How to Read ML Research Papers

Beyond CS 181 - Lecture 8

Anna Trella

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Welcome Back!

Who thinks reading research papers are difficult?

Who thinks reading research papers are difficult?

Why is reading research papers difficult?

Story Time

I was told to read a 56 page paper on Reinforcement Learning

Story Time

I was told to read a 56 page paper on Reinforcement Learning

Motivation

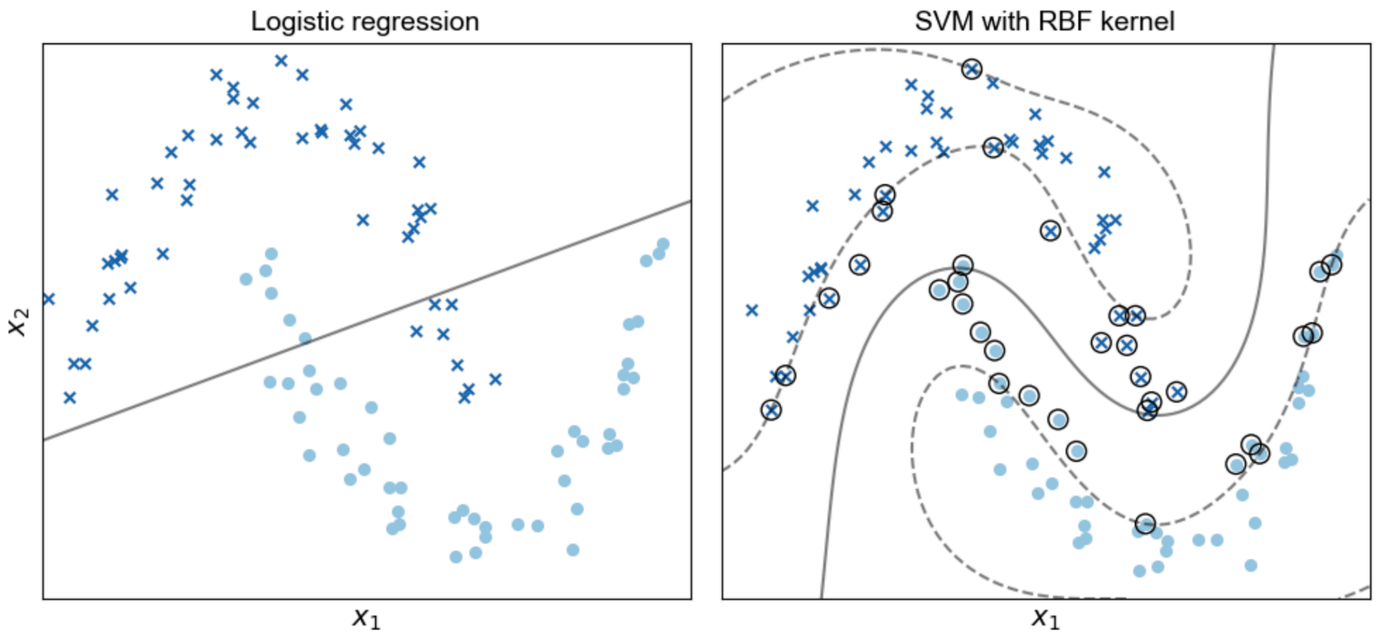
Reading papers means you can build off of someone else's work!

- Papers can offer a new method or a better way of doing an old method (Method Papers)
- Reading papers can give you a good review of existing literature (Review Papers)
- Papers can offer a framework or "best practices" for a general topic (Position or Framework Papers)

Method Papers

I have a new method that's better than all the other methods!

$$k(\mathbf{x}, \mathbf{y}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle_{\mathcal{V}} \approx \mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{y}).$$



