Deep Probability Estimation

Carlos Fernandez-Granda

www.cims.nyu.edu/~cfgranda

Acknowledgements

Research partially supported by NSF and NIH

Joint work with

- Sheng Liu
- Aakash Kaku
- Weicheng Zhu
- Matan Leibovich
- Sreyas Mohan
- Boyang Yu
- Haoxiang Huang
- Laure Zanna
- Narges Razavian
- Jonathan Niles-Weed

Probability Estimation

Goal: Estimate probability of uncertain events from high-dimensional input (images, videos)

Probability Estimation

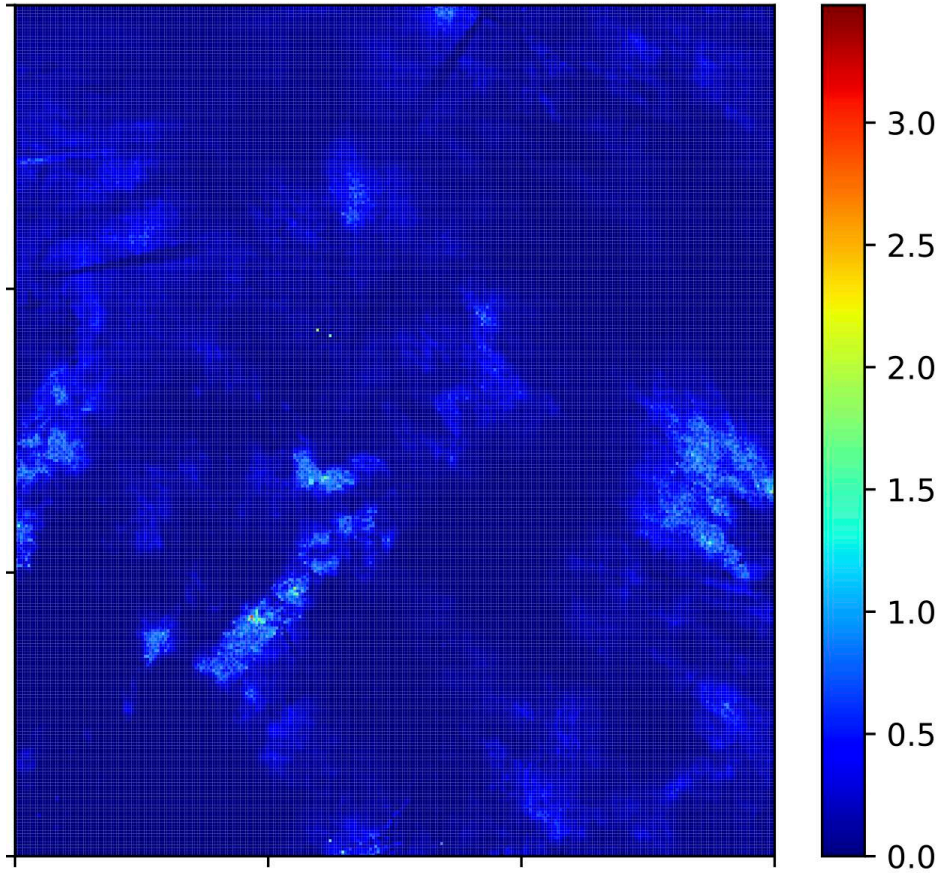
Goal: Estimate probability of uncertain events from high-dimensional input (images, videos)

Not equivalent to classification because of inherent uncertainty

Weather Forecasting

Event of interest: Will it rain?

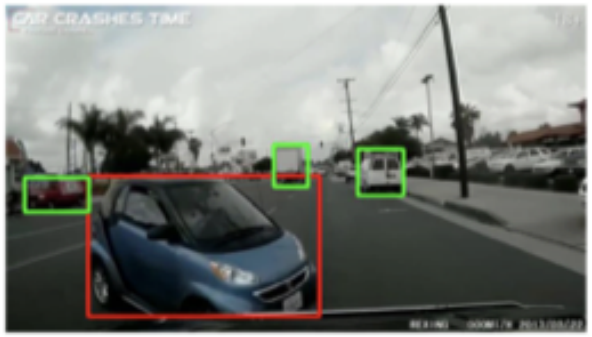
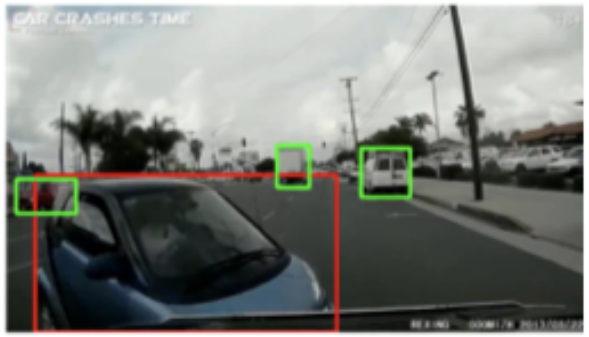
Input: Map of past precipitation



Collision Prediction

Event of interest: Will there be a car crash?

Input: Dashboard video

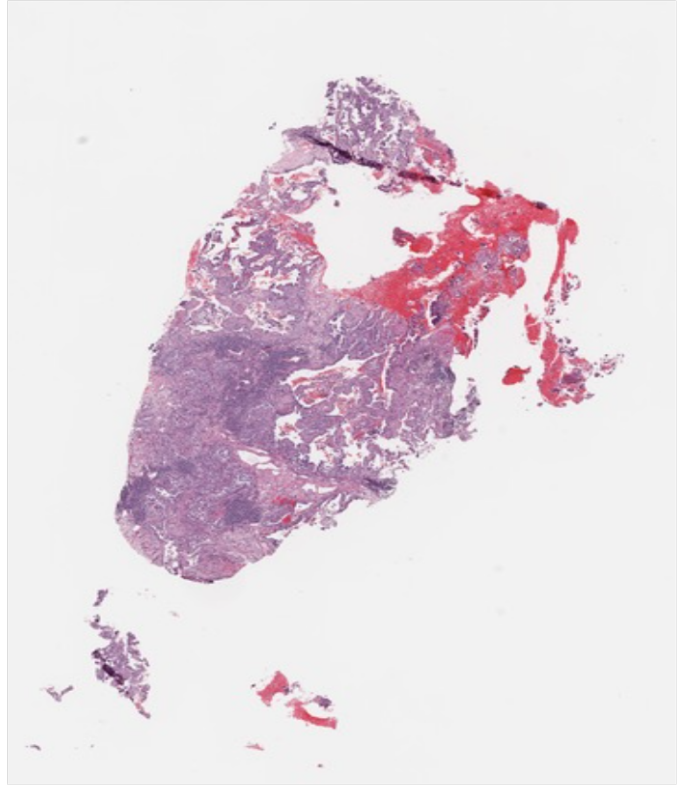




Crash to not crash: Learn to identify dangerous vehicles using a simulator. Kim, H., Lee, K., Hwang, G., and Suh, C. AAAI 2019

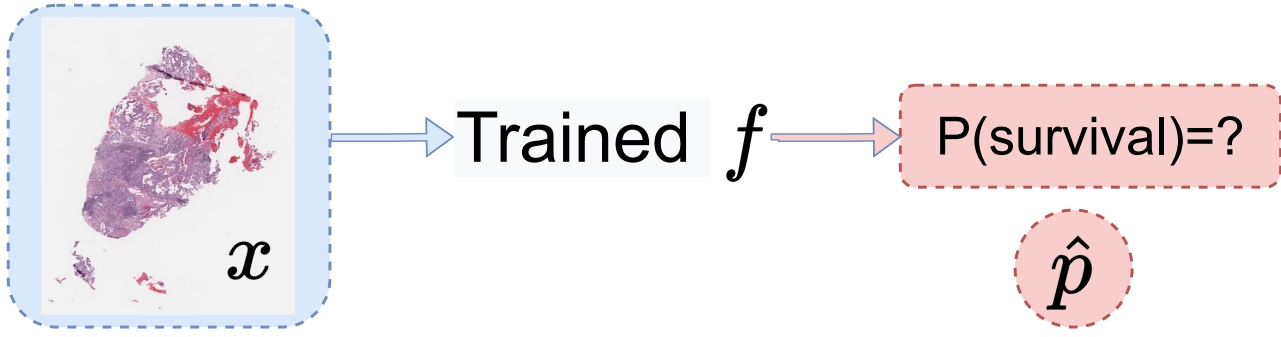
Survival Prediction

Event of interest: Will a cancer patient die in the next 5 years?

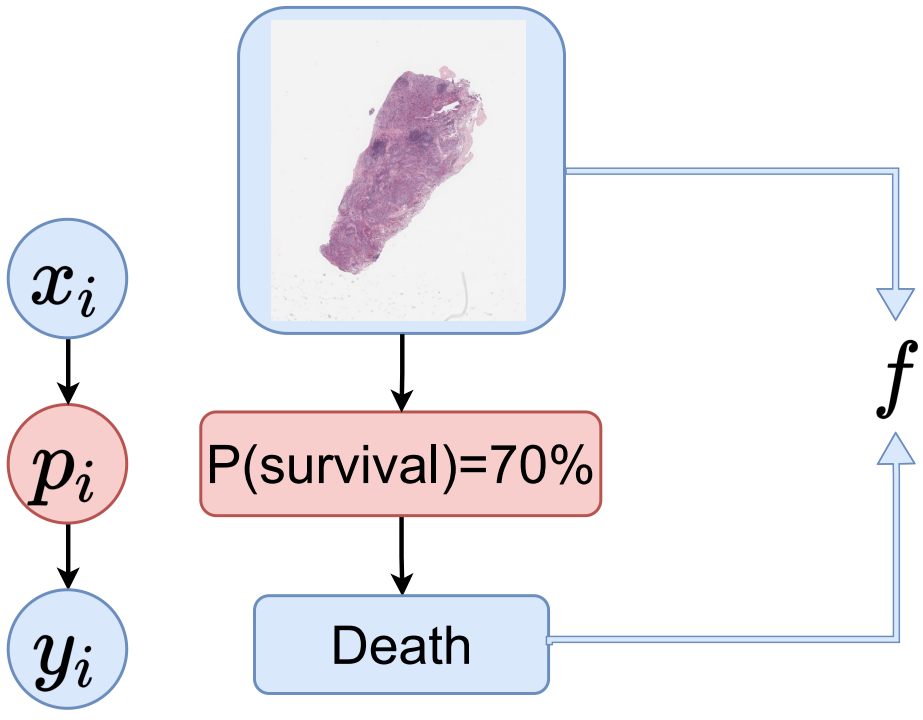
Input: Histopathology slide



Probability estimation via deep learning



Training



Evaluation

How do we evaluate whether estimated probabilities are calibrated?

