On the origins of data

Dan Navarro
University of Adelaide
The lab has been busy lately, and I really wanted to talk about all their good work...
Drew has cool stuff looking at the kind of evidence people prefer to learn from…

Hendrickson, Navarro & Perfors (under revision). Sensitivity to hypothesis size during information search. *Decision*
Steve has cool stuff looking at how people learn (and use) “admissible” stimulus transformations.

Langsford, Navarro, Perfors & Hendrickson (under review). Transformation learning and its effect on similarity. *JEP:LMC*
Lauren has some scarily effective ideas about how the analysis of clinical trials could be done better...

For every participant $i$

- $a(i) \sim \text{bern}(0.5)$ #condition

- $z(i) \sim \text{dbern}(p(j))$ #got better or not

For $j$

- $p(j) \sim \text{dbeta}(1,1)$
- $\text{Effect Mean}(j) \sim \text{dunif}(0,100)$

Initial Scale
Wai Keen thinks much of the semi-supervised learning literature is missing the point...

Sean has an awesome rant about how Bayesian cognitive modeling ought to work.

Simon has a semantic network model for predicting similarities between very unrelated words

De Deyne, Navarro, Perfors, Storms (submitted?). Structure at every scale: A semantic network account of the similarities between very unrelated concepts. *JEP:G*
Navarro, Newell & Schulze (under revision). Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology*

Navarro & Kemp (in preparation). None of the above: A Bayesian account of the section of novel categories. *Psych Review?*
In the end, I had to ignore most people and concentrate on one line of work... :-(

**Hendrickson**, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations.


**Drew Hendrickson**, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations


So... I want to build a smart machine, and I want it to do human-like inductive reasoning
So… I want to build a smart machine, and I want it to do human-like inductive reasoning.

I want it to have common sense.
So I have to ask…
Why isn’t inductive inference simple?
Why isn’t inductive inference simple?

\[ P(h|d) = \frac{P(d|h)P(h)}{P(d)} \]
A simple learning rule...

\[ P(h|d) = \frac{P(d|h)P(h)}{P(d)} \]
... hides a lot of complexity

Other possible worlds

True state of the world

Observed data

Inferred state of the world

Other learners

Other factors

“sampling”

“learning”
And what this means is that even “simple” problems become surprisingly tricky…

\[
P(h|d) = \frac{P(d|h)P(h)}{P(d)}
\]
Sampling assumptions in simple generalisation problems

Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations.


Perfors, Navarro & Shafto (in preparation). Stronger evidence isn’t always better: The role of social inference in evidence selection and interpretation. Previously rejected from *JEP:G*
“tufa”
| “tufa” |
“tufa”
Why **should** generalizations become narrower with more positive examples?
There is a puzzle here…

"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
Okay let’s reason like a falsificationist...

Here are some objects
And seem plausible a priori hypotheses for the extension of a novel category.
The first labelled object eliminates some hypotheses
... and two more
Generalization to these items should be the same in both cases. It is not
Ockham’s razor: the smaller hypothesis provides a simpler explanation for why all the observed tufas look so damn similar.
Generalizations should tighten around the positive exemplars as the sample size increases.
A tale of two Bayesians

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

\[ P(h|d) = \frac{P(d|h)P(h)}{P(d)} \]
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 is the learner’s theory about the problem they’re solving

\[ P(h|x) \propto P(x|h)P(h) \]
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 very simple theories…

Weak sampling:

“select an item at random and then provide the category label”
Two very simple theories…

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

Strong sampling:
“make sure you pick an item that actually belongs to the target category”
... produce two different learning rules

**Weak sampling:**

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

**Strong sampling:**

\[
P(x|h) = \begin{cases} 
\frac{1}{|h|} & \text{if } x \in h \\
0 & \text{otherwise}
\end{cases}
\]
And 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
And a series of 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
And a series of 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
And a series of 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:
Looks like strong sampling...

Vong, et al (2013) - probability judgment with “toy” task
Looks like strong sampling...

