Distributed Machine Learning

Algorithms and Open Source Implementations

Martin Jaggi
ETH Zurich

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SPINNINGBYTES

SDS | 2015, the 2nd Swiss Workshop on Data Science, 12th June 2015

monthly

Zürich Machine Learning and Data Science

[Link to Website]

Machine Learning?
AI?

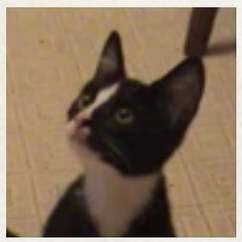
(Prediction)

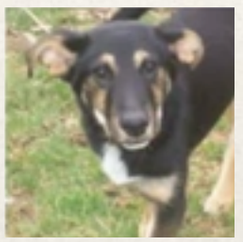
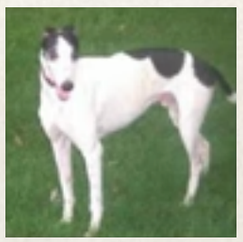
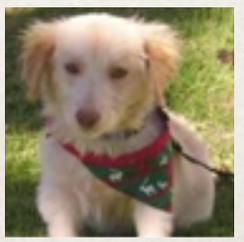
Classification & Regression

image source

Classification

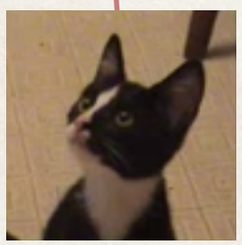
Training data

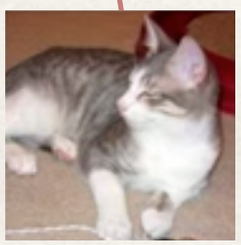
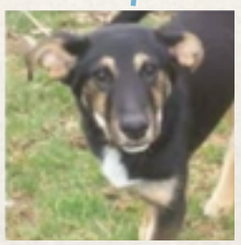
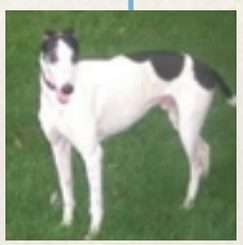
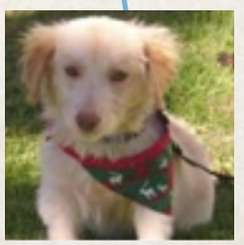


Classification

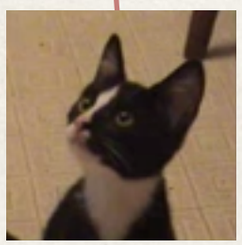
Training data

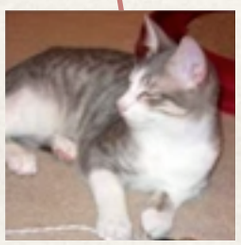
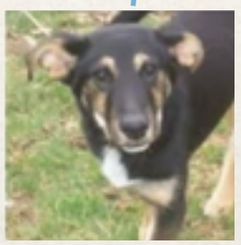
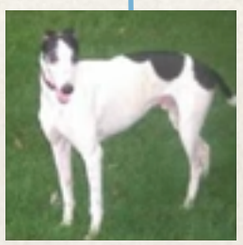
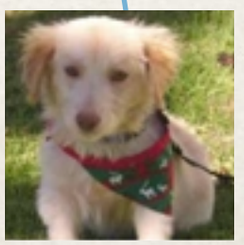


Classification

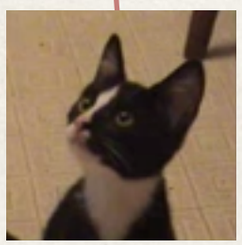
Training data

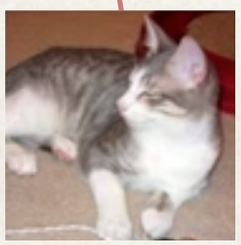
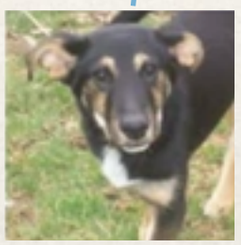
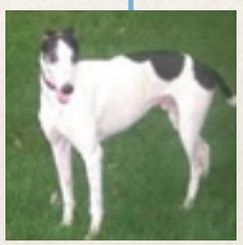
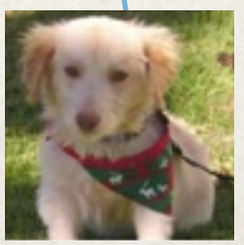


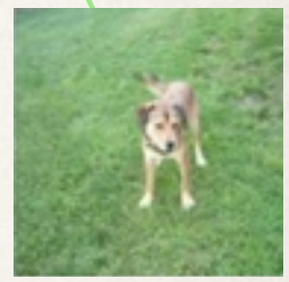
Classification

Training data

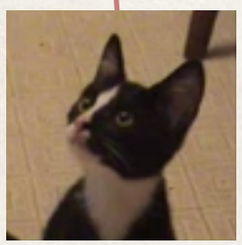


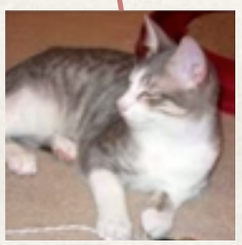
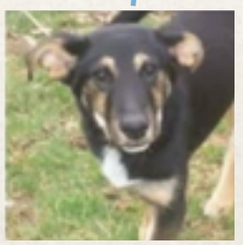
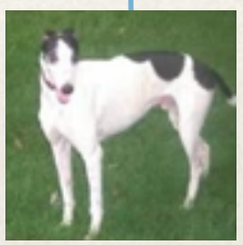
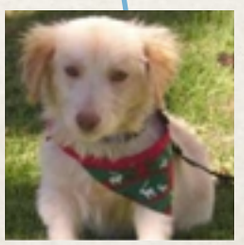





Classification



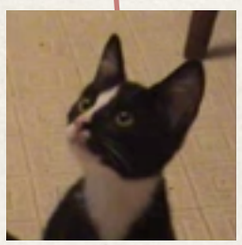
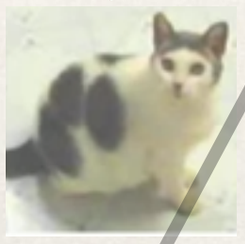
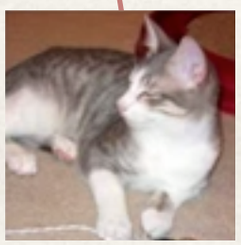
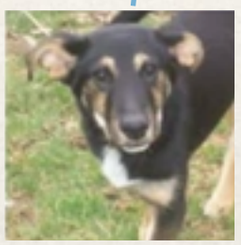
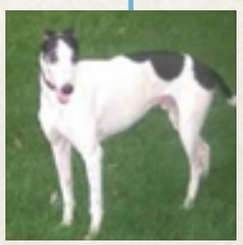
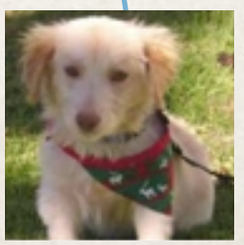
Training data



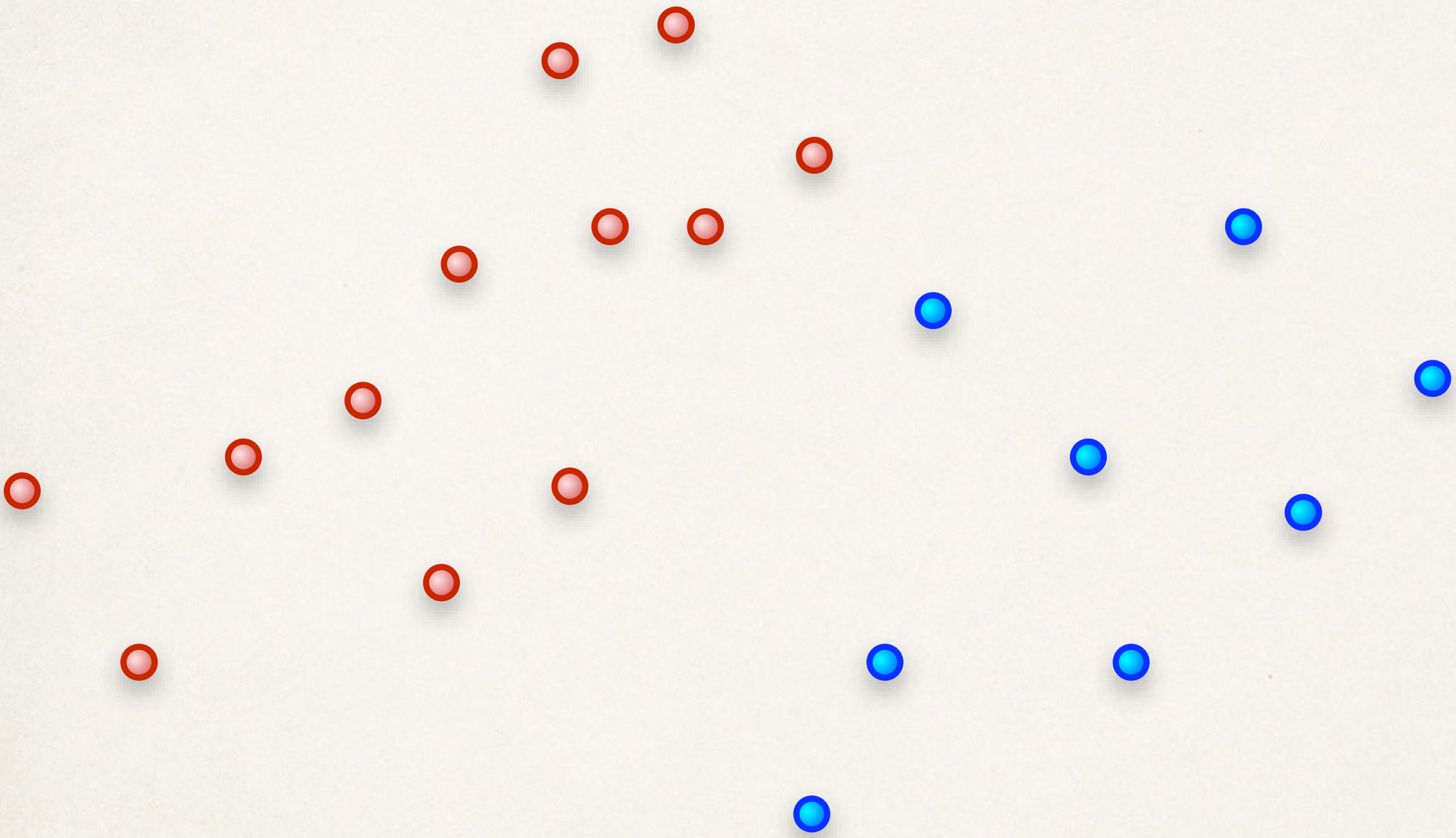





The Training Algorithm

Training data

The Training Algorithm



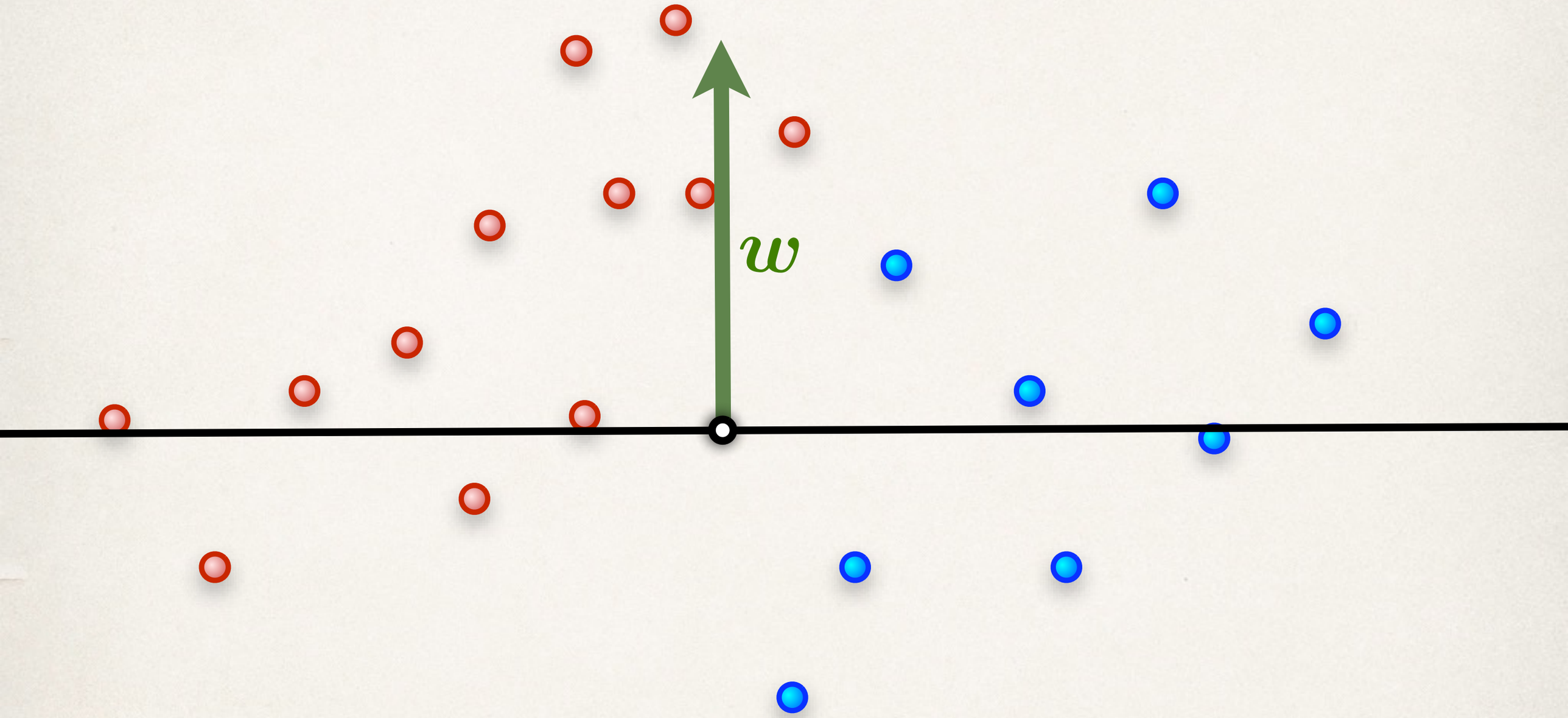
Perceptron

(Rosenblatt 1957)

