Introduction to Big Data Optimization

Peter Richtárik

EPSRC Fellow in Mathematical Sciences
(Mathematical Underpinnings of OR)

OR58 - Portsmouth - September 6-8, 2016

OPERATIONAL RESEARCH

THE SCIENCE OF ~~BETTER~~

BEST

Outline

1. Data Science, Big Data & Optimization
2. Applications
3. Methods

Part 1
Data Science, Big Data & Optimization

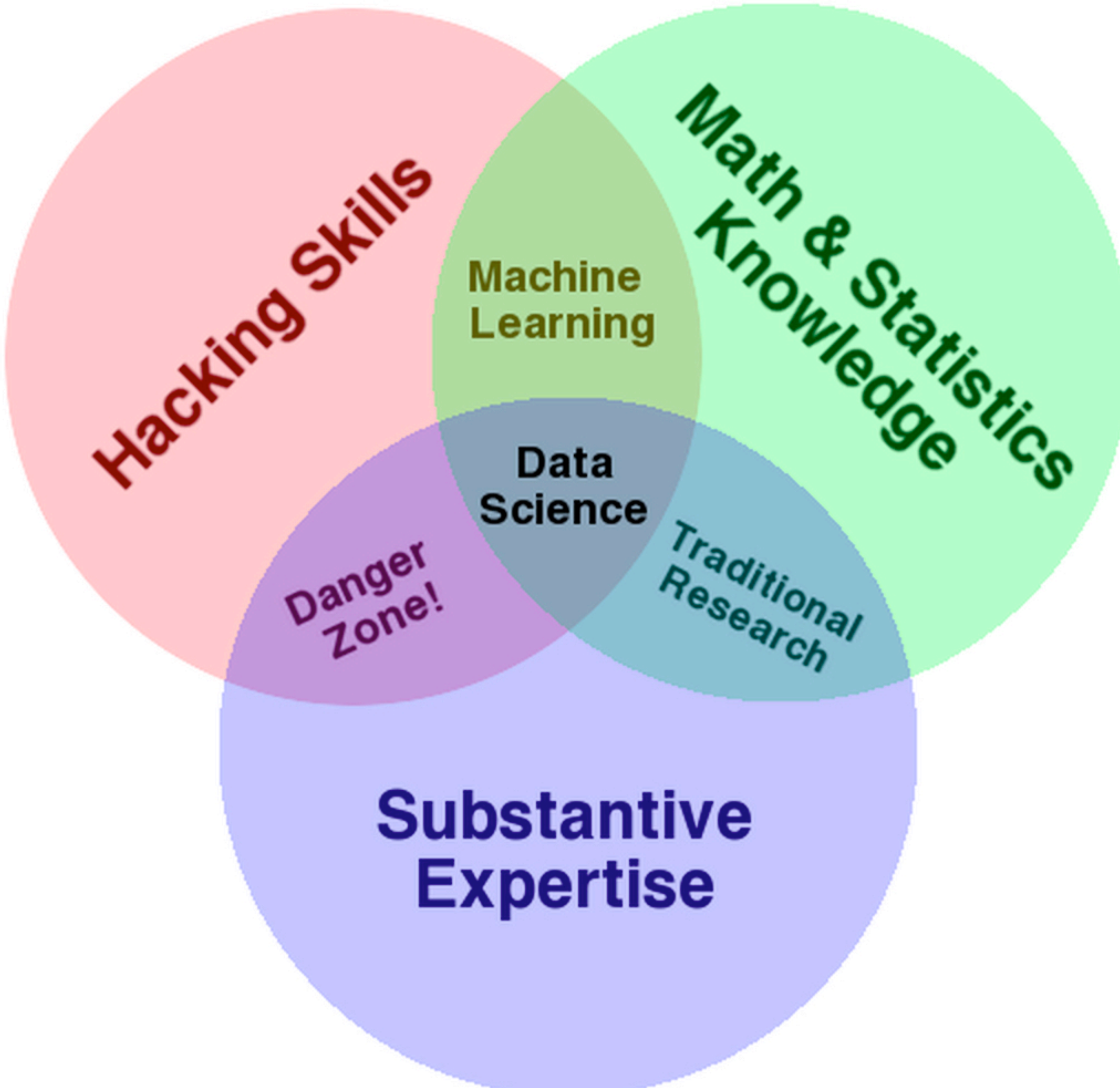
Data Science and Machine Learning

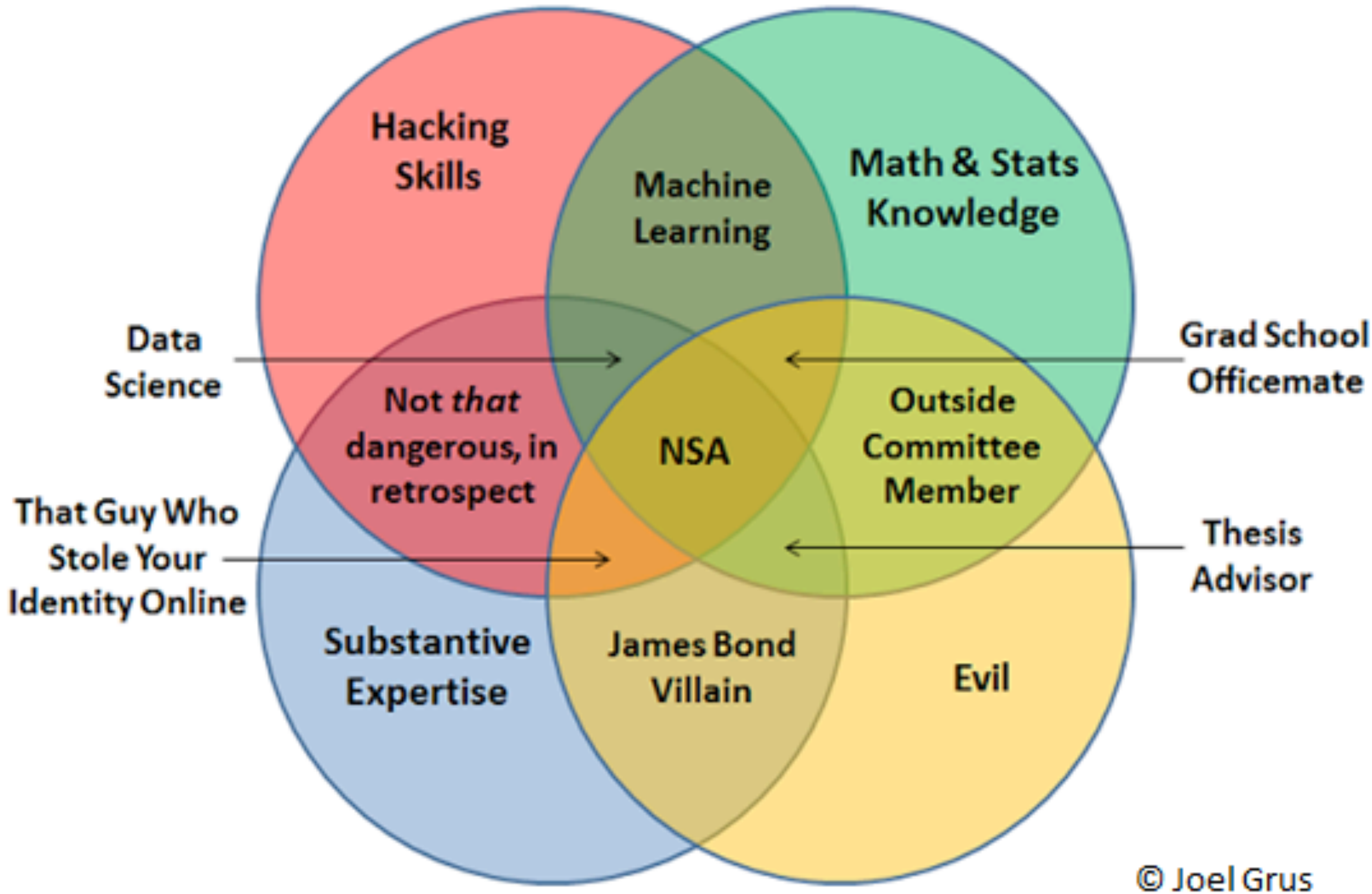
Data: Anything collected/recorded in digital form of potential value

- Text, music, video, images, scans, databases, health records, tax data, email, online clicks, tweets, blogs, …
- Usually modelled statisically, or as a signal

Data Science: Extraction of knowledge from data

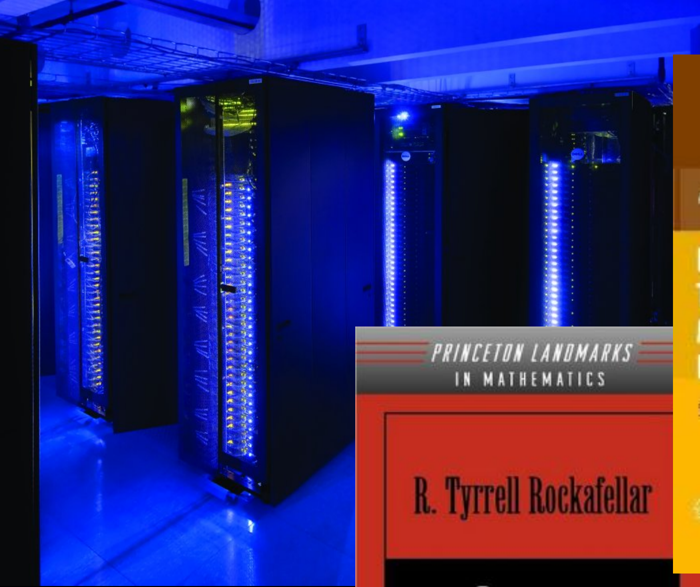
Machine Learning: Automated learning from available data to make predictions & decisions about unseen data


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

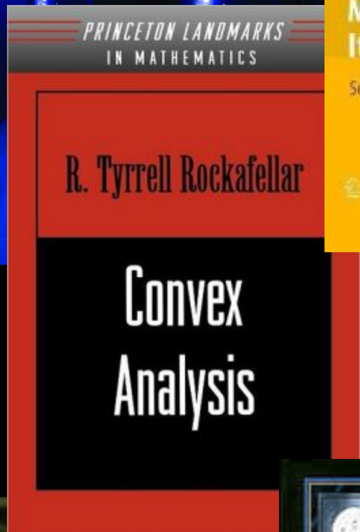


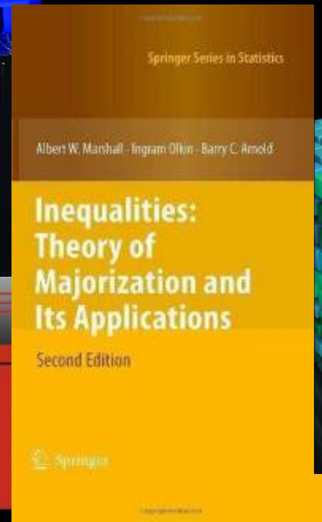
Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil
© Joel Grus

