Uncertainty and Interpretability in Modern ML Algorithms

Anthony Khong
25 July 2018

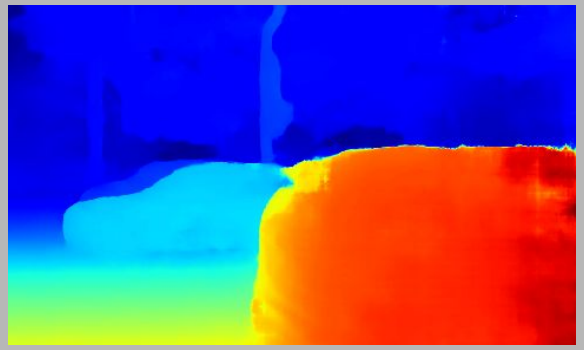
Plan of Attack

1. Why Do We Care About Uncertainty?
2. Types of Uncertainty
3. Modelling Uncertainty in Modern ML
4. Characterising Model Interpretability
5. Unboxing the Black Box

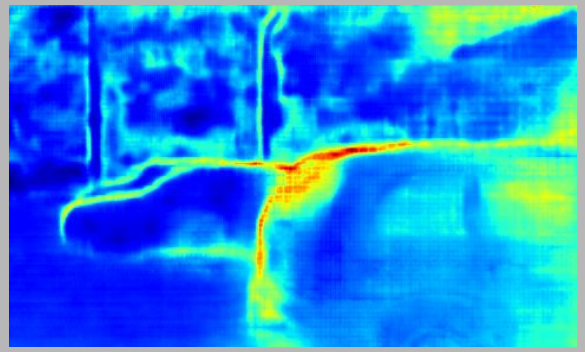
1. Why Do We Care About Uncertainty?

