CSE598: Operationalizing Deep Learning: A Sociotechnical Perspective

Ransalu Senanayake
School of Computing and Augmented Intelligence
Arizona State University

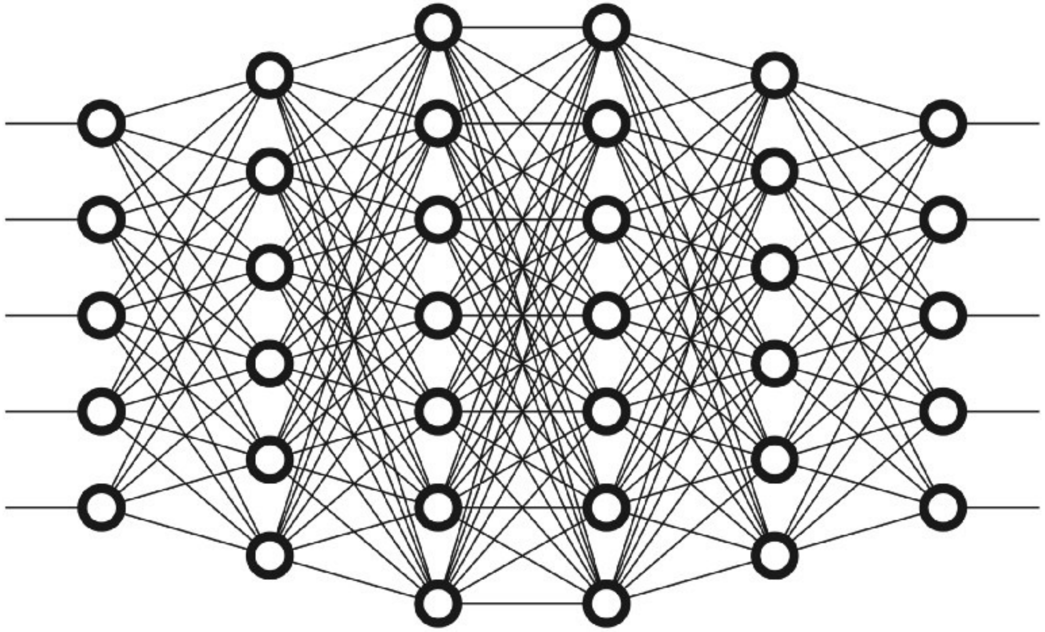
LENS Lab

www.ransalu.com/lens-lab ransalu@asu.edu


WAYMO


R
IVÁN BARRAGAN
AUTOPILOT DIDN'T REACT OR GIVE ANY WARNINGS, SO IVÁN TOOK OVER AND APPLIED THE BRAKES.
HE DOESN'T THINK AUTOPILOT HAD A CHANCE TO REACT, AS THE ACCIDENT WAS TOO FAR IN THE DISTANCE.
KAPWING

We Don’t Know



Because We Don’t Know How Black Box ML Models Work

Chandler, Arizona: Waymo Collision

Tempe, Arizona: Uber Hit And Run

Mountain View, California: Tesla Model X Crash

Driver killed in self-driving car accident for first time

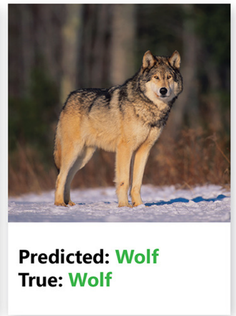
Nation Jun 30, 2016 5:24 PM EST

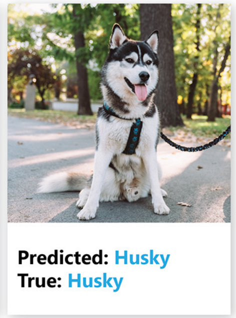
“The high ride height of the trailer of the truck combined with its positioning across the road and the rare circumstances of the impact caused the Model S to pass under the trailer” - Tesla

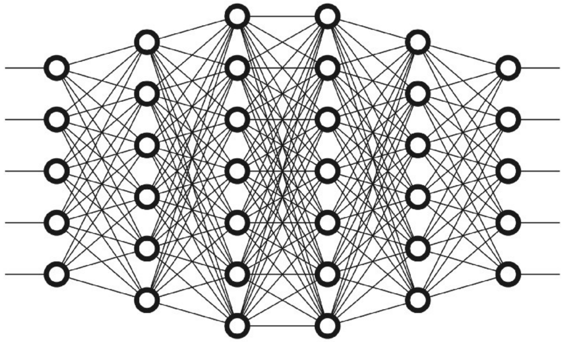
How Do we Train Deep Neural Networks?

