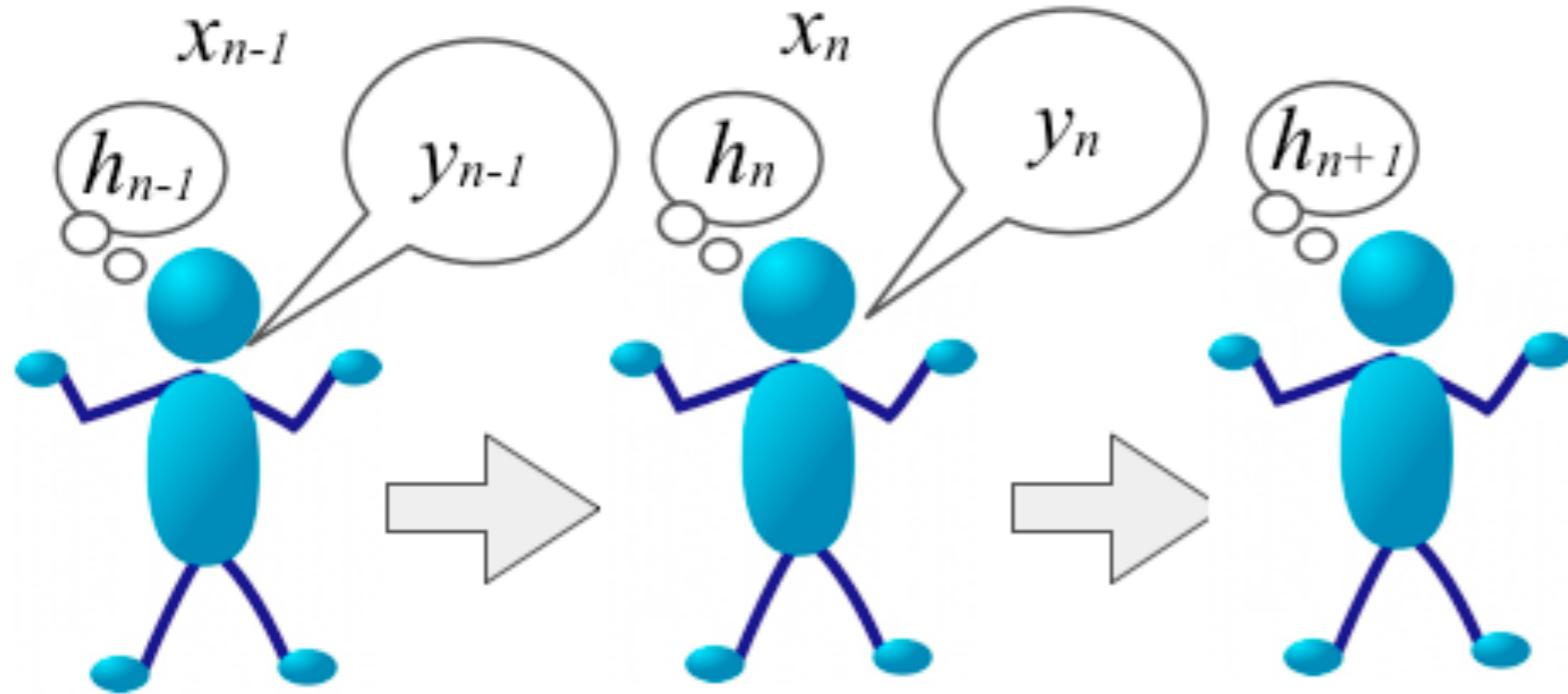
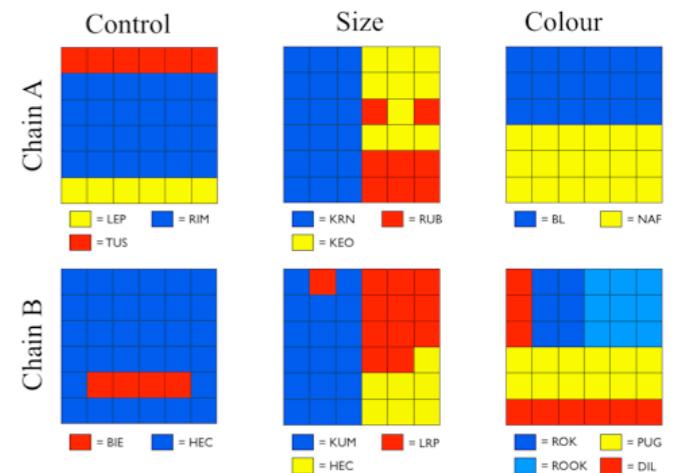
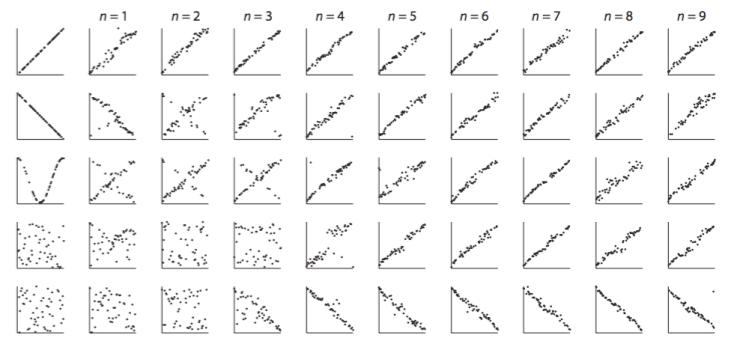
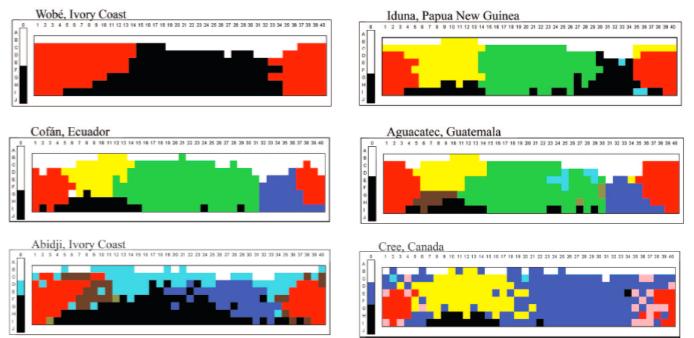


# Computational Cognitive Science



Lecture 14: Iterated learning  
(continued)



# Last time

---

- We started looking a little about how to study things that change over time. As a first stab, we began looking at conceptual change /evolution over time



**Ancient Greece/Rome:**  
- marriages arranged  
- affairs (for men) okay,  
including with young boys  
- not usually for love



**medieval times:**  
- still usually economic  
- often involved dowries  
- women were property  
- church involved more



**19th century times:**  
- occasionally for love  
- often not cross-racial  
- women sometimes can  
keep property



**1950s etc**  
- often for love  
- cross racial sometimes ok  
- women "in the home"

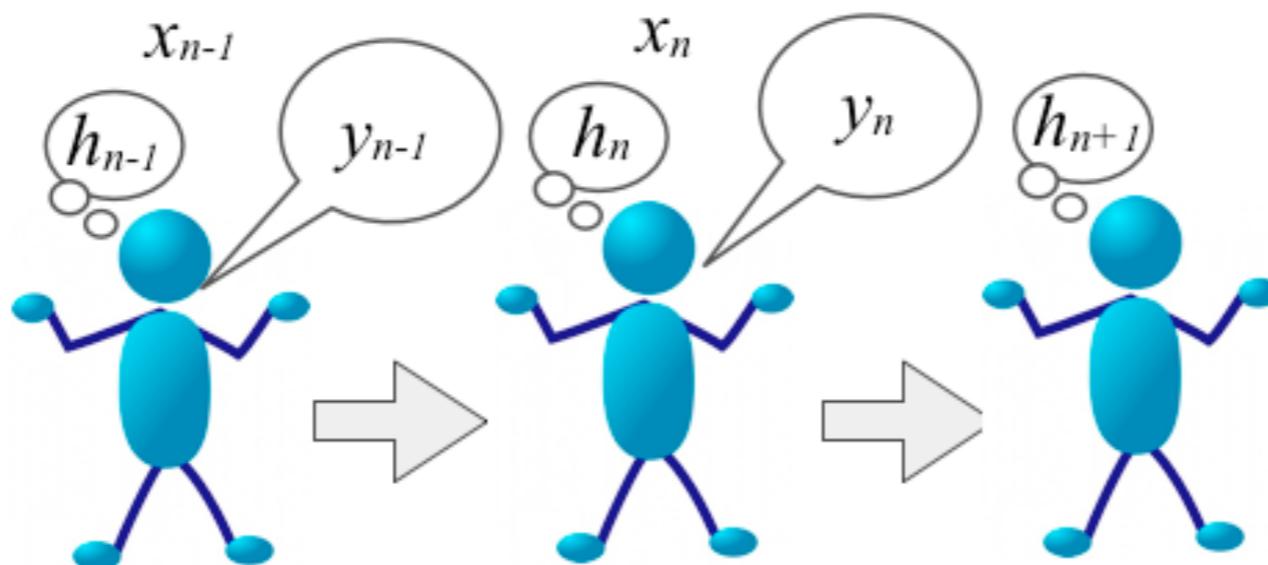


**today:**  
- usually for love  
- cross-racial okay  
- same-sex sometimes ok  
- women wield much  
more economic power

# Last time

---

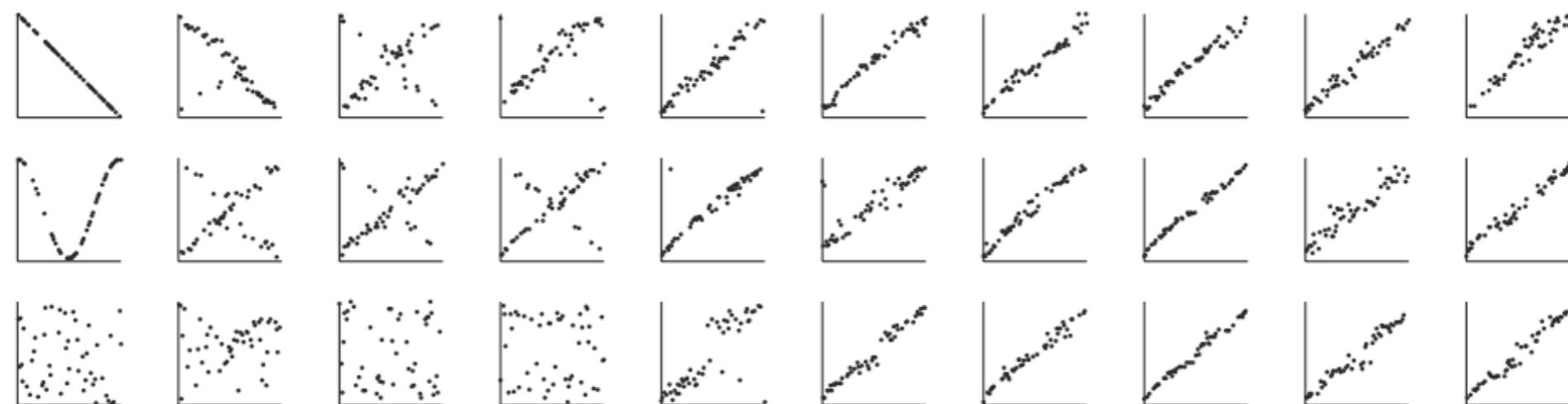
- ▶ We started looking a little about how to study things that change over time. As a first stab, we began looking at conceptual change /evolution over time
- ▶ This is modeled as chains of learners who pass each other information, and are individually Bayesian in how they learn from the previous one



# Last time

---

- ▶ We started looking a little about how to study things that change over time. As a first stab, we began looking at conceptual change /evolution over time
- ▶ This is modeled as chains of learners who pass each other information, and are individually Bayesian in how they learn from the previous one
- ▶ The main prediction, that the stationary distribution of the chain reflects (only) prior probability, was borne out experimentally



# Last time

---

- ▶ We started looking a little about how to study things that change over time. As a first stab, we began looking at conceptual change /evolution over time
- ▶ This is modeled as chains of learners who pass each other information, and are individually Bayesian in how they learn from the previous one
- ▶ The main prediction, that the stationary distribution of the chain reflects (only) prior probability, was borne out experimentally

But how (and in what ways) are these results dependent on certain assumptions we had to make?

