On the origins of data:  
A Bayesian perspective on inductive generalisation, categorisation and reasoning  

Dan Navarro
How do people acquire new knowledge?
(categorisation & reasoning)

How do we make choices in an uncertain world?
(judgment & decision making)

How should psychologists analyse our data?
(math psych & statistics)
How do people acquire new knowledge?
(categorisation & reasoning)

How do we make choices in an uncertain world?
(judgment & decision making)

How should psychologists analyse our data?
(math psych & statistics)
What criteria should be used to evaluate psychological models?

How should psychologists analyse our data?

What criteria should be used to evaluate psychological models?

What tools should we teach to undergrads?

How should psychologists analyse our data?

Navarro (2015)
What tools should we teach to undergrads?

What criteria should be used to evaluate psychological models?

What should we use as our “default” models when our data sets are messy?

How should psychologists analyse our data?

Kennedy, Navarro, Perfors & Briggs (in prep)
What tools should we teach to undergrads?

What criteria should be used to evaluate psychological models?

What should we use as our “default” models when our data sets are messy?

How should psychologists analyse our data?

Should we all be Bayesian?
How do people acquire new knowledge?

(categorisation & reasoning)

How should psychologists analyse our data?

(math psych & statistics)

How do we make choices in an uncertain world?

(judgment & decision making)
How do we make choices in an uncertain world?

How do people aggregate prior knowledge with statistical evidence?

Tauber, Navarro, Perfors & Steyvers (under revision). Psych Review
How do people evaluate the quality of evidence?

How do we make choices in an uncertain world?

How do people evaluate the quality of evidence?

How do we make choices in an uncertain world?

When do the statistics of the task matter?

How do people aggregate prior knowledge with statistical evidence?

How do we make choices in an uncertain world?

What psychological processes control how people regulate evidence accrual and choice?

How do people evaluate the quality of evidence?

When do the statistics of the task matter?

How do people aggregate prior knowledge with statistical evidence?

Navarro, Newell & Schulze (2016). Cogn Psych
How do people acquire new knowledge?
(categorisation & reasoning)

How should psychologists analyse our data?
(math psych & statistics)

How do we make choices in an uncertain world?
(judgment & decision making)
How do people acquire new knowledge?

What mental representations underpin human reasoning?

Tauber, Navarro, Perfors & Steyvers (under revision). *Psych. Review*
Navarro & Griffiths (2008). *Neural Computation*
Navarro & Lee (2004) *PB&R*
Can we use “big data” to study semantic representation?

How do people acquire new knowledge?

What mental representations underpin human reasoning?
Can we use “big data” to study semantic representation?

How do people acquire new knowledge?

What kinds of inductive biases guide human learning and reasoning?

What mental representations underpin human reasoning?
Inductive bias
Why do people generalise one way and not another?
Different biases are learned for different ontological kinds

<table>
<thead>
<tr>
<th>Class</th>
<th>Category</th>
<th>Instance</th>
<th>Shape</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
<td>Cat</td>
<td>1</td>
<td>Quadruped</td>
<td>Tabby</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Quadruped</td>
<td>White</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Quadruped</td>
<td>Grey</td>
</tr>
<tr>
<td></td>
<td>Ball</td>
<td>1</td>
<td>Sphere</td>
<td>Blue</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Sphere</td>
<td>Red</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Sphere</td>
<td>Green</td>
</tr>
<tr>
<td>Substance</td>
<td>Mud</td>
<td>1</td>
<td>Pile</td>
<td>Brown</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Puddle</td>
<td>Brown</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Spherical</td>
<td>Brown</td>
</tr>
<tr>
<td></td>
<td>Flour</td>
<td>1</td>
<td>Pile</td>
<td>White</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Cloud</td>
<td>White</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Scatter</td>
<td>White</td>
</tr>
</tbody>
</table>
Biases influence the discovery of new categories

Navarro & Kemp (under revision). Psych. Review
Biases constrain & shape learning

Limited capacity statistical learning?

Biases shape information search

When confirmation is uninformative, people prefer to search for disconfirmatory evidence.

“Confirmatory evidence” is informative when hypotheses are consistent with only a minority of possible observations.

Hendrickson, Perfors & Navarro (2016) Decision
Why study inductive bias using computational models?
In the case where \( n = 1 \) we observe that,

\[
\int_{\mathcal{R}} P(x_1, x_1 \in r_t \mid r_t = r) \, dr = \int_0^{z_l} \int_{z_u}^1 P(x_1 \mid r_t = [l, u]) \, du \, dl
\]

\[
= \int_0^{z_l} \int_{z_u}^1 (u - l)^{-1} \, du \, dl
\]

\[
= \int_0^{z_l} [\ln(u - l)]_{z_u}^1 \, dl
\]

\[
= \int_0^{z_l} \ln(1 - l) - \ln(z_u - l) \, dl
\]

\[
= [(l - 1) \ln(1 - l) - l]_0^{z_l} - [(l - z_u) \ln(z_u - l) - l]_0^{z_l}
\]

\[
= ((z_l - 1) \ln(1 - z_l) - z_l) - ((z_l - z_u) \ln(z_u - z_l) - z_l + z_u \ln z_u)
\]

\[
= (z_u - z_l) \ln(z_u - z_l) - (1 - z_l) \ln(1 - z_l) - z_u \ln z_u
\]  

