Interpretability and functional transparency

Tommi Jaakkola

in collaboration with
David Alvarez Melis, Guang-He Lee, et al.

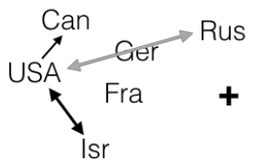
Many facets of “interpretability”

Many facets of "interpretability"

uncover
causal mechanisms

UN resolution

\+



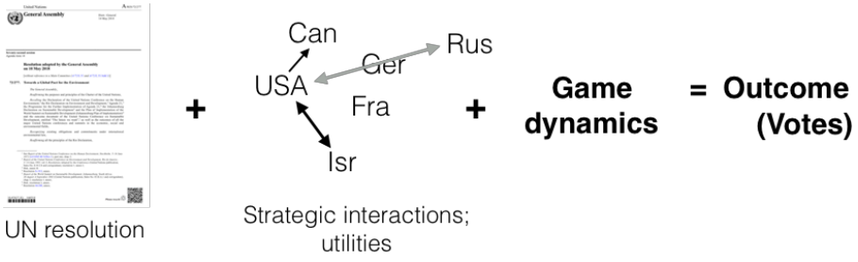
Strategic interactions; utilities

\+ **Game dynamics** = **Outcome (Votes)**

[Garg et al. 2018]

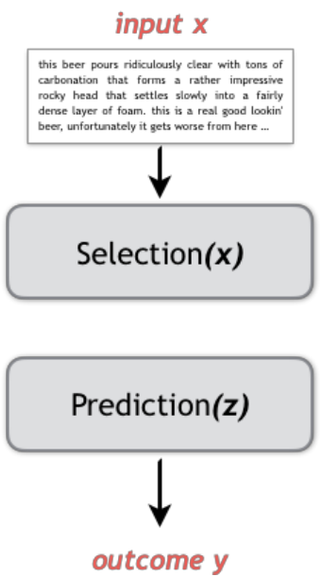
Many facets of "interpretability"

uncover causal mechanisms



[Garg et al. 2018]

learn to highlight relevance



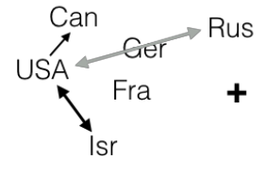
[Lei et al. 2016; Jin et al. 2017]

Many facets of "interpretability"

uncover causal mechanisms

UN resolution

\+

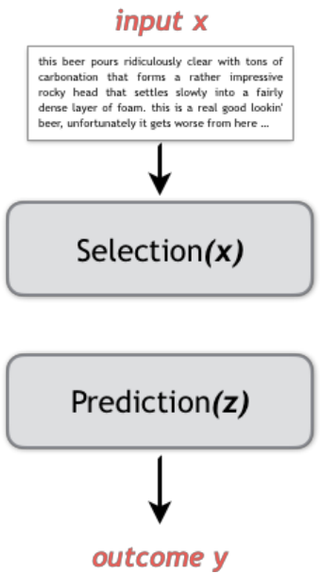


Strategic interactions; utilities

\+ **Game dynamics** = **Outcome (Votes)**

[Garg et al. 2018]

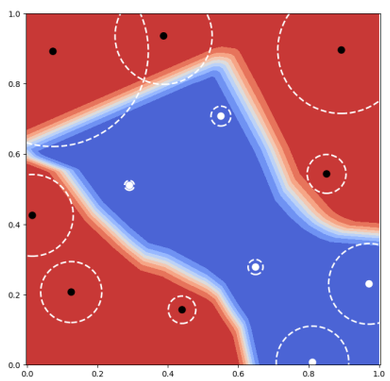
learn to highlight relevance



[Lei et al. 2016; Jin et al. 2017]

learn functional transparency

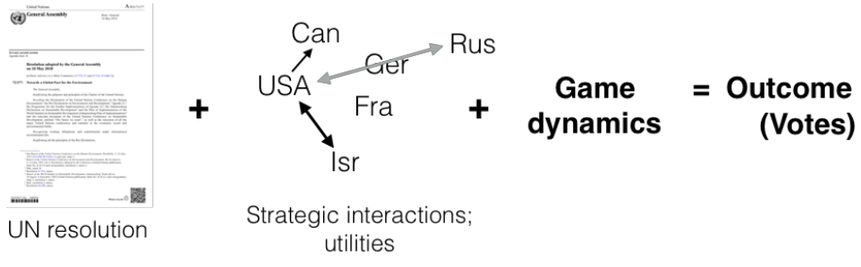
$$f(\text{doc}) = \theta(\text{doc}) \bullet h(\text{doc})$$



[Lee et al. 2018; Alvarez et al. 2018]

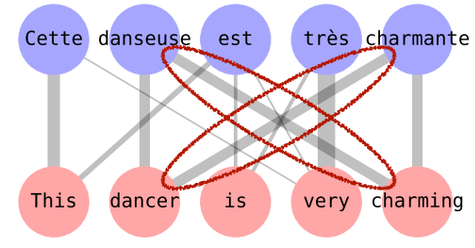
Many facets of "interpretability"

uncover causal mechanisms



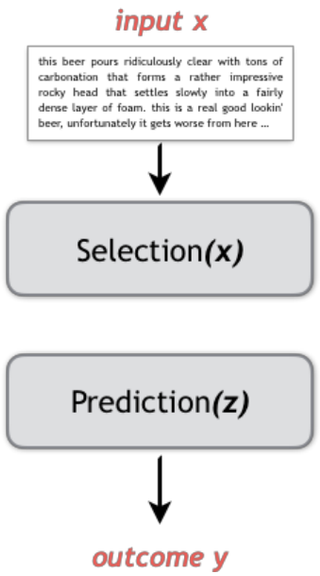
[Garg et al. 2018]

summarize by causal relations



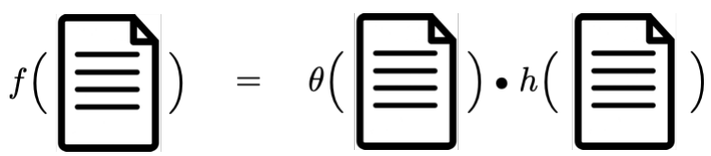
[Alvarez et al. 2017]

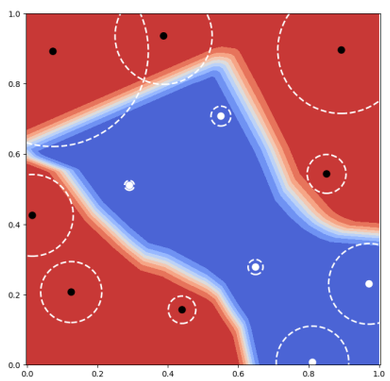
learn to highlight relevance



[Lei et al. 2016; Jin et al. 2017]

learn functional transparency

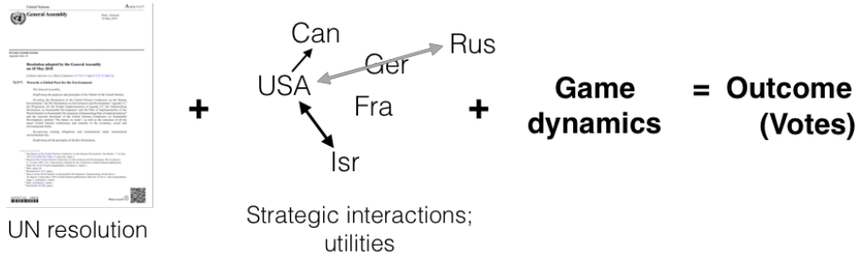




[Lee et al. 2018;
Alvarez et al. 2018]

Many facets of "interpretability"

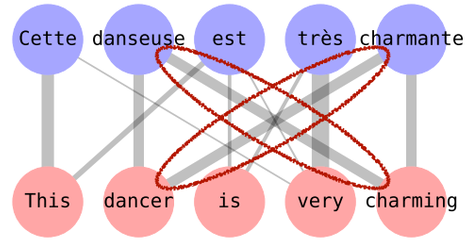
uncover causal mechanisms



[Garg et al. 2018]