Hendrickson, et al (in prep) - probability judgments
Looks like strong sampling…

Hendrickson, et al (in prep) - categorisation data
But there are individual differences:

Sensitivity to sample size in simple generalisation

Insensitivity to sample size in simple generalisation

Navarro et al (2012)
And there are **task** differences:

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

Hendrickson, et al (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, et al (in prep)
• The tightening effect predicted by strong sampling does happen

• But there are differences across individuals and across tasks

• The task differences make sense if you assume people are forming theories about how the experiment(er) designed the task

• This starts to feel like social cognition…
Relevance, social cognition and inductive reasoning

Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations


Perfors, Navarro & Shafto (in preparation). Stronger evidence isn’t always better: The role of social inference in evidence selection and interpretation. Previously rejected from *JEP:G*
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.

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

- False
- True (60% certain)

---

**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
Grizzly Bears $\rightarrow$ Lions
Grizzly Bears + Black Bears $\rightarrow$ Lions

Adding the “Black Bears” premise weakens the argument?
Grizzly Bears → Lions
Grizzly Bears + Black Bears → Lions

Tigers → Ferrets
Tigers + Lions → Ferrets

Same thing with the “Lions” premise
Conversely, the “Chimpanzee” premise strengthens the argument here.
Grizzly Bears + Black Bears
Tigers + Lions
Orangutans + Chimpanzees

In all cases the additional premise concentrates beliefs around a target category, e.g. bears, cats, primates
Is this “tightening” effect related to the tightening in the “tufa” generalisation problem?
Is this “tightening” effect related to the tightening in the “tufa” generalisation problem?

Does it depend on the learner’s theory about how the argument was constructed?
Is this “tightening” effect related to the tightening in the “tufa” generalisation problem?

Does it depend on the learner’s theory about how the argument was constructed?

Can we produce qualitative shifts in people’s reasoning by manipulating their theory about how the argument was made?
<table>
<thead>
<tr>
<th>Cover story?</th>
<th>Relevant cover story, Relevant fillers</th>
<th>Neutral cover story, Relevant fillers</th>
<th>Neutral cover story, Random fillers</th>
<th>Random cover story, Random fillers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous experience? (filler trials)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cover story manipulation

- **Relevant**: people were told that the “additional” premise was chosen by a helpful teacher.
- **Neutral**: people were told nothing about how the second premise was generated.
- **Random**: people were told that the second premise was selected at random from the set of true facts.
Eagles → Doves
Elephants → Deer
Kangaroos → Wombats

Three “filler” arguments
Eagles → Doves
Elephants → Deer
Kangaroos → Wombats

Eagles + Hawks → Doves
Elephants + Cows → Deer
Kangaroos + Koalas → Wombats

... with a relevant second premise

(positive premises from the same category suggest strong sampling)
Eagles $\rightarrow$ Doves
Elephants $\rightarrow$ Deer
Kangaroos $\rightarrow$ Wombats

Eagles + Hawks $\rightarrow$ Doves
Elephants + Cows $\rightarrow$ Deer
Kangaroos + Koalas $\rightarrow$ Wombats

Eagles - Tortoises $\rightarrow$ Doves
Elephants + Anteaters $\rightarrow$ Deer
Kangaroos - Flamingos $\rightarrow$ Wombats

... or a random one

(negated premises unrelated to the topic suggest weak sampling)
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>+CHIMPANZEEES</td>
<td>+CHIMPANZEE</td>
</tr>
</tbody>
</table>
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

orangutans
gorillas

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

Change in argument strength

Control

orangutans
gorillas

orangutans
gorillas
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

grizzly bears
black bears
lions

grizzly bears
lions
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

Condition

Tigers
Lions
Ferrets
(Bayesian) data analysis: hypothesis tests for order restricted models

