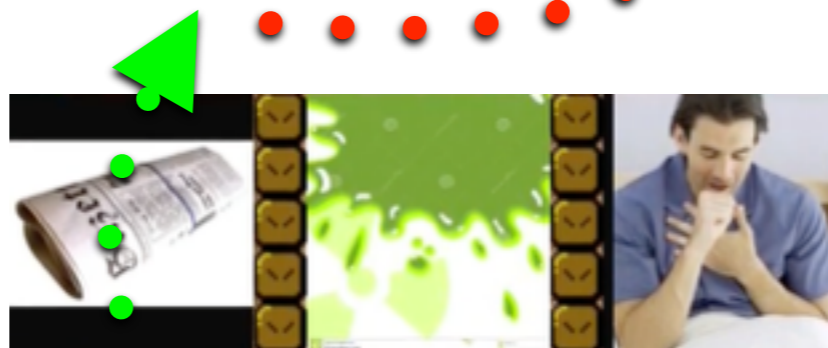


Accuracy: 100%

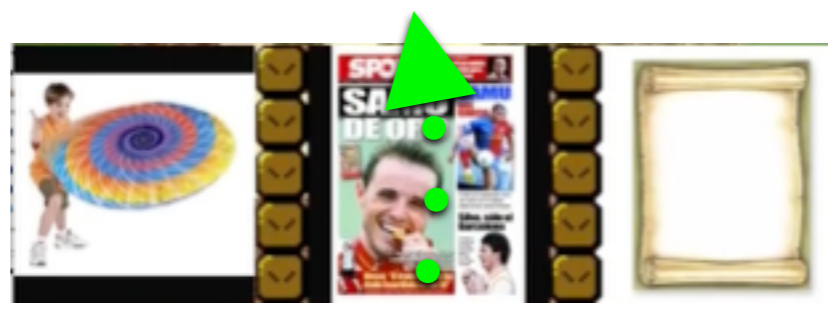
First three gates are always **golden**



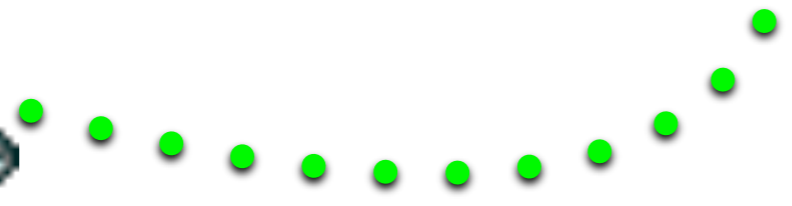
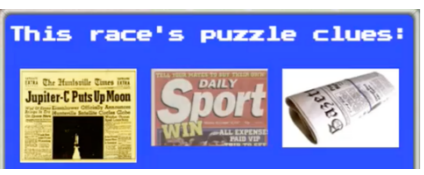
Accuracy: 66%