Support-Vektor-Maschine

(Cortes & Vapnik 1995)

The Training Algorithm



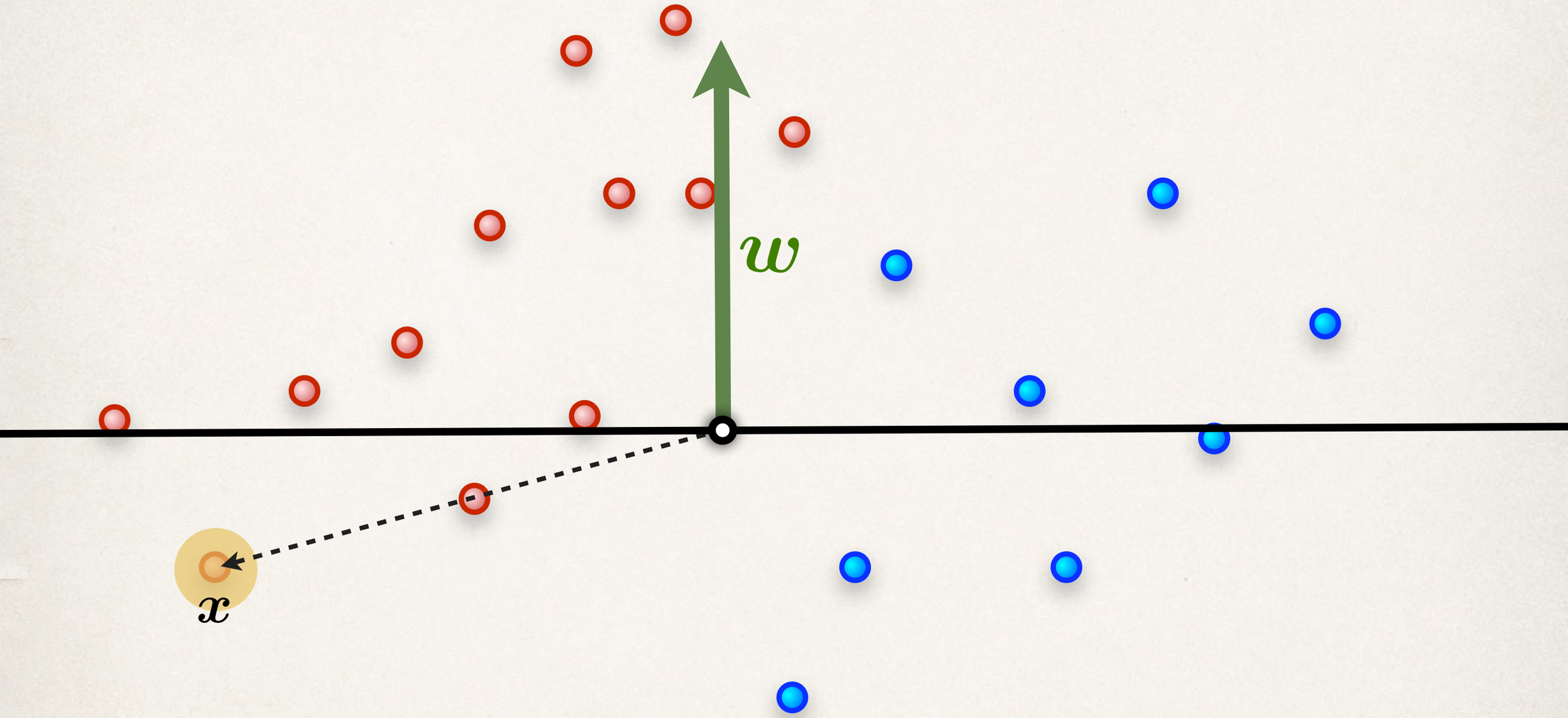
Perceptron

(Rosenblatt 1957)

Support-Vektor-Maschine

(Cortes & Vapnik 1995)

The Training Algorithm



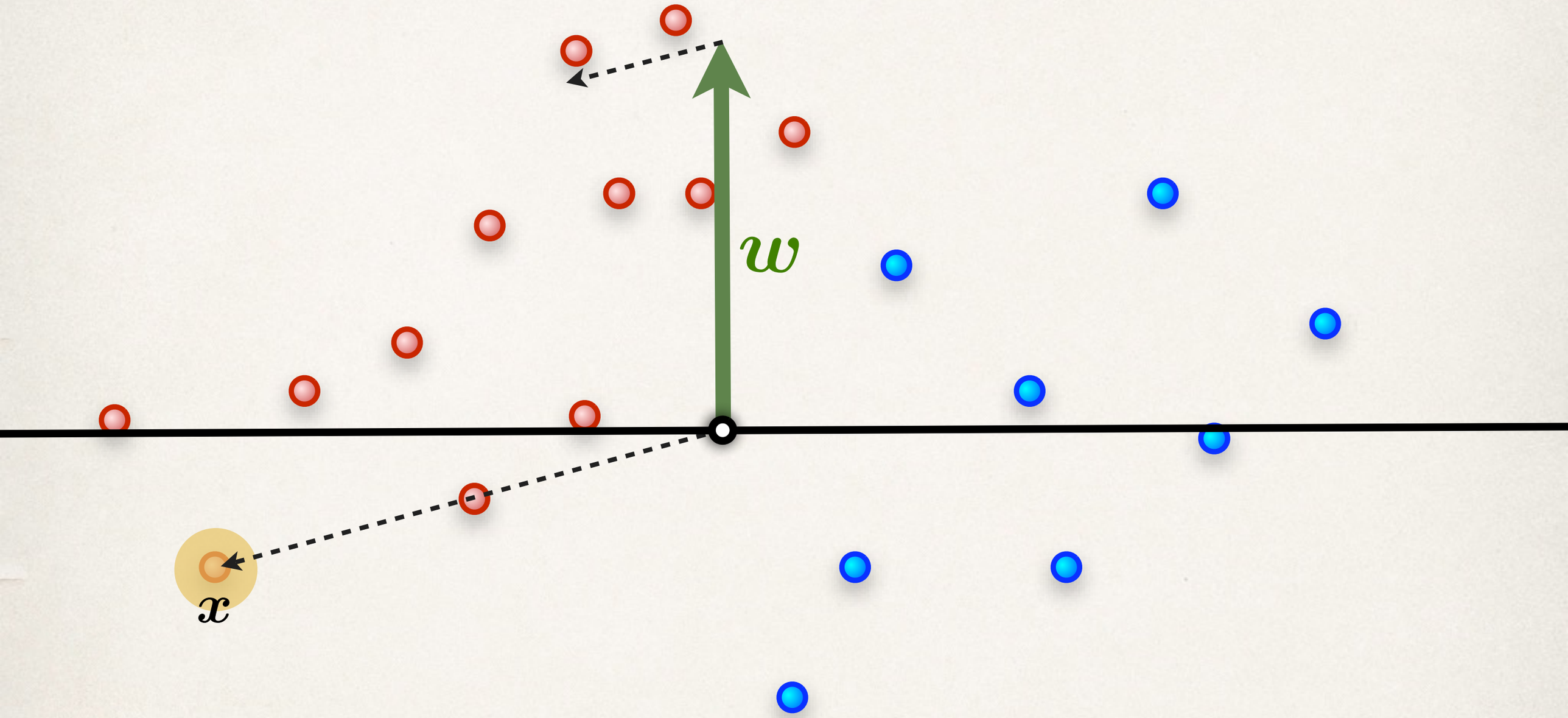
Perceptron

(Rosenblatt 1957)

Support-Vektor-Maschine

(Cortes & Vapnik 1995)

The Training Algorithm



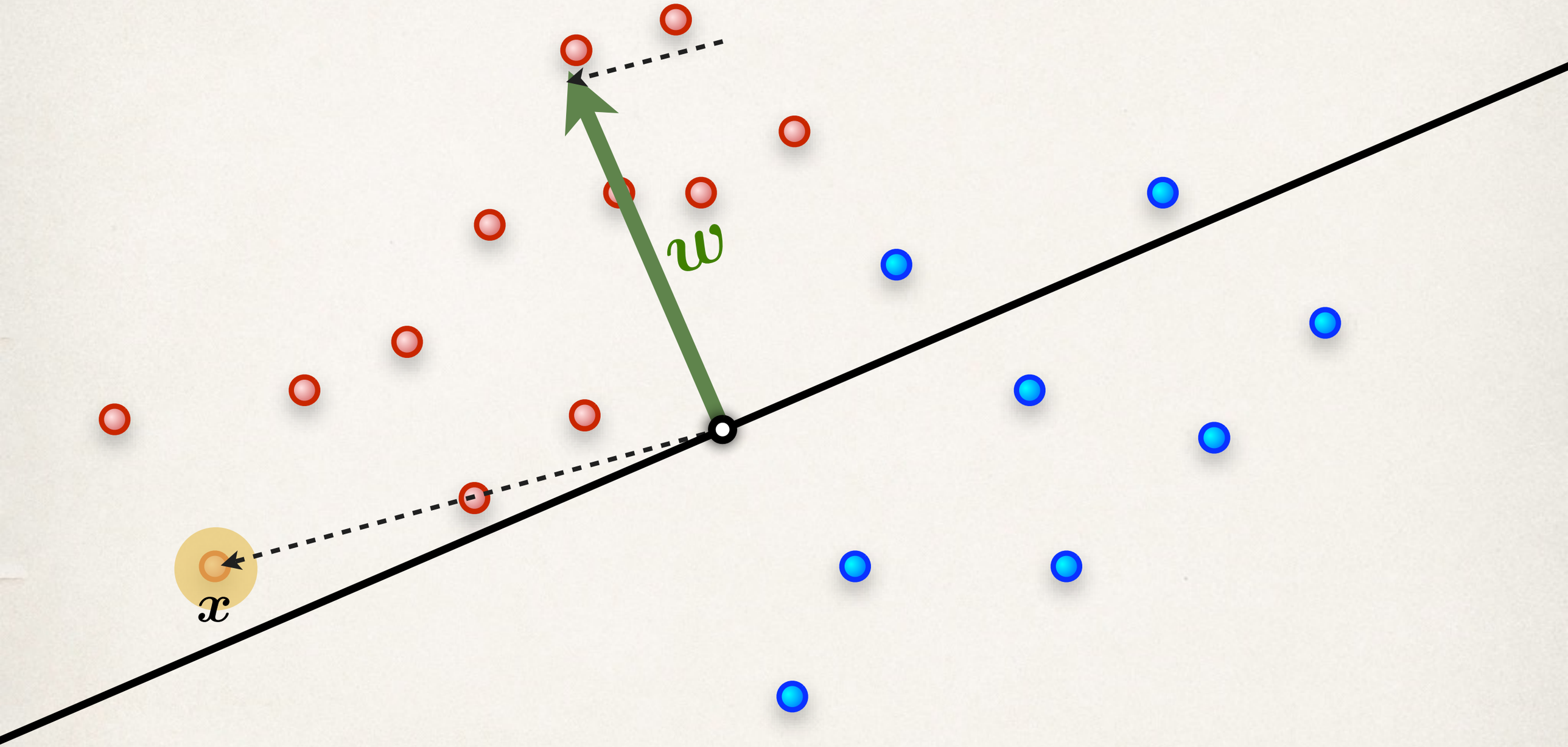
Perceptron

(Rosenblatt 1957)

Support-Vektor-Maschine

(Cortes & Vapnik 1995)

