Statistical Learning in Complex Prediction Spaces: What Do We Know?

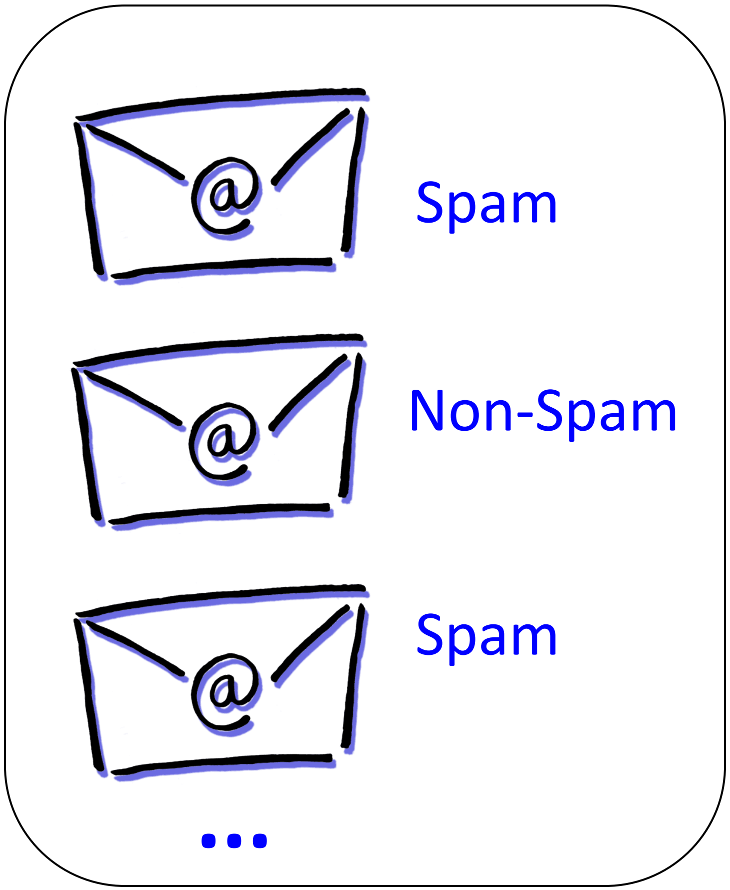
Shivani Agarwal

Department of Computer Science & Automation
Indian Institute of Science

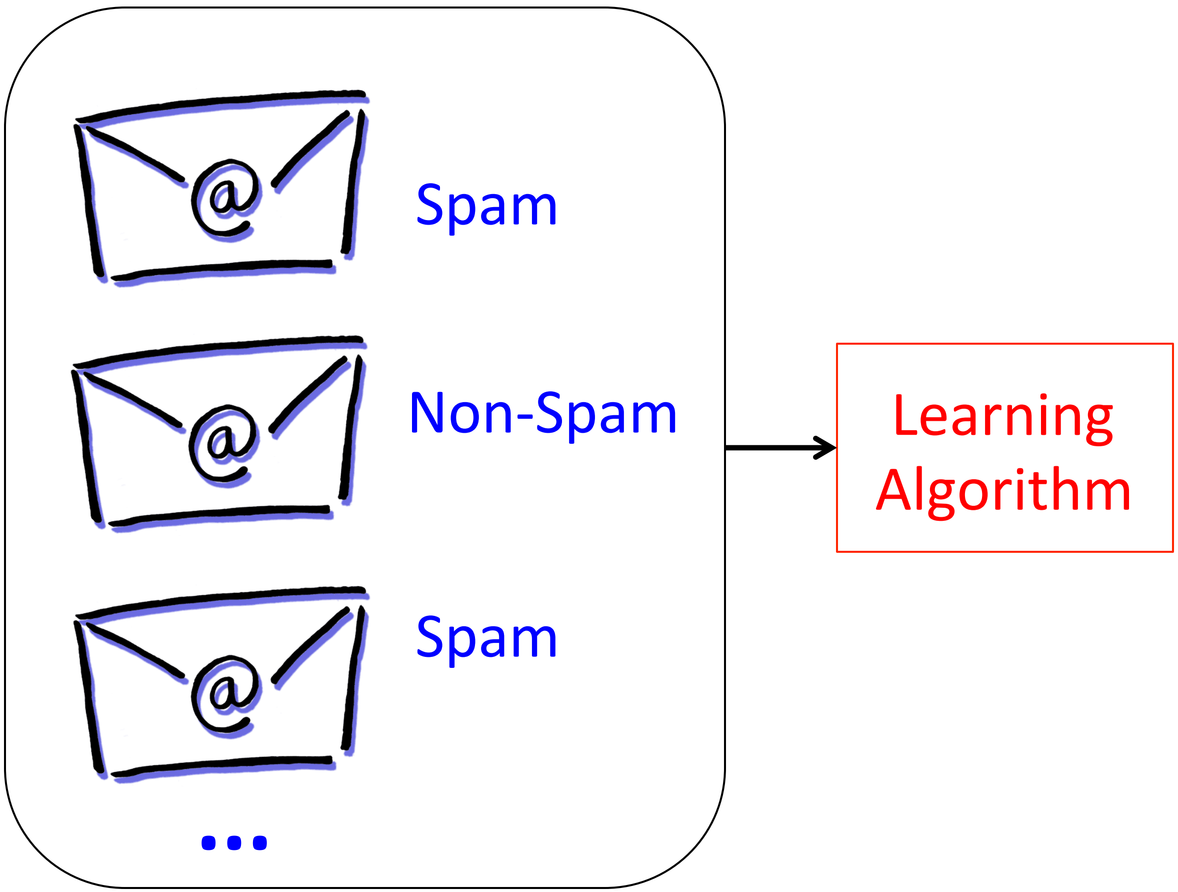
January 2015

‘Simple’ prediction problems

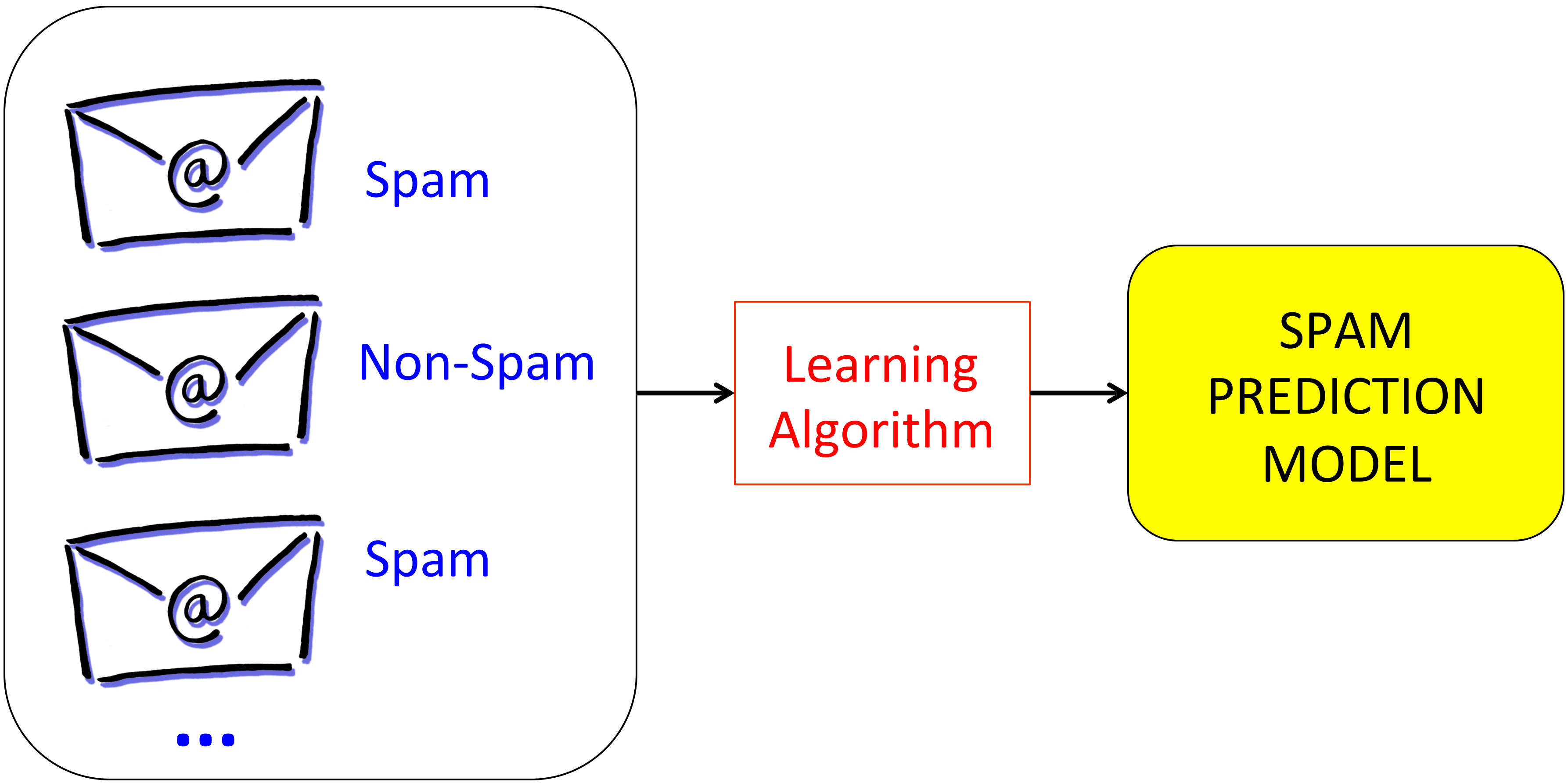
Example: Email Spam Filter



Example: Email Spam Filter



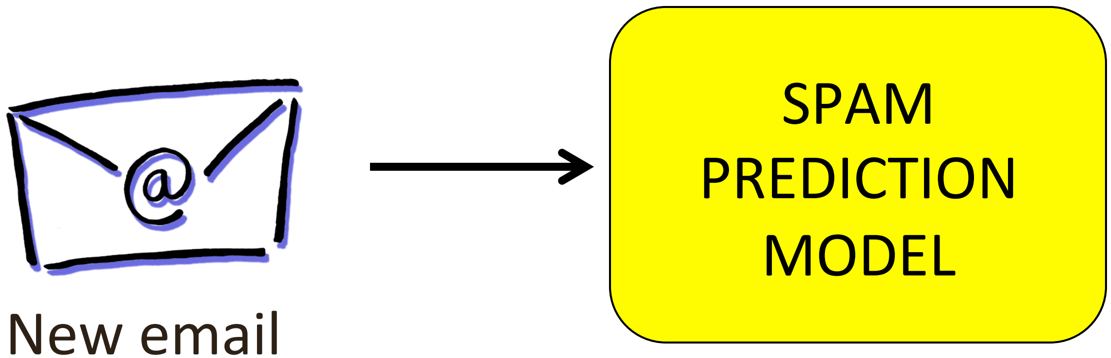
Example: Email Spam Filter



Example: Email Spam Filter

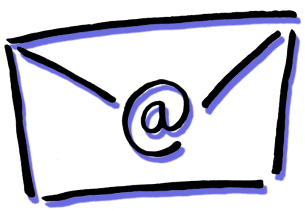
SPAM
PREDICTION
MODEL

Example: Email Spam Filter



New email

Example: Email Spam Filter



New email

SPAM PREDICTION MODEL

Prediction: Spam or Non-Spam?

Example: Tumor Classification

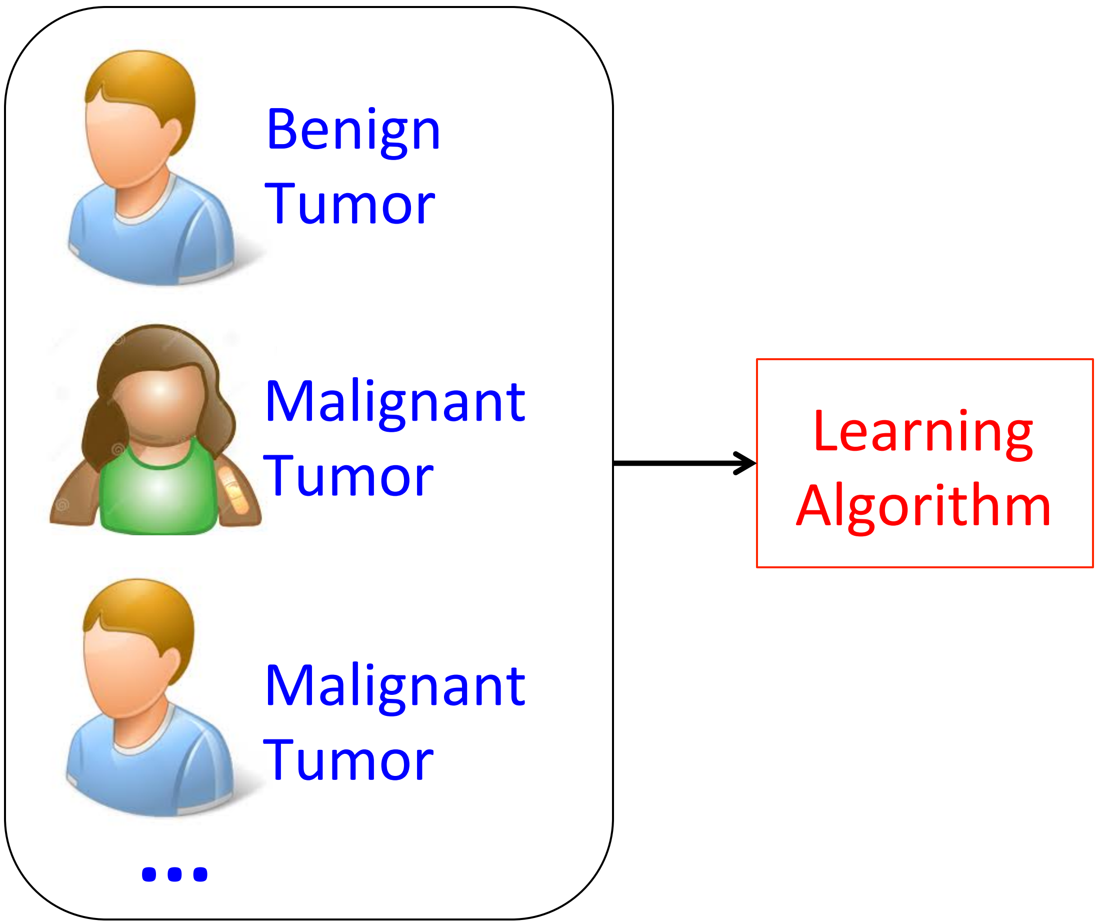
Benign Tumor

Malignant Tumor

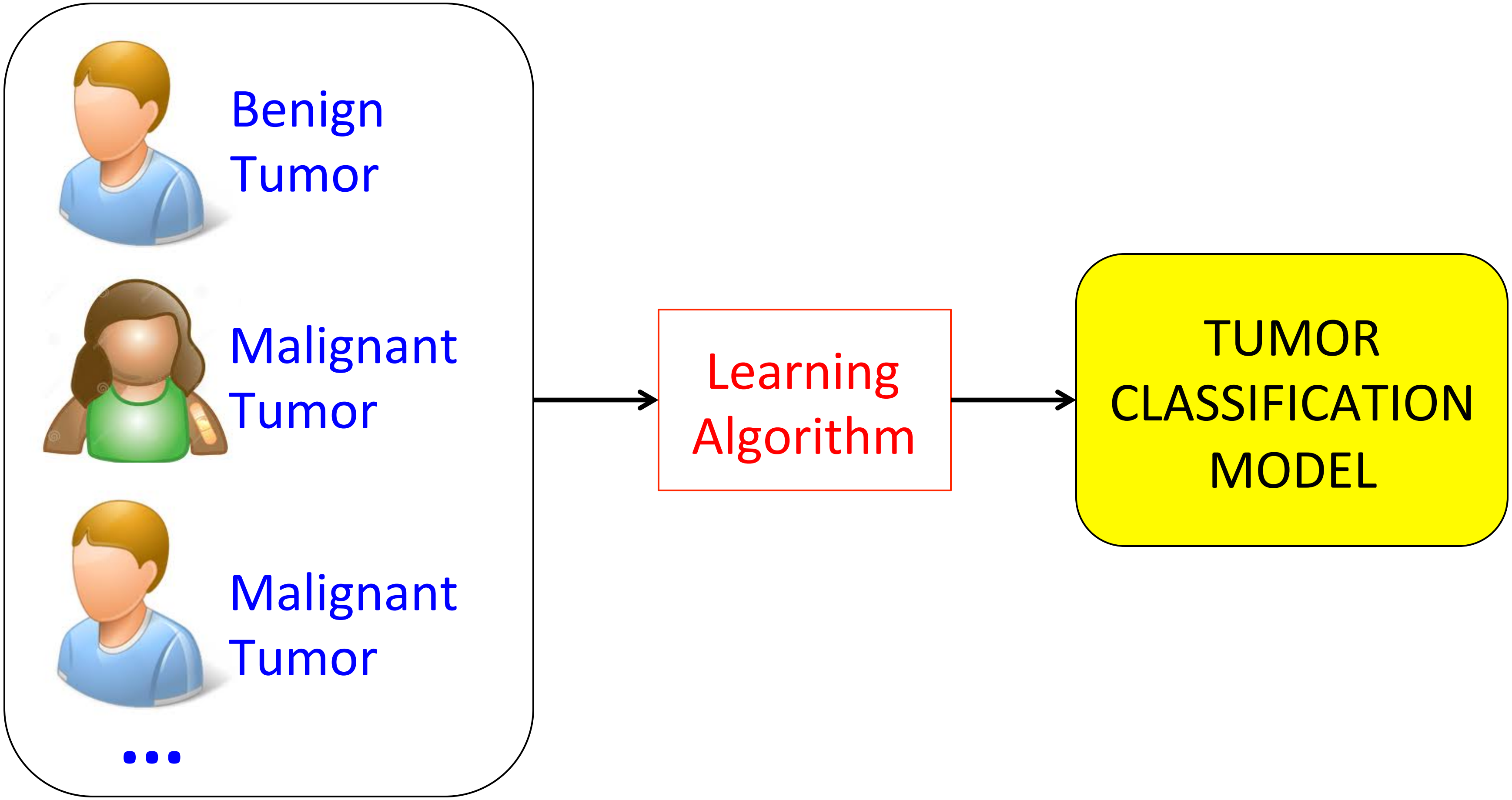
Malignant Tumor

...

Example: Tumor Classification



Example: Tumor Classification



Example: Tumor Classification

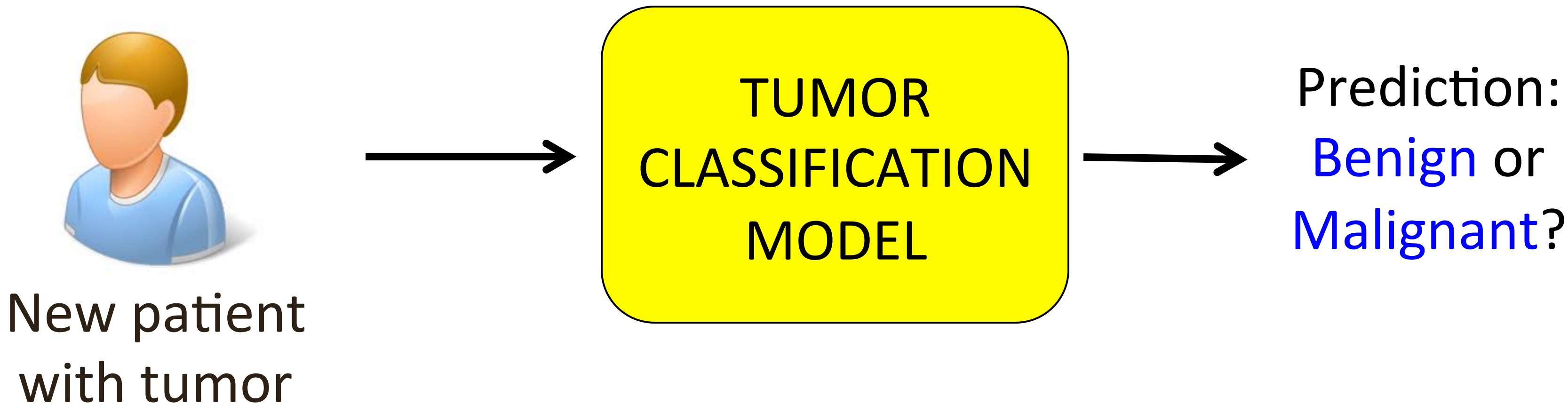
TUMOR
CLASSIFICATION
MODEL

Example: Tumor Classification

New patient
with tumor

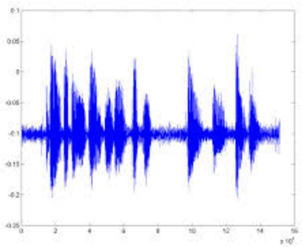
→ TUMOR CLASSIFICATION MODEL

Example: Tumor Classification

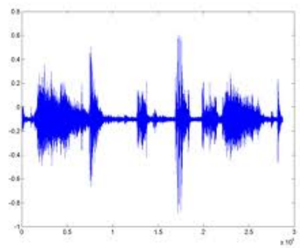


Complex prediction problems

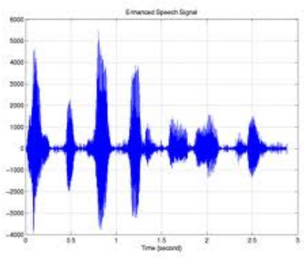
Example: Speech Recognition



"How are you?"



