Verteiltes Machine Learning:

Klassifikation und Regression auf grossen Datenmengen

Martin Jaggi
ETH Zurich

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SPINNINGBYTES

Big Data Workshop: Squeezing more out of Data, FFHS, 11th June 2015 *monthly*

Zürich Machine Learning and Data Science

Meetup®

[Link to Website]

Maschinelles Lernen?

(Vorhersage)

Klassifikation & Regression

Maschinelles Lernen?

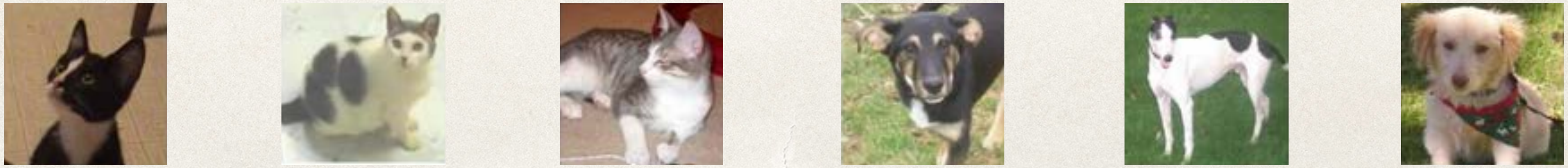
(Vorhersage)

Klassifikation & ~~Regression~~

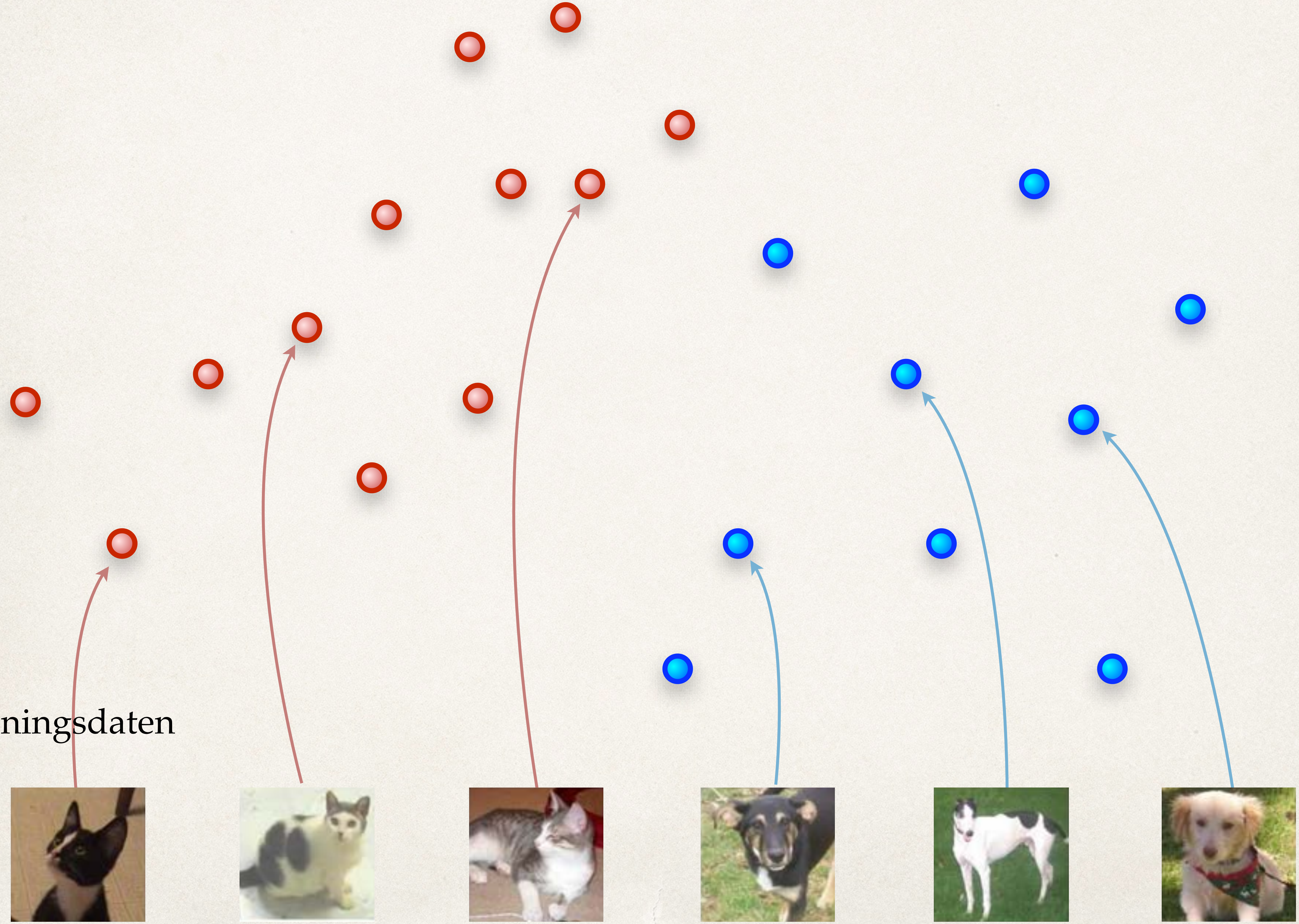
CAT
vs
DOG

Klassifikation

Trainingsdaten

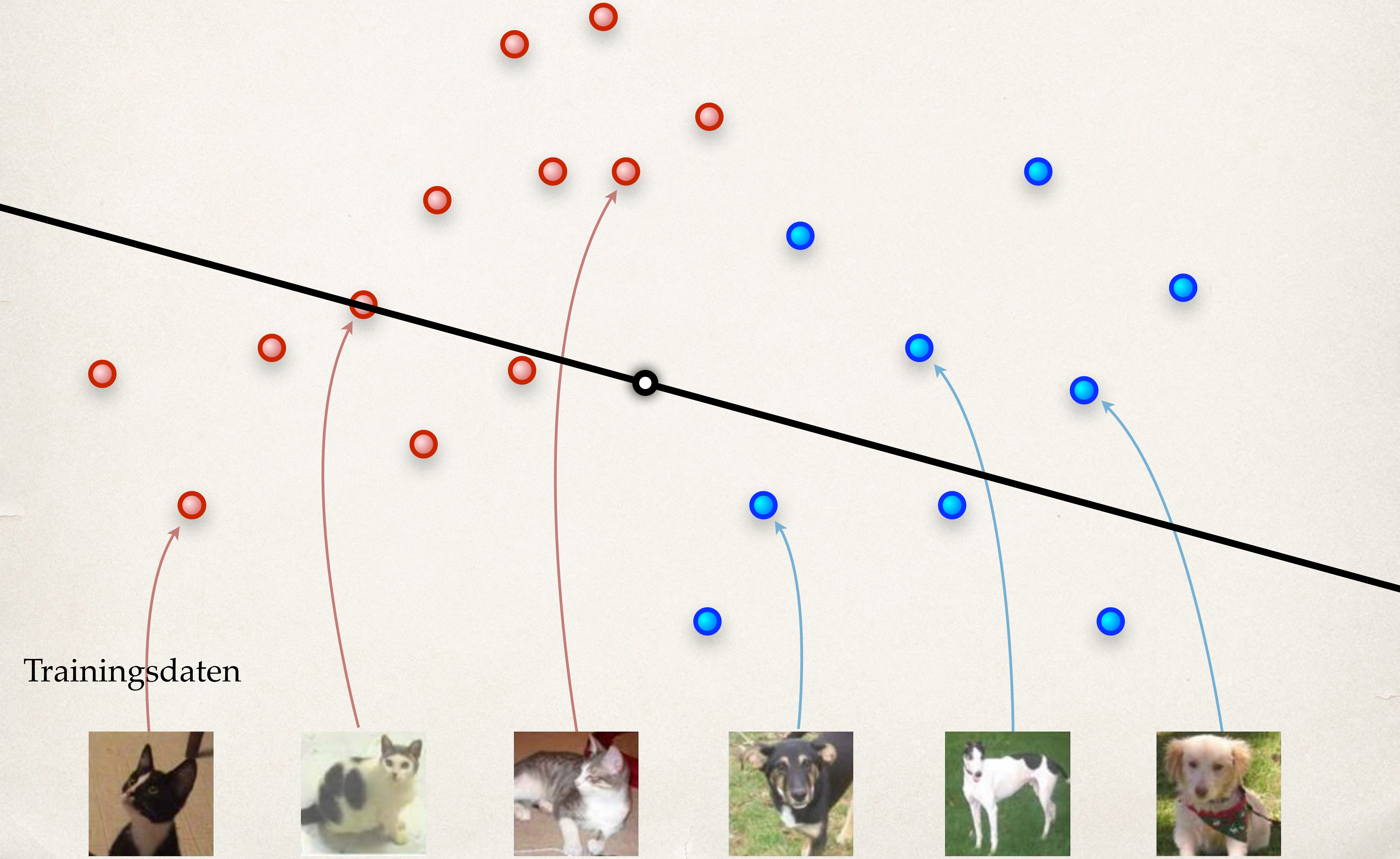


Klassifikation



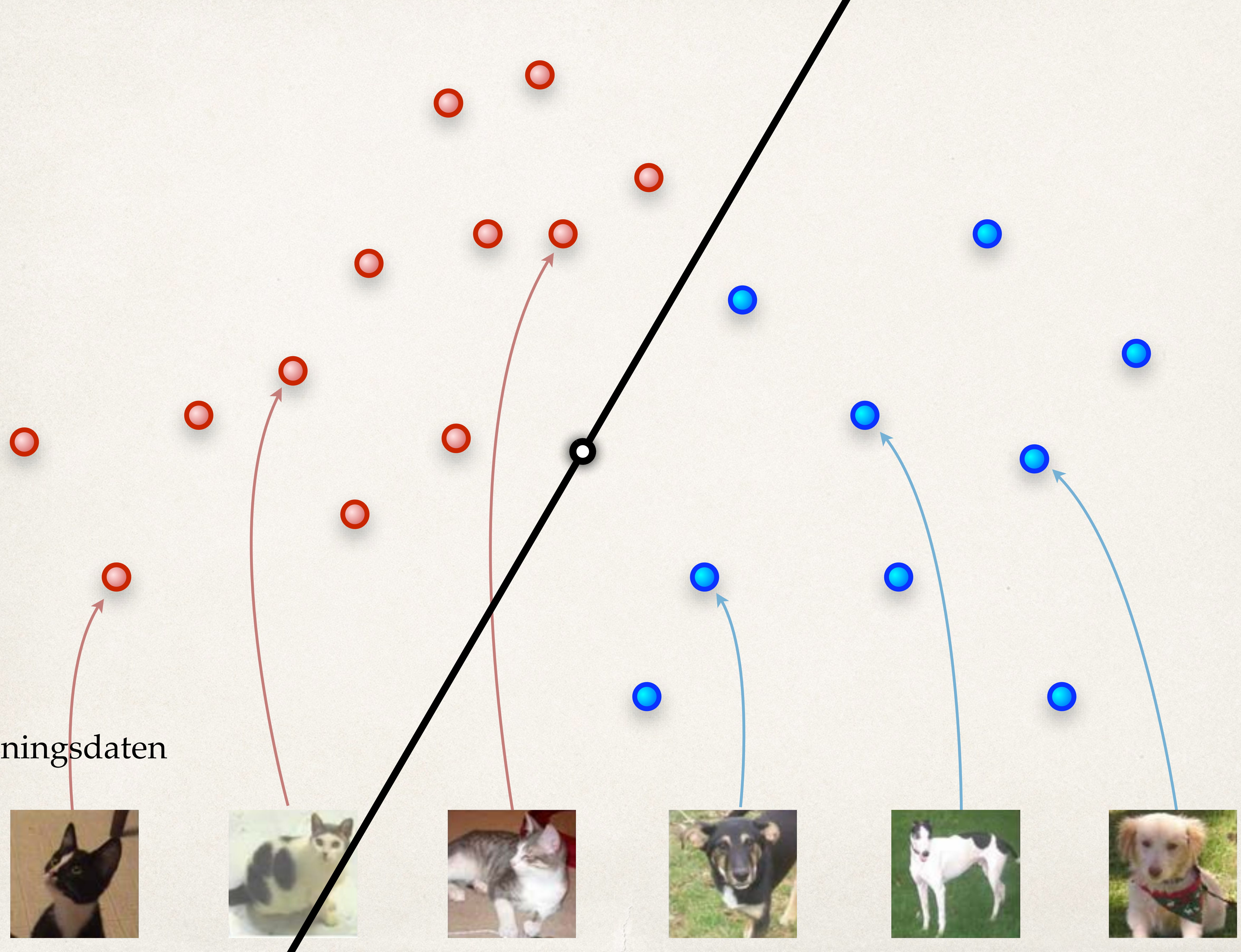
Trainingsdaten

Klassifikation



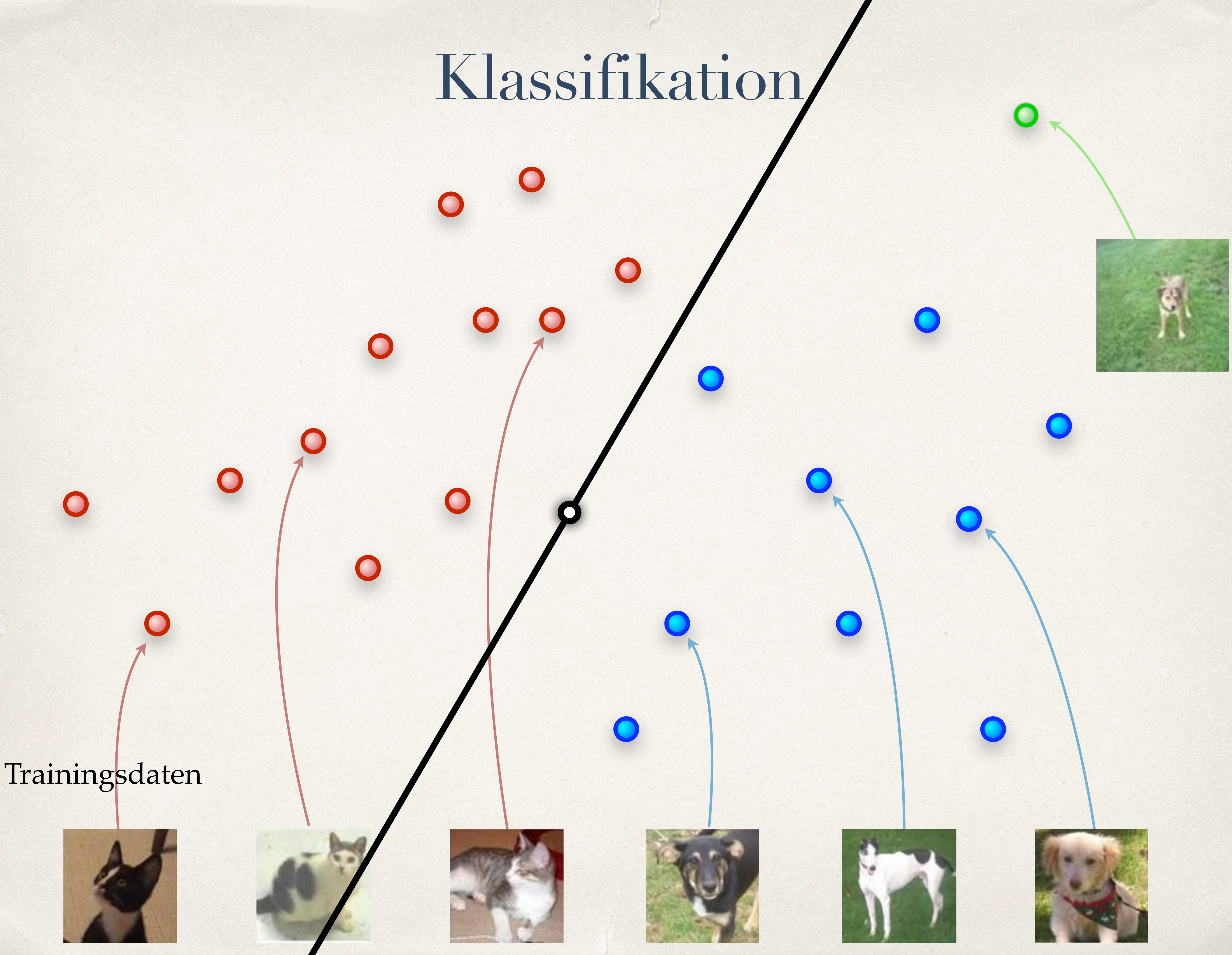
Trainingsdaten

Klassifikation

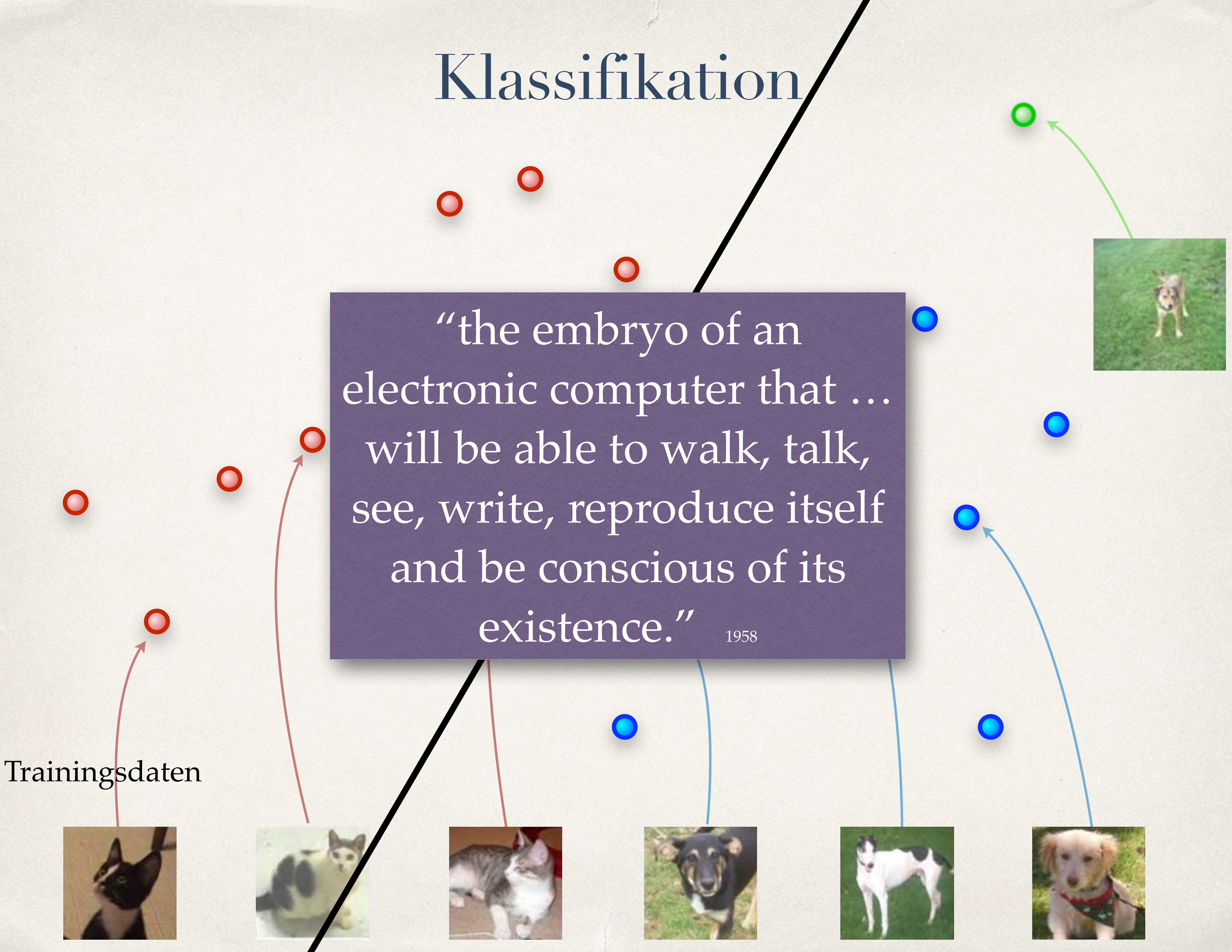


Trainingsdaten

Klassifikation



Klassifikation

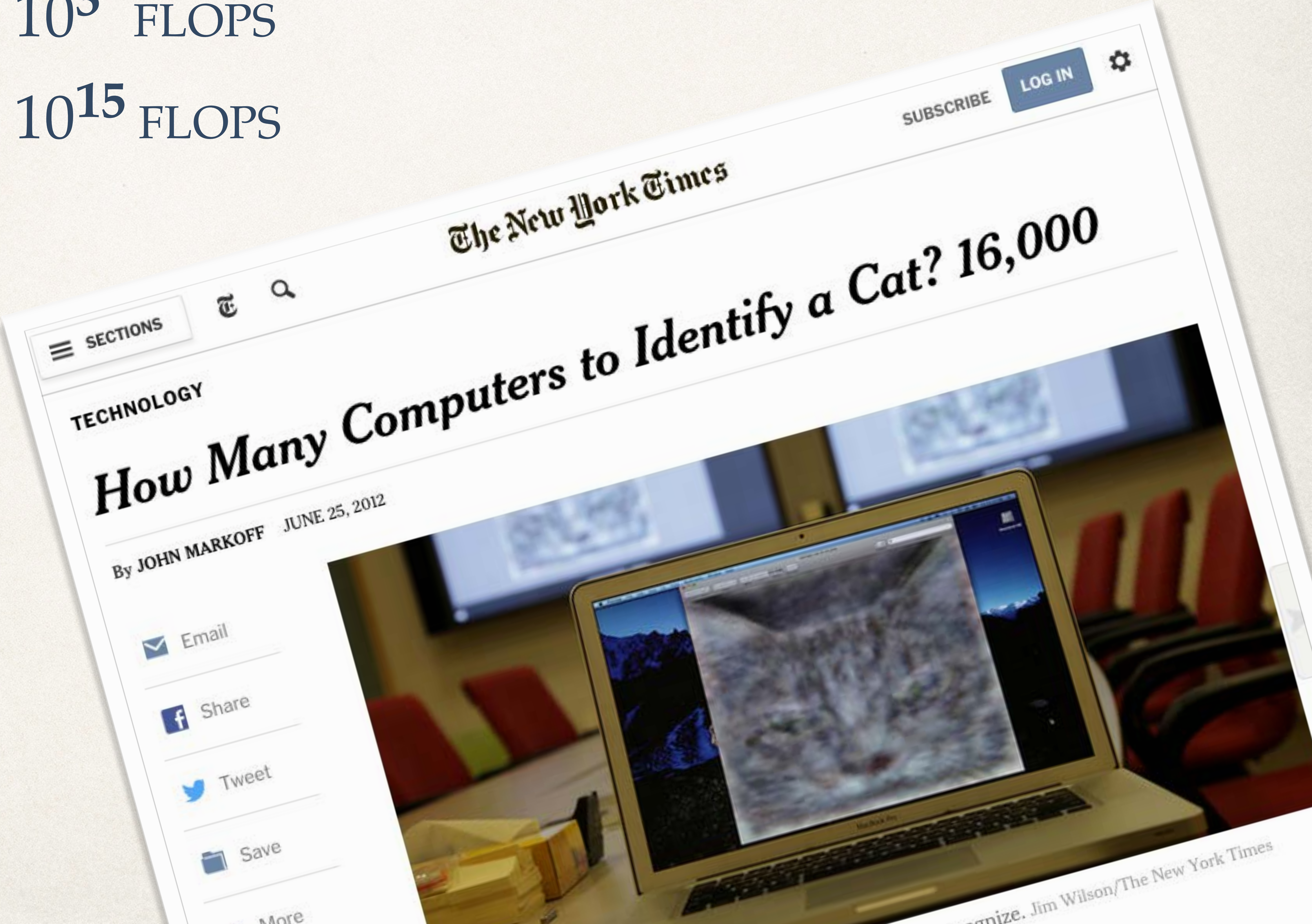


> “the embryo of an electronic computer that … will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.” 1958

Computing Performance:

1950s: 10^3 FLOPS

2010s: 10^{15} FLOPS



TECHNOLOGY

How Many Computers to Identify a Cat? 16,000

By JOHN MARKOFF JUNE 25, 2012

gnize. Jim Wilson/The New York Times

Maschinelles Lernen?

Einige aktuelle Anwendungen / Big Data

Bild-Daten

- Astronomie
- Gesichtserkennung
- 2D + 3D Medizin
- (Hand)schrift-Erkennung
- Bilderkennung
- self-driving cars

Bild-Daten

- Astronomie
- Gesichtserkennung
- 2D + 3D Medizin
- (Hand)schrift-Erkennung
- Bilderkennung
- self-driving cars

Bild-Daten

- Astronomie
- Gesichtserkennung
- 2D + 3D Medizin
- (Hand)schrift-Erkennung
- Bilderkennung
- self-driving cars

Bild-Daten

- Astronomie
- Gesichtserkennung
- 2D + 3D Medizin
- (Hand)schrift-Erkennung
- Bilderkennung
- self-driving cars

kaggle

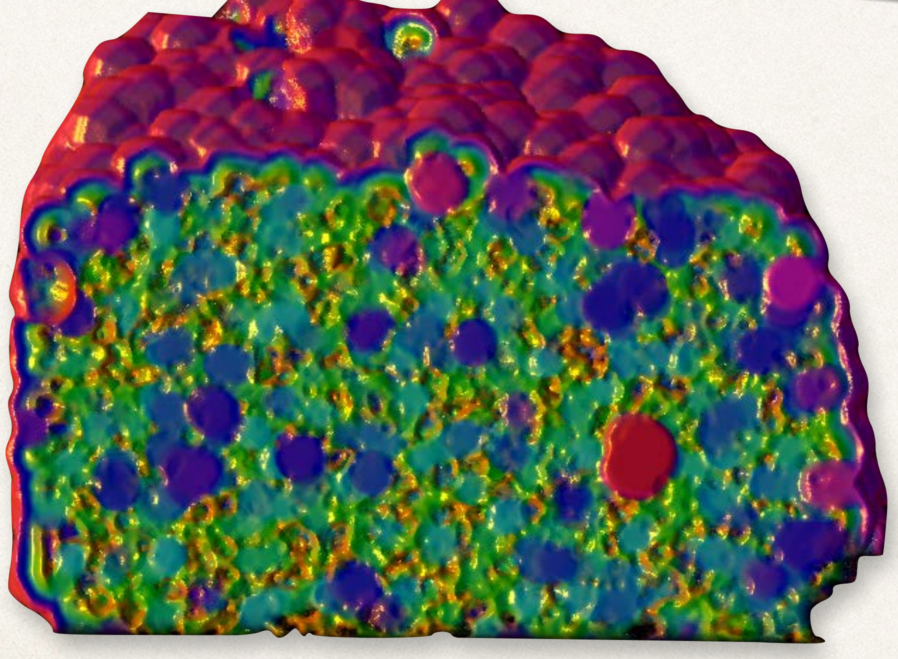


Bild-Daten

- Astronomie
- Gesichtserkennung
- 2D + 3D Medizin
- (Hand)schrift-Erkennung
- Bilderkennung
- self-driving cars