The Training Algorithm



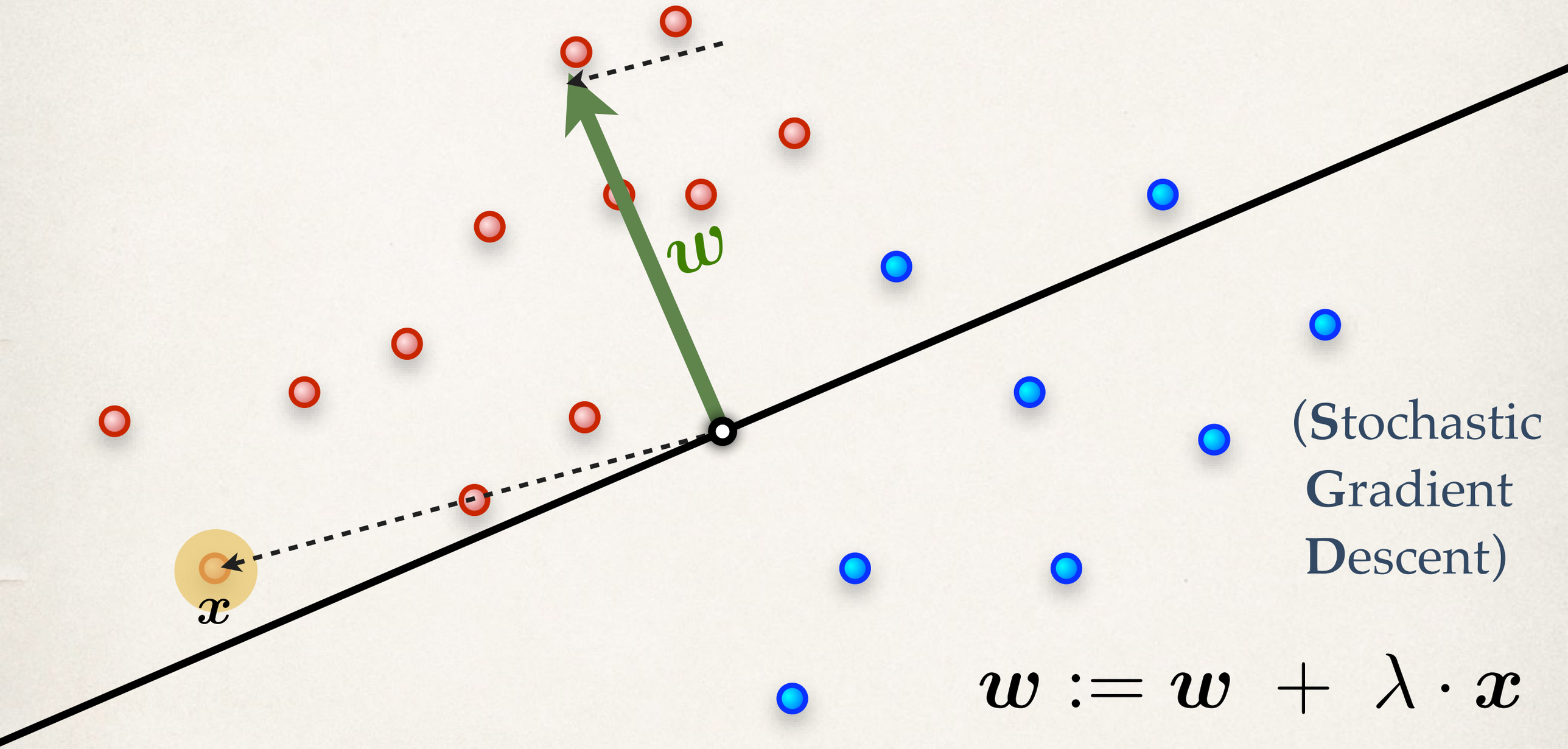
Perceptron

(Rosenblatt 1957)

Support-Vektor-Maschine

(Cortes & Vapnik 1995)

The Training Algorithm



(Stochastic Gradient Descent)

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

Perceptron

(Rosenblatt 1957)

Support-Vektor-Maschine

(Cortes & Vapnik 1995)

What has changed?

1950s: 10^{3} FLOPS

2010s: 10^{15} FLOPS

What has changed?

1950s: 10^3 FLOPS

2010s: 10^{15} FLOPS

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

What has changed?

1950s: 10^3 FLOPS

2010s: 10^{15} FLOPS

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

The New York Times

SECTIONS

TECHNOLOGY

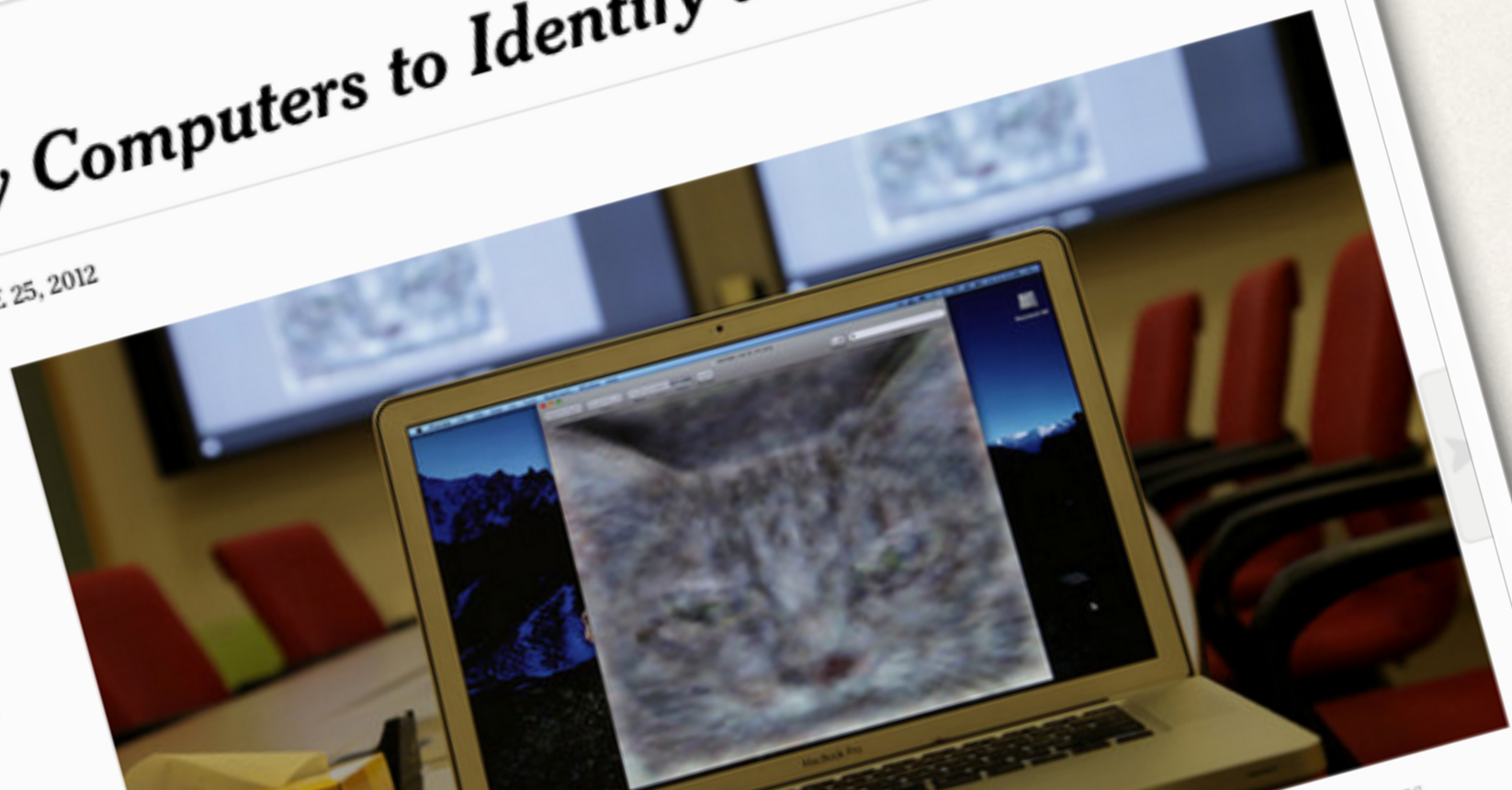
How Many Computers to Identify a Cat? 16,000

By JOHN MARKOFF JUNE 25, 2012

Email

Share

Tweet



Machine Learning?

Some Applications w/ Big Data

Classification & Regression

Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

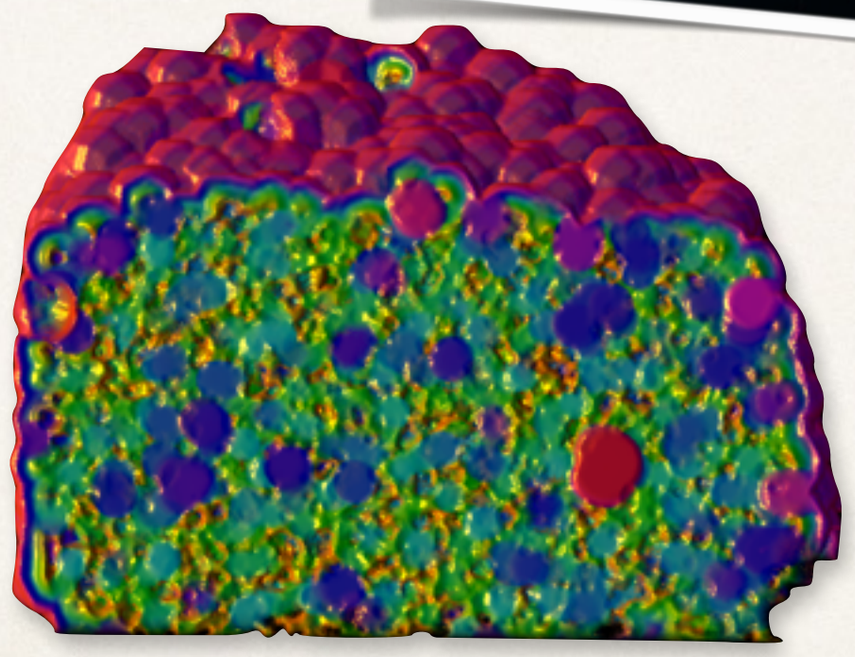


Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

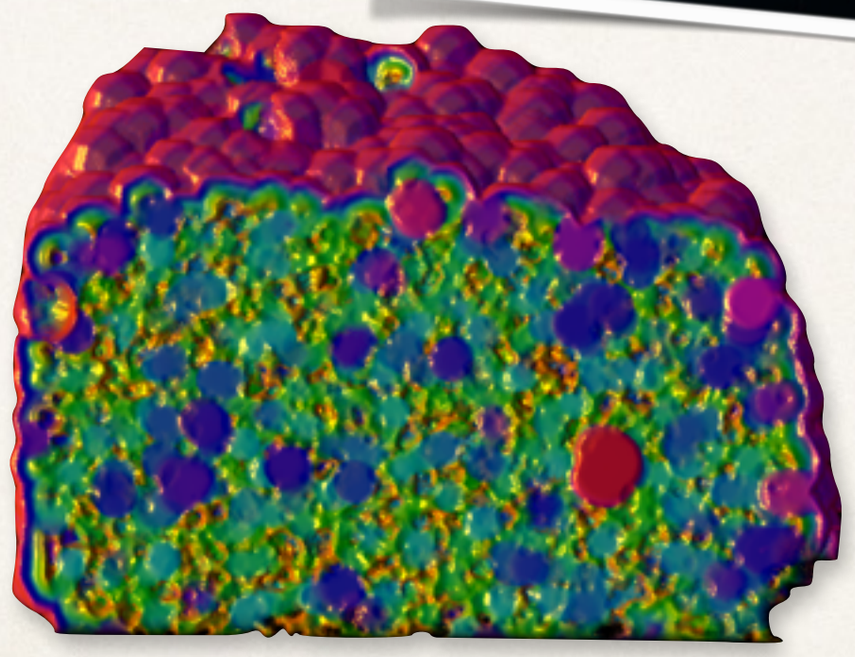


Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

[how-old.net](how-old.net)

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

Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

[how-old.net](how-old.net)

Text Data

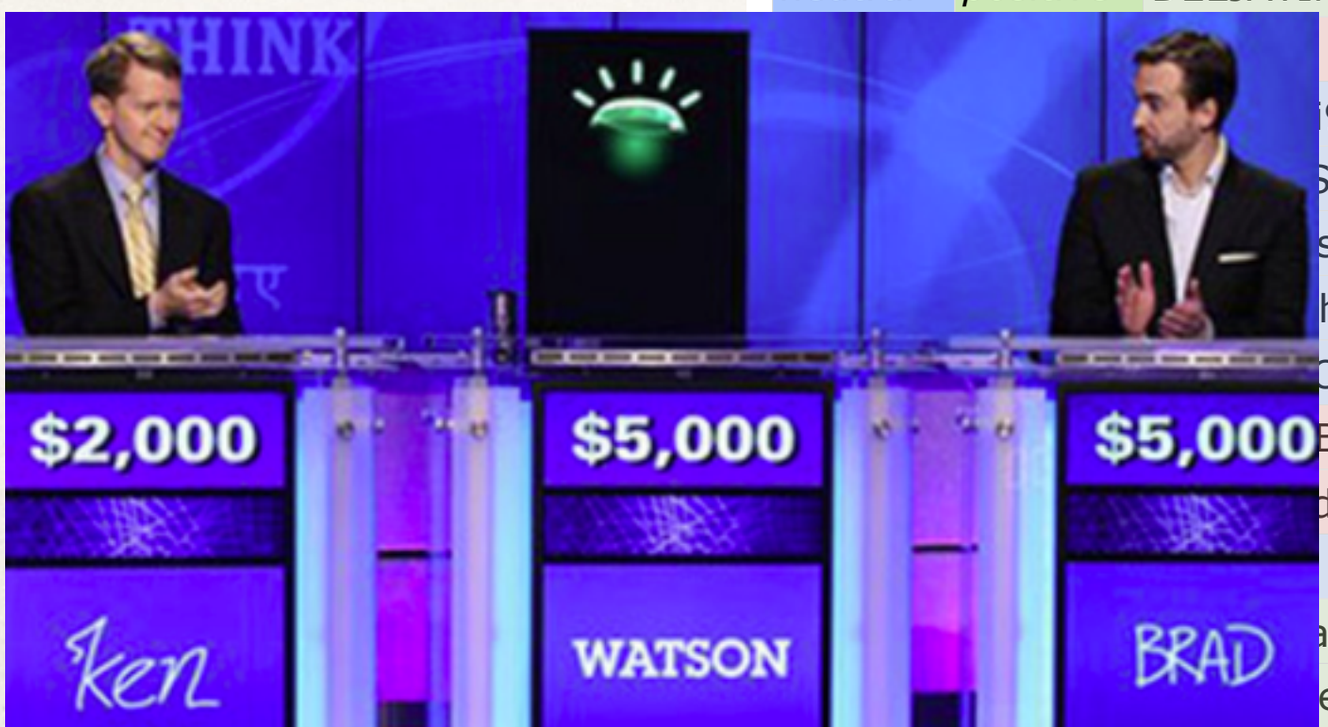
- Spam Detection
- User Content
- Medical Data