(24)

Applying the same procedure as before yields the expression

\[
P(y \in r_t \mid x_1, x_1 \in r_t) = \begin{cases} 
\frac{(z_u - y) \ln(z_u - y) - (1 - y) \ln(1 - y) - z_u \ln z_u}{(z_u - z_l) \ln(z_u - z_l) - (1 - z_l) \ln(1 - z_l) - z_u \ln z_u} & \text{if } y < z_l \\
1 & \text{if } z_l \leq y \leq z_u \\
\frac{(y - z_l) \ln(y - z_l) - (1 - z_l) \ln(1 - z_l) - y \ln y}{(z_u - z_l) \ln(z_u - z_l) - (1 - z_l) \ln(1 - z_l) - z_u \ln z_u} & \text{if } z_u < y
\end{cases}
\]

(25)

In this case, however, the expression can be further simplified since \( z_l = z_u = x_1 \):

\[
P(y \in r_t \mid x_1, x_1 \in r_t) = \begin{cases} 
\frac{(1 - y) \ln(1 - y) + x_1 \ln x_1 - (x - y) \ln(x_1 - y)}{(1 - x_1) \ln(1 - x_1) + x_1 \ln x_1} & \text{if } y < x_1 \\
1 & \text{if } y = x_1 \\
\frac{(1 - x_1) \ln(1 - x_1) + y \ln y - (y - x_1) \ln(y - x_1)}{(1 - x_1) \ln(1 - x_1) + x_1 \ln x_1} & \text{if } x_1 < y
\end{cases}
\]

(26)

(Obviously, this expression could be derived directly, rather than found as a special case of the previous expression.)
In the case where \( n = 1 \) we observe that,

\[
\int_{\mathcal{R}} P(x_1, x_1 \in r_t \mid r_t = r) \ dr = \int_0^{z_l} \int_{z_u}^1 P(x_1 \mid r_t = [l, u]) \ du \ dl
\]

\[
= \int_0^{z_l} \int_{z_u}^1 (u - l)^{-1} \ du \ dl
\]

Applying the same procedure.

\[
P(y \in r_t \mid x_1, x_1 \in r_t) = \begin{cases} 
\frac{(1 - y) \ln(1 - y) + x_1 \ln x_1 - (x - y) \ln(x_1 - y)}{(1 - x_1) \ln(1 - x_1) + x_1 \ln x_1} & \text{if } y < x_1 \\
1 & \text{if } y = x_1 \\
\frac{(1 - x_1) \ln(1 - x_1) + y \ln y - (y - x_1) \ln(y - x_1)}{(1 - x_1) \ln(1 - x_1) + x_1 \ln x_1} & \text{if } x_1 < y 
\end{cases}
\]

(Obviously, this expression could be derived directly, rather than found as a special case of the more general formula.)
Computational models make it easier to be precise about one’s theories.

Categorisation is sort of related to similarity I guess?

Categorisation probability is proportional to the sum of similarities to previous exemplars.
Formal descriptions of human inductive biases are useful for machine learning:

- Inferring intention from actions
- Understanding the relevance of utterances to context
- Constructing categories from instances

Examples:
- Triangles are jerks
- "I'm not driving"
- Teapot death star?
Machine agents need to interact with humans, so they need to understand us changing speech patterns when the speech recognition system doesn’t work autonomous vehicles need to understand human drivers
Sampling assumptions in inductive generalisation
very dissimilar things are probably not tufas
it’s such a weird coincidence that ALL THREE have the SAME shape, right?
Generalisations narrow as this “coincidence” becomes suspicious

“tufas”
"The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis."

- R.A. Fisher
Let’s play… “Do what Fisher says”
Which things are tufas?

- helical things?
- seashell things?
- creepy flowers?
- mushroom heads?
- botanical radio telescopes?
- anything long and narrow

They're all tufas!
Observe one tufa and falsify…

- helical things?
- seashell things?
- creepy flowers?
- mushroom heads?
- botanical radio telescopes?
- anything long and narrow

They’re all tufas!
The next two tufas falsify no hypotheses, so we learn nothing?

helical things?

creepy flowers?

mushroom heads?

seashell things?

botanical radio telescopes?

they’re all tufas!

anything long and narrow
Mr. Ockham wishes to discuss tufas with you...
These hypotheses do not require me to believe a bizarre coincidence as to why the only observed tufas are so bloody similar.
For these to be plausible, I require an additional explanation as to why the only tufas I have seen are flower-like.

they’re all tufas!

anything long and narrow
An Ockhamist reasoner prefers the smallest hypotheses consistent with the observations.

- helical things?
- seashell things?
- mushroom heads?
- creepy flowers?
- botanical radio telescopes?
- anything long and narrow

They’re all tufas!
Are these fundamentally distinct?
Or can we express them in a common framework?

Bayes #1?

