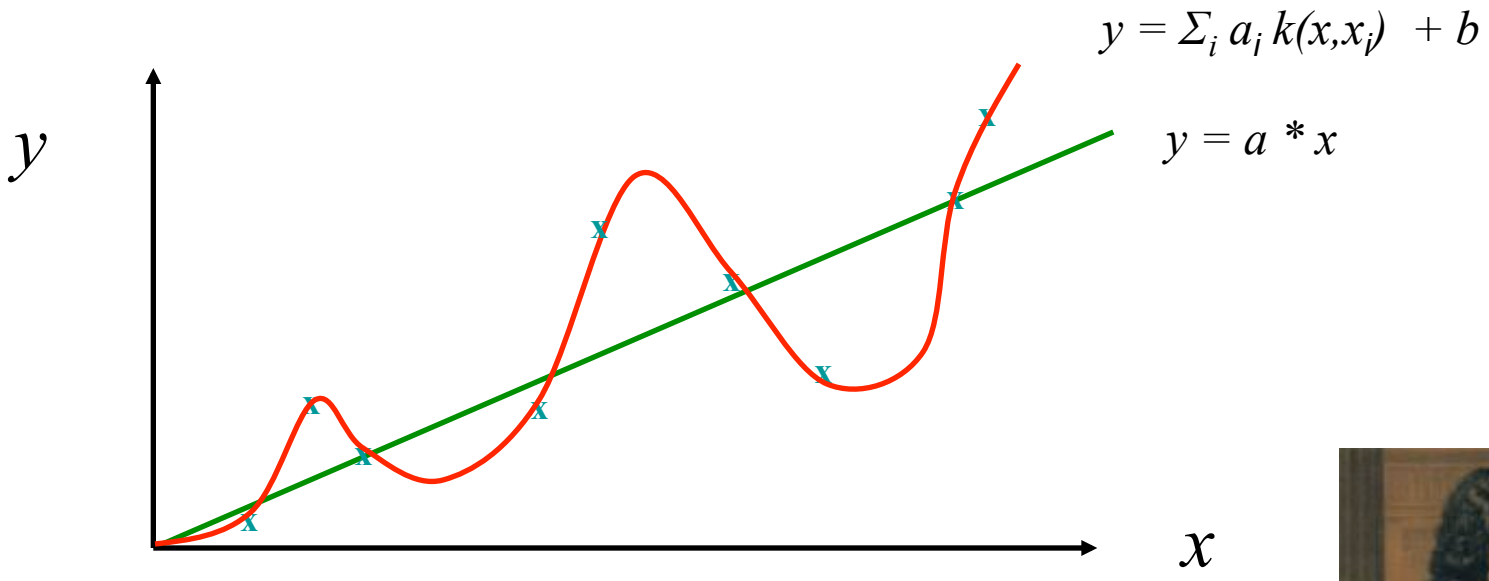
Statistical Learning and Kernel Methods

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Tübingen, Germany

http://www.tuebingen.mpg.de/~bs

Empirical Inference

- Drawing conclusions from empirical data (observations, measurements)
- Example 1: scientific inference



Leibniz, Weyl, Chaitin

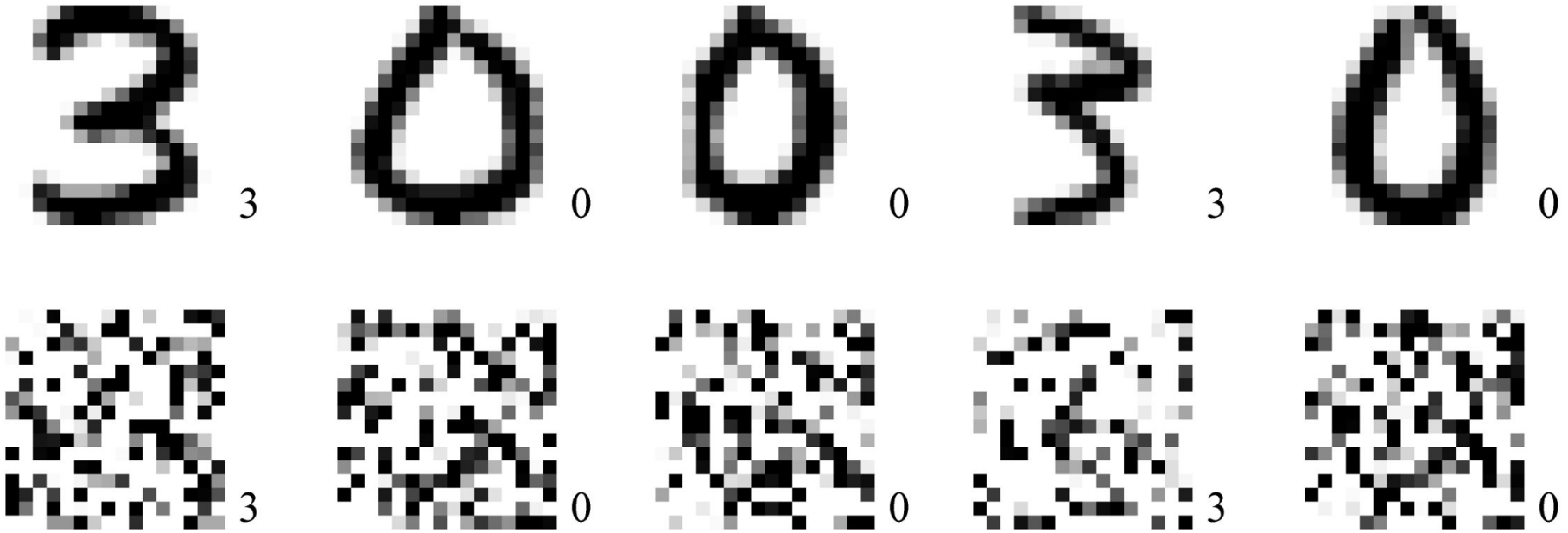
Empirical Inference

- Drawing conclusions from empirical data (observations, measurements)
- Example 1: scientific inference

"If your experiment needs statistics [inference]*,*
you ought to have done a better experiment." (Rutherford)

Empirical Inference, II

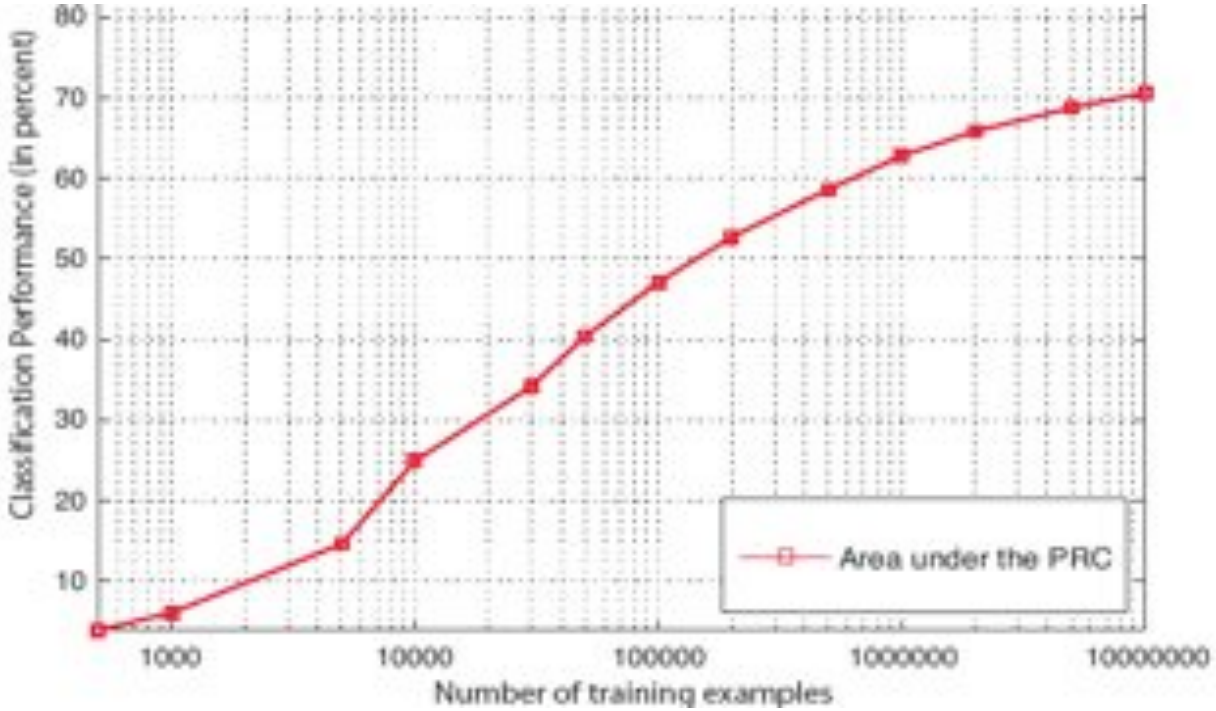
- Example 2: perception



"The brain is nothing but a statistical decision organ"

(H. Barlow)

Hard Inference Problems



Sonnenburg, Rätsch, Schäfer, Schölkopf, 2006, Journal of Machine Learning Research

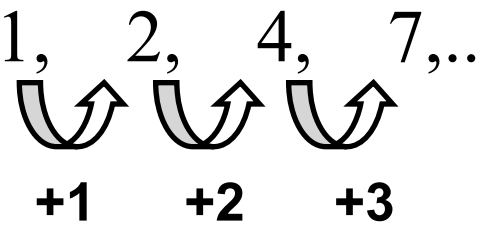
Task: classify human DNA sequence locations into {acceptor splice site, decoy} using 15 Million sequences of length 141, and a Multiple-Kernel Support Vector Machines.

PRC = Precision-Recall-Curve, fraction of correct positive predictions among all positively predicted cases

- High dimensionality – *consider many factors simultaneously to find the regularity*
- Complex regularities – *nonlinear, nonstationary, etc.*
- Little prior knowledge – *e.g., no mechanistic models for the data*
- Need large data sets – *processing requires computers and automatic inference methods*

Generalization

- *observe*
- What's next?



- 1,2,4,7,11,16,…: $a_{n+1}=a_n+n$ ("lazy caterer's sequence")
- 1,2,4,7,12,20,…: $a_{n+2}=a_{n+1}+a_n+1$
- 1,2,4,7,13,24,…: "Tribonacci"-sequence
- 1,2,4,7,14,28: divisors of 28
- 1,2,4,7,1,1,5,…: decimal expansions of π=3,14159… and e=2,718… interleaved *(thanks to O. Bousquet)*
- **The On-Line Encyclopedia of Integer Sequences**: >600 hits…

Generalization, II

- Question: which continuation is correct ("generalizes")?
- *Answer:* there's no way to tell *("induction problem")*

- Question of statistical learning theory: how to come up with a law that generalizes *("demarcation problem")*

  [i.e.: a law that will probably do almost as well in the future as it has done in the past]

Learning problem

Data $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}$

$(x_i, y_i) \sim \mathrm{P}(x, y).$

Goal: find $f : \mathcal{X} \to \mathcal{Y}$

with minimal "risk" $R[f] = \int_{\mathcal{X}} l(f(x), y)\, d\mathrm{P}(x, y)$

Cost function $l(f(x), y)$

Special case $\mathcal{Y} = \{\pm 1\}, l(f(x), z) = \frac{1}{2}|f(x) - y|$:

"2-class-classification"

2-class classification *(Vapnik & Chervonenkis)*

Learn $f : \mathcal{X} \to \{\pm 1\}$ based on *m* observations

$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}$ generated i.i.d. from some $\mathrm{P}(x, y)$

Goal: **minimize expected error** ("risk")

$$R[f] = \int \frac{1}{2} |f(x) - y| \, dP(x, y)$$

V. Vapnik

Problem: P is unknown.

Induction principle: **minimize training error** ("empirical risk")

$$R_{emp}[f] = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} |f(x_i) - y_i|$$

over some class of functions. Q: is this "consistent"?