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 fes
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'm

Text Data

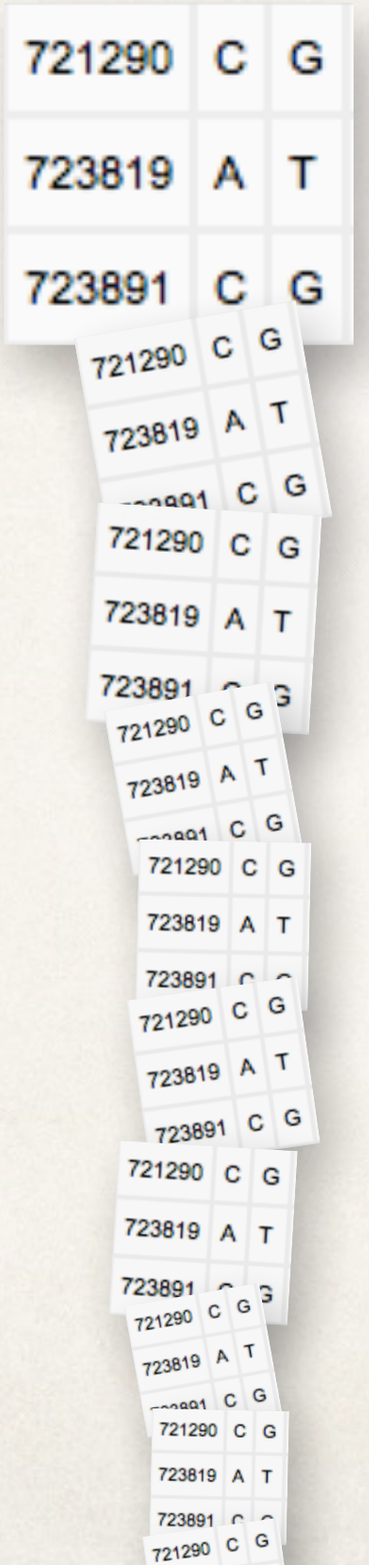
- Spam Detection
- User Content
- Medical Data

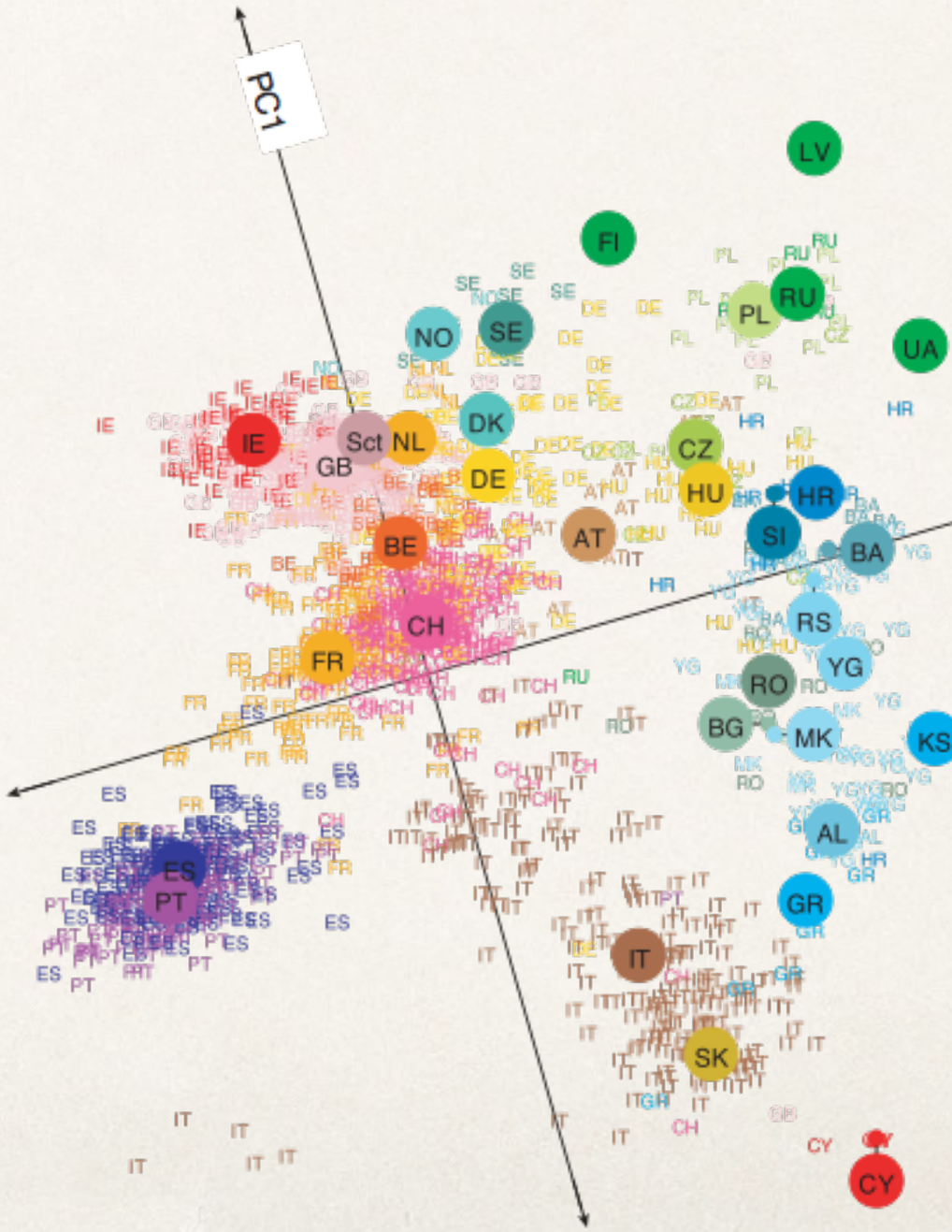
negative	*neutral*	But i wanna wear my Concords tomorrow though but i don't
positive	*neutral*	Gonna watch Grey's Anatomy all day today and tomorrow(:
negative	*neutral*	@CoachVac heey do you know anything about UVA's fallll fe
neutral	*neutral*	@DustyEf when that sun is high in that Texas sky, I'll be bu
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

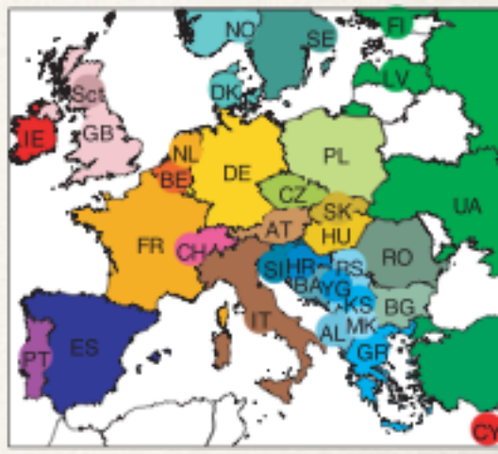


Santorum signing that was scheduled for tomorrow
9 lol yep looks like it! Was after El Clasico on Sunda
Stoke on Trent for the 2nd time today!
s Varsity Basketball Game tomorrow at 6:00 pm Th
htweights @Young__Assassin VS @jamievarner set
O_WEEEE slide thru sometime this weekend ill have
B618 Sure absolutely-- 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'm

Medical: Genetic Data



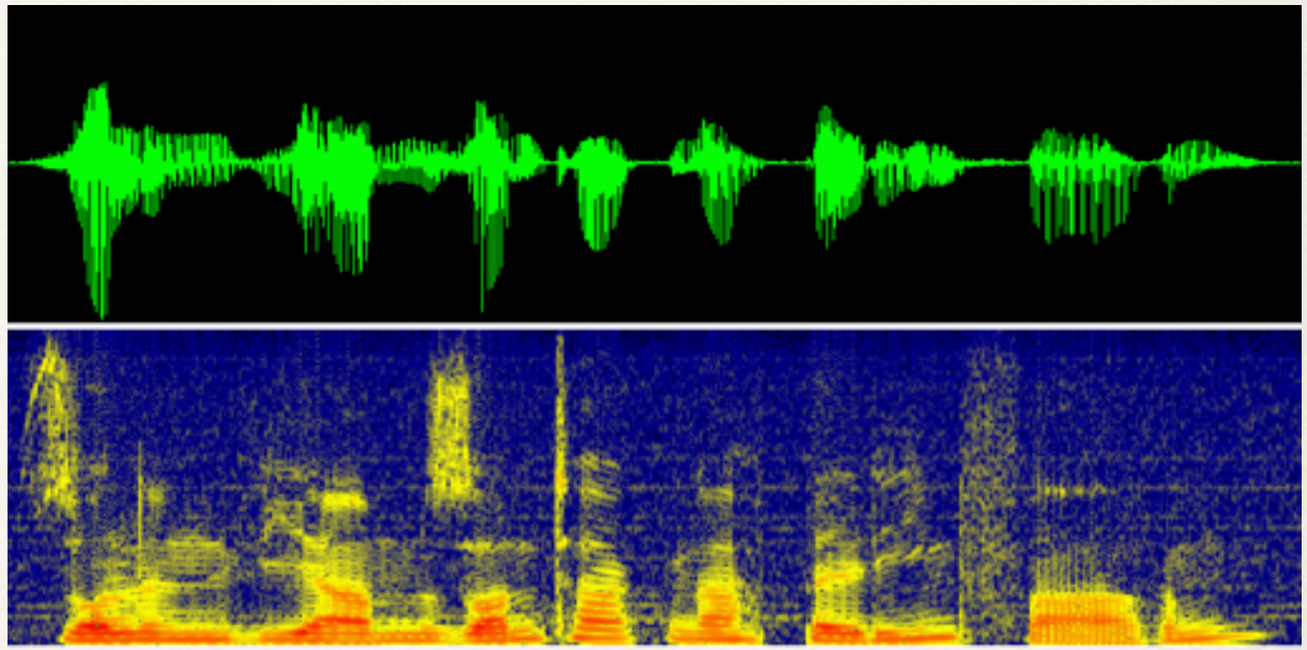




nature paper

Audio Data

- Hearing aids
- Voice Recognition
- Automatic Translation



Numerical / Sensor Data

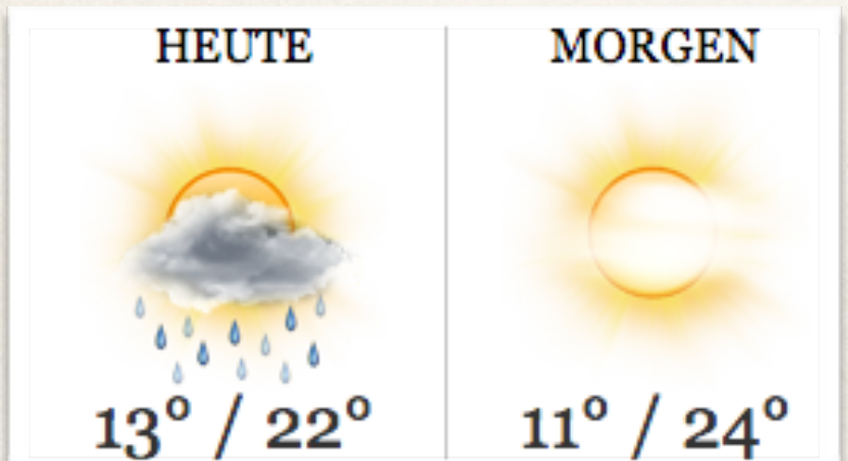
- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics

Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics

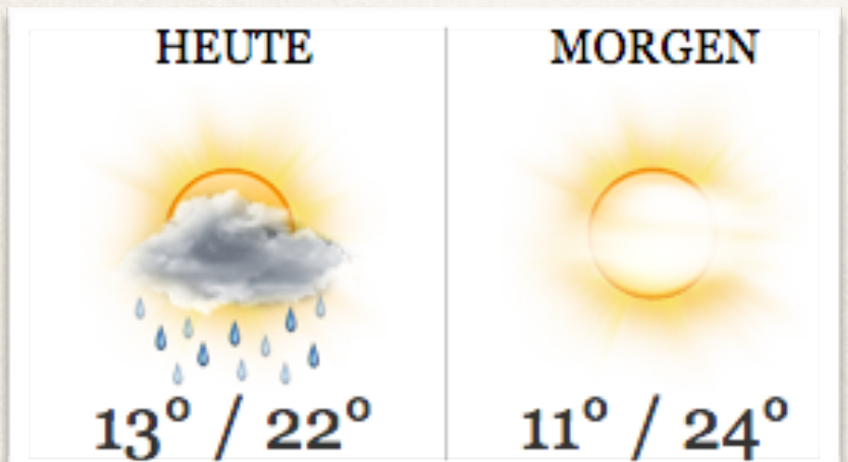
Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics



