Modeling language and vision

Micha Elsner
The Ohio State University

Let’s listen...

“How to find the Andromeda galaxy”

https://youtu.be/clfjPvaXGIs?t=63

Listen for 30 seconds: what is the speaker saying?

Is this what he said?

...and if you actually look at Cassiopeia and you look down below the W, do you see that little fuzzy thing right there?

That is the Andromeda galaxy, which it actually says is 2.4 million light years away…

A long way toward the back.

Or is it this?

...and if **we** actually look at Cassiopeia and you look down below the W, ~~do~~ you see that little fuzzy thing right there?

That is**...** the Andromeda galaxy, **which is actually saying that it is… does it give you distance? Doesn't seem to be giving you distance, but it's basically** 2.4 million light years away…

~~A~~ long way toward ~~the~~ back.

Well, it’s both, right?

Computational language generation typically ignores words like “basically” and “um”

As engineers, we don’t need to produce these words (Siri doesn’t say “um”)

As psycholinguists...

Those “ums”, pauses and restarts serve important speech functions

All languages have them

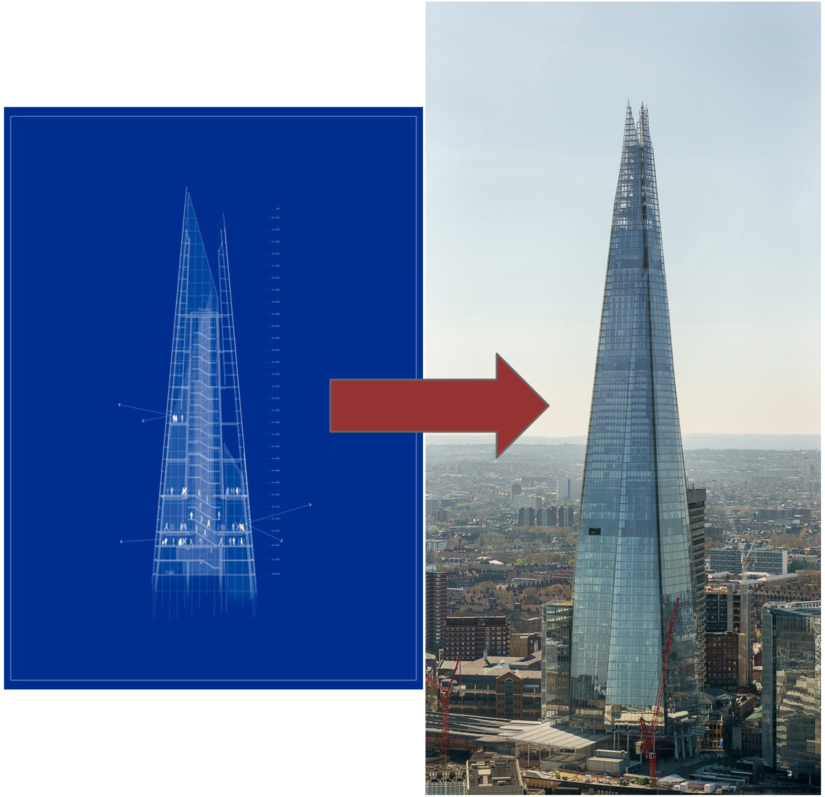
All speakers use them

They have a lot to tell us about how language is created in the human mind

Psycholinguistics

Connects the study of language as an abstract structure (phonology, morphology, syntax)...

To language as concrete reality: how particular utterances are produced and interpreted in real time



This talk

How **computer models of the mind** can help us understand how it works

And where those pauses and disfluencies come from

Computational cognitive models

AI as a “model organism” for:

- Learning
  - How do babies learn language by listening?
- Perception
  - How do we recognize faces?
  - What acoustic cues help us recognize words?
- Decision-making
  - How do we assess risk and reward?

In the first part of the talk, we'll see data primarily from the lab,

as we try to understand what's going on

In the second, we'll build a model of the human speaker using AI,

and use it to test our hypotheses

The team

Alasdair Clarke:
Psychology, University of Essex
Ph.D. in Mathematics

expert on the visual system

Hannah Rohde:
Linguistics and the English Language, University of Edinburgh
Ph.D. in Linguistics

expert on discourse pragmatics

Micha Elsner:
Linguistics, The Ohio State University
Ph.D. in Computer Science

expert on AI and computational models

What are pauses for?

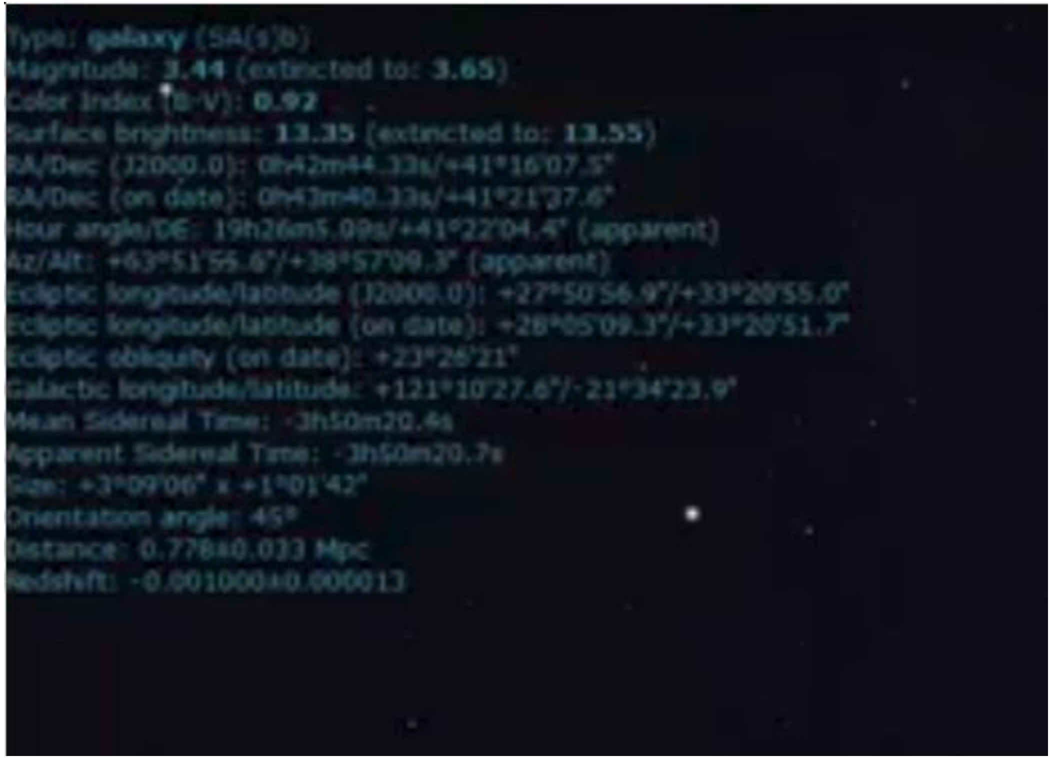
You pause because you’re still thinking.

You **fill in** the pause with words like “um” (or “basically”, “like”, “well”, and others) because you still want to talk.

Where delays come from

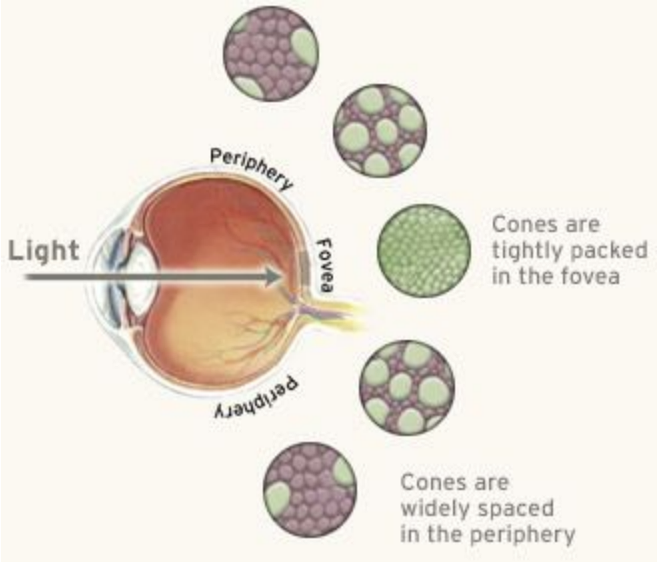
Why does it take so long to figure out that distance is missing?

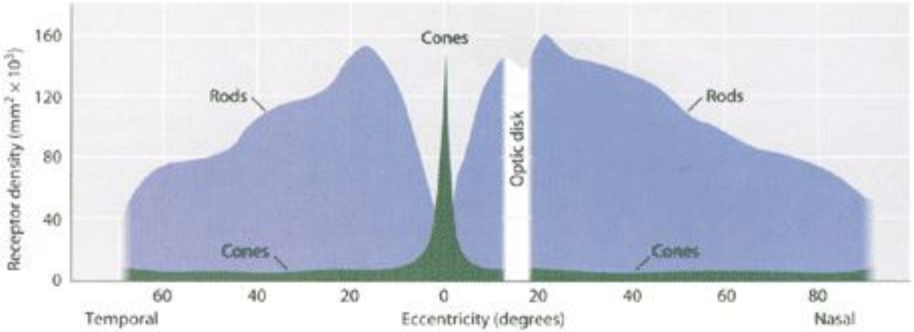
It’s hard to see!



http://www.webexhibits.org/colorart/ag.html
curated: Michael Douma

Vision is hard





So, you see the world like this

Demo by Geisler and Perry at UT

Your eye has to move

To create the illusion of detailed vision across the entire visual field, your eye moves around…

These movements are called **saccades**

They happen about every 200 milliseconds

Watch the moving eye

An **eyetracker** is a camera pointed at the pupil of your eye

Using the tracker, we can see exactly where you're looking:

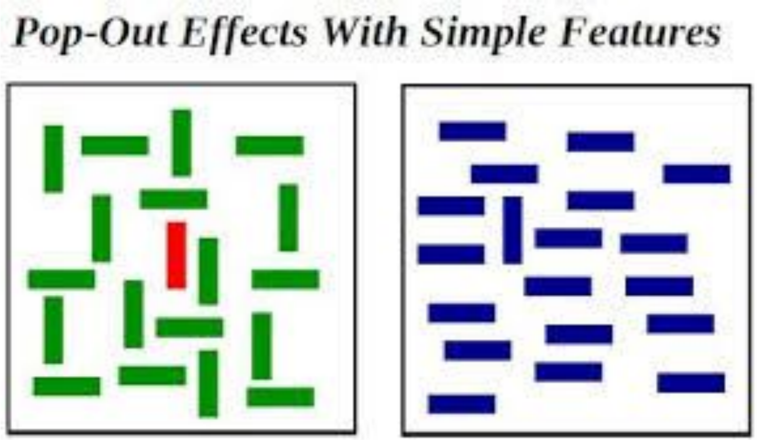
https://youtu.be/0_KaltdTkEM?t=79

Visual search

In these images, find the odd one out:

Visual search

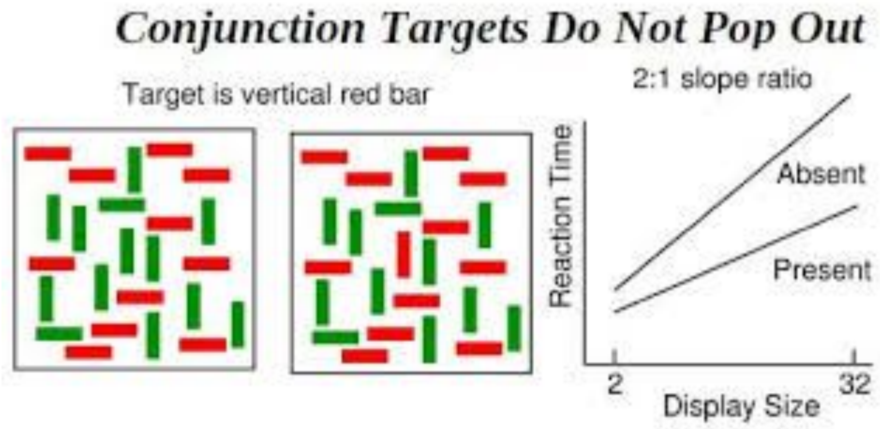
In these images, find the odd one out:



img: Univ. Melbourne
research by Anne Treisman

Visual search

Which of these images has a vertical red bar?



