Distributed Machine Learning and Text Analysis

Martin Jaggi
EPFL
mlo.epfl.ch

31th Jan 2017

Machine Learning Methods to Analyze Large-Scale Data

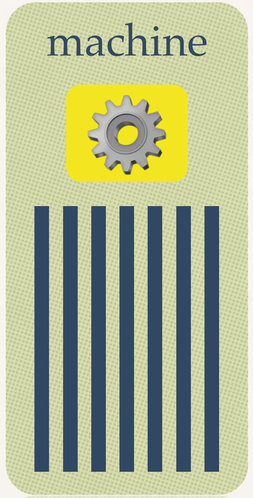
Machine Learning

Optimization

Systems

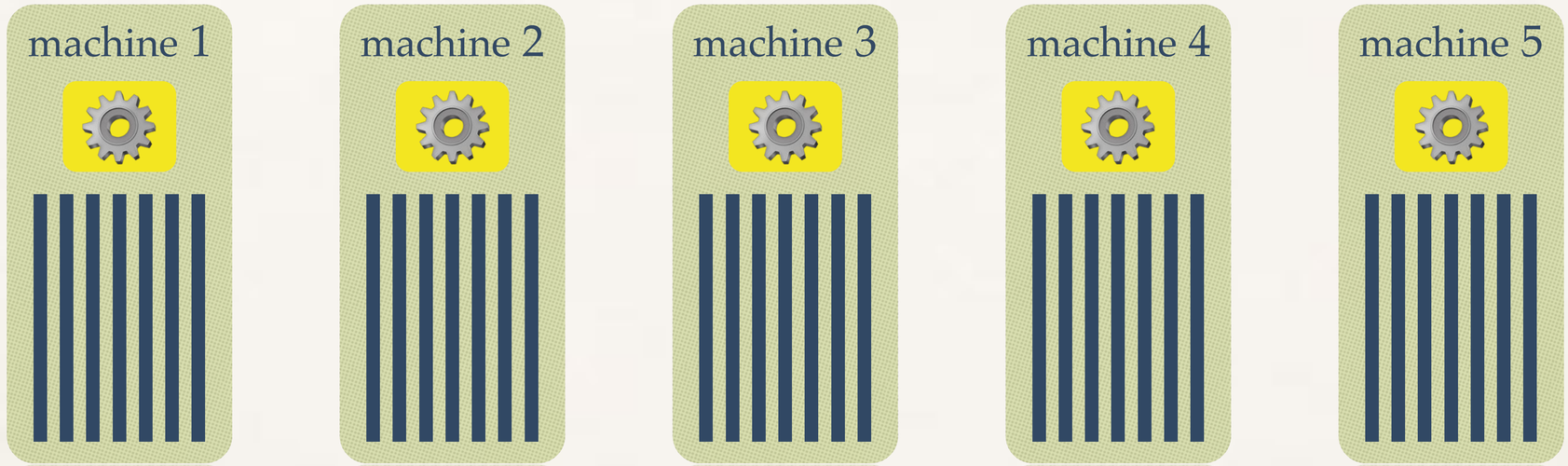
Applications

Machine Learning Systems

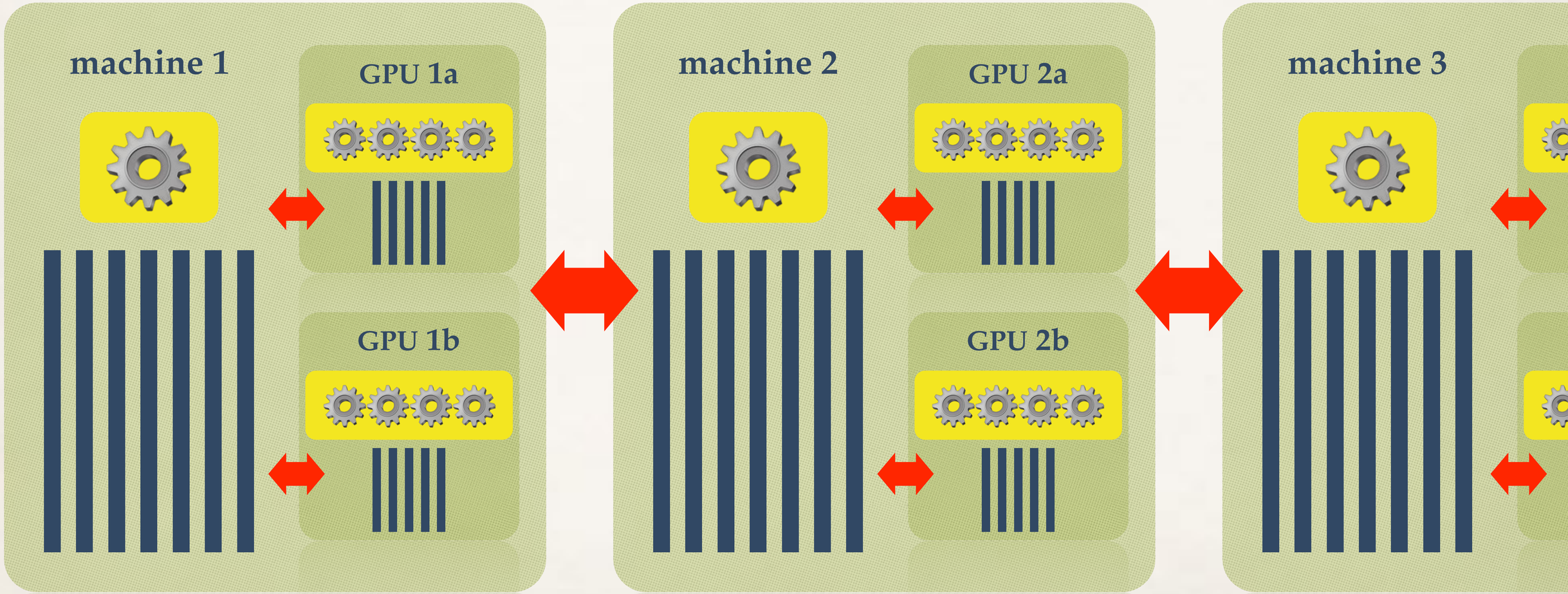


Machine Learning Systems

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



Machine Learning Systems



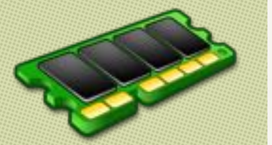