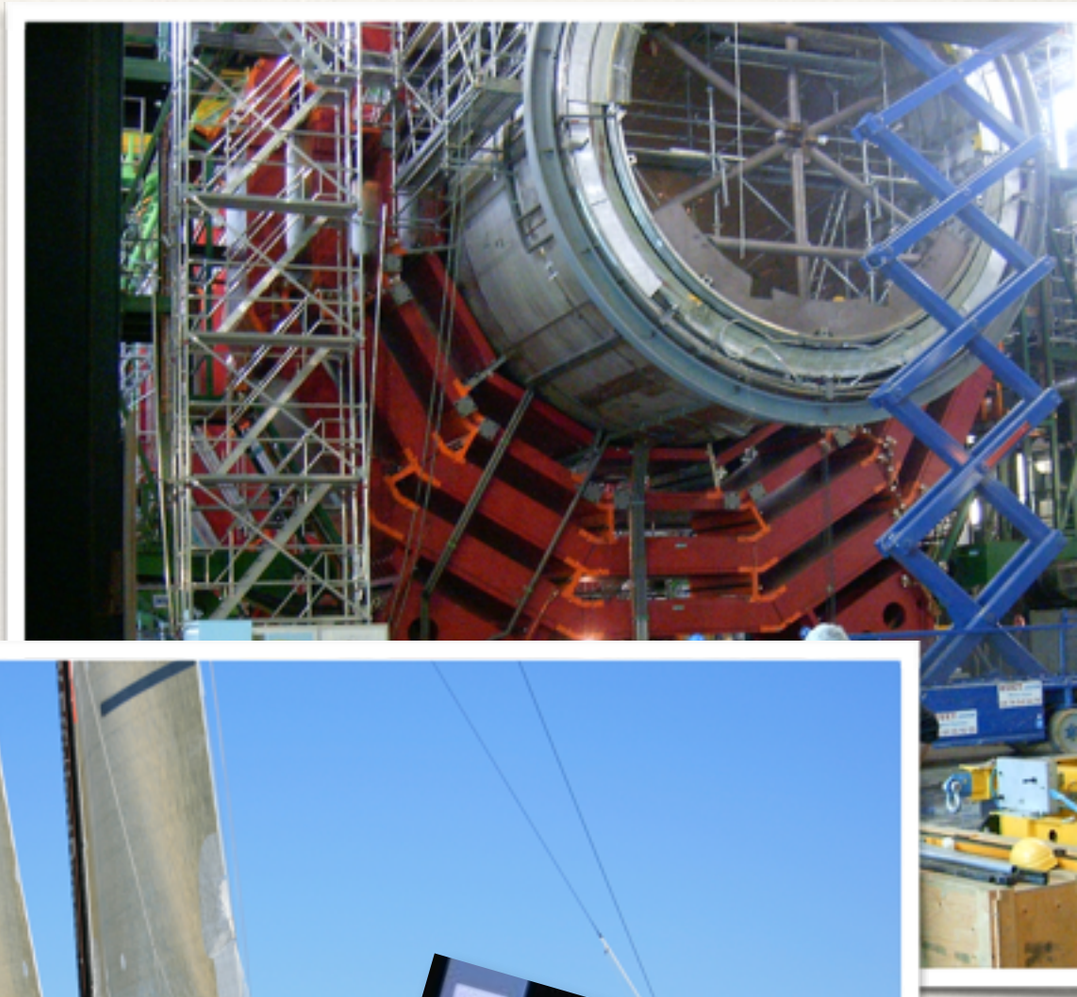
Numerical / Sensor Data

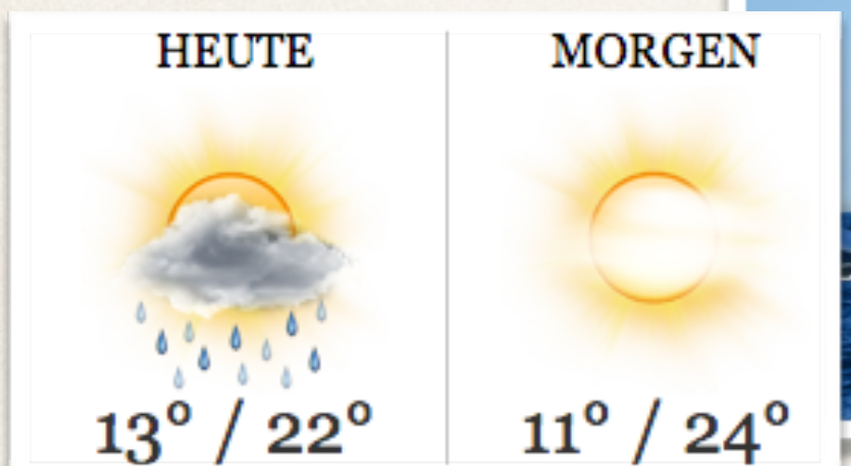
- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics



Numerical / Sensor Data

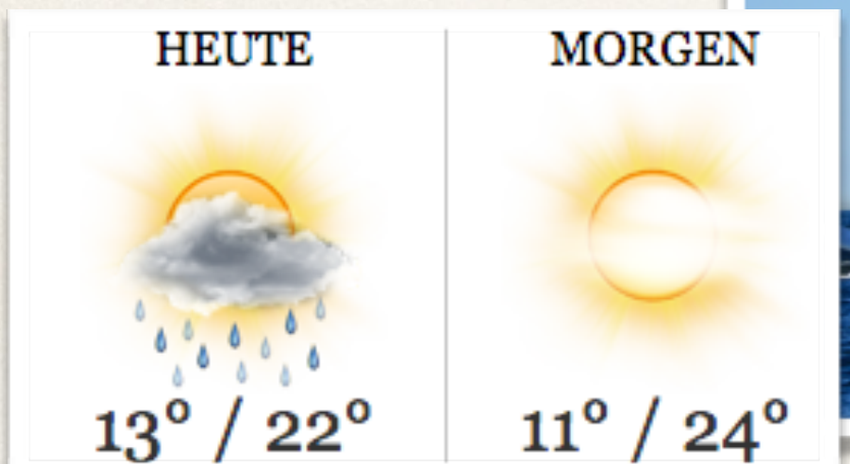
- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics





Numerical / Sensor Data

- Cern (Higgs Particle)
- Fitness Trackers
- Weather Forecast
- Robotics



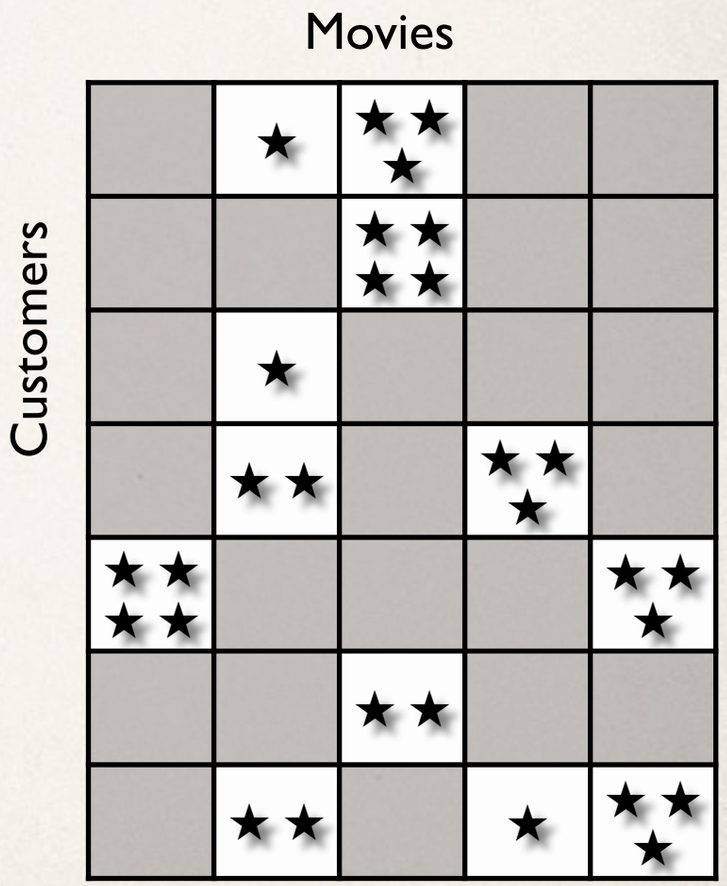
Internet Data

- Advertizing
- Recommender Systems

Internet Data

- Advertizing
- Recommender Systems

NETFLIX

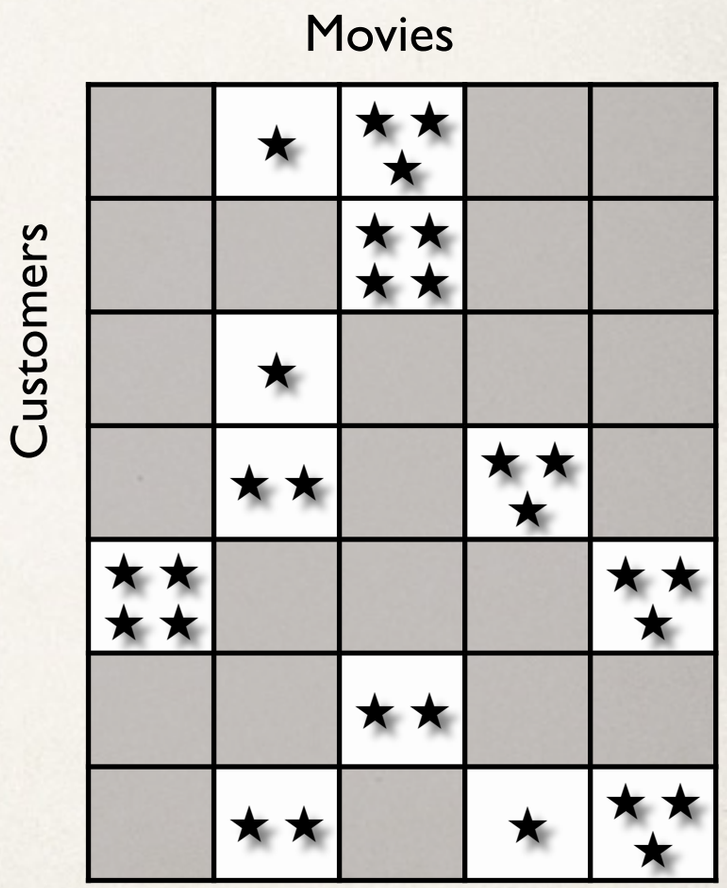


Internet Data

- Advertizing
- Recommender Systems

NETFLIX

amazon.com



Insurance & Finance

- Business-Analytics
- Targeted Advertizing
- Fraud
- Risk
- Customer Relations
- Marketing

Your turn!

Getting Started with Machine Learning

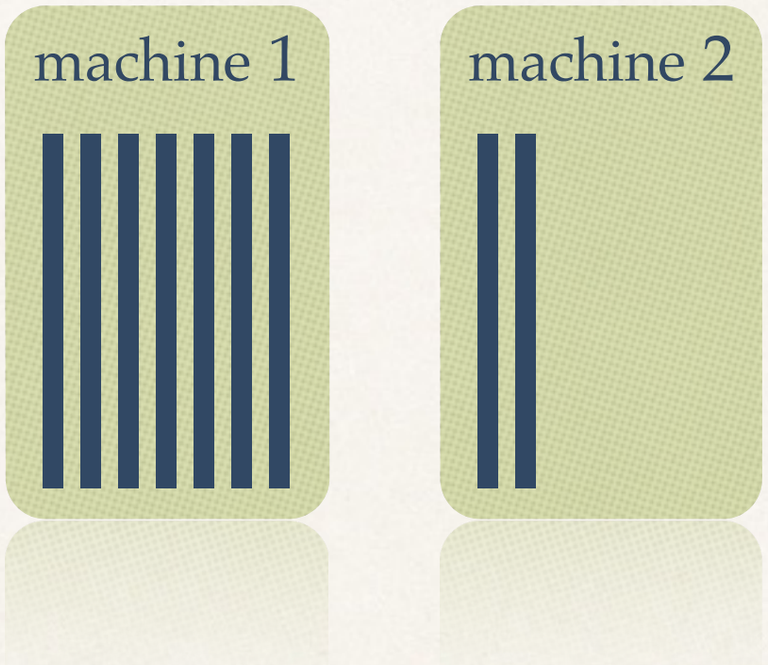
- scikit learn  *(python)*
- real data applications: kaggle.com

Distributed Machine Learning

What if the data does not fit onto one computer anymore?