Case Study I: Self-Driving Cars

Image

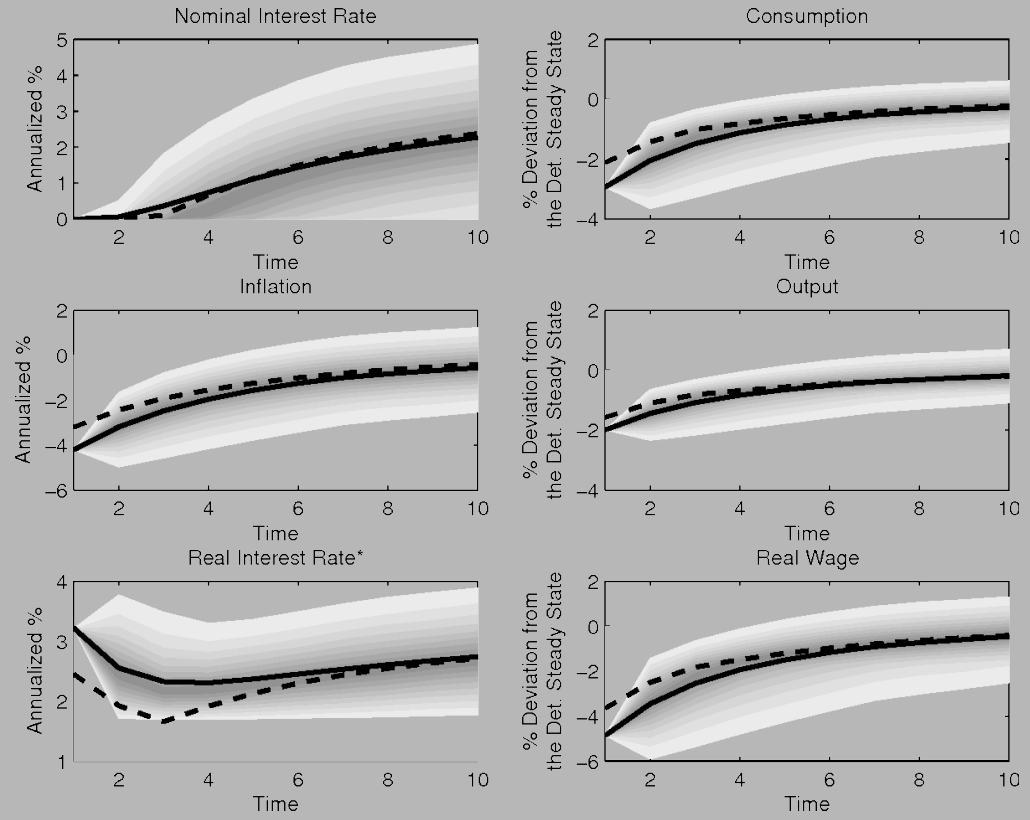


Depth Prediction

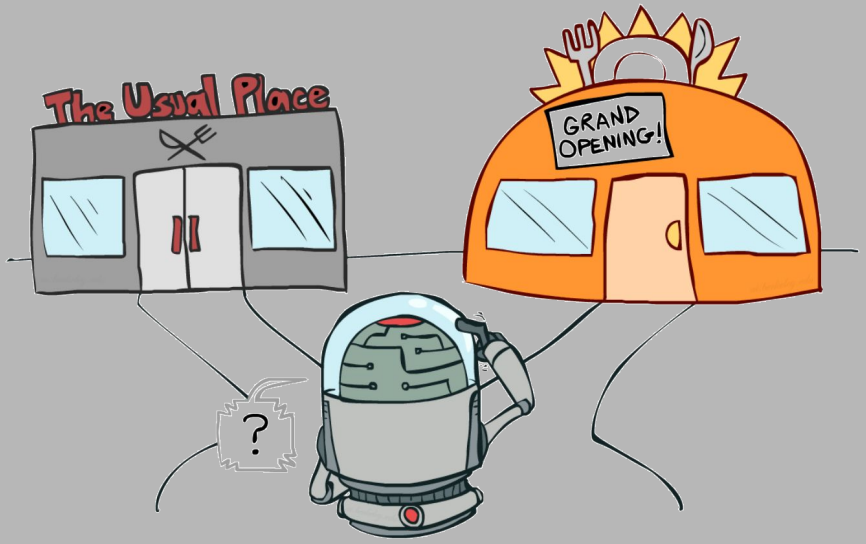


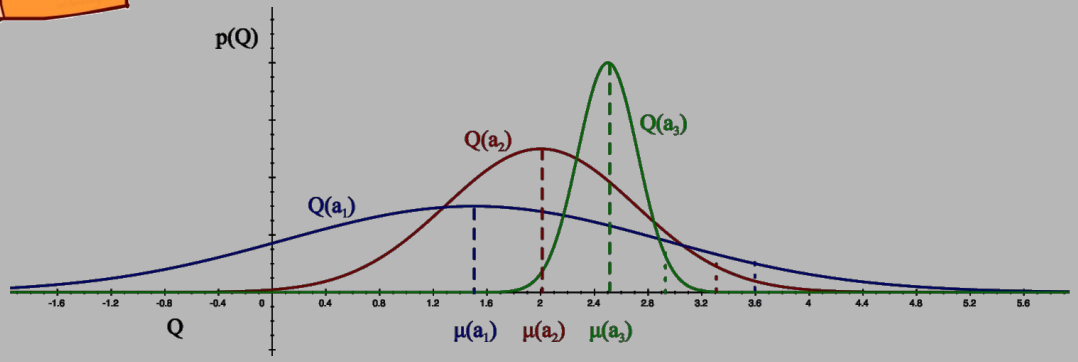
Depth Uncertainty

Case Study II: Policy Uncertainty



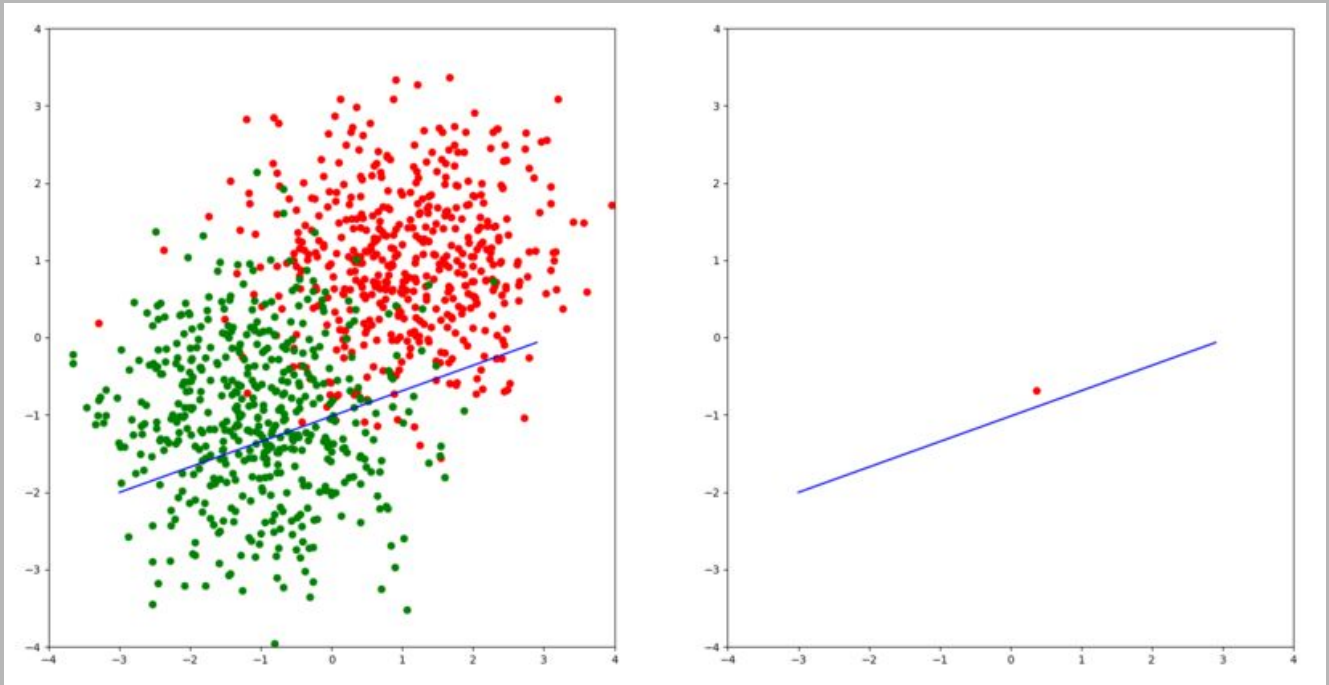
Bonus I: Multi-Armed Bandits





"A Modern Bayesian Look at Multi-Armed Bandit"
Scott (2010)

Bonus II: Active Learning



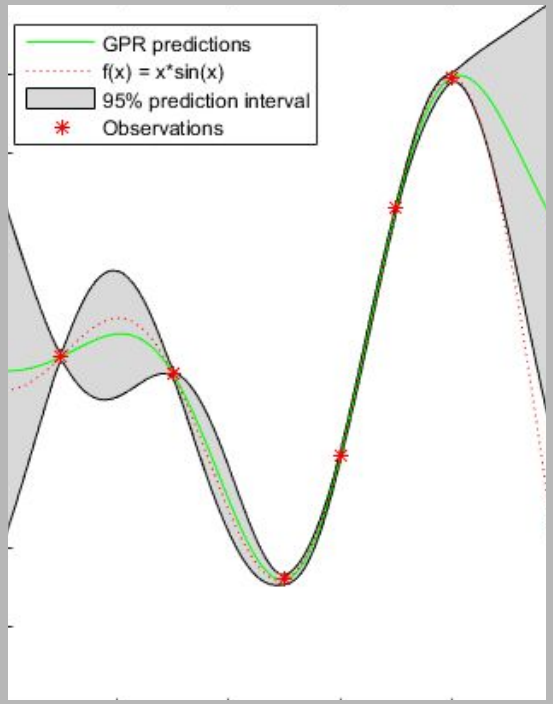
"Active Learning"
Settles (2012)

2. Types of Uncertainty

Epistemic

Uncertainty over data generation.

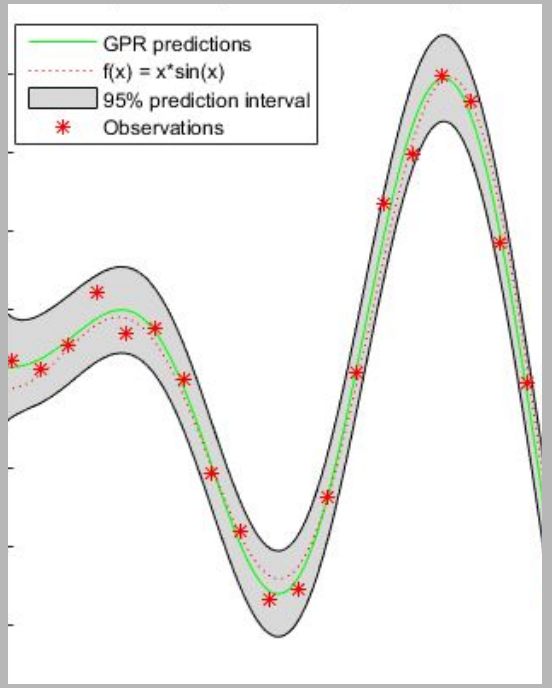
→ parameter/model uncertainty



Aleatoric

Uncertainty inherent to the system.

→ exogenous uncertainty.