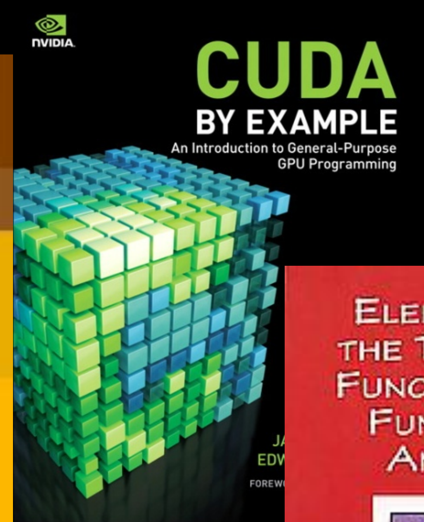
Tools

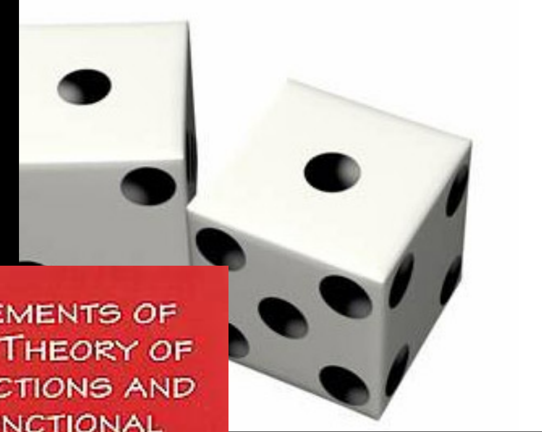


HPC

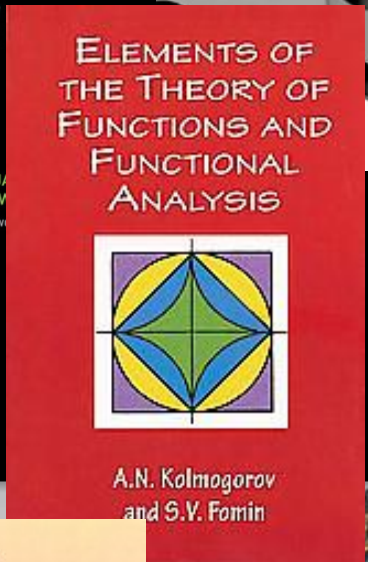




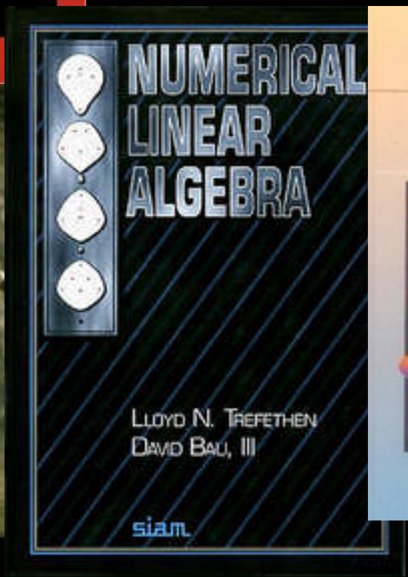


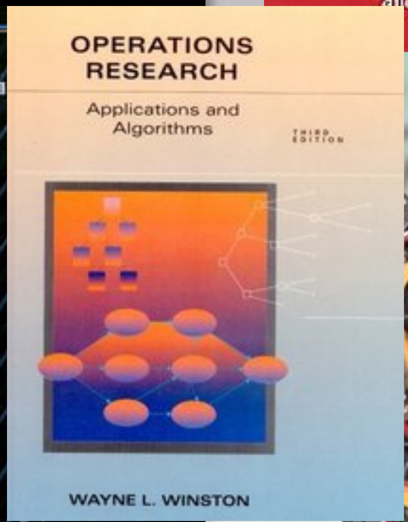


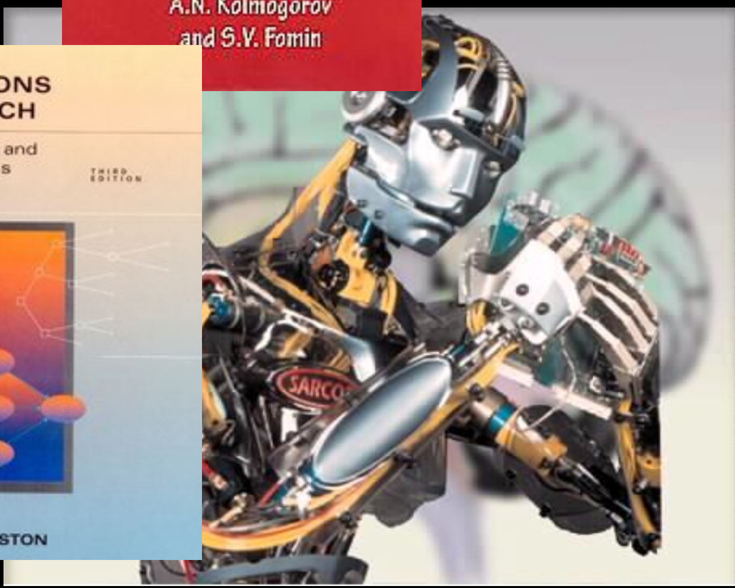
Probability



Matrix Theory







Machine Learning

Translational Research

EPSRC
Engineering and Physical Sciences Research Council

Foundational Research

UNIVERSITY OF CAMBRIDGE

THE ALAN TURING INSTITUTE

WARWICK
THE UNIVERSITY OF WARWICK

THE UNIVERSITY of EDINBURGH

UCL

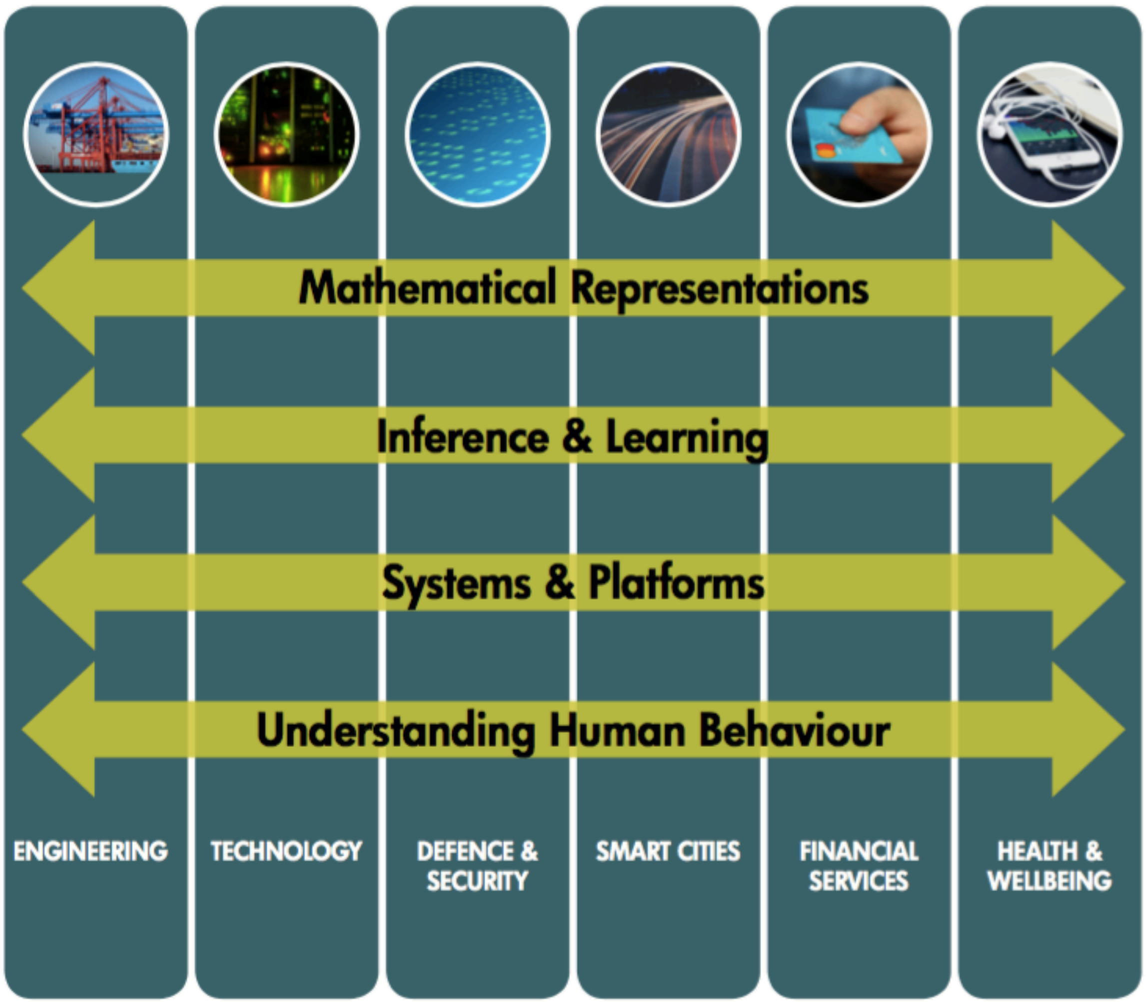
UK's national institute for data science

Leading Public Conversation

UNIVERSITY OF OXFORD

Training the Next Generation

Strategic Priorities of the ATI



4 Key Capabilities

6 Priority Sectors for Translational Research

Big Data

Too much hype?

"Big data opens the door to a new approach to understanding the world and making decisions" (New York Times, 2013)

"Don't be colonized by the Americans with their big data, colonize them" (Cathal MacSwiney Brugha, 8.9.2016)

Data that can't be stored on a "typical system" or analyzed via "normal procedures"

What to do with huge quantities of data?

- New models
- New algorithms

ALAN COCHRANE TAXIS
38 HARRYSMUIR GARDENS
THANK YOU

TERMINAL ID: ****3577
MERCHANT ID: **********21661

MASTERCARD
************6875 ICC
PAN SEQ NO: 12
AID: A0000000041010

SALE
AMOUNT £100009.96

*** CUSTOMER COPY ***

TRANSACTION VOID

DATE: 05/03/13 TIME: 16:00

Optimization & Big Data

Conference Series

- Established & run in Edinburgh in 2012, 2013, 2015, 2017

Optimization plays a key role in big data analysis

- Machine Learning = Stochastic Optimization (Srebro)
- Optimization is used to train ML models
- Optimization used in discovering new data representations
- Optimization used in turning extracted knowledge into action

Optimization Objective in Big Data Problems

Objective is formed from collected data, and hence is not a "precise object"

- Low to medium accuracy solutions are fine!
- What methods can find rough solutions quickly?

Objective often simple

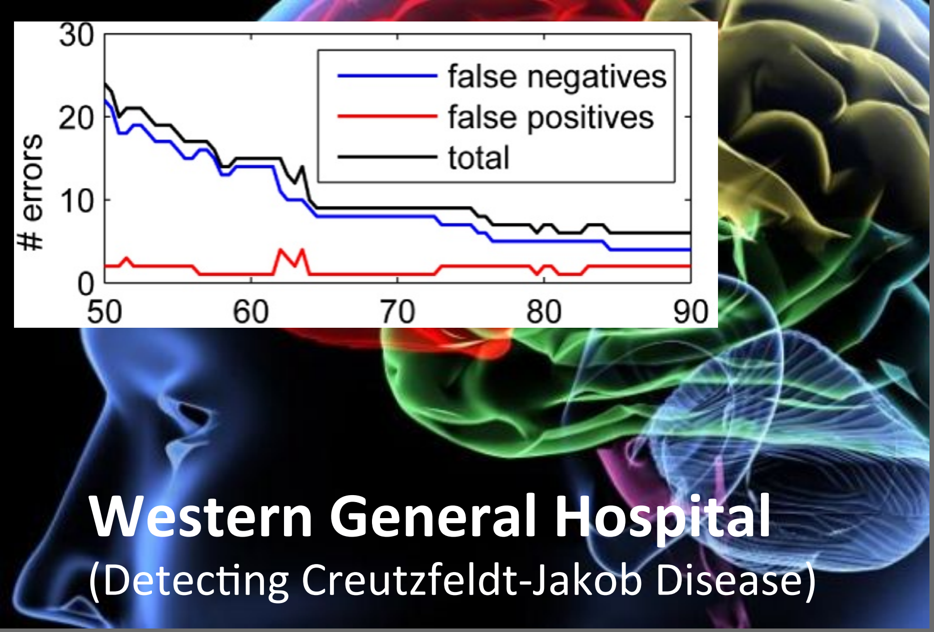
- The more data we have, the less modeling we should do: "the model is in the data"
- Typically: Data-fitting term + Prior knowledge term

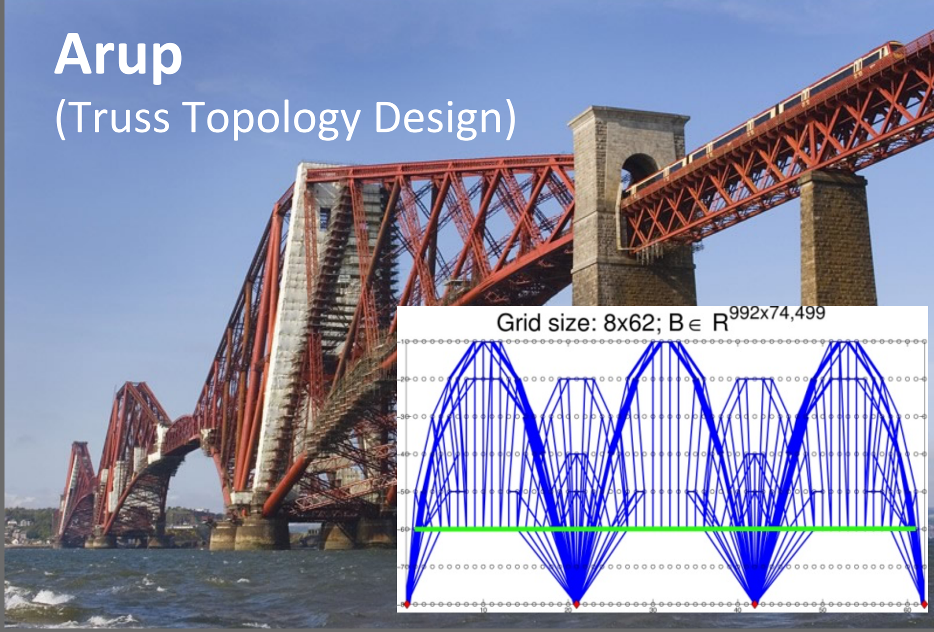
$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_2^2$$

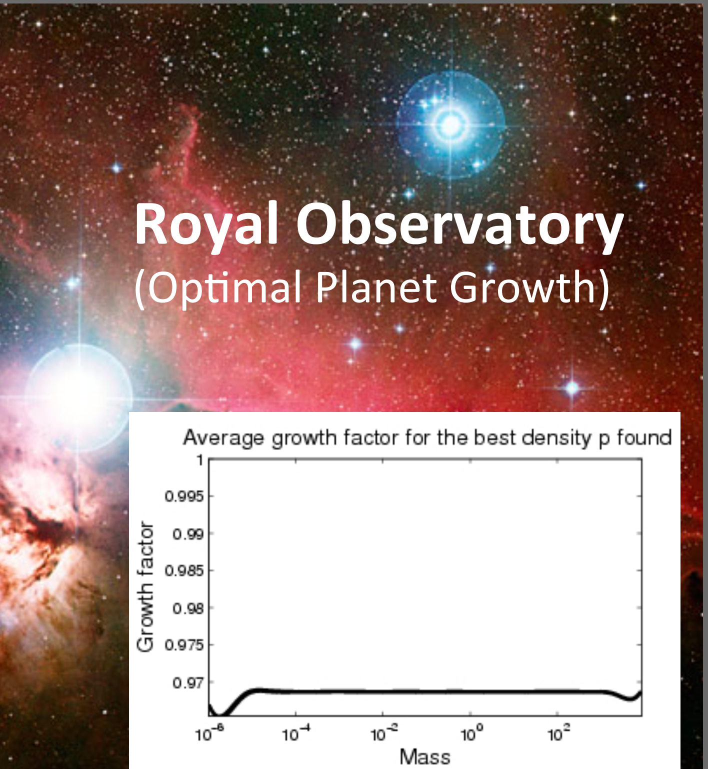
Part 2
Applications

Application Areas

- Natural language processing
  - speech recognition
- Text processing
  - text prediction, recognition, machine translation, spam filtering
- Image & video processing
  - deblurring, denoising, inpainting, face detection and recognition
- Social networks
  - community detection, geo-tagging of tweets
- Public records analysis
  - tax data, financial records, health records
- Online advertising
  - ad allocation, ad pricing
- Scientific measurements
  - truss topology design, inverse problems, data assimilation, gene expression analysis,


Western General Hospital
(Detecting Creutzfeldt-Jakob Disease)
errors
false negatives
false positives
total
0
10
20
30
50
60
70
80
90


Arup
(Truss Topology Design)
Grid size: 8x62; B ∈ R^992x74,499


Royal Observatory
(Optimal Planet Growth)
Average growth factor for the best density p found
Growth factor
Mass

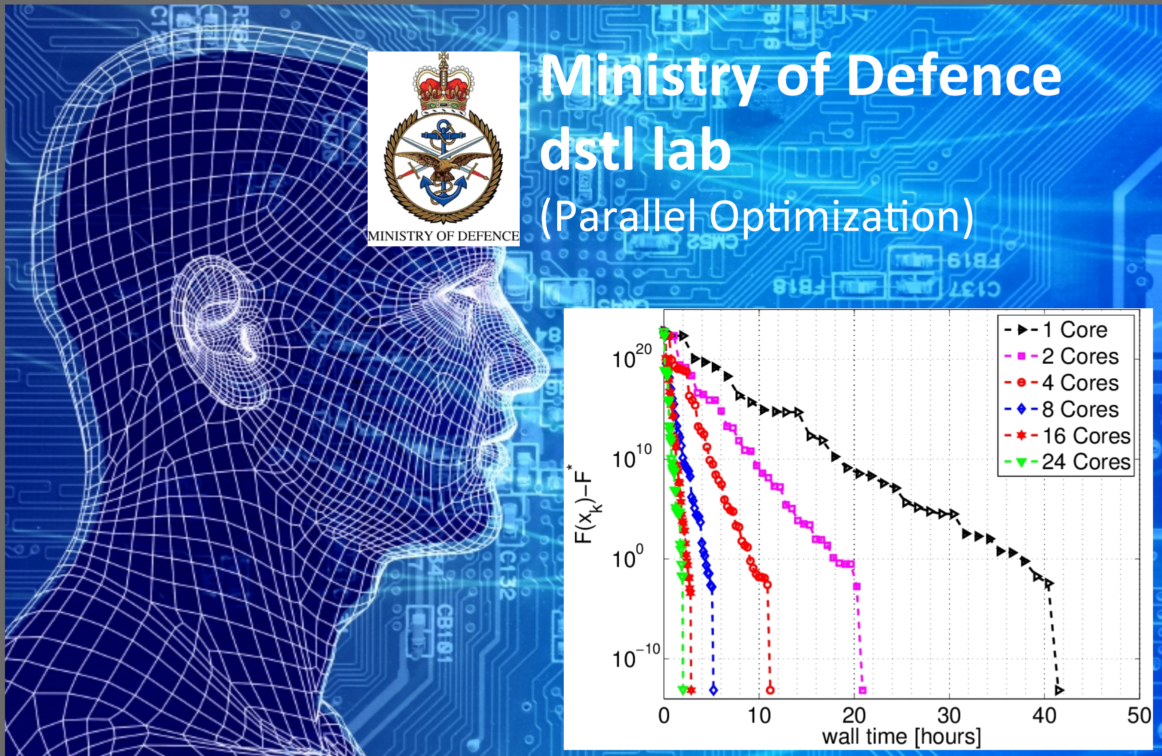
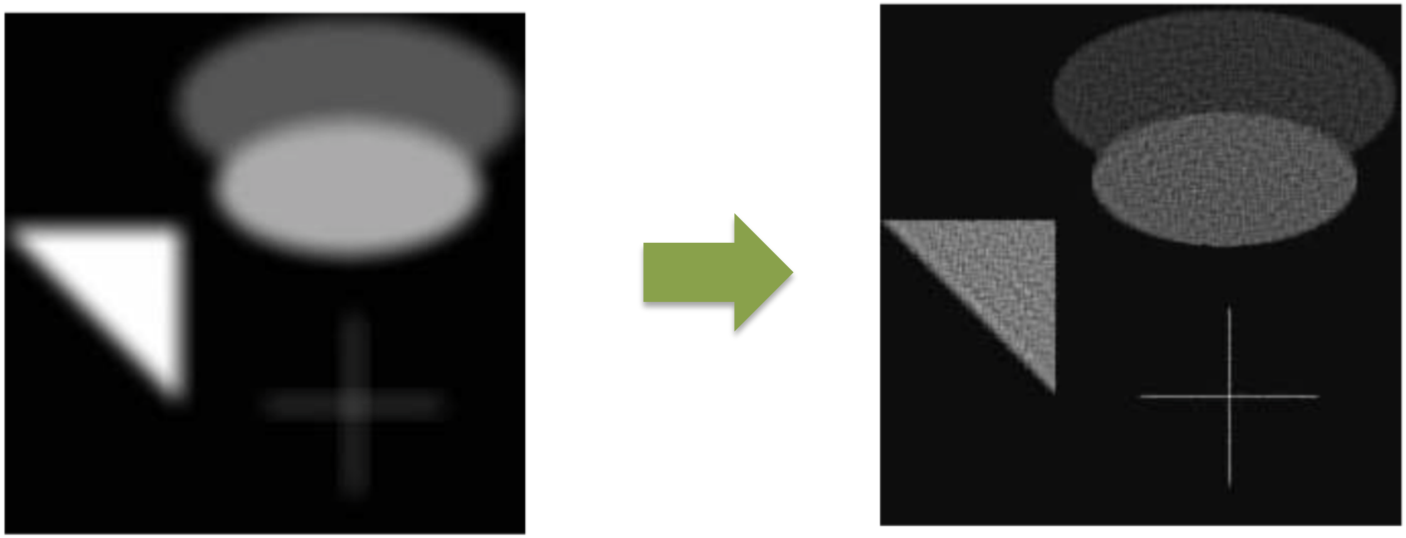

