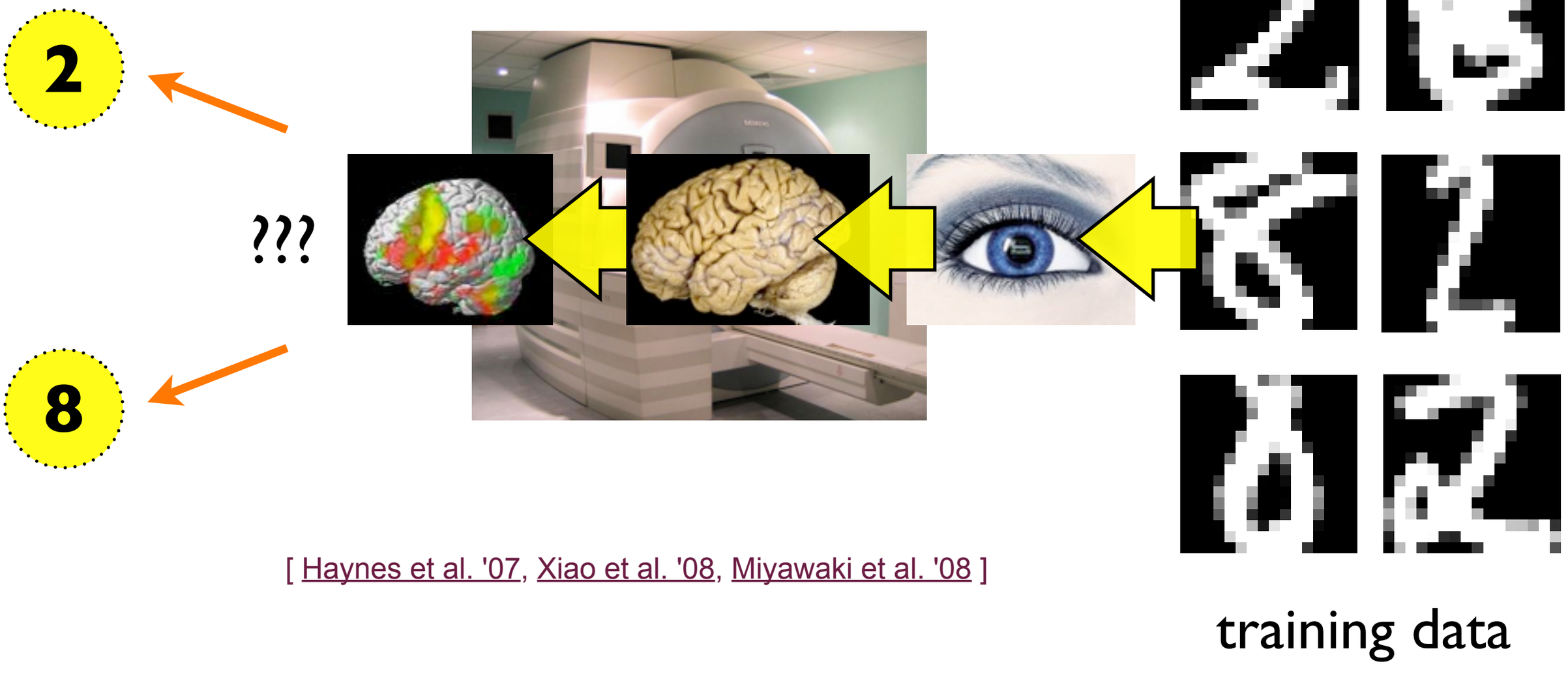
Kernels, Margins, Coresets and The Fight of High- vs. Low-Dimensional Spaces

