Unpacking additivity biases reveal the class of sampling algorithm for word predictability

Yutong Zhang and Matthew Husband

University of Oxford

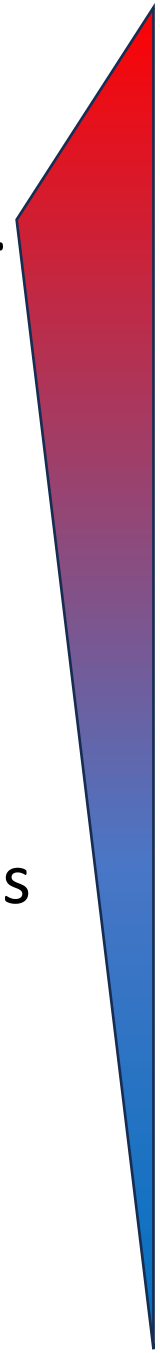
Jordan learned a lot about cars from his…

“dad” is *predictable* given this context.

- Cloze: 0.45 (Peelle et al 2020)
- GPT-2 surprisal: 3.76

As are many other potential continuations

Question: How do we arrive at ‘dad’?



	cloze *(Peelle et al. 2020)*	suprisal *(-log(GPT-2))*
experience		
childhood		
dad	0.45	3.76
youth		
mom		
uncle		
teacher		
…		
studies		
…		
magazines		
…		
guessing		
bananas		
judiciary		

Roadmap

- Word Predictability as Bayesian Inference
- Classes of Sampling Algorithms for Approximate Bayesian Inference
- Biases in Human Judgement and Decision Making
- Initial Study examining Biases in Word Predictability
- Discussion & Conclusion

Word Predictability

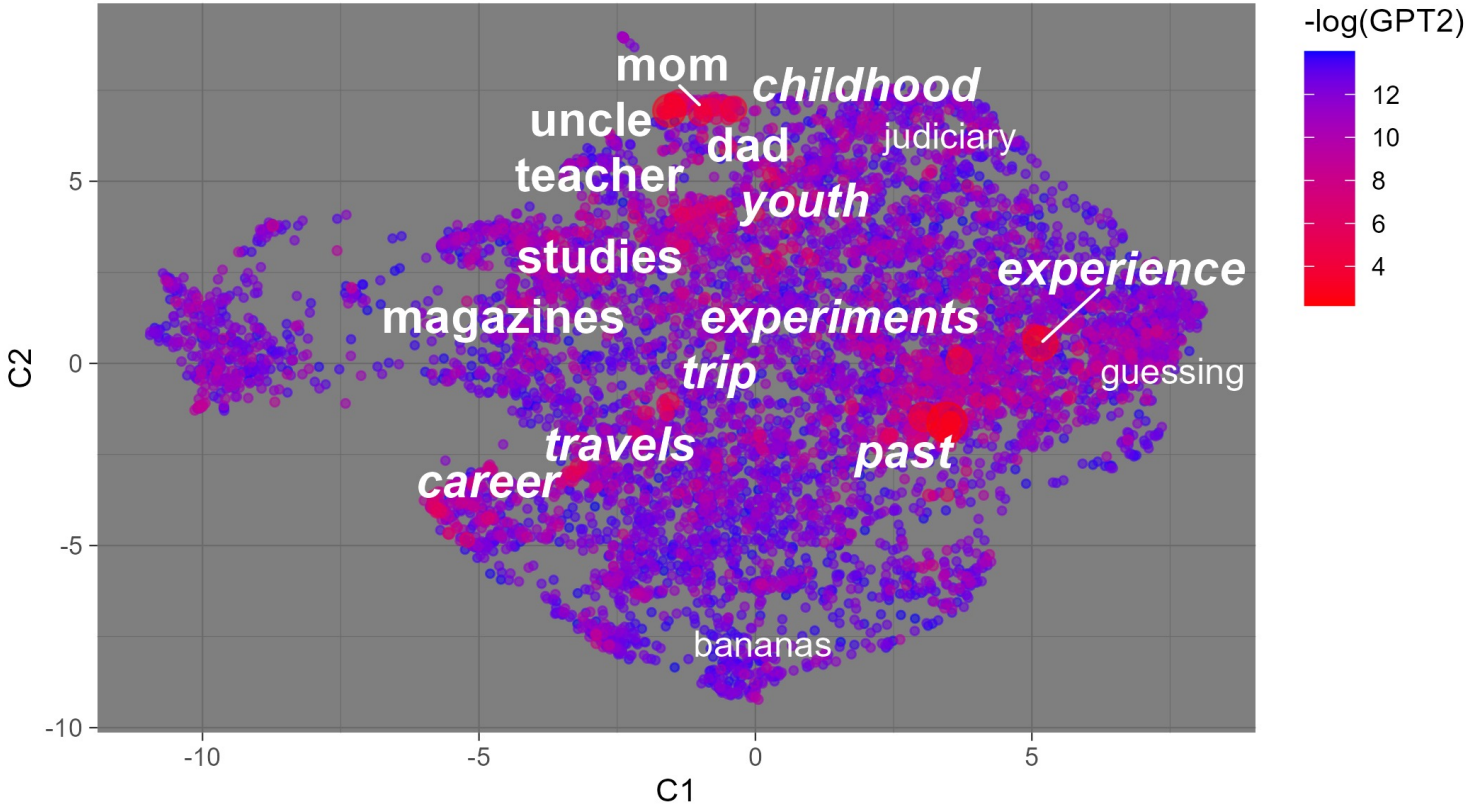
- The difficulty of comprehending a word is related to its predictability in context.

- Computational accounts of predictability
  - Predictability reflects Bayesian updating of a prior probability distribution over lexical items, raising or lowering the probability of any lexical item given input (Kuperberg & Jaeger, 2016)
  - Likely a form of probabilistic inference, e.g. Surprisal Theory (Shain et al 2024)

- But such accounts face computational complexity (Kwisthout & van Rooij, 2020)
  - Can be computationally intractable to calculate Bayes' Rule.
  - Especially for large and complex space.

Word predictability – a complex space!

- Jordan learned a lot about cars from his…



Word embedding: GLoVe 50 filtered for the 30,000 most common English words, UMAP to 2 dimensions for visualization.

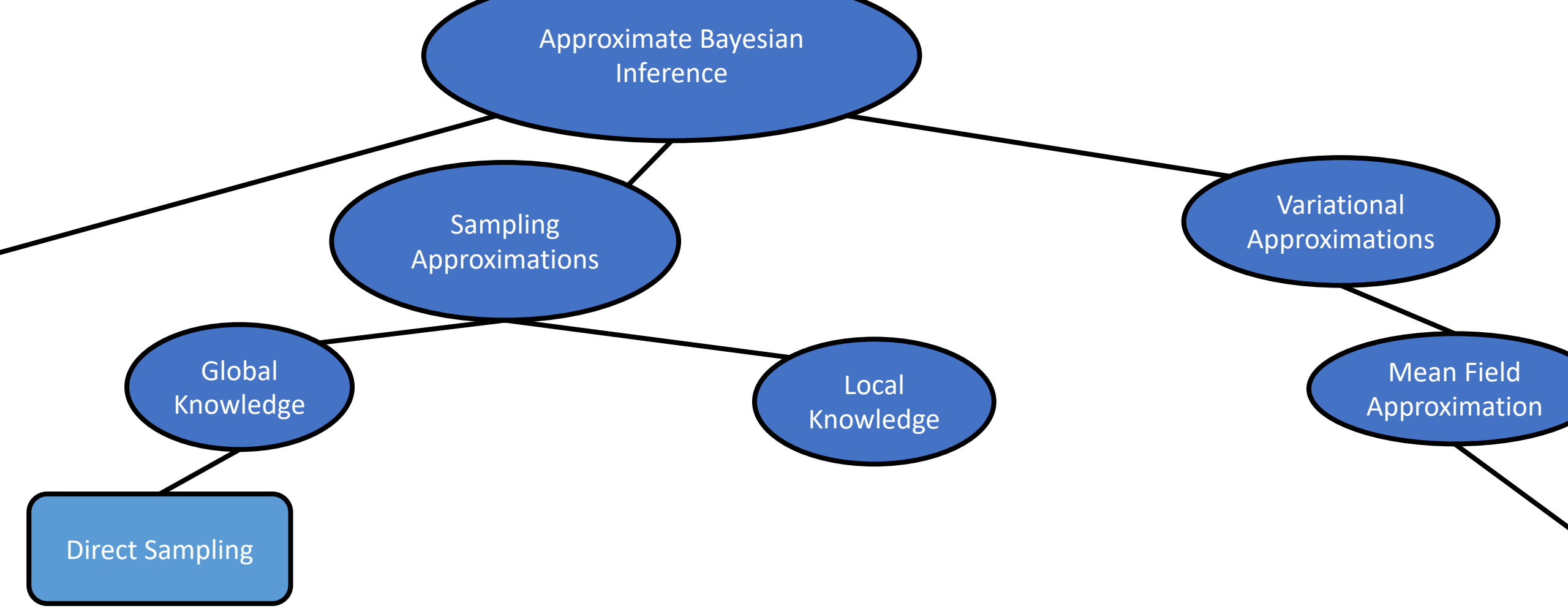
Surprisal calculated from GPT-2

A Puzzle

- Comprehenders appear to be Bayesian about word predictability.