# Today's plan

---

- ▶ Evidence for conceptual evolution
  - Inevitable given noisy transmission
  - Historical record
  - Cultural variation
- ▶ A model of conceptual change over time
  - Iterated learning model: basic idea
  - Mathematical proof and corresponding intuition
- ▶ Experimental evidence for iterated learning models
  - Function learning
  - Language
- ➡ Limitations and extensions to the iterated learning model
  - changing learner
  - changing producer
  - changing how hypotheses map onto the world

# Extending iterated learning

---

- ▶ The iterated learning model as it stands now contains a lot of different assumptions. We can explore the effect of each of them!
  - What learners do with the data (i.e., how they infer hypotheses)
  - How producers produce the data
  - How hypotheses relate to the events in the world

# Extending iterated learning

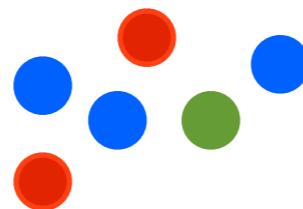
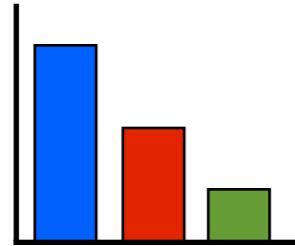
---

- ▶ The iterated learning model as it stands now contains a lot of different assumptions. We can explore the effect of each of them!
  - ➔ What learners do with the data (i.e., how they infer hypotheses)
    - How producers produce the data
    - How hypotheses relate to the events in the world

# Extending iterated learning

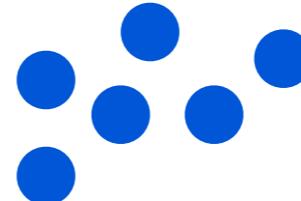
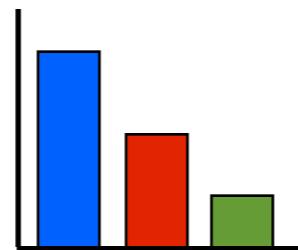
---

- ▶ So far iterated learning assumes that learners select a grammar by sampling it from the posterior distribution over hypotheses



$$P(h_{n+1}|x_n, y_n) = \frac{P(y_n|x_n, h_{n+1})P(h_{n+1})}{\sum_{h \in \mathcal{H}} P(y_n|x_n, h)P(h)}$$

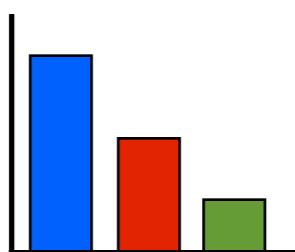
- ▶ An alternative is that learners select the hypothesis that is most probable. This is called **MAP learning** (maximum a posteriori)



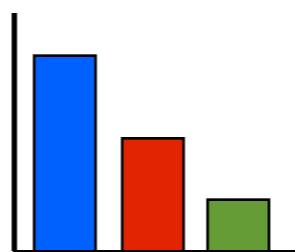
- ▶ If people are MAP learners, how does this change the convergence behaviour of the chain?

# MAP learning

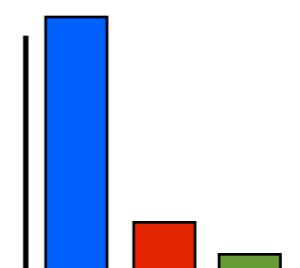
- ▶ If people are MAP learners, it is now much more complicated.
- ▶ Overall, the distribution of hypotheses still reflects the ordering of hypotheses in the prior, but differences are magnified -- the *a priori* most likely hypothesis will be overrepresented



Prior

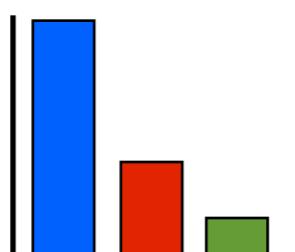


Stationary distribution  
normally

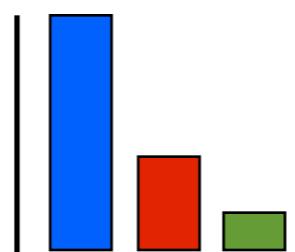


MAP stationary  
distribution

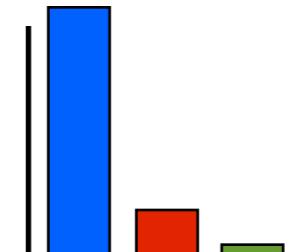
- ▶ This means the same stationary distribution can result from different priors



Prior



Stationary distribution  
normally

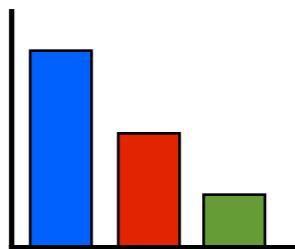


MAP stationary  
distribution

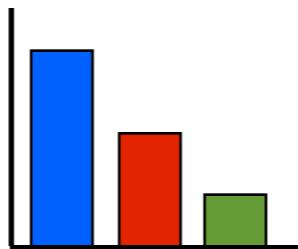
# MAP learning

---

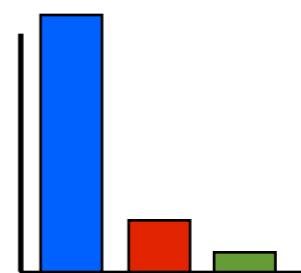
- ▶ If people are MAP learners, it is now much more complicated.
- ▶ Overall, the distribution of hypotheses still reflects the ordering of hypotheses in the prior, but differences are magnified -- the *a priori* most likely hypothesis will be overrepresented



Prior



Stationary distribution  
normally



MAP stationary  
distribution

- ▶ This means the same stationary distribution can result from different priors
- ▶ In addition, changing transmission factors (like the amount of data) can result in convergence to a different stationary distribution

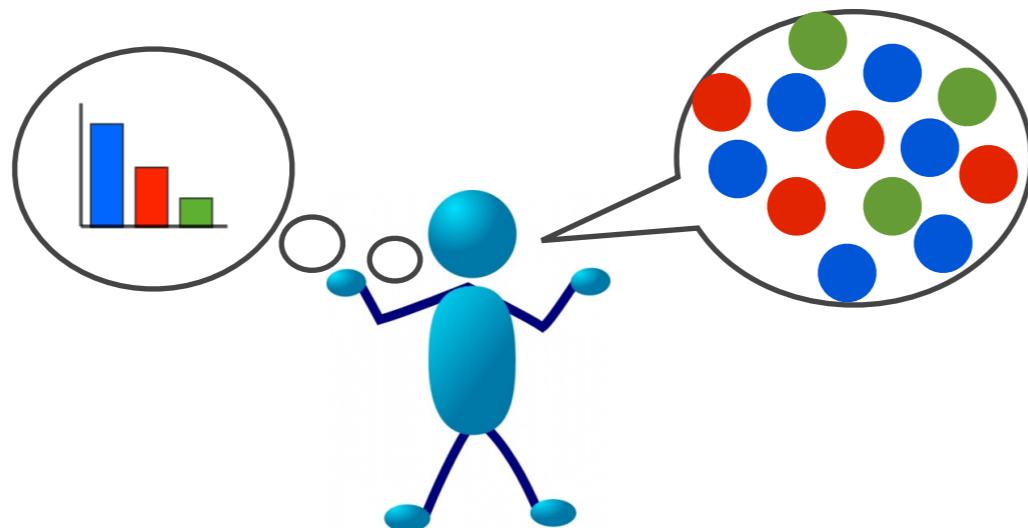
# Extending iterated learning

---

- ▶ The iterated learning model as it stands now contains a lot of different assumptions. We can explore the effect of each of them!
  - What learners do with the data (i.e., how they infer hypotheses)
    - ➡ How producers produce the data
  - How hypotheses relate to the events in the world

# Changing assumptions about the producer

- ▶ Originally we assumed that people produce data by drawing it at random from their inferred hypothesis



$$P(y_{n+1} | x_{n+1}, h_{n+1})$$

- ▶ But in real life, at least sometimes, people can actually just *describe their hypothesis*.



$$P(h_n | d_n)$$

formally, we just assume each agent's prior is the posterior of the previous agent

# Changing assumptions about the producer

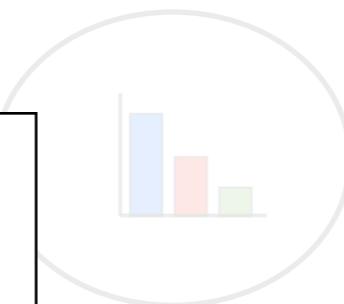
- ▶ Originally we assumed that people produce data by drawing it at random from their inferred hypothesis



$$P(y_{n+1} | x_{n+1}, h_{n+1})$$

- ▶ But in real life, at least sometimes, people can actually just *describe their hypothesis*.

just the same as assuming it  
is one agent who saw all of  
the data over time



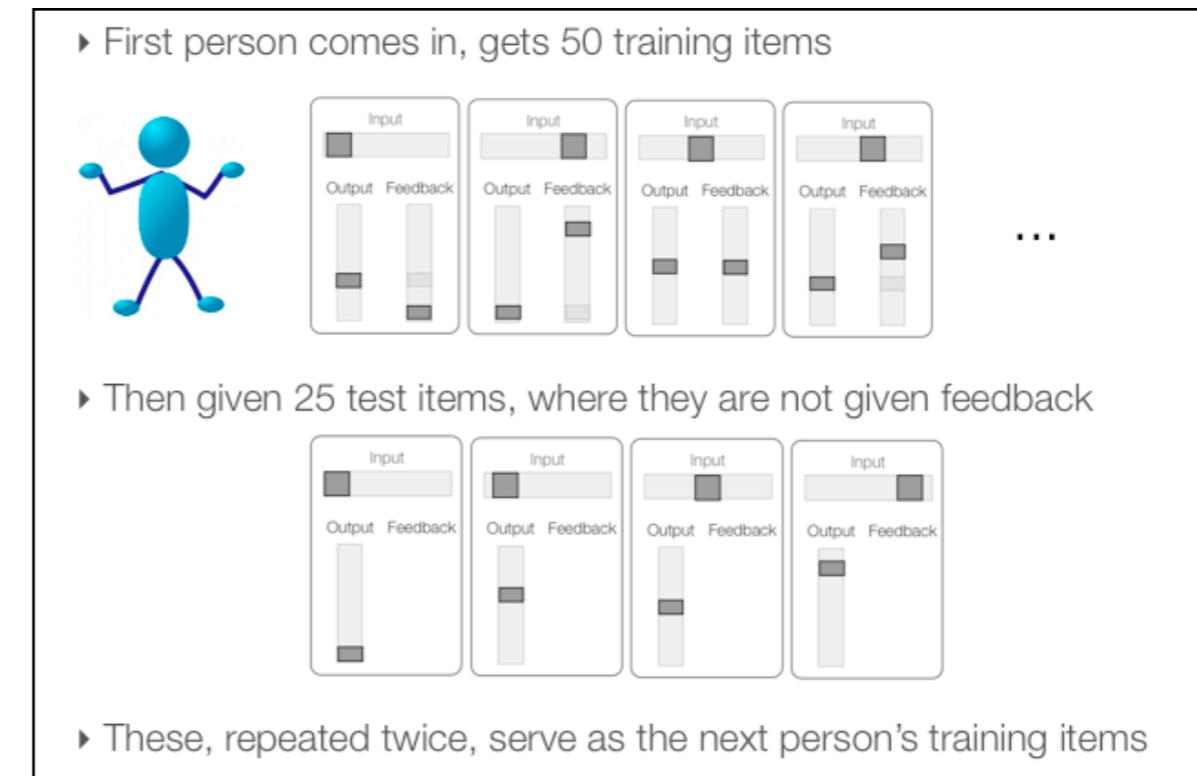
$$P(h_n | d_n)$$

formally, we just assume each  
agent's prior is the posterior of  
the previous agent

# Changing assumptions about the producer

- ▶ Experimental test: vary what information people are allowed to pass along

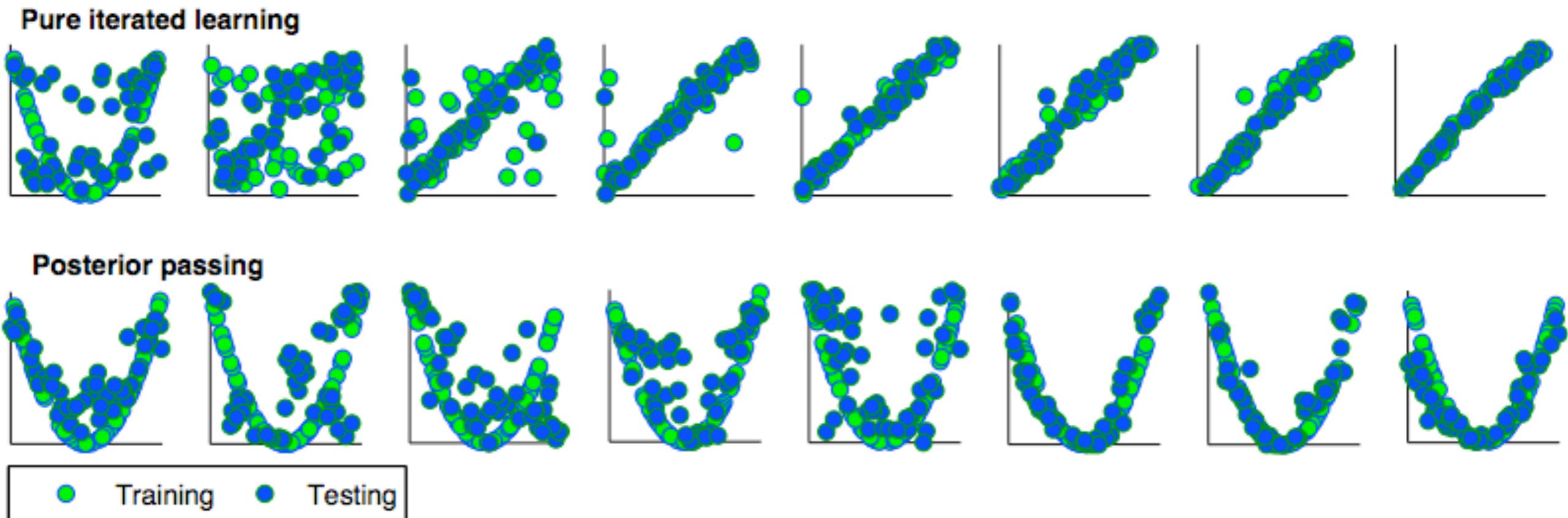
**Control:** typical frequency learning experiment



**Posterior passing:** at end, participants can describe what they thought the function was

This seemed like when the training item is really high or really low, the test is at the top. If the training is in the middle the test is at the bottom.