MINISTRY OF DEFENCE
Ministry of Defence
dstl lab
(Parallel Optimization)
1 Core
2 Cores
4 Cores
8 Cores
16 Cores
24 Cores
F(x_k)−F*
wall time [hours]

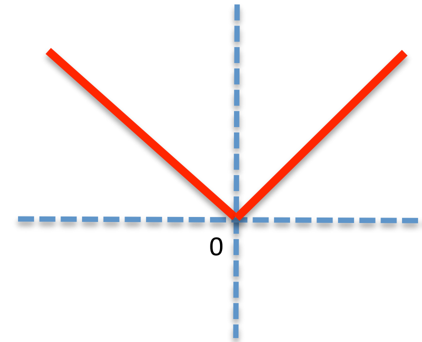
Image Deblurring



Amir Beck and Marc Teboulle. **A Fast Iterative Shrinking-Thresholding Algorithm for Linear Inverse Problems.** *SIAM J. Imaging Sciences* 2(1), 183-202, 2009

Jakub Konečný, Jie Liu, P.R., Martin Takáč. **Mini-Batch Semi-Stochastic Gradient Descent in the Proximal Setting.** *IEEE Journal of Selected Topics in Signal Processing* 10(2), 242-255, 2016

Image Deblurring: "LASSO" Problem



blurred image

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$$

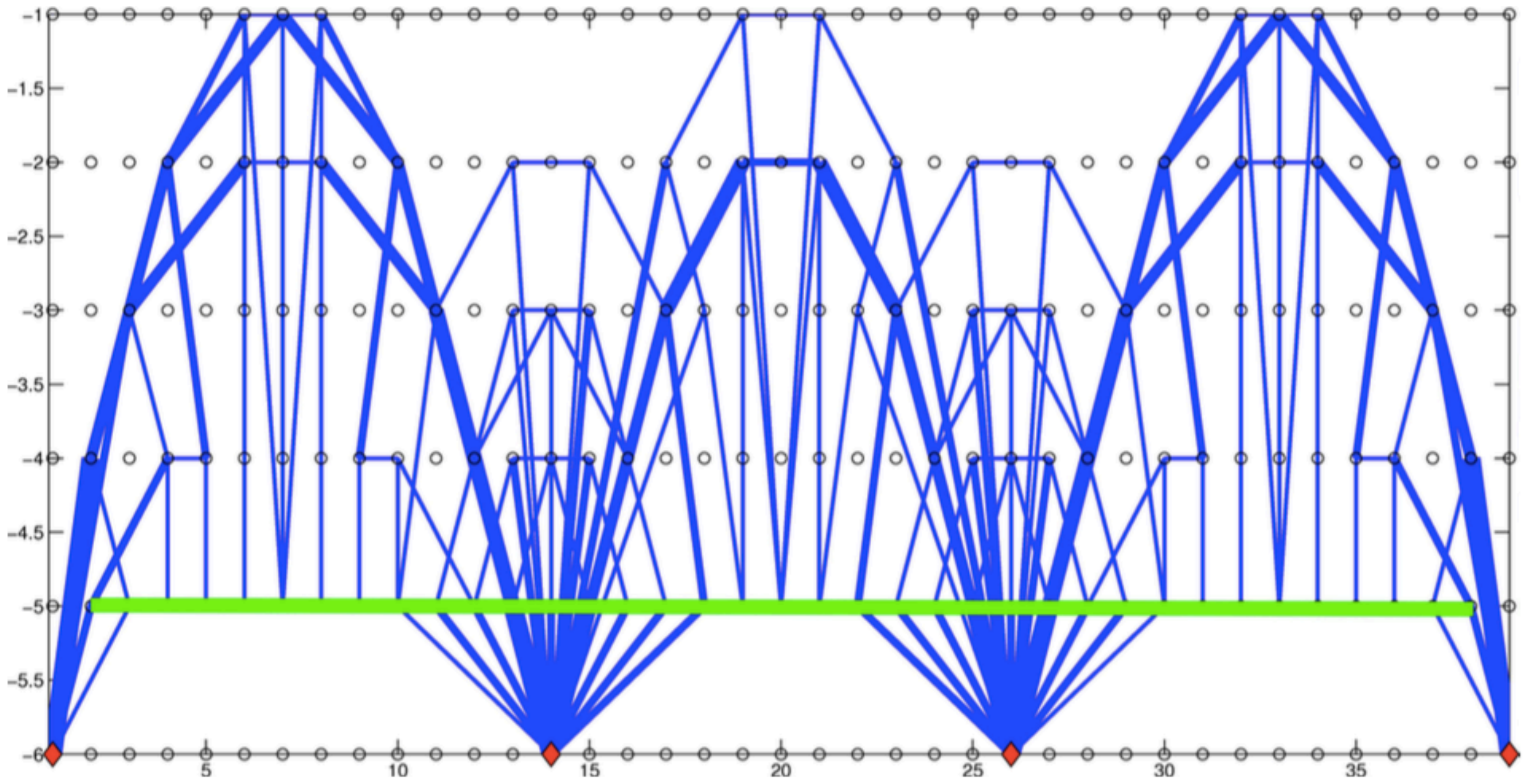
image

pixels in the image

Blurring matrix multiplied by a wavelet basis matrix

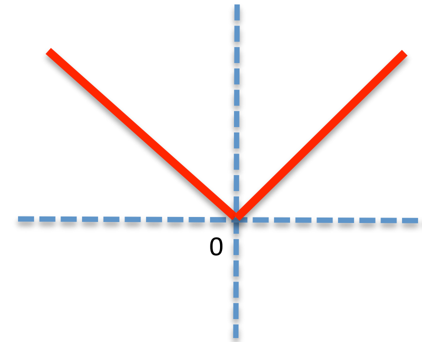
Encourages sparsity in the wavelet basis

Truss Topology Design



P.R. and Martin Takáč. **Efficient Serial and Parallel Coordinate Descent Methods for Huge-Scale Truss Topology Design.** *Operations Research Proceedings*, pp 27-32, 2012

Truss Topology Design: "LASSO" Problem



Encodes all potential bars

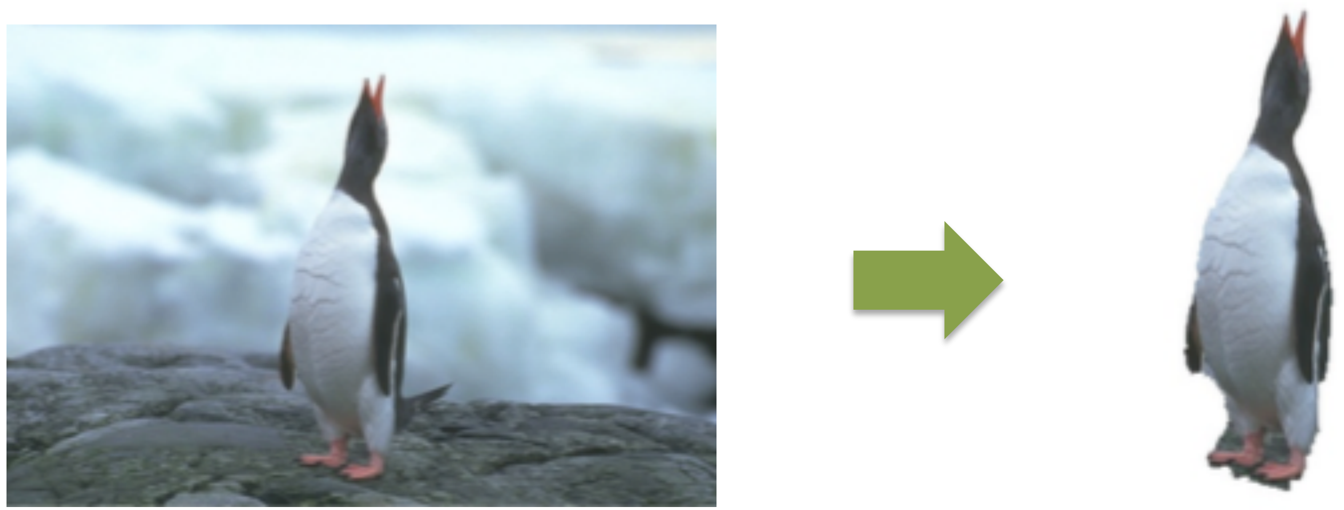
$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$$

potential bars (quadratic in mesh size)

Least-squares (convex, smooth, quadratic)

L1 norm (convex, nonsmooth, but "simple")

Image Segmentation



Olivier Fercoq and P.R. **Accelerated, Parallel and Proximal Coordinate Descent.** *SIAM Journal on Optimization* 25(4), 1997-2023, 2015

Alina Ene and Huy L. Nguyen. **Random Coordinate Descent Methods for Minimizing Decomposable Submodular Functions.** *ICML* 2015

Image Segmentation: Reformulated Submodular Optimization

$$\text{minimize} \quad \frac{1}{2}\left\|\sum_{i=1}^{n} x^i\right\|^2$$

$$\text{subject to} \quad x^i \in P^i, \; i = 1, 2, \ldots, n$$

Smooth, convex, quadratic

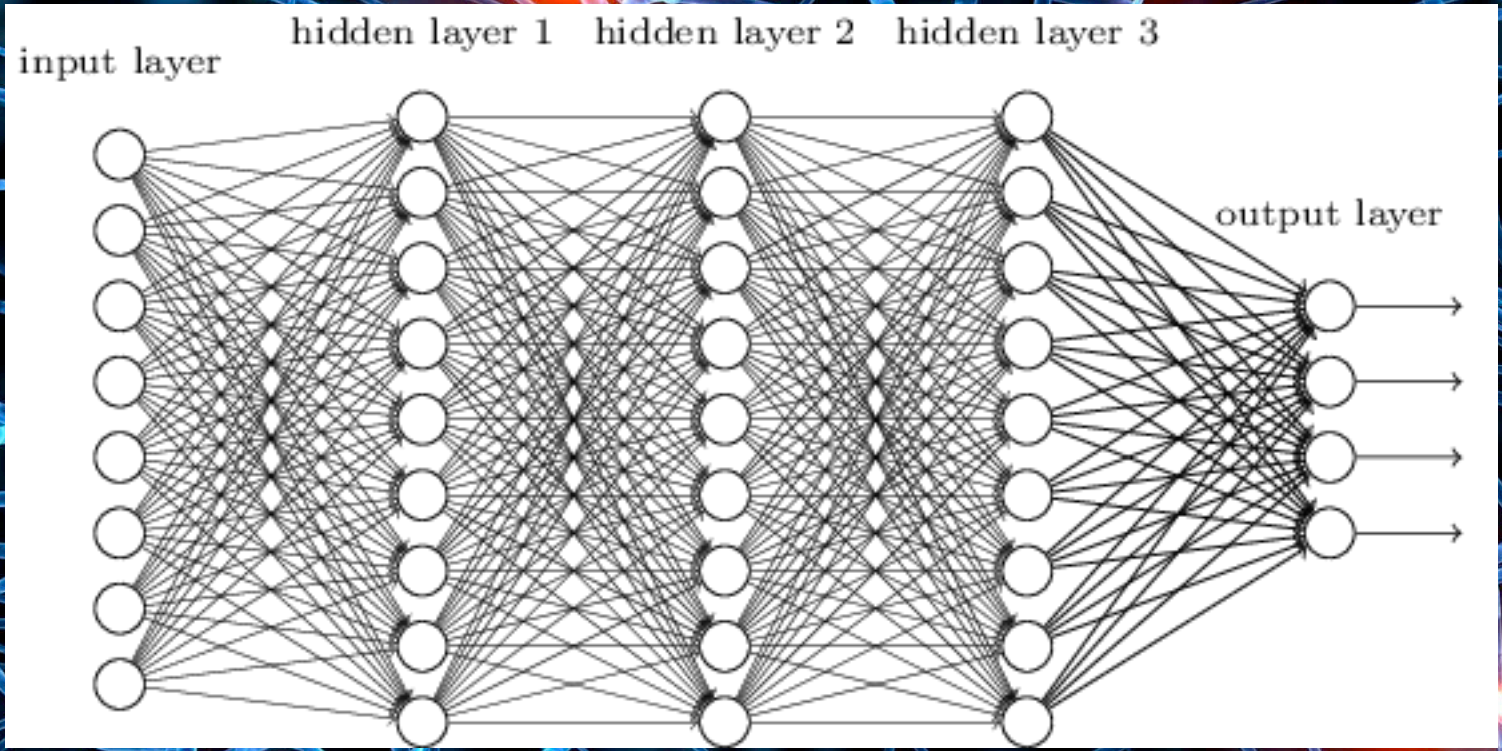
polytope

grows with the image size

Predicting Expert Moves in Go

Silver et al. **Mastering the Game of Go with Deep Neural Networks and Tree Search.** *Nature* 529, pp 484–489, 2016