Regression Task with No Uncertainty

$$\underset{N\times 1}{\boldsymbol{y}} = \underset{N\times K}{f(\boldsymbol{X})} + \underset{N\times 1}{\boldsymbol{\varepsilon}}$$

$$f^* = \arg\min_{f\in\mathbb{F}} L(\boldsymbol{y}, f(\boldsymbol{X}))$$

Regression with Learned Variance

$$y_i = \mu(\boldsymbol{x}_i) + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}\Big(0, \sigma^2(\boldsymbol{x}_i)\Big) \;\forall i$$

$$f^* = \arg\max_{\mu,\sigma^2} \Big\{ \sum_i \ell(y_i; \mu(\boldsymbol{x}_i), \sigma^2(\boldsymbol{x}_i)) \Big\}$$

What about model/parameter uncertainty?

Good Ol' Linear Regression Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_K)$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

$$\hat{\sigma}^2 = \frac{1}{N-K} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

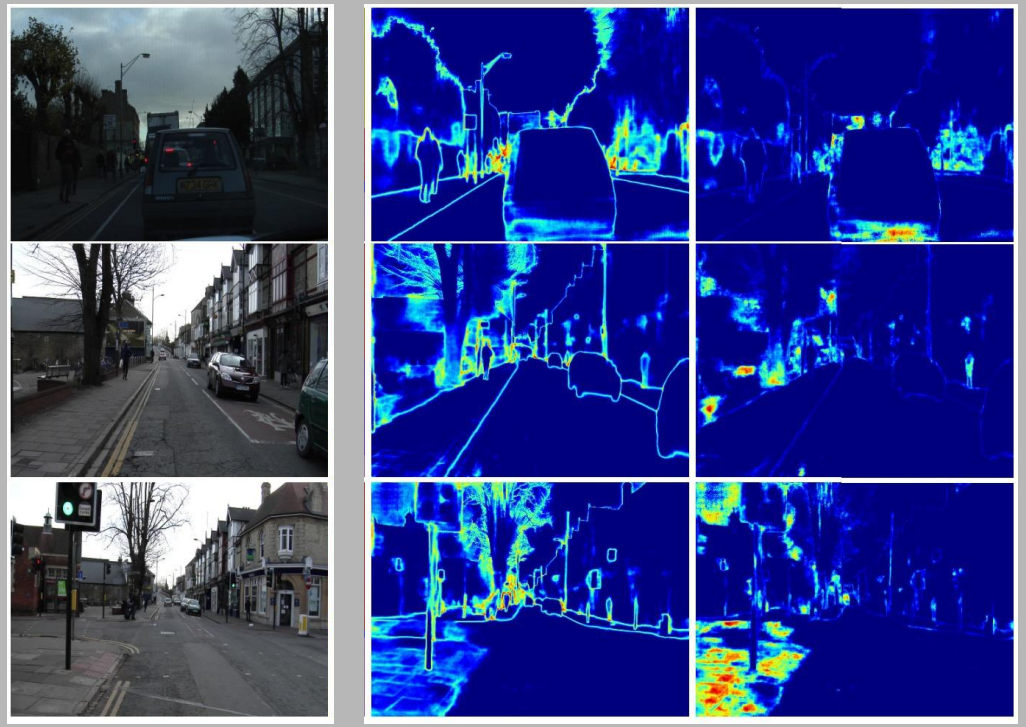
$$\Rightarrow \hat{\mathbb{V}}(y_* - \boldsymbol{x}_*^T \hat{\boldsymbol{\beta}}) = ?$$

What Exactly Did We Lose?

$$\hat{\sigma}^2 = \frac{1}{N-K}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

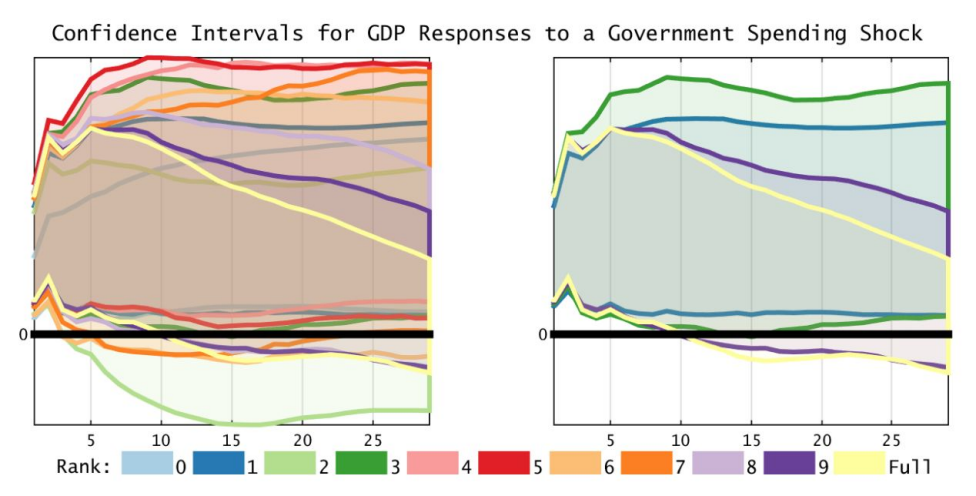
$$\hat{\mathbb{V}}(y_* - \boldsymbol{x}_*^T\hat{\boldsymbol{\beta}}) = \underbrace{\hat{\sigma}^2\boldsymbol{x}_*^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_*}_{\text{Epistemic}} + \underbrace{\hat{\sigma}^2}_{\text{Aleatoric}}$$

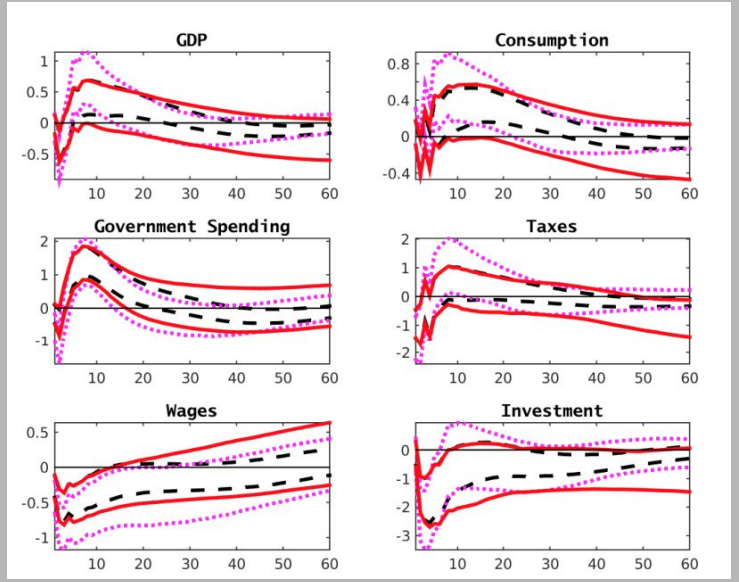
Can We Ignore Model Uncertainty?