<table>
<thead>
<tr>
<th>Model</th>
<th>Order restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO EFFECT</td>
<td>$\mu_1 = \mu_2 = \mu_3 = \mu_4$</td>
</tr>
<tr>
<td>FILLERS ONLY</td>
<td>$\mu_1 = \mu_2 &lt; \mu_3 = \mu_4$</td>
</tr>
<tr>
<td>STORY ONLY</td>
<td>$\mu_1 &lt; \mu_2 = \mu_3 &lt; \mu_4$</td>
</tr>
<tr>
<td>BOTH</td>
<td>$\mu_1 &lt; \mu_2 &lt; \mu_3 &lt; \mu_4$</td>
</tr>
<tr>
<td>RANDOM EFFECT</td>
<td>$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$</td>
</tr>
</tbody>
</table>
Clear effect of cover story on targets, possibly also an effect of filler type

<table>
<thead>
<tr>
<th>Model</th>
<th>Order restrictions</th>
<th>Bayes Factor ( : NO EFFECT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Target 1</td>
</tr>
<tr>
<td><strong>NO EFFECT</strong></td>
<td>$\mu_1 = \mu_2 = \mu_3 = \mu_4$</td>
<td>-</td>
</tr>
<tr>
<td><strong>FILLERS ONLY</strong></td>
<td>$\mu_1 = \mu_2 &lt; \mu_3 = \mu_4$</td>
<td>740:1</td>
</tr>
<tr>
<td><strong>STORY ONLY</strong></td>
<td>$\mu_1 &lt; \mu_2 = \mu_3 &lt; \mu_4$</td>
<td>4,100:1</td>
</tr>
<tr>
<td><strong>BOTH</strong></td>
<td>$\mu_1 &lt; \mu_2 &lt; \mu_3 &lt; \mu_4$</td>
<td>2,900:1</td>
</tr>
<tr>
<td><strong>RANDOM EFFECT</strong></td>
<td>$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$</td>
<td>520:1</td>
</tr>
</tbody>
</table>
Null effect for the control item

<table>
<thead>
<tr>
<th>Model</th>
<th>Order restrictions</th>
<th>Bayes Factor ( : NO EFFECT )</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO EFFECT</td>
<td>$\mu_1 = \mu_2 = \mu_3 = \mu_4$</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>FILLERS ONLY</td>
<td>$\mu_1 = \mu_2 &lt; \mu_3 = \mu_4$</td>
<td></td>
<td>$&lt; 1 : 1$</td>
</tr>
<tr>
<td>STORY ONLY</td>
<td>$\mu_1 &lt; \mu_2 = \mu_3 &lt; \mu_4$</td>
<td></td>
<td>$&lt; 1 : 1$</td>
</tr>
<tr>
<td>BOTH</td>
<td>$\mu_1 &lt; \mu_2 &lt; \mu_3 &lt; \mu_4$</td>
<td></td>
<td>$&lt; 1 : 1$</td>
</tr>
<tr>
<td>RANDOM EFFECT</td>
<td>$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$</td>
<td></td>
<td>$&lt; 1 : 1$</td>
</tr>
</tbody>
</table>
Should we model this as a difference between two Bayesian learners?

A weakly sampling falsificationist

A strongly sampling Ockhamist
Or posit a continuum of Bayesians?

θ = 0  θ = 0.33  θ = 0.67  θ = 1

“weak”  “strong”
And what shall our Bayesians use for their hypothesis space and priors?
Assume any subset of items is a legitimate hypothesis, with weights inferred from similarity judgments.

* 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
There are many high weighted features involving these three, but overall the prior puts the bears together more often.
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.
Does the model work???
Model fits

Mixed Sampling

\[ \theta = 0.31 \quad \theta = 0.22 \]
\[ \theta = 0.11 \quad \theta = 0 \]

Empirical data

Condition

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

• Bayesian models explain the reversal as a shift in the sampling assumption

• This is encouraging, so…
How to take a helpful hint...
(the curious power of negative evidence)

Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations.