Evaluation

How do we evaluate whether estimated probabilities are calibrated?

Among examples with estimated probability 0.7, are 70% equal to one?

Evaluation

How do we evaluate whether estimated probabilities are calibrated?

Among examples with estimated probability 0.7, are 70% equal to one?

Bin points according to estimated probability

Evaluation

How do we evaluate whether estimated probabilities are calibrated?

Among examples with estimated probability 0.7, are 70% equal to one?

Bin points according to estimated probability

Compare estimated probability to empirical probability in each bin

Evaluation

How do we evaluate whether estimated probabilities are calibrated?

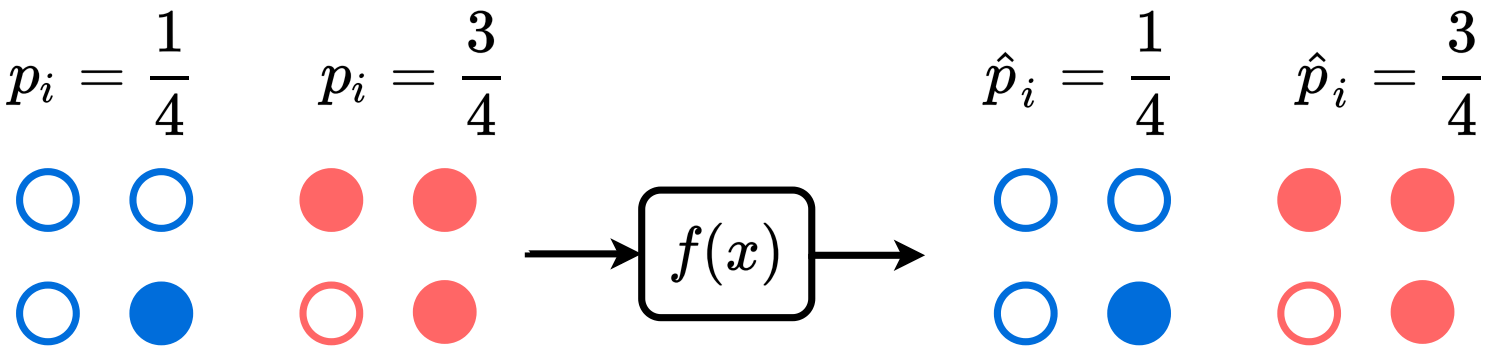
Among examples with estimated probability 0.7, are 70% equal to one?

Bin points according to estimated probability

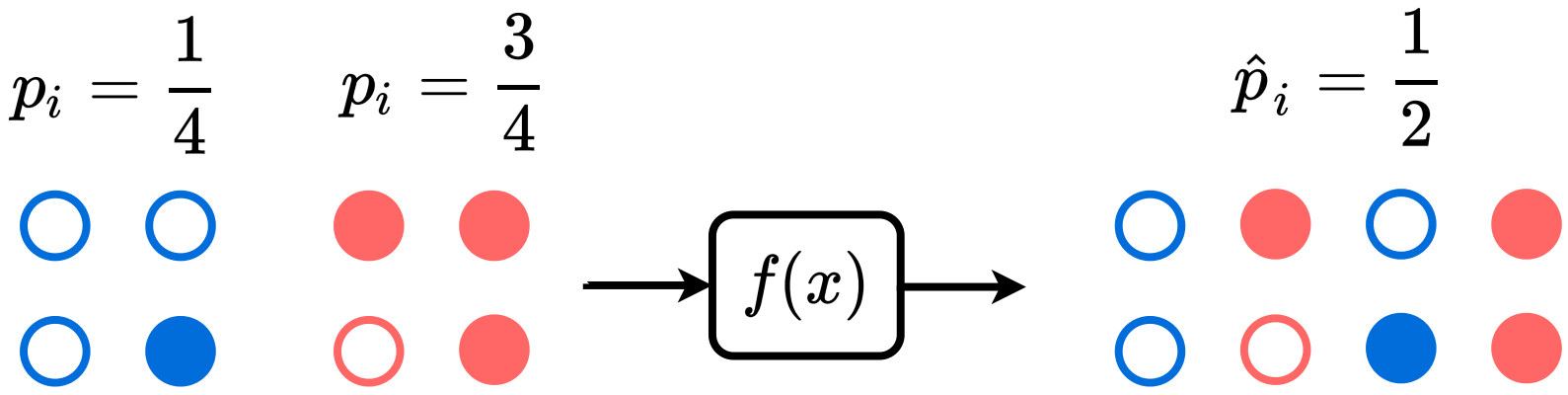
Compare estimated probability to empirical probability in each bin

Is calibration enough?

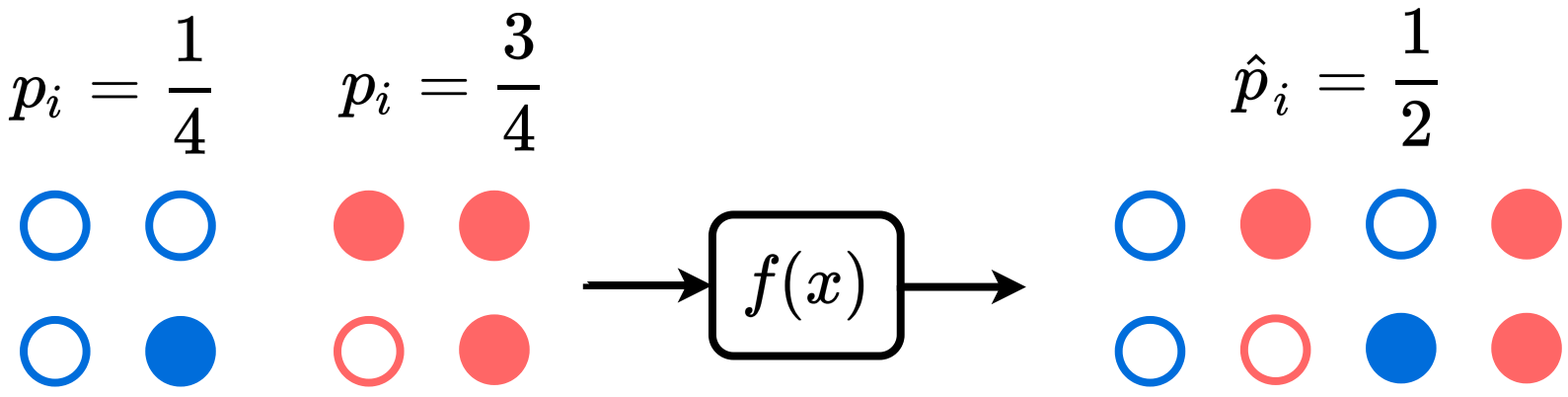
Example



Example



Example



Good calibration does not imply good probability estimation

Synthetic Dataset: Face-based Risk Prediction

x_i

Synthetic Dataset: Face-based Risk Prediction

x_i	Age z_i
	20

Synthetic Dataset: Face-based Risk Prediction

x_i	Age z_i	$p_i = \frac{z_i}{100}$
	20	0.2

Synthetic Dataset: Face-based Risk Prediction

x_i	Age z_i	$p_i = \frac{z_i}{100}$	y_i
	20	0.2	0

Synthetic Dataset: Face-based Risk Prediction

x_i	Age z_i	$p_i = \frac{z_i}{100}$	y_i
	20	0.2	0
			

Synthetic Dataset: Face-based Risk Prediction

x_i	Age z_i	$p_i = \frac{z_i}{100}$	y_i
	20	0.2	0
	70		

Synthetic Dataset: Face-based Risk Prediction

x_i	Age z_i	$p_i = \frac{z_i}{100}$	y_i
	20	0.2	0
	70	0.7	

Synthetic Dataset: Face-based Risk Prediction

x_i	Age z_i	$p_i = \frac{z_i}{100}$	y_i
	20	0.2	0
	70	0.7	1

Synthetic Dataset: Face-based Risk Prediction

x_i	Age z_i	$p_i = \frac{z_i}{100}$	y_i
	20	0.2	0
	70	0.7	1

Face-Age dataset from *Age Progression/Regression by Conditional Adversarial Autoencoder*. Zhang, Z., Song, Y., and Qi, H., CVPR 2017

Cross-entropy

Standard cross-entropy loss is a *proper scoring rule*

Cross-entropy

Standard cross-entropy loss is a *proper scoring rule*

Probabilities estimated by minimizing cross entropy are well calibrated in an infinite data regime