Aleatoric **Epistemic**

Epistemic Uncertainty in Impulse Responses





"Inference for Impulse Responses under Model Uncertainty"
Lieb et al. (2018)

3. Modelling Uncertainty in Modern ML

The Classical Gold Standard - MLE

$$\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{\theta}) \rightarrow \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \left\{ \log p(\boldsymbol{y}|\boldsymbol{\theta}) \right\}$$

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \mathcal{I}^{-1}(\boldsymbol{\theta})) \rightarrow \hat{\mathbb{V}}(\boldsymbol{\theta}) = -\left[\frac{\partial^2 \log p(\boldsymbol{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right]^{-1}$$

Sometimes Possible for Some Black Boxes

2.1. *The Infinitesimal Jackknife for Random Forests.* In order to estimate $\sigma^2(\hat{y})$, we use the infinitesimal jackknife (or non-parametric delta-method) estimator $\widehat{V}_{IJ}$ for bagging introduced by Efron [15]. This estimator can be computed using a particularly simple formula:

$$\widehat{V}_{IJ}\left(x;\, Z_1,\, ...,\, Z_n\right) = \sum_{i=1}^{n} \mathrm{Cov}_*\left[T\left(x;\, Z_1^*,\, ...,\, Z_n^*\right),\, N_i^*\right], \tag{5}$$

"Asymptotic Theory of Random Forests"
Wager (2014)

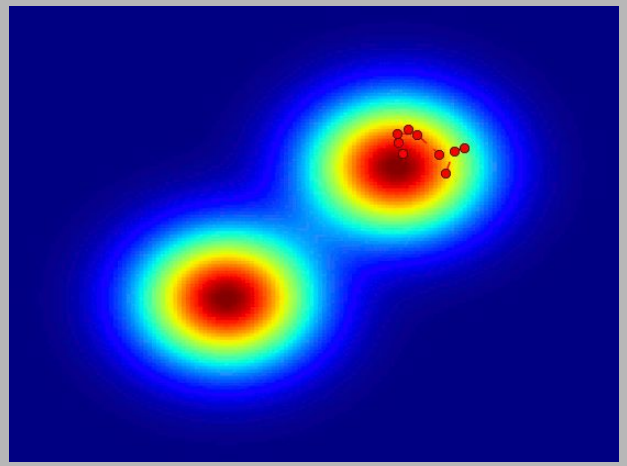
Bayesian Inference

$$\begin{aligned} \text{Likelihood:} \quad & \mathcal{D} \sim p(\mathcal{D}|\boldsymbol{\theta}) \\ \text{Prior:} \quad & \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \\ \text{Posterior:} \quad & \pi(\boldsymbol{\theta}|\mathcal{D}) \propto \pi(\boldsymbol{\theta}) p(\mathcal{D}|\boldsymbol{\theta}) \end{aligned}$$

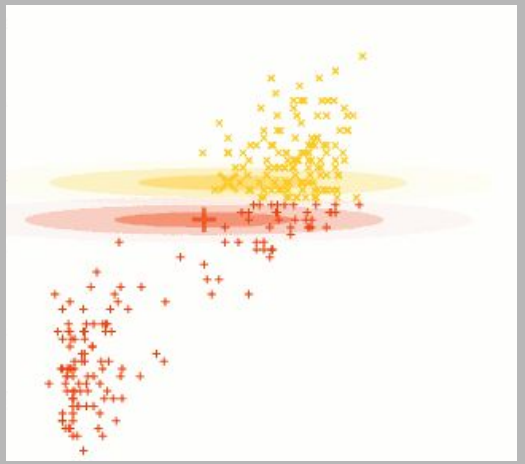
Posterior Computations

$$\phi = \int_{\boldsymbol{\theta} \in \Theta} \phi(\boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta}$$