Perfors, Navarro & Shafto (in preparation). Stronger evidence isn’t always better: The role of social inference in evidence selection and interpretation. Previously rejected from *JEP:G*
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 $\rightarrow$ Black

$\neg$Black $\rightarrow$ $\neg$Raven

(raven, black)

($\neg$black, $\neg$raven)
<table>
<thead>
<tr>
<th>Raven</th>
<th>¬Raven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td></td>
</tr>
<tr>
<td>¬Black</td>
<td>???</td>
</tr>
<tr>
<td>Raven</td>
<td>~\text{Raven}</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
</tr>
<tr>
<td>Black</td>
<td><img src="image1" alt="Raven" /> <img src="image2" alt="Shoe" /> <img src="image3" alt="Car" /> <img src="image4" alt="Sofa" /> <img src="image5" alt="Objects" /></td>
</tr>
<tr>
<td>~\text{Black}</td>
<td><img src="image6" alt="Objects" /> <img src="image7" alt="Shoe" /> <img src="image8" alt="Hat" /> <img src="image9" alt="Shoe" /> <img src="image10" alt="Pepper" /> <img src="image11" alt="Bottle" /> <img src="image12" alt="Shoe" /> <img src="image13" alt="Sunglasses" /></td>
</tr>
</tbody>
</table>
Category size/frequency matters, theoretically & empirically

• Positive (labelled) categories are small
  • Oaksford & Chater (1998), Navarro & Perfors (2011), etc.

• 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 non-zero but negligible 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>music</th>
<th>¬music</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td><img src="image" alt="Beethoven" /></td>
</tr>
<tr>
<td>¬alpha</td>
<td><img src="image" alt="Rocks" /></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
Adding Metallica as a negative example has a small effect (yay!)
OKAY WTF HUMANS I HATE YOU ALL.
| classical music | all music | all sound |

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…

Mozart+

![Bar chart showing generation frequency for different categories: base, sub, category, super, with additional labels for just Mozart, classical music, all music, all sounds.](chart_image)
Well, let’s ask them what they think the true extension of the property is…
And there it is.
(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>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>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>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>Silverlake</td>
<td>book page</td>
</tr>
<tr>
<td>WATER BODIES</td>
<td>Atlantic</td>
<td>television</td>
<td>crying child</td>
<td>Mediterranean</td>
<td>violin</td>
<td>olive oil</td>
</tr>
<tr>
<td>WIND</td>
<td>flute</td>
<td>guitar</td>
<td>grass blades</td>
<td>clarinet</td>
<td>apples</td>
<td>door</td>
</tr>
<tr>
<td>FRUIT</td>
<td>strawberries</td>
<td>banana’s</td>
<td>elephants</td>
<td>cranberries</td>
<td>cranberries</td>
<td>oak leaves</td>
</tr>
<tr>
<td>WATER BIRDS</td>
<td>ducks</td>
<td>sparrows</td>
<td>seagulls</td>
<td>blackbirds</td>
<td>seagulls</td>
<td>camels</td>
</tr>
<tr>
<td>INSECTS</td>
<td>moths</td>
<td>spiders</td>
<td>flies</td>
<td>centipede</td>
<td>flies</td>
<td>goldfish</td>
</tr>
<tr>
<td>POLAR ANIMALS</td>
<td>polar bears</td>
<td>deer</td>
<td>pinguins</td>
<td>parakeet</td>
<td>ant</td>
<td>ant</td>
</tr>
</tbody>
</table>

plus we ran an entire pseudo-replication with different items

<table>
<thead>
<tr>
<th>topic</th>
<th>subcat A</th>
<th>premises</th>
<th>cat C</th>
<th>A-member</th>
<th>conclusions B-member</th>
<th>C-member</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAMMALS</td>
<td>dog (+)</td>
<td>magpie (-)</td>
<td>wolf</td>
<td>donkey</td>
<td>blackbird</td>
<td></td>
</tr>
<tr>
<td>BIRDS</td>
<td>crow (+)</td>
<td>tuna fish (-)</td>
<td>raven</td>
<td>swan</td>
<td>halibot</td>
<td></td>
</tr>
<tr>
<td>FISH</td>
<td>salmon (+)</td>
<td>lizzard (-)</td>
<td>codfish</td>
<td>goldfish</td>
<td>snake</td>
<td></td>
</tr>
<tr>
<td>INSECTS</td>
<td>bee (+)</td>
<td>sparrow (-)</td>
<td>ant</td>
<td>cricket</td>
<td>pigeon</td>
<td></td>
</tr>
</tbody>
</table>
(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?