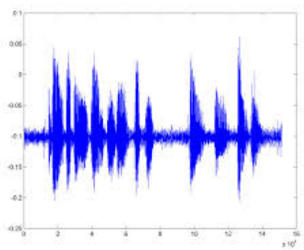
"I would like to order some pizza."



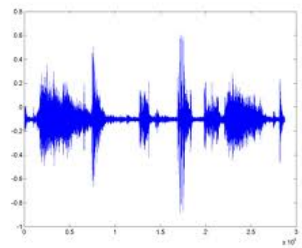
"Today is a gorgeous day for playing tennis."

...

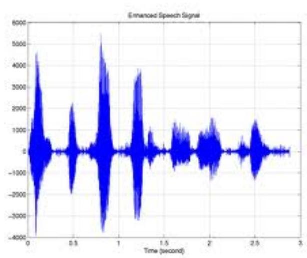
Example: Speech Recognition



"How are you?"



"I would like to order some pizza."

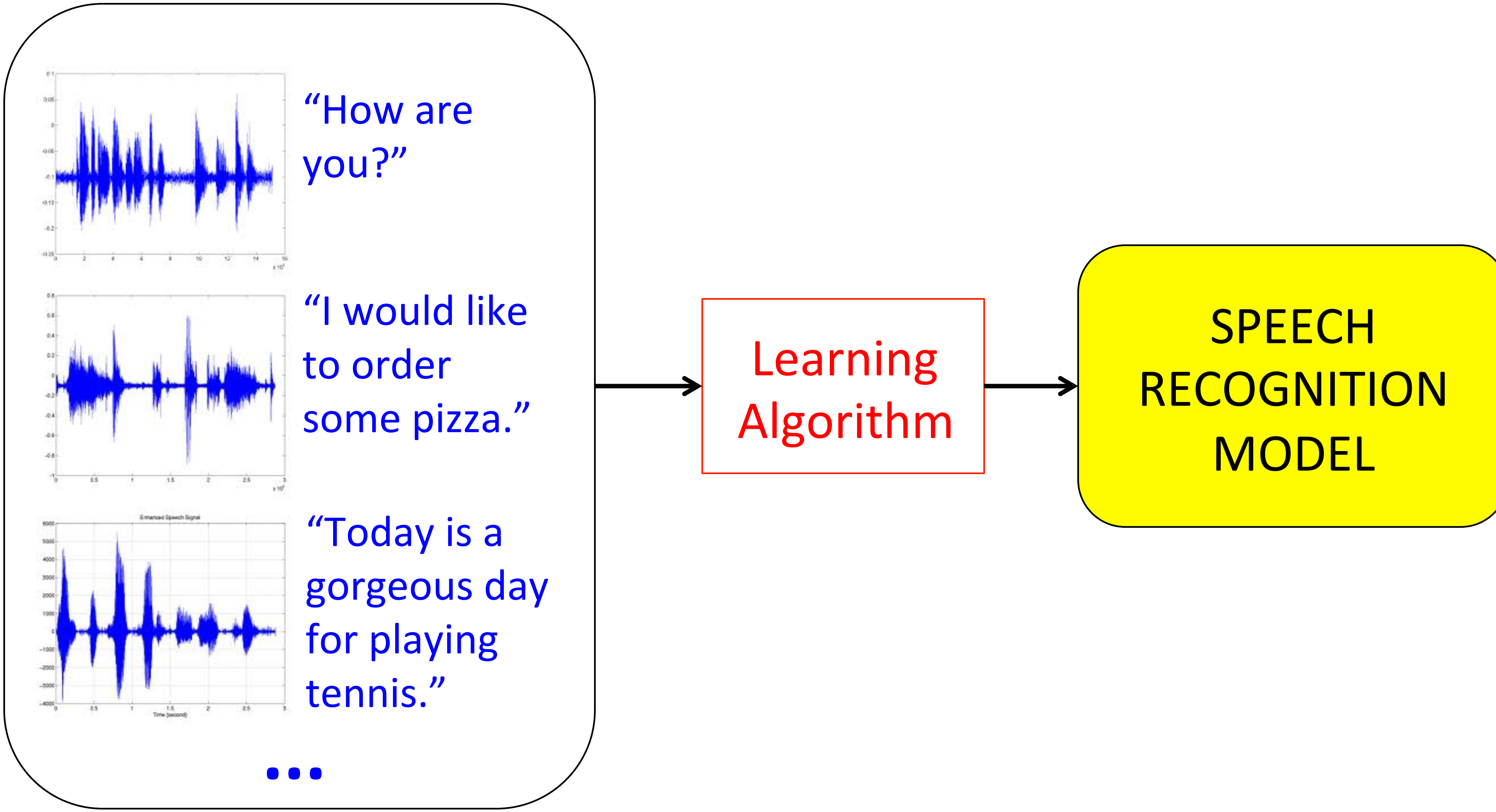


"Today is a gorgeous day for playing tennis."

...

Learning Algorithm

Example: Speech Recognition



Example: Speech Recognition

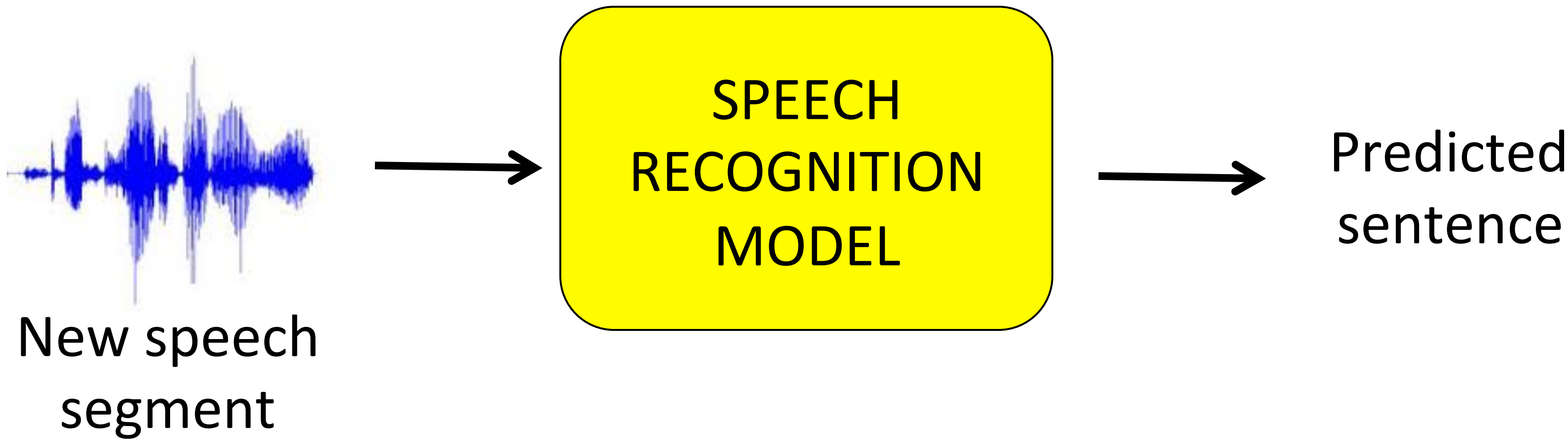
SPEECH
RECOGNITION
MODEL

Example: Speech Recognition

New speech segment

SPEECH RECOGNITION MODEL

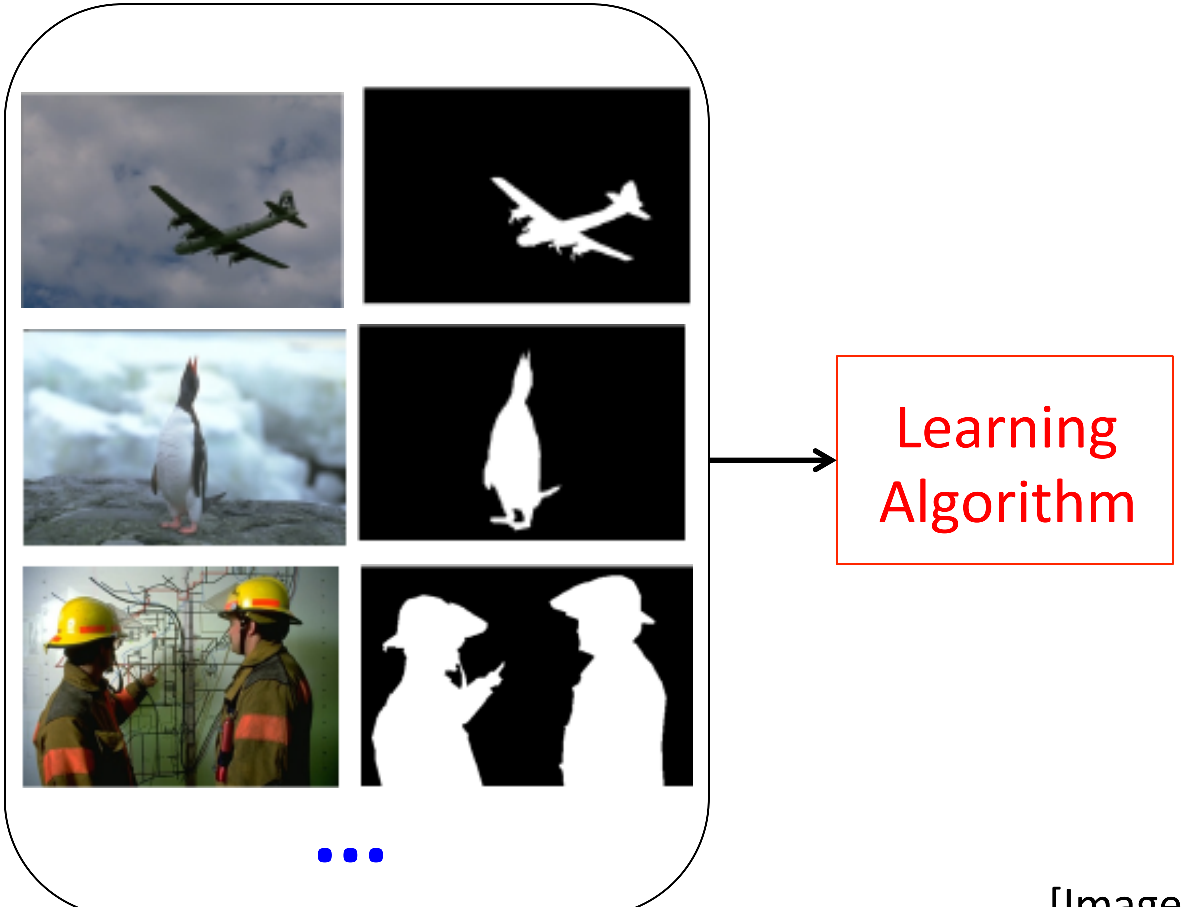
Example: Speech Recognition



Example: Image Segmentation

[Images taken from Li et al, ICME 2013]

Example: Image Segmentation



[Images taken from Li et al, ICME 2013]

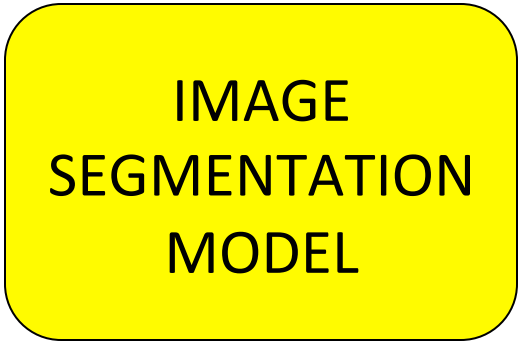
Example: Image Segmentation

Learning Algorithm

IMAGE SEGMENTATION MODEL

[Images taken from Li et al, ICME 2013]