Bayes #2?
Bayes’ rule: \[ P(h|x) \propto P(x|h)P(h) \]
A Bayesian “scores” hypotheses by asking how likely they think it is that we data $x$ would be if hypothesis $h$ were true?

$$P(h|x) \propto P(x|h)P(h)$$
The likelihood supplies the learner with a theory about the data generating process

\[ P(h|x) \propto P(x|h)P(h) \]
Different theories, different learning

\[ P(h|d) = \frac{P(d|h)P(h)}{P(d)} \]

\[ P(h|d) = \frac{P(d|h)P(h)}{P(d)} \]
Two simple theories about the data generating mechanism...

Weak sampling:

“select an item at random and then provide the category label”
Two simple theories about the data generating mechanism...

Weak sampling:

“select an item at random and then provide the category label”

Strong sampling:

“select items from the target category”
\[ P(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \]

Weak sampling:

\[ P(x|h) \propto \begin{cases} 1 & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \]

\[ \implies \]

\[ P(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \]

Strong sampling:

\[ \implies \]

\[ P(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \]
And yield qualitatively different behaviour

Weak sampling:
Act like a falsificationist

Strong sampling:
Apply Ockham’s razor: prefer small/simple hypotheses
Here’s the testable prediction about generalisation gradients...

weak sampling

strong sampling

Tenenbaum & Griffiths (2001)
Here’s the testable prediction about generalisation gradients…

weak sampling

strong sampling

Tenenbaum & Griffiths (2001)
Looks like strong sampling…

Looks like strong sampling…

Hendrickson, Perfors & Navarro (in prep)
Looks like strong sampling…
But there are individual differences:

Sensitivity to sample size in simple generalisation

Insensitivity to sample size in simple generalisation

And there are **task** differences:

“Concept learning” designs where people see positive examples from one category produce the strong sampling “tightening” effect

Hendrickson, Perfors & Navarro (in prep)
And there are **task** differences:

```
“Concept learning” designs where people see positive examples from one category produce the strong sampling “tightening” effect

“Classification” designs where people see labelled examples from two categories show no tightening, only a weak base rate effect (in the opposite direction)
```

Hendrickson, Perfors & Navarro (in prep)
Manipulating the sampling assumption in an inductive reasoning task

Property induction tasks

**Grizzly Bears** produce the hormone TH-L2.

Do **Lions** produce the hormone TH-L2?
Property induction tasks

**GRIZZLY BEARS** produce the hormone TH-L2.

Do **LIONS** produce the hormone TH-L2?

- False
- True (60% certain)

Done

**GRIZZLY BEARS** produce the hormone TH-L2.

**BLACK BEARS** produce the hormone TH-L2.

Do **LIONS** produce the hormone TH-L2?

- False (65% certain)
- True

Done
#1 Grizzly Bears → Lions
#2 Grizzly Bears + Black Bears → Lions

#2 … adding the “black bears” premise doesn’t falsify any (reasonable) hypothesis

#1 … it’s a suspicious coincidence that both exemplars are bears
Weak sampling: If the premises are selected at random \textit{from the set of all true facts}, it’s just bad luck that the two exemplars happen to be bears and it is not relevant to the inductive inference.

Strong sampling: If the premises are selected by sampling \textit{from a target category} then the fact that both exemplars are both bears is entirely relevant and very informative.

#1 Grizzly Bears $\rightarrow$ Lions
#2 Grizzly Bears $+$ Black Bears $\rightarrow$ Lions
<table>
<thead>
<tr>
<th>Cover story?</th>
<th>Previous experience? (filler trials)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant cover story, Relevant fillers</td>
<td>Neutral cover story, Neutral fillers</td>
</tr>
<tr>
<td>Neutral cover story, Relevant fillers</td>
<td>Neutral cover story, Random fillers</td>
</tr>
<tr>
<td></td>
<td>Random cover story, Random fillers</td>
</tr>
</tbody>
</table>
Change in argument strength

Condition
- Both Relevant
- Relevant Fillers
- Random Fillers
- Both Random
Target 1
Target 2
Control

−0.2
−0.1
0.0
0.1
0.2

Change in argument strength

Condition

- Both Relevant
- Relevant Fillers
- Random Fillers
- Both Random

orangutans
chimpanzees
gorillas

Control

orangutans
gorillas
Target 2  Control

Change in argument strength

-0.2 -0.1 0.0 0.1 0.2

Condition
- Both Relevant
- Relevant Fillers
- Random Fillers
- Both Random

black bears
lions

golden bears
lions
Change in argument strength

<table>
<thead>
<tr>
<th>Condition</th>
<th>Target 1</th>
<th>Target 2</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Relevant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevant Fillers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Fillers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Random</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **tigers**
- **lions**
- **ferrets**
Is this consistent with Bayesian generalisation?
Yep.