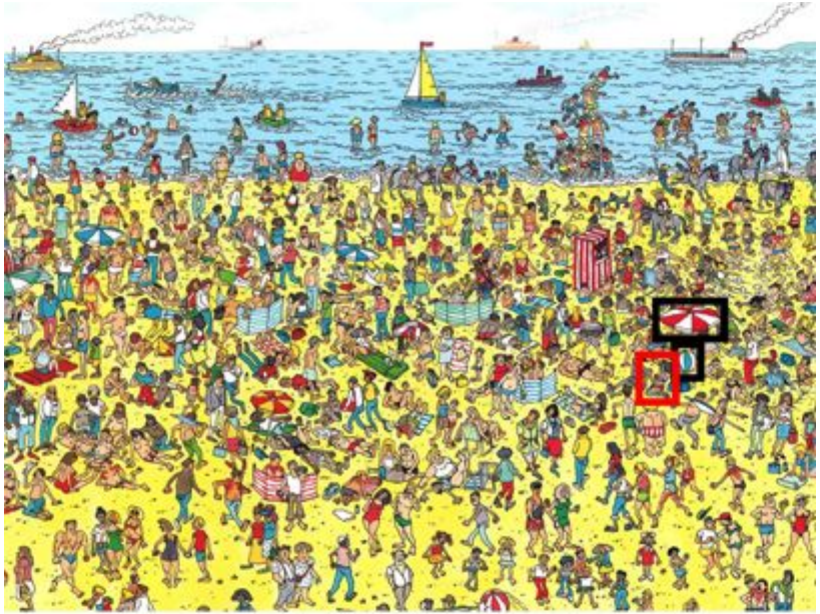
img: Univ. Melbourne
research by Anne Treisman

The Waldo studies

Clarke, Elsner, Rohde 2015, 2013
Elsner, Rohde, Clarke 2014
Manjuan Duan, Elsner, Marie-Catherine de Marneffe 2013

We found that:

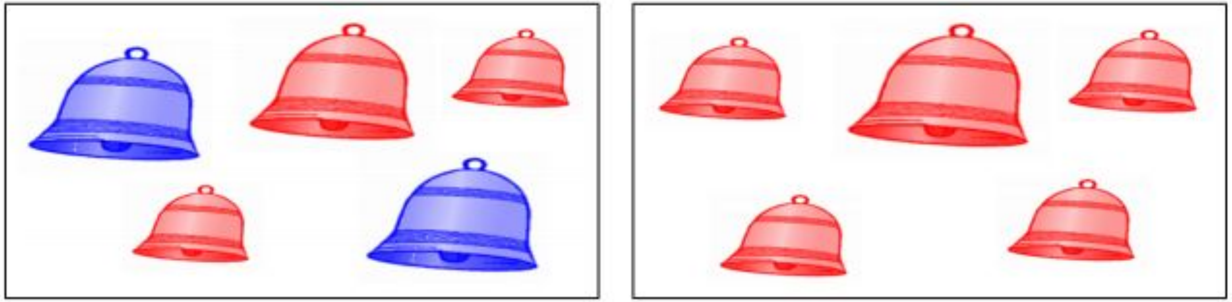
- Visual processing determines how much you say
    - As well as syntax, choice of determiners
- These choices help listeners to find the target faster



Find <est obj="imgID1">the red and white umbrella</est>. Then find <est obj="imgID2">the blue and white beach ball</est>. Below and to the left <lmark obj="imgID2" rel="targ"/> is <targ>a dark skinned woman with a red bathing suit</targ>.

Gatt’s experiments

Reference production as search: The impact of domain size on the production of distinguishing descriptions, 2016
Albert Gatt, Emiel Krahmer, Kees van Deemter, Roger van Gompel

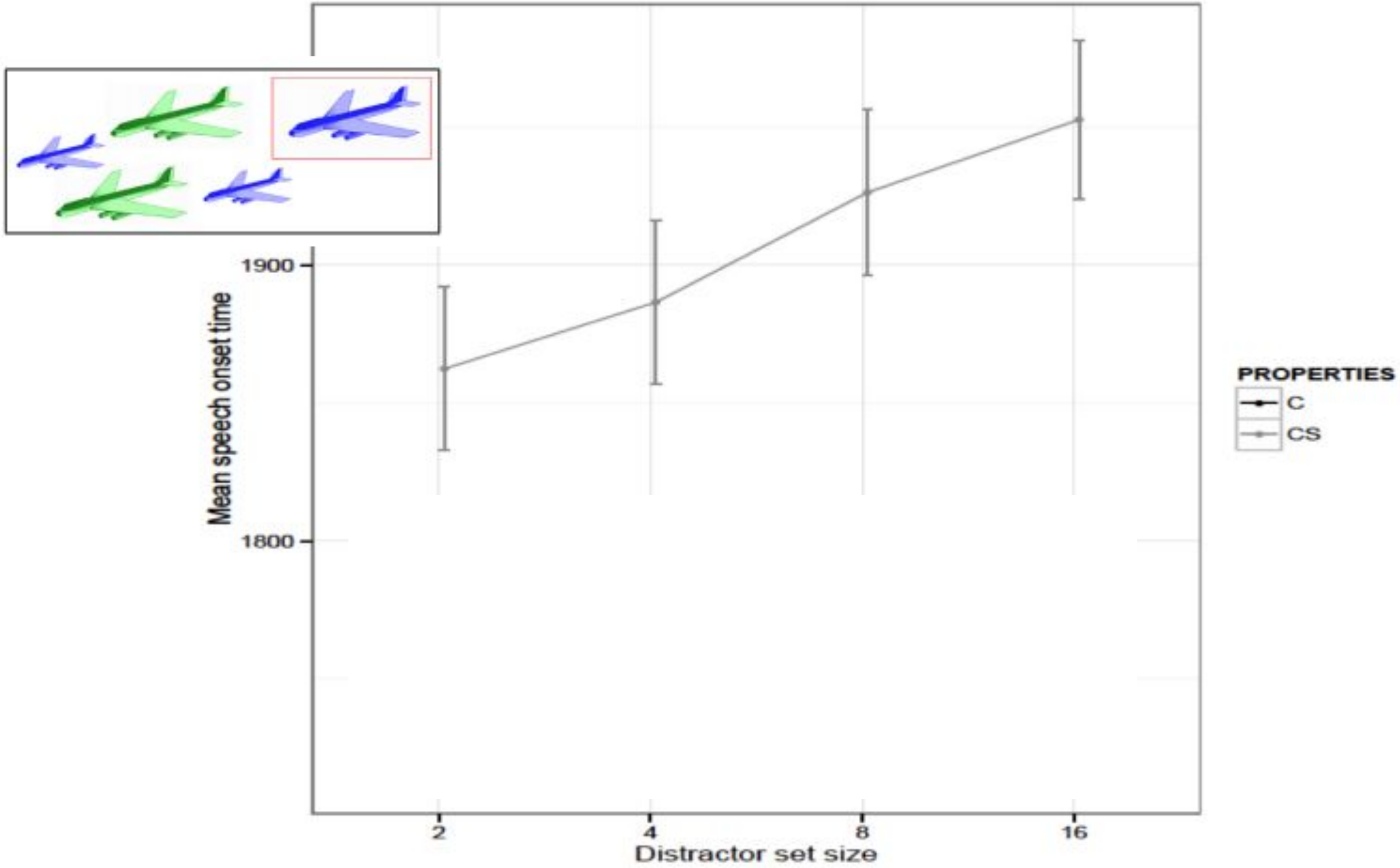


(b) A *large red bell* among large blue and small red distractors

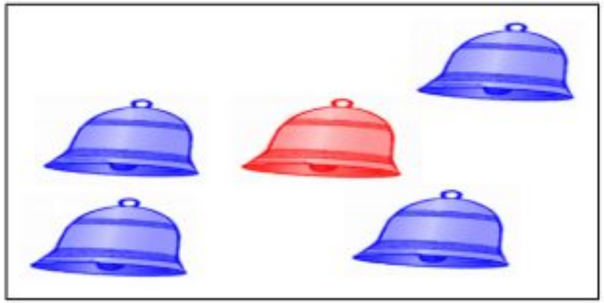
(c) A *large bell* among smaller distractors

Gatt varied the number of bells in the scene...

Time
before
speech
onset



But some cases are easy

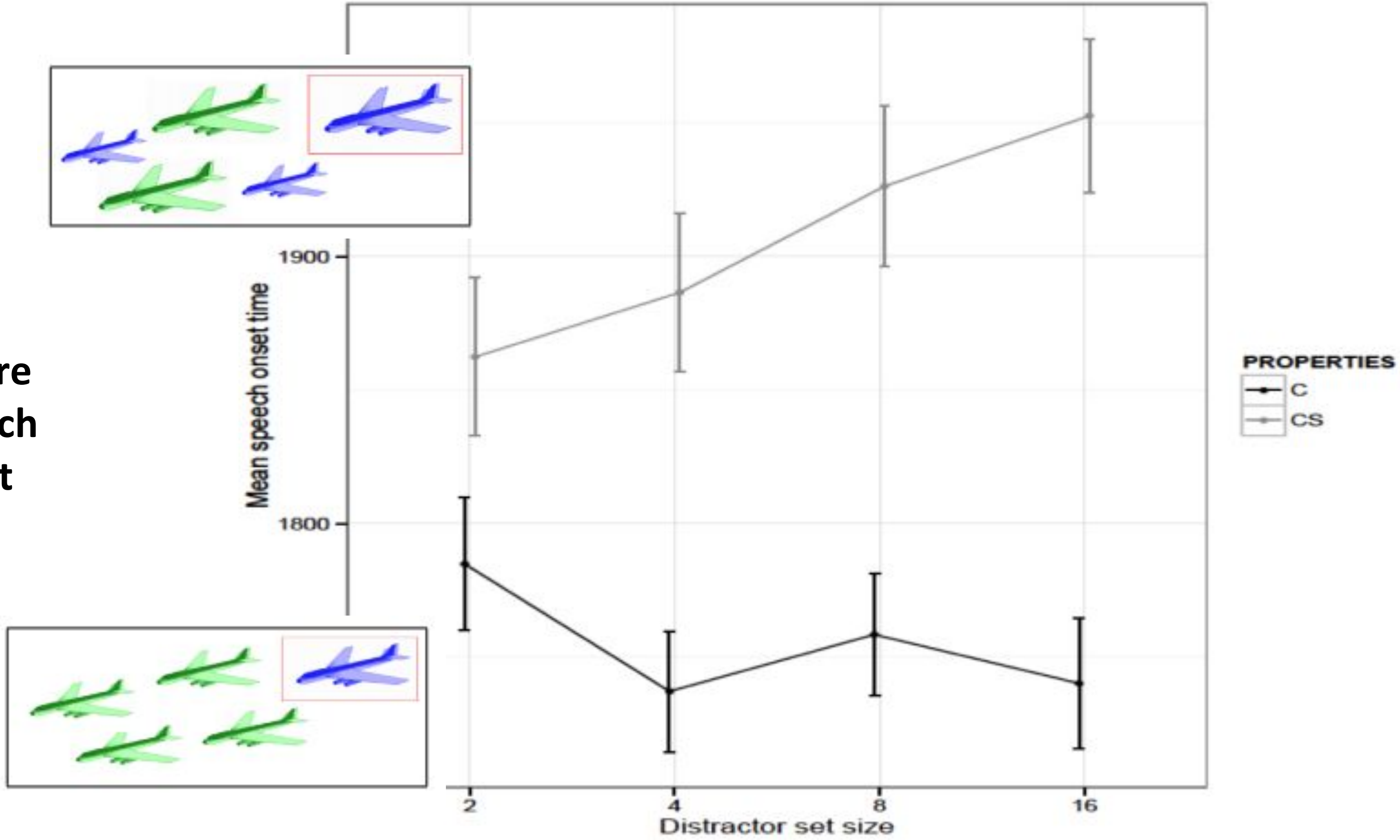


(a) A *red bell* among blue distrac-

tors

Gatt varied the number of bells in the scene...