Mittagsseminar
Martin Jaggi, 26.3.2009

Support Vector Machines

Some weird applications

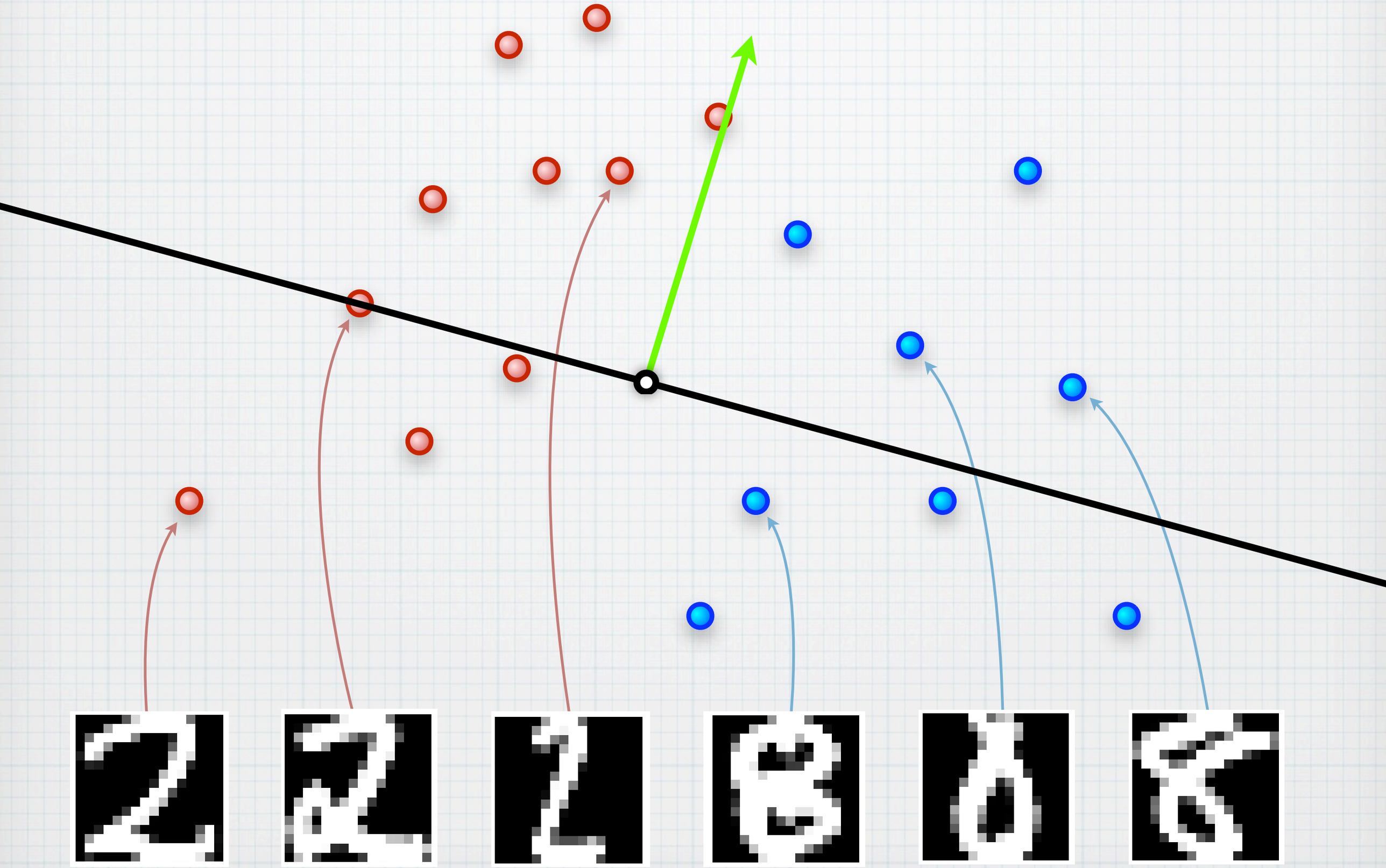


[Haynes et al. '07, Xiao et al. '08, Miyawaki et al. '08]

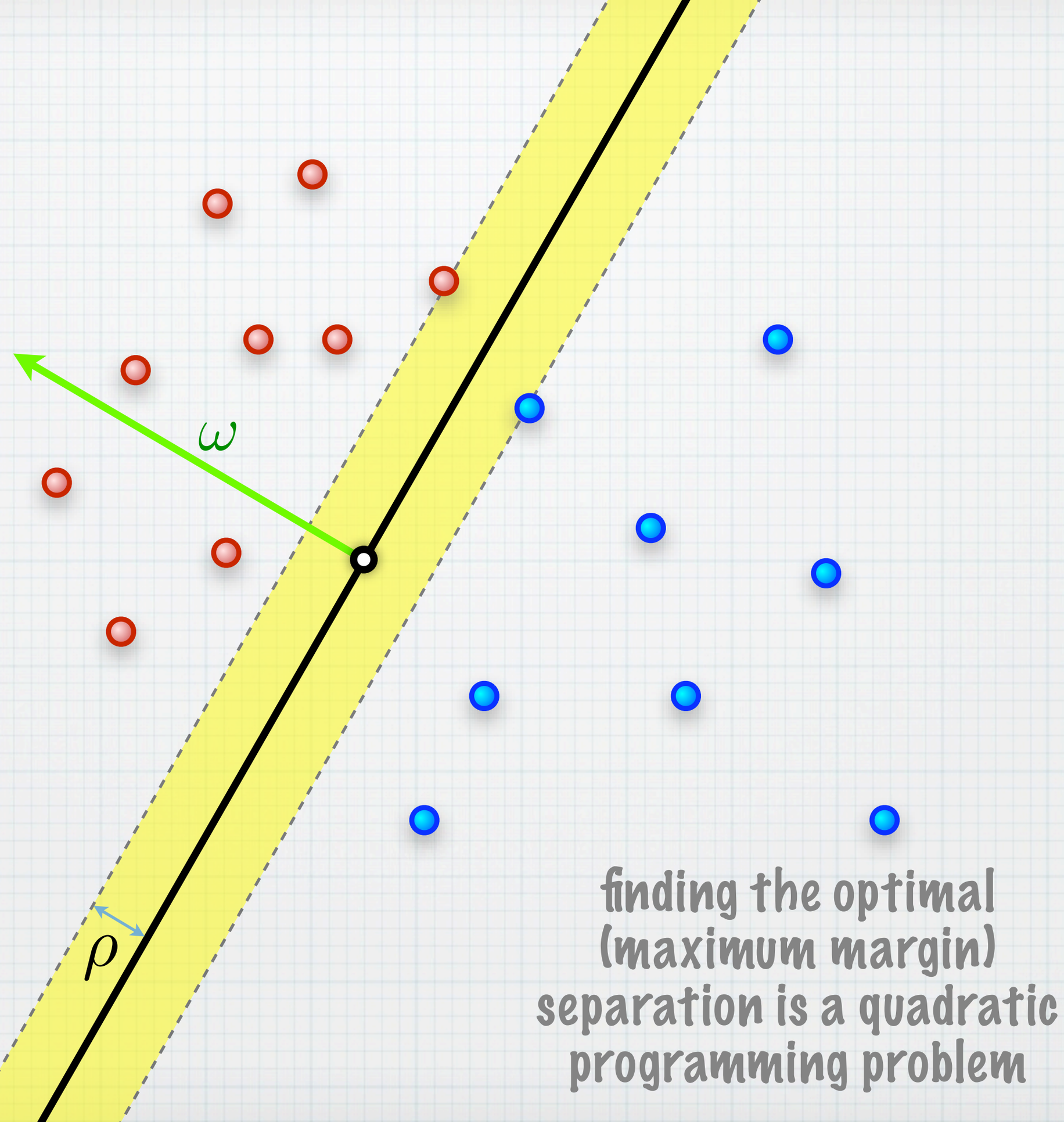
ACM Kanellakis Theory and Practice Award