Empirical data

Mixed Sampling

- $\theta = 0.31$
- $\theta = 0.22$
- $\theta = 0.11$
- $\theta = 0$

Condition
- Both Relevant
- Relevant Fillers
- Random Fillers
- Both Random

Change in argument strength

Target 1: $\theta = 0.31$
Target 2: $\theta = 0.22$
Control: $\theta = 0.11$

Empirical data: Yep.
Inductive reasoning depends on more than mere facts

- It’s not just about the evidence that facts provide for a conclusion, it’s also about how you think those facts were put together

- Bayesian models explain this as a shift in the sampling assumption
How to take a helpful hint…
(the curious power of negative evidence)

Voorspoels, Navarro, Perfors, Ransom & Storms (2015). *Cognitive Psychology*
You want to infer whether all ravens are black. Which of these observations is more helpful?
Law of contraposition makes these two statements logically equivalent.

Raven → Black
¬Black → ¬Raven
Okaaaay…. apparently these are the same?

Raven → Black
¬Black → ¬Raven

(raven, black)
(¬black, ¬raven)
<table>
<thead>
<tr>
<th>Black</th>
<th>Raven</th>
<th>¬Raven</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Raven" /></td>
<td><img src="image2.png" alt="Black" /></td>
<td>???</td>
</tr>
<tr>
<td><img src="image3.png" alt="¬Black" /></td>
<td><img src="image4.png" alt="¬Raven" /></td>
<td>???</td>
</tr>
<tr>
<td>Raven</td>
<td>¬Raven</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td><img src="image1.png" alt="Images" /> <img src="image2.png" alt="Images" /> <img src="image3.png" alt="Images" /> <img src="image4.png" alt="Images" /></td>
<td></td>
</tr>
<tr>
<td>¬Black</td>
<td><img src="image5.png" alt="Images" /> <img src="image6.png" alt="Images" /> <img src="image7.png" alt="Images" /> <img src="image8.png" alt="Images" /></td>
<td></td>
</tr>
</tbody>
</table>
Category size/frequency matters, theoretically & empirically

- Positive (labelled) categories are small

- Sampling from a small category is more powerful

- People treat positive evidence as more informative than negative evidence
  - Wason (1960, 1968), many many others…
  - So it all makes sense! And…
Paradox resolved!

A black raven is very informative

A non-black non-raven has very modest evidentiary value
So we’ll just some empirical work, with some *obviously* predictable results…

Mozart produces alpha waves

The sound of a falling rock does not
<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>¬music</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td><img src="image" alt="Beethoven" /></td>
<td><img src="image" alt="Green Shoes" /></td>
</tr>
<tr>
<td>¬alpha</td>
<td>![???]</td>
<td>This ought to be about as utterly useless as the green shoes thing</td>
</tr>
</tbody>
</table>
Okay, we start by telling people that Mozart does produce alpha waves…
... and they generalise in a way that seems terribly sensible

Bach
Nirvana
waterfall

+Mozart
Adding Metallica as a negative example has a modest, sensible effect on inferences about Nirvana.
sigh.

Bach | Nirvana | waterfall

judged likelihood

+Mozart

-Falling rock
three relevant hypotheses for the extension of the alpha waves property
positive example of classical music means people strongly endorse the narrow category
but add a negative observation from a distant category and you get a huge belief revision?
Apparently people make a (pragmatic?) inference that the negative observation is used to demarcate the category boundary.
Well, let’s ask them what they think the true extension of the property is…
Well, let’s ask them what they think the true extension of the property is…

![Graph showing generation frequency with categories such as base, sub, category, and super for just Mozart, classical music, all music, and all sounds.](image)
The negative observation shifts belief to the largest category that excludes it.
(aside: the actual experiment used many different arguments)

<table>
<thead>
<tr>
<th>topic</th>
<th>subcat A</th>
<th>premises subcat B</th>
<th>cat C</th>
<th>A-member</th>
<th>conclusions B-member</th>
<th>C-member</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUSIC</td>
<td>Mozart</td>
<td>Metallica</td>
<td>falling rock</td>
<td>Bach</td>
<td>Nirvana</td>
<td>waterfall</td>
</tr>
<tr>
<td>PAINTERS</td>
<td>Rubens</td>
<td>Dahli</td>
<td>woodcarver</td>
<td>Van Eyck</td>
<td>Warhol</td>
<td>sculpturer</td>
</tr>
<tr>
<td>PUBLIC FIGURES</td>
<td>actors</td>
<td>librarians</td>
<td>moles</td>
<td>politicians</td>
<td>programmers</td>
<td>pheasants</td>
</tr>
<tr>
<td>SHIPS</td>
<td>freight ships</td>
<td>hovercrafts</td>
<td>cars</td>
<td>cruise ships</td>
<td>sail boats</td>
<td>rocks</td>
</tr>
<tr>
<td>GLASS</td>
<td>window glass</td>
<td>bottle glass</td>
<td>art glass</td>
<td>car glass</td>
<td>drinking glass</td>
<td>jewelry glass</td>
</tr>
<tr>
<td>DISPLAYS</td>
<td>LCD</td>
<td>television</td>
<td>paintings</td>
<td>plasma</td>
<td>traffic signs</td>
<td>book page</td>
</tr>
<tr>
<td>WATER BODIES</td>
<td>Atlantic</td>
<td>Balaton</td>
<td>mustard gass</td>
<td>Mediterranean</td>
<td>Silverlake</td>
<td>olive oil</td>
</tr>
<tr>
<td>WIND</td>
<td>flute</td>
<td>guitar</td>
<td>crying child</td>
<td>clarinet</td>
<td>violin</td>
<td>door</td>
</tr>
<tr>
<td>FRUIT</td>
<td>strawberries</td>
<td>banana’s</td>
<td>grass blades</td>
<td>cranberries</td>
<td>apples</td>
<td>oak leaves</td>
</tr>
<tr>
<td>WATER BIRDS</td>
<td>ducks</td>
<td>sparrows</td>
<td>elephants</td>
<td>seagulls</td>
<td>blackbirds</td>
<td>camels</td>
</tr>
<tr>
<td>INSECTS</td>
<td>moths</td>
<td>spiders</td>
<td>lizards</td>
<td>flies</td>
<td>centipede</td>
<td>goldfish</td>
</tr>
<tr>
<td>POLAR ANIMALS</td>
<td>polar bears</td>
<td>deer</td>
<td>sow bug</td>
<td>pinguins</td>
<td>parakeet</td>
<td>ant</td>
</tr>
</tbody>
</table>
(aside: the actual experiment used many different arguments)