No, it predicts a null effect.
Okay, does the “strong sampling” model capture the effect?

Weak sampling

Strong sampling

Yes, but the effect is much smaller than the empirical one

(people are out-Bayesing Bayes??)
Well, here’s a model that gets the effect size right…
But Bayes is going to need a fancier hat…

Pedagogical sampling
Weak sampling

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.

Weak sampling

Strong sampling

Pedagogical / persuasive sampling
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 weak sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>premise 1</td>
</tr>
<tr>
<td>EXAMPLE</td>
</tr>
<tr>
<td>TRIAL 1</td>
</tr>
<tr>
<td>TRIAL 2</td>
</tr>
<tr>
<td>FILLER</td>
</tr>
<tr>
<td>FILLER</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>fillers helpful sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>premise 1</td>
</tr>
<tr>
<td>EXAMPLE</td>
</tr>
<tr>
<td>TRIAL 1</td>
</tr>
<tr>
<td>TRIAL 2</td>
</tr>
<tr>
<td>FILLER</td>
</tr>
<tr>
<td>FILLER</td>
</tr>
</tbody>
</table>

200 participants on MTurk
Adding negative evidence as a “hint” produces the effect, as before
Presenting it as an arbitrary fact makes the effect vanish...
• 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?
Let’s make the social aspect explicit:

The role of goals and social reasoning when aggregating expert opinions

Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations


Perfors, Navarro & Shafto (in preparation). Stronger evidence isn’t always better: The role of social inference in evidence selection and interpretation. Previously rejected from *JEP:G*
You’re a journalist writing an article about expert opinions about climate change…
You’re a journalist writing an article about expert opinions about climate change...

93% likely
95% likely
97% likely
99% likely
99% likely
92% likely
91% likely
89% likely
You’re a journalist writing an article about expert opinions about climate change…
Your editor says the article only has room for (at most) three quotes. Who to choose??

Here’s your full distribution of expert opinion.
| 93 | 95 | 97 | 92 |
| 99 | 91 | 89 | 5  |

Do you quote only from the consensus?
(maximises distributional similarity)
Or do you include the dissenter?
("full spectrum" but terrible approximation)
Some empirical data. Even when outnumbered 11 to 1, most people choose to quote the contrarian.
A hypothesis space of possible expert distributions
A hypothesis space of possible journalistic agendas

“Helpful” → Communicate the true distribution

“Bias high” → Communicate a distribution with highest/lowest mean

“Bias low”
Select evidence to manipulate the reader’s beliefs
Bayesian writer

Select evidence to manipulate the reader’s beliefs

Bayesian reader

Guess the true distribution **AND**
infer the journalistic agenda
So what does a Bayesian reader infer about the Bayesian writer?

(I’ll assume uniform priors over possible agendas and over possible distributional hypotheses)
Quoting one expert only looks suspicious

“Dude, you only quoted one person BUT you had room for more? Very suspicious…”
Anything less than maximum number of experts causes a deterioration of trust.
But when all the quoted experts agree, the reader thinks you’re probably biased.
You can increase the reader’s trust by including the contrarian
A Bayesian journalist who cares about their reputation has a strong motivation to pursue “he says she says” journalism.
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.
Oh, and we have a heap of other data and modelling on this too, but I have no time…
A few final thoughts about human reasoning 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 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 detect novelty and invariances in a dynamic world.
We need to read the intention of other agents
Understanding human common sense reasoning requires something a lot richer
Thanks!