The law of large numbers

For all $f \in \mathcal{F}$ and $\epsilon > 0$

$$\lim_{m \to \infty} \mathrm{P}\{|R[f] - R_{\mathrm{emp}}[f]| > \epsilon\} = 0$$

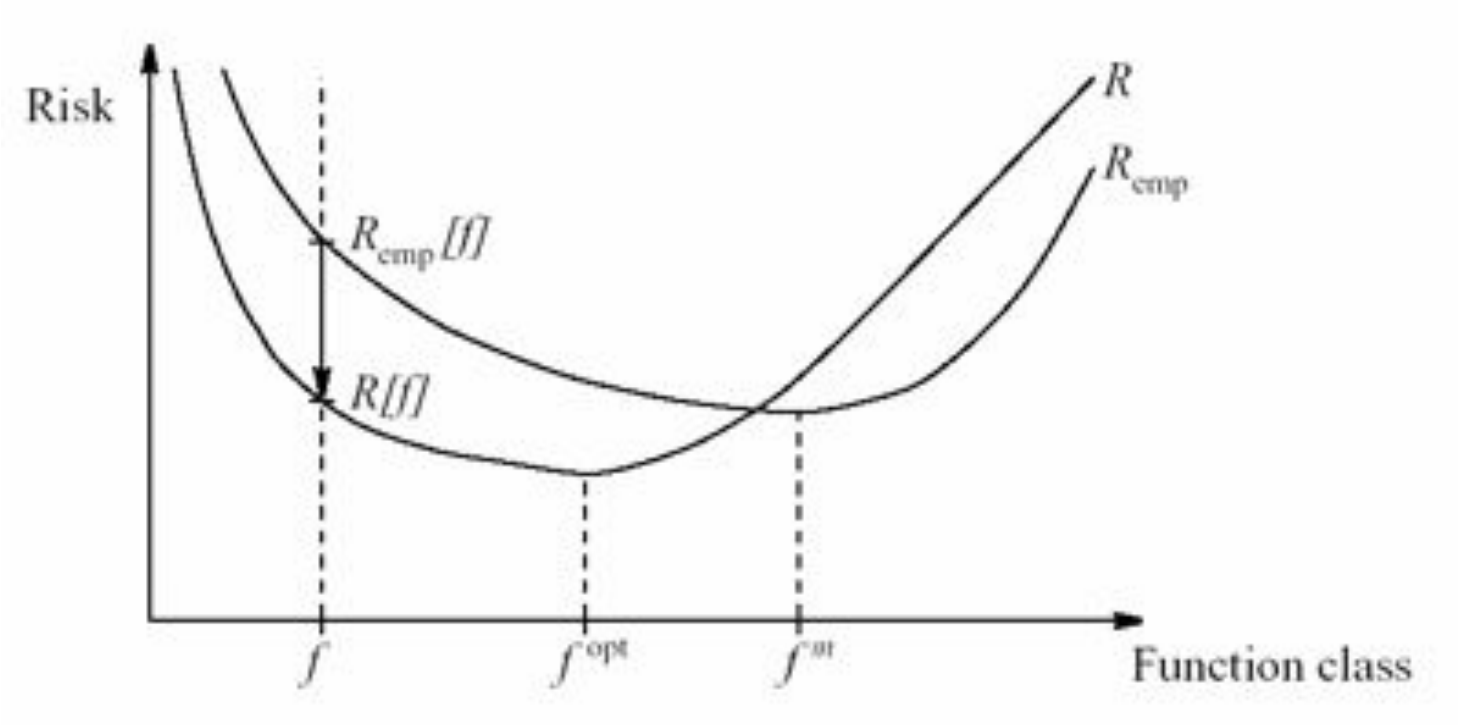
Does this imply "consistency" of empirical risk minimization (optimality in the limit)?

No – need a uniform law of large numbers:

For all $\epsilon > 0$

$$\lim_{m \to \infty} \mathrm{P}\{\sup_{f \in \mathcal{F}} (R[f] - R_{\mathrm{emp}}[f]) > \epsilon\} = 0$$

Consistency and uniform convergence



The Importance of the Set of Functions

What about allowing *all* functions from $\mathcal{X}$ to $\{\pm 1\}$?

Training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathcal{X} \times \{\pm 1\}$
Test patterns $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{\bar{m}} \in \mathcal{X}$,
such that $\{\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{\bar{m}}\} \cap \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} = \{\}$.

For any f there exists f^* s.t.:

1. $f^*(\mathbf{x}_i) = f(\mathbf{x}_i)$ for all i
2. $f^*(\bar{\mathbf{x}}_j) \neq f(\bar{\mathbf{x}}_j)$ for all j.

Based on the training set alone, there is *no* means of choosing which one is better. On the test set, however, they give *opposite* results. There is 'no free lunch' [25, 60].

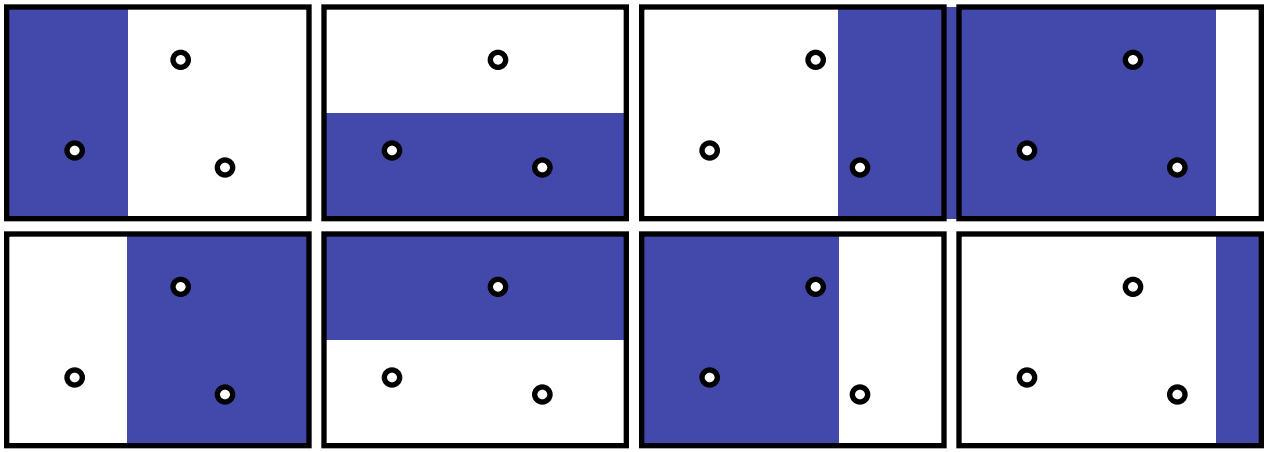
$\longrightarrow$ a restriction must be placed on the *functions* that we allow

Vapnik-Chervonenkis (VC) dimension

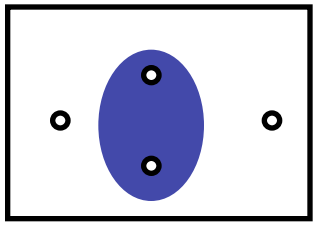
ERM is consistent for all probability distributions, provided that the ***VC dimension*** of the function class is finite.

VC dimension h = maximal number of points which can be classified in all possible ways using functions from the class.

Linear classifiers on R^2: $h = 3$:



non-falsifiable



falsifiable

… on R^d: $h = d + 1$

… with margin of separation: $h < const./\text{margin}^2$

VC Risk bound

$$R[f] \leq R_{\text{emp}}[f] + \sqrt{\frac{h\left(\log\frac{2m}{h}+1\right) - \log(\delta/4)}{m}}$$

with probab. 1-δ for all f

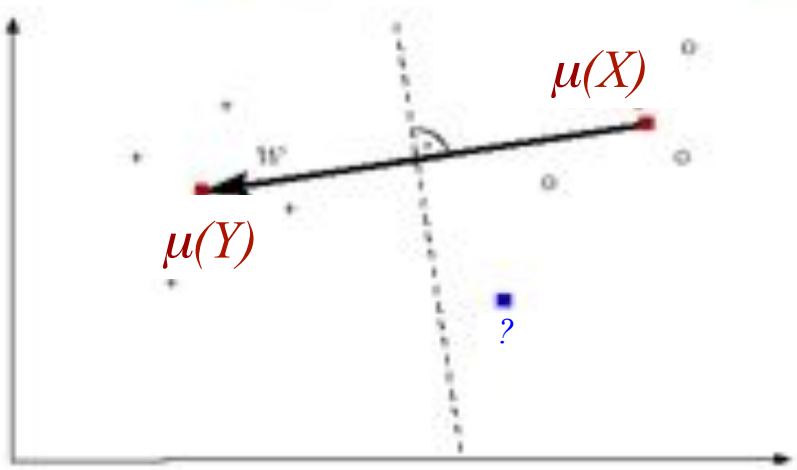
If both training error R_{emp} and VC dimension h (compared to the number of observations m) are small, then test error R is small.

-> depends on the function class.

Example of a Pattern Recognition Algorithm

Idea:
classify points according to which of the two **class means** is closer:

$$\mu(X) = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \mu(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i$$



- Decision function: hyperplane with normal vector $w := \mu(X)-\mu(Y)$
- How about problems that are not linearly separable?

Feature Spaces

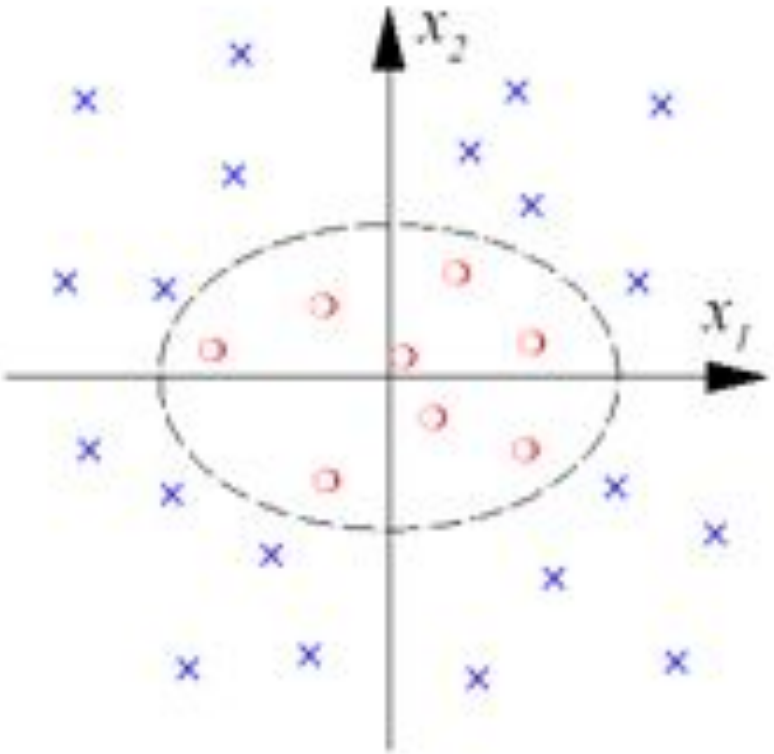
Preprocess the inputs with

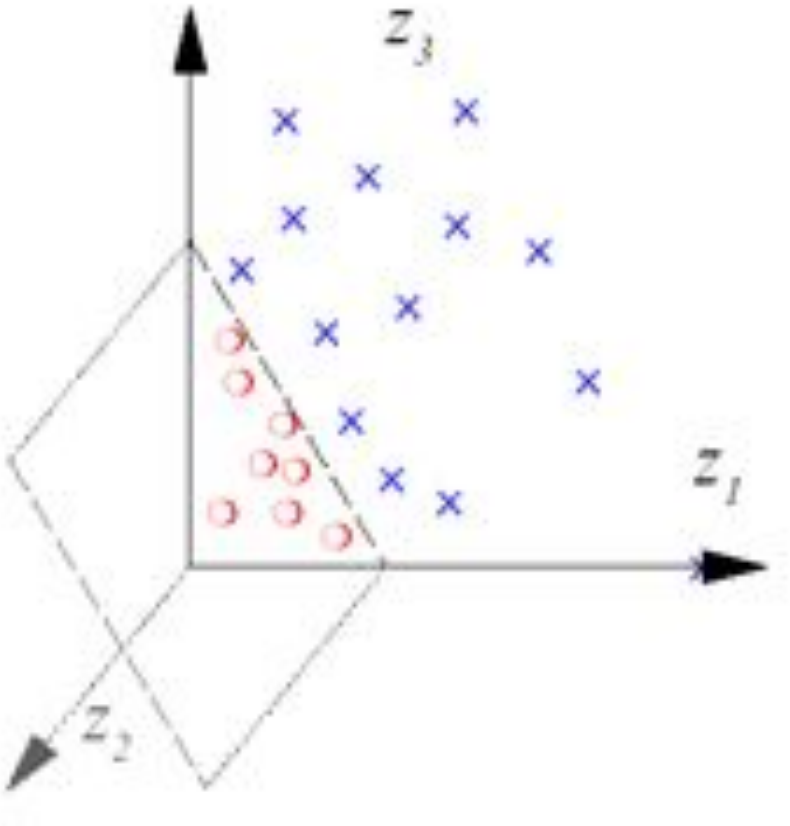
$$\Phi : \mathcal{X} \to \mathcal{H}$$
$$x \mapsto \Phi(x),$$

where $\mathcal{H}$ is a dot product space, and learn the mapping from $\Phi(x)$ to y.