<table>
<thead>
<tr>
<th>topic</th>
<th>subcat A</th>
<th>premises subcat B</th>
<th>cat C</th>
<th>A-member</th>
<th>conclusions B-member</th>
<th>conclusions C-member</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUSIC</td>
<td>Mozart</td>
<td>Metallica</td>
<td>falling rock</td>
<td>Bach</td>
<td>Nirvana</td>
<td>waterfall</td>
</tr>
<tr>
<td>PAINTERS</td>
<td>Rubens</td>
<td>Dahlia</td>
<td>woodcarver</td>
<td>Van Eyck</td>
<td>Warhol</td>
<td>sculpturer</td>
</tr>
<tr>
<td>PUBLIC FIGURES</td>
<td>actors</td>
<td>librarians</td>
<td>moles</td>
<td>politicians</td>
<td>programmers</td>
<td>pheasants</td>
</tr>
<tr>
<td>SHIPS</td>
<td>freight ships</td>
<td>hovercrafts</td>
<td>cars</td>
<td>cruise ships</td>
<td>sail boats</td>
<td>rocks</td>
</tr>
<tr>
<td>GLASS</td>
<td>window glass</td>
<td>television</td>
<td>art glass</td>
<td>car glass</td>
<td>drinking glass</td>
<td>jewelry glass</td>
</tr>
<tr>
<td>DISPLAYS</td>
<td>LCD</td>
<td>Balaton</td>
<td>mustard gass</td>
<td>plasma</td>
<td>Mediterranean</td>
<td>Silverlake</td>
</tr>
<tr>
<td>WATER BODIES</td>
<td>Atlantic</td>
<td>flute</td>
<td>crying child</td>
<td>clarinet</td>
<td>violin</td>
<td>book page</td>
</tr>
<tr>
<td>WIND</td>
<td>strawberry</td>
<td>guitar</td>
<td>grass blades</td>
<td>cranberries</td>
<td>apples</td>
<td>olive oil</td>
</tr>
<tr>
<td>FRUIT</td>
<td>bananas</td>
<td>banana’s</td>
<td>elephants</td>
<td>seagulls</td>
<td>blackbirds</td>
<td>door</td>
</tr>
<tr>
<td>WATER BIRDS</td>
<td>ducks</td>
<td>sparrows</td>
<td>lizards</td>
<td>flies</td>
<td>centipede</td>
<td>oak leaves</td>
</tr>
<tr>
<td>INSECTS</td>
<td>moths</td>
<td>spiders</td>
<td>sow bug</td>
<td>pinguins</td>
<td>parakeet</td>
<td>camels</td>
</tr>
<tr>
<td>POLAR ANIMALS</td>
<td>polar bears</td>
<td>deer</td>
<td></td>
<td></td>
<td></td>
<td>goldfish</td>
</tr>
</tbody>
</table>

plus we ran an entire pseudo-replication with different items
(and yes, the replication worked)
(and yes, the replication worked)

The big question is how to account for the results...
Does the **weak** sampling model capture the effect?

Weak sampling

- **Nope.**
Okay, does the “strong sampling” model capture the effect?

Meh.

(Out-Bayesing Bayes??!)
Here’s a model that gets the effect size right…
But we’re going to need a bigger hat.
Weak sampling

An argument consists of random true statements about the world
An argument consists of random true statements about the world

An argument consists of randomly selected facts particular to a target category
An argument consists of randomly selected facts particular to a target category.

An argument consists of randomly true statements about the world.

An argument consists of purposefully chosen facts designed to convince an intelligent reasoner of the truth of some proposition.
The data $x$ selected by the communicator…

$P(x|h) \propto P(h|x)^\alpha$

… is designed to maximise the learner’s posterior degree of belief in hypothesis $h$. 
If that’s right, then the same manipulation we used in the previous study should work...

If the negative example is perceived as a “helpful hint” we should continue to get the effect.