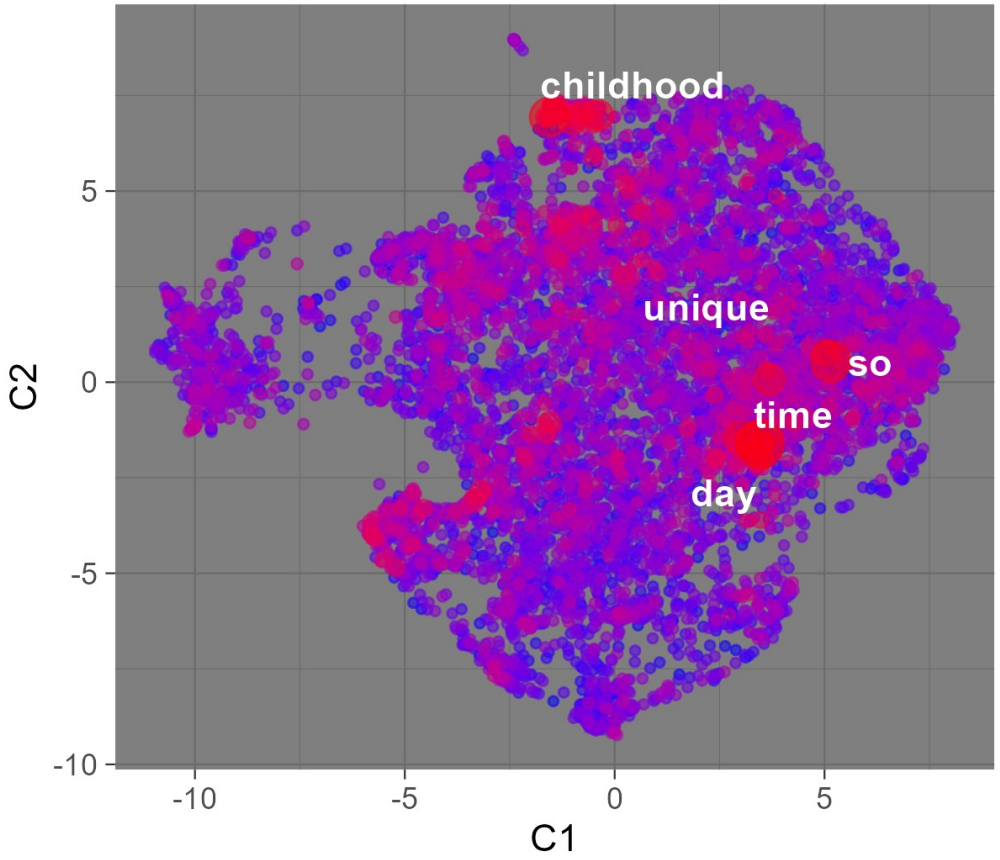
- BUT! Probabilistic (Bayesian) inference is likely to not be tractable in real time given the limits of our cognitive resources.

- The Question: How then does comprehension make use of probabilistic inference for word predictability if such inferences not tractable in real time under resource limitations?
  - Responding to this question seems likely to open up new avenues for research.


Approximate Bayesian Inference
Sampling Approximations
Variational Approximations
Global Knowledge
Local Knowledge
Mean Field Approximation
Direct Sampling

Direct Sampling

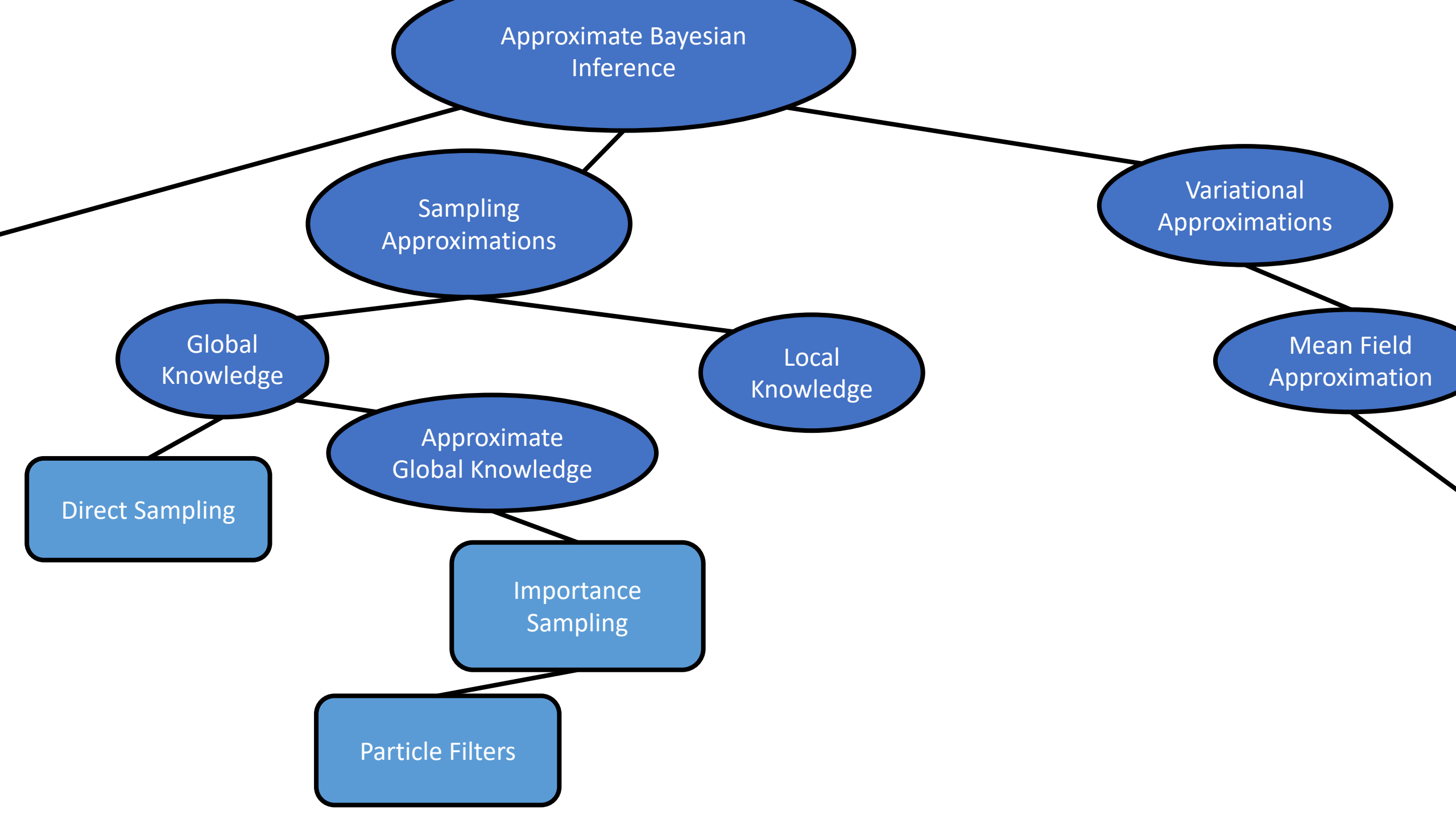
Jordan learned a lot about cars from his…



childhood
unique
so
time
day

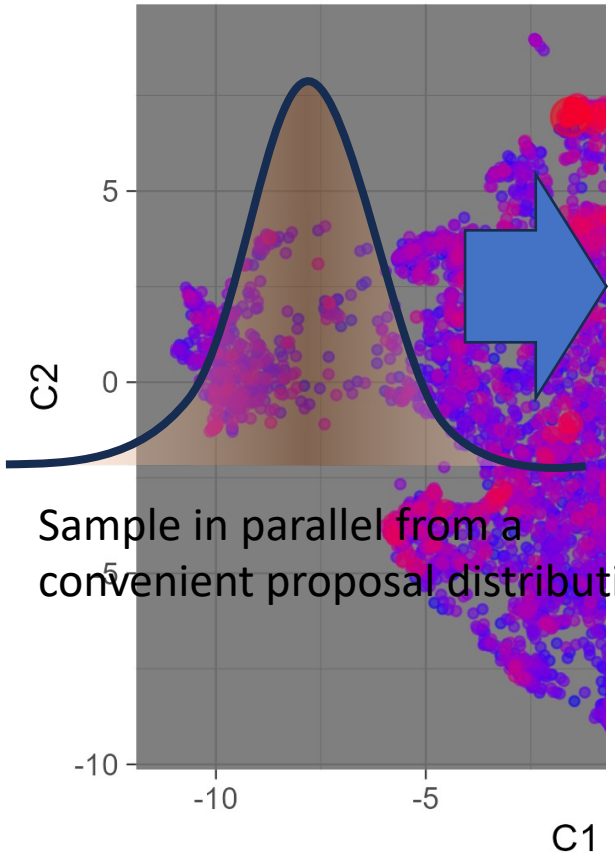
Average surprisal
Average location
Average cosine similarity to target
…