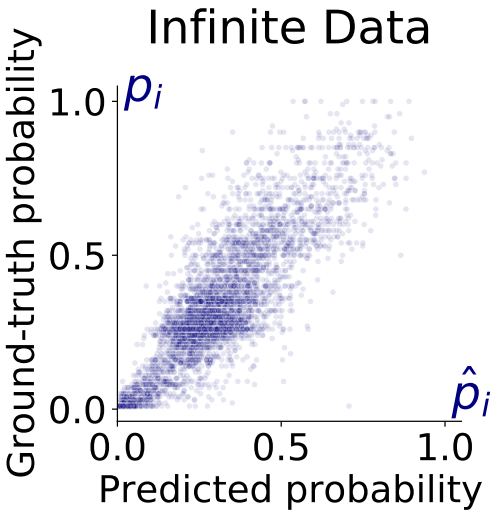
Cross-entropy

Standard cross-entropy loss is a *proper scoring rule*

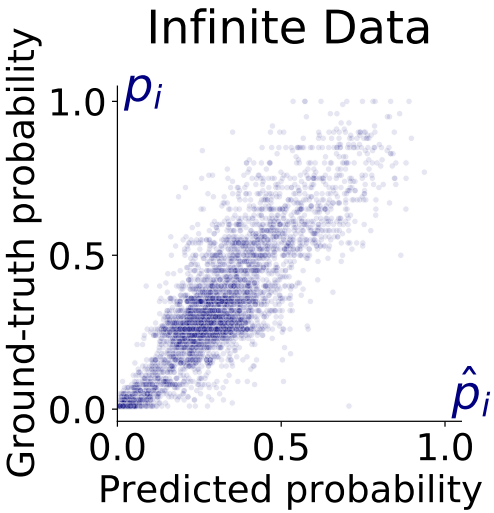
Probabilities estimated by minimizing cross entropy are well calibrated in an infinite data regime

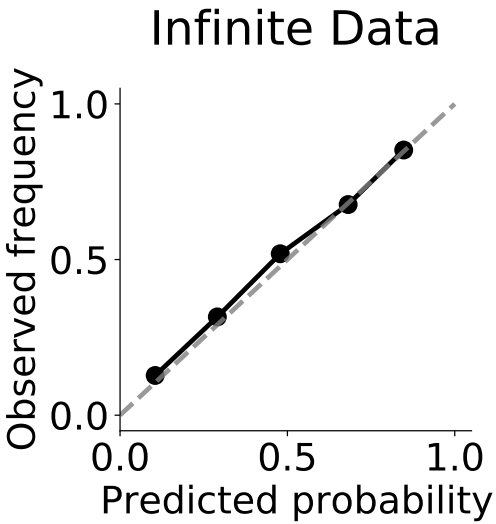
To simulate this, we resample y_i at each epoch based on p_i

Face-based Risk Prediction



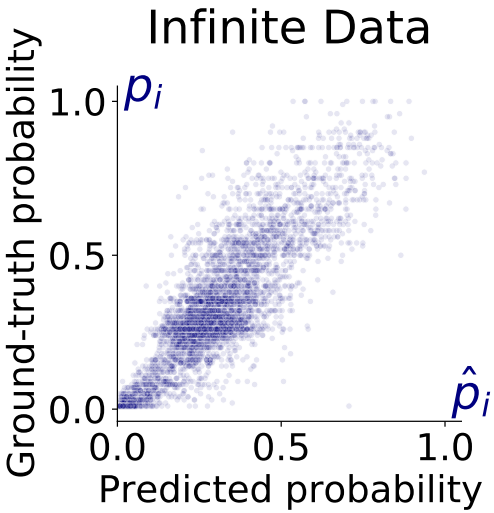
Face-based Risk Prediction

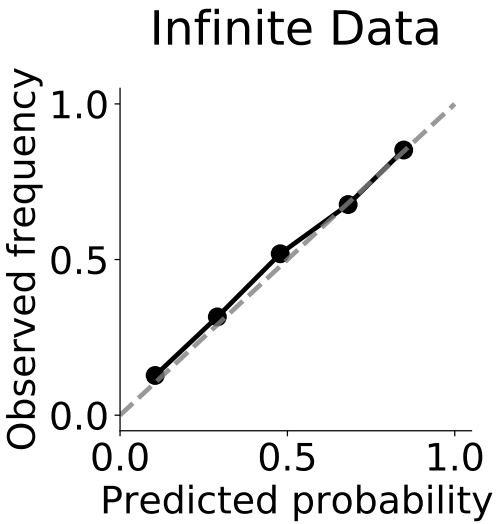




Face-based Risk Prediction

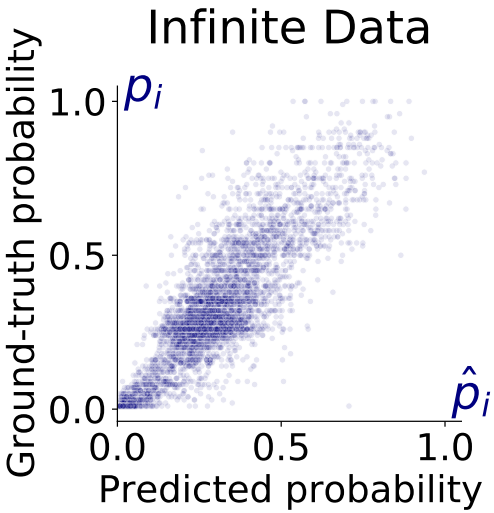
What happens if dataset is finite?

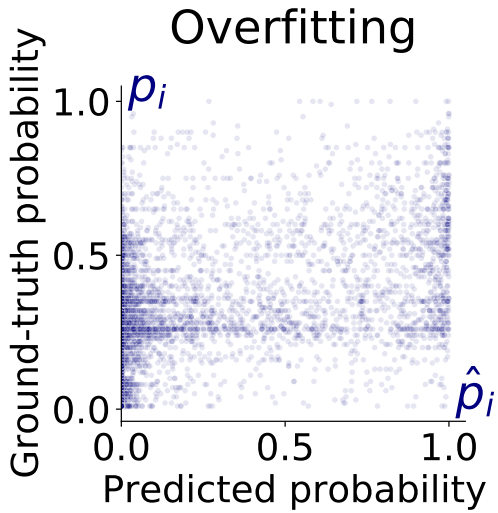


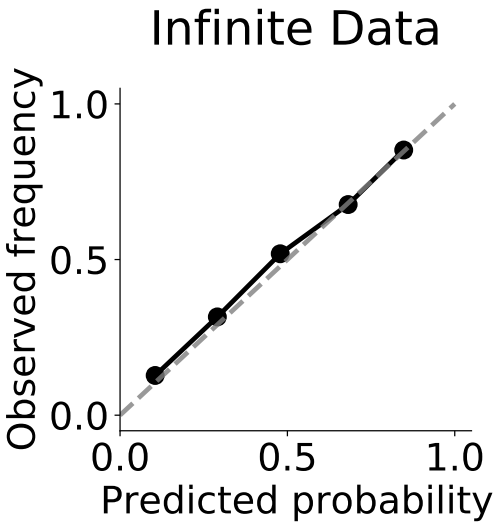


Face-based Risk Prediction

What happens if dataset is finite?

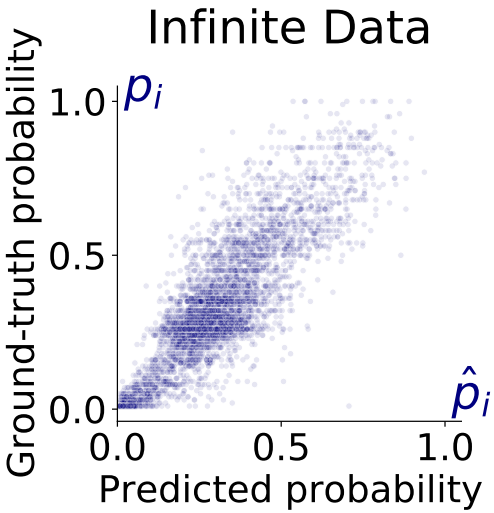


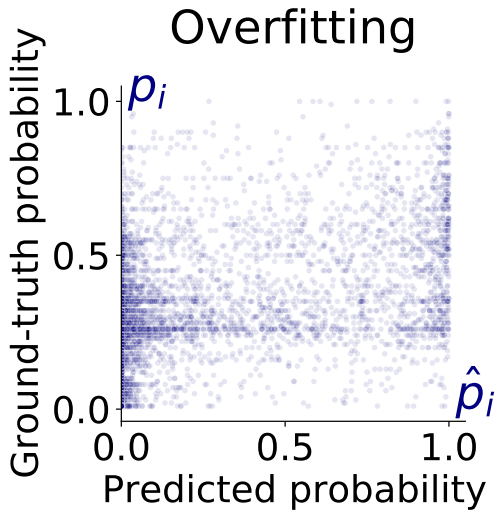


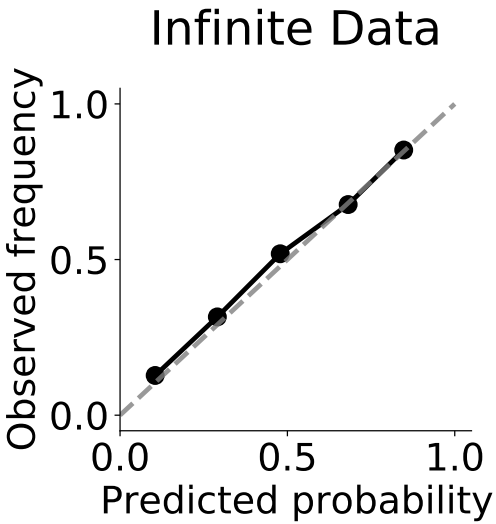


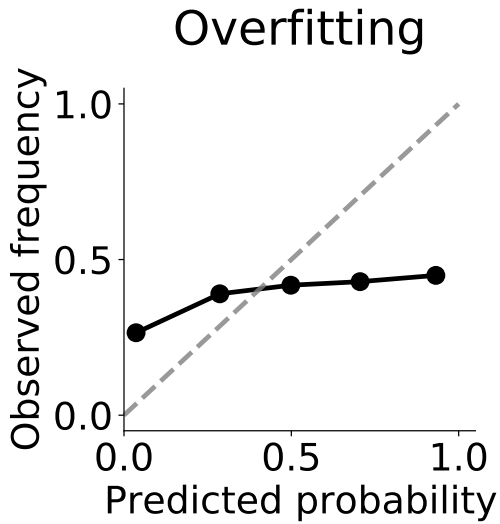
Face-based Risk Prediction

What happens if dataset is finite?