Challenge 1

The Cost of Communication

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

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

  100 ns



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

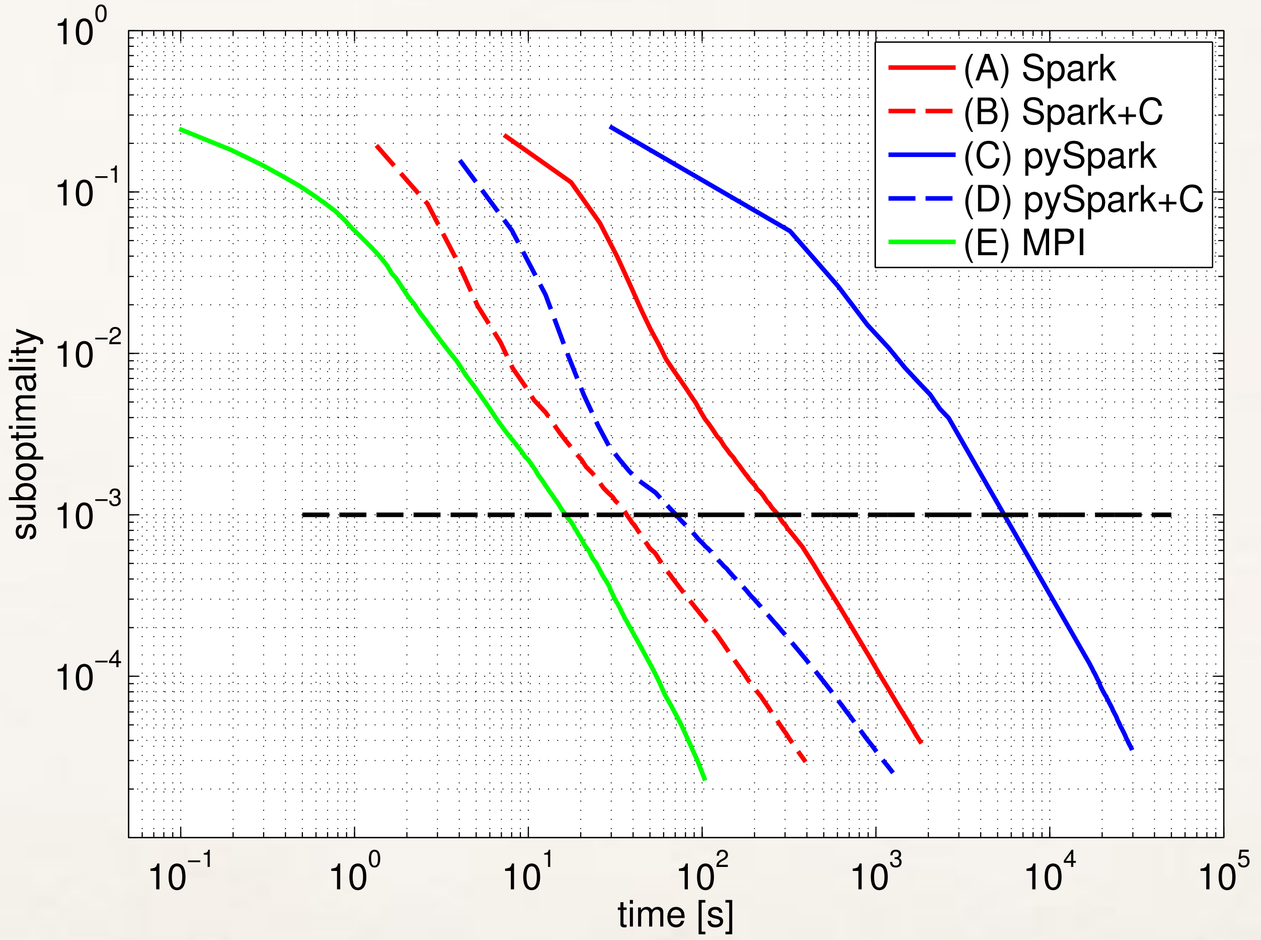
  500'000 ns

- Typical Map-Reduce iteration

  10'000'000'000 ns

Challenge 1

The Cost of Communication



High-Performance Distributed Machine Learning using Apache SPARK
Dünner et al. 2016, arxiv.org/abs/1612.01437