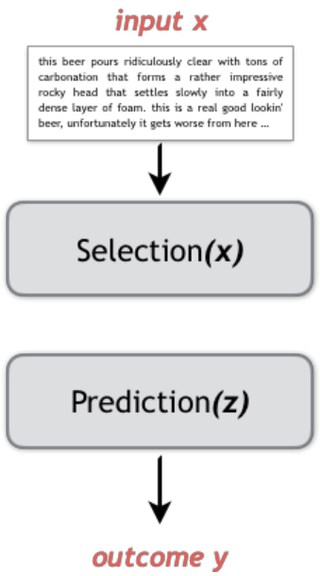
from phenomena to models

summarize by causal relations



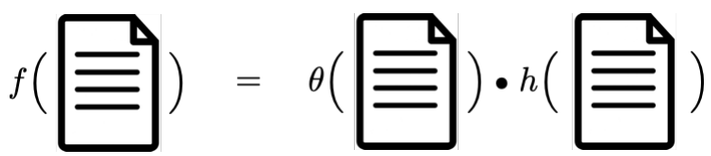
[Alvarez et al. 2017]

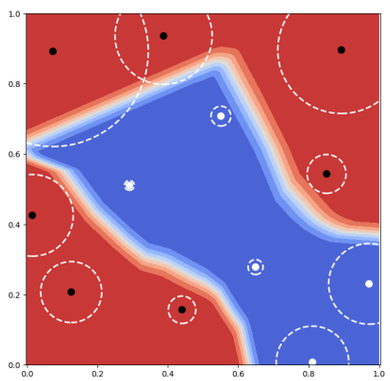
learn to highlight relevance



[Lei et al. 2016; Jin et al. 2017]

learn functional transparency

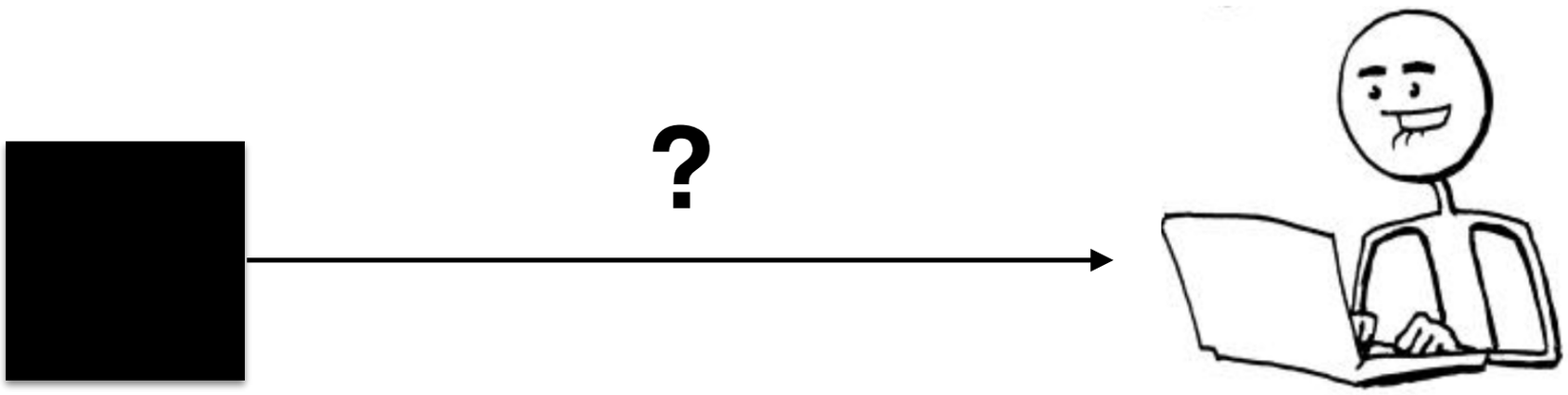




[Lee et al. 2018;
Alvarez et al. 2018]

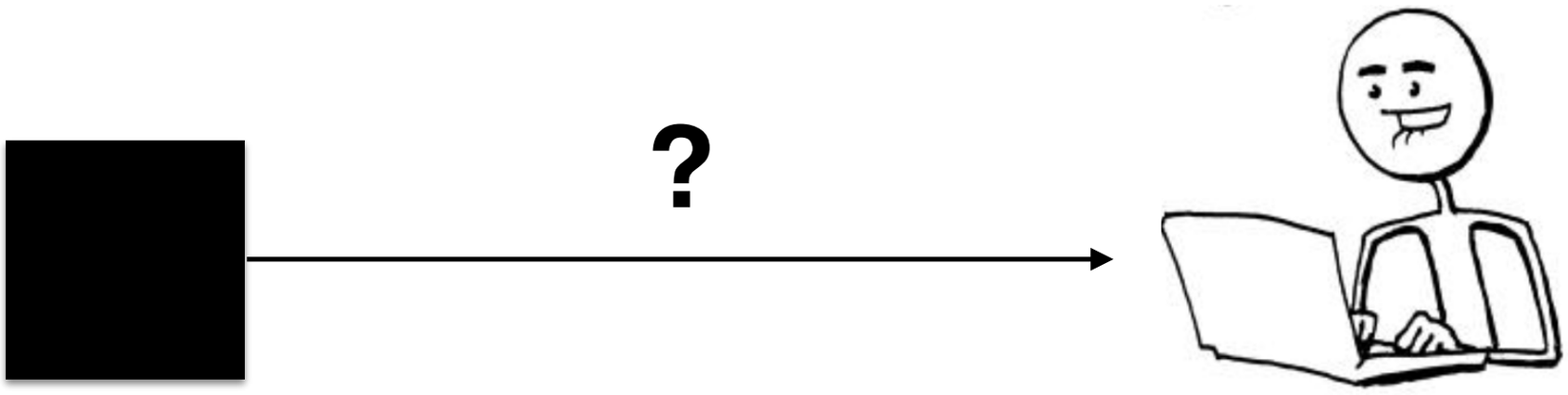
“Interpretability”

- (Human) interpretability
  - features (that make sense)
  - relevance (what information is used to make a decision)
  - reasoning (mechanism used to arrive at the decision)
  - etc.



“Interpretability”

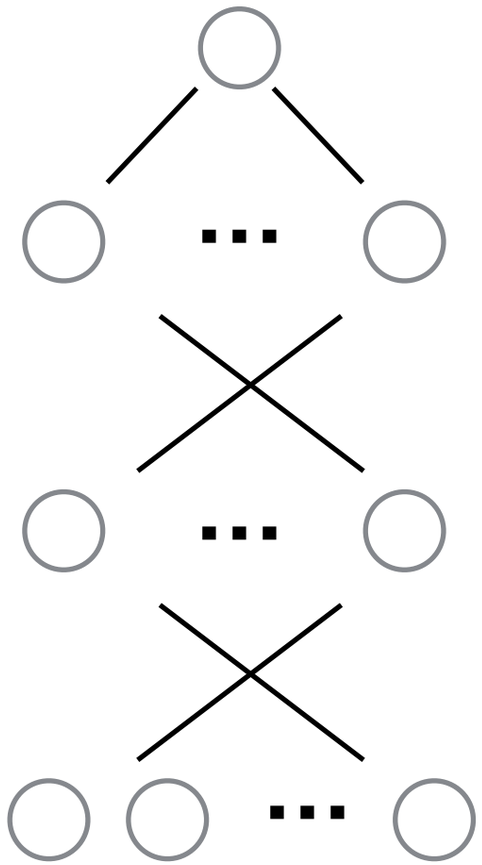
- (Human) interpretability
  - features (that make sense)
  - relevance (what information is used to make a decision)
  - reasoning (mechanism used to arrive at the decision)
  - etc.



