DTY 1 – NIPS 2018 – TUTORIALS 03/12/18

TUTORIAL 1 – Scalable Bayesian Inference – D. Dunson (Duke)

- Interpretable uncertainty quantification / characterization (UQ)
  → Bayesian stats ⇒ one prob. model perspective ⇒ (outcomes and uncert.)
- Var. Inf. ⇒ another way of doing UQ → very constrained posterior estimate
  ↳ vs. robustness of two flexible models
  ↳ less robust inferences as sample size n grows ⇒ (very) significant

Notation Bayes: $\pi_n(\theta|y^{(n)}) = \dfrac{\pi(\theta)\,L(y^{(n)}|\theta)}{\int \pi(\theta)\,L(y^{(n)}|\theta)\,d\theta}$, denominator $L(y^{(n)})$ → ML/Evidence

→ characterizes uncertainty in θ or $f(\theta)$ ↳ e.g. NN (interpretability)

- Problem: approx. of intractable high-dim. integral
  → 'doubly-intractable': also prior (LH not given)
  → 'intractable' / non-conjugate models
  ↳ use tractable class of distr. to approx.:

① Bayesian CLT (von Mises): $\pi_n(\theta|y^{(n)}) \approx N(\hat{\mu}_n, \Sigma_n)$ as $n \to \infty$
(large sample Gaussian approx.)
  → ass. smooth + differentiable likelihood
  → true θ_0 in interior of param space

② Laplace approx.: Good job at 1st + 2nd moment approx.

- Alternative: Define approx. class of $\pi(\theta)$
  → E.g. element of exp. family ⇒ minimize divergence from (KL, α, Bregman)
  → VI basically: throwing away intractable term ⇒ lost Bayesian UQ!
    ↳ don't know how accurate
    ↳ Fix-ups ⇒ Giordano, Broderick, Jordan (2015)

- MCMC methods ⇒ sequential algo ⇒ obtain correlated draws from posterior
  → samples can be used in very many ways! ⇒ flexible inference
    ↳ summaries of posterior of any fct. $f(\theta)$
  → constructs MC with stationary distr. $\pi_n(\theta \mid y^{(n)})$ ⇒ tractable level
  → MH algo: ① sample proposal: $\theta^* \sim g(\theta^{(s-1)})$ ⇒ extends MC

  ② accept: $\theta^{(s)} = \theta^*$ w.p. $\min\left\{1, \frac{\pi(\theta^*) L(y^{(n)} \mid \theta^*)}{\pi(\theta^{(s-1)}) L(y^{(n)} \mid \theta^{(s-1)})} \frac{g(\theta^{(s-1)})}{g(\theta^*)}\right\}$

  → how to choose g?! ⇒ Gibbs: draw subsets from θ from cond.
  ⇒ Random Walk: Tunable variance
  ⇒ HMC/Langevin: Exploit gradient info

- Scalability

COST PER ITERATION ① bottleneck

  → cost of sampling ⊕ evaluate acceptance prob. (→ eval. LH)
  → similar to most optimization procedures

CONVERGENCE PROBLEMS ② bottleneck

  → No indep. samples ⇒ auto-correlated / slow mixing
  → MLE/mode should not explored with simple algo

- Big N solutions

① Embarrassingly parallel (EP) MCMC: Run in parallel on small subsets
② Approximate MCMC: approx. exp. to evaluate transition kernels
  → Trade-off: accuracy vs. computational budget
③ C-Bayes: conditioning on observed data being in small neighborhood
④ Hybrid MCMC:

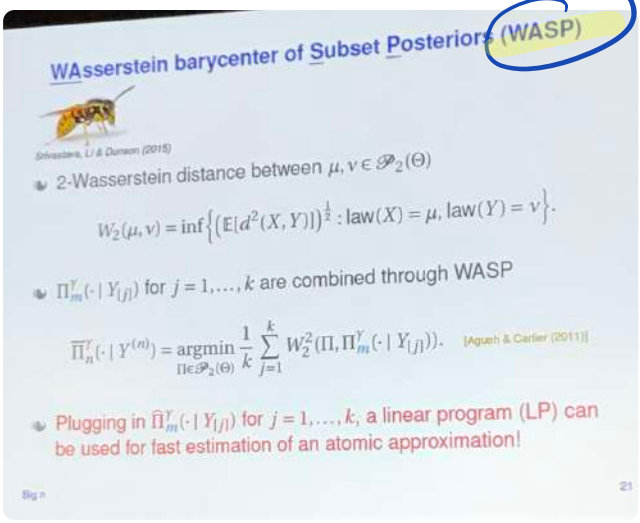
① EP-MCMC: Combine noisy subset posteriors