Go: Training a Neural Network



Go: Training a Neural Network

$$\min_{x \in \mathbb{R}^d} \left\{ F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

average loss

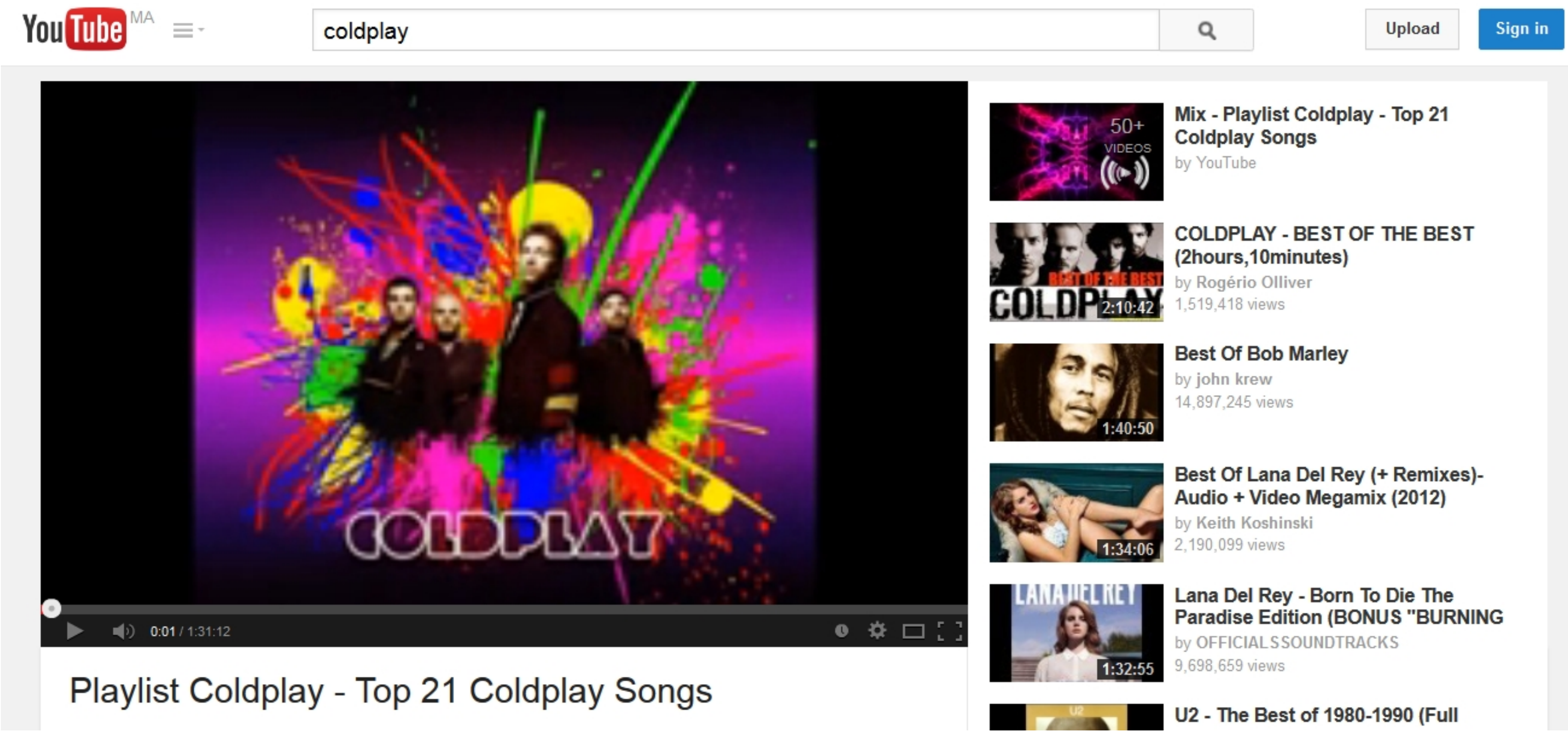
n is big
(# of training
data points)

parameters
representing the
network

loss functions

Face Detection

Recommender Systems



Geotagging Tweets

Cornell University Library

arXiv.org > cs > arXiv:1404.7152

Search or Ar

Computer Science > Social and Information Networks

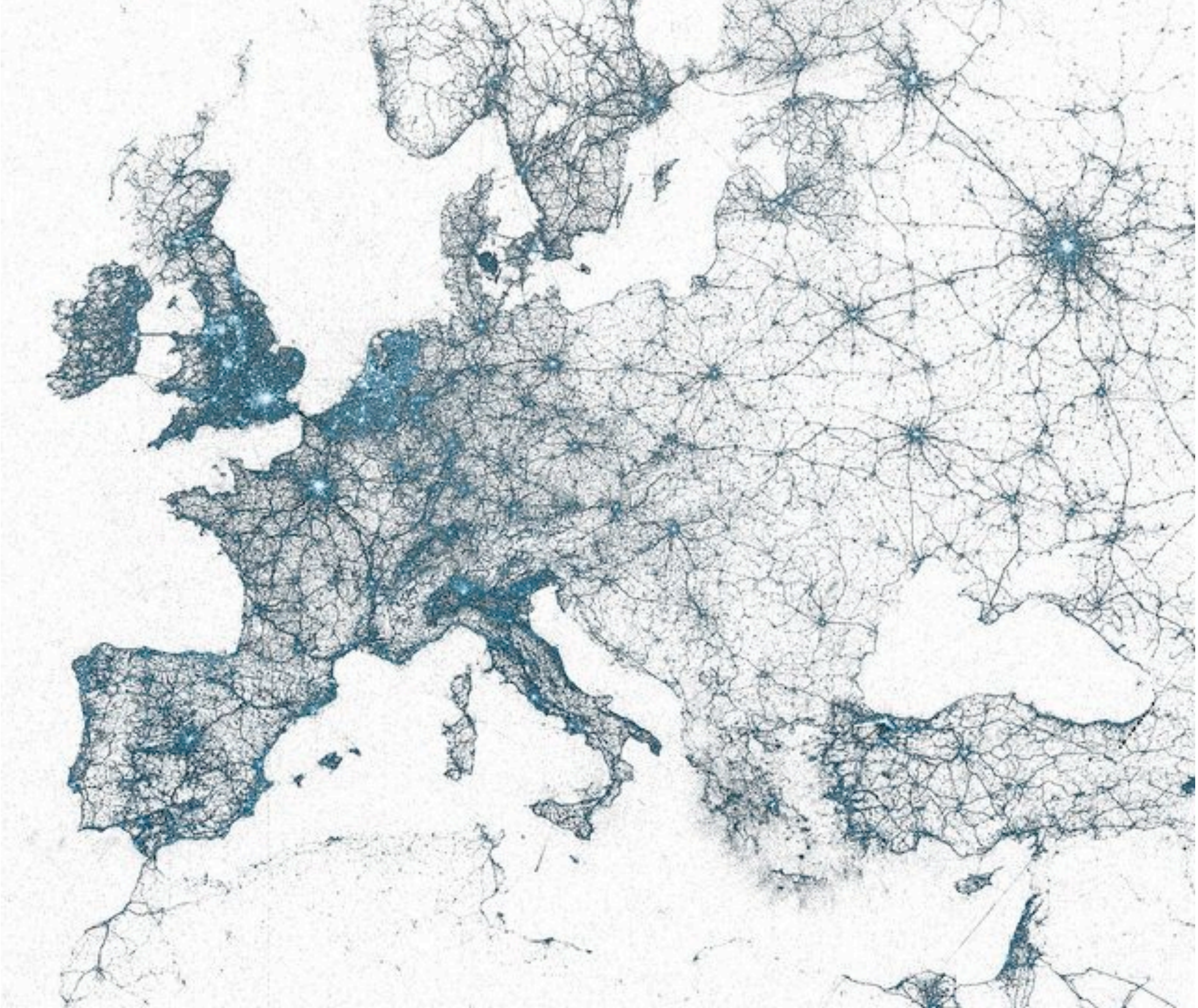
Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization

Ryan Compton, David Jurgens, David Allen

(Submitted on 28 Apr 2014)

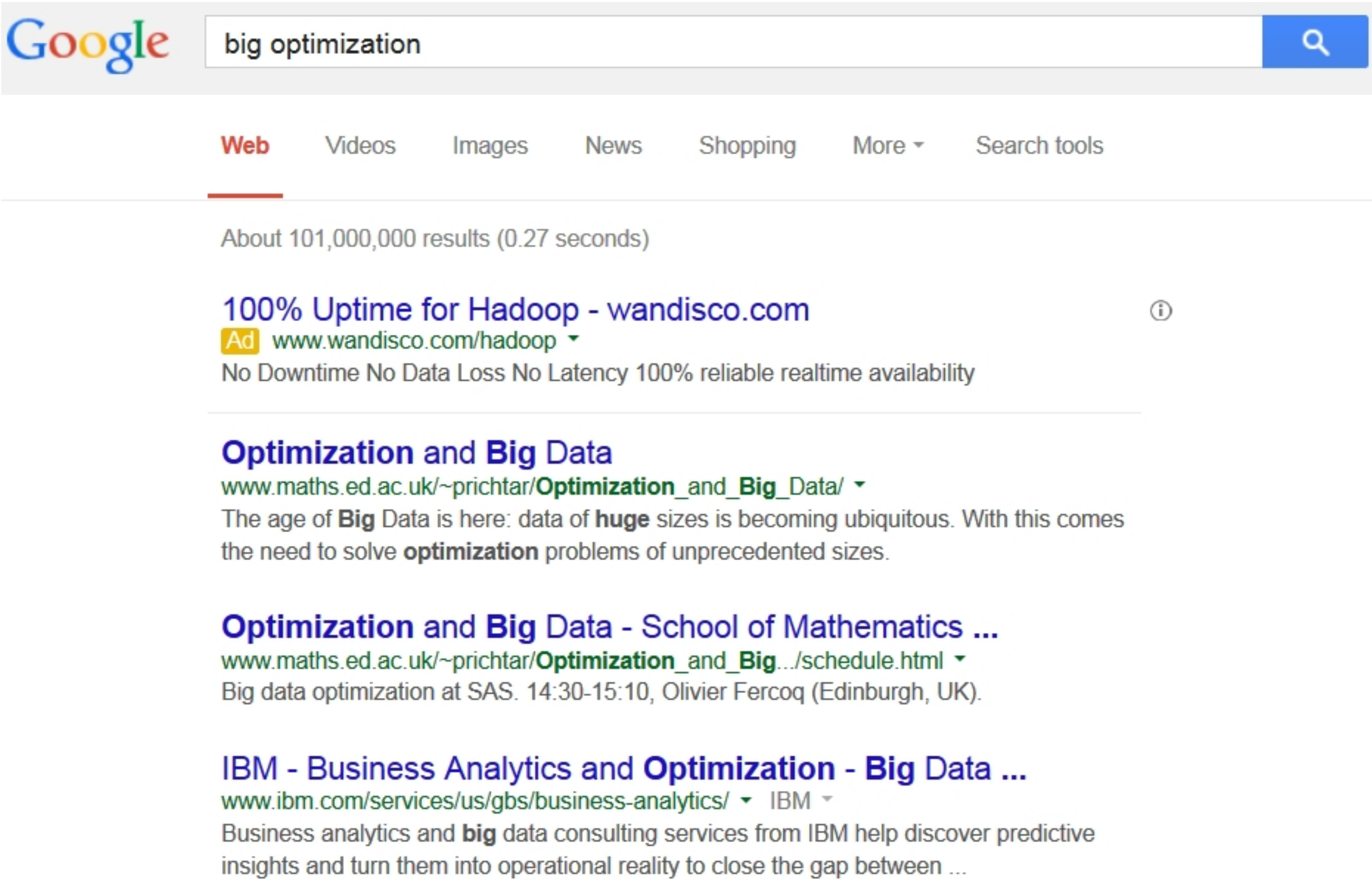
Geographically annotated social media is extremely valuable for modern information retrieval. However, when researchers can only access publicly-visible data, one quickly finds that social media users rarely publish location information. In this work, we provide a method which can geolocate the overwhelming majority of active Twitter users, independent of their location sharing preferences, using only publicly-visible Twitter data.

Our method infers an unknown user's location by examining their friend's locations. We frame the geotagging problem as an optimization over a social network with a total variation-based objective and provide a scalable and distributed algorithm for its solution. Furthermore, we show how a robust estimate of the geographic dispersion of each user's ego network can be used as a per-user accuracy measure, allowing us to discard poor location inferences and



Spam Filtering

Ranking


Google
big optimization
Web
Videos
Images
News
Shopping
More
Search tools
About 101,000,000 results (0.27 seconds)
100% Uptime for Hadoop - wandisco.com
Ad
www.wandisco.com/hadoop
No Downtime No Data Loss No Latency 100% reliable realtime availability
Optimization and Big Data
www.maths.ed.ac.uk/~prichtar/Optimization_and_Big_Data/
The age of Big Data is here: data of huge sizes is becoming ubiquitous. With this comes the need to solve optimization problems of unprecedented sizes.
Optimization and Big Data - School of Mathematics ...
www.maths.ed.ac.uk/~prichtar/Optimization_and_Big.../schedule.html
Big data optimization at SAS. 14:30-15:10, Olivier Fercoq (Edinburgh, UK).
IBM - Business Analytics and Optimization - Big Data ...
www.ibm.com/services/us/gbs/business-analytics/
IBM
Business analytics and big data consulting services from IBM help discover predictive insights and turn them into operational reality to close the gap between ...

Application in Focus

Training Linear Predictors

"Predict based on past observations"

Statistical Nature of Data

$$(A_i, y_i) \sim Distribution$$

DATA

The New York Times

WEDNESDAY, MAY 10, 2006

Above, finalists from Nick at Nite's "Funniest Mom in America 2." Right, Darlene Westgor, the 2005 winner.

A Mom's Riff on the Parodies of Parenthood

Continued From First Arts Page

$$A_i \in \mathbb{R}^{d \times m}$$

LABEL

"politics"

$$y_i \in \mathbb{R}^m$$

Prediction of Labels from Data

Find $w \in \mathbb{R}^d$ ← Linear predictor

such that when a (data, label) pair is drawn from the distribution

$$(A_i, y_i) \sim Distribution$$

then

$$A_i^\top w \approx y_i$$

Predicted label ($A_i^\top w$)

True label (y_i)

Measure of Success

$$loss(a, b)$$

Predicted label

True label

We want the expected loss (=risk) to be small:

$$\mathbf{E}\left[loss(A_i^\top w, y_i)\right]$$

$$(A_i, y_i) \sim Distribution$$

Replace Expectation by Average

Draw i.i.d. data samples from the distribution

$$(A_1, y_1), (A_2, y_2), \ldots, (A_n, y_n) \sim Distribution$$

Output predictor which minimizes the empirical risk:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} loss(A_i^\top w, y_i)$$

Minimize the Average of a Large Number of Functions

n is big

$$\min_{x \in \mathbb{R}^d} \left\{ F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