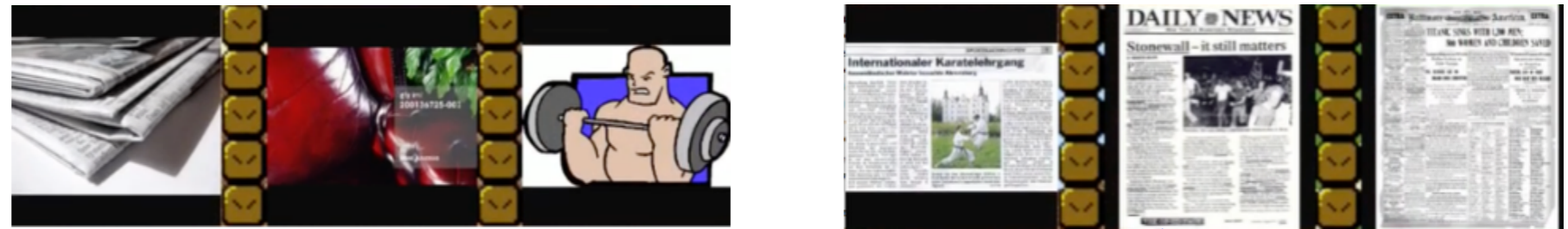


Accuracy: 100%

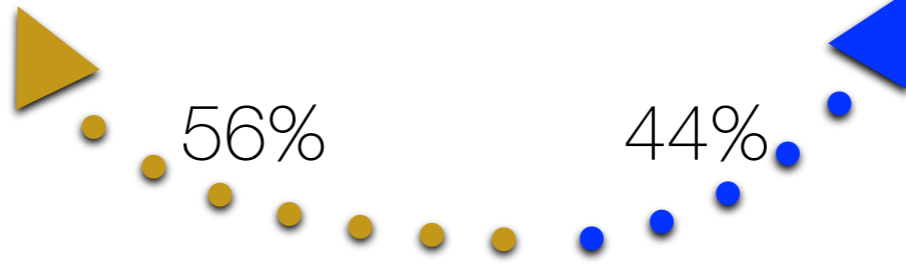


Accuracy: 100%

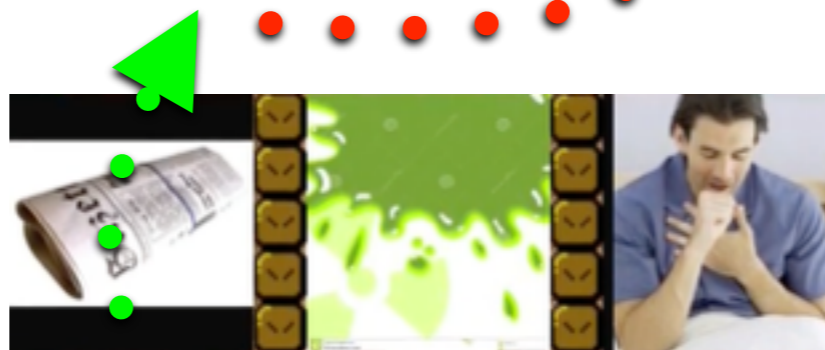




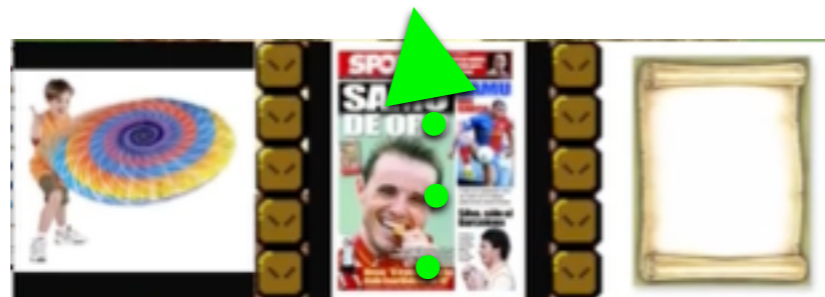
Show a **mystery** gate with probability = $0.66 \times \text{Accuracy}$



Accuracy: 66%

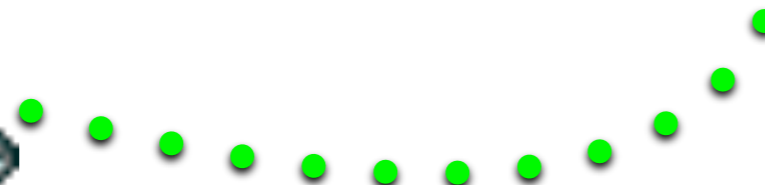
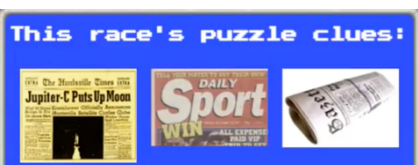


Accuracy: 100%

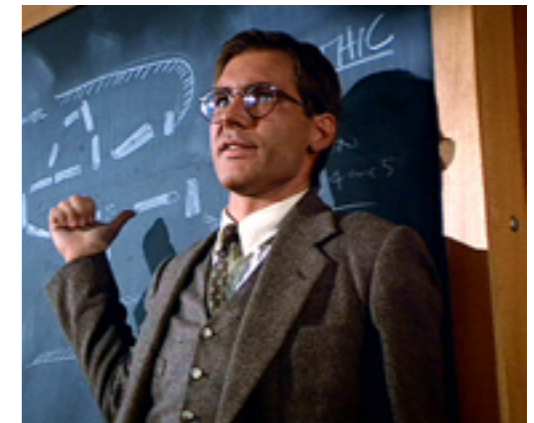


Accuracy: 100%

First three gates are always **golden**



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

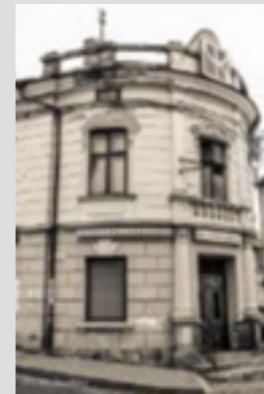
How does PuzzleRacer
compare in quality with
Crowdsourcing?