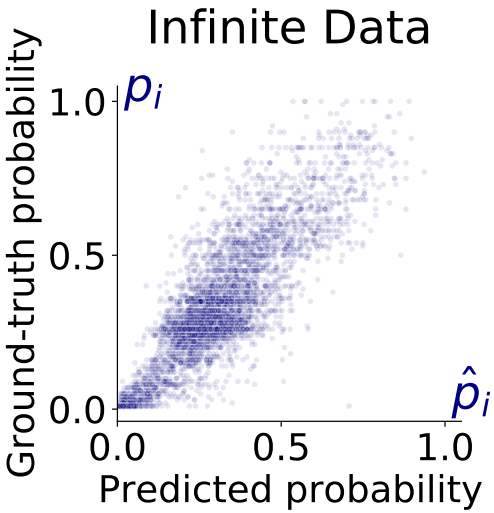


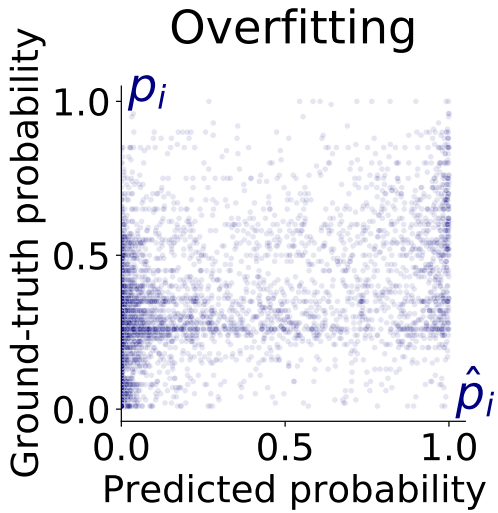


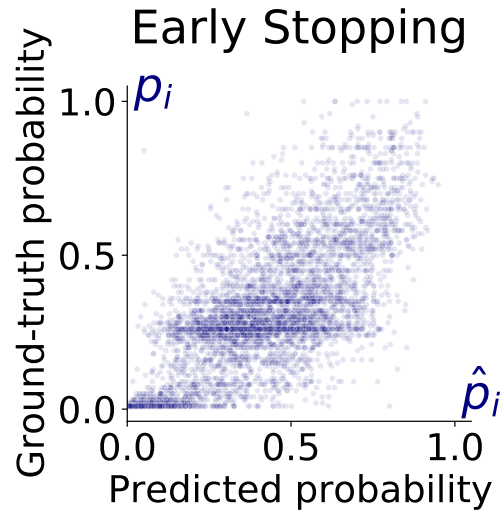


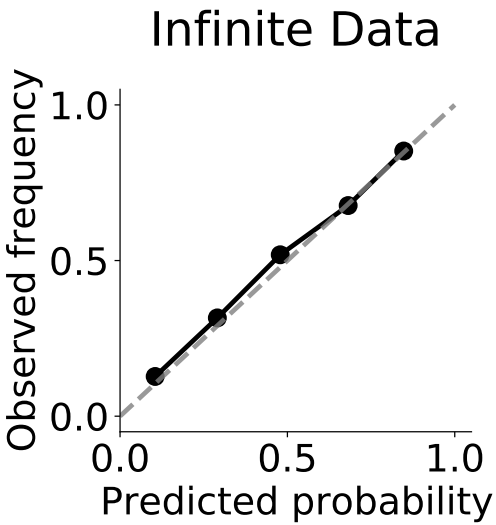
Face-based Risk Prediction

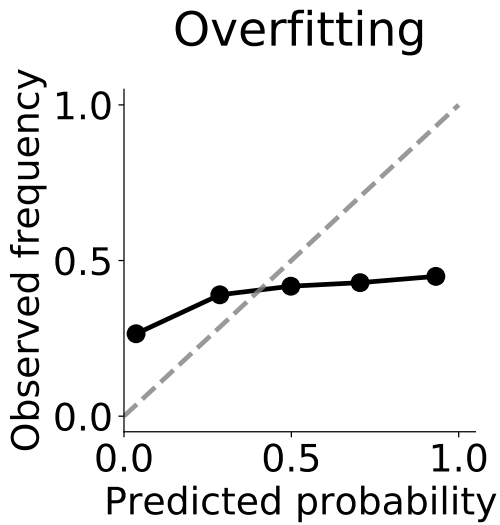
What happens if dataset is finite?





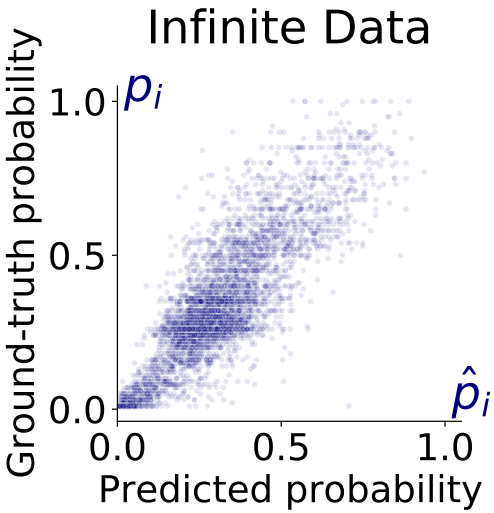


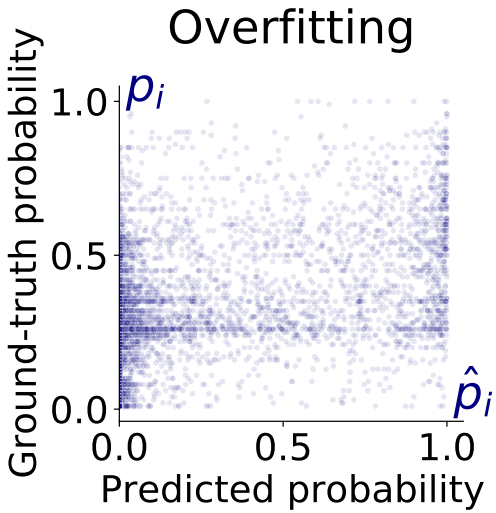


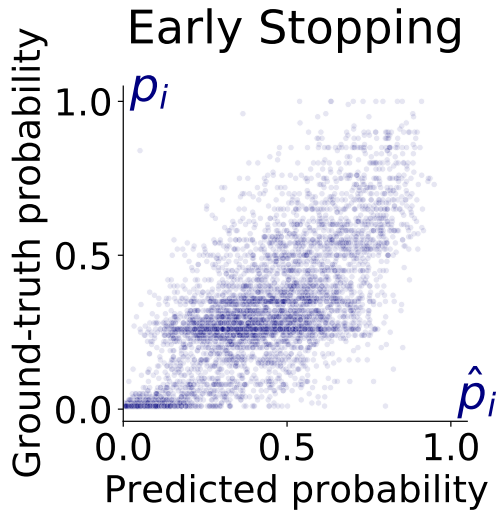


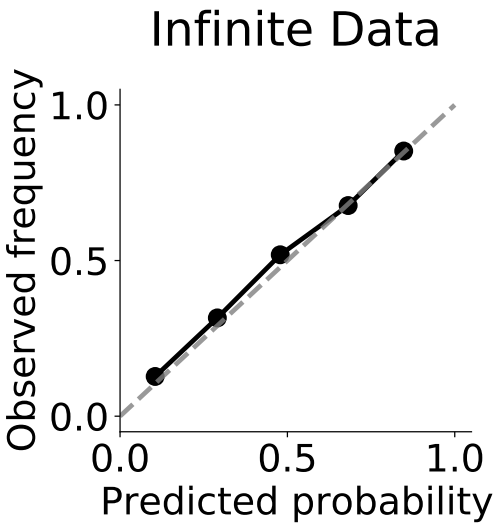
Face-based Risk Prediction

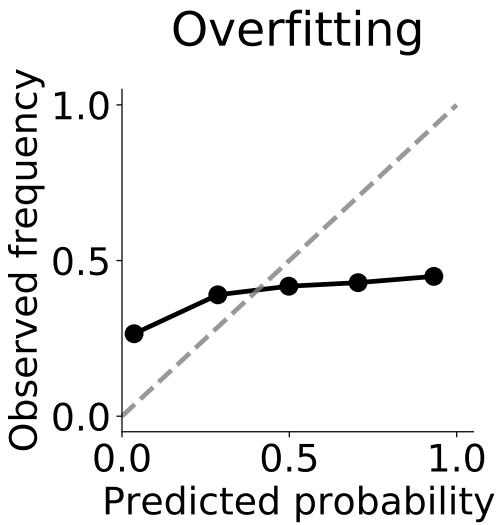
What happens if dataset is finite?