	Topic	Awarded to
1996	Public key cryptography	Adleman, Diffie, Hellman, Merkle, Rivest, Shamir
1997	Data compression	Lempel, Ziv
1998	Algorithmic model checking	Bryant, Clarke, Emerson, McMillan
1999	Splay tree data structure	Sleator, Tarjan
2000	Polynomial time interior point methods for LP	Karmarkar
2001	Genome sequencing algorithms	Myers
2002	Constrained channel coding	Franaszek
2003	Randomized primality tests	Miller, Rabin, Solovay, Strassen
2004	AdaBoost machine learning algorithm	Freund, Schapire
2005	Formal verification	Holzmann, Kurshan, Vardi, Wolper
2006	Logic synthesis and simulation of electronic systems	Brayton
2007	Gröbner bases for computer algebra	Buchberger
2008	Support vector machine	Cortes, Vapnik

$\mathbb{R}^{256}$



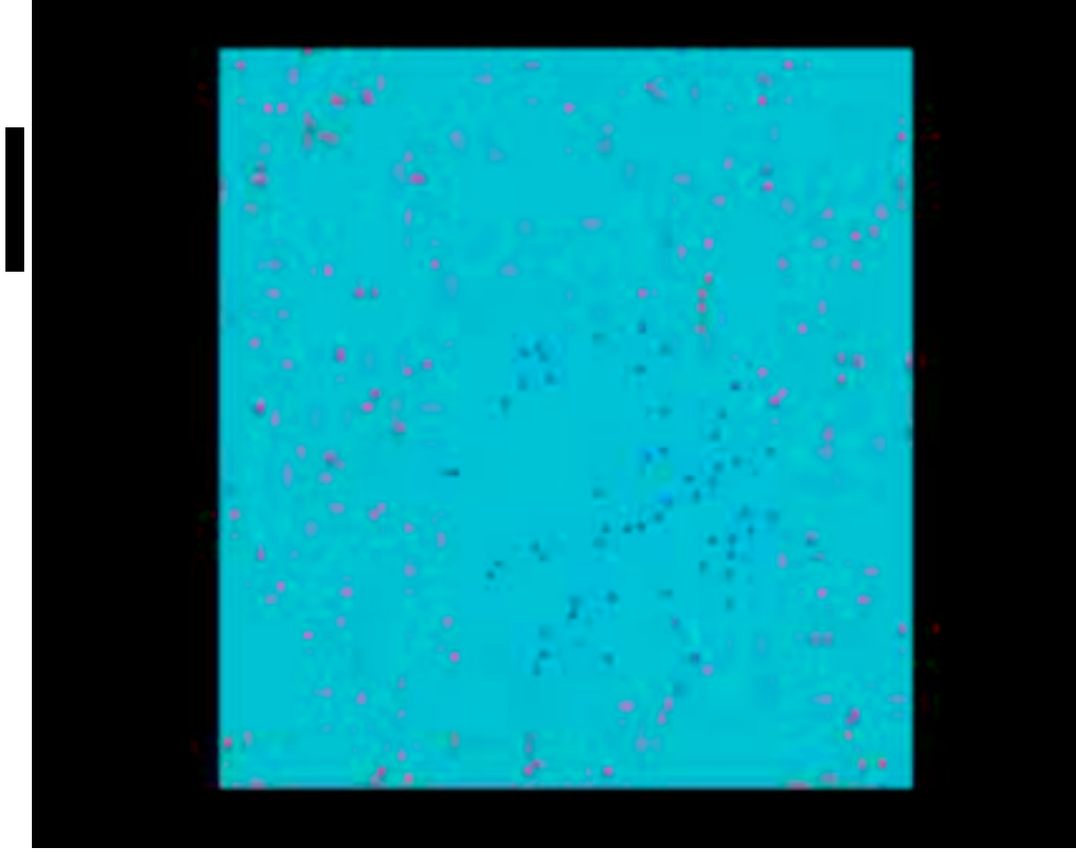
$\mathbb{R}^{256}$



finding the optimal (maximum margin) separation is a quadratic programming problem

The Kernel

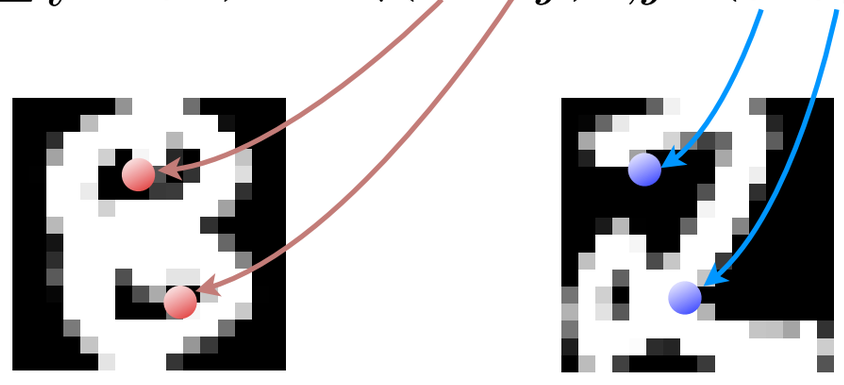
$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$



Polynomial Kernel

$$K(x, y) = \langle x, y \rangle^d, \quad d \in \mathbb{N}$$

$$\langle x, y \rangle^2 = (\sum_i x_i y_i)(\sum_i x_i y_i) = \langle (x_i x_j)_{i,j}, (y_i y_j)_{i,j} \rangle$$