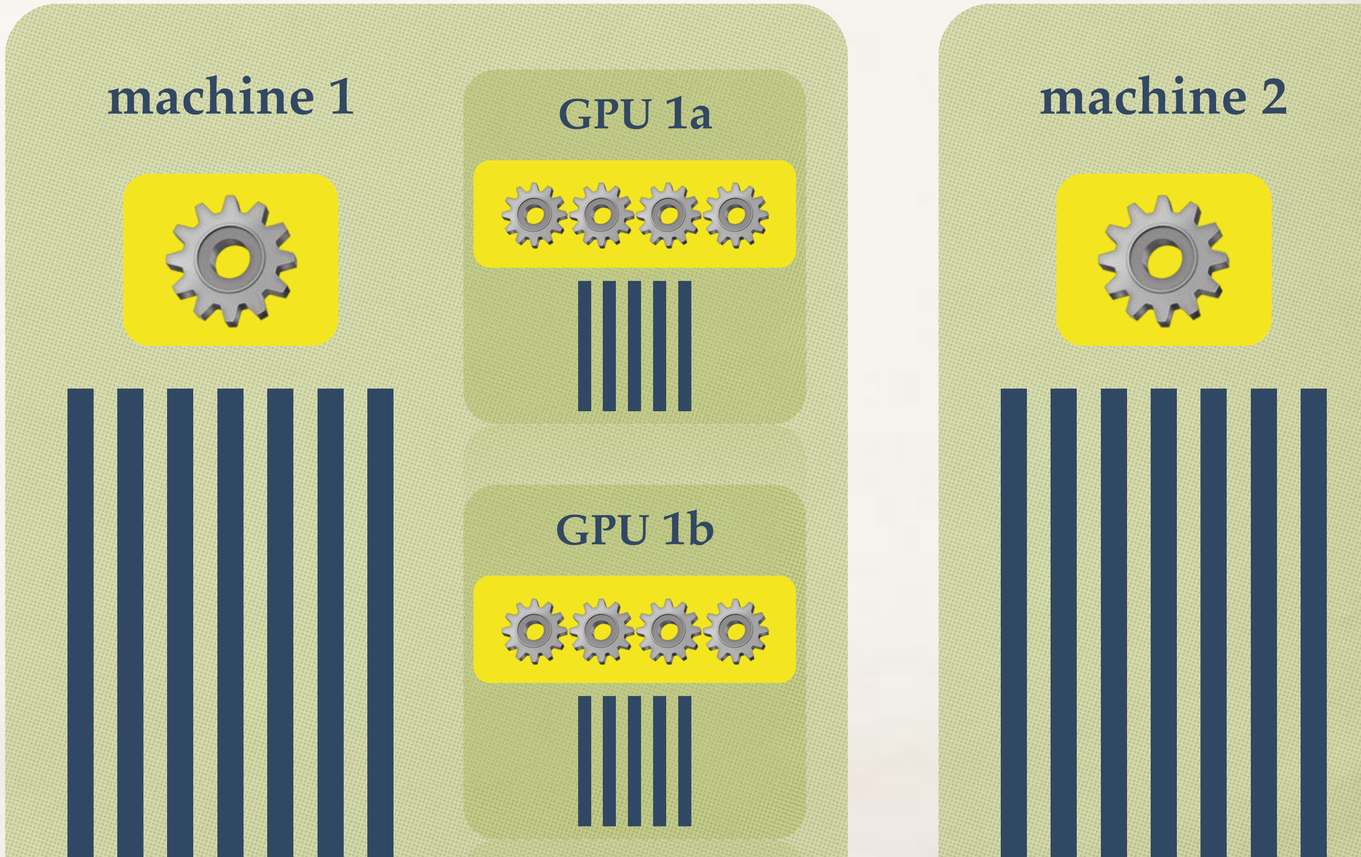
Challenge 2

Usability - Parallel Coding is Hard Single Machine Solvers are Fast