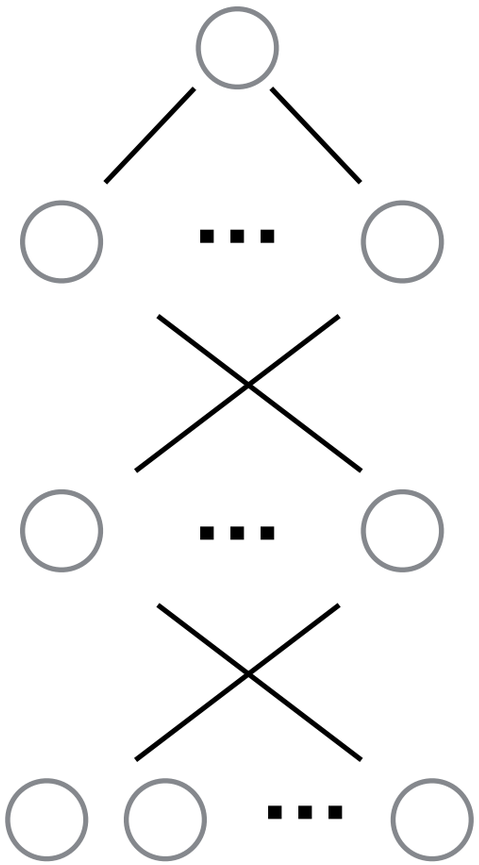
- Functional transparency
  - guaranteed properties, including robustness