Example: Image Segmentation



Example: Image Segmentation

New image

IMAGE
SEGMENTATION
MODEL

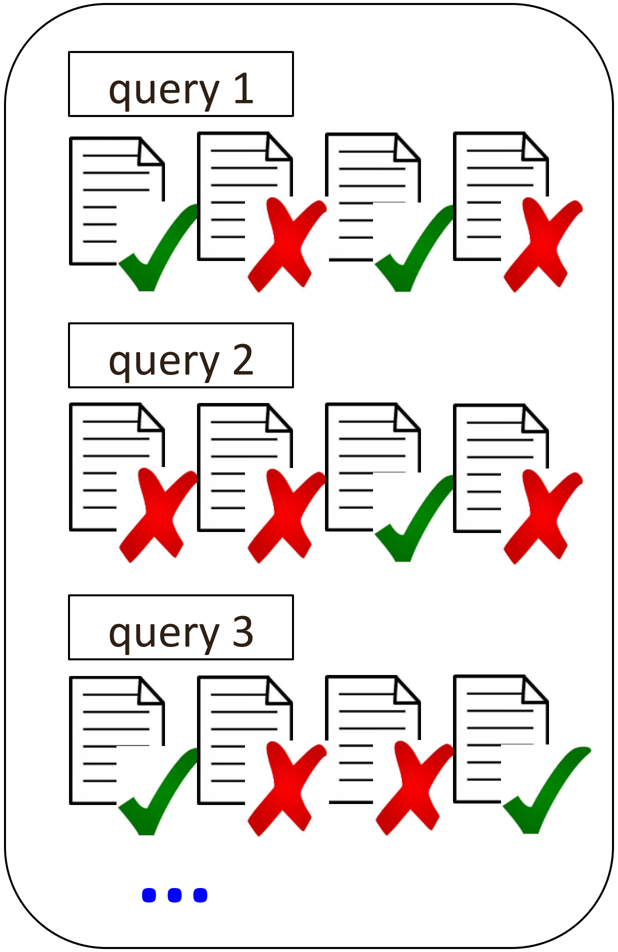
Example: Image Segmentation

New image

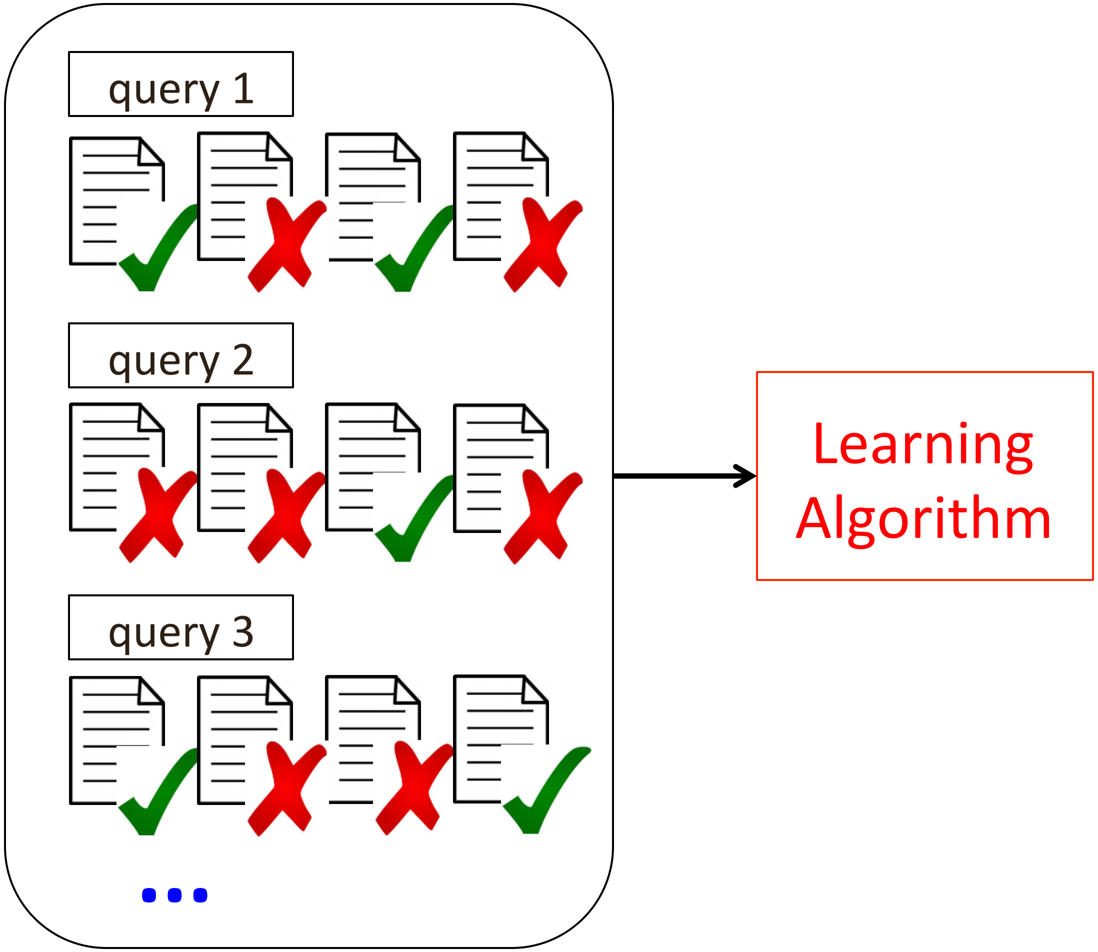
IMAGE
SEGMENTATION
MODEL

Predicted segmentation

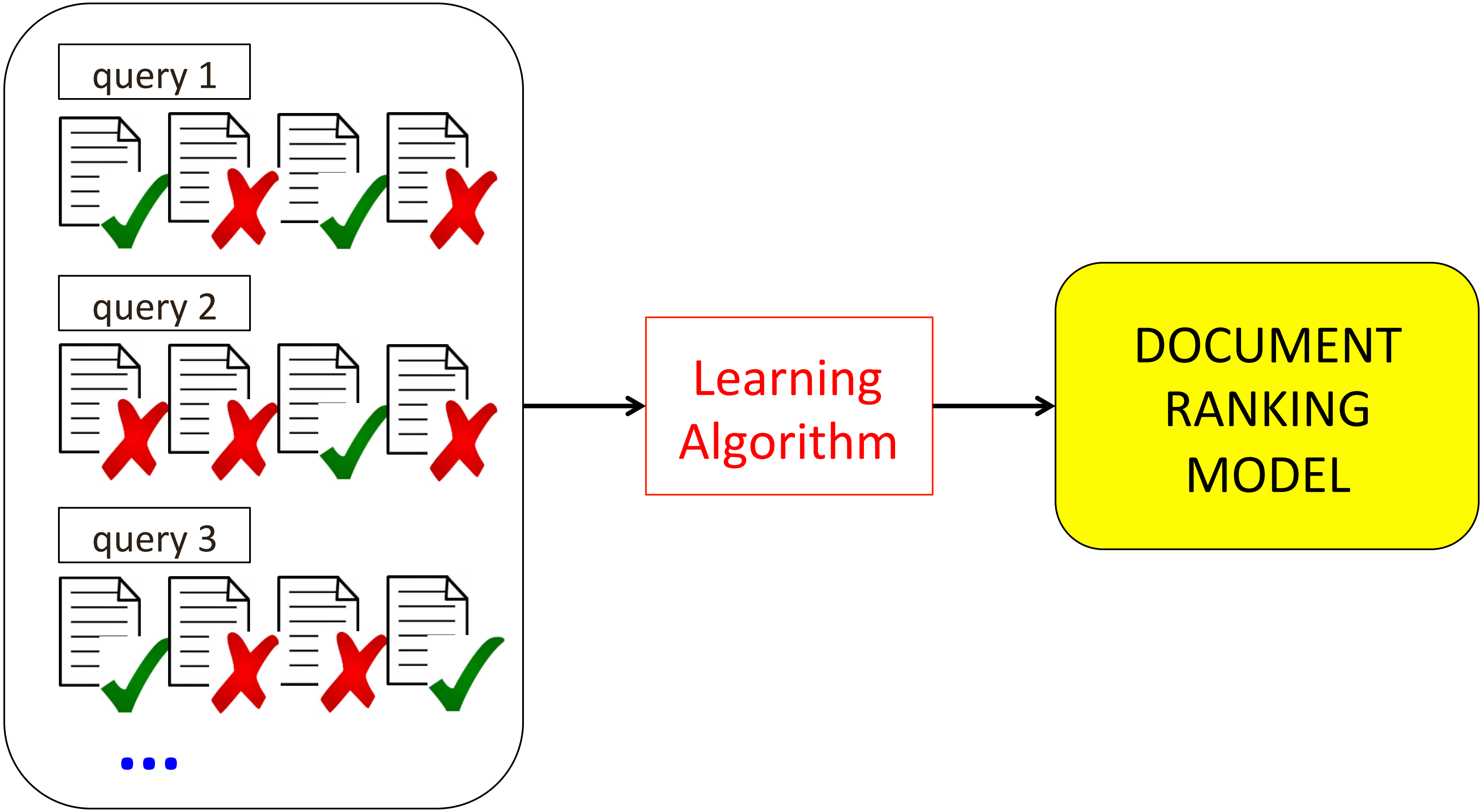
Example: Document Ranking



Example: Document Ranking

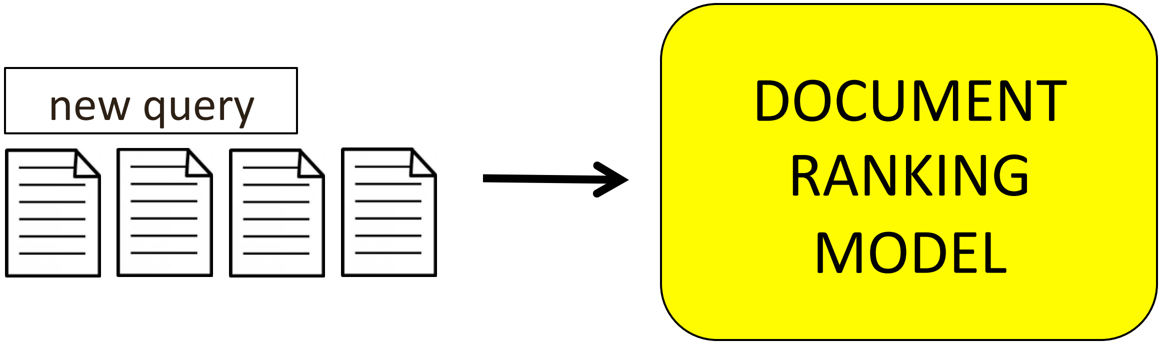


Example: Document Ranking



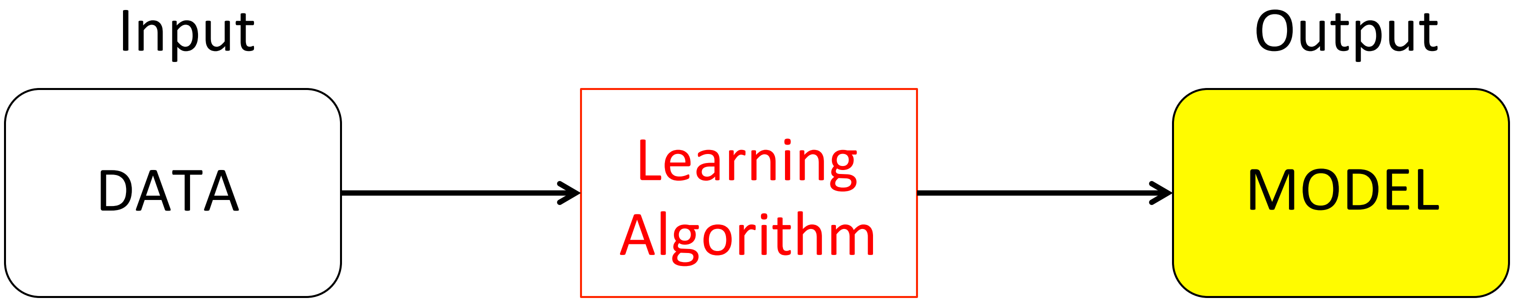
Example: Document Ranking

Example: Document Ranking

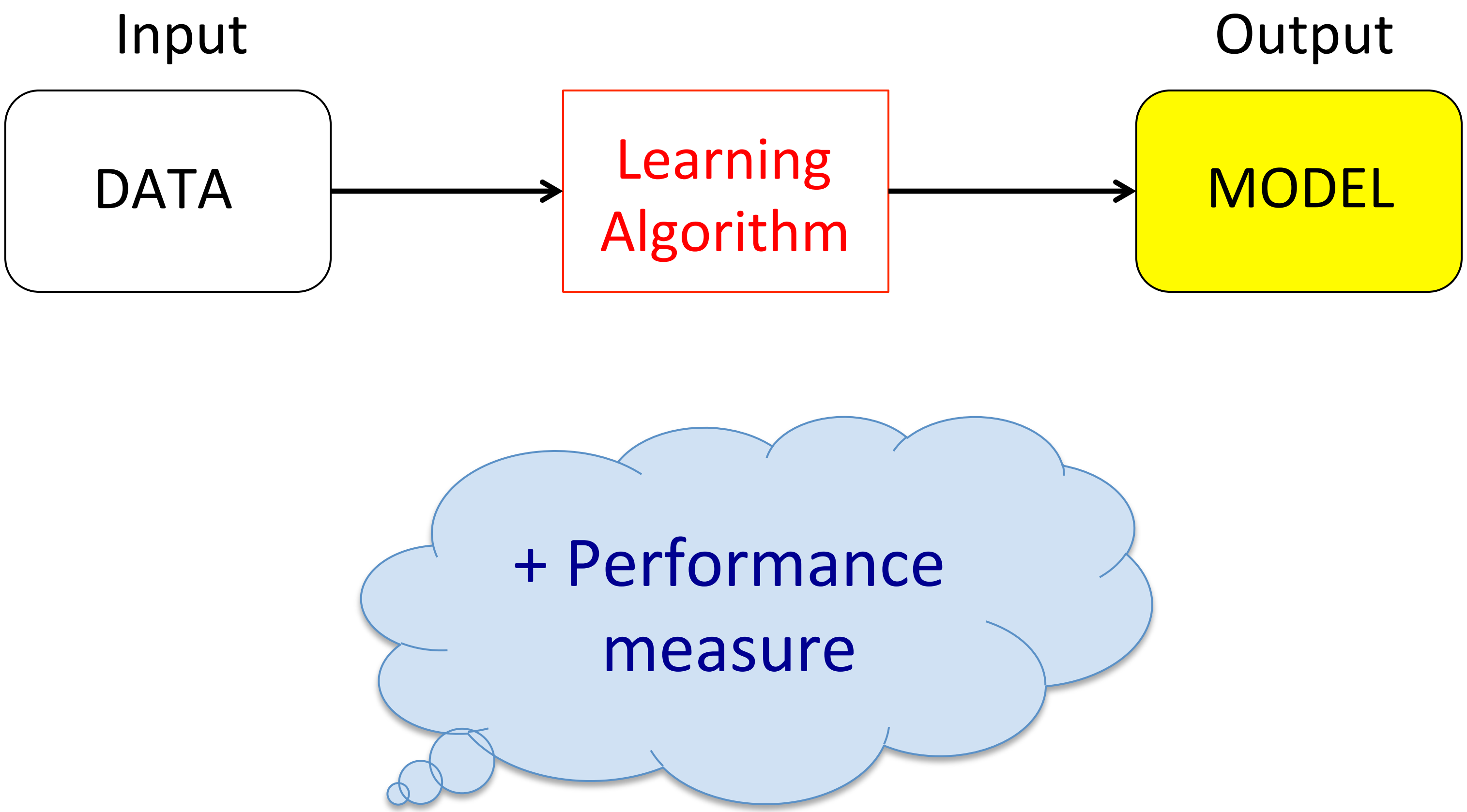


Example: Document Ranking

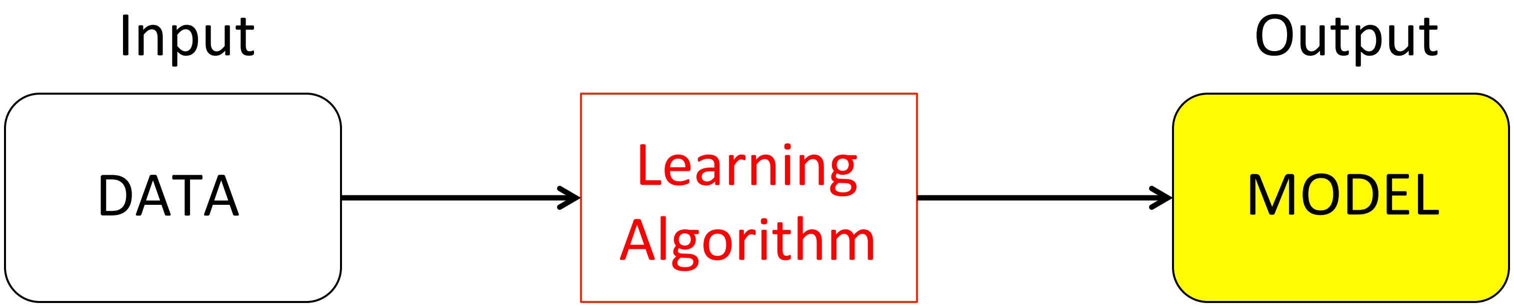
Learning



Learning

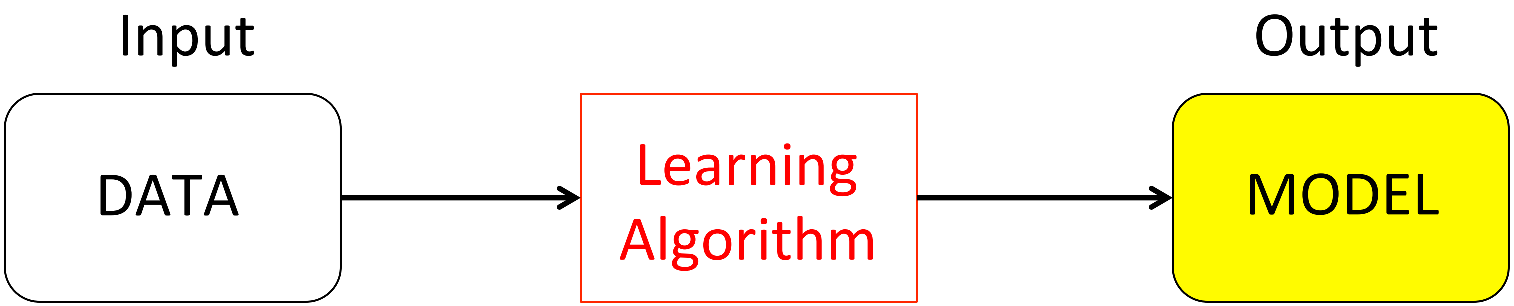

Input
Output
DATA
Learning
Algorithm
MODEL
+ Performance
measure

Supervised Learning



$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$$

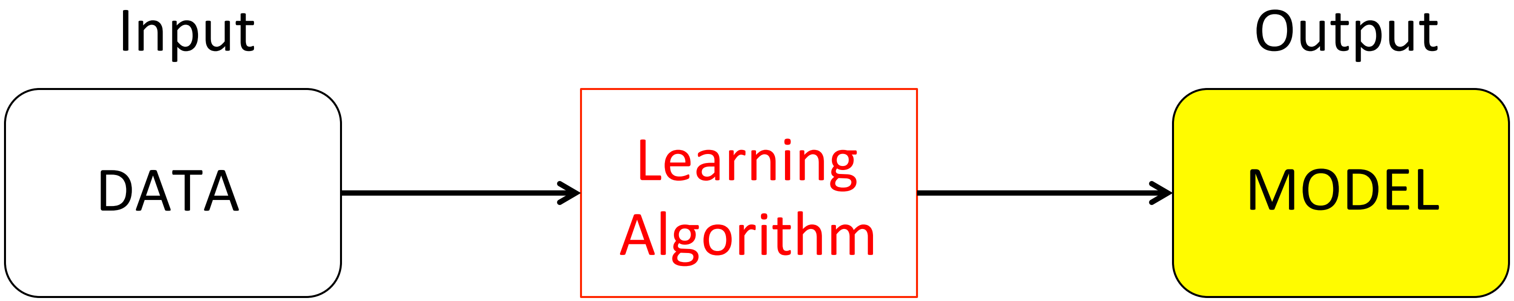
Supervised Learning



$$S = ((x_1, y_1), \ldots, (x_m, y_m))$$
$$\in (X \times Y)^m$$

Instance space

Supervised Learning



$$S = ((x_1, y_1), \ldots, (x_m, y_m))$$

$$\in (X \times Y)^m$$

Instance space

Label space

Supervised Learning

Input

DATA → Learning Algorithm → MODEL

Output

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$$

Instance space

Label space

$$h_S : X \to \hat{Y}$$

Supervised Learning

Input

Output

DATA → Learning Algorithm → MODEL

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$$

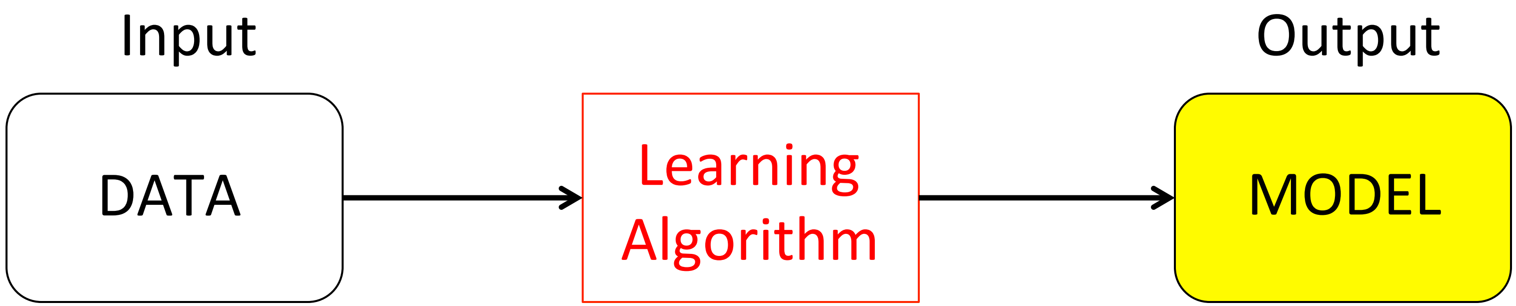
$$h_S : X \rightarrow \hat{Y}$$

Instance space

Label space

Prediction space

Binary Classification

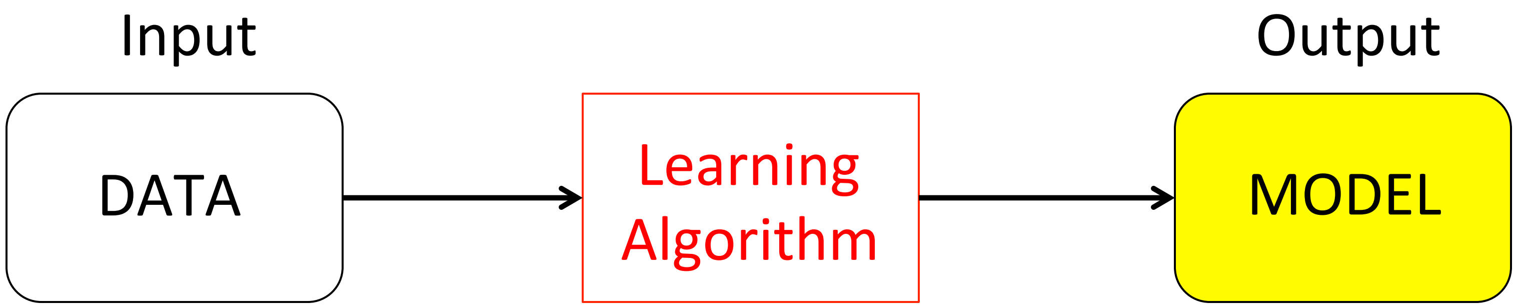


$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$$

$$h_S : X \to \hat{Y}$$

$$Y = \hat{Y} = \{\pm 1\}$$

Prediction in Complex Spaces

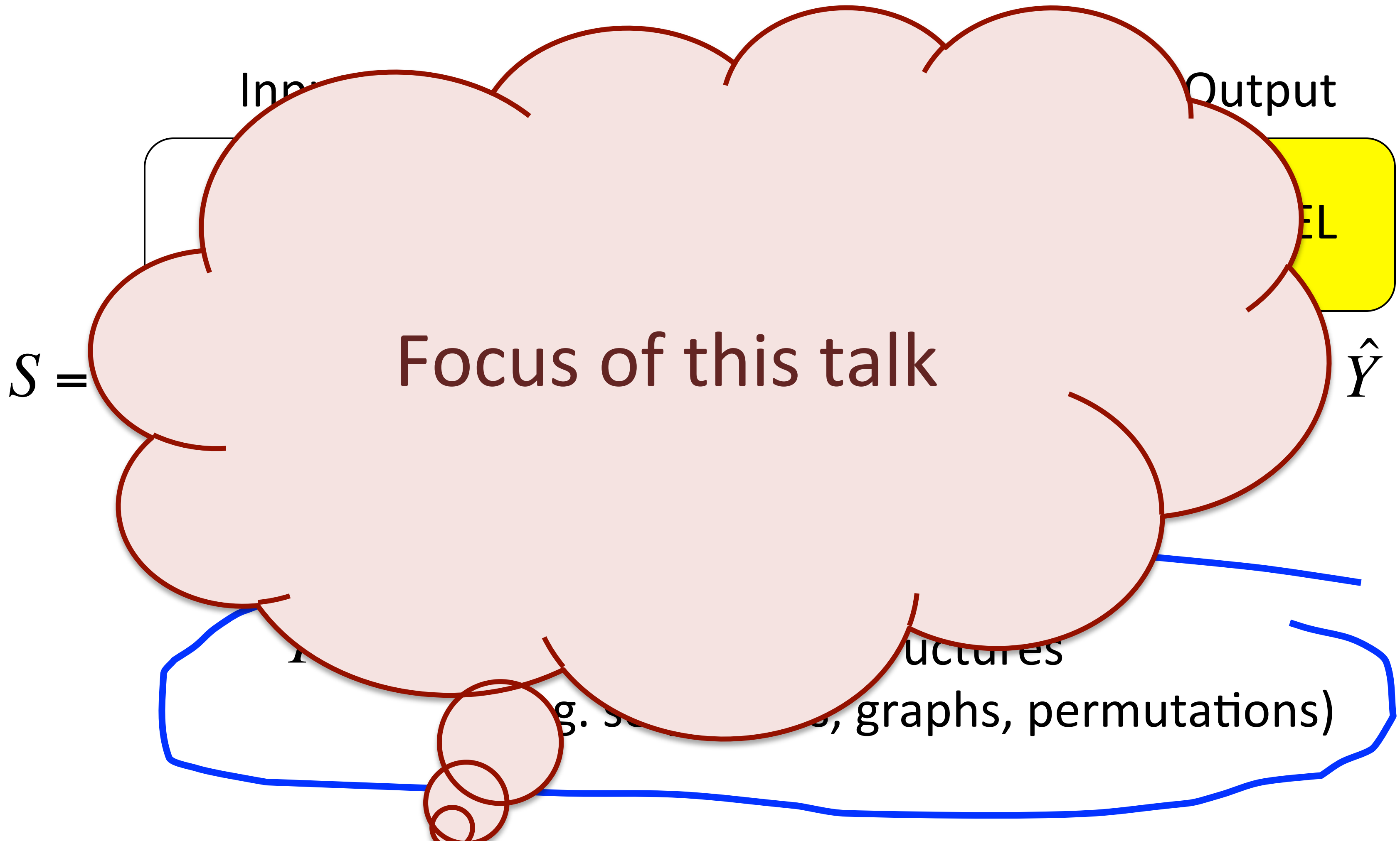


$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$$

$$h_S : X \to \hat{Y}$$

$Y = \hat{Y} =$ sets of complex structures
(e.g. sequences, graphs, permutations)

Prediction in Complex Spaces



Loss Function

Performance in supervised learning is often measured via a (label-dependent) loss function:

$$\ell : Y \times \hat{Y} \rightarrow R_+$$

Loss Function

Performance in supervised learning is often measured via a (label-dependent) loss function:

$$\ell : Y \times \hat{Y} \rightarrow R_+$$

$\ell(y, \hat{y}) =$ `loss' incurred on predicting $\hat{y}$ when true label is y

Loss Matrix

$$\left|Y\right| = n, \ \left|\hat{Y}\right| = k$$

$$\mathbf{L} = \begin{array}{c} \\ 1 \\ 2 \\ \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & & k \end{array} \\ \begin{bmatrix} \ell(1,1) & \ell(1,2) & \cdots & \ell(1,k) \\ \ell(2,1) & \ell(2,2) & \cdots & \ell(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ \ell(n,1) & \ell(n,2) & \cdots & \ell(n,k) \end{bmatrix} \end{array}$$

Example: Binary 0-1 Classification

$$Y = \hat{Y} = \{\pm 1\}$$
$$n = k = 2$$

$$\mathbf{L}^{0\text{-}1} = \begin{array}{c c} & \begin{array}{cc} -1 & +1 \end{array} \\ \begin{array}{c} -1 \\ +1 \end{array} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$$

Example: Multiclass 0-1 Classification

$$Y = \hat{Y} = [n]$$

$$n = k > 2$$

$$\mathbf{L}^{0\text{-}1} = \begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & & n \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ \\ n \end{array} & \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 1 & \cdots & 0 \end{bmatrix} \end{array}$$

Example: Multiclass 0-1 Classification

$$Y = \hat{Y} = [n]$$

$$n = k > 2$$

$$n = 5$$

$$\mathbf{L}^{0\text{-}1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Example: Sequence Prediction with Hamming Loss

$$Y = \hat{Y} = \{0,1\}^r$$

$$n = k = 2^r$$

$r = 3$

$$\mathbf{L}^{\mathrm{Ham}} = \begin{array}{c} \\ 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \\ 1 & 0 & 2 & 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 2 & 2 & 1 \\ 1 & 2 & 2 & 3 & 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 2 & 1 & 0 & 2 & 1 \\ 2 & 3 & 1 & 2 & 1 & 2 & 0 & 1 \\ 3 & 2 & 2 & 1 & 2 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Example: Document Ranking with Pairwise Disagreement Loss

$$Y = \{0,1\}^r,\ \hat{Y} = S_r$$

$$n = 2^r,\ \ k = r!$$

$r = 3$

$$\mathbf{L}^{\mathrm{PD}} = \begin{array}{cc} & \begin{matrix} \begin{pmatrix}1\\2\\3\end{pmatrix} & \begin{pmatrix}1\\3\\2\end{pmatrix} & \begin{pmatrix}2\\3\\1\end{pmatrix} & \begin{pmatrix}2\\1\\3\end{pmatrix} & \begin{pmatrix}3\\1\\2\end{pmatrix} & \begin{pmatrix}3\\2\\1\end{pmatrix} \end{matrix} \\ \begin{matrix} \text{✗✗✗} \\ \text{✗✗✓} \\ \text{✗✓✗} \\ \text{✗✓✓} \\ \text{✓✗✗} \\ \text{✓✗✓} \\ \text{✓✓✗} \\ \text{✓✓✓} \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 2 & 0 & 0 \\ 1 & 2 & 0 & 0 & 2 & 1 \\ 2 & 2 & 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 0 & 2 & 2 & 0 & 1 \\ 0 & 1 & 1 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Formal Setup

Instance space X

Label space $Y = \{1,\ldots,n\} = [n]$

Prediction space $\hat{Y} = \{1,\ldots,k\} = [k]$

Loss matrix $\mathbf{L} \in R_+^{n\times k}$

Formal Setup

Instance space X

Label space $Y = \{1, \ldots, n\} = [n]$

Prediction space $\hat{Y} = \{1, \ldots, k\} = [k]$

Loss matrix $\mathbf{L} \in R_+^{n \times k}$

Goal: Given training sample

$$S = ((x_1, y_1), \ldots (x_m, y_m)) \in (X \times [n])^m,$$

learn prediction model $h_S : X \to [k]$

What is a Good Prediction Model $h : X \rightarrow [k]$?