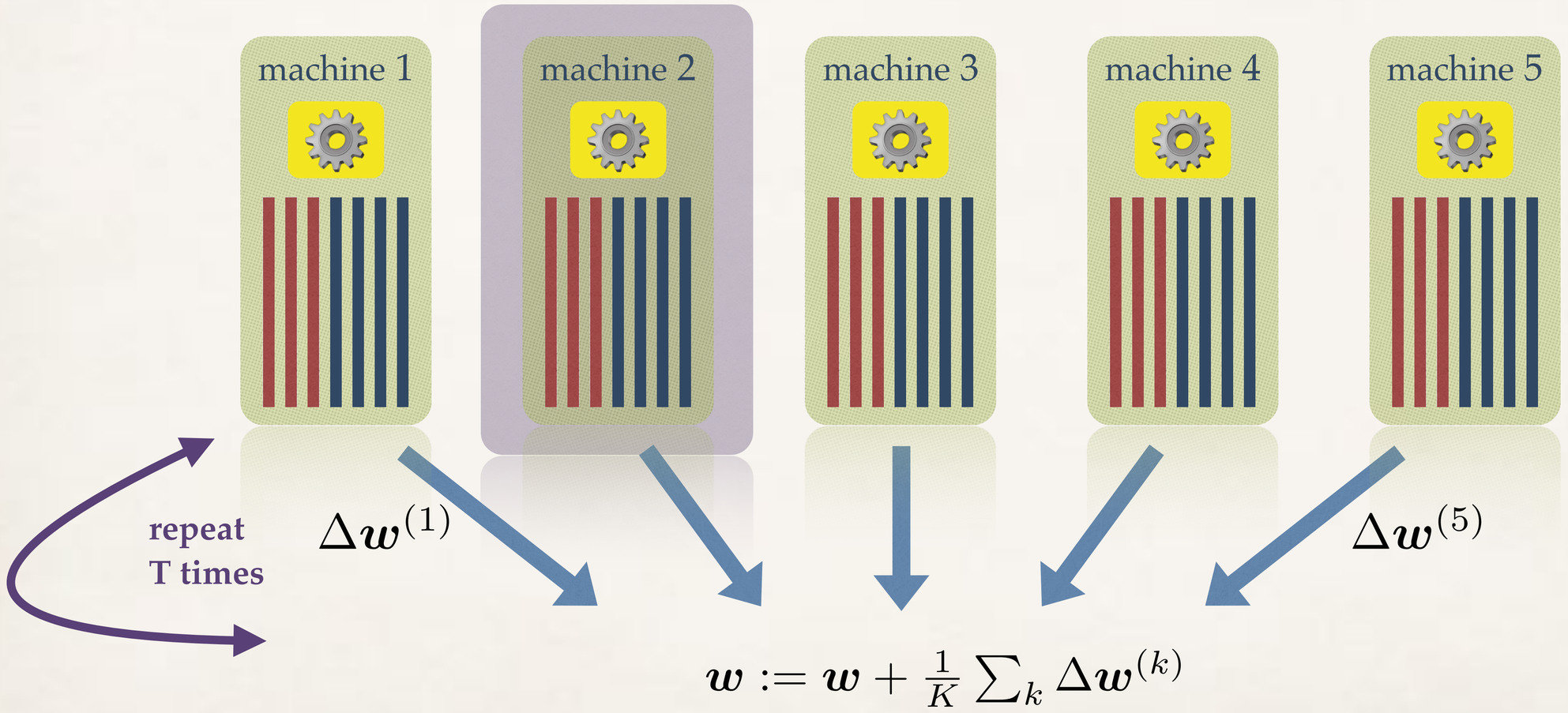
- no **reusability** of good single machine algorithms

Challenge 3

Data Locality - Which data in which memory?