Gauss Kernel

$$K(x, y) = e^{-\frac{1}{2\sigma^2}\|x-y\|^2}$$

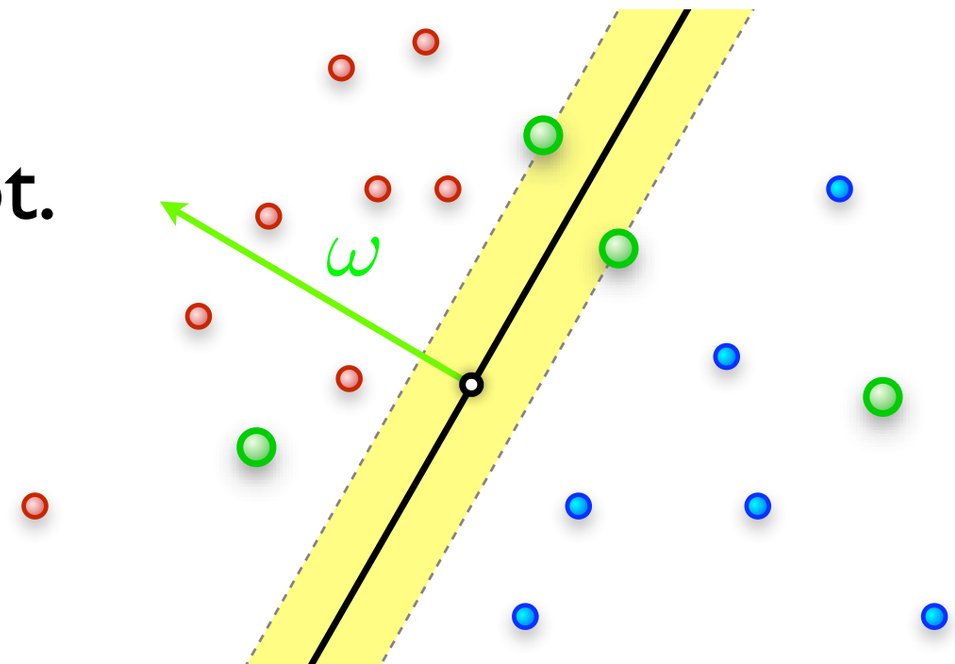
The Kernel Trick

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

- We only have black-box access to scalar products.

- The solution vector ω is built as a linear combination of the points (hopefully very few points).

- Most geometric algorithms work no matter if a kernel is used or not.



Why does it work well?

☐ „Because it finds a very ‚sophisticated' separating hyperplane in a very cool very high-dimensional space"

☑ „Because no matter which space (or kernel) is used, *if* the points are separable by large margin, then this is an easy and low-dimensional problem anyway"

Aim of this talk:

Any number of points in an arbitrary dimensional space can efficiently be projected down to a $\Theta\left(\frac{1}{\rho^2}\right)$ dimensional space, preserving separability by large margin.

	randomized	deterministic
upper bound	$O\left(\frac{1}{\rho^2}\right)$	$O\left(\frac{1}{\rho^2}\right)$
lower bound	$\Omega\left(\frac{1}{\rho^2}\right)$	

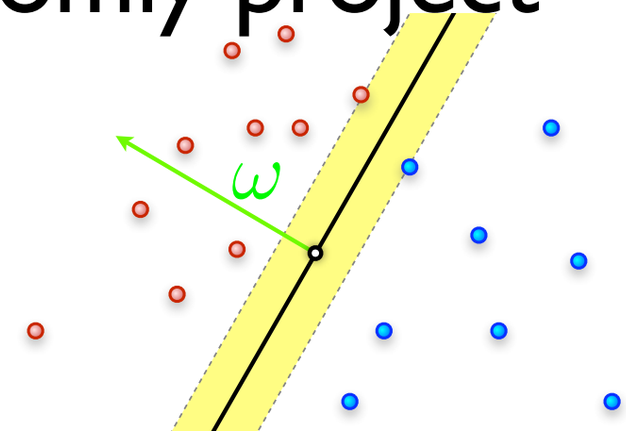
If we have found such a low-dimensional space then we can forget about the high-dimensional original space.

„Feature Selection“

A low-dimensional Interpretation

Random Projections

Are the points still separable if we randomly project down to a low-dimensional space?



Trying to apply the Johnson-Lindenstrauss lemma

For a random projection P from any $\mathbb{R}^{\text{high}}$ down to $\mathbb{R}^d$, it holds that

$$\mathbb{P}\Big[\begin{array}{c}(1-\varepsilon)\|x-y\|^2 \\ \wedge \\ \|Px-Py\|^2 \\ \wedge \\ (1+\varepsilon)\|x-y\|^2\end{array}\Big] \geq 1 - 2e^{-(\varepsilon^2-\varepsilon^3)\frac{d}{4}}$$

Projecting down to dimension $d := \Theta\left(\frac{\log n}{\rho^2}\right)$ would preserve separation.

Problem: Such mappings are not computable in our case, as we have no ortho-normal basis of our space.