Sample in parallel from the posterior by probability

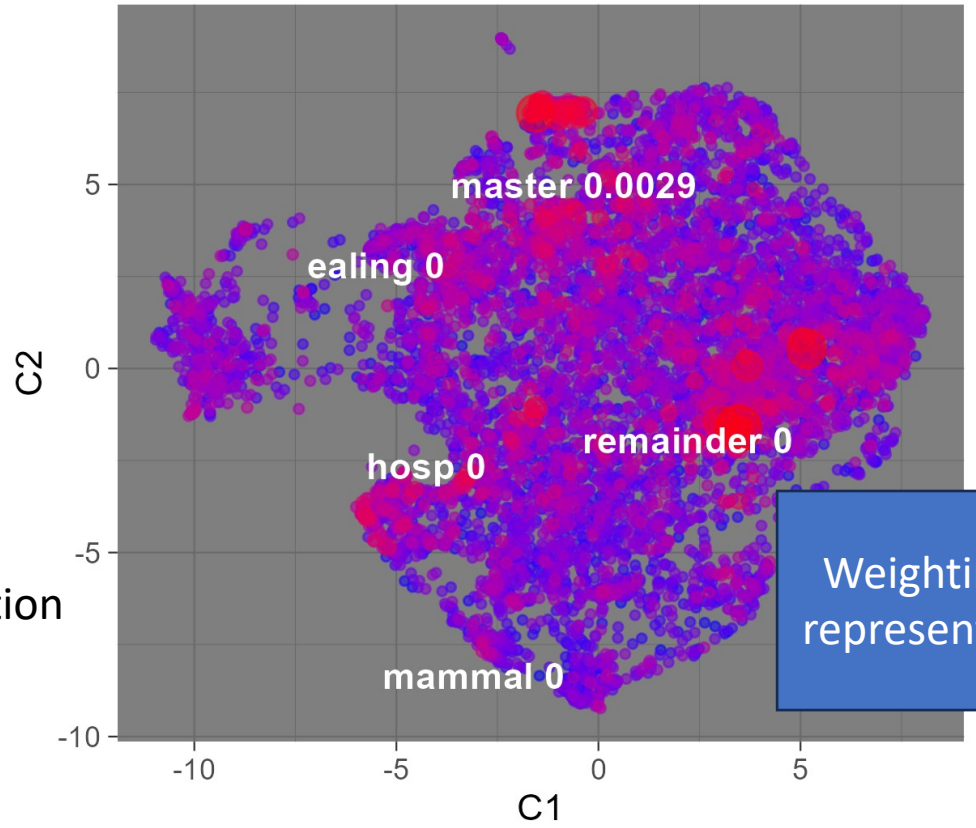

Approximate Bayesian Inference
Sampling Approximations
Variational Approximations
Global Knowledge
Local Knowledge
Mean Field Approximation
Direct Sampling
Approximate Global Knowledge
Importance Sampling
Particle Filters

Importance Sampling

Jordan learned a lot about cars from his…



Sample in parallel from a convenient proposal distribution

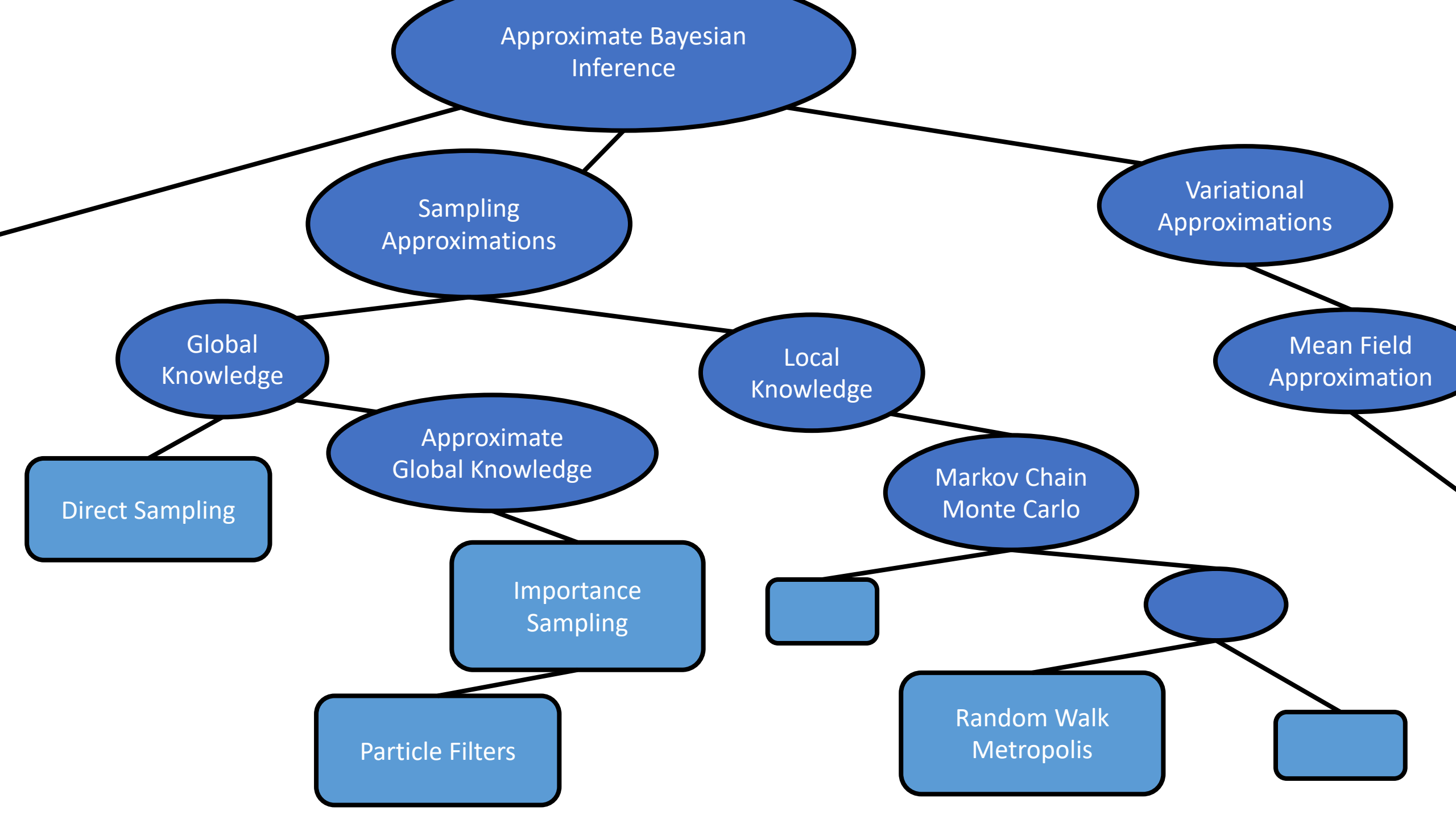


Weight samples by differences between proposal probability and posterior probability