how-old.net

How-Old.net

HOW OLD DO I LOOK? #HowOldRobot

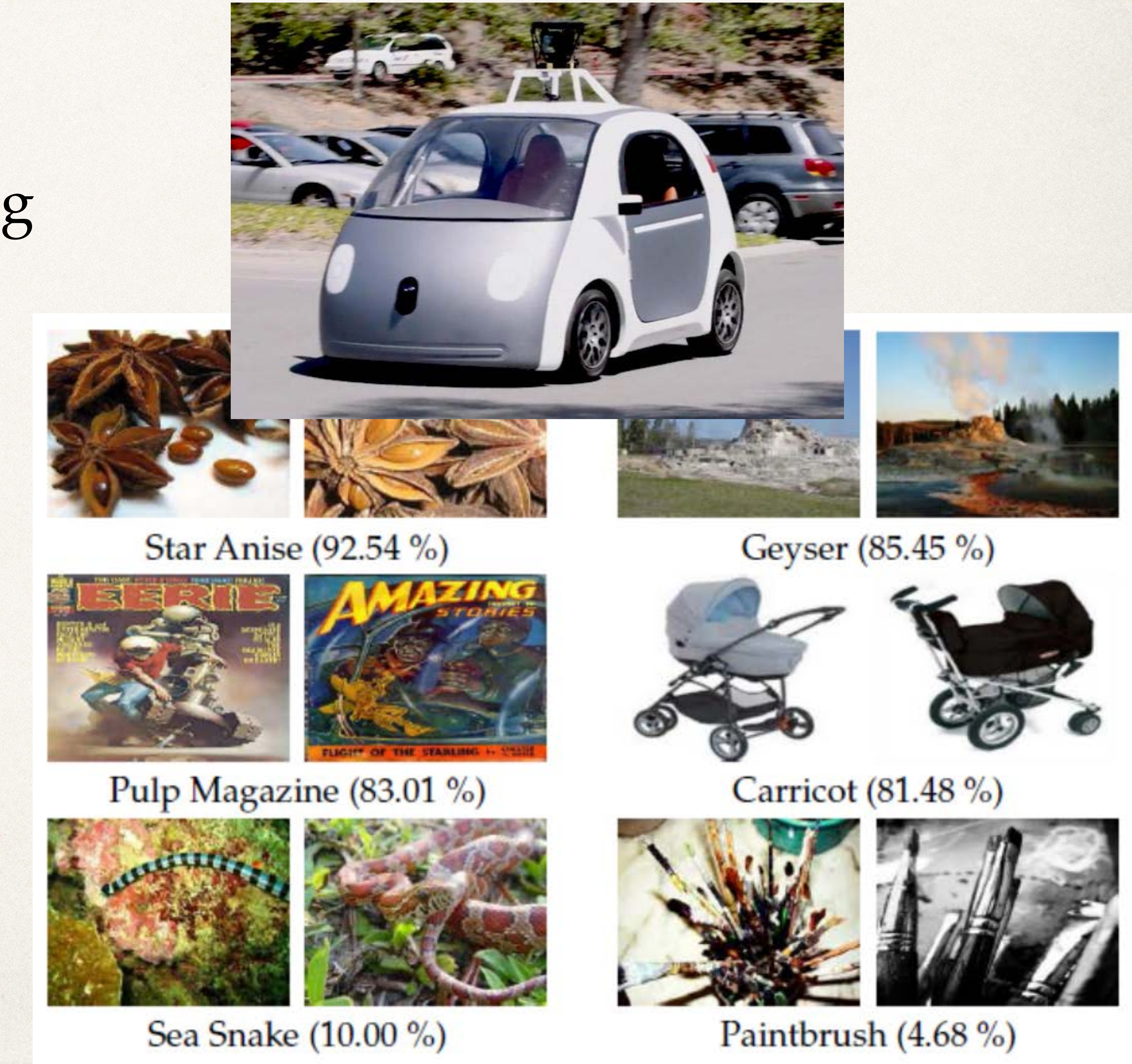


Bild-Daten

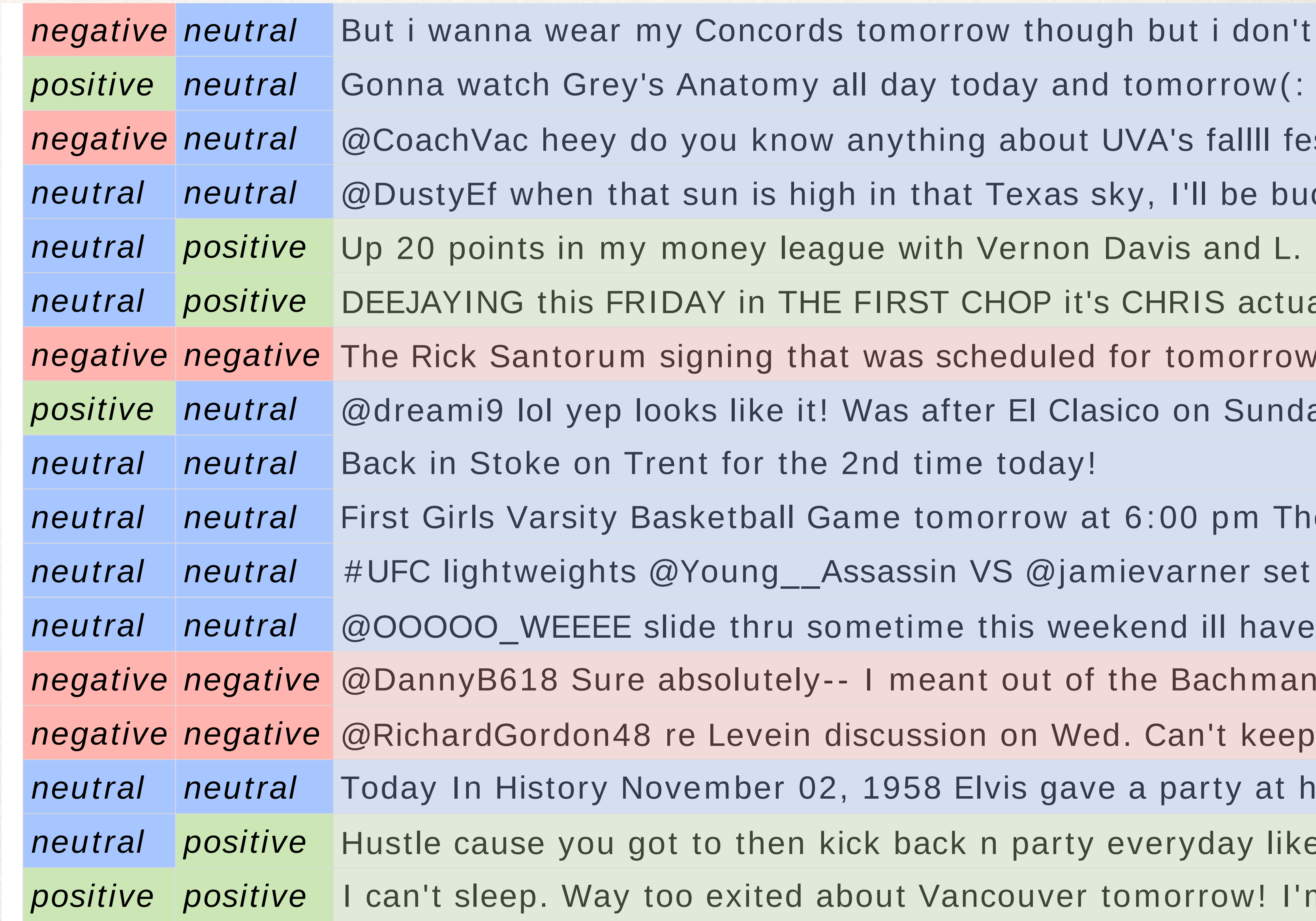
- Astronomie
- Gesichtserkennung
- 2D + 3D Medizin
- (Hand)schrift-Erkennung
- Bilderkennung
- self-driving cars

how-old.net

How-Old.net
HOW OLD DO I LOOK? #HowOldRobot

Text-Daten

- Spam
- Internet-Daten
- Medizin: Gendaten

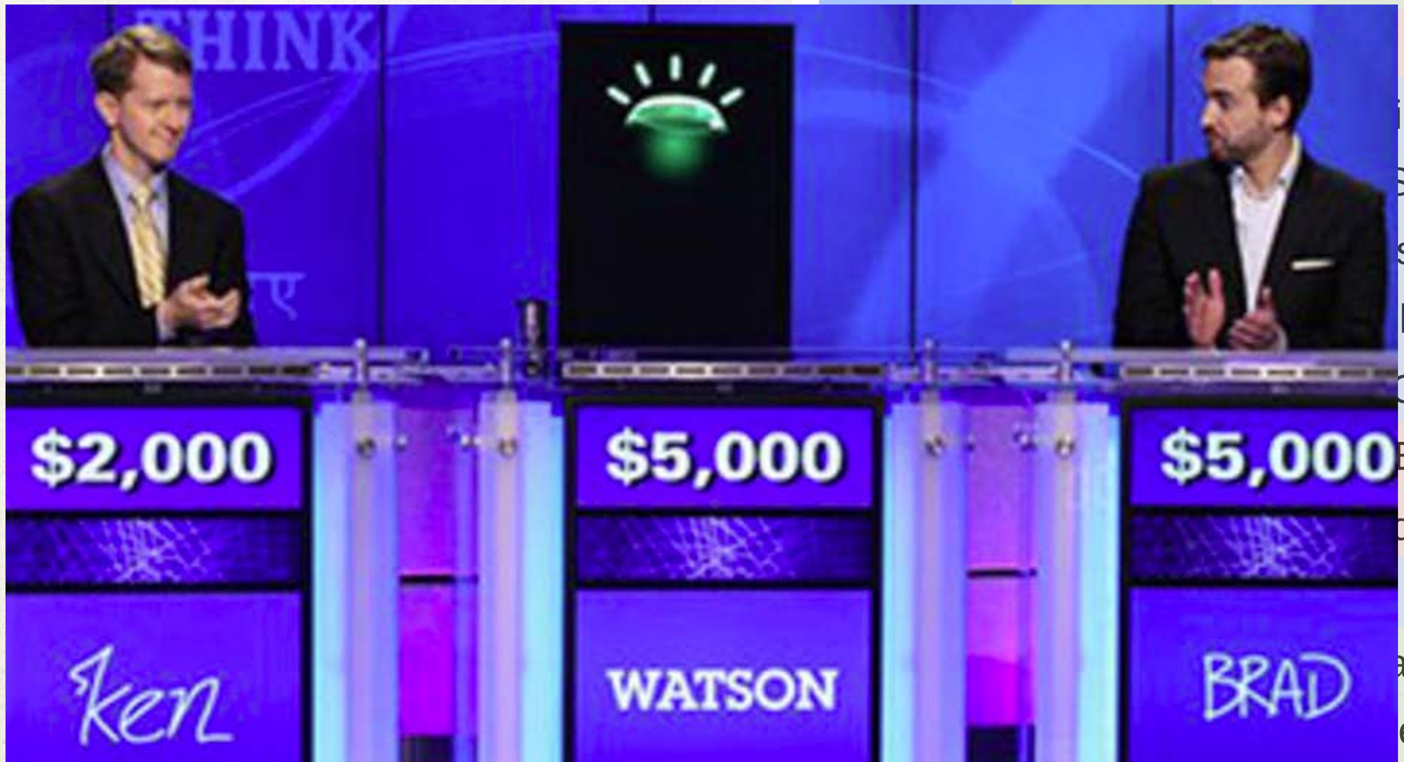


negative	*neutral*	But i wanna wear my Concords tomorrow though but i don't
positive	*neutral*	Gonna watch Grey's Anatomy all day today and tomorrow(:
negative	*neutral*	@CoachVac heey do you know anything about UVA's fallll fe
neutral	*neutral*	@DustyEf when that sun is high in that Texas sky, I'll be buc
neutral	*positive*	Up 20 points in my money league with Vernon Davis and L.
neutral	*positive*	DEEJAYING this FRIDAY in THE FIRST CHOP it's CHRIS actua
negative	*negative*	The Rick Santorum signing that was scheduled for tomorrow
positive	*neutral*	@dreami9 lol yep looks like it! Was after El Clasico on Sunda
neutral	*neutral*	Back in Stoke on Trent for the 2nd time today!
neutral	*neutral*	First Girls Varsity Basketball Game tomorrow at 6:00 pm Th
neutral	*neutral*	#UFC lightweights @Young__Assassin VS @jamievarner set
neutral	*neutral*	@OOOOO_WEEEE slide thru sometime this weekend ill have
negative	*negative*	@DannyB618 Sure absolutely-- I meant out of the Bachman
negative	*negative*	@RichardGordon48 re Levein discussion on Wed. Can't keep
neutral	*neutral*	Today In History November 02, 1958 Elvis gave a party at h
neutral	*positive*	Hustle cause you got to then kick back n party everyday like
positive	*positive*	I can't sleep. Way too exited about Vancouver tomorrow! I'r

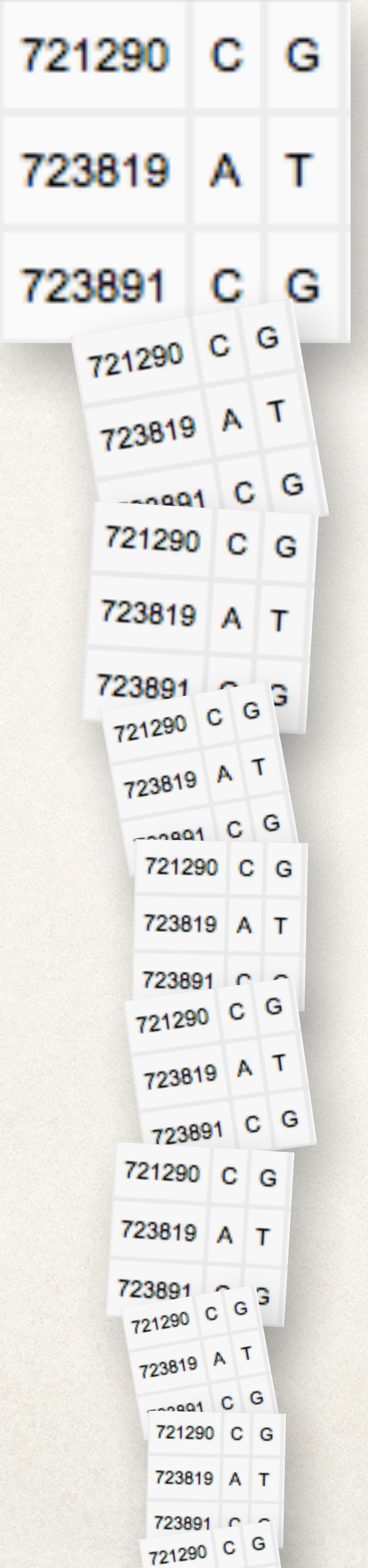
Text-Daten

- Spam
- Internet-Daten
- Medizin: Gendaten

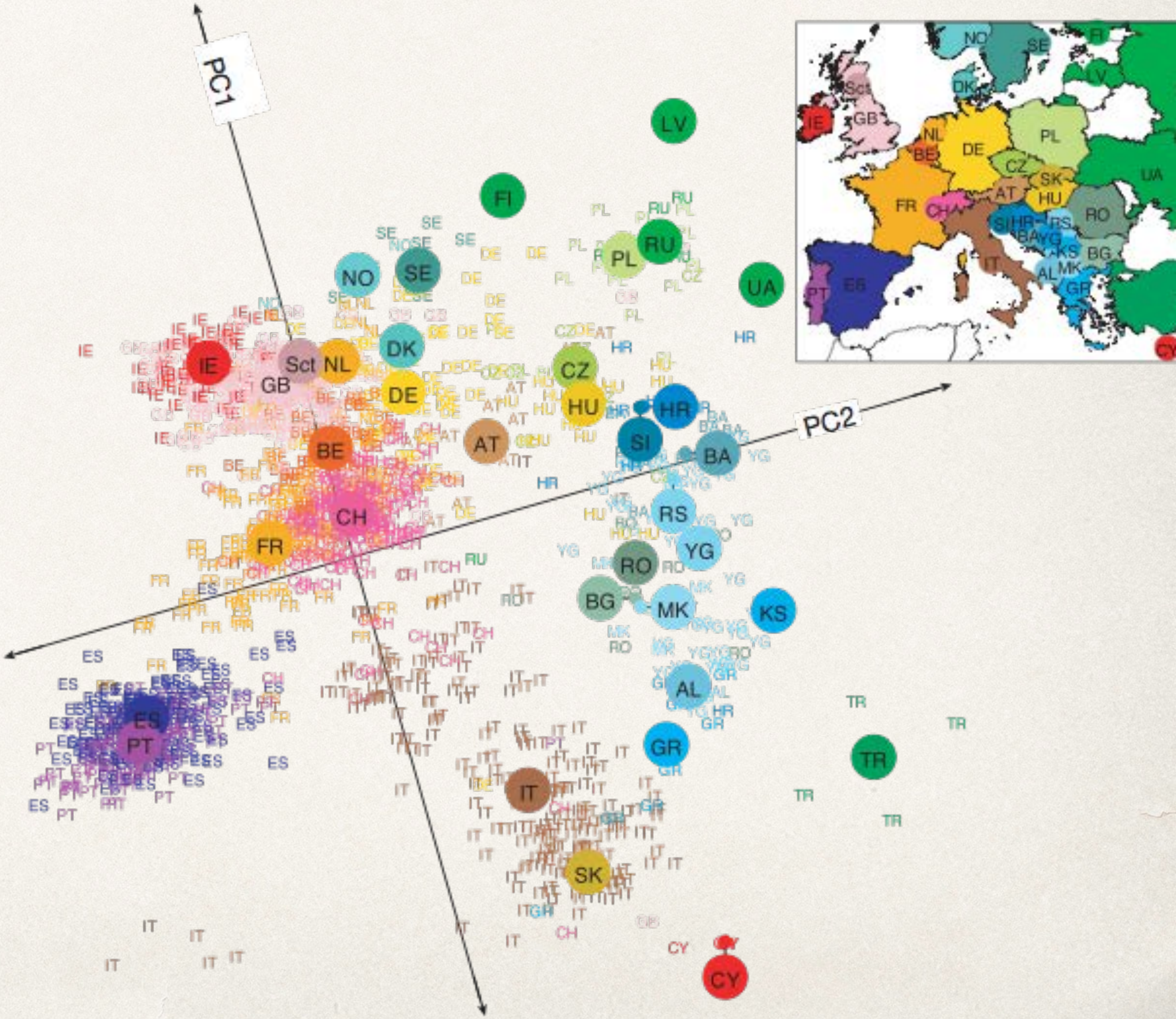
negative	neutral	But i wanna wear my Concords tomorrow though but i don't
positive	neutral	Gonna watch Grey's Anatomy all day today and tomorrow(:
negative	neutral	@CoachVac heey do you know anything about UVA's fallll fe
neutral	neutral	@DustyEf when that sun is high in that Texas sky, I'll be buc
neutral	positive	Up 20 points in my money league with Vernon Davis and L.
neutral	positive	DEEJAYING this FRIDAY in THE FIRST CHOP it's CHRIS actua
		Santorum signing that was scheduled for tomorrow
		i9 lol yep looks like it! Was after El Clasico on Sunda
		Stoke on Trent for the 2nd time today!
		s Varsity Basketball Game tomorrow at 6:00 pm Th
		htweights @Young__Assassin VS @jamievarner set
		O_WEEEE slide thru sometime this weekend ill have
		3618 Sure absolutely-- I meant out of the Bachman
		dGordon48 re Levein discussion on Wed. Can't keep
		History November 02, 1958 Elvis gave a party at h
		ause you got to then kick back n party everyday like
		eep. Way too exited about Vancouver tomorrow! I'r



Medizin: Analyse von Gen-Daten



721290	C	G
723819	A	T
723891	C	G

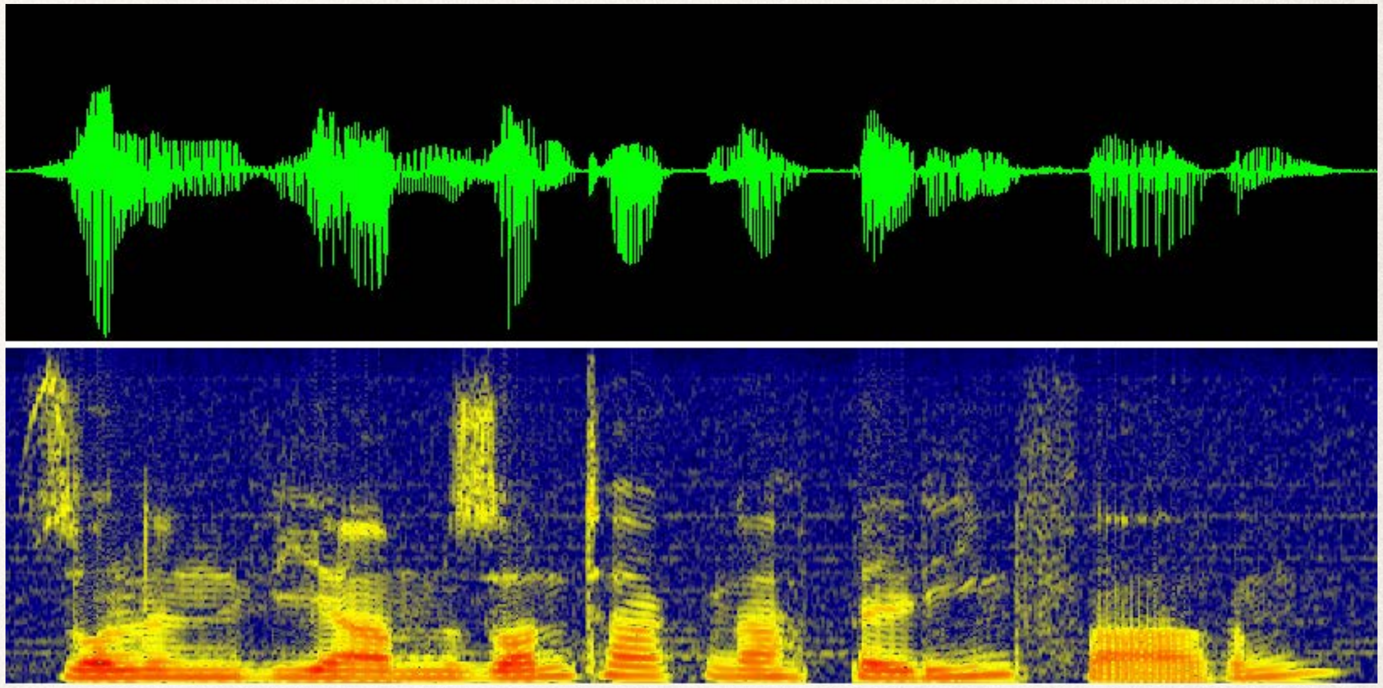


Audio-Daten

- Hörgeräte
- Spracherkennung
- Automatische Übersetzung

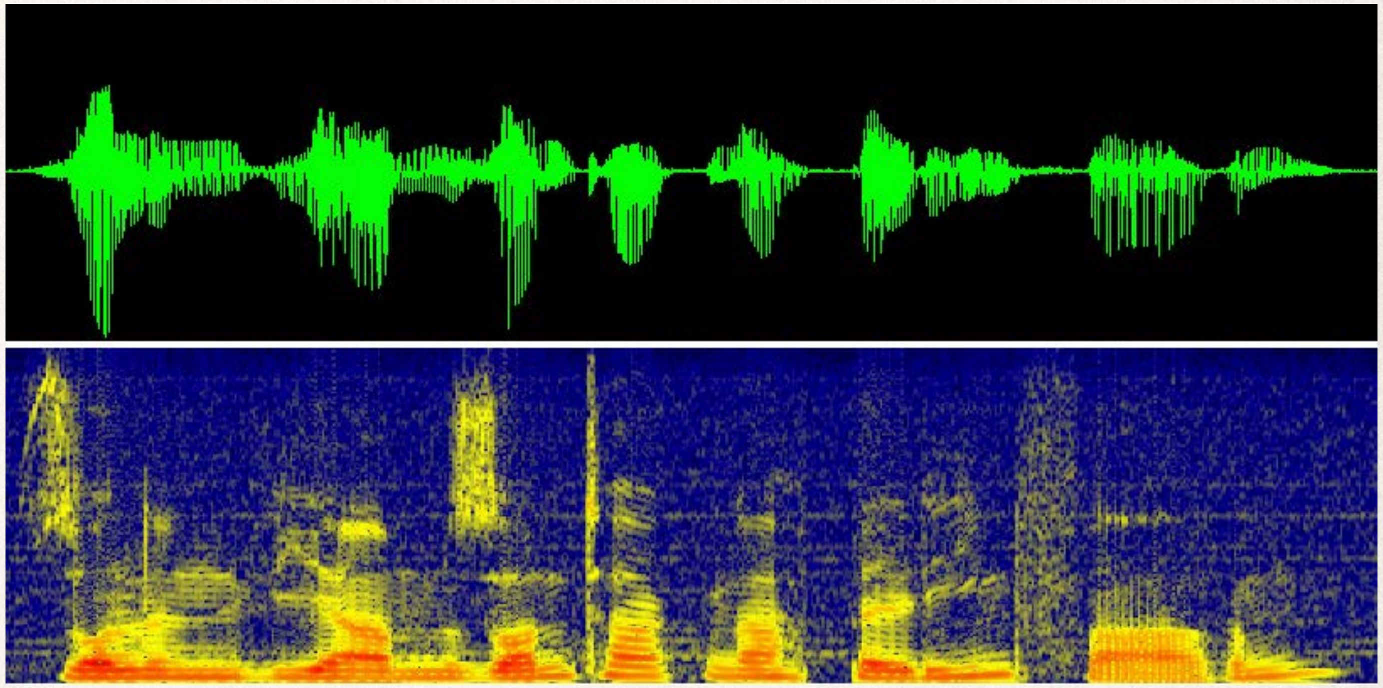
Audio-Daten

- Hörgeräte
- Spracherkennung
- Automatische Übersetzung



Audio-Daten

- Hörgeräte
- Spracherkennung
- Automatische Übersetzung



Audio-Daten

- Hörgeräte
- Spracherkennung
- Automatische Übersetzung

MARINEXPLORE
Cornell University

Numerische / Sensor-Daten

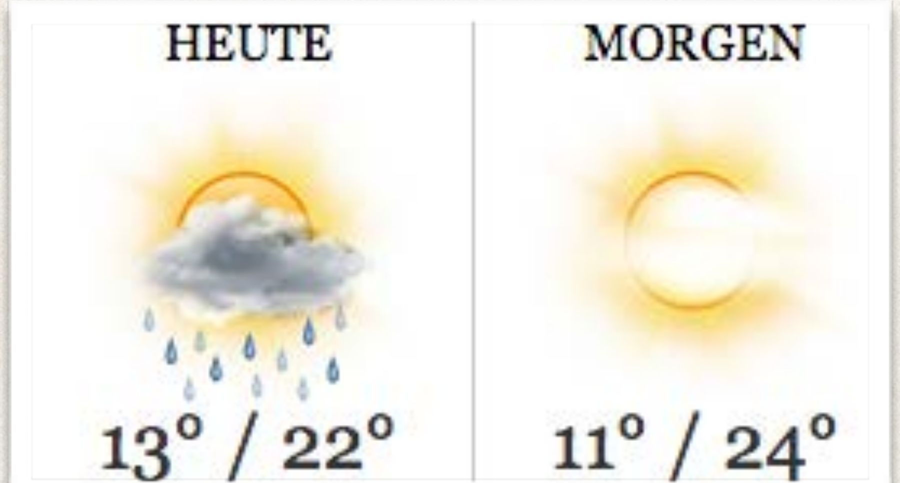
- Cern (Higgs Teilchen)
- Fitness-Armband
- Wetter-Vorhersage
- Segeln
- Robotik

Numerische / Sensor-Daten

- Cern (Higgs Teilchen)
- Fitness-Armband
- Wetter-Vorhersage
- Segeln
- Robotik

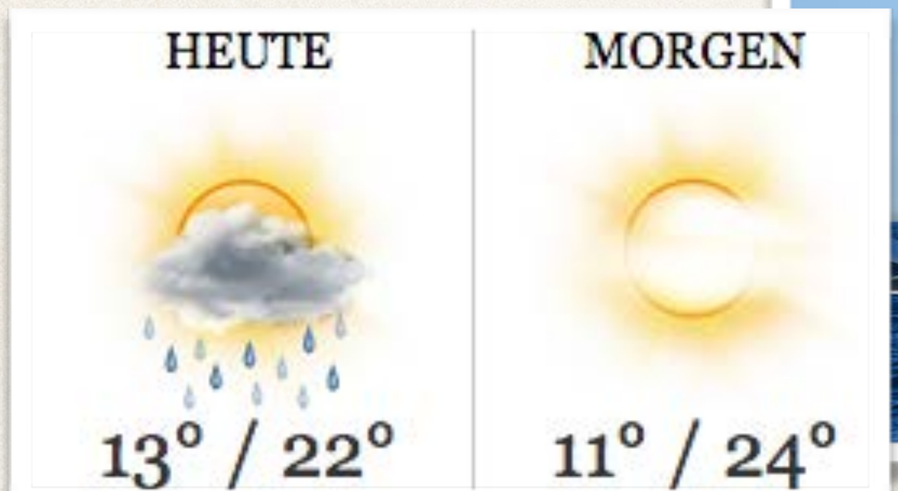
Numerische / Sensor-Daten

- Cern (Higgs Teilchen)
- Fitness-Armband
- Wetter-Vorhersage
- Segeln
- Robotik



Numerische / Sensor-Daten

- Cern (Higgs Teilchen)
- Fitness-Armband
- Wetter-Vorhersage
- Segeln
- Robotik



HEUTE	MORGEN
13° / 22°	11° / 24°

Numerische / Sensor-Daten

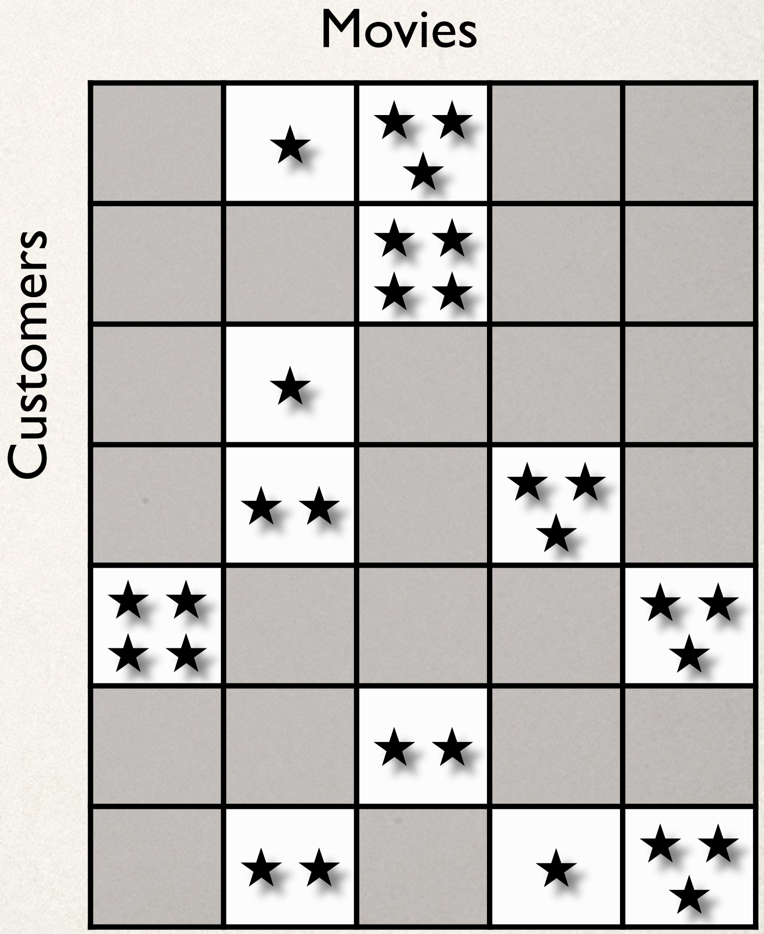
- Cern (Higgs Teilchen)
- Fitness-Armband
- Wetter-Vorhersage
- Segeln
- Robotik