Time before speech onset



Why model?

At a high level, we can guess that these effects come from visual processing…

But outside of carefully controlled stimuli, it's impossible to tell how strong these visual effects might be

AI to the rescue?

Perhaps we could use a *model* of a person?

In the rest of the talk, I'll show a proof-of-concept:

A model capable of replicating Gatt's result using machine learning

But that might later be applied in more realistic stimuli

Back to the basics

Before we start modeling anything complicated, let's go over some building blocks

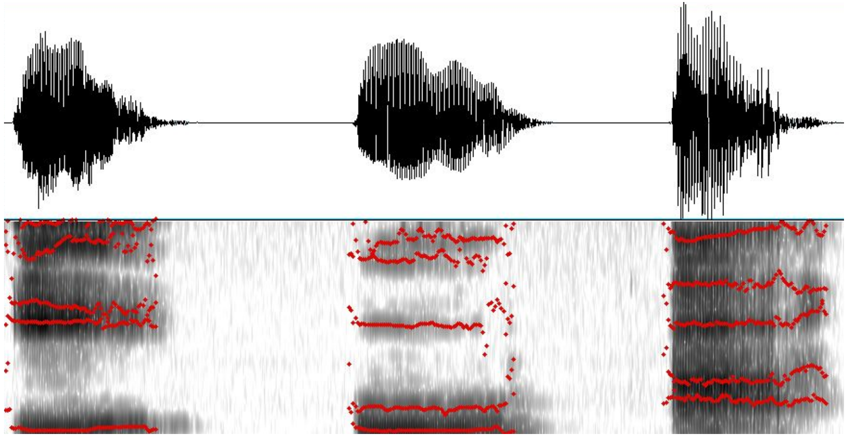
- **Machine learning**: use **training data** to discover a **decision-making function** which can make **predictions** about **unseen data**

Speech recognition: oversimplified

Three English vowels

Amplitude

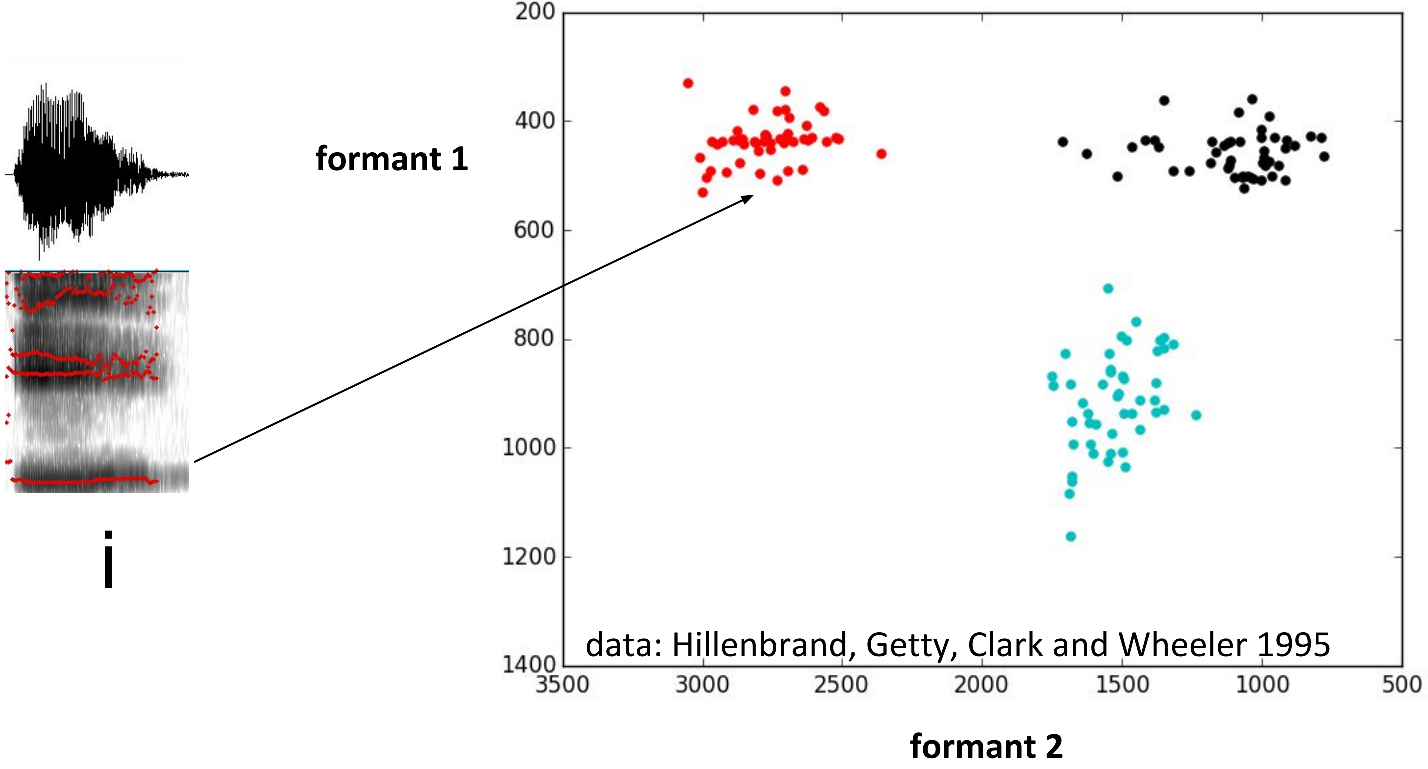
Frequency



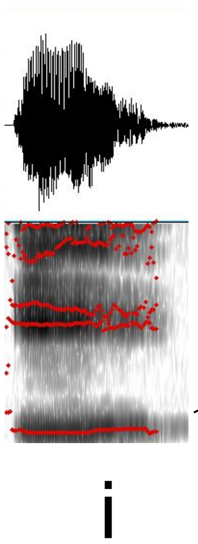
i u ɑ

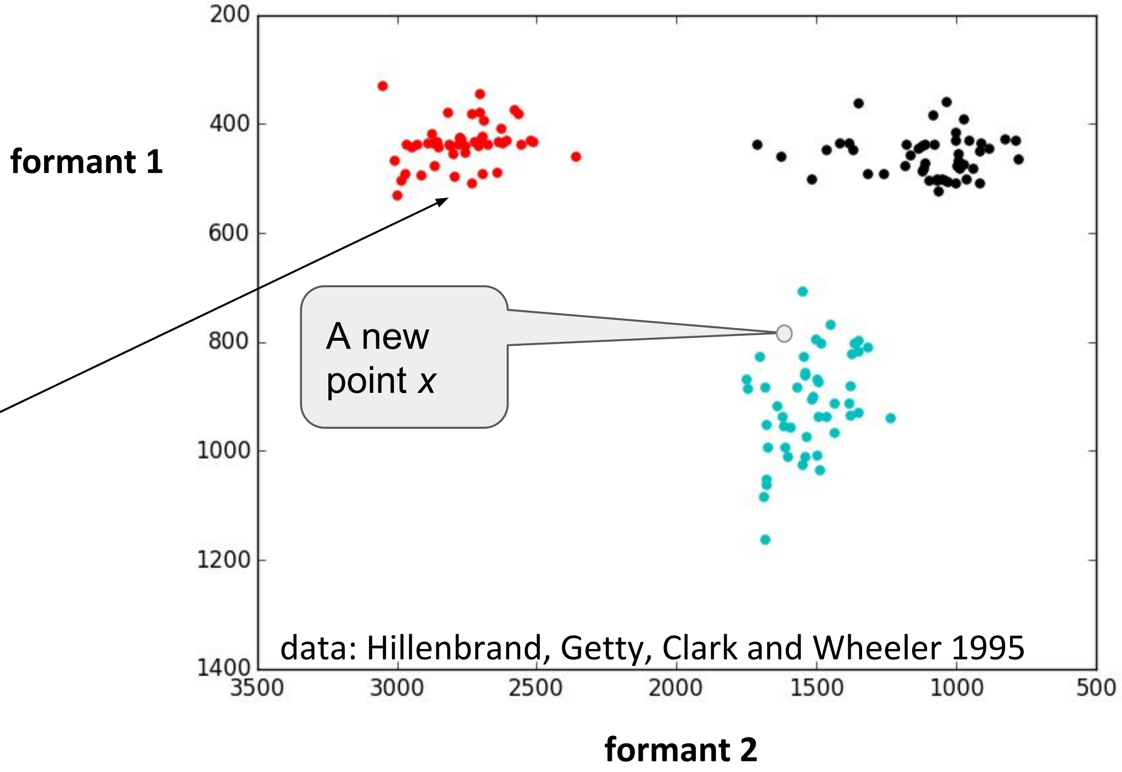
Harmonics are **formants.** Lowest two formants identify the vowel.

Training data



Training data





A linear model

We can write down what we know about the speech dataset as a **linear equation** with unknown coefficients:

Some numbers we don't know (*parameters*)

$$y = w_1 x_1 + w_2 x_2 + b$$

The *score* (which class is this point in)

The coordinates of unknown point *x*

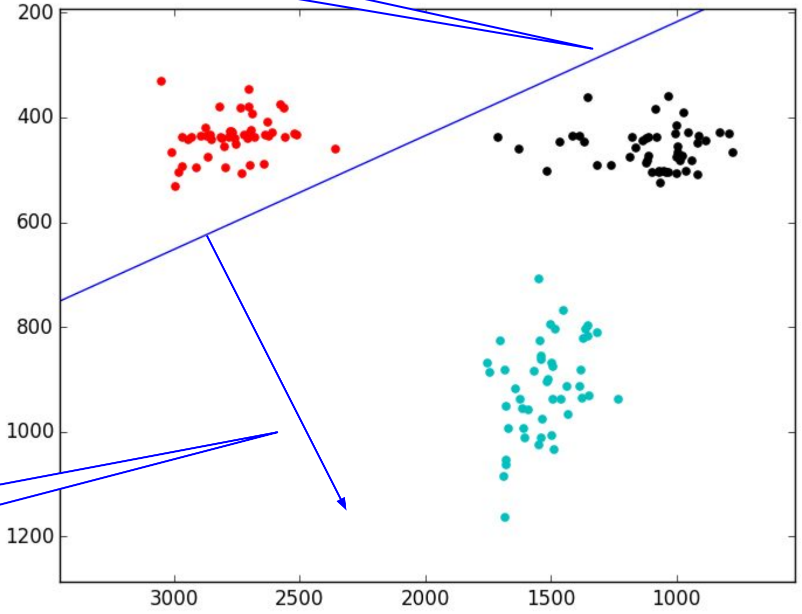
A linear model

x: input features
y: predicted category
w, b: parameters of the learned classifier