Molding for transparency

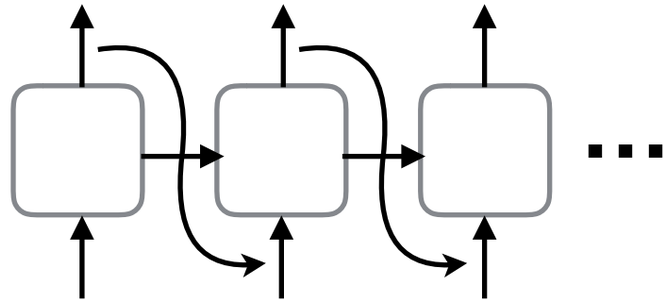
- deep locally linear models



- ReLU networks with large linear regions

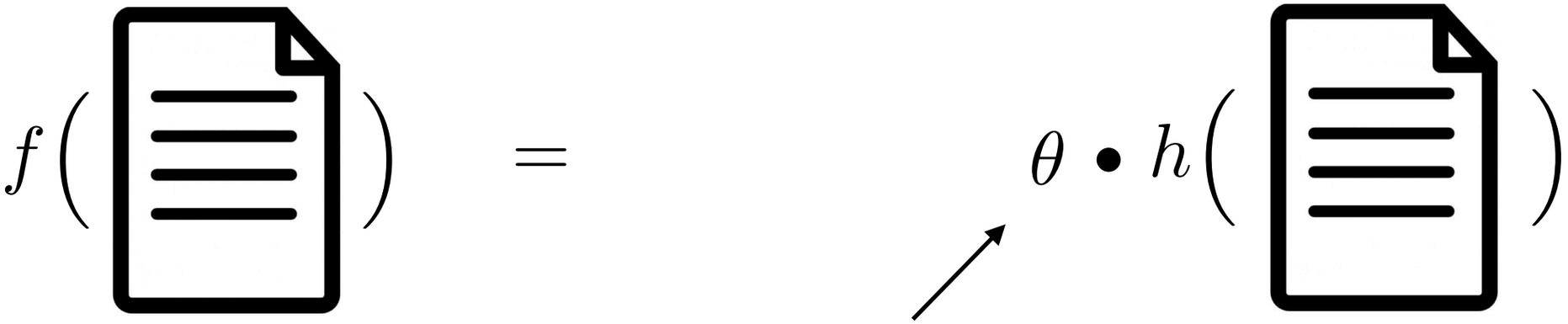


- temporal models with desired local behavior



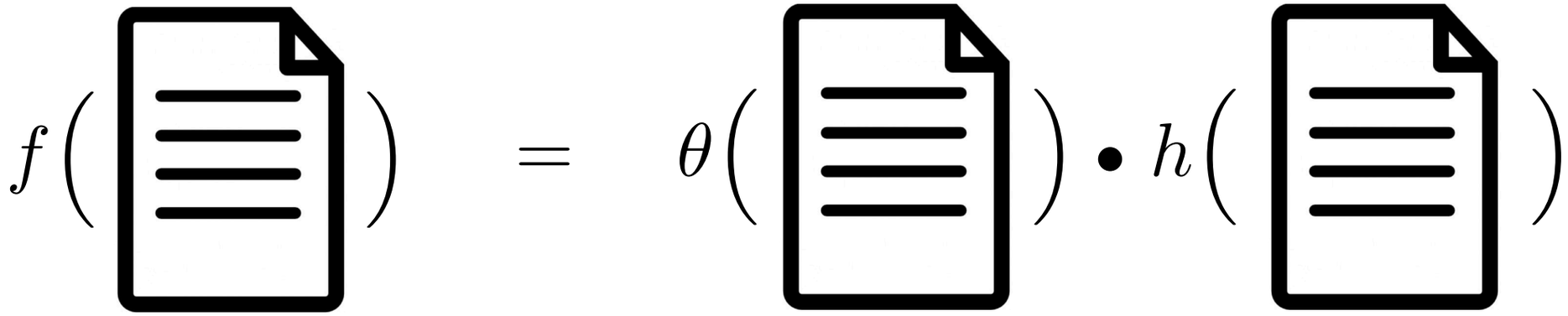
Deep locally linear models

- A self-explaining architecture from a deep linear model



Deep locally linear models

- A self-explaining architecture from a deep linear model



A deep linear model | deep linear coefficients | simple feature transformation

- Arbitrarily powerful, but not (linearly) interpretable

Deep locally linear models

- A self-explaining architecture from a deep linear model

$$f(\quad) = \theta(\quad) \cdot h(\quad)$$

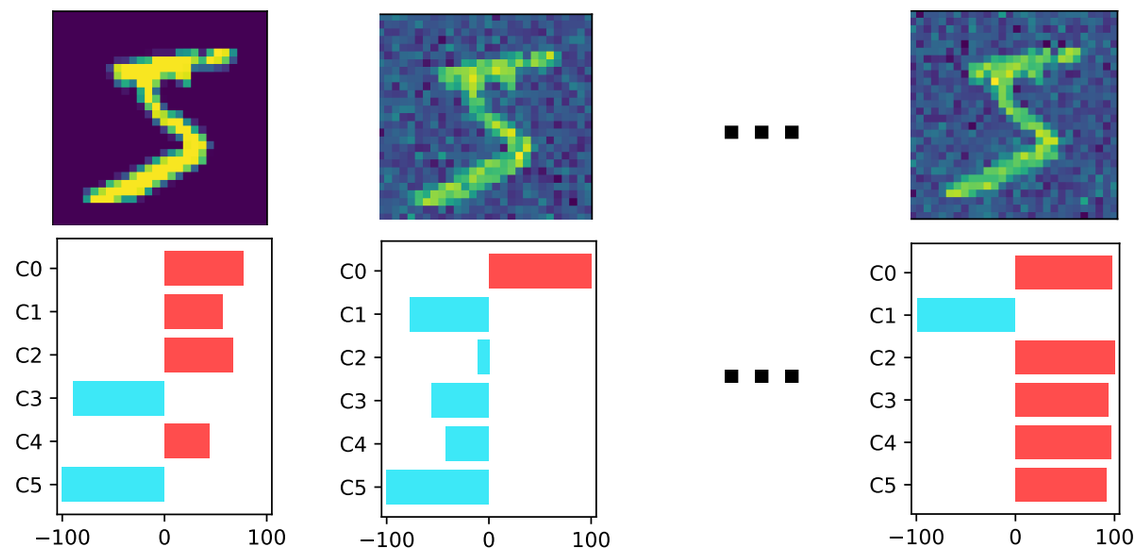
A deep linear model

deep linear coefficients

simple feature transformation

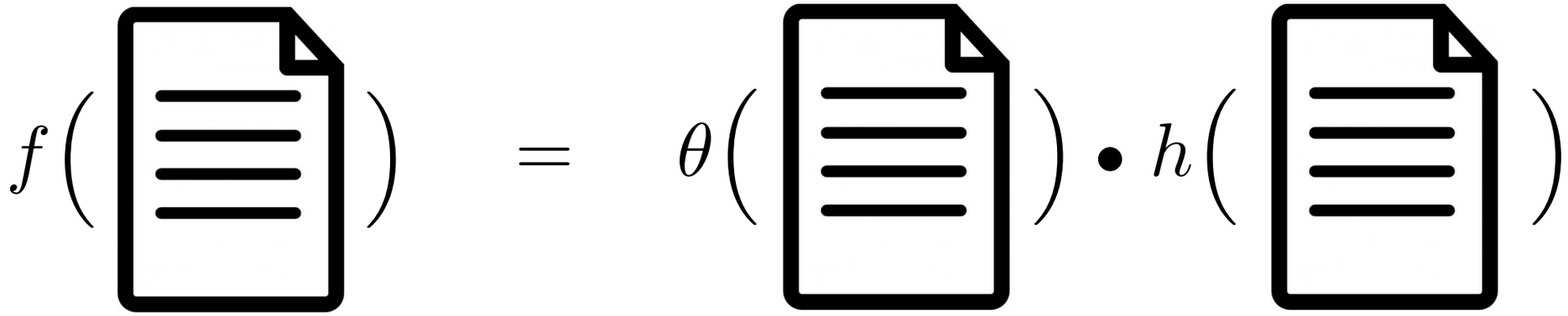
- Arbitrarily powerful, but not (linearly) interpretable

deep linear coefficients



Deep locally linear models

- A self-explaining architecture from a deep linear model



A deep linear model | deep linear coefficients | simple feature transformation

- We can regularize the model for local interpretability

$$R(\theta) = \|\nabla f(x) - \theta(x)^T J_{h;x}\|^2$$

(Alvarez et al. 2018)

Deep locally linear models

- A self-explaining architecture from a deep linear model

$$f(\quad) = \theta(\quad) \cdot h(\quad)$$

A deep linear model

deep linear coefficients

simple feature transformation

- We can regularize the model for local interpretability

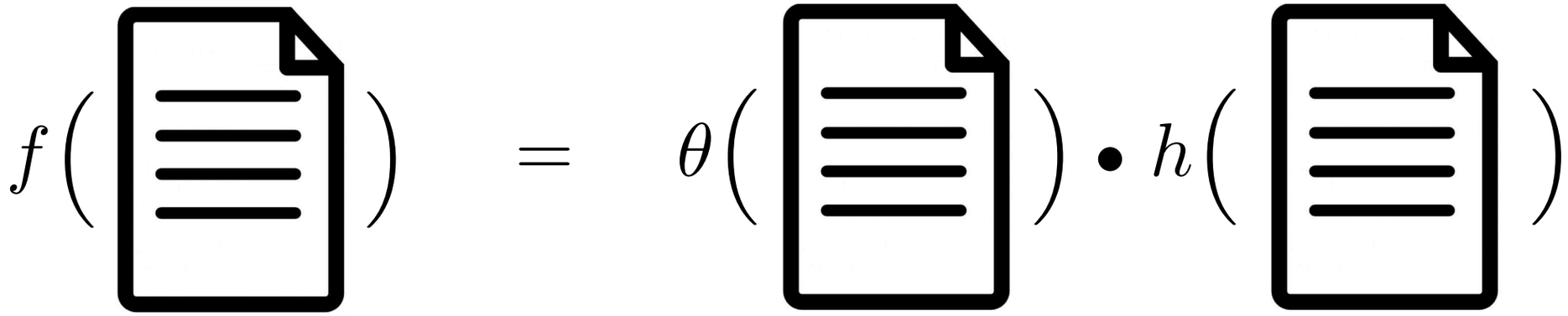
$$R(\theta) = \|\nabla f(x) - \theta(x)^T J_{h;x}\|^2$$

locally linear "witness"

(Alvarez et al. 2018)

Deep locally linear models

- A self-explaining architecture from a deep linear model



A deep linear model | deep linear coefficients | simple feature transformation

- We can regularize the model for local interpretability

$$R(\theta) = \|\nabla f(x) - \theta(x)^T J_{h;x}\|^2$$

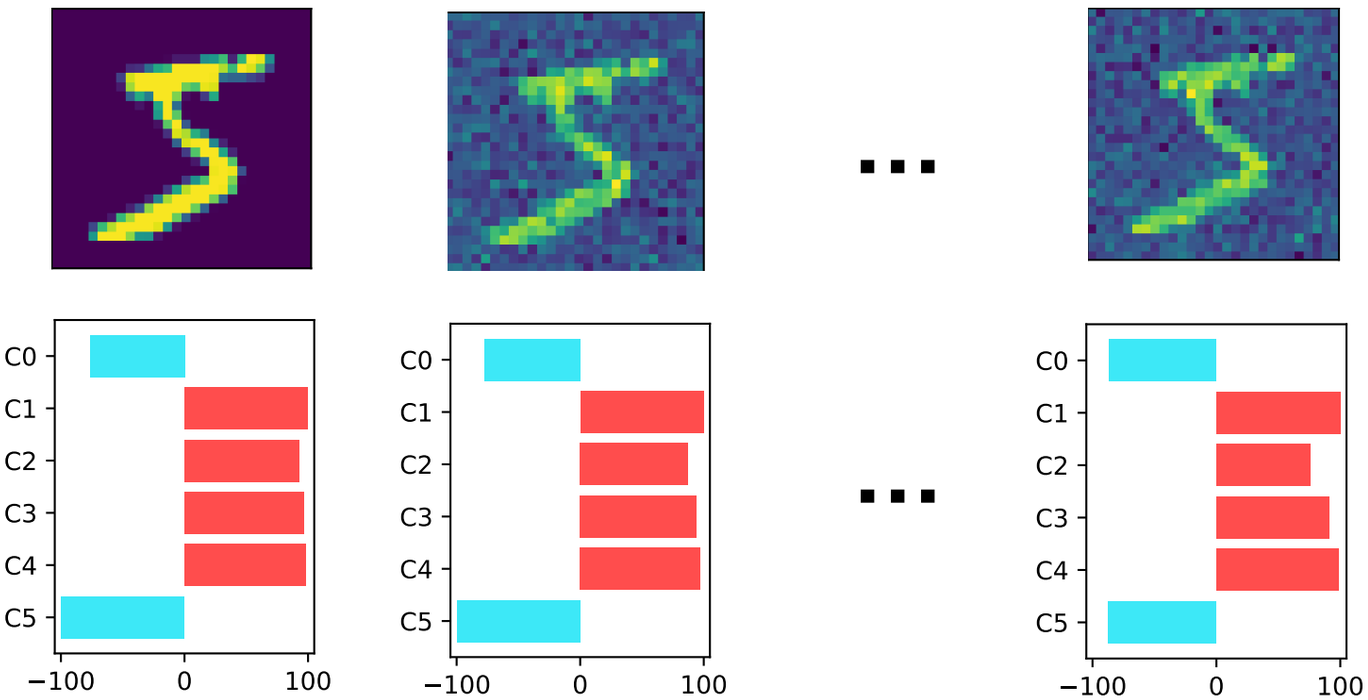
- Generalizable beyond linear (monotone, separable)

(Alvarez et al. 2018)

