

What is Probabilistic Reasoning and why shall you care?

antonio vergari (he/him)



@tetradosse

4-5th Mar 2024 - Advanced Probabilistic Modeling - University of Trento

april

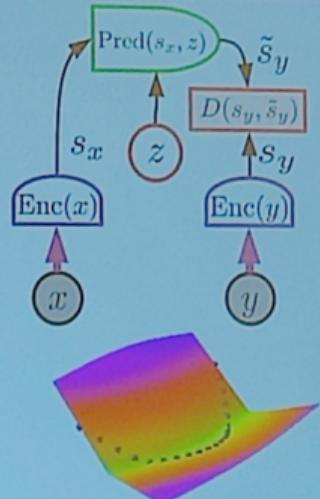
*april is
probably a
recursive
identifier of a
lab*

april

*about
probabilities
integrals &
logic*

Recommendations:

- ▶ Abandon generative models
 - ▶ in favor joint-embedding architectures
- ▶ Abandon probabilistic model
 - ▶ in favor of energy-based models
- ▶ Abandon contrastive methods
 - ▶ in favor of regularized methods
- ▶ Abandon Reinforcement Learning
 - ▶ In favor of model-predictive control
 - ▶ Use RL only when planning doesn't yield the predicted outcome, to adjust the world model or the critic.

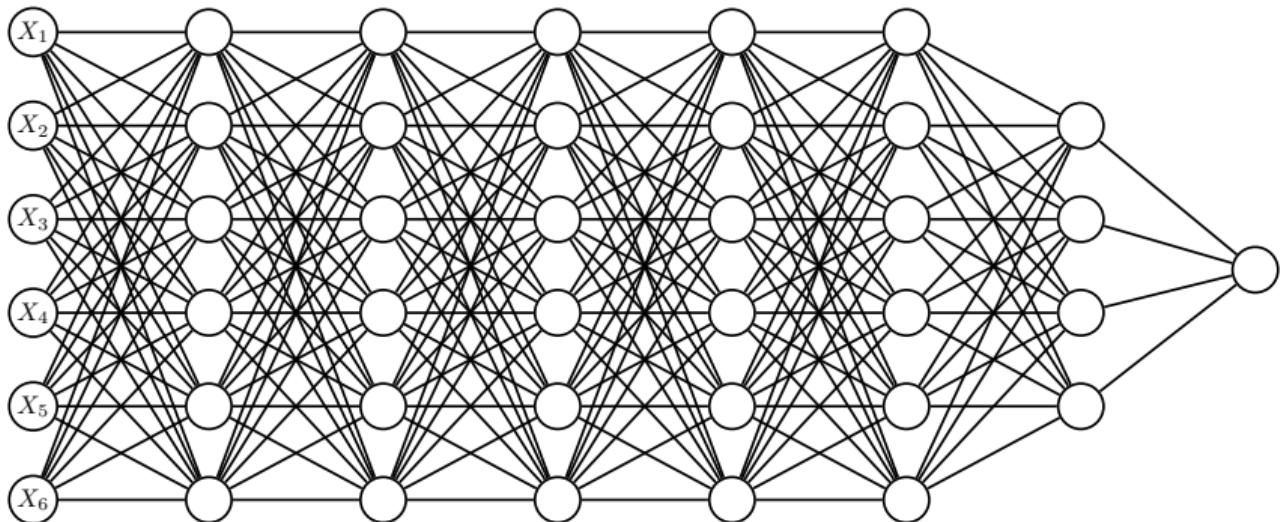


why shall we care about probabilities?

why *deep generative models*

Computational graphs that encode expressive distributions $p_{\mathbf{m}}(\mathbf{X})$

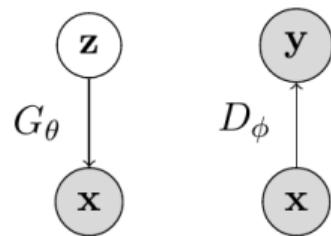
⇒ by stacking layers of “latent” units



why *deep generative models*

Computational graphs that encode expressive distributions $p_{\mathbf{m}}(\mathbf{X})$

⇒ by stacking layers of “latent” units



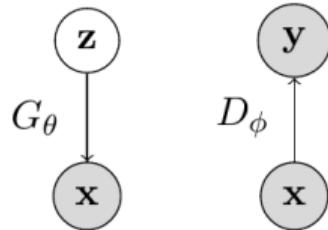
GANs

[Goodfellow et al. 2014]

why *deep generative models*

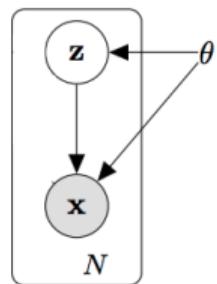
Computational graphs that encode expressive distributions $p_{\mathbf{m}}(\mathbf{X})$

\Rightarrow by stacking layers of “latent” units



GANs

[Goodfellow et al. 2014]



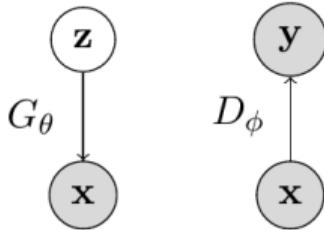
VAEs/Diffusion

[Kingma and Welling 2014]

why deep generative models

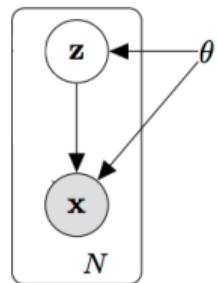
Computational graphs that encode expressive distributions $p_{\mathbf{m}}(\mathbf{X})$

\Rightarrow by stacking layers of “latent” units



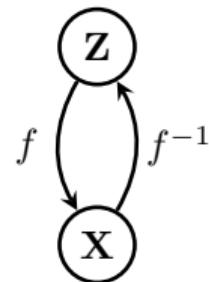
GANs

[Goodfellow et al. 2014]



VAEs/Diffusion

[Kingma and Welling 2014]



Cont/Disc Flows

[Papamakarios et al. 2019]

why *deep generative models* ...

[PROS] they encode *expressive* distributions with *millions of parameters*

+

they *scale learning* to high-dimensional, large datasets via *GPU*

...and why they lack ***complex reasoning***

[PROS] they encode *expressive* distributions with *millions of parameters*

+

they *scale learning* to high-dimensional, large datasets

[CONS] they are *limited to simple reasoning* routine (lack flexibility)

+

their computations come with *little or no guarantees* (unreliable)

Deep generative models

+

***flexible and reliable advanced probabilistic
reasoning?***

What is advanced reasoning?

or the inherent trade-off of tractability vs. expressiveness

wait!

the basics first

The “**product**” rule:

$$p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = p(\mathbf{X})p(\mathbf{Y}, \mathbf{Z} \mid \mathbf{X}) = p(\mathbf{X})p(\mathbf{Y} \mid \mathbf{X})p(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y})$$

wait!

the basics first

The “**product**” rule:

$$p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = p(\mathbf{X})p(\mathbf{Y}, \mathbf{Z} \mid \mathbf{X}) = p(\mathbf{X})p(\mathbf{Y} \mid \mathbf{X})p(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y})$$

The “**sum**” rule:

$$p(\mathbf{X}, \mathbf{Y}) = \text{???}$$

wait!

the basics first

The “**product**” rule:

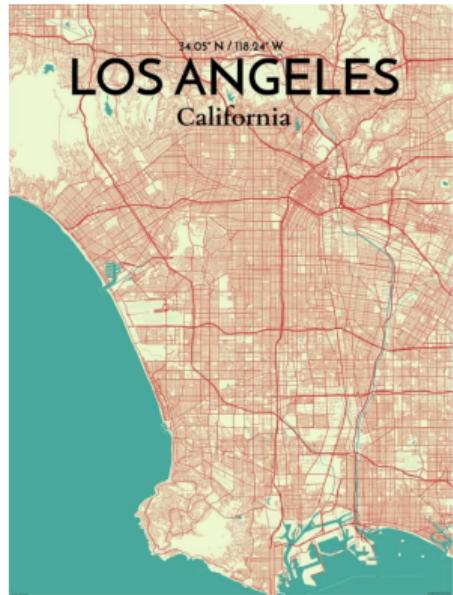
$$p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = p(\mathbf{X})p(\mathbf{Y}, \mathbf{Z} \mid \mathbf{X}) = p(\mathbf{X})p(\mathbf{Y} \mid \mathbf{X})p(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y})$$

The “**sum**” rule:

$$p(\mathbf{X}, \mathbf{Y}) = \int_{\text{dom}(\mathbf{z})} p(\mathbf{X}, \mathbf{Y}, \mathbf{z}) d\mathbf{Z}$$

Why probabilistic inference?

q₁: *What is the probability that today is a Monday and there is a traffic jam on Westwood Blvd.?*

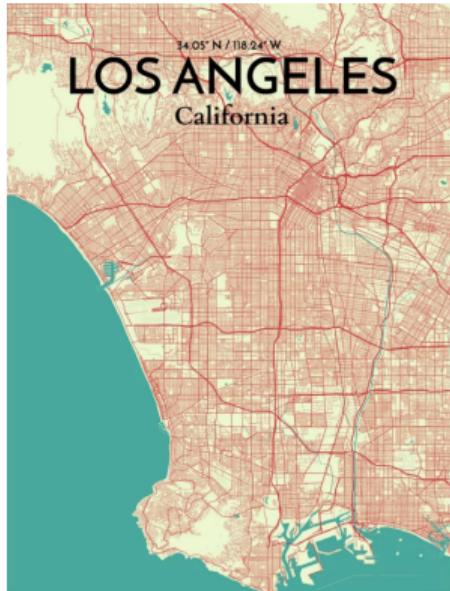


© fineartamerica.com

Why probabilistic inference?

q₁: *What is the probability that today is a Monday and there is a traffic jam on Westwood Blvd.?*

q₂: *Which day is most likely to have a traffic jam on my route to campus?*



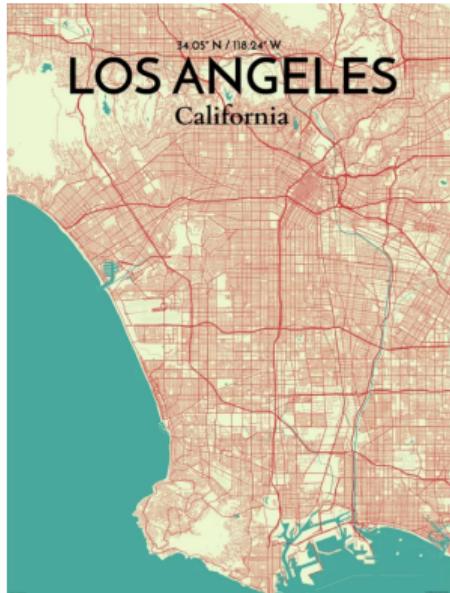
© fineartamerica.com

Why probabilistic inference?

q₁: *What is the probability that today is a Monday and there is a traffic jam on Westwood Blvd.?*

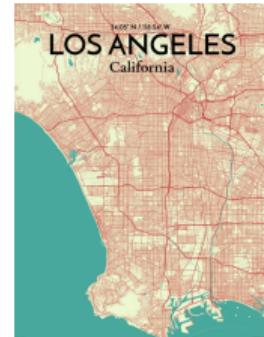
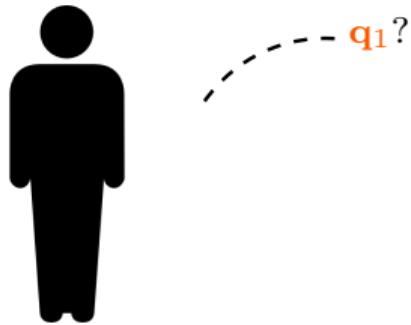
q₂: *Which day is most likely to have a traffic jam on my route to campus?*

How to answer several of these **probabilistic queries**?



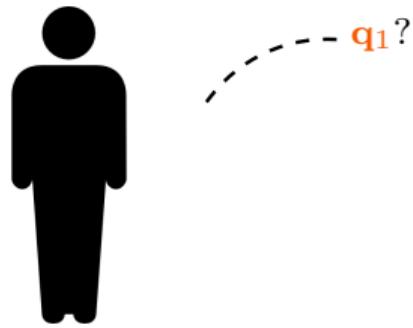
© fineartamerica.com

“What is the most likely street to have a traffic jam at 12.00?”



answering queries...

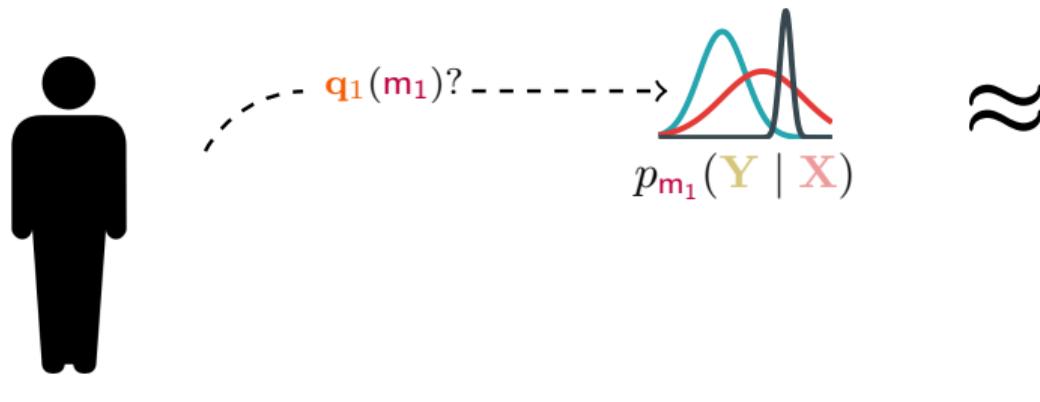
“What is the most likely street to have a traffic jam at 12.00?”



	X^1	X^2	X^3	X^4	X^5
x_8	■				
x_7					
x_6	■				
x_5	■				
x_4					
x_3	■				
x_2	■				
x_1	■	■	■	■	■

answering queries...

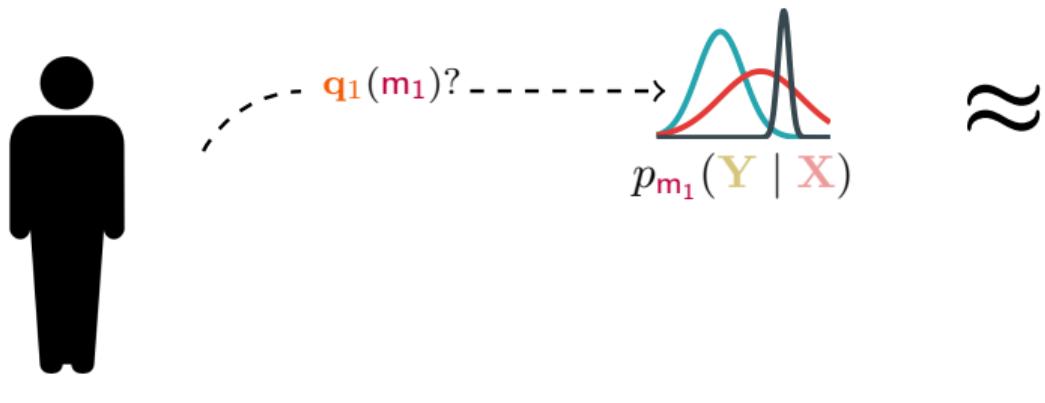
*"What is the most likely **street** to have a traffic jam at **12.00**?"*



	X^1	X^2	X^3	X^4	X^5
x_8	Red		Red		Yellow
x_7	Red		Red		Yellow
x_6	Red			Red	Yellow
x_5	Red		Red		Yellow
x_4	Red		Red		Yellow
x_3	Grey				Grey
x_2	Grey				Grey
x_1	Grey				Grey

...by fitting predictive models!

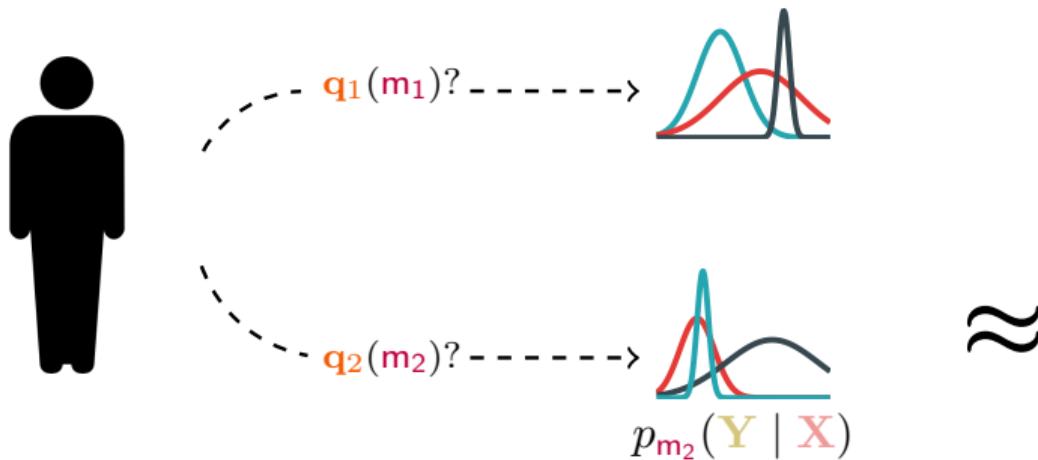
*“What is the most likely **street** to have a traffic jam at **12.00**?“*



	X^1	X^2	X^3	X^4	X^5
x_8	Red				Yellow
x_7	Red				Yellow
x_6	Red				Yellow
x_5	Red				Yellow
x_4	Red				Yellow
x_3	Grey				Grey
x_2	Grey				Grey
x_1	Grey				Grey

~~...by fitting predictive models!~~

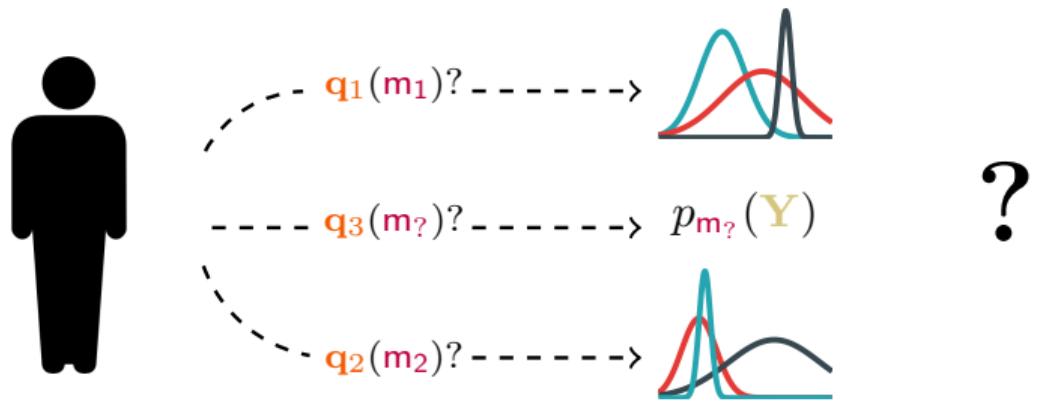
*"What is the most likely **time** to see a traffic jam at **Sunset Blvd.**?"*



	X^1	X^2	X^3	X^4	X^5
x_8					
x_7					
x_6	red	red	yellow	red	
x_5					
x_4	red	red	yellow	red	
x_3					
x_2					
x_1	red	red	yellow	red	

~~...by fitting predictive models!~~

"What is the probability of a traffic jam on Westwood Blvd. on Monday?"