Discriminative model vs. Generative model

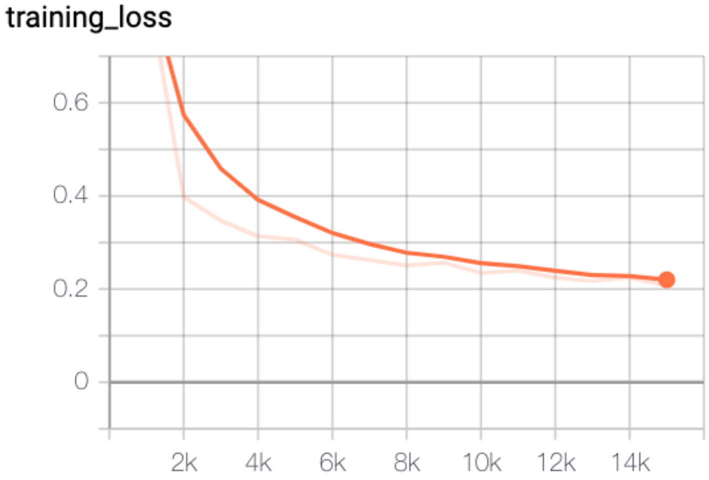
$p(y|x)$ $p(y, x)$







Training with backpropagation to minimize the error



Predicted: Wolf
True: Wolf

Predicted: Husky
True: Husky

Predicted: Husky
True: Husky

Predicted: Wolf
True: Wolf

Predicted: Wolf
True: Wolf

Predicted: Wolf
True: Wolf

Predicted: Husky
True: Wolf

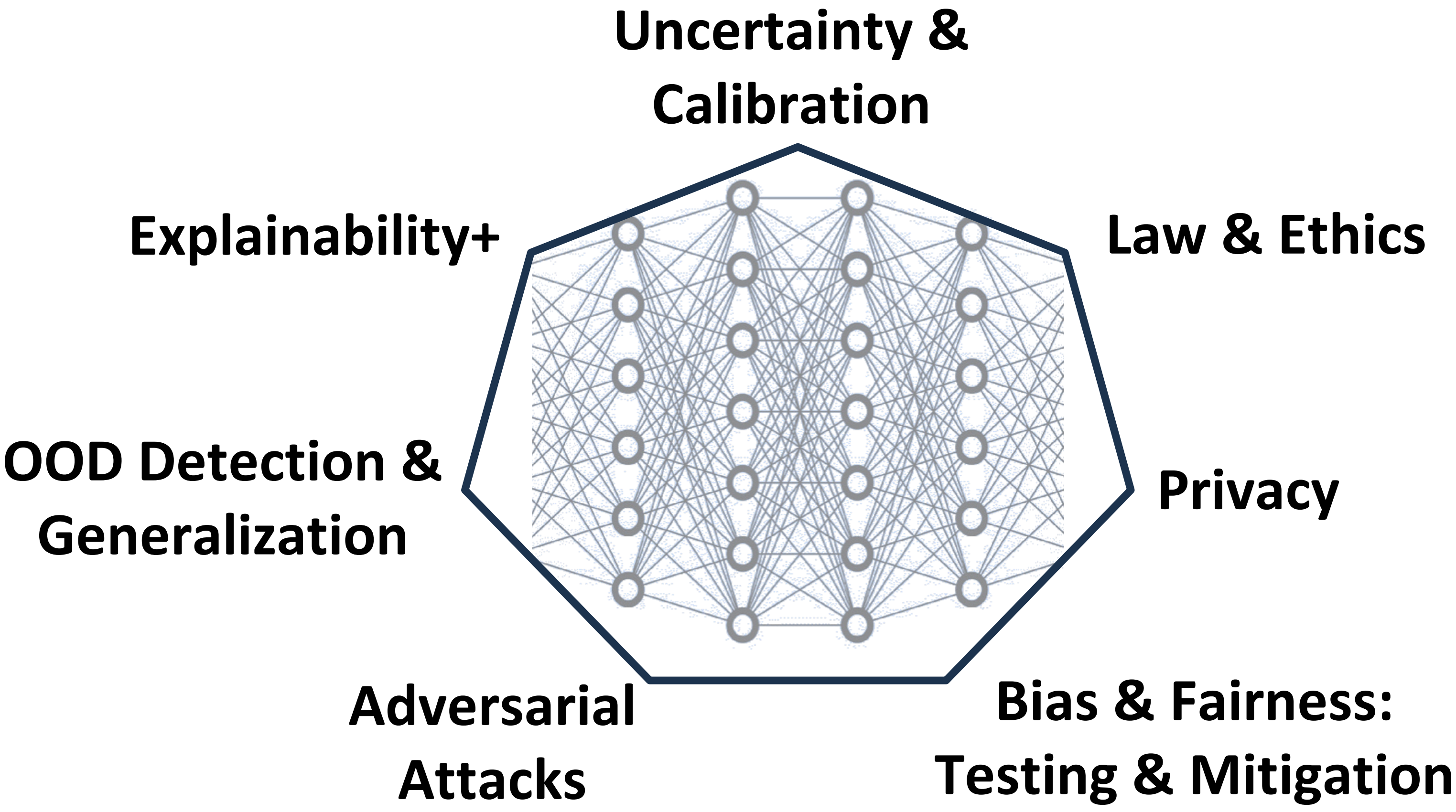
Predicted: Wolf
True: Wolf

Predicted: Wolf
True: Husky

Predicted: Husky
True: Husky

https://news.microsoft.com/source/features/ai/azure-responsible-machine-learning/

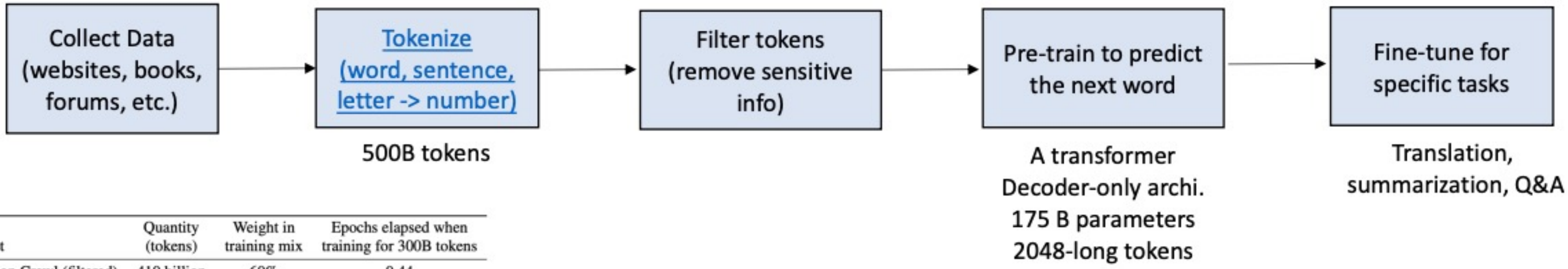
Is That Enough?



Application 1 (NLP)

Sentence Completion & Review Generation

How Has GPT3 Been Trained?



Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Social bias
(historical bias, life bias, etc.)

Algorithmic bias
(dataset bias, model bias, etc.)

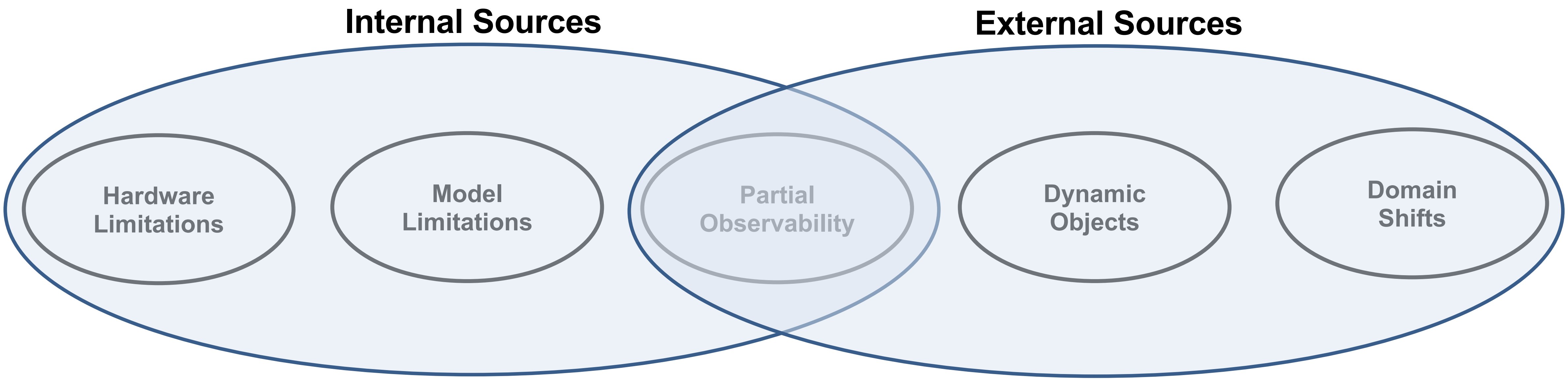
Generative models in Vision

Application 2 (Robotics)