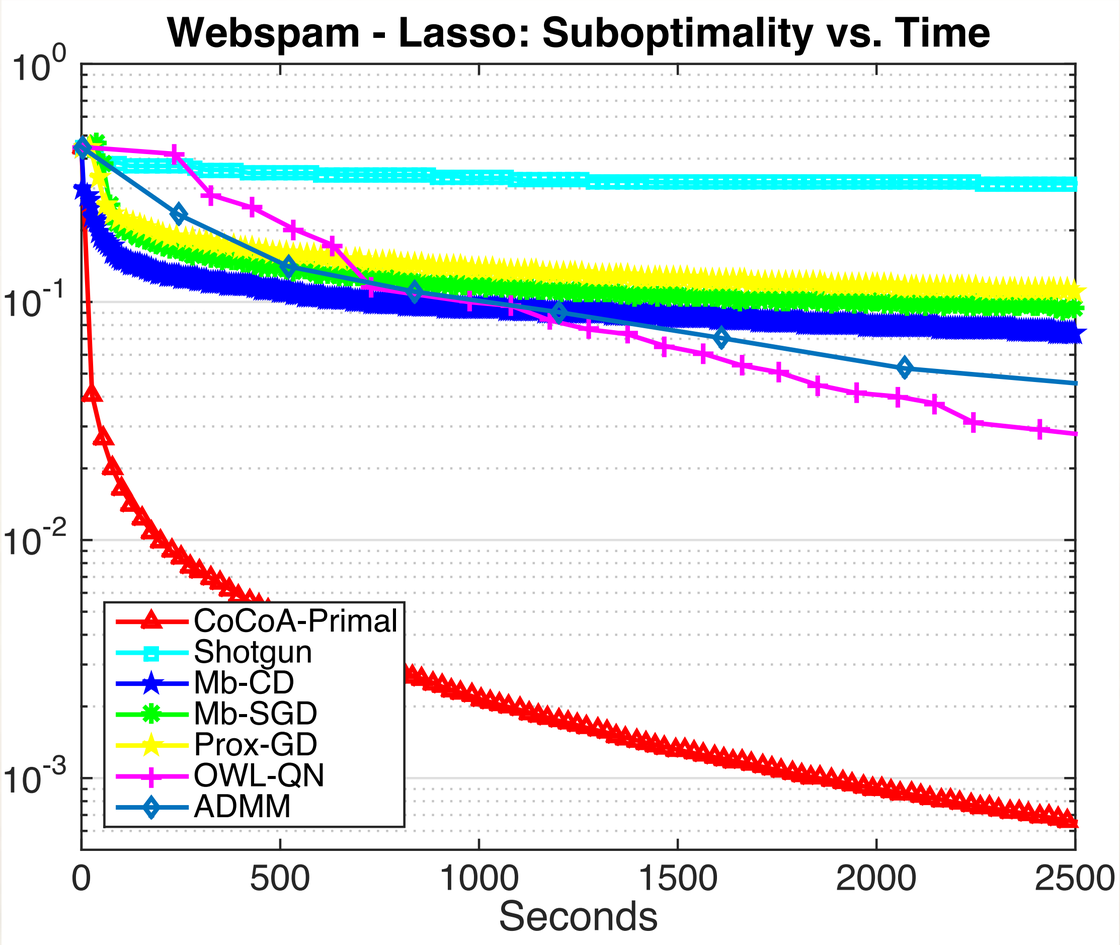
CoCoA - Communication Efficient Distributed Optimization



Experiments

Sparse Linear Regression

Dataset	Training	Features	Sparsity
url	2,396,130	3,231,961	3.5e-3%
epsilon	400,000	2,000	100%
kddb	19,264,097	29,890,095	9.8e-5%
webspam	350,000	16,609,143	0.02%



NIPS 2014, ICML 2015,
arxiv.org/abs/1611.02189

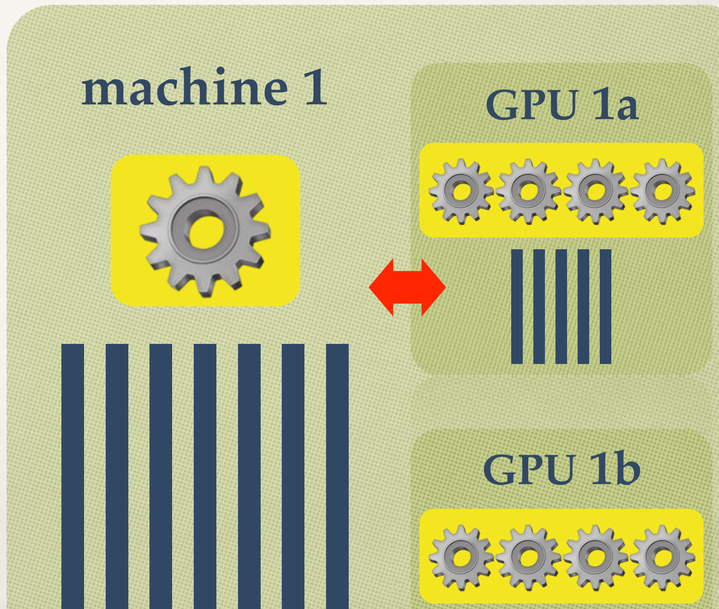
Spark Code:
github.com/gingsmith/proxcocoa

+ TensorFlow
+ Apache Flink

Summary

- improve **usability** of large-scale ML
- full **adaptivity** to the communication cost, **fault tolerance**
- **re-usability** of good single machine solvers
- **accuracy** certificates

AIStats 2017



Open Research

- *multi-level approach* on heterogenous systems
- training neural network models

Project:

Distributed Machine Learning Benchmark

Goal:

Public and Reproducible

Comparison of Distributed Solvers

github.com/mlbench/mlbench

Apache

Google

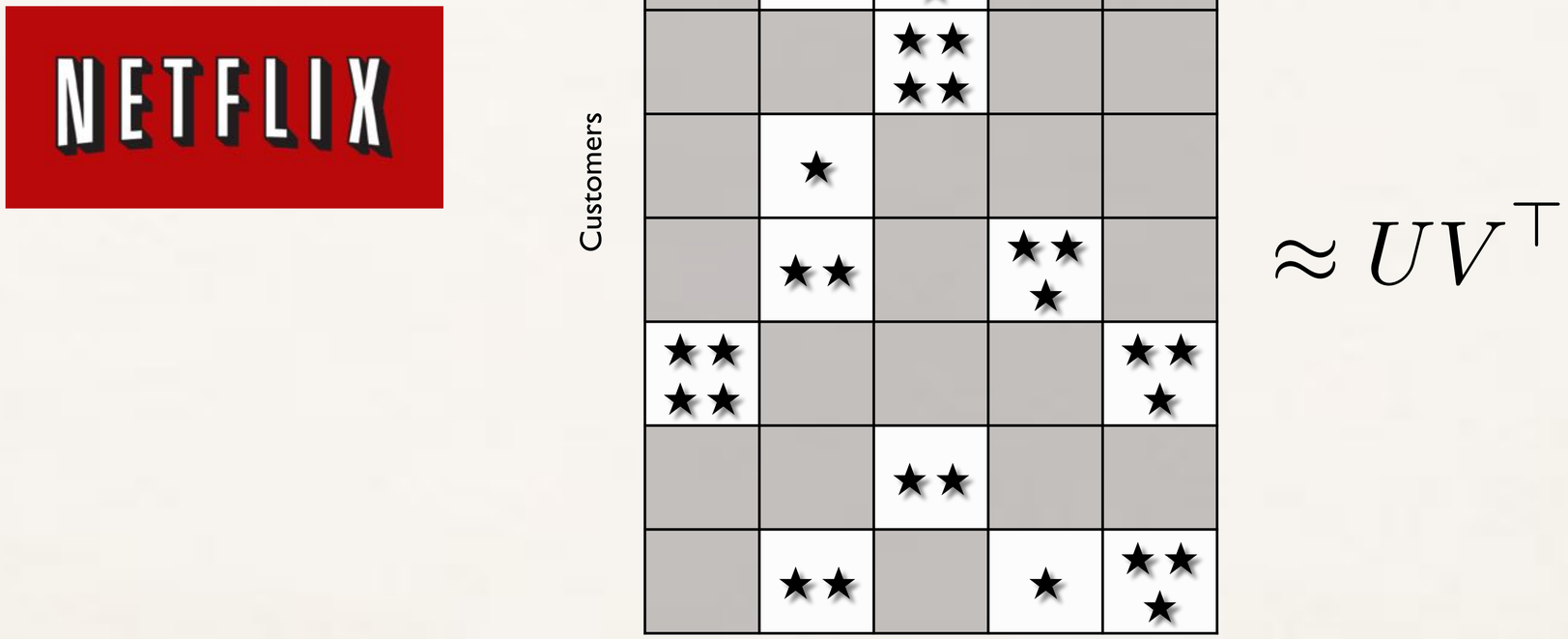
Apache

HPC

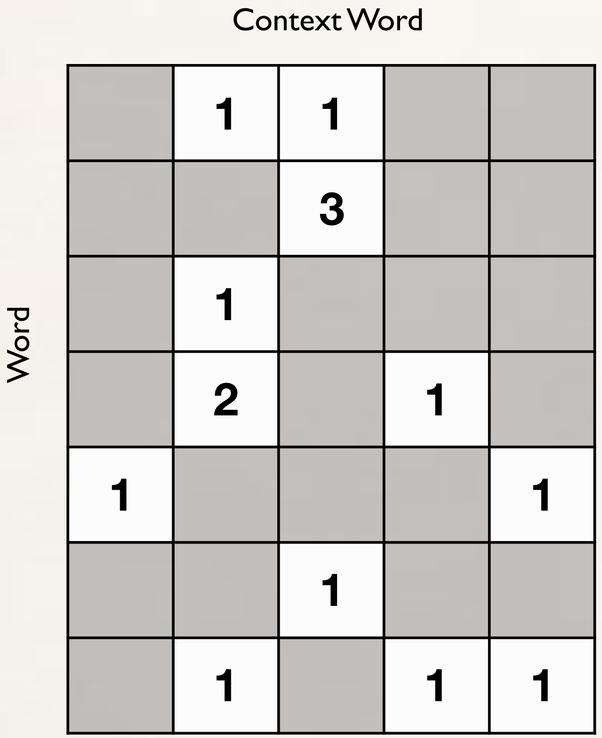
Matrix Factorizations

$$\min_{\boldsymbol{U},\boldsymbol{V}} f(\boldsymbol{U}\boldsymbol{V}^\top)$$

from Recommender Systems



to Word Representations



explain co-occurence *i,j*
by means of

$$\boldsymbol{v}_i^\top \boldsymbol{v}_j$$

Word Representations

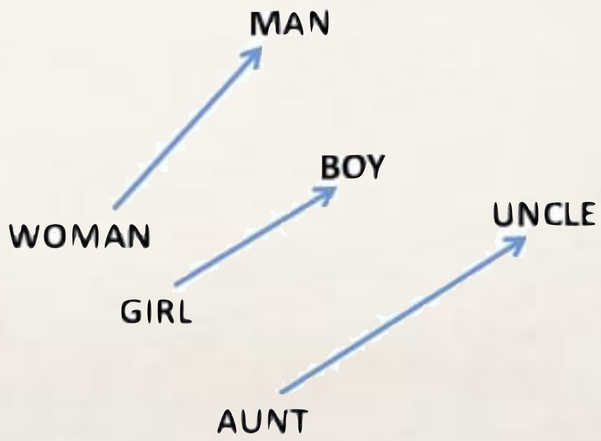
Context Word