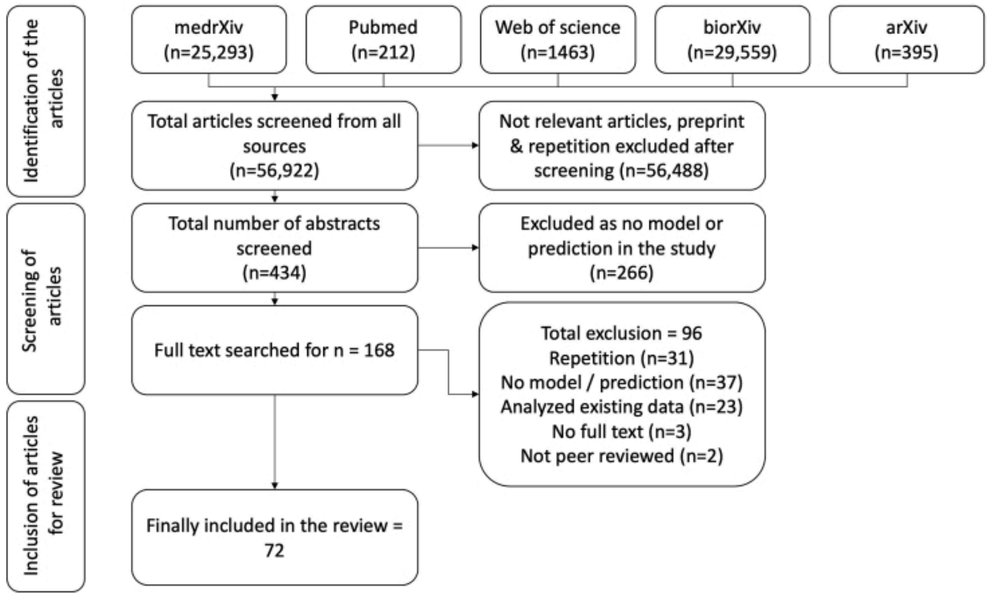
Random Features for Large-Scale Kernel Machines (Rahimi and Recht, 2007)
https://gregorygundersen.com/blog/2019/12/23/random-fourier-features/

Survey Papers

I did a survey of the literature so you don't have to.

Table 1: Assessment of the collected papers on the key properties, which are important for readily comparing and comprehending the differences and limitations of different counterfactual algorithms. Papers are sorted chronologically. Details about the full table is given in appendix A.

	Assumptions		Optimization amortization		CF attributes			CF opt. problem attributes	
Paper	Model access	Model domain	Amortized Inference	Multiple CF	Sparsity	Data manifold	Causal relation	Feature preference	Categorical dist. func
[72]	Black-box	Agnostic	No	No	Changes iteratively	No	No	Yes	-
[111]	Gradients	Differentiable	No	No	L1	No	No	No	-
[104]	Complete	Tree ensemble	No	No	No	No	No	No	-
[74]	Black-box	Agnostic	No	No	L0 and post-hoc	No	No	No	-
[57]	Black-box	Agnostic	No	Yes	Flips min. split nodes	No	No	No	Indicator



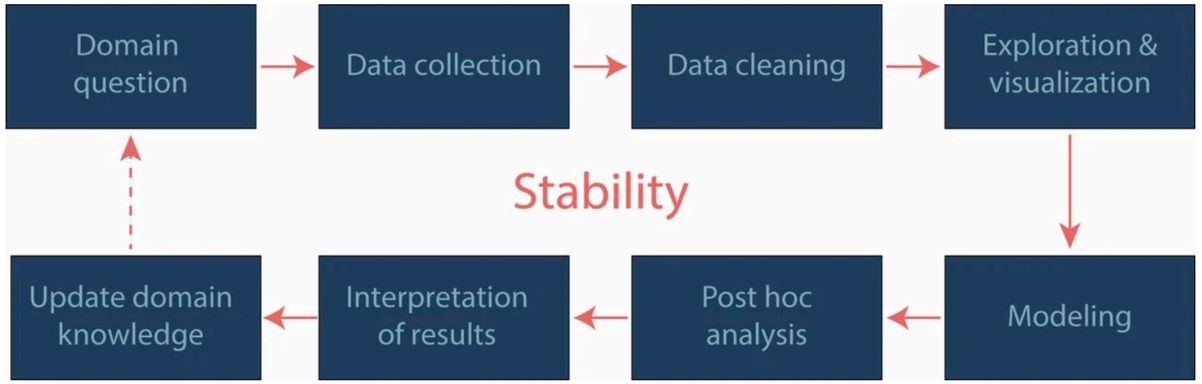
Counterfactual Explanations for Machine Learning: A Review (Verma et. al., 2020)

Systematic review of predictive mathematical models of COVID-19 epidemic (Shankar etl.al., 2021)

Position Papers

I found a problem or gap and I'm making an arguement that folks should focus on this in the future

Fig. 1.