$$P(y|x) = \frac{1}{1 + exp(w \cdot x + b)}$$

Decision boundary

formant 1



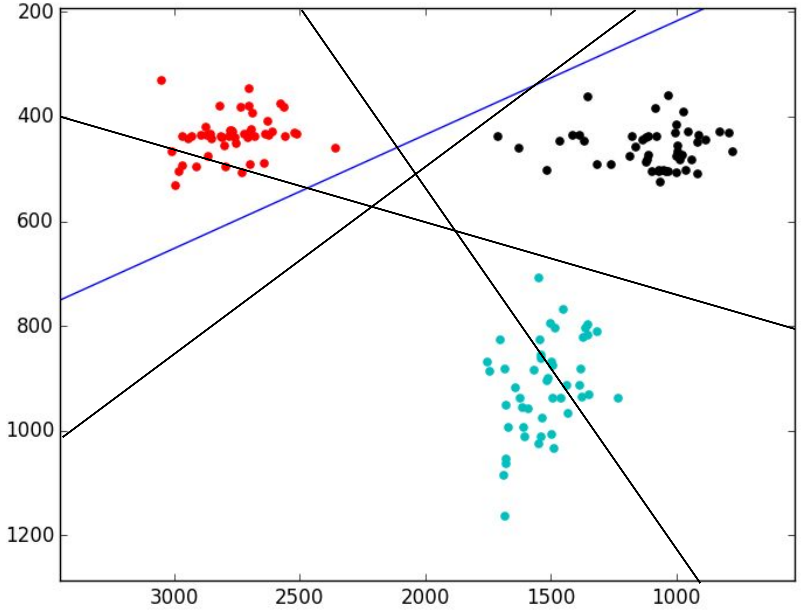
formant 2

Direction of *w*:
Probability of an *i* vowel falls off exponentially

Obviously, we need the right parameters

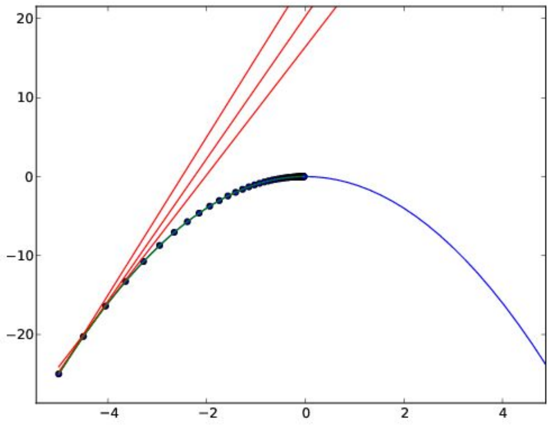
We don't know how any given equation will actually work on *x*

But we can figure out how well it works on the training data!

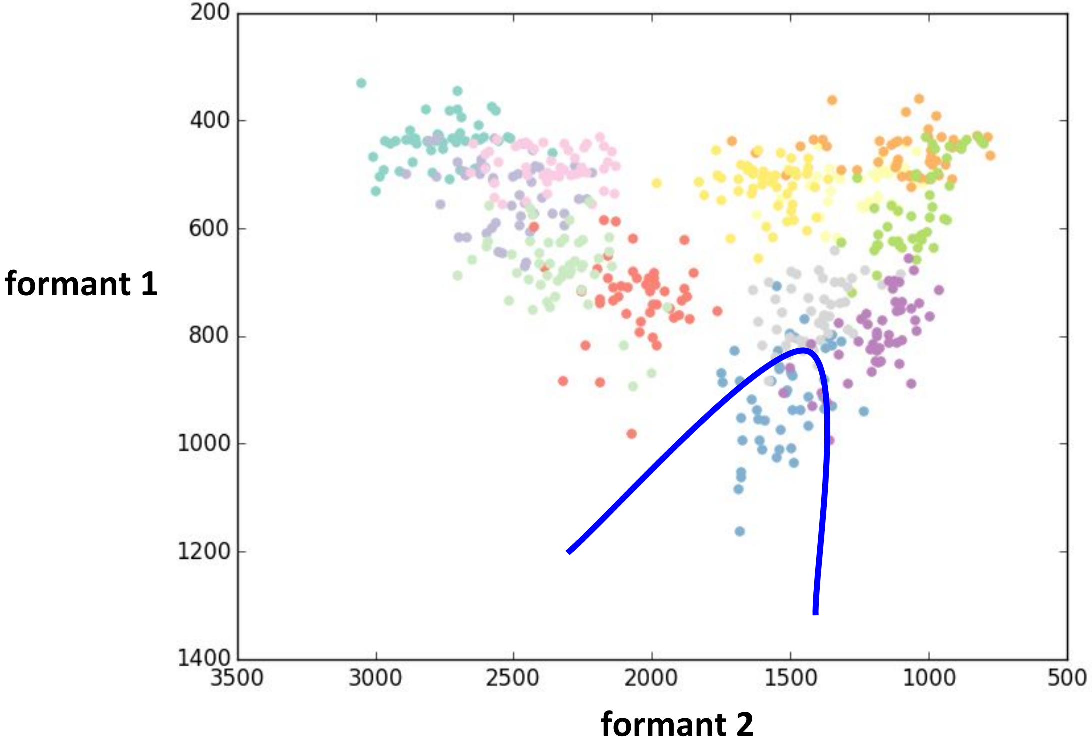


Optimization

- Write down an approximate function for the training error
- Take the derivative
  - Actually, the computer does this
- Find a minimum by hill-climbing (gradient ascent)



But real life is highly non-linear



We could add these terms as features

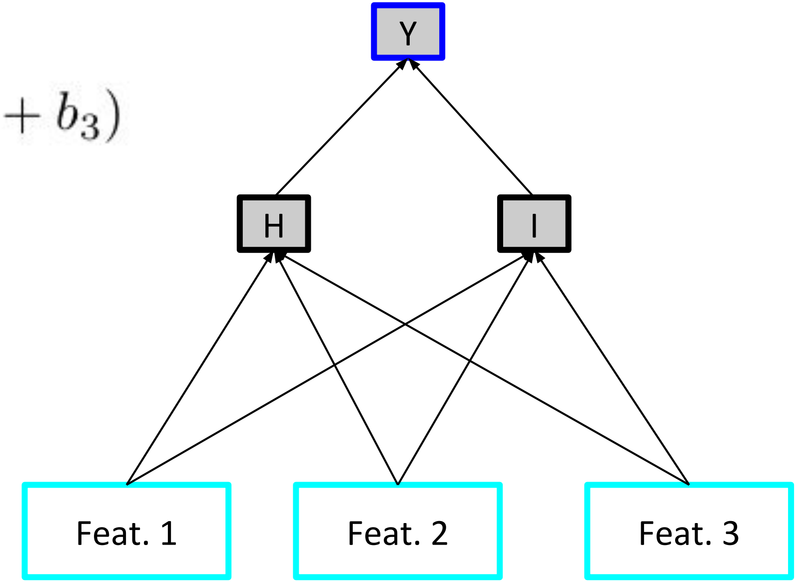
$$P(y|x) = \frac{1}{1 + exp(w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2 + w_6 x_3 + \ldots + b)}$$

But estimating these parameters requires tons of **data**

And increases our ability to learn **spurious correlations** which **don’t generalize** beyond the training data

Limit the number of interactions

$$H(x) = F(w_1 \cdot x + b_1)$$
$$I(x) = F(w_2 \cdot x + b_2)$$
$$P(y|x) = F(w_3 \cdot [H(x), I(x)] + b_3)$$



Model is allowed only two intermediate variables…

But these can summarize any combination of features 1, 2 and 3

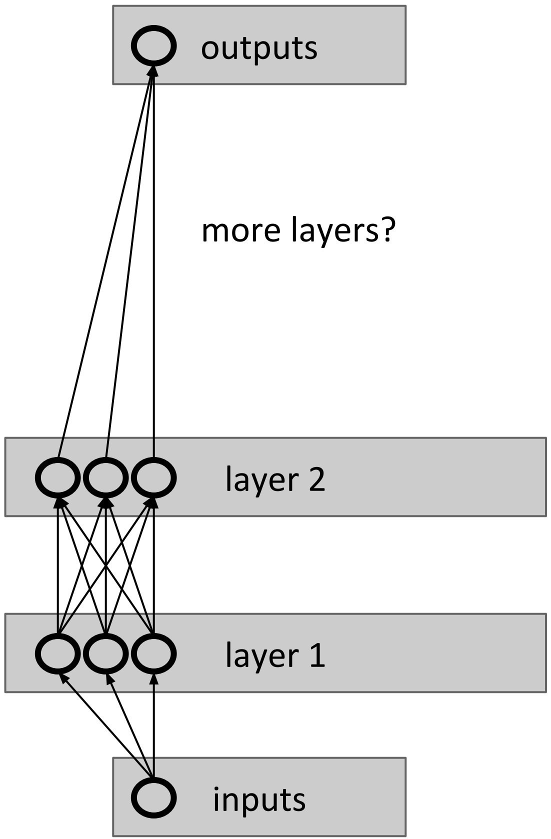
Final decision Y is *non*-linear

Multilayer network

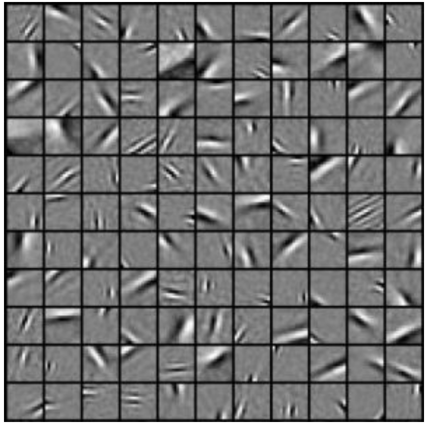
More complicated network topologies are common…

Take advantage of structure in the data:

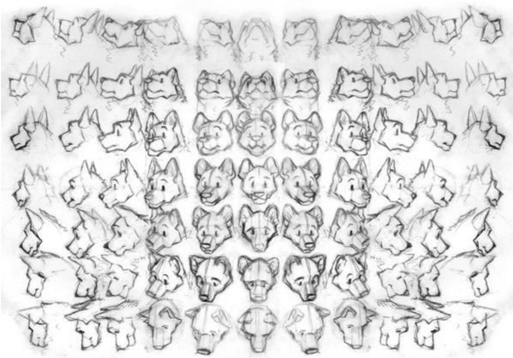
- Temporal (speech ms. by ms.)
- Spatial (nearby pixels in image)
- Source of data (my voice vs. yours)
- Confounding factors (lighting, orientation)



Networks for vision



Low-level receptors

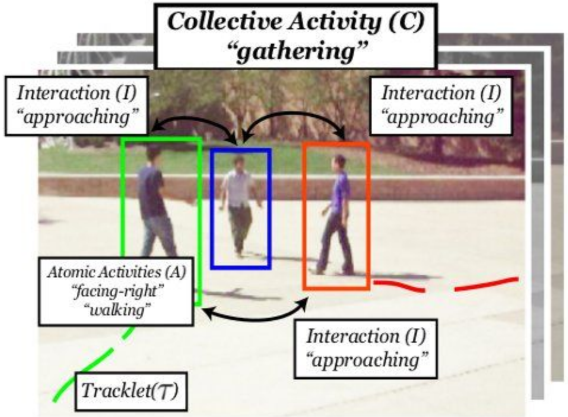


Objects

img: miru3192 on twitter

Hierarchical structure

Choi and Savarese 2012



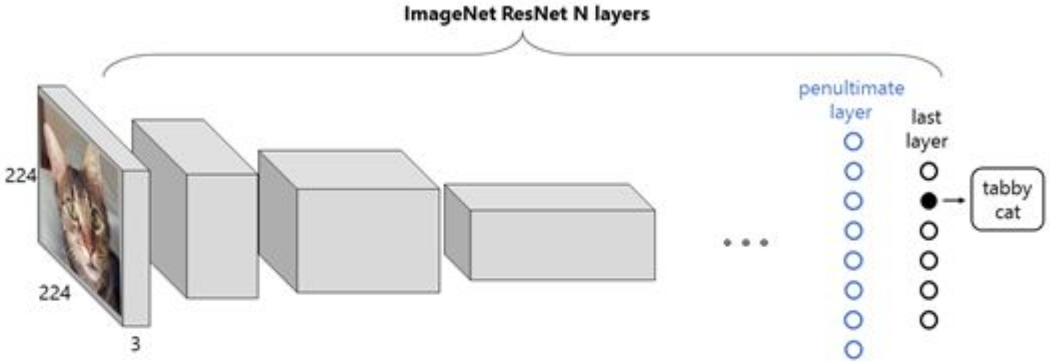
Convolutional neural nets

Basic idea 1: apply the **same learned feature extractor** to each patch in an image