Example: All Degree 2 Monomials

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$





Polynomial Kernels

$$\begin{aligned}\langle \Phi(x), \Phi(x') \rangle &= (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)(x'^2_1, \sqrt{2}\, x'_1 x'_2, x'^2_2)^\top \\ &= (x_1 x'_1 + x_2 x'_2)^2 \\ &= \langle x, x' \rangle^2 \\ &= : k(x, x')\end{aligned}$$

$\longrightarrow$ the dot product in $\mathcal{H}$ can be computed from the dot product in $\mathbb{R}^2$

More generally: for $x, x' \in \mathbb{R}^N$, $d \in \mathbb{N}$,

$$\langle x, x' \rangle^d = \left(\sum_{j=1}^{N} x_j \cdot x'_j \right)^d = \sum_{j_1, \ldots, j_d = 1}^{N} x_{j_1} \cdot \cdots \cdot x_{j_d} \cdot x'_{j_1} \cdot \cdots \cdot x'_{j_d} = \langle \Phi(x), \Phi(x') \rangle$$

Positive Definite Kernels

Let $\mathcal{X}$ be a nonempty set. The following two are equivalent:

- k is *positive definite (pd)*, i.e., k is symmetric, and for
  - any set of training points $x_1, \ldots, x_m \in \mathcal{X}$ and
  - any $a_1, \ldots, a_m \in \mathbb{R}$

  we have

$$\sum_{i,j} a_i a_j K_{ij} \geq 0, \quad \text{where } K_{ij} := k(x_i, x_j)$$

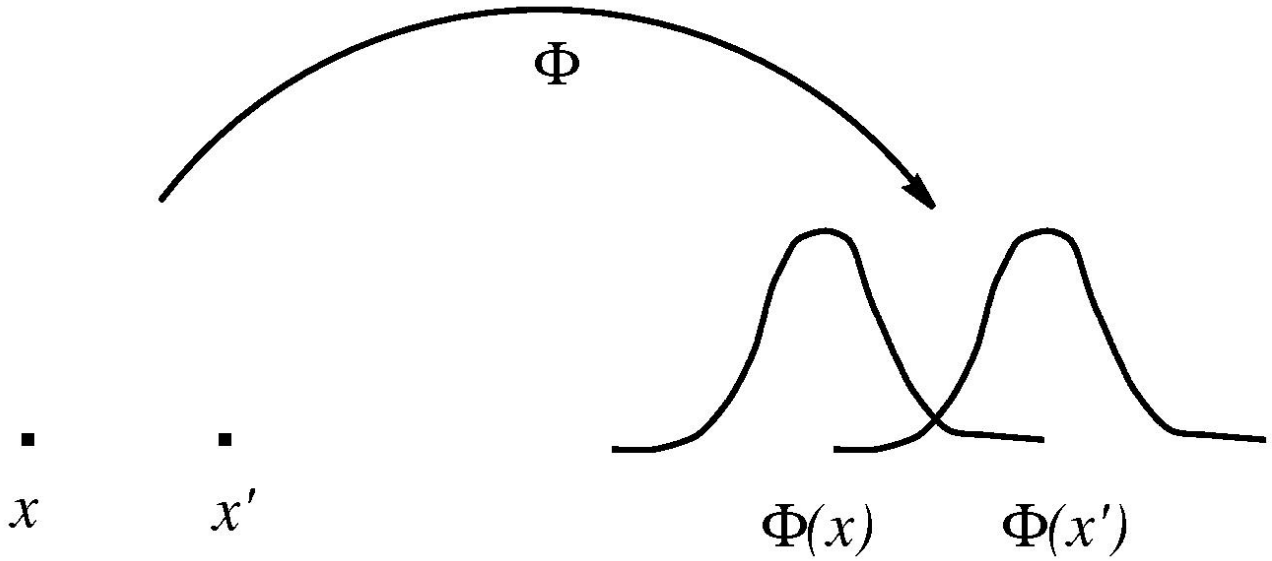
- there exists a map Φ into a dot product space $\mathcal{H}$ such that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

$\mathcal{H}$ is a so-called *reproducing kernel Hilbert space.* *(RKHS)*

If for pairwise distinct points, $\Sigma=0$ iff all $a_i = 0$, call *k strictly p.d.*

Constructing the RKHS



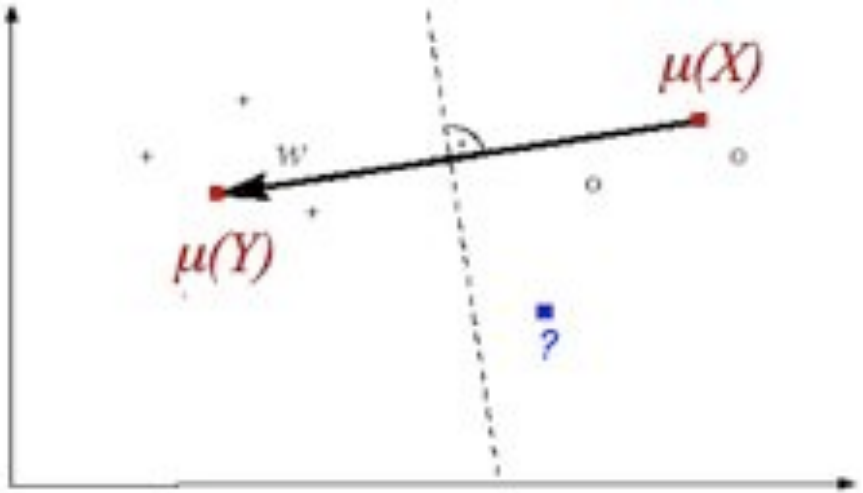
$x \mapsto \Phi(x) := k(x, .)$ *(Aronszajn, 1950)*

e.g., Gaussian kernel $k(x, x') = e^{-\frac{\|x - x'\|^2}{2\,\sigma^2}}$ (or some other positive definite k)

Take linear hull and define a dot product $\langle \Phi(x), \Phi(x') \rangle := k(x, x')$

Point evaluation: $f(x) = \langle f, k(x, .) \rangle$ — *"reproducing kernel Hilbert space"*

Pattern Recognition Algorithm in the RKHS



$\mathcal{X}$ compact subset of a separable metric space, $m, n \in \mathbb{N}$.

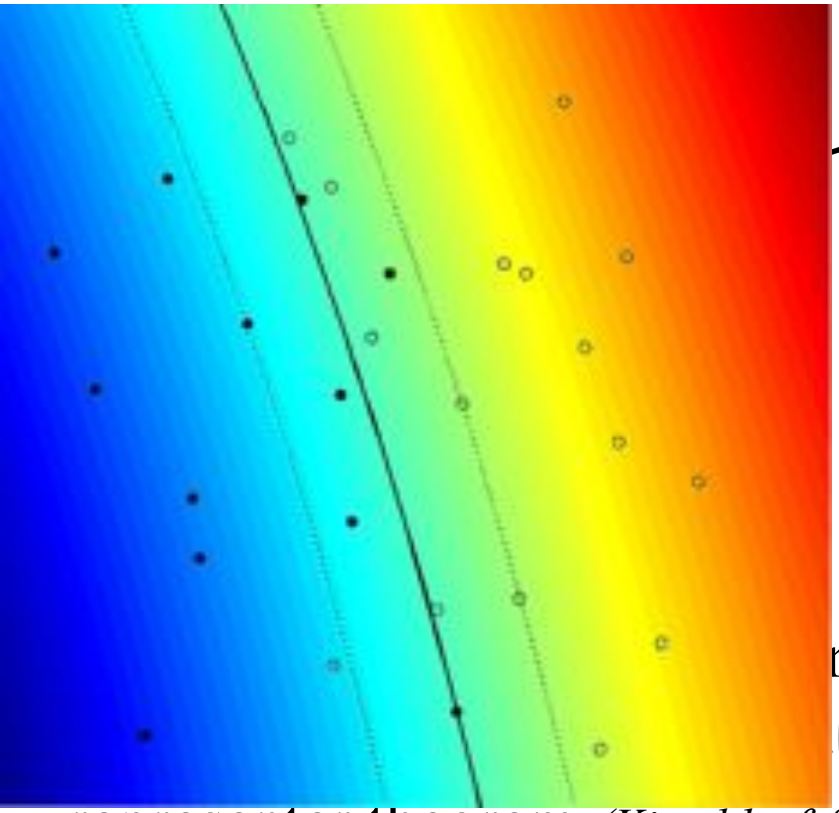
Positive class $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}$
Negative class $Y := \{y_1, \ldots, y_n\} \subset \mathcal{X}$

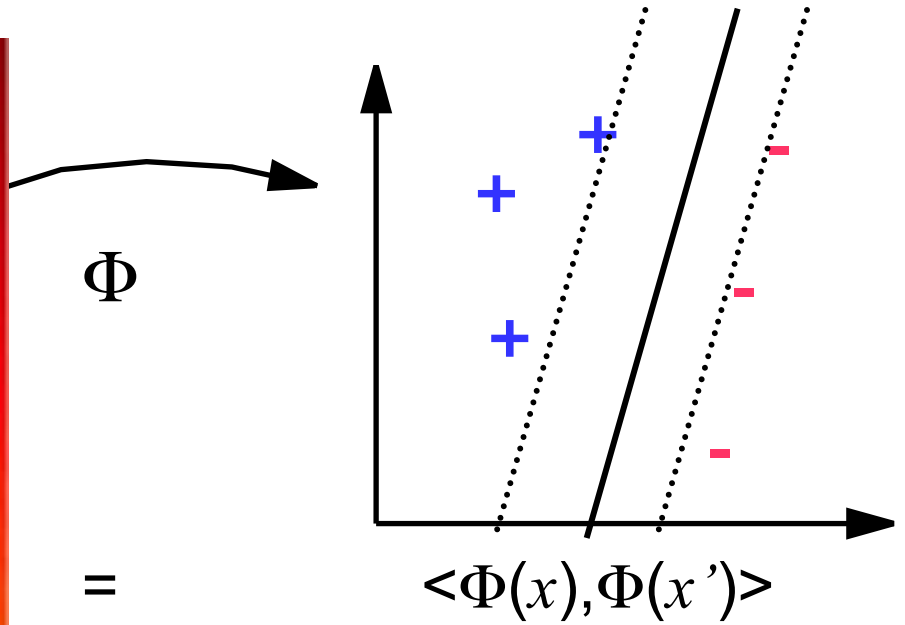
RKHS means $\mu(X) = \frac{1}{m}\sum_{i=1}^{m} k(x_i, \cdot)$, $\mu(Y) = \frac{1}{n}\sum_{i=1}^{n} k(y_i, \cdot)$.

Leads to a **Parzen windows** plug-in classifier:

$$f(z) = sgn\left(\frac{1}{m}\sum_{i=1}^{m} k(z, x_i) - \frac{1}{n}\sum_{i=1}^{n} k(z, y_i) + b\right)$$

Support Vector Machines *(Boser, Guyon, Vapnik 1992; Cortes & Vapnik 1995)*





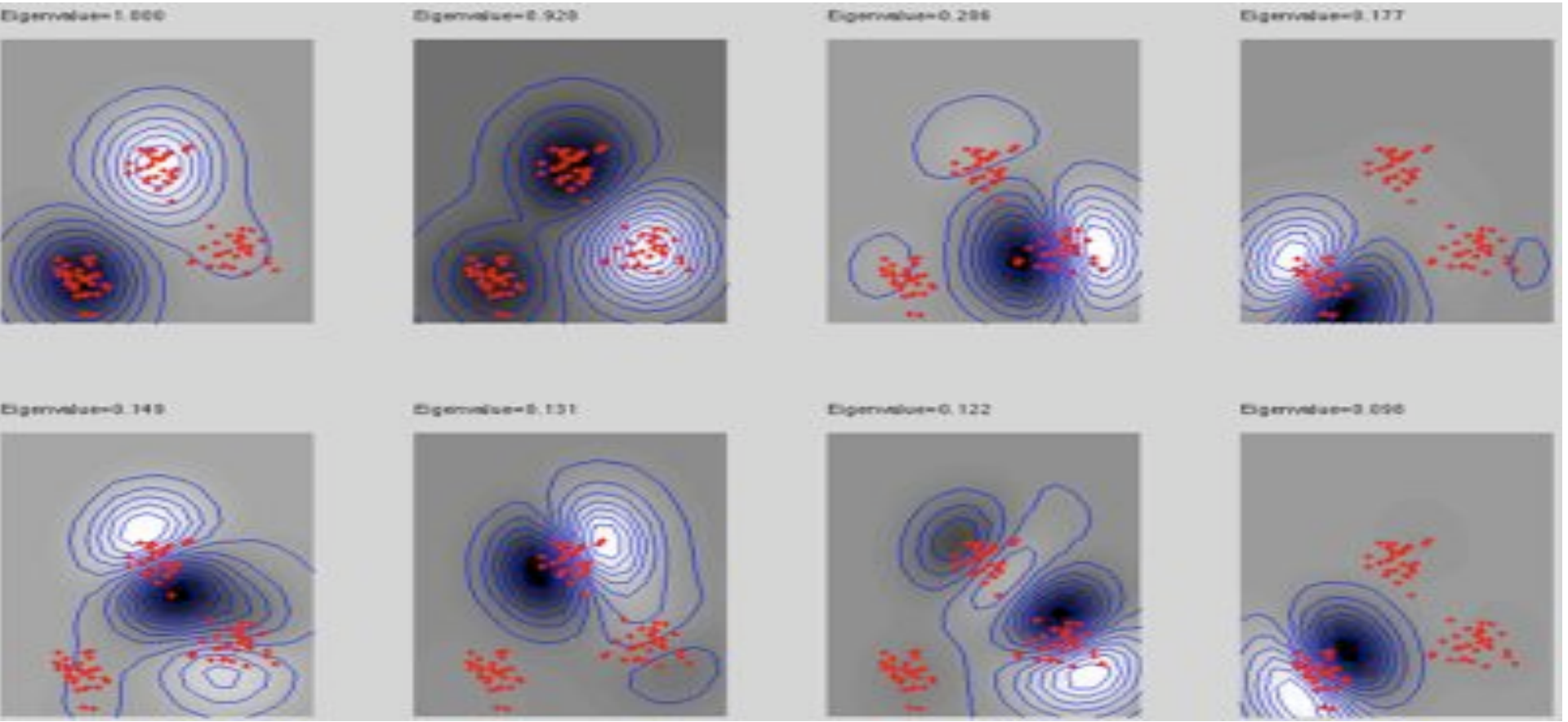
n terms of SVs:

$$(\sum_i \lambda_i k(x_i, x) + b)$$

representer theorem *(Kimeldorf & Wahba 1971, Schölkopf et al. 2000)*

- unique solution found by convex QP

Kernel PCA *(Schölkopf, Smola, Müller 1998)*



Contains LLE, Laplacian Eigenmap, and (in the limit) Isomap as special cases with data dependent kernels *(Ham et al. 2004)*

Spectral clustering

K similarity matrix; $D_{ii} = \sum_j K_{ij}$

Normalized cuts (Shi & Malik, 2000):
– map inputs to corresponding entries of the second smallest eigenvector of the normalized Laplacian

$$L = I - D^{-1/2} K D^{-1/2}.$$

– Partition them based on these values.