→ typically reduces noise ⇒ stochastic approx. → as in SGD

↳ $\pi_m(\theta \mid y^{[k]})$ ⇒ γ weights the respective subset posterior

→ Min Wasserstein = solution to discrete optimal transport problem

↳ linear program solver



→ Posterior Interval Estimation (PIE)

↳ 1-d

↳ WASP: quantiles = simple averages of quantiles of sub.p.

↳ run MCMC for subsets and average quantiles! (for any fct. of interest)

⇔ bag of little bootstraps

- Robustness: If model is not exact ⇒ model complexity will explode / grow larger than optimal! → model misspecification!

→ Coarsened posterior: idealized vs. observed data (coarsened)

$\pi(\theta \mid X_{1:n} = x_{1:n})$ vs. $\pi(\theta \mid d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < R)$

↳ model misspecification

↳ $R \sim \exp(\alpha)$ ⇒ power posterior approx.!

if d is chosen to be relative entropy

→ can show robustness to small perturbations to coll. posteriors

→ MCMC for misspecified model versions! // also robust to local misspecification

- Big P solutions ⇒ Variable selection for $p >> n$

① Independent Screening ⇒ select via univariate regression p-values & control for FDR ⇒ Benjamini-Hochberg

② Penalized Strategy ⇒ Introduce small bias to reduce variance significantly

↳ horseshoe, generalised double Pareto, Dirichlet laplace: strong shrinkage and 0/no shrinkage of terms.

$$\beta_j \overset{iid}{\sim} N(0, \psi_j \lambda) \; ; \; \psi_j \sim f \; , \; \lambda \sim g$$

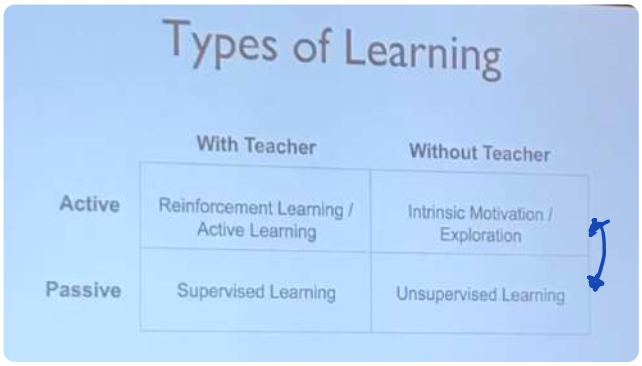
ψ_j: local scale; λ: global scale

→ Problem: Need aggressive shrinkage to combat high-dim. but what if value is not sparse!

- Modular Bayes ⇒ Specify distn and build model for each part!
  → don't have to model everything jointly!
  → separate modelling ⊕ computation
  ↳ shared theme Mixture Models → Bayesian Mix-Prom. Testing Method
- Application to brain network modelling ⇒ Creativity testing!
  ↳ intersubject connectivity in frontal lobe!
- Vision → form of 'Bayesian assistant'
  → 'these question can't be answered' ⇒ But there can be ...

TUTORIAL 2 - Unsupervised Deep Learning - A. Courville (DeepMind), M. Ranzato (FB)

- Active vs. passive // Teacher vs. no teacher → not necessarily binary but continuum!



→ No teacher: rewards / targets based to obtain/define
↳ How to determine reward structure
↳ Unsupervised ⇔ 'human' learning
⇒ rapid generalization to new tasks!
↳ transfer learning ⇒ retrain existing structure
⇒ Problem: Not just or even generalizable
⇒ Hypothesis: Not exist structure (info in tasks
Unsupervised > Supervised > RL
↳ Amount of info

→ LEARN SKILLS AND NOT TASKS!