(Animation by Erik Reppel:

https://hackernoon.com/visualizing-parts-of-convolutional-neural-networks-using-keras-and-cats-5cc01b214e59)

Basic idea 2: repeat to create a “deep” architecture with layers of abstraction

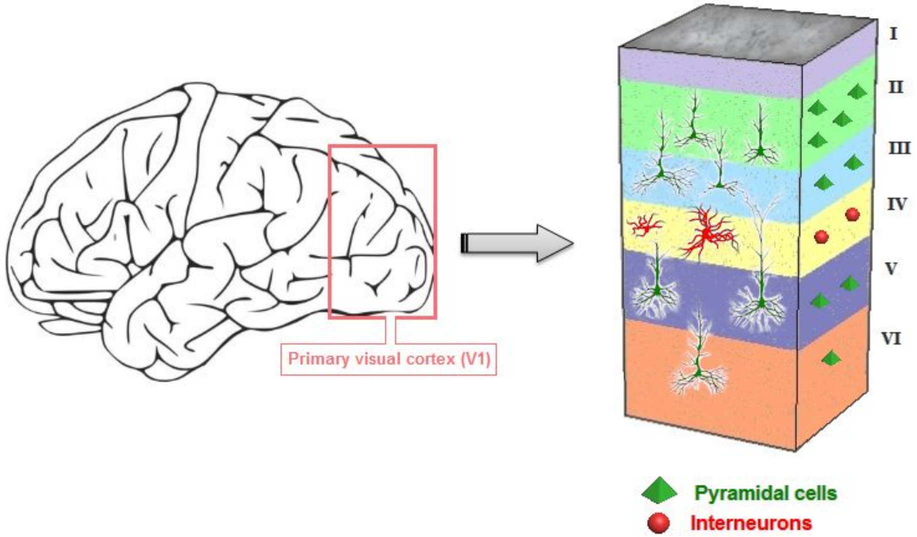


https://blogs.technet.microsoft.com/machinelearning/

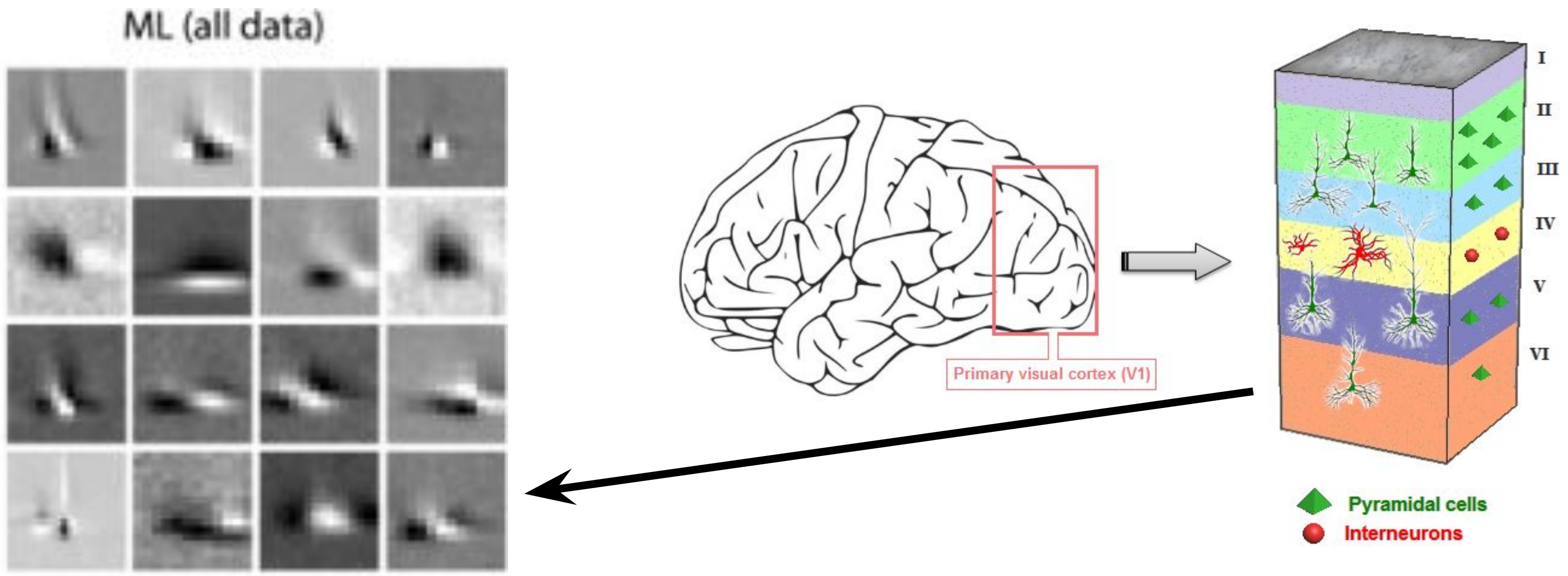
Neural nets and the brain

The visual cortex processes input in layers…

Lower layers detect “low-level” features; higher layers are more abstract



Adaptation and Neuronal Network in Visual Cortex
Lyes Bachatene, Vishal Bharmauria and Stéphane Molotchnikoff

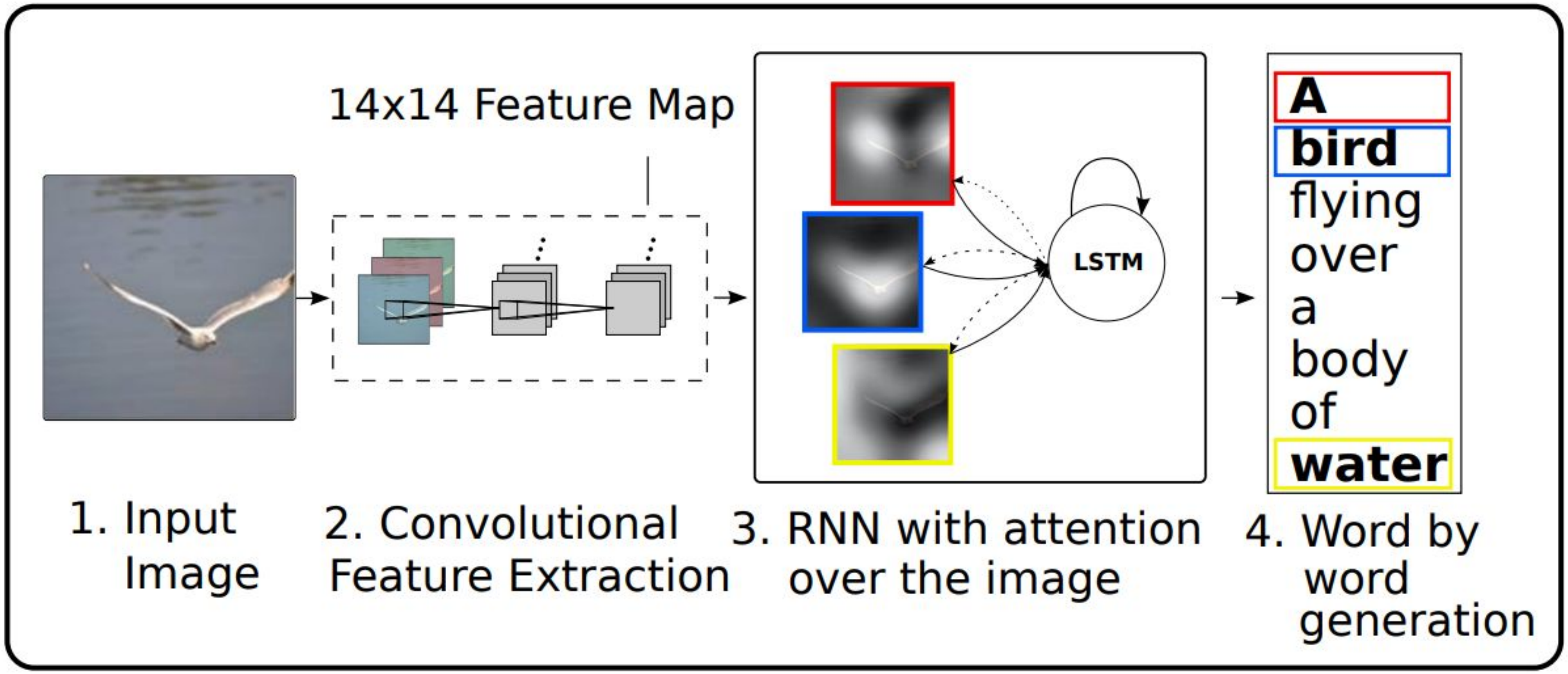


Receptive Field Inference with Localized Priors
Mijung Park and Jonathan Pillow

Adaptation and Neuronal Network in Visual Cortex
Lyes Bachatene, Vishal Bharmauria and Stéphane Molotchnikoff

Captioning with CNNs

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." ICML 2015.



Learning to look

following an approach from Volodomyr Mnih, Nicolas Hees and Alex Graves, 2014

Captioning systems see the whole image at once…

But we’re trying to model a human speaker

So, we need to give it a steerable focal point and let it learn where to look

We use reinforcement learning

As recently used to play video games… and beat the world champion of Go:

https://www.youtube.com/watch?v=V1eYniJ0Rnk

Brief overview of the model

The model has focal and peripheral vision

At every step, it moves the focus point…

And then decides whether to utter a word…

And then which word to say

We’ll see the details in a minute...

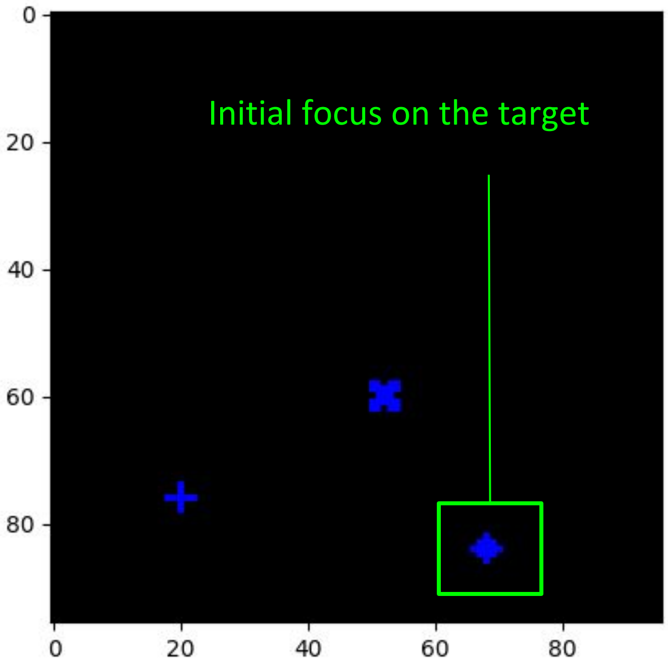
Some results!