Part 3
Methods

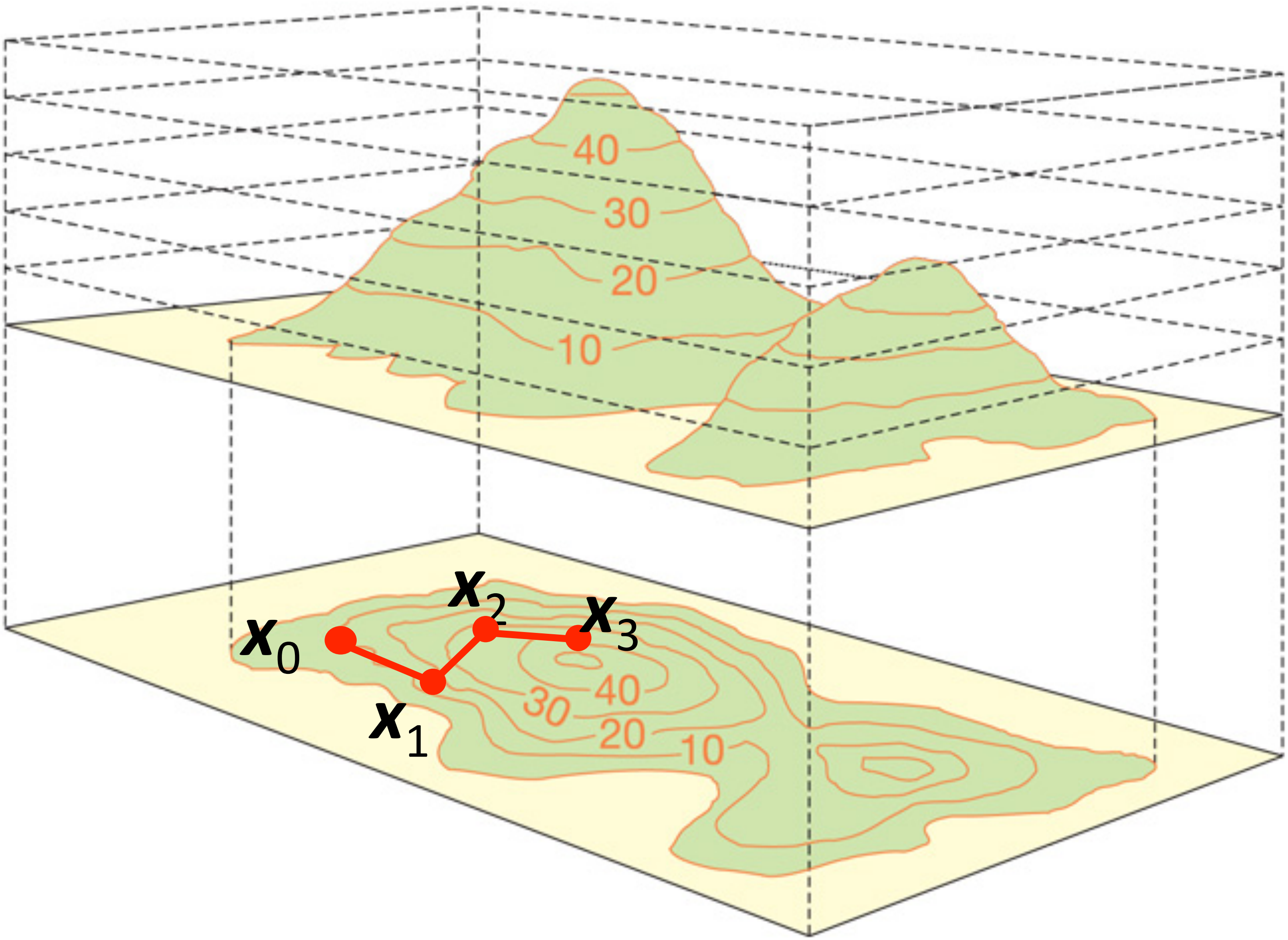
Optimization with Big Data

= Extreme* Mountain Climbing

* in a billion dimensional space on a foggy day

God's Algorithm = Teleportation

Mortals Have to Walk…



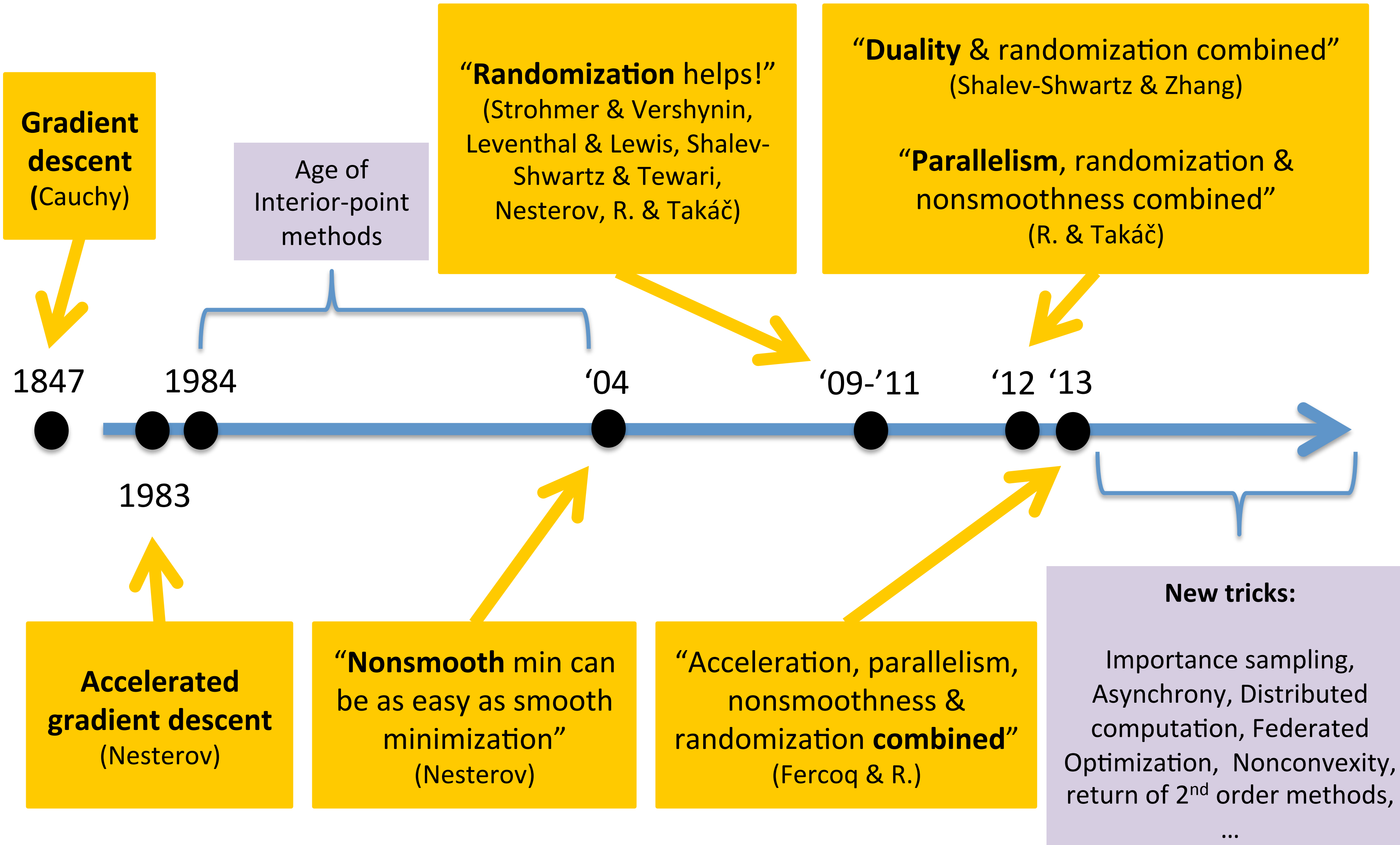
Algorithmic Tricks

1. Gradient descent
2. Handling nonsmoothness via the proximal trick
3. Acceleration
4. Randomized decomposition
5. Parallelism/Minibatching & Sparsity
6. Distributed computation
7. Importance sampling

All these tricks can be combined!

There are more tricks: duality, variance reduction, asynchrony, curvature, …

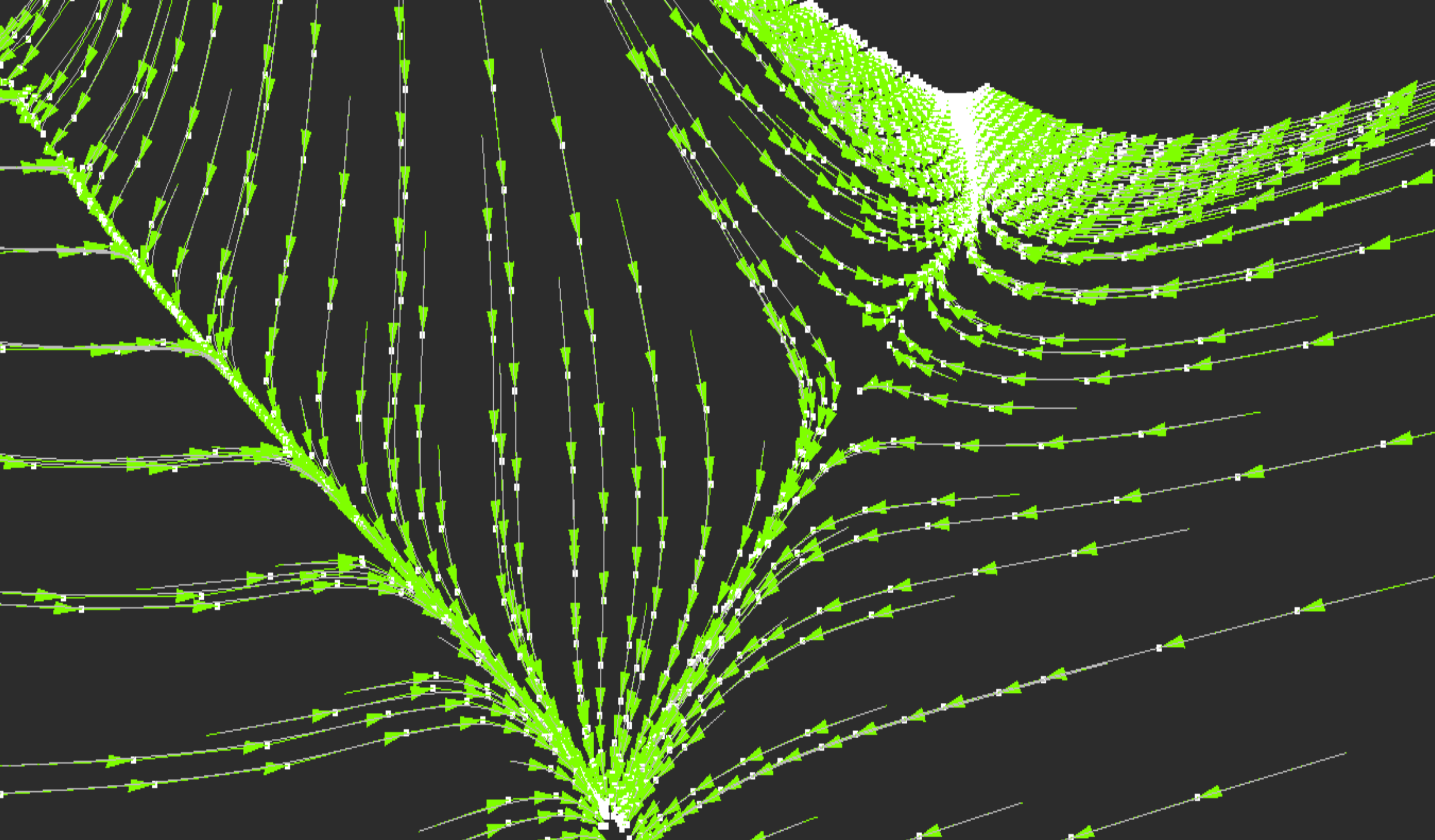
Brief, Biased and Severely Incomplete History of Big Data Optimization



Tool 1

Gradient Descent (1847)

"Just follow a ball rolling down the hill"



Augustin Cauchy

Méthode générale pour la résolution des systèmes d'équations simultanées, *pp. 536–538,* 1847

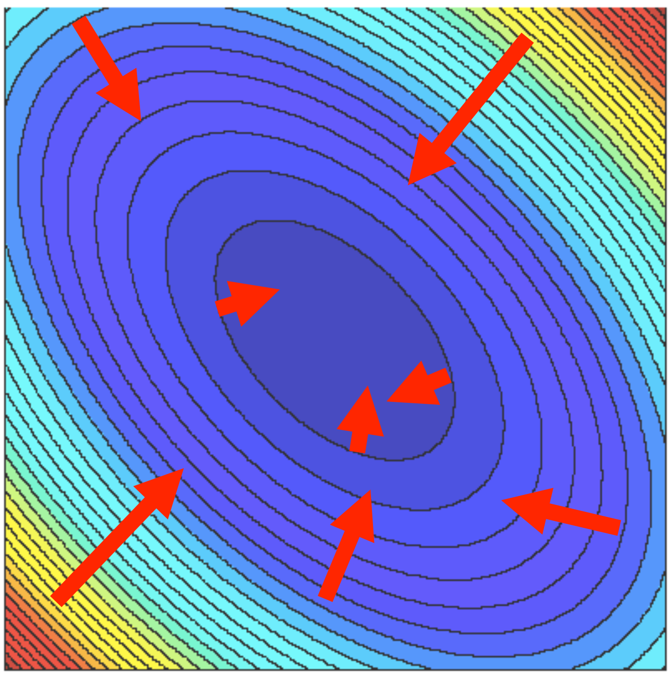
The Problem

$$\min_{x \in \mathbb{R}^d} F(x)$$

Convex, smooth

Gradient Descent (GD)

$$x_{k+1} = x_k - \frac{1}{L}\nabla F(x_k)$$



iterations

condition number of *F*

$$k \geq \frac{L}{\mu}\log(1/\epsilon)$$

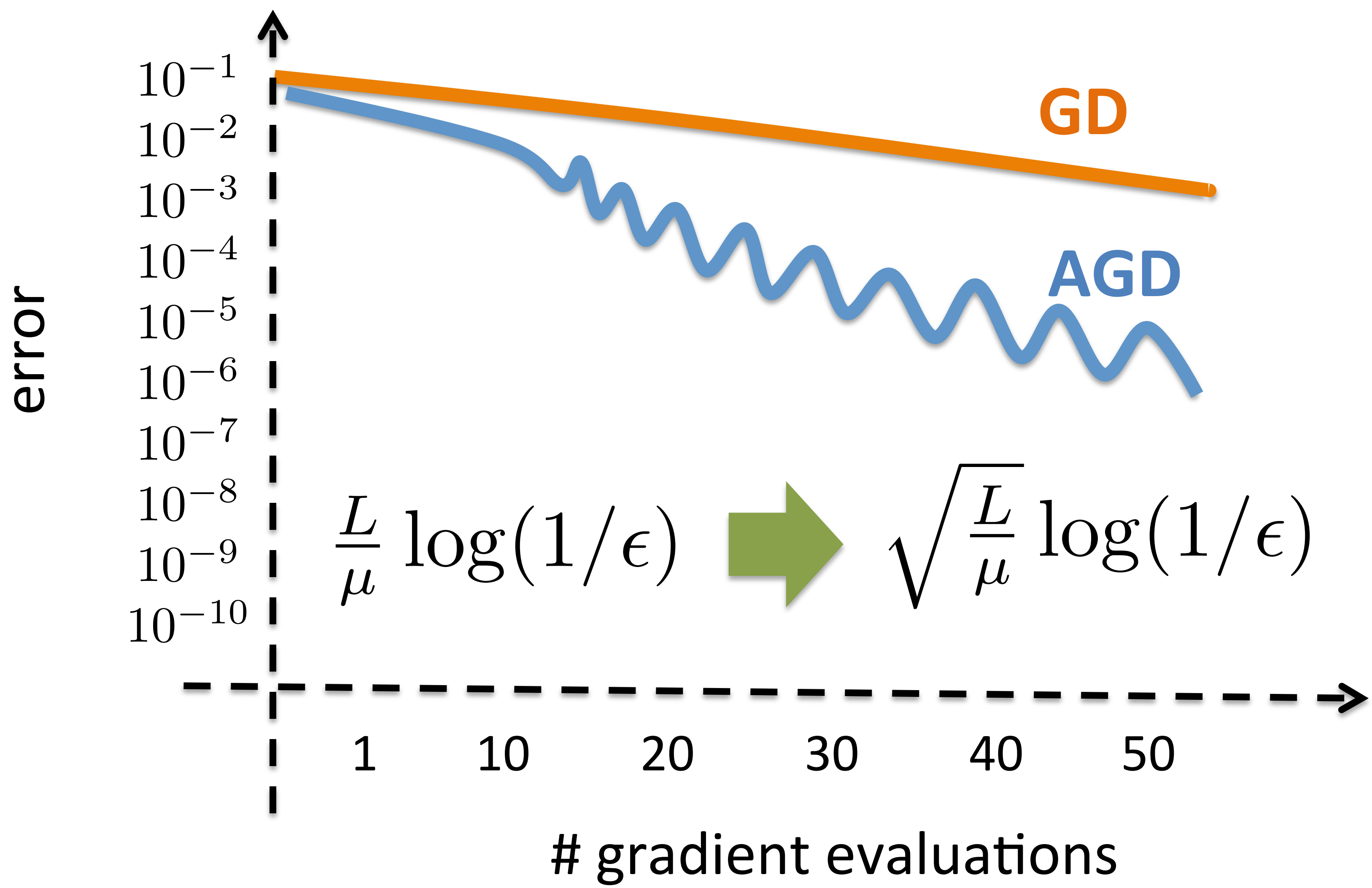
$$F(x_k) - F(x_*) \leq \epsilon$$

Tool 2

Acceleration (1983/2003)

"Gradient descent can be made much faster!"