Markov Chain Monte Carlo

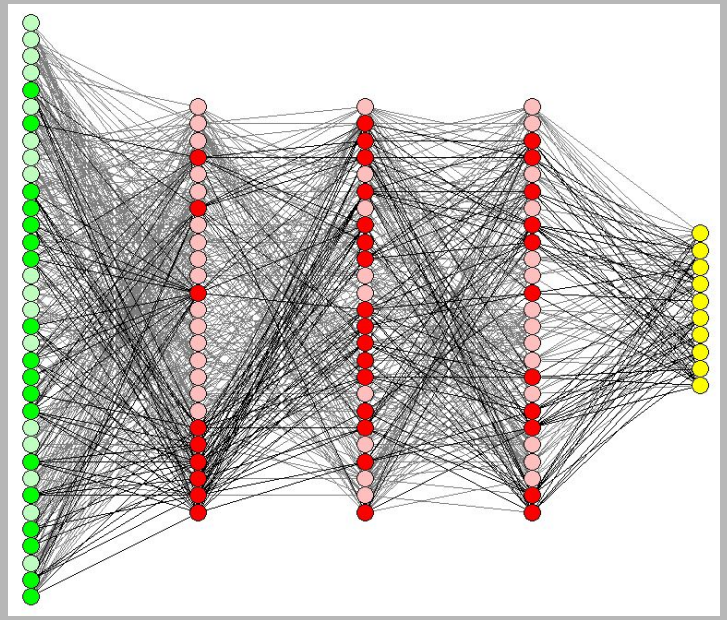


Variational Inference

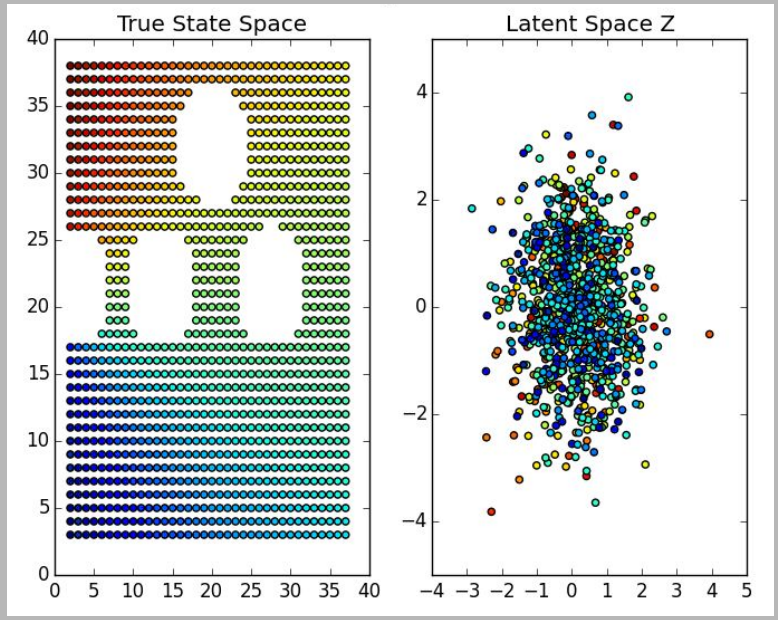


Case Study I: Monte Carlo Dropout

Dropout

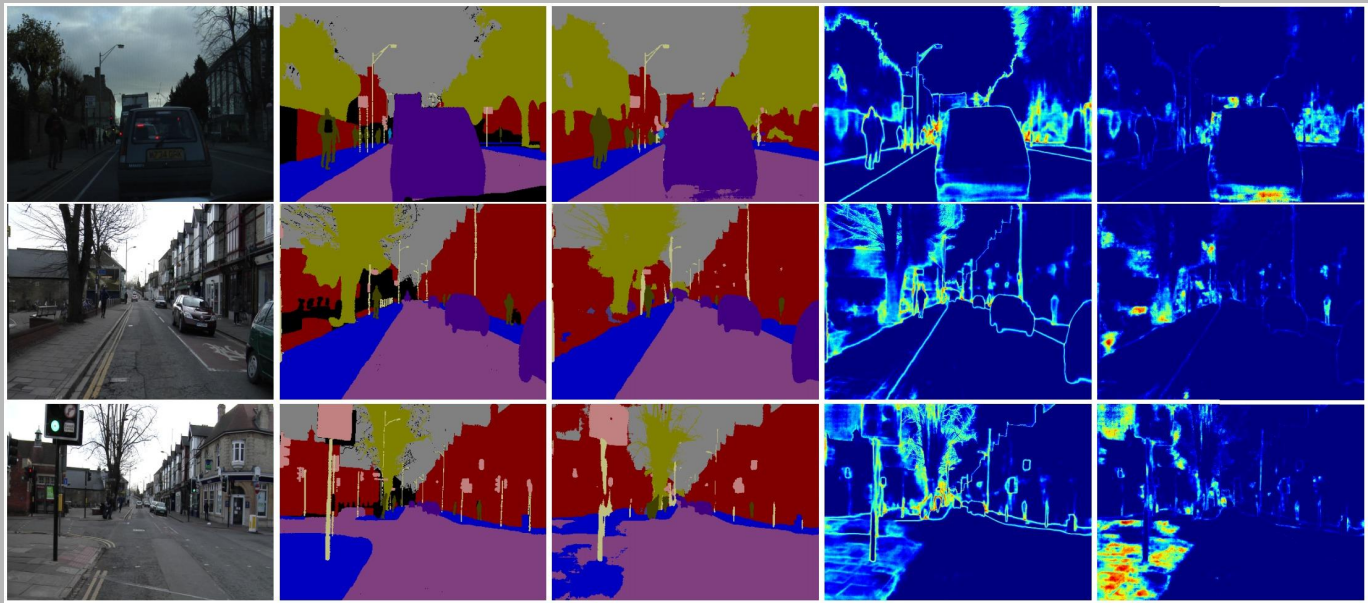


Variational Inference



"Dropout as Bayesian Approximation"
Gal et al. (2015)

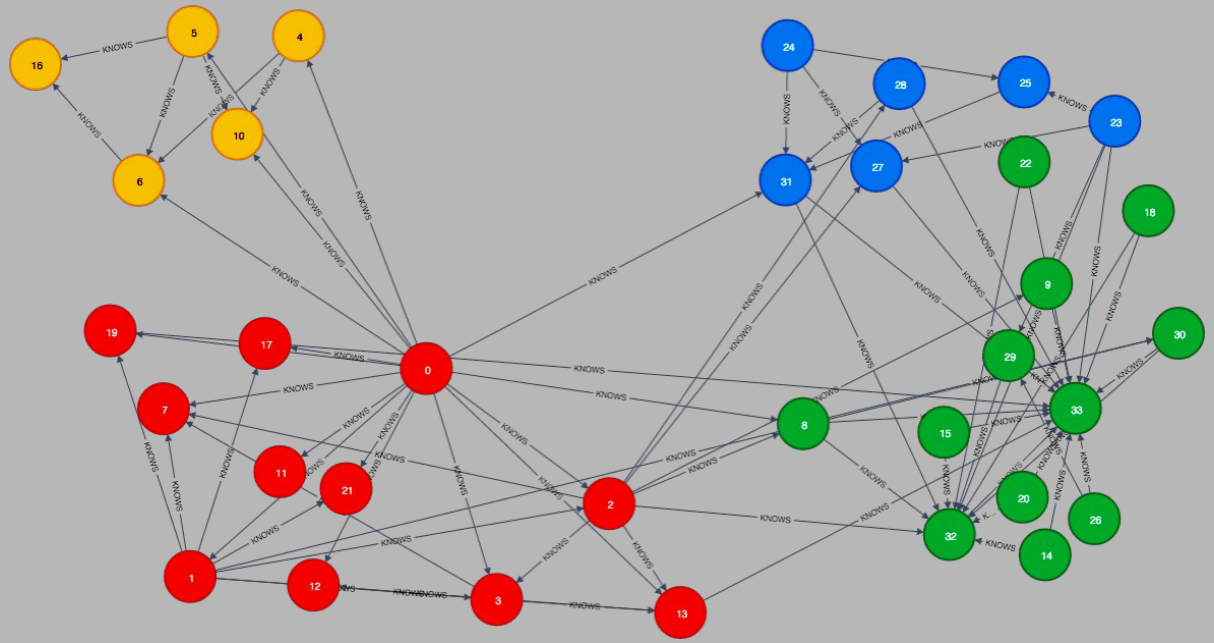
Case Study I: Monte Carlo Dropout (Cont.)



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

"What Uncertainties Do We Need in BDL for CV?"
Kendall et al. (2017)

Case Study II: Clustering with Dirichlet Processes



Case Study II: Clustering with Dirichlet Processes

“Lecture Notes on Bayesian Nonparametrics”
Orbanz (2014)

Challenges of Bayesian Inference

1. **Computationally intensive:**
    a. Computational power and memory can be expensive.
    b. Challenging real-time computations.

2. **Complex computations:**
    a. High barrier to entry for beginners.
    b. Are people willing to trust it?

4. Characterising Model Interpretability

Why Do We Even Care?