Perception to Decision-Making

source: https://www.youtube.com/watch?v=-Q7YM8IlivU

Sources of Uncertainty



Sources of Uncertainty

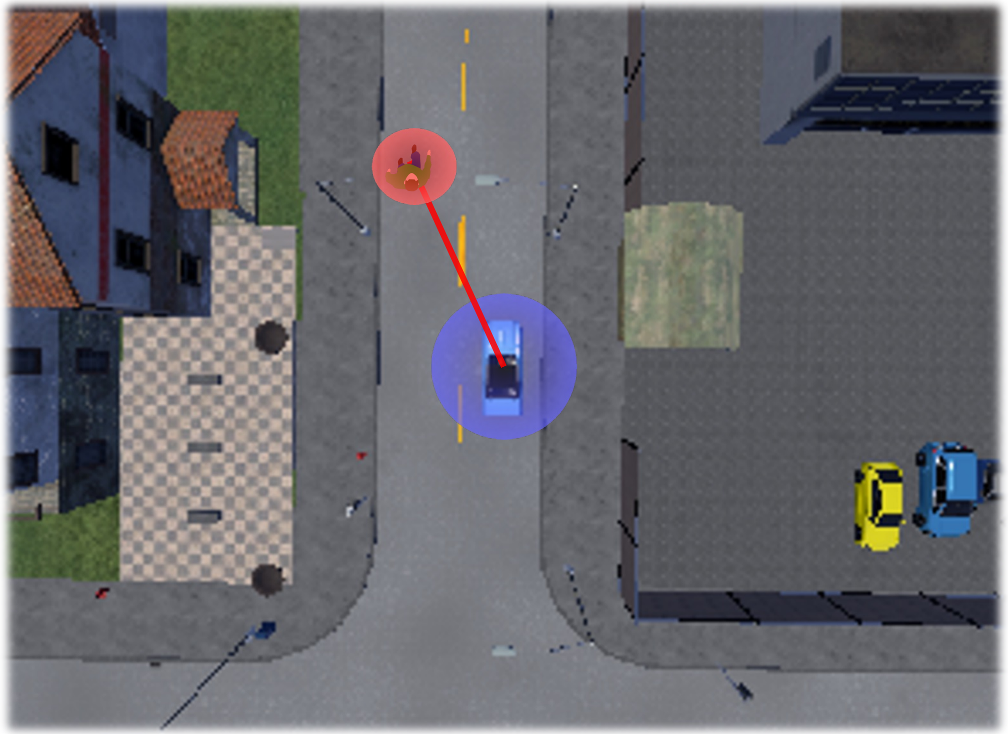
Hardware Limitations

Model Limitations

Partial Observability

Dynamic Objects

Domain Shifts



Sources of Uncertainty

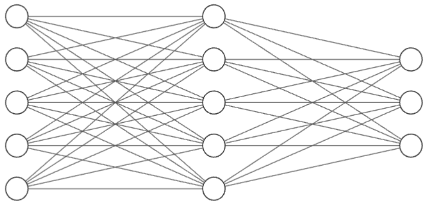
Hardware Limitations

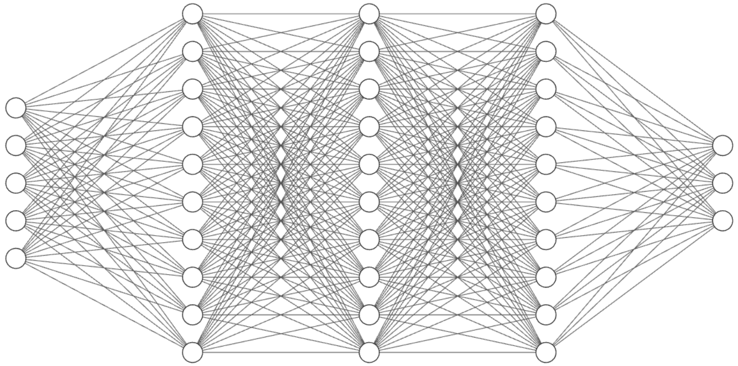
Model Limitations

Partial Observability

Dynamic Objects

Domain Shifts





Sources of Uncertainty

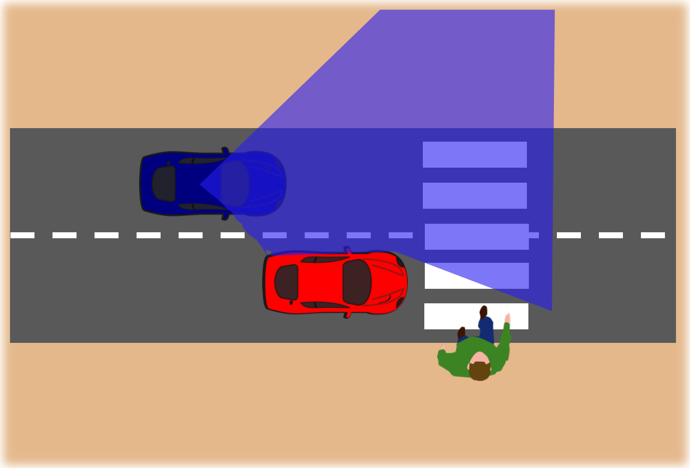
Hardware Limitations

Model Limitations

Partial Observability

Dynamic Objects

Domain Shifts



Sources of Uncertainty

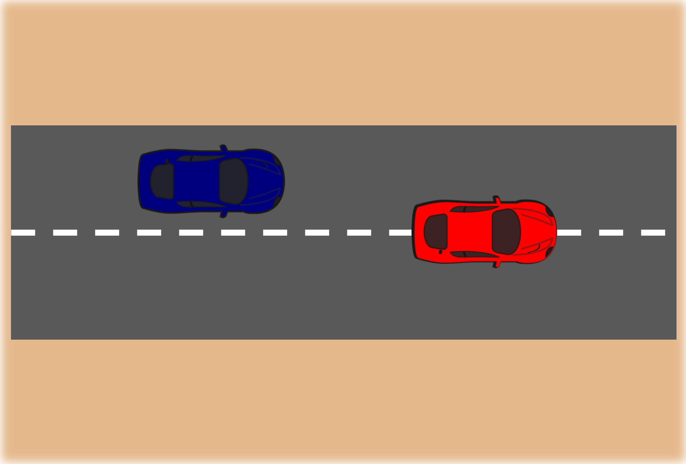
Hardware Limitations

Model Limitations

Partial Observability

Dynamic Objects

Domain Shifts



Sources of Uncertainty

Hardware Limitations

Model Limitations

Partial Observability

Dynamic Objects

Domain Shifts

During training

At deployment

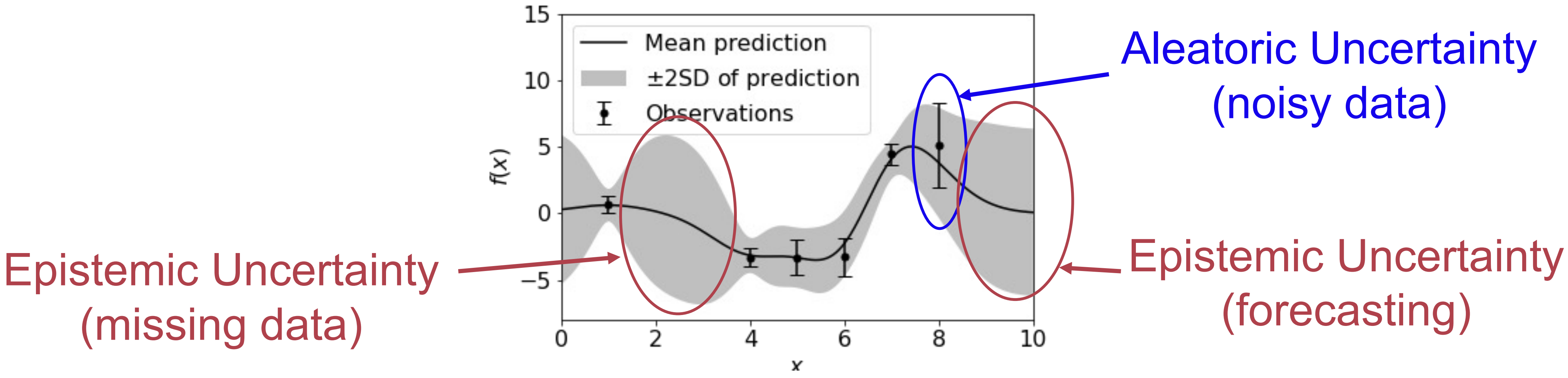
source: https://www.bbc.com/news/technology-40416606, https://bearclawindustries.com/dash-cam-full-hd/