Acceleration Works (Mysteriously)



Acceleration

- Reignited interest in gradient methods
- Usage in all areas of data science (called momentum in deep neural networks literature)
- Oscilation can be tamed (e.g., by restarting)
- Can be combined with other tricks
  - Duality [Shai-Shalev Shwartz & Zhang 2013]
  - Randomized decomposition, Parallelism, Proximal trick [Fercoq & R 2013]

Yurii Nesterov
Introductory lectures on convex optimization: a basic course
Kluwer, Boston, 2003

Yurii Nesterov
A method for unconstrained convex minimization problem with the rate of convergence O(1 / k^2)
Soviet Math. Doclady 269, 543-547, 1983

Tool 3

Proximal Trick (2004)

"Some nonsmooth problems are as easy as smooth problems"

The Problem

$$\min_{x \in \mathbb{R}^d} F(x) + G(x)$$

Convex, smooth

Convex, nonsmooth

Proximal Gradient Descent (PGD)

STEP 1: Pretend there is no *G*

$$z_{k+1} = x_k - \frac{1}{L}\nabla F(x_k)$$

STEP 2: Take a "proximal" step with respect to *G*

$$x_{k+1} = \arg\min_x \frac{1}{2}\|x - z_{k+1}\|^2 + \frac{1}{L}G(x)$$

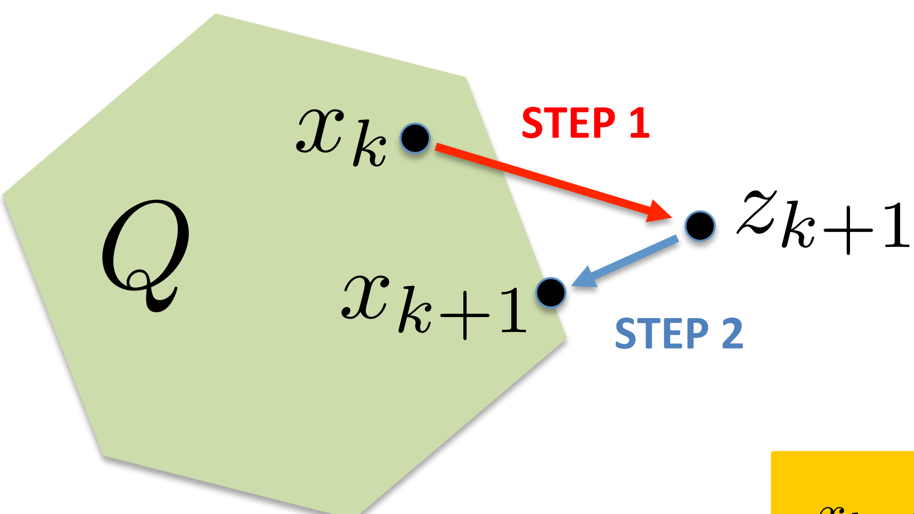
1. Gradient Descent is a special case for $G = 0$
2. Even though this is a nonsmooth problem,
   # steps is the same as for Gradient Descent!!! ← $\frac{L}{\mu}\log(1/\epsilon)$
3. Efficient if Step 2 is easy to do

Example: Projected Gradient Descent

$$\min_{x \in Q} F(x) \quad \Leftrightarrow \quad \min_x F(x) + G(x)$$

Convex set

$$G(x) = \begin{cases} 0 & x \in Q \\ +\infty & x \notin Q \end{cases}$$



$$z_{k+1} = x_k - \frac{1}{L}\nabla F(x_k)$$

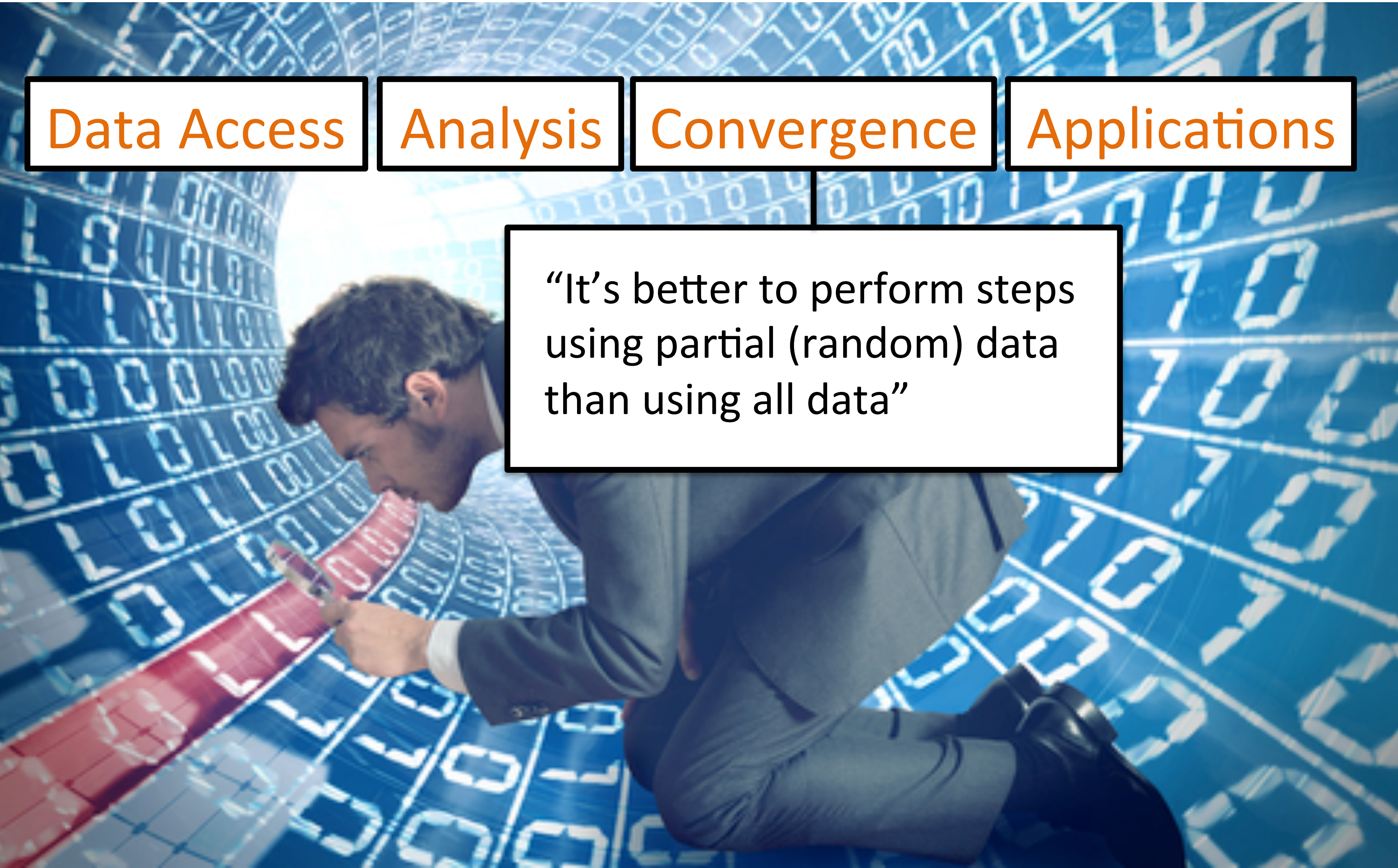
$$x_{k+1} = \arg\min_x \frac{1}{2}\|x - z_{k+1}\|^2 + \frac{1}{L}G(x)$$

Tool 4

Randomized Decomposition

"Doing many simple decisions is better than doing a few smart ones"

Why Randomize?



Stochastic Gradient Descent

H. Robbins and S. Monro
A Stochastic Approximation Method
Annals of Mathematical Statistics 22, pp. 400–407, 1951

The Problem

n is big

$$\min_{x \in \mathbb{R}^d} \left\{ F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

Stochastic Gradient Descent (SGD)

stepsize

- Update rule:

$$x_{k+1} = x_k - h_k \nabla f_i(x_k)$$

$\mathbb{E}[\nabla f_i(x)] = \nabla F(x)$

i = chosen uniformly at random

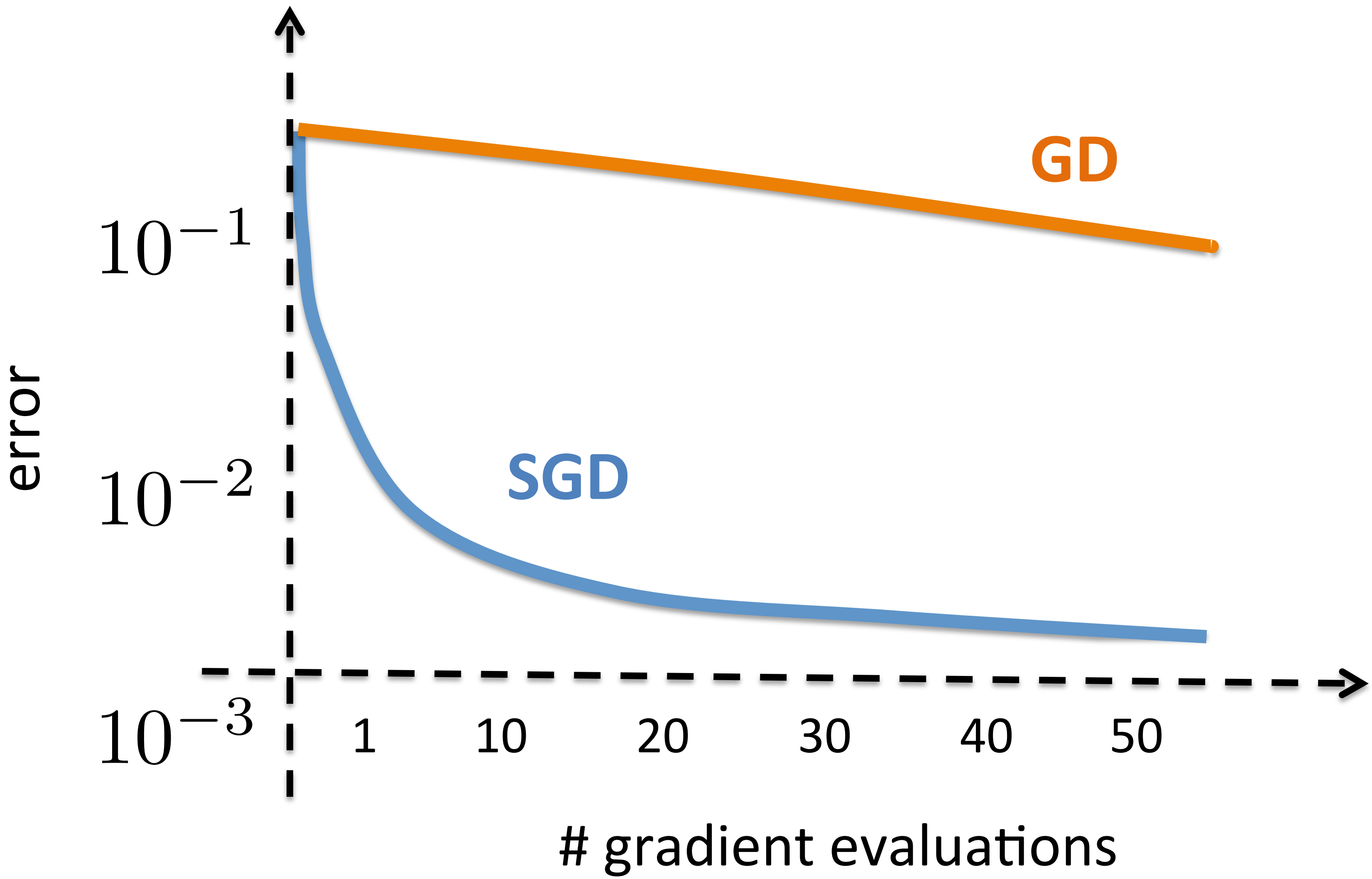
- Complexity:

$$\mathcal{O}\left(\frac{L}{\mu}\frac{1}{\epsilon}\right)$$

- Cost of a single iteration: 1

stochastic gradient evaluations

Stochastic Gradient Descent vs Gradient Descent



2014 OR Society Doctoral Prize

Randomized Coordinate Descent

P.R. and Martin Takáč
Iteration Complexity of Randomized Block Coordinate Descent Methods for Minimizing a Composite Function
Mathematical Programming 144(2), 1-38, 2014