master	0.0029
ealing	0.000
hosp	0.000
mammal	0.000
remainder	0.000

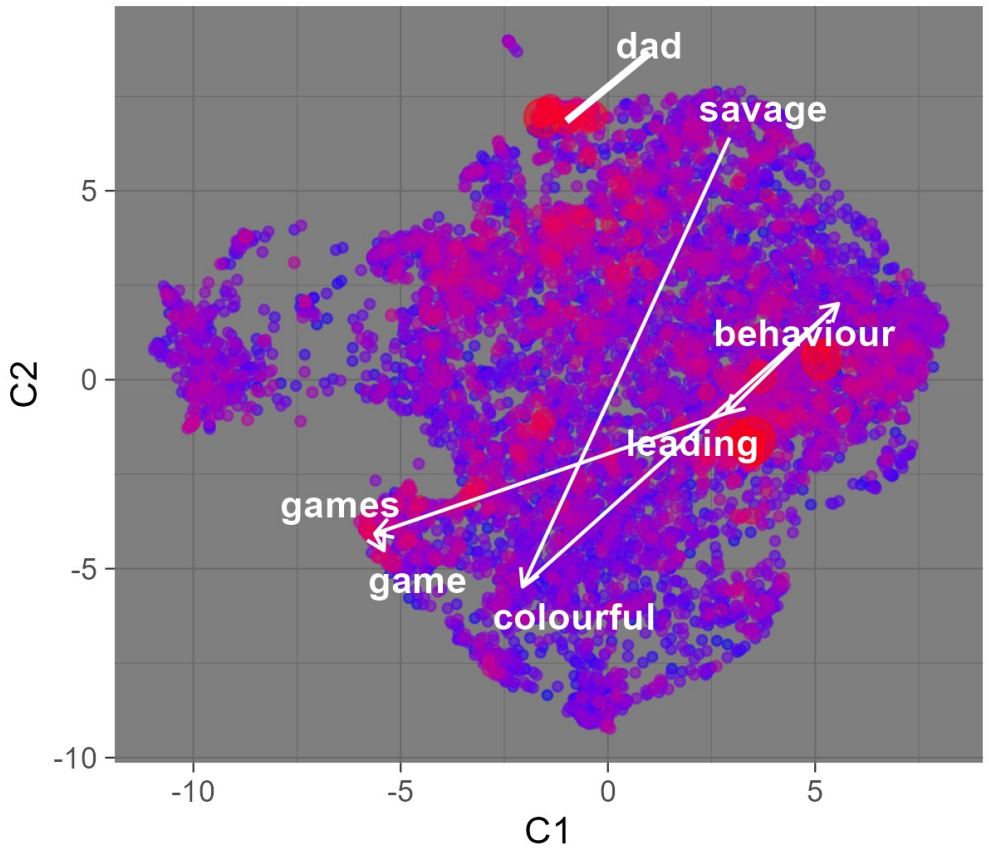
Weighting tends to over-represent typical examples

verage surprisal
Average location
Average cosine similarity to target
…


Approximate Bayesian Inference
Sampling Approximations
Variational Approximations
Global Knowledge
Local Knowledge
Mean Field Approximation
Direct Sampling
Approximate Global Knowledge
Markov Chain Monte Carlo
Importance Sampling
Particle Filters
Random Walk Metropolis

MCMC Sampling

Jordan learned a lot about cars from his…



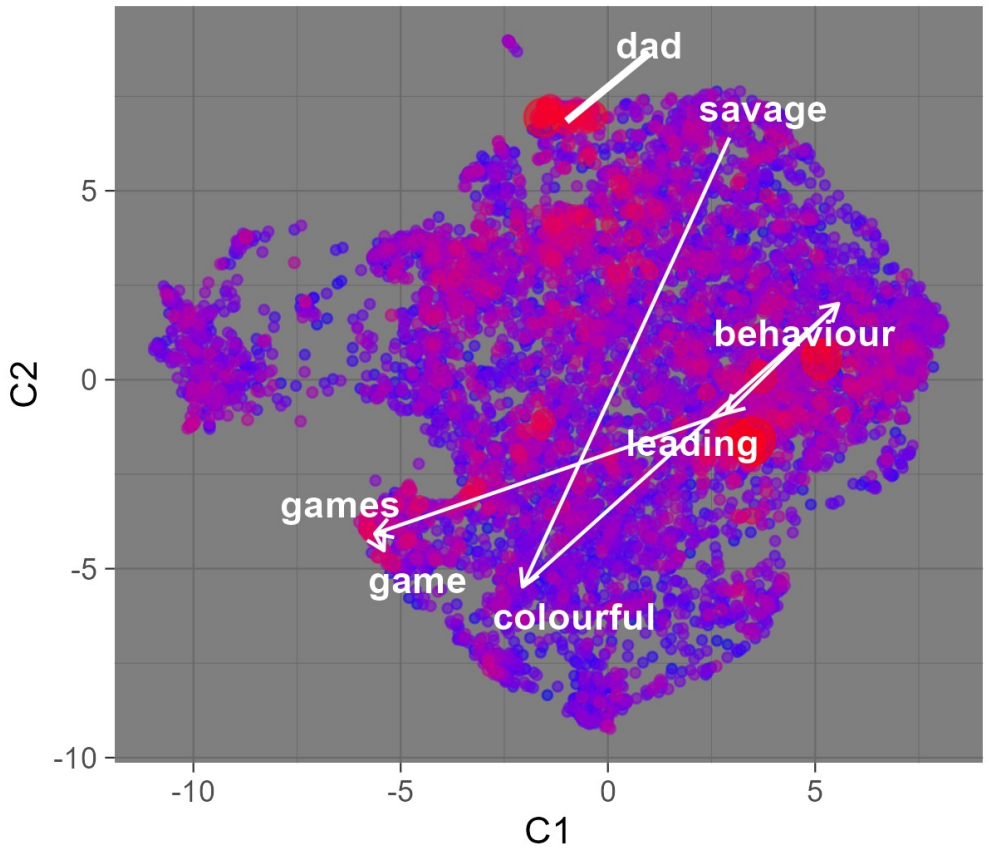
savage 1
colourful 2
behaviour 3
leading 4
games 5
game 6

Average surprisal
Average location
Average cosine similarity to target
…

Initiate a chain with a sample. Propose a new sample; accept/reject that sample given its relative probability to the current sample in the chain

MCMC Sampling

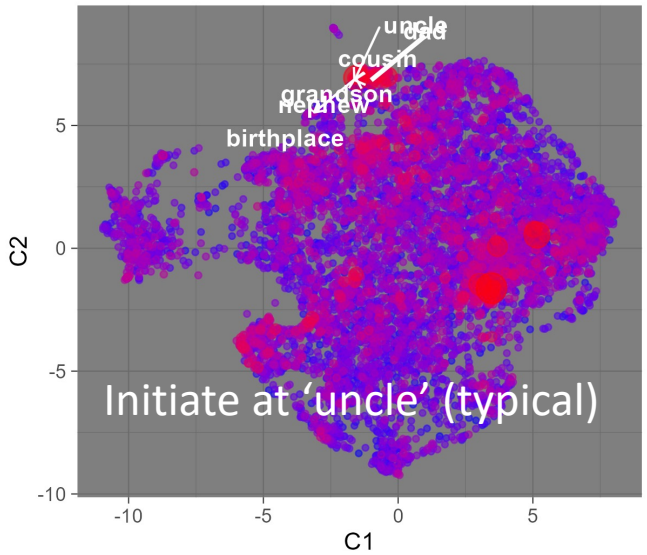
Jordan learned a lot about cars from his…

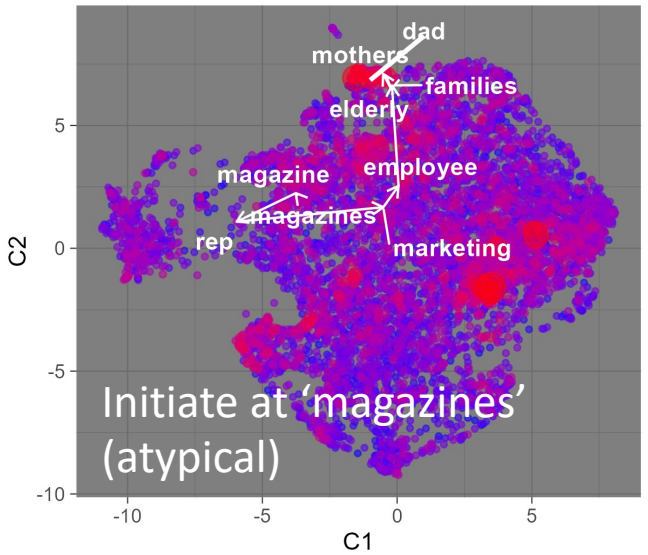


savage 1
colourful 2
behaviour 3
leading 4
games 5
game 6

Average surprisal
Average location
Average cosine similarity to target
…

Initiate a chain with a sample. Propose a new sample; accept/reject that sample given its relative probability to the current sample in the chain





Sampling algorithm	Knowledge	Sytematic Bias	Areas of explanation
Direct Sampling	Global	Stochastic behavior	Probability matching be… 2014); Exploration–exp… (Gershman, 2018; Speeke… …nk & Konstantinidis, 2015)
Importance Sampling	Approx. Global	Subadditivity (Over-represents typical samples)	Perceptual stimuli reproduc… memory (Shi et al., 2010); S… dependence and working m… in category learning (Lloyd et a… …2019; Sanborn et al., 2010)
MCMC Sampling	Local	Subadditivity and Superadditivity (Over-represents /under-represent typical examples when initiated with a typical/atypical example)	Anchoring bias (Lieder et al., 2018); Biases in probability judgments (Dasgupta et al., 2017; Sanborn & Chater, 2016);