Solution: An alternative random projection:

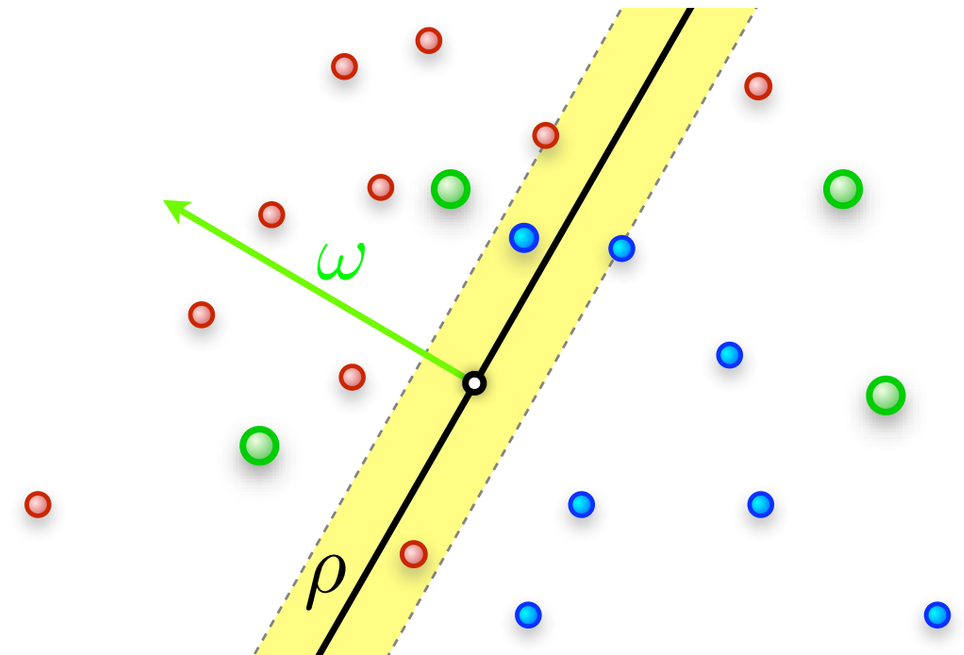
Theorem [Balcan,Blum,Vempala '06]

$0 < \alpha, \beta < 1$

Randomly pick a $d := \frac{8}{1-\alpha}\left(\frac{1}{\rho^2} + \log\frac{1}{1-\beta}\right)$ -subset of from our point set.

Let ω be a random unit vector in the span of these d points. Then with probability $\geq \beta$, ω separates at least an α -fraction of all points by margin at least $\rho/2$.

(Use Gram-Schmidt on to get an ONB)