Recreate the Puzzle Racer annotation task on CrowdFlower



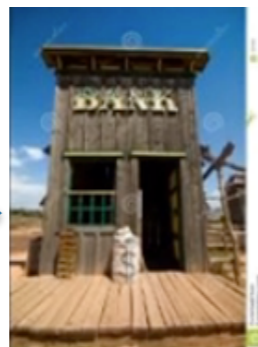
Given the three example images in the instructions, which of the following images most resembles underlying idea?

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%

=

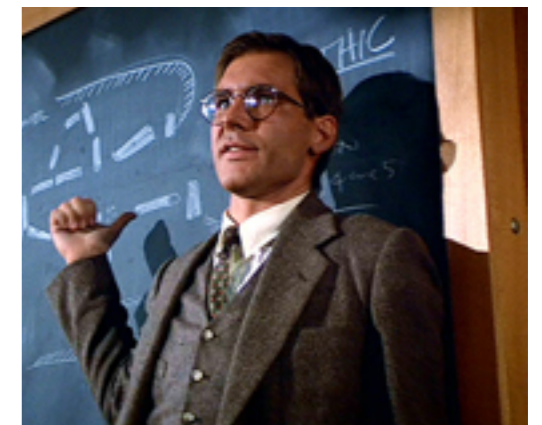


79%

14%

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

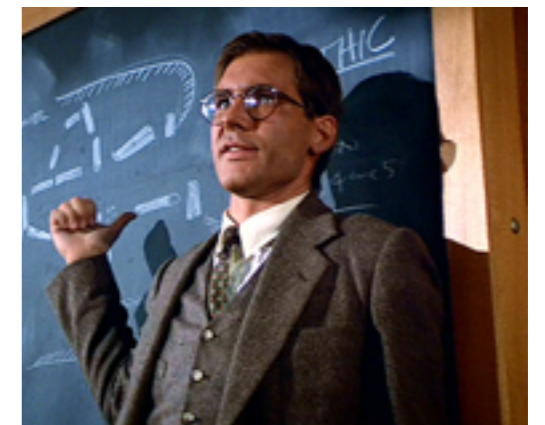
Does it work?



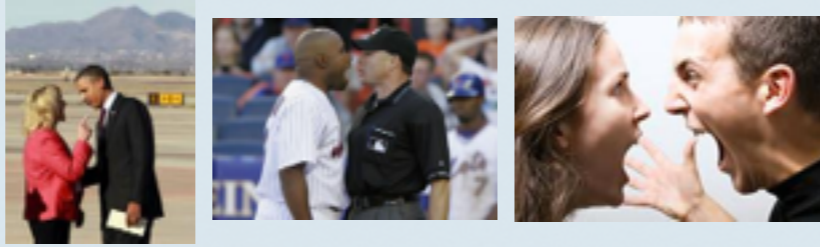
No statistically-significant difference in quality between Puzzle Racer-created and Expert-selected images