Types of Uncertainties

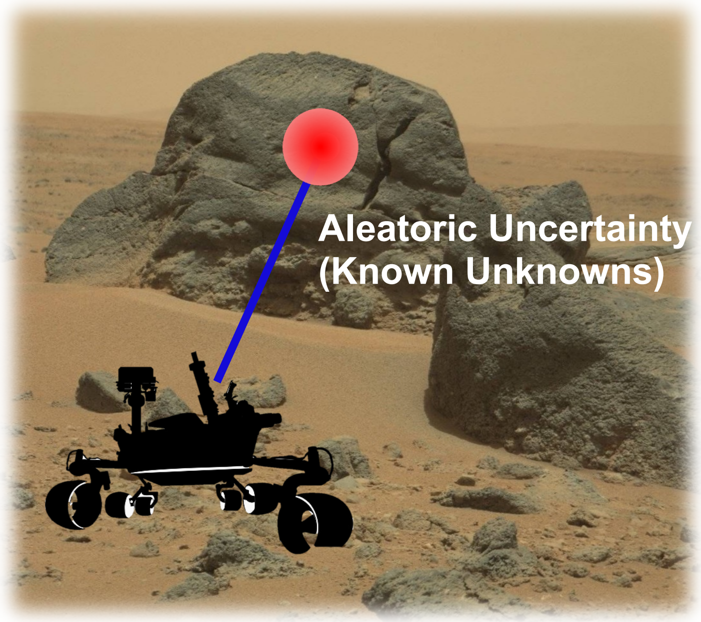
Unknowns

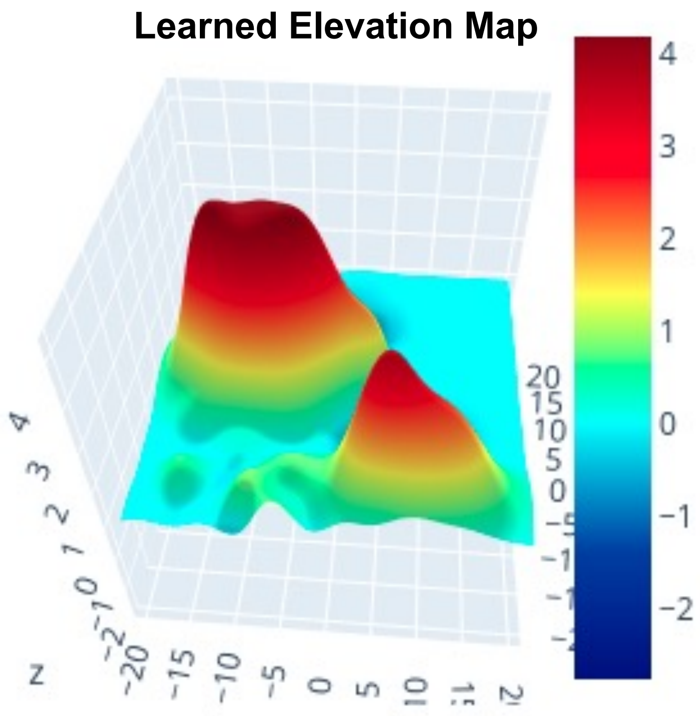
Known Unknowns
(Aleatoric Uncertainty)

Unknown Unknowns
(Epistemic Uncertainty)

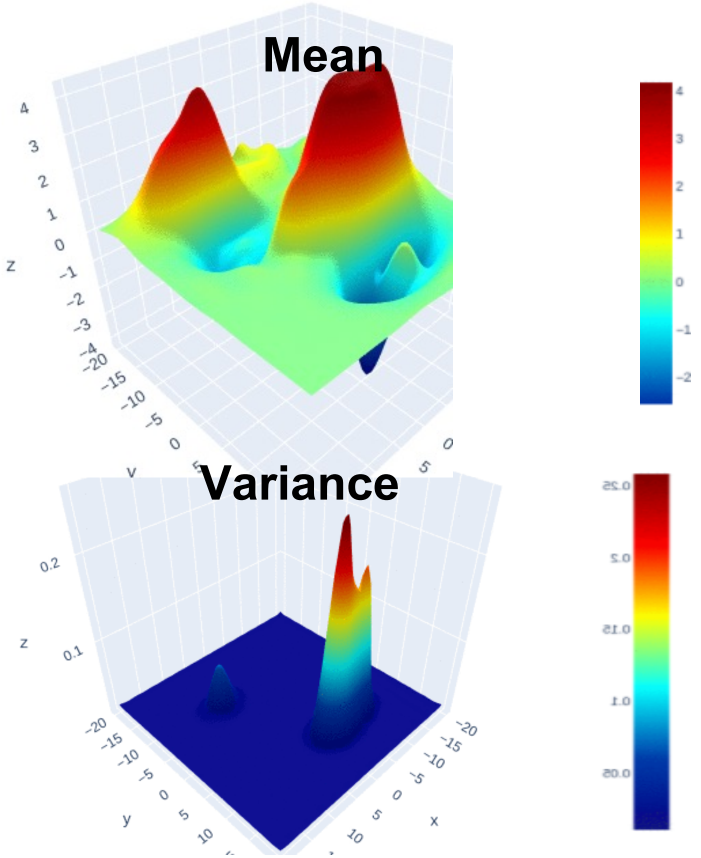


Known Unknowns (a.k.a. Aleatoric Uncertainty)

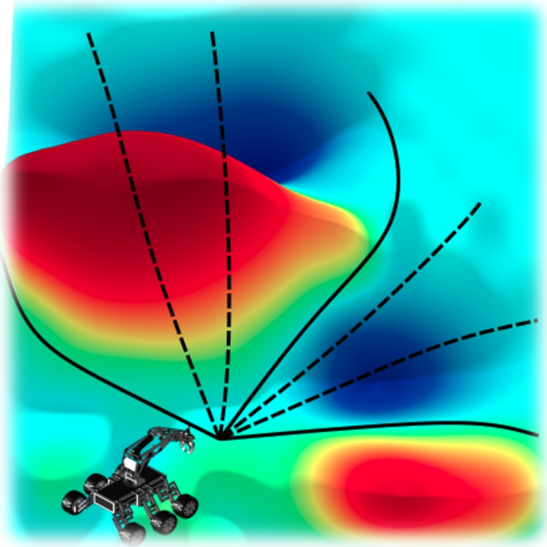




Unknown Unknowns (a.k.a. Epistemic Uncertainty)

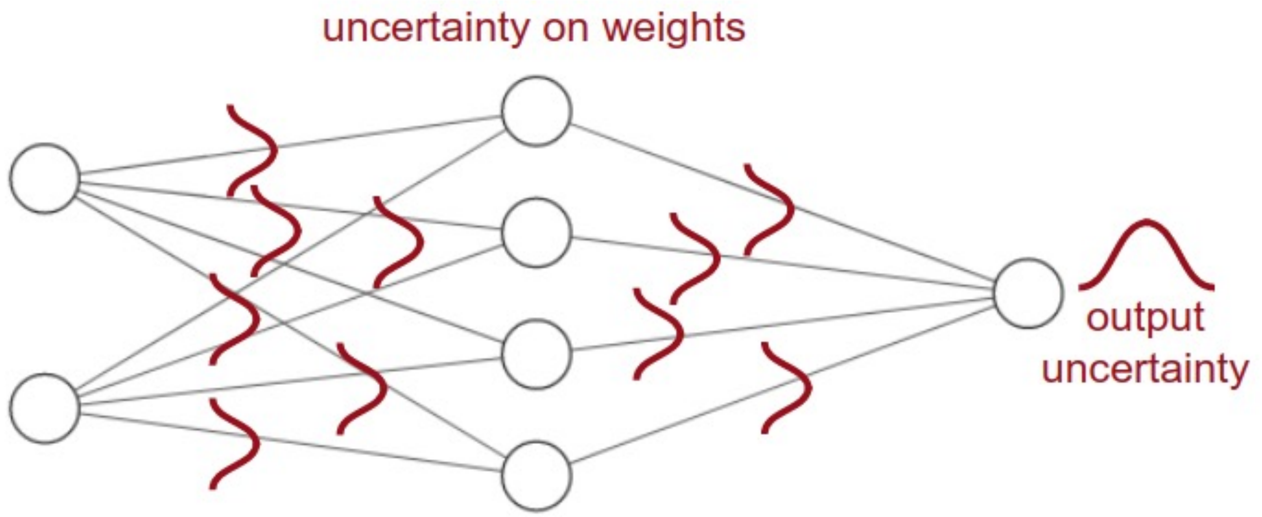


$3 billion mission!