Hoover, et al. (2023) *Open Mind*

Levy, Reali, & Griffiths (2008) *Adv. In NeuroIPS*

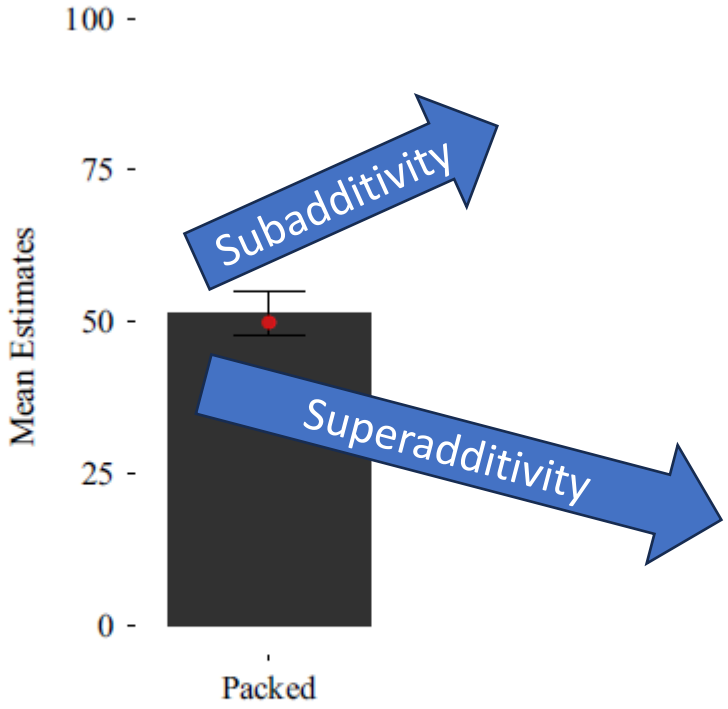
Question: Are there systematic biases in word predictability, and can we use it to narrow down the class of sampling algorithm that supports it?

Outside of language comprehension, the probabilistic inference computations underlying judgment and decision making in human cognition are known to be systematically biased.

Probability judgments and unpacking

- I see a table.

- What is the probability that I also see…
  - an object starting with the letter C? [packed]
  - a chair, computer, curtain, or any other object starting with the letter C? [unpacked – typical]
  - a cannon, cow, canoe, or any other object starting with the letter C? [unpacked – atypical]



Summary

- Unpacking context impacts probability judgments
  - Judgments are *subadditive* when unpacked examples are typical, leading to an increase in overall probability

    (Dasgupta et al 2017; Hadjchristidis et al 1999; Fox & Tversky 1998; Tversky & Koehler 1994).
  - Judgments are *superadditive* when unpacked examples are atypical, leading to a decrease in the overall probability judgment

    (Dasgupta et al 2017; Dougherty & Hunter 2003 ; Sloman et al 2004).

- Human Probability Judgment shows systematic biases.
  - Theses are proposed to arise from Approximate Probabilistic Inference algorithms e.g. consistent with the MCMC class (Dasgupta et al 2017).

Back to word predictability

- If comprehension is also a form of Approximate Probabilistic Inference, does it show biases similar to those found in judgment and decision making?

- If so, do those biases diagnose which class of sampling algorithms is a more appropriate fit for word predictability?

Methods

- 60 Participants; 120 Items; 3 conditions

- Range of target word cloze: 0.30 to 0.70 (Peelle et al. 2020)
- Unpacking: Typical unpacking was a high cloze word; atypical unpacking was a low cloze word

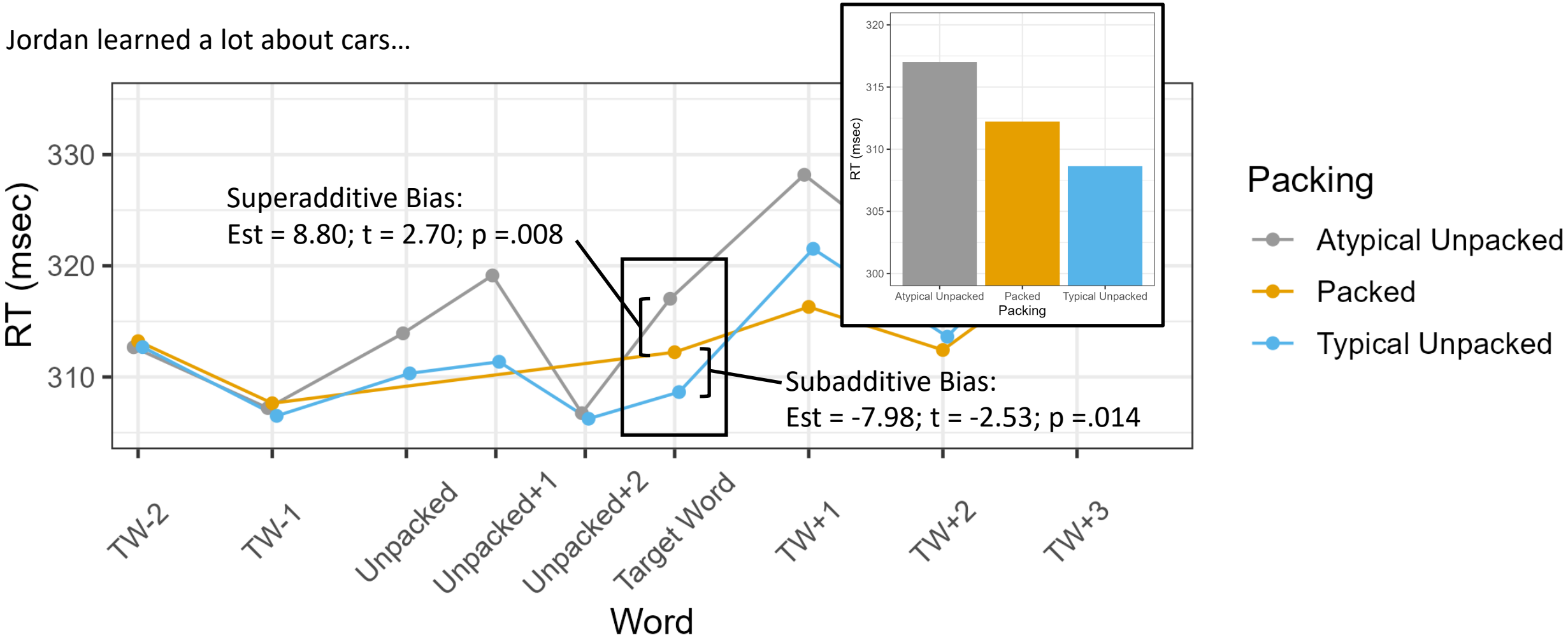
Cloze	Context	Target Word
0.30	The boy never accomplished anything since he lacked the	motivation to pursue his goals
0.51	Her friend said the story was nothing but	lies that could be seen through easily.
0.70	The bride spent weeks shopping for her	dress for the upcoming wedding.

- Also measured Word2Vec and LSA similarity between the unpacked word and target word.

Predictions

Bias	Description	Reading Time Prediction
Subadditivity	Perceived predictability of a word is *higher* when the context is unpacked to *typical* examples.	Faster than packed baseline
Superadditivity	Perceived predictability of a word is *lower* when the context is unpacked to *atypical* examples.	Slower than packed baseline

Jordan learned a lot about cars…



…from his uncle magazines and his dad during the last summer vacation.

Is this just priming?

- LSA between unpacked word and target word
  - Inclusion of centered LSA scores did not interact with Typicality
  - or reduce the effect of Typicality significantly.

	Est	t	p
Typicality	8.38	3.07	.003
LSA	0.78	0.06	.951
Typicality:LSA	8.83	0.42	.674

- Word2Vec between unpacked word and target word
  - Inclusion of centered word2vec scores reduced the significant of Typicality,
  - but did not significantly interact with Typicality and did not improve model fit ($\chi^2 = 1.203$; $p = .273$).
  - Typicality was still significant with word2vec as a covariate

	Est	t	p
Typicality	5.14	0.89	.375
w2v	16.70	1.53	.128
Typicality:w2v	23.01	1.09	.275

	Est	t	p
Typicality	10.40	3.23	.001
w2v	10.97	1.14	.254

What about surprisal?