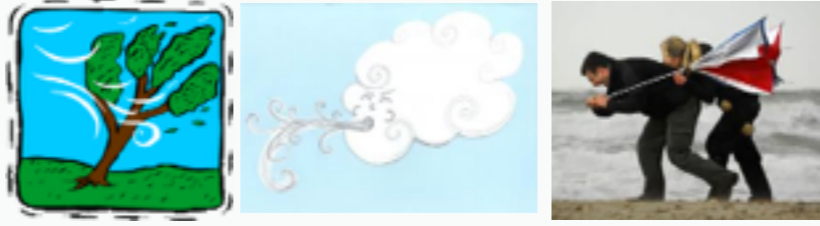
=



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

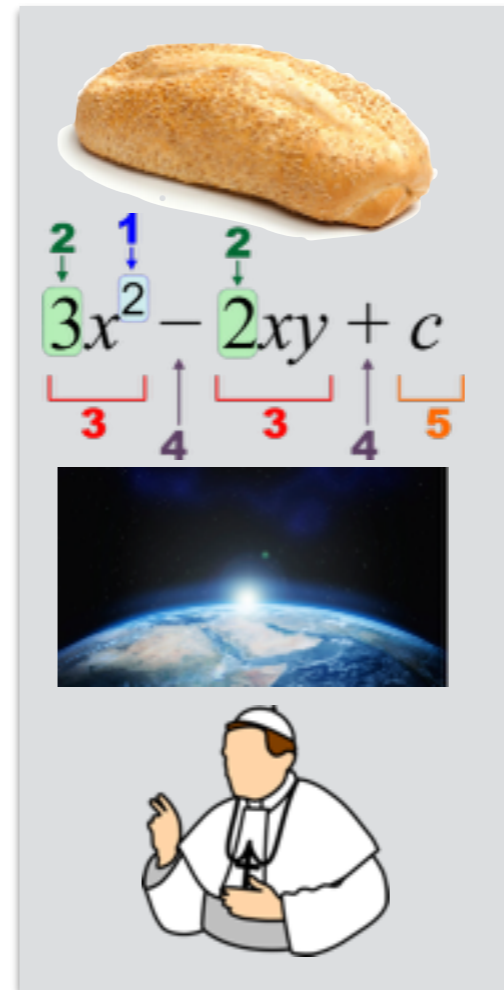
Show one picture for each of the n senses

Diagram illustrating the concept of showing one picture for each of the n senses. The diagram features a large left-facing curly bracket on the left and a large right-facing curly bracket on the right, both spanning the vertical extent of the central content. The central content is organized into two columns. The left column contains four items: a bass clef symbol, a photograph of a man with a beard singing into a microphone, a photograph of a striped bass fish, and a photograph of a red electric guitar. The right column contains four items: a photograph of a loaf of bread, a mathematical expression $3x^2 - 2xy + c$ with various colored numbers (2, 1, 2, 3, 4, 3, 4, 5) and arrows pointing to different parts of the expression, a photograph of the Earth from space, and a cartoon illustration of a man in a white lab coat and cap holding a clipboard.

Include n pictures from random senses

She plays the bass

Show one picture for each of the n senses



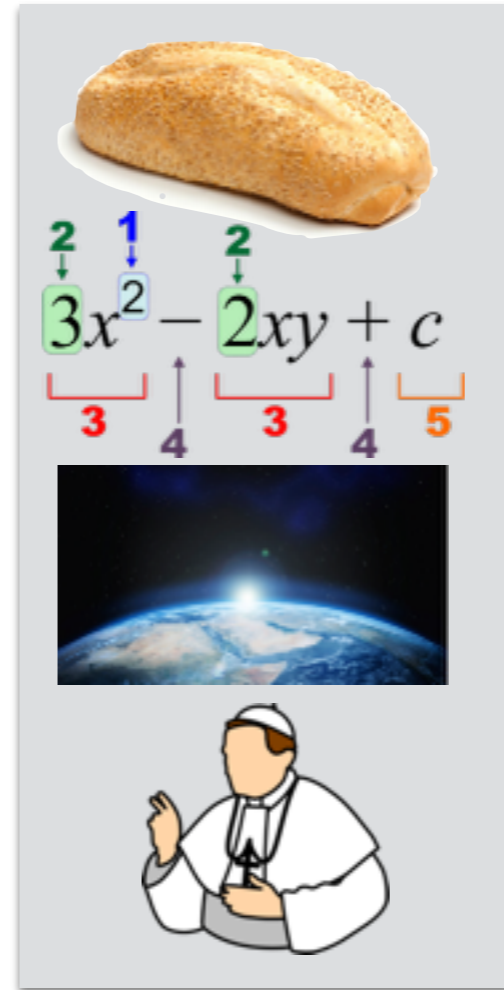
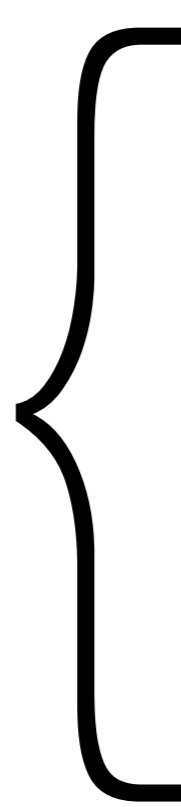
Include n pictures from random senses