The data science life cycle. The PCS workflow considers predictability, computability, and stability at every step, with a strong emphasis on stability.

Inference methods designed to capture structure in the posterior do not necessarily produce better approximations of the true posterior. In our experiments, we do not see that methods using a richer divergence metric or structured variational family are able to better capture the ground truth posterior. This is likely due to the fact that the true

Veridical Data Science (Yu and Kumbier, 2020)
Quality of Uncertainty Quantification for Bayesian Neural Network Inference (Yao, Pan et. al., 2019)

ML Paper Organization

- Motivation and Potential Impact (Introduction and Abstract)
- Related Work and Contributions (Introduction and Related Works)
- Method
- Experiments
- Future Work and Broader Impact (Conclusion)

Motivation and Potential Impact

- What is the problem? Why is the problem important?
- Why is the problem difficult to solve?

A Motivating Example. Consider a pre-existing clinical dataset of patients (agents). Our goal is to design a personalized treatment plan (policy) for each patient moving forward. Notable challenges include the following: First, each patient only provides a single trajectory of their medical history. Second, each patient is heterogeneous in that they may have a varied response for a given treatment under similar medical conditions; further, the underlying reason for this heterogeneous response across patients is likely unknown and thus not recorded in the dataset. Third, in the absence of an accurate personalized "forecasting" model for a patient's medical outcome given a treatment plan, the treatment assigned is likely to be sub-optimal. This is particularly true for complicated medical conditions like T-Cell Lymphoma. We aim to address the challenges laid out above (offline scarce data, heterogeneity, and sub-optimal policies) so as to develop a personalized "forecasting" model for each patient under any given treatment plan. Doing so will then naturally enable ideal personalized treatment policies for each patient.

Key question. Tackling a scenario like the one described above in a principled manner is the focus of this work. In particular, we seek to answer the following question:

"Can we leverage scarce, offline data collected across heterogeneous agents under unknown, sub-optimal policies to learn a personalized policy for each agent?"

PerSim: Data-efficient Offline Reinforcement Learning with Heterogeneous Agents via Personalized Simulators (Agarwal et. al., 2021)

Related Works and Contribution

- What are their main contributions?
- What have other people tried before?
- What does this paper offer that is different from previous work?

Our Contribution In this work, we provide a novel inferential method for adaptively collected data that makes progress towards addressing the above challenges. Specifically we provide a method for constructing valid confidence regions with the following properties:

1. **Applicable to Non-Markovian Environments:** Our inference methods are valid even when user rewards and contexts are non-stationary and highly dependent over time. They are applicable to both Markovian and non-Markovian environments.
2. **Applicable to Datasets Collected by Algorithms that Learn Across Users:** Adaptive sampling algorithms that learn using the data of multiple users can potentially learn faster, but induce dependence between the collected user data trajectories. Our inference method is applicable to datasets collected by such algorithms because it accounts for induced dependence.
3. **Algorithm Agnostic:** We assume the adaptive sampling algorithm uses policies in a parametric policy class that is sufficiently smooth in the policy parameter and explores sufficiently. Besides these conditions, the validity of our inference method is not affected by potential misspecification of the adaptive sampling algorithm, e.g., if the RL algorithm incorrectly assumes a linear model for the reward or mistakenly assumes that the environment is Markovian, the validity of our inference method is not affected.

Statistical Inference After Adaptive Sampling in Non-Markovian Environments (Zhang et. al., 2022)

Method

- Problem Statement
- Definitions, Assumptions, and Theorems
- Explain their procedure or algorithm