Word

	1	1		
		3		
	1			
	2		1	
1				1
		1		
	1		1	1

$$\approx \quad \boldsymbol{V} \quad \boldsymbol{V}^\top$$

SVD, PLSA etc
word2vec, gloVe



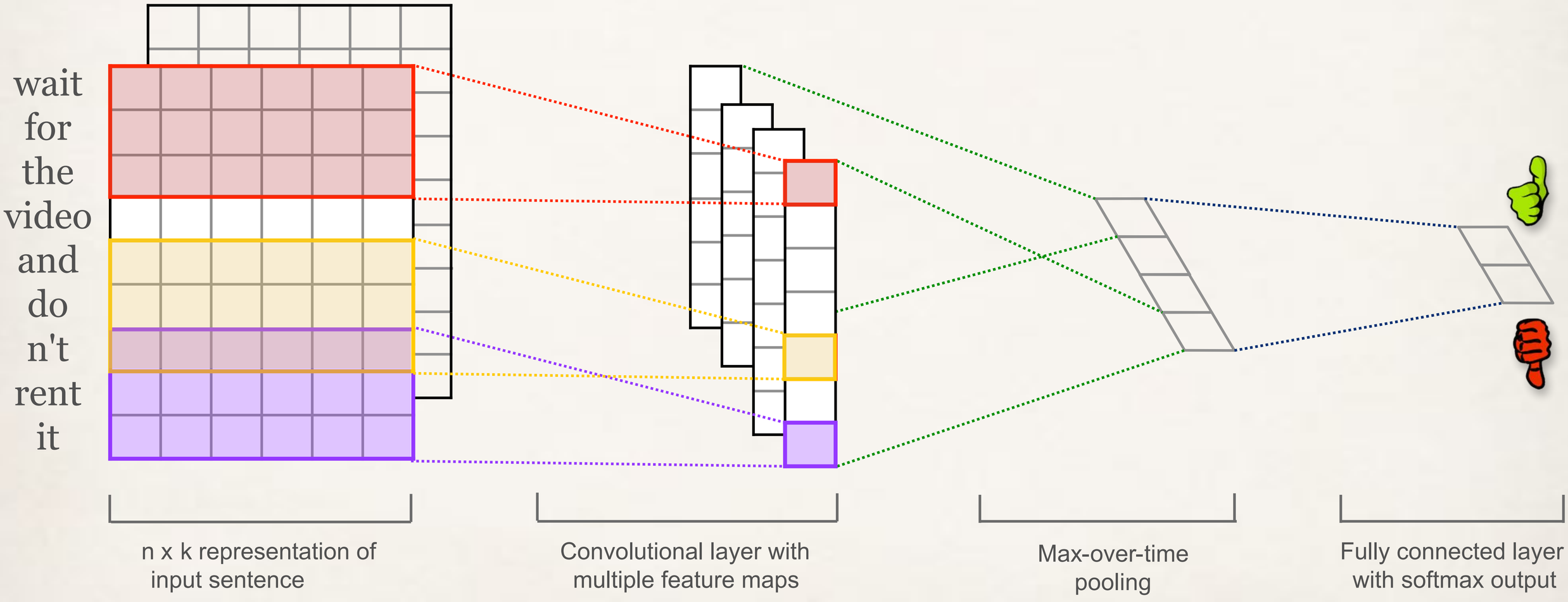
Text Representation Learning

- How to represent a sequence of words?

Text Representation Learning

- Neural Computers, Attention
- Recurrent Networks (such as LSTM)
- Convolutional Neural Networks (CNN)
- paragraph2vec / doc2vec
- Matrix Factorizations, FastText

Convolutional Neural Network (CNN)



adapted from [Kim 2014]