Deep locally linear models

- A self-explaining architecture from a deep linear model

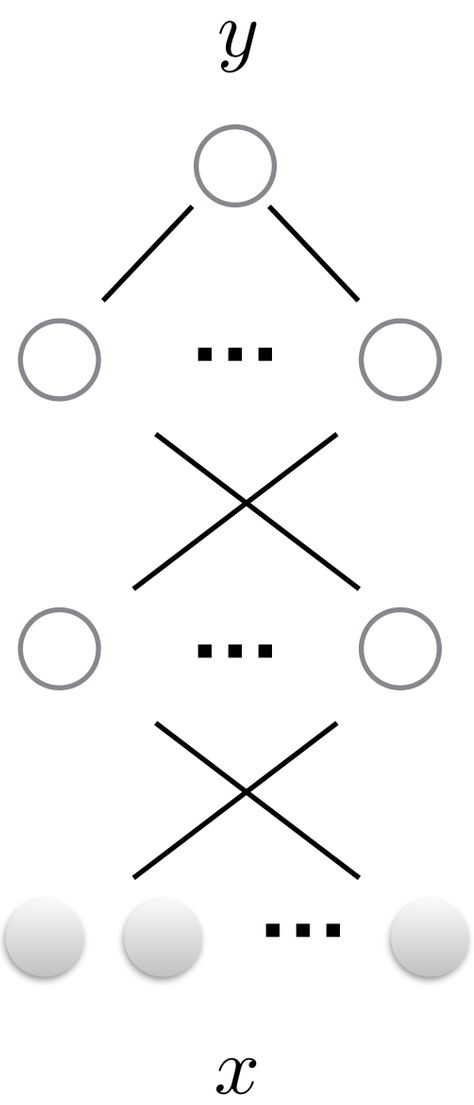
locally interpretable deep linear coefficients



[Alvarez et al. 2018]

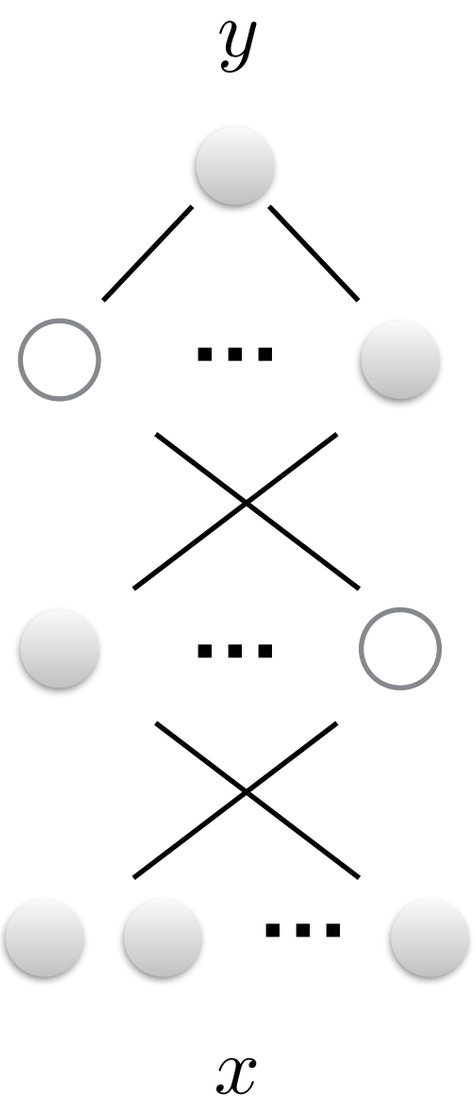
Expanding linear regions

- E.g., a ReLU network (locally linear)



Expanding linear regions

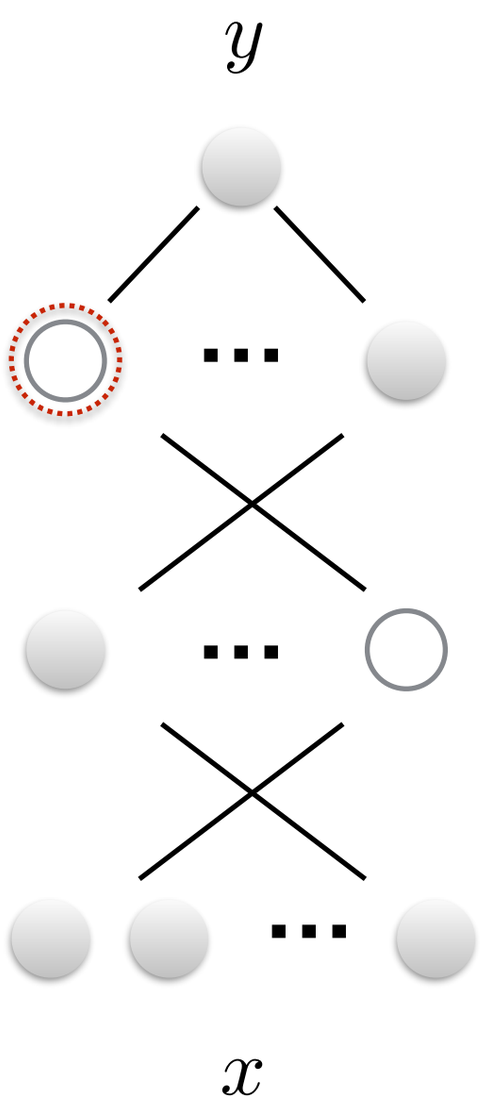
- E.g., a ReLU network (locally linear)



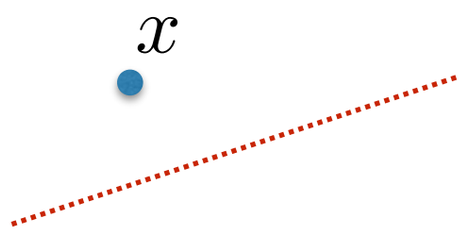
  - "activation pattern"

Expanding linear regions

- E.g., a ReLU network (locally linear)

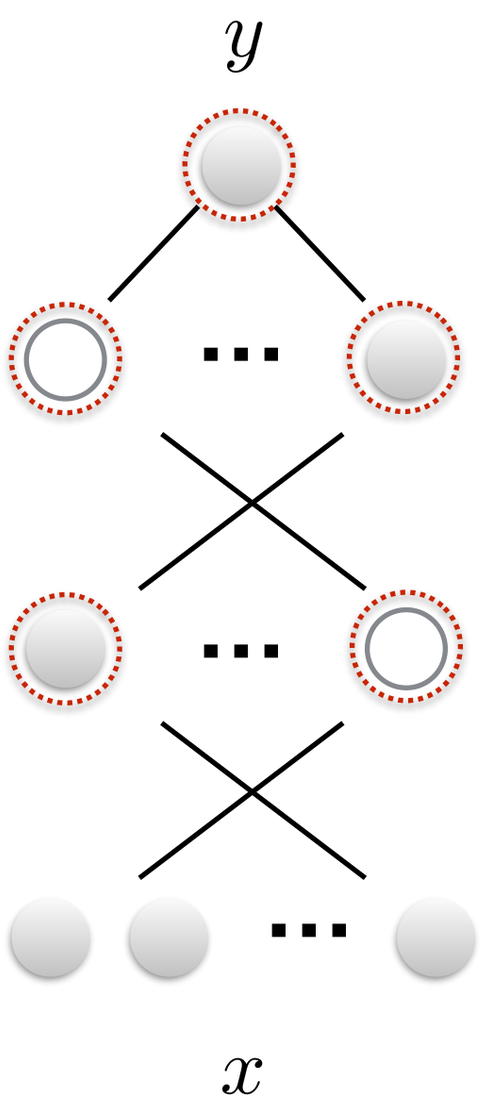


- "activation pattern"

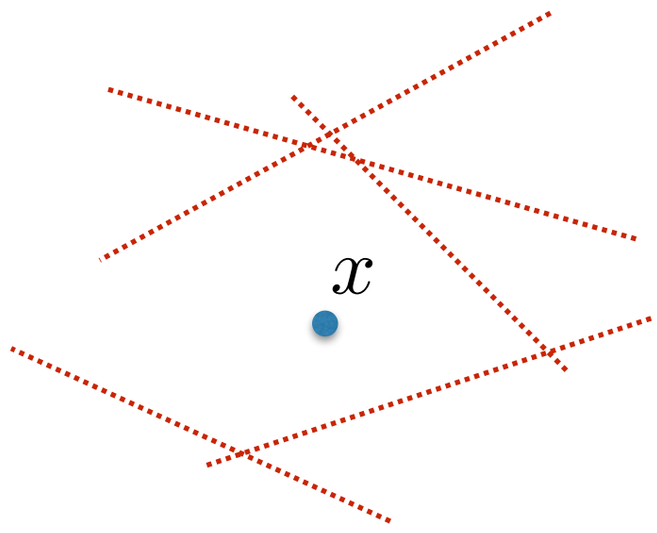


Expanding linear regions

- E.g., a ReLU network (locally linear)