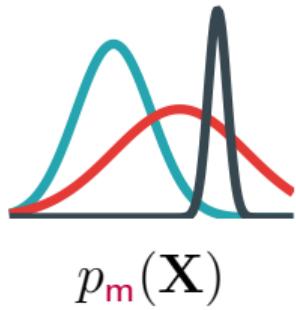


	X^1	X^2	X^3	X^4	X^5
x_8	Gold	Gold			
x_7	Gold	Gold			
x_6	Gold	Gold			
x_5	Gold				
x_4	Gold				
x_3	Gold	Gold			
x_2	Gold	Gold			
x_1	Gold	Gold			

~~...by fitting predictive models!~~



q₁(m)?
q₂(m)?
...
q_k(m)?



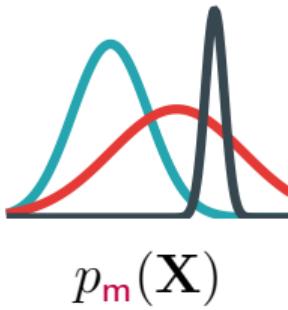
≈

	X ¹	X ²	X ³	X ⁴	X ⁵
x ₈					
x ₇					
x ₆					
x ₅					
x ₄					
x ₃					
x ₂					
x ₁					

...by fitting generative models!



q₁(m)?
q₂(m)?
...
q_k(m)?



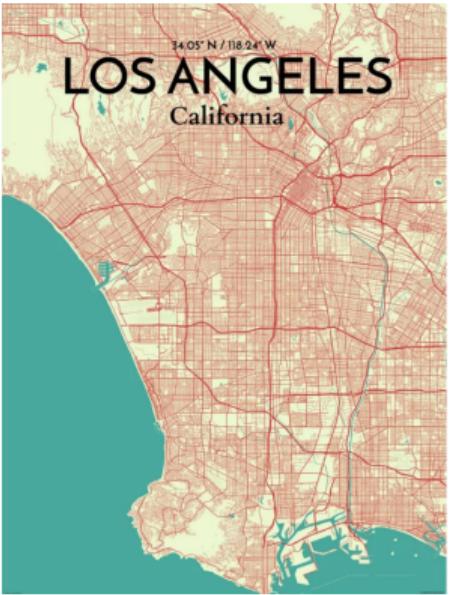
\approx

	X^1	X^2	X^3	X^4	X^5
x_8	Red				Green
x_7	Red	Red	Red		Green
x_6	Blue	Yellow	Yellow		Green
x_5	Blue	Blue	Blue		Green
x_4	Blue	Blue	Blue	Green	Green
x_3	Orange	Orange	Orange	Orange	Orange
x_2	Brown	Brown	Brown	Brown	Brown
x_1	Brown	Brown	Brown	Brown	Brown

...e.g. exploratory data analysis

Why probabilistic inference?

q₁: *What is the probability that today is a Monday and there is a traffic jam on Westwood Blvd.?*



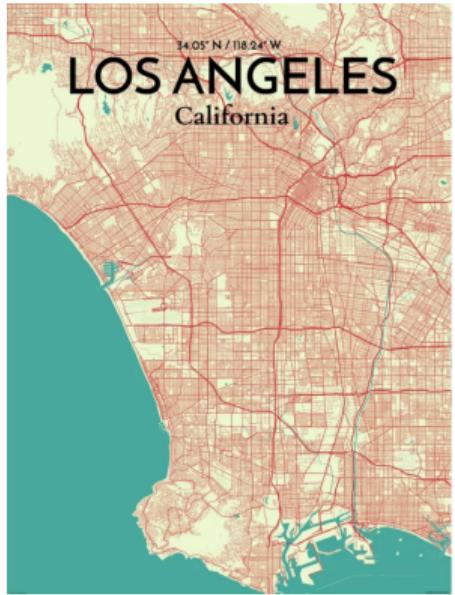
© fineartamerica.com

Why probabilistic inference?

q₁: What is the probability that today is a Monday and there is a traffic jam on Westwood Blvd.?

$$\mathbf{X} = \{\text{Day}, \text{Time}, \text{Jam}_{\text{Str1}}, \text{Jam}_{\text{Str2}}, \dots, \text{Jam}_{\text{StrN}}\}$$

$$\mathbf{q_1(m)} = p_{\mathbf{m}}(\text{Day} = \text{Mon}, \text{Jam}_{\text{Westwood}} = 1)$$



© fineartamerica.com

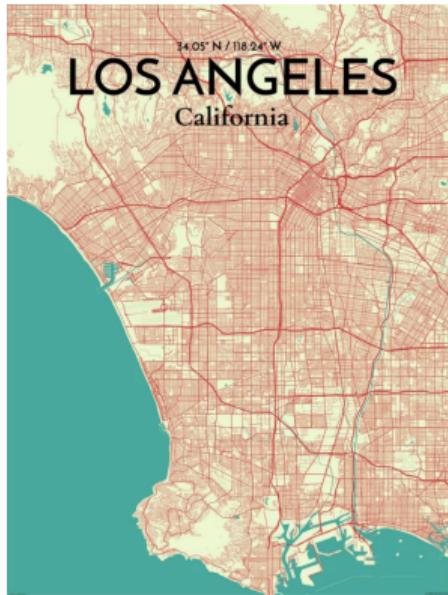
Why probabilistic inference?

q₁: What is the probability that today is a Monday and there is a traffic jam on Westwood Blvd.?

$$X = \{\text{Day}, \text{Time}, \text{Jam}_{\text{Str}1}, \text{Jam}_{\text{Str}2}, \dots, \text{Jam}_{\text{Str}N}\}$$

$$q_1(\mathbf{m}) = p_{\mathbf{m}}(\text{Day} = \text{Mon}, \text{Jam}_{\text{Westwood}} = 1)$$

⇒ **marginals** (the sum rule)



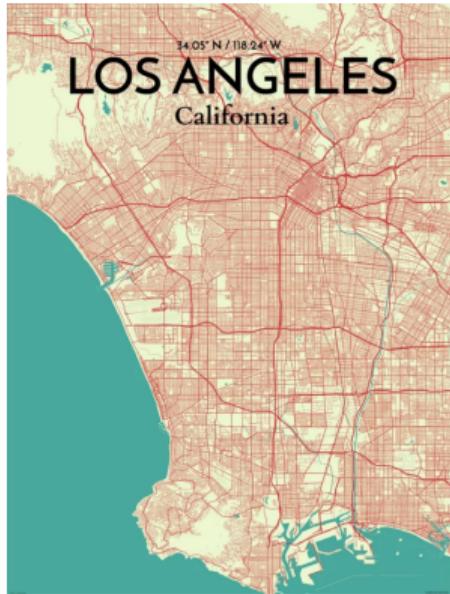
© fineartamerica.com

Why probabilistic inference?

q₂: Which day is most likely to have a traffic jam on my route to campus?

$$\mathbf{X} = \{\text{Day}, \text{Time}, \text{Jam}_{\text{Str1}}, \text{Jam}_{\text{Str2}}, \dots, \text{Jam}_{\text{StrN}}\}$$

$$\mathbf{q_2(m)} = \operatorname{argmax}_d p_{\mathbf{m}}(\text{Day} = d \wedge \bigvee_{i \in \text{route}} \text{Jam}_{\text{Str}i})$$



© fineartamerica.com

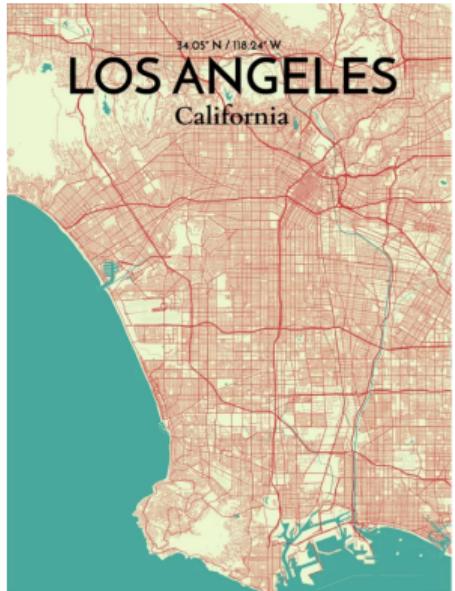
Why probabilistic inference?

q₂: Which day is most likely to have a traffic jam on my route to campus?

$$\mathbf{X} = \{\text{Day}, \text{Time}, \text{Jam}_{\text{Str1}}, \text{Jam}_{\text{Str2}}, \dots, \text{Jam}_{\text{StrN}}\}$$

$$\mathbf{q}_2(\mathbf{m}) = \operatorname{argmax}_d p_{\mathbf{m}}(\text{Day} = d \wedge \bigvee_{i \in \text{route}} \text{Jam}_{\text{Str}i})$$

⇒ *marginals + MAP + logical events*



© fineartamerica.com

Tractable Probabilistic Inference

A class of queries \mathcal{Q} is tractable on a family of probabilistic models \mathcal{M} iff for any query $\mathbf{q} \in \mathcal{Q}$ and model $\mathbf{m} \in \mathcal{M}$ exactly computing $\mathbf{q}(\mathbf{m})$ runs in time $O(\text{poly}(|\mathbf{m}|))$.

⇒ **model-centric** definition...

Tractable Probabilistic Inference

A class of queries \mathcal{Q} is tractable on a family of probabilistic models \mathcal{M} iff for any query $\mathbf{q} \in \mathcal{Q}$ and model $\mathbf{m} \in \mathcal{M}$ exactly computing $\mathbf{q}(\mathbf{m})$ runs in time $O(\text{poly}(|\mathbf{m}|))$.

\Rightarrow **model-centric** definition...

\Rightarrow ...and **query-centric**: Tractability is not a universal property!

Tractable Probabilistic Inference

A class of queries \mathcal{Q} is tractable on a family of probabilistic models \mathcal{M} iff for any query $\mathbf{q} \in \mathcal{Q}$ and model $\mathbf{m} \in \mathcal{M}$ exactly computing $\mathbf{q}(\mathbf{m})$ runs in time $O(\text{poly}(|\mathbf{m}|))$.

⇒ **model-centric** definition...

⇒ ...and **query-centric**: Tractability is not a universal property!

⇒ often poly will in fact be **linear**!

Tractable Probabilistic Inference

A class of queries \mathcal{Q} is tractable on a family of probabilistic models \mathcal{M} iff for any query $\mathbf{q} \in \mathcal{Q}$ and model $\mathbf{m} \in \mathcal{M}$ exactly computing $\mathbf{q}(\mathbf{m})$ runs in time $O(\text{poly}(|\mathbf{m}|))$.

⇒ **model-centric** definition...

⇒ ...and **query-centric**: Tractability is not a universal property!

⇒ often poly will in fact be **linear**!

⇒ Note: if $|\mathbf{m}| \in O(\text{poly}(|\mathbf{X}|))$, then query time is $O(\text{poly}(|\mathbf{X}|))$.

Tractable Probabilistic Inference

A class of queries \mathcal{Q} is tractable on a family of probabilistic models \mathcal{M} iff for any query $\mathbf{q} \in \mathcal{Q}$ and model $\mathbf{m} \in \mathcal{M}$ exactly computing $\mathbf{q}(\mathbf{m})$ runs in time $O(\text{poly}(|\mathbf{m}|))$.

⇒ **model-centric** definition...

⇒ ...and **query-centric**: Tractability is not a universal property!

⇒ often poly will in fact be **linear**!

⇒ Note: if $|\mathbf{m}| \in O(\text{poly}(|\mathbf{X}|))$, then query time is $O(\text{poly}(|\mathbf{X}|))$.

⇒ Why **exactness**? Highest guarantee possible!

why tractable models?

exactness can be crucial in safety-driven applications



guarantee constraint satisfaction

[Ahmed et al. 2022]



estimation error is bounded (0)

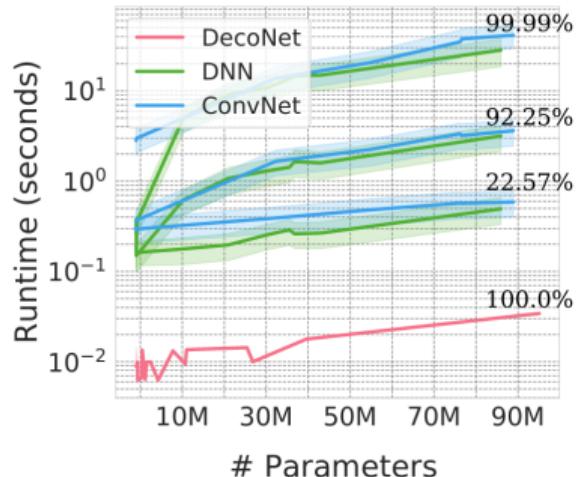
[Choi 2022]

why tractable models?

they can be much faster than intractable ones!

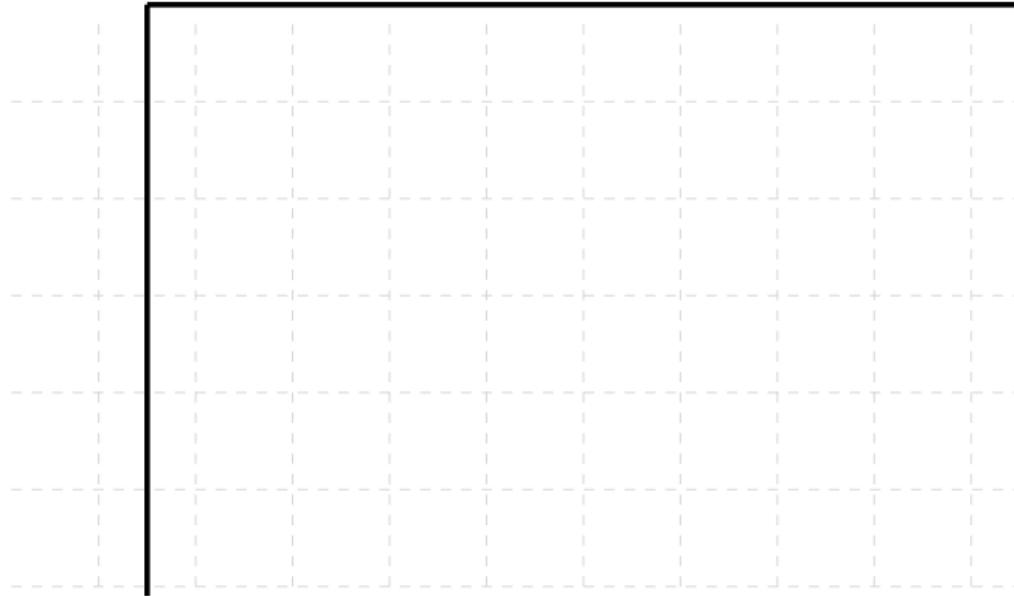
Method	MNIST (10,000 test images)		
	Theoretical bpd	Comp. bpd	En- & decoding time
PC (small)	1.26	1.30	53
PC (large)	1.20	1.24	168
IDF	1.90	1.96	880
BitSwap	1.27	1.31	904

[Liu, Mandt, and Broeck 2022]



[Subramani et al. 2021]

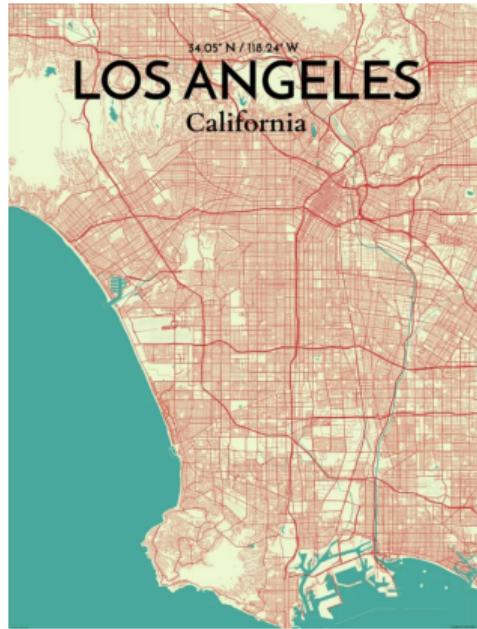
\mathcal{M} $\mathcal{Q}:$



tractable bands

Complete evidence (EVI)

q₃: *What is the probability that today is a Monday at 12.00 and there is a traffic jam only on Westwood Blvd.?*



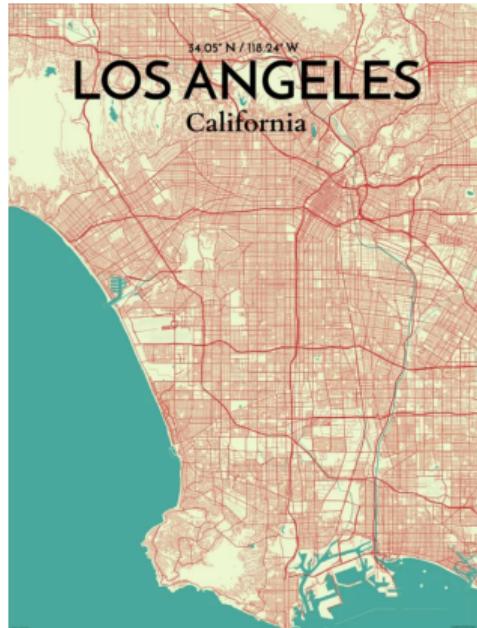
© fineartamerica.com

Complete evidence (EVI)

q₃: *What is the probability that today is a Monday at 12.00 and there is a traffic jam only on Westwood Blvd.?*

$$\mathbf{X} = \{\text{Day}, \text{Time}, \text{Jam}_{\text{Westwood}}, \text{Jam}_{\text{Str2}}, \dots, \text{Jam}_{\text{StrN}}\}$$

$$\mathbf{q}_3(\mathbf{m}) = p_{\mathbf{m}}(\mathbf{X} = \{\text{Mon}, 12.00, 1, 0, \dots, 0\})$$



© fineartamerica.com

Complete evidence (EVI)

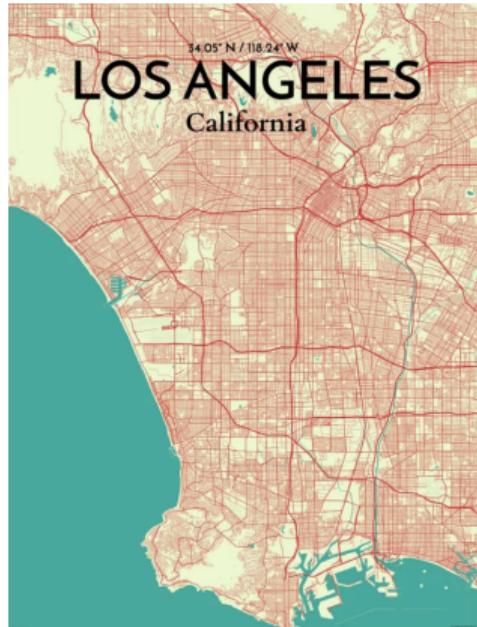
q₃: *What is the probability that today is a Monday at 12.00 and there is a traffic jam only on Westwood Blvd.?*

$$\mathbf{X} = \{\text{Day}, \text{Time}, \text{Jam}_{\text{Westwood}}, \text{Jam}_{\text{Str2}}, \dots, \text{Jam}_{\text{StrN}}\}$$

$$\mathbf{q}_3(\mathbf{m}) = p_{\mathbf{m}}(\mathbf{X} = \{\text{Mon}, 12.00, 1, 0, \dots, 0\})$$

...fundamental in ***maximum likelihood learning***

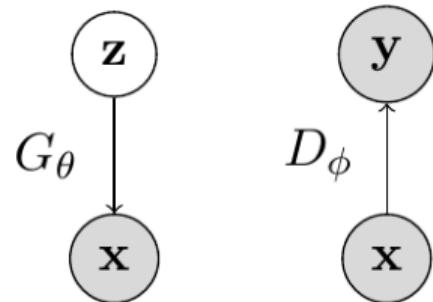
$$\theta_{\mathbf{m}}^{\text{MLE}} = \operatorname{argmax}_{\theta} \prod_{\mathbf{x} \in \mathcal{D}} p_{\mathbf{m}}(\mathbf{x}; \theta)$$



© fineartamerica.com

Generative Adversarial Networks

$$\min_{\theta} \max_{\phi} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D_{\phi}(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - D_{\phi}(G_{\theta}(\mathbf{z})))]$$



~~Generative Adversarial Networks~~

$$\min_{\theta} \max_{\phi} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D_{\phi}(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - D_{\phi}(G_{\theta}(\mathbf{z})))]$$

no explicit likelihood!