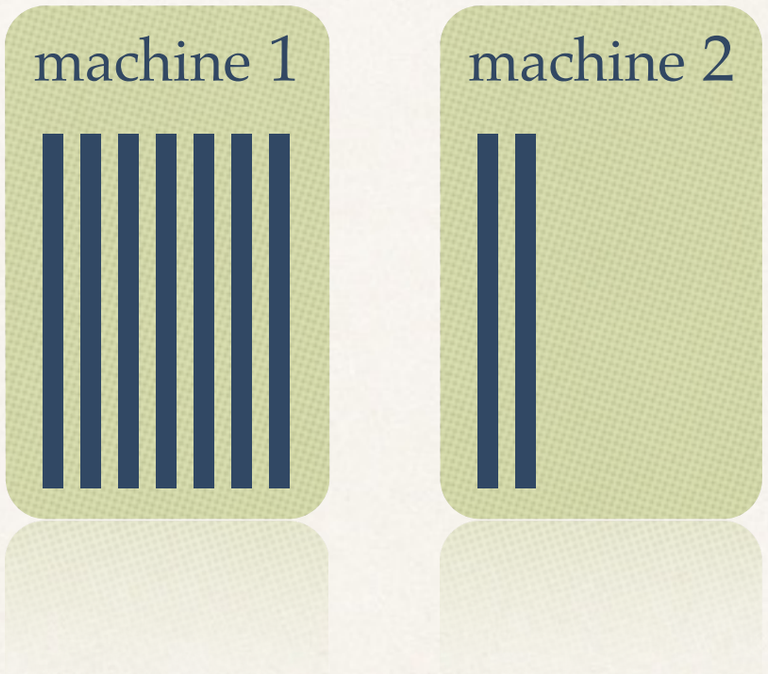
Distributed Machine Learning

What if the data does not fit onto one computer anymore?



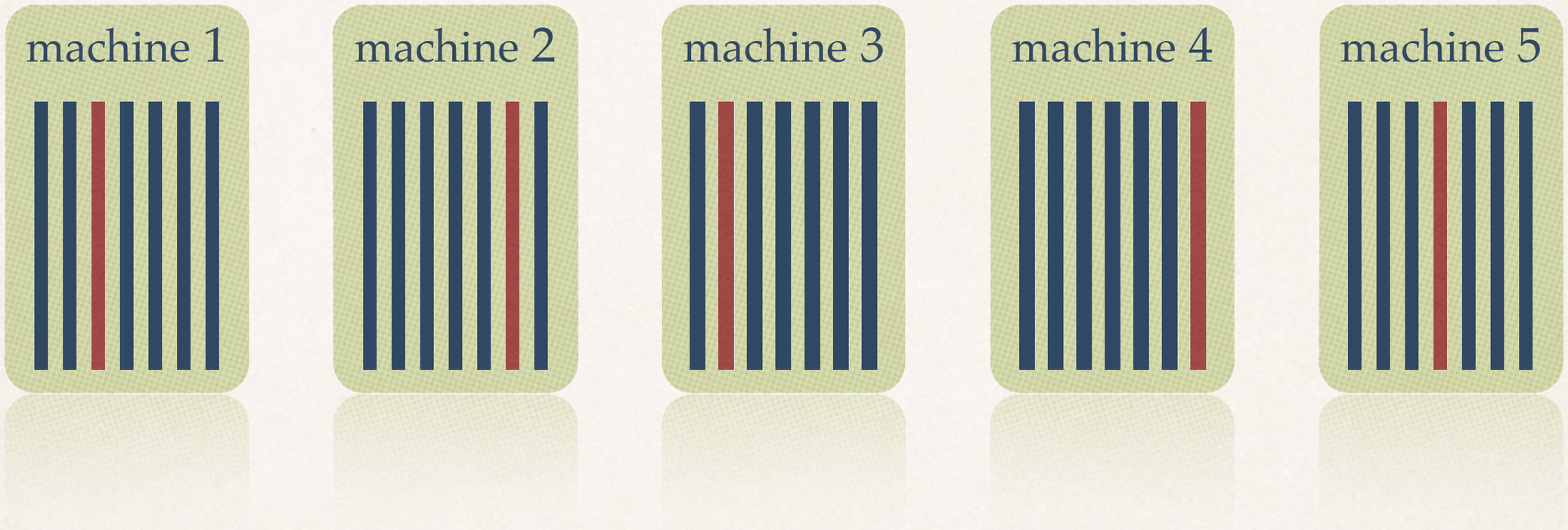
Distributed Machine Learning

What if the data does not fit onto one computer anymore?

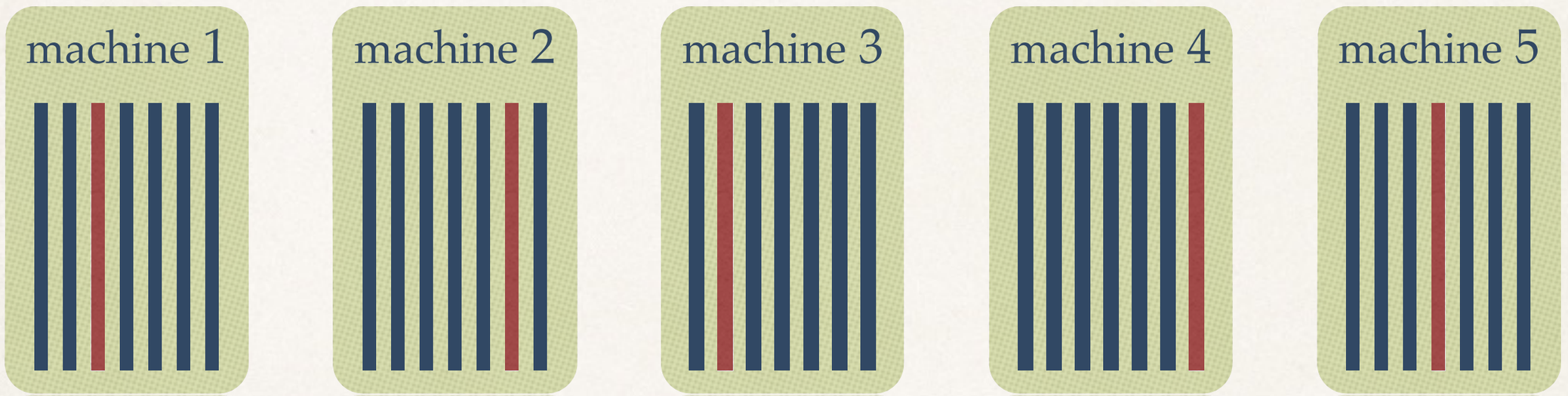


Does More Data Help?

Distributed Machine Learning



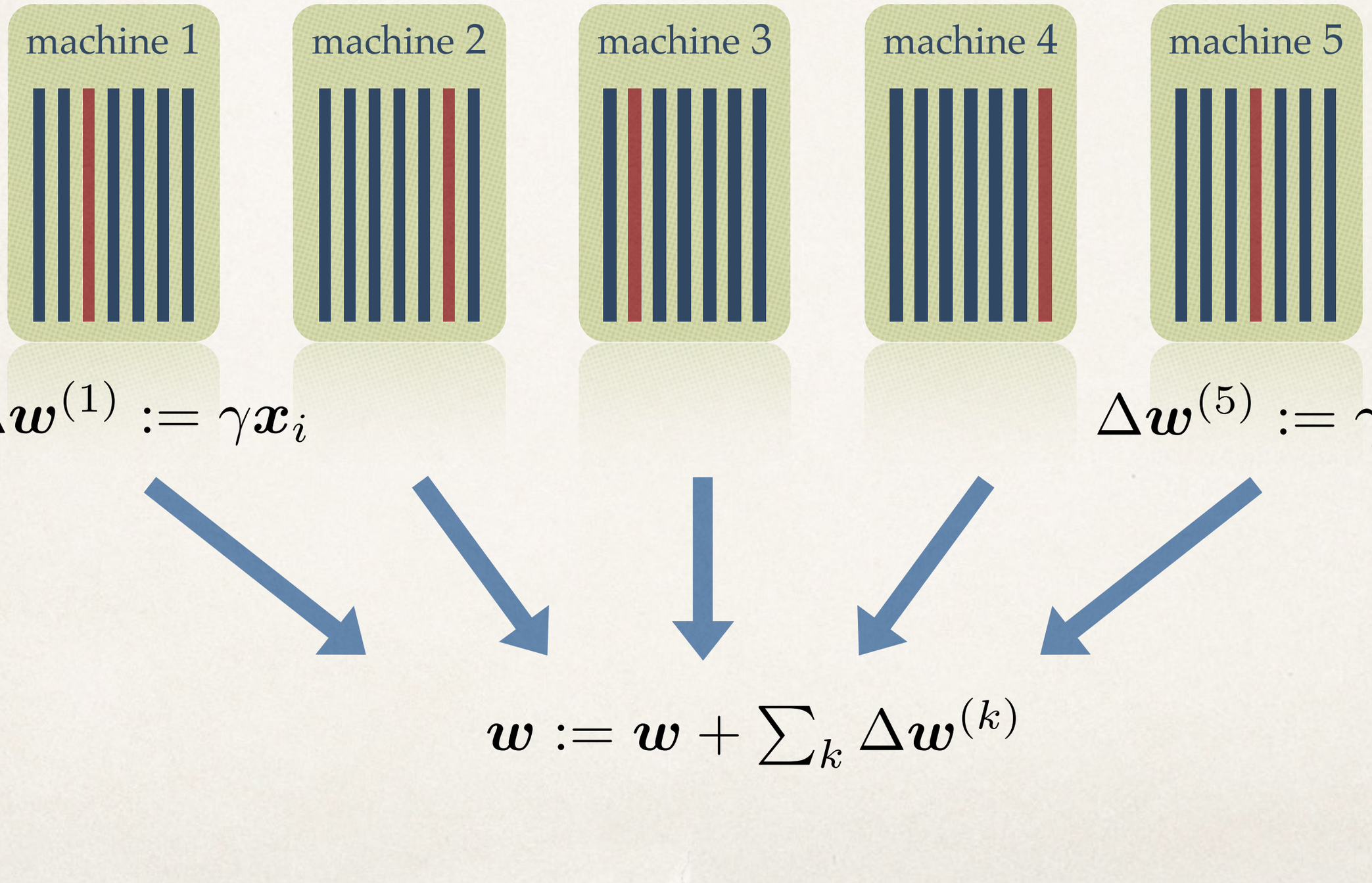
Distributed Machine Learning



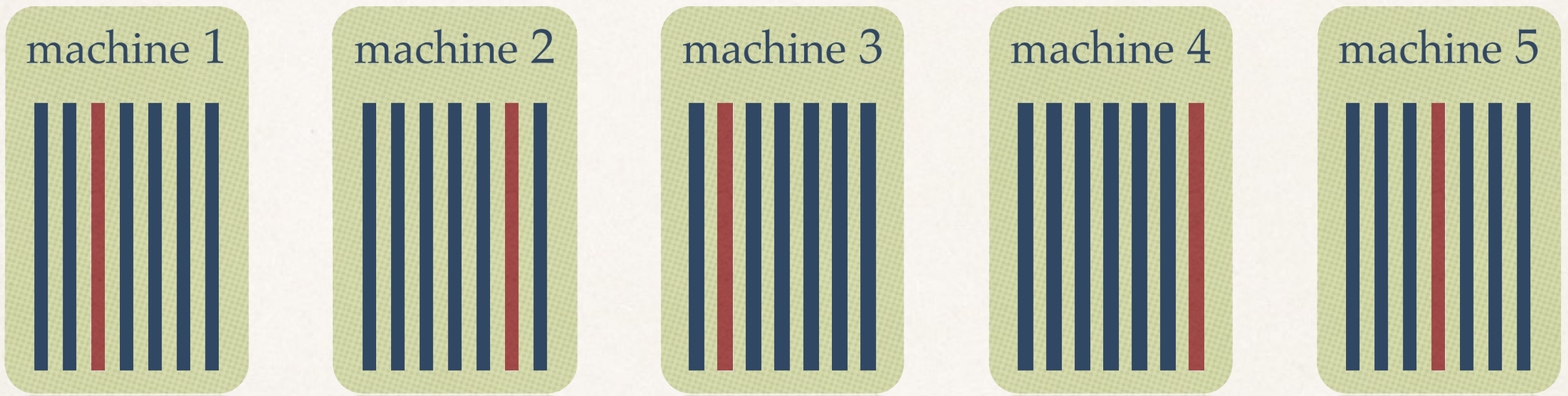
$\Delta \boldsymbol{w}^{(1)} := \gamma \boldsymbol{x}_i$

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

Distributed Machine Learning



Distributed Machine Learning



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

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

repeat many times

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

Problem 1

The Cost of Communication

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

- Reading v from Memory (RAM)