- Supervised learning outputs info in the targets not in the inputs!
- Overfitting not problem when flexibility in terms of bits vs. # weights/params

- Density modelling ⇒ ML on data instead of targets → learn from DGP!

$\mathcal{D} = \{x\}$, $L(\mathcal{D}) = \sum_{x \in \mathcal{D}} -\log p(x)$

→ Problem: Want to learn everything ⇒ Focus on relative beliefs to be useful!
1. Curse of Dimensionality
2. Not all bits created equal ⇒ low-level details (pixel variables)
3. How to use the learned structure ⇒ representation learning

- Generative Model ⇒ Visual inspection allows us to see what model has learned! ↳ imagine/simulate

- Intrinsic Motivation ⇒ incentivize curiosity
→ driven by compression progress - Schmidhuber (2008)
- Empowered agents ⇒ max. MI between actions and consequences!
↳ max. control of dynamics!

① Autoregressive Models

- split high-dim data into sequence of small pieces ⇒ predict each piece from those before ⇒ slow (sampling)
  ↳ conditioning on refresh state → LSTM / GRU

⇒ $D = \{x\}$, $p(x) = \prod_{t=1}^{T} p(x_t | x_{<t})$, $L(D) = \sum_{x \in D} \sum_{t=1}^{T} - \log p(x_t | x_{<t})$

(+) - simple to define ⇒ pick ordering, easy to sample ⇒ sequential feeding of network output to next time step → (Dreamy)

(−) - very expensive ⇒ training via parallelization, but necessarily slow because of sequential generation!
- Order dependence! // Reader pretty → only known one-step forget
  ↳ to longer dependence generation ?!

- WaveNets, PixelRNN ⇒ language model for images ↳ softmax output for each channel
  ↳ local vs global structure
  ↳ can be improved via latent conditioning!

② Representation Learning

- Internal language to describe data! ⇒ language can emerge from tasks ⇒ Wittgenstein
  ↳ grammar ?!
- E.g. C. Olah Distill Visualizability ⇒ visual vocabulary
- Unsupervised L. should more general than task constrained!
- Want network that 'describes' → re-use to plan, reason, generalize
- Autoencoder ⇒ Variational version → information cost of reconstruction!
- Associated Compression models → latent manifold structure from high-dim
  ↳ Final representation with down stream tasks
- Contrastive Predictive Coding → van den Oord et al
  ↳ Noise contrastive estimation → Mutual info (CPC)

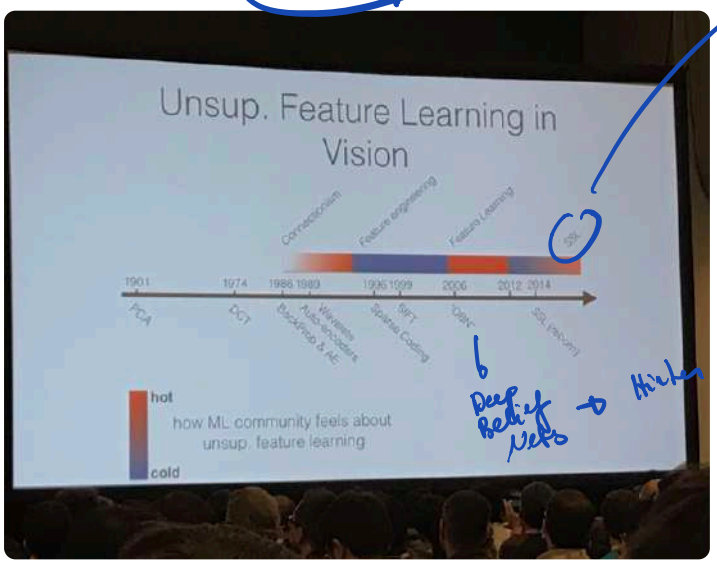
③ Unsupervised RL

- Auxiliary tasks ⇒ max. reward + min unsupervised tasks!