Internet-Daten

- Werbung
- Empfehlungssysteme

Internet-Daten

- Werbung
- Empfehlungssysteme



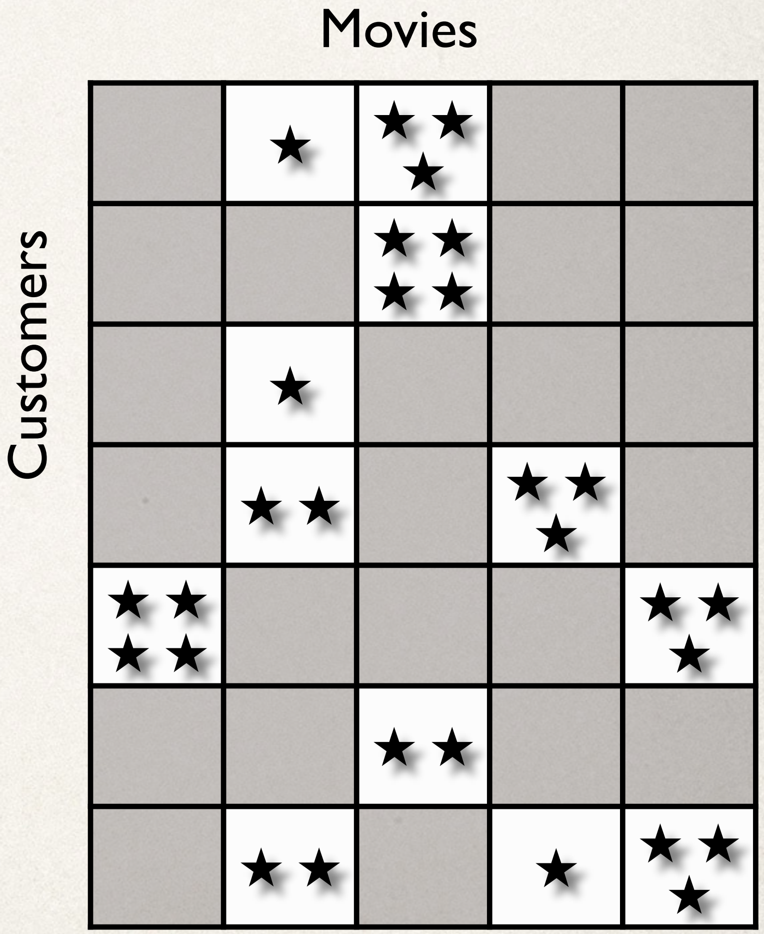
	★	★★★		
		★★★★		
	★			
	★★		★★★	
★★★★				★★★
		★★		
	★★		★	★★★

Internet-Daten

- Werbung
- Empfehlungssysteme

NETFLIX

amazon.com

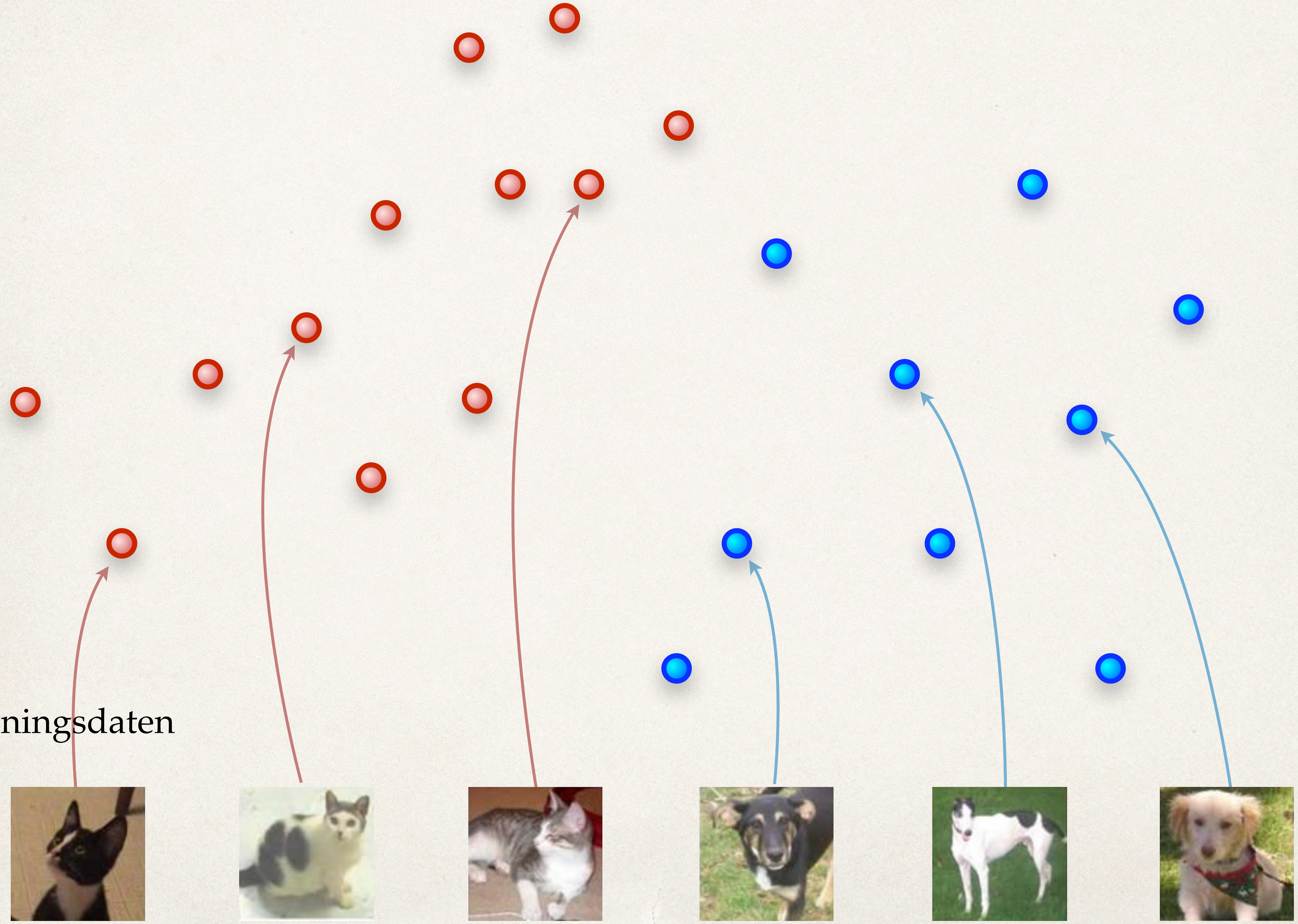


	Movies				
Customers		★	★★★		
			★★★★		
		★			
		★★		★★★	
	★★★★				★★★
			★★		
		★★		★	★★★

Versicherungen & Finanzwelt

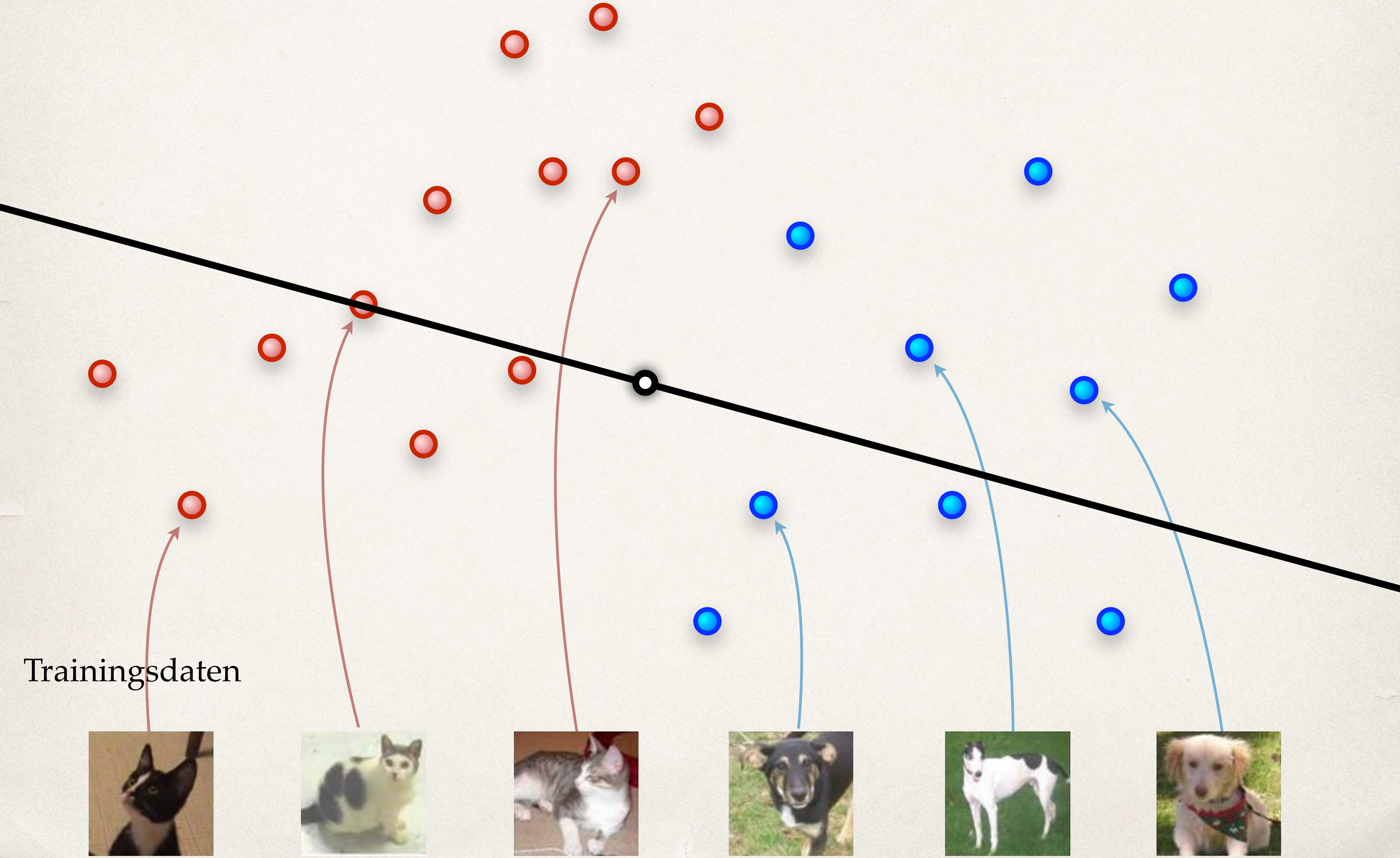
- Business-Analytics
- Werbung
- Kreditkarten-Betrug
- Versicherungs-Risiko
- Kundenbindung

Klassifikation



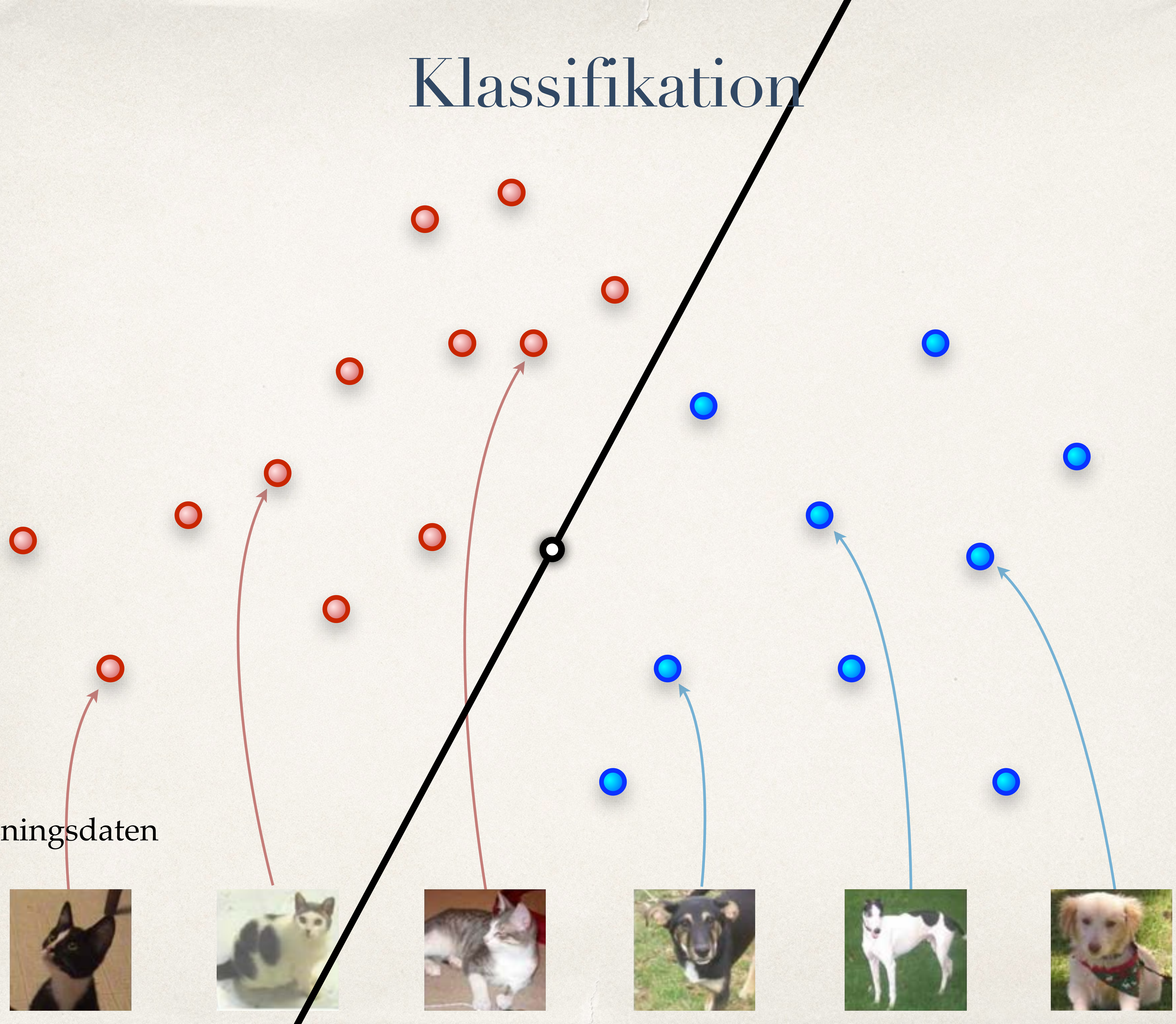
Trainingsdaten

Klassifikation



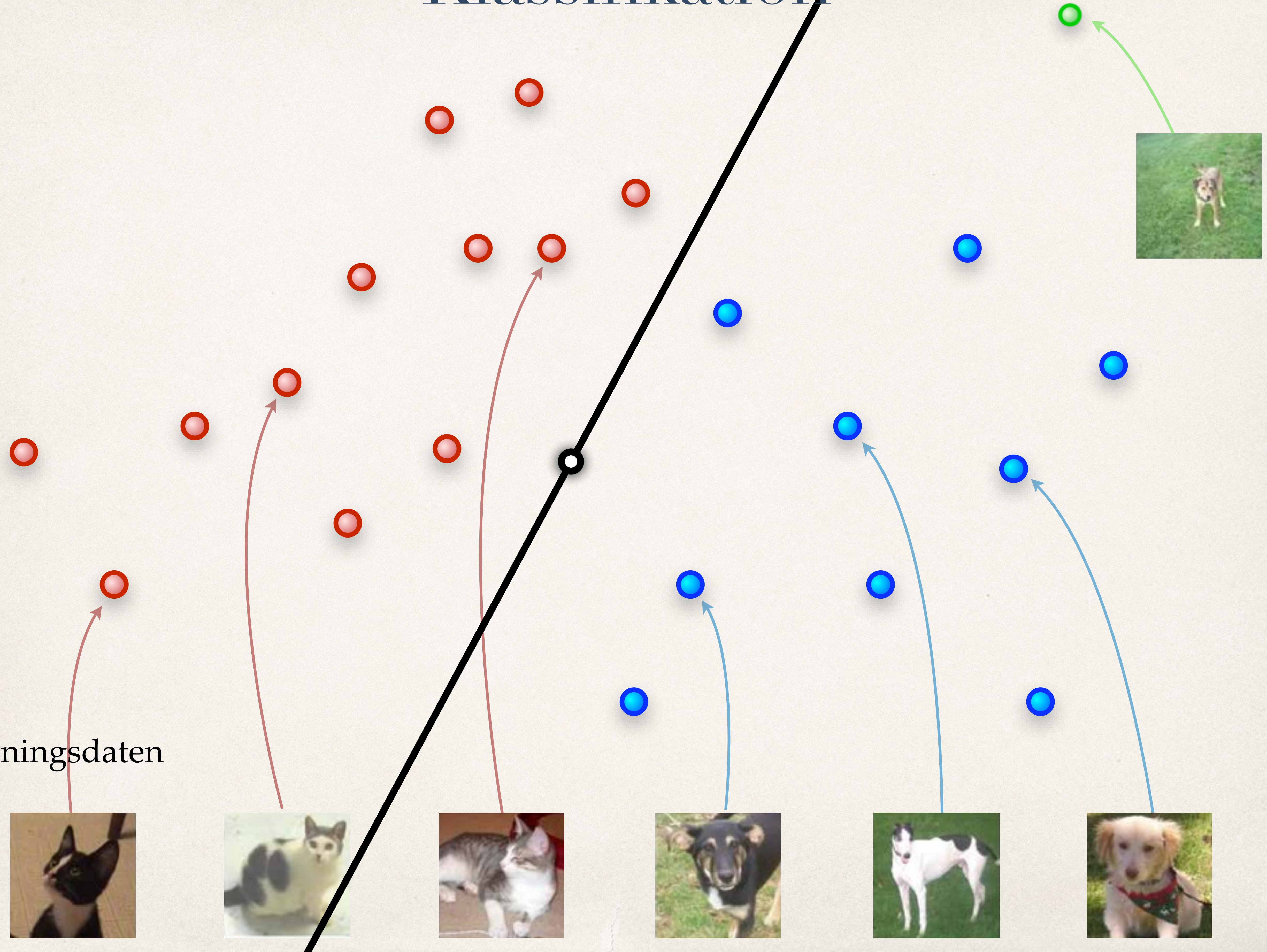
Trainingsdaten

Klassifikation



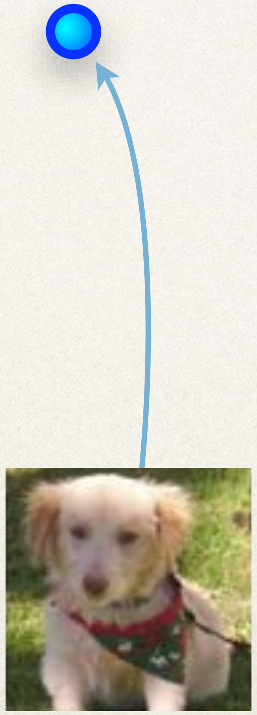
Trainingsdaten

Klassifikation

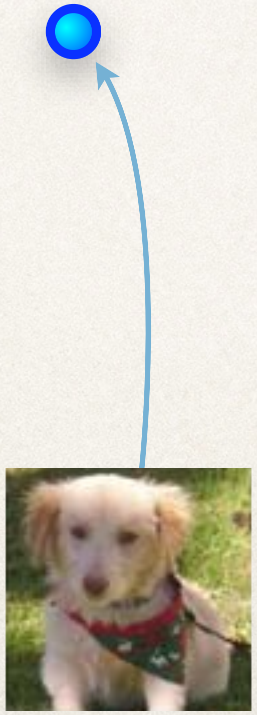


Trainingsdaten

Von Daten zu geometrischen Punkten



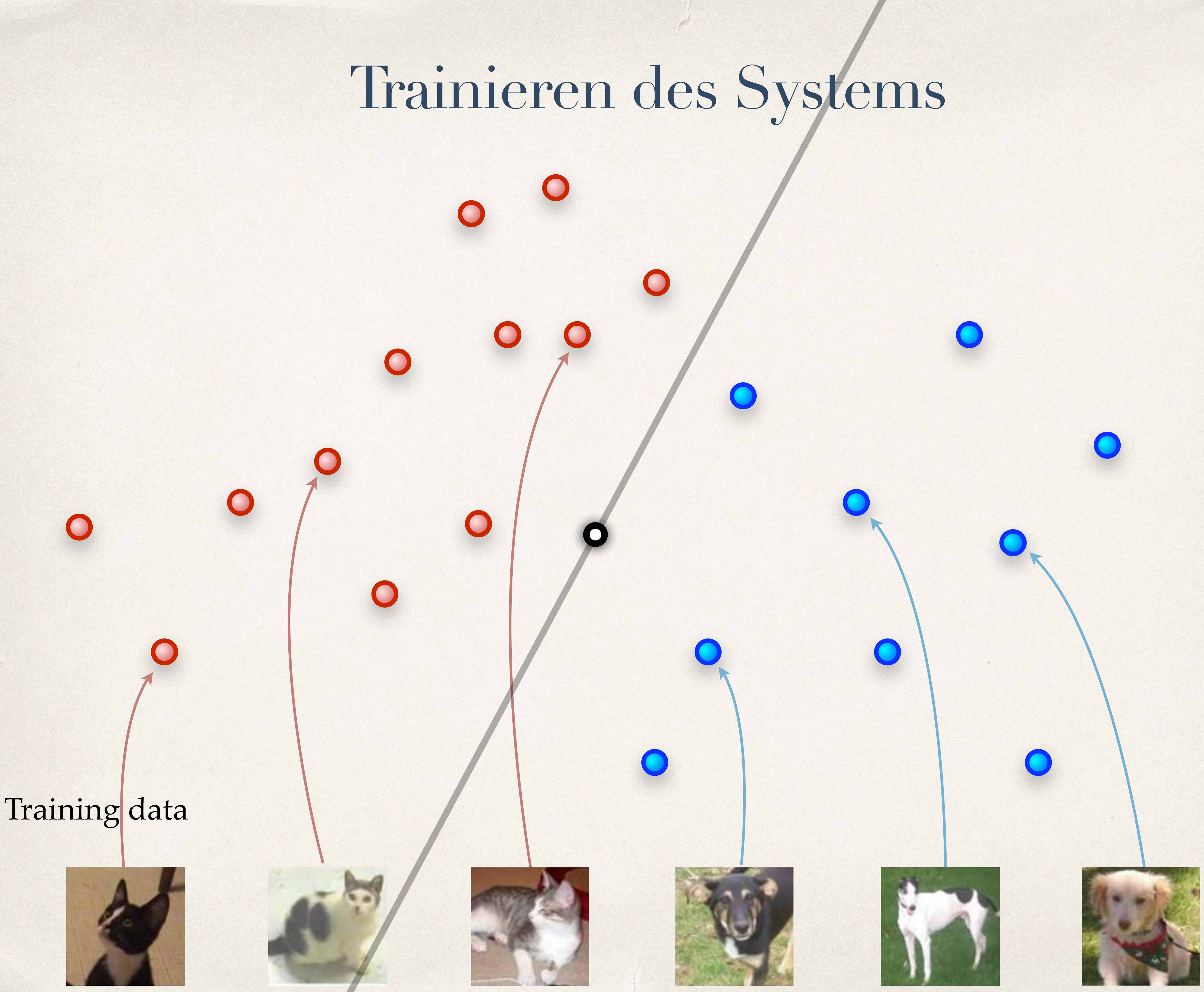
Von Daten zu geometrischen Punkten



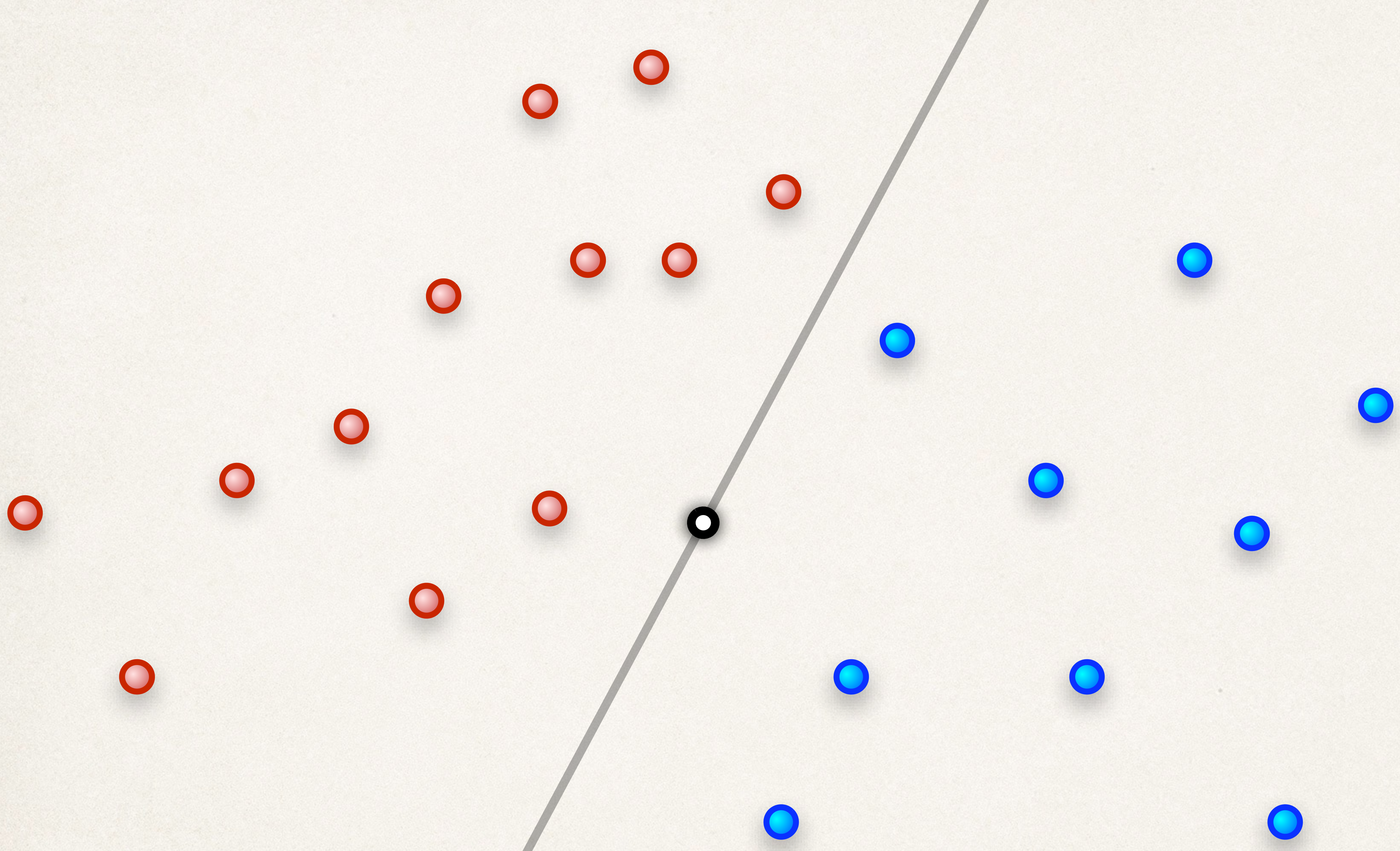
$$\begin{array}{|c|}\hline 1 \\ \hline 0.5 \\ \hline 0.8 \\ \hline 0.7 \\ \hline 0.8 \\ \hline 0 \\ \hline 0.2 \\ \hline \end{array} \; = \boldsymbol{x}$$

$$\vdots$$

Trainieren des Systems



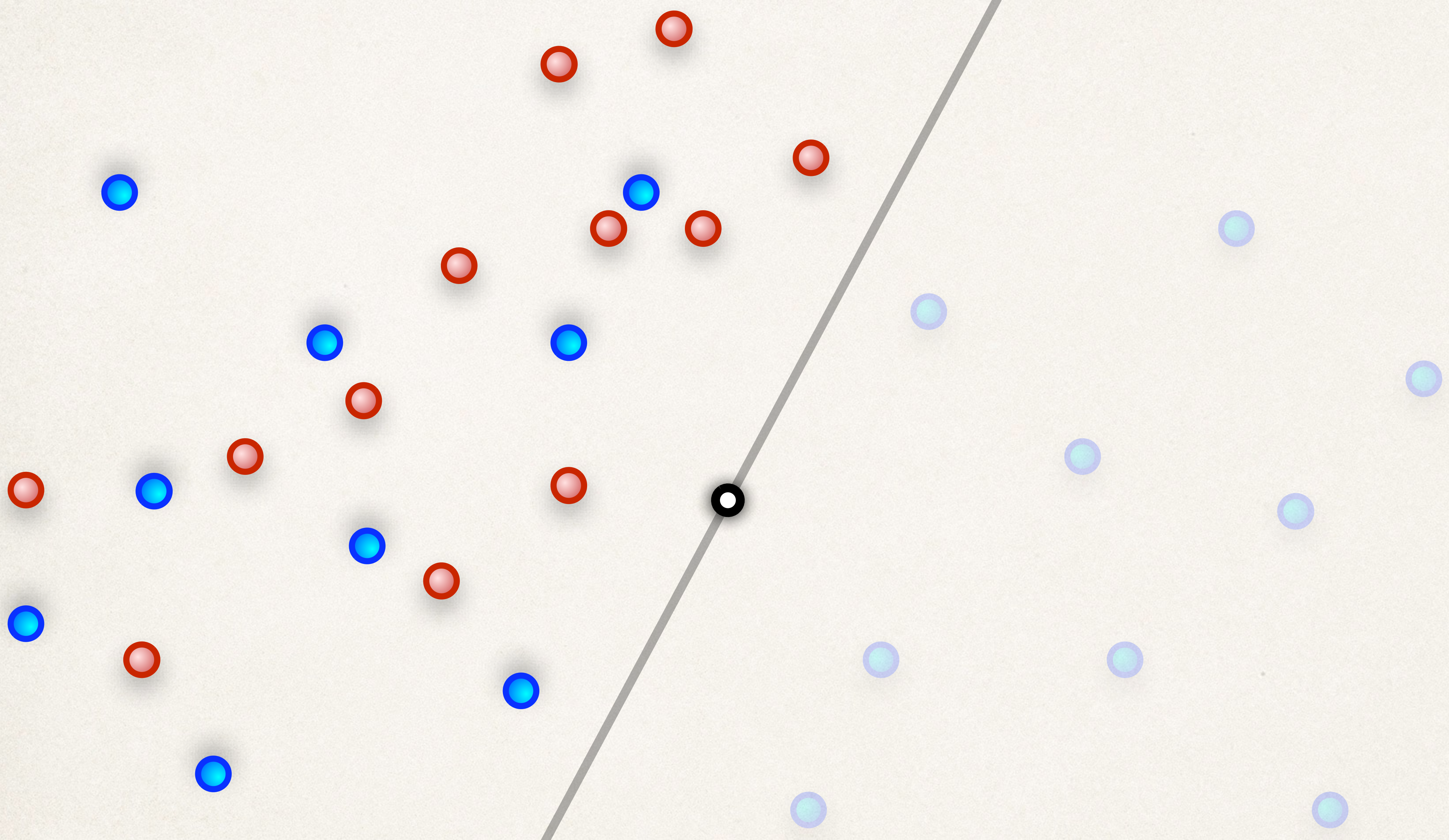
Trainieren des Systems



Perzeptron

(Rosenblatt 1957)