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

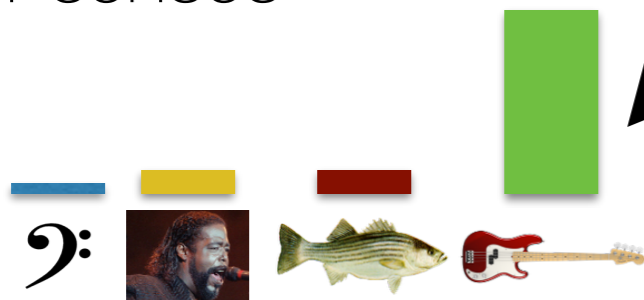
She plays the bass

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.

Does it work?

Direct comparison with Wordrobe, a WSD game



The screenshot shows the Senses WSD game interface. At the top left, there is a red circular logo with a white 'S' and the word 'Senses' next to it. To the right of the logo, it says 'Questions left until drawer is completed: 5'. In the top right corner, there is a small teal circle with a white question mark. The main text of the game is 'Authorities say the accident **occurred** Saturday, near the town of Veligonda in southern Andhra Pradesh state.' Below this text is a list of three radio button options: 'come to pass (synonyms: happen, hap, go on, pass off, pass, fall out, come about, take place)', 'come to one's mind - suggest itself', and 'to be found to exist'. At the bottom left, there is a slider labeled 'Place your bet: low' on the left and 'high' on the right, with a blue bar and a silver knob. At the bottom right, there are two buttons: 'answer' and 'skip'.

Senses Questions left until drawer is completed: 5

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

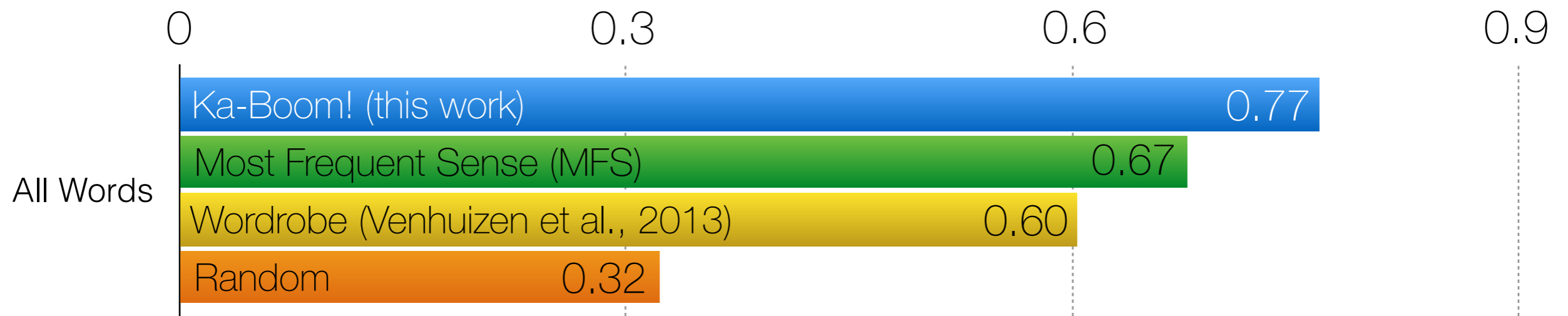
- come to pass (synonyms: happen, hap, go on, pass off, pass, fall out, come about, take place)
- come to one's mind - suggest itself
- to be found to exist