Unknown Unknowns (a.k.a. Epistemic Uncertainty)

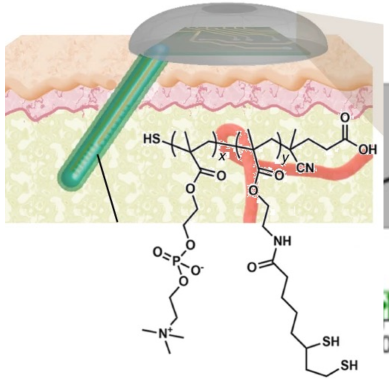
$$y = \underbrace{w_1 \mathrm{x} + w_0}_{f_w(x)} + \epsilon$$

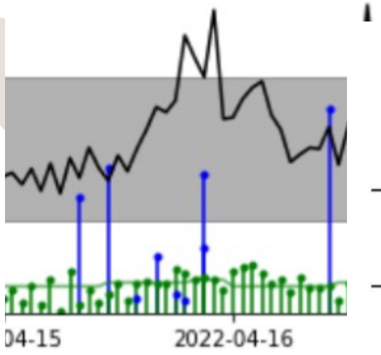


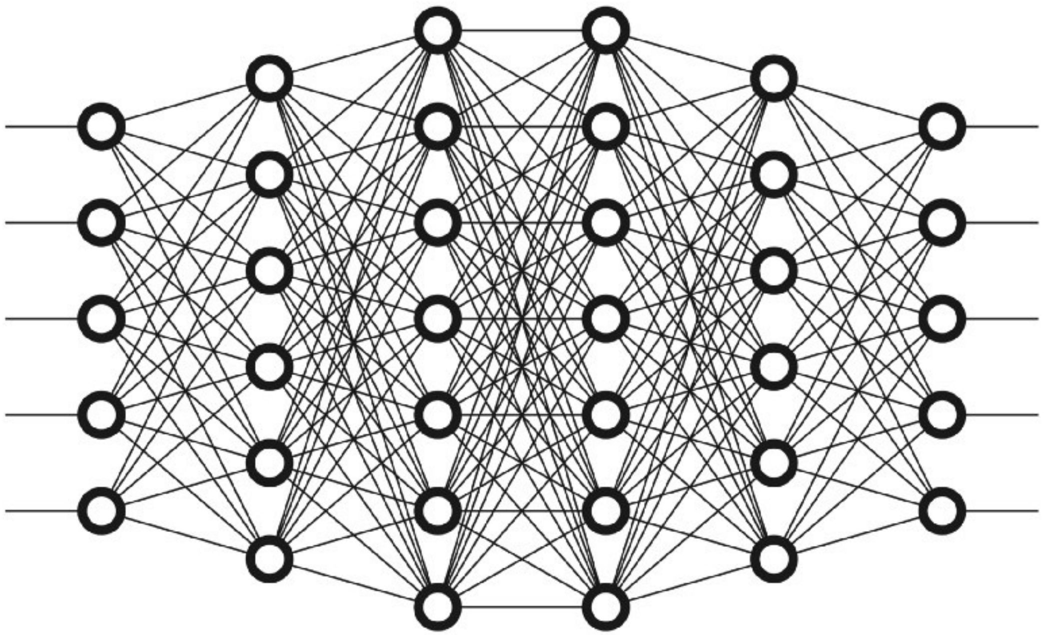
It’s okay to not know something. However, we ought to know what we don’t know.

Application 3 (Healthcare)

Personalized Healthcare

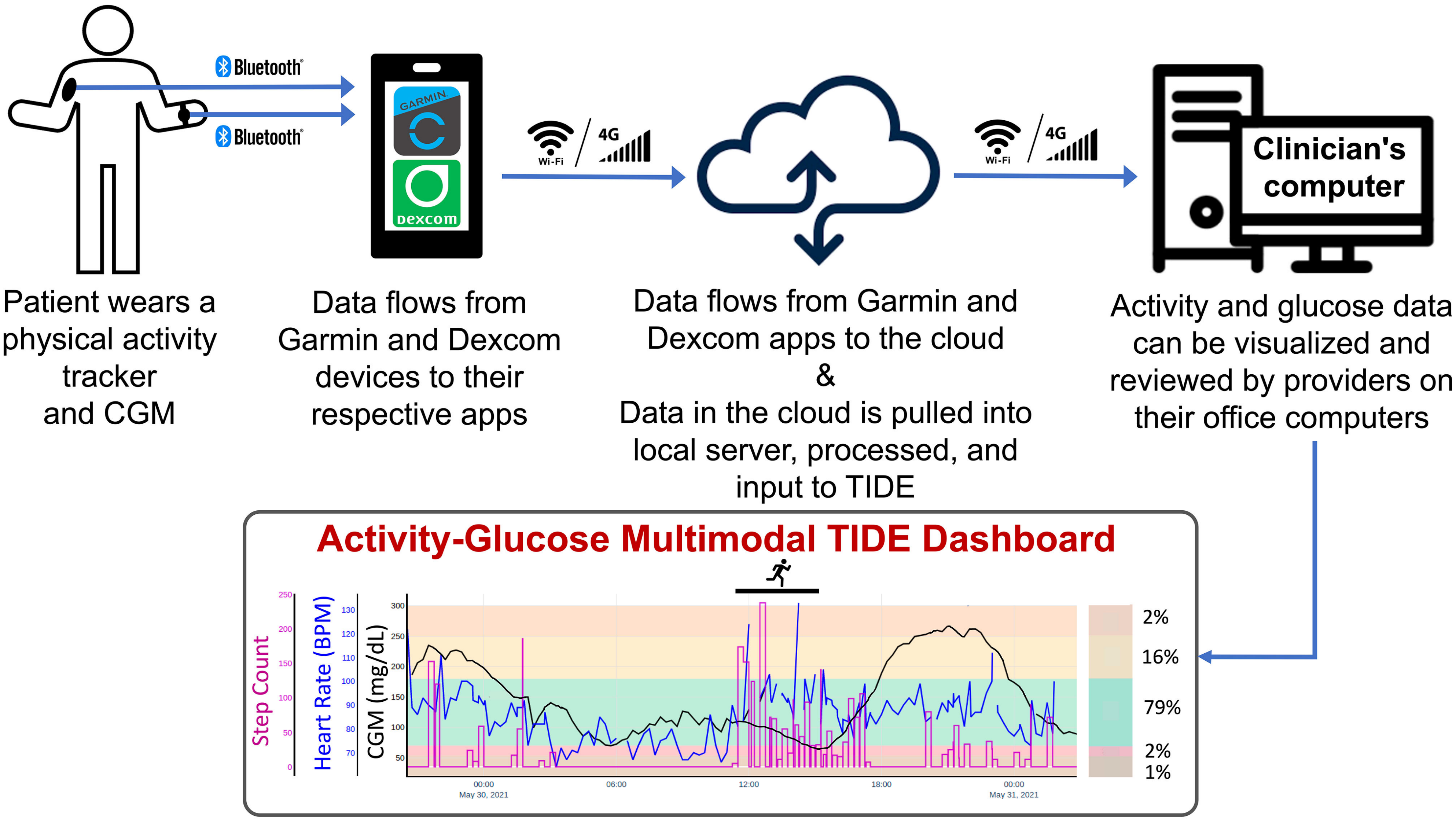






Output:
This patient has hypoglycemia

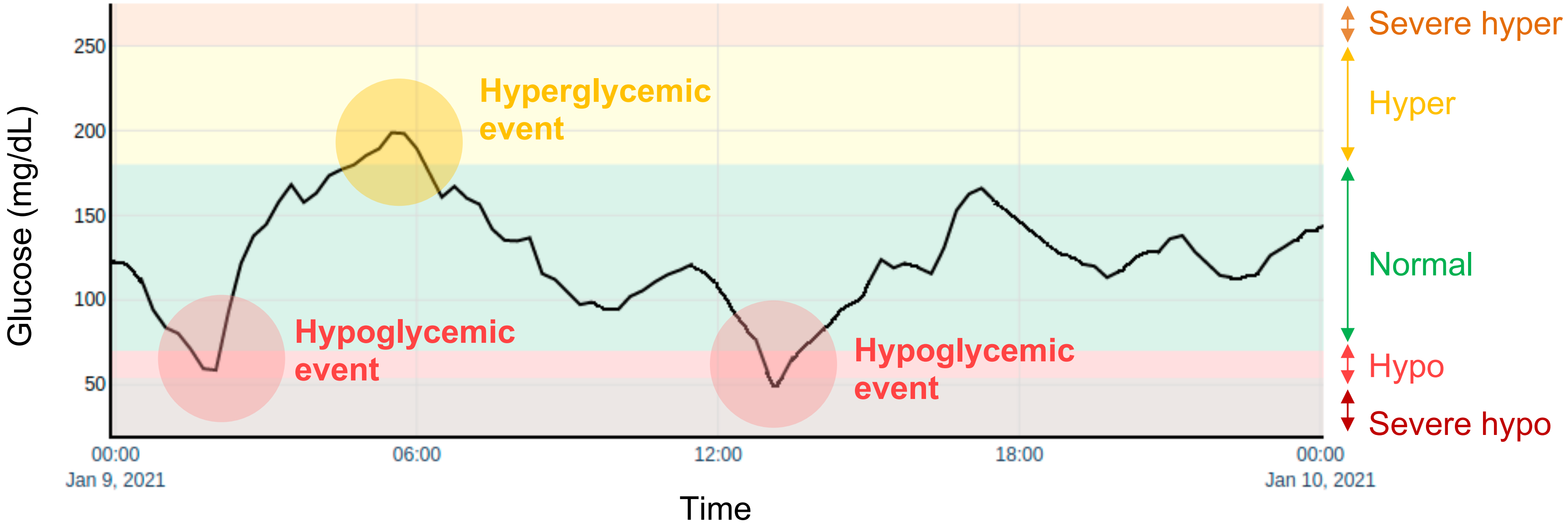
Clinicians need to understand!