Current version is trained on artificial scenes, with captions generated by a simple rule-based strategy:

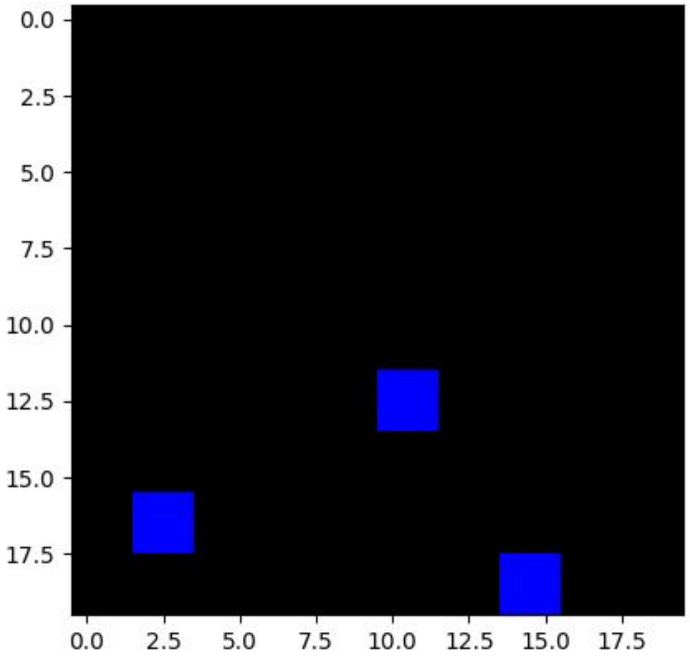
Say whether the target is **unique** or **non-unique**

Model setup

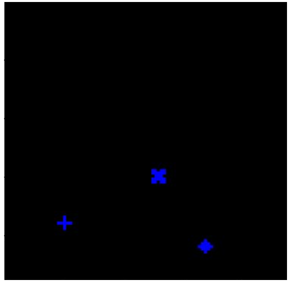
The actual scene



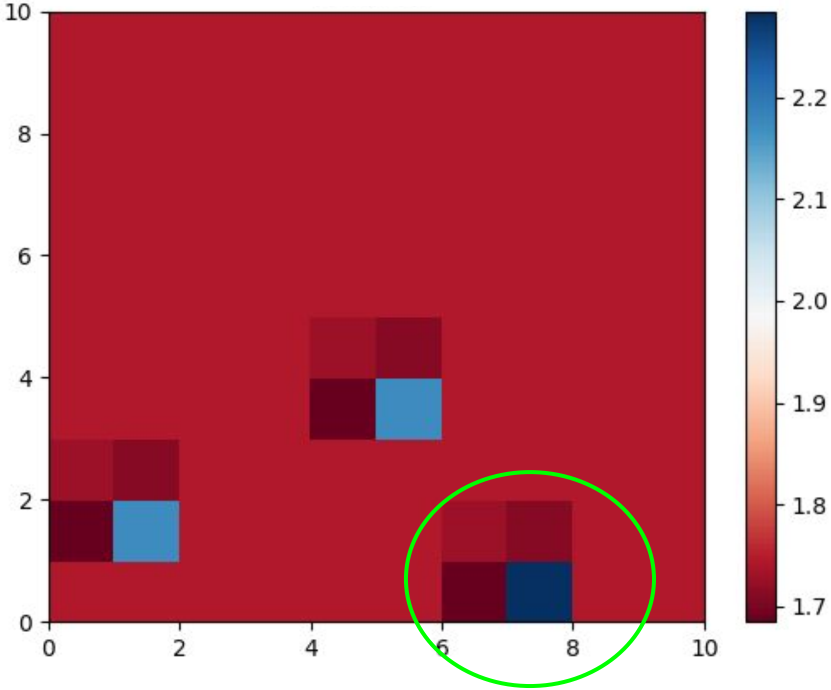
Simulated peripheral vision



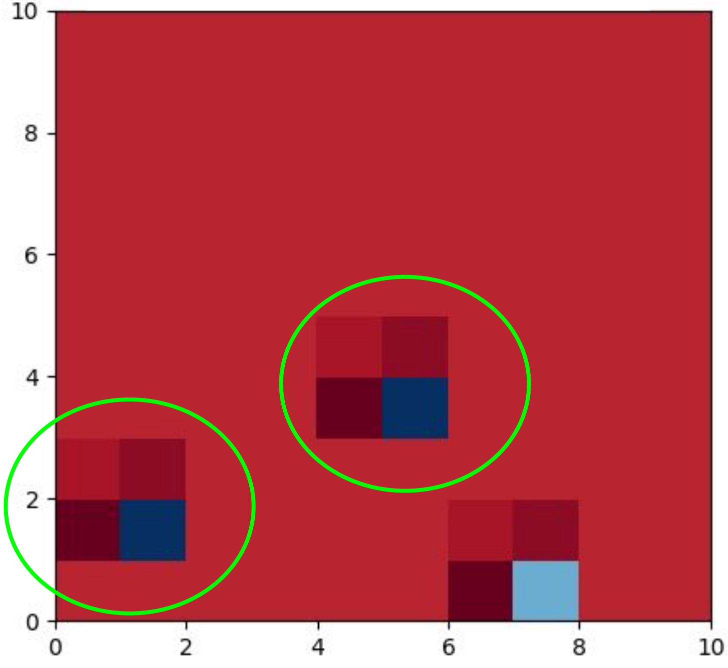
Gaze track



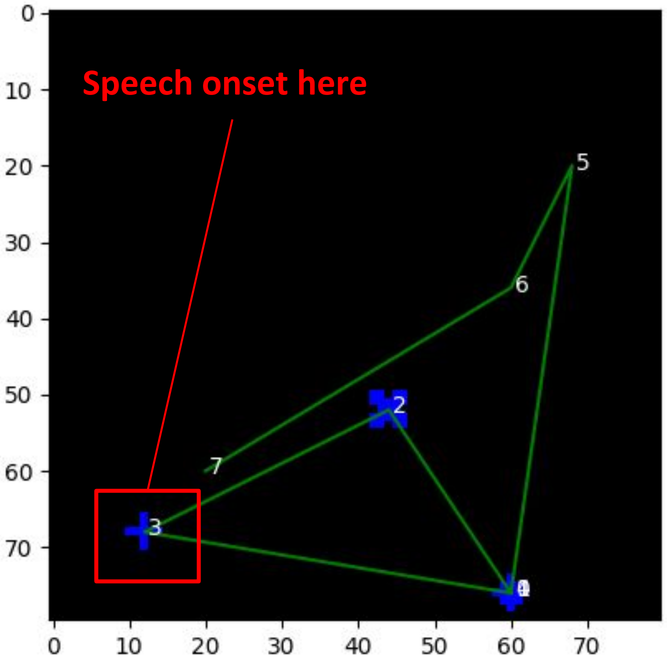
Where does the model want to look first?



What will it do next?



The outcome

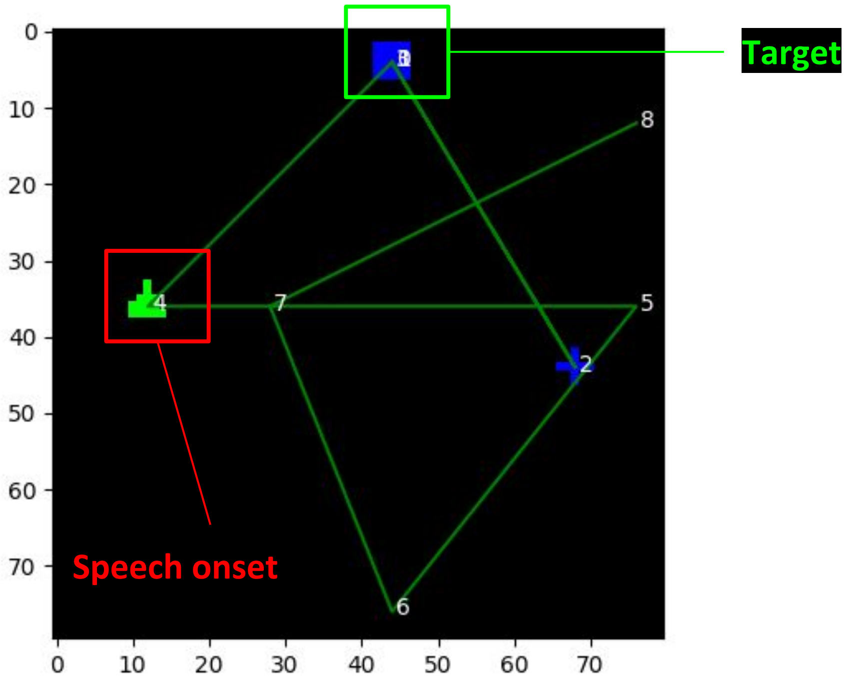


Caption: `unique blue diamond`

What if the image is multicolor?

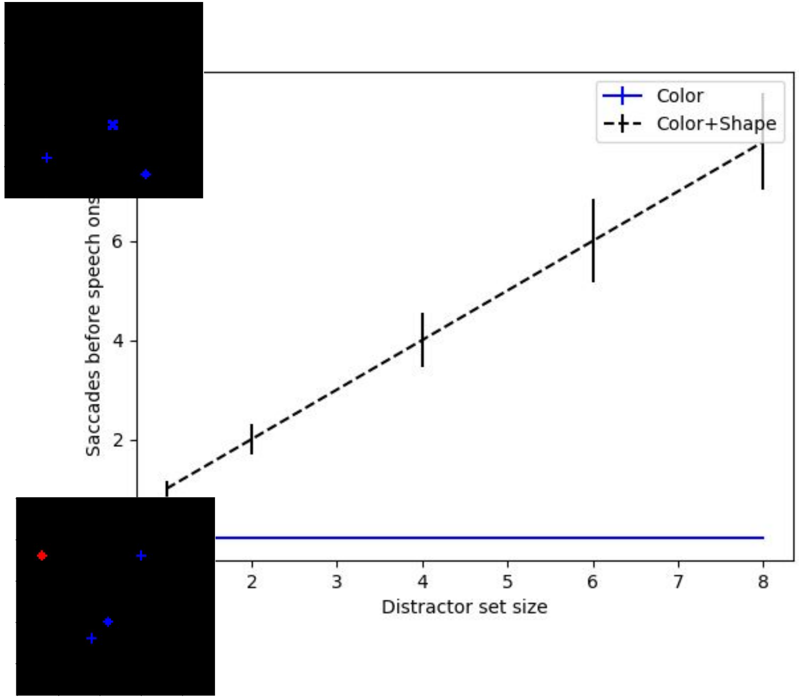
The system has learned to look at the other blue shape first…

But only sometimes, and it hasn't learned to ignore the green one.

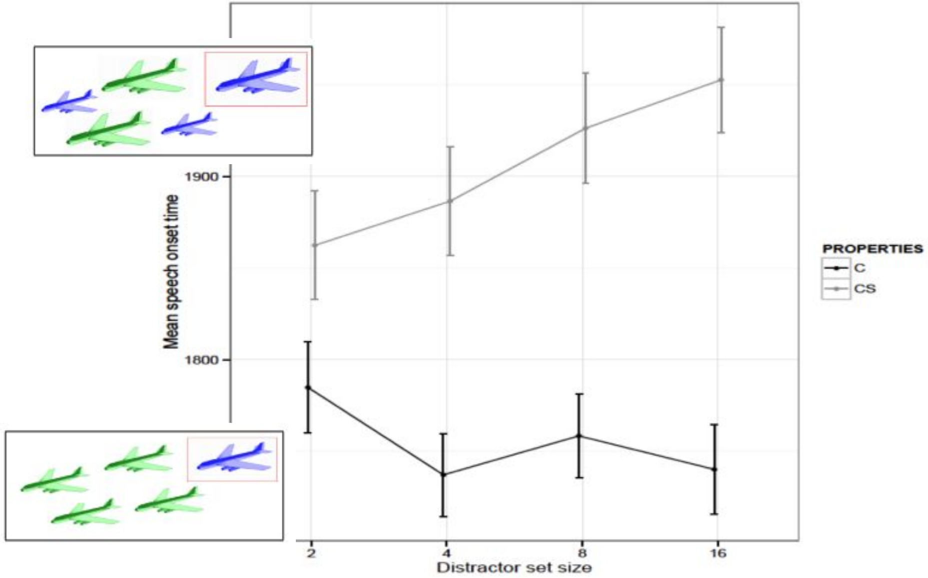


Onset times

Model onset times (saccades)