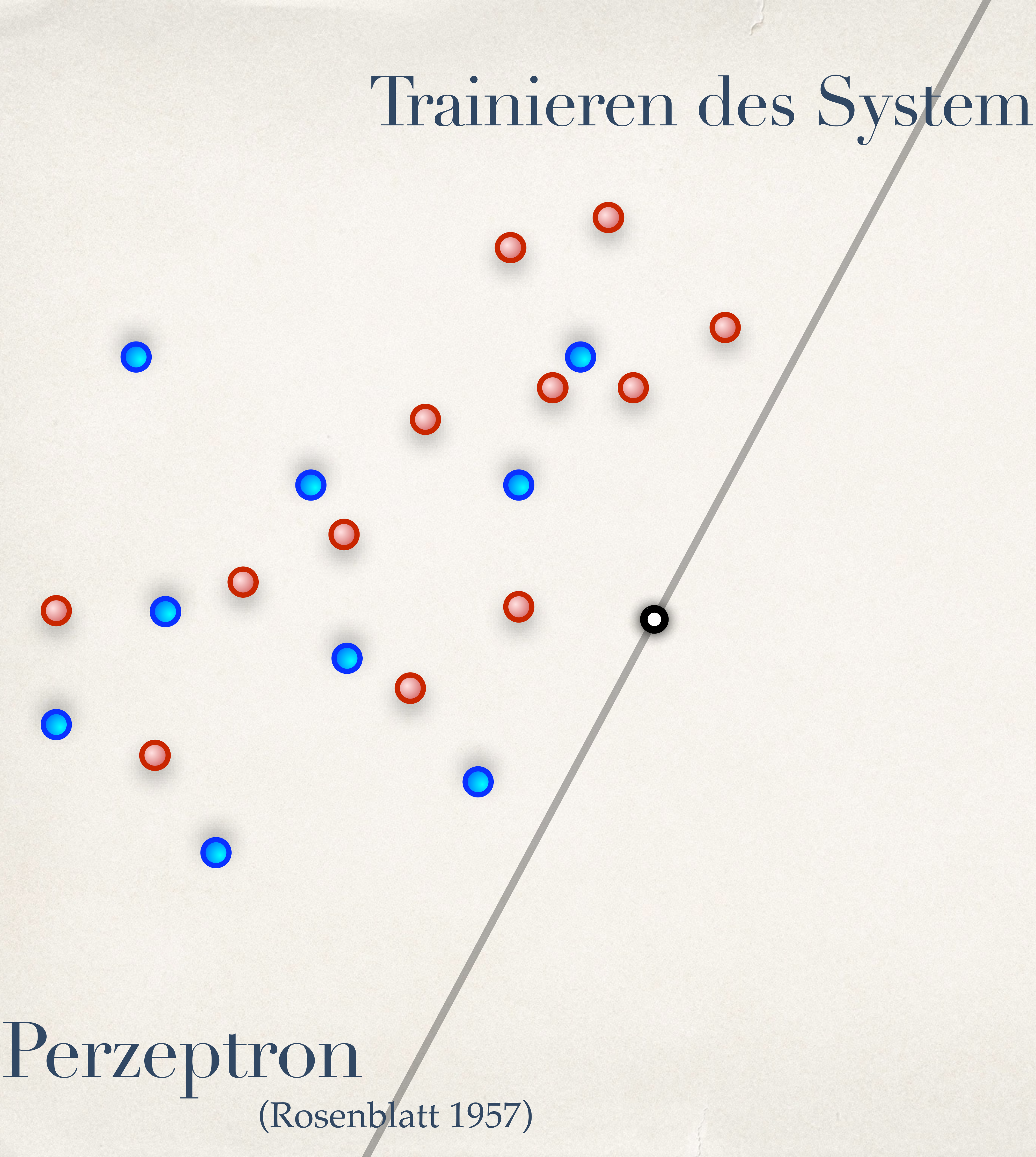
Trainieren des Systems



Perzeptron

(Rosenblatt 1957)

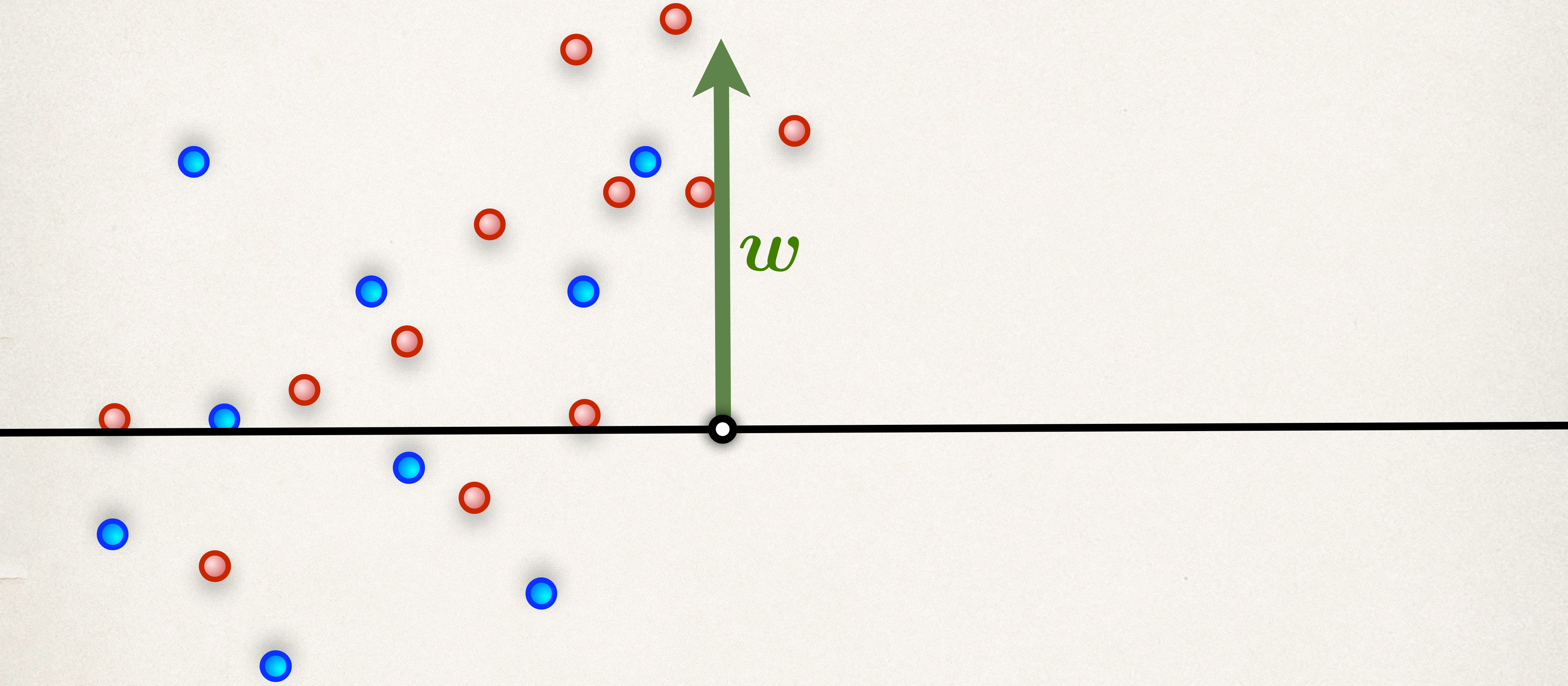
Trainieren des Systems



Perzeptron

(Rosenblatt 1957)

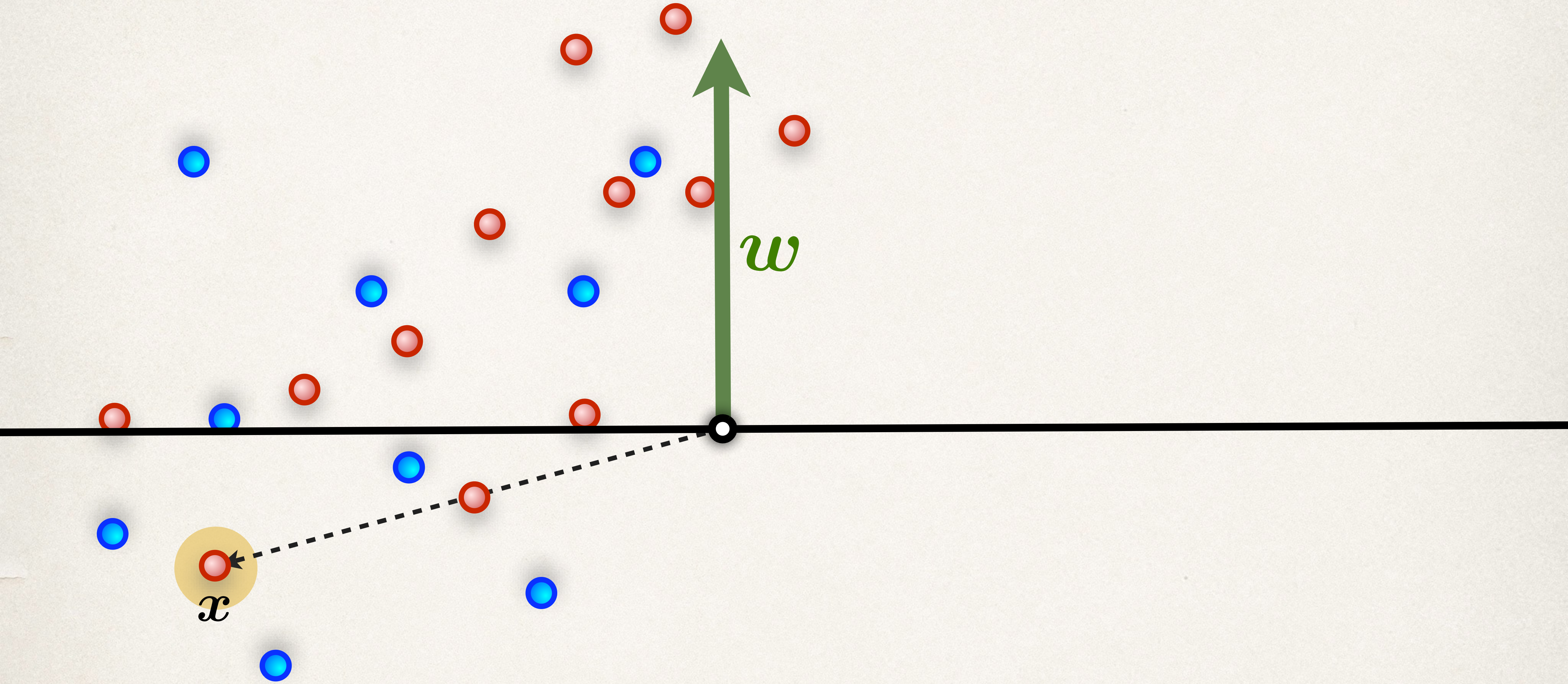
Trainieren des Systems



Perzeptron

(Rosenblatt 1957)

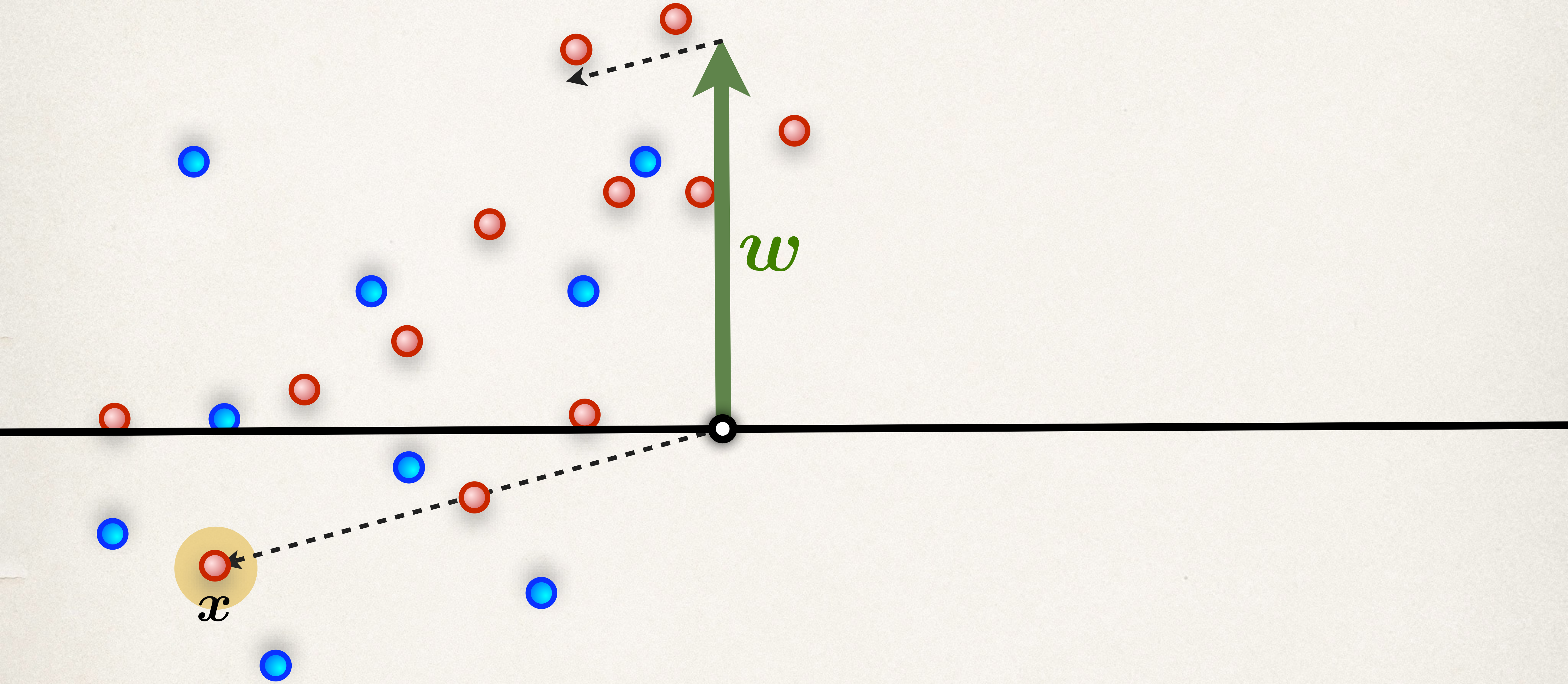
Trainieren des Systems



Perzeptron

(Rosenblatt 1957)

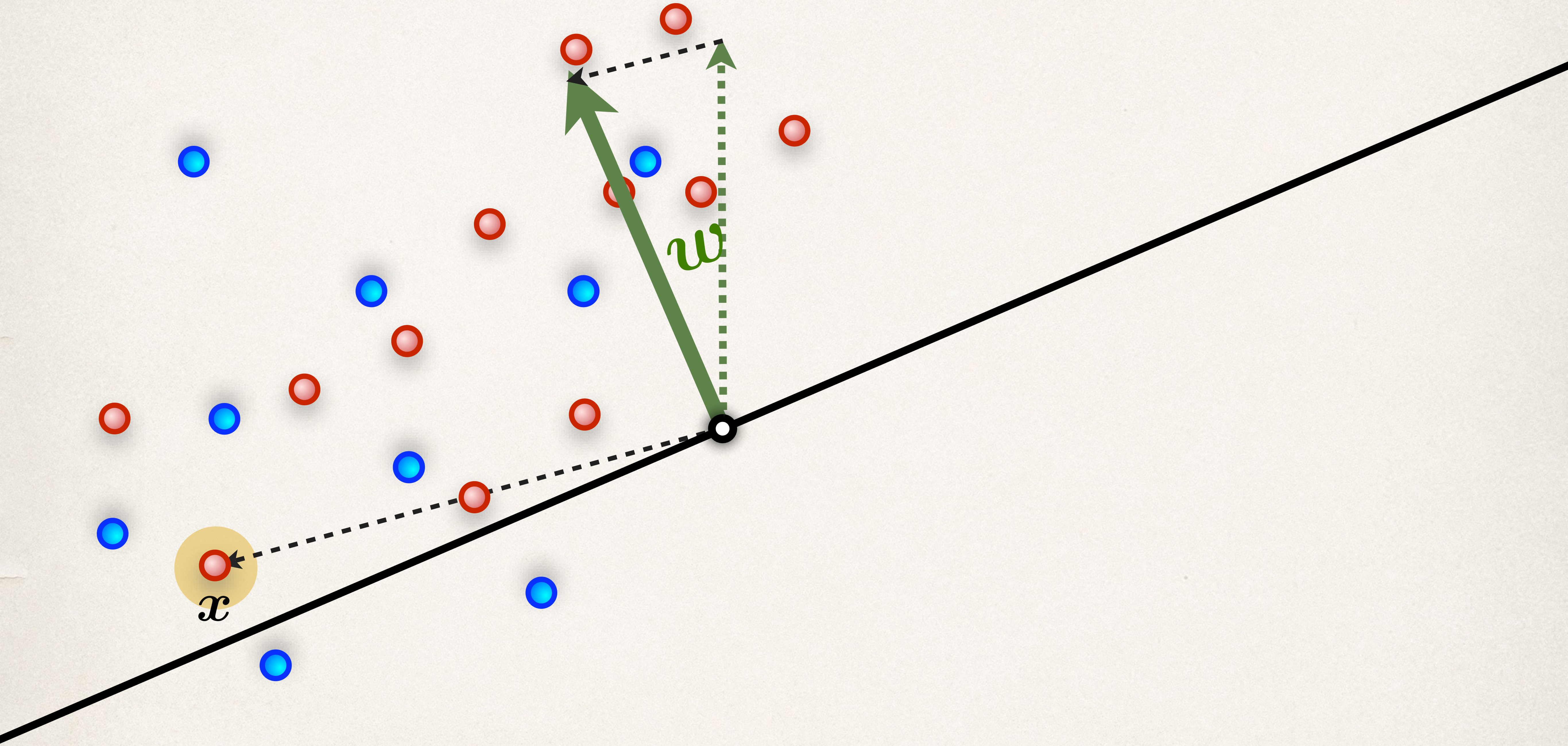
Trainieren des Systems



Perzeptron

(Rosenblatt 1957)

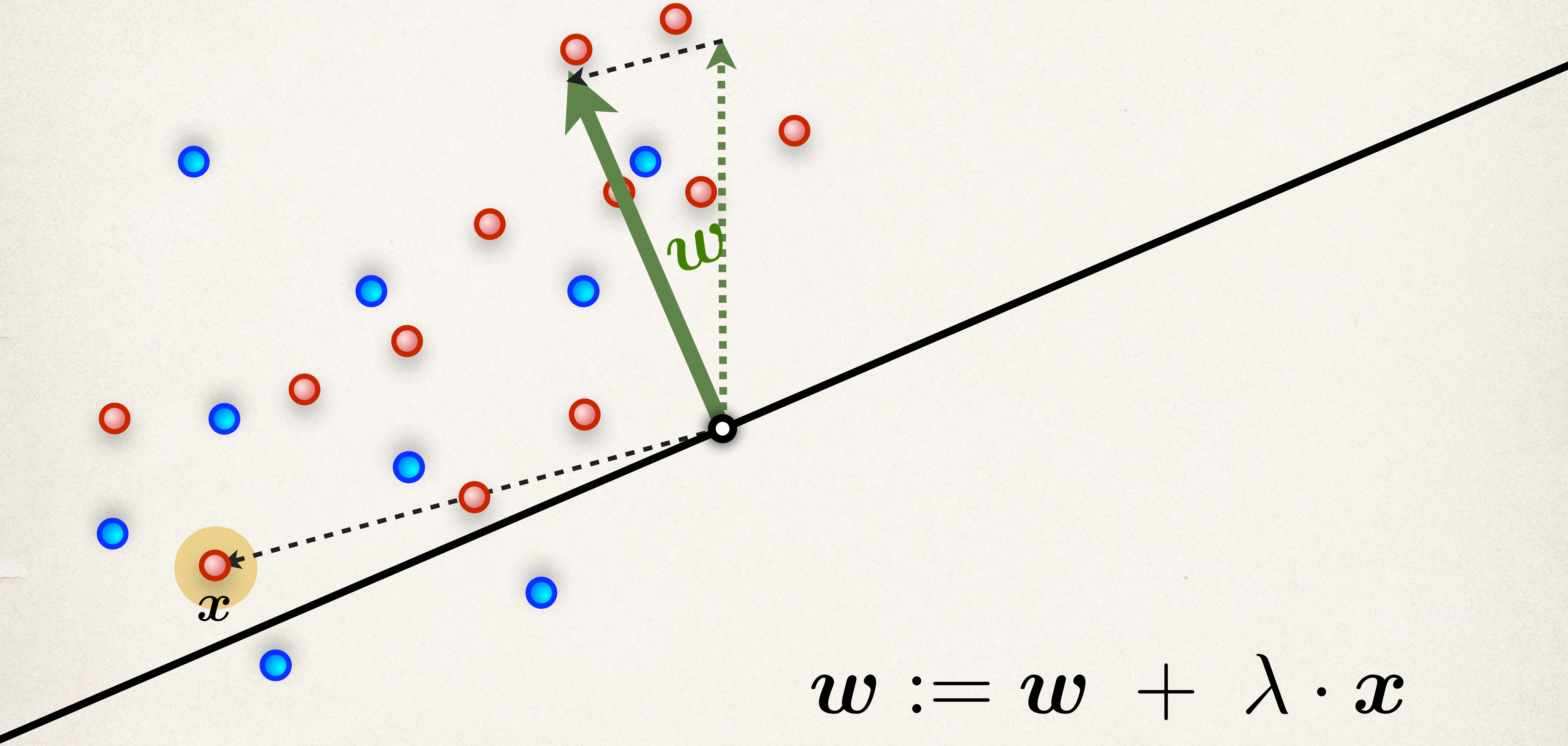
Trainieren des Systems



Perzeptron

(Rosenblatt 1957)

Trainieren des Systems

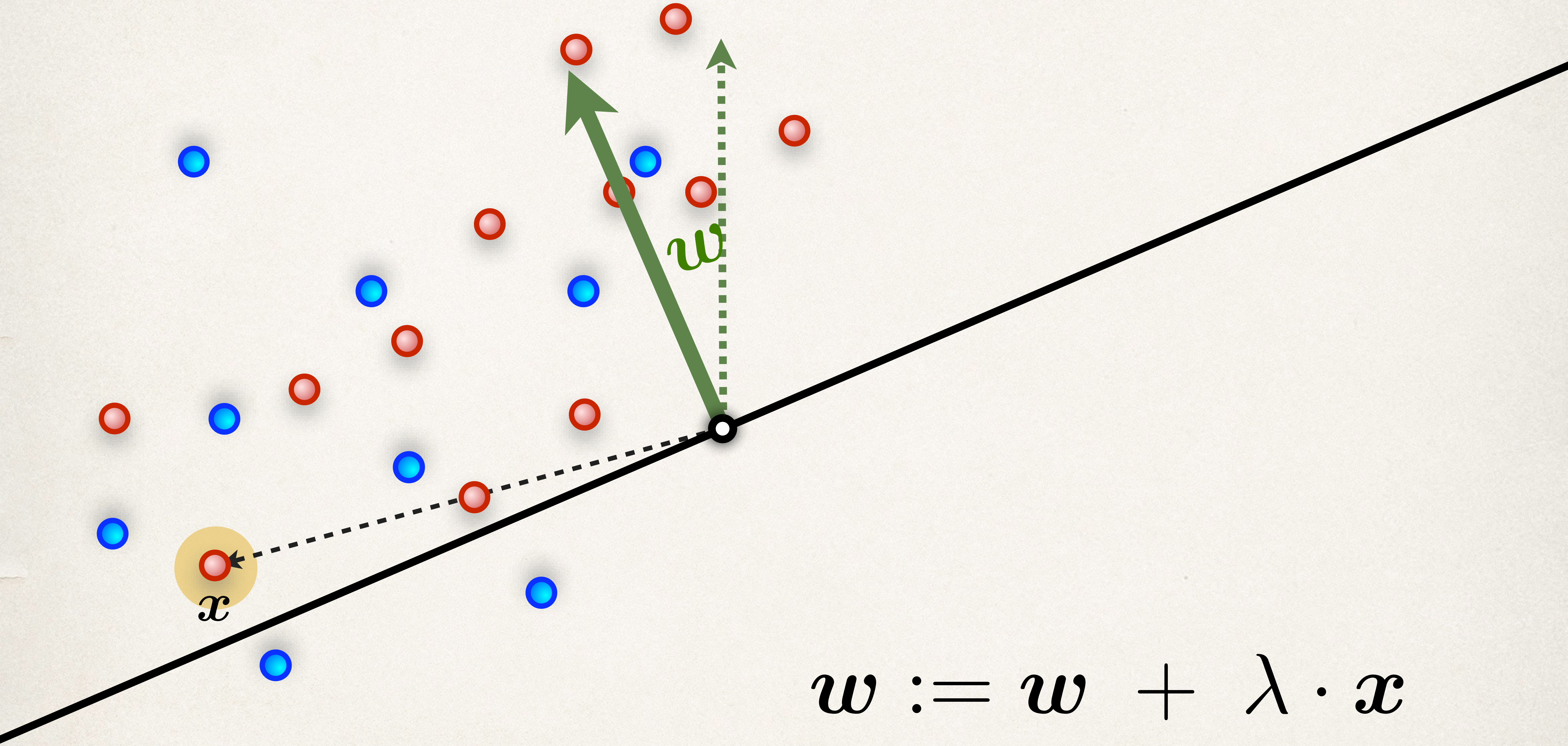


$$w := w + \lambda \cdot x$$

Perzeptron

(Rosenblatt 1957)

Trainieren des Systems

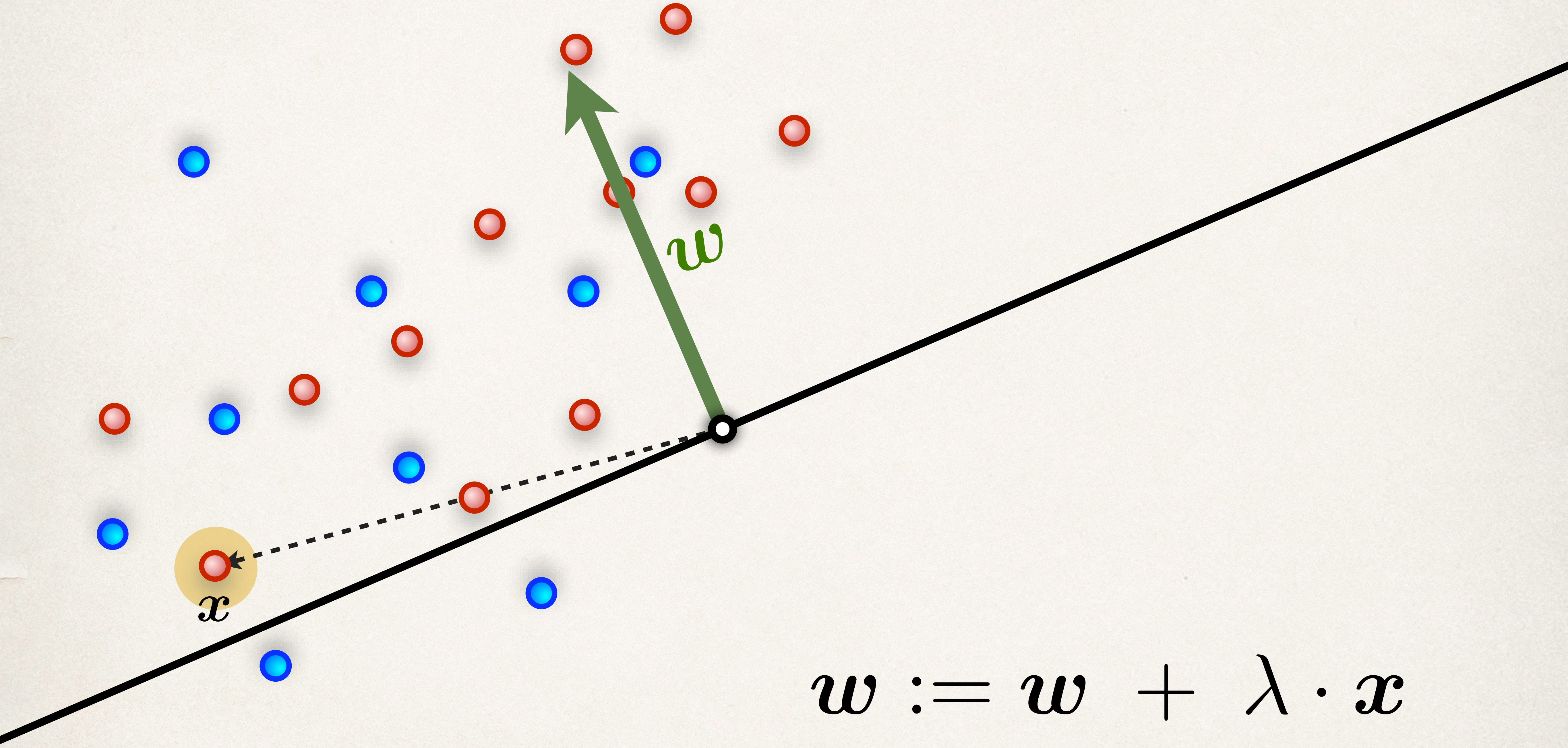


$$w := w \; + \; \lambda \cdot x$$

Perzeptron

(Rosenblatt 1957)