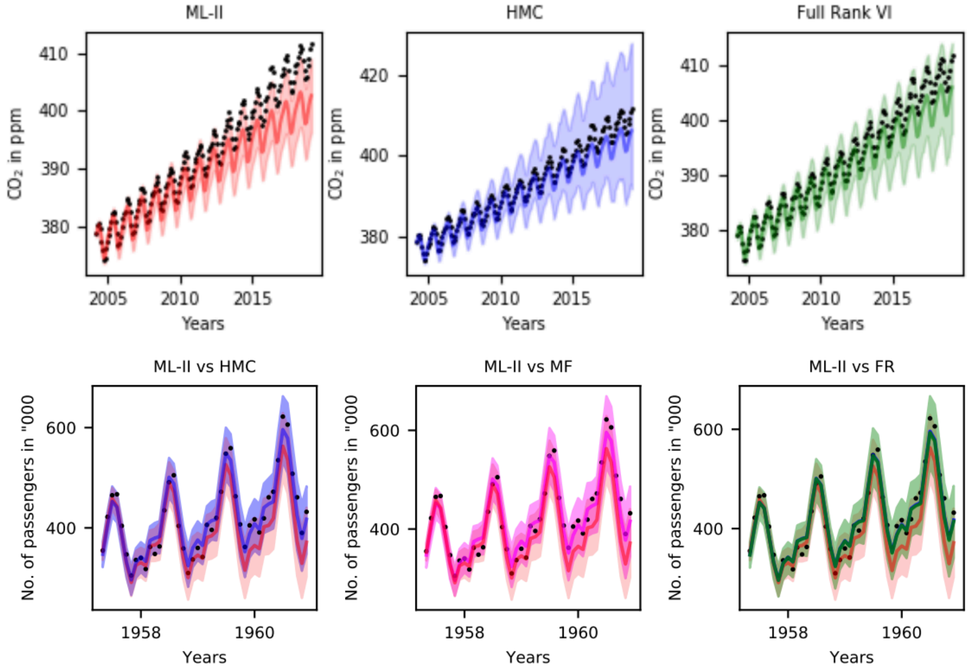
4 PerSim Algorithm

We now detail our proposed algorithm which is composed of two steps: (i) build a personalized simulator for each agent n given offline observations $\mathcal{D}$, which is comprised of a single trajectory per agent; (ii) learn a personalized decision-making policy using MPC.

PerSim: Data-efficient Offline Reinforcement Learning with Heterogeneous Agents via Personalized Simulators (Agarwal et. al., 2021)

Experiments

- What are the takeaways of the experiments they run?
- Do there experiments support their claim?



Data set	CO_2		Wine		Concrete		Airline	
Inputs	$N = 732, d = 1$		$N = 1599, d = 11$		$N = 1030, d = 8$		$N = 144, d = 1$	
Hyperparameters	$\theta = 11$		$\theta = 13$		$\theta = 10$		$\theta = 6$	
Inference Scheme	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD
ML-II	4.230 (0.18)	3.03	0.65 (0.02)	0.98	6.12 (0.39)	3.19	21.08 (2.64)	4.62
HMC (NUTS)	2.37 (0.10)	2.53	0.65 (0.02)	0.97	5.47 (0.38)	3.06	16.47 (2.34)	4.31
Mean-field VI	2.74 (0.12)	2.05	0.65 (0.02)	0.97	5.55 (0.38)	3.07	16.86 (2.49)	4.36
Full Rank VI	2.56 (0.12)	1.99	0.64 (0.02)	0.97	5.52 (0.35)	3.17	16.78 (2.47)	4.34

Table 1: A comparison of approximate inference schemes for fully Bayesian GPR. For both metrics lower is better, the value in parenthesis denotes standard error of the RMSE.

Approximate Inference for Fully Bayesian Gaussian Process Regression (Lachland and Rasmussen, 2020)

Experiments

Data set	CO_2		Wine		Concrete		Airline	
Inputs	$N = 732, d = 1$		$N = 1599, d = 11$		$N = 1030, d = 8$		$N = 144, d = 1$	
Hyperparameters	$\theta = 11$		$\theta = 13$		$\theta = 10$		$\theta = 6$	
Inference Scheme	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD
ML-II	4.230 (0.18)	3.03	0.65 (0.02)	0.98	6.12 (0.39)	3.19	21.08 (2.64)	4.62
HMC (NUTS)	2.37 (0.10)	2.53	0.65 (0.02)	0.97	5.47 (0.38)	3.06	16.47 (2.34)	4.31
Mean-field VI	2.74 (0.12)	2.05	0.65 (0.02)	0.97	5.55 (0.38)	3.07	16.86 (2.49)	4.36
Full Rank VI	2.56 (0.12)	1.99	0.64 (0.02)	0.97	5.52 (0.35)	3.17	16.78 (2.47)	4.34

Table 1: A comparison of approximate inference schemes for fully Bayesian GPR. For both metrics lower is better, the value in parenthesis denotes standard error of the RMSE.