# Posterior passing changes the distribution!



- ▶ Interesting implications for cultural evolution -- our ability to explicitly pass information along (by teaching) effectively makes humanity able to learn from hundreds or thousands of years of experience, rather than one lifetime. This is called the **cultural ratchet**.

# Extending iterated learning

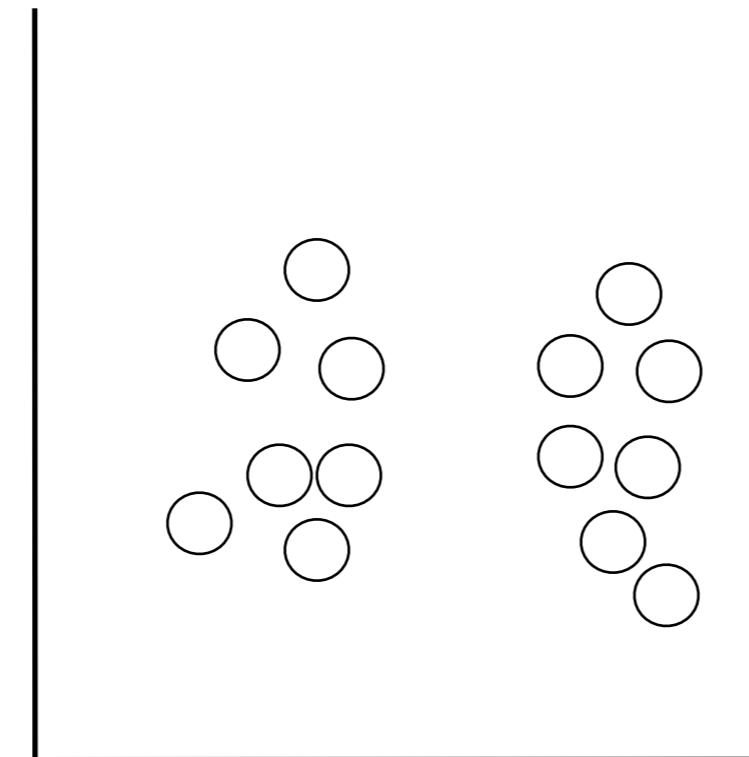
---

- ▶ The iterated learning model as it stands now contains a lot of different assumptions. We can explore the effect of each of them!
  - What learners do with the data (i.e., how they infer hypotheses)
  - How producers produce the data
- ➡ How hypotheses relate to the events in the world

# A counterintuitive thing

---

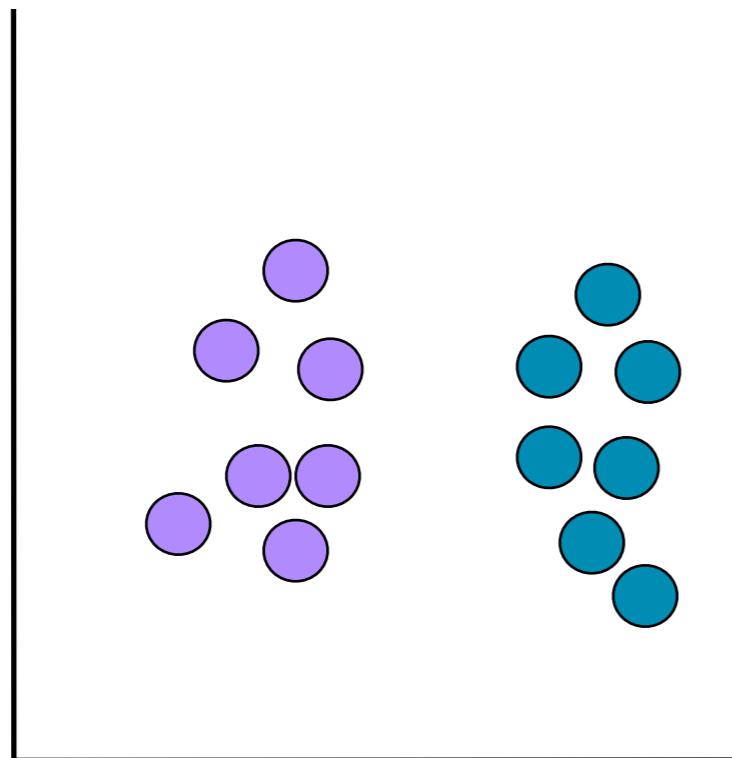
- ▶ Suppose these are the events in the world...



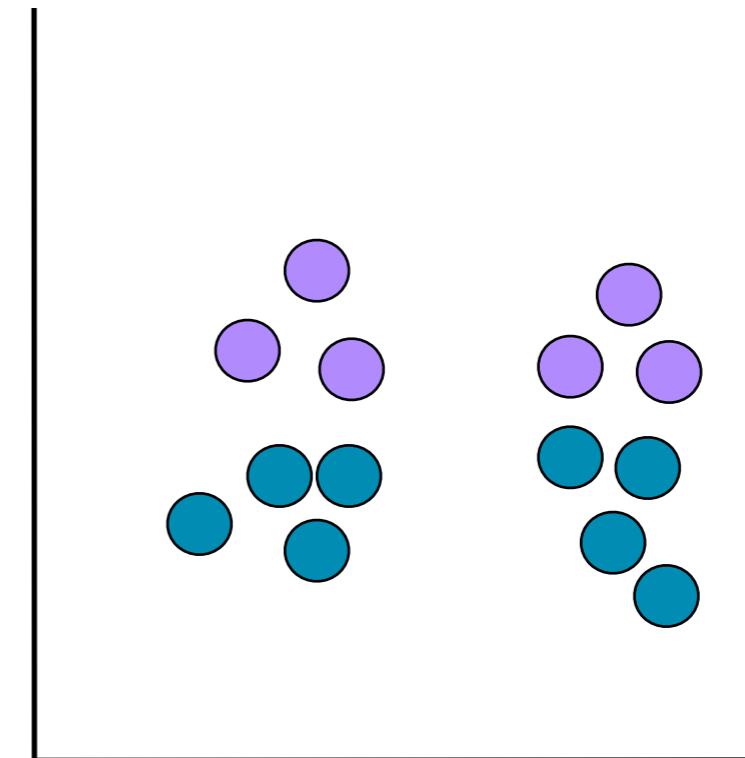
# A counterintuitive thing

---

- As we've seen many times, some ways of classifying these events seem natural and others not-so-natural



More natural

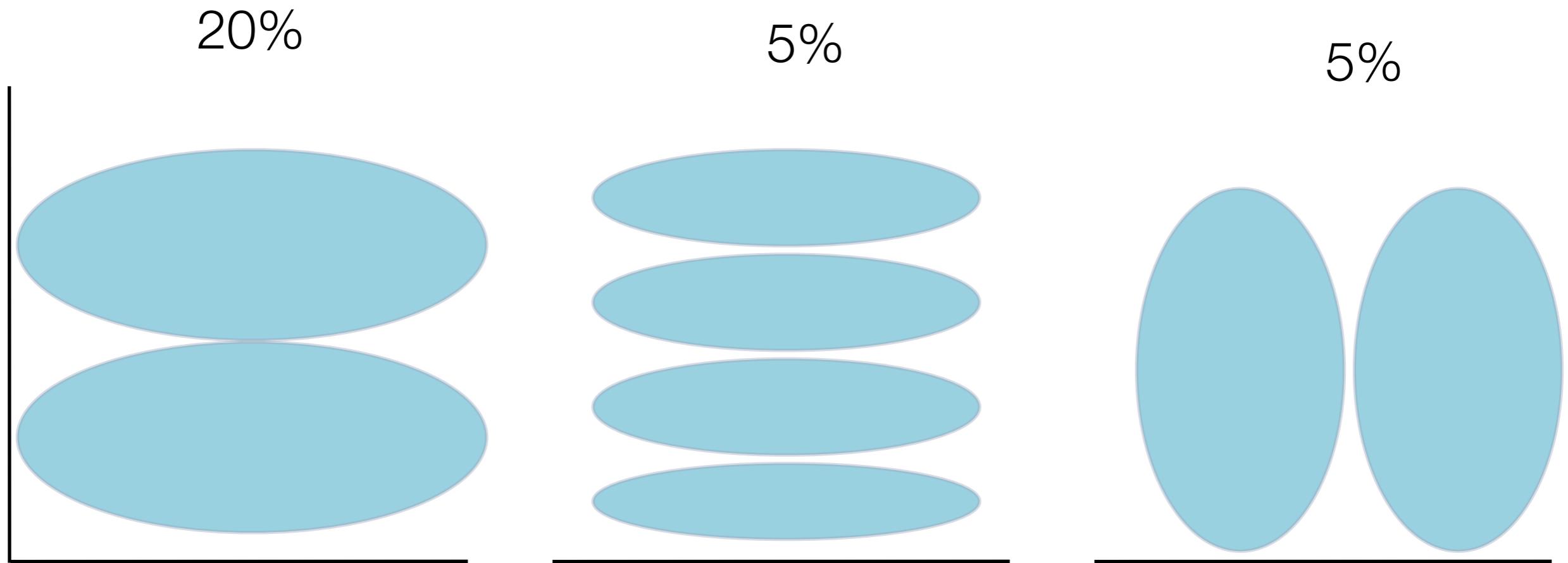


Less natural

# A counterintuitive thing

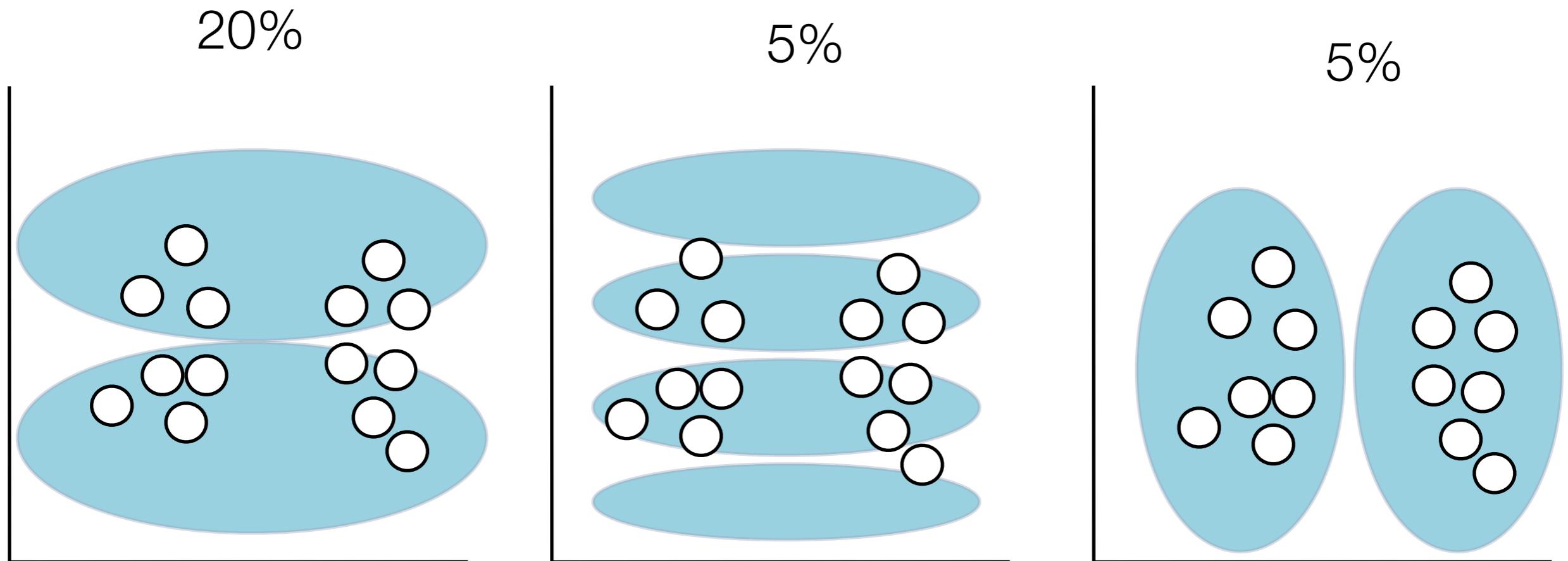
---

- ▶ Yet convergence to the prior suggests that the stationary distribution should occur independently of what the actual structure of events in the world is!