  100 ns

- Sending v to another Machine

  500'000 ns

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

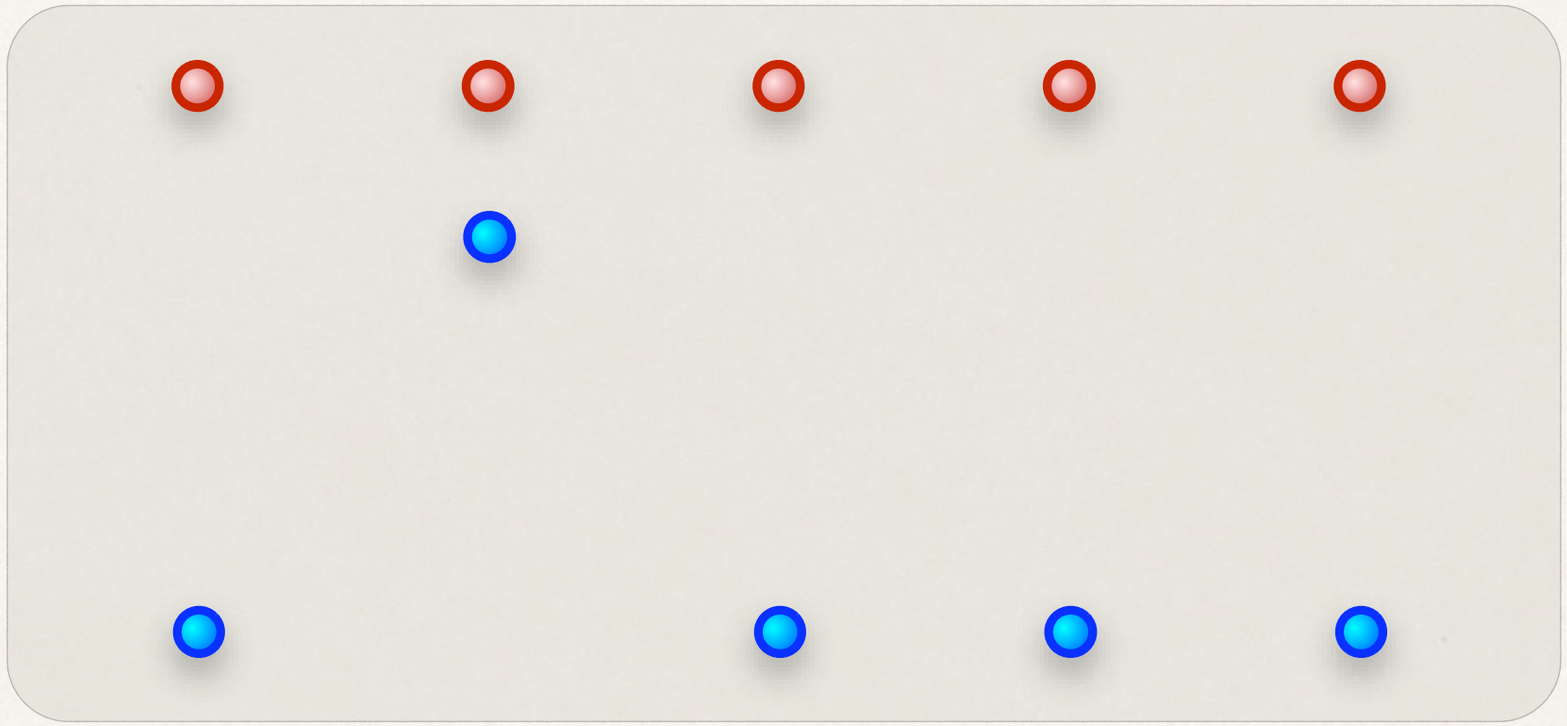
  10'000'000'000 ns

Problem 2

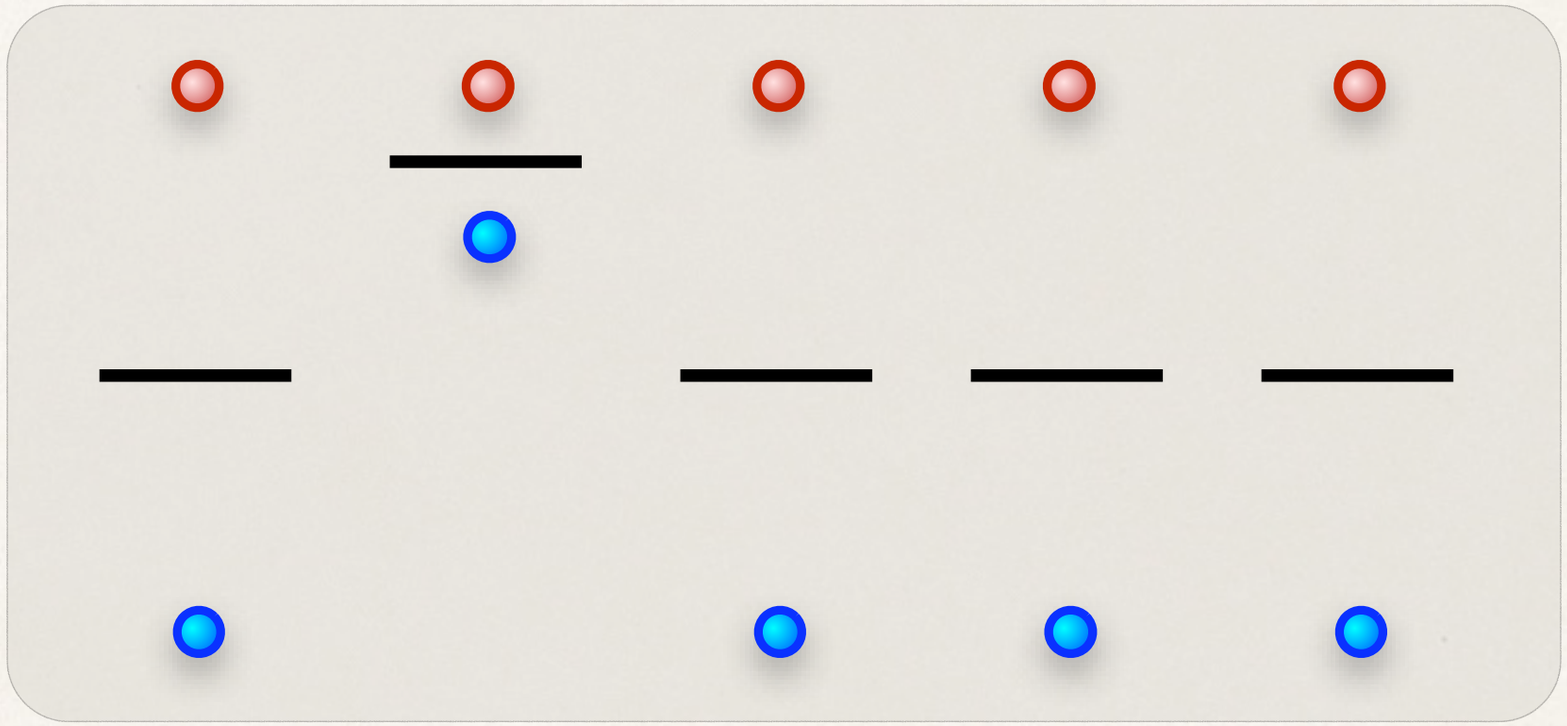
- Parallel Algorithms are Hard
- Single Machine Solvers are Fast