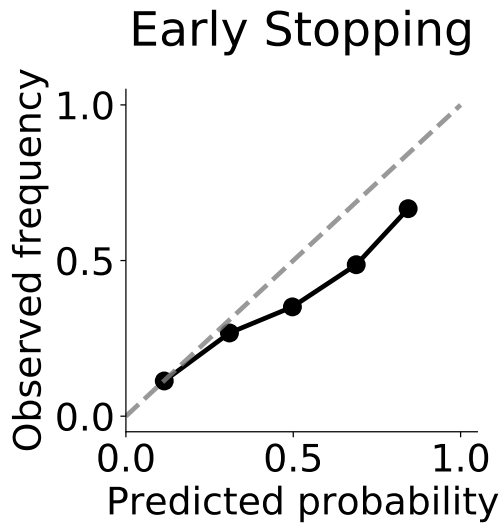












Fundamental phenomenon

For finite datasets, models memorize labels

Fundamental phenomenon

For finite datasets, models memorize labels

Understanding deep learning requires rethinking generalization. Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. ICLR 2017

Fundamental phenomenon

For finite datasets, models memorize labels

Understanding deep learning requires rethinking generalization. Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. ICLR 2017

During early learning stage probabilities are well calibrated

Fundamental phenomenon

For finite datasets, models memorize labels

Understanding deep learning requires rethinking generalization. Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. ICLR 2017

During early learning stage probabilities are well calibrated

Early-learning regularization prevents memorization of noisy labels. Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. NeurIPS 2020

Fundamental phenomenon

For finite datasets, models memorize labels

Understanding deep learning requires rethinking generalization. Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. ICLR 2017

During early learning stage probabilities are well calibrated

Early-learning regularization prevents memorization of noisy labels. Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. NeurIPS 2020

This occurs even for linear models!

Calibrated Probability Estimation (CaPE)

Goal: Improve model *while remaining calibrated*

Calibrated Probability Estimation (CaPE)

Goal: Improve model *while remaining calibrated*

1. Minimize cross-entropy loss until memorization begins

Calibrated Probability Estimation (CaPE)

Goal: Improve model *while remaining calibrated*

1. Minimize cross-entropy loss until memorization begins

2. Alternate between:

Calibrated Probability Estimation (CaPE)

Goal: Improve model *while remaining calibrated*

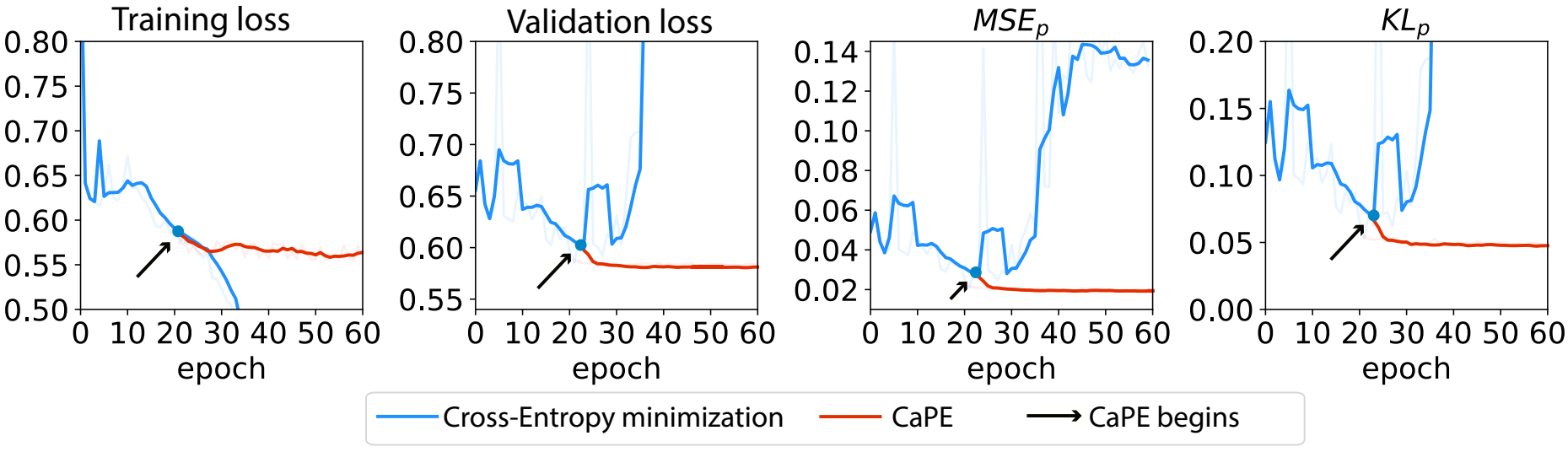
1. Minimize cross-entropy loss until memorization begins
2. Alternate between:
   - Enforcing calibration

Calibrated Probability Estimation (CaPE)

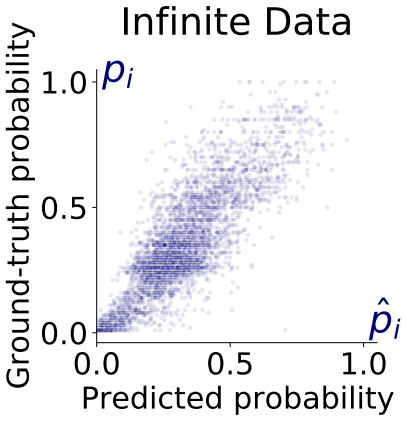
Goal: Improve model *while remaining calibrated*

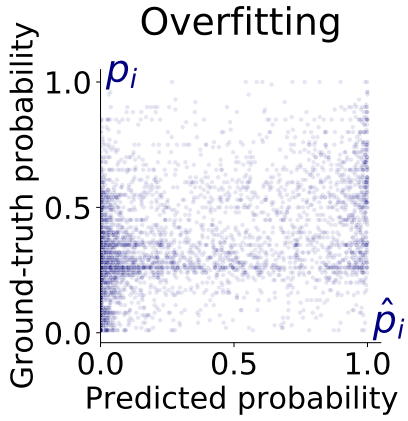
1. Minimize cross-entropy loss until memorization begins
2. Alternate between:
   - Enforcing calibration
   - Minimizing cross entropy