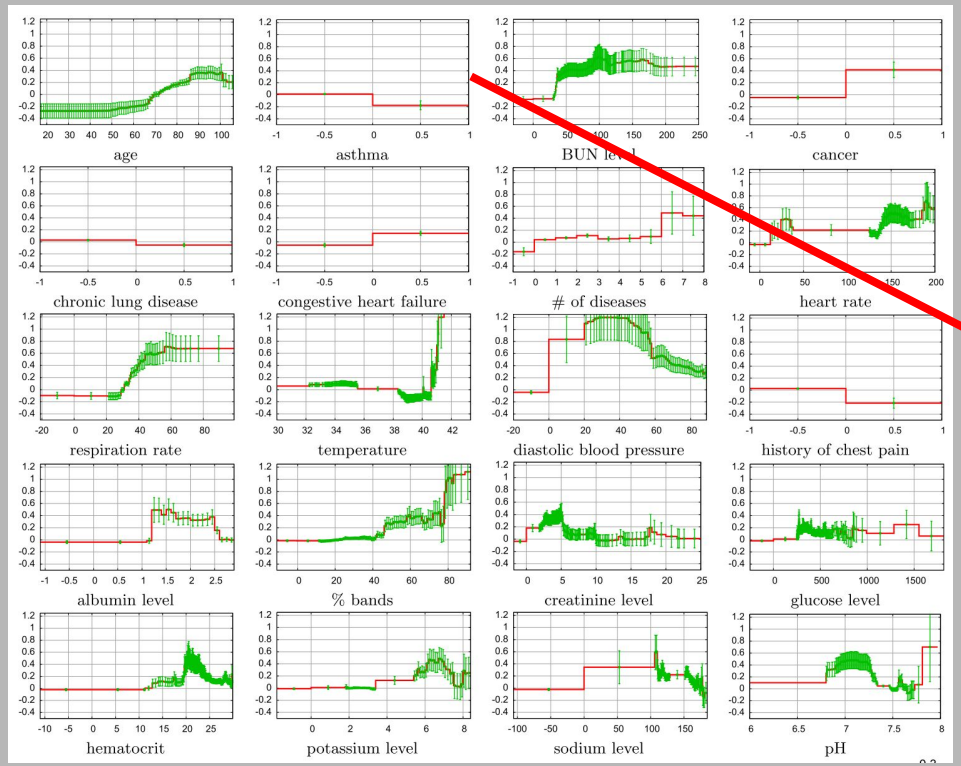
- **Trust:**
  - → Without trust, adoption rate will be low.
  - → Can we legally deploy a black box?

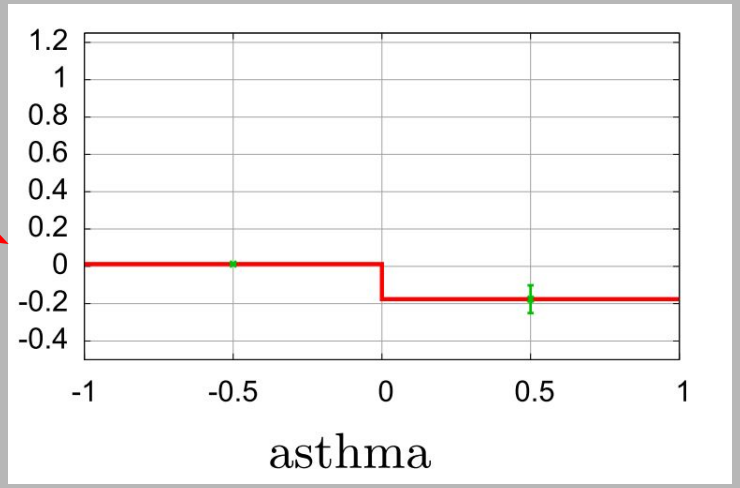
- **Causality**:
  - → Are we capturing the real policy effects?
  - → More understanding equals better “debuggability”.

- **External validity:**
  - → Do we introduce feedback loops?
  - → Real pattern or data leak?

“The Mythos of Model Interpretability”
Lipton et al. (2016)

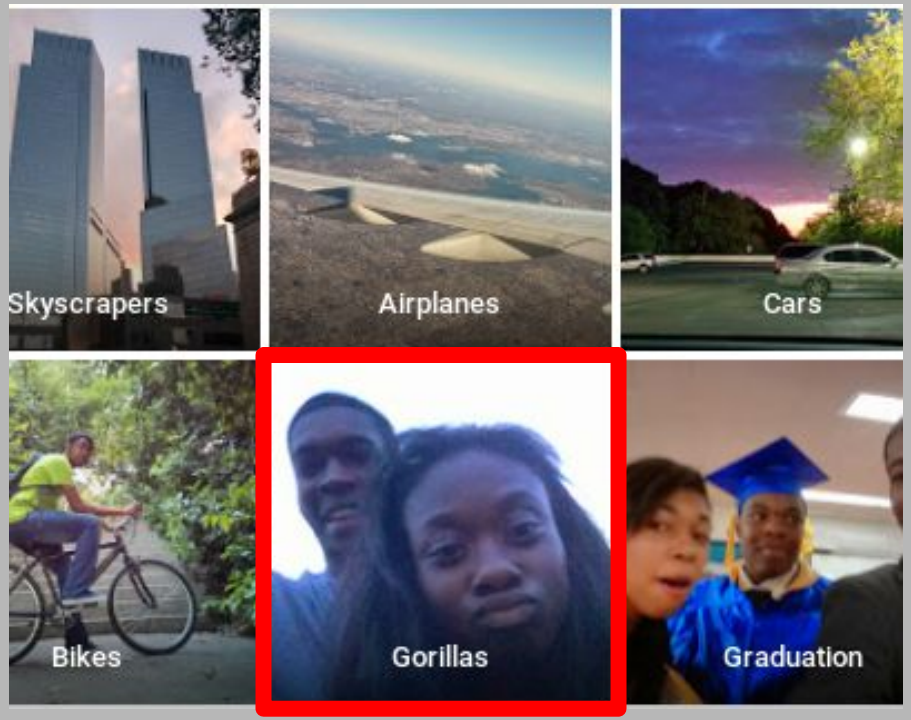
Case Study I: Medical Treatment for Pneumonia





"Intelligible Models for Healthcare"
Caruana et al. (2015)

Case Study II: Google's "Racist" Algorithm



Interpretations matter because we cannot encapsulate our objectives into a mathematical functions.

How Do We Characterise Interpretability?

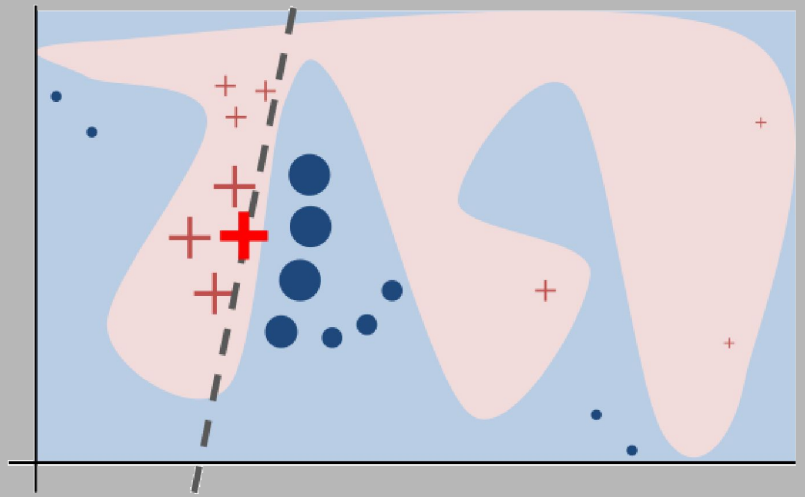
- **Transparency:**
  → “I can simulate the algorithm in my head.”
  → “I can break the algorithm down into smaller intuitive pieces.”