Meila & Shi (2001):
– map inputs to entries of largest eigenvectors of

$$D^{-1} K$$

– continue with k-means

Kernel PCA (1998):
α^n the nth eigenvector of K, with eigenvalue λ_n
– map test point x to RKHS, project on largest eigenvectors of K, normalized by $\lambda^{-1/2}$:

$$\langle V^n, k(x,.)\rangle = \lambda_n^{-1/2} \sum_{i=1}^{m} \alpha_i^n \langle k(x_i,.), k(x,.)\rangle = \lambda_n^{-1/2} \sum_{i=1}^{m} \alpha_i^n k(x_i, x)$$

Link to kernel PCA

projection of a *training point* x_t onto the nth eigenvector equals

$$\lambda_n^{-1/2}(K\alpha^n)_t = \lambda_n^{1/2}\alpha_t^n.$$

– the coefficient vector α^n contains the projections of the training set

– for a connected graph, the normalized Laplacian has a single 0 eigenvalue. Its (pseudo-)inverse is known as the discrete Green's function of the diffusion process on the graph governed by L. It can be viewed as a kernel matrix, encoding the dot product implying the commute time metric *(Ham, Lee, Mika, Schölkopf, 2004)*.

– the kPCA matrix is centered, and thus has a single eigenvalue 0 (for strictly p.d. kernel) that corresponds to the 0 eigenvalue of the normalized Laplacian.

– inversion inverts the order of the remaining eigenvalues.

The Empirical Kernel Map

Recall the feature map

$$\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$$
$$x \mapsto k(., x).$$

- each point is represented by its similarity to *all* other points
- how about representing it by its similarity to a *sample* of points?

Consider

$$\Phi_m : \mathcal{X} \to \mathbb{R}^m$$
$$x \mapsto k(., x)|_{(x_1,\ldots,x_m)} = (k(x_1, x), \ldots, k(x_m, x))^\top$$

- $\Phi_m(x_1), \ldots, \Phi_m(x_m)$ contain *all* necessary information about $\Phi(x_1), \ldots, \Phi(x_m)$
- the Gram matrix $G_{ij} := \left\langle \Phi_m(x_i), \Phi_m(x_j) \right\rangle$ satisfies $G = K^2$ where $K_{ij} = k(x_i, x_j)$
- modify Φ_m to

$$\Phi_m^w : \mathcal{X} \to \mathbb{R}^m$$
$$x \mapsto K^{-\frac{1}{2}}(k(x_1, x), \ldots, k(x_m, x))^\top$$

- this "whitened" map ("kernel PCA map") satifies

$$\left\langle \Phi_m^w(x_i), \Phi_m^w(x_j) \right\rangle = k(x_i, x_j)$$

for all $i, j = 1, \ldots, m$.

The Representer Theorem

Theorem 7 *Given: a p.d. kernel k on $\mathcal{X} \times \mathcal{X}$, a training set $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$, a strictly monotonic increasing real-valued function Ω on $[0, \infty[$, and an arbitrary cost function $c : (\mathcal{X} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{\infty\}$*

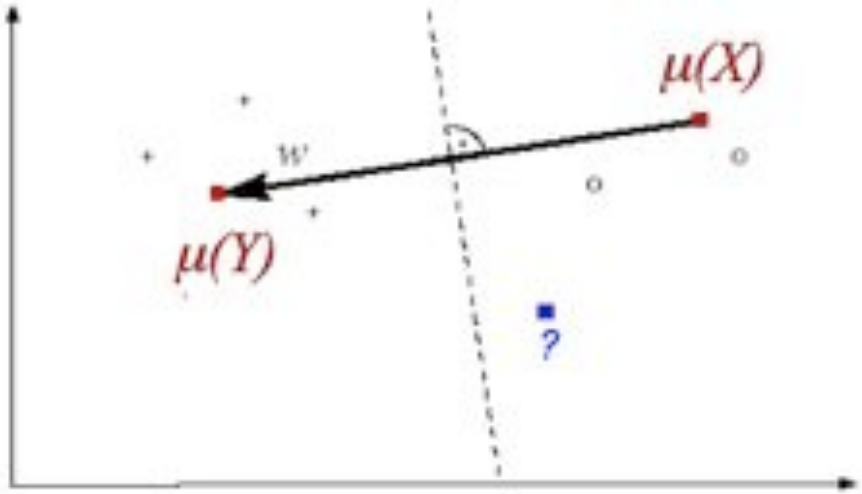
Any $f \in \mathcal{H}_k$ minimizing the regularized risk functional

$$c\left((x_1, y_1, f(x_1)), \ldots, (x_m, y_m, f(x_m))\right) + \Omega\left(\|f\|\right) \qquad (3)$$

admits a representation of the form

$$f(.) = \sum_{i=1}^{m} \alpha_i k(x_i, .).$$

- the kernel corresponds to
  - a similarity measure for the data, or
  - a (linear) representation of the data, or
  - a hypothesis space for learning,



What if $\mu(X) = \mu(Y)$?

When do the means coincide?

$k(x, x') = \langle x, x' \rangle$:	the means coincide
$k(x, x') = (\langle x, x' \rangle + 1)^d$:	all empirical moments up to order d coincide
k strictly pd:	$X = Y$.

The mean "remembers" each point that contributed to it.

Proposition 1 *Assume that k is strictly pd, and for all i, j, $x_i \neq x_j$, and $y_i \neq y_j$. If for some $\alpha_i, \beta_j \in \mathbb{R} - \{0\}$, we have*

$$\sum_{i=1}^{m} \alpha_i k(x_i, .) = \sum_{j=1}^{n} \beta_j k(y_j, .), \tag{1}$$

then $X = Y$.

Proof (by contradiction): W.l.o.g., assume that $x_1 \notin Y$. Subtract $\sum_{j=1}^{n} \beta_j k(y_j, .)$ from (1), and make it a sum over distinct points, to get

$$0 = \sum_i \gamma_i k(z_i, .),$$

where $z_1 = x_1, \gamma_1 = \alpha_1 \neq 0$, and $z_2, \cdots \in X \cup Y - \{x_1\}$, $\gamma_2, \cdots \in \mathbb{R}$. Take the dot product with $\sum_j \gamma_j k(z_j, .)$, using $\langle k(z_i, .), k(z_j, .) \rangle = k(z_i, z_j)$, to get

$$0 = \sum_{ij} \gamma_i \gamma_j k(z_i, z_j),$$

with $\gamma \neq 0$, hence k cannot be strictly pd.

The mean map for samples

$$\mu\colon X = (x_1, \ldots, x_m) \mapsto \frac{1}{m}\sum_{i=1}^{m} k(x_i, \cdot)$$

satisfies

$$\langle \mu(X), f \rangle = \langle \frac{1}{m}\sum_{i=1}^{m} k(x_i, \cdot), f \rangle = \frac{1}{m}\sum_{i=1}^{m} f(x_i)$$

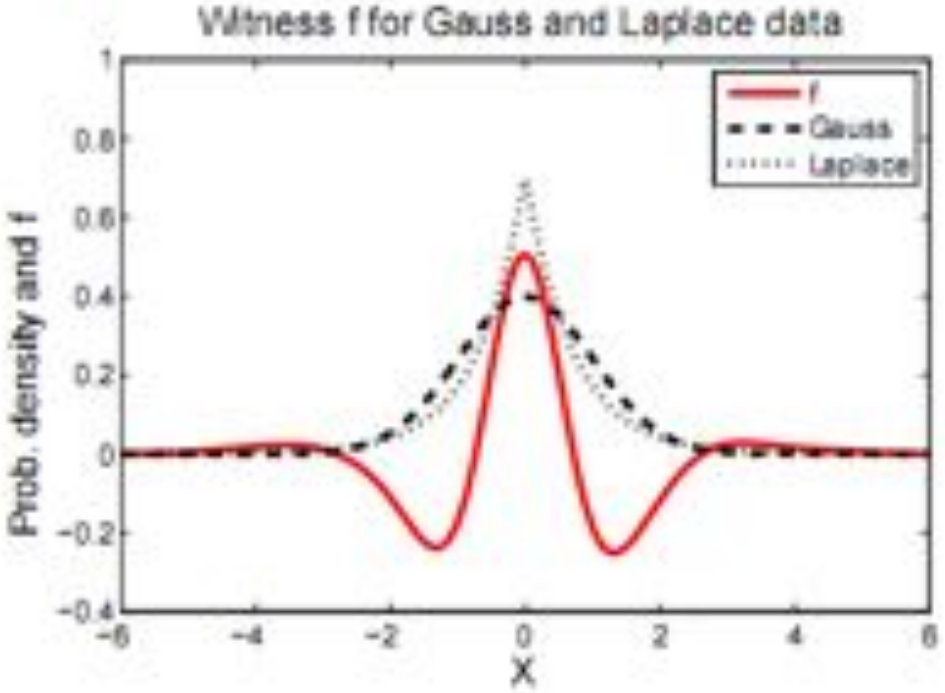
and

$$\|\mu(X) - \mu(Y)\| = \sup_{\|f\| \leq 1} |\langle \mu(X) - \mu(Y), f \rangle| = \sup_{\|f\| \leq 1} \left| \frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(y_i) \right|.$$

- $\nu(X) \neq \mu(Y) \Longleftrightarrow$ can find a function distinguishing the samples
- If k is strictly p.d., this is equivalent to $X \neq Y$; i.e., we can always distinguish distinct samples.

Witness function

$f = \frac{\mu(X)-\mu(Y)}{\|\mu(X)-\mu(Y)\|}$, thus $f(x) \propto \langle \mu(X) - \mu(Y), k(x,.) \rangle$):



(done in the Gaussian RKHS)

The mean map for measures

Assumptions:

- p, q Borel probability measures
- $\mathbf{E}_{x,x' \sim p}[k(x, x')]$, $\mathbf{E}_{x,x' \sim q}[k(x, x')] < \infty$
  (follows if we assume $\|k(x, .)\| \leq M < \infty$)

Define

$$\mu : p \mapsto \mathbf{E}_{x \sim p}[k(x, \cdot)].$$

Note

$$\langle \mu(p), f \rangle = \mathbf{E}_{x \sim p}[f(x)]$$

and

$$\|\mu(p) - \mu(q)\| = \sup_{\|f\| \leq 1} \left| \mathbf{E}_{x \sim p}[f(x)] - \mathbf{E}_{x \sim q}[f(x)] \right|.$$

Under which conditions is μ injective?

Smola, Gretton, Song, Schölkopf, ALT 2007;
Fukumizu, Gretton, Sun, Schölkopf, NIPS 2007

Theorem 2 *[Fortet and Mourier (1953); Dudley (2002)]*

$$p = q \iff \sup_{f \in C(\mathcal{X})} \left| \mathbf{E}_{x \sim p}(f(x)) - \mathbf{E}_{x \sim q}(f(x)) \right| = 0,$$

where $C(\mathcal{X})$ is the space of continuous bounded functions on $\mathcal{X}$.

Combine this with

$$\|\mu(p) - \mu(q)\| = \sup_{\|f\| \leq 1} \left| \mathbf{E}_{x \sim p}[f(x)] - \mathbf{E}_{x \sim q}[f(x)] \right|$$

Replace $C(\mathcal{X})$ by the unit ball in an RKHS that is dense in $C(\mathcal{X})$ — universal kernel [51], e.g., Gaussian.

Theorem 3 *[Gretton et al. (2007)] If k is universal, then*

$$p = q \iff \|\mu(p) - \mu(q)\| = 0.$$

- for $k(x, x') = e^{\langle x, x' \rangle}$ we recover the *moment generating function* of a RV x with distribution p:

$$M_p(.) = \mathbf{E}_{x \sim p} \left[e^{\langle x, \cdot \rangle} \right]$$

- for $k(x, x') = e^{i\langle x, x' \rangle}$ we recover the *characteristic function*:

$$M_p(.) = \mathbf{E}_{x \sim p} \left[e^{i\langle x, \cdot \rangle} \right].$$

- μ is invertible on its image
  $\mathcal{M} = \{\mu(p) \mid p \text{ is a probability distribution}\}$
  (the "marginal polytope", *Wainwright & Jordan, 2003*)

An injective μ provides us with a convenient metric on probability distributions, which can be used to check whether two distributions are different.

Construct estimators for $\|\mu(p) - \mu(q)\|^2$ for various applications.

Application 1: Two-sample problem *(Gretton et al., 2007)*

X, Y i.i.d. m-samples from p, q, respectively.

$$\begin{aligned}\|\mu(p) - \mu(q)\|^2 =& \mathbf{E}_{x,x' \sim p}\left[k(x,x')\right] - 2\mathbf{E}_{x \sim p, y \sim q}\left[k(x,y)\right] + \mathbf{E}_{y,y' \sim q}\left[k(y,y')\right] \\ =& \mathbf{E}_{x,x' \sim p, y,y' \sim q}\left[h((x,y),(x',y'))\right]\end{aligned}$$

with

$$h((x,y),(x',y')) := k(x,x') - k(x,y') - k(y,x') + k(y,y').$$

Define

$$\begin{aligned}D(p,q)^2 &:= \mathbf{E}_{x,x' \sim p, y,y' \sim q} h((x,y),(x',y')) \\ \hat{D}(X,Y)^2 &:= \frac{1}{m(m-1)} \sum_{i \neq j} h((x_i,y_i),(x_j,y_j)).\end{aligned}$$

$\hat{D}(X,Y)^2$ is an unbiased estimator of $D(p,q)^2$.
It's easy to compute, and works on structured data.

Kernel Independence Testing

k bounded universal p.d. kernel; p, q Borel probability measures

Kernel mean map

$$\mu : p \mapsto \mathbf{E}_{x\sim p}[k(x, .)]$$

is injective.

Corollary: $x \perp\!\!\!\perp y \iff \Delta := \|\mu(p_{xy}) - \mu(p_x \times p_y)\| = 0.$

Link to cross-covariance: For $k((x, y), (x', y')) = k_x(x, x')k_y(y, y')$:
Δ^2 = squared HS-norm of cross-covariance operator between the two RKHSes.