If it is construed as an arbitrary fact, the effect should vanish.
<table>
<thead>
<tr>
<th>topics</th>
<th>premise 1 (+)</th>
<th>premise 2 (-)</th>
<th>A-member</th>
<th>B-member</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUSIC</td>
<td>Mozart</td>
<td>waterfall</td>
<td>Bach</td>
<td>Nirvana</td>
</tr>
<tr>
<td>FRUIT</td>
<td>strawberries</td>
<td>grass blades</td>
<td>blackberry</td>
<td>apple</td>
</tr>
<tr>
<td>BIRDS</td>
<td>ducks</td>
<td>elephants</td>
<td>swan</td>
<td>blackbird</td>
</tr>
<tr>
<td>TYPES OF WATER</td>
<td>Atlantic ocean</td>
<td>tap water</td>
<td>Mediterranean</td>
<td>Lake Balaton</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>fillers</th>
<th>weak sampling</th>
<th>conclusion 1</th>
<th>conclusion 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXAMPLE</td>
<td>sheep (+)</td>
<td>dogs (-)</td>
<td>horses</td>
</tr>
<tr>
<td>TRIAL 1</td>
<td>aluminium (+)</td>
<td>lead (+)</td>
<td>copper</td>
</tr>
<tr>
<td>TRIAL 2</td>
<td>Earth (+)</td>
<td>weather satellite (-)</td>
<td>Uranus</td>
</tr>
<tr>
<td>FILLER</td>
<td>physicists (+)</td>
<td>engineers (+)</td>
<td>mathematicians</td>
</tr>
<tr>
<td>FILLER</td>
<td>cobras (+)</td>
<td>iguanas (-)</td>
<td>pythons</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>fillers</th>
<th>helpful sampling</th>
<th>conclusion 1</th>
<th>conclusion 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXAMPLE</td>
<td>sheep (+)</td>
<td>cows (+)</td>
<td>horses</td>
</tr>
<tr>
<td>TRIAL 1</td>
<td>aluminium (+)</td>
<td>brass (-)</td>
<td>copper</td>
</tr>
<tr>
<td>TRIAL 2</td>
<td>Earth (+)</td>
<td>Mars (+)</td>
<td>Uranus</td>
</tr>
<tr>
<td>FILLER</td>
<td>cobras (+)</td>
<td>pythons (-)</td>
<td>vipers</td>
</tr>
<tr>
<td>FILLER</td>
<td>physicists (+)</td>
<td>mathematicians (+)</td>
<td>chemists</td>
</tr>
</tbody>
</table>

200 participants on MTurk
Negative evidence framed as a “hint” produces the effect
Arbitrary negative evidence does not
• The social aspect to inductive reasoning is central
  • By default, people seem to “read” an inductive argument as if it were put together for a purpose

• Pedagogical sampling as normative standard
  • In real life, arguments aren’t collections of facts
  • They’re acts of persuasion
  • If so, shouldn’t “normative” accounts reflect that?
Inductive biases reflect the structure of the mind and the environment

Languages carve the world into categories in different ways.

### English
- **ON**
  - put magnet on refrigerator
  - put cup on table
  - put hat on
  - put ring on finger
  - put top on pen

- **IN**
  - button a button
  - put cassette in case
  - close tightly latching drawer
  - put book in fitted box-covers
  - put piece in puzzle

- **apple in bowl
  - put book in bag

### Korean
- **ON**
  - put magnet on refrigerator
  - put cup on table
  - put hat on

- **IN**
  - button a button
  - put cassette in case
  - close tightly latching drawer
  - put book in fitted box-covers
  - put piece in puzzle

- **apple in bowl
  - put book in bag

### Korean
- **PWUCHITA**
- **NOHTA**
- **SSUTA**
- **KKITA**
- **NEHTA**
How do naming systems evolve?

- the structure of the world being communicated about
- cognitive biases of the learners
- the dynamics of communication
The usual idea in categorisation: stimulus distribution (i.e., the world) shapes inferences.
Language evolution as “iterated learning” places the inductive bias within the learner...

A sequence of Bayesian learners, each learning from the language output of the last one, and then generating the input for the next one... converges to the learner’s prior.

Griffiths & Kalish (2005, 2007)
This is *weird*. Suppose this is my prior:
And these are entities that need names
Convergence to the prior is *absurd*.

Your language should do this? *(high prior)*

Your language should not do this? *(low prior)*
Where did things go astray?  
The sampling assumption…

\[ \ell = P(y|x) \]

A language provides labels \( y \) for entities \( x \)

It says **nothing** about which entities \( x \) will be observed

Griffiths & Kalish (2005, 2007)

“laser”

My language has a word for laser… does that really tell me **nothing at all** about my chances of encountering one?
Where did things go astray?
The **sampling assumption**…

\[ \ell = P(y|x) \]

A language provides labels \( y \) for entities \( x \)

It says nothing about which entities \( x \) will be observed

Griffiths & Kalish (2005, 2007)

\[ \ell = P(y, x) \]

A language provides labels \( y \) for entities \( x \), but it **also** makes assumptions about which entities \( x \) are likely to appear