Practical advice

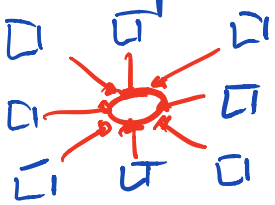
- Visual representations

history



Deep Belief Nets → thriller

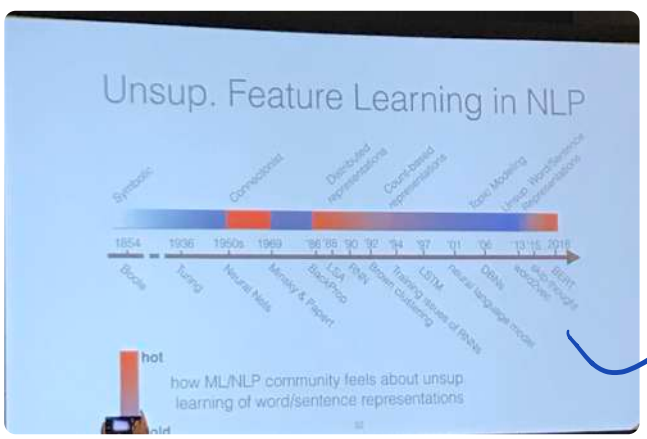
→ Self-supervised learning: ↗ domain knowledge
* Context prediction



transfer properties of to supervised problem!

→ Learning by clustering:
Random filters have selective property
↓
Init random CNN
↳ ① Extract features ⇒ pseudo-classes
② Train network on pseudo-class assignments

- Text representations
→ word carries a lot more info than pixel
↳ dissociate signal ⇒ hard search ↔ any modality modelling
→ word2vec: Meaning of word is determined by its context

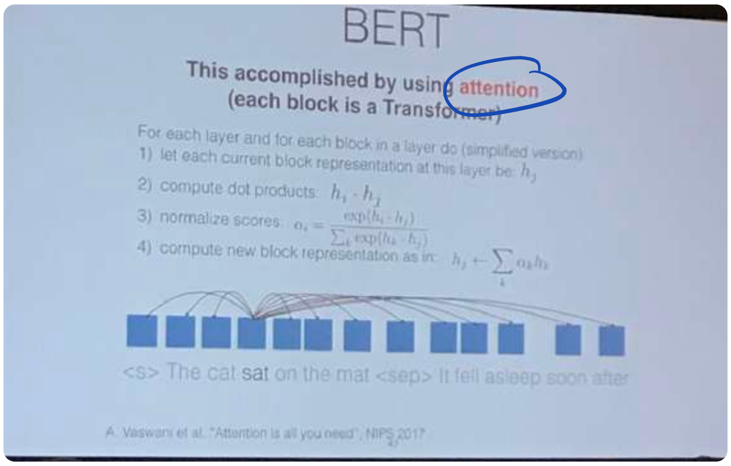


word-embeddings

two texts mean similar things if they have similar effect
↳ dissociate to continuous variables!

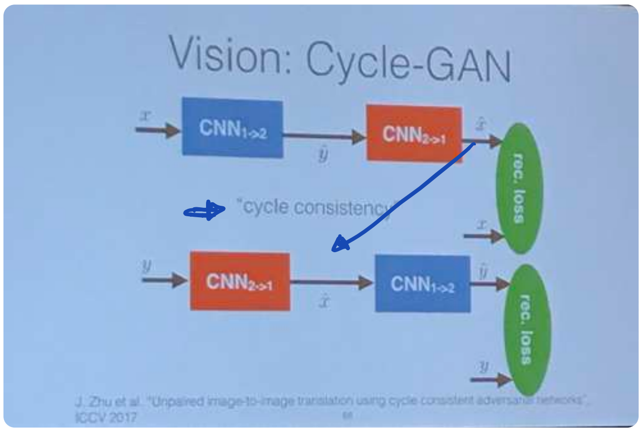
→ BERT: sentences ⇒ consist of very comp. blocks ⇒ transformers
↳ training by hiding blocks and predicting ⇒ 'de-noising'
↳ GLUE benchmark (11 tasks)

- Vision Sampling
→ GANs, autoregressive, OLD ⇒ crucial: architecture!