Estimator $\frac{1}{n}tr[K_X K_Y]$, where K_X is the centered Gram matrix of $\{x_1, \ldots, x_n\}$ (likewise, K_Y).
(Gretton, Herbrich, Smola, Bousquet, Schölkopf, 2005; Gretton et al., 2008)

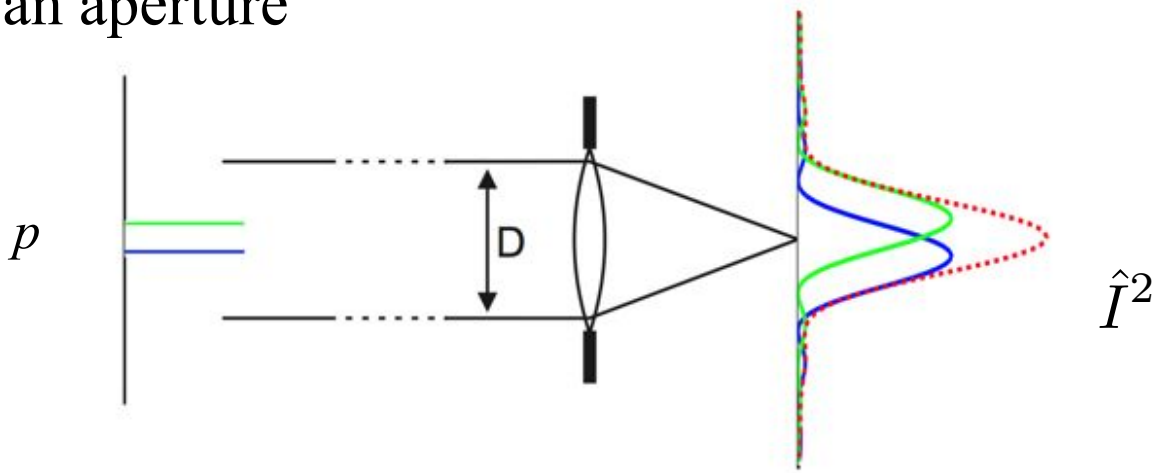
Link to Kernel ICA *(Bach & Jordan, 2002)*:
$x \perp\!\!\!\perp y$ iff $\sup_{f,g \in \text{RHKS unit balls}} \text{cov}(f(x), g(y)) = 0$

Shift-Invariant Optical Realization

$\mu : p \mapsto \mathbf{E}_{x\sim p}[k(x - .)]$

Fourier imaging through an aperture

- p source of incoherent light
- I indicator function of an aperture of width D



- in *Fraunhofer diffraction*, the intensity image is $\propto p * \hat{I}^2$
- set $k := \hat{I}^2$ (this is p.d. by Bochner's theorem)
- then the image equals $\mu(p)$
- this μ is not invertible (since k is not universal) — "diffraction limit"
- if we restrict the input domain to distributions with compact support, it is invertible no matter how small $D > 0$

(Schölkopf, Sriperumbudur, Gretton, Fukumizu 2008; Harmeling, Hirsch, Schölkopf 2013)

Kernels as Green's Functions

- in this case, the kernel is the point response of a linear optical system
- more generally, the kernel k can be viewed as the Green's function of P^*P, where P is a regularization operator such that the RKHS norm can be written as $\|f\|_k = \|Pf\|$
- for instance, the Gaussian kernel corresponds to a regularization operator which computes an infinite series of derivatives of f
- for translation-invariant kernels, P can be written as a multiplication operator in Fourier space, amplifying high frequencies and thus penalizing them in $\|Pf\|$

Poggio & Girosi 1990; Schölkopf & Smola 2002; Hofmann, Schölkopf, Smola 2008

Non-Injectivity of Fourier Imaging

• assume: densities exist, kernel shift invariant, *k(x,y) = k(x-y),* all Fourier transforms exist. Note that μ is invertible iff

$$\mathbf{E}_{x\sim p}[k(x,\cdot)] = \mathbf{E}_{x\sim q}[k(x,\cdot)] \implies p = q$$

$$\int k(x-y)p(y)\,dy = \int k(x-y)q(y)\,dy \implies p = q$$

$$\text{i.e.,} \quad \hat{k}(\hat{p}-\hat{q}) = 0 \implies p = q$$

(Sriperumbudur, Fukumizu, Gretton, Lanckriet, Schölkopf, COLT 2008)

• E.g.: μ is invertible if $\hat{k}$ has full support.
• this is not the case for $\hat{k} = I * I$.

More precisely,

$$\|\mu(p) - \mu(q)\| = \|F^{-1}[(\bar{\hat{p}} - \bar{\hat{q}})\hat{k}]\|$$

where $\hat{k}$ is the nonnegative finite measure corresponding to k via Bochner's theorem, and $\hat{p}, \hat{q}$ are the characteristic functions of the Borel measures p, q. Thus μ is invertible for the class of all nonnegative measures if $\hat{k}$ has full support.

Injectivity of Fourier Imaging with Prior Knowledge

• How about if $\hat{k}$ does not have full support, but nonempty interior (e.g., $\hat{k} = I * I$) ?
• in that case, μ is invertible for all distributions with compact support, by Schwartz-Paley-Wiener *(Sriperumbudur, Fukumizu, Gretton, Lanckriet, Schölkopf, COLT 2008).*

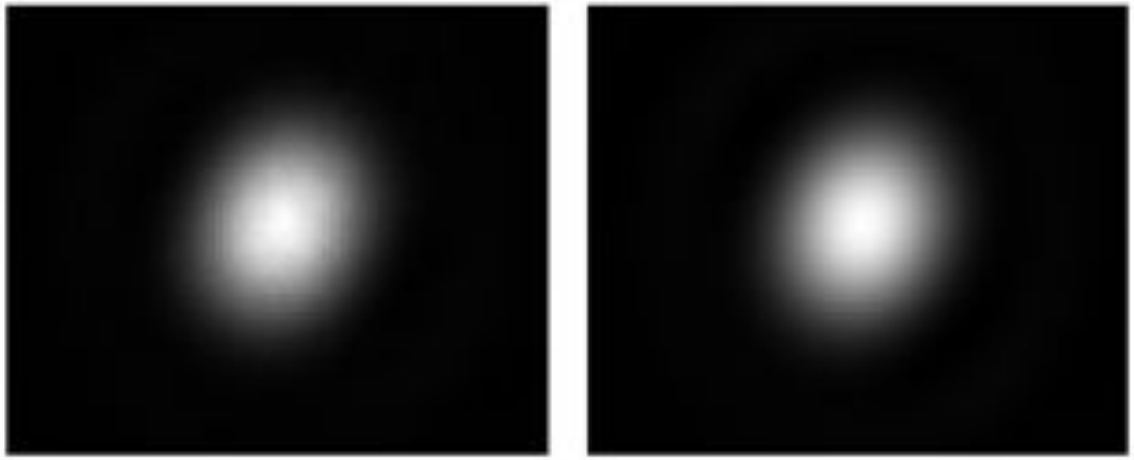
• The Fraunhofer diffraction aperture imaging process is *not* invertible for the class of all light sources, but it is if we restrict the class (e.g., to compact support).

Algorithmic Method

Harmeling, Hirsch, Schölkopf, CVPR 2013

• exploit nonegativity of image, and bounded support of object

double star; distance= 0.5*Rayleigh



reconstruction passed through forward model

reconstruction

"ground truth" recorded with larger aperture

"imitate the superficial exterior of a process or system without having any understanding of the underlying substance".
(source: http://philosophyisfashionable.blogspot.com/)

"cargo cult"

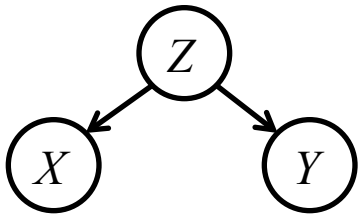
- for prediction in the IID setting, imitating the exterior of a process is enough (i.e., can disregard causal structure)
- anything else can benefit from causal learning

Thanks to P. Laskov.

Statistical Implications of Causality

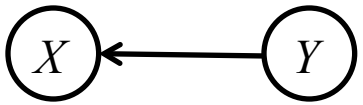
Reichenbach's
Common Cause Principle
links **causality** and **probability:**

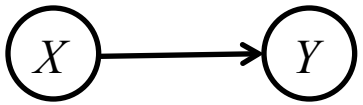
(i) if X and Y are statistically dependent, then there is a Z causally influencing both;



special cases:

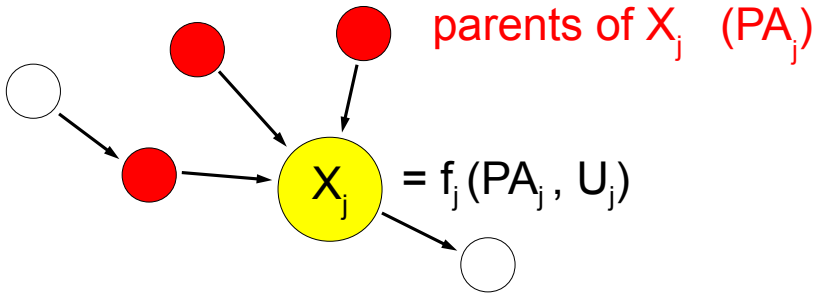
(ii) Z screens X and Y from each other (given Z, the observables X and Y become independent)





Functional Causal Model *(Pearl et al.)*

- Set of observables $X_1, \ldots, X_n$
- directed acyclic graph G with vertices $X_1, \ldots, X_n$
- Semantics: parents = direct causes
- $X_i = f_i(\text{ParentsOf}_i, \text{Noise}_i)$, with jointly independent $\text{Noise}_1, \ldots, \text{Noise}_n$.



- entails $p(X_1, \ldots, X_n)$ with particular conditional independence structure

  Question: Can we recover G from p?

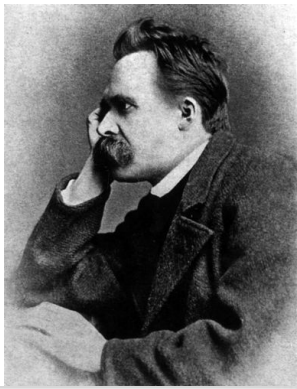
  Answer: under certain assumptions, can recover an equivalence class containing the correct G using conditional independence testing.

  Problem: does not work in the simplest case. *Below: two ideas.*

Friedrich Nietzsche's

TWILIGHT OF THE IDOLS

or How to Philosophize with a Hammer



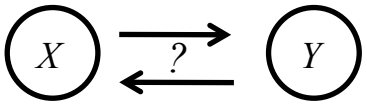
Translated, with commentary, by R.J. Hollingdale

[. . .]

The Four Great Errors

1

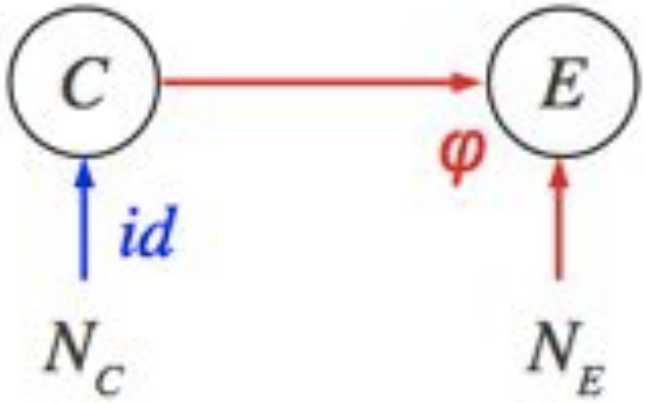
The error of mistaking cause for consequence. - There is no more dangerous error than that of *mistaking the consequence for the cause*: I call it reason's intrinsic form of corruption. Nonetheless, this error is among the most ancient and most recent



Independence of input and mechanism

Causal structure:

C cause
E effect
N noise
φ mechanism

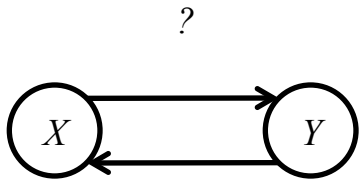


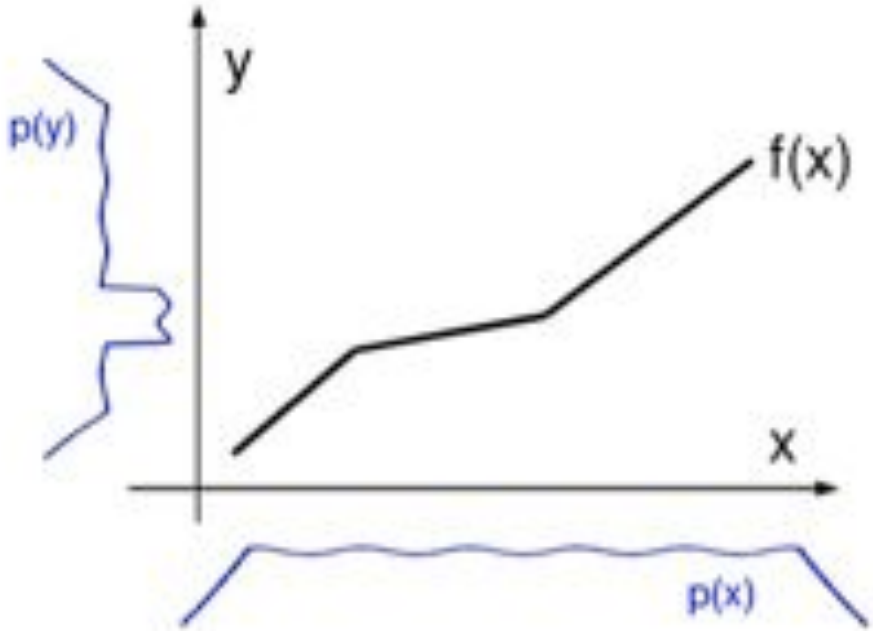
Assumption:
$p(C)$ and $p(E|C)$ are "independent"

Janzing & Schölkopf, IEEE Trans. Inf. Theory, 2010; cf. also Lemeire & Dirkx, 2007

Inferring deterministic causal relations

- Does not require noise
- Assumption: $y = f(x)$ with invertible f





Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf: Inferring deterministic causal relations, *UAI* 2010

Causal independence implies anticausal dependence

Assume that f is a monotonously increasing bijection of $[0,1]$.
View p_x and $\log f'$ as RVs on the prob. space $[0,1]$ w. Lebesgue measure.