⇒ *adversarial training instead of MLE*

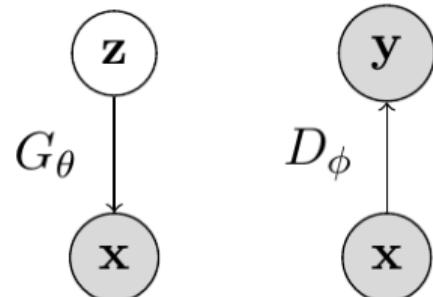
⇒ *no tractable EVI*

good sample quality

⇒ *but lots of samples needed for MC*

unstable training

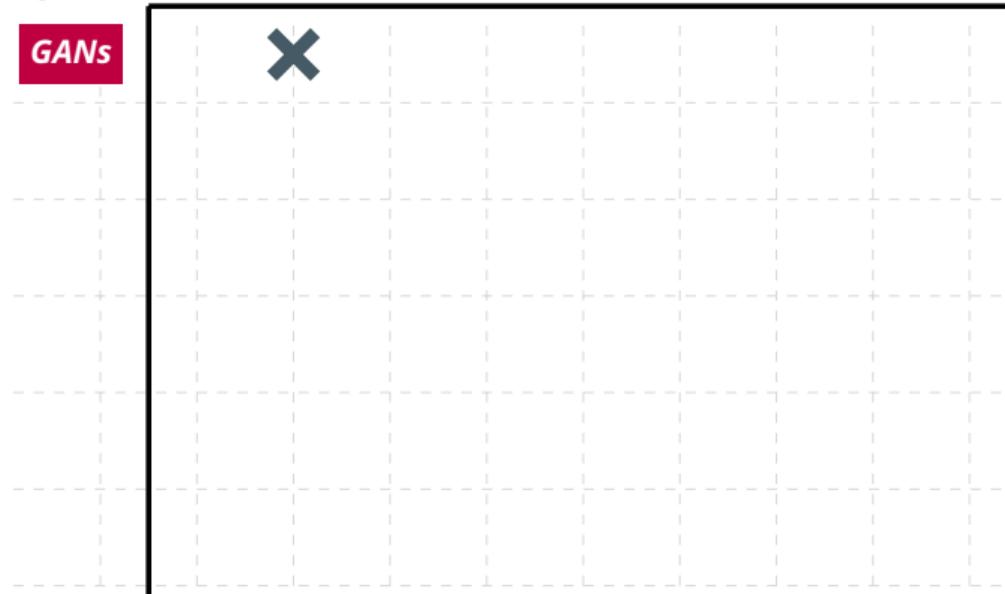
⇒ *mode collapse*



\mathcal{M}

$\mathcal{Q}:$ **EVI**

GANs

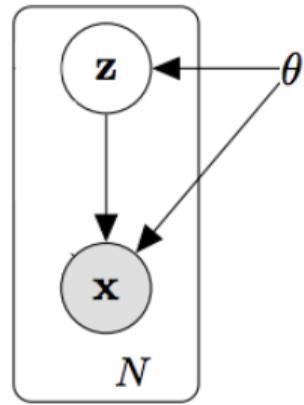


tractable bands

Variational Autoencoders

$$p_{\theta}(\mathbf{x}) = \int p_{\theta}(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})d\mathbf{z}$$

an explicit likelihood model!



Rezende, Mohamed, and Wierstra, "Stochastic backprop. and approximate inference in deep generative models", *arXiv preprint arXiv:1401.4082*, 2014

Kingma and Welling, "Auto-Encoding Variational Bayes",, 2014

Variational Autoencoders

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} | \mathbf{x})} [\log p_\theta(\mathbf{x} | \mathbf{z})] - \text{KL}(q_\phi(\mathbf{z} | \mathbf{x}) || p(\mathbf{z}))$$

an explicit likelihood model!

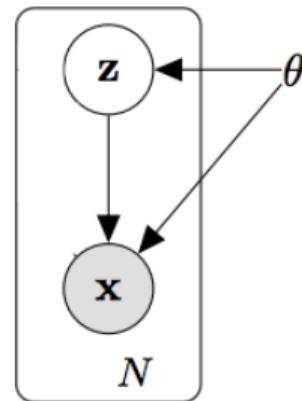
... but computing $\log p_\theta(\mathbf{x})$ is intractable

⇒ *an infinite and uncountable mixture*

⇒ *no tractable EVI*

we need to optimize the ELBO...

⇒ *which is “tricky” [Alemi et al. 2017; Dai and Wipf 2019; Ghosh et al. 2019]*



~~Energy Based Models~~

$$p_{\theta}(\mathbf{x}) = e^{-E(\mathbf{x}; \theta)} / Z$$

an explicit likelihood model!

... but computing the partition function

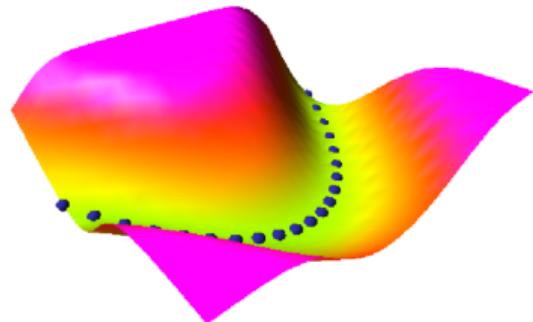
$Z = \int e^{-E(\mathbf{x}; \theta)} d\mathbf{X}$ is intractable

\Rightarrow no tractable EVI

alternative ways to train them than MLE...

\Rightarrow e.g., (denoising) score/ratio matching

[Song and Kingma 2021]



\mathcal{M}

$\mathcal{Q}:$

EVI

GANs



VAEs



tractable bands

Normalizing flows

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{Z}}(f^{-1}(\mathbf{x})) \left| \det \left(\frac{\delta f^{-1}}{\delta \mathbf{x}} \right) \right|$$

an explicit likelihood!

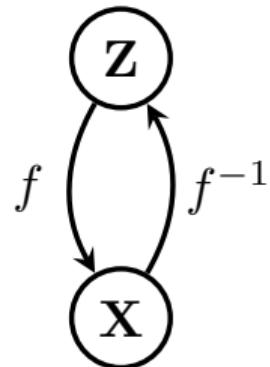
...plus structured Jacobians

\Rightarrow tractable EVI queries!

many neural variants

RealNVP [Dinh, Sohl-Dickstein, and Bengio 2016], MAF [Papamakarios, Pavlakou, and Murray 2017]

MADE [Germain et al. 2015],
PixelRNN [Oord, Kalchbrenner, and Kavukcuoglu 2016]



Normalizing flows

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{Z}}(f^{-1}(\mathbf{x})) \left| \det \left(\frac{\delta f^{-1}}{\delta \mathbf{x}} \right) \right|$$

an explicit likelihood!

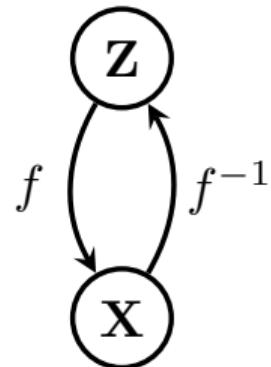
...plus structured Jacobians

⇒ *tractable EVI queries!*

many neural variants

RealNVP [Dinh, Sohl-Dickstein, and Bengio 2016], MAF [Papamakarios, Pavlakou, and Murray 2017]

MADE [Germain et al. 2015],
PixelRNN [Oord, Kalchbrenner, and Kavukcuoglu 2016]



Normalizing flows

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{Z}}(f^{-1}(\mathbf{x})) \left| \det \left(\frac{\delta f^{-1}}{\delta \mathbf{x}} \right) \right|$$

an explicit likelihood!

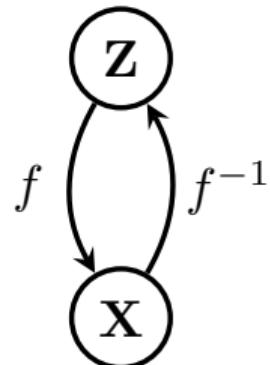
...plus structured Jacobians

⇒ *tractable EVI queries!*

many neural variants

RealNVP [Dinh, Sohl-Dickstein, and Bengio 2016], MAF [Papamakarios, Pavlakou, and Murray 2017]

MADE [Germain et al. 2015],
PixelRNN [Oord, Kalchbrenner, and Kavukcuoglu 2016]



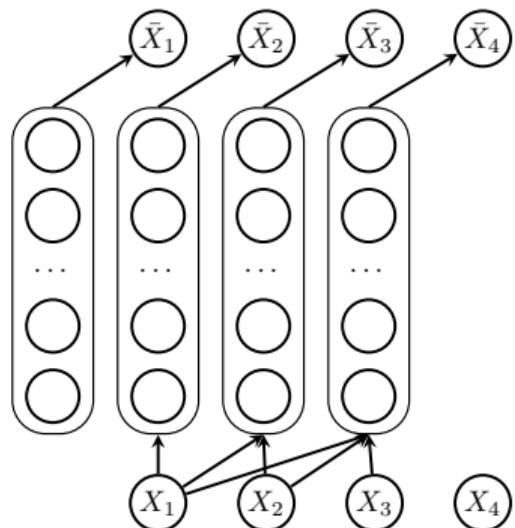
Autoregressive models

$$p_{\theta}(\mathbf{x}) = \prod_i p_{\theta}(x_i \mid \mathbf{x}_{<i})$$

an explicit likelihood model via the product rule
a special case of normalizing flows!

⇒ *the Zs are the noise used to sample*

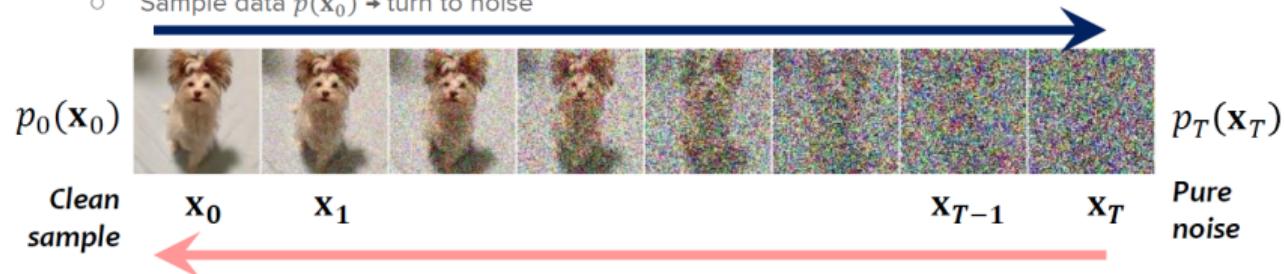
[Papamakarios et al. 2021]



~~Diffusion models~~

and continuous (time) normalizing flows

- Sample data $p(\mathbf{x}_0) \rightarrow$ turn to noise

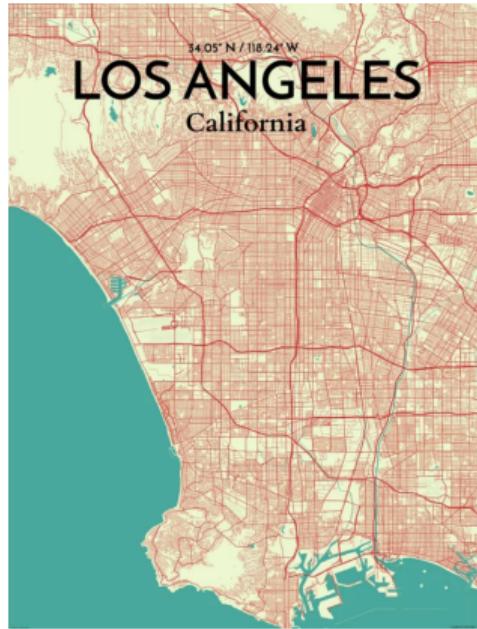


continuous-time flow/diffusion process as a P/SDE

⇒ no close form for the likelihood in general
⇒ no tractable EVI

Marginal queries (MAR)

q₁: What is the probability that today is a Monday ~~at 12:00~~ and there is a traffic jam ~~only~~ on Westwood Blvd.?

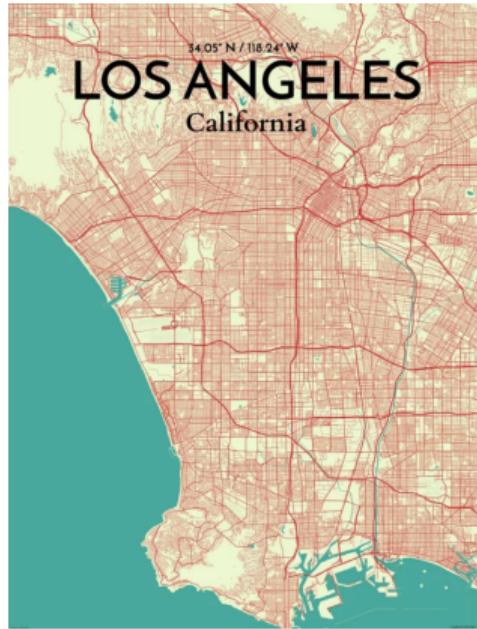


© fineartamerica.com

Marginal queries (MAR)

q₁: What is the probability that today is a Monday at 12:00 and there is a traffic jam only on Westwood Blvd.?

$$q_1(\mathbf{m}) = p_{\mathbf{m}}(\text{Day} = \text{Mon}, \text{Jam}_{\text{Westwood}} = 1)$$



© fineartamerica.com

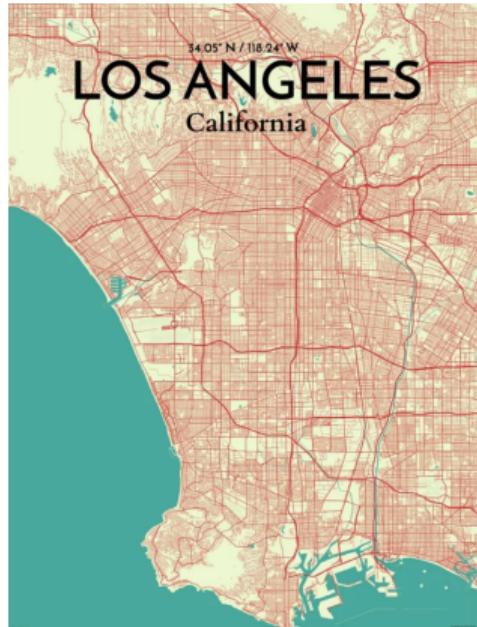
Marginal queries (MAR)

q₁: What is the probability that today is a Monday at 12:00 and there is a traffic jam only on Westwood Blvd.?

$$q_1(\mathbf{m}) = p_{\mathbf{m}}(\text{Day} = \text{Mon}, \text{Jam}_{\text{Westwood}} = 1)$$

General: $p_{\mathbf{m}}(\mathbf{e}) = \int p_{\mathbf{m}}(\mathbf{e}, \mathbf{H}) d\mathbf{H}$

where $\mathbf{E} \subset \mathbf{X}$, $\mathbf{H} = \mathbf{X} \setminus \mathbf{E}$



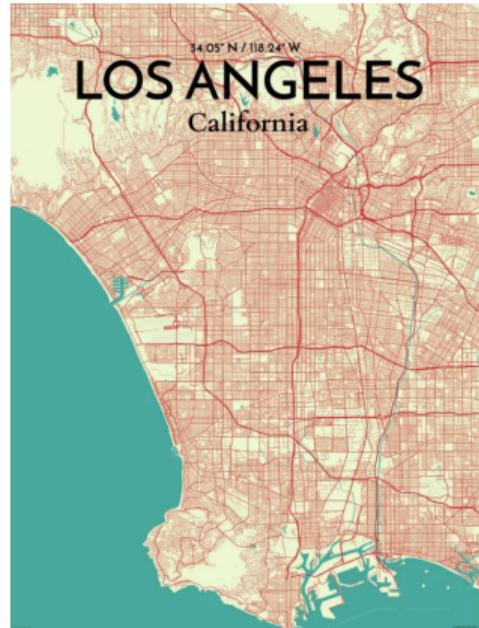
Marginal queries (MAR)

q₁: What is the probability that today is a Monday at 12:00 and there is a traffic jam only on Westwood Blvd.?

$$q_1(\mathbf{m}) = p_{\mathbf{m}}(\text{Day} = \text{Mon}, \text{Jam}_{\text{Westwood}} = 1)$$

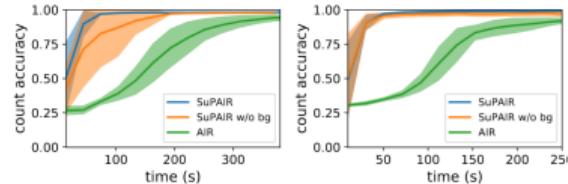
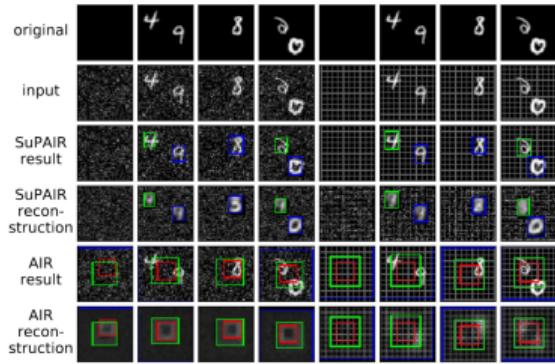
tractable MAR \Rightarrow tractable **conditional queries**
(CON):

$$p_{\mathbf{m}}(\mathbf{q} \mid \mathbf{e}) = \frac{p_{\mathbf{m}}(\mathbf{q}, \mathbf{e})}{p_{\mathbf{m}}(\mathbf{e})}$$



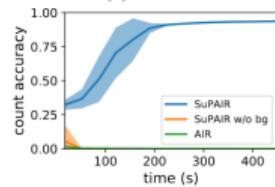
© fineartamerica.com

Tractable MAR : scene understanding

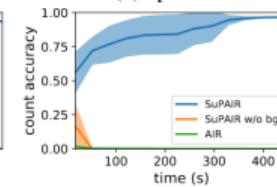


(a) MNIST

(b) Sprites



(c) Noisy MNIST



(d) Grid MNIST

East and exact marginalization over unseen or “do not care” parts in the scene
Stelzner, Peharz, and Kersting, “Faster Attend-Infer-Repeat with Tractable Probabilistic Models”,
ICML, 2019

Kossen et al., “Structured Object-Aware Physics Prediction for Video Modeling and Planning”,
arXiv preprint arXiv:1910.02425, 2019

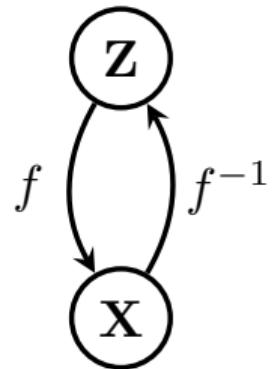
Normalizing flows

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{Z}}(f^{-1}(\mathbf{x})) \left| \det \left(\frac{\delta f^{-1}}{\delta \mathbf{x}} \right) \right|$$

an explicit likelihood!