Perfors & Navarro (2011, 2014)
Leads to different ideas about learning

\[ \ell = P(y|x) \]

A language provides labels \( y \) for entities \( x \)

It says nothing about which entities \( x \) will be observed

\[ \ell = P(y, x) \]

A language provides labels \( y \) for entities \( x \), but it also makes assumptions about which entities \( x \) are likely to appear

Labels converge to the prior distribution

Labels converge to the expected posterior given the entities (sort of)
Three stimulus domains

Attend to size
Three stimulus domains

Attend to size

Attend to colour
Three stimulus domains

Control

Attend to size

Attend to colour
An iterated learning design

<table>
<thead>
<tr>
<th>LUM</th>
<th>YOP</th>
<th>BIR</th>
<th>SUF</th>
<th>GUG</th>
<th>POF</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIC</td>
<td>RIN</td>
<td>NER</td>
<td>LEP</td>
<td>VOD</td>
<td>CIP</td>
</tr>
<tr>
<td>ROK</td>
<td>DUT</td>
<td>KUN</td>
<td>TAF</td>
<td>MIG</td>
<td>DIL</td>
</tr>
<tr>
<td>BAP</td>
<td>VAM</td>
<td>MAK</td>
<td>PAG</td>
<td>FOD</td>
<td>YEM</td>
</tr>
<tr>
<td>TUS</td>
<td>HEC</td>
<td>DOX</td>
<td>CEN</td>
<td>RUP</td>
<td>KOT</td>
</tr>
<tr>
<td>FIM</td>
<td>SOR</td>
<td>PIK</td>
<td>BOM</td>
<td>YIB</td>
<td>MEF</td>
</tr>
</tbody>
</table>

Responses from the participant become the target labels for the next participant.

- **Training**: see a random subset of those labels (SEEN) twice each
- **Testing**: half SEEN, half UNSEEN, asked to generate label (no feedback)
- **Training**: see a random subset of those labels (SEEN) twice each
- **Testing**: half SEEN, half UNSEEN, asked to generate label (no feedback)
- **Training**: see a random subset of those labels (SEEN) twice each
- **Final Test**: all stimuli, asked to generate label (no feedback)
What patterns do we expect?

Control

Size

Colour
That’s exactly what we find
That’s exactly what we find

Quantifying the fit:

<table>
<thead>
<tr>
<th></th>
<th>Expected Size</th>
<th>Expected Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>-0.0204</td>
<td>0.0618</td>
</tr>
<tr>
<td>Size</td>
<td>0.704</td>
<td>0.079</td>
</tr>
<tr>
<td>Colour</td>
<td>0.065</td>
<td>0.696</td>
</tr>
</tbody>
</table>

(*Adjusted Rand index compared to expected patterns for size and colour; 1=identical; 0=random)
Inductive biases are complicated

- Convergence to the “prior” is deeply suspicious
  - It has been interpreted badly in the literature … taken to mean that the structure of actually existing languages transparently reveal the cognitive biases / universal grammar
  - But this is very wrong

- Convergence to “something a bit like a posterior conditioned on expected evidence from the environment” is less sexy
  - But it does have the advantage of being true
  - Well, maybe…
Measuring inductive bias is tricky…

Navarro, Kary, Donkin, Perfors, Brown?? ("in prep")
Iterated learning only works when people share the same biases:
... and I can prove it.

The mixed chain misrepresents the mixture of learners

Suppose it were true that “iterated learning reveals the prior” holds even in the presence of mixed learners. Suppose that our two learner types agree on the hypothesis space (i.e., shared values of \( p \) and \( q \)) above, and disagree only in the prior probability of hypothesis 1, denoted \( b_1 \) for learner type 1 and \( b_2 \) for learner type 2. If each learner type converges to their own prior (as they do in a homogenous chain), then the observable probability that learner type 1 produces \( x = 1 \) is given by the prior predictive probability:

\[
P(x = 1|c = 1) = \sum_{h \in \{h_1, h_2\}} P(x = 1|h_1)P(h_1|c = 1) = pb_1 + q(1 - b_1)
\]

and equivalently

\[
P(x = 1|c = 2) = pb_2 + q(1 - b_2)
\]

Accordingly, since the proportion of learners of type 1 in the mixed chain is \( w \) the marginal probability of a success should be the mixture of prior predictives:

\[
P(x = 1) = wP(x = 1|c = 1) + (1 - w)P(x = 1|c = 2)
\]

\[
= w(pb_1 + q(1 - b_1)) + (1 - w)(pb_2 + q(1 - b_2))
\]

However, substitution into the expressions above reveals that this is far from the case:

\[
P(x = 1) = \frac{u}{u + d} = \left(1 + \frac{P(x_i = 0|x_{i-1} = 1)}{P(x_i = 1|x_{i-1} = 0)}\right)^{-1}
\]

\[
= \left(1 + \frac{wP(x_i = 0|x_{i-1} = 1, c_i = 1) + (1 - w)P(x_i = 0|x_{i-1} = 1, c_i = 2)}{wP(x_i = 1|x_{i-1} = 0, c_i = 1) + (1 - w)P(x_i = 1|x_{i-1} = 0, c_i = 2)}\right)^{-1}
\]

\[
= \left(1 + \frac{w(p(1-p)b_2 + q(1-q)(1-b_2))}{pb_2 + q(1-b_2)} + (1 - w)(p(1-P)b_2 + q(1-q)(1-b_2))}{pb_2 + q(1-b_2)} + (1 - w)(p(1-P)b_2 + q(1-q)(1-b_2))\right)^{-1}
\]

Suffice it to say, these are not at all the same except in special cases.
But it’s more fun to run dumb experiments that give stupid answers…

Americans *claim* to be totally ignorant about the Oz election…
But it’s more fun to run dumb experiments that give stupid answers...