Approximate Inference for Fully Bayesian Gaussian Process Regression (Lachland and Rasmussen, 2020)

Experiments

Baseline datatsets

Data set	CO_2		Wine		Concrete		Airline	
Inputs	$N = 732, d = 1$		$N = 1599, d = 11$		$N = 1030, d = 8$		$N = 144, d = 1$	
Hyperparameters	$\theta = 11$		$\theta = 13$		$\theta = 10$		$\theta = 6$	
Inference Scheme	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD
ML-II	4.230 (0.18)	3.03	0.65 (0.02)	0.98	6.12 (0.39)	3.19	21.08 (2.64)	4.62
HMC (NUTS)	2.37 (0.10)	2.53	0.65 (0.02)	0.97	5.47 (0.38)	3.06	16.47 (2.34)	4.31
Mean-field VI	2.74 (0.12)	2.05	0.65 (0.02)	0.97	5.55 (0.38)	3.07	16.86 (2.49)	4.36
Full Rank VI	2.56 (0.12)	1.99	0.64 (0.02)	0.97	5.52 (0.35)	3.17	16.78 (2.47)	4.34

Table 1: A comparison of approximate inference schemes for fully Bayesian GPR. For both metrics lower is better, the value in parenthesis denotes standard error of the RMSE.

Approximate Inference for Fully Bayesian Gaussian Process Regression (Lachland and Rasmussen, 2020)

Experiments

Data set	CO_2		Wine		Concrete		Airline	
Inputs	$N = 732, d = 1$		$N = 1599, d = 11$		$N = 1030, d = 8$		$N = 144, d = 1$	
Hyperparameters	$\theta = 11$		$\theta = 13$		$\theta = 10$		$\theta = 6$	
Inference Scheme	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD
ML-II	4.230 (0.18)	3.03	0.65 (0.02)	0.98	6.12 (0.39)	3.19	21.08 (2.64)	4.62
HMC (NUTS)	2.37 (0.10)	2.53	0.65 (0.02)	0.97	5.47 (0.38)	3.06	16.47 (2.34)	4.31
Mean-field VI	2.74 (0.12)	2.05	0.65 (0.02)	0.97	5.55 (0.38)	3.07	16.86 (2.49)	4.36
Full Rank VI	2.56 (0.12)	1.99	0.64 (0.02)	0.97	5.52 (0.35)	3.17	16.78 (2.47)	4.34

Method they're trying to beat

Table 1: A comparison of approximate inference schemes for fully Bayesian GPR. For both metrics lower is better, the value in parenthesis denotes standard error of the RMSE.

Approximate Inference for Fully Bayesian Gaussian Process Regression (Lachland and Rasmussen, 2020)

Experiments

Data set	CO_2		Wine		Concrete		Airline	
Inputs	$N = 732, d = 1$		$N = 1599, d = 11$		$N = 1030, d = 8$		$N = 144, d = 1$	
Hyperparameters	$\theta = 11$		$\theta = 13$		$\theta = 10$		$\theta = 6$	
Inference Scheme	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD	RMSE	NLPD
ML-II	4.230 (0.18)	3.03	0.65 (0.02)	0.98	6.12 (0.39)	3.19	21.08 (2.64)	4.62
HMC (NUTS)	2.37 (0.10)	2.53	0.65 (0.02)	0.97	5.47 (0.38)	3.06	16.47 (2.34)	4.31
Mean-field VI	2.74 (0.12)	2.05	0.65 (0.02)	0.97	5.55 (0.38)	3.07	16.86 (2.49)	4.36
Full Rank VI	2.56 (0.12)	1.99	0.64 (0.02)	0.97	5.52 (0.35)	3.17	16.78 (2.47)	4.34

Variations of their method

Table 1: A comparison of approximate inference schemes for fully Bayesian GPR. For both metrics lower is better, the value in parenthesis denotes standard error of the RMSE.