Should minimize target loss on new instances

What is a Good Prediction Model $h : X \rightarrow [k]$?

Should minimize target loss on new instances

Assume all instance-label pairs drawn i.i.d. from a probability distribution D on $X \times [n]$.

What is a Good Prediction Model $h : X \rightarrow [k]$?

Should minimize target loss on new instances

Assume all instance-label pairs drawn i.i.d. from a probability distribution D on $X \times [n]$.

$$\mathrm{er}_D^{\mathbf{L}}[h] = E_{(x,y)\sim D}[\ell(y, h(x))]$$

Generalization $\mathbf{L}$-error (or $\mathbf{L}$-risk) of h (w.r.t. D)

Bayes Error and Regret

Bayes $\mathbf{L}$-error for D:

$$\mathrm{er}_D^{\mathbf{L},*} = \inf_{h:X\to[k]} \mathrm{er}_D^{\mathbf{L}}[h]$$

Bayes Error and Regret

Bayes $\mathbf{L}$-error for D:

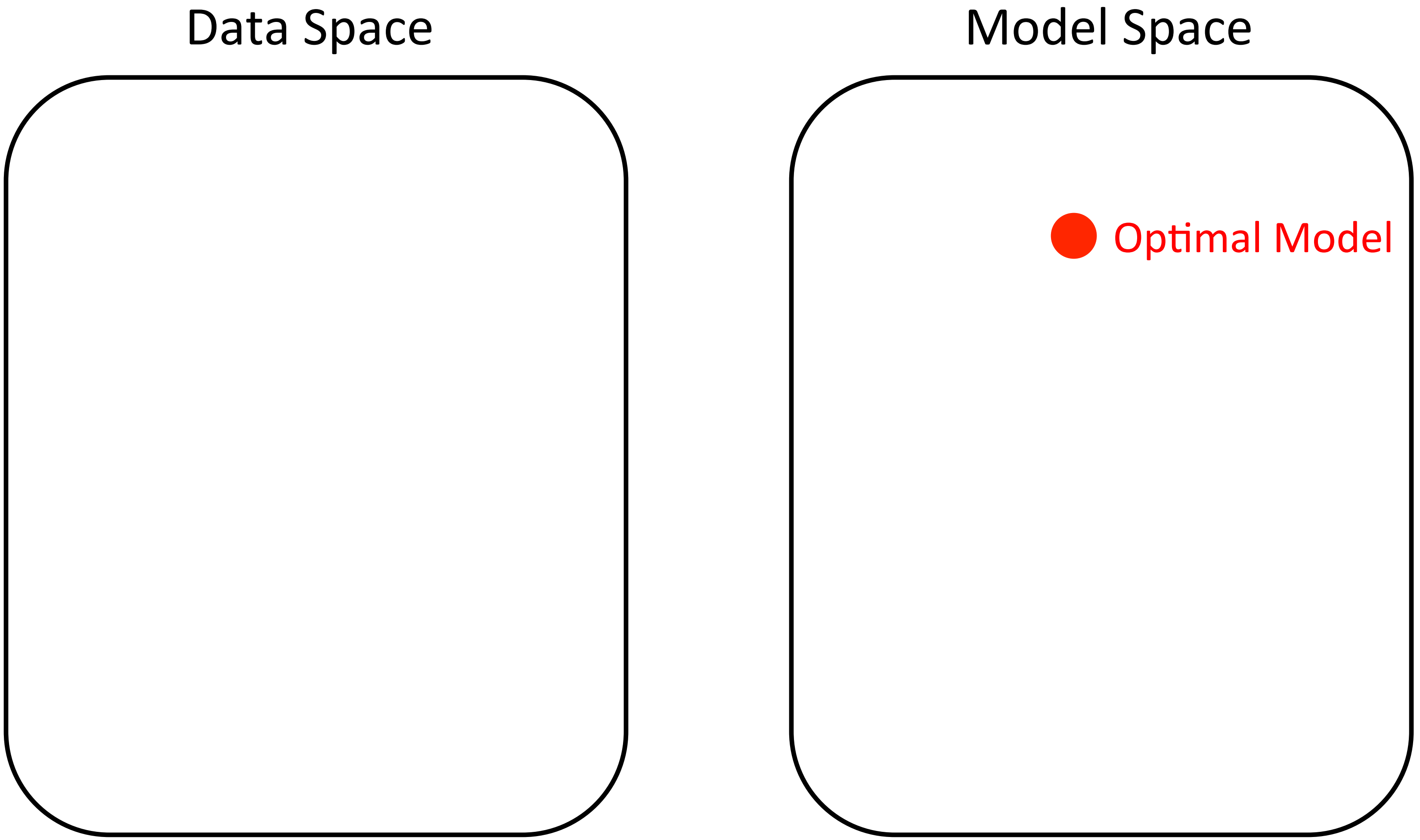
$$\mathrm{er}_D^{\mathbf{L},*} = \inf_{h:X\to[k]} \mathrm{er}_D^{\mathbf{L}}[h]$$

$\mathbf{L}$-regret of h w.r.t. D:

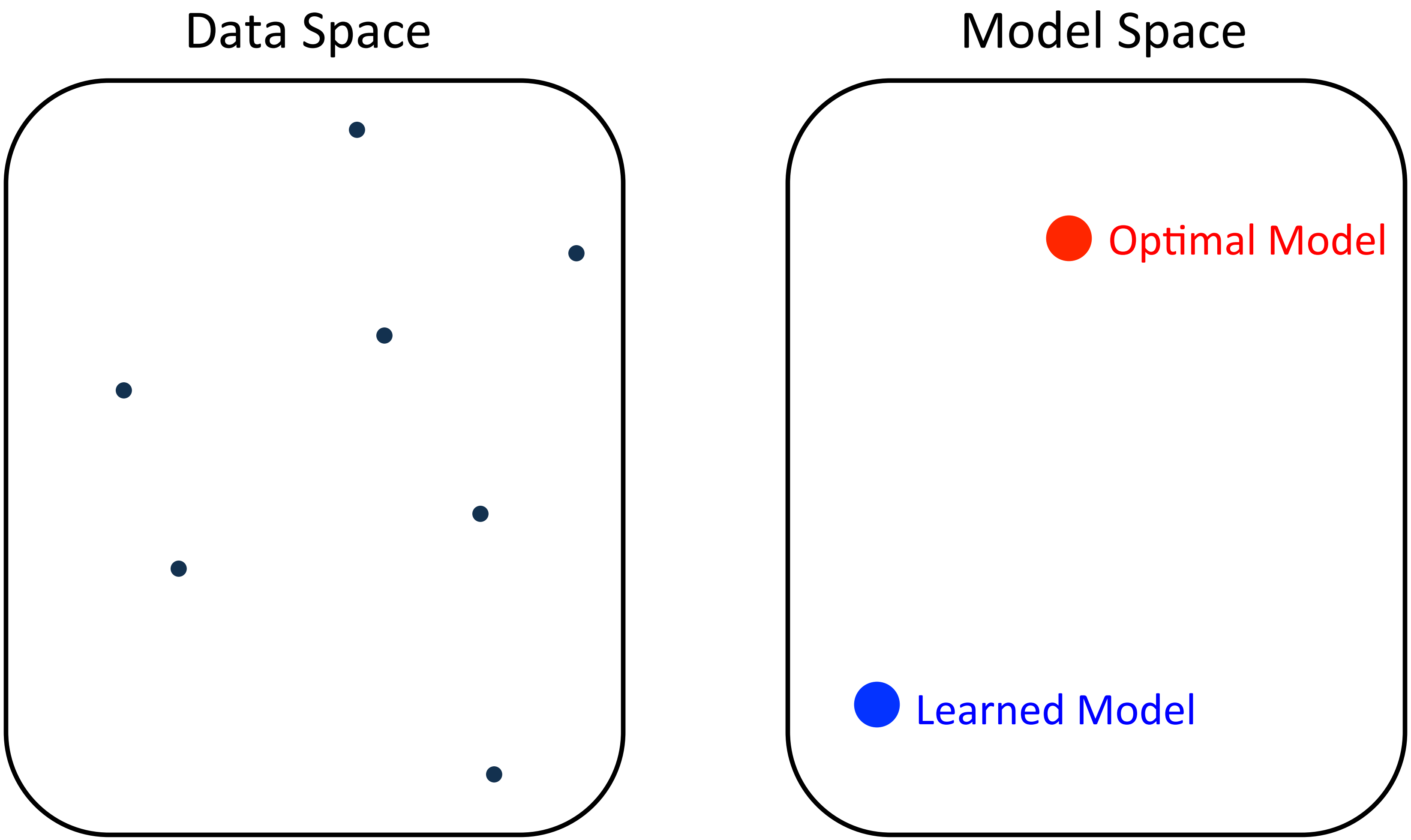
$$\mathrm{regret}_D^{\mathbf{L}}[h] = \mathrm{er}_D^{\mathbf{L}}[h] - \mathrm{er}_D^{\mathbf{L},*}$$

What is a Good Learning Algorithm?

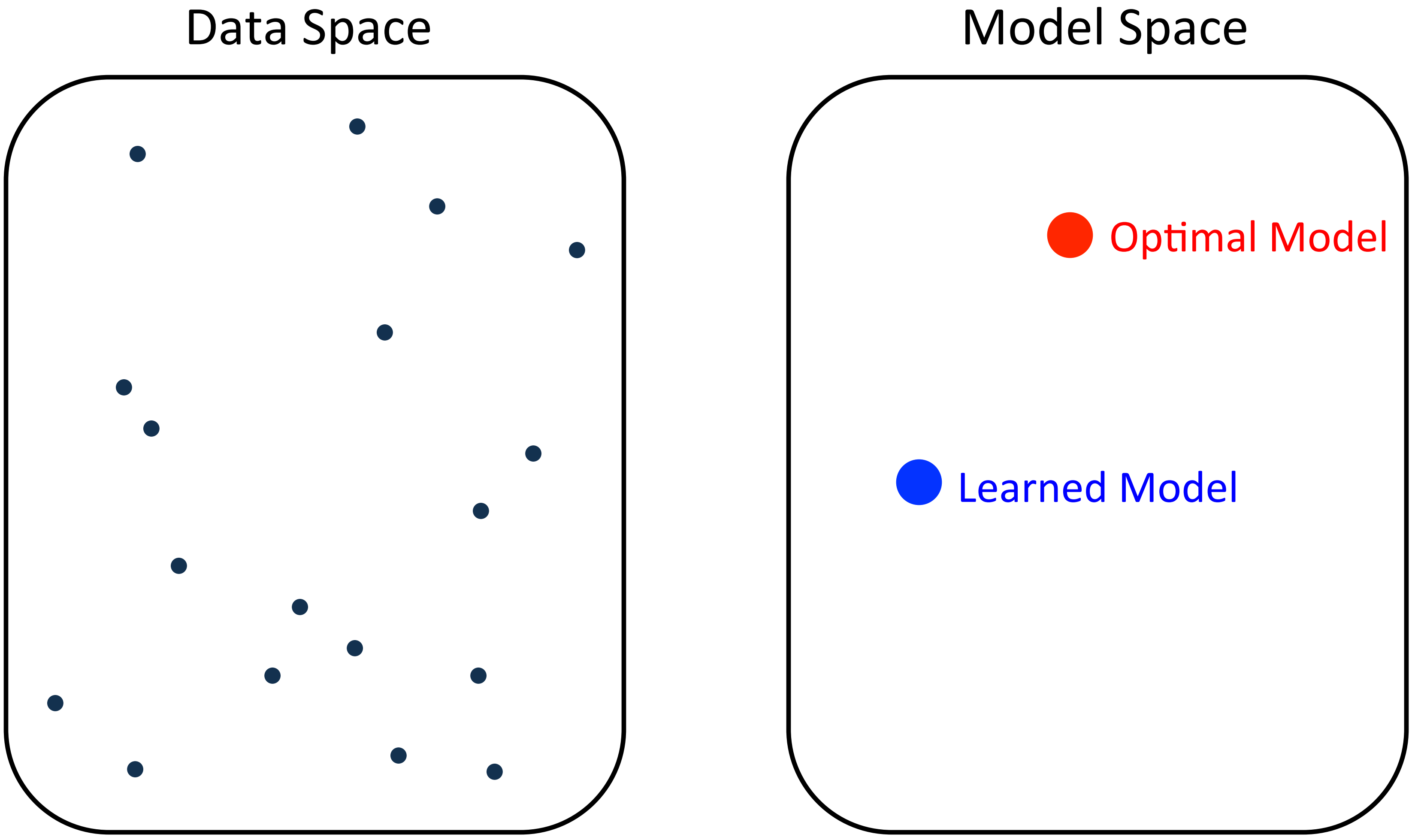
What is a Good Learning Algorithm?



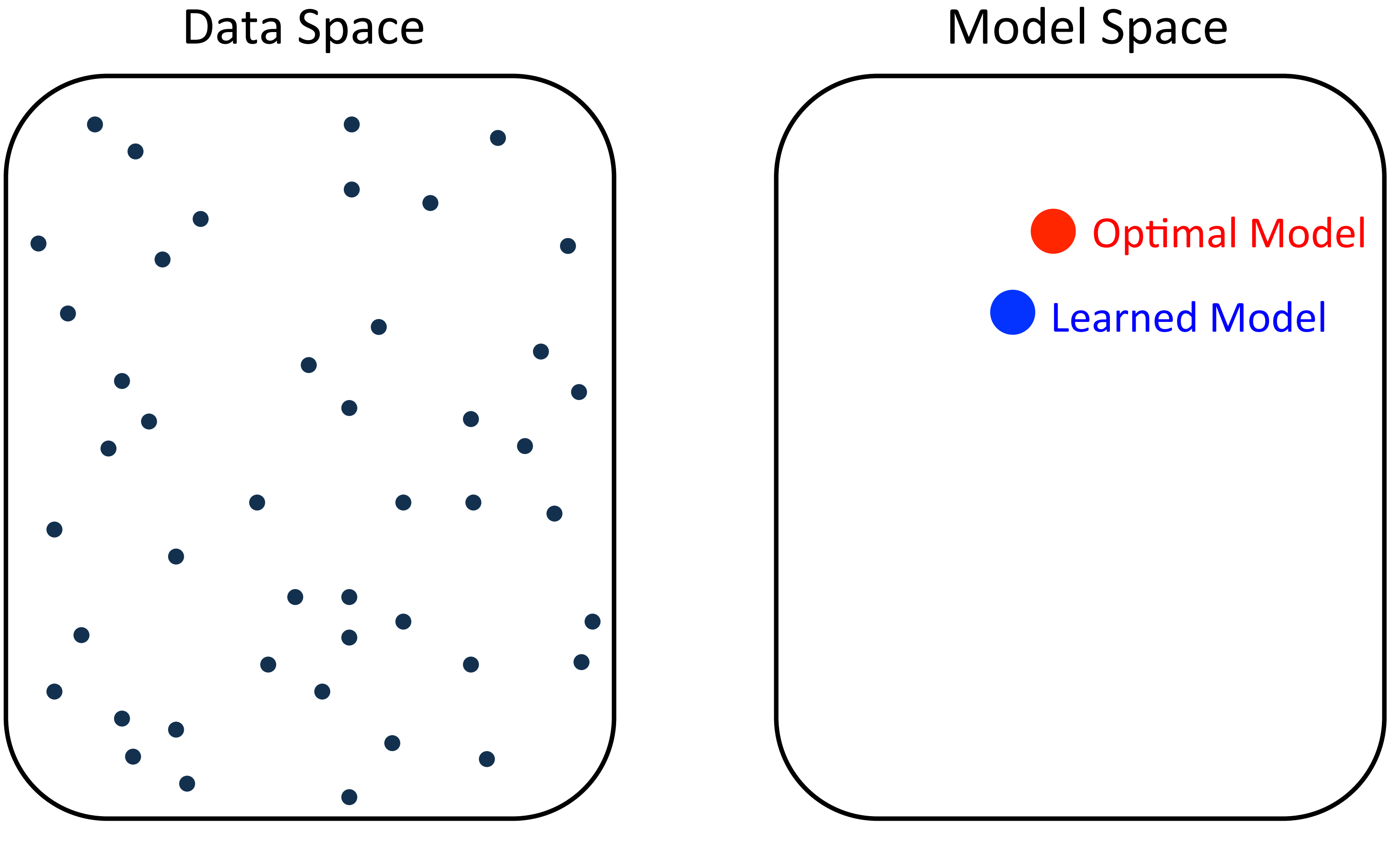
What is a Good Learning Algorithm?



What is a Good Learning Algorithm?



What is a Good Learning Algorithm?



Statistical Consistency

Bayes $\mathbf{L}$-consistent w.r.t. D if for $S \sim D^m$:

$$\mathrm{regret}_D^{\mathbf{L}}[h_S] \xrightarrow{P} 0 \quad \text{as} \quad m \to \infty.$$

Statistical Consistency

Universally Bayes $\mathbf{L}$-consistent if
Bayes $\mathbf{L}$-consistent w.r.t. *all* distributions D

How do we design a statistically consistent learning algorithm for a given loss matrix $\mathbf{L}$?

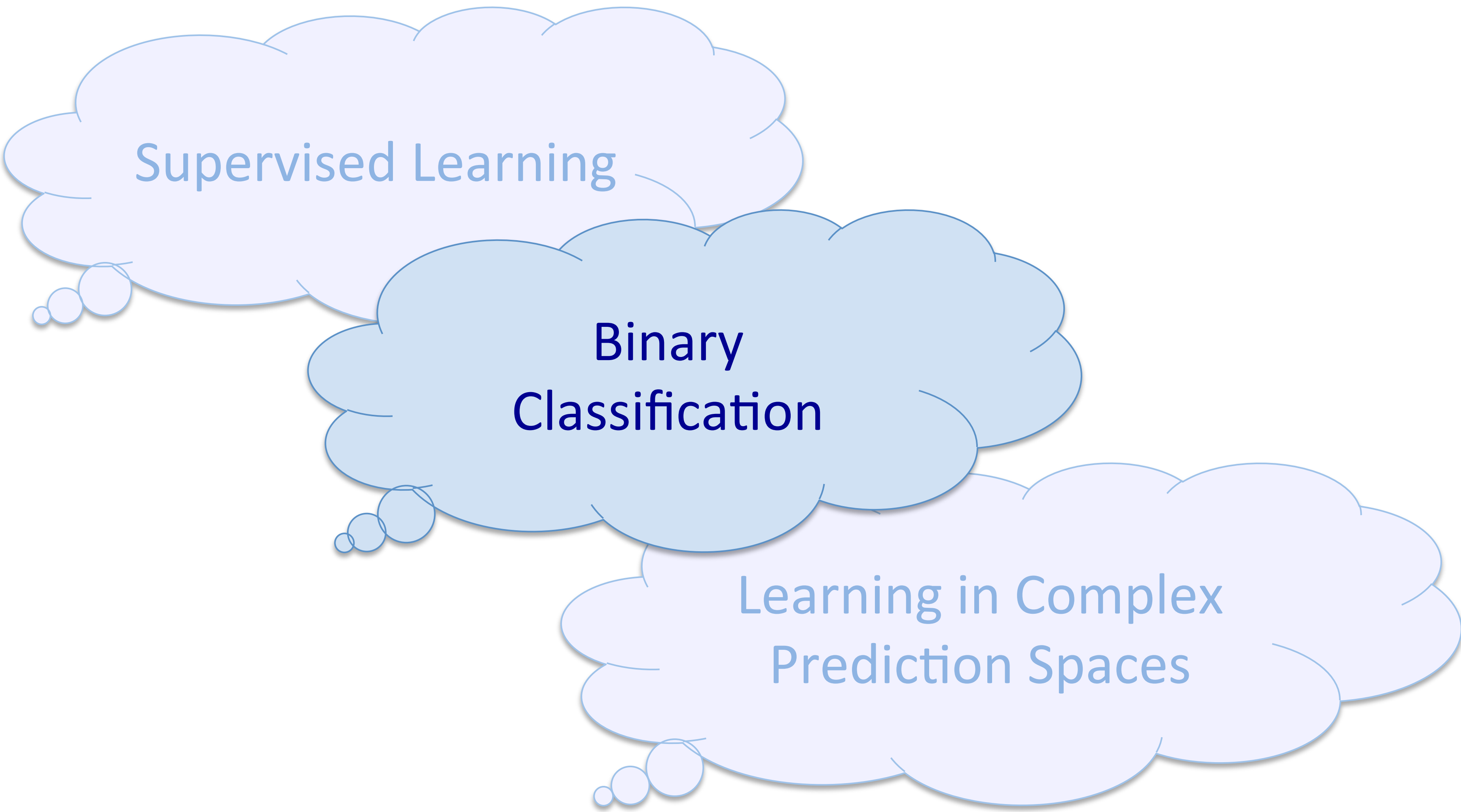
Road Map

Supervised Learning

Binary Classification

Learning in Complex Prediction Spaces

Road Map


Supervised Learning
Binary Classification
Learning in Complex Prediction Spaces

Empirical Risk Minimization (ERM)

Let H be some class of functions from X to $\{\pm 1\}$.