Trainieren des Systems

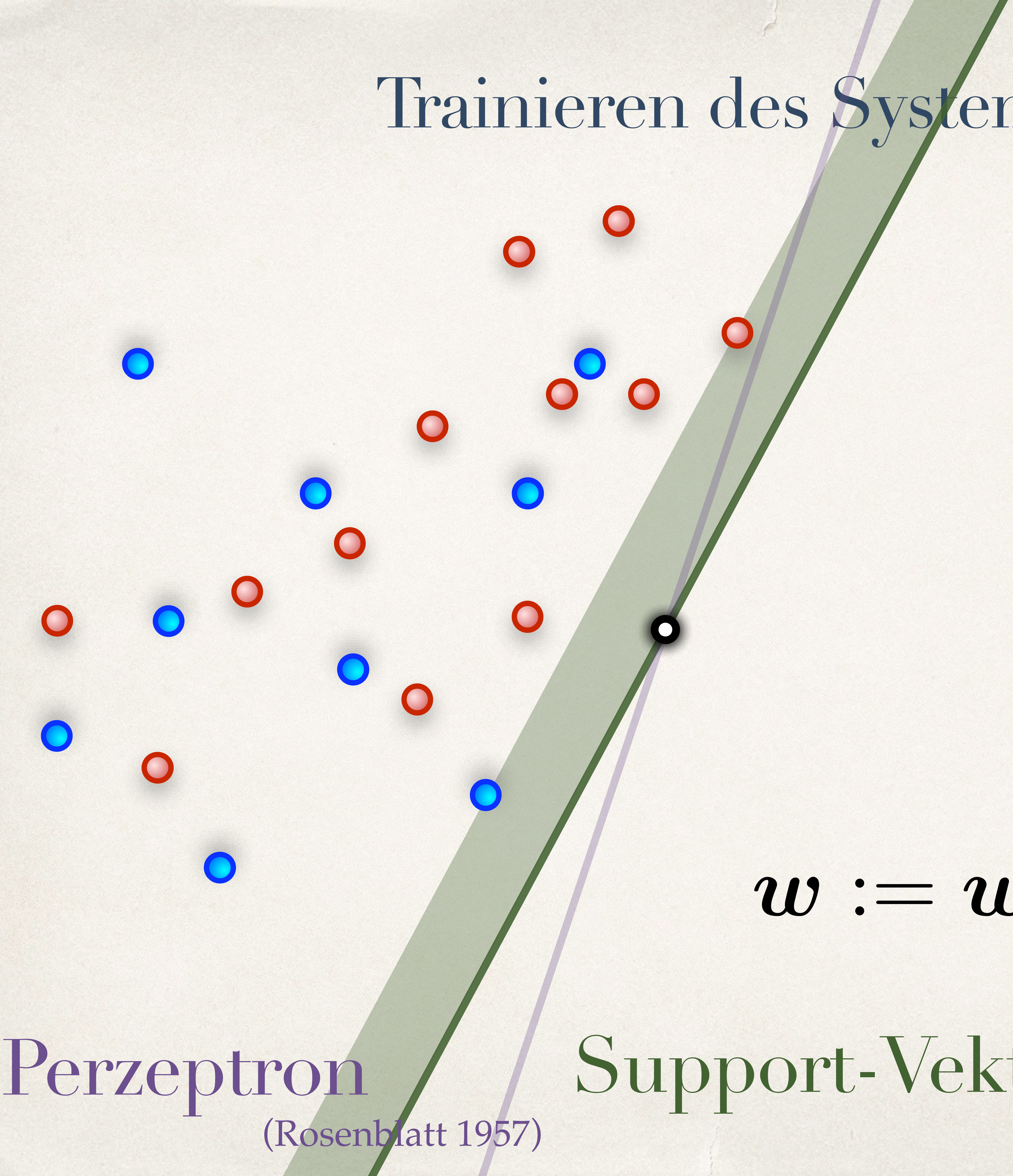


$$\boldsymbol{w} := \boldsymbol{w} + \lambda \cdot \boldsymbol{x}$$

Perzeptron

(Rosenblatt 1957)

Trainieren des Systems



$$\boldsymbol{w} := \boldsymbol{w} + \lambda \cdot \boldsymbol{x}$$

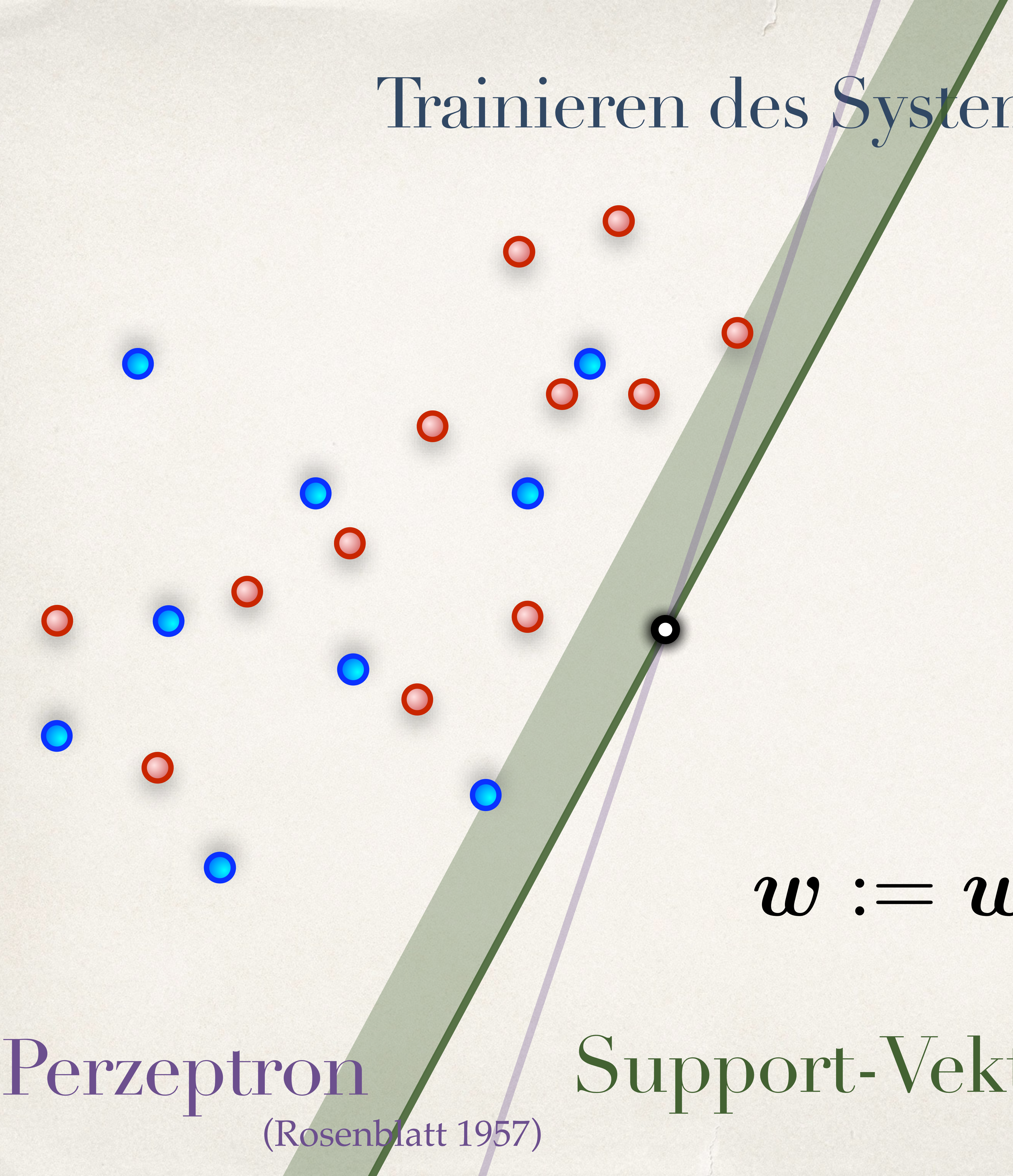
Perzeptron

(Rosenblatt 1957)

Support-Vektor-Maschine

(Cortes & Vapnik 1995)

Trainieren des Systems



$$\boldsymbol{w} := \boldsymbol{w} + \lambda \cdot \boldsymbol{x}$$

Perzeptron

(Rosenblatt 1957)

Support-Vektor-Maschine

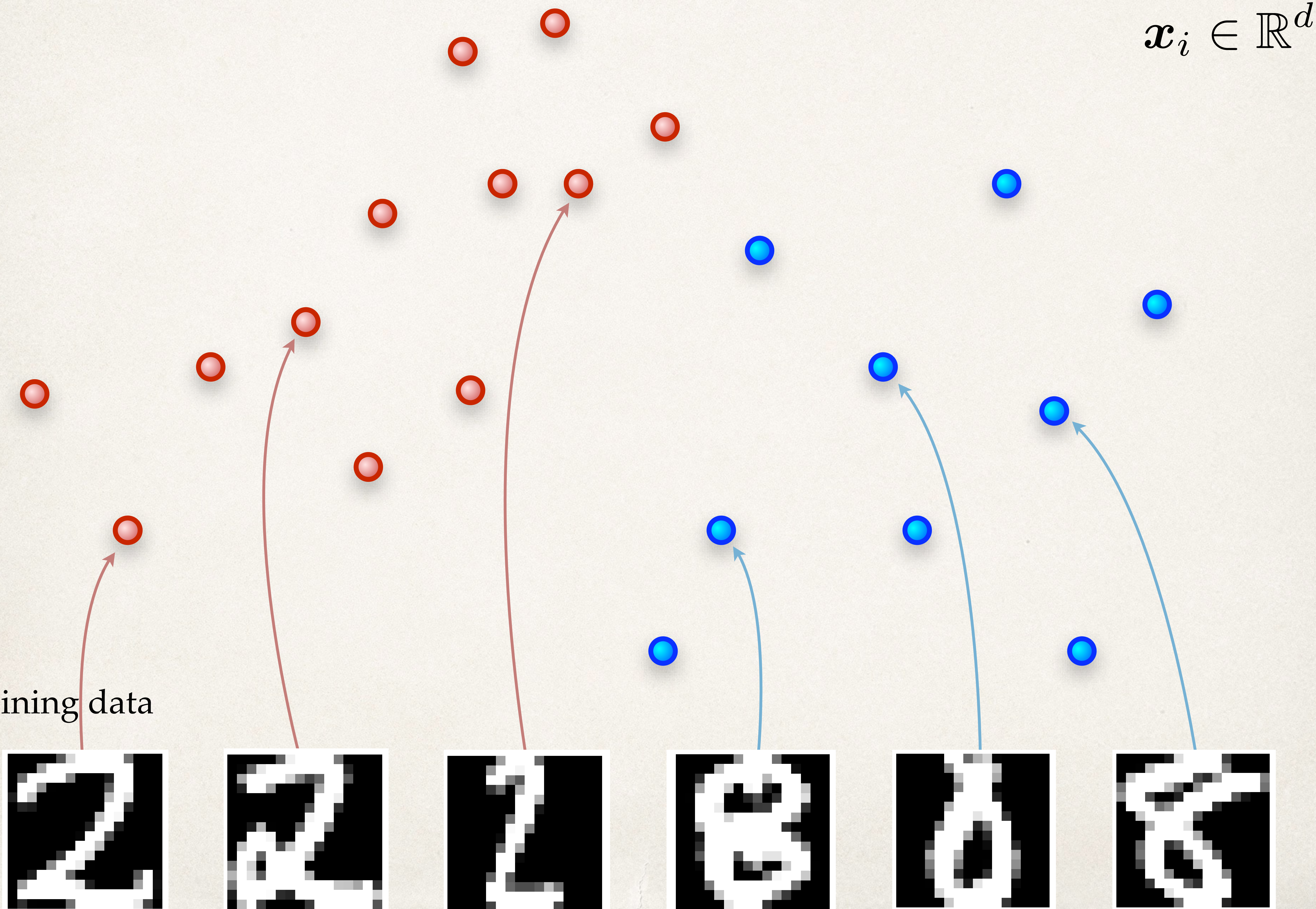
(Cortes & Vapnik 1995)

Training Linear Classifiers

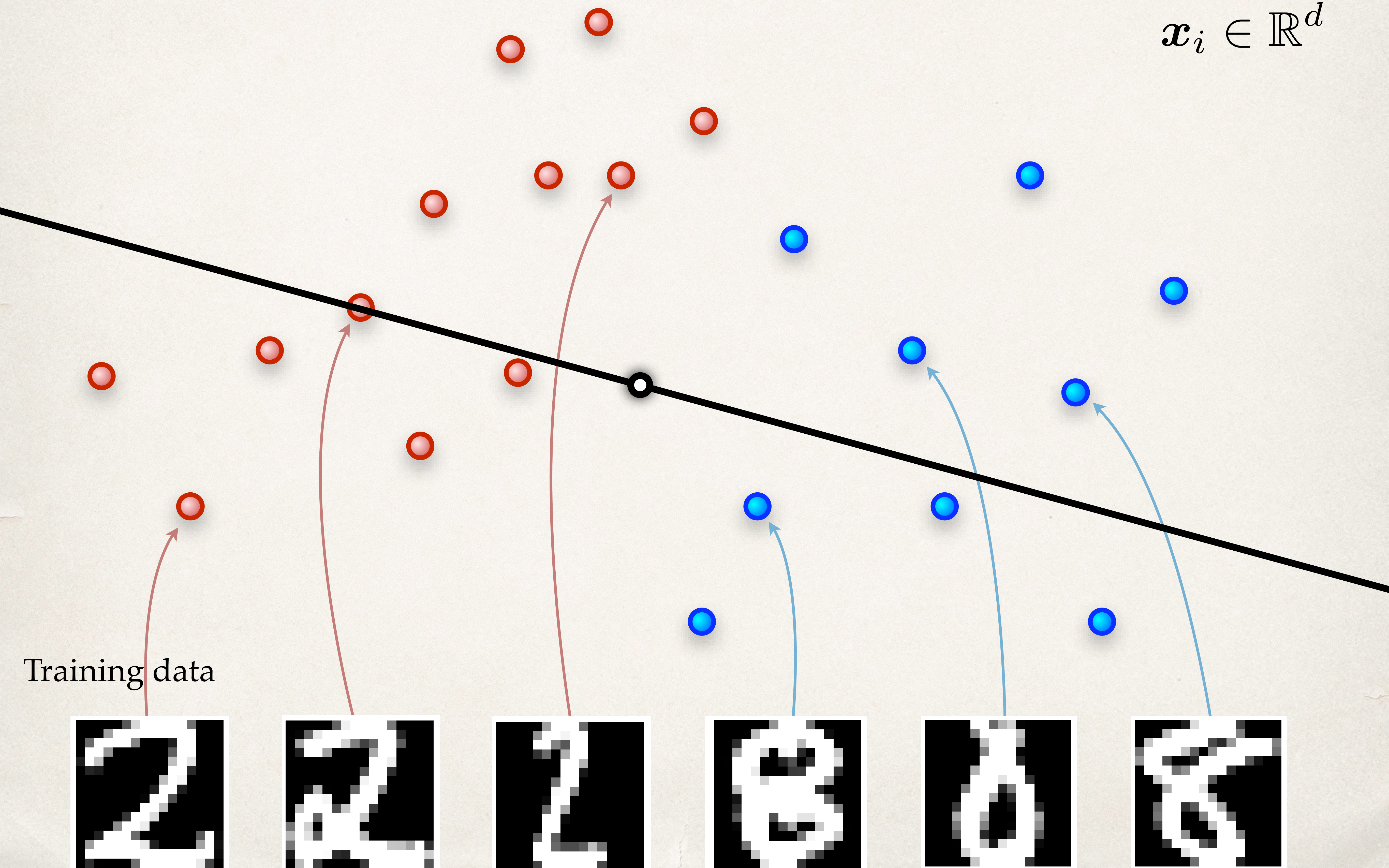
$\boldsymbol{x}_i \in \mathbb{R}^d$

Training data

Training Linear Classifiers



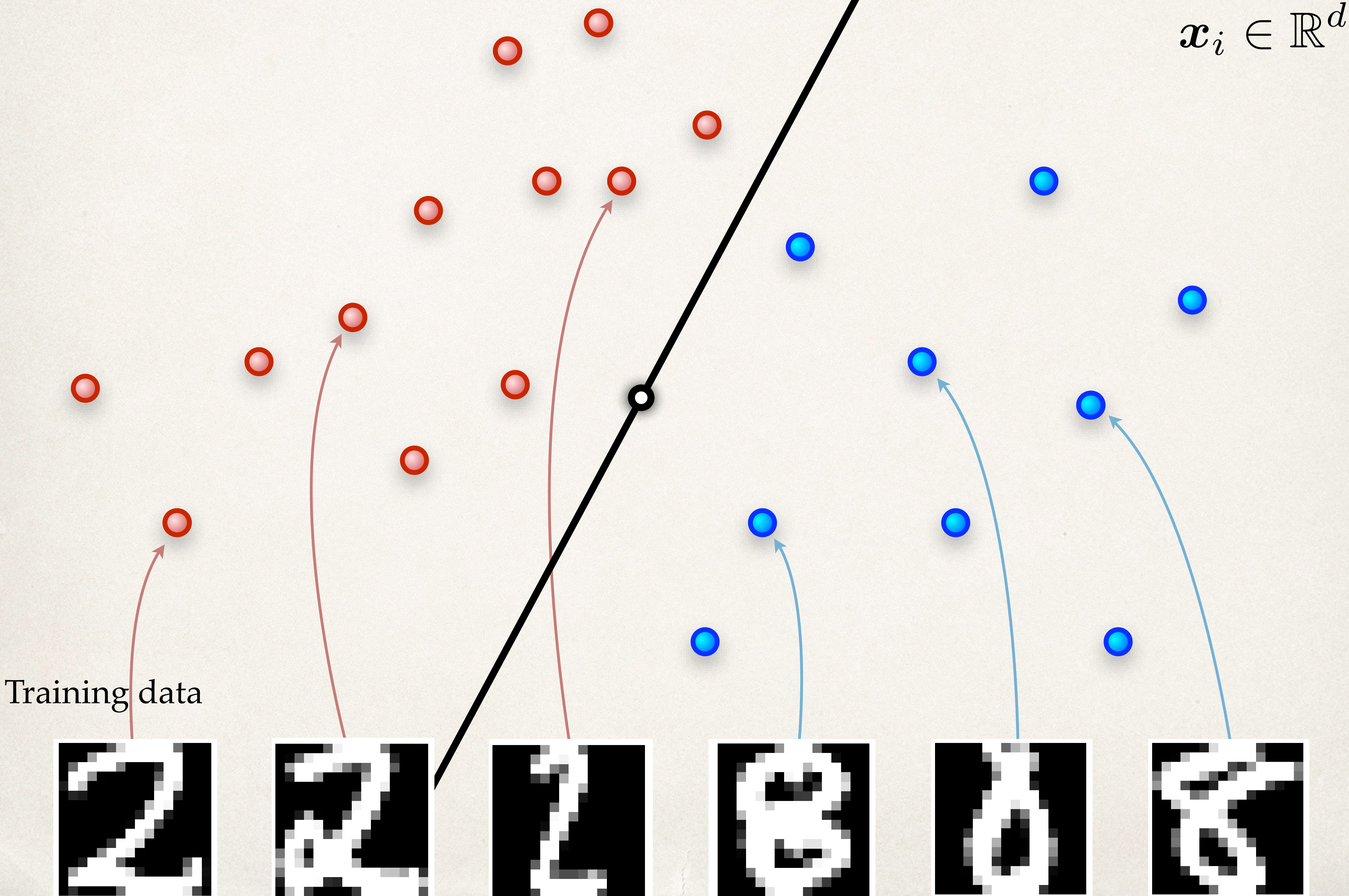
Training Linear Classifiers



$\boldsymbol{x}_i \in \mathbb{R}^d$

Training data

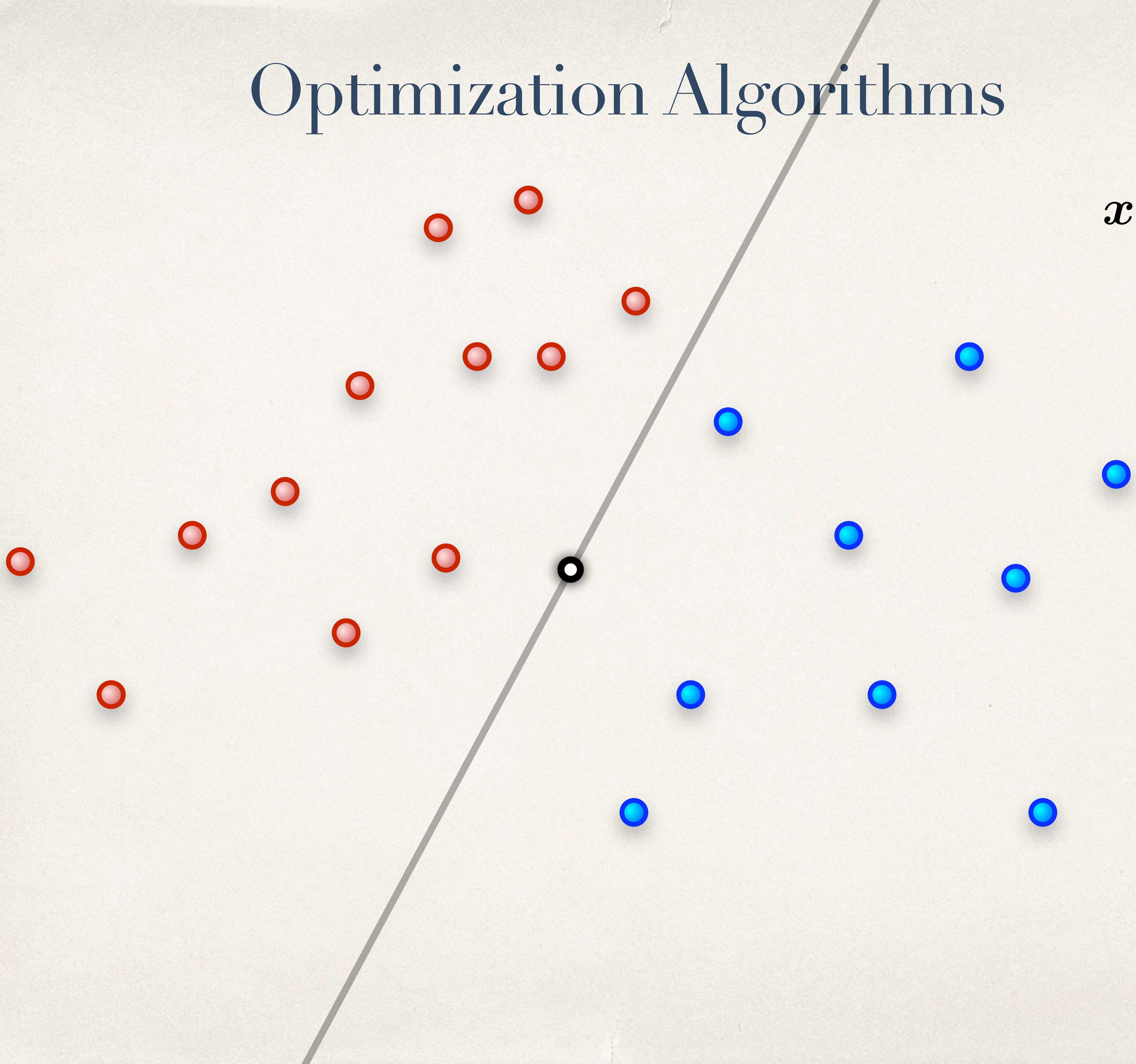
Training Linear Classifiers



$\boldsymbol{x}_i \in \mathbb{R}^d$

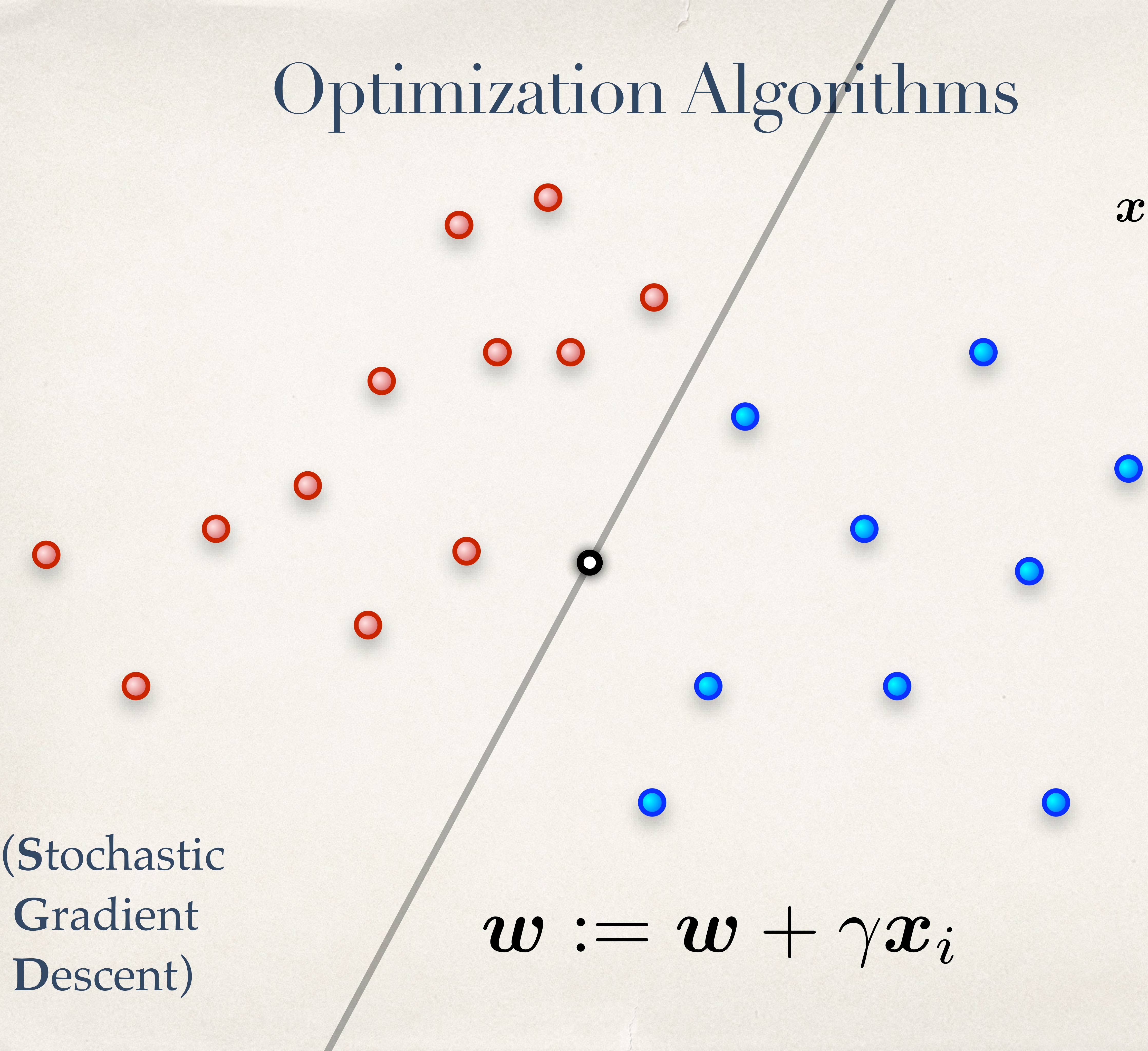
Training data

Optimization Algorithms



$\boldsymbol{x}_i \in \mathbb{R}^d$

Optimization Algorithms

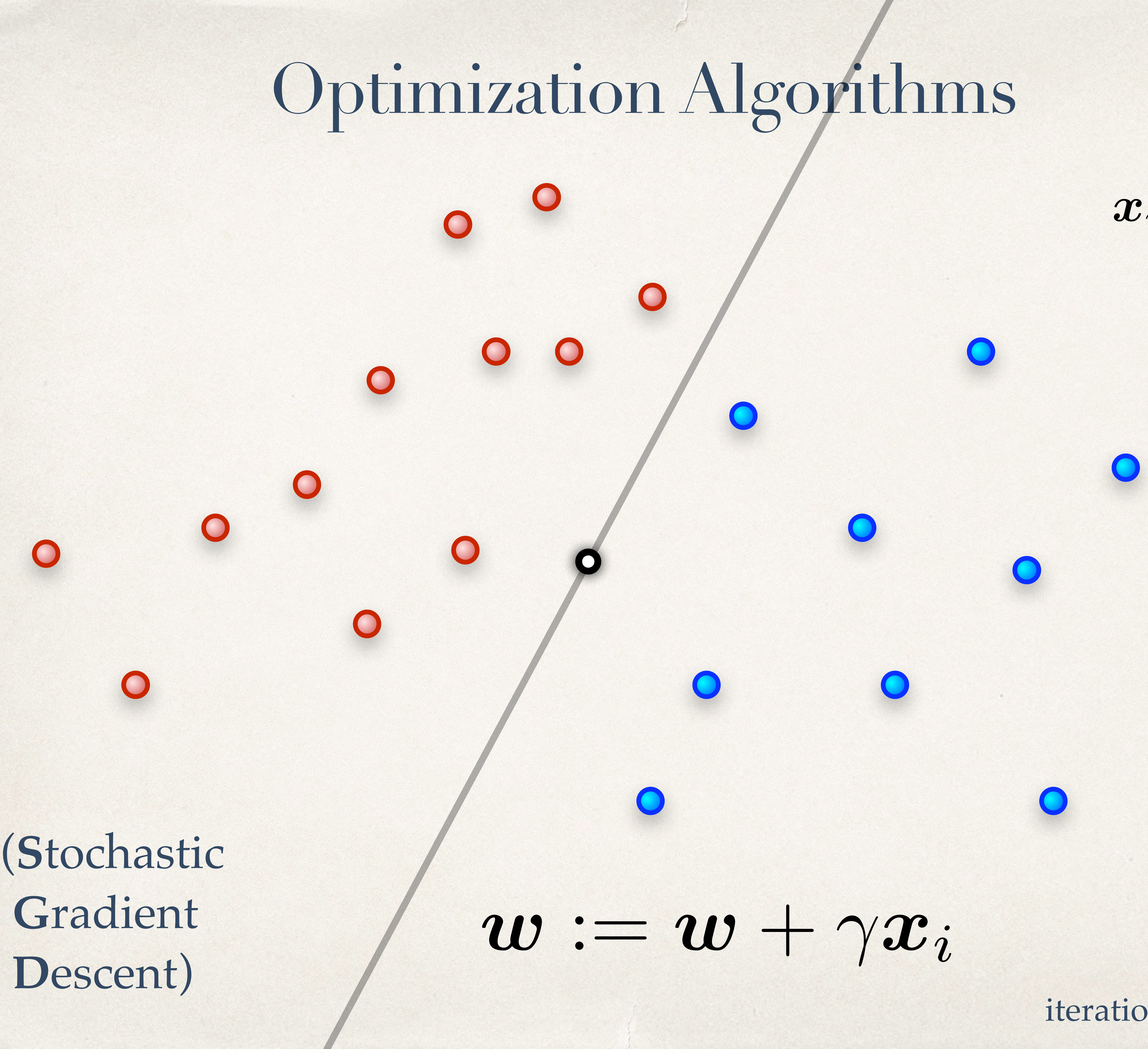


$x_i \in \mathbb{R}^d$

(Stochastic
Gradient
Descent)