Postulate (independence of mechanism and input):

$$\mathrm{Cov}\,(\log f', p_x) = 0$$

Note: this is equivalent to

$$\int_0^1 \log f'(x) p(x) dx = \int_0^1 \log f'(x) dx,$$

since

$$\mathrm{Cov}\,(\log f', p_x) = E\,[\log f' \cdot p_x] - E\,[\log f']\,E\,[p_x] = E\,[\log f' \cdot p_x] - E\,[\log f'].$$

Proposition:

$$\mathrm{Cov}\,(\log f^{-1'}, p_y) \geq 0$$

with equality iff $f = Id$.

u_x, u_y uniform densities for x, y

v_x, v_y densities for x, y induced by transforming u_y, u_x via f^{-1} and f

Equivalent formulations of the postulate:

Additivity of Entropy:

$$S(p_y) - S(p_x) = S(v_y) - S(u_x)$$

Orthogonality (information geometric):

$$D(p_x \,\|\, v_x) = D(p_x \,\|\, u_x) + D(u_x \,\|\, v_x)$$

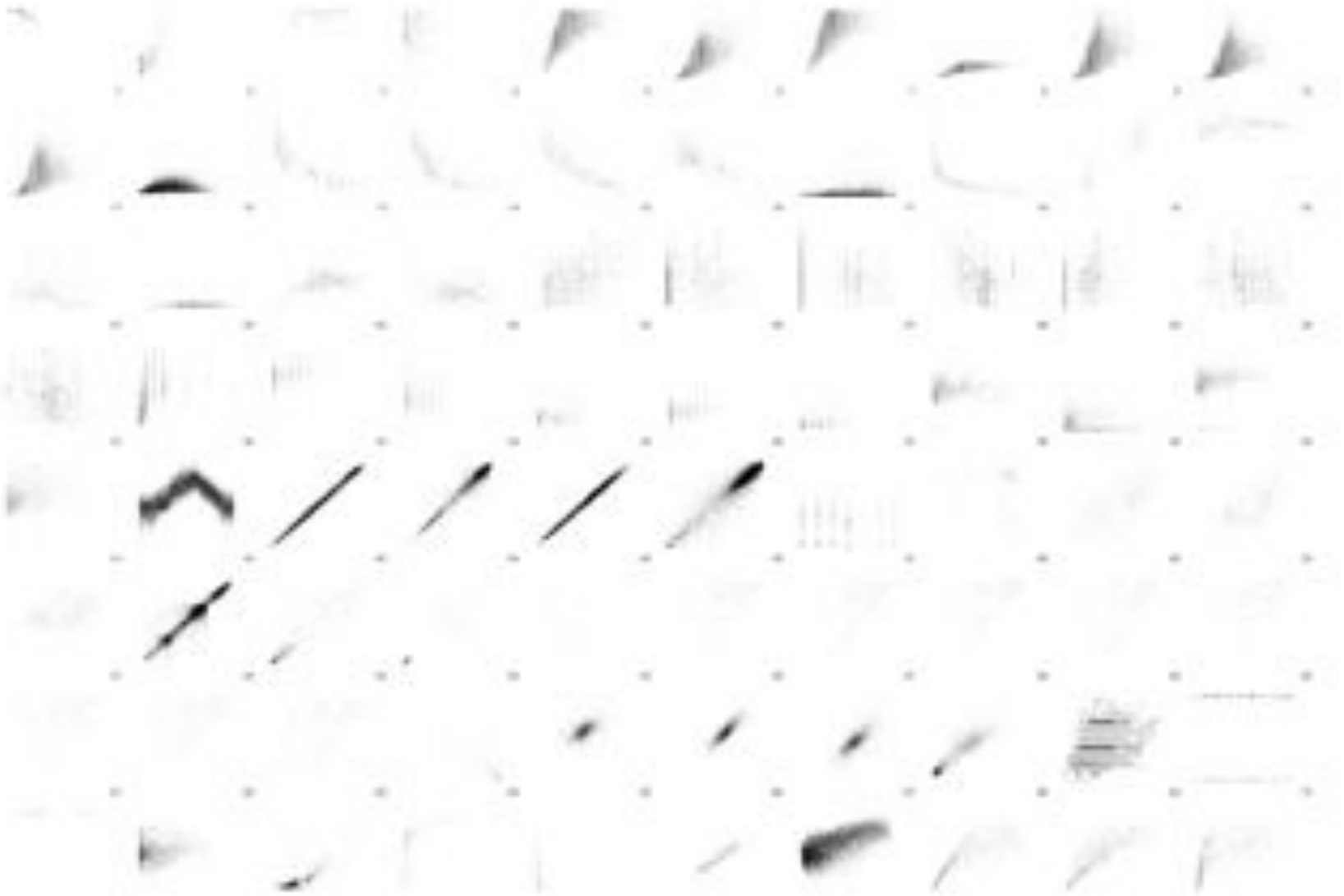
which can be rewritten as

$$D(p_y \,\|\, u_y) = D(p_x \,\|\, u_x) + D(v_y \,\|\, u_y)$$

Interpretation:

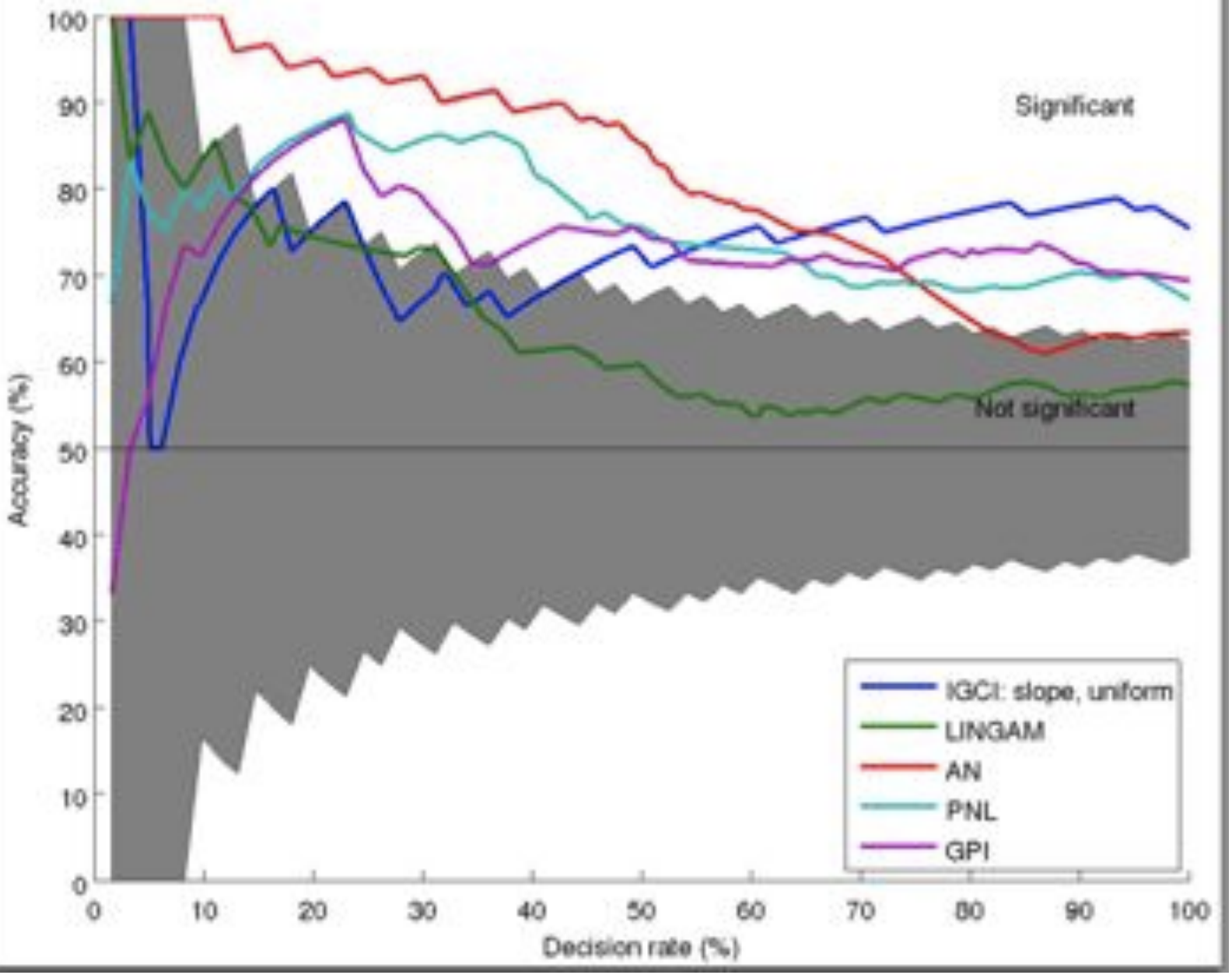
irregularity of p_y = irregularity of p_x + irregularity introduced by f

80 Cause-Effect Pairs



80 Cause-Effect Pairs − Examples

	var 1	var 2	dataset	ground truth
pair0001	Altitude	Temperature	DWD	→
pair0005	Age (Rings)	Length	Abalone	→
pair0012	Age	Wage per hour	census income	→
pair0025	cement	compressive strength	concrete_data	→
pair0033	daily alcohol consumption	mcv mean corpuscular volume	liver disorders	→
pair0040	Age	diastolic blood pressure	pima indian	→
pair0042	day	temperature	B. Janzing	→
pair0047	#cars/24h	specific days	traffic	←
pair0064	drinking water access	infant mortality rate	UNdata	→
pair0068	bytes sent	open http connections	P. Daniusis	←
pair0069	inside room temperature	outside temperature	J. M. Mooij	←
pair0070	parameter	sex	Bülthoff	→
pair0072	sunspot area	global mean temperature	sunspot data	→
pair0074	GNI per capita	life expectancy at birth	UNdata	→
pair0078	PPFD (Photosynth. Photon Flux)	NEP (Net Ecosystem Productivity)	Moffat A. M.	→



IGCI:
Deterministic
Method

LINGAM:
Shimizu et al.,
2006

AN:
Additive Noise
Model (nonlinear)

PNL:
AN with post-
nonlinearity

GPI:
Mooij et al.,
2010

Further Applications of Causal Inference

1. *Grosse-Wentrup, Schölkopf, and Hill*, Causal Influence of Gamma Oscillations on the Sensorimotor Rhythm. *NeuroImage*, 2011

2. *Grosse-Wentrup & Schölkopf*, High Gamma-Power Predicts Performance in Sensorimotor-Rhythm Brain-Computer Interfaces. *J. Neural Engineering*, 2012
   (2011 International BCI Research Award)

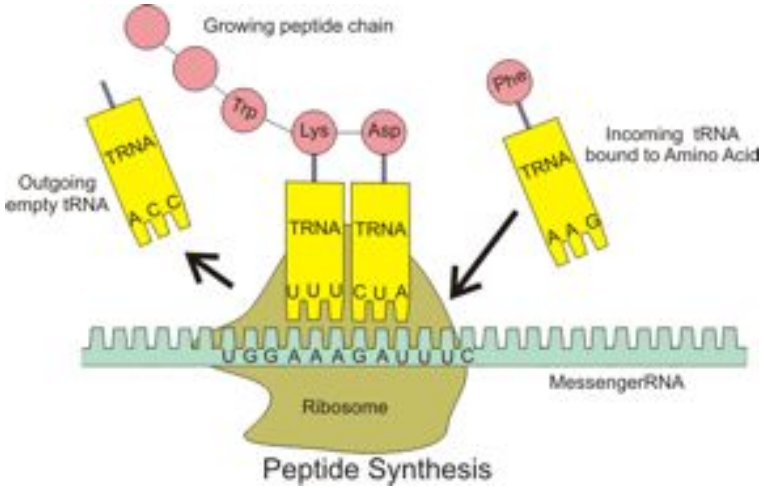
3. *Besserve, Janzing, Logothetis & Schölkopf*, Finding dependencies between frequencies with the kernel cross-spectral density, *Intl. Conf. Acoustics, Speech and Signal Processing*, 2011

4. *Besserve, Schölkopf, Logothetis & Panzeri*, Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis. *J. Computational Neuroscience*, 2010