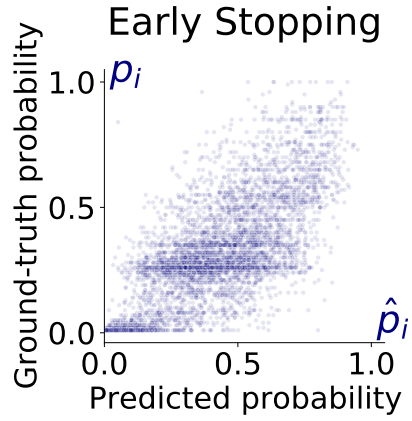
Calibrated Probability Estimation

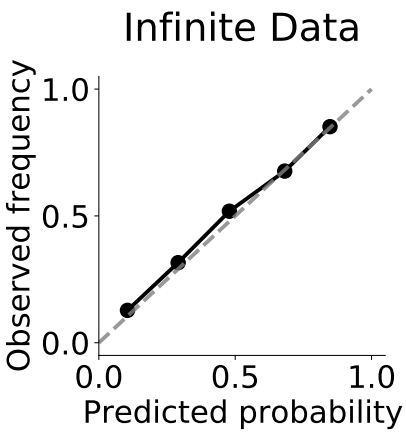


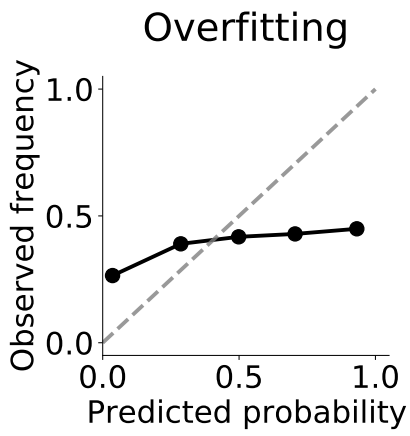
Face-based Risk Prediction

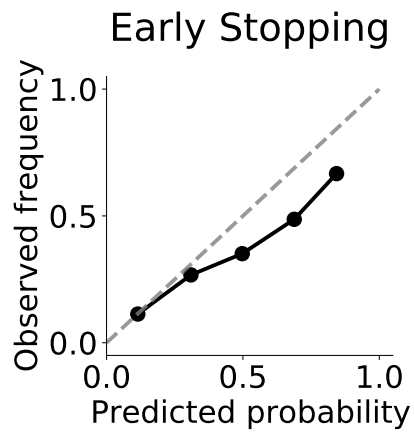




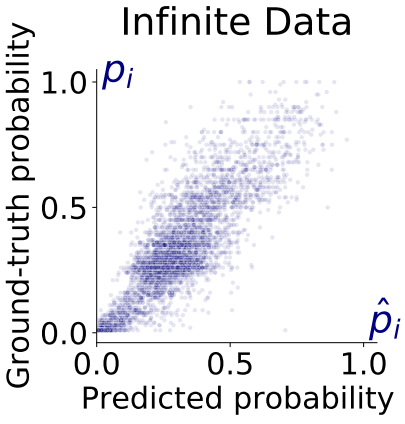


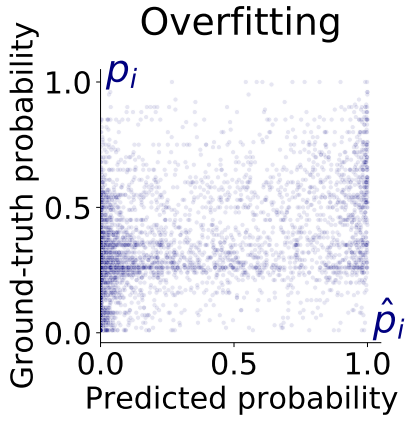


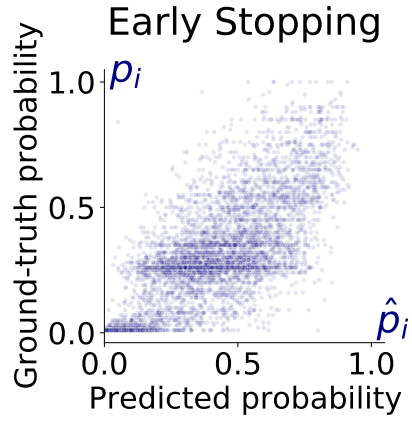


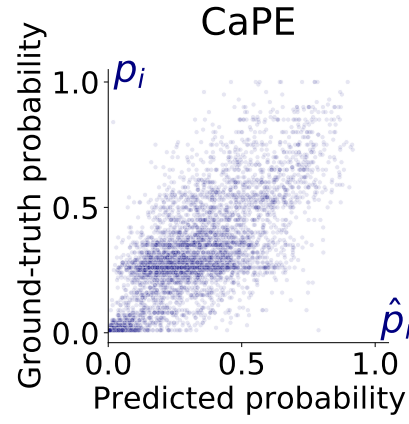


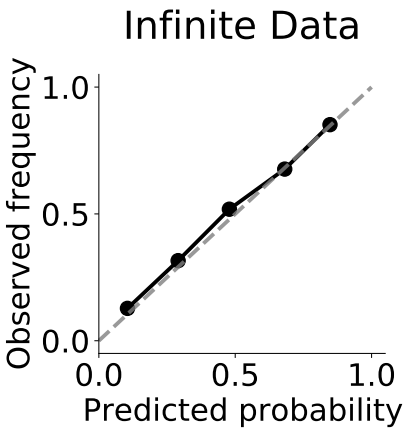
Face-based Risk Prediction

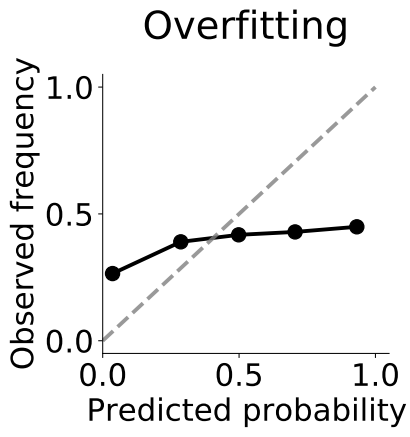


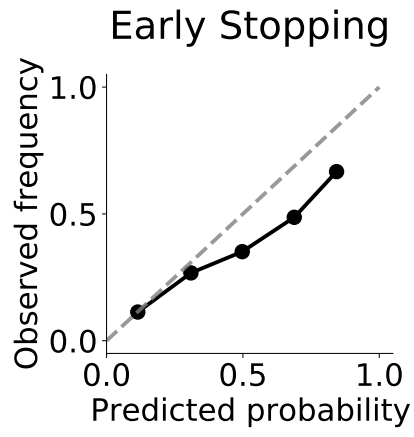




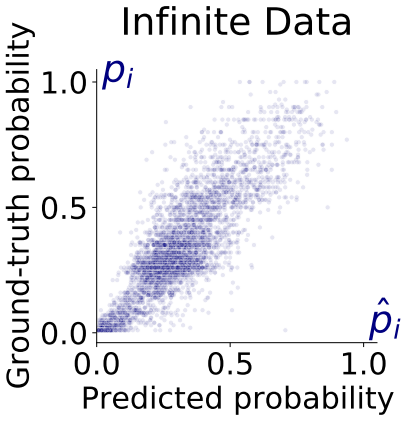


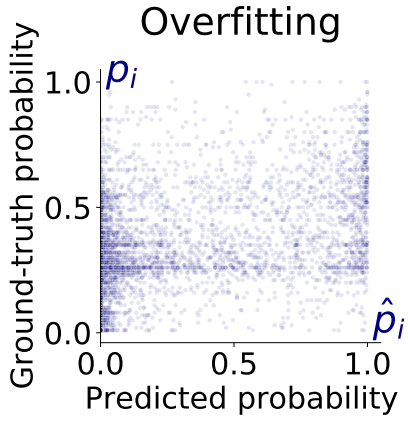


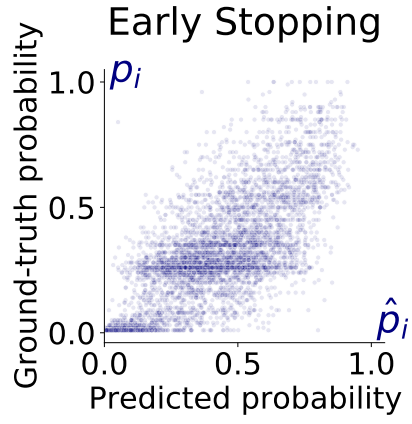


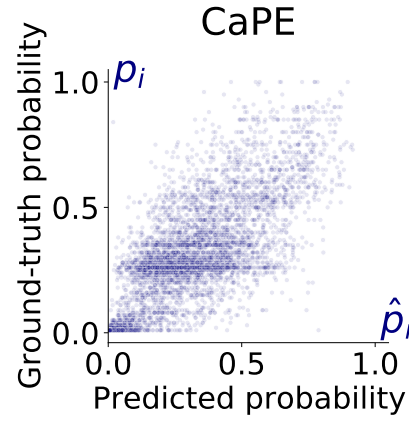


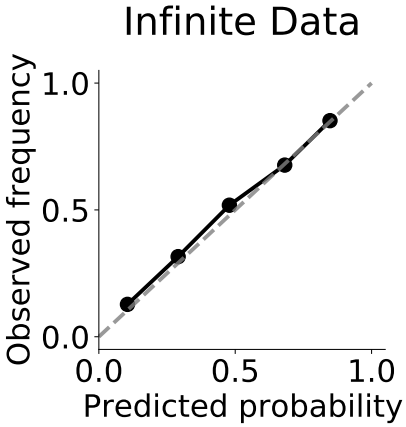
Face-based Risk Prediction

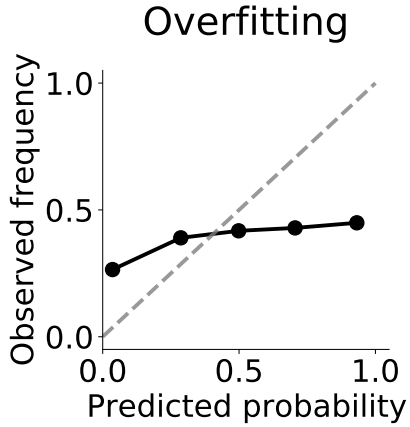


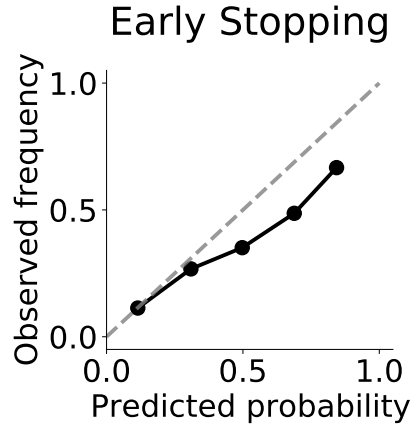


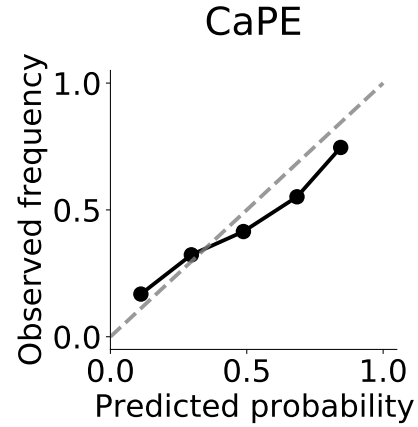












Existing approaches

Mostly focus on calibration in classification (i.e. no inherent uncertainty)

Existing approaches

Mostly focus on calibration in classification (i.e. no inherent uncertainty)

They can still be applied to probability estimation!

Existing approaches

Mostly focus on calibration in classification (i.e. no inherent uncertainty)

They can still be applied to probability estimation!

- Post-processing: Temperature scaling (Guo et al. 2017), Platt scaling (Platt 1999), and Dirichlet calibration (Kull et al. 2019)

Existing approaches

Mostly focus on calibration in classification (i.e. no inherent uncertainty)

They can still be applied to probability estimation!

- Post-processing: Temperature scaling (Guo et al. 2017), Platt scaling (Platt 1999), and Dirichlet calibration (Kull et al. 2019)
- Bayesian / ensembling: Mix-n-Match (Zhang et al. 2020), Deep Ensemble (Lakshminarayanan et al. 2017)

Existing approaches

Mostly focus on calibration in classification (i.e. no inherent uncertainty)

They can still be applied to probability estimation!