...plus structured Jacobians

⇒ *tractable EVI queries!*



~~Normalizing flows~~

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{Z}}(f^{-1}(\mathbf{x})) \left| \det \left(\frac{\delta f^{-1}}{\delta \mathbf{x}} \right) \right|$$

an explicit likelihood!

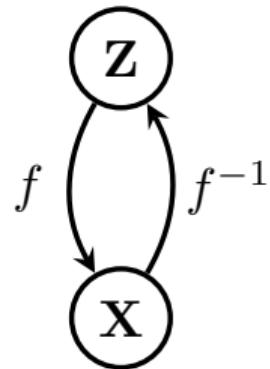
...plus structured Jacobians

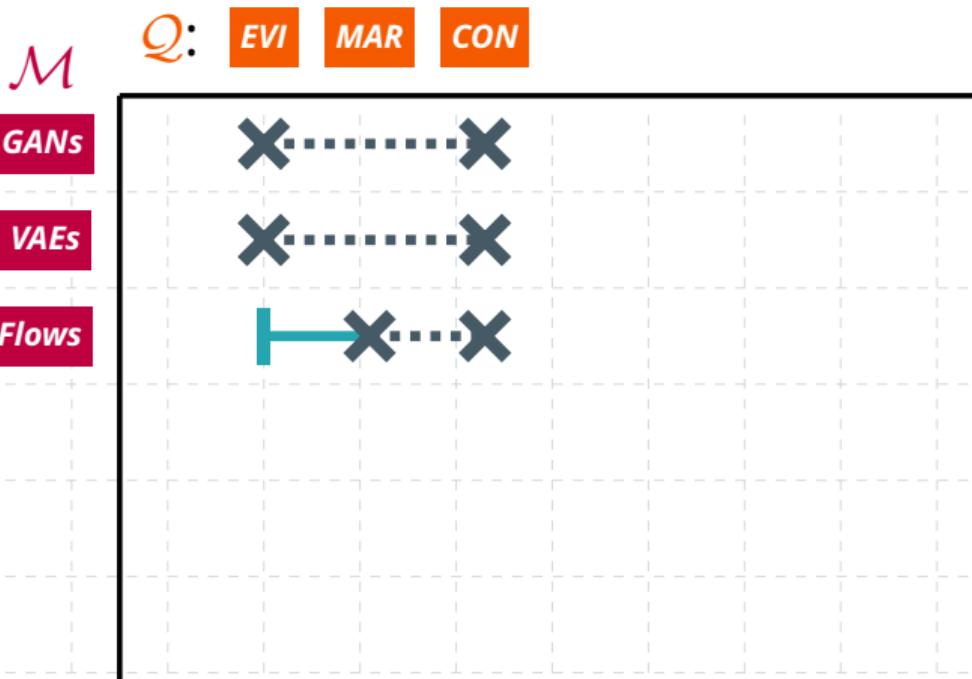
⇒ *tractable EVI queries!*

MAR is generally intractable:

we cannot easily integrate over f

⇒ *unless f is “simple”, e.g. identity*





tractable bands

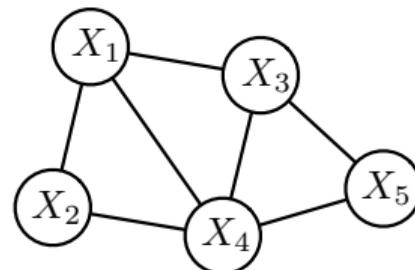
Probabilistic Graphical Models (PGMs)

Declarative semantics: a clean separation of modeling assumptions from inference

Nodes: random variables

Edges: dependencies

+



Inference: conditioning [Darwiche 2001; Sang, Beame, and

Kautz 2005]

elimination [Zhang and Poole 1994; Dechter

1998]

message passing [Yedidya, Freeman, and Weiss

MOGWAI

H~~A~~**PGMs** WILL NEVER DIE,
BUT YOU WILL.



Complexity of MAR on PGMs

Exact complexity: Computing MAR and CON is *#P-hard*

⇒ [Cooper 1990; Roth 1996]

Approximation complexity: Computing MAR and CON approximately within a relative error of $2^{n^{1-\epsilon}}$ for any fixed ϵ is *NP-hard*

⇒ [Dagum and Luby 1993; Roth 1996]

Why? Treewidth!

Treewidth:

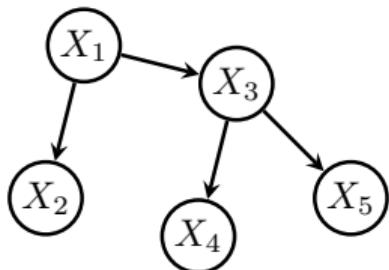
Informally, how tree-like is the graphical model **m**?

Formally, the minimum width of any tree-decomposition of **m**.

Fixed-parameter tractable: MAR and CON on a graphical model **m** with treewidth w take time $O(|\mathbf{X}| \cdot 2^w)$, which is linear for fixed width w

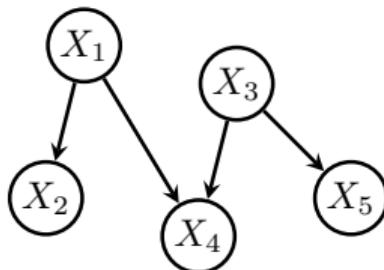
[Dechter 1998; Koller and Friedman 2009]. \Rightarrow what about bounding the treewidth by design?

Low-treewidth PGMS



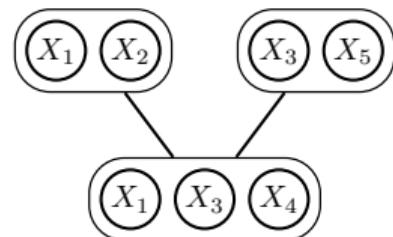
Trees

[Meilă and Jordan 2000]



Polytrees

[Dasgupta 1999]



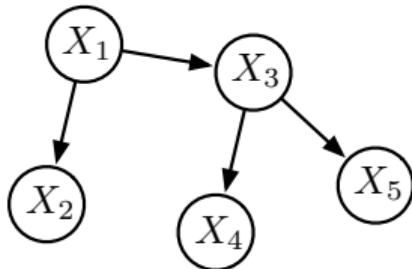
Thin Junction trees

[Bach and Jordan 2001]

If treewidth is bounded (e.g. ≈ 20), exact MAR and CON inference is possible in practice

Tree distributions

A **tree-structured BN** [Meilă and Jordan 2000] where each $X_i \in \mathbf{X}$ has *at most* one parent Pa_{X_i} .

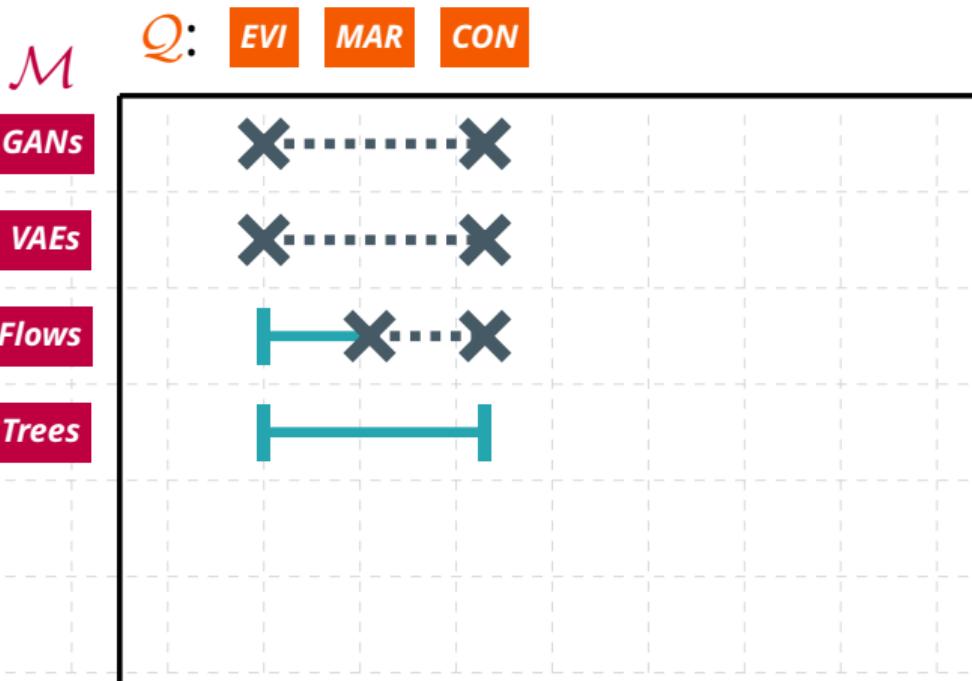


$$p(\mathbf{X}) = \prod_{i=1}^n p(x_i | \text{Pa}_{x_i})$$

Exact querying: EVI, MAR, CON tasks *linear* for trees: $O(|\mathbf{X}|)$

Exact learning from d examples takes $O(|\mathbf{X}|^2 \cdot d)$ with the classical Chow-Liu algorithm¹

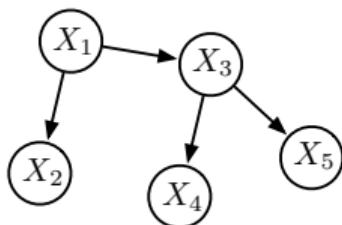
¹ Chow and Liu, "Approximating discrete probability distributions with dependence trees", 1968



tractable bands

What do we lose?

Expressiveness: Ability to represent rich and complex classes of distributions

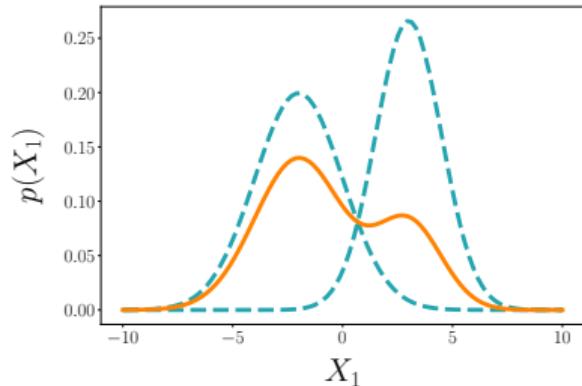


Bounded-treewidth PGMs lose the ability to represent *all possible distributions* ...

Cohen, Sharir, and Shashua, "On the expressive power of deep learning: A tensor analysis", 2016
Martens and Medabalimi, "On the Expressive Efficiency of Sum Product Networks", CoRR, 2014

Mixtures

Mixtures as a convex combination of k (simpler) probabilistic models

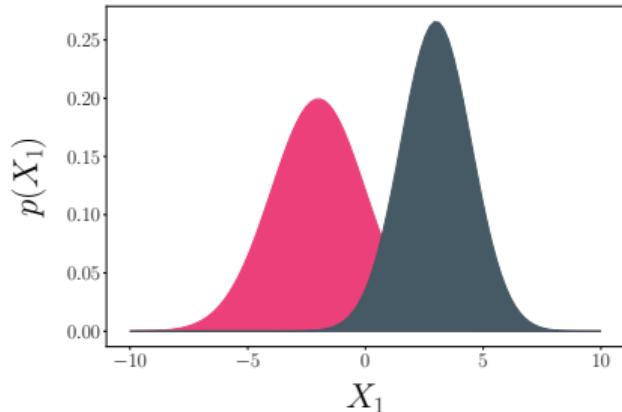


$$p(X) = w_1 \cdot p_1(X) + w_2 \cdot p_2(X)$$

EVI, MAR, CON queries scale linearly in k

Mixtures

Mixtures as a convex combination of k (simpler) probabilistic models



$$p(X) = p(Z = 1) \cdot p_1(X|Z = 1) + p(Z = 2) \cdot p_2(X|Z = 2)$$

Mixtures are marginalizing a **categorical latent variable** Z with k values

⇒ increased expressiveness

Expressiveness and efficiency

Expressiveness: Ability to represent rich and effective classes of functions

⇒ *mixture of Gaussians can approximate any distribution!*

Expressiveness and efficiency

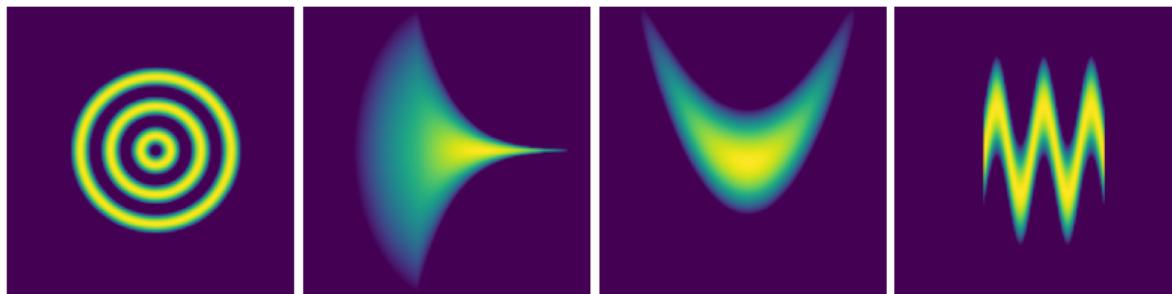
Expressiveness: Ability to represent rich and effective classes of functions

⇒ *mixture of Gaussians can approximate any distribution!*

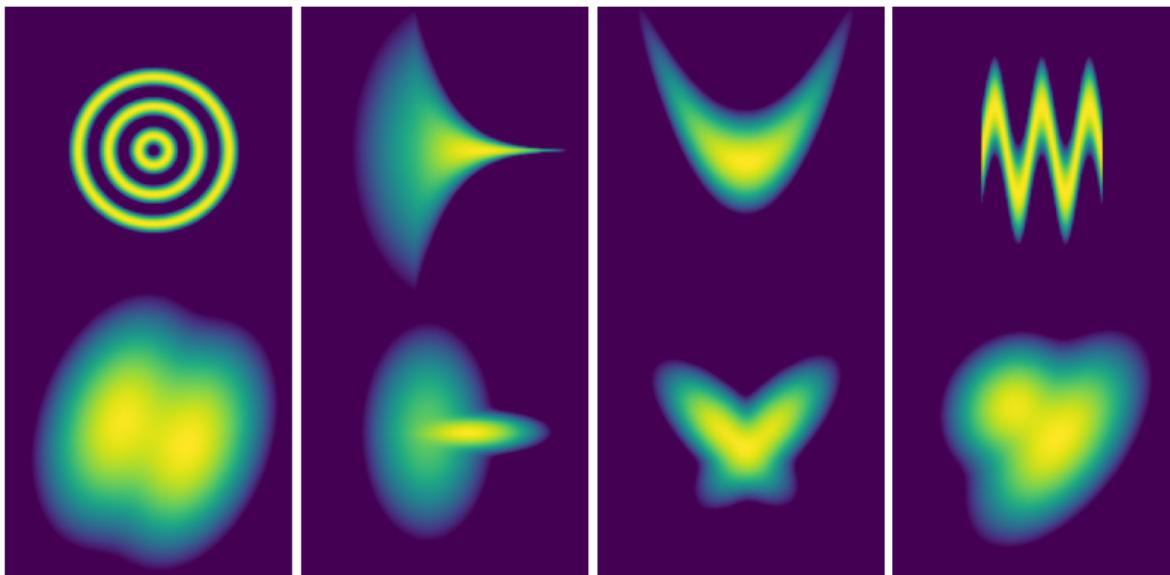
Expressive efficiency (succinctness) Ability to represent rich and effective classes of functions **compactly**

⇒ *but how many components does a Gaussian mixture need?*

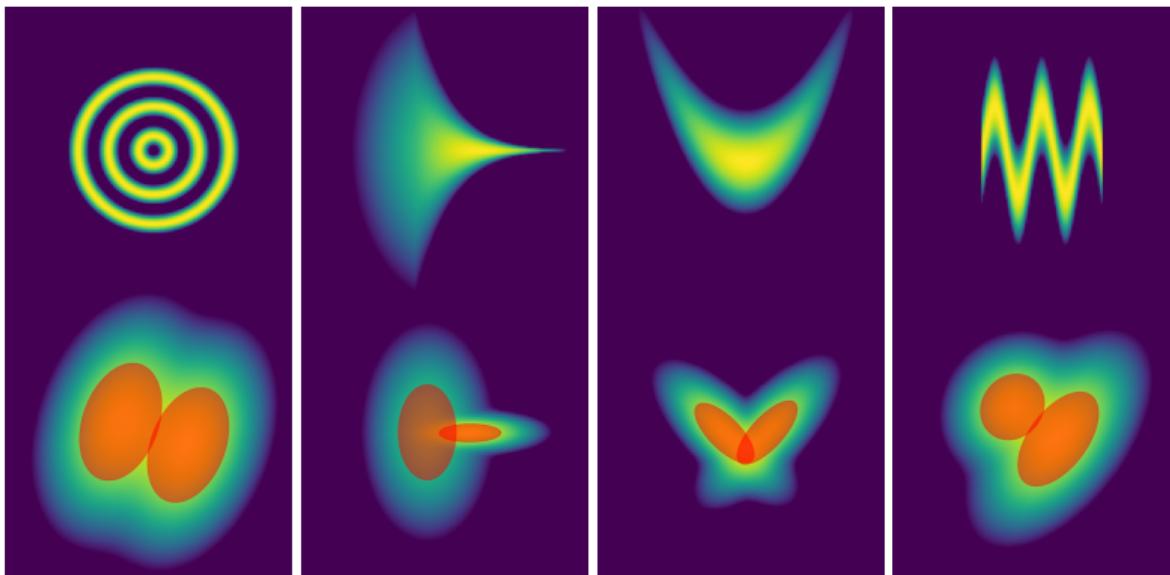
How expressive efficient are mixtures?