$$\boldsymbol{w} := \boldsymbol{w} + \gamma \boldsymbol{x}_i$$

Optimization Algorithms



$$\boldsymbol{x}_i \in \mathbb{R}^d$$

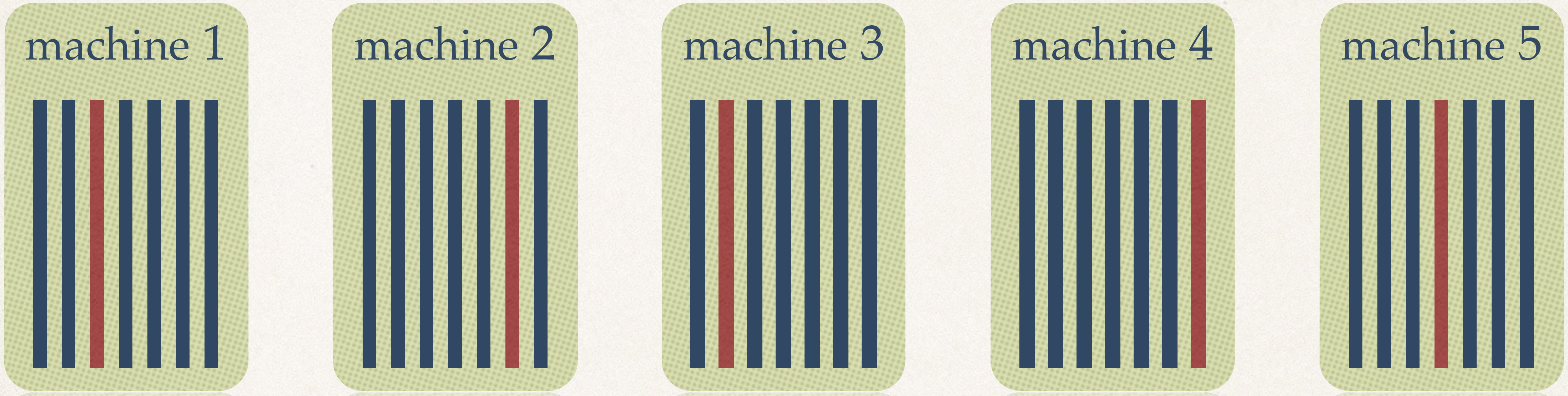
(Stochastic Gradient Descent)

$$\boldsymbol{w} := \boldsymbol{w} + \gamma \boldsymbol{x}_i$$

iteration cost: O(d)

Distributed Optimization

$$\boldsymbol{x}_i \in \mathbb{R}^d$$

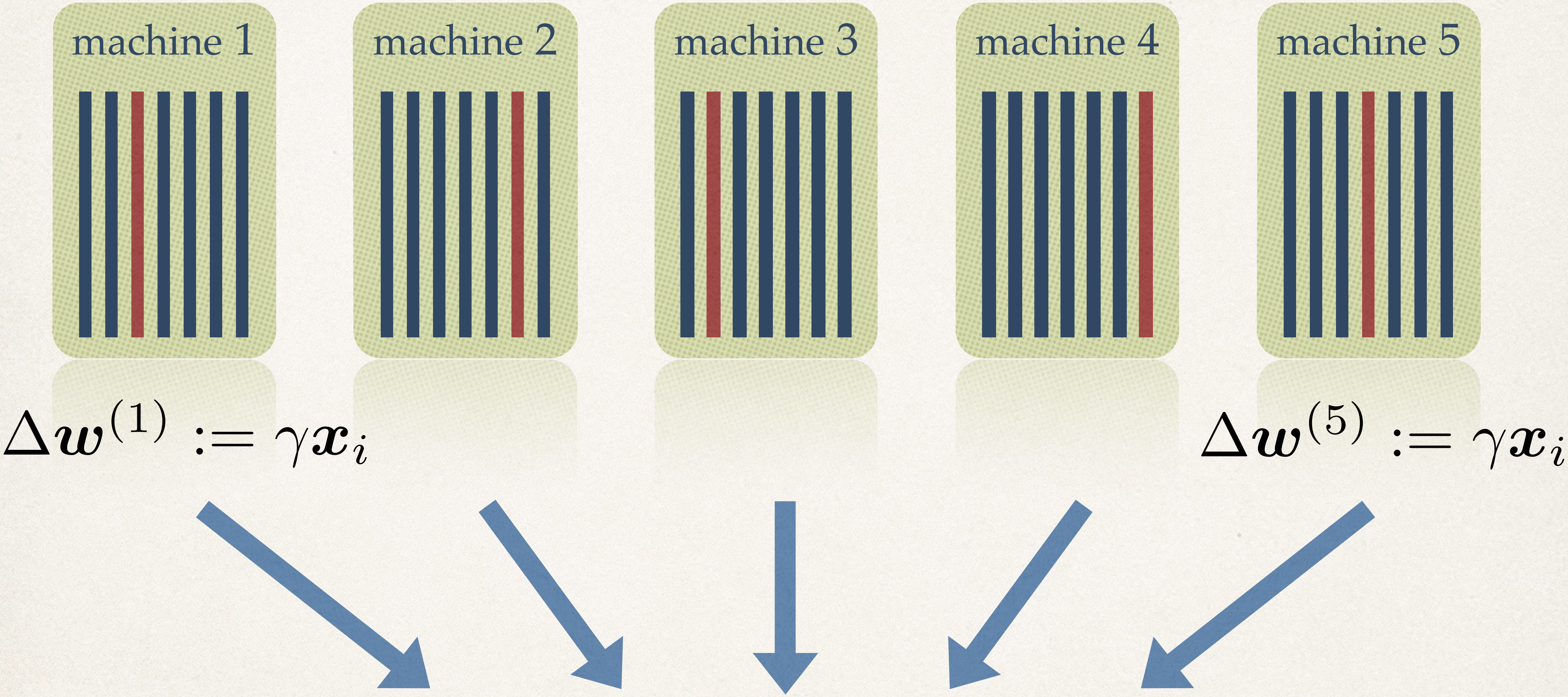


$$\Delta \boldsymbol{w}^{(1)} := \gamma \boldsymbol{x}_i$$

$$\Delta \boldsymbol{w}^{(5)} := \gamma \boldsymbol{x}_i$$

Distributed Optimization

$\boldsymbol{x}_i \in \mathbb{R}^d$



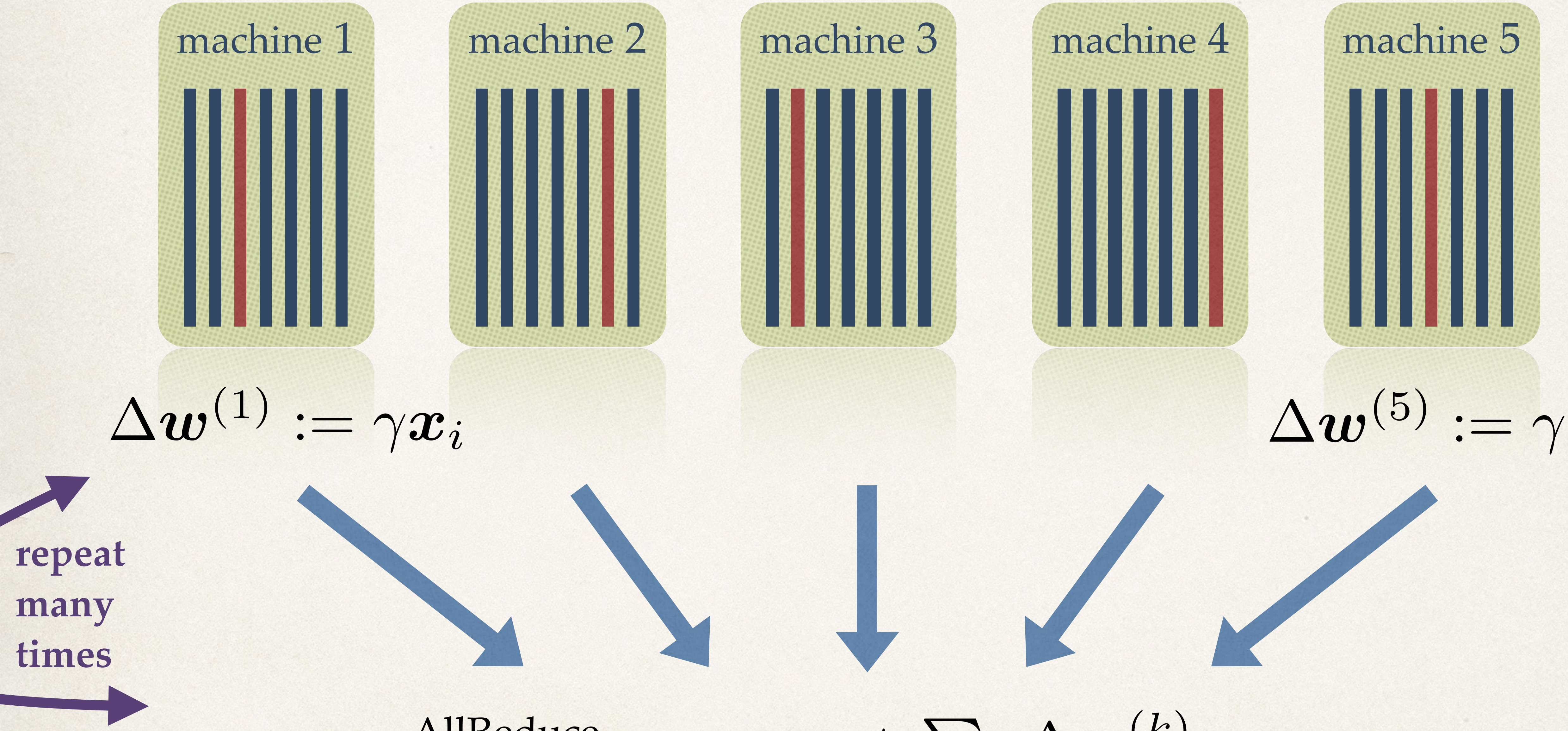
$\Delta\boldsymbol{w}^{(1)} := \gamma\boldsymbol{x}_i$

$\Delta\boldsymbol{w}^{(5)} := \gamma\boldsymbol{x}_i$

AllReduce $\boldsymbol{w} := \boldsymbol{w} + \sum_k \Delta\boldsymbol{w}^{(k)}$

Distributed Optimization

$\boldsymbol{x}_i \in \mathbb{R}^d$



$\Delta \boldsymbol{w}^{(1)} := \gamma \boldsymbol{x}_i$

$\Delta \boldsymbol{w}^{(5)} := \gamma \boldsymbol{x}_i$

repeat
many
times

AllReduce $\boldsymbol{w} := \boldsymbol{w} + \sum_k \Delta \boldsymbol{w}^{(k)}$

Naive Distributed SGD

The Cost of Communication

$\boldsymbol{v} \in \mathbb{R}^{100}$

- Reading $\boldsymbol{v}$ from Memory (RAM)

  100 ns

- Sending $\boldsymbol{v}$ to another Machine

  500’000 ns

- One Typical Map-Reduce Iteration *(Hadoop)*

  10’000’000’000 ns

“Big Data Analytics” Applications

Classification

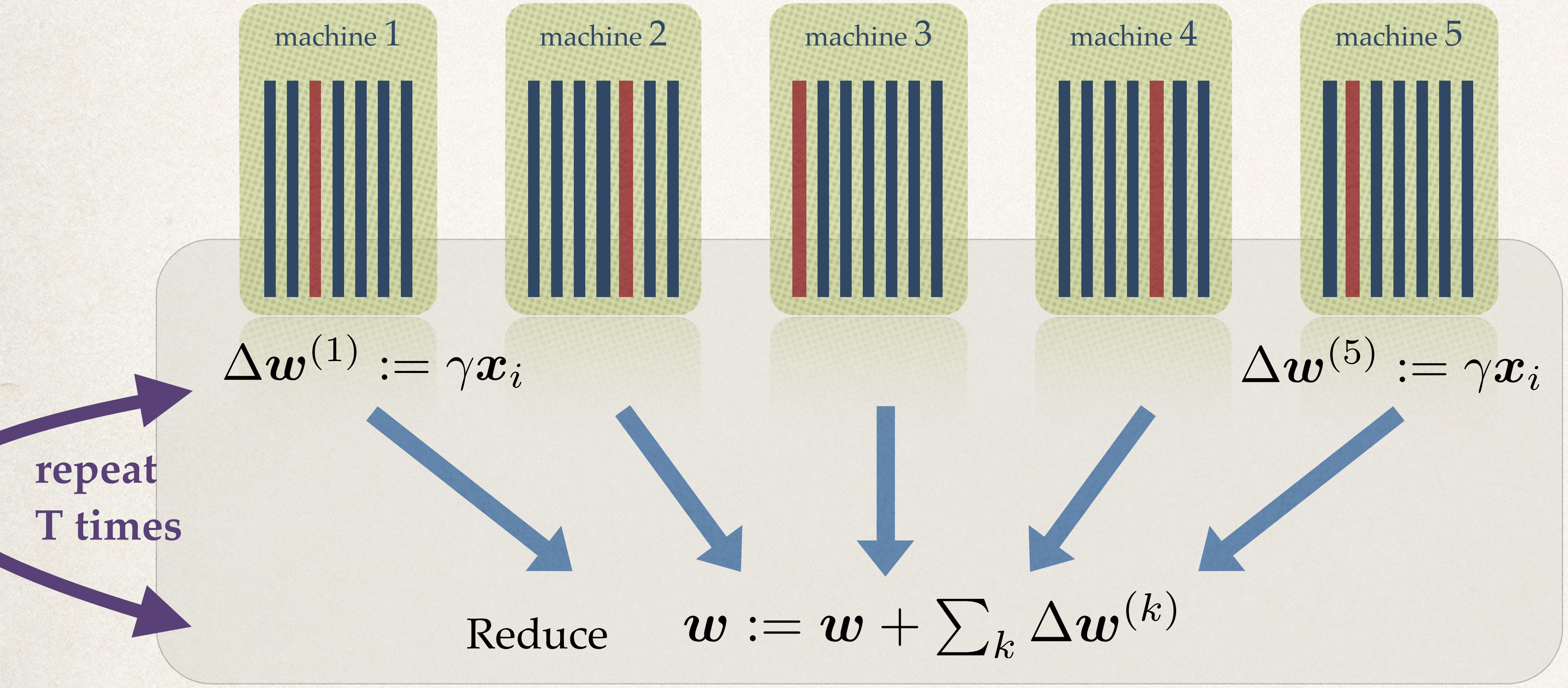
- Support Vector Machine *(SVM) (L1,L2)*
- Logistic Regression *(L1,L2)*
- Structured Prediction *(L1,L2)*

Regression

- Ridge Regression
- Least Squares variants *(L1,L2):*
- Lasso, Elastic-Net *(Feature Selection, Compressed Sensing)*

$$\min_{w \in \mathbb{R}^d} \left[P(\boldsymbol{w}) := \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{w}^T \boldsymbol{x}_i) \right]$$

Distributed Optimization



$\Delta \boldsymbol{w}^{(1)} := \gamma \boldsymbol{x}_i$

$\Delta \boldsymbol{w}^{(5)} := \gamma \boldsymbol{x}_i$

repeat T times

Reduce $\boldsymbol{w} := \boldsymbol{w} + \sum_k \Delta \boldsymbol{w}^{(k)}$

Naive Distributed SGD

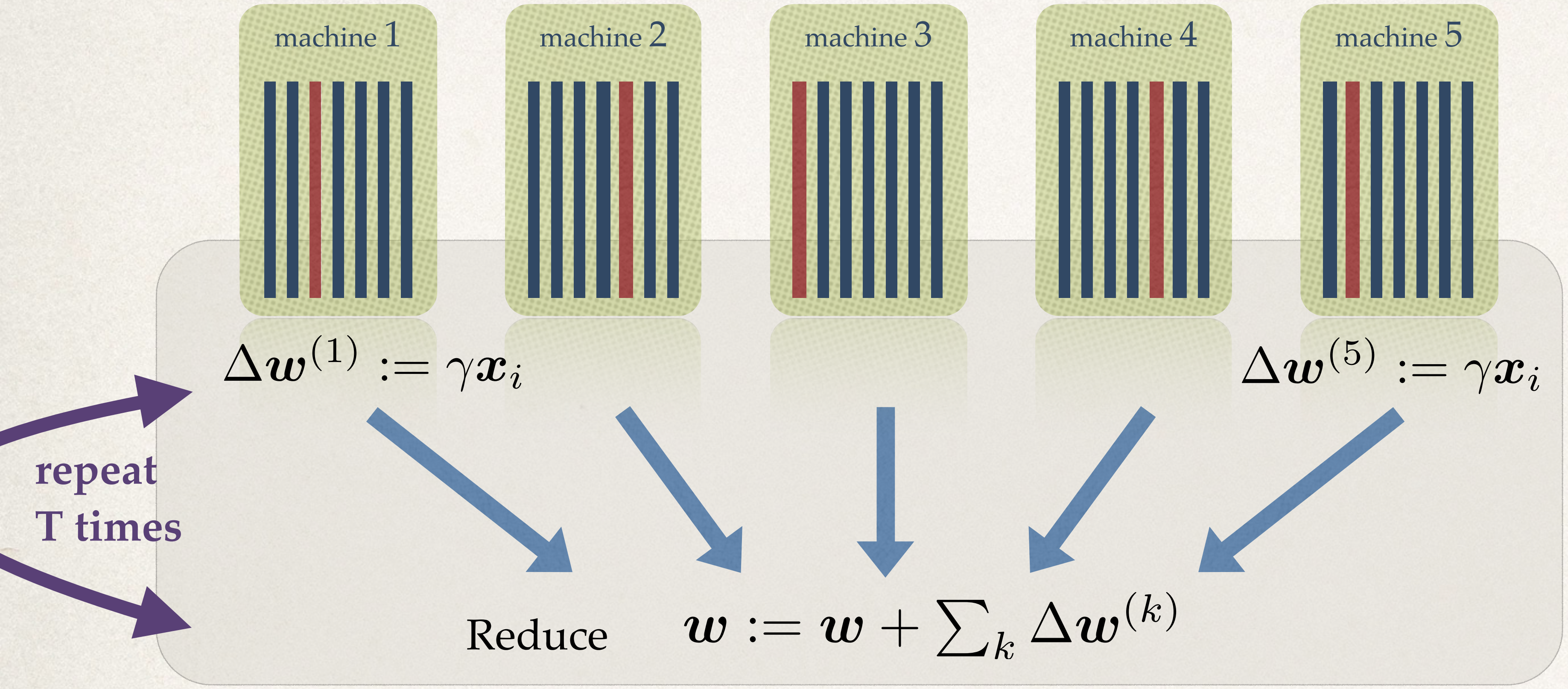
#local datapoints read: T

#communications: T

convergence: ✓

"always communicate"

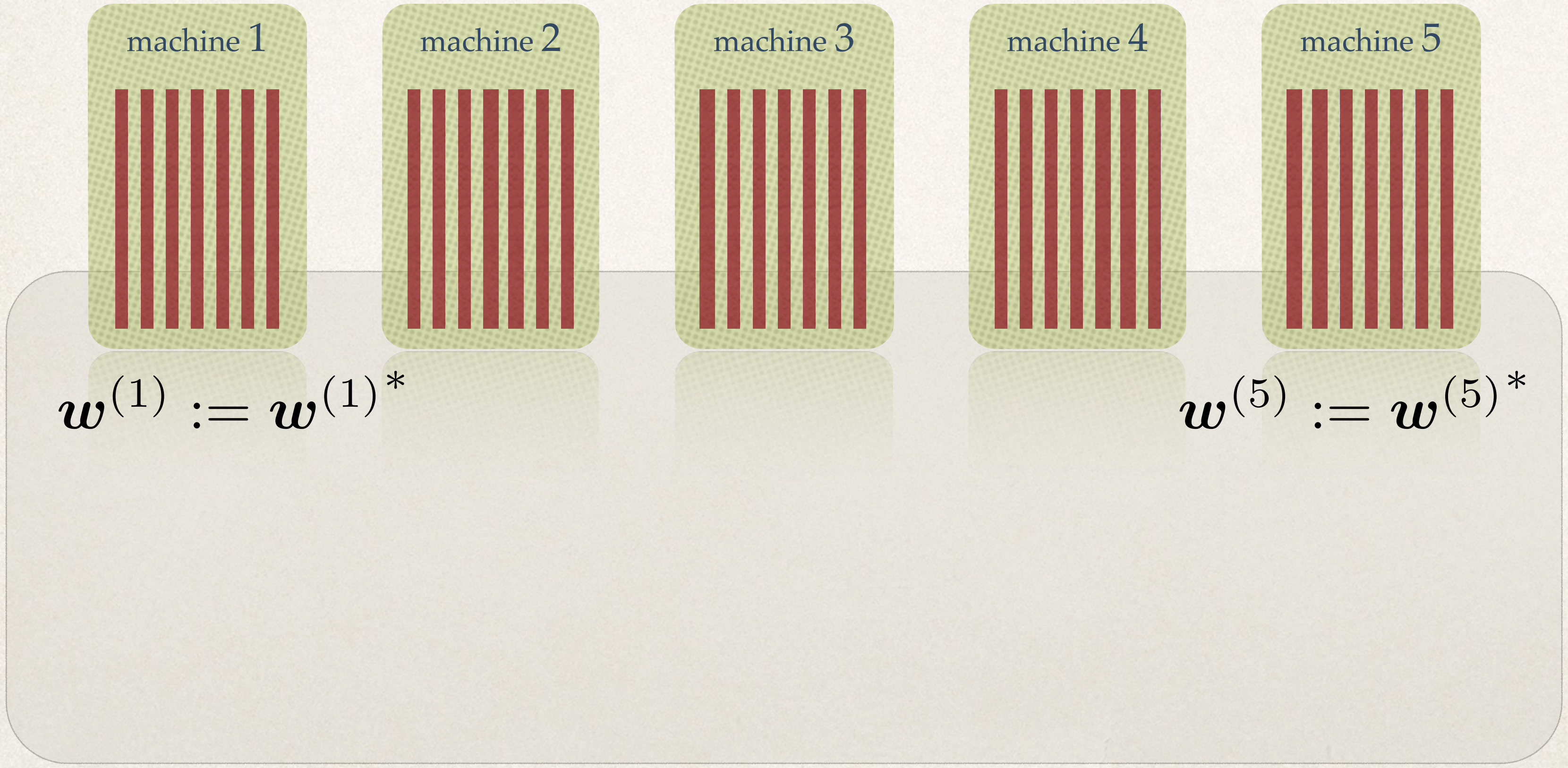
Communication: Always / Never



$\Delta \boldsymbol{w}^{(1)} := \gamma \boldsymbol{x}_i$

$\Delta \boldsymbol{w}^{(5)} := \gamma \boldsymbol{x}_i$

$\boldsymbol{w} := \boldsymbol{w} + \sum_k \Delta \boldsymbol{w}^{(k)}$



$\boldsymbol{w}^{(1)} := \boldsymbol{w}^{(1)*}$

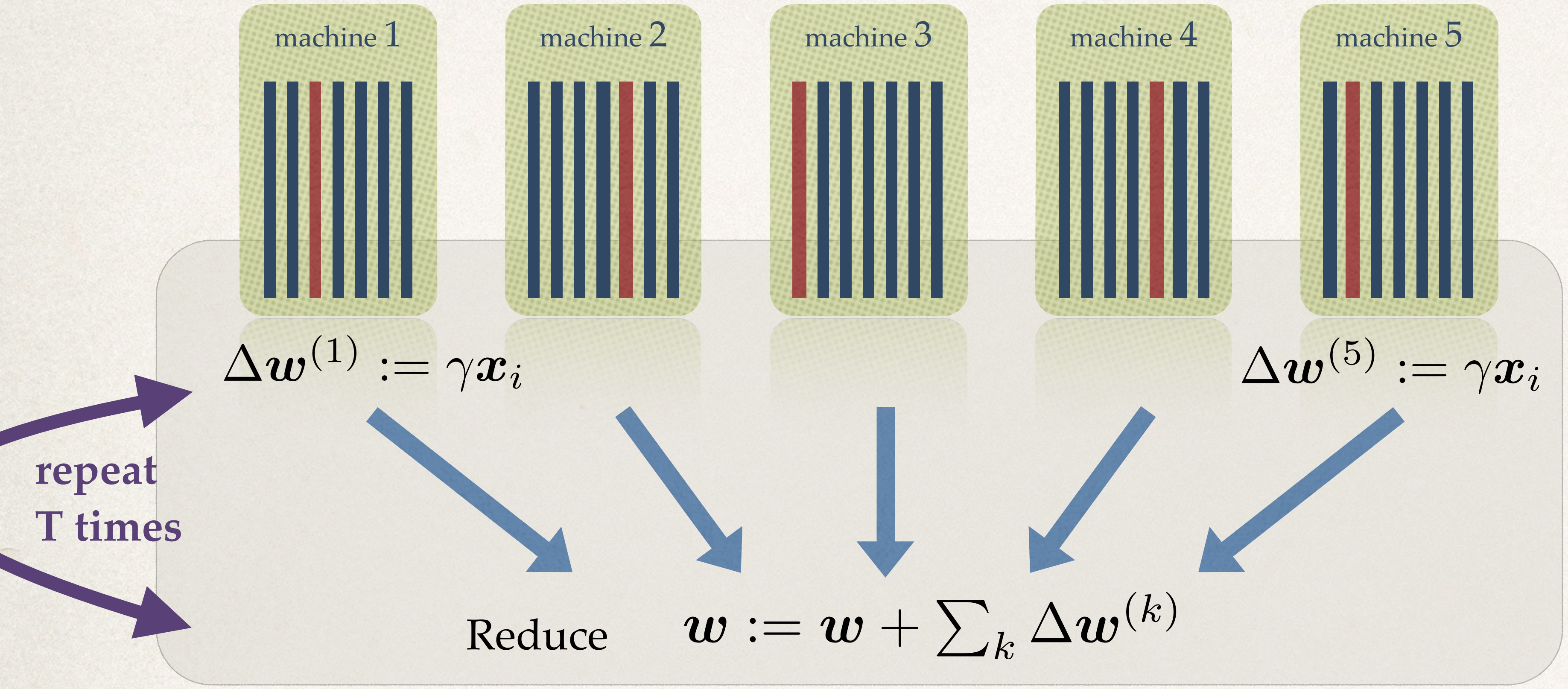
$\boldsymbol{w}^{(5)} := \boldsymbol{w}^{(5)*}$