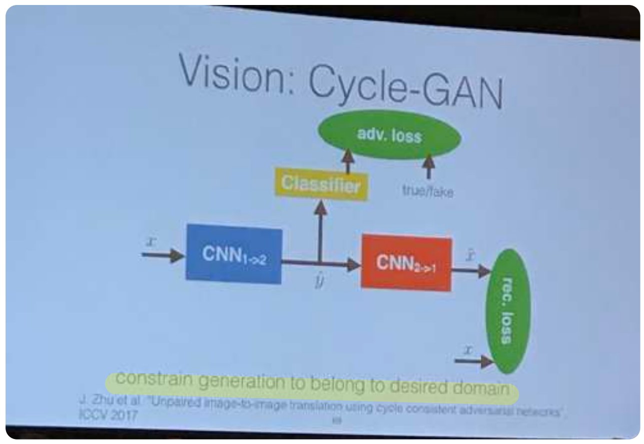


- Text Sampling
  - → Autoregressive, Retrieval-based, comb. of both

- Learning to map between domains
  - → Analogies
  - → Zhu et al (2017) ⇒ Cycle-GAN:



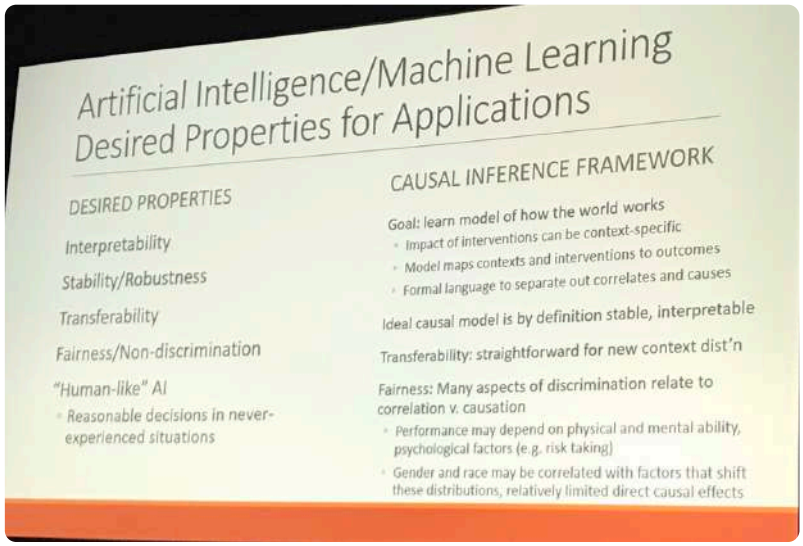
constraint ⇐



→ Unsupervised Machine Translation: MUSE → for words
① Learn separate word embedding per language
② Learn joint space via adversarial training + refinement
↳ Does not work for sentences ⇒ seq2seq as in cycle-GAN w. constraint

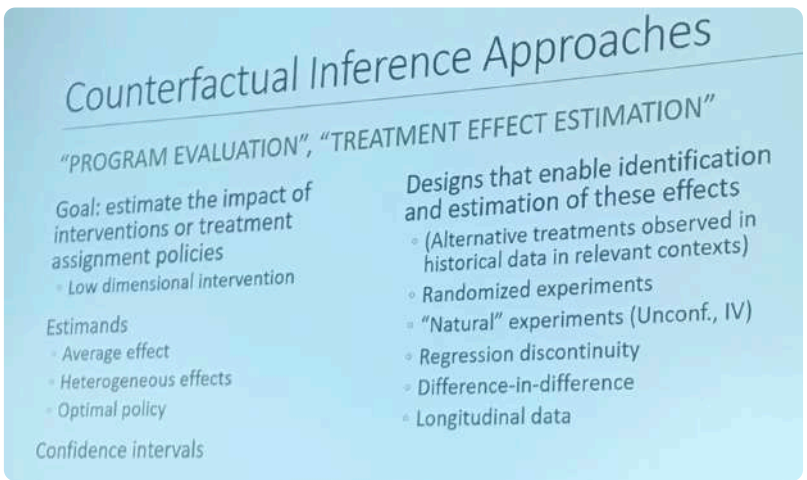
- Open Research
  ① Definition of metrics & down-stream tasks & task-specific metrics
  ② General principles ⇒ Energy-based models → energy: contrastive fct., lower where data has high density ⇒ hard to learn fct.
  ↳ what are positive/neg. samples?! ⇔ bottlenecks