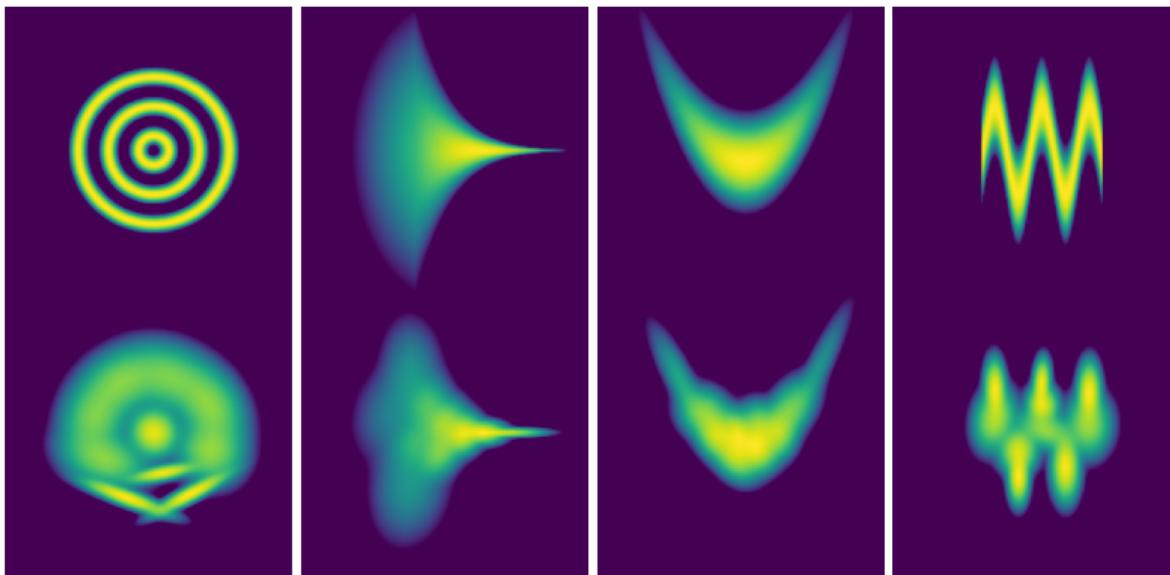
How expressive efficient are mixtures?



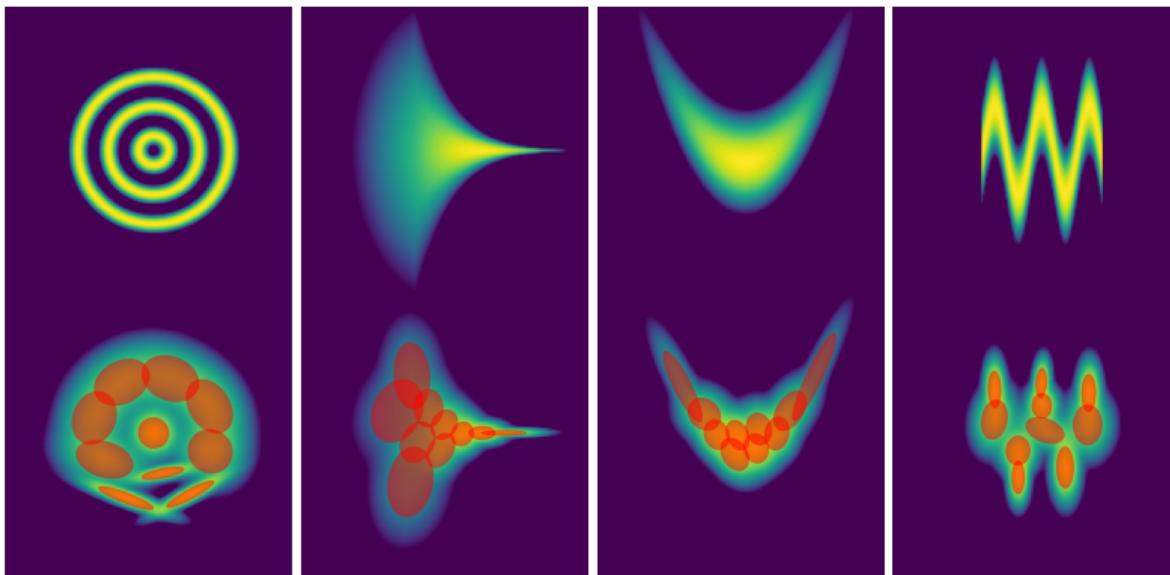
How expressive efficient are mixtures?



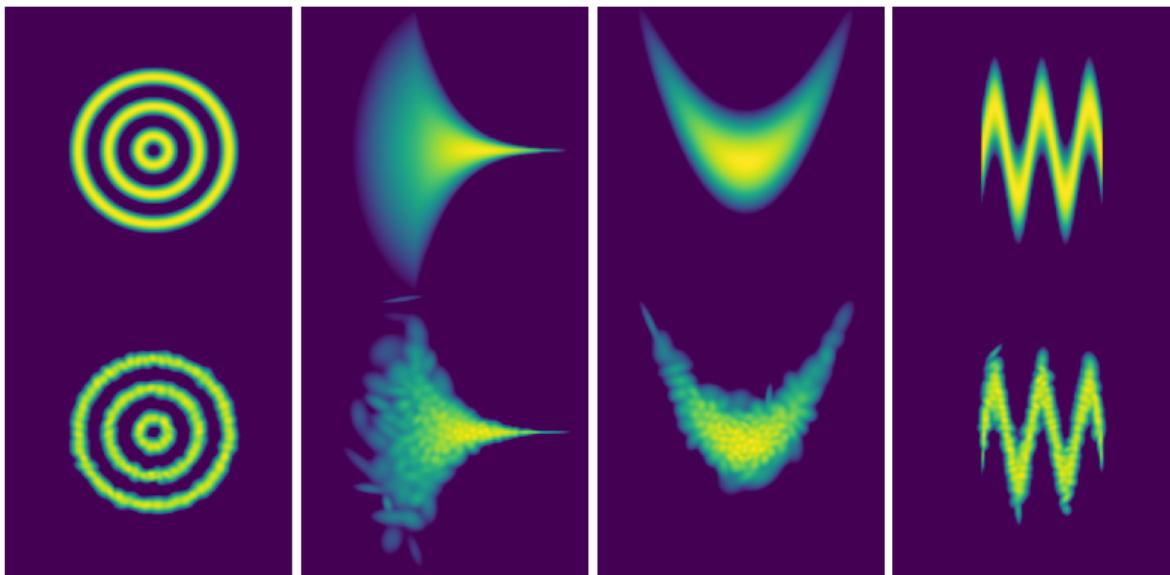
How expressive efficient are mixtures?



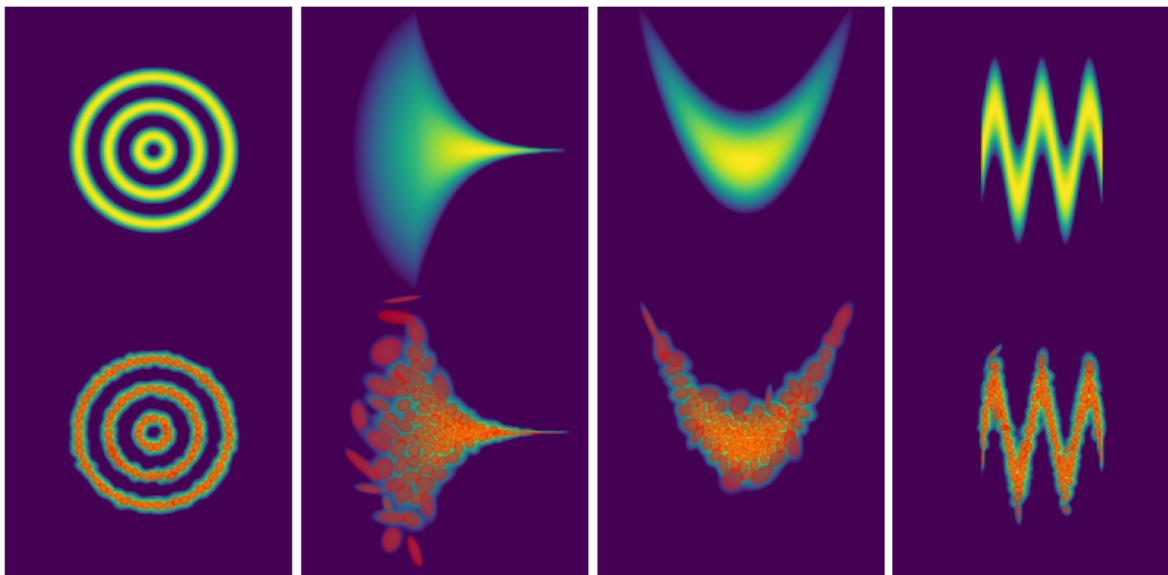
How expressive efficient are mixtures?



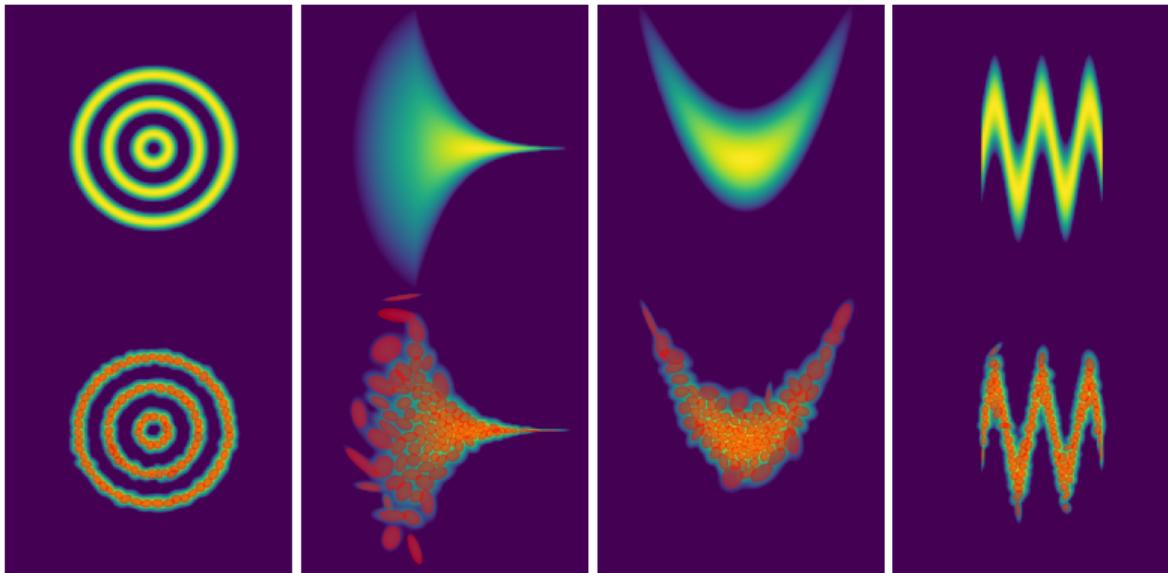
How expressive efficient are mixtures?



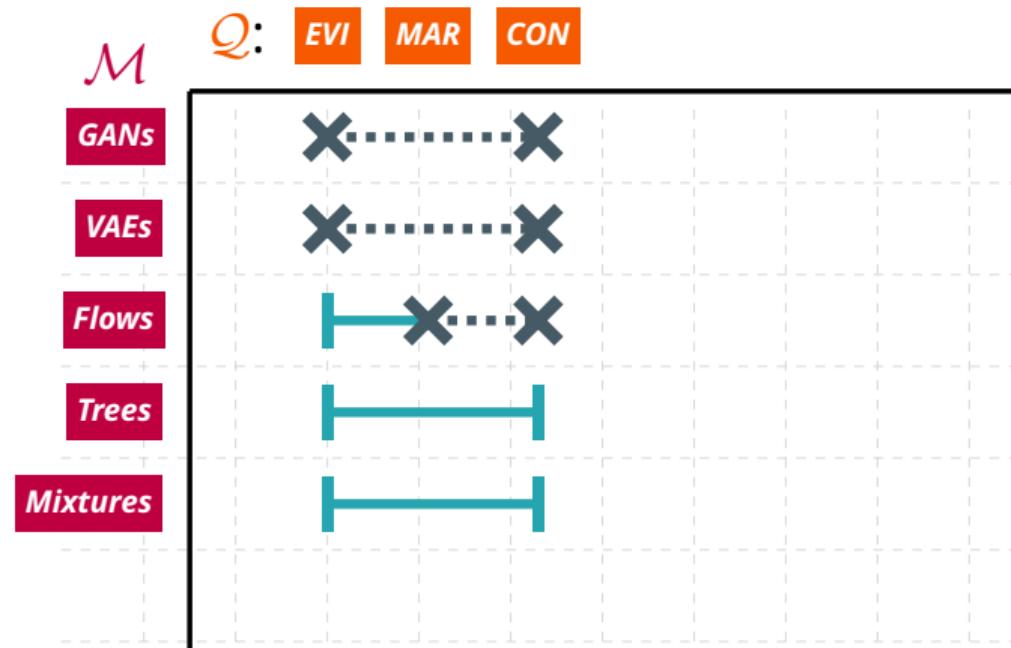
How expressive efficient are mixtures?



How expressive efficient are mixtures?



⇒ *stack mixtures like in deep generative models* 43/67

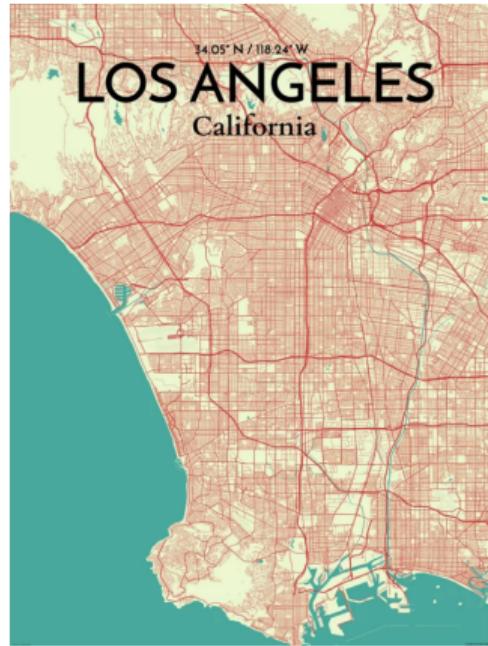


tractable bands

Maximum A Posteriori (MAP)

aka Most Probable Explanation (MPE)

q₅: *Which combination of roads is most likely to be jammed on Monday at 9am?*



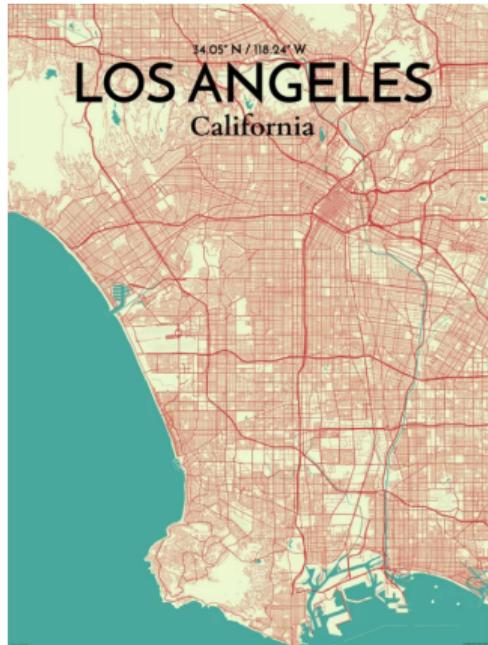
© fineartamerica.com

Maximum A Posteriori (MAP)

aka Most Probable Explanation (MPE)

q₅: Which combination of roads is most likely to be jammed on Monday at 9am?

$$q_5(\mathbf{m}) = \operatorname{argmax}_{\mathbf{j}} p_{\mathbf{m}}(\mathbf{j}_1, \mathbf{j}_2, \dots \mid \text{Day}=\text{M}, \text{Time}=9)$$



© fineartamerica.com

Maximum A Posteriori (MAP)

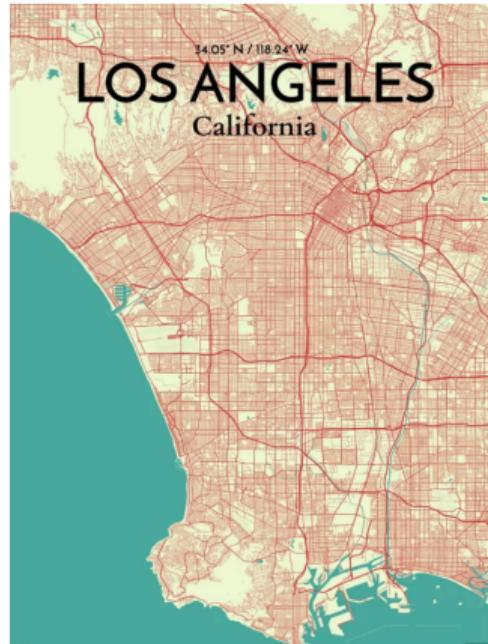
aka Most Probable Explanation (MPE)

q₅: Which combination of roads is most likely to be jammed on Monday at 9am?

$$q_5(\mathbf{m}) = \operatorname{argmax}_{\mathbf{j}} p_{\mathbf{m}}(\mathbf{j}_1, \mathbf{j}_2, \dots \mid \text{Day}=\text{M}, \text{Time}=9)$$

General: $\operatorname{argmax}_{\mathbf{q}} p_{\mathbf{m}}(\mathbf{q} \mid \mathbf{e})$

where $\mathbf{Q} \cup \mathbf{E} = \mathbf{X}$



© fineartamerica.com

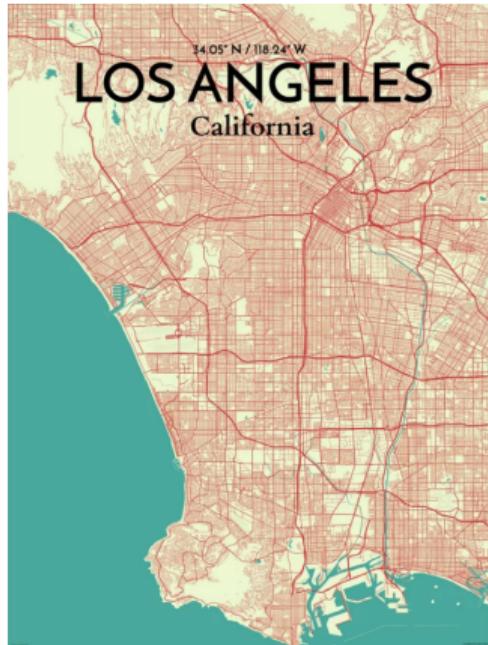
Maximum A Posteriori (MAP)

aka Most Probable Explanation (MPE)

q₅: Which combination of roads is most likely to be jammed on Monday at 9am?