Naive Distributed SGD

# local datapoints read:	T
# communications:	T
convergence:	✓

"always communicate"

Communication: Always / Never



machine 1 machine 2 machine 3 machine 4 machine 5

$\Delta \boldsymbol{w}^{(1)} := \gamma \boldsymbol{x}_i$ $\Delta \boldsymbol{w}^{(5)} := \gamma \boldsymbol{x}_i$

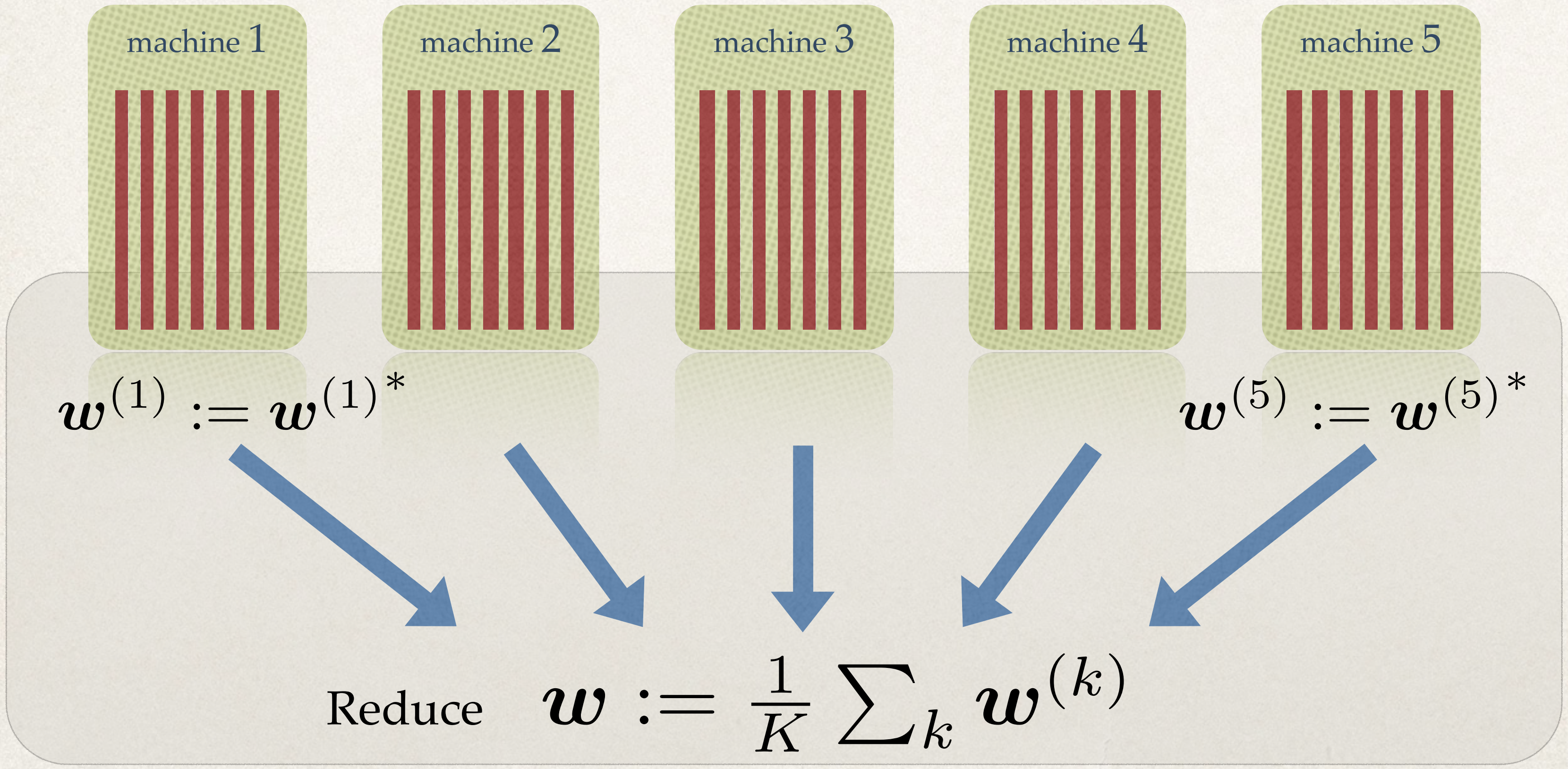
repeat T times

Reduce $\boldsymbol{w} := \boldsymbol{w} + \sum_k \Delta \boldsymbol{w}^{(k)}$

Naive Distributed SGD

# local datapoints read:	T
# communications:	T
convergence:	✓

“always communicate”

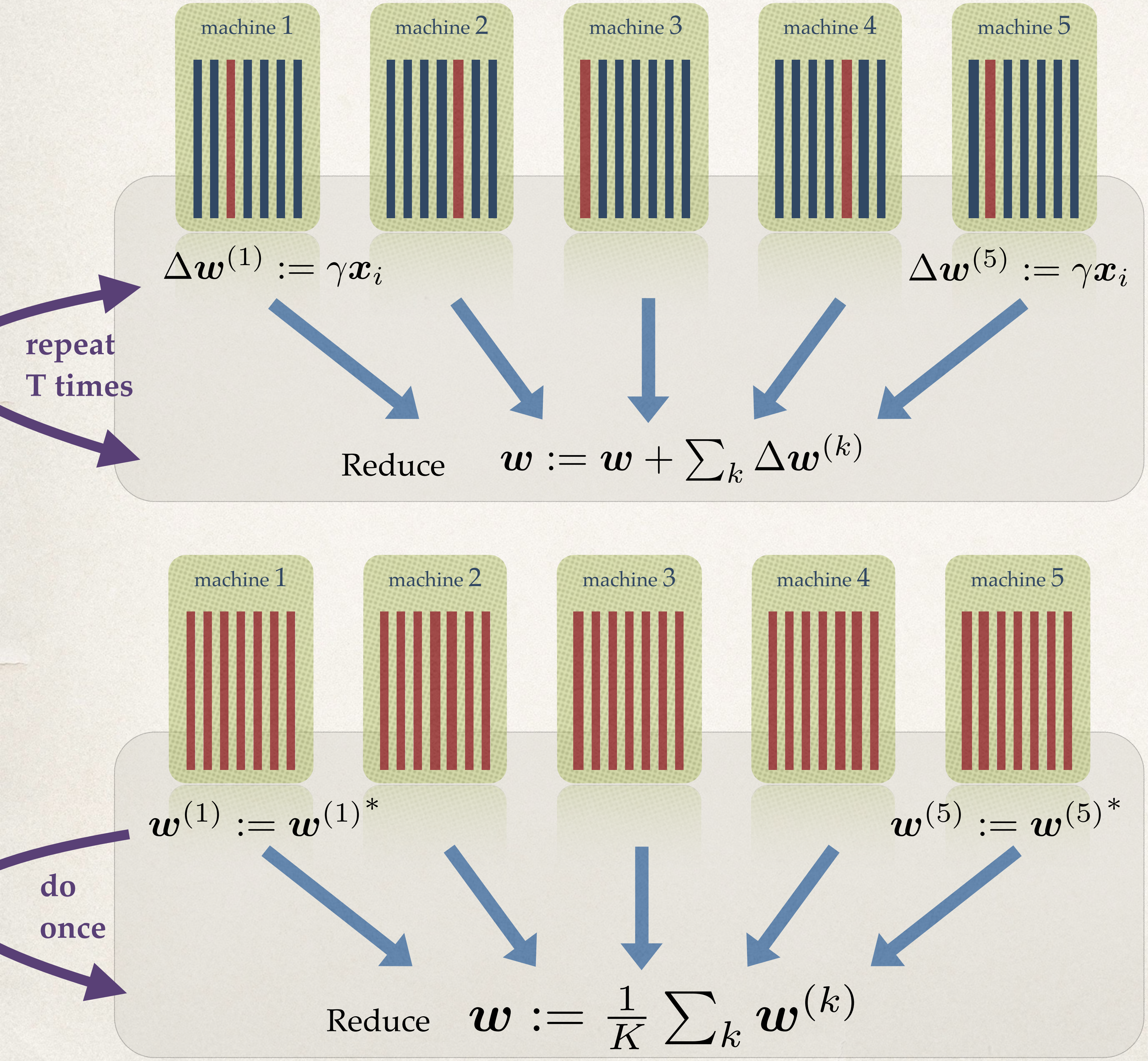


machine 1 machine 2 machine 3 machine 4 machine 5

$\boldsymbol{w}^{(1)} := \boldsymbol{w}^{(1)*}$ $\boldsymbol{w}^{(5)} := \boldsymbol{w}^{(5)*}$

Reduce $\boldsymbol{w} := \frac{1}{K} \sum_k \boldsymbol{w}^{(k)}$

Communication: Always / Never



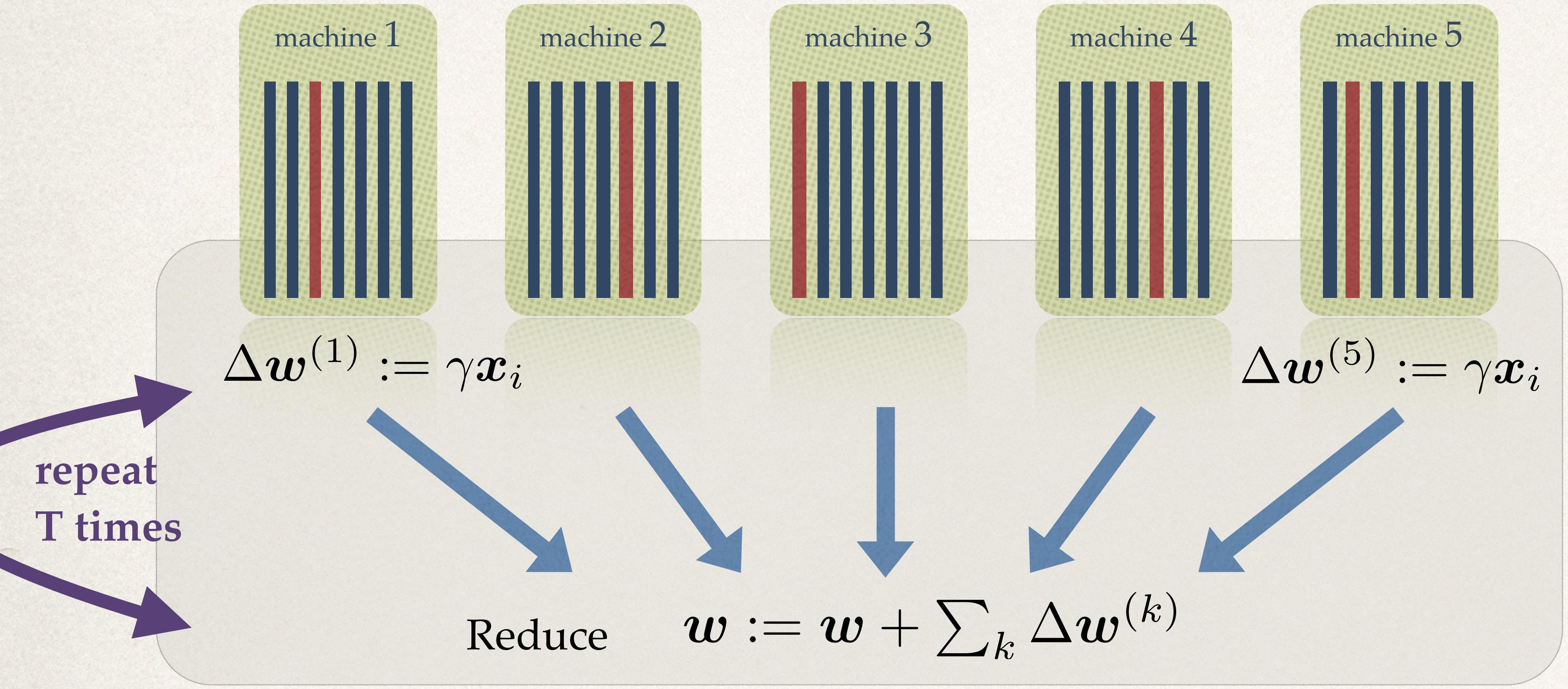
machine 1
machine 2
machine 3
machine 4
machine 5
$\Delta \boldsymbol{w}^{(1)} := \gamma \boldsymbol{x}_i$
$\Delta \boldsymbol{w}^{(5)} := \gamma \boldsymbol{x}_i$
repeat T times
Reduce $\boldsymbol{w} := \boldsymbol{w} + \sum_k \Delta \boldsymbol{w}^{(k)}$
machine 1
machine 2
machine 3
machine 4
machine 5
$\boldsymbol{w}^{(1)} := \boldsymbol{w}^{(1)*}$
$\boldsymbol{w}^{(5)} := \boldsymbol{w}^{(5)*}$
do once
Reduce $\boldsymbol{w} := \frac{1}{K} \sum_k \boldsymbol{w}^{(k)}$

Naive Distributed SGD

# local datapoints read:	T
# communications:	T
convergence:	✓

"always communicate"

Communication: Always / Never



machine 1 machine 2 machine 3 machine 4 machine 5

$\Delta \boldsymbol{w}^{(1)} := \gamma \boldsymbol{x}_i$

$\Delta \boldsymbol{w}^{(5)} := \gamma \boldsymbol{x}_i$

repeat T times

Reduce $\boldsymbol{w} := \boldsymbol{w} + \sum_k \Delta \boldsymbol{w}^{(k)}$

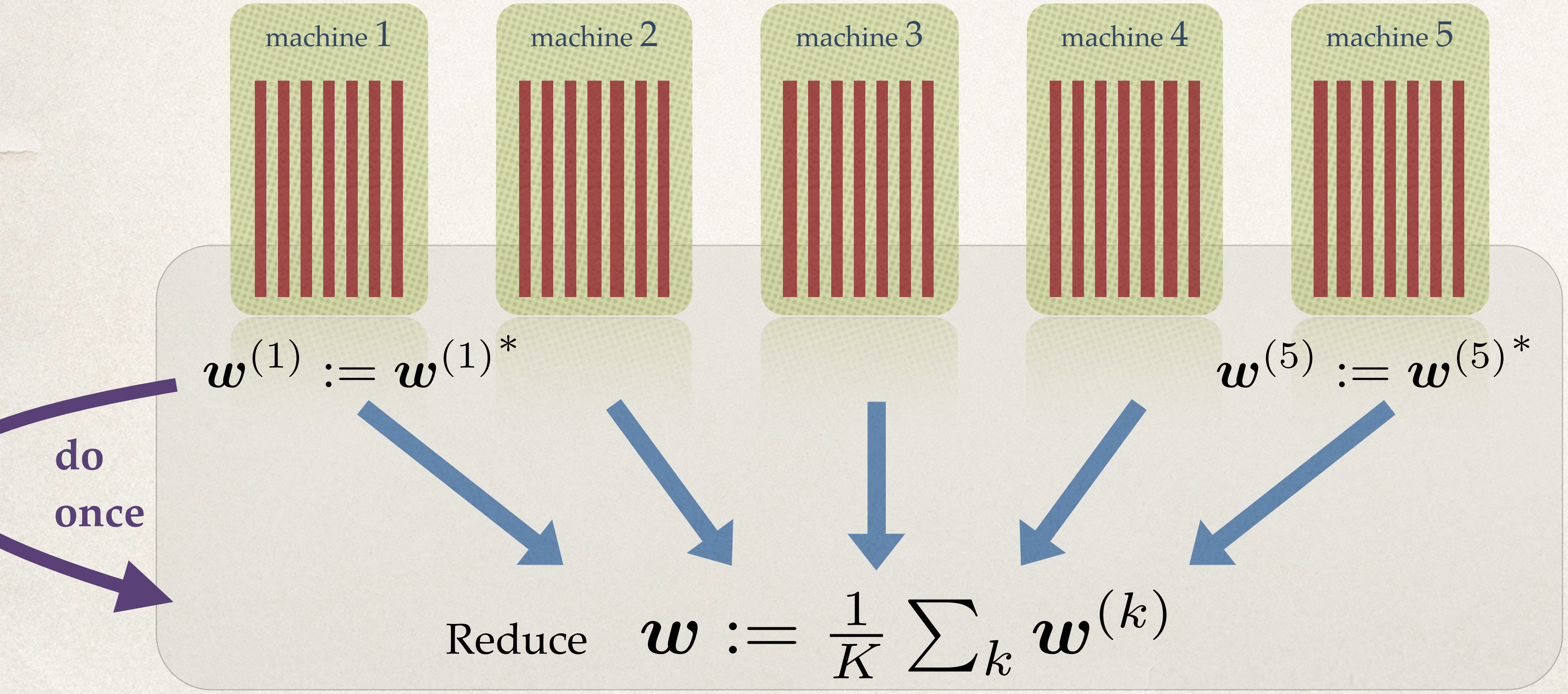
Naive Distributed SGD

# local datapoints read: T

# communications: T

convergence: ✓

"always communicate"



machine 1 machine 2 machine 3 machine 4 machine 5

$\boldsymbol{w}^{(1)} := \boldsymbol{w}^{(1)*}$

$\boldsymbol{w}^{(5)} := \boldsymbol{w}^{(5)*}$

do once

Reduce $\boldsymbol{w} := \frac{1}{K} \sum_k \boldsymbol{w}^{(k)}$

One-Shot Averaged Distributed Optimization

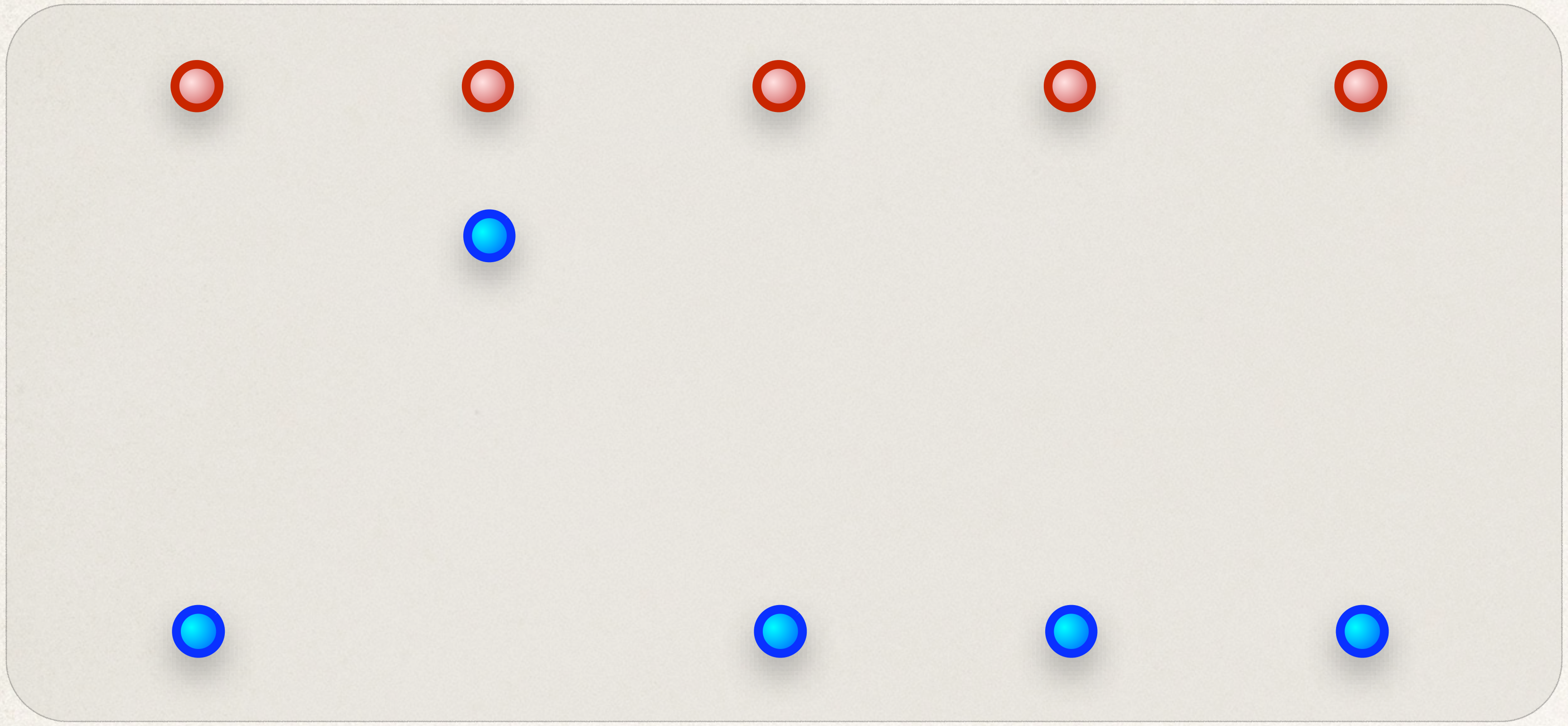
# local datapoints read: T

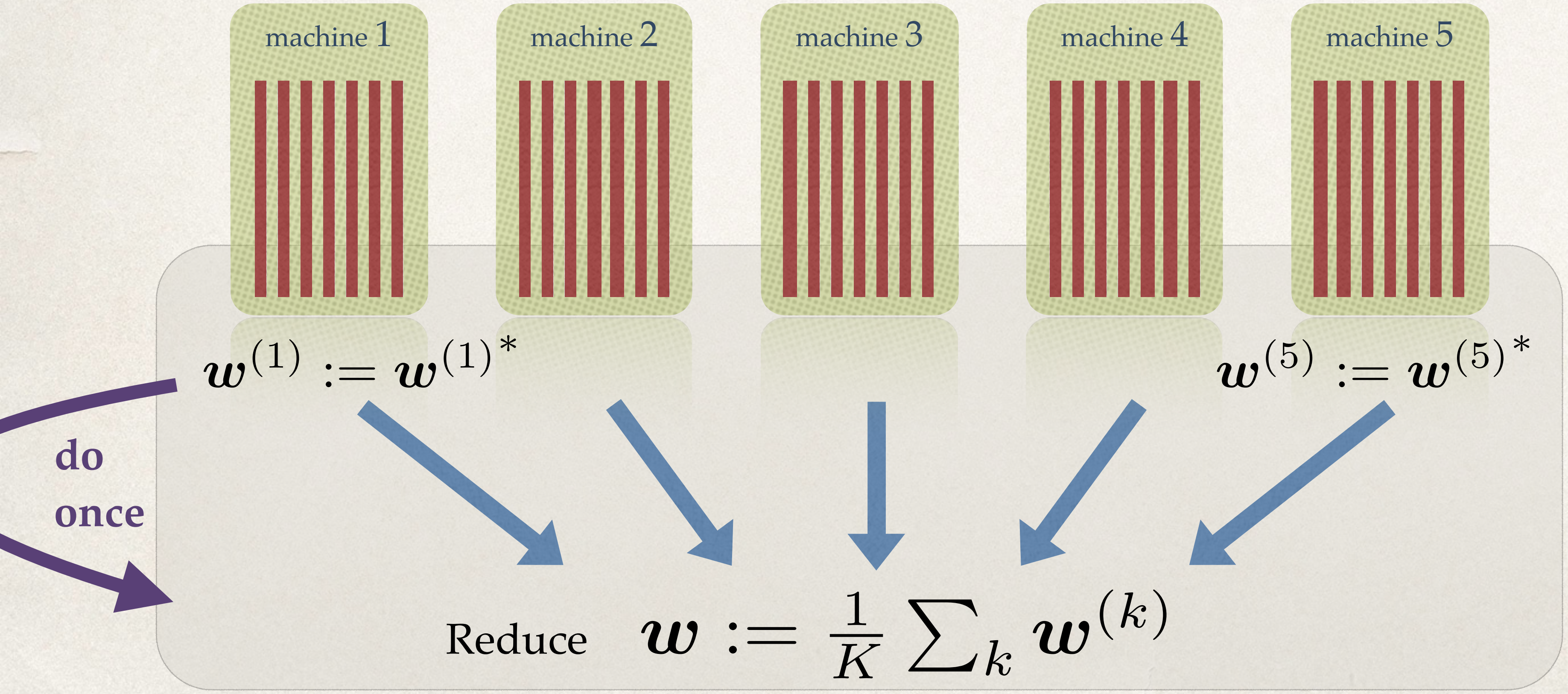
# communications: 1

convergence: ✗

"never communicate"

One-Shot Averaging Does Not Work





machine 1

machine 2

machine 3

machine 4

machine 5

$\boldsymbol{w}^{(1)} := \boldsymbol{w}^{(1)^*}$

$\boldsymbol{w}^{(5)} := \boldsymbol{w}^{(5)^*}$

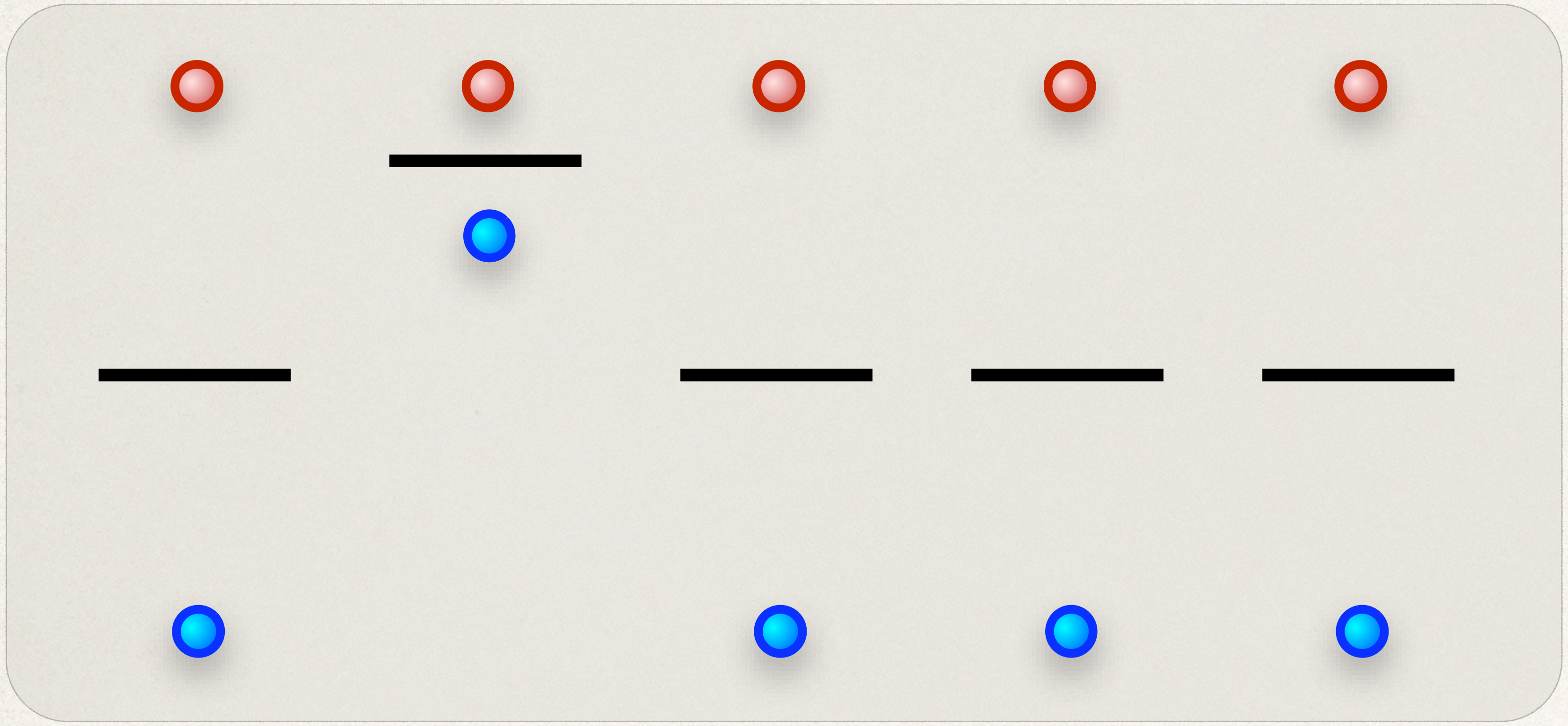
do once

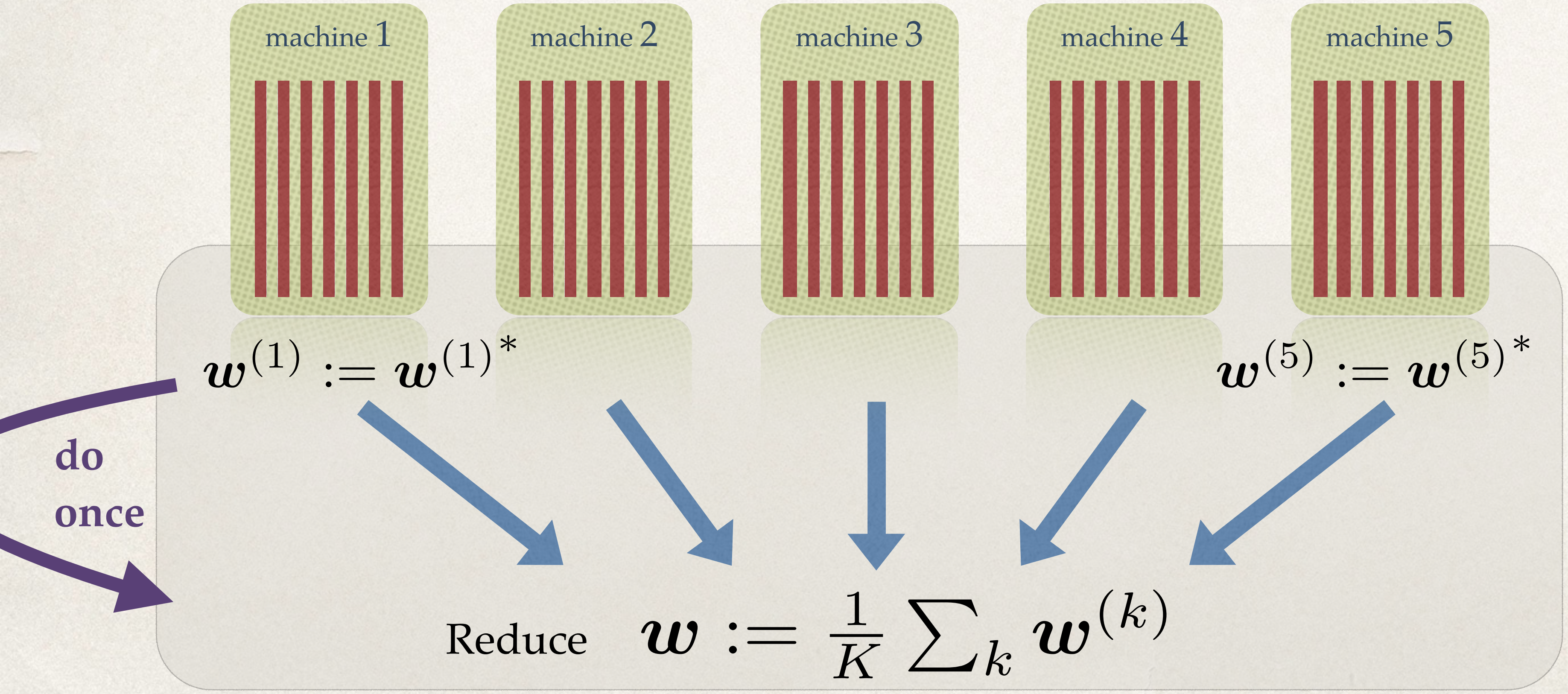
Reduce $\boldsymbol{w} := \frac{1}{K} \sum_k \boldsymbol{w}^{(k)}$