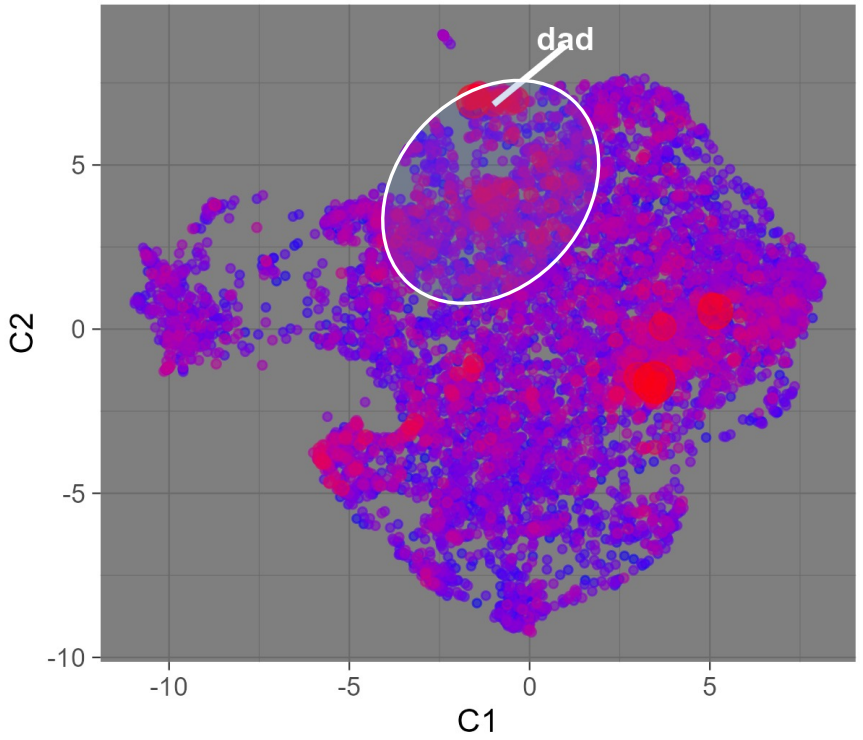
- Surprisal of the target word in each condition was measured with GPT-2.

- Surprisal did not interact with Unpacking Typicality
  - and did not improve model fit ($\chi^2 = 2.760$; $p = .252$)
  - Typicality remained significant with GPT-2 Surprisal as a covariate.

	Est	t	p
Typical	-9.46	-2.66	.009
Atypical	9.28	2.60	.011
GPT2	-4.63	-0.15	.882
Typical:GPT2	86.77	1.05	.300
Atypical:GPT2	-64.86	-0.63	.533

Unpacked Typical

Jordan learned a lot about cars from his…



uncle

Category includes uncle

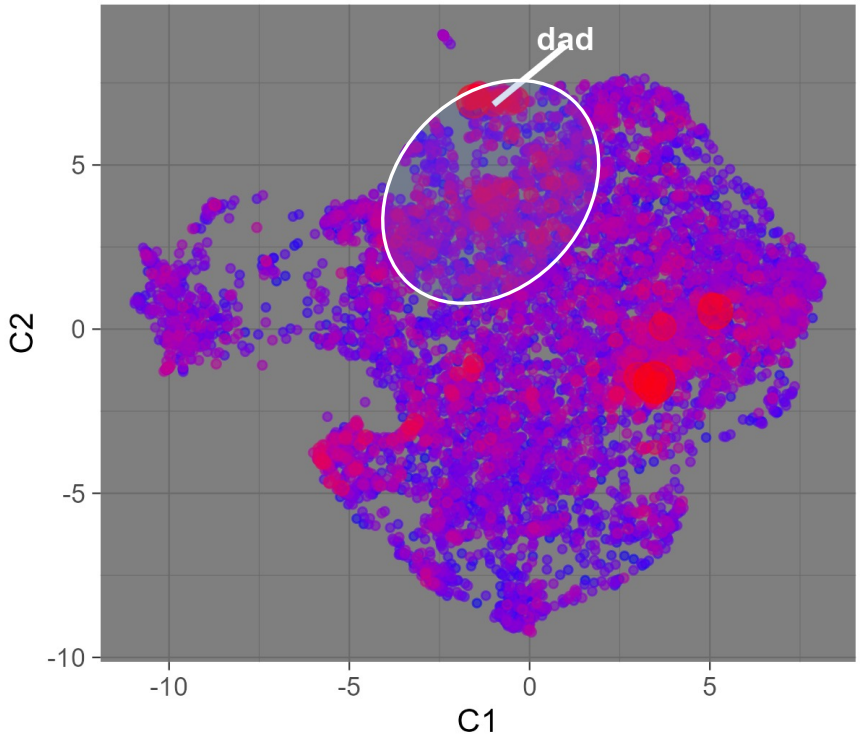
Expectation: Category of ways Jordan learned about cars

and his dad

Predictability of “dad” increases over baseline

Unpacked Atypical

Jordan learned a lot about cars from his…



magazines

Category includes magazines

Expectation: Category of ways Jordan learned about cars

and his dad

Predictability of “dad” decreases below baseline

Word predictability by MCMC?

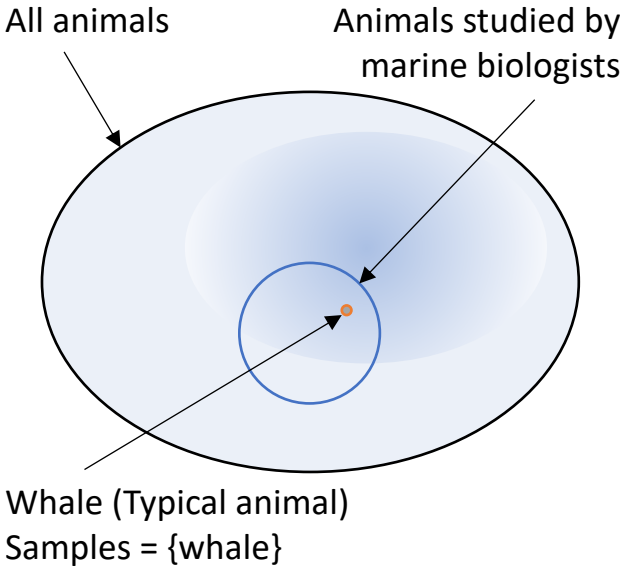
Why should language work this way?

- Local knowledge
  - Only the relative probabilities of the current sample and the proposal are needed (Sanborn & Chater 2016; Stewart et al 2006)
- Makes appropriate use of prior context to directly anchor a sample chain
  - Such chains might even be reused for further computation down stream (Dasgupta et al 2018)
- Computational efficient
  - MCMC sounds fancy, but it has a straightforward implementation and is even biologically plausible (Buesing et al 2011; Moreno-Bote et al., 2011; Pecevski et al 2011)

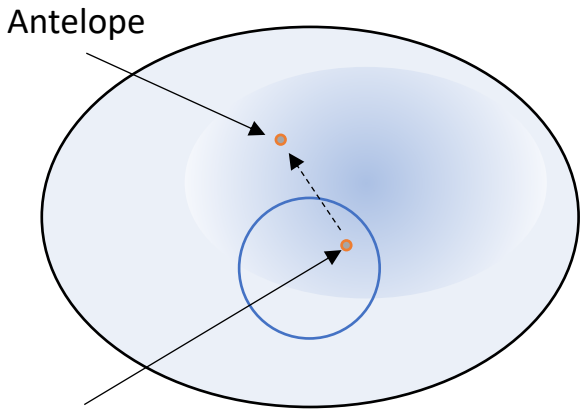
Conclusions

- Comprehenders may be Bayesian about word predictability.
  - But they are unlikely to be able to fully execute the computations involved due to resource limitations.
- Probabilistic inference can be approximated by a variety of algorithms
  - These algorithms have certain biases when resource limited.
- What biases do we see during human language comprehension?
  - Some evidence today that word predictability shows both subadditive and superadditive biases.
  - To the extent that sampling algorithms are appropriate models for cognition, Markov Chain Monte Carlo (MCMC) Sampling provides a better fit than Direct or Importance Sampling.