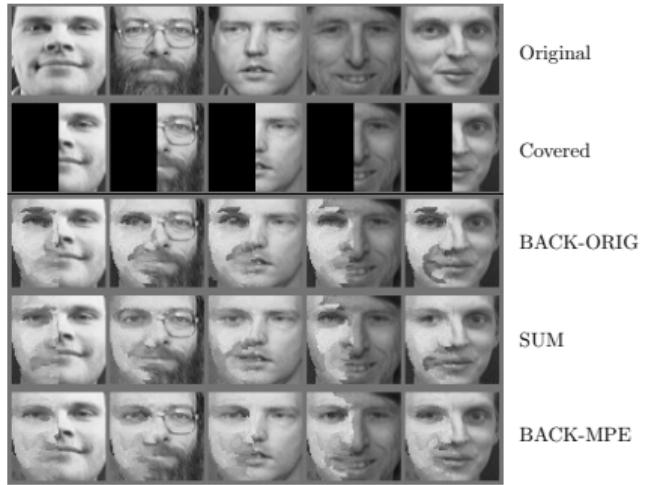
...***intractable*** for latent variable models!

$$\begin{aligned}\max_{\mathbf{q}} p_{\mathbf{m}}(\mathbf{q} \mid \mathbf{e}) &= \max_{\mathbf{q}} \sum_{\mathbf{z}} p_{\mathbf{m}}(\mathbf{q}, \mathbf{z} \mid \mathbf{e}) \\ &\neq \sum_{\mathbf{z}} \max_{\mathbf{q}} p_{\mathbf{m}}(\mathbf{q}, \mathbf{z} \mid \mathbf{e})\end{aligned}$$



© fineartamerica.com

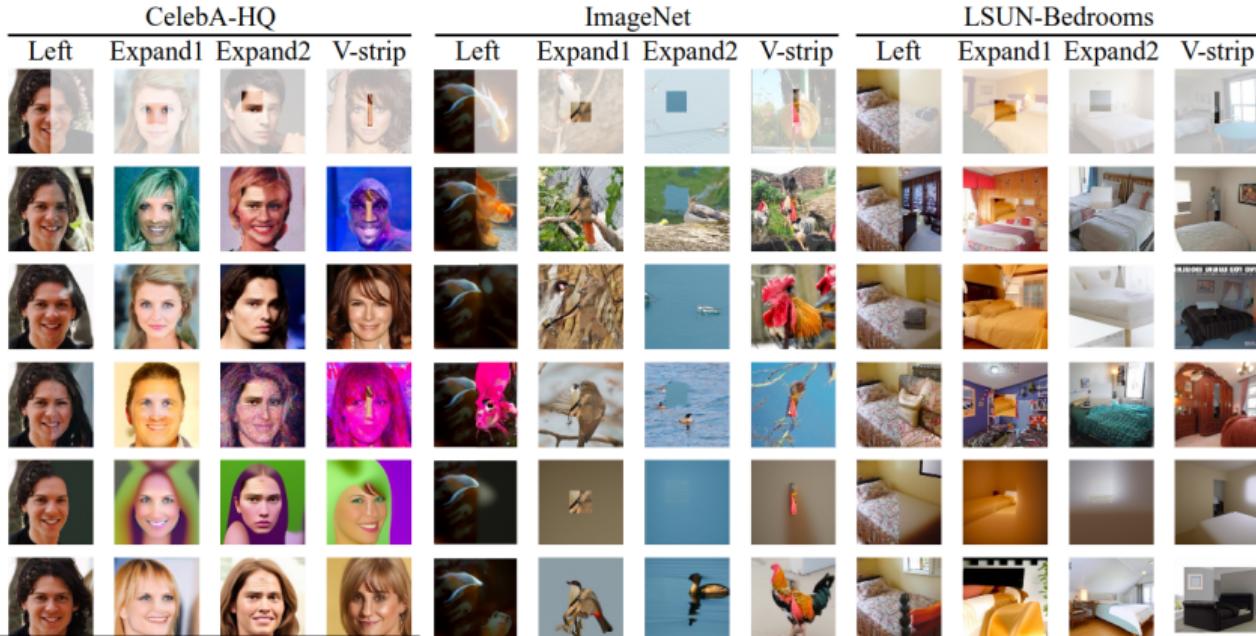
MAP inference : image inpainting



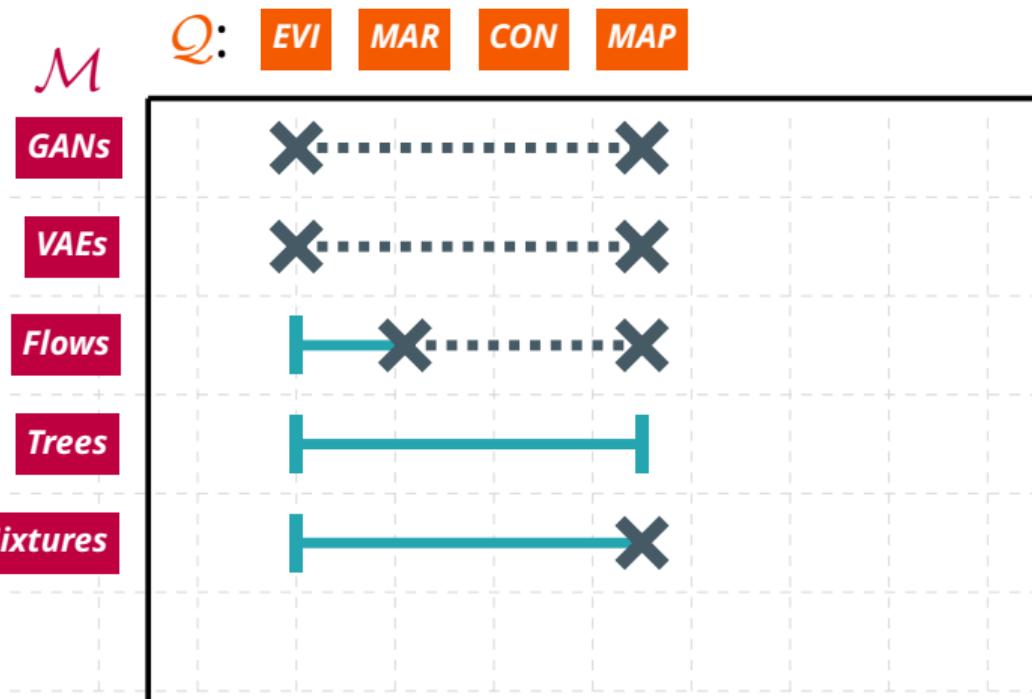
Predicting ***arbitrary patches***
given a ***single*** model
without the need of retraining.

Poon and Domingos, "Sum-Product Networks: a New Deep Architecture", UAI 2011, 2011
Sguerra and Cozman, "Image classification using sum-product networks for autonomous flight of
micro aerial vehicles", , 2016

MAP inference : image inpainting



Liu, Niepert, and Broeck, "Image Inpainting via Tractable Steering of Diffusion Models", , 2024

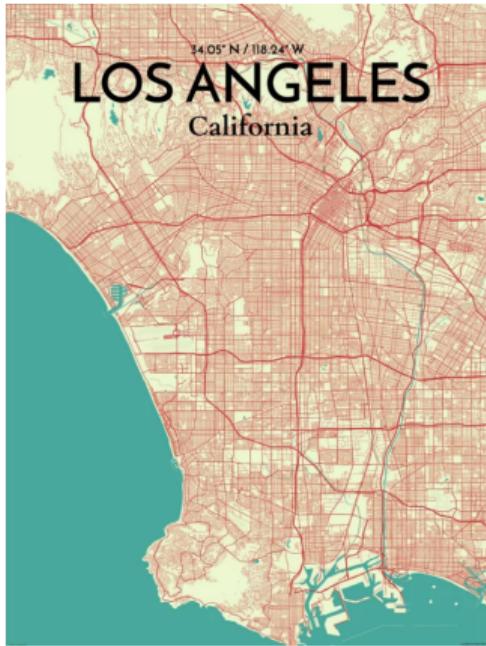


tractable bands

Marginal MAP (MMAP)

aka Bayesian Network MAP

q₆: Which combination of roads is most likely to be jammed ~~on Monday~~ at 9am?



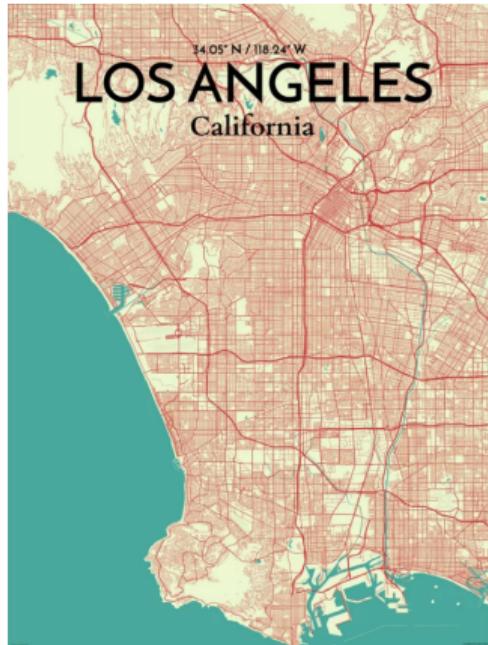
© fineartamerica.com

Marginal MAP (MMAP)

aka Bayesian Network MAP

q₆: Which combination of roads is most likely to be jammed ~~on Monday~~ at 9am?

$$\mathbf{q}_6(\mathbf{m}) = \operatorname{argmax}_{\mathbf{j}} p_{\mathbf{m}}(\mathbf{j}_1, \mathbf{j}_2, \dots \mid \text{Time}=9)$$



© fineartamerica.com

Marginal MAP (MMAP)

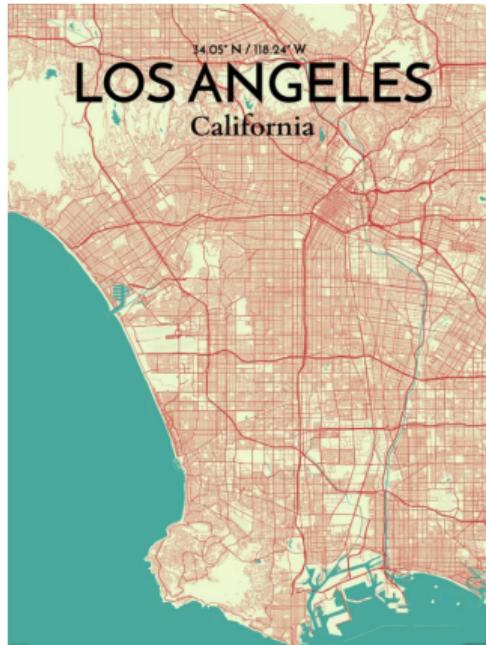
aka Bayesian Network MAP

q₆: Which combination of roads is most likely to be jammed ~~on Monday~~ at 9am?

$$q_6(\mathbf{m}) = \operatorname{argmax}_{\mathbf{j}} p_{\mathbf{m}}(\mathbf{j}_1, \mathbf{j}_2, \dots | \text{Time}=9)$$

General: $\operatorname{argmax}_{\mathbf{q}} p_{\mathbf{m}}(\mathbf{q} | \mathbf{e})$
 $= \operatorname{argmax}_{\mathbf{q}} \sum_{\mathbf{h}} p_{\mathbf{m}}(\mathbf{q}, \mathbf{h} | \mathbf{e})$

where $\mathbf{Q} \cup \mathbf{H} \cup \mathbf{E} = \mathbf{X}$



Marginal MAP (MMAP)

aka Bayesian Network MAP

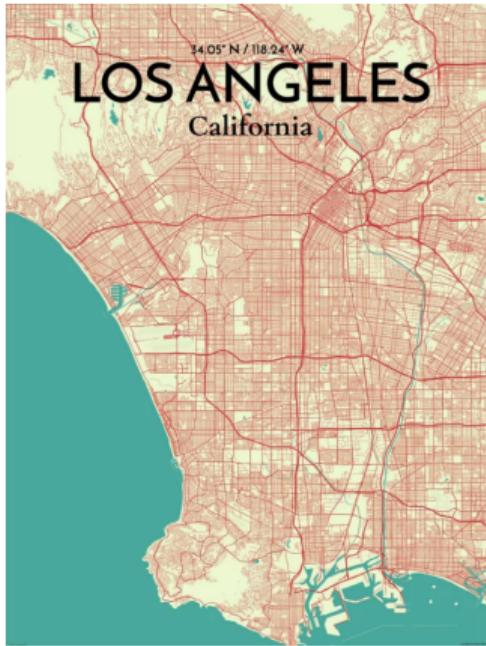
q₆: Which combination of roads is most likely to be jammed ~~on Monday~~ at 9am?

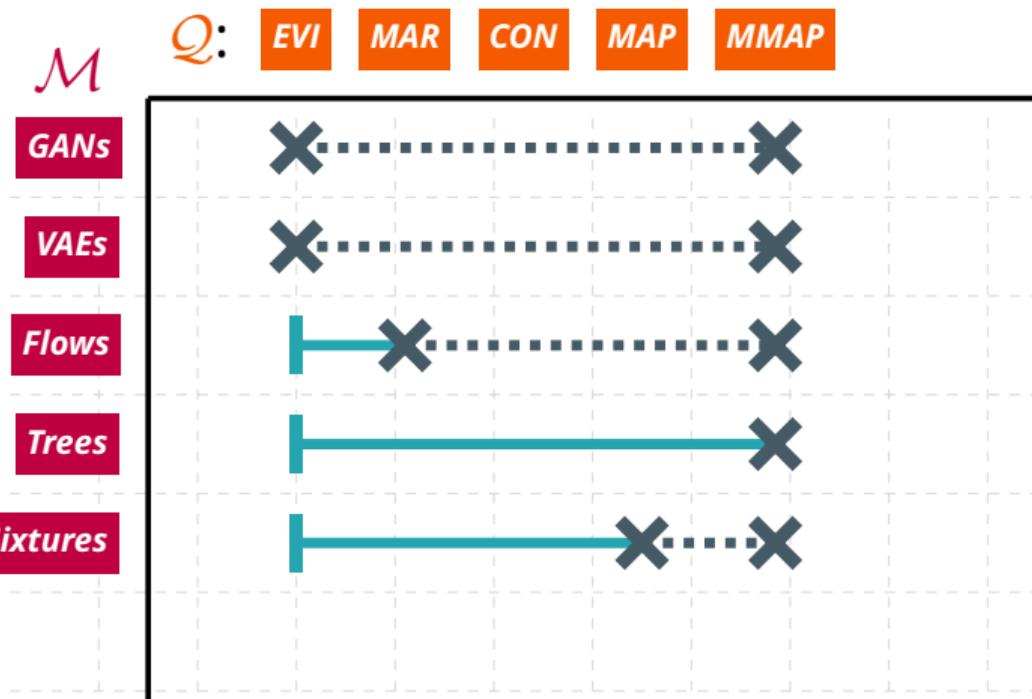
$$q_6(\mathbf{m}) = \operatorname{argmax}_{\mathbf{j}} p_{\mathbf{m}}(\mathbf{j}_1, \mathbf{j}_2, \dots \mid \text{Time} = 9)$$

⇒ NP^{PP}-complete [Park and Darwiche 2006]

⇒ NP-hard for trees [de Campos 2011]

⇒ NP-hard even for Naive Bayes [ibid.]

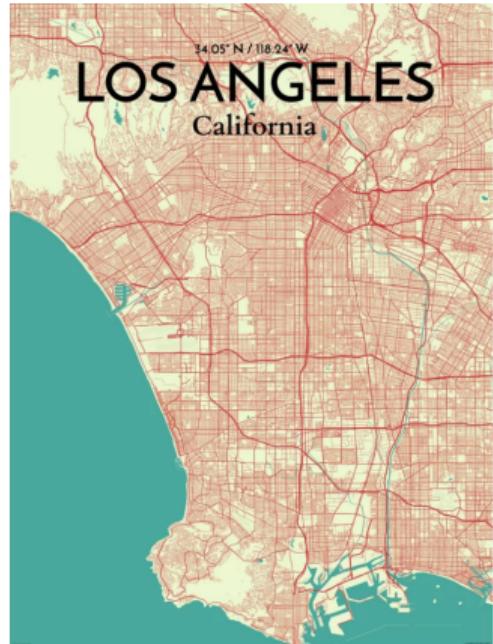




tractable bands

Advanced queries

q₂: *Which day is most likely to have a traffic jam on my route to campus?*



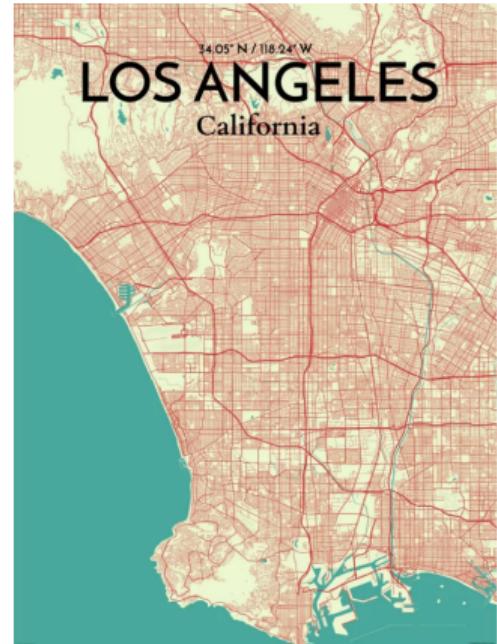
© fineartamerica.com

Advanced queries

q₂: Which day is most likely to have a traffic jam on my route to campus?

$$q_2(\mathbf{m}) = \operatorname{argmax}_d p_{\mathbf{m}}(\text{Day} = d \wedge \bigvee_{i \in \text{route}} \text{Jam}_{\text{Str } i})$$

\Rightarrow **marginals + MAP + logical events**

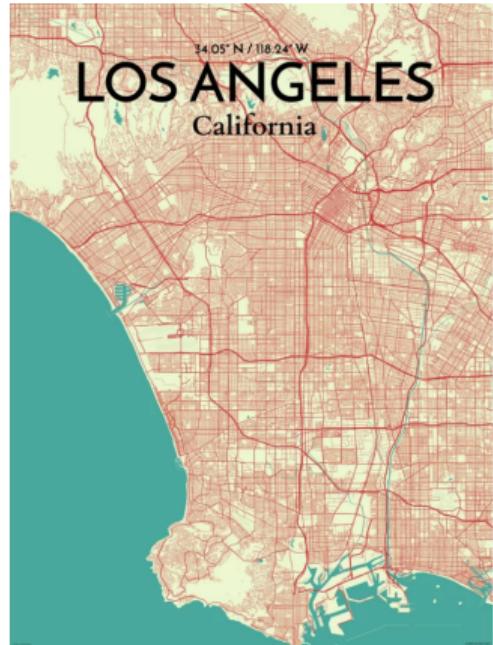


© fineartamerica.com

Advanced queries

q₂: *Which day is most likely to have a traffic jam on my route to campus?*

q₇: *What is the probability of seeing more traffic jams in Westwood than Hollywood?*

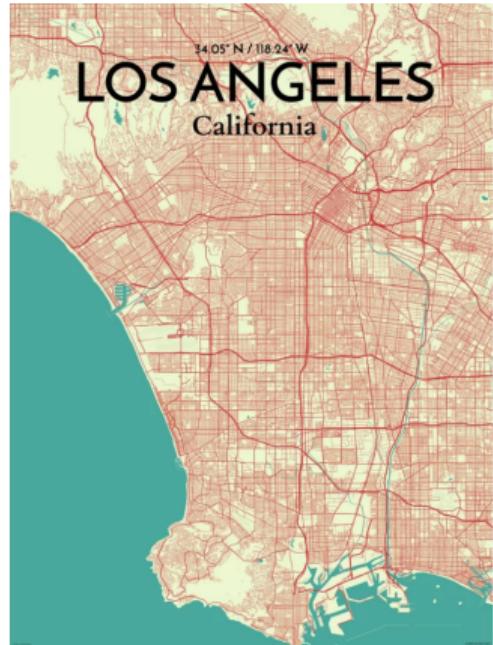


Advanced queries

q₂: Which day is most likely to have a traffic jam on my route to campus?

q₇: What is the probability of seeing more traffic jams in Westwood than Hollywood?