S → ERM in H →

Empirical Risk Minimization (ERM)

Let H be some class of functions from X to $\{\pm 1\}$.

S → ERM in H → $$h_S \in \operatorname*{argmin}_{h \in H} \sum_{i=1}^{m} \underbrace{\mathbf{1}(h(x_i) \neq y_i)}_{\text{0-1 loss}}$$

Empirical Risk Minimization (ERM)

Let H be some class of functions from X to $\{\pm 1\}$.

S $\longrightarrow$ ERM in H $\longrightarrow$

$$h_S \in \operatorname*{argmin}_{h \in H} \sum_{i=1}^{m} \underbrace{\mathbf{1}(h(x_i) \neq y_i)}_{\text{0-1 loss}}$$

For suitable H, universally 0-1 consistent in H;
suitable extensions can be made universally Bayes 0-1 consistent

Empirical Risk Minimization (ERM)

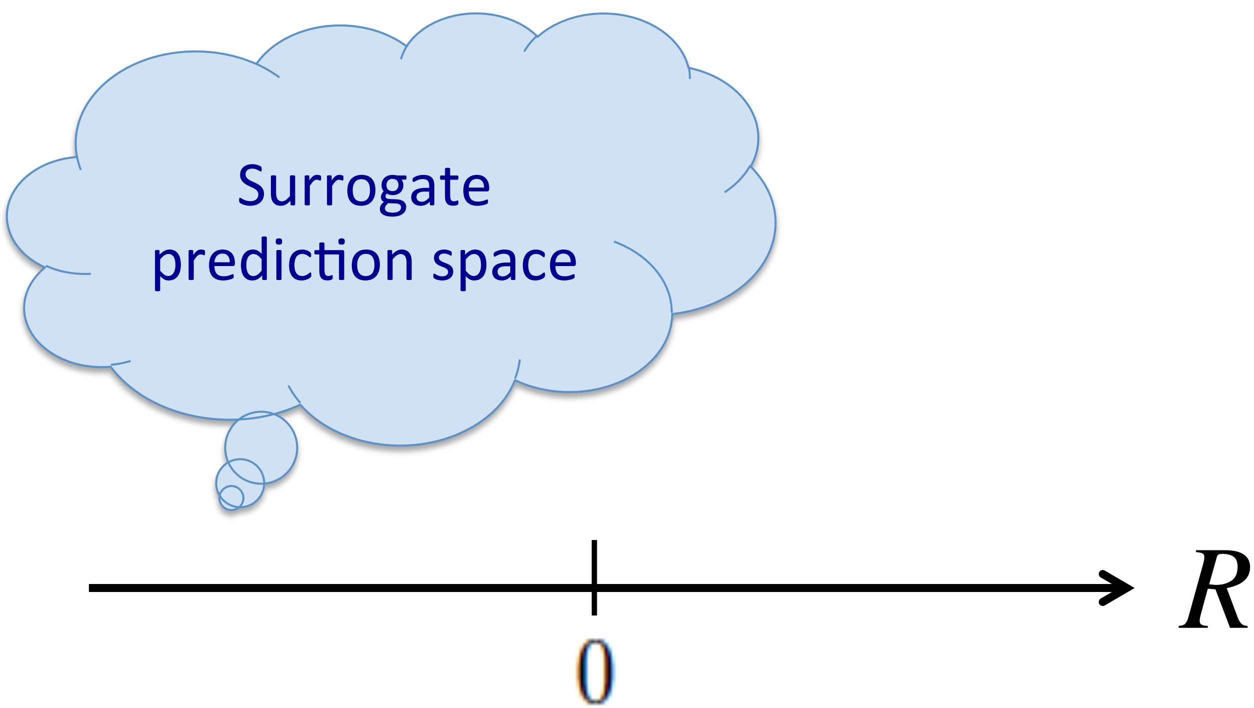
Let H be some class of functions from X to $\{\pm 1\}$.

$$S \longrightarrow \boxed{\text{ERM in } H} \longrightarrow h_S \in \operatorname*{argmin}_{h \in H} \sum_{i=1}^{m} \underbrace{\mathbf{1}(h(x_i) \neq y_i)}_{\text{0-1 loss}}$$

✓ For suitable H, universally 0-1 consistent in H;
suitable extensions can be made universally Bayes 0-1 consistent

✗ Computationally hard!

Surrogate Risk Minimization



Surrogate Risk Minimization

$$\psi : \{\pm 1\} \times R \to R_+$$

Surrogate loss function

Surrogate Risk Minimization

$$\psi : \{\pm 1\} \times R \rightarrow R_+$$

Surrogate loss function

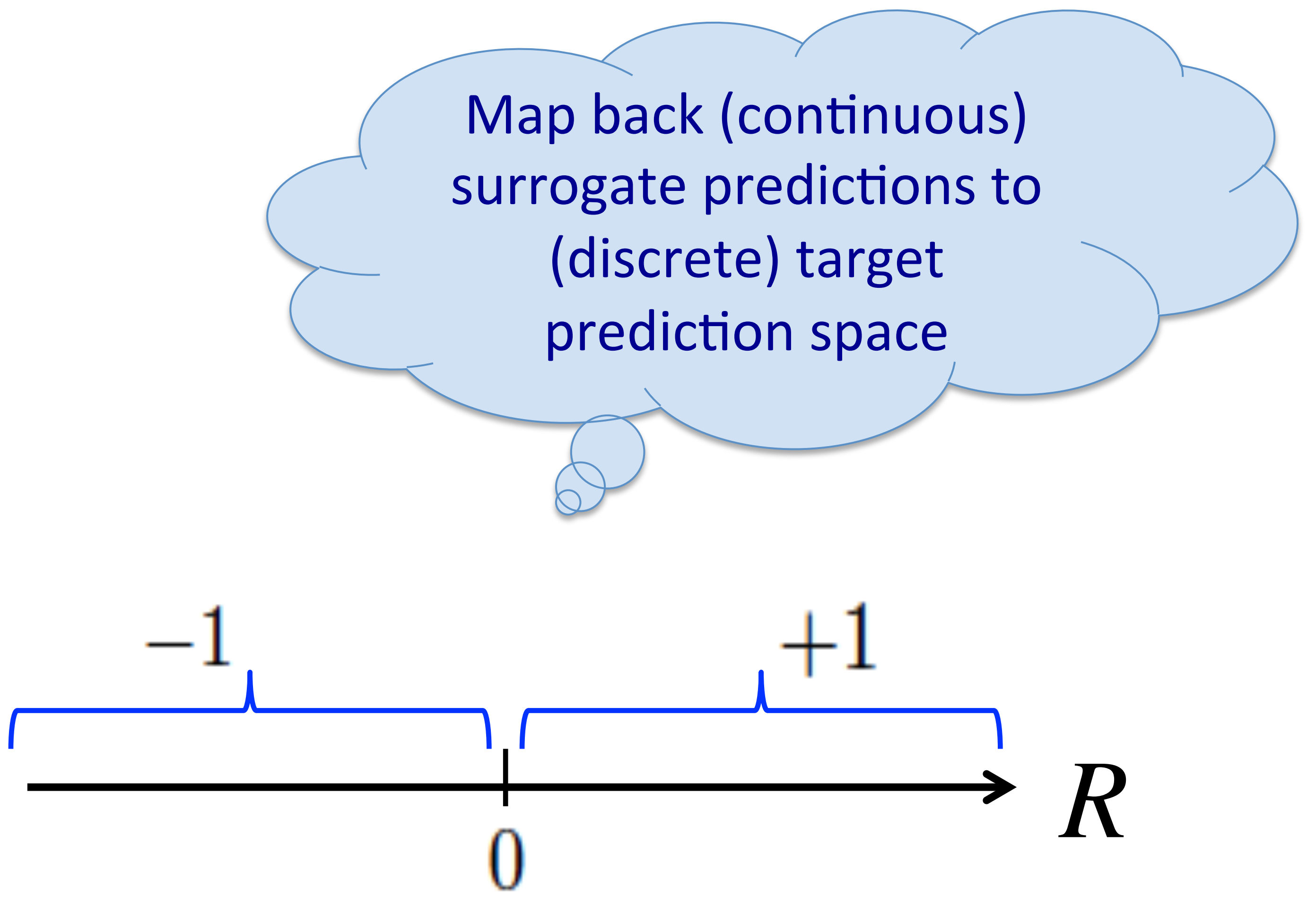
Surrogate Risk Minimization

$$\min_{f} \sum_{i=1}^{m} \psi(y_i, f(x_i))$$

Functions mapping
X to R

Surrogate optimization problem (convex for suitable surrogate loss)

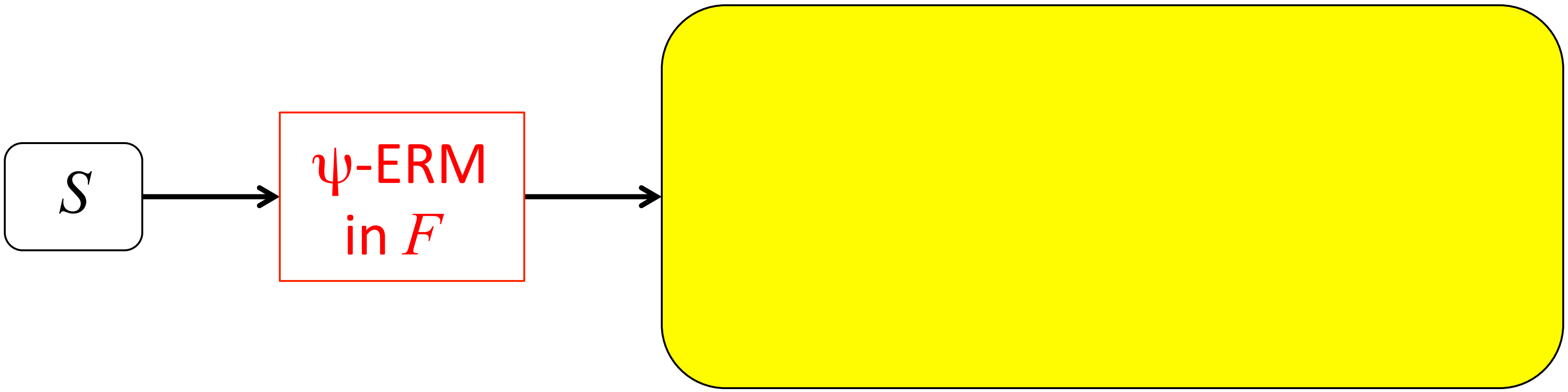
Surrogate Risk Minimization



Surrogate Risk Minimization

Let $\psi : \{\pm 1\} \times R \to R_+$.

Let F be some class of functions from X to R.



Surrogate Risk Minimization

Let $\psi : \{\pm 1\} \times R \rightarrow R_+$.

Let F be some class of functions from X to R.

S → ψ-ERM in F →

$$f_S \in \operatorname*{argmin}_{f \in F} \sum_{i=1}^{m} \underbrace{\psi(y_i, f(x_i))}_{\psi\text{-loss}}$$

Surrogate Risk Minimization

Let $\psi : \{\pm 1\} \times R \to R_+$.

Let F be some class of functions from X to R.

S → ψ-ERM in F →

$$f_S \in \operatorname*{argmin}_{f \in F} \sum_{i=1}^{m} \underbrace{\psi(y_i, f(x_i))}_{\psi\text{-loss}}$$

$$h_S = \operatorname{sign} \circ f_S$$

Surrogate Risk Minimization

Let $\psi : \{\pm 1\} \times R \to R_+$.

Let F be some class of functions from X to R.

S → ψ-ERM in F →

$$f_S \in \operatorname*{argmin}_{f \in F} \sum_{i=1}^{m} \underbrace{\psi(y_i, f(x_i))}_{\psi\text{-loss}}$$

$$h_S = \mathrm{sign} \circ f_S$$

✓ For convex ψ and suitable F, computationally efficient!

Example: Support Vector Machines

$$S \longrightarrow \boxed{\text{SVM}} \longrightarrow$$

$$f_S \in \operatorname*{argmin}_{f\in F} \sum_{i=1}^{m} (1 - y_i f(x_i))_+$$

$$h_S = \operatorname{sign} \circ f_S$$

Hinge surrogate

Example: Logistic Regression

S → Logistic Regression →

$$f_S \in \operatorname*{argmin}_{f \in F} \sum_{i=1}^{m} \ln(1 + e^{-y_i f(x_i)})$$

$$h_S = \operatorname{sign} \circ f_S$$

Logistic surrogate

Example: AdaBoost

S → AdaBoost →

$$f_S \in \operatorname*{argmin}_{f \in F} \sum_{i=1}^{m} e^{-y_i f(x_i)}$$

$$h_S = \mathrm{sign} \circ f_S$$

Exponential surrogate

Surrogate Risk Minimization

Let $\psi : \{\pm 1\} \times R \to R_+$.

Let F be some class of functions from X to R.

$$S \longrightarrow \boxed{\psi\text{-ERM in } F} \longrightarrow$$

$$f_S \in \operatorname*{argmin}_{f \in F} \sum_{i=1}^{m} \underbrace{\psi(y_i, f(x_i))}_{\psi\text{-loss}}$$

$$h_S = \operatorname{sign} \circ f_S$$

✓ For convex ψ and suitable F, computationally efficient!

Surrogate Risk Minimization

Let $\psi : \{\pm 1\} \times R \to R_+$.

Let F be some class of functions from X to R.

S → ψ-ERM in F →

$$f_S \in \operatorname*{argmin}_{f \in F} \sum_{i=1}^{m} \underbrace{\psi(y_i, f(x_i))}_{\psi\text{-loss}}$$

$$h_S = \operatorname{sign} \circ f_S$$

For convex ψ and suitable F, computationally efficient!

? For suitable F, universally ψ-consistent in F;
suitable extensions can be made universally Bayes ψ-consistent

Classification-Calibrated Surrogates

Theorem. If ψ is `classification-calibrated', then

Bayes ψ-consistency $\Longrightarrow$ Bayes 0-1 consistency (after applying sign)

[Bartlett et al., 2006]

Classification-Calibrated Surrogates

Theorem. If ψ is `classification-calibrated', then

Bayes ψ-consistency $\Longrightarrow$ Bayes 0-1 consistency (after applying sign)

✓ Hinge

[Bartlett et al., 2006]

Classification-Calibrated Surrogates

Theorem. If ψ is `classification-calibrated', then

Bayes ψ-consistency $\Longrightarrow$ Bayes 0-1 consistency (after applying sign)

✓ Hinge ✓ Logistic

[Bartlett et al., 2006]

Classification-Calibrated Surrogates

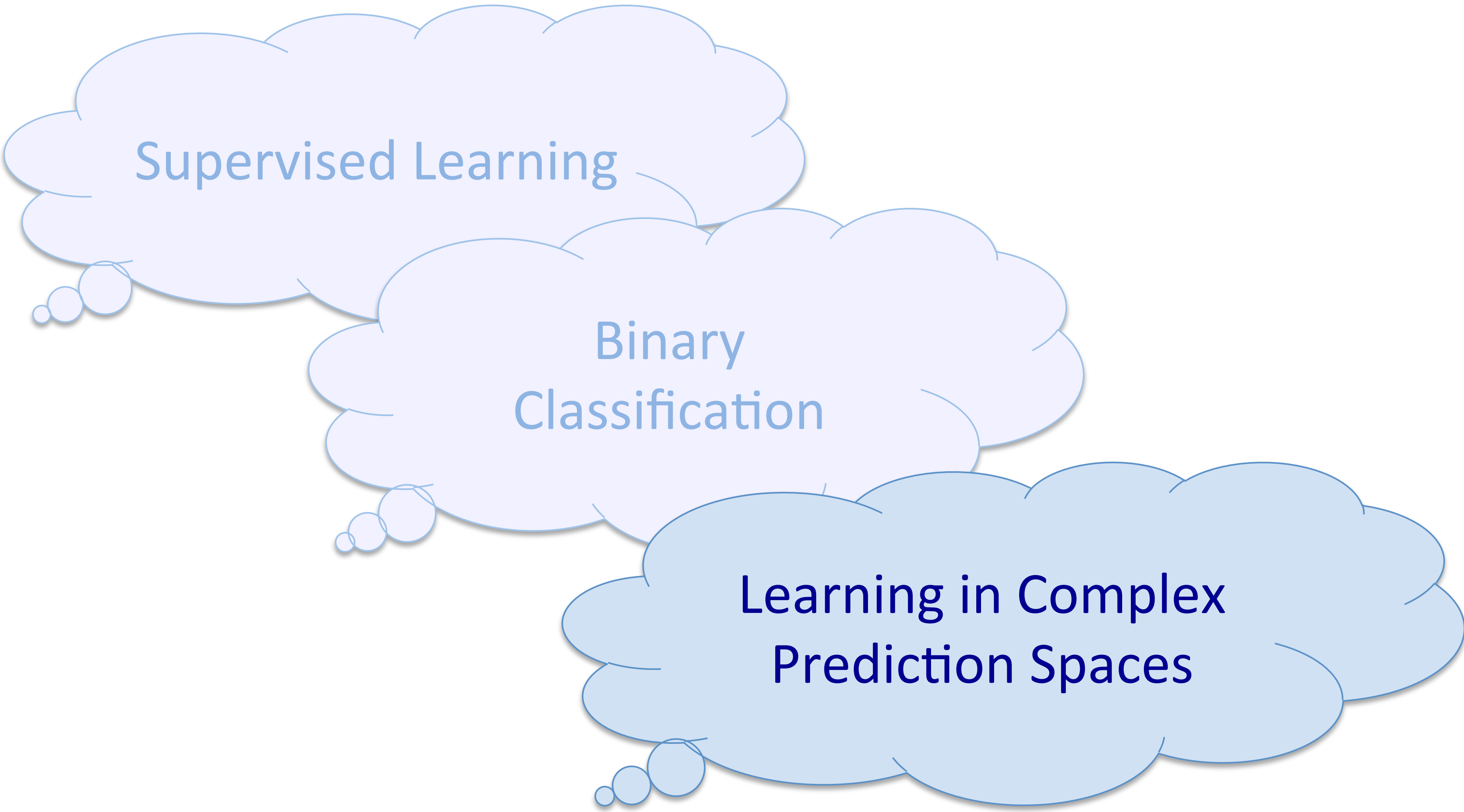
Theorem. If ψ is `classification-calibrated', then

Bayes ψ-consistency $\Longrightarrow$ Bayes 0-1 consistency (after applying sign)

✓ Hinge ✓ Logistic ✓ Exponential

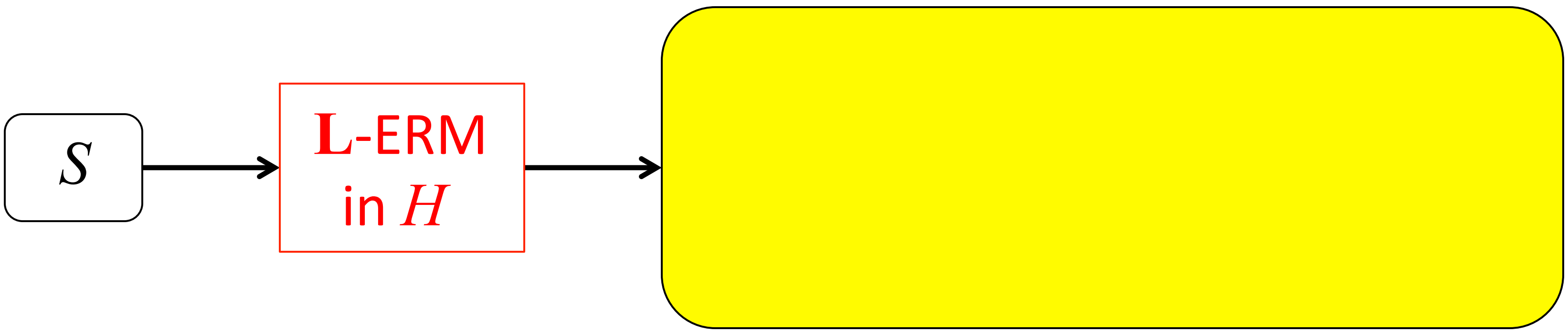
[Bartlett et al., 2006]

Road Map


Supervised Learning
Binary Classification
Learning in Complex Prediction Spaces

Empirical Risk Minimization (ERM)

Let H be some class of functions from X to $[k]$.



Empirical Risk Minimization (ERM)

Let H be some class of functions from X to $[k]$.