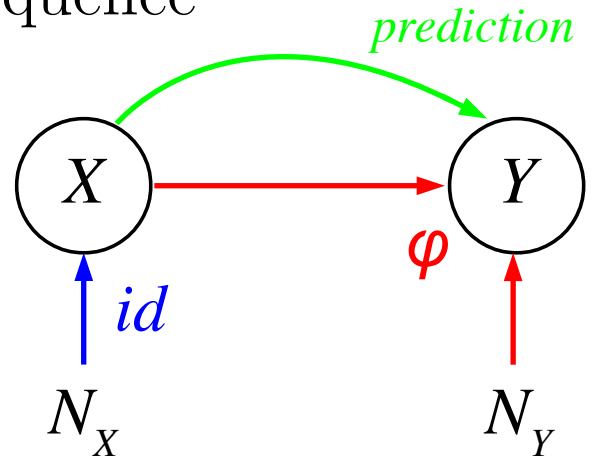
Causal Learning and Anticausal Learning

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij, *ICML* 2012

- example 1: predict gene from mRNA sequence

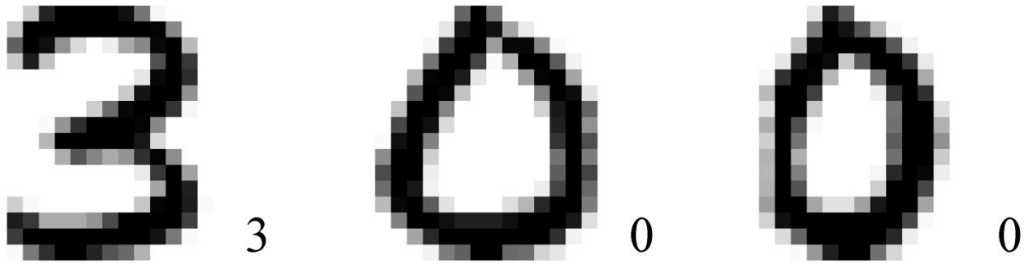


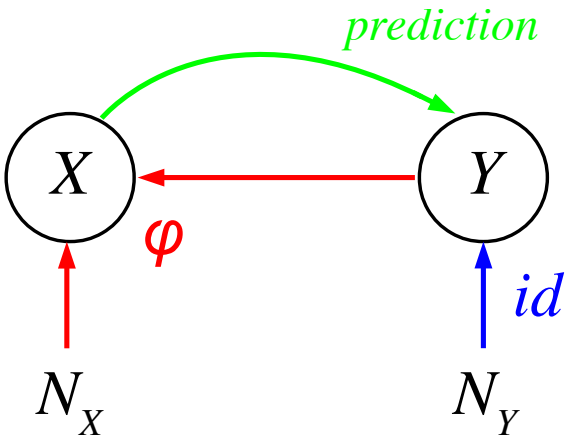
Source: http://commons.wikimedia.org/wiki/File:Peptide_syn.png



causal mechanism φ

- example 2: predict class membership from handwritten digit



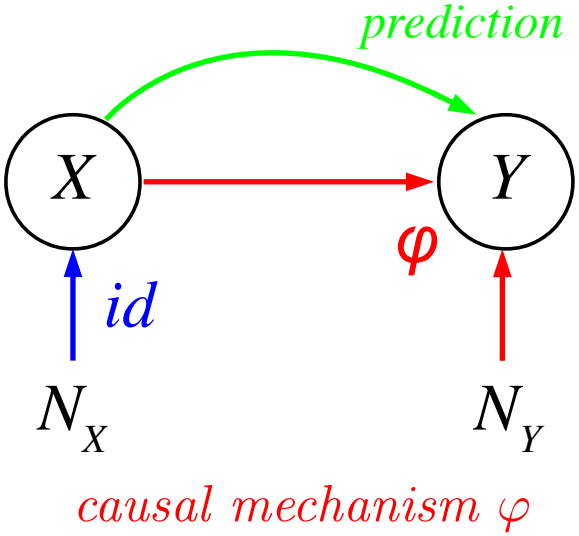


Covariate Shift and Semi-Supervised Learning

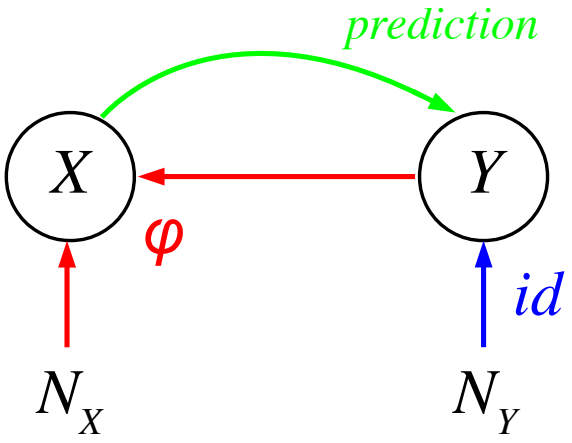
Assumption: $p(C)$ and mechanism $p(E|C)$ are "independent"

Goal: learn $X \mapsto Y$, i.e., estimate (properties of) $p(Y|X)$

- covariate shift (i.e., $p(X)$ changes): mechanism $p(Y|X)$ is unaffected by assumption
- semi-supervised learning: impossible, since $p(X)$ contains no information about $p(Y|X)$
- transfer learning (N_X, N_Y change, φ not): could be done by additive noise model with conditionally independent noise



- $p(X)$ **changes**: need to decide if change is due to mechanism $p(X|Y)$ or cause distribution $p(Y)$ (sometimes: by deconvolution)
- **semi-supervised learning**: possible, since $p(X)$ contains information about $p(Y|X)$ — e.g., cluster assumption.
- transfer learning: as above



(cf. Storkey, 2009)

Semi-Supervised Learning *(Schölkopf et al., ICML 2012)*

- Known SSL assumptions link $p(X)$ to $p(Y|X)$:
  - *Cluster assumption*: points in same cluster of $p(X)$ have the same Y
  - *Low density separation assumption*: $p(Y|X)$ should cross 0.5 in an area where $p(X)$ is small
  - *Semi-supervised smoothness assumption*: $\mathrm{E}(Y|X)$ should be smooth where $p(X)$ is large

- Next slides: experimental analysis

SSL Book Benchmark Datasets – Chapelle et al. (2006)

Table 1. Categorization of eight benchmark datasets as Anticausal/Confounded, Causal or Unclear

Category	Dataset
Anticausal/ Confounded	g241c: the class causes the 241 features.
	g241d: the class (binary) and the features are confounded by a variable with 4 states.
	Digit1: the positive or negative angle and the features are confounded by the variable of continuous angle.
	USPS: the class and the features are confounded by the 10-state variable of all digits.
	COIL: the six-state class and the features are confounded by the 24-state variable of all objects.
Causal	SecStr: the amino acid is the cause of the secondary structure.
Unclear	BCI, Text: Unclear which is the cause and which the effect.

UCI Datasets used in SSL benchmark – Guo et al., 2010

Table 2. Categorization of 26 UCI datasets as Anticausal/Confounded, Causal or Unclear

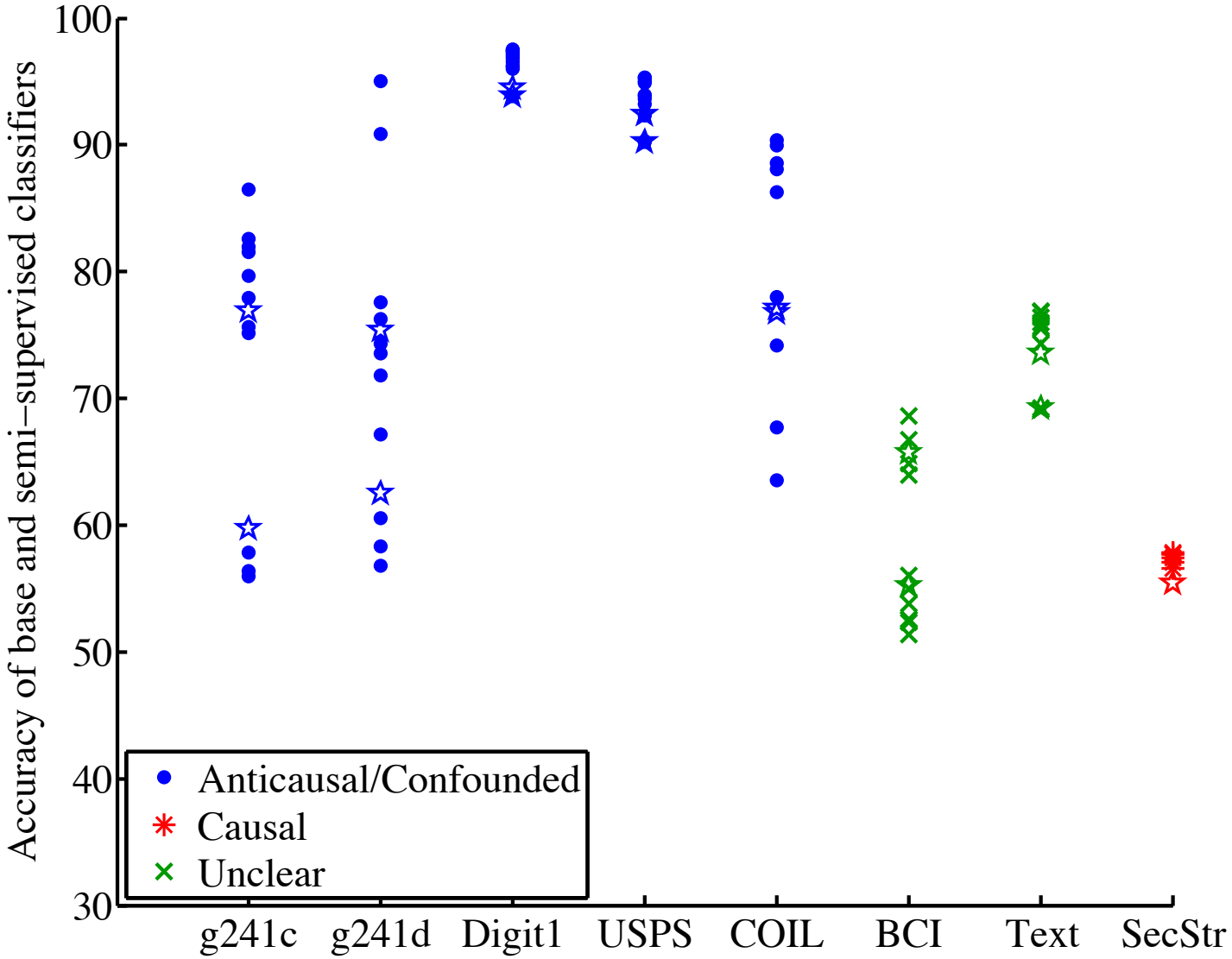
Categ.	Dataset
Anticausal/Confoumded	Breast Cancer Wisconsin: the class of the tumor (benign or malignant) causes some of the features of the tumor (e.g., thickness, size, shape etc.).
	Diabetes: whether or not a person has diabetes affects some of the features (e.g., glucose concentration, blood pressure), but also is an effect of some others (e.g. age, number of times pregnant).
	Hepatitis: the class (die or survive) and many of the features (e.g., fatigue, anorexia, liver big) are confounded by the presence or absence of hepatitis. Some of the features, however, may also cause death.
	Iris: the size of the plant is an effect of the category it belongs to.
	Labor: cyclic causal relationships: good or bad labor relations can cause or be caused by many features (e.g., wage increase, number of working hours per week, number of paid vacation days, employer's help during employee 's long term disability). Moreover, the features and the class may be confounded by elements of the character of the employer and the employee (e.g., ability to cooperate).
	Letter: the class (letter) is a cause of the produced image of the letter.
	Mushroom: the attributes of the mushroom (shape, size) and the class (edible or poisonous) are confounded by the taxonomy of the mushroom (23 species).
	Image Segmentation: the class of the image is the cause of the features of the image.
	Sonar, Mines vs. Rocks: the class (Mine or Rock) causes the sonar signals.
	Vehicle: the class of the vehicle causes the features of its silhouette.
	Vote: this dataset may contain causal, anticausal, confounded and cyclic causal relations. E.g., having handicapped infants or being part of religious groups in school can cause one's vote, being democrat or republican can causally influence whether one supports Nicaraguan contras, immigration may have a cyclic causal relation with the class. Crime and the class may be confounded, e.g., by the environment in which one grew up.
	Vowel: the class (vowel) causes the features.
	Wave: the class of the wave causes its attributes.
Causal	Balance Scale: the features (weight and distance) cause the class.
	Chess (King-Rook vs. King-Pawn): the board-description causally influences whether white will win.
	Splice: the DNA sequence causes the splice sites.
Unclear	Breast-C, Colic, Sick, Ionosphere, Heart, Credit Approval were unclear to us. In some of the datasets, it is unclear whether the class label may have been generated or defined based on the features (e.g., Ionoshpere, Credit Approval, Sick).