⇒ **counts + group comparison**



© fineartamerica.com

Advanced queries

q₂: *Which day is most likely to have a traffic jam on my route to campus?*

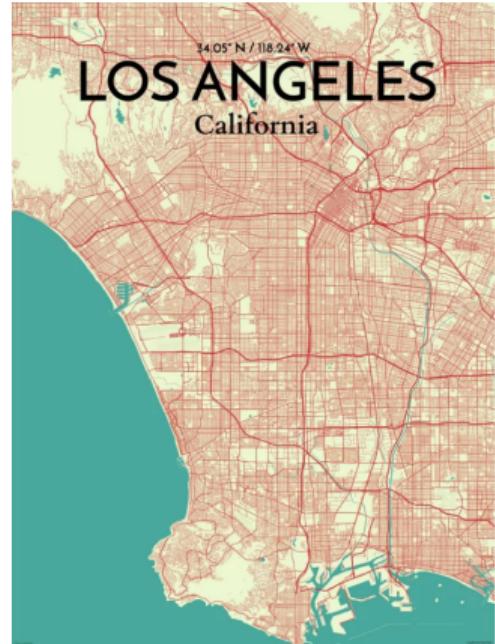
q₇: *What is the probability of seeing more traffic jams in Westwood than Hollywood?*

and more:

expected classification agreement

[Oztok, Choi, and Darwiche 2016; Choi, Darwiche, and Broeck 2017; Choi and Broeck 2018]

expected predictions [Khosravi et al. 2019]



© fineartamerica.com

more complex reasoning



q₁

"What is the **expected prediction** for a patient with unavailable records?"

q₂

"How **fair** is the prediction when a certain protected attribute changes?"

q₃

"Can we certify no **adversarial examples** exist?"

*...asking **queries** to a ML model*

more complex reasoning



q₁ $\mathbb{E}_{\mathbf{x}_m \sim p(\mathbf{X}_m | \mathbf{x}_o)} [f(\mathbf{x}_o, \mathbf{x}_m)]$
(expected prediction)

q₂ $\mathbb{E}_{\mathbf{x}_c \sim p(\mathbf{X}_c | X_s=0)} [f_0(\mathbf{x}_c)] -$
 $\mathbb{E}_{\mathbf{x}_c \sim p(\mathbf{X}_c | X_s=1)} [f_1(\mathbf{x}_c)]$
(fairness)

q₃ $\mathbb{E}_{\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_D)} [f(\mathbf{x} + \mathbf{e})]$
(adversarial robust.)

...into math expressions

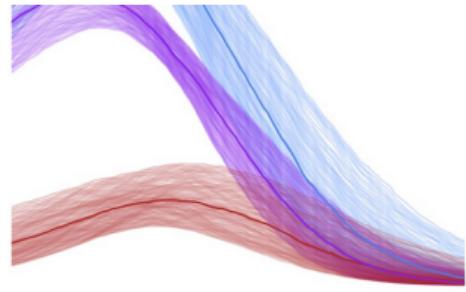
more complex reasoning



neuro-symbolic AI



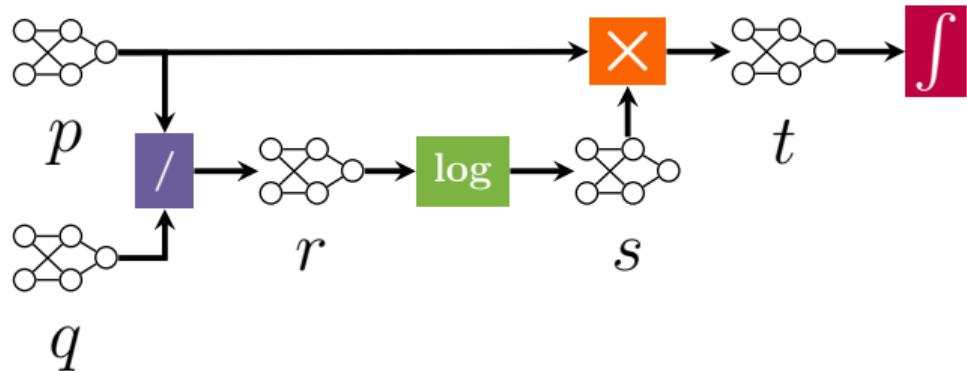
probabilistic programming



*computing uncertainties
(Bayesian inference)*

...and more application scenarios

$$\int p(\mathbf{x}) \times \log \left(p(\mathbf{x}) / q(\mathbf{x}) \right) d\mathbf{X}$$



build a LEGO-like query calculus!

	Query	Tract. Conditions	Hardness
CROSS ENTROPY	$-\int p(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{X}$	Cmp, q Det	#P-hard w/o Det
SHANNON ENTROPY	$-\sum p(\mathbf{x}) \log p(\mathbf{x})$	Sm, Dec, Det	coNP-hard w/o Det
RÉNYI ENTROPY	$(1 - \alpha)^{-1} \log \int p^\alpha(\mathbf{x}) d\mathbf{X}, \alpha \in \mathbb{N}$	SD	#P-hard w/o SD
MUTUAL INFORMATION	$(1 - \alpha)^{-1} \log \int p^\alpha(\mathbf{x}) d\mathbf{X}, \alpha \in \mathbb{R}_+$	Sm, Dec, Det	#P-hard w/o Det
KULLBACK-LEIBLER DIV.	$\int p(\mathbf{x}, \mathbf{y}) \log(p(\mathbf{x}, \mathbf{y})/(p(\mathbf{x})p(\mathbf{y})))$	Sm, SD, Det*	coNP-hard w/o SD
RÉNYI'S ALPHA DIV.	$\int p(\mathbf{x}) \log(p(\mathbf{x})/q(\mathbf{x})) d\mathbf{X}$	Cmp, Det	#P-hard w/o Det
ITAKURA-SAITO DIV.	$(1 - \alpha)^{-1} \log \int p^\alpha(\mathbf{x}) q^{1-\alpha}(\mathbf{x}) d\mathbf{X}, \alpha \in \mathbb{N}$	Cmp, q Det	#P-hard w/o Det
CAUCHY-SCHWARZ DIV.	$(1 - \alpha)^{-1} \log \int p^\alpha(\mathbf{x}) q^{1-\alpha}(\mathbf{x}) d\mathbf{X}, \alpha \in \mathbb{R}$	Cmp, Det	#P-hard w/o Det
ITAKURA-SAITO DIV.	$\int [p(\mathbf{x})/q(\mathbf{x}) - \log(p(\mathbf{x})/q(\mathbf{x})) - 1] d\mathbf{X}$	Cmp, Det	#P-hard w/o Det
CAUCHY-SCHWARZ DIV.	$-\log \frac{\int p(\mathbf{x}) q(\mathbf{x}) d\mathbf{X}}{\sqrt{\int p^2(\mathbf{x}) d\mathbf{X} \int q^2(\mathbf{x}) d\mathbf{X}}}$	Cmp	#P-hard w/o Cmp
SQUARED LOSS	$\int (p(\mathbf{x}) - q(\mathbf{x}))^2 d\mathbf{X}$	Cmp	#P-hard w/o Cmp

compositionally derive the tractability of many more queries

Query	Tract. Conditions	Hardness
CROSS ENTROPY	$-\int p(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{X}$	Cmp, q Det #P-hard w/o Det
SHANNON ENTROPY	$-\sum p(\mathbf{x}) \log p(\mathbf{x})$	Sm, Dec, Det coNP-hard w/o Det
RÉNYI ENTROPY	$(1 - \alpha)^{-1} \log \int p^\alpha(\mathbf{x}) d\mathbf{X}, \alpha \in \mathbb{N}$ $(1 - \alpha)^{-1} \log \int p^\alpha(\mathbf{x}) d\mathbf{X}, \alpha \in \mathbb{R}_+$	SD #P-hard w/o SD
MUTUAL INFORMATION	$\int p(\mathbf{x}, \mathbf{y}) \log(p(\mathbf{x}, \mathbf{y})/(p(\mathbf{x})p(\mathbf{y})))$	Sm, Dec, Det* #P-hard w/o Det
KULLBACK-LEIBLER DIV.	$\int p(\mathbf{x}) \log(p(\mathbf{x})/q(\mathbf{x})) d\mathbf{X}$	Sm, SD, Det* coNP-hard w/o SD
RÉNYI'S ALPHA DIV.	$(1 - \alpha)^{-1} \log \int p^\alpha(\mathbf{x}) q^{1-\alpha}(\mathbf{x}) d\mathbf{X}, \alpha \in \mathbb{N}$ $(1 - \alpha)^{-1} \log \int p^\alpha(\mathbf{x}) q^{1-\alpha}(\mathbf{x}) d\mathbf{X}, \alpha \in \mathbb{R}$	Cmp, Det #P-hard w/o Det
ITAKURA-SAITO DIV.	$\int [p(\mathbf{x})/q(\mathbf{x}) - \log(p(\mathbf{x})/q(\mathbf{x})) - 1] d\mathbf{X}$	Cmp, q Det #P-hard w/o Det
CAUCHY-SCHWARZ DIV.	$-\log \frac{\int p(\mathbf{x}) q(\mathbf{x}) d\mathbf{X}}{\sqrt{\int p^2(\mathbf{x}) d\mathbf{X} \int q^2(\mathbf{x}) d\mathbf{X}}}$	Cmp, Det #P-hard w/o Det
SQUARED LOSS	$\int (p(\mathbf{x}) - q(\mathbf{x}))^2 d\mathbf{X}$	Cmp #P-hard w/o Cmp

and prove hardness when some input properties are not satisfied

\mathcal{M}

\mathcal{Q} :

EVI

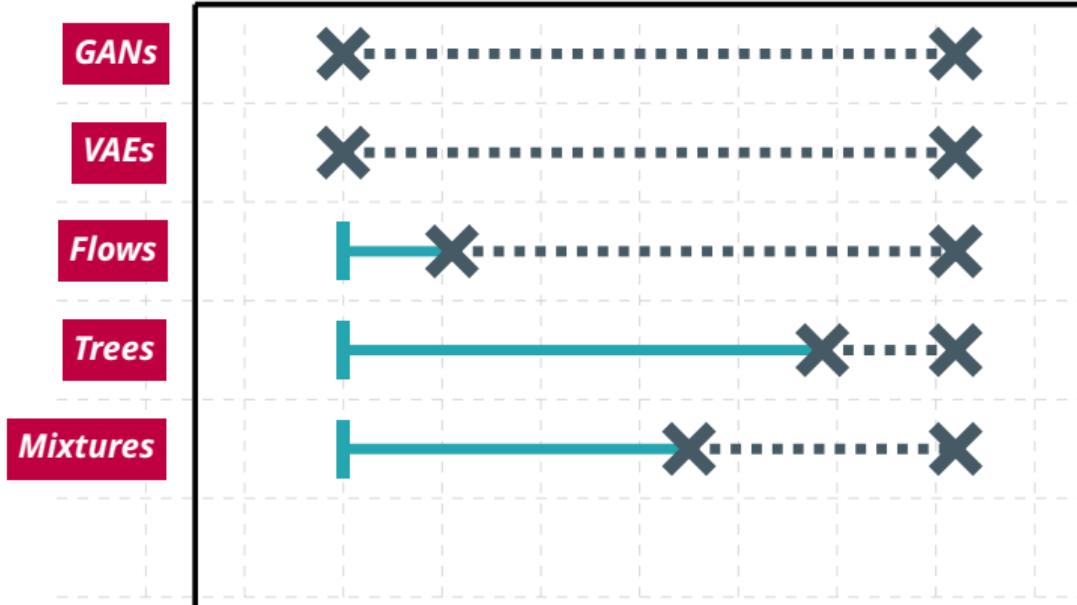
MAR

CON

MAP

MMAP

ADV



tractable bands

\mathcal{M}

\mathcal{Q} :

EVI

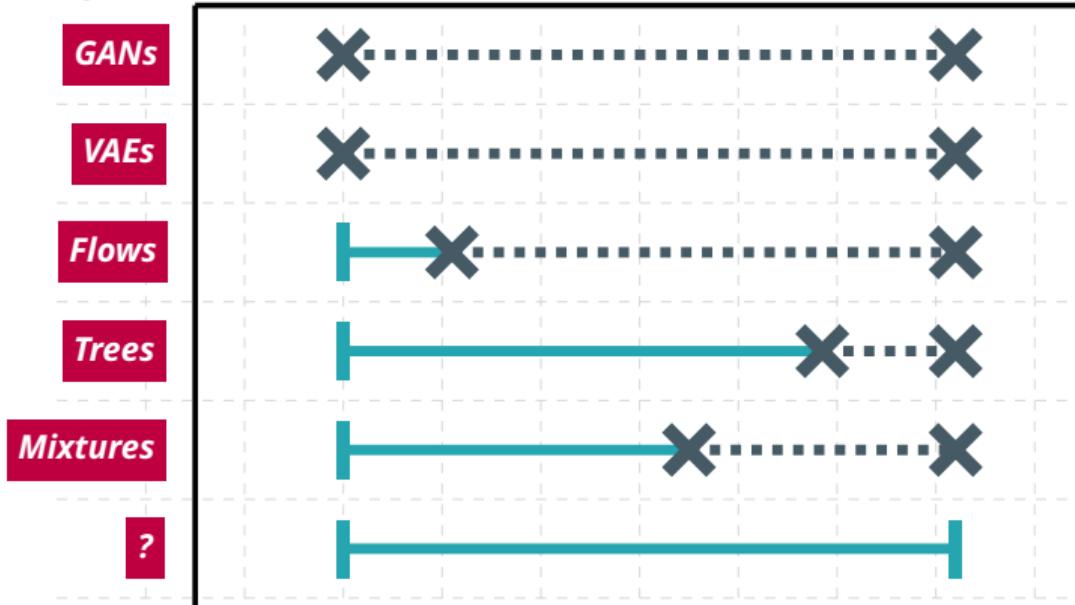
MAR

CON

MAP

MMAP

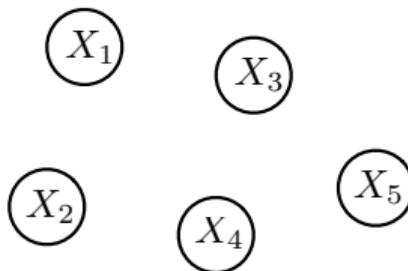
ADV



tractable bands

Fully factorized models

A completely disconnected graph. Example: Product of Bernoullis (PoBs)



$$p(\mathbf{x}) = \prod_{i=1}^n p(x_i)$$

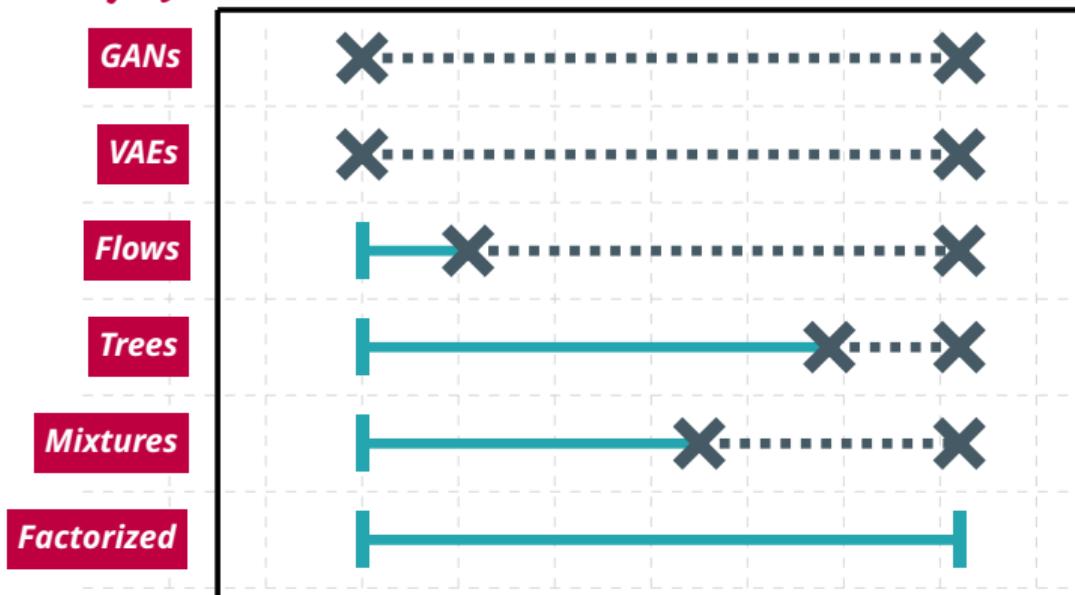
Complete evidence, marginals and MAP, MMAP inference is *linear*!