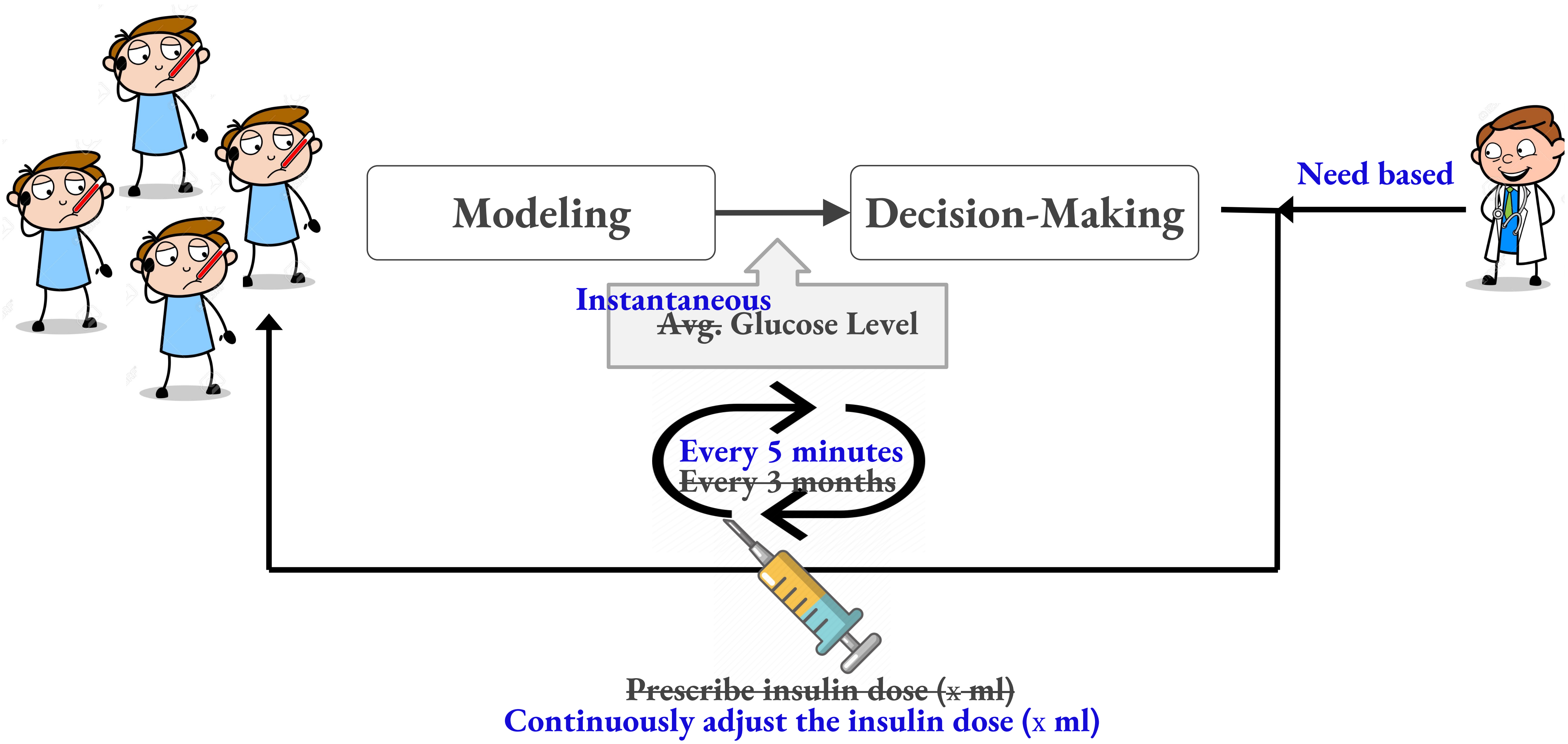
Adding Glycemic and Physical Activity Metrics to a Multimodal Algorithm-Enabled Decision-Support Tool for Type 1 Diabetes Care: Keys to Implementation and Opportunities
DP Zaharieva, R Senanayake, C Brown, B Watkins, G Loving, P Prahalad

Type 1 Diabetes

The 3-month avg glucose (Hb1Ac) misses important events



Challenges


Modeling
Decision-Making
Need based
Instantaneous
~~Avg.~~ Glucose Level
Every 5 minutes
~~Every 3 months~~
~~Prescribe insulin dose (x ml)~~
Continuously adjust the insulin dose (x ml)

Training…

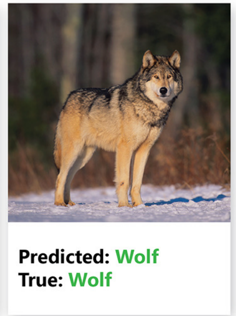
How Do we Train Deep Neural Networks?

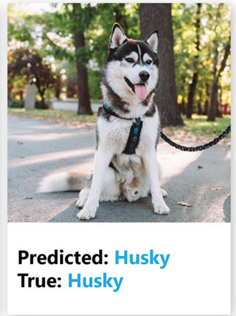
Discriminative model vs. Generative model