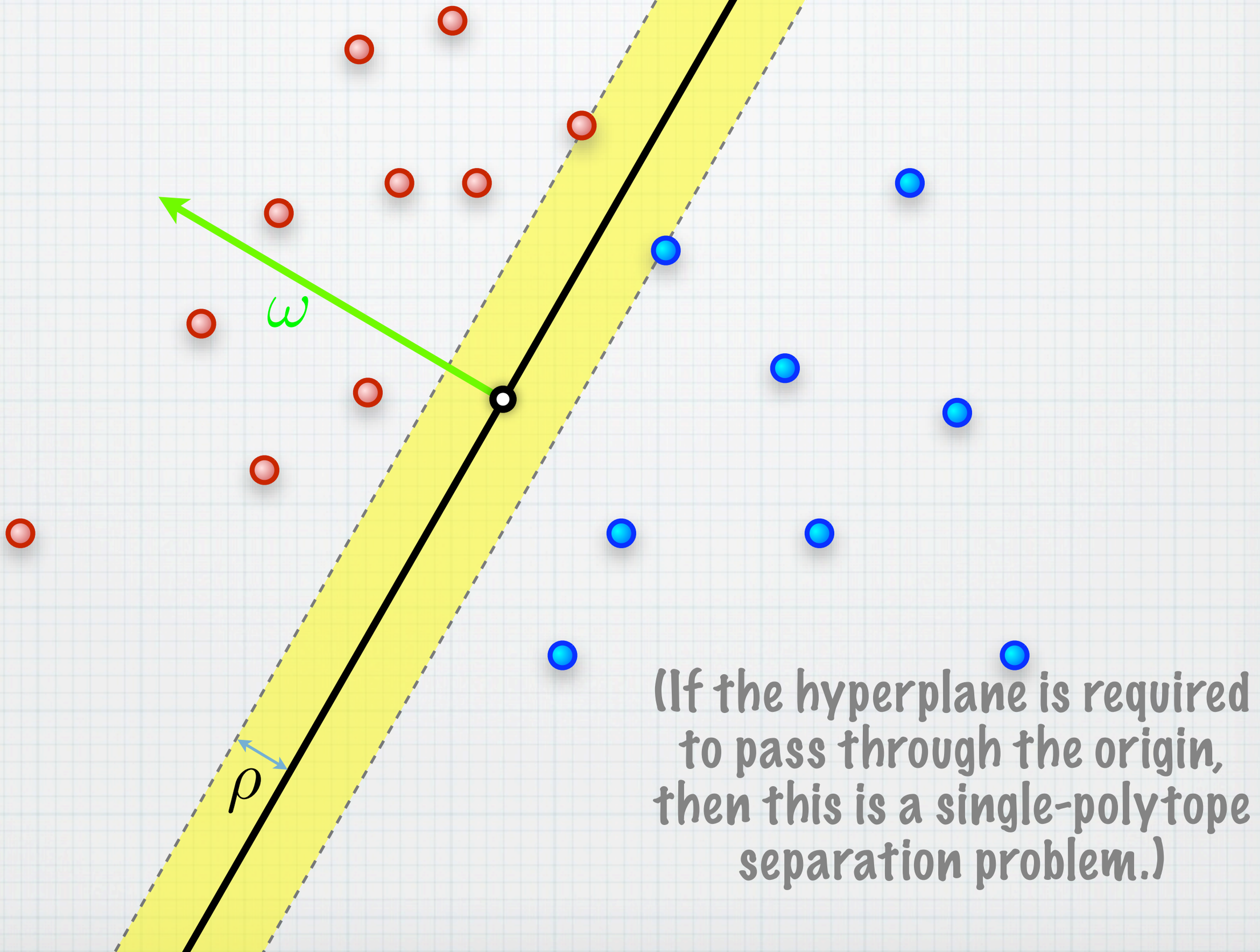


Another low-dimensional Interpretation

Coresets

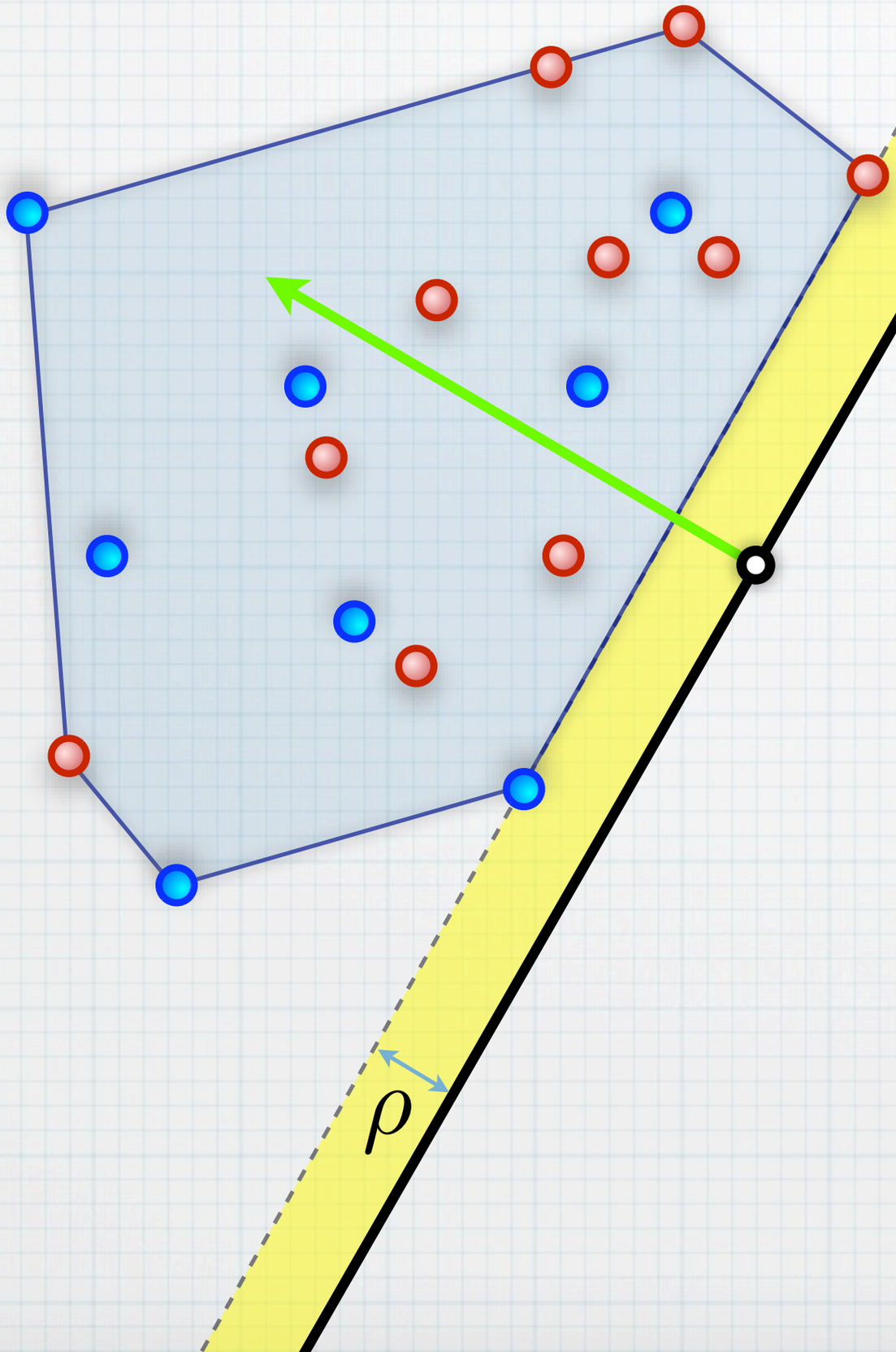
Is a very small subset of the points always sufficient to represent a good solution?

One or Two Polytopes?



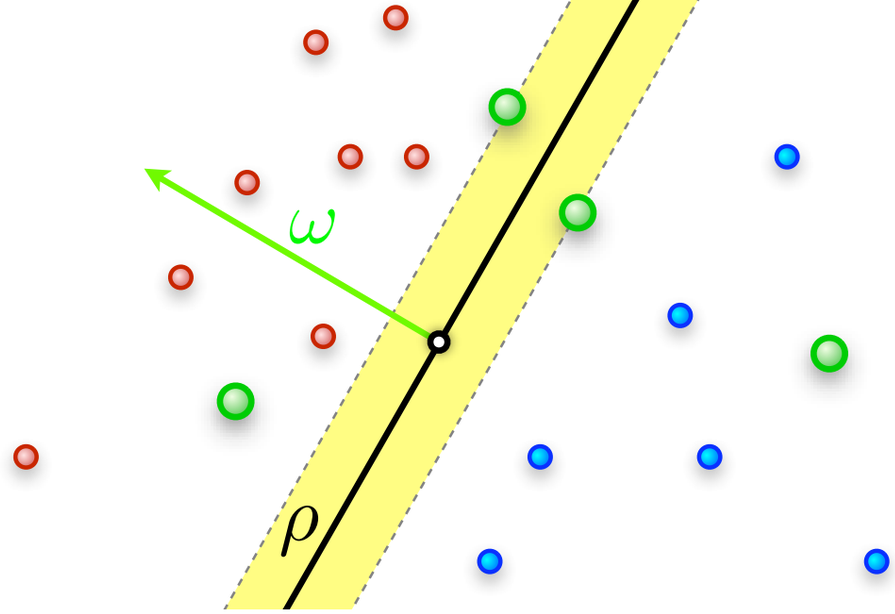
(If the hyperplane is required to pass through the origin, then this is a single-polytope separation problem.)