Datasets, co-regularized LS regression – Brefeld et al., 2006

Table 3. Categorization of 31 datasets (described in the paragraph "Semi-supervised regression") as Anticausal/Confounded, Causal or Unclear

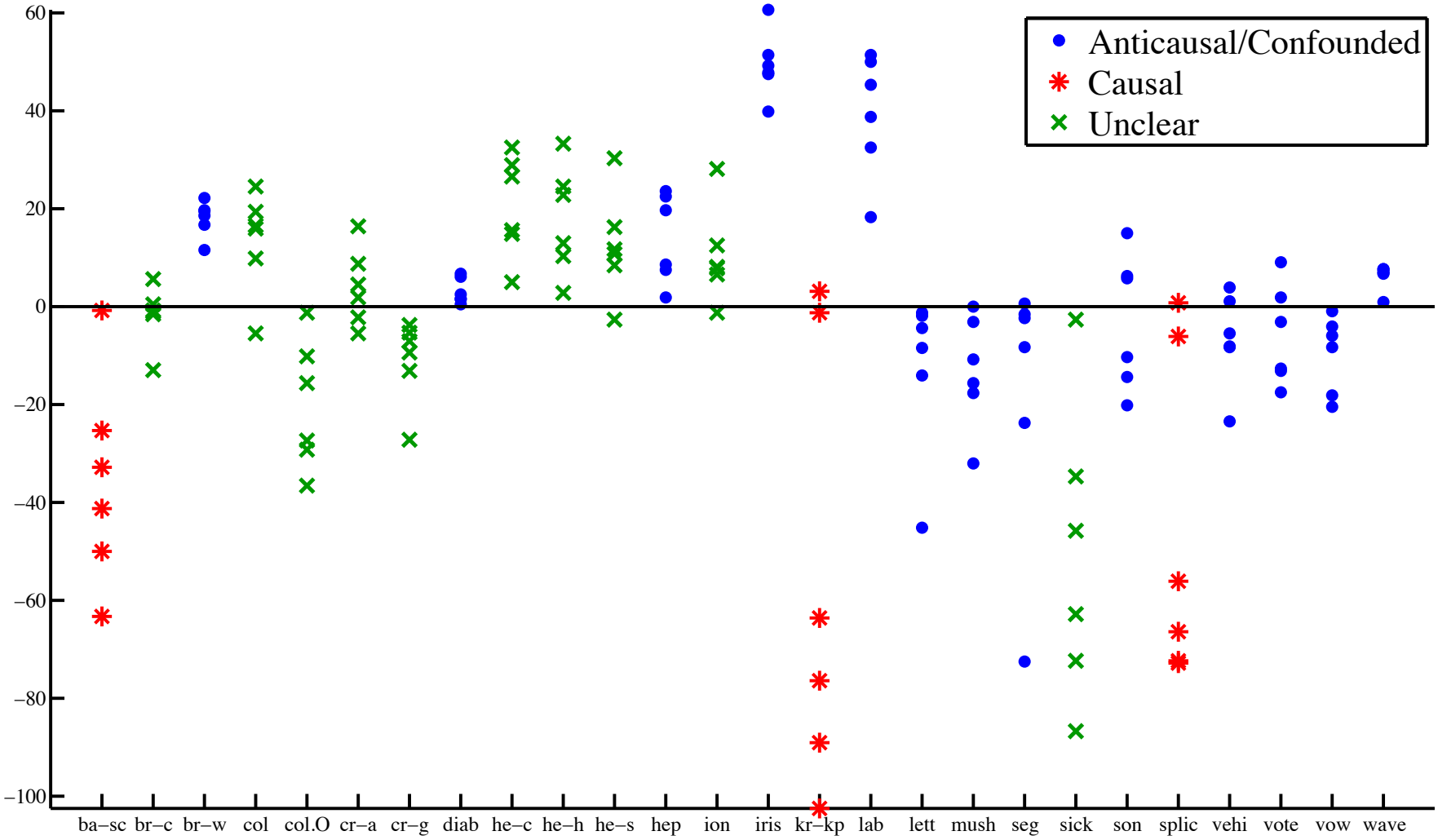
Categ.	Dataset	Target variable	Remark
Anticausal/Confounded	breastTumor	tumor size	causing predictors such as inv-nodes and deg-malig
	cholesterol	cholesterol	causing predictors such as resting blood pressure and fasting blood sugar
	cleveland	presence of heart disease in the patient	causing predictors such as chest pain type, resting blood pressure, and fasting blood sugar
	lowbwt	birth weight	causing the predictor indicating low birth weight
	pbc	histologic stage of disease	causing predictors such as Serum bilirubin, Prothrombin time, and Albumin
	pollution	age-adjusted mortality rate per 100,000	causing the predictor number of 1960 SMSA population aged 65 or older
	wisconsin	time to recur of breast cancer	causing predictors such as perimeter, smoothness, and concavity
Causal	autoMpg	city-cycle fuel consumption in miles per gallon	caused by predictors such as horsepower and weight
	cpu	cpu relative performance	caused by predictors such as machine cycle time, maximum main memory, and cache memory
	fishcatch	fish weight	caused by predictors such as fish length and fish width
	housing	housing values in suburbs of Boston	caused by predictors such as pupil-teacher ratio and nitric oxides concentration
	machine_cpu	cpu relative performance	see remark on "cpu"
	meta	normalized prediction error	caused by predictors such as number of examples, number of attributes, and entropy of classes
	pwLinear	value of piecewise linear function	caused by all 10 involved predictors
	sensory	wine quality	caused by predictors such as trellis
	servo	rise time of a servomechanism	caused by predictors such as gain settings and choices of mechanical linkages
Unclear	auto93 (target: midrange price of cars); bodyfat (target: percentage of body fat); autoHorse (target: price of cars); autoPrice (target: price of cars); baskball (target: points scored per minute); cloud (target: period rainfalls in the east target); echoMonths (target: number of months patient survived); fruitfly (target: longevity of mail fruitflies); pharynx (target: patient survival); pyrim (quantitative structure activity relationships); sleep (target: total sleep in hours per day); stock (target: price of one particular stock); strike (target: strike volume); triazines (target: activity); veteran (survival in days)		

Benchmark Datasets of *Chapelle et al. (2006)*



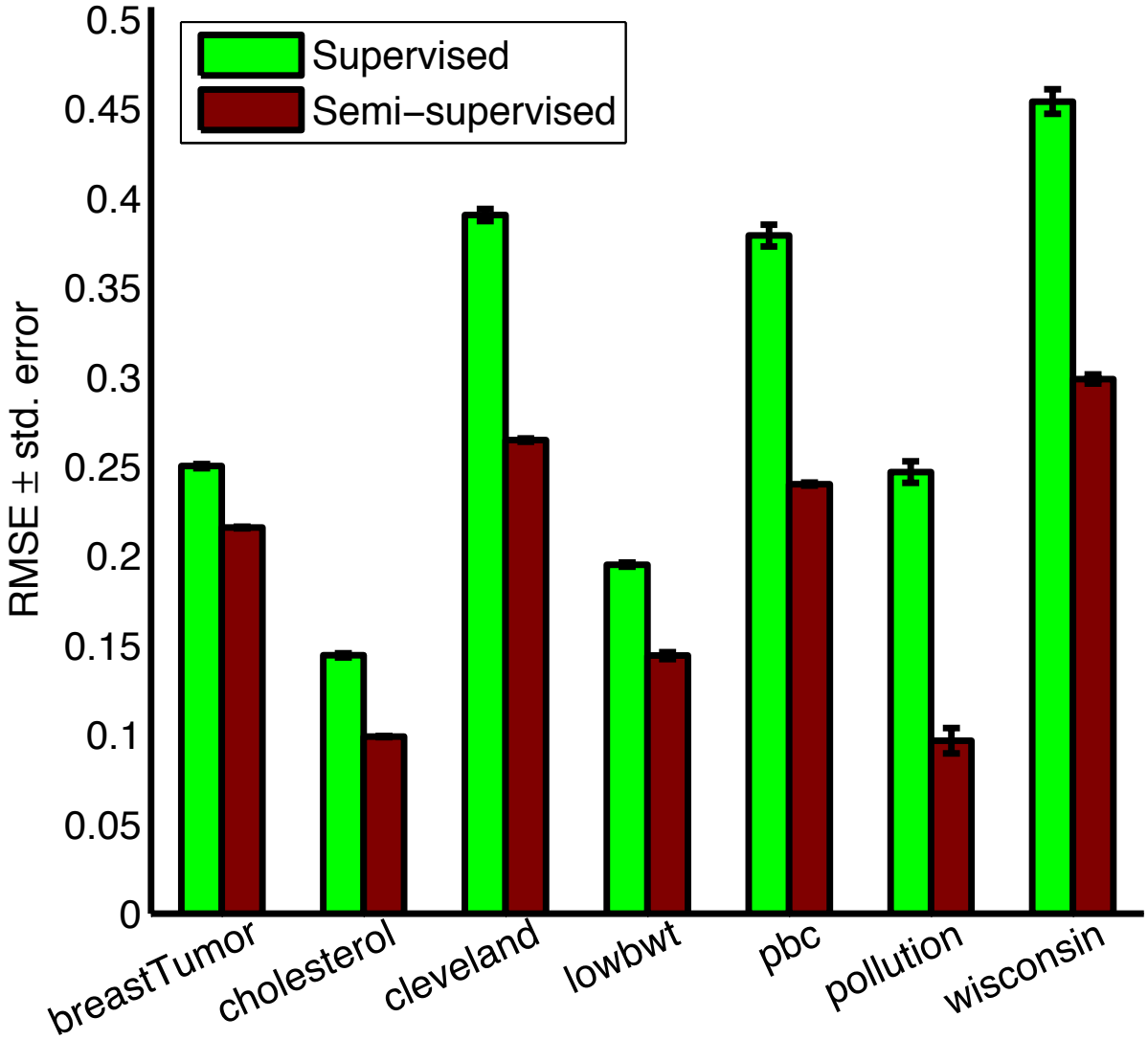
Asterisk = 1-NN, SVM

Self-training does not help for causal problems (cf. *Guo et al., 2010)*

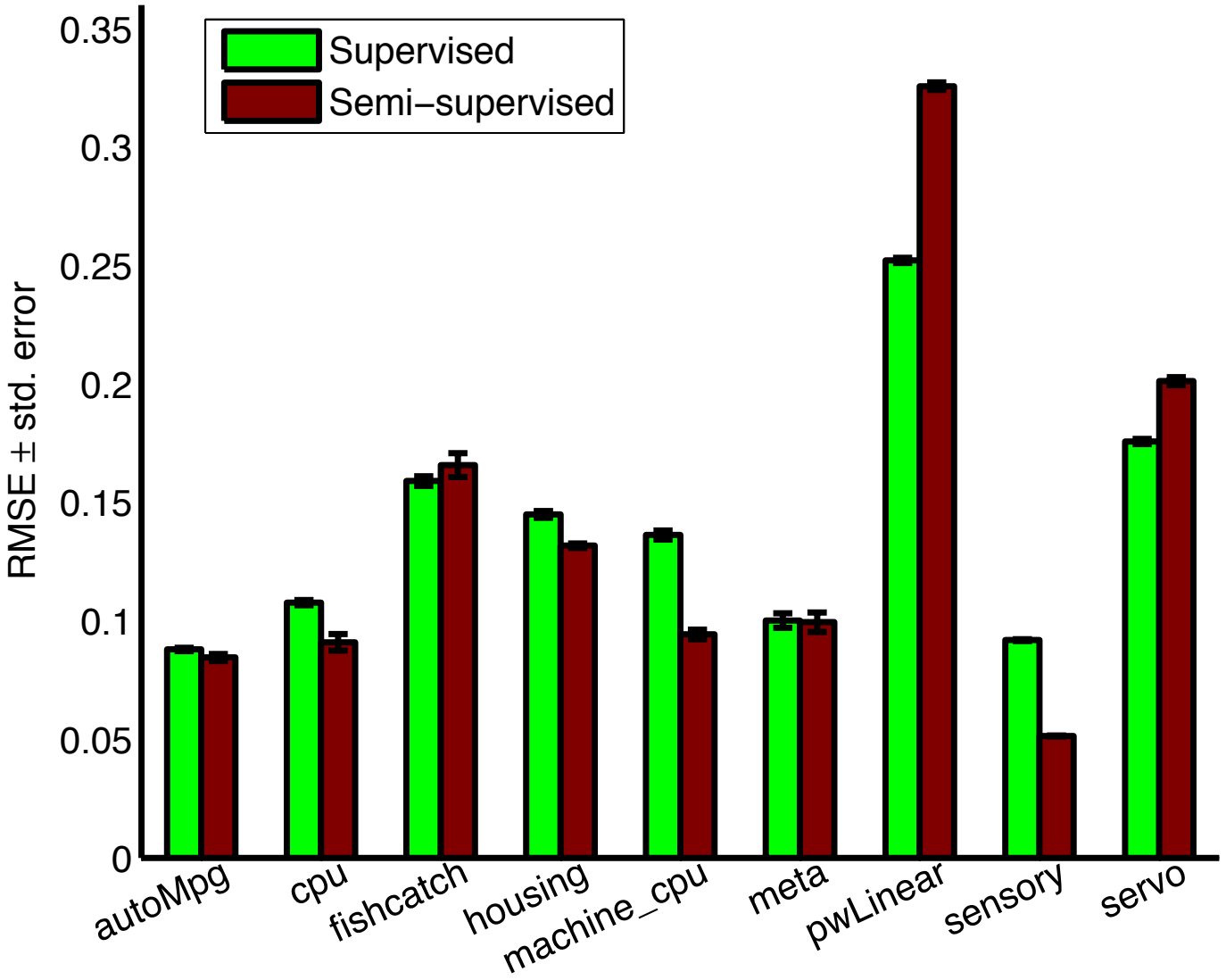


Relative error decrease = (error(base) –error(self-train)) / error(base)

Co-regularization helps for the **anticausal** problems of *Brefeld et al., 2006*



Co-regularizarion hardly helps for the **causal** problems of *Brefeld et al., 2006*



Causality




Dominik Janzing, Jonas Peters, Kun Zhang, Joris Mooij





Moritz Grosse-Wentrup, Michel Besserve, Olivier Stegle, Eleni Sgouritsa, Jakob Zscheischler

Image Deconvolution



Michael Hirsch, Stefan Harmeling, Christian Schuler,

Kernel Means




Arthur Gretton, Kenji Fukumizu, Alex Smola, Bharath Sriperumbudur

The purpose […] is to identify frontiers for collaborative research integrating

(a) mathematical and computational modeling of human cognition with
(b) machine learning and machine intelligence

[…] as an additional objective of this meeting, we are asked to consider the following from the perspective of the computational cognition community:

(a) identify the major obstacles to progress in understanding the brain and
(b) discuss theoretical and experimental approaches to overcome these obstacles

BRAIN project: $1e8

- "give scientists the tools to get a dynamic picture of the brain and better understand how we think, learning, and remember
- possible outcomes:
- Parkinson
- reduce language barriers through technological advances in how computers interface with human thought
- PTSD, brain injuries in war veterans (50% DARPA)
- high-tech jobs

BLUE BRAIN project: EUR 1e9

- Reconstructing the brain piece by piece and building a virtual brain in a supercomputer
- new understanding of the brain and a better understanding of neurological diseases.