S $\longrightarrow$ **L**-ERM in H $\longrightarrow$

$$h_S \in \operatorname*{argmin}_{h \in H} \sum_{i=1}^{m} \underbrace{\ell(y_i, h(x_i))}_{\text{Target loss}}$$

Empirical Risk Minimization (ERM)

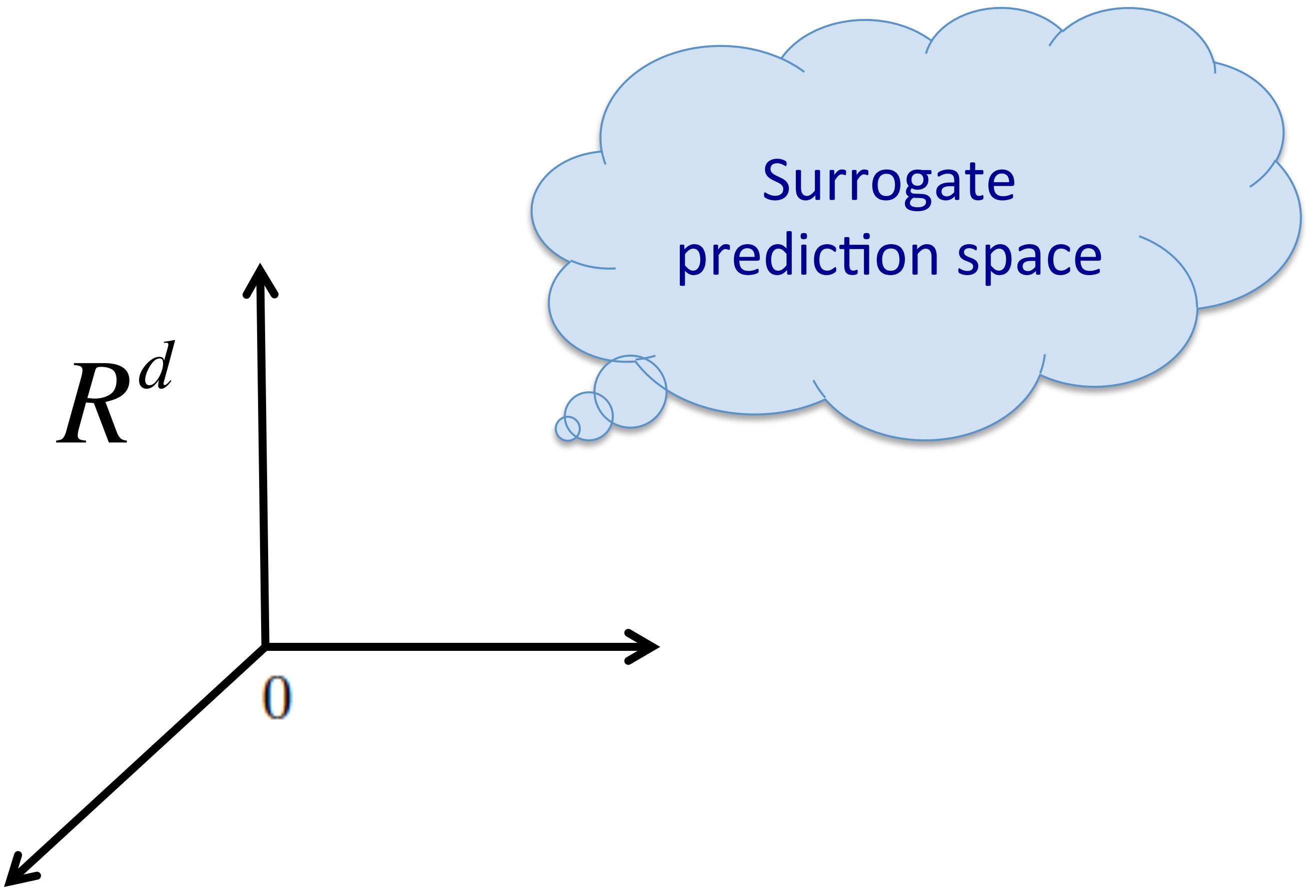
Let H be some class of functions from X to $[k]$.

$$S \longrightarrow \boxed{\begin{array}{c}\mathbf{L}\text{-ERM}\\ \text{in } H\end{array}} \longrightarrow h_S \in \operatorname*{argmin}_{h\in H} \sum_{i=1}^{m} \underbrace{\ell(y_i, h(x_i))}_{\text{Target loss}}$$

✓ For suitable H, universally $\mathbf{L}$-consistent in H;
suitable extensions can be made universally Bayes $\mathbf{L}$-consistent

✗ Computationally hard!

Surrogate Risk Minimization



Surrogate Risk Minimization

$$\psi : [n] \times R^d \rightarrow R_+$$

Surrogate loss function

Surrogate Risk Minimization

$$\psi : [n] \times R^d \to R_+$$

Surrogate loss function

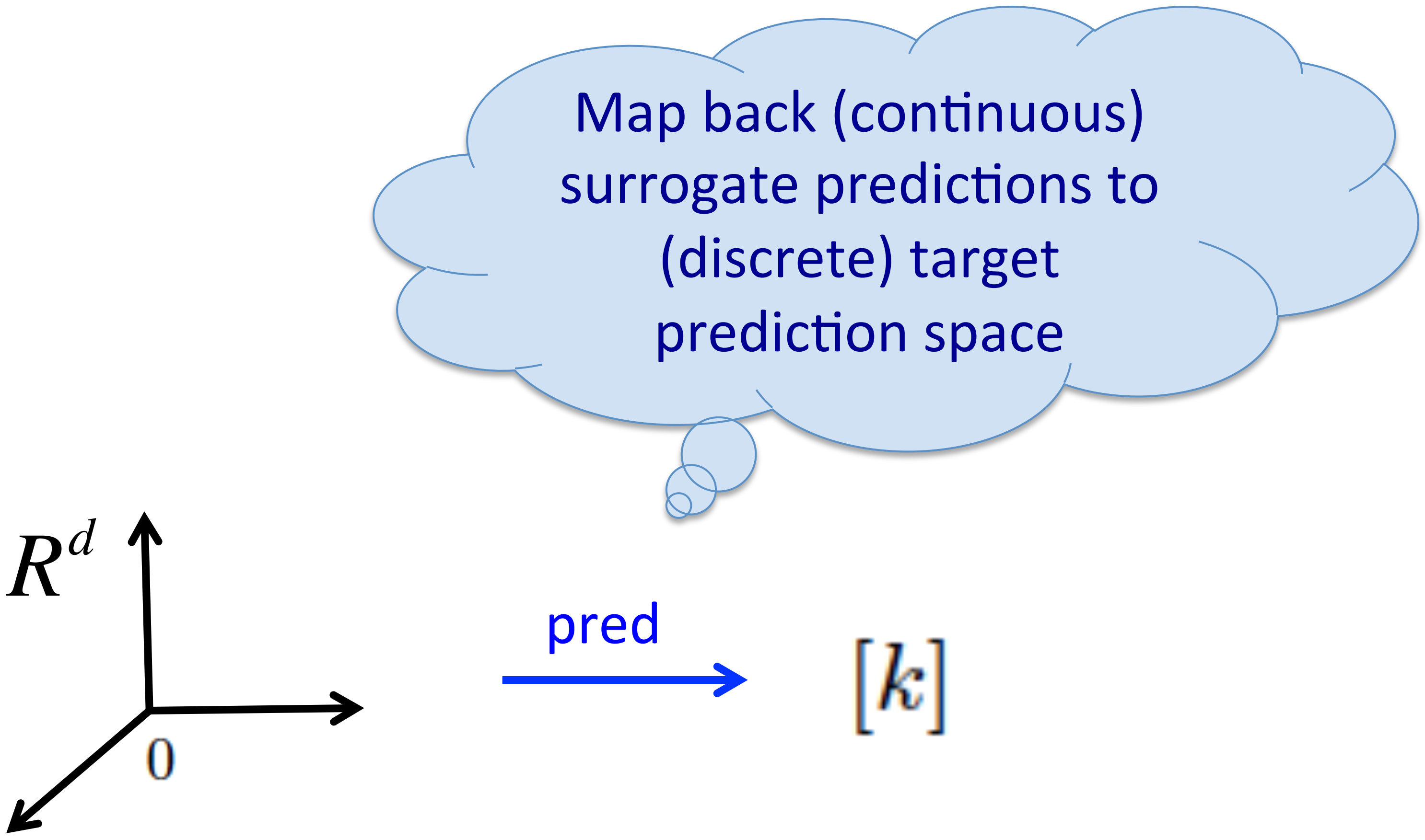
Surrogate Risk Minimization

$$\min_{\mathbf{f}} \sum_{i=1}^{m} \psi(y_i, \mathbf{f}(x_i))$$

Functions mapping
X to R^d

Surrogate optimization problem (convex for suitable surrogate loss)

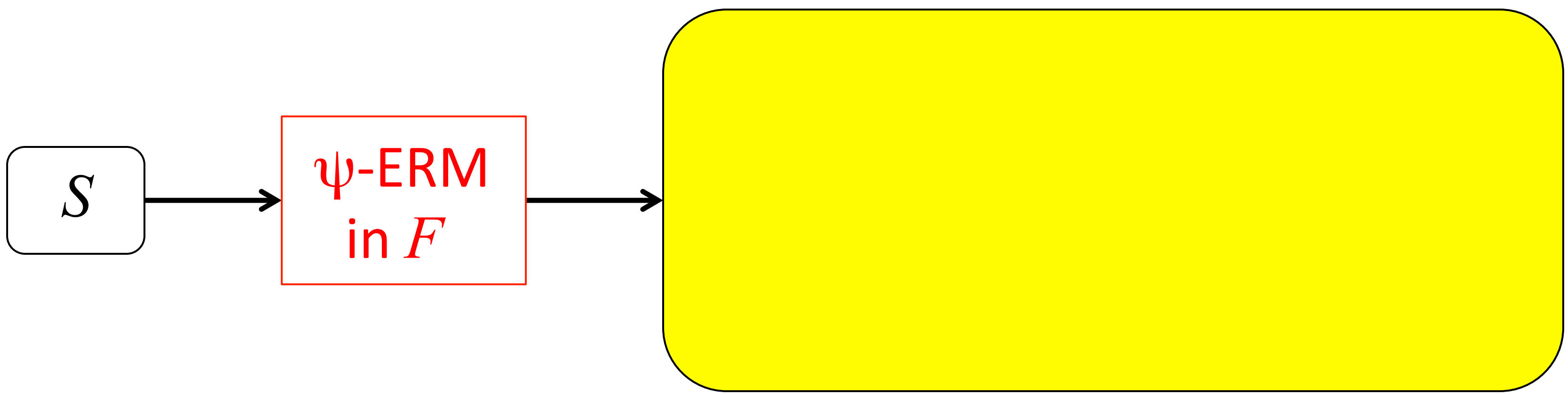
Surrogate Risk Minimization



Surrogate Risk Minimization

Let $\psi : [n] \times R^d \to R_+,\ \text{pred} : R^d \to [k].$

Let F be some class of functions from X to R^d.



Surrogate Risk Minimization

Let $\psi : [n] \times R^d \rightarrow R_+, \ \text{pred} : R^d \rightarrow [k].$

Let F be some class of functions from X to R^d.

S

ψ-ERM in F

$$\mathbf{f}_S \in \operatorname*{argmin}_{\mathbf{f} \in F} \sum_{i=1}^{m} \underbrace{\psi(y_i, \mathbf{f}(x_i))}_{\psi\text{-loss}}$$

Surrogate Risk Minimization

Let $\psi : [n] \times R^d \to R_+, \ \mathrm{pred} : R^d \to [k]$.

Let F be some class of functions from X to R^d.

$S \longrightarrow$ ψ-ERM in F $\longrightarrow$

$$\mathbf{f}_S \in \operatorname*{argmin}_{\mathbf{f}\in F} \sum_{i=1}^{m} \underbrace{\psi(y_i, \mathbf{f}(x_i))}_{\psi\text{-loss}}$$

$$h_S = \mathrm{pred} \circ \mathbf{f}_S$$

Surrogate Risk Minimization

Let $\psi : [n] \times R^d \rightarrow R_+$, $\mathrm{pred} : R^d \rightarrow [k]$.

Let F be some class of functions from X to R^d.

S

ψ-ERM in F

$$\mathbf{f}_S \in \underset{\mathbf{f} \in F}{\mathrm{argmin}} \sum_{i=1}^{m} \underbrace{\psi(y_i, \mathbf{f}(x_i))}_{\psi\text{-loss}}$$

$$h_S = \mathrm{pred} \circ \mathbf{f}_S$$

✓ For convex ψ and suitable F, computationally efficient!

? For suitable F, universally ψ-consistent in F;
suitable extensions can be made universally Bayes ψ-consistent

$\mathbf{L}$-Calibrated Surrogates

Theorem. If ψ is `$\mathbf{L}$-calibrated', then

Bayes ψ-consistency $\implies$ Bayes $\mathbf{L}$-consistency
(after applying some pred)

[Zhang, 2004; Tewari & Bartlett, 2007; Ramaswamy & Agarwal, 2012]

How do we design convex calibrated surrogates for a given loss matrix $\mathbf{L}$?

Recent Work on Convex Calibrated Surrogates for Specific Target Losses $\mathbf{L}$

Multiclass 0-1 Loss

Zhang, 2004; Tewari & Bartlett, 2007

Various Document (Subset) Ranking Losses

Cossock & Zhang, 2008; Xia et al, 2008; Duchi et al, 2010; Ravikumar et al, 2011; Buffoni et al, 2011; Lan et al, 2012; Calauzenes et al, 2012

Multilabel Losses

Gao & Zhou, 2011; Dembczynski et al, 2011

Our Work

Unified Approach for Designing
Convex Calibrated Surrogates for
General Loss Matrices

Our Work

When is a surrogate
calibrated for $\mathbf{L}$?

ch for Designing
Co Calibrated Surrogates for
General Loss Matrices

Our Work

When is a surrogate calibrated for $\mathbf{L}$?

What is the smallest dimension d that supports a convex calibrated surrogate for $\mathbf{L}$?

Our Work

When is a surrogate calibrated for $\mathbf{L}$?

What is the smallest dimension d that supports a convex calibrated surrogate for $\mathbf{L}$?

Can we design explicit low-dimensional surrogates for $\mathbf{L}$?

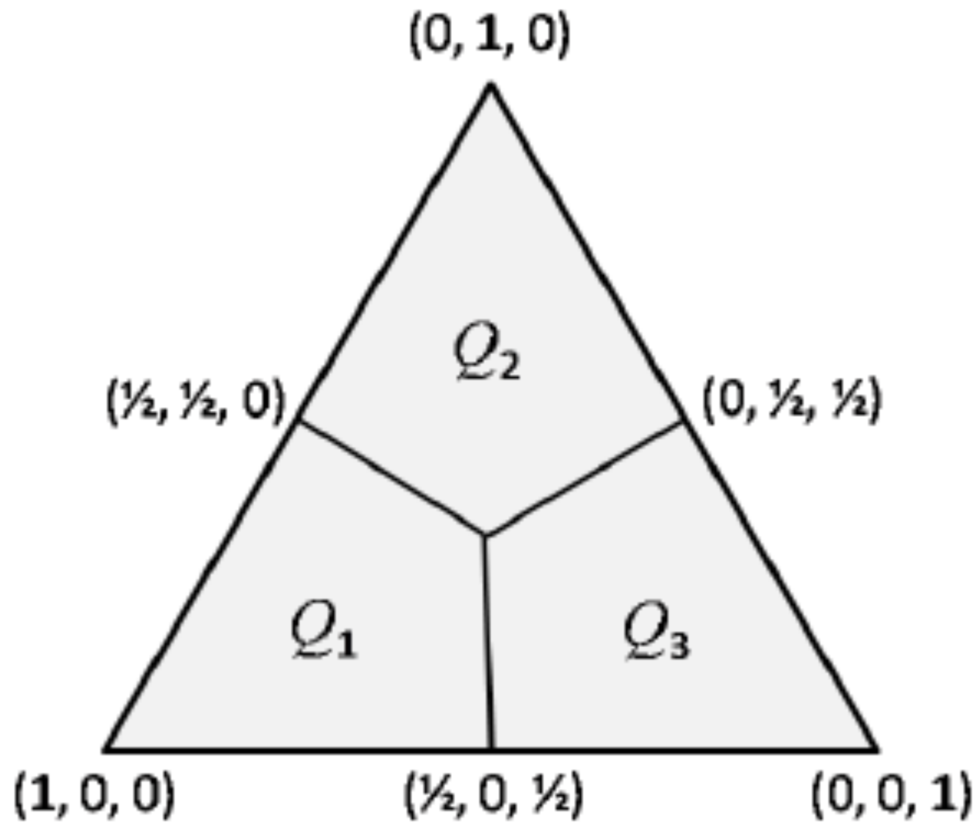
Our Work

When is a surrogate calibrated for $\mathbf{L}$?

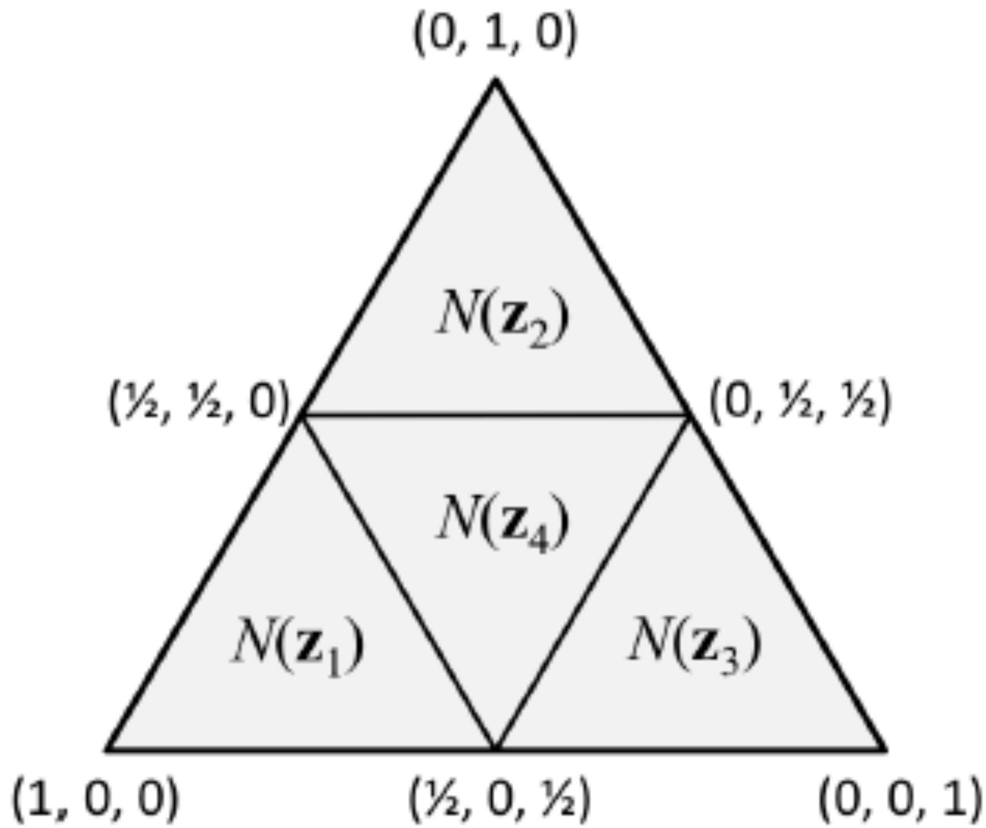
What is the smallest dimension d that supports a convex calibrated surrogate for $\mathbf{L}$?

Can we design explicit low-dimensional surrogates for $\mathbf{L}$?

Trigger Probability Sets of Loss $\mathbf{L}$



Positive Normal Sets of Surrogate ψ



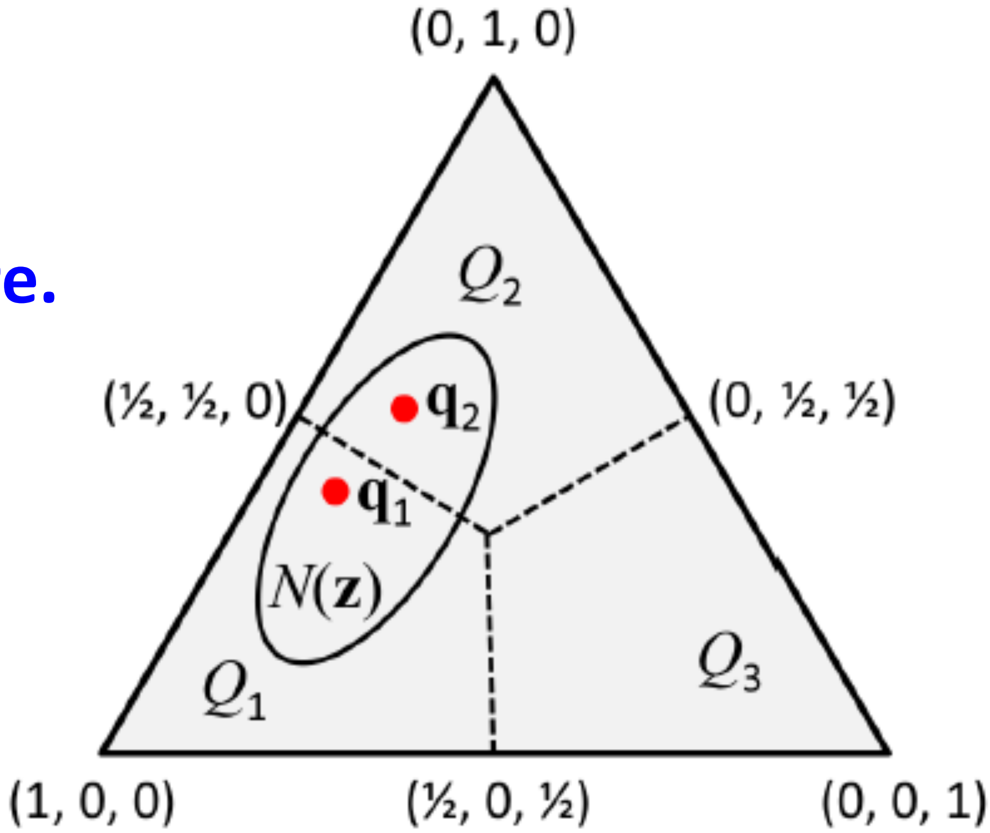
Necessary Condition for Calibration

Theorem. If ψ is $\mathbf{L}$-calibrated, then every positive normal set of ψ must be contained in some trigger probability set of $\mathbf{L}$.