$p(y|x)$ $\quad\quad\quad\quad$ $p(y,x)$

Optimization

Training with backpropagation
to minimize the error

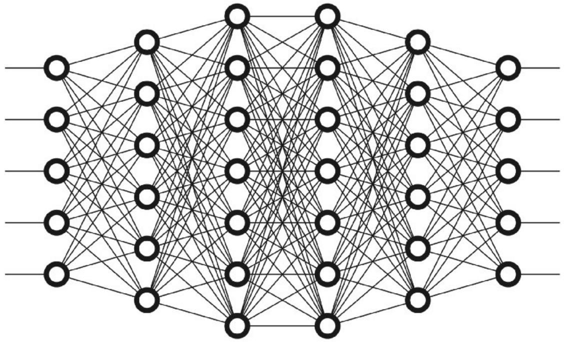


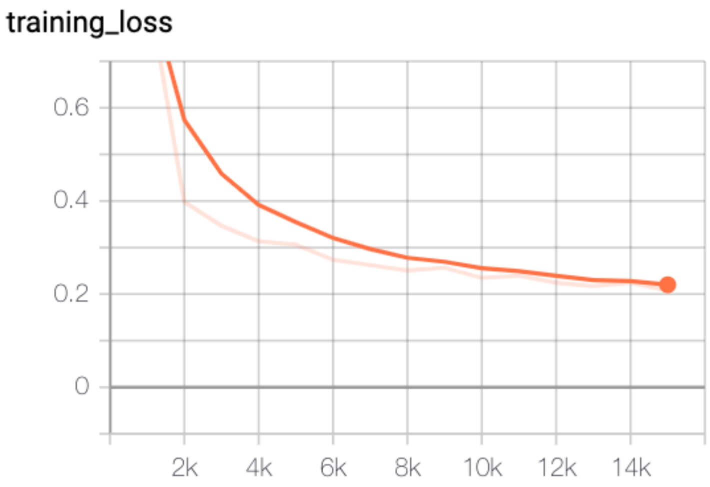


Dataset

Data

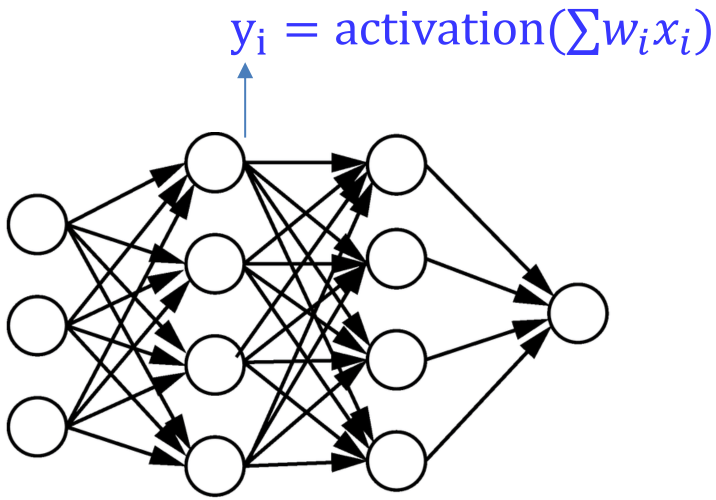
Model





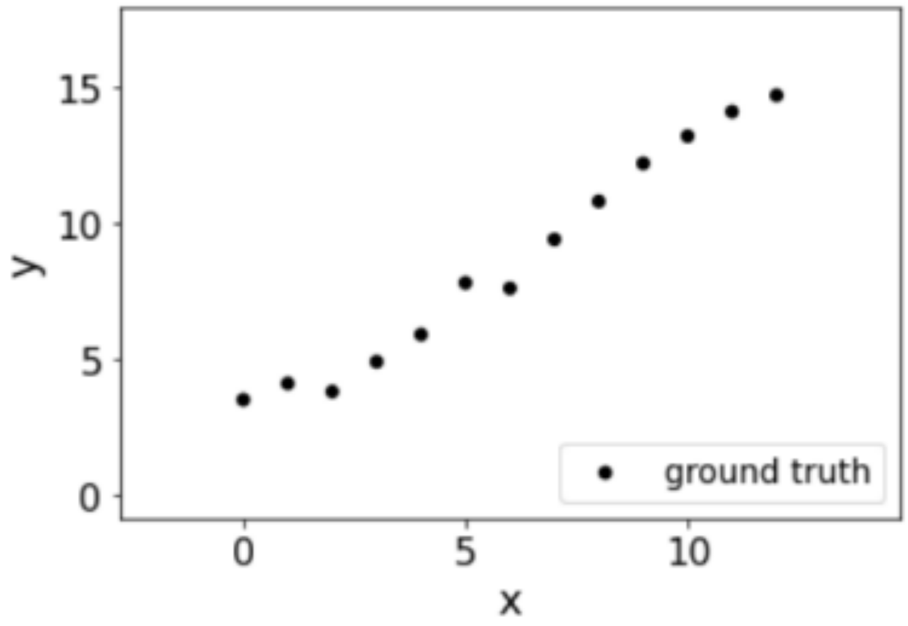
Evaluation

Why Does a Neural Network Work So Well?



1) (Non)linear regression of (non)linear regressions view
2) Manifold view

Linear Regression: Problem



Objective: Predict outputs for unknown inputs $\mathbf{x}_q$, given the training data $\{\mathbf{x}_i, y_i\}_{i=1}^N$.

$$y = f(\mathbf{x}) + e, \quad e \sim \mathcal{N}(0, \sigma^2)$$

Linear Regression: Model

$$y = \theta_0 + \theta_1 x^{(1)} + \theta_2 x^{(3)} + \theta_3 x^{(3)} + \ldots + \theta_D x^{(D)} + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(D)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(D)} \\ 1 & x_3^{(1)} & \cdots & x_3^{(D)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N^{(1)} & \cdots & x_N^{(D)} \end{bmatrix} \begin{bmatrix} \theta^{(0)} \\ \theta^{(1)} \\ \vdots \\ \theta^{(D)} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_N \end{bmatrix}$$

$$\mathbf{y} = X\boldsymbol{\theta} + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

Linear Regression: Solution 1
Ordinary Least Square (OLS)

$$\boldsymbol{\theta}^*_{OLS} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{e}\|_2^2 = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$$

$$\begin{aligned} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 &= (\mathbf{y} - X\boldsymbol{\theta})^\top(\mathbf{y} - X\boldsymbol{\theta}) \\ &= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top X\boldsymbol{\theta} - \boldsymbol{\theta}^\top X^\top\mathbf{y} + \boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} \\ &= \mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\theta}^\top X^\top\mathbf{y} + \boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} \end{aligned}$$

$$(AB)^\top = B^\top A^\top$$

$$\theta^\top X^\top \mathbf{y} = \text{scalar} = (\theta^\top X^\top \mathbf{y})^\top = \mathbf{y}^\top X\theta$$

$$\frac{\partial}{\partial\boldsymbol{\theta}}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 = -2X^\top\mathbf{y} + 2X^\top X\boldsymbol{\theta}$$

By setting the derivative to *zero*,

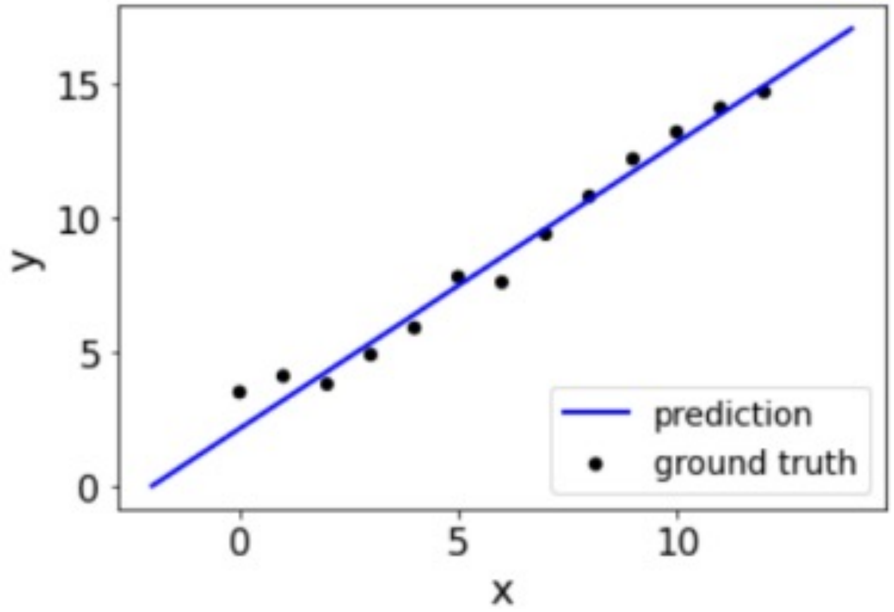
$$-2X^\top\mathbf{y} + 2X^\top X\boldsymbol{\theta}^*_{OLS} = 0 \Rightarrow X^\top\mathbf{y} = X^\top X\boldsymbol{\theta}^*_{OLS}$$

$$\boxed{\boldsymbol{\theta}^*_{OLS} = (X^\top X)^{-1}X^\top\mathbf{y}}$$

Linear Regression: Predictions/Querying

Now, for unknown query input $\mathbf{x}_q$, the output can be estimated as,

$$\mathbf{y}_q = X_q \boldsymbol{\theta}^*$$



$$y_q = \underbrace{\theta_0}_{=2.4} + \underbrace{\theta_1}_{=1.03} x_q$$

Linear Regression: Solution 2
Maximum Likelihood Estimate (MLE)