One or Two Polytopes?



(If the hyperplane is required to pass through the origin, then this is a single-polytope separation problem.)

Coresets for polytope distance:



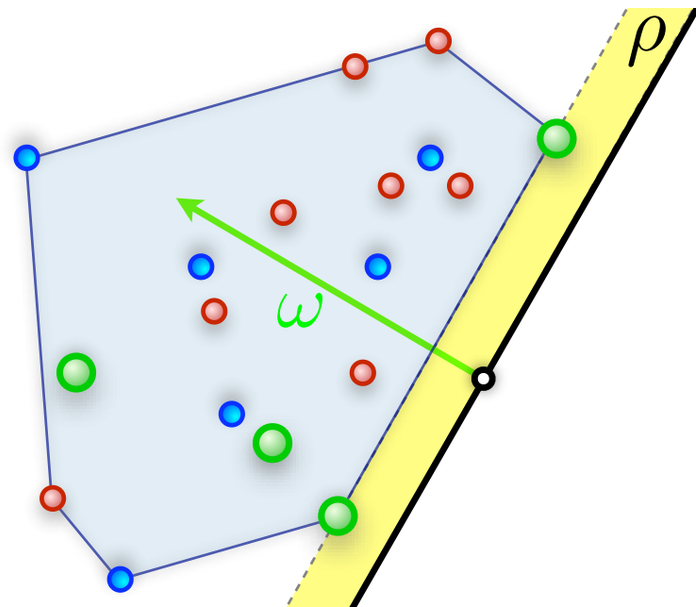
Theorem [Clarkson '08 , Gärtner, Jaggi '09]

Given the point set, a simple greedy algorithm finds a linear combination ω of just $2\left\lceil \frac{4D^2}{\rho^2} \right\rceil$ points, separating all points by margin at least $\rho/2$.

(Where $D \leq 2$ is the diameter of the polytope)

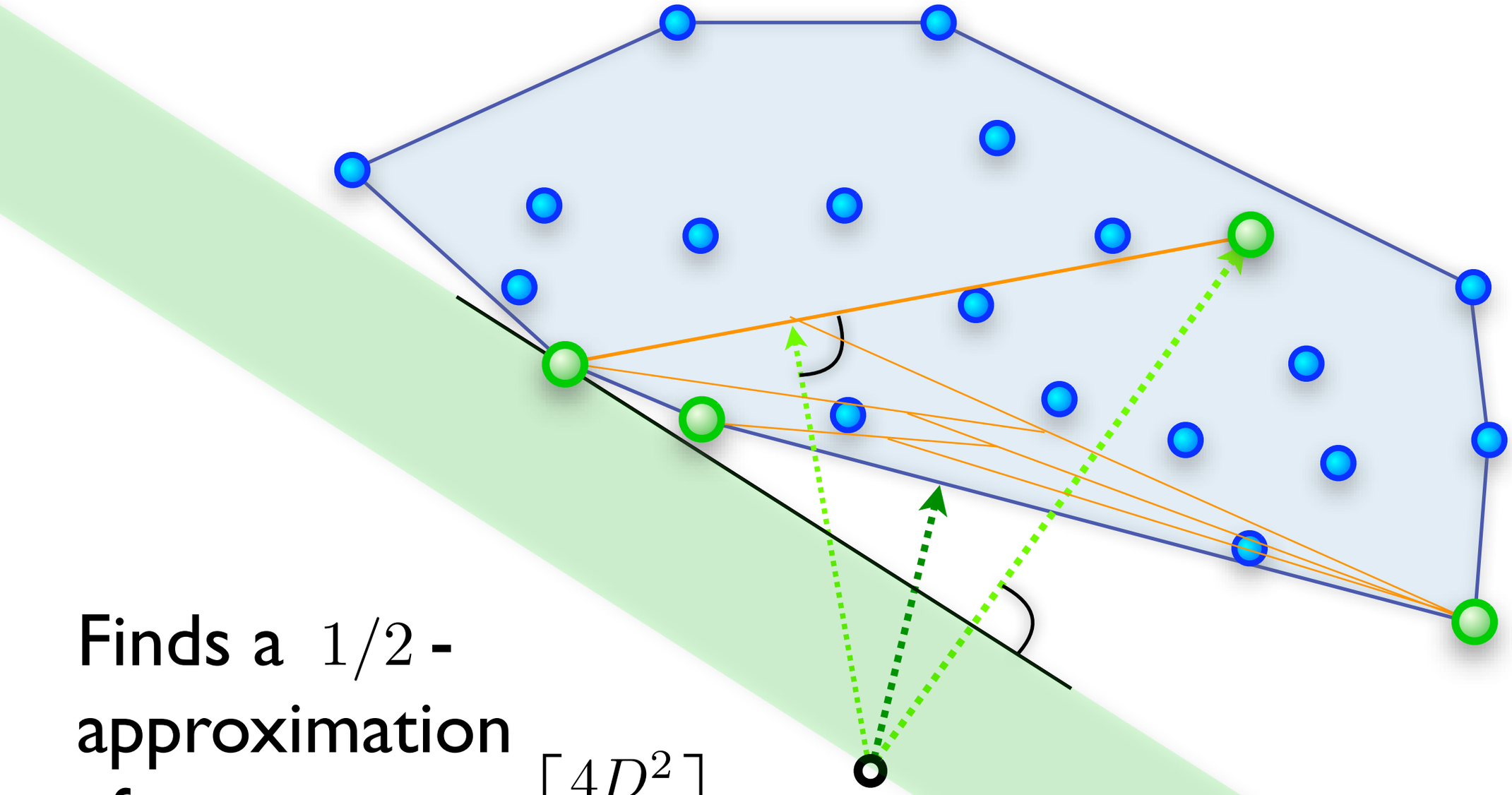
Theorem

Existence: $\left\lceil \frac{2D^2}{\rho^2} \right\rceil$ points always suffice.