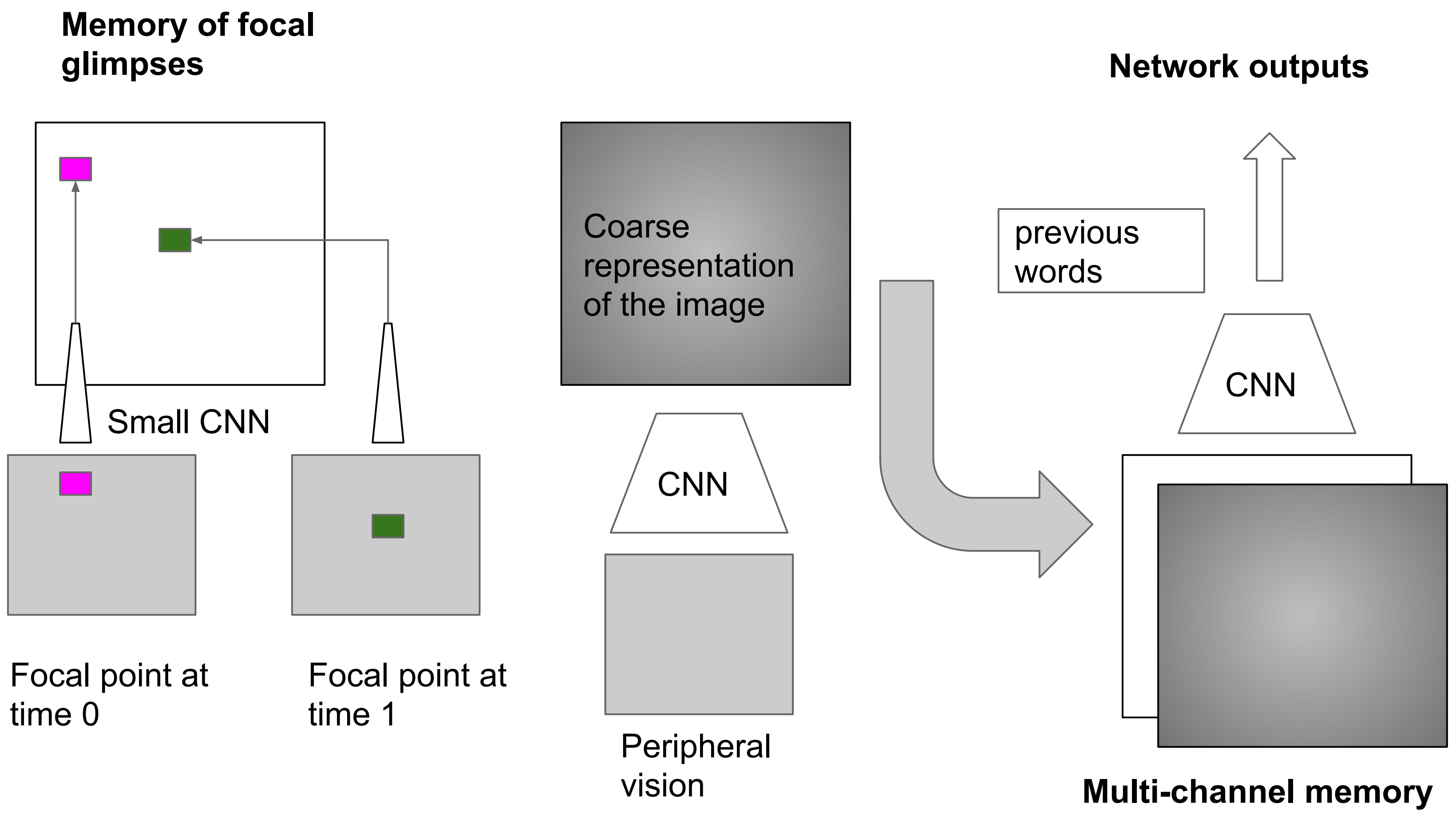
Gatt’s onset times (milliseconds)



Architecture

The model’s “memory” represents visual space

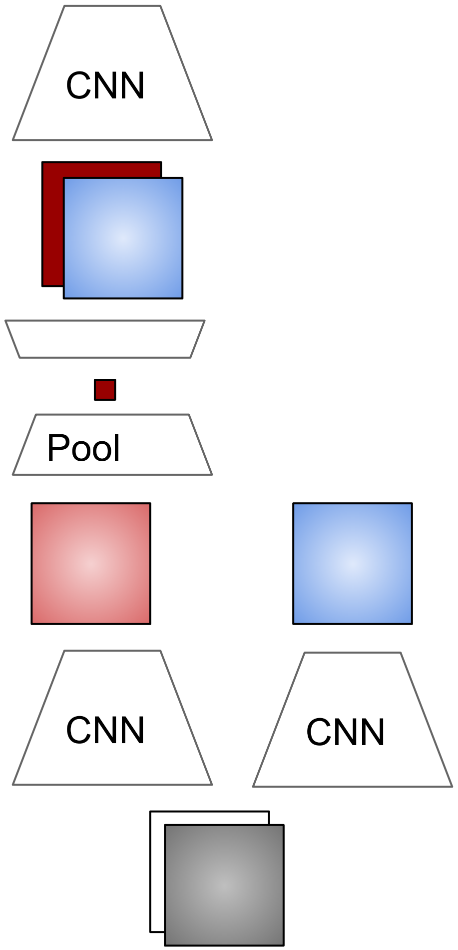
Using a convolutional architecture


Memory of focal glimpses
Small CNN
Focal point at time 0
Focal point at time 1
Coarse representation of the image
CNN
Peripheral vision
previous words
CNN
Network outputs
Multi-channel memory

Issues in network design

How many intermediate layers?

- Currently, using two intermediate convolutional layers with 128 filters
- Intermediate pooling layer allows global information to affect local decisions



Q-learning: quick review

To decide what to do,

$$Q(a_t \mid s_t) = r_t + \text{argmax } Q(a_{t+1}, s_{t+1})$$

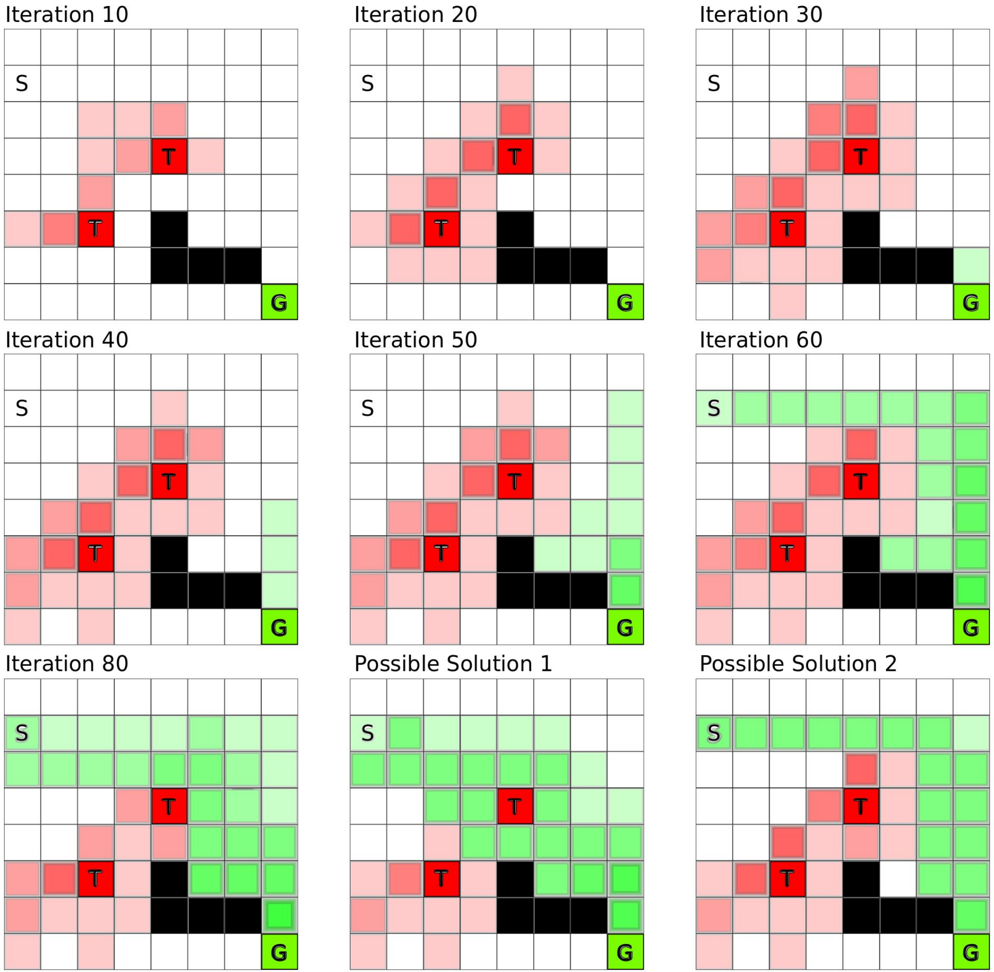
***Q*uality of the current action** in current **state** depends on **local reward** plus **expected future reward**

We don't know the future reward, but we can approximate it:

If our estimate of the **quality** of the **state we end up in** is correct,

Then we can use that to recursively estimate the current reward

Deep Learning in a Nutshell: Reinforcement Learning
By Tim Dettmers
https://devblogs.nvidia.com/deep-learning-nutshell-reinforcement-learning/



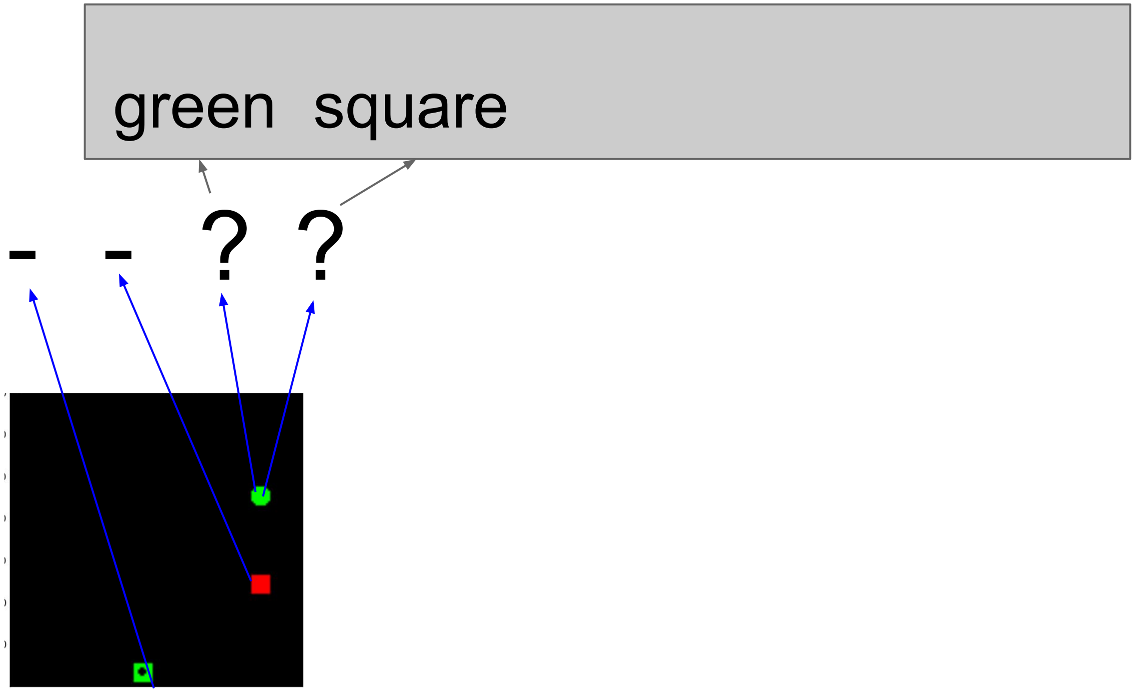
Hybrid Q/supervised architecture

Q-learning introduces opportunities for error, and doesn’t work as well as supervised learning

How to hybridize?