Alternatively, the same $\boldsymbol{\theta}^*$ can be obtained by maximizing the likelihood,

$$
\begin{aligned}
\boldsymbol{\theta}^*_{ML} &= \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|X, \boldsymbol{\theta}) \quad \text{(maximize likelihood)} \\
&= \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{y}|X, \boldsymbol{\theta}) \\
&= \arg\min_{\boldsymbol{\theta}} - \log p(\mathbf{y}|X, \boldsymbol{\theta}) \quad \text{(minimize negative log likelihood (NLL))} \\
&= \arg\min_{\boldsymbol{\theta}} (\mathbf{y} - X\boldsymbol{\theta})^\top (\mathbf{y} - X\boldsymbol{\theta}) \\
&= \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \\
&= \boldsymbol{\theta}^*_{OLS}
\end{aligned}
$$

$$
p(\mathbf{y}|X, \theta) = \frac{1}{\sqrt{(2\pi)^{D+1}|\sigma^2 I|}} \exp\left(-\frac{1}{2}(\mathbf{y} - X\theta)^\top (\sigma^2 I)^{-1} (\mathbf{y} - X\theta)\right)
$$

Linear Regression: Solution 3

Maximum-A-Posterior (MAP)

Alternatively, $\boldsymbol{\theta}^*$ can be obtained by maximizing the the posterior distribution (i.e. the mode of the posterior distribution),

$$
\begin{aligned}
\boldsymbol{\theta}^*_{MAP} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|X, \mathbf{y}) \quad \textit{maximum-a-posteriori (MAP)} \\
&= \arg\max_{\boldsymbol{\theta}} \frac{p(\mathbf{y}|X, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \\
&= \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|X, \boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \underbrace{\log p(\mathbf{y}|X, \boldsymbol{\theta})}_{\text{log likelihood}} + \underbrace{\log p(\boldsymbol{\theta})}_{\text{log prior}}
\end{aligned}
$$

$$
\underbrace{p(\boldsymbol{\theta}|X, \mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y}|X, \boldsymbol{\theta})}^{\text{likelihood}} \times \overbrace{p(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{y})}_{\text{marginal likelihood}}}
$$

Linear Regression: Solution 3
Maximum-A-Posterior (MAP)

- The log-prior acts as a regularizer (penalizer) and prevents over-fitting/multicollinearity
- If $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \sigma_0^2 I)$,

$$\boxed{\boldsymbol{\theta}^*_{MAP} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y}}, \quad \lambda = \sigma^2 / \sigma_0^2$$

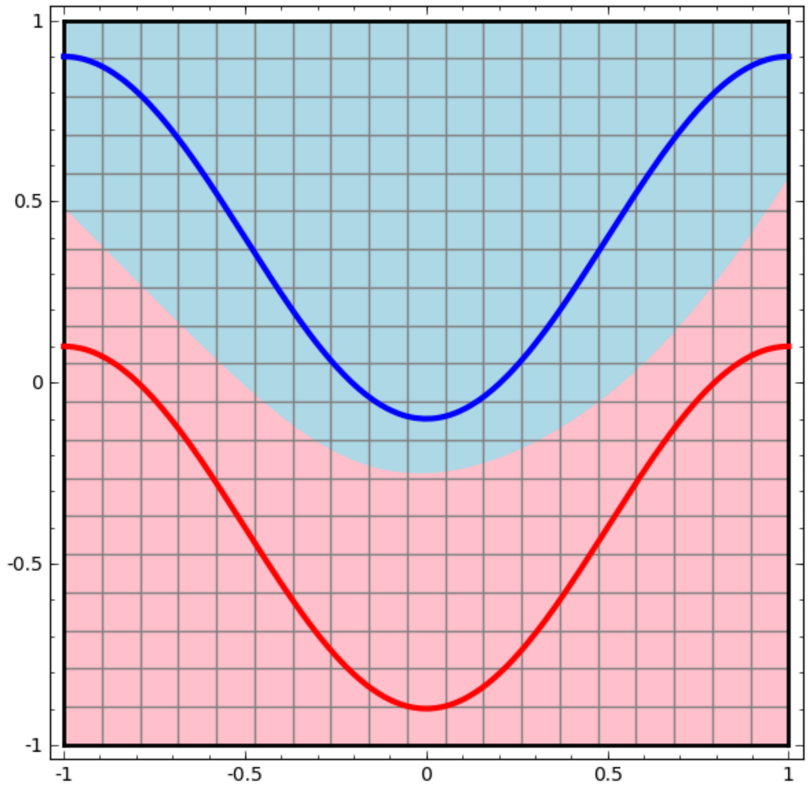
  - This is equivalent to ridge regression (Tikhonov or L2 regularization)
  - λI term improves the numerical stability of the inversion

---

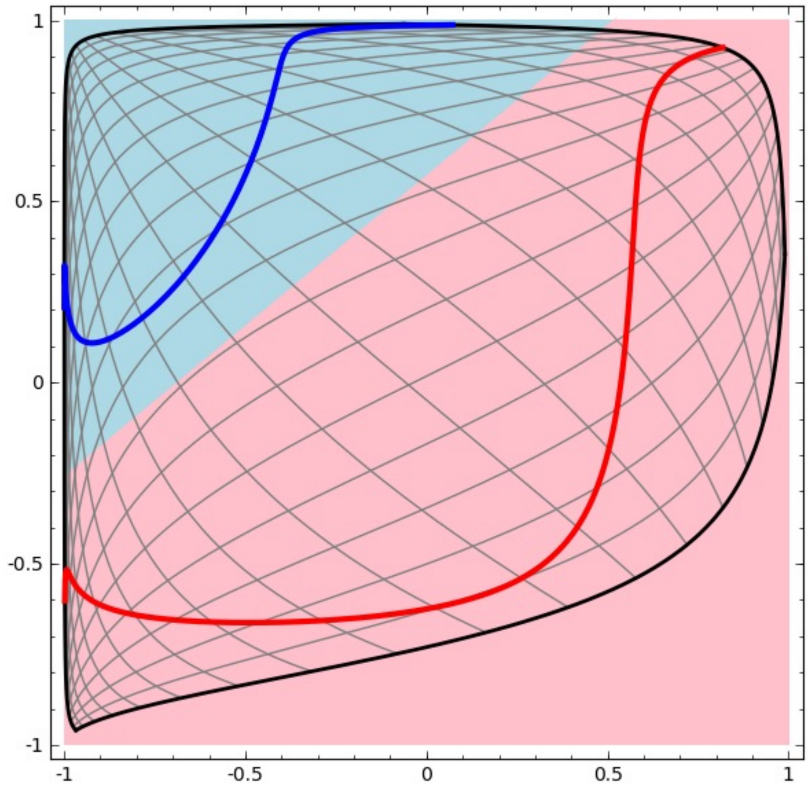
$$\boldsymbol{\theta}^*_{ML} = (X^\top X)^{-1} X^\top \mathbf{y}$$

Neural Networks: Manifolds View

https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/



$\text{activation}(\sum w_i x_i)$



Metrics

Confusion Matrix

		Predicted condition	
	Total population = P + N	Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

Confusion Matrix

		Predicted condition			
	Total population $= P + N$	**Predicted Positive (PP)**	**Predicted Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP), hit**	**False negative (FN), type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP), type I error, false alarm, overestimation**	**True negative (TN), correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2 PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Wikipedia article has the best summary table I've seen so far

Precision vs. Recall

- Precision (TP/TP+FP): Accuracy of the positive predictions.
- Recall/sensitivity (TP/TP+FN): Ability to capture/retrieve all positive samples. Useful when missing positives is a bad thing.
- F1 score: $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$
- Sometimes working with precision or recall separately makes more sense than using F1 (e.g., in a medical diagnosis, we might need to focus more on maximizing precision)
- When there is an imbalance in the number of ground truth positive and negative samples, F1 score will be biased
- For different classification thresholds, we can plot the precision-recall curve. Generally, if we decrease the classification threshold (say, from 0.5 to 0.3), TP+FP go up (something remotely looks like a positive will now be a positive, whether correct or not). This will increase fall positives (because false positives typically lie around 0.5) which will then decrease precision.

Spam vs. ham
Cancer diagnosis