One-Shot Averaged Distributed Optimization

# local datapoints read:	T
# communications:	1
convergence:	✗

One-Shot Averaging Does Not Work





$\boldsymbol{w}^{(1)} := \boldsymbol{w}^{(1)*}$

$\boldsymbol{w}^{(5)} := \boldsymbol{w}^{(5)*}$

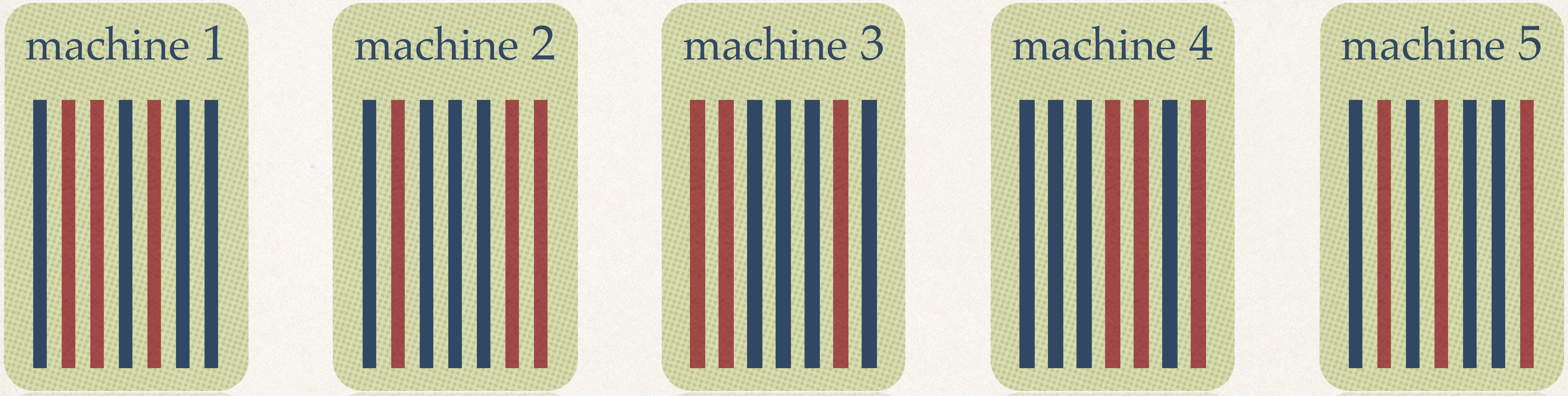
Reduce $\boldsymbol{w} := \frac{1}{K} \sum_k \boldsymbol{w}^{(k)}$

One-Shot Averaged Distributed Optimization

# local datapoints read:	T
# communications:	1
convergence:	✗

$\boldsymbol{w}_{(\boldsymbol{\alpha})} := A\boldsymbol{\alpha}$

Communication Efficient Distributed *Dual* Coordinate Ascent

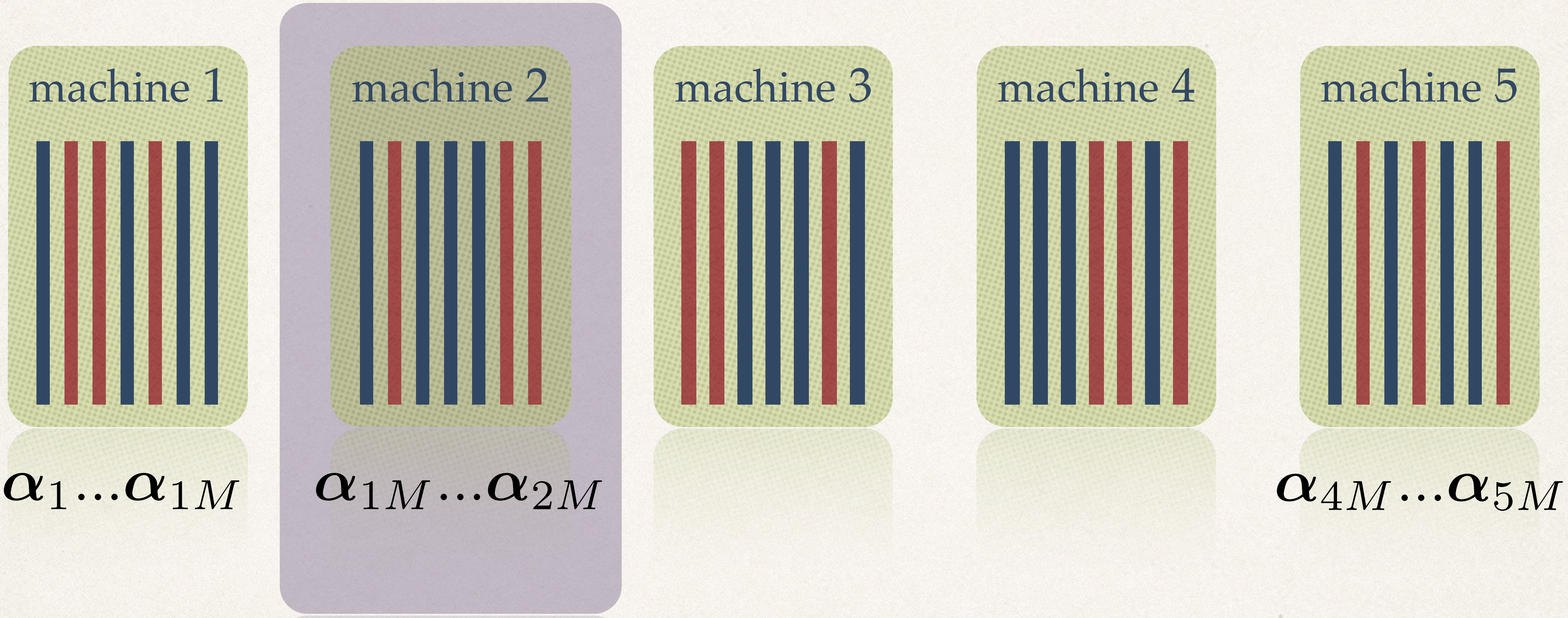


$\boldsymbol{\alpha}_1 \dots \boldsymbol{\alpha}_{1M}$ $\boldsymbol{\alpha}_{1M} \dots \boldsymbol{\alpha}_{2M}$ $\boldsymbol{\alpha}_{4M} \dots \boldsymbol{\alpha}_{5M}$

CoCoA

$\boldsymbol{w}_{(\boldsymbol{\alpha})} := A\boldsymbol{\alpha}$

Communication Efficient Distributed *Dual* Coordinate Ascent



machine 1

machine 2

machine 3

machine 4

machine 5

$\boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_{1M}$

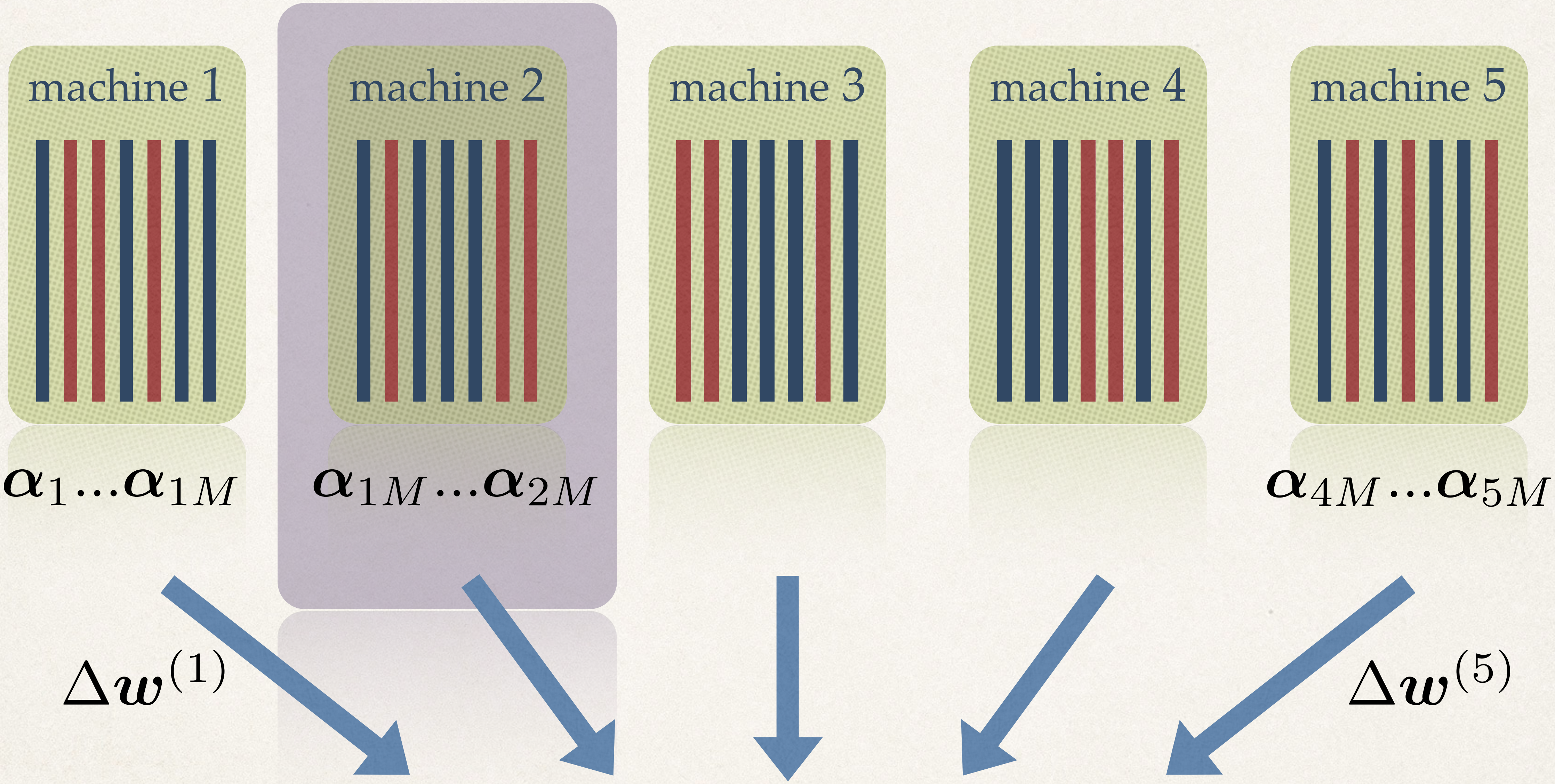
$\boldsymbol{\alpha}_{1M} \ldots \boldsymbol{\alpha}_{2M}$

$\boldsymbol{\alpha}_{4M} \ldots \boldsymbol{\alpha}_{5M}$

CoCoA

$\boldsymbol{w}_{(\boldsymbol{\alpha})} := A\boldsymbol{\alpha}$

Communication Efficient Distributed *Dual* Coordinate Ascent


machine 1
machine 2
machine 3
machine 4
machine 5
$\boldsymbol{\alpha}_1 ... \boldsymbol{\alpha}_{1M}$
$\boldsymbol{\alpha}_{1M} ... \boldsymbol{\alpha}_{2M}$
$\boldsymbol{\alpha}_{4M} ... \boldsymbol{\alpha}_{5M}$
$\Delta\boldsymbol{w}^{(1)}$
$\Delta\boldsymbol{w}^{(5)}$

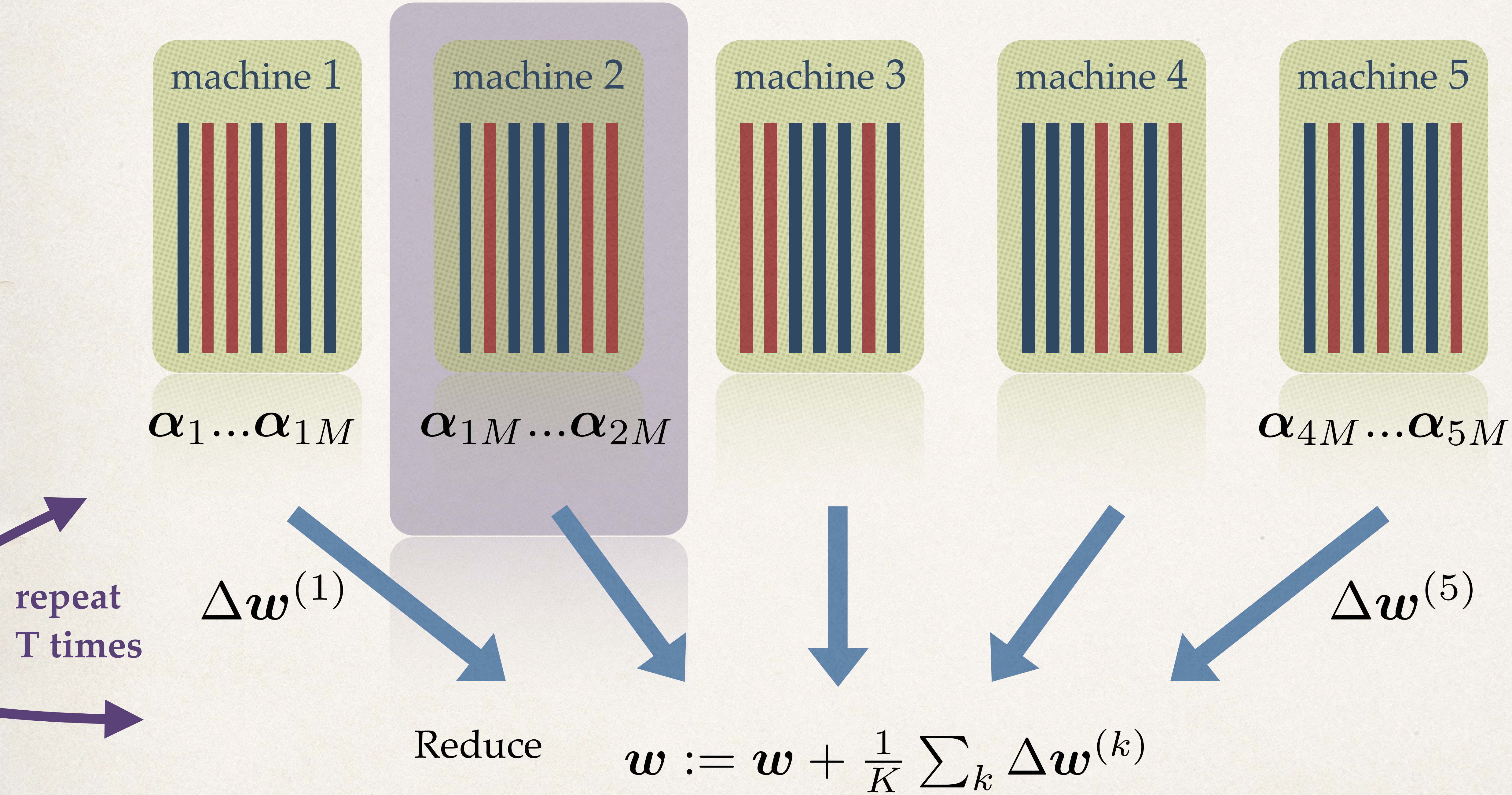
Reduce

$$\boldsymbol{w} := \boldsymbol{w} + \tfrac{1}{K} \textstyle\sum_k \Delta\boldsymbol{w}^{(k)}$$

CoCoA

$\boldsymbol{w}_{(\boldsymbol{\alpha})} := A\boldsymbol{\alpha}$

Communication Efficient Distributed *Dual* Coordinate Ascent



machine 1

machine 2

machine 3

machine 4

machine 5

$\boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_{1M}$

$\boldsymbol{\alpha}_{1M} \ldots \boldsymbol{\alpha}_{2M}$

$\boldsymbol{\alpha}_{4M} \ldots \boldsymbol{\alpha}_{5M}$

repeat
T times

$\Delta \boldsymbol{w}^{(1)}$

$\Delta \boldsymbol{w}^{(5)}$

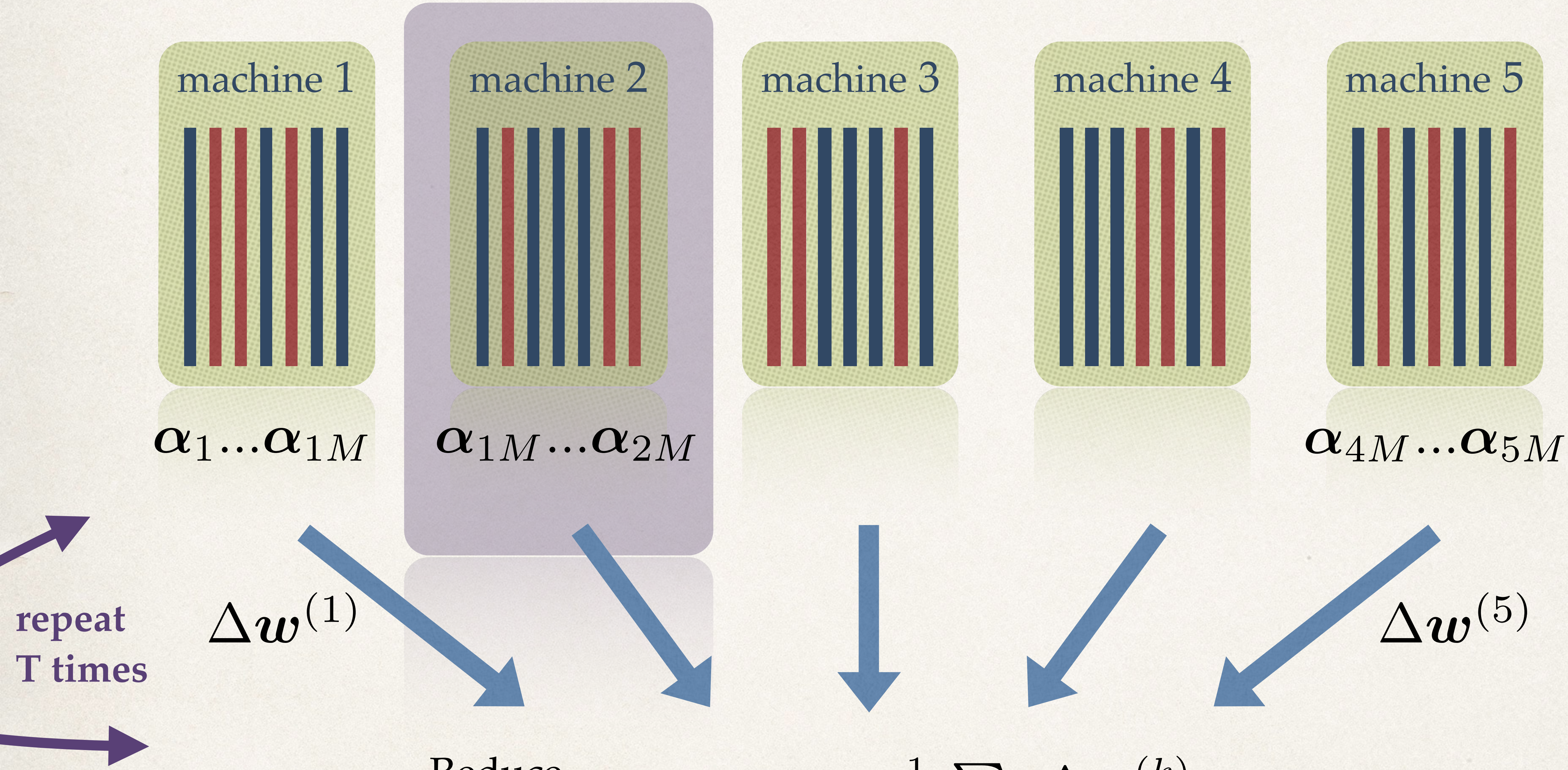
Reduce

$$\boldsymbol{w} := \boldsymbol{w} + \frac{1}{K} \sum_k \Delta \boldsymbol{w}^{(k)}$$

CoCoA

$\boldsymbol{w}_{(\boldsymbol{\alpha})} := A\boldsymbol{\alpha}$

Communication Efficient Distributed *Dual* Coordinate Ascent



machine 1

machine 2

machine 3

machine 4

machine 5

$\boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_{1M}$

$\boldsymbol{\alpha}_{1M} \ldots \boldsymbol{\alpha}_{2M}$

$\boldsymbol{\alpha}_{4M} \ldots \boldsymbol{\alpha}_{5M}$

repeat
T times

$\Delta \boldsymbol{w}^{(1)}$

$\Delta \boldsymbol{w}^{(5)}$

Reduce

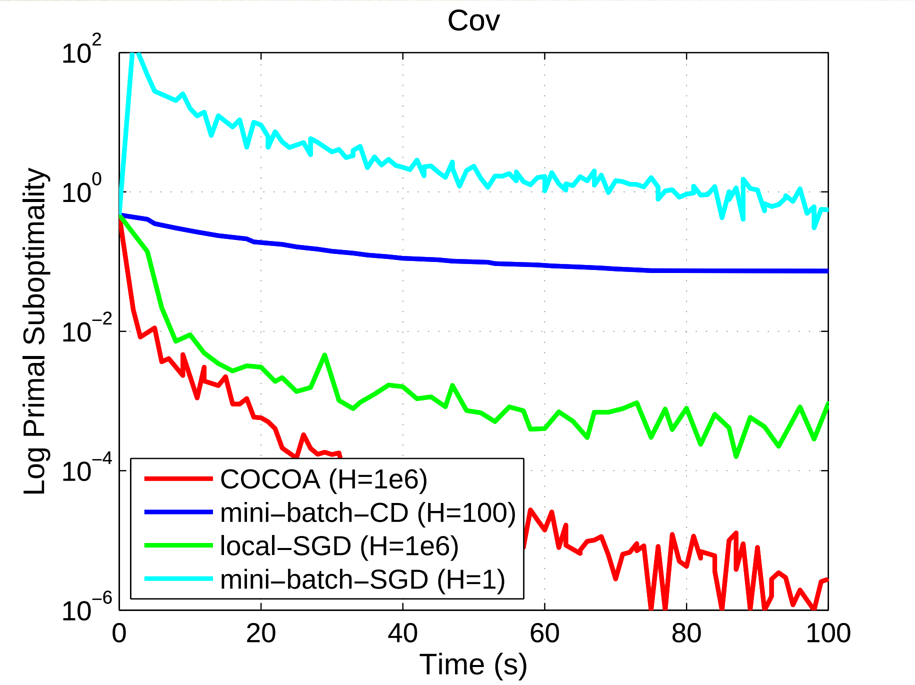
$$\boldsymbol{w} := \boldsymbol{w} + \tfrac{1}{K} \textstyle\sum_k \Delta \boldsymbol{w}^{(k)}$$

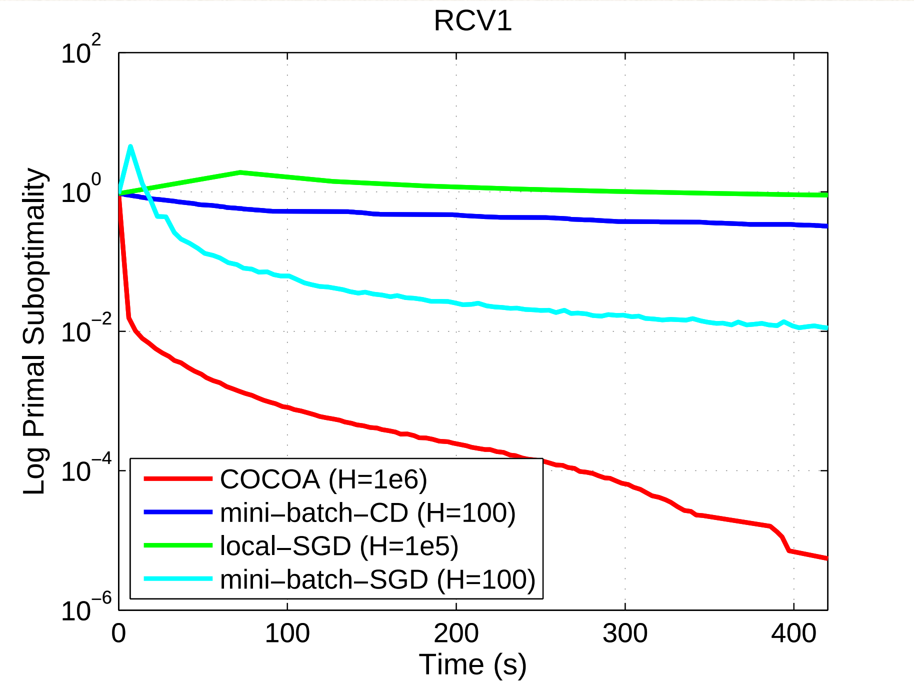
CoCoA

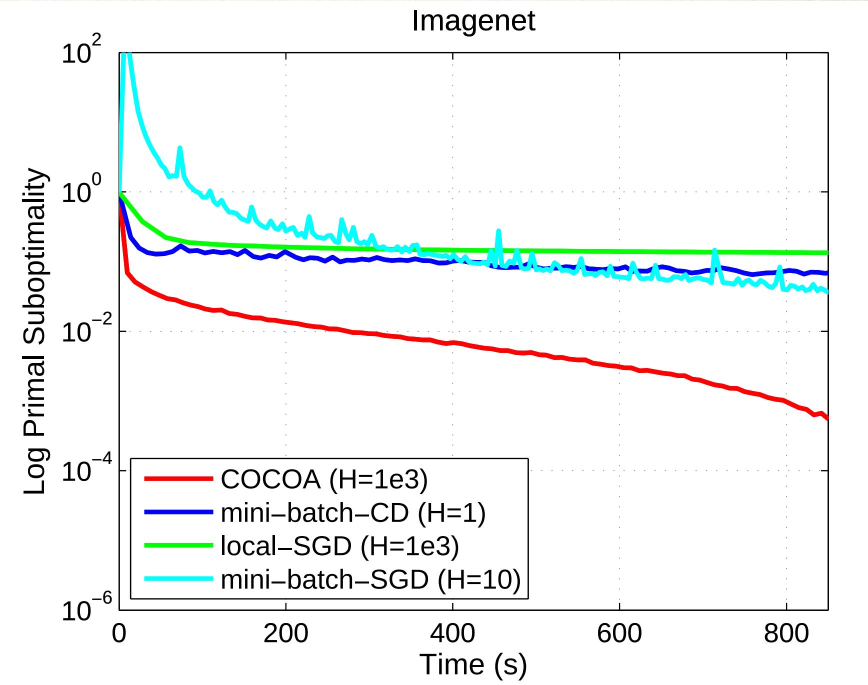
# local datapoints read:	TH
# communications:	T
convergence:	✓

Experiments

Dataset	**Training** n	**Features** d	**Sparsity**	λ	**Workers** K
cov	522,911	54	22.22%	1e-6	4
rcv1	677,399	47,236	0.16%	1e-6	8
imagenet	32,751	160,000	100%	1e-5	32







dissolve *struct*

Open Source Library for
Large Scale Machine Learning

built on Spark

DATA ANALYTICS LAB

Open Source

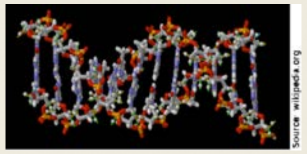
Applications:

Text

- Parsing
- POS tagging, chunking
- sentence alignment
- named entity recognition

Biology

Protein structure & function prediction



Vision

Horse Segmentation, OCR

more?

- Scene understanding
- object localization & recog.

Your Application?

Getting Started with Machine Learning

Does More Data Help?

- scikit learn
- kaggle.com

Thanks

“Communication-Efficient Distributed Dual Coordinate Ascent”

CoCoA paper (NIPS 2014)

CoCoA+ paper (ICML 2015)

Spark code is available on *github*

joint work with Virginia Smith, Martin Takáč, Chenxin Ma, Simone Forte, Tribhuvanesh Orekondy, Jonathan Terhorst, Sanjay Krishnan, Aurelien Lucchi, Peter Richtarik, Thomas Hofmann, Michael I. Jordan