Place your bet: low high

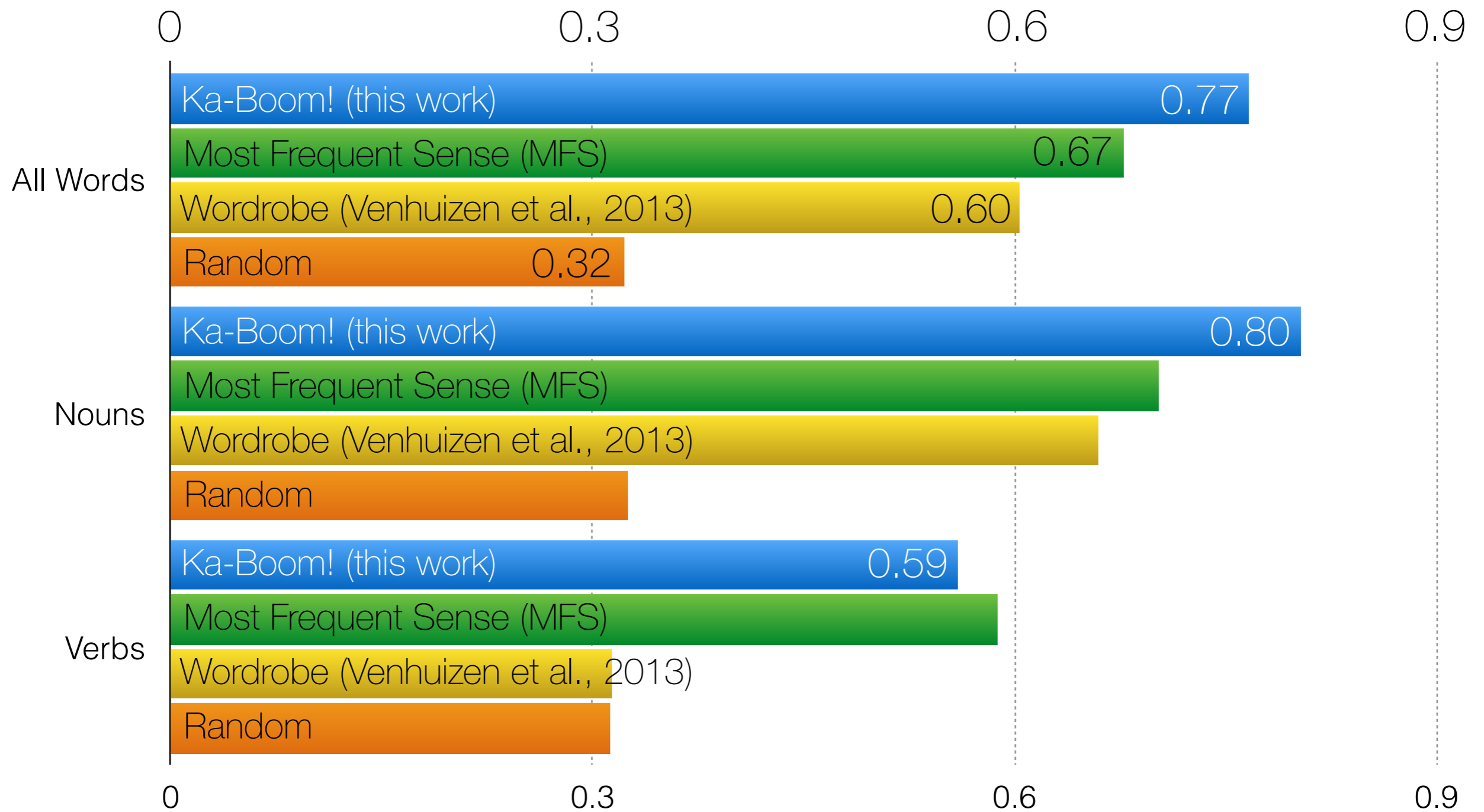
answer skip

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

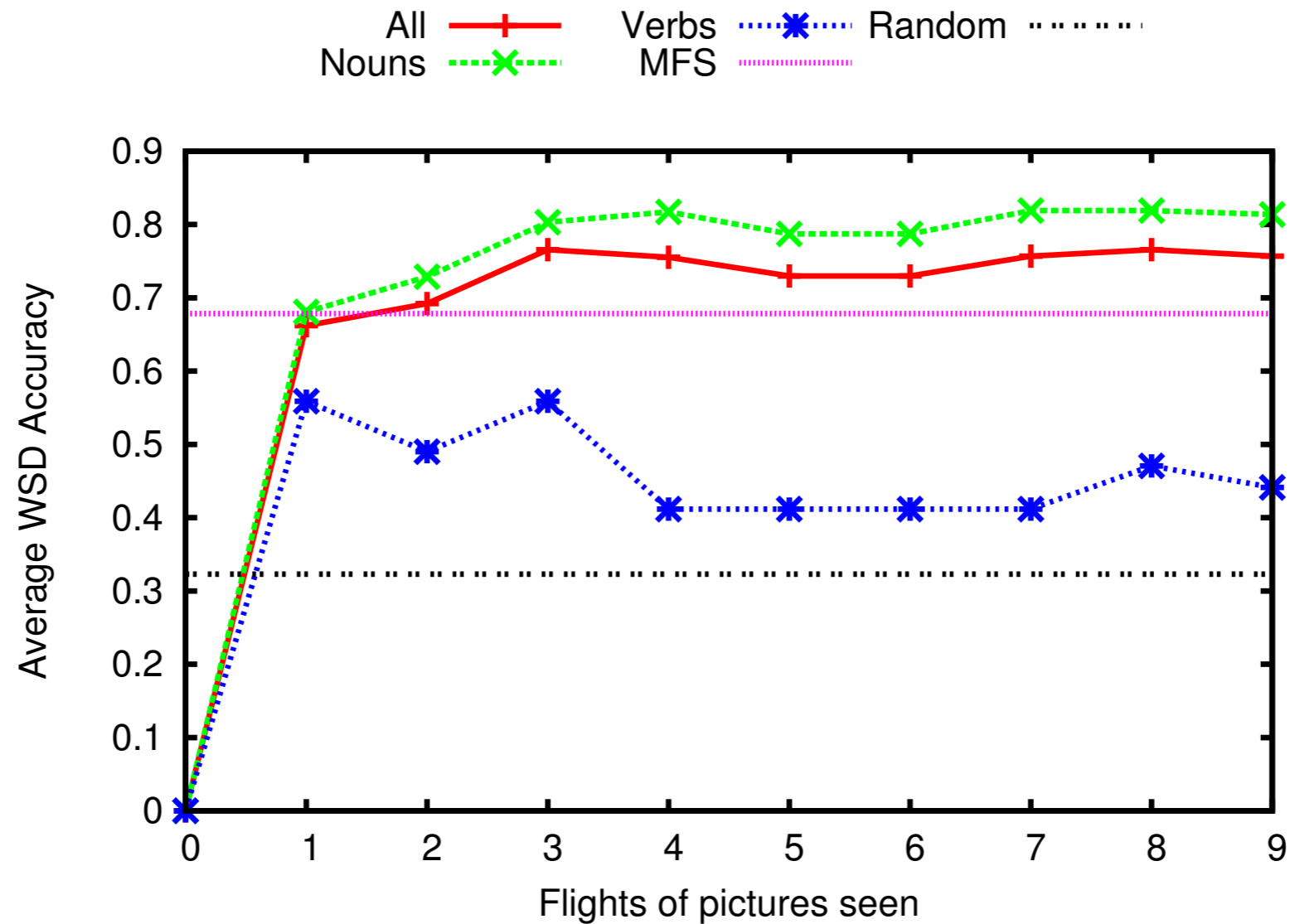
Disambiguation Accuracy



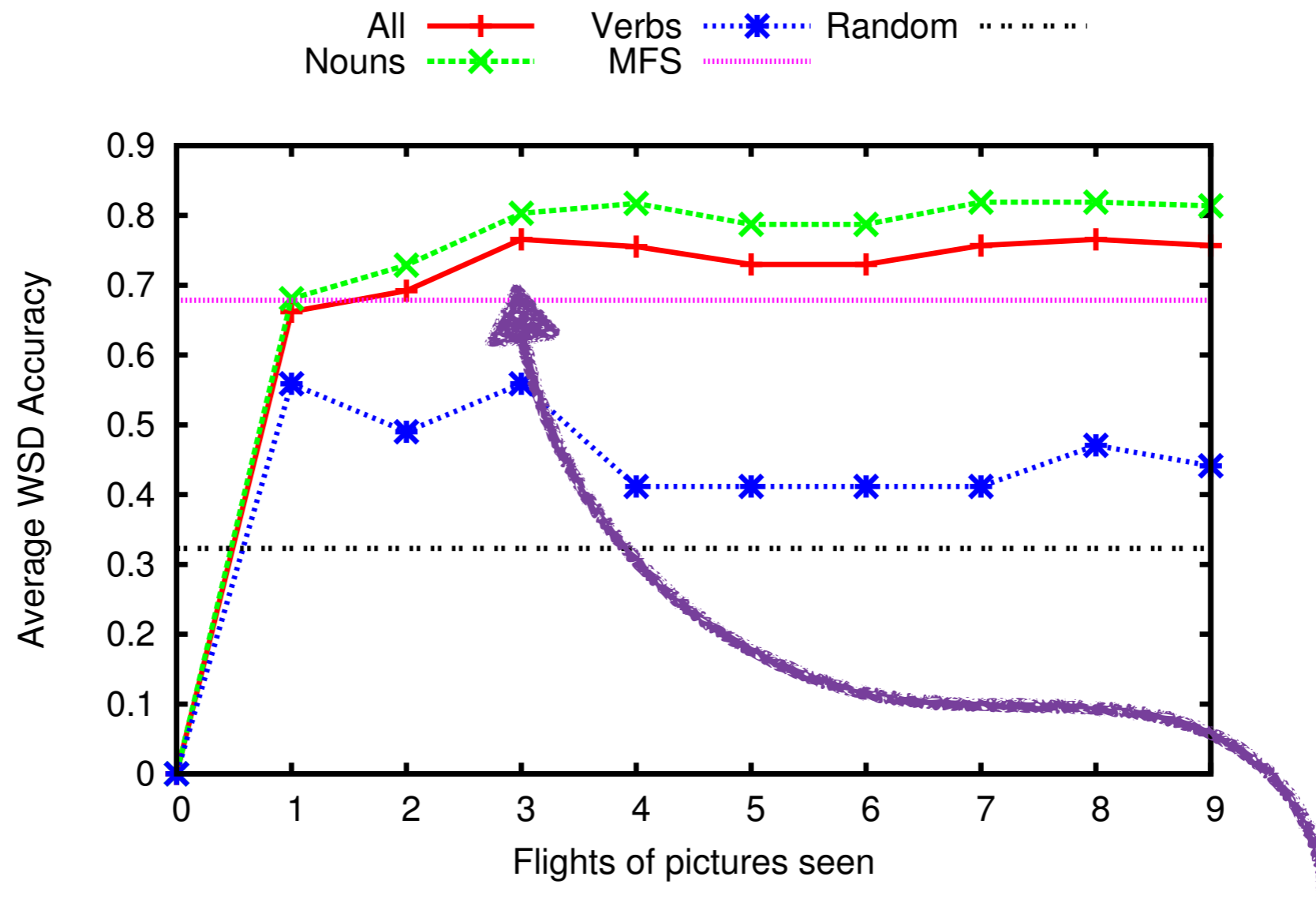
Disambiguation Accuracy