- Post-processing: Temperature scaling (Guo et al. 2017), Platt scaling (Platt 1999), and Dirichlet calibration (Kull et al. 2019)
- Bayesian / ensembling: Mix-n-Match (Zhang et al. 2020), Deep Ensemble (Lakshminarayanan et al. 2017)
- Modified cost function: Focal loss (Mukhoti et al. 2020), entropy-maximizing loss (Pereyra et al. 2017), and kernel mean embeddings (Kumar et al., 2018)

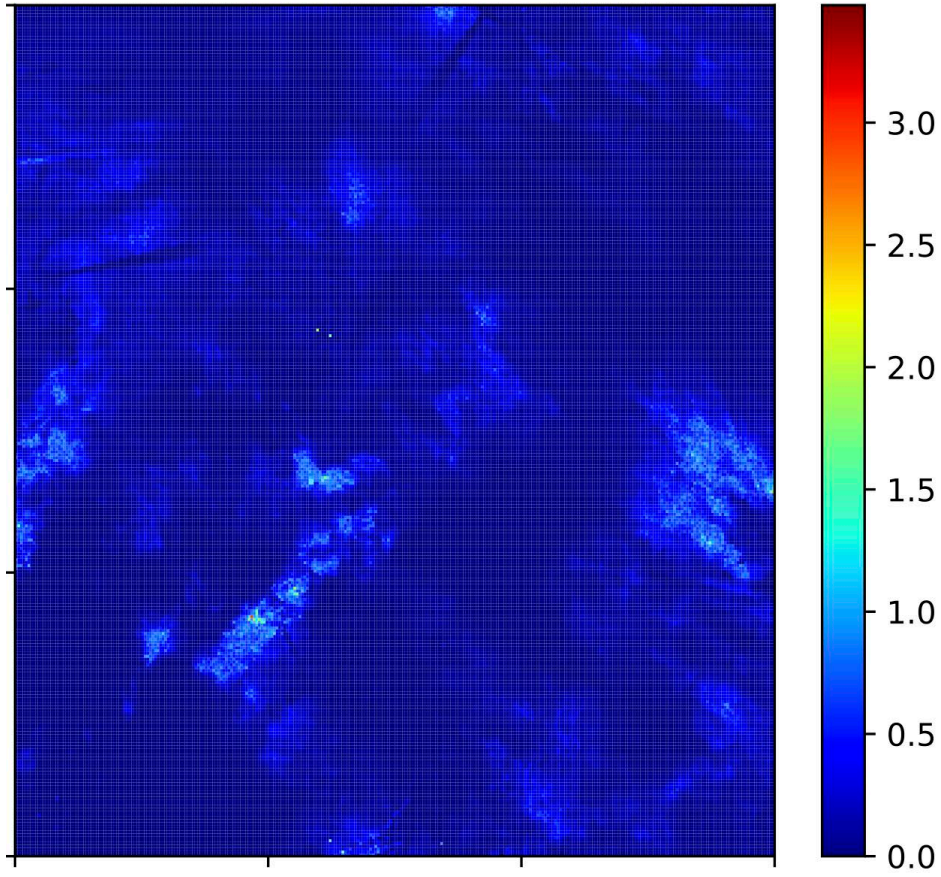
Results

Method ($\times 10^{-2}$)	Cancer Survival AUC	Cancer Survival ECE	Cancer Survival Brier	Weather forecasting AUC	Weather forecasting ECE	Weather forecasting Brier	Collision Prediction AUC	Collision Prediction ECE	Collision Prediction Brier
CE early-stop	58.88	12.25	23.96	77.64	10.91	20.57	85.68	4.36	8.59
Temperature	58.88	12.07	23.73	77.64	8.66	20.21	85.68	4.56	8.51
Platt Scaling	58.91	10.28	23.33	77.65	6.97	19.53	85.76	3.04	8.23
Dirichlet Cal.	49.89	13.83	24.08	77.51	14.29	21.89	83.36	5.78	8.78
Mix-n-match	58.88	12.16	23.67	77.64	8.65	20.21	85.68	4.40	8.52
Focal Loss	55.02	12.15	23.31	76.18	8.32	20.27	82.21	9.07	9.82
Entropy Reg.	56.29	11.73	23.62	79.01	10.53	19.77	83.15	14.54	11.10
MMCE Reg.	48.45	11.84	23.73	76.69	8.46	20.12	85.18	**2.94**	8.48
Deep Ens.	52.46	9.99	23.47	**79.86**	7.41	18.82	85.27	3.15	8.55
CaPE (bin)	**61.44**	12.31	23.20	78.99	5.16	**18.37**	85.70	3.16	8.18
CaPE (kern.)	61.22	**9.48**	**23.18**	79.00	**5.08**	18.39	**85.95**	3.22	**8.13**

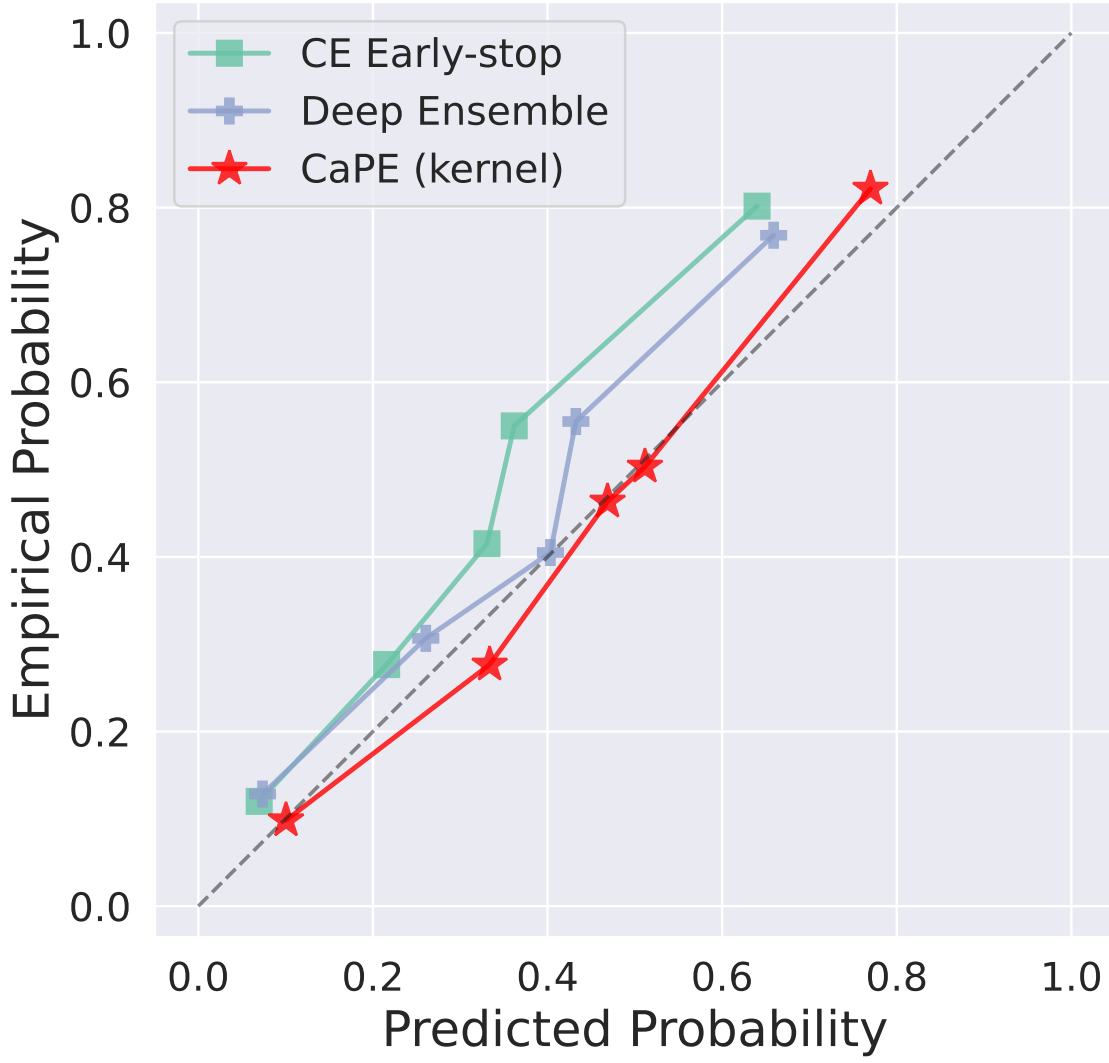
Weather Forecasting

Event of interest: Will it rain?

Input: Map of past precipitation



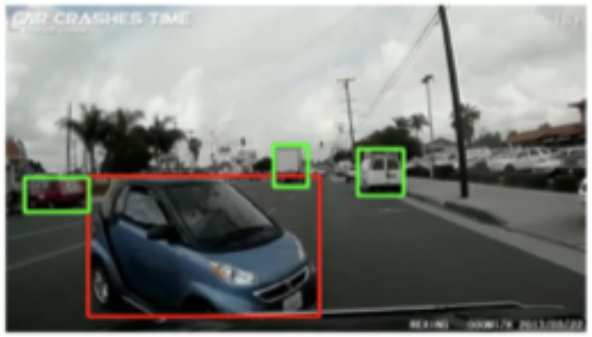
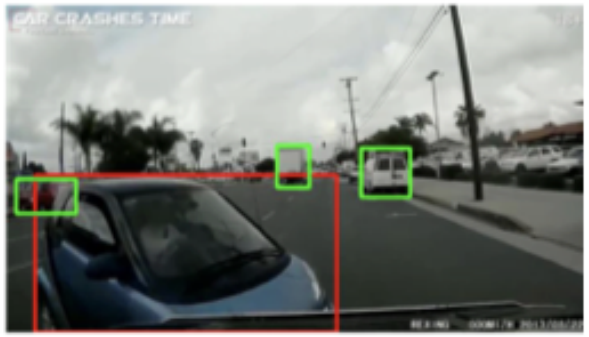
Weather Forecasting