# A counterintuitive thing

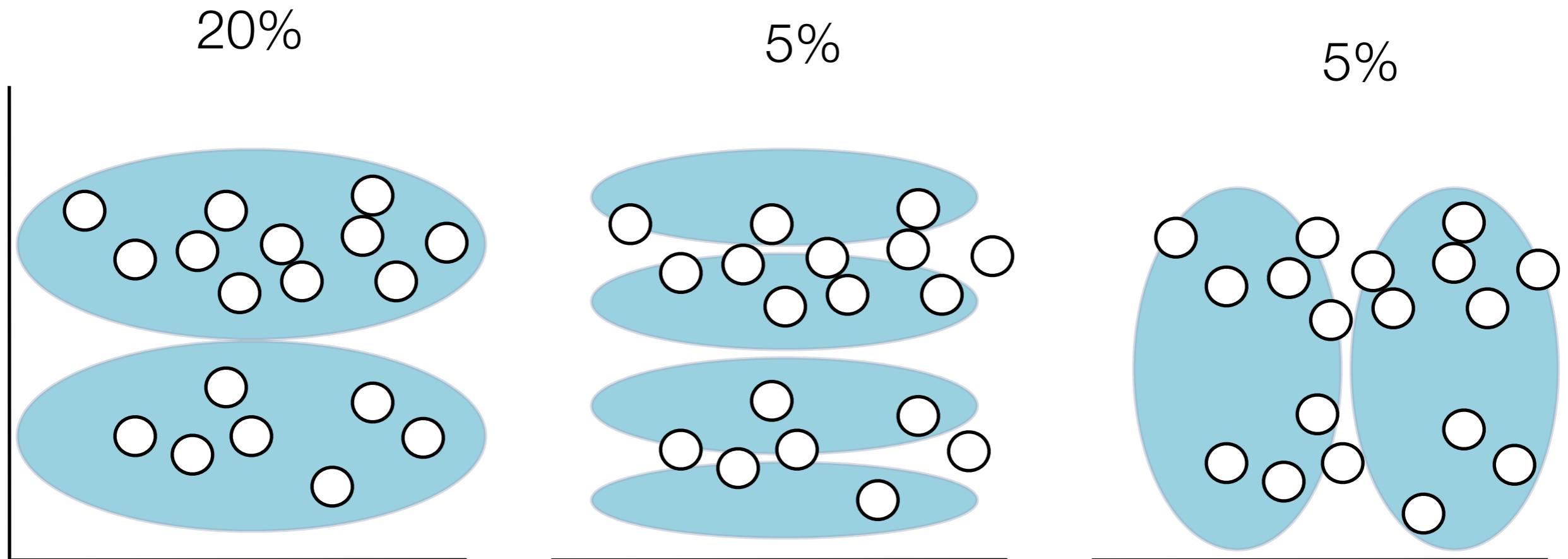
- ▶ Yet convergence to the prior suggests that the stationary distribution should occur independently of what the actual structure of events in the world is!



# A counterintuitive thing

---

- ▶ Yet convergence to the prior suggests that the stationary distribution should occur independently of what the actual structure of events in the world is!



# A counterintuitive thing

---

- ▶ Yet convergence to the prior suggests that the stationary distribution should occur independently of what the actual structure of events in the world is!

This seems wrong.

# Previous learning step...

---

- ▶ **Learning step:** learner  $n+1$  sees  $x_n$  (from previous person) and computes a posterior distribution over  $h_{n+1}$  according to Bayes' Rule

$$P(h_{n+1}|x_n, y_n) = \frac{P(y_n|x_n, h_{n+1})P(h_{n+1})}{\sum_{h \in \mathcal{H}} P(y_n|x_n, h)P(h)}$$

- ▶ This form of the equation follows from the assumption that each hypothesis  $h$  makes no assumptions about which events are more likely
- ▶ But, that might be silly in many situations. We might instead presume that hypotheses carry with them some assumptions about the event structure in the world:

$$P(h_{n+1}|x_n, y_n) = \frac{P(y_n|x_n, h_{n+1})P(h_{n+1}|x_n)}{\sum_{h \in \mathcal{H}} P(y_n|x_n, h)P(h|x_n)}$$

# Rederiving with the new equation...

- The resulting Markov Chain has a stationary distribution that now converges to  $\pi(h) = \sum_x P(h|x)Q(x)$

$$\begin{aligned}\pi(h_{n+1}) &= \sum_{h_n} P(h_{n+1}|h_n)\pi(h_n) \\ &= \sum_x \sum_y \sum_{h_n} P(h_{n+1}|x,y)P(y|x,h_n)Q(x)\pi(h_n) \\ &= \sum_x \sum_y \sum_{h_n} P(h_{n+1}|x,y)P(y|x,h_n)Q(x) \sum_{x'} P(h_n|x')Q(x') \\ &\approx \sum_x \sum_y \sum_{h_n} P(h_{n+1}|x,y)P(y|x,h_n)Q(x)P(h_n|x) \\ &= \sum_x \sum_y P(h_{n+1}|x,y)Q(x) \sum_{h_n} P(y|x,h_n)P(h_n|x) \\ &= \sum_x \sum_y P(h_{n+1}|x,y)Q(x)P(y|x) \\ &= \sum_x Q(x) \sum_y P(h_{n+1}|x,y)P(y|x) \\ &= \sum_x Q(x)P(h_{n+1}|x) \\ &= \pi(h_{n+1})\end{aligned}$$

Must assume that  
 $P(h|x) \approx E_{Q(x')}[P(h|x')] = \sum_{x'} P(h|x')Q(x')$ .  
which basically comes down to assuming  
that  $x$  is a representative draw from  $Q(x)$

# What does this mean?

---

- ▶ Cultural evolution should converge on the languages / concepts that are likely given the structure of the world and the posterior distribution over those languages / concepts
  - Incorporates the structure of the world
  - Prior probability of concepts / languages
  - Also their likelihood given the data produced
- ▶ Previous experiments didn't really manipulate event structure ( $x$ ).. only the  $x,y$  pairing (as in function learning).
- ▶ A prediction here is that manipulating event structure only should still strongly affect what is learned!

# Test: learning a language

---

- ▶ Event structure, as before, is a set of meanings, but this time simpler: 36 possible, varying on two dimensions

■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■

Control

■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■

Size

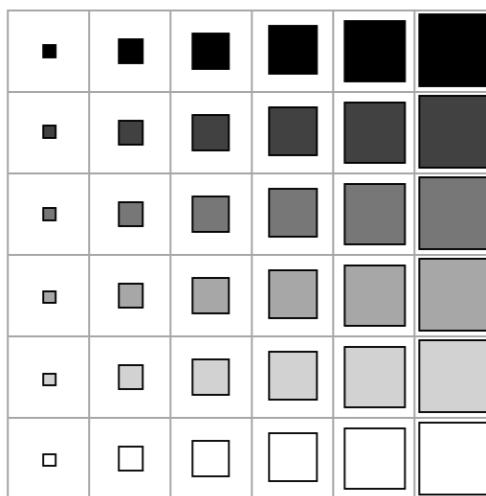
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■
■	■	■	■	■	■	■

Colour

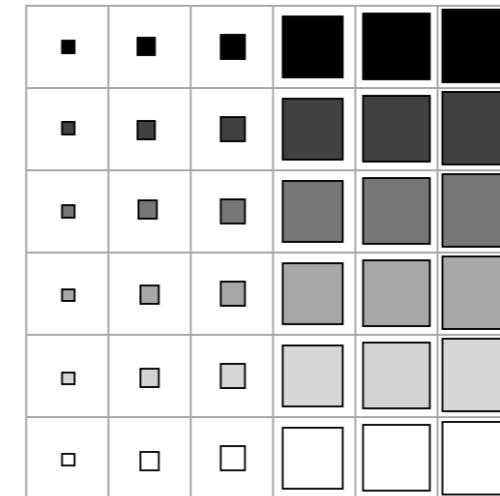
- ▶ Events are paired with a label (for first person, it's random)
- ▶ After training on these, the person is shown events and has to generate the label themselves
- ▶ The next person is given the previous person's labels to use as their training data

# Test: learning a language

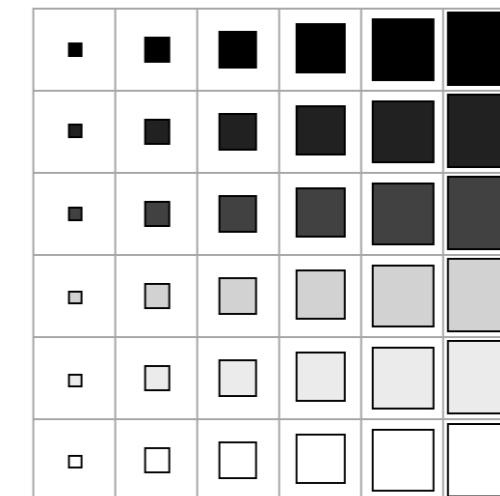
- ▶ Event structure, as before, is a set of meanings, but this time simpler: 36 possible, varying on two dimensions