Results

- 1st Place **SemEval** 2016 Competition
  - Convolutional NN

  - ETH Master Theses

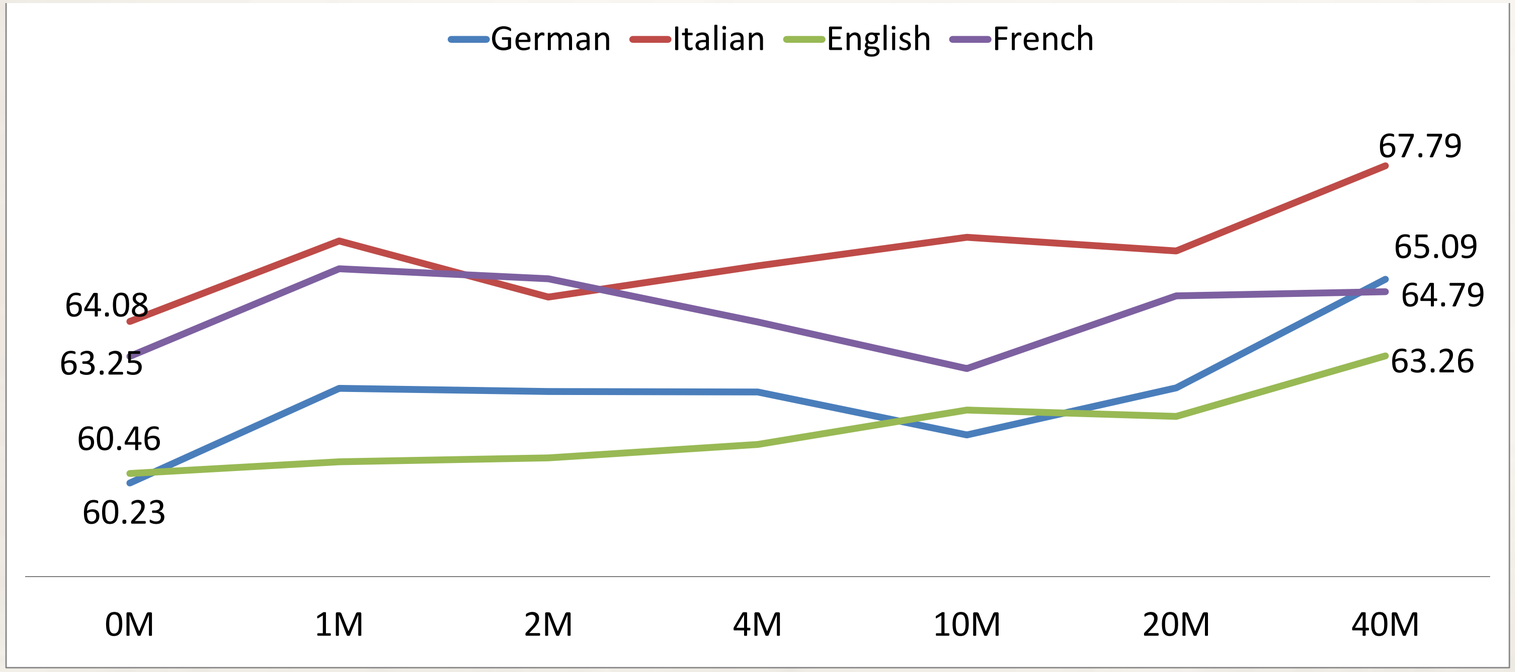
    Jan Deriu

    & Maurice Gonzenbach

negative	*neutral*	But i wanna wear my Concords tomorrow though but i don't feel like it
positive	*neutral*	Gonna watch Grey's Anatomy all day today and tomorrow(:
negative	*neutral*	@CoachVac heey do you know anything about UVA's fallll fest loll they
neutral	*neutral*	@DustyEf when that sun is high in that Texas sky, I'll be buckin it to co
neutral	*positive*	Up 20 points in my money league with Vernon Davis and L. Fitz still to
neutral	*positive*	DEEJAYING this FRIDAY in THE FIRST CHOP it's CHRIS actual SMITH w
negative	*negative*	The Rick Santorum signing that was scheduled for tomorrow at the Boo
positive	*neutral*	@dreami9 lol yep looks like it! Was after El Clasico on Sunday. I didn't
neutral	*neutral*	Back in Stoke on Trent for the 2nd time today!
neutral	*neutral*	First Girls Varsity Basketball Game tomorrow at 6:00 pm Then Football
neutral	*neutral*	#UFC lightweights @Young__Assassin VS @jamievarner set for TUF 16
neutral	*neutral*	@OOOOO_WEEEE slide thru sometime this weekend ill have somethin
negative	*negative*	@DannyB618 Sure absolutely-- I meant out of the Bachmann, Perry, S
negative	*negative*	@RichardGordon48 re Levein discussion on Wed. Can't keep changing
neutral	*neutral*	Today In History November 02, 1958 Elvis gave a party at his hotel bef
neutral	*positive*	Hustle cause you got to then kick back n party everyday like its Fri
positive	*positive*	I can't sleep. Way too exited about Vancouver tomorrow! I'm like a kid

Distant Supervision

- millions of tweets containing :) or :(



SemEval 2016,
WWW 2017

Text Representation Learning - **Unsupervised**?

- Neural Computers, Attention
- Recurrent Networks (such as LSTM)
- Convolutional Neural Networks (CNN)
- paragraph2vec / doc2vec
- Matrix Factorizations, FastText

Unsupervised?

- modify supervised model to predict next word
- negative sampling
- FastText

$$\min_{\mathbf{U},\mathbf{V}} \mathcal{L}(\mathbf{U},\mathbf{V}) := \sum_{\mathbf{s}_n \text{ a sentence}} f(y_n \mathbf{U}\mathbf{V}^\top \mathbf{s}_n)$$

- large datasets, distributed training

[Joulin et al., 2016;
Bojanowski et al., 2016]

Thanks!

mlo.epfl.ch

Virginia Smith, Simone Forte, Chenxin Ma, Martin Takac, Michael I. Jordan, Celestine Dünner, Jan Deriu, Maurice Gonzenbach, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, Matteo Pagliardini, Gupta Prakhar