Collision Prediction

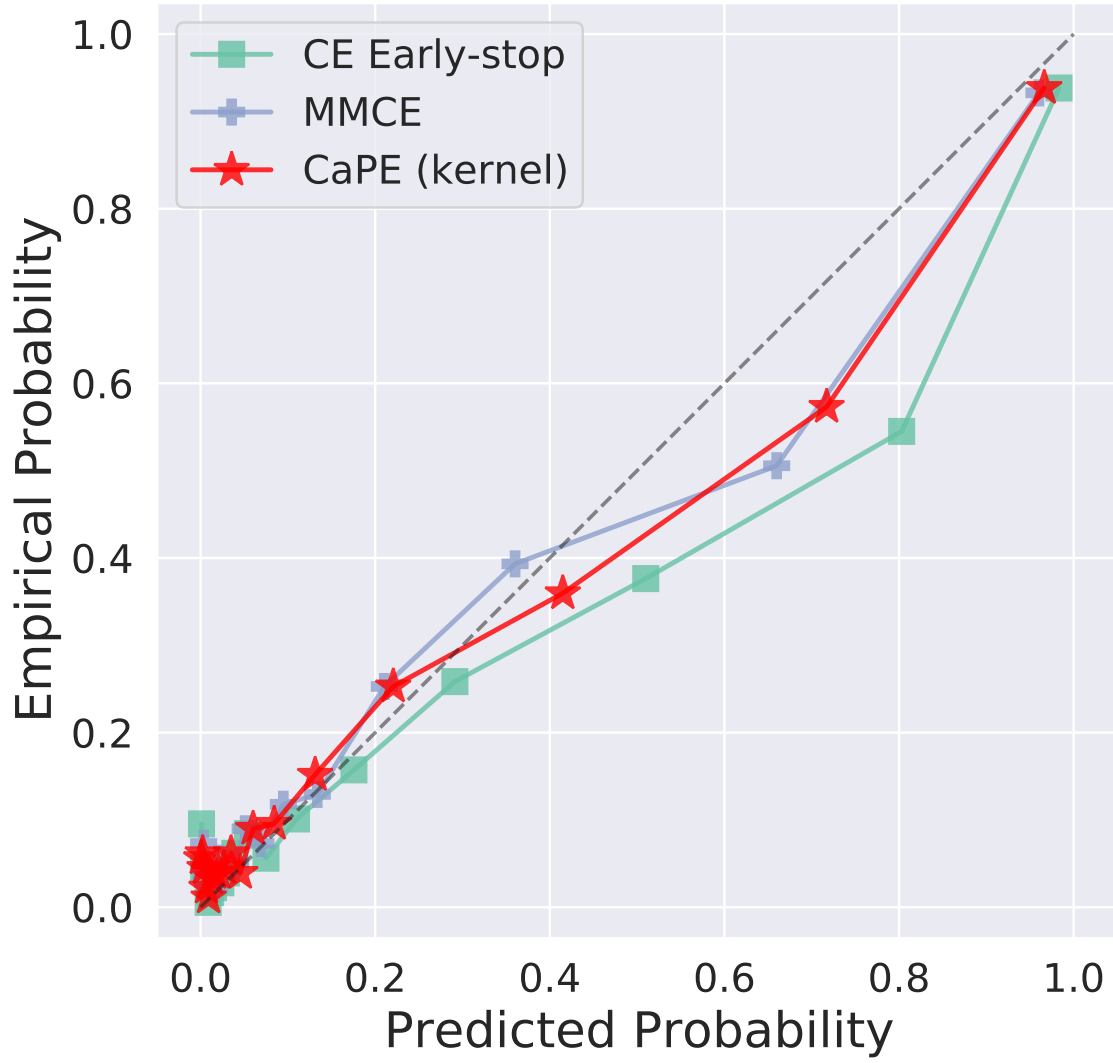
Event of interest: Will there be a collision?

Input: Dashboard video

Crash to not crash: Learn to identify dangerous vehicles using a simulator. Kim, H., Lee, K., Hwang, G., and Suh, C. AAAI 2019

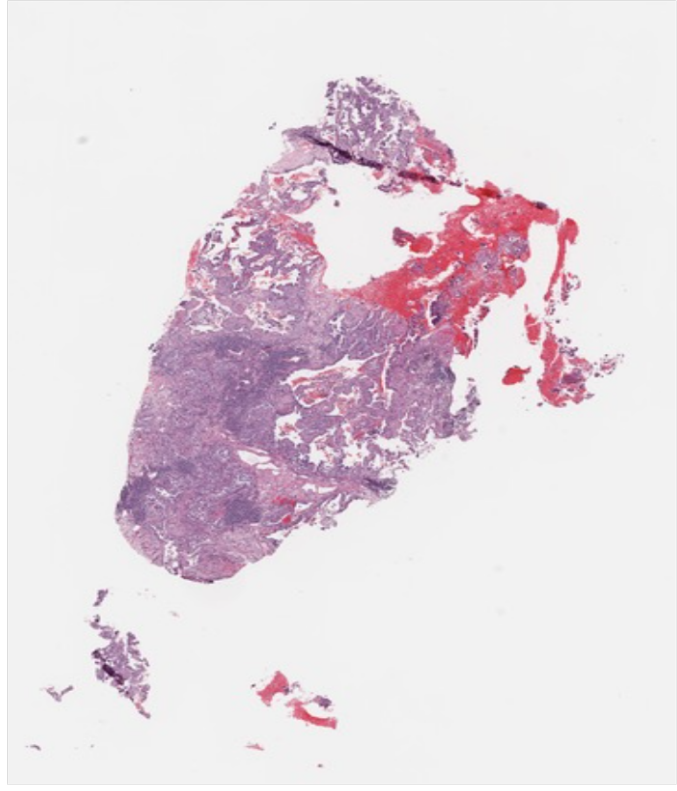
Collision Prediction



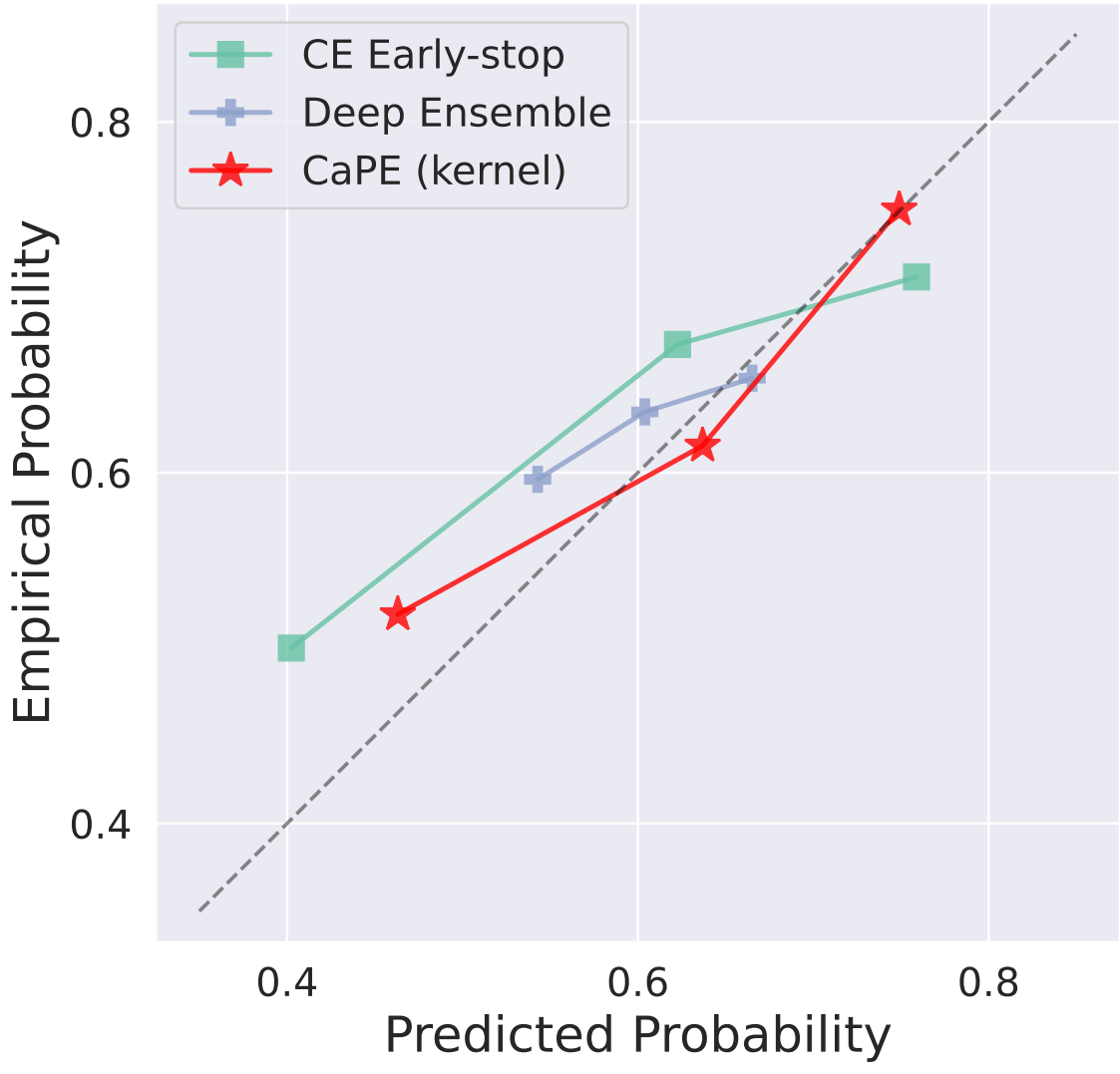
Survival Prediction

Event of interest: Will a cancer patient die in the next 5 years?

Input: Histopathology slide



Survival Forecasting



Conclusions

- Deep networks can be effective for probability estimation

Conclusions

- Deep networks can be effective for probability estimation
- Evaluation is tricky! It is important to think carefully about metrics

Conclusions

- Deep networks can be effective for probability estimation
- Evaluation is tricky! It is important to think carefully about metrics
- Exploiting early learning and enforcing calibration can improve probability estimates

Conclusions

- Deep networks can be effective for probability estimation
- Evaluation is tricky! It is important to think carefully about metrics
- Exploiting early learning and enforcing calibration can improve probability estimates
- More benchmark datasets are needed!

For more information

Deep Probability Estimation

S. Liu, A. Kaku, W. Zhu, M. Leibovich, S. Mohan, B. Yu, H. Huang, L. Zanna, N. Razavian, J. Niles-Weed, C. Fernandez-Granda