INFORMS Computing Society Best Student Paper Prize (runner up), 2012

2014 OR Society Doctoral Prize

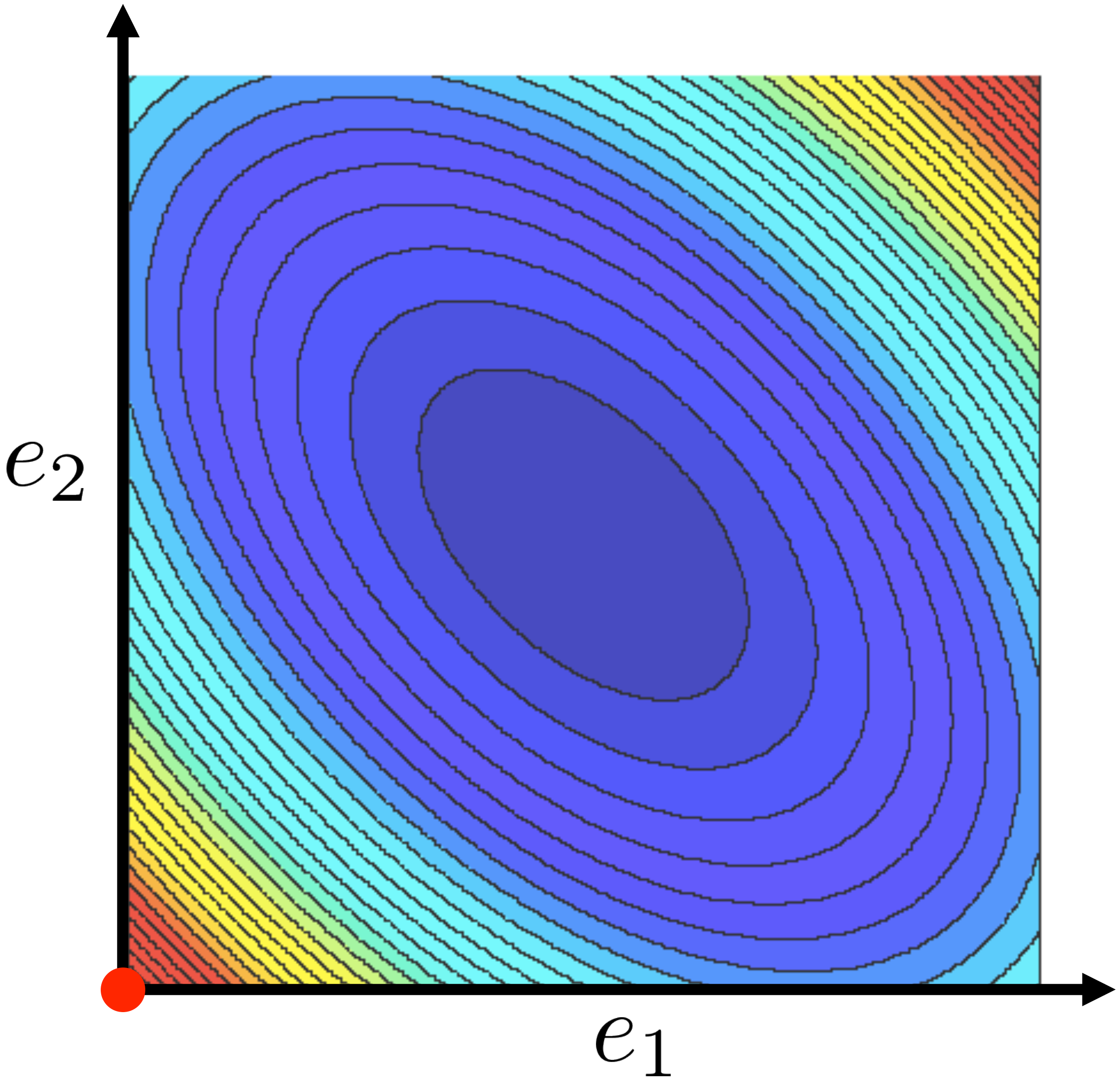
The Problem

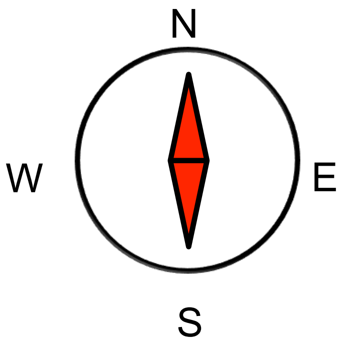
$$\min_{x \in \mathbb{R}^n} F(x)$$

Size of *x* is BIG

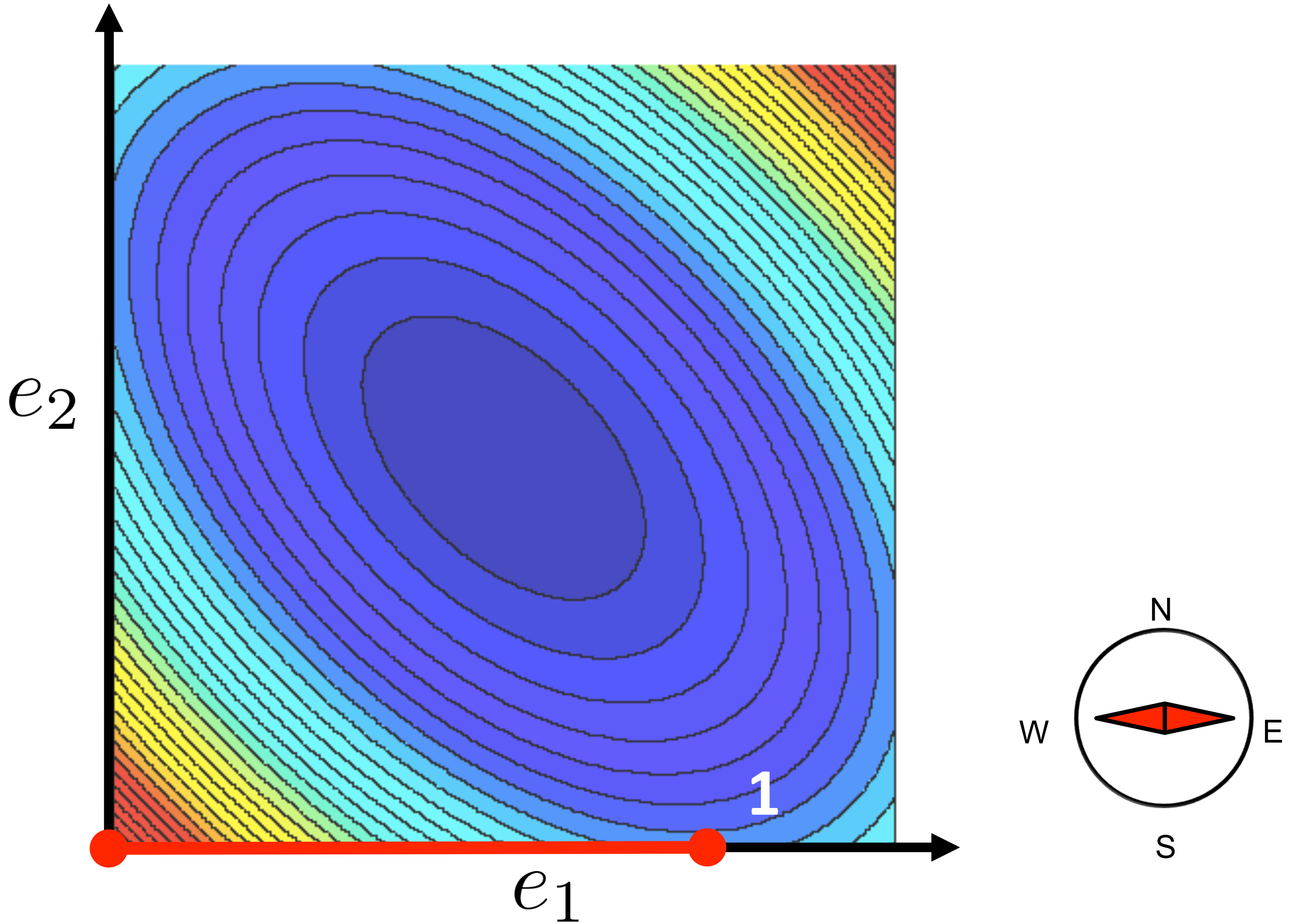
Convex, smooth

Randomized Coordinate Descent in 2D

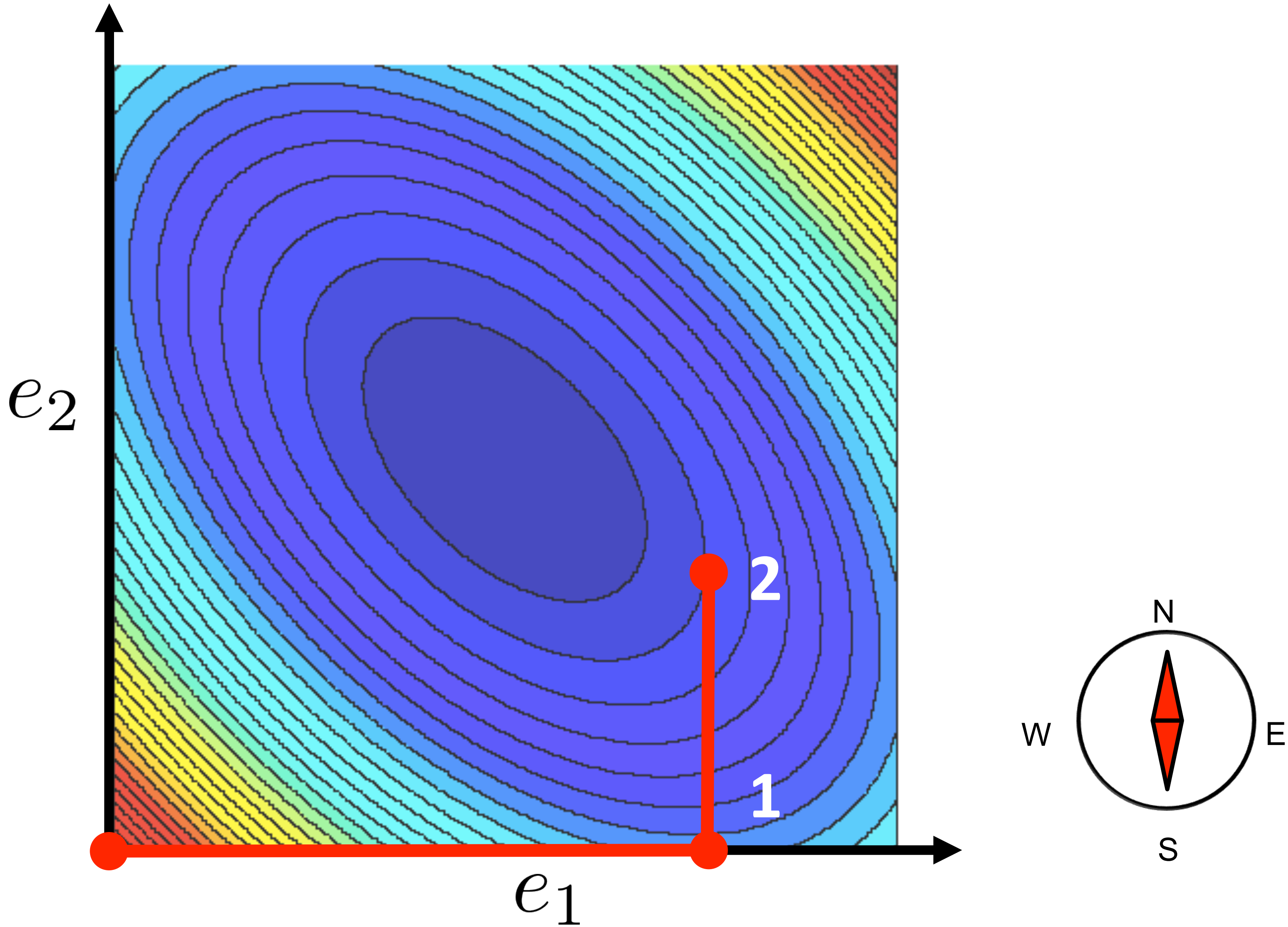




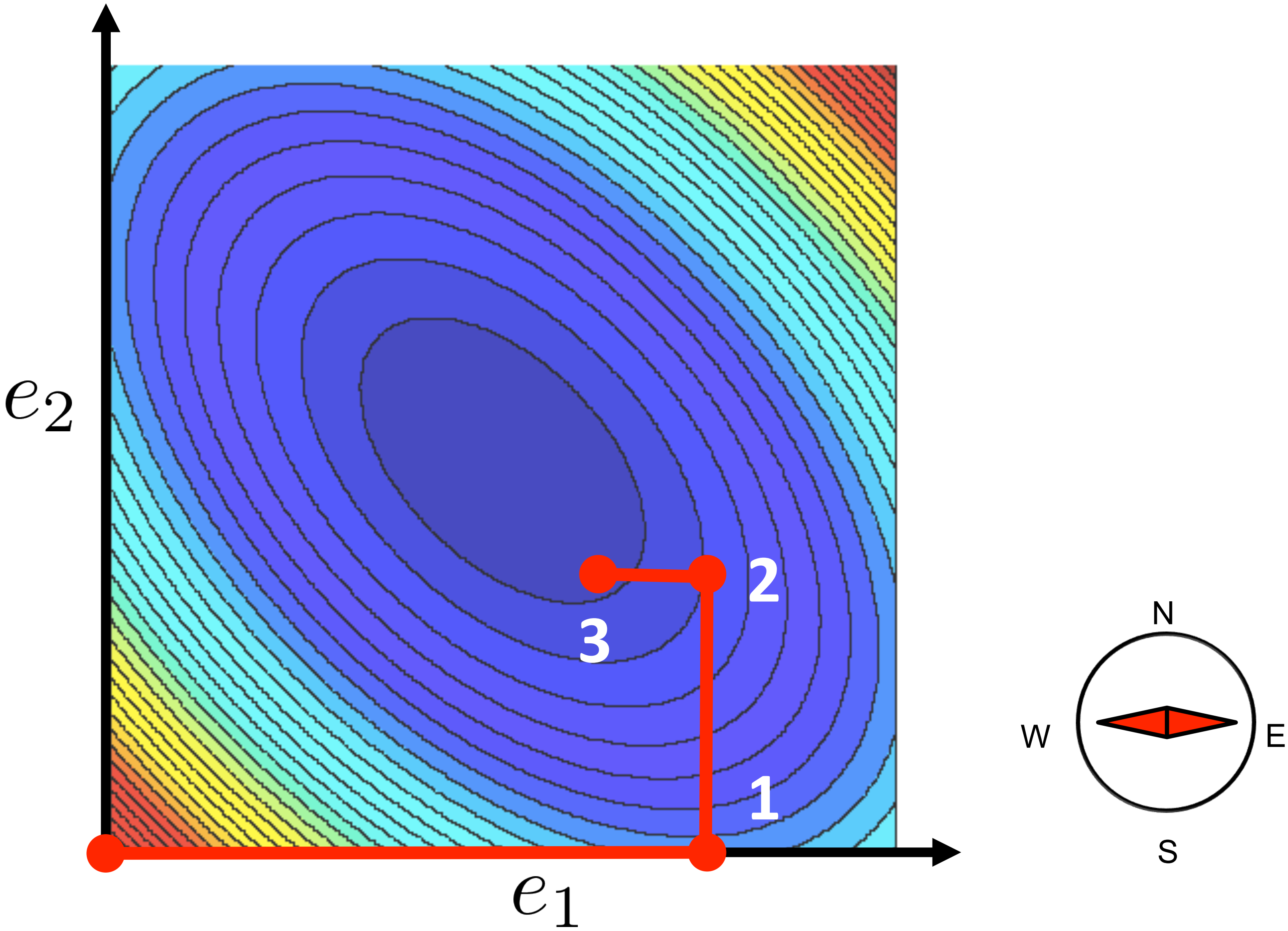
Randomized Coordinate Descent in 2D



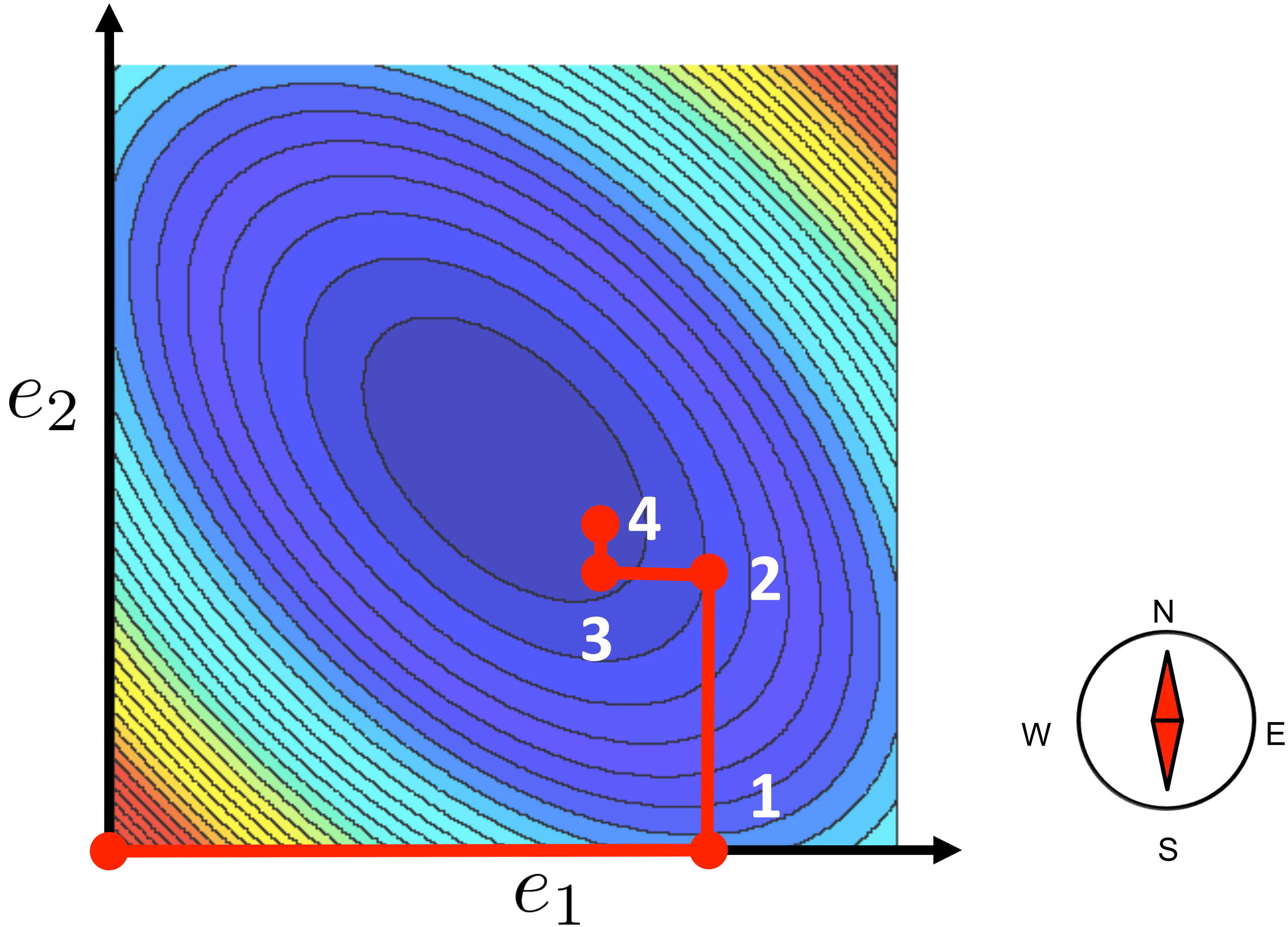
Randomized Coordinate Descent in 2D



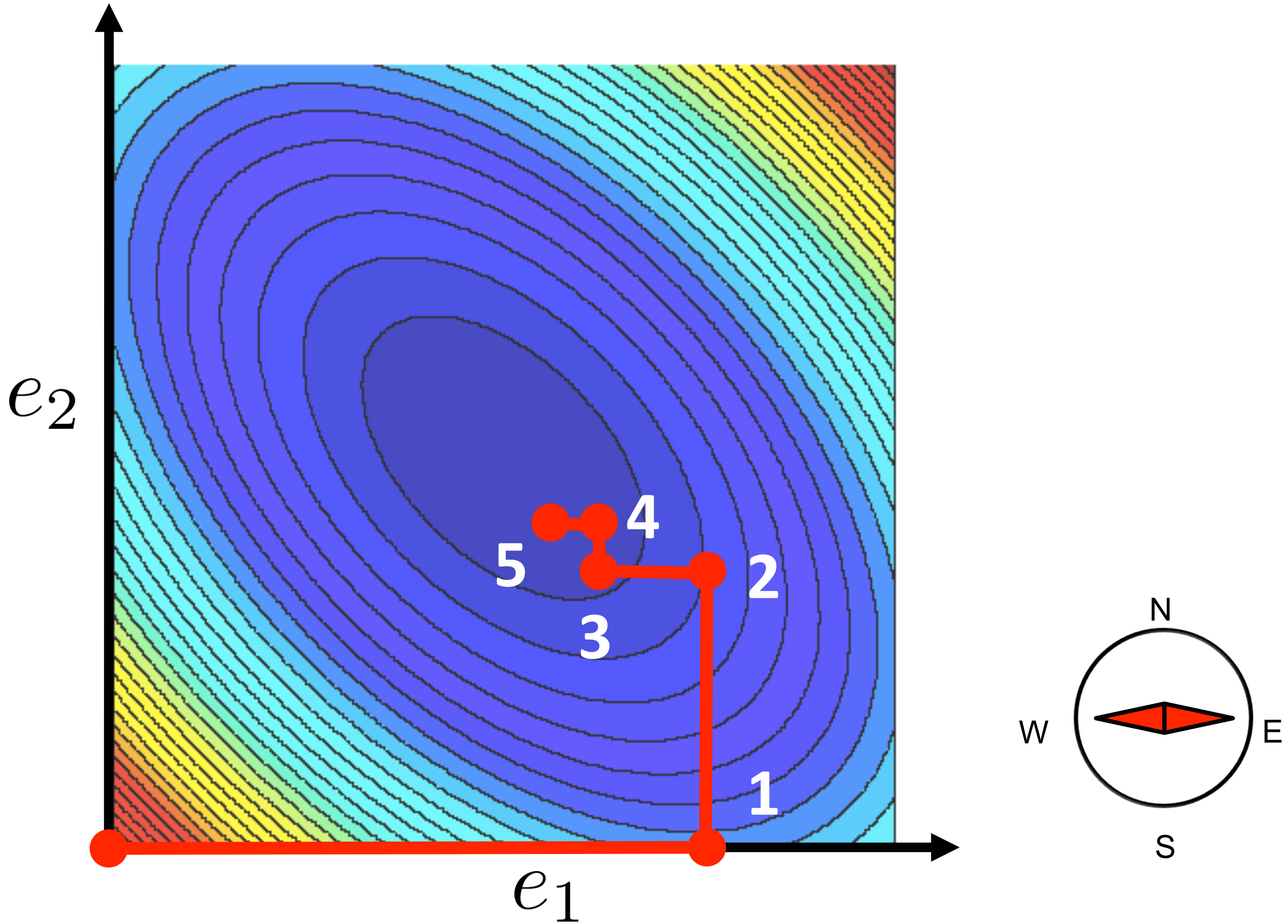
Randomized Coordinate Descent in 2D