- no **reusability** of good single machine algorithms

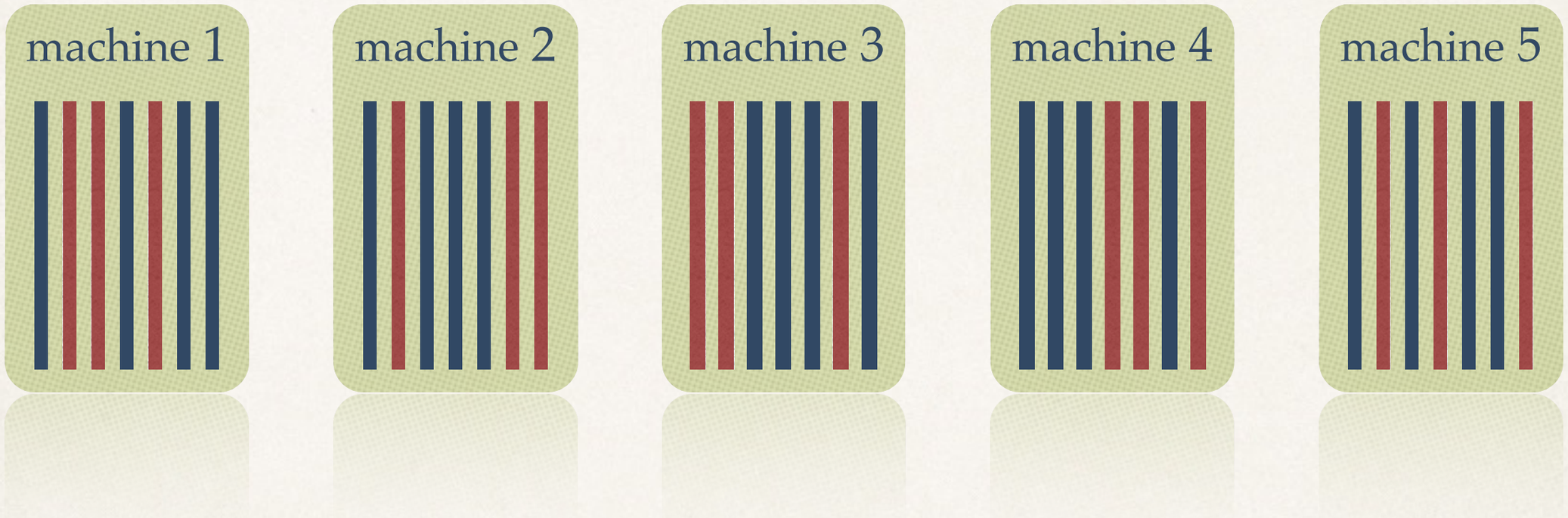
Model Averaging Does Not Work



Model Averaging Does Not Work

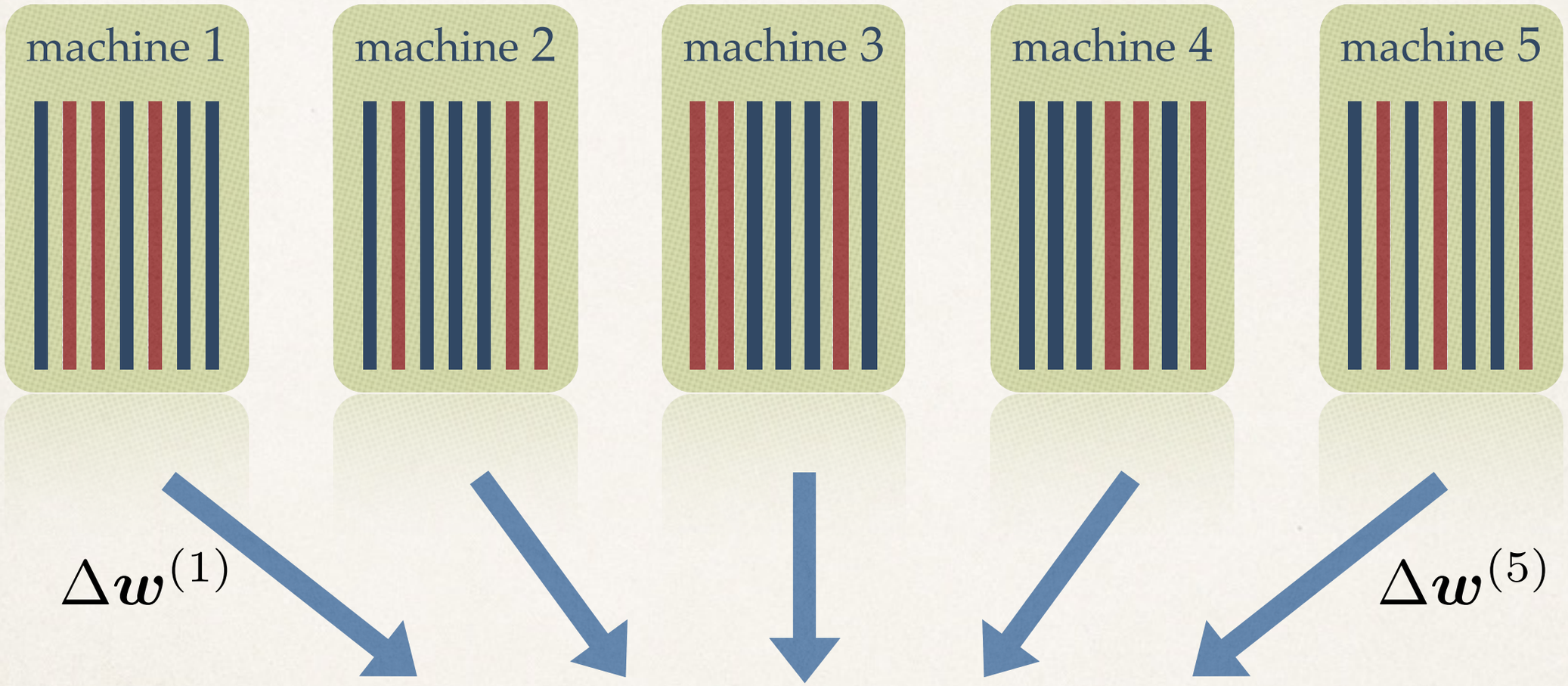


Communication Efficient Distributed Dual Coordinate Ascent



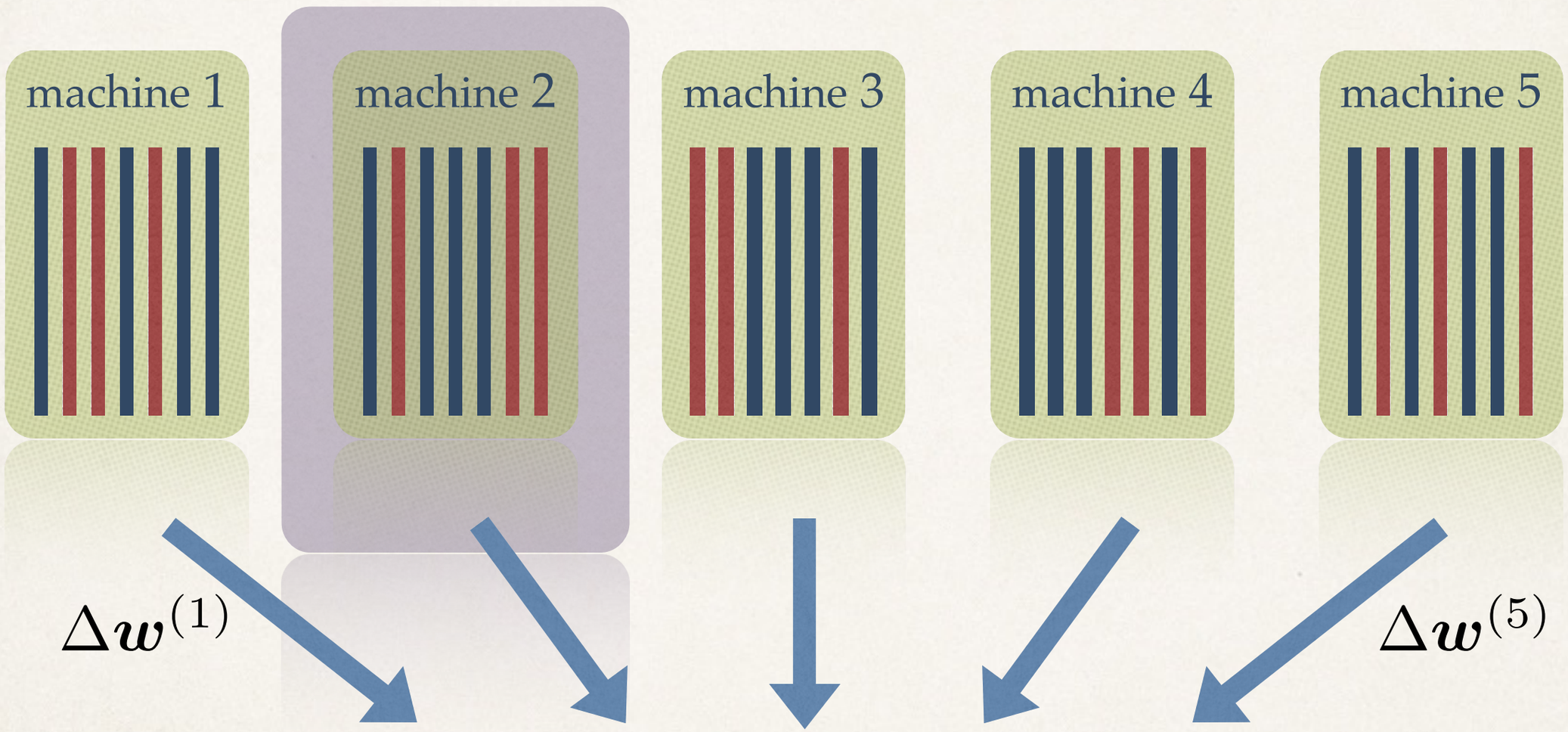
CoCoA

Communication Efficient Distributed Dual Coordinate Ascent



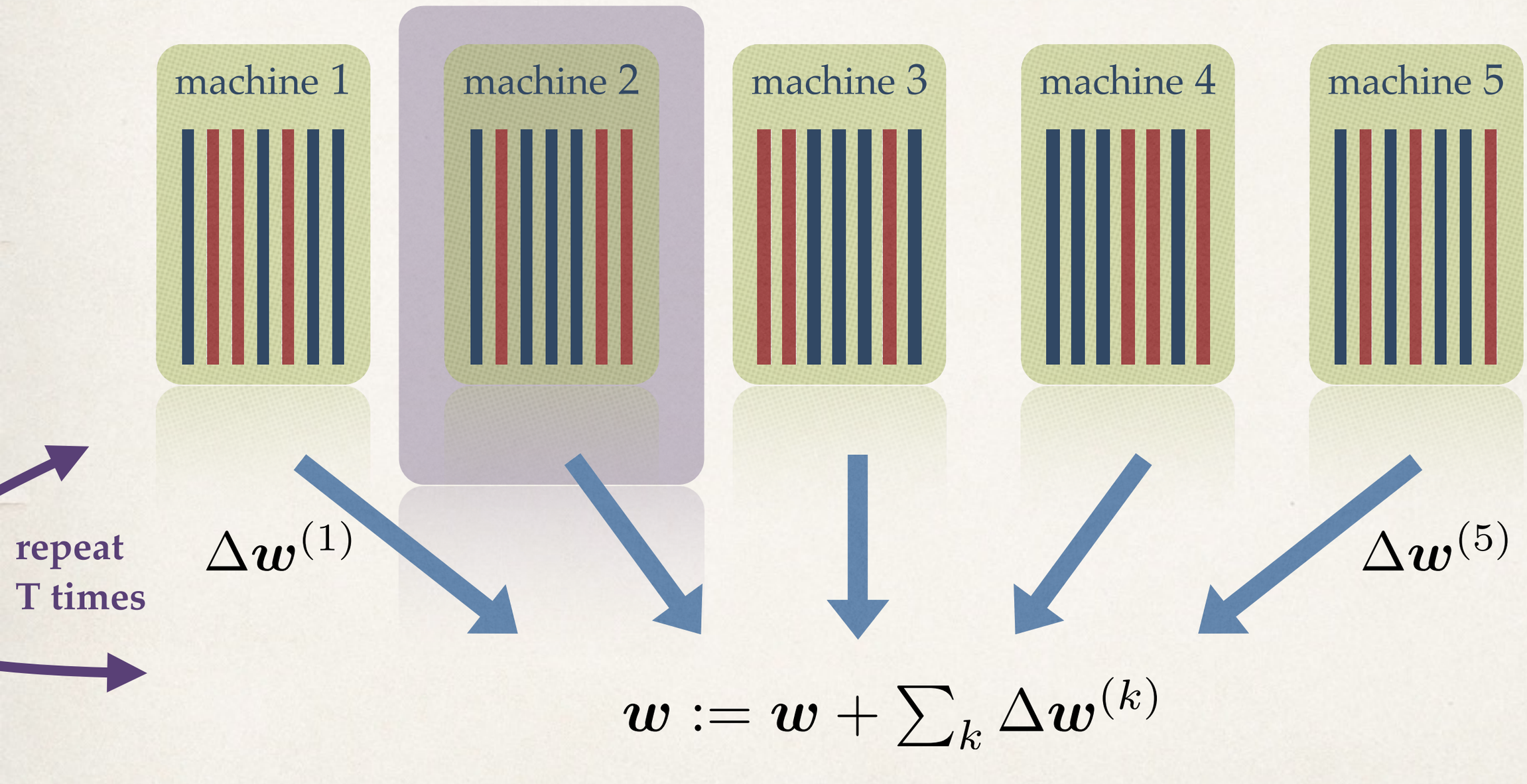
CoCoA

Communication Efficient Distributed Dual Coordinate Ascent



CoCoA

Communication Efficient Distributed Dual Coordinate Ascent



CoCoA

“Big Data Analytics” Applications

Classification

Support Vector Machine *(SVM)*

Logistic Regression

Structured Prediction

Regression

Ridge Regression

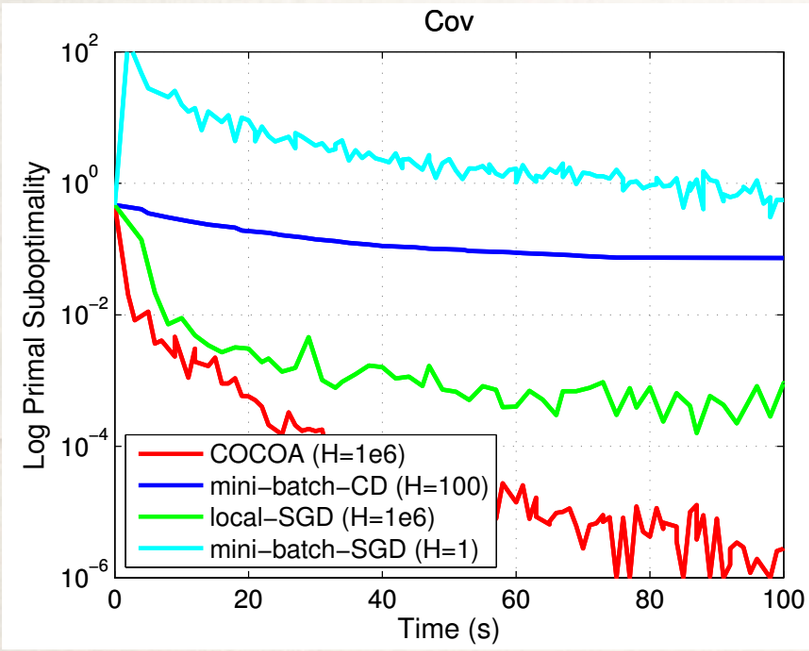
Sparse Least Squares variants

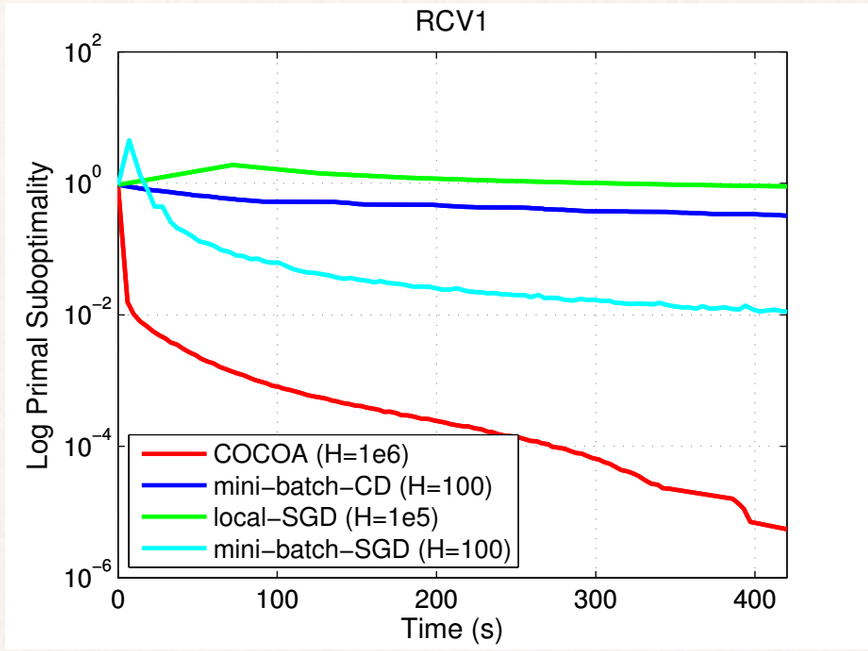
Lasso, Elastic-Net *(Feature Selection, Compressed Sensing)*

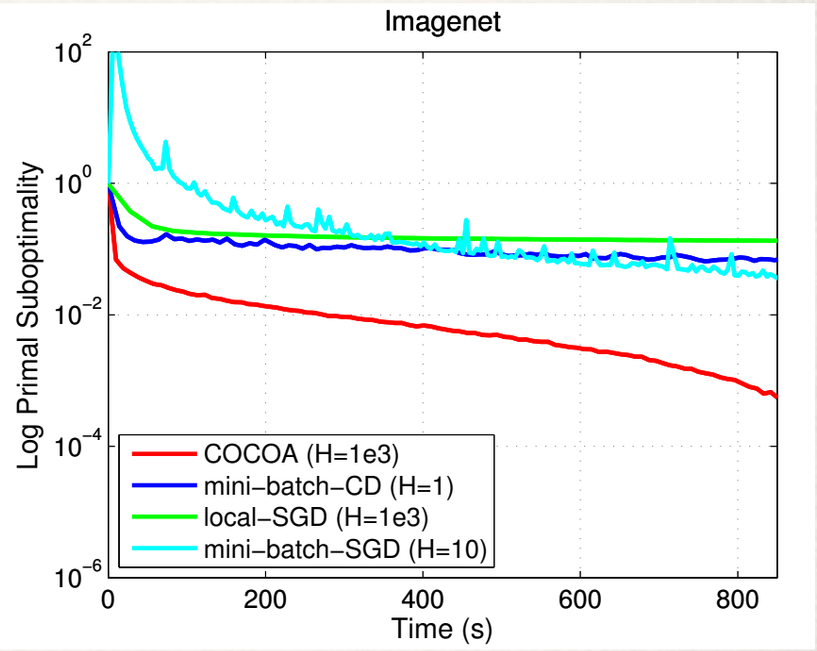
$$\min_{w \in \mathbb{R}^d} \quad \left[P(\boldsymbol{w}) := \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{w}^T \boldsymbol{x}_i) \right]$$

Experiments vs Spark MLlib

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







Thanks

Email: *m.jaggi@gmail.com*

Web: *da.inf.ethz.ch*

spinningbytes.com

all our 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

ZüriML Meetup