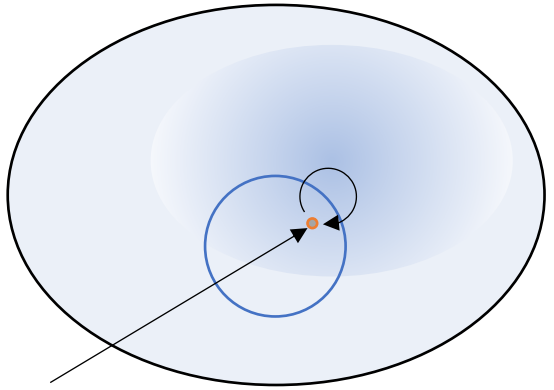
a) Different initialization points



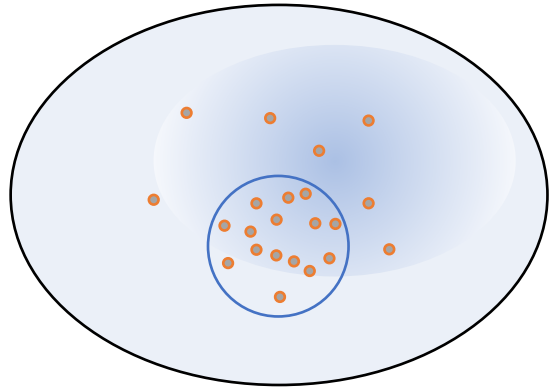
b) New proposal made

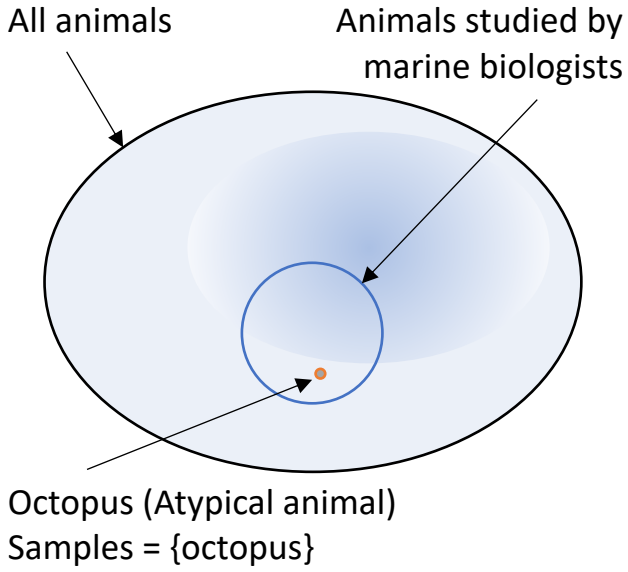


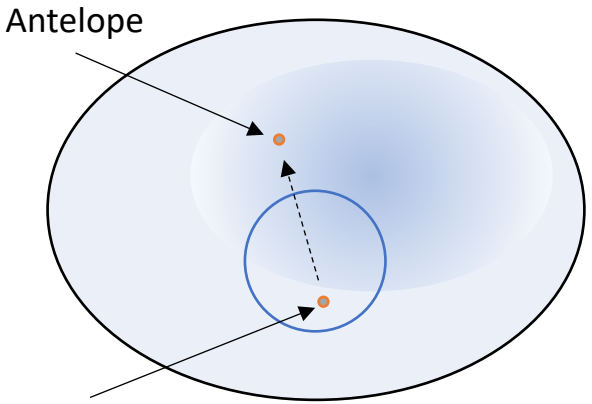
c) Proposal is rejected/accepted

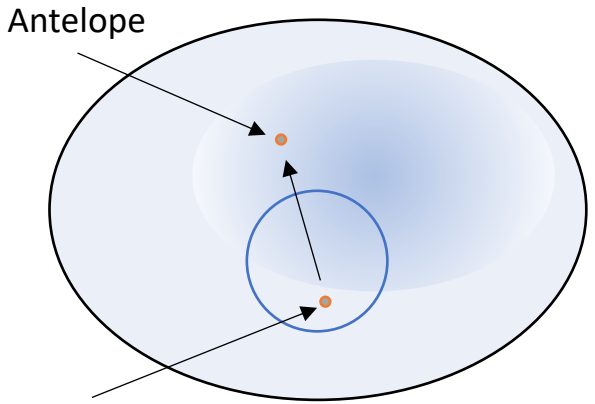


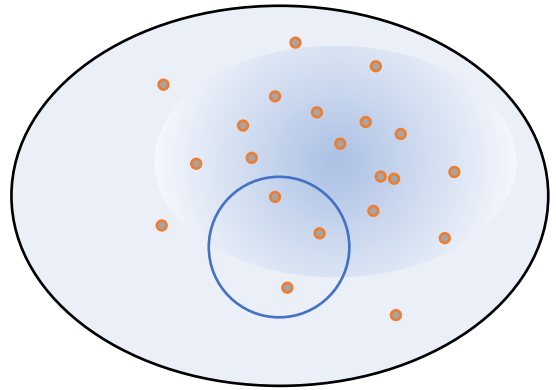
d) Over-/under-estimation



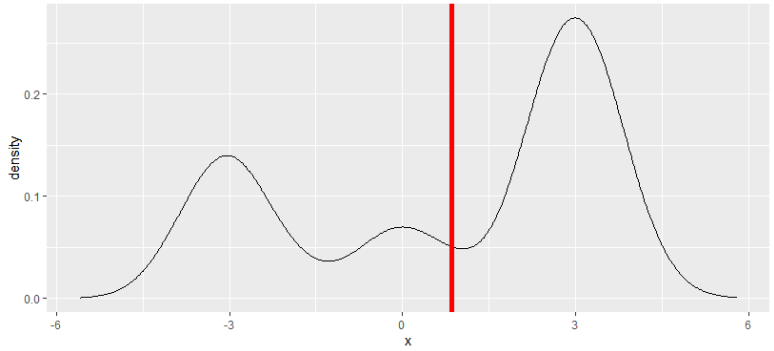


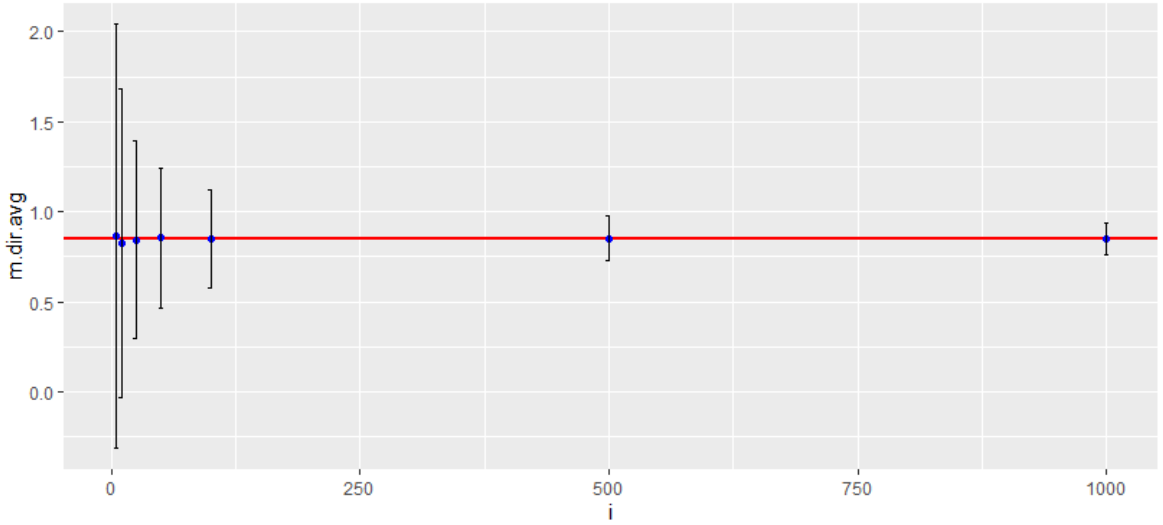






Direct Sampling

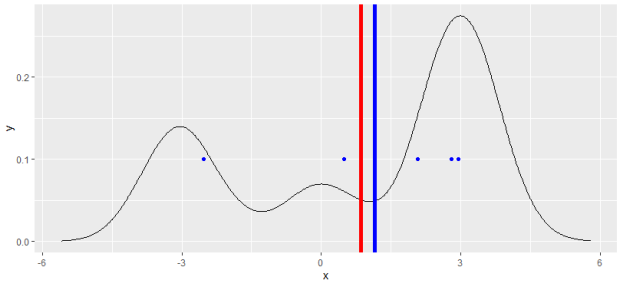




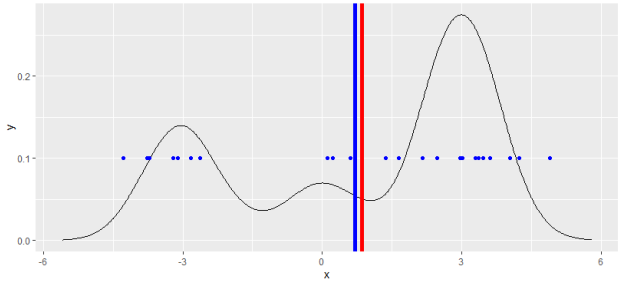
Take samples directly from the posterior distribution.

- Need to have global knowledge of the posterior
- No systematic bias with small numbers of samples – just added noise.