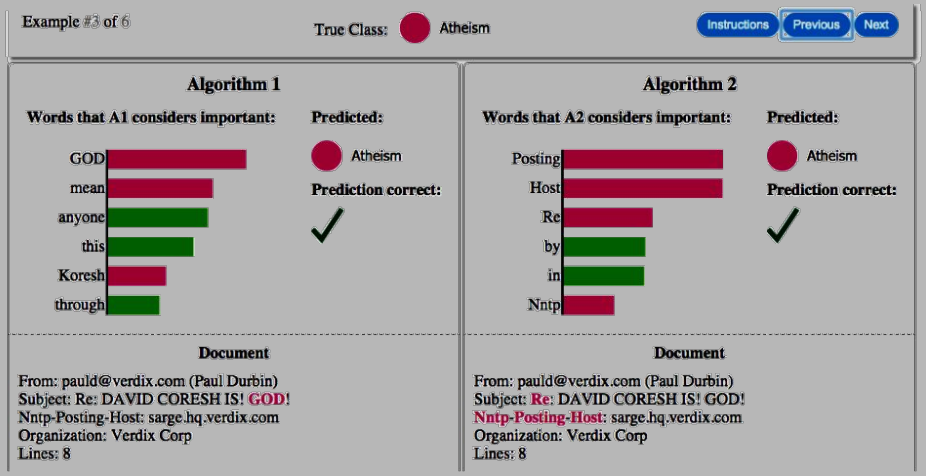
- **Post-hoc explicability**:
  → “I can tell you why the model behaved that way.”
  → “I know of other instances where the model behaved that way.”

“The Mythos of Model Interpretability”
Lipton et al. (2016)

5. Unboxing the Black Box

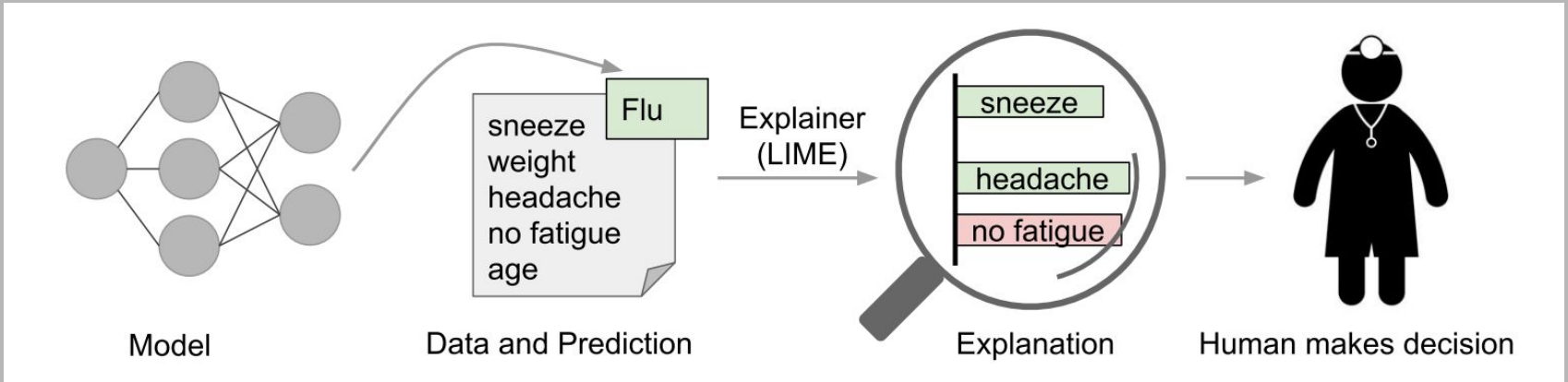
Case Study I: LIME





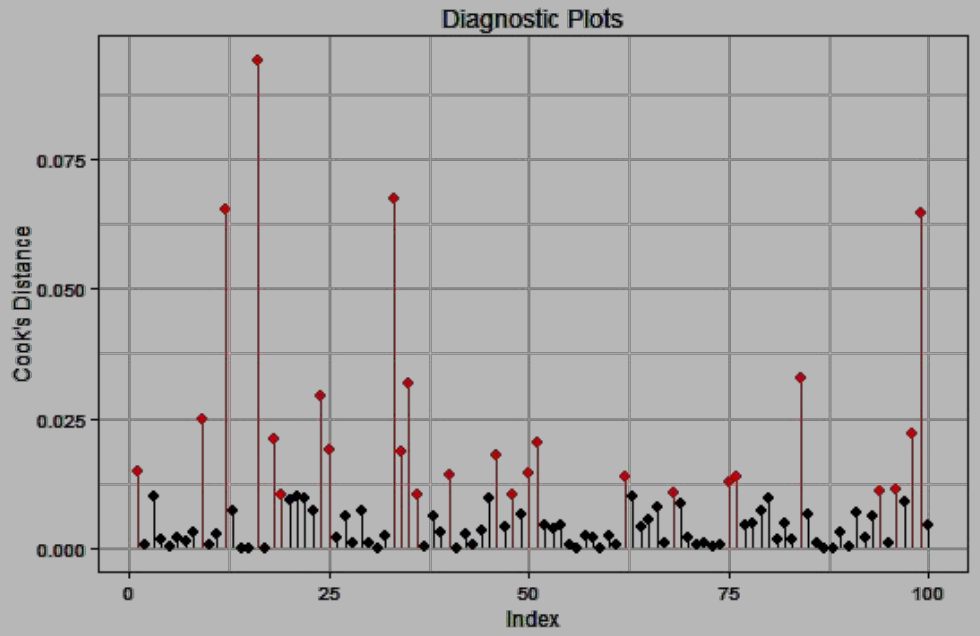
"Why Should I Trust You?"
Ribeiro et al. (2016)