⇒ *but definitely not expressive...*

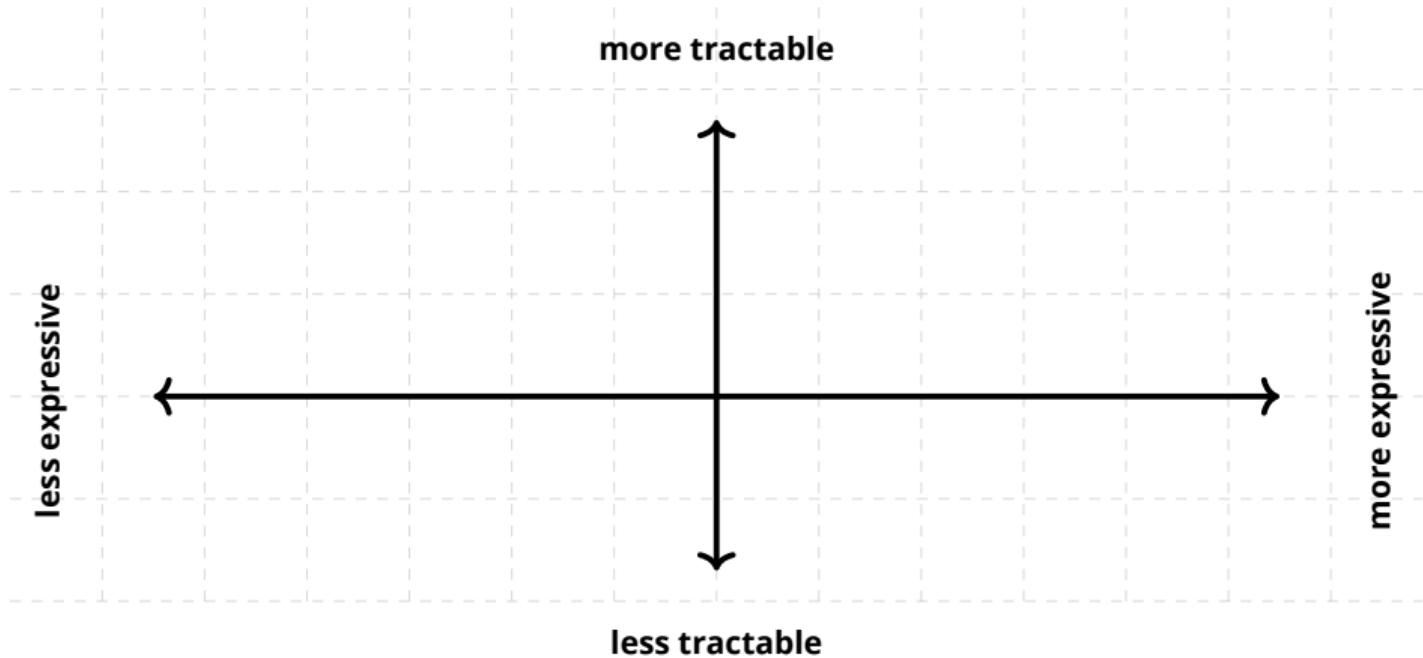
\mathcal{M}

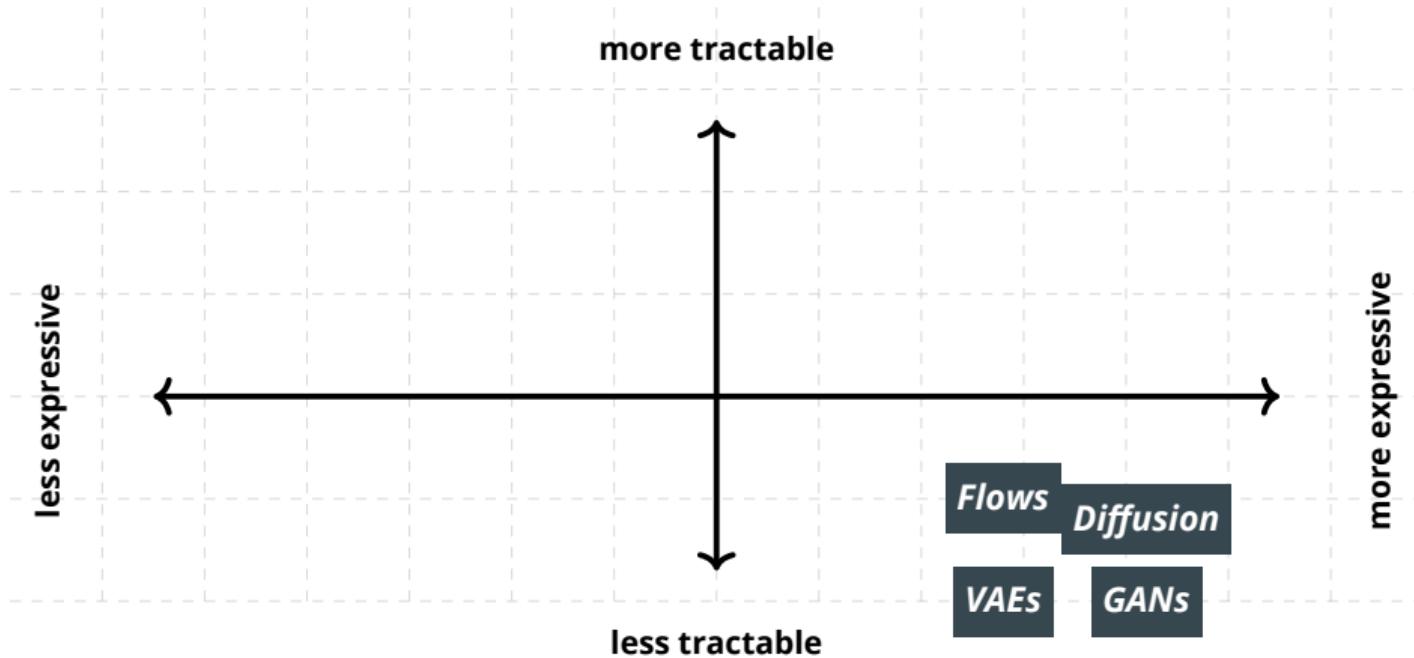
\mathcal{Q} :

EVI MAR CON MAP MMAP ADV

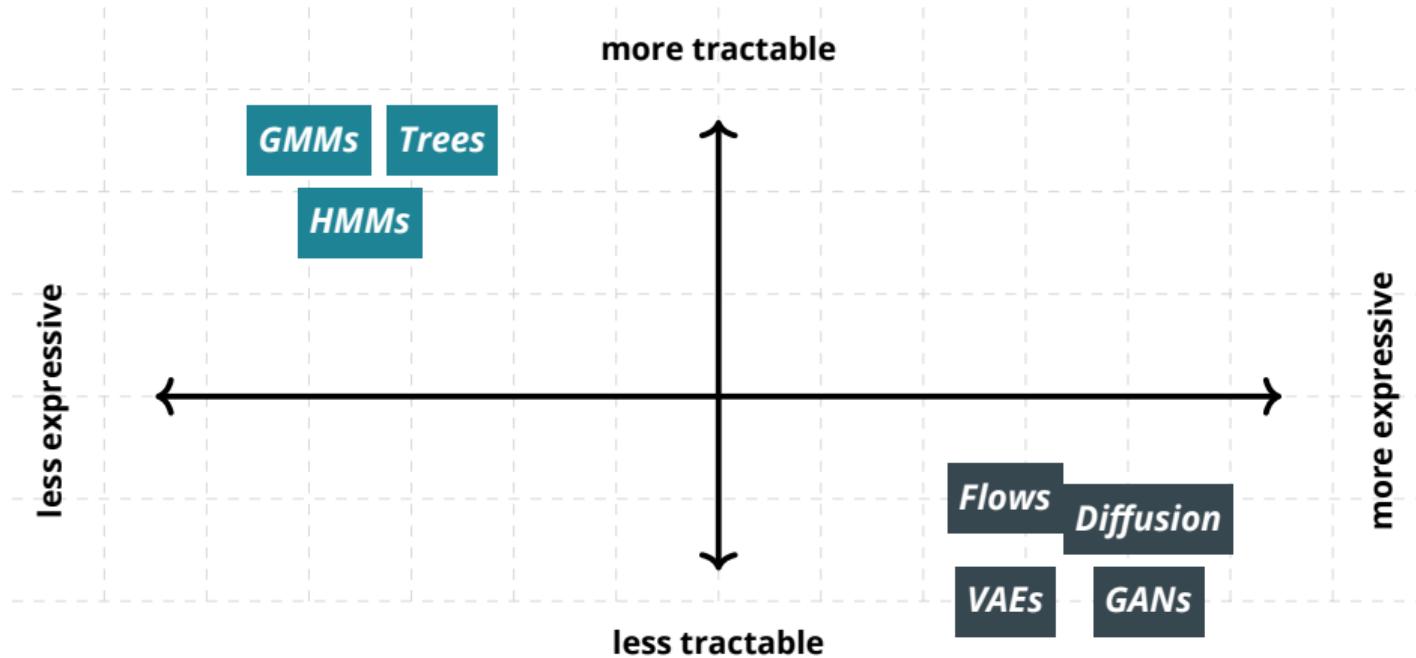


tractable bands

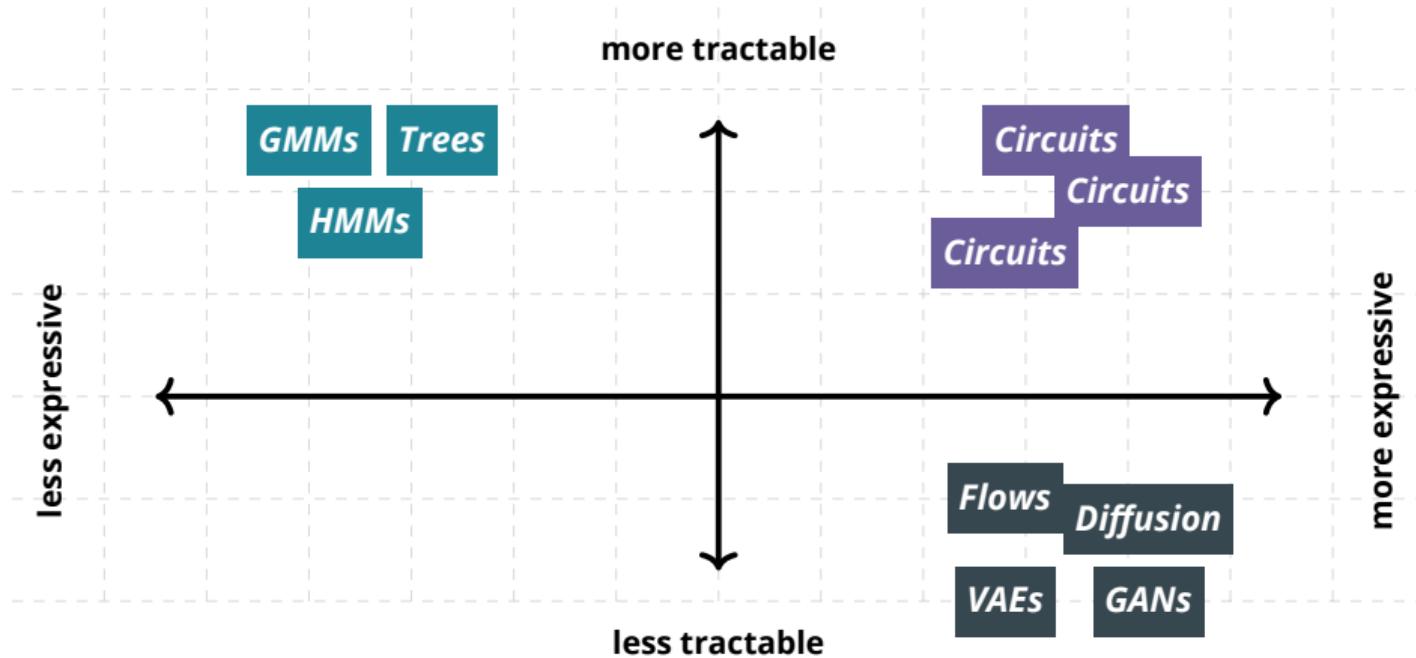




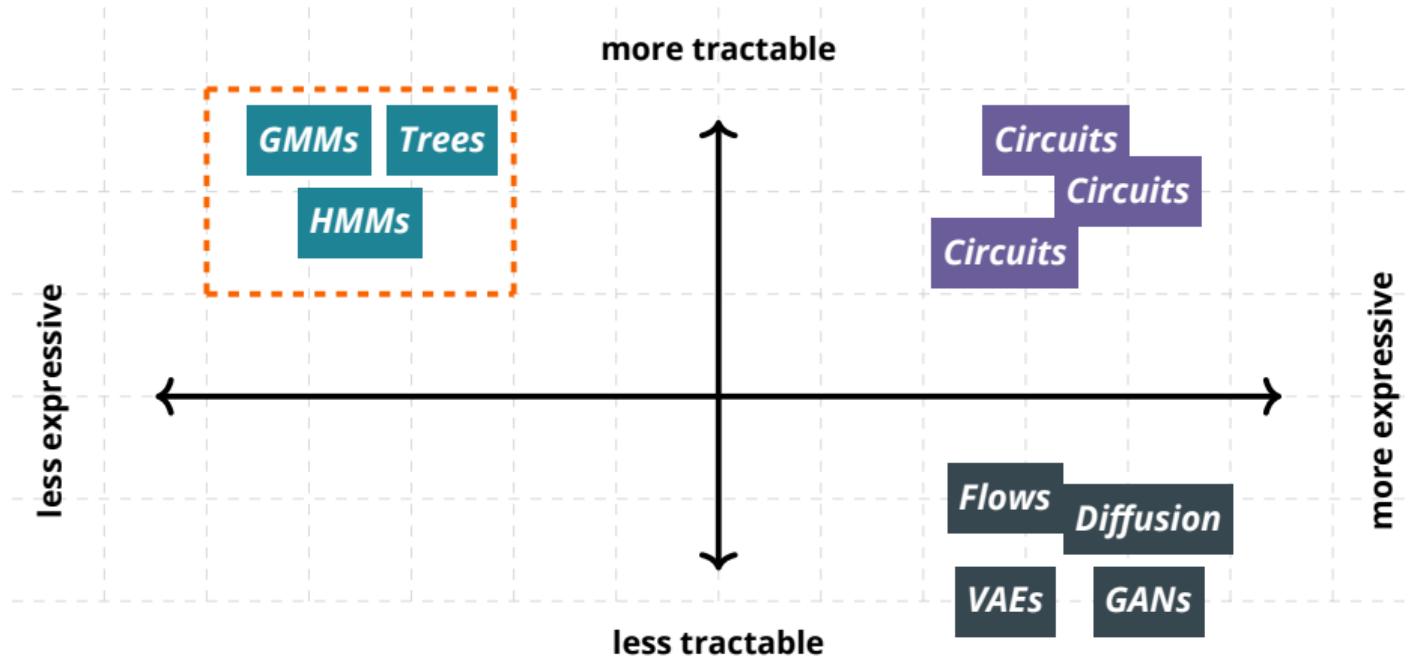
Expressive models are not much tractable...



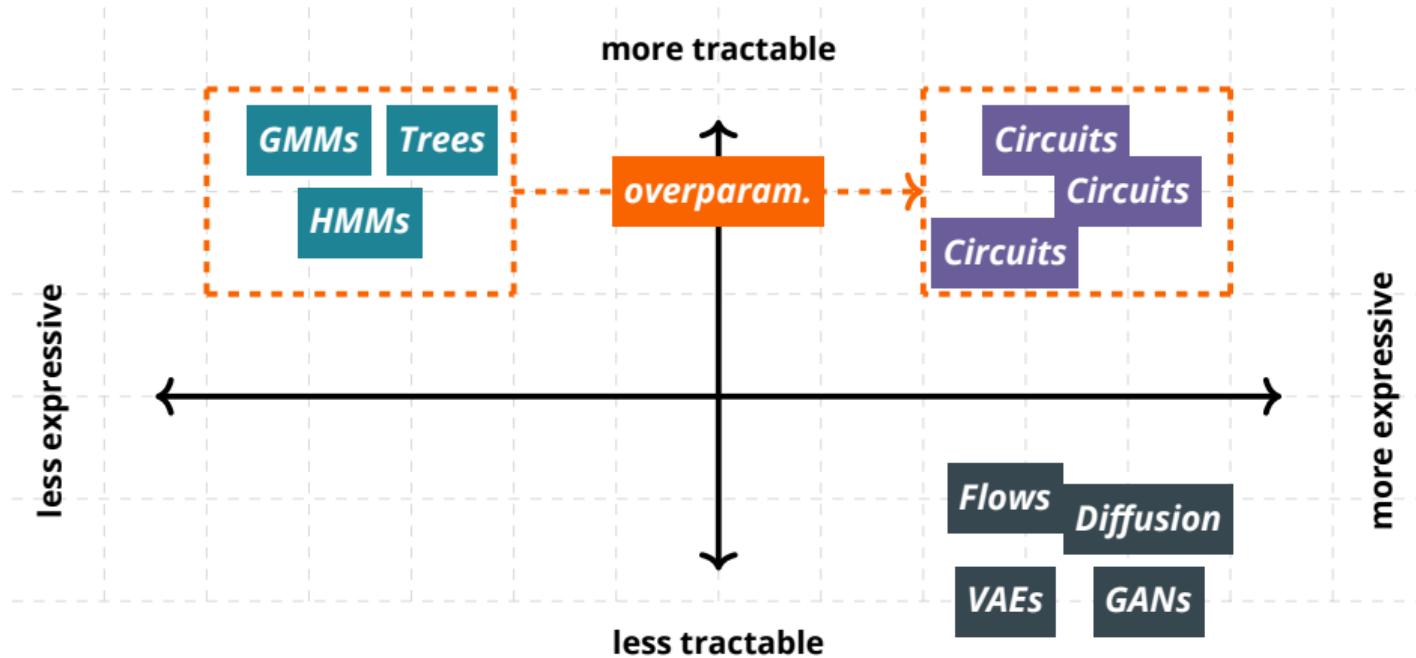
Tractable models are not that expressive...



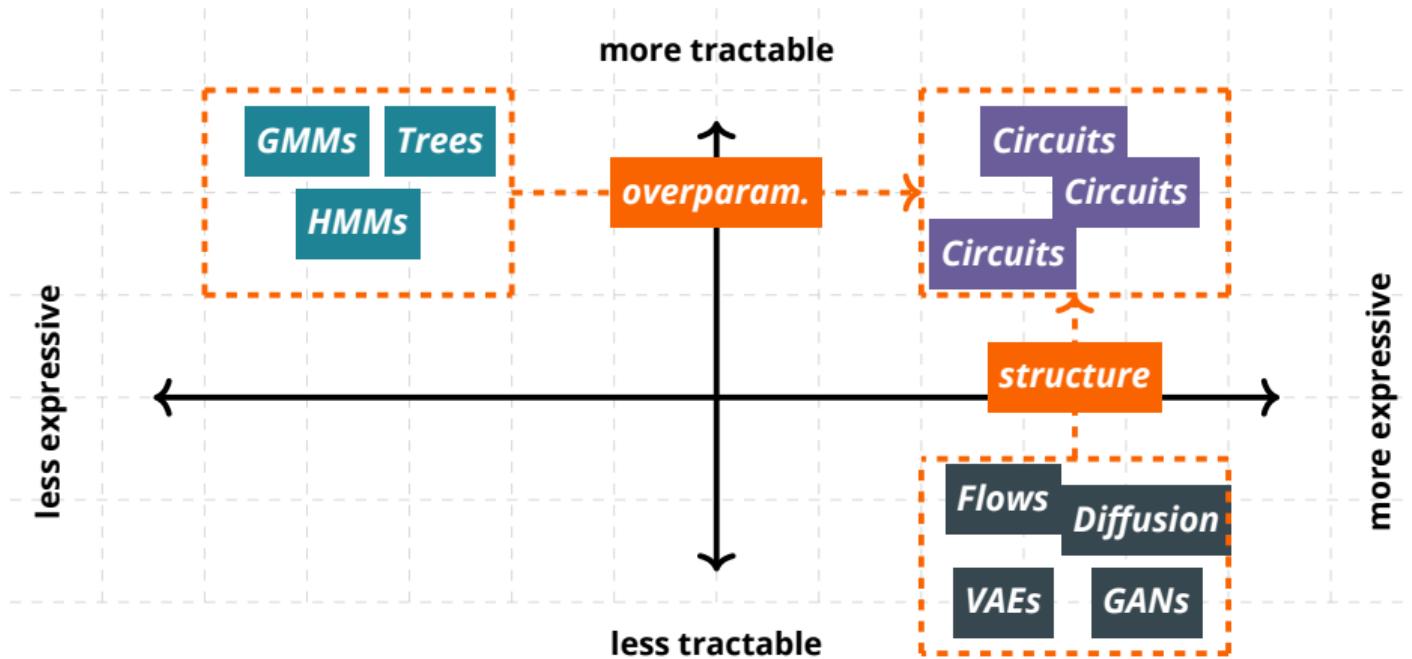
Circuits can be both expressive and tractable!



Start simple...



then make it more expressive!



impose structure!