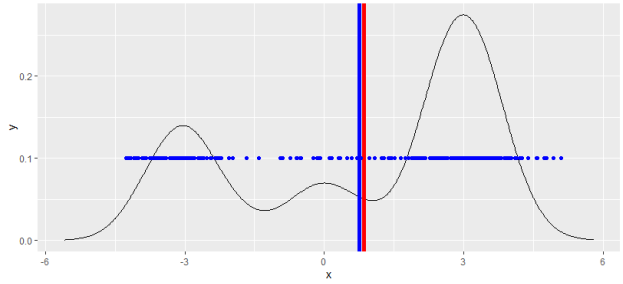
N=5



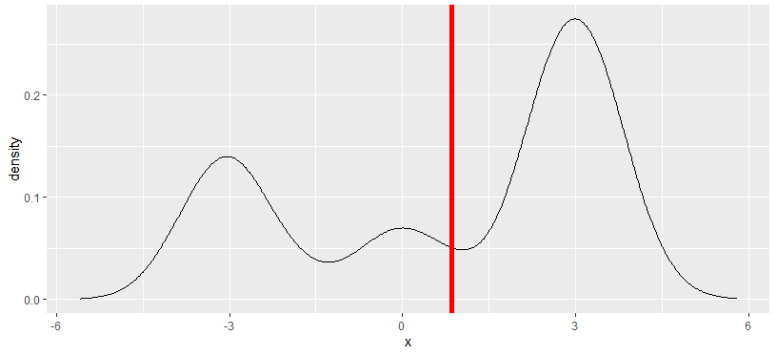
N=25

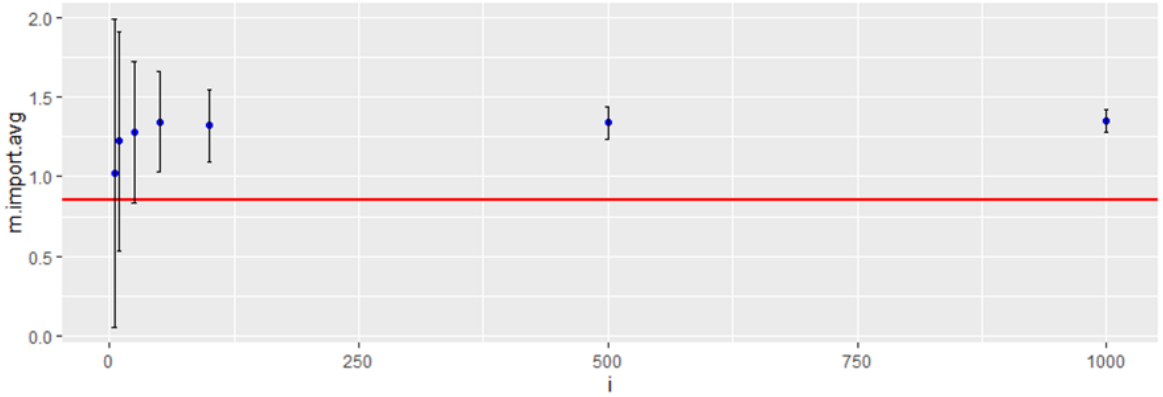


N=250



Importance Sampling

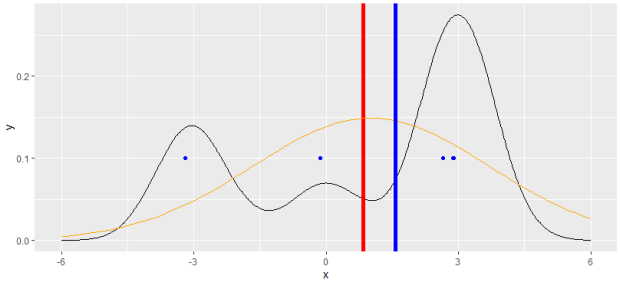




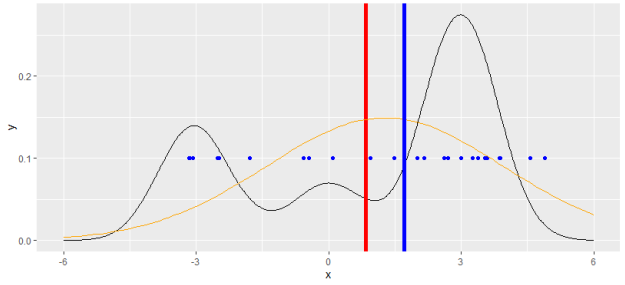
Take samples from the a proposal distribution that is (hopefully) related to the posterior distribution.

- Need to have approximate global knowledge of the posterior
- *Subadditivity*: Can be bias by more typical values.

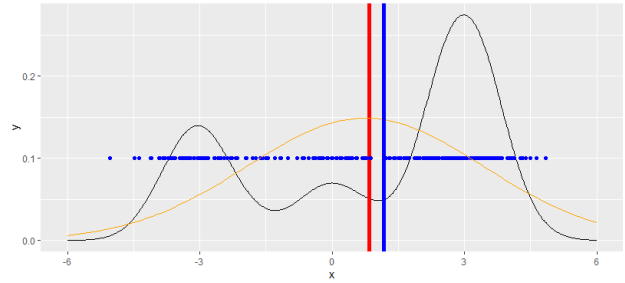
N=5



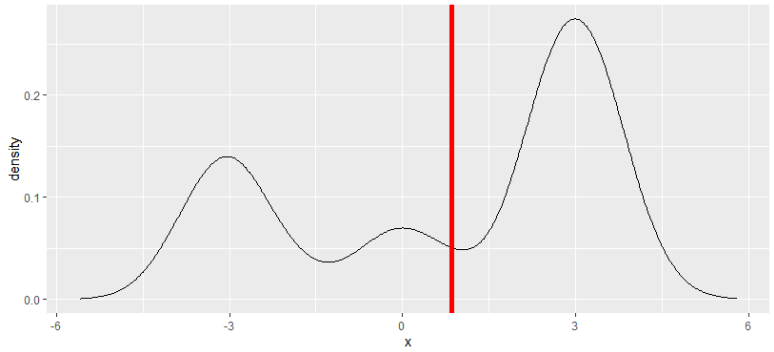
N=25

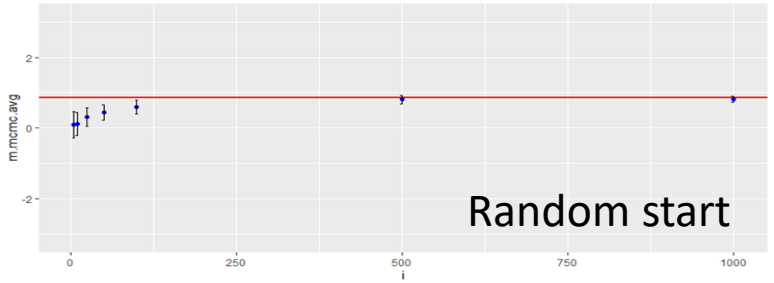


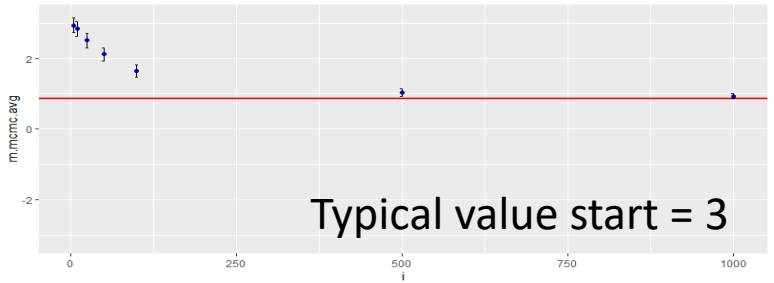
N=250

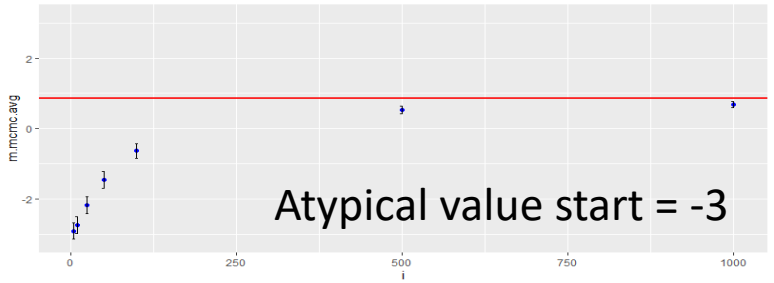


MCMC Sampling





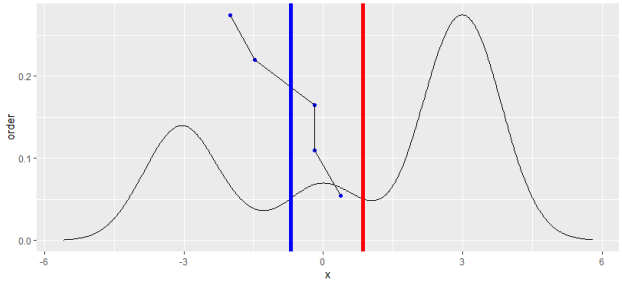




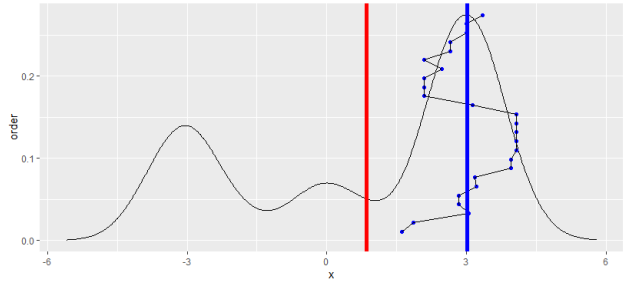
Initiate with one sample, propose a second, and compare their probabilities.

- Need to only have local knowledge of the posterior.
- *Subadditivity*: Can be bias by starting at a typical value.
- *Superadditivity*: Can be bias by starting at a more atypical value.

N=5



N=25



N=250