Case Study I: LIME

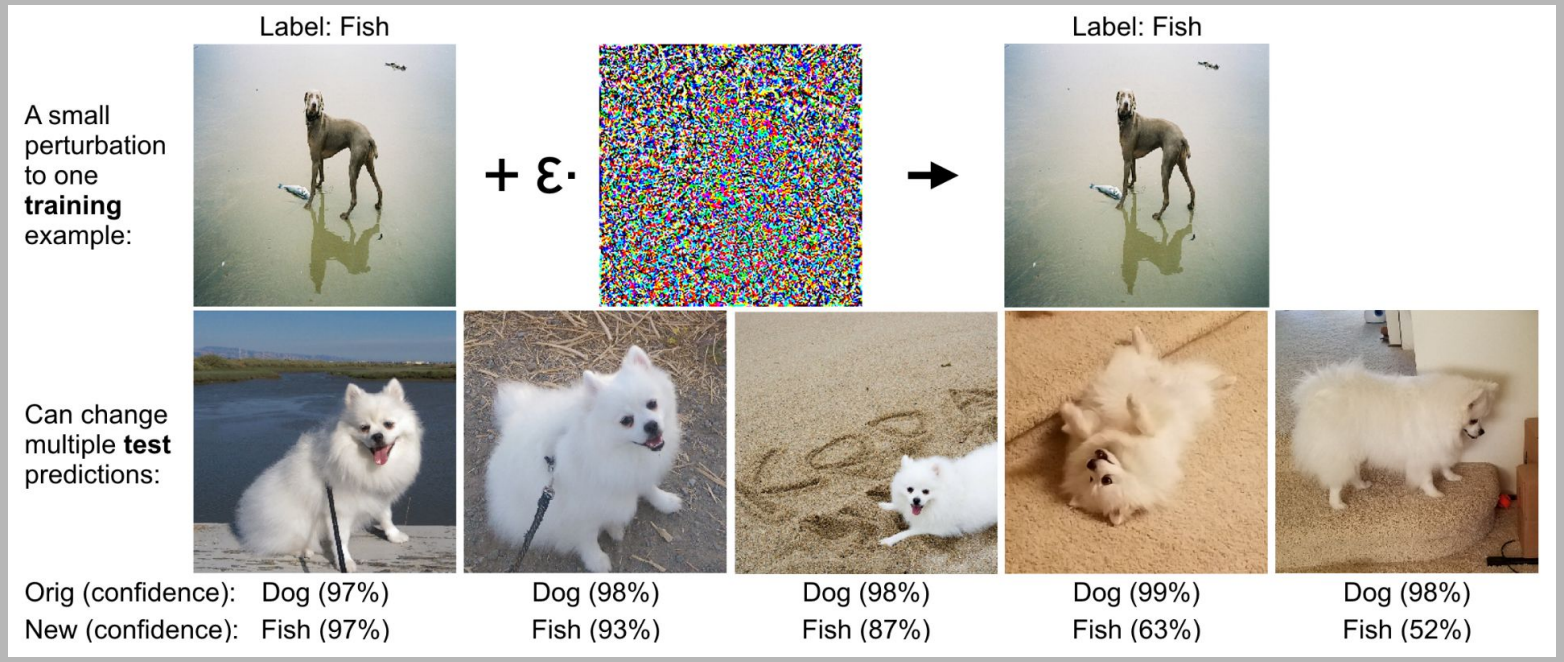


"Why Should I Trust You?"
Ribeiro et al. (2016)

Case Study II: Influence Functions



Case Study II: Influence Functions



"Understanding Black-box Predictions via Influence Functions"
Koh and Liang (2017)

Conclusion

- Modelling uncertainty and interpretability are extremely valuable:
  - → Uncertainties are required to fully inform decisions.
  - → Black boxes become blockers in many cases.

- Epistemic uncertainty in modern ML is challenging:
  - → Many existing ML methods do not fully account for epistemic uncertainties.
  - → Bayesian inference provides a general, principled framework to estimate uncertainty.

- Interpretability is a multi-faceted concept:
  - → It all boils down to not being able to write our objectives mathematically.
  - → We can aim for transparency and post-hoc explicability.

References

- Athey, Susan. "The impact of machine learning on economics." Economics of Artificial Intelligence. University of Chicago Press, 2017.
- Balan, Anoop Korattikara, et al. "Bayesian dark knowledge." Advances in Neural Information Processing Systems. 2015.
- Blanchard, Olivier, and Roberto Perotti. "An empirical characterization of the dynamic effects of changes in government spending and taxes on output." the Quarterly Journal of economics 117.4 (2002): 1329-1368.
- Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American Statistical Association 112.518 (2017): 859-877.
- Blundell, Charles, et al. "Weight uncertainty in neural networks." arXiv preprint arXiv:1505.05424 (2015).
- Bui, Thang, et al. "Deep gaussian processes for regression using approximate expectation propagation." International Conference on Machine Learning. 2016.
- Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.
- Ching, Travers, et al. "Opportunities and obstacles for deep learning in biology and medicine." Journal of The Royal Society Interface 15.141 (2018): 20170387.
- Gal, Yarin. "Uncertainty in deep learning." University of Cambridge (2016).
- Gal, Yarin, and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Insights and applications." Deep Learning Workshop, ICML. Vol. 1. 2015.

References

- Gershman, Samuel J., and David M. Blei. "A tutorial on Bayesian nonparametric models." Journal of Mathematical Psychology 56.1 (2012): 1-12.
- Green, Peter J., and David I. Hastie. "Reversible jump MCMC." Genetics 155.3 (2009): 1391-1403.
- Kendall, Alex, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." arXiv preprint arXiv:1705.07115 3 (2017).
- Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." Advances in neural information processing systems. 2017.
- Kindermans, Pieter-Jan, et al. "The (Un) reliability of saliency methods." arXiv preprint arXiv:1711.00867 (2017).
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
- Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." arXiv preprint arXiv:1703.04730 (2017).
- Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533 (2016).
- Lieb, Lenard, and Stephan Smeekes. "Inference for Impulse Responses under Model Uncertainty." arXiv preprint arXiv:1709.09583 (2017).
- Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490 (2016).
- Neal, Radford M. "MCMC using Hamiltonian dynamics." Handbook of Markov Chain Monte Carlo 2.11 (2011): 2.

References

- Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
- Olah, Chris, et al. "The building blocks of interpretability." Distill 3.3 (2018): e10.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.
- Russo, Daniel, et al. "A tutorial on thompson sampling." arXiv preprint arXiv:1707.02038 (2017).
- Scott, Steven L. "A modern Bayesian look at the multi-armed bandit." Applied Stochastic Models in Business and Industry 26.6 (2010): 639-658.
- Settles, Burr. "Active learning." Synthesis Lectures on Artificial Intelligence and Machine Learning 6.1 (2012): 1-114.
- Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014): 1929-1958.
- Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." arXiv preprint arXiv:1703.01365 (2017).
- Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." Journal of the American Statistical Association just-accepted (2017).

References

- Welling, Max, and Yee W. Teh. "Bayesian learning via stochastic gradient Langevin dynamics." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.
- Yeomans, Mike, et al. "Making sense of recommendations." Preprint at http://scholar. harvard. edu/files/sendhil/files/recommenders55_01. pdf (2016).