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

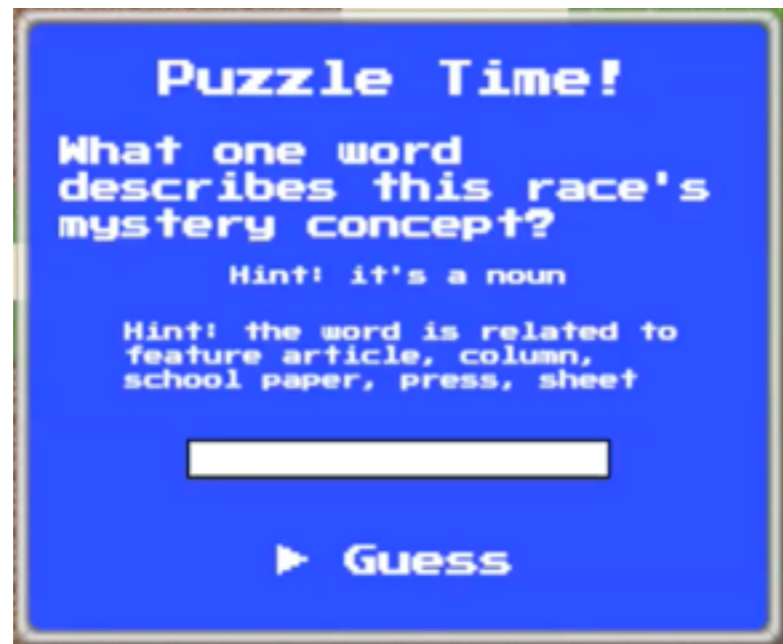


How long did players take to converge on the right sense?



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

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
- Game were slower than crowdsourcing
 - But only because we didn't have a ready pool of players

Don't gamify your tasks, taskify your games!



David Jurgens

jurgens@stanford.edu

Roberto Navigli

navigli@di.uniroma1.it

Games

<http://knowledgeforge.org/>



ERC Starting Grant
MultiJEDI No. 259234