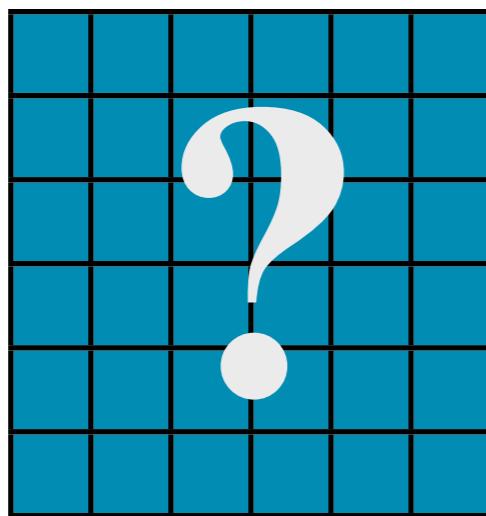
## Control



# Size



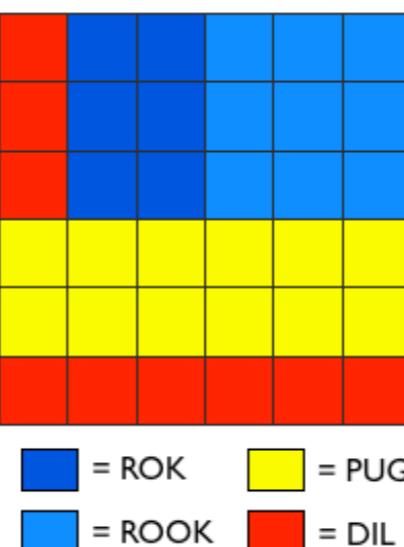
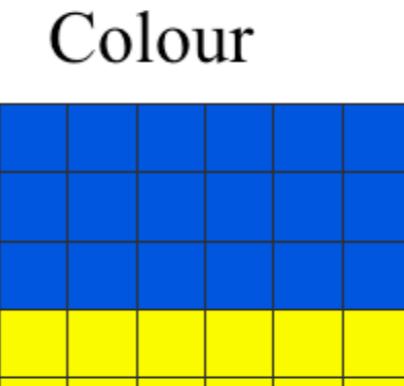
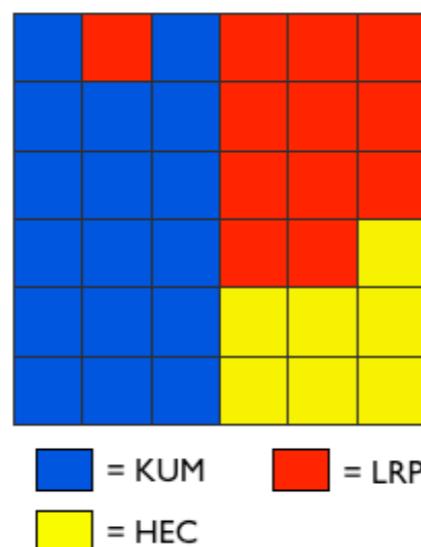
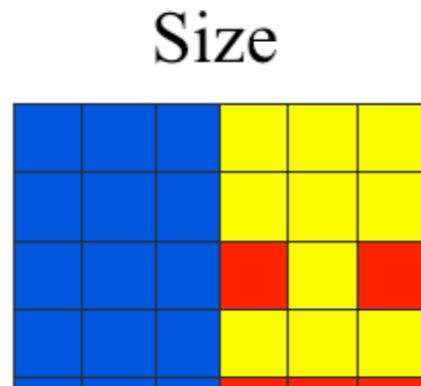
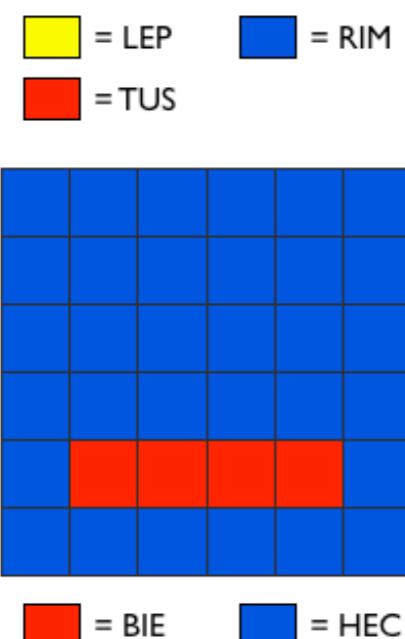
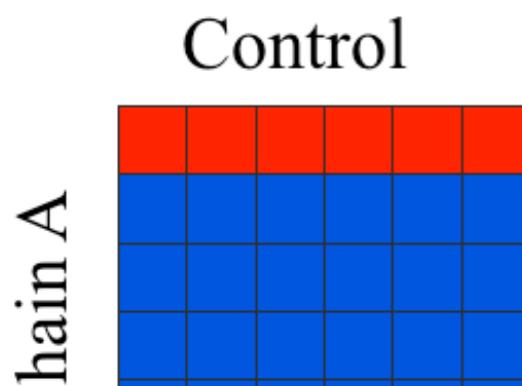
# Colour



W <sub>1</sub>	W <sub>1</sub>	W <sub>1</sub>	W <sub>2</sub>	W <sub>2</sub>	W <sub>2</sub>
W <sub>1</sub>	W <sub>1</sub>	W <sub>1</sub>	W <sub>2</sub>	W <sub>2</sub>	W <sub>2</sub>
W <sub>1</sub>	W <sub>1</sub>	W <sub>1</sub>	W <sub>2</sub>	W <sub>2</sub>	W <sub>2</sub>
W <sub>1</sub>	W <sub>1</sub>	W <sub>1</sub>	W <sub>2</sub>	W <sub>2</sub>	W <sub>2</sub>
W <sub>1</sub>	W <sub>1</sub>	W <sub>1</sub>	W <sub>2</sub>	W <sub>2</sub>	W <sub>2</sub>

# Test: learning a language

- First, we can look at the final languages in each chain:



Quantify with the adjusted Rand Index

(compared to canonical size and canonical colour;  
1=identical; 0=random)

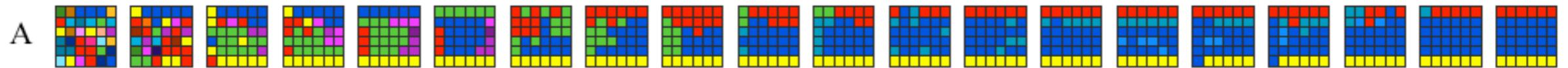
	Canonical Size	Canonical Colour
Control	-0.0204	0.0618
Size	0.704	0.079
Colour	0.065	0.696

# Test: learning a language

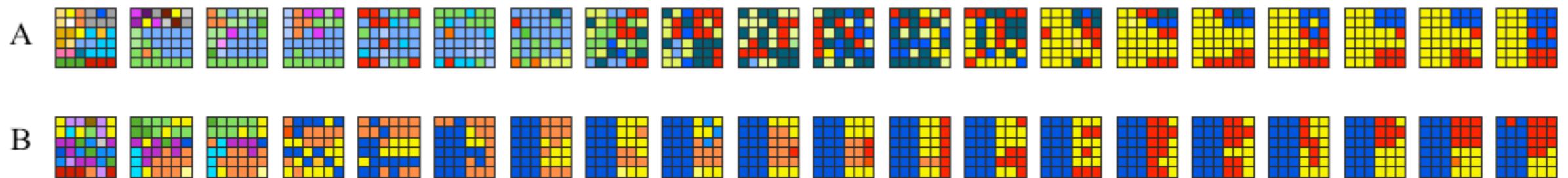
---

- ▶ The full chains tell the same story - gradual evolution to match the structure of the world

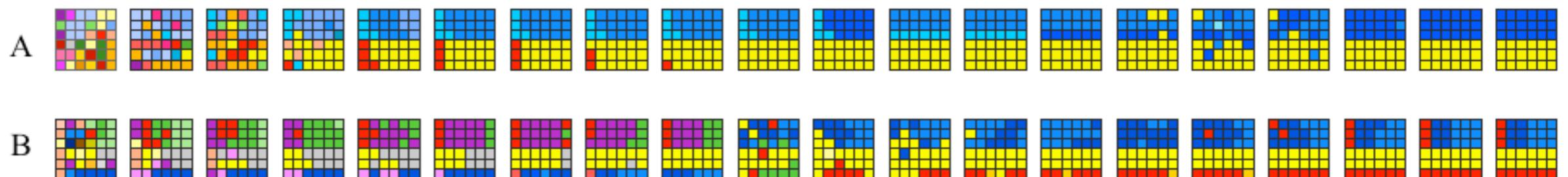
**Control**



**Size**



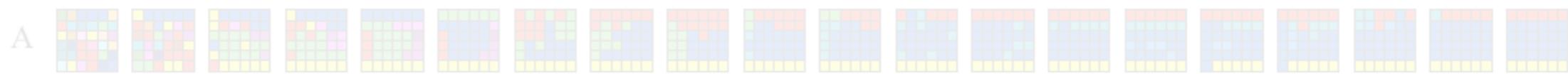
**Colour**



# Test: learning a language

- ▶ The full chains tell the same story - gradual evolution to match the structure of the world

Control



Size

A

B

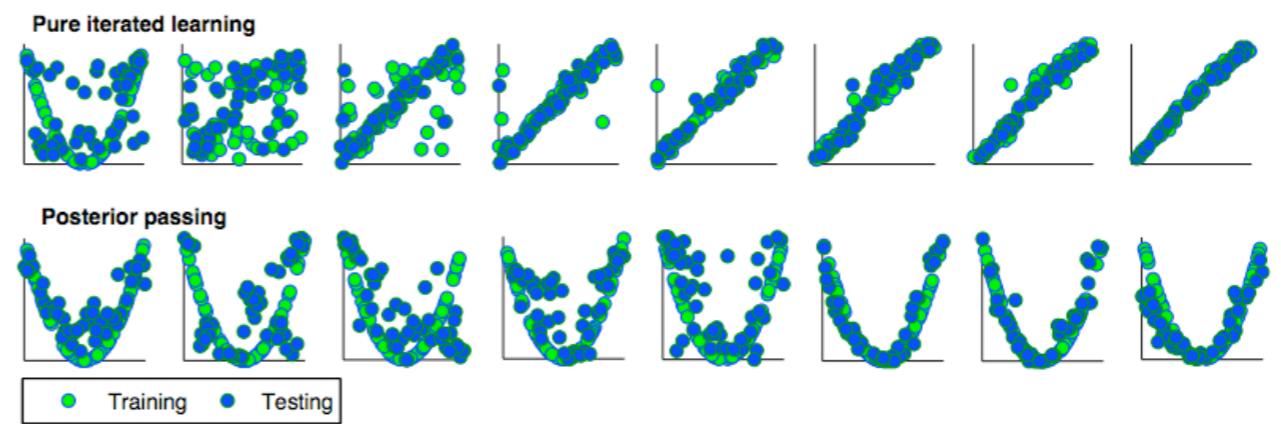
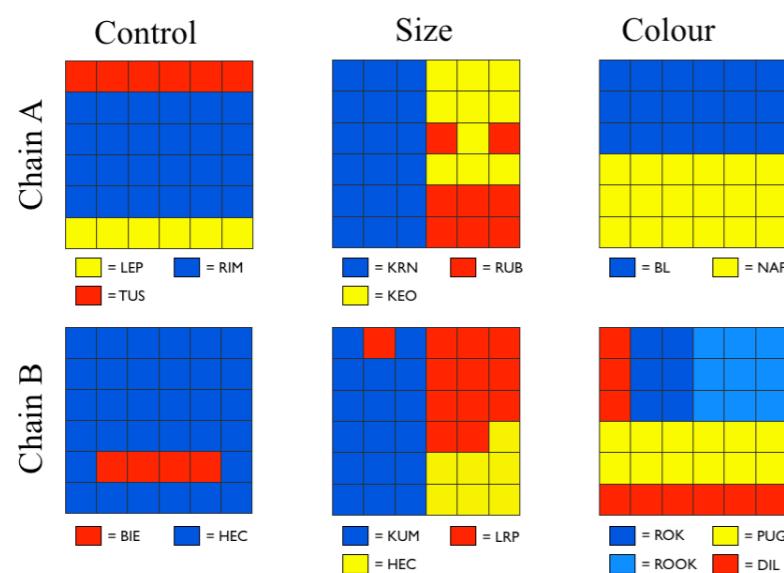
Note too that it's not *just* the structure of the world that matters. There is clearly a role for prior biases as well, in particular a bias for learnability -- otherwise you would have one word for each meaning

Colour

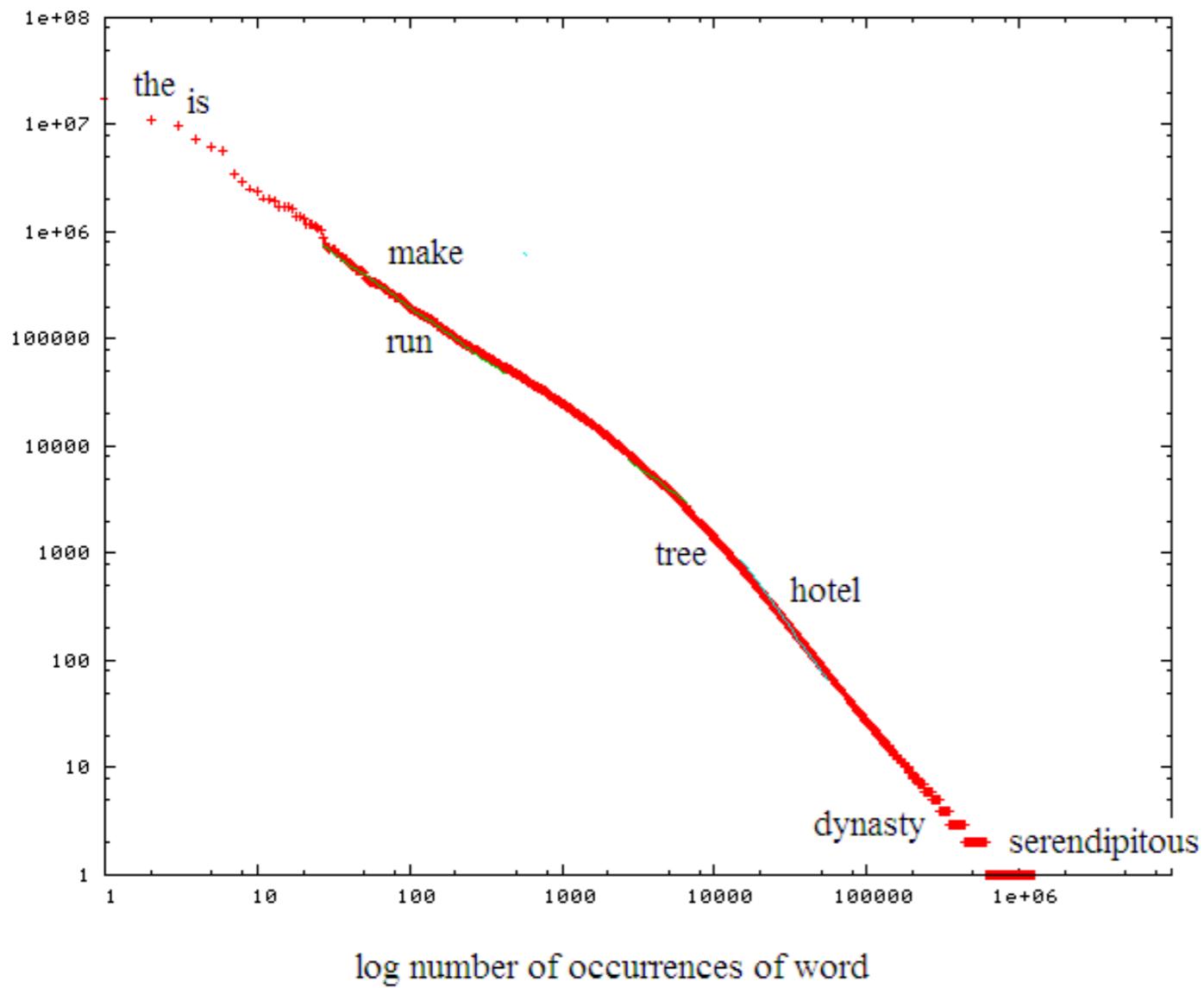


# Summary

- ▶ We can model conceptual change / evolution as chains of learners who pass each other information, and are individually Bayesian in how they learn from the previous one
- ▶ The main prediction, that the stationary distribution of the chain reflects (only) prior probability, was borne out experimentally
- ▶ Changing the assumptions changes the results in interesting ways