[Gilbert '66]

Gilbert‘s Algorithm

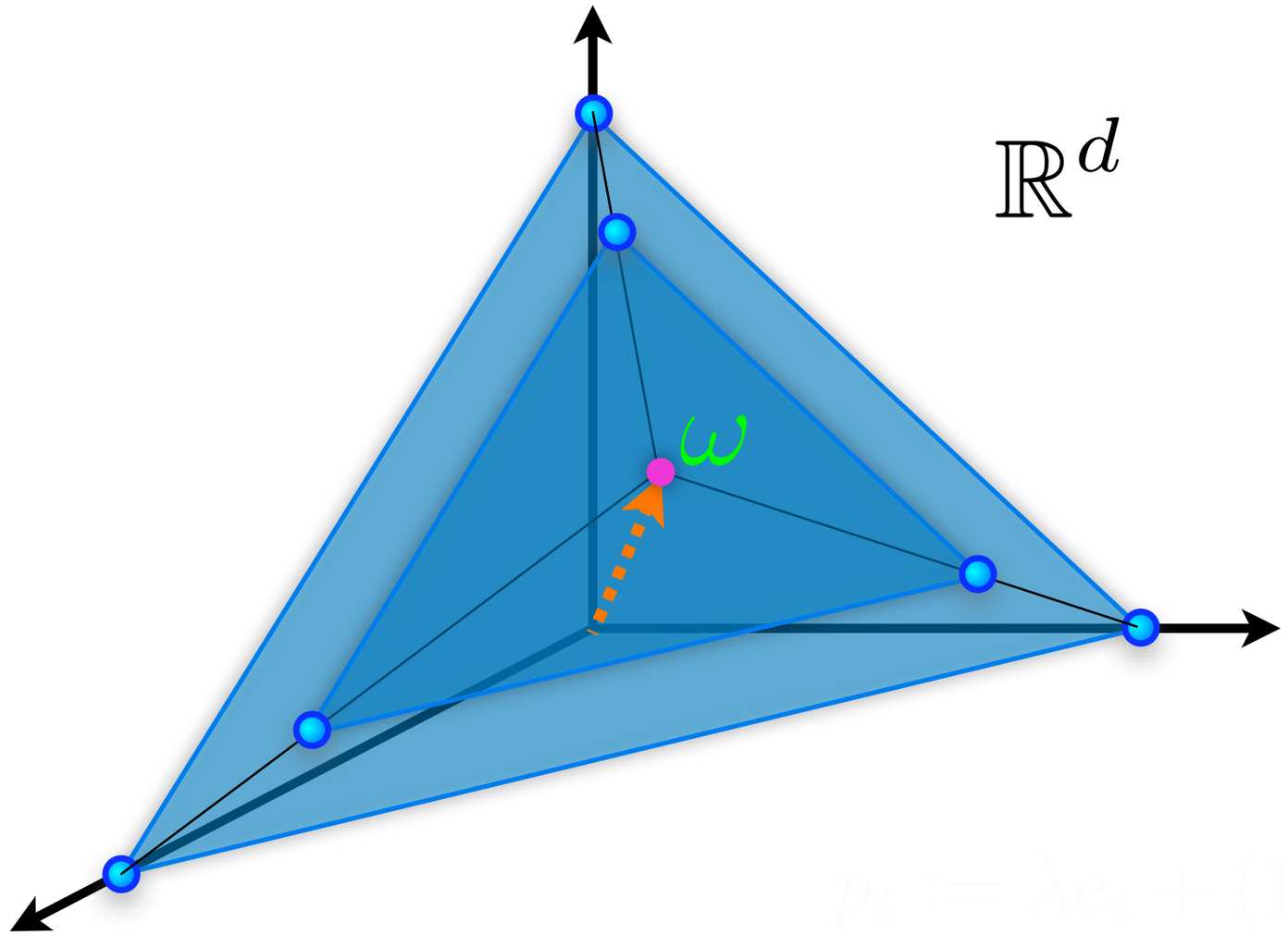


Finds a $1/2$ - approximation after at most $2\left\lceil \frac{4D^2}{\rho^2} \right\rceil$ many steps.

[Clarkson '08 , Gärtner, Jaggi '09]

Lower bound

There are arbitrary sized point sets for which any linear combination that separates with at least $\rho/2$, needs at least $\left\lceil \frac{D^2}{\rho^2} \right\rceil$ many points.



[Gärtner, Jaggi '09]

Conclusions

	randomized	deterministic
upper bound	$O\left(\frac{1}{\rho^2}\right)$	$O\left(\frac{1}{\rho^2}\right)$
lower bound	$\Omega\left(\frac{1}{\rho^2}\right)$	

- If we have a large margin, then finding a good separating hyperplane is a *low*-dimensional problem.
- Nearly matching upper and lower bounds for the dimension that is necessary.
- Finding a good kernel is roughly the same as *feature selection* (finding a good new low-dimensional feature space).