[Ramaswamy & Agarwal, 2012; 2014]

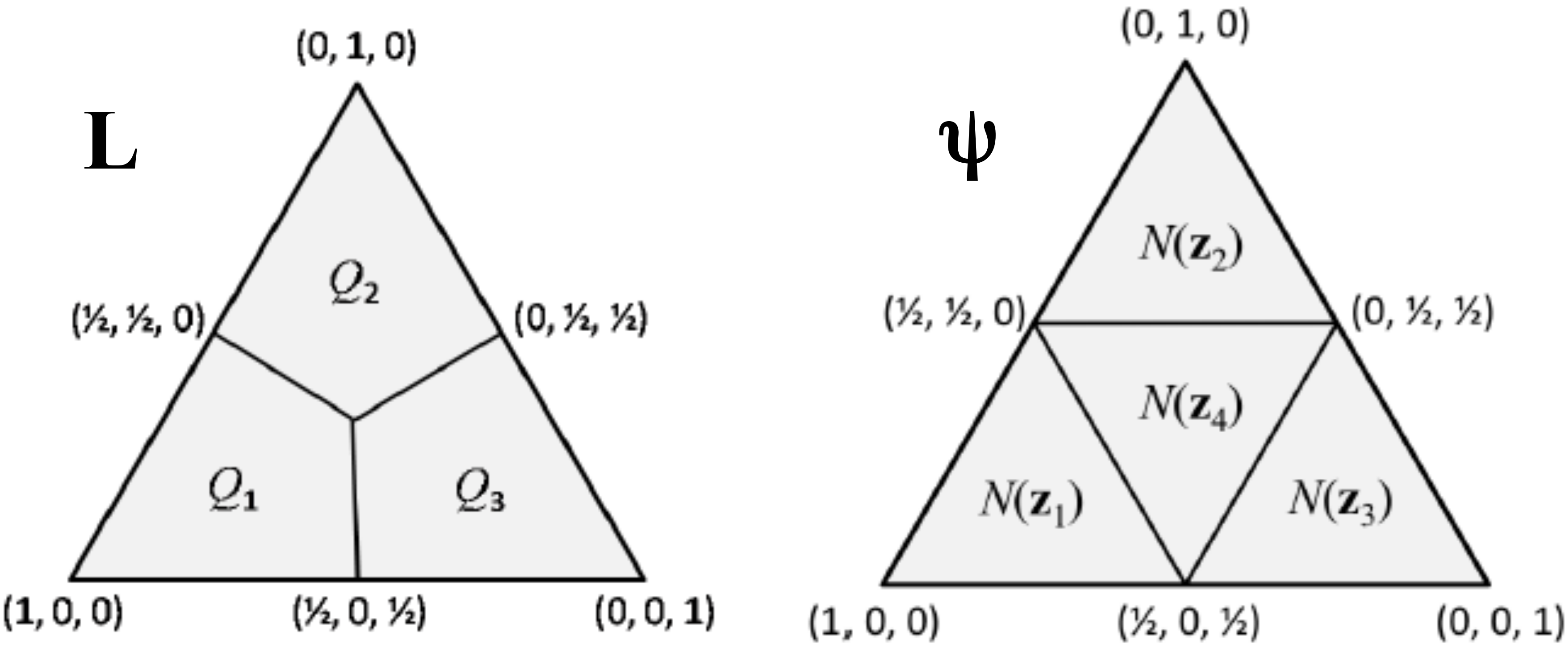
Necessary Condition for Calibration

Proof by picture.



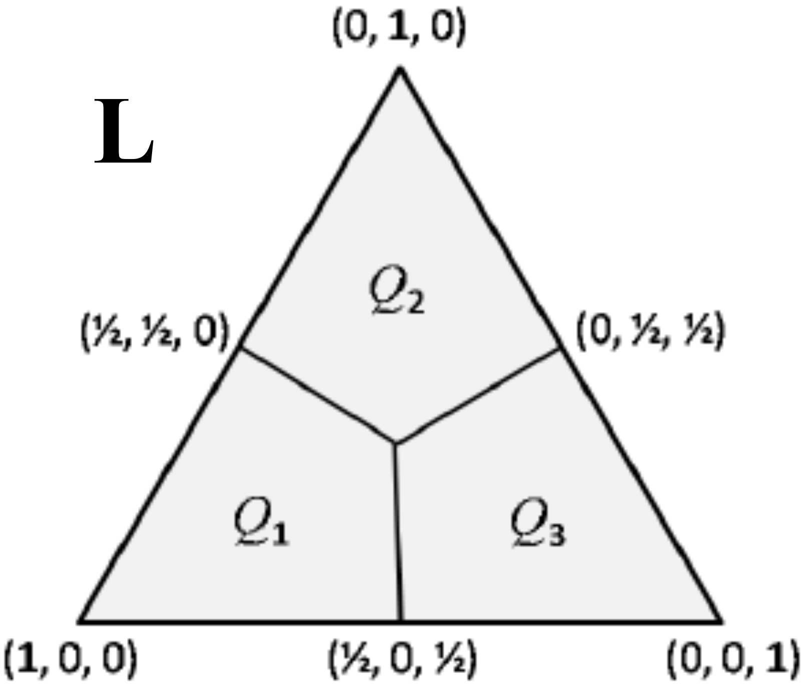
[Ramaswamy & Agarwal, 2012; 2014]

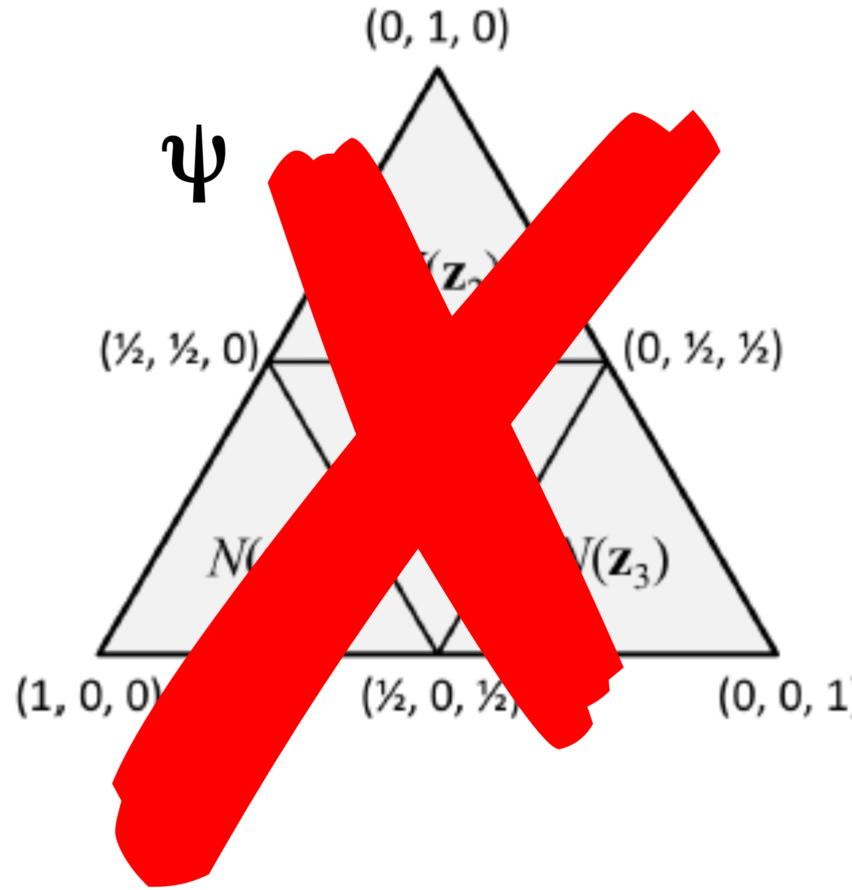
Example



[Ramaswamy & Agarwal, 2012; 2014]

Example





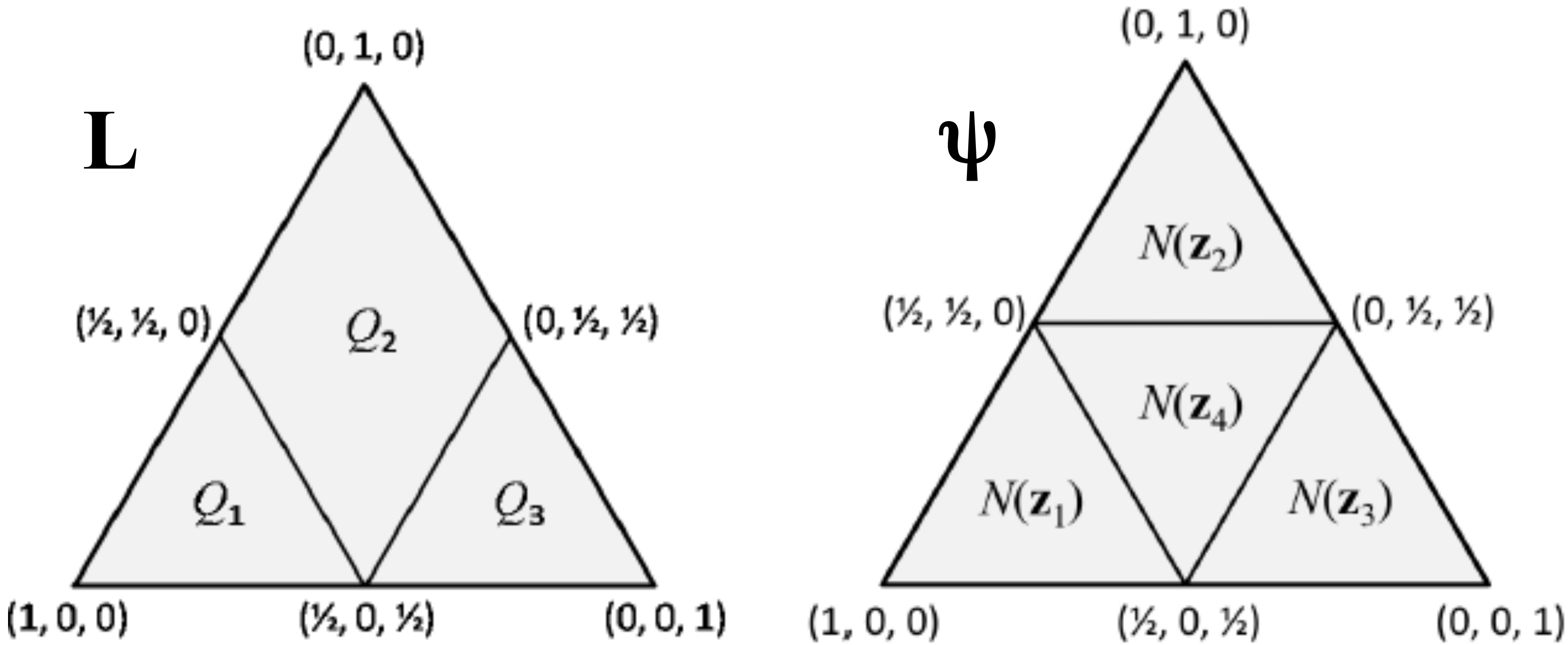
[Ramaswamy & Agarwal, 2012; 2014]

Sufficient Condition for Calibration

Theorem. If there is a finite collection of positive normal sets of ψ that are each contained in some trigger probability set of $\mathbf{L}$ and that collectively cover the simplex, then ψ is $\mathbf{L}$-calibrated.

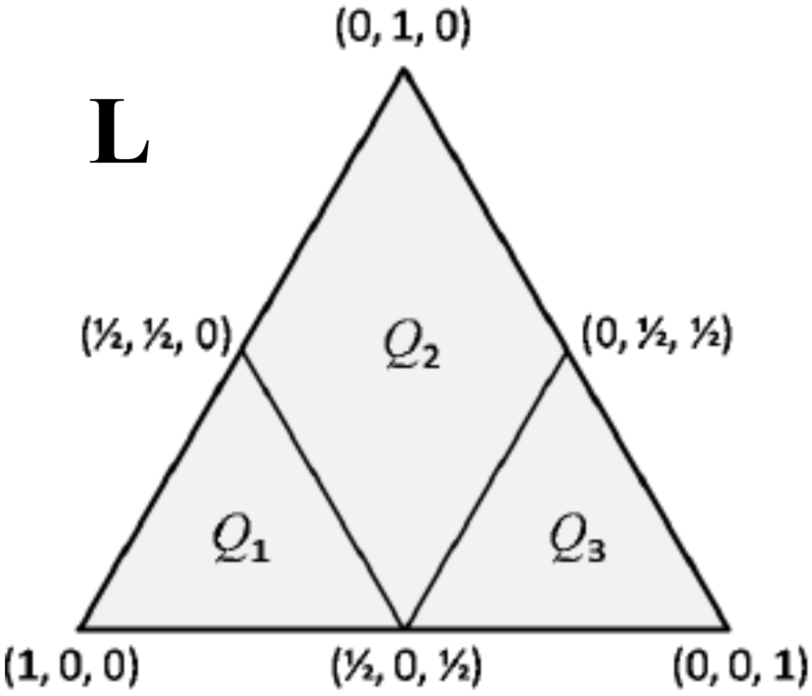
[Ramaswamy & Agarwal, 2012; 2014]

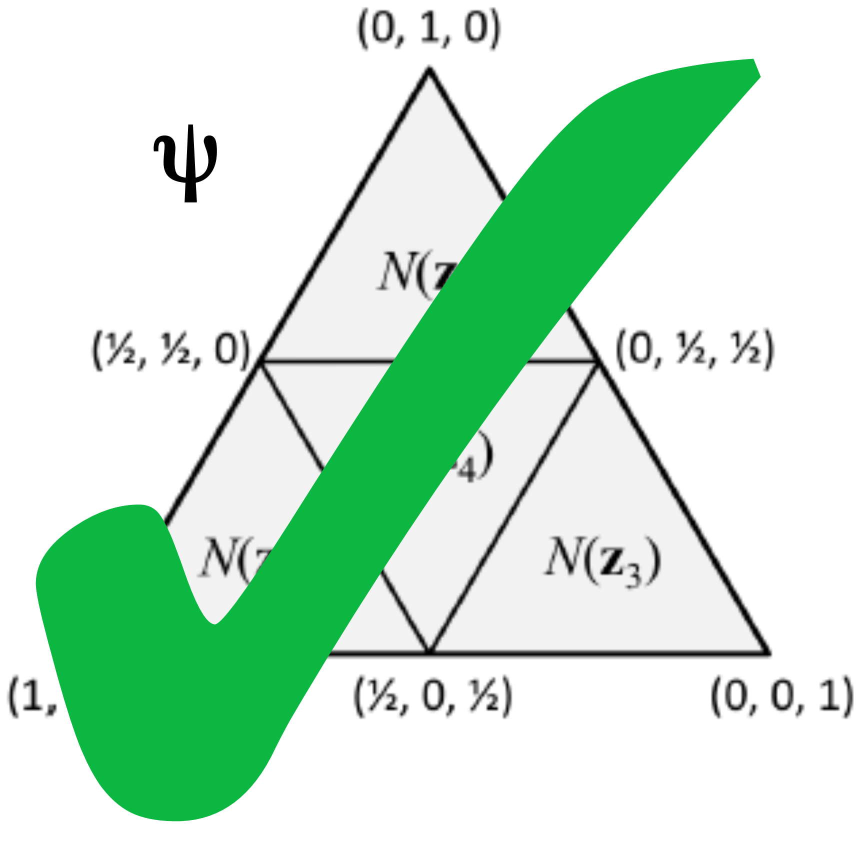
Example



[Ramaswamy & Agarwal, 2012; 2014]

Example





[Ramaswamy & Agarwal, 2012; 2014]

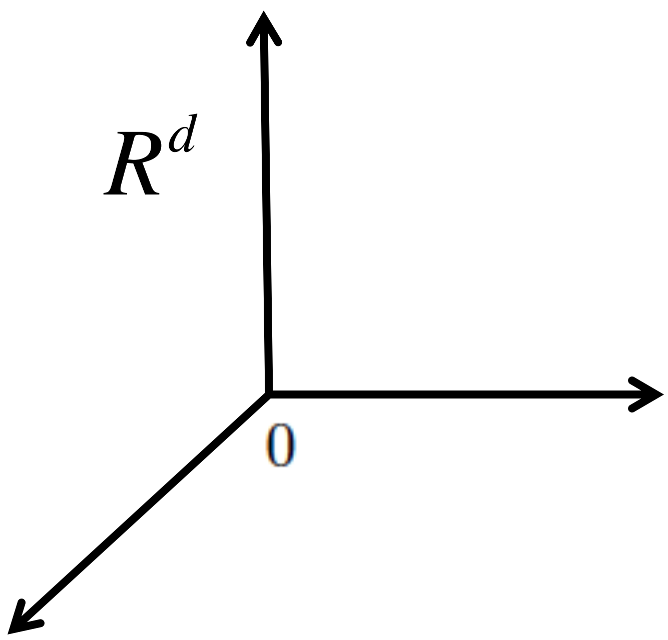
Our Work

When is a surrogate calibrated for $\mathbf{L}$?

What is the smallest dimension d that supports a convex calibrated surrogate for $\mathbf{L}$?

Can we design explicit low-dimensional surrogates for $\mathbf{L}$?