# Computational Cognitive Science



Lecture 15: Sequential learning  
with n-grams

Bigram frequencies

$P(w_2|w_1)$

$w_1$	The	Man	Ate	Old	Fruit	Who	Was
The	0	0	1.0	0	0	1.0	0
Man	0.33	0	0	1.0	0	0	0
Ate	0	0	0	0	0	1.0	0
Old	0.33	0	0	0	0	0	0
Fruit	0.33	0	0	0	0	0	0
Who	0	0.5	0	0	0	0	0
Was	0	0.5	0	0	0	0	0

$X_t = \text{colour at time } t$   
 $S = \{\text{Red}, \text{Blue}\}$

$T$

$\text{Red} \quad \text{Blue}$

$\text{Red} \quad p(R|R) \quad p(R|B)$

$\text{Blue} \quad p(B|R) \quad P(B|B)$

$X_t = \text{word at time } t$   
 $S = \{\text{the, old, man, was, ...}\}$

$t-2 \quad t-1 \quad t \quad t+1$

$\text{the} \rightarrow \text{old} \rightarrow \text{man} \rightarrow \text{was}$

why is Australia so

why is Australia so - Google Search

why is australia so expensive

why is australia so hot

why is australia so great

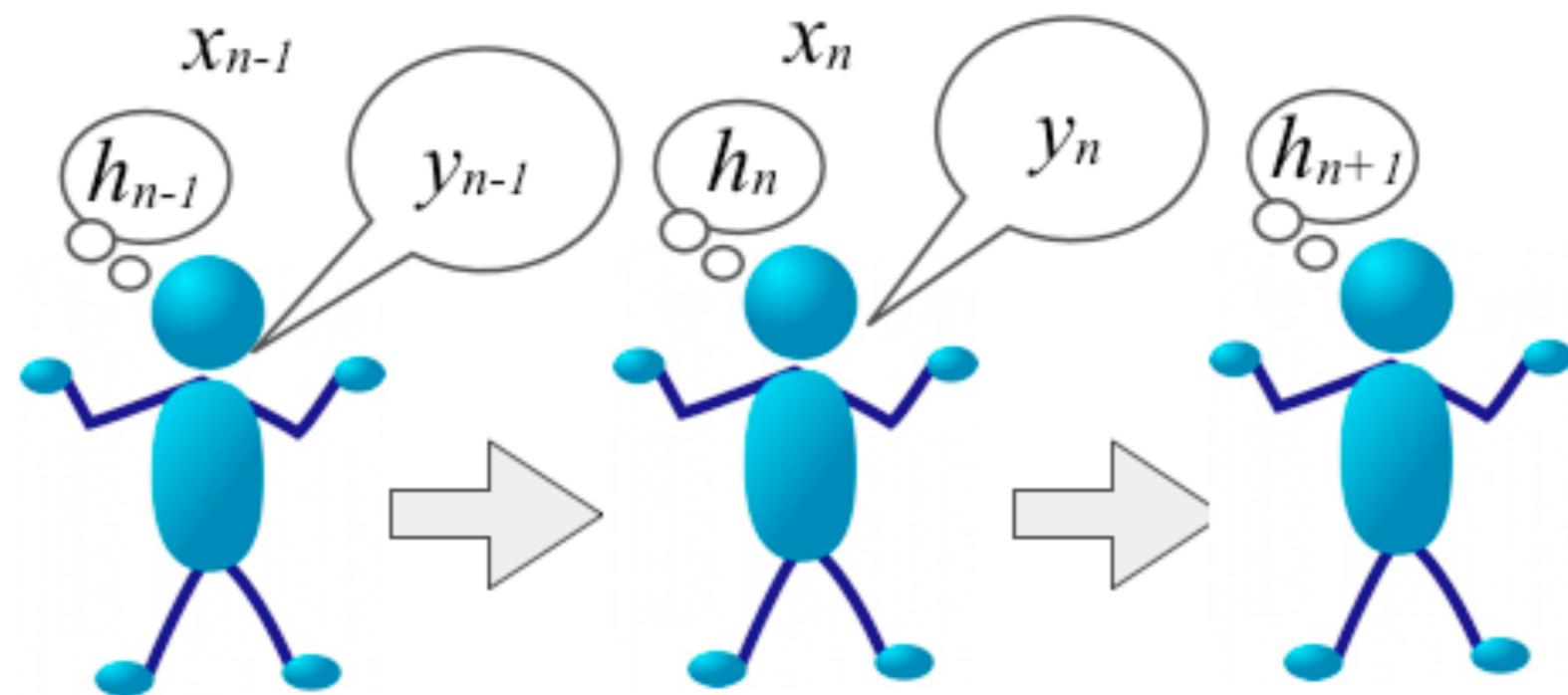
why is australia so dry

why is australia so boring

# Why should we care about sequence learning?

---

- ▶ So far we've been looking at understanding how concepts can change over a long time span, based on the sequence of generations or states that change goes through



# Why should we care about sequence learning?

---

- ▶ So far we've been looking at understanding how concepts can change over a long time span, based on the sequence of generations or states that change goes through
- ▶ But much in day-to-day life we are faced with sequences that occur on a much smaller time scale, and we need to learn about them



what is that melody?

# Why should we care about sequence learning?

---

- ▶ So far we've been looking at understanding how concepts can change over a long time span, based on the sequence of generations or states that change goes through
- ▶ But much in day-to-day life we are faced with sequences that occur on a much smaller time scale, and we need to learn about them

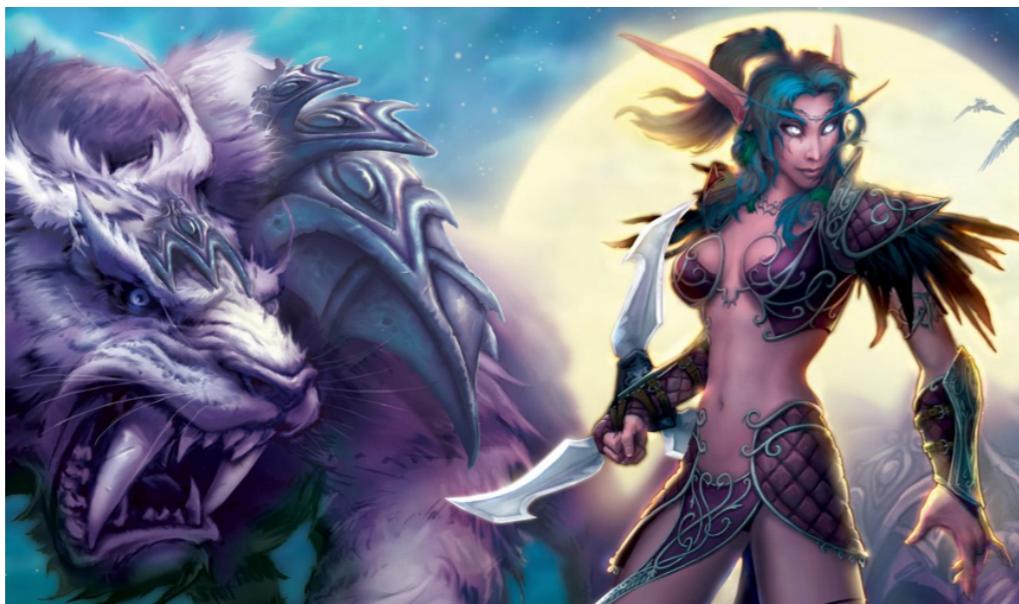


what is she going to  
say next?

# Why should we care about sequence learning?

---

- ▶ So far we've been looking at understanding how concepts can change over a long time span, based on the sequence of generations or states that change goes through
- ▶ But much in day-to-day life we are faced with sequences that occur on a much smaller time scale, and we need to learn about them



where is the monster  
going to go next?

# Why should we care about sequence learning?

---

- ▶ So far we've been looking at understanding how concepts can change over a long time span, based on the sequence of generations or states that change goes through
- ▶ But much in day-to-day life we are faced with sequences that occur on a much smaller time scale, and we need to learn about them



is he going to make  
his next goal?

# Why should we care about sequence learning?

---

- ▶ So far we've been looking at understanding how concepts can change over a long time span, based on the sequence of generations or states that change goes through
- ▶ But much in day-to-day life we are faced with sequences that occur on a much smaller time scale, and we need to learn about them



what is the plan for  
the day?

# Why should we care about sequence learning?

---



All of these involve learning what kinds of sequences of actions there are, and which actions tend to follow which others

We can therefore apply similar models to learning all of them, although the specific questions we care about in each will differ

# Plan for the rest of the lectures

---

- ▶ Today: a simple model for sequence learning ( $n$ -grams)
  - Description of the approach
  - Application to natural language processing
  - The problem of overfitting
- ▶ Tomorrow 1: applications of  $n$ -gram models
  - A solution to the problem of overfitting
  - Word segmentation
  - Nonadjacent learning
  - What about more complex structure?
- ▶ Tomorrow 2: extending  $n$ -grams (HMMs)
  - Computing likelihood of observations
  - Inferring the hidden state sequence
  - Finding the best HMM (if time)

# Plan for the rest of the lectures

---

→ Today: a simple model for sequence learning ( $n$ -grams)

- Description of the approach
- Application to natural language processing
- The problem of overfitting

► Tomorrow 1: applications of  $n$ -gram models

- A solution to the problem of overfitting
- Word segmentation
- Nonadjacent learning
- What about more complex structure?

► Tomorrow 2: extending  $n$ -grams (HMMs)

- Computing likelihood of observations
- Inferring the hidden state sequence
- Finding the best HMM (if time)

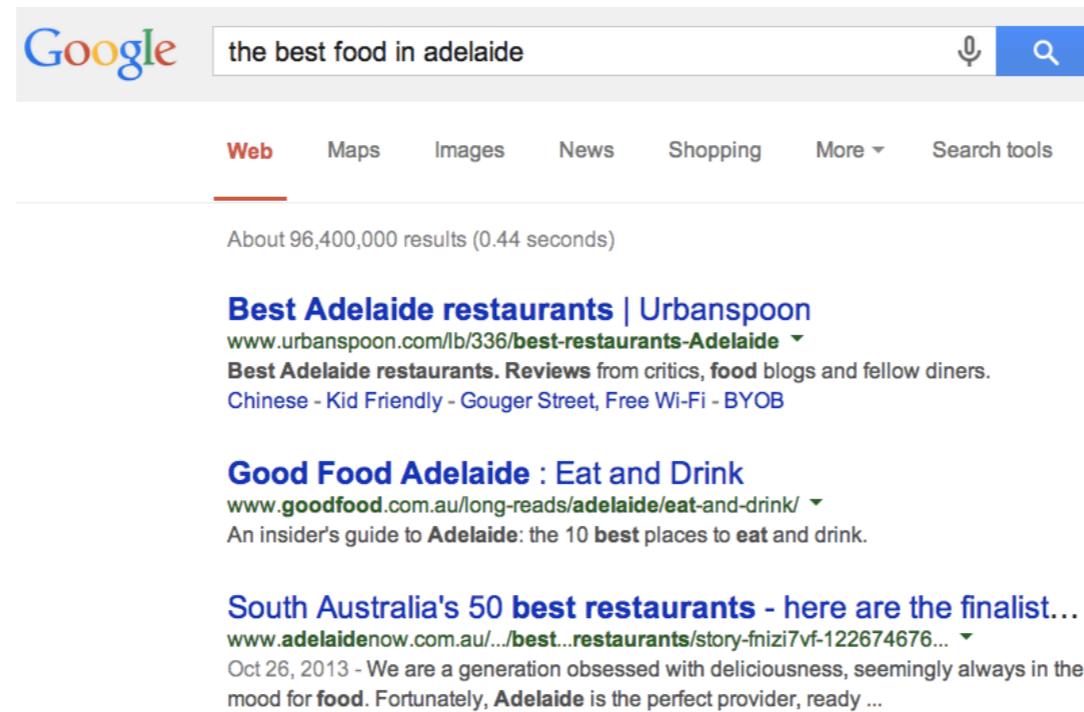
# Natural language processing

---

- ▶ A very important kind of sequential knowledge: language

the itsy bitsy spider went up the water spout...