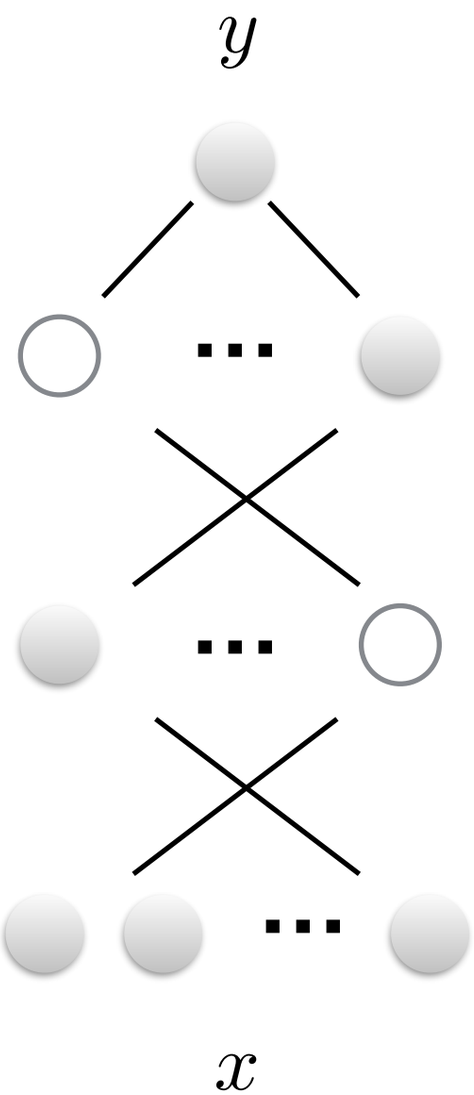


- "activation pattern"

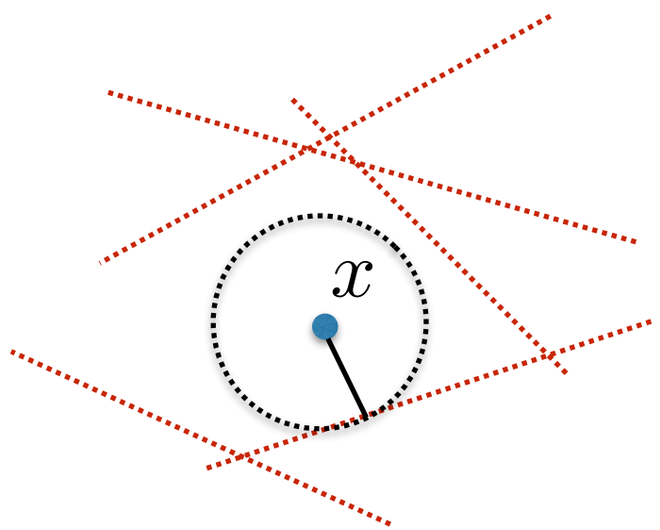


Expanding linear regions

- E.g., a ReLU network (locally linear)

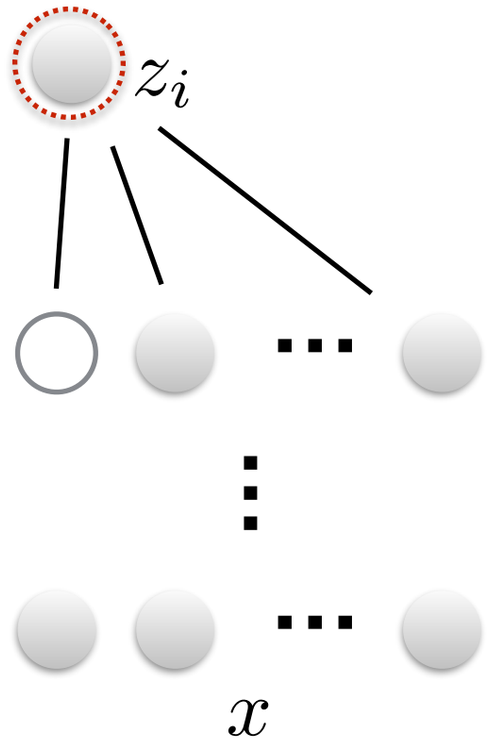


- "activation pattern"
- we can learn the network so as to encourage large linear regions (gradient stability)