Convex Calibration Dimension

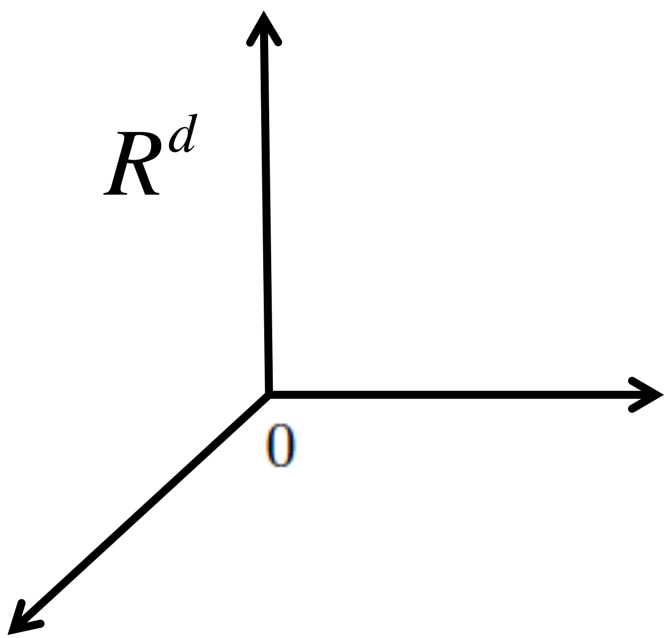


$\mathrm{CCdim}(\mathbf{L})$ = smallest dimension d that supports a convex calibrated surrogate for $\mathbf{L}$

[Ramaswamy & Agarwal, 2012; 2014]

Convex Calibration Dimension

CCdim($\mathbf{L}$) = smallest dimension d that supports a convex calibrated surrogate for $\mathbf{L}$



CCdim($\mathbf{L}$) $\leq$ n-1

[Ramaswamy & Agarwal, 2012; 2014]

Upper Bound on Convex Calibration Dimension

Theorem.

$$\mathrm{CCdim}(\mathbf{L}) \;\leq\; \mathrm{rank}(\mathbf{L})$$

[Ramaswamy & Agarwal, 2012; 2014]

Lower Bound on Convex Calibration Dimension

Theorem. For losses $\mathbf{L}$ whose columns can be obtained from one another by permuting entries,

$$\mathrm{CCdim}(\mathbf{L}) \geq \mathrm{rank}(\mathbf{L}) - 2$$

[Ramaswamy & Agarwal, 2012; 2014]

Example: Multiclass 0-1 Classification

$$Y = \hat{Y} = [n]$$

$$n = k > 2$$

$$\mathbf{L}^{0\text{-}1} = \begin{array}{c|ccccc|}
 & 1 & 2 & 3 & & n \\
\hline
1 & 0 & 1 & 1 & \cdots & 1 \\
2 & 1 & 0 & 1 & \cdots & 1 \\
3 & 1 & 1 & 0 & \cdots & 1 \\
 & \vdots & \vdots & \vdots & & \vdots \\
n & 1 & 1 & 1 & \cdots & 0
\end{array}$$

Example: Multiclass 0-1 Classification

$$Y = \hat{Y} = [n]$$

$$n = k > 2$$

$$\mathrm{rank}(\mathbf{L}^{0\text{-}1}) = n$$

$$\begin{array}{cc} n & \begin{bmatrix} 1 & 1 & 1 & \cdots & 0 \end{bmatrix} \end{array}$$

Example: Sequence Prediction with Hamming Loss

$$Y = \hat{Y} = \{0,1\}^r$$

$$n = k = 2^r$$

$r = 3$

$$\mathbf{L}^{\mathrm{Ham}} = \begin{array}{c} \\ 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \\ 1 & 0 & 2 & 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 2 & 2 & 1 \\ 1 & 2 & 2 & 3 & 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 2 & 1 & 0 & 2 & 1 \\ 2 & 3 & 1 & 2 & 1 & 2 & 0 & 1 \\ 3 & 2 & 2 & 1 & 2 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Example: Sequence Prediction with Hamming Loss

$Y = \hat{Y} = \{0,1\}^r$

$n = k = 2$

$$\mathrm{rank}(\mathbf{L}^{\mathrm{Ham}}) = r$$

$\mathbf{L}^{\mathrm{H}}$

							0	111
								3
							3	2
						3	1	2
						2	2	1
				3	0	1	1	2
101	2	1	3	2	1	0	2	1
110	2	3	1	2	1	2	0	1
111	3	2	2	1	2	1	1	0

Example: Document Ranking with Pairwise Disagreement Loss

$Y = \{0,1\}^r, \; \hat{Y} = S_r$

$n = 2^r, \; k = r!$

$r = 3$

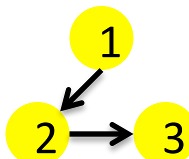
$$\mathbf{L}^{\mathrm{PD}} = \begin{array}{c} \\ \times\times\times \\ \times\times\checkmark \\ \times\checkmark\times \\ \times\checkmark\checkmark \\ \checkmark\times\times \\ \checkmark\times\checkmark \\ \checkmark\checkmark\times \\ \checkmark\checkmark\checkmark \end{array} \begin{array}{c} \begin{array}{cccccc} \begin{pmatrix}1\\2\\3\end{pmatrix} & \begin{pmatrix}1\\3\\2\end{pmatrix} & \begin{pmatrix}2\\3\\1\end{pmatrix} & \begin{pmatrix}2\\1\\3\end{pmatrix} & \begin{pmatrix}3\\1\\2\end{pmatrix} & \begin{pmatrix}3\\2\\1\end{pmatrix} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 2 & 0 & 0 \\ 1 & 2 & 0 & 0 & 2 & 1 \\ 2 & 2 & 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 0 & 2 & 2 & 0 & 1 \\ 0 & 1 & 1 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

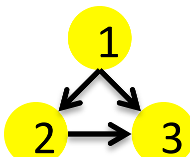
Example: Document Ranking with Pairwise Disagreement Loss

$$Y = G_r,\ \hat{Y} = S_r$$

$$n = |G_r|,\ k = r!$$

$$r = 3$$





$$\mathbf{L}^{\mathrm{PD}} = \begin{array}{cccccc} \begin{pmatrix}1\\2\\3\end{pmatrix} & \begin{pmatrix}1\\3\\2\end{pmatrix} & \begin{pmatrix}2\\3\\1\end{pmatrix} & \begin{pmatrix}2\\1\\3\end{pmatrix} & \begin{pmatrix}3\\1\\2\end{pmatrix} & \begin{pmatrix}3\\2\\1\end{pmatrix} \\ \end{array}$$

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 2 \\ & & & \vdots & & \\ 0 & 1 & 2 & 1 & 2 & 3 \end{bmatrix}$$

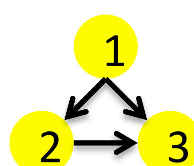
Example: Document Ranking with Pairwise Disagreement Loss

$Y = G_r, \ \hat{Y} = S_r$

$n = |G_r|$

$$\mathrm{rank}(\mathbf{L}^{\mathrm{PD}}) = \Theta(r^2)$$

$$\mathbf{L}^{\mathrm{PD}} = \begin{bmatrix} & \vdots & & & \vdots & \\ 0 & 1 & 2 & 1 & 2 & 3 \end{bmatrix}$$



Application: Stronger Versions of Recent Results on Non-Existence of Convex Calibrated Surrogates

(Duchi et al, 2010; Calauzenes et al, 2012): no convex calibrated surrogates for $\mathbf{L}^{\mathrm{PD}}$ in $\leq r$ dimensions

Application: Stronger Versions of Recent Results on Non-Existence of Convex Calibrated Surrogates

(Duchi et al, 2010; Calauzenes et al, 2012): no convex calibrated surrogates for $\mathbf{L}^{\mathrm{PD}}$ in $\leq r$ dimensions

Our results: no convex calibrated surrogates for $\mathbf{L}^{\mathrm{PD}}$ in $< r(r\text{-}1)/2 - 2$ dimensions!

Our Work

When is a surrogate calibrated for $\mathbf{L}$?

What is the smallest dimension d that supports a convex calibrated surrogate for $\mathbf{L}$?

Can we design explicit low-dimensional surrogates for $\mathbf{L}$?

Explicit Convex Calibrated Least Squares Surrogate with $d = \mathrm{rank}(\mathbf{L})$

$$\Big[\quad \mathbf{L} \quad \Big]_{n\times k} = \Big[\mathbf{A} \Big]_{n\times d} \Big[\quad \mathbf{B} \quad \Big]_{d\times k}$$

[Ramaswamy, Agarwal & Tewari, 2013]

Explicit Convex Calibrated Least Squares Surrogate with $d = \mathrm{rank}(\mathbf{L})$

$$\psi^{\mathrm{LS}}(y, \mathbf{u}) = \sum_{j=1}^{d} (u_j - A_{yj})^2$$

$$\mathrm{pred}^{\mathrm{LS}}(\mathbf{u}) \in \underset{\hat{y} \in [k]}{\mathrm{argmin}} \sum_{j=1}^{d} u_j B_{j\hat{y}}$$

[Ramaswamy, Agarwal & Tewari, 2013]

Explicit Convex Calibrated Output Code Based Surrogate with d = rank($\mathbf{L}$)

$$\Big[\quad \mathbf{L} \quad\Big]_{n\times k} = \Big[\mathbf{A}\Big]_{n\times d}\Big[\quad \mathbf{B} \quad\Big]_{d\times k}$$

[Ramaswamy, Babu, Agarwal & Williamson, 2014]

Explicit Convex Calibrated Output Code Based Surrogate with $d = \mathrm{rank}(\mathbf{L})$

$$\psi^{\mathrm{OC}}(y,\mathbf{u}) = \sum_{j=1}^{d} \left(C_{yj}\phi(1,u_j) + (1-C_{yj})\phi(-1,u_j) \right)$$

$$\mathrm{pred}^{\mathrm{OC}}(\mathbf{u}) \in \operatorname*{argmin}_{\hat{y}\in[k]} \sum_{j=1}^{d} \lambda^{-1}(u_j)\beta_{j\hat{y}}$$

[Ramaswamy, Babu, Agarwal & Williamson, 2014]

Explicit Convex Calibrated Output Code Based Surrogate with $d = \mathrm{rank}(\mathbf{L})$

$$\psi^{\mathrm{OC}}(y, \mathbf{u}) = \sum_{j=1}^{d} \left(C_{yj}\,\phi(1, u_j) + (1 - C_{yj})\,\phi(-1, u_j) \right)$$

Strictly proper composite
binary surrogate

[Ramaswamy, Babu, Agarwal & Williamson, 2014]

Explicit Convex Calibrated Output Code Based Surrogate with $d = \mathrm{rank}(\mathbf{L})$

$$\psi^{\mathrm{OC}}(y,\mathbf{u}) = \sum_{j=1}^{d} \left(C_{yj}\,\phi(1,u_j) + (1 - C_{yj})\,\phi(-1,u_j) \right)$$

Code matrix
constructed from $\mathbf{A}$

[Ramaswamy, Babu, Agarwal & Williamson, 2014]

Explicit Convex Calibrated Output Code Based Surrogate with $d = \mathrm{rank}(\mathbf{L})$

Inverse link
associated with ϕ

$$\mathrm{pred}^{\mathrm{OC}}(\mathbf{u}) \in \operatorname*{argmin}_{\hat{y}\in[k]} \sum_{j=1}^{d} \lambda^{-1}(u_j)\beta_{j\hat{y}}$$

[Ramaswamy, Babu, Agarwal & Williamson, 2014]

Explicit Convex Calibrated Output Code Based Surrogate with $d = \mathrm{rank}(\mathbf{L})$

Obtained from $\mathbf{B}$

$$\mathrm{pred}^{\mathrm{OC}}(\mathbf{u}) \in \underset{\hat{y}\in[k]}{\mathrm{argmin}} \sum_{j=1}^{d} \lambda^{-1}(u_j)\beta_{j\hat{y}}$$

[Ramaswamy, Babu, Agarwal & Williamson, 2014]

Current/Future Directions

Relaxing universal consistency requirement

Relaxing exact consistency requirement

Complex performance measures

Current/Future Directions

Relaxing universal consistency requirement

Relaxing exact consistency requirement

Complex performance measures

Surrogates for Multiclass 0-1 Classification

Popular Crammer-Singer surrogate:

$$\psi^{\mathrm{CS}}(y,\mathbf{u}) = \max_{\hat{y}\in[k]}(1-(u_y - u_{\hat{y}}))_+$$

$$\mathrm{pred}^{\mathrm{CS}}(\mathbf{u}) \in \operatorname*{arg\,max}_{\hat{y}\in[k]} u_{\hat{y}}$$

n-dimensional surrogate

[Crammer & Singer, 2001]

Surrogates for Multiclass 0-1 Classification

Crammer-Singer algorithm is 0-1 consistent only for distributions D satisfying \`dominant-label' condition

[Zhang, 2004; Tewari & Bartlett, 2007]

Surrogates for Multiclass 0-1 Classification

Under dominant-label condition, $\log n$ dimensions suffice!

[Ramaswamy, Tewari & Agarwal, 2015]

Surrogates for Document Ranking with Pairwise Disagreement Loss

Least Squares surrogate:

$\Theta(r^2)$-dimensional surrogate

$$\psi^{\mathrm{LS}}(y,\mathbf{u}) = \sum_{i=1}^{r}\sum_{j\neq i}(u_{ij} - y_{ij})^2$$

$$\mathrm{pred}^{\mathrm{LS}}(\mathbf{u}) \in \operatorname*{argmin}_{\sigma\in S_r} \sum_{i=1}^{r}\sum_{j\neq i} u_{ij}\mathbf{1}(\sigma(i) > \sigma(j))$$

[Ramaswamy, Agarwal & Tewari, 2013]

Surrogates for Document Ranking with Pairwise Disagreement Loss

Resulting algorithm is universally consistent, BUT…

[Ramaswamy, Agarwal & Tewari, 2013]

Surrogates for Document Ranking with Pairwise Disagreement Loss

Least Squares surrogate:

$$\psi^{\mathrm{LS}}(y,\mathbf{u}) = \sum_{i=1}^{r}\sum_{j\neq i}(u_{ij} - y_{ij})^2$$

$$\mathrm{pred}^{\mathrm{LS}}(\mathbf{u}) \in \operatorname*{argmin}_{\sigma\in S_r} \sum_{i=1}^{r}\sum_{j\neq i} u_{ij}\mathbf{1}(\sigma(i) > \sigma(j))$$

MWFAS !!!

[Ramaswamy, Agarwal & Tewari, 2013]

Surrogates for Document Ranking with Pairwise Disagreement Loss

Efficient implementation
under `DAG' condition

[Ramaswamy, Agarwal & Tewari, 2013]

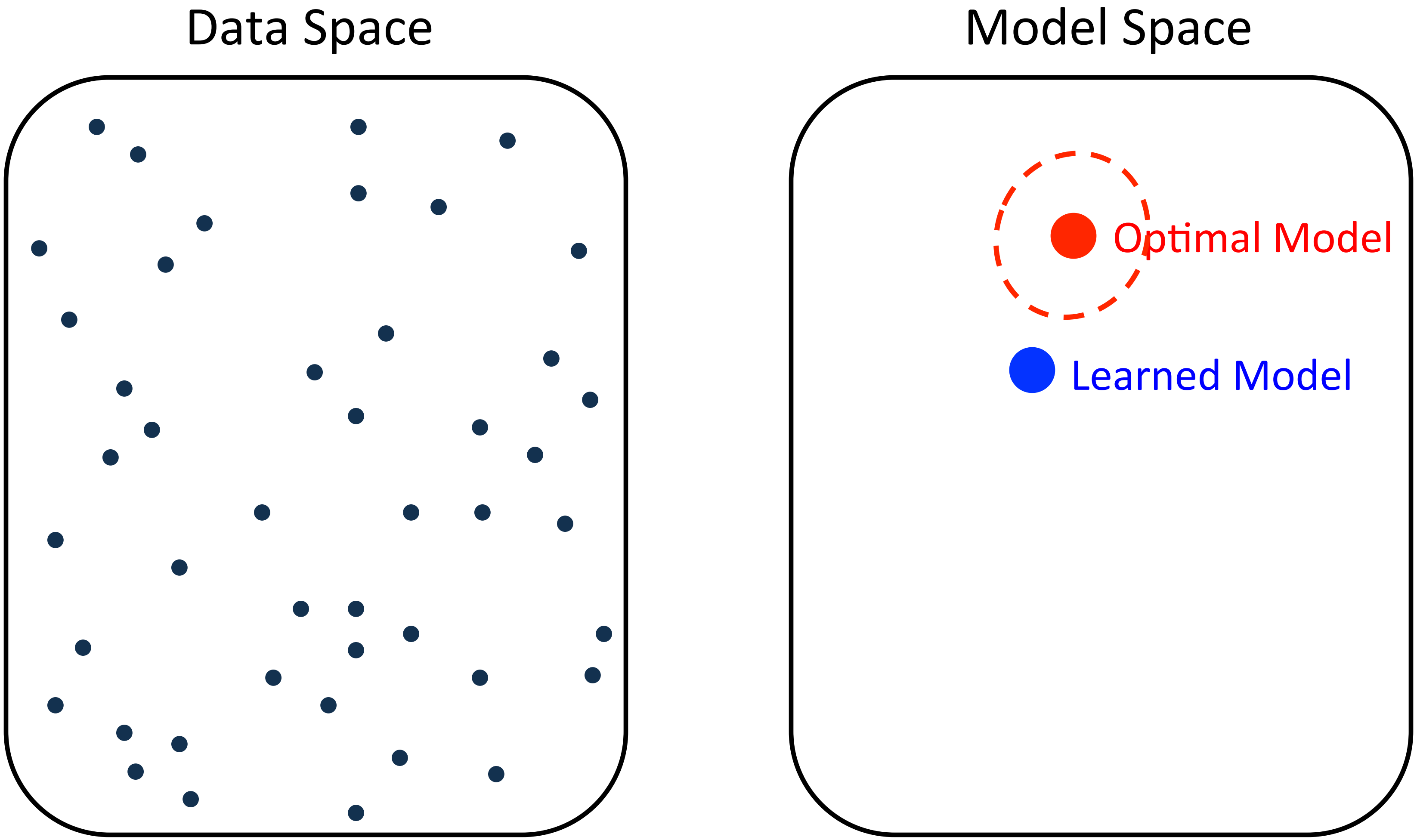
Current/Future Directions

Relaxing universal consistency requirement

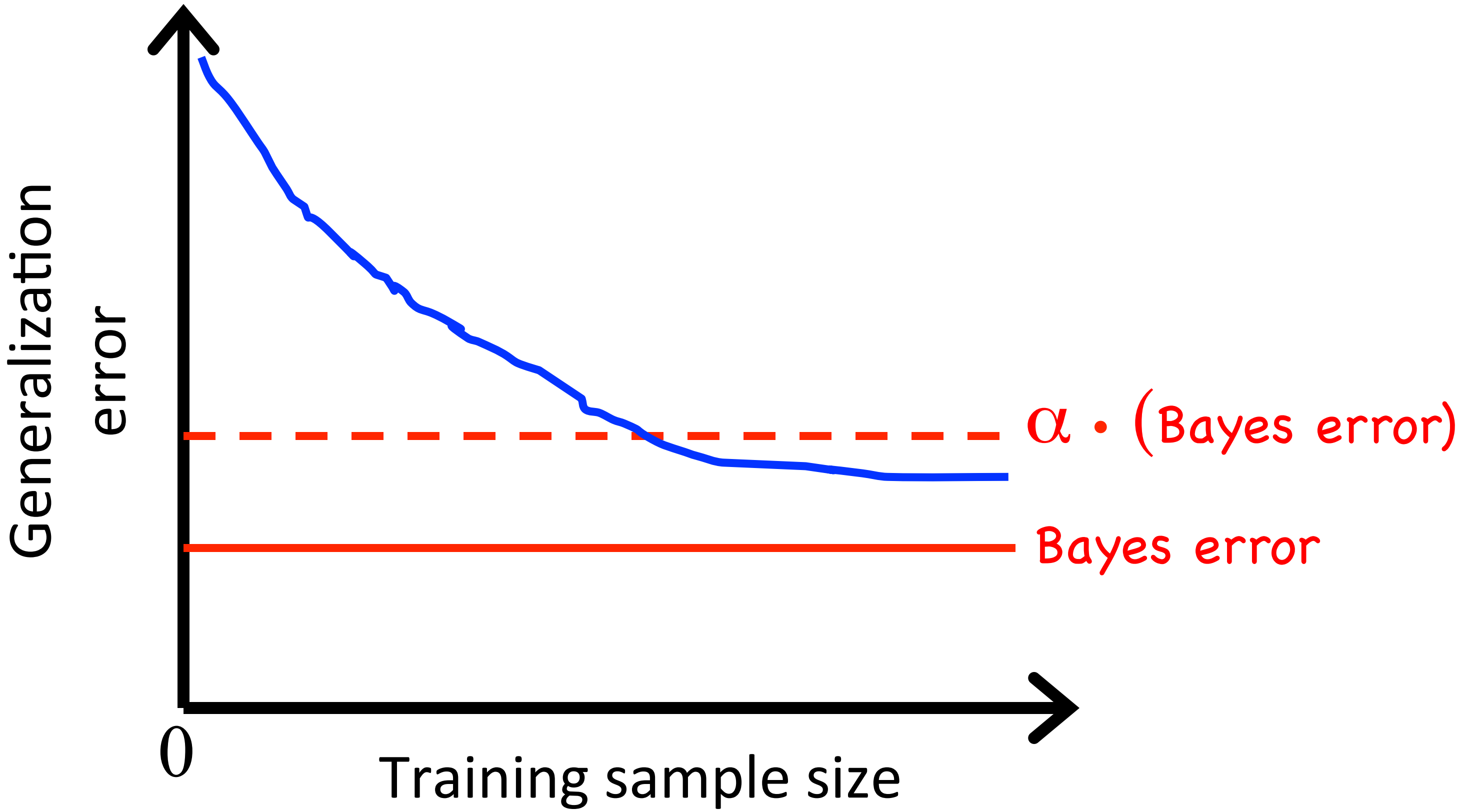
Relaxing exact consistency requirement

Complex performance measures

Approximate Consistency



Approximate Consistency via Approximately Calibrated Surrogates



[Ramaswamy & Agarwal, in progress]

Current/Future Directions

Relaxing universal consistency requirement

Relaxing exact consistency requirement

Complex performance measures

Example: F-Measure

$$Y = \hat{Y} = \{\pm 1\}$$

$$F_D[h] = \frac{2 \cdot \mathrm{Prec}_D[h] \cdot \mathrm{Rec}_D[h]}{\mathrm{Prec}_D[h] + \mathrm{Rec}_D[h]}$$

[Narasimhan, Vaish & Agarwal, 2014;
Narasimhan, Ramaswamy & Agarwal, 2015]

Summary

Given a learning problem defined by a loss matrix $\mathbf{L}$, need $\mathbf{L}$-consistent learning algorithms

Summary

Given a learning problem defined by a loss matrix $\mathbf{L}$, need $\mathbf{L}$-consistent learning algorithms

Many popular learning algorithms minimize a convex surrogate loss

Summary

Given a learning problem defined by a loss matrix $\mathbf{L}$, need $\mathbf{L}$-consistent learning algorithms

Many popular learning algorithms minimize a convex surrogate loss

Our work:
General methodology for designing convex calibrated surrogates for any loss $\mathbf{L}$

Open Questions

General characterization of convex calibration dimension of a loss $\mathbf{L}$?

Open Questions

General characterization of convex calibration dimension of a loss $\mathbf{L}$?

Convex calibrated surrogates with $d = \mathrm{CCdim}(\mathbf{L})$?

Open Questions

General characterization of convex calibration dimension of a loss $\mathbf{L}$?

Convex calibrated surrogates with $d = \mathrm{CCdim}(\mathbf{L})$?

Other structure in loss matrices?

Acknowledgments

Harikrishna Narasimhan

Rohit Vaish

Balaji S. Babu

Harish G. Ramaswamy

Robert C. Williamson

Ambuj Tewari

सत्यमेव जयते

Department of
Science & Technology
Government of India

IUSSTF

Indo-US Science &
Technology Forum