- ▶ The techniques we will be talking about are a simplified version of techniques used by many companies and researchers who want to use and manipulate text

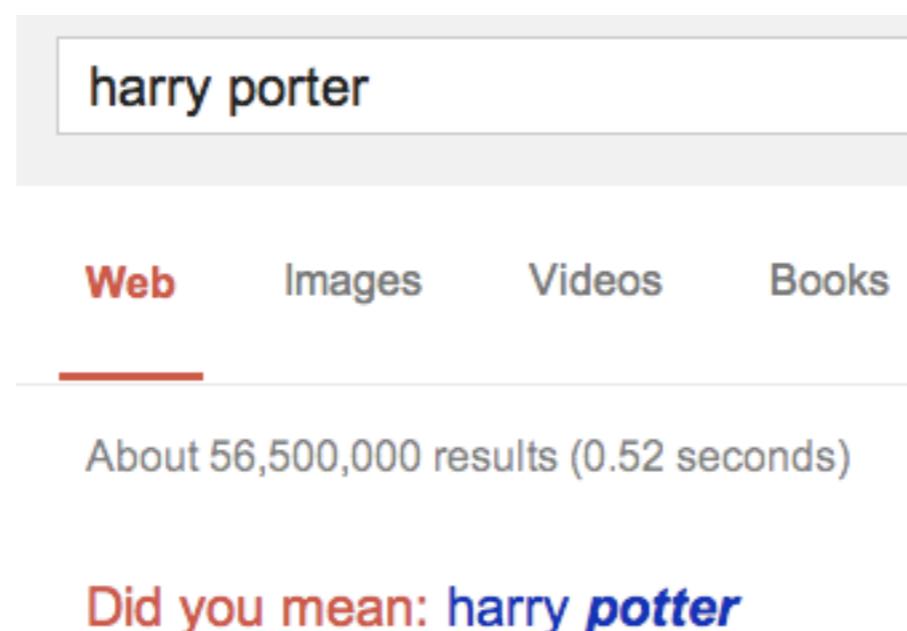


A screenshot of a Google search results page. The search query "the best food in adelaide" is entered in the search bar. Below the search bar, there are tabs for "Web", "Maps", "Images", "News", "Shopping", "More", and "Search tools". The "Web" tab is selected. Below the tabs, it says "About 96,400,000 results (0.44 seconds)". The first result is a link to "Best Adelaide restaurants | Urbanspoon" with the URL "www.urbanspoon.com/lb/336/best-restaurants-Adelaide". The description below the link says "Best Adelaide restaurants. Reviews from critics, food blogs and fellow diners. Chinese - Kid Friendly - Gouger Street, Free Wi-Fi - BYOB". The second result is a link to "Good Food Adelaide : Eat and Drink" with the URL "www.goodfood.com.au/long-reads/adelaide/eat-and-drink/". The description says "An insider's guide to Adelaide: the 10 best places to eat and drink.". The third result is a link to "South Australia's 50 best restaurants - here are the finalist..." with the URL "www.adelaidenow.com.au/.../best...restaurants/story-fnizi7vf-122674676...". The description says "Oct 26, 2013 - We are a generation obsessed with deliciousness, seemingly always in the mood for food. Fortunately, Adelaide is the perfect provider, ready ...".

# Natural language processing

---

- ▶ A very important kind of sequential knowledge: language
  - the itsy bitsy spider went up the water spout...
- ▶ The techniques we will be talking about are a simplified version of techniques used by many companies and researchers who want to use and manipulate text



# Natural language processing

---

- ▶ A very important kind of sequential knowledge: language

the itsy bitsy spider went up the water spout...

- ▶ The techniques we will be talking about are a simplified version of techniques used by many companies and researchers who want to use and manipulate text

<input type="checkbox"/>		Ashley Ogletree	Boner Extension Is Possible And Safe lea
<input type="checkbox"/>		「黄金のKABU手法」 小野寺	あなたが「株」で損をしてる原因を教えまし
<input type="checkbox"/>		小野寺典夫 mag2 0000258547	あなたが「株」で損をしてる原因を教えまし
<input type="checkbox"/>		小野寺典夫 mag2 0000258547	「黄金のKABU手法」（公式）：「1万時間
<input type="checkbox"/>		「黄金のKABU手法」 小野寺	「黄金のKABU手法」（公式）：「1万時間
<input type="checkbox"/>		Barack Obama	Hey, Amy - The top Republican in the House
<input type="checkbox"/>		International Work-Confe.	Time Series Conference - ITISE 2014, Gran

# Natural language processing

- ▶ A very important kind of sequential knowledge: language

the itsy bitsy spider went up the water spout...

- ▶ The techniques we will be talking about are a simplified version of techniques used by many companies and researchers who want to use and manipulate text

Translate

Korean English Chinese Detect language ▾

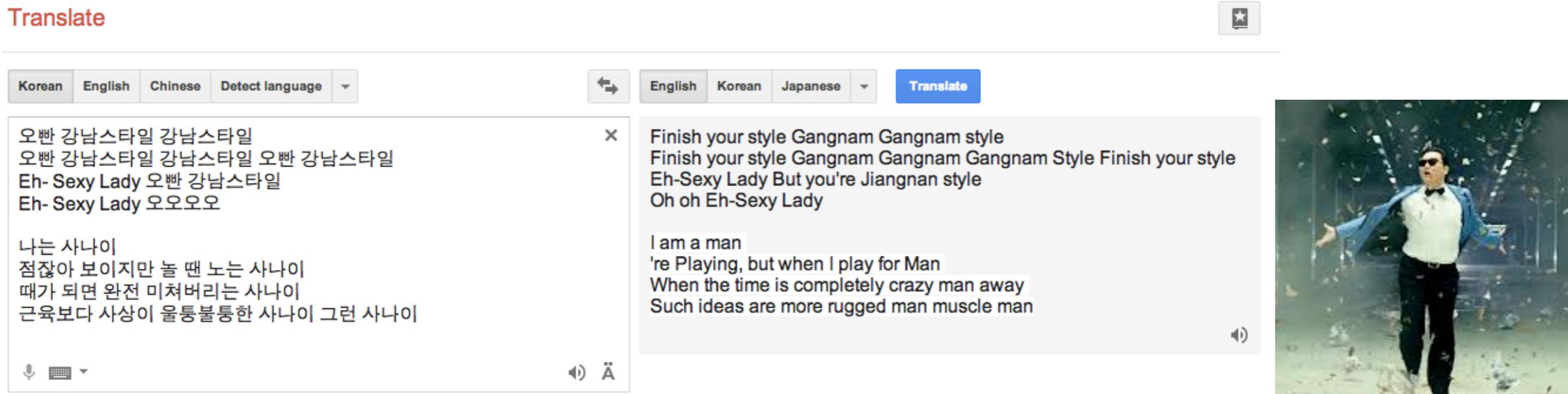
English Korean Japanese ▾ Translate

오빤 강남스타일 강남스타일  
오빤 강남스타일 강남스타일 오빤 강남스타일  
Eh- Sexy Lady 오빤 강남스타일  
Eh- Sexy Lady 오오오오

나는 사나이  
점잖아 보이지만 놀 땐 노는 사나이  
때가 되면 완전 미쳐버리는 사나이  
근육보다 사상이 울퉁불퉁한 사나이 그런 사나이

Finish your style Gangnam Gangnam style  
Finish your style Gangnam Gangnam Style Finish your style  
Eh-Sexy Lady But you're Jiangnan style  
Oh oh Eh-Sexy Lady

I am a man  
're Playing, but when I play for Man  
When the time is completely crazy man away  
Such ideas are more rugged man muscle man



# Natural language processing

---

- ▶ A very important kind of sequential knowledge: language

the itsy bitsy spider went up the water spout...

- ▶ The techniques we will be talking about are a simplified version of techniques used by many companies and researchers who want to use and manipulate text

A screenshot of a search interface. The search bar at the top contains the query "why is Australia so". Below the search bar, a list of suggestions and results is displayed. The first result is highlighted with a blue background and shows the full query "why is Australia so - Google Search". Below this, seven other suggestions are listed, each starting with a magnifying glass icon and followed by a truncated version of the query: "why is australia so expensive", "why is australia so hot", "why is australia so great", "why is australia so dry", and "why is australia so boring".

A screenshot of a search interface. The search bar at the top contains the query "why is America so". Below the search bar, a list of suggestions and results is displayed. The first result is highlighted with a blue background and shows the full query "why is America so - Google Search". Below this, six other suggestions are listed, each starting with a magnifying glass icon and followed by a truncated version of the query: "why is america so stupid", "why is america so religious", "why is america so violent", "why is america so rich", and "why is america so cheap".

# Natural language processing

---

- ▶ A very important kind of sequential knowledge: language
  - the itsy bitsy spider went up the water spout...
- ▶ The techniques we will be talking about are a simplified version of techniques used by many companies and researchers who want to use and manipulate text
- ▶ Cognitive scientists are interested in the same tools, because people may rely on something like them to solve similar problems

# Corpora

---

- ▶ Data = a corpus (plural: corpora), which is a (set of) text files used as input to your algorithm
- ▶ In cognitive science, where we're interested in the data people actually hear, it is often transcribed speech (often to children)

```
yuwanttusid6bUk  
&nd6d0gi  
yuwanttulUk&tDIS  
lUk&tDIS  
h&v6drINK  
tekItQt
```

```
you want to see the book?  
and a doggie!  
you want to look at this?  
look at this!  
have a drink  
take it out
```

Corpus of child-directed speech transcribed into an  
ASCII version of phonetic notation

# Corpora

---

- ▶ Data = a corpus (plural: corpora), which is a (set of) text files used as input to your algorithm
- ▶ In cognitive science, where we're interested in the data people actually hear, it is often transcribed speech (often to children)
- ▶ In NLP, it is more often just gigabytes of text documents from the web, samples from articles / books / etc

**Scorching heat to return to southern Australia, total fire bans in place**

Updated Tue 28 Jan 2014, 11:48am AEDT

**Residents across much of southern Australia are bracing for another heatwave, with temperatures forecast to reach into the 40s in some areas today.**

Total fire bans have been issued across South Australia, Victoria and Tasmania ahead of the extreme heat.

Adelaide's maximum temperature today is expected to be 41 degrees Celsius, with 40C on Friday, 41C on Saturday and 40C on Sunday.

A catastrophic fire danger rating has been issued for the state's lower south-east.

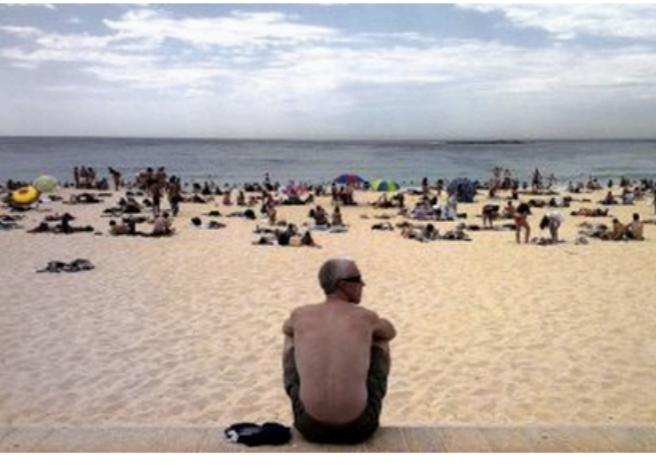


PHOTO: Temperatures forecast to reach into the 40s in some areas today. (ABC News: Amy Simmons)

# Corpora

---

- ▶ There are a lot of hidden complexities involved in getting corpora to the point that we can get useful information about the language from them -- we'll skip over a lot of these

## Periods

end of sentence?  
Mr., S.A., \$12.30

## Apostrophes

it's, the boy's  
one word or two?