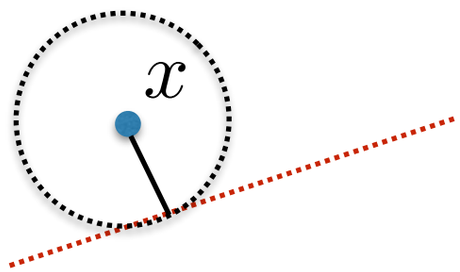


Expanding linear regions

- We can aim to maximize the margin for each neuron

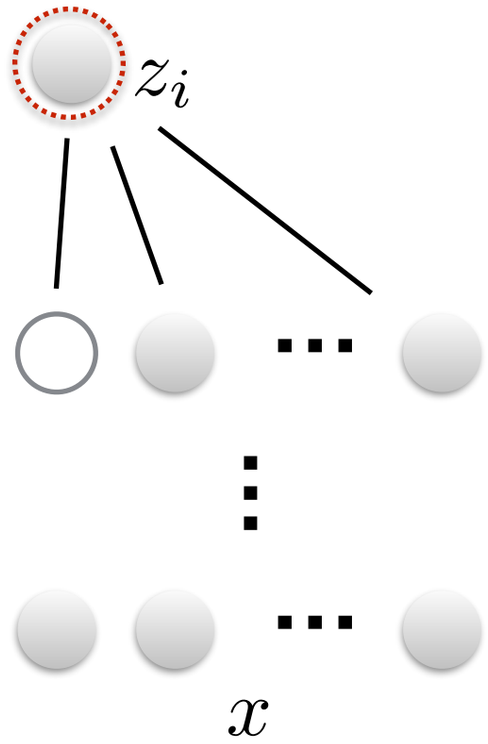


effective linear weights $\nabla_x z_i$



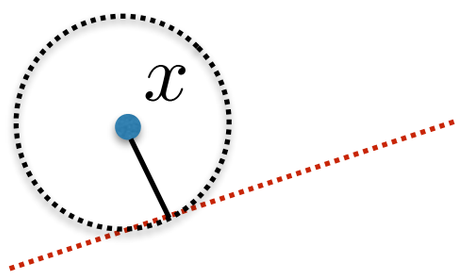
Expanding linear regions

- We can aim to maximize the margin for each neuron



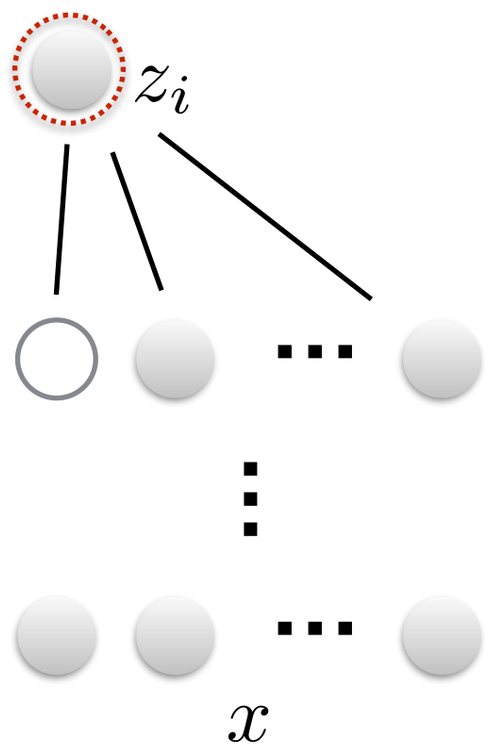
effective linear weights $\nabla_x z_i$

margin $\dfrac{|z_i|}{\|\nabla_x z_i\|}$



Expanding linear regions

- We can aim to maximize the margin for each neuron

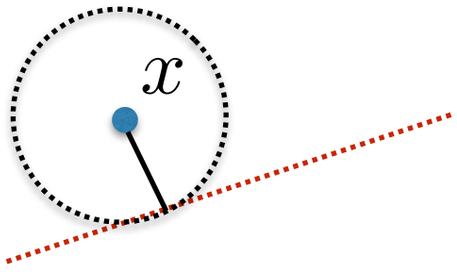


effective linear weights $\nabla_x z_i$

margin $\dfrac{|z_i|}{\|\nabla_x z_i\|}$

relaxed margin regularizer

$$\|\nabla_x z_i\|^2 + C \max(0, 1 - |z_i|)$$

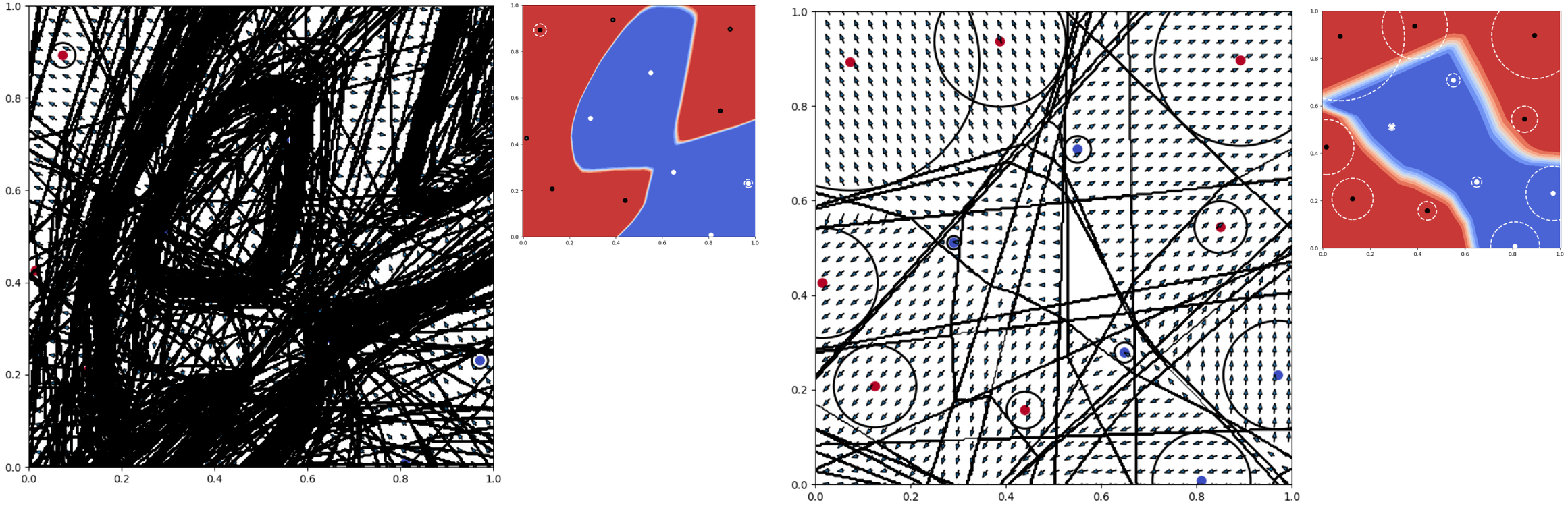


Expanding linear regions

- We maximize a relaxed margin loss

$$\min_{\theta} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \mathcal{L}(f_\theta(\mathbf{x}),\mathbf{y}) + \frac{\lambda}{|\hat{\mathcal{I}}(\mathbf{x},\gamma)|} \sum_{(i,j)\in\hat{\mathcal{I}}(\mathbf{x},\gamma)} \left[\|\nabla_{\mathbf{x}} \mathbf{z}_j^i\|_2^2 + C\max(0, 1 - |\mathbf{z}_j^i|) \right]$$

- A toy example

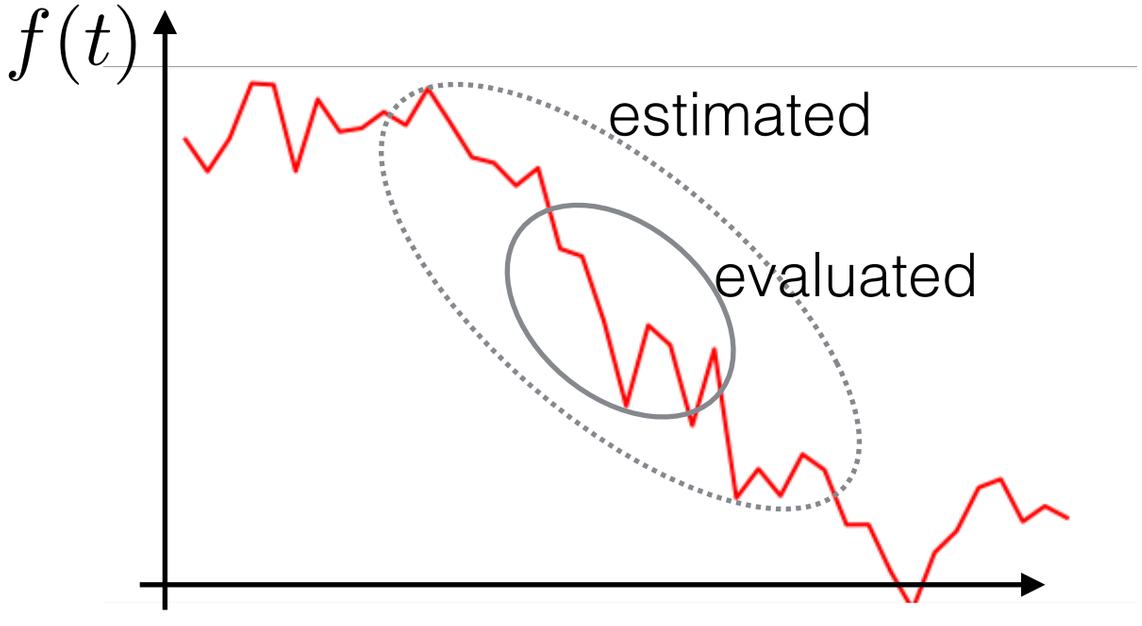


[Lee et al. 2018]

Molding temporal models

- Introducing a local "explainer" as a witness of desired local behavior
- For example:

deep temporal models
that are locally ARMA
(witness: ARMA)



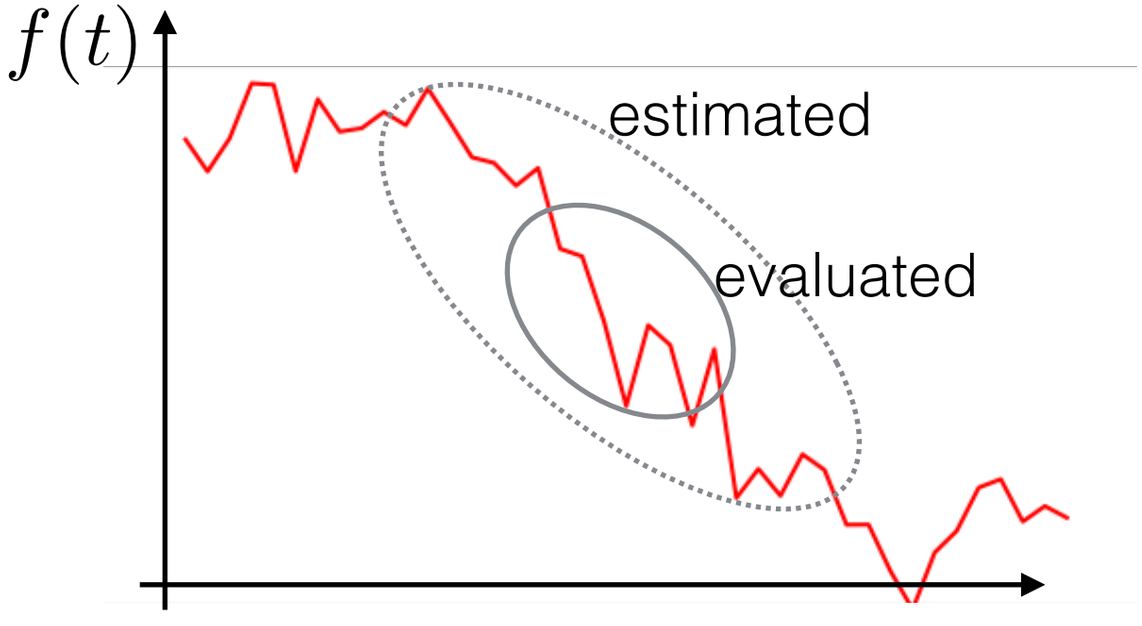
local witness

$$g(t) = \hat{a}_1\, g(t-1) + \hat{a}_2\, g(t-2) + \epsilon$$

Molding temporal models

- Introducing a local "explainer" as a witness of desired local behavior
- For example:

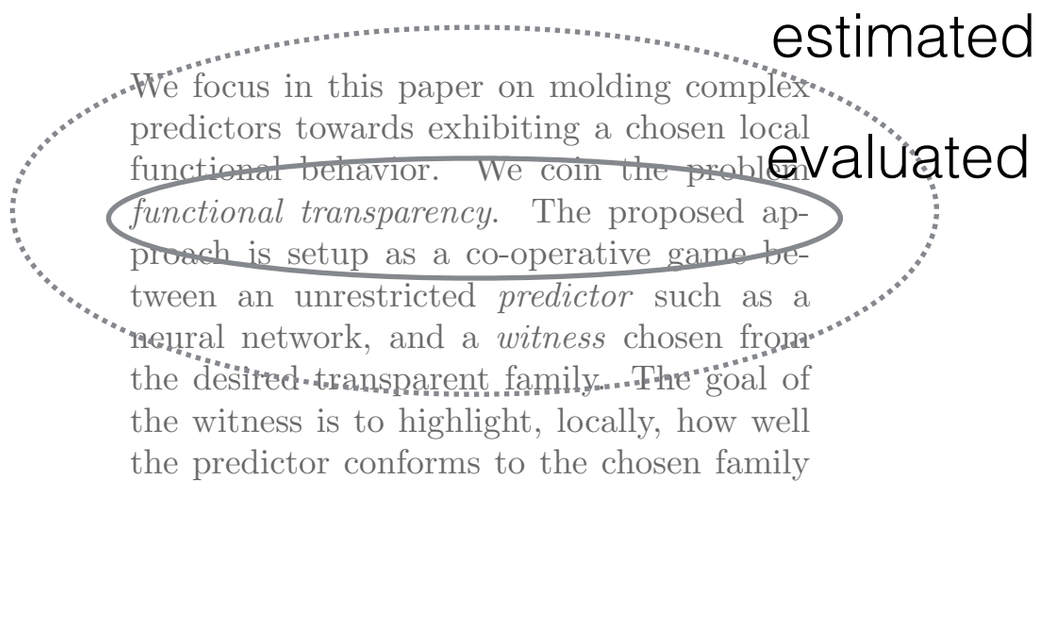
deep temporal models that are locally ARMA (witness: ARMA)



local witness

$$g(t) = \hat{a}_1\, g(t-1) + \hat{a}_2\, g(t-2) + \epsilon_t$$

deep sequence models that are locally bigram (witness: bigram)



local witness

$$\hat{P}(w_t | w_{t-1})$$

A co-operative witness...

- We can mold a complex function to agree locally with the corresponding local witness

$$\hat{f} \leftarrow \arg\min_{f} \sum_{i=1}^{n} \left[\mathcal{L}(f(x_i), y_i) + \lambda d(f(x_i), \hat{g}(x_i)) \right]$$

loss on observations

discrepancy with the local witness

$$\hat{g} \leftarrow \arg\min_{g} \sum_{x_j \in B(x_i)} d(\hat{f}(x_j), g(x_j))$$

witness at x_i

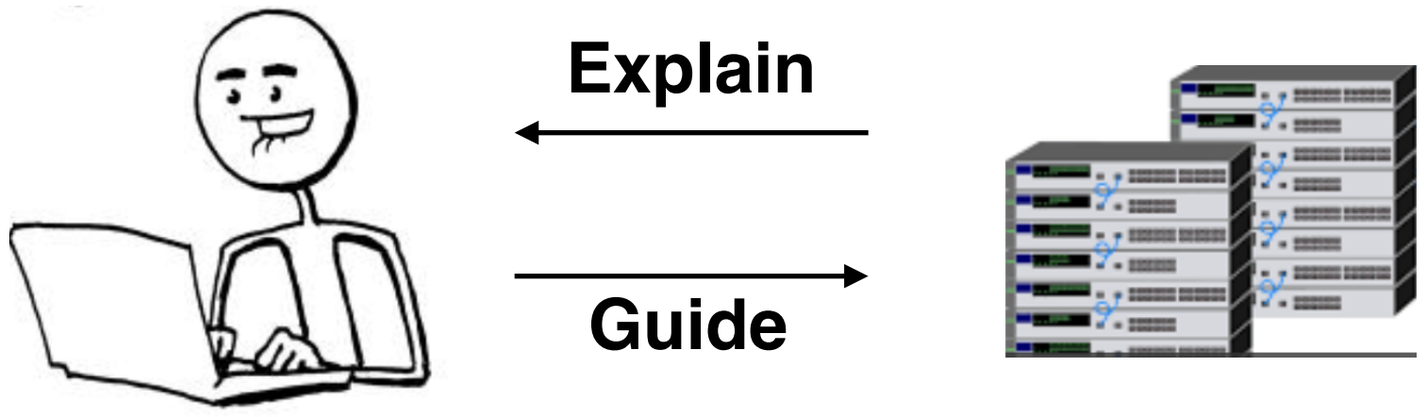
locally tailored witness

- (an asymmetric game, information sets do not agree)

[Lee et al. 2018]

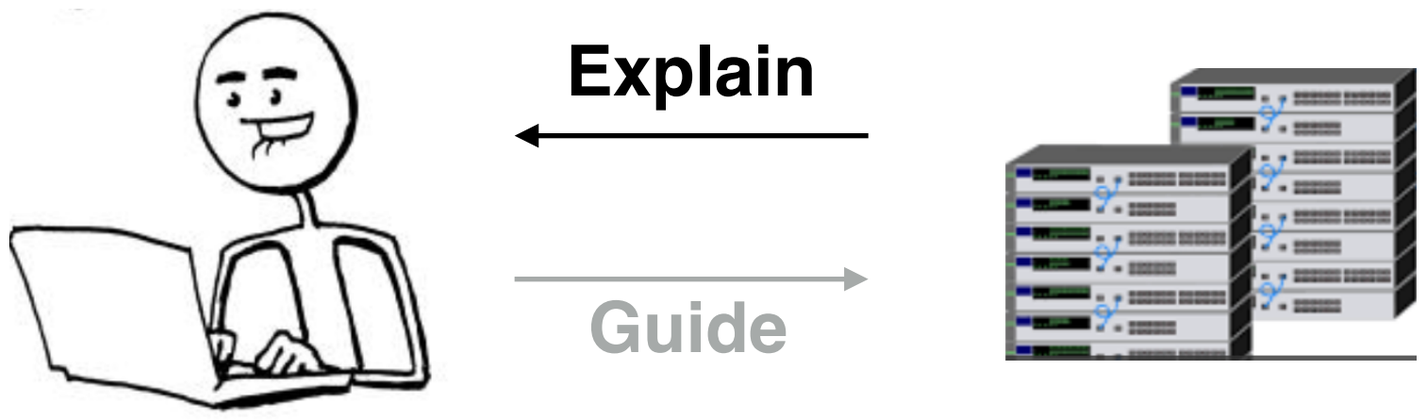
Interpretability - the broader view

- The overall goal is about two-way communication, more formal view of "interpretability"



Interpretability - the broader view

- The overall goal is about two-way communication, more formal view of “interpretability”



- **Self-explaining models**: models are trained to exhibit desirable properties (causal, functional, relevance, etc)
- **Multi-resolution explanations**