Approximate Inference for Fully Bayesian Gaussian Process Regression (Lachland and Rasmussen, 2020)

Future Work

- Authors may acknowledge the drawbacks or gaps in their work.
- Also gives a good idea of future research directions that could be interesting to explore!

Issues around choosing a model for decision-making also need to be explored more. Here we chose an algorithm for contextual multi-armed bandit problems, as this algorithm may be particularly suitable for mobile health studies. There is a lack of research that compares the effectiveness of different RL models and assesses what kind of problems within mobile health they should be applied to.

Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions (Figueroa, 2021)

Appendix

- Sometimes the main paper is an advertisement for the appendix and the most helpful parts are in the appendix.

C Detail of variational inference

Prior Recall the prior over $\tau, \boldsymbol{\lambda}, \phi_\tau, \boldsymbol{\phi_\lambda}$ after reparameterization is

$$\begin{aligned}
\tau^2|\phi_\tau &\sim \mathrm{IG}(\tau^2|1/2, \phi_\tau^{-1}),\\
\phi_\tau &\sim \mathrm{IG}(\phi_\tau|1/2, A^{-1}),\\
\lambda_i^2|\phi_{\lambda_i} &\sim \mathrm{IG}(\lambda_i^2|1/2, \phi_{\lambda_i}^{-1}), \quad i = 1\dots m,\\
\phi_{\lambda_i} &\sim \mathrm{IG}(\phi_{\lambda_i}|1/2, B^{-1}).
\end{aligned}$$

Variational distribution The variational distributions of τ and λ_i are in the form of log normal distribution.

$$\begin{aligned}
q(\tau^2) &= \mathrm{Lognormal}(\tau^2|m_\tau, \sigma_\tau^2)\\
q(\lambda_i^2) &= \mathrm{Lognormal}(\lambda_i^2|m_{\lambda_i}, \sigma_{\lambda_i}^2), \quad i = 1\dots m.
\end{aligned}$$

On the other hand, the variational distributions of the auxiliary variables ϕ_τ and ϕ_{λ_i} remain as Inverse Gamma distributions

$$\begin{aligned}
q(\phi_\tau) &= \mathrm{IG}(\phi_\tau|s_\tau, r_\tau)\\
q(\phi_{\lambda_i}) &= \mathrm{IG}(\phi_{\lambda_i}|s_{\lambda_i}, r_{\lambda_i}), \quad i = 1\dots m.
\end{aligned}$$

KL divergence As $\mathbf{w} = \{\tau, \boldsymbol{\lambda}, \phi_\tau, \boldsymbol{\phi_\lambda}\}$, the KL divergence $\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}))$ becomes

$$\begin{aligned}
&\mathrm{KL}\left(q(\tau^2)q(\phi_\tau)\prod_i q(\lambda_i^2)q(\phi_{\lambda_i})||p(\tau^2|\phi_\tau)p(\phi_\tau)\prod_i p(\lambda_i|\phi_{\lambda_i})p(\lambda_i)\right)\\
=&H[q(\tau^2)] + H[q(\phi_\tau)] + \sum_i H[q(\lambda_i^2)] + \sum_i H[q(\phi_{\lambda_i})] + \qquad (9)\\
&\mathbb{E}_{q(\tau^2)q(\phi_\tau)}[\log p(\tau^2|\phi_\tau)] + \mathbb{E}_{q(\phi_\tau)}[\log p(\phi_\tau)] + \sum_i \mathbb{E}_{q(\lambda_i)q(\phi_{\lambda_i})}[\log p(\lambda_i^2|\phi_{\lambda_i})] + \sum_i \mathbb{E}_{q(\phi_{\lambda_i})}[\log p(\phi_{\lambda_i})].
\end{aligned}$$

where $H[\cdot]$ denotes the entropy of a distribution.

Learning Compositional Sparse Gaussian Processes with a Shrinkage Prior (Tong, 2021)

One Approach

“Thinking, Fast and Slow”

- **Have a goal:** “I want to _____. I am reading this paper for _____.”
- **Spend 30 minutes surveying the paper:**
  - Abstract and Introduction
  - Do the authors do what they say they do?
  - Decide if it’s worth diving deeper