## Morphology

runs, running, run  
1 word or 3?

## Homographs

saw vs saw

(in spoken corpora, homophones)

# Basic problems in language modelling

---

- ▶ What sequences of words are allowable or frequent?

hey! how are you?

pickled crysanthemums

- ▶ Given a sequence, what word(s) would you expect to come next?

I don't want to date you, I'd rather just be \_\_\_\_\_

- ▶ Models called **n-gram models** are among the simplest ways to answer both of these problems (as well as in many non-language contexts as well, as we'll see tomorrow)

# N-grams: tracking clusters of words

---

I'd rather just be \_\_\_\_\_

# N-grams: tracking clusters of words

W<sub>n-3</sub> W<sub>n-2</sub> W<sub>n-1</sub> W<sub>n</sub> \_\_\_\_\_

- Essentially, we want to estimate the following probability function

$$p(w_n | w_1, \dots, w_{n-1})$$

1-gram (unigram)	2-gram (bigram)	3-gram (trigram)	...
$p(w_n)$	$p(w_n   w_{n-1})$	$p(w_n   w_{n-1}, w_{n-2})$	

# N-grams: tracking clusters of words

W<sub>n-3</sub> W<sub>n-2</sub> W<sub>n-1</sub> W<sub>n</sub> \_\_\_\_\_

- Essentially, we want to estimate the following probability function

$$p(w_n | w_1, \dots, w_{n-1})$$

I'd rather just be *friends*

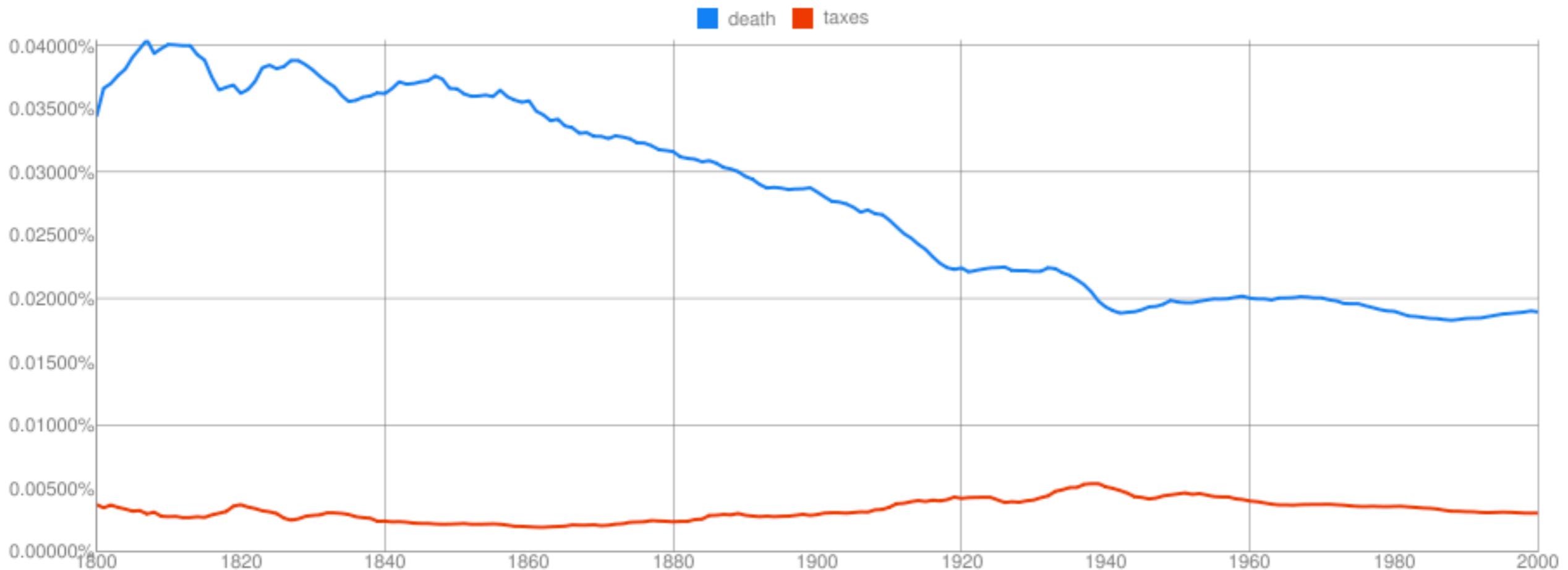
1-gram (unigram)	2-gram (bigram)	3-gram (trigram)	...
$P(\text{friends})$	$P(\text{friends}   \text{be})$	$P(\text{friends}   \text{just be})$	

# N-grams are fun!

## Google labs Books Ngram Viewer

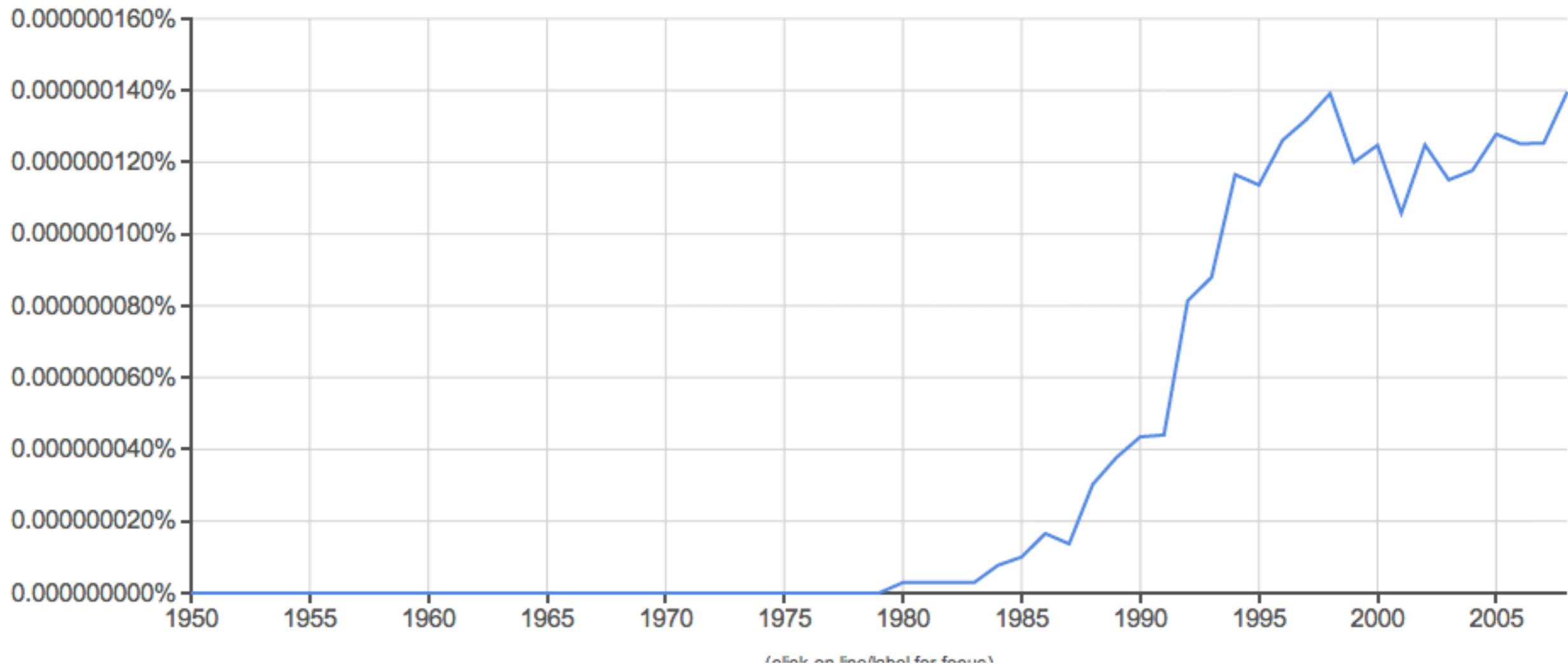
Graph these **case-sensitive** comma-separated phrases:

between  and  from the corpus  with smoothing of .



# N-grams are fun!

Graph these comma-separated phrases: computational cognitive science  case-insensitive  
between 1950 and 2008 from the corpus English with smoothing of 3 . **Search lots of books**



# N-grams are fun!

Graph these comma-separated phrases:   case-insensitive  
between  and  from the corpus  with smoothing of  .



# N-grams are fun!

---

- ▶ You can also use these models to generate sentences rather than just predict them

## Bigram

- ✓ Don't you get on the front of the bathroom?
- ✓ Ernie hesitated to Father Feeney, for me he said, “Come sit”, he should bear the soul.
- ✓ And Bill Doolin and known magic might be as training?
- ✓ The previous night, means by the initial cost increases, the official representative government in address for manned bombers to gallop was Roman nose on a million of a couple of the book of purpose we fail to a cowhand was heard.

# N-grams are fun!

---

- ▶ You can also use these models to generate sentences rather than just predict them

## Trigram

- ✓ It wasn't just Japan, either, because the rain came more heavily, and men in an involuntorial fashion.
- ✓ These two were blacked out, as it did when Eisenhower appointed him Ambassador to Moscow.
- ✓ On chemical grounds it should be remitted with your naked eye at night in South Africa, beats it by taking the raw sewage into the Austrian branch of Multnomah Bank, also was ordered by a public servant but not much to remove restrictions that, it requires only written notice that he was glad the fat that extended beyond himself, something which, consequently, he was counting on the cheek.

# N-grams are fun!

---

- ▶ You can also use these models to generate sentences rather than just predict them

## 4-gram

- ✓ A Yankee sergeant gave the following description of his sweetheart: "My girl is none of your business", she said, then turned and attacked Morgan who became greatly outnumbered and had to fasten it on someone.
- ✓ When words can be used in almost any classroom today can be found amongst us without a correct version of the welfare state in England still allows wild scope for all kinds of problems for which she was moving, her method for keeping an escort from departing too early was unique.
- ✓ But, again, we have no clocks and nobody cares enough to count days or to make calendars and there's not much climate here, so none of us knew anything but filth and poverty.

# N-grams are fun!

---

- ▶ You can also use these models to generate sentences rather than just predict them

## 5-gram

- ✓ Nevertheless, it remained one of the great singers of our time but she is one of the best-gaited pacers on the grounds.
- ✓ These are few and seemingly disjointed data, but they illustrate the important fact that fundamental alterations in conditioned reactions occur in a variety of expressions in your country denied not only the existence of this conflict but it was elaborated even further with an incredible semantic dexterity.
- ✓ The queen afterward keeps incubating and guarding her eggs like a mother hen, taking a sip from time to time in ratable proportions, on account of the crazy tourists.

# N-grams: tracking clusters of words

---

- ▶ For both generation and prediction, higher  $n$  is better!
- ▶ Both are extremely straightforward given the n-gram probabilities

## Two kinds of probabilities

1. Probability of a word or series of words

$$p(w_1, \dots, w_n)$$

2. Probability of a word given a previous word or series of words

$$p(w_n | w_1, \dots, w_{n-1})$$

The equations are distinct (except in the unigram case)

# Summary

---

- ▶ Sequence learning is an important problem in cognition, and language is a clear example of when this is relevant



# Summary

---

- ▶ Sequence learning is an important problem in cognition, and language is a clear example of when this is relevant
- ▶  $n$ -gram models, which calculate the probability of an item given the previous  $n-1$  items, are widely used in natural language processing to address this problem.

The screenshot shows a search interface with a query bar containing "why is Australia so". Below the query bar, a list of search suggestions and results is displayed. The first result is highlighted in blue: "why is Australia so - Google Search". Below this, there are seven other suggestions, each starting with a magnifying glass icon and followed by a bolded search term:  
- why is australia so expensive  
- why is australia so hot  
- why is australia so great  
- why is australia so dry  
- why is australia so boring

The screenshot shows a search interface with a query bar containing "why is America so". Below the query bar, a list of search suggestions and results is displayed. The first result is highlighted in blue: "why is America so - Google Search". Below this, there are seven other suggestions, each starting with a magnifying glass icon and followed by a bolded search term:  
- why is america so stupid  
- why is america so religious  
- why is america so violent  
- why is america so rich  
- why is america so cheap

# Additional references (not required)

---

## N-gram models

- ▶ Manning, C., & Schütze, H. (1999). Foundations of statistical natural language processing. Chapter 5: 191-203