... and a “pure” iterated learning chain “reveals” a preference for Gordon Brown ...
But it’s more fun to run dumb experiments that give stupid answers…

But if we mix some UNSW students into the chain the Americans now appear to back Malcolm Turnbull
A few final thoughts about human induction and Bayesian reasoning
Traditional accounts of learning and inference specify norms that implicitly rely on something like falsificationist reasoning.
But why?

... it **only** makes sense when evidence is selected in an arbitrary and random fashion.
In real life, isn’t \textbf{ANYTHING ELSE} a more reasonable theory for the origin of the data???
“Common sense” inference requires people to learn from complex (and smart) data sources...
We need to disentangle facts from agendas
We need to detect trickery
We need to know when to reject the rules/concepts we’re given
We need to read the intention of other agents.
Understanding human common sense reasoning requires something a lot richer
Thanks!
“Expert opinions” about the likelihood that climate change is happening

If an article can only fit three quotes, most people choose to include the dissenter

A Bayesian journalist who cares about their reputation has a strong motivation to pursue “he says she says” journalism

Because a Bayesian reader can’t tell the difference between journalistic bias and expert consensus
### Constrained host

<table>
<thead>
<tr>
<th>Prize location, ( l )</th>
<th>Host choice, ( o )</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

### Random host

<table>
<thead>
<tr>
<th>Prize location, ( l )</th>
<th>Host choice, ( o )</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td></td>
</tr>
</tbody>
</table>

### Uninformative host

<table>
<thead>
<tr>
<th>Prize location, ( l )</th>
<th>Host choice, ( o )</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>( \emptyset )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

### Malicious host

<table>
<thead>
<tr>
<th>Prize location, ( l )</th>
<th>Host choice, ( o )</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>( \emptyset )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### Helpful host

<table>
<thead>
<tr>
<th>Prize location, ( l )</th>
<th>Host choice, ( o )</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>( \emptyset )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

### Inferred Intention

- **Cooperative**
- **Random**
- **Antagonistic**
- **Original MHP**

### Similarity

- **Rated Similar**
- **Same Choice**
- **Same Inference**
Experimental tests

- Navarro, Dry & Lee (2012):
  - Two experiments, stimuli varied on one dimension
  - N=22 & N=20 undergraduates
  - Non traditional stimulus presentation
  - Response measure: Probability judgments

- Vong, Hendrickson, Perfors & Navarro (2013)
  - As above, but with N=318 workers on AMT

- Hendrickson, Perfors & Navarro (in preparation)
  - One experiment (N=470) on AMT
  - Participants shown traditional categorisation stimuli (below)
  - Response measures: probability judgment & categorisation decisions

stimuli:
Hypotheses inferred from a separate data set

*This illustration only shows a high-weighted sets that contain at least two animals. The actual prior assigned non-zero prior probability to every possible subset of the set of all animals that appeared in the task. Qualitative features of model predictions are robust to the specific choice of prior: anything even semi-reasonable seems to work.
(Chimpanzee, Gorilla, Orangutan)
(Lions, Tigers) but not Ferrets
The prior explains why there are structural differences between the targets and the control.

The likelihood describes how “adding more premises” can have different effects across conditions.
Stimulus ordering was fixed and designed to ensure that fillers (mostly) preceded targets:

<table>
<thead>
<tr>
<th>First generalisation</th>
<th>Helpful</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAGLES → DOVES</td>
<td>+HAWKS</td>
<td>−TORTOISES</td>
</tr>
<tr>
<td>ELEPHANTS → DEERS</td>
<td>+COWS</td>
<td>+ANTEATERS</td>
</tr>
<tr>
<td>TIGERS → FERRETS</td>
<td>+LIONS</td>
<td>+LIONS</td>
</tr>
<tr>
<td>KANGAROOS → WOMBATS</td>
<td>+KOALAS</td>
<td>−FLAMINGOS</td>
</tr>
<tr>
<td>GRIZZLY BEARS → LIONS</td>
<td>+BLACK BEARS</td>
<td>+BLACK BEARS</td>
</tr>
<tr>
<td>ORANGUTANS → GORILLAS</td>
<td>+CHIMPANZEES</td>
<td>+CHIMPANZEES</td>
</tr>
</tbody>
</table>

Participants were 296 people recruited through MTurk
Here’s the complete data set...

Size

Iteration →

A

B
Here’s the complete data set...
Here’s the complete data set...

Control

A

B

Size

A

B

Colour

A

B