TUTORIAL 3 – Counterfactual Inference – S. Athey (Stanford)

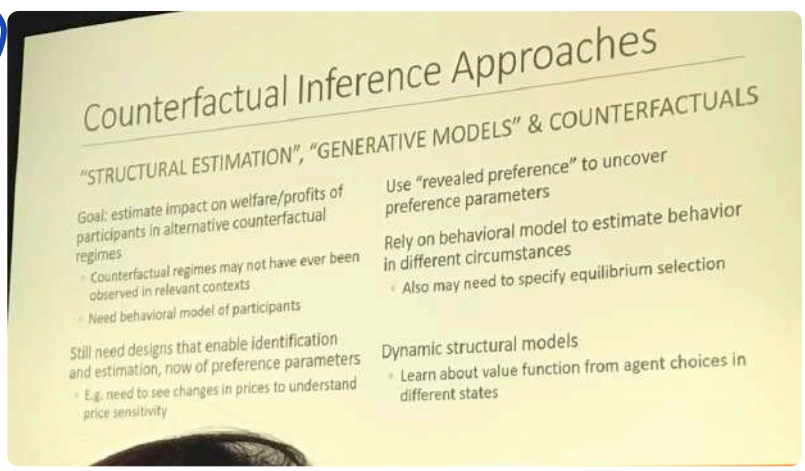


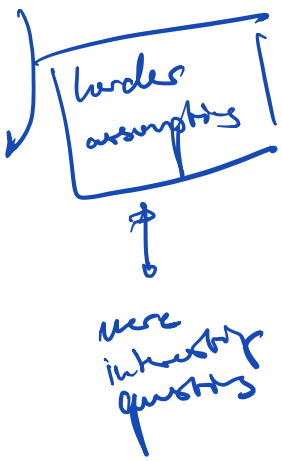
- Example: Contextual Bandits → context influences what best arm is!
- ↕ to fundamentally counterfactual problem!
- Program Eval., Treatment effect estimation
- Randomized experiments / A/B Test.

(I)

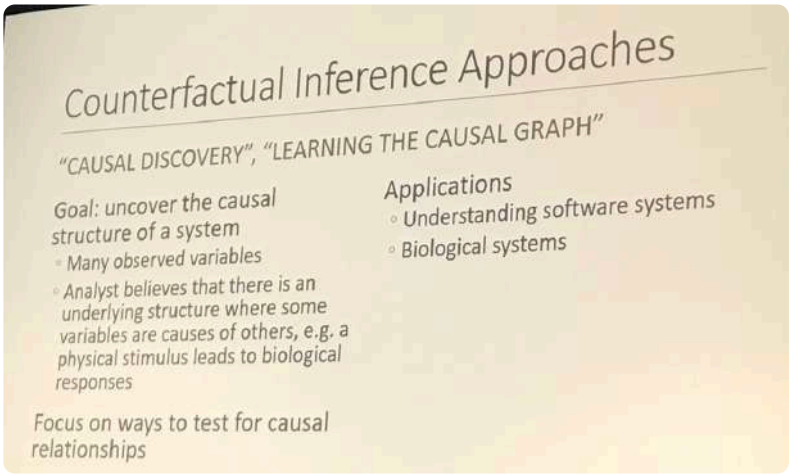


(II)





III



↓ Judea Pearl
↳ a lot more complex discovery than what is often possible in Econometrics

- Identification ⇒ Can Q be answered with infinite data? → Estimation with finite data
- Regularization ⇒ Address omitted variable bias
- Semi-Parametric efficiency theory ⇒ Cross-fitting / out-of-bag estimates of nuisance functions

⇒ orthogonal scores / double robustness

Estimating Average Treatment Effects under Unconfoundedness

$Y_i(1), Y_i(0) \perp W_i \mid X_i$ • Unconfoundedness

- Double Robust: Reweighting by propensity scores ⊕ Outcome modelling

- Predictive model with regularization might omit small variables which together moderate full effect

// Left to continue!