- Supervised learning as initial policy
- Mixed supervised / reinforcement objective throughout training
- **Supervised components in reinforcement architecture**

Division of labor



Supervised word prediction

Unsupervised saccades and decision about when to speak

Reward function

Pretty simple:

- -.1 for every blank
- 1 for every correct word
- -3 for the first wrong word (and that state is final)
    - I’ve also implemented versions with error recovery to study speech disfluencies

Q-learning setup

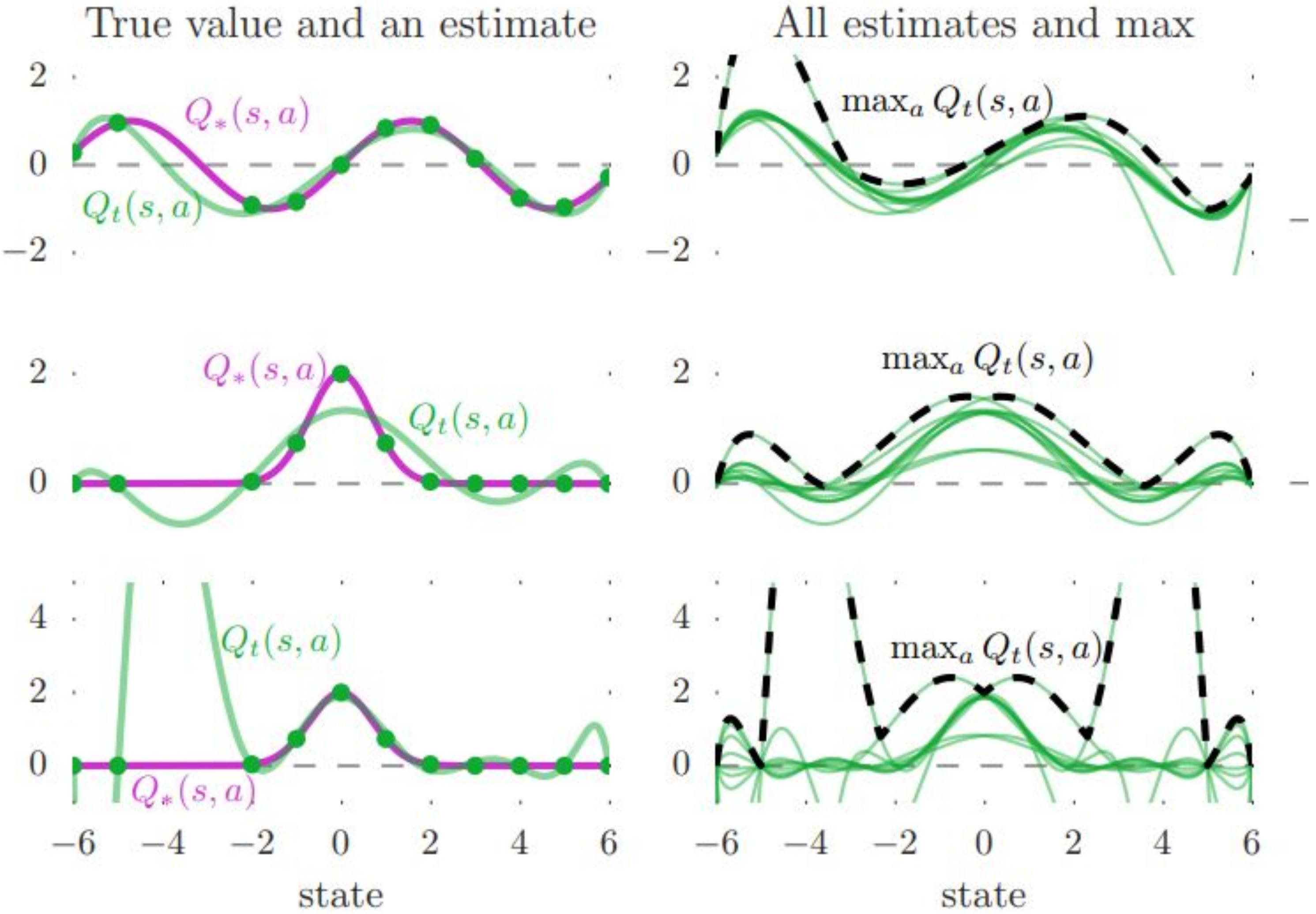
Huge issues with bias in standard Q-learning recurrence:

$$Q(a_t \mid s_t) = r_t + \text{argmax } Q(a_{t+1}, s_{t+1})$$

Why can this go wrong when *Q* is a neural function approximator?

What happens?

Deep Reinforcement Learning with Double Q-learning:
Hado van Hasselt, Arthur Guez, David Silver



Q-learning setup

Replaced learning rule with double-Q-learning-esque:

$$Q(s_t) = r_t + V(s_{t+1})$$

Where *V* is a separate network

Still had issues, so switched to training *V* to predict the empirical rewards:

$$V(s_t) = \sum_{i=t} r_i$$

This estimate has too high variance for Go games, but it’s unbiased...

Concrete results

In these simple images, the model's captions are 95% correct

And, as shown above, it replicates Gatt's onset timing predictions

In another experiment, I show that the model can produce some human-like disfluencies ("red big square") as a byproduct of learning to recover from errors

Future work

Photorealistic images will require a better “visual system” with deeper CNNs

I’d also like to study more complex captioning strategies

Conclusions

Pauses, fillers and mistakes can teach us about how language works in the mind

To understand visual language, you have to think about the visual system

Computer models can help us test theories about complex behavior

Disfluency

So far, we’ve looked at silent pauses…

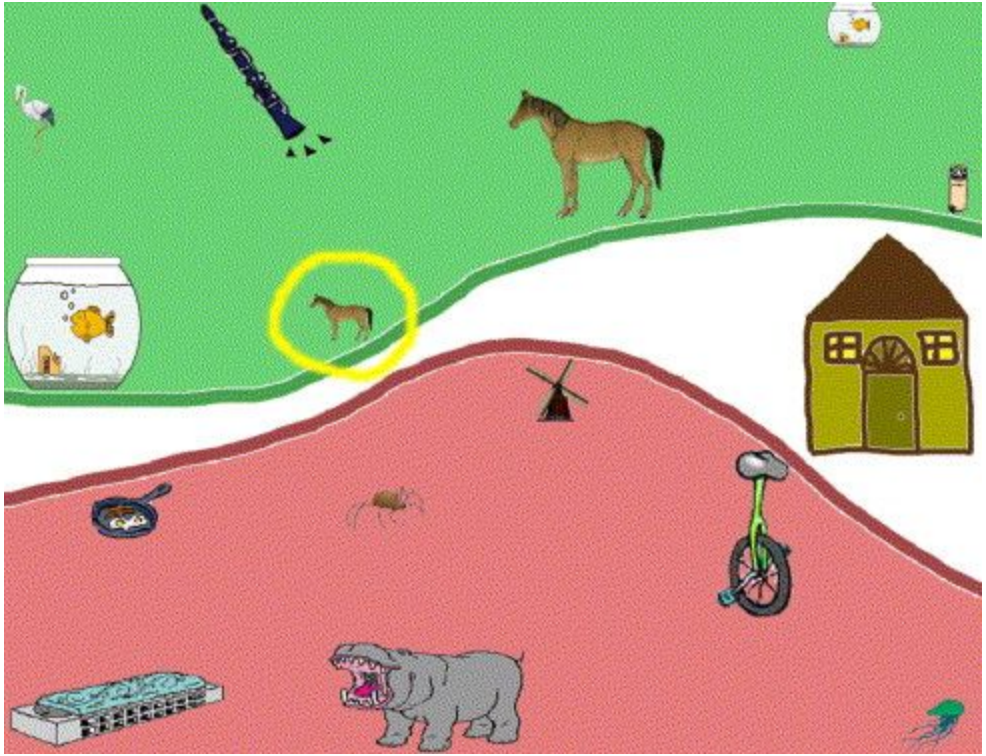
What about filled pauses (“um”)?

And outright rephrasing:

(“the galaxy is… saying it’s…)

Vision as a source of speech errors

Is this a “small horse” or a “horse”?
When would you expect one vs the other?



Watching the eyes when talking about size: An investigation of message formulation and utterance planning
Sarah Brown-Schmidt, Michael Tanenhaus

Speakers use the adj. adaptively

“small” used 72% of the time when there was a different-sized horse

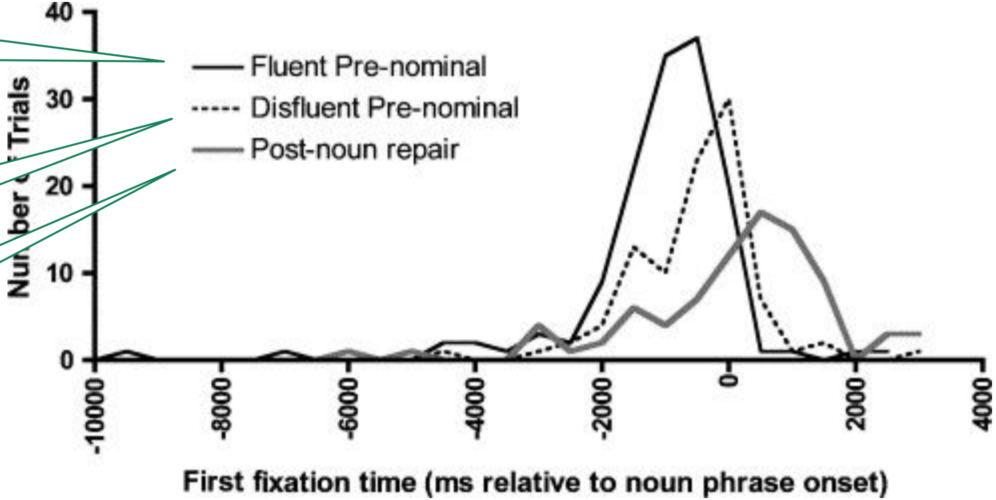
- Of these, 62% had normal modifier order:
  - “The small horse”
- 37% had speech repairs:
  - “The horse… OH, the small one”

Eye track: first look at large horse

“the small horse”

“the… uh... small horse”

“the horse… uh… small one”



Can we relate this to our data?

At a high level, we can guess that these errors come from visual processing…

But without carefully controlled stimuli like Brown-Schmidt’s, understanding individual errors is way too expensive and time-consuming

Pause length

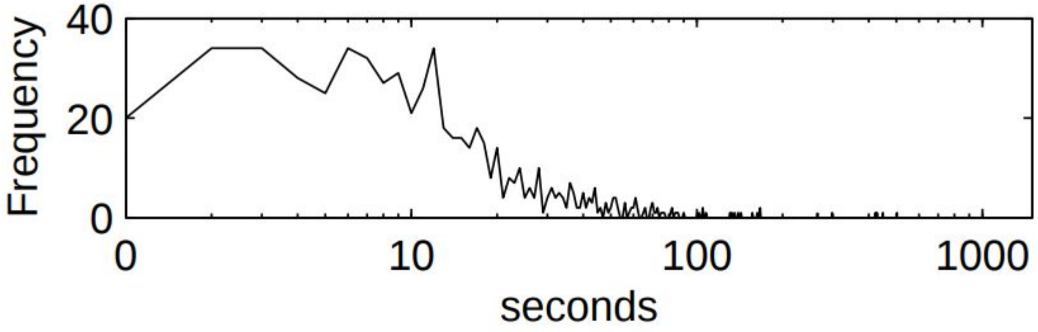
How long is the typical pause?

How long can you pause before another person will start talking?

Back in grad school, I did this for IMs

Online, people sometimes wait a very long time...

Elsner and Charniak 2008



In real conversations...

How long is the typical pause?

Sacks and Schlegoff '78 find that **no perceptible gap** is the most common!

How long can you pause before another person will start talking?

Denny '85: if you make eye contact, around 1-1.5 seconds

One particular effect

from studies by Judith Degen:
images from talk at OSU, 2017

What would you call this?

One particular effect

How about this?

Adjectives mark contrasts

People don’t usually say “yellow banana” unless there’s another colored banana around

But to figure out the contrast, you need to do some searching...