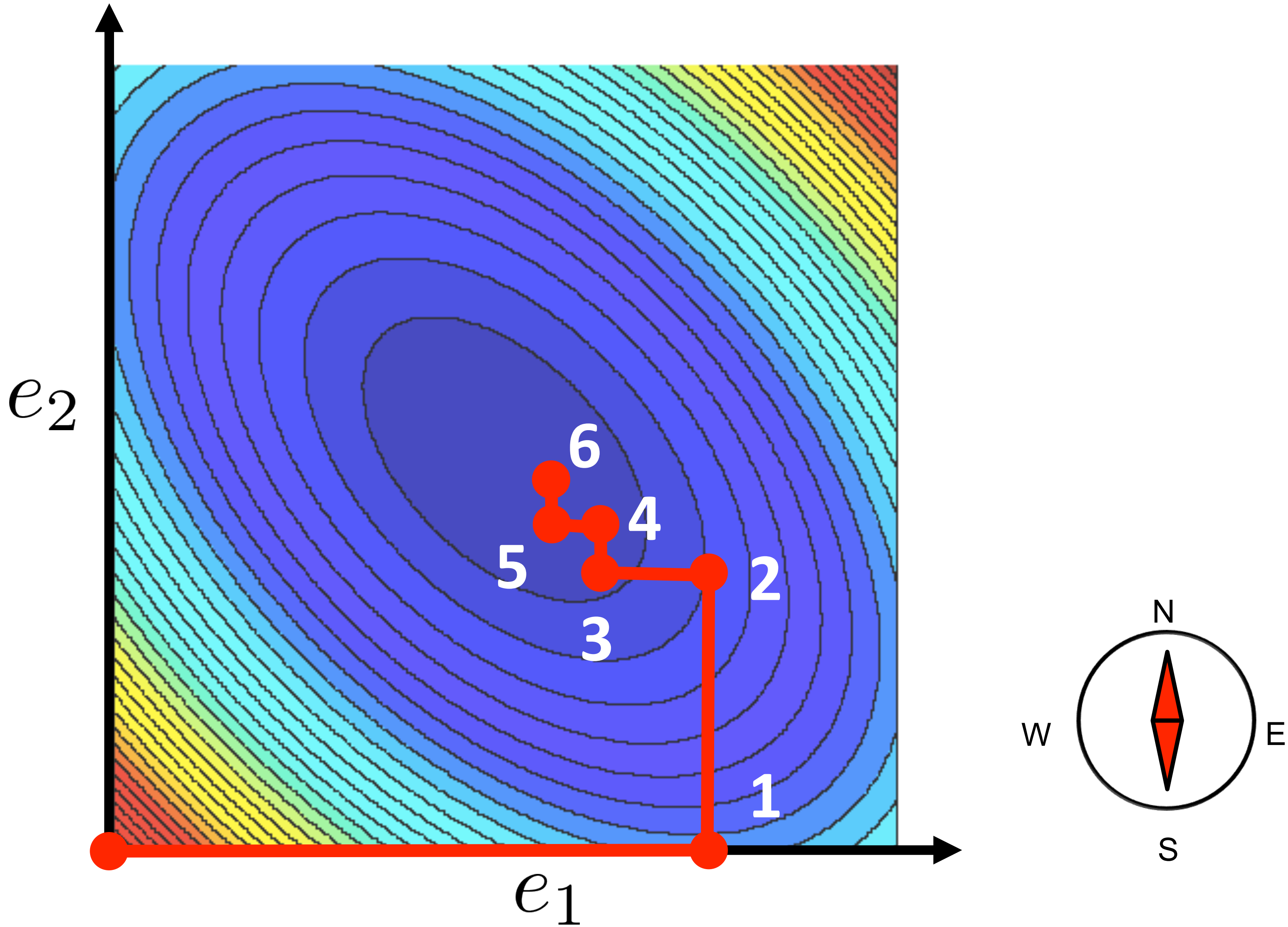
Randomized Coordinate Descent in 2D



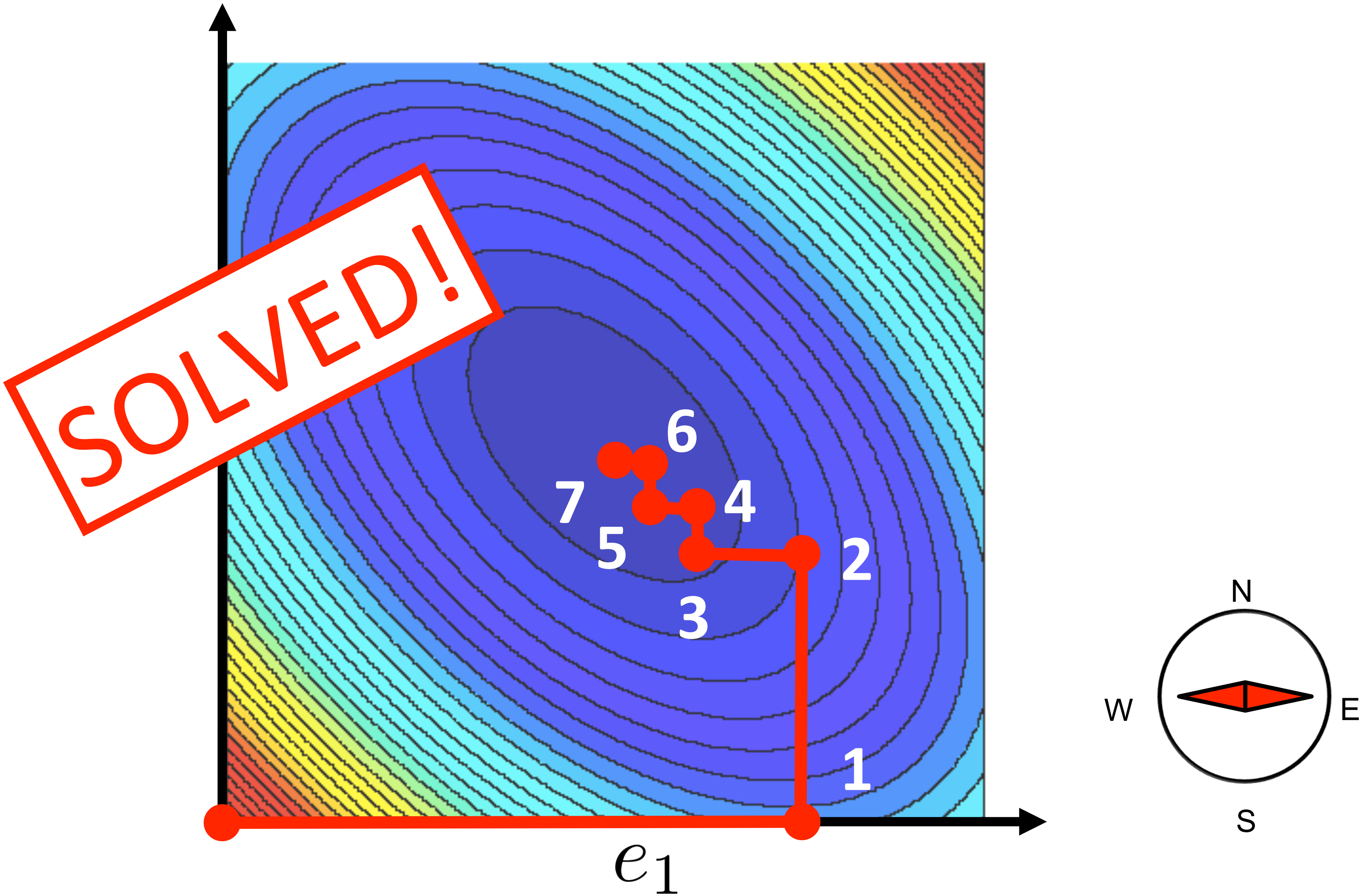
Randomized Coordinate Descent in 2D



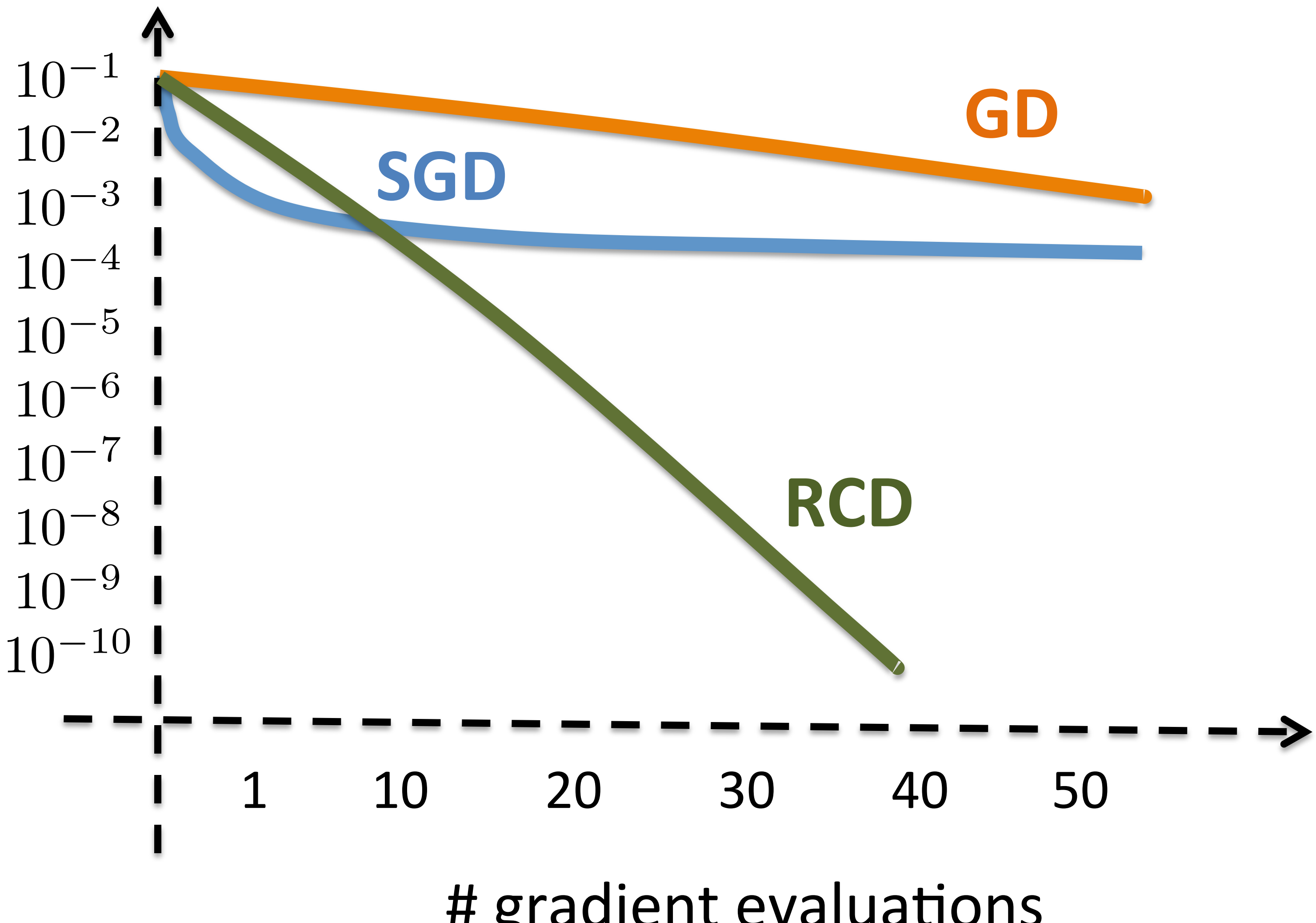
Randomized Coordinate Descent in 2D



Randomized Coordinate Descent in 2D



Randomized Coordinate Descent



1 Billion Rows & 100 Million Variables

$$A \in \mathbf{R}^{10^9 \times 10^8}$$

k/n	$F(x_k) - F^*$	# nonzeros in x_k	time [s]
0.01	$< 10^{18}$	18,486	1.32
9.35	$< 10^{14}$	99,837,255	1294.72
11.97	$< 10^{13}$	99,567,891	1657.32
14.78	$< 10^{12}$	98,630,735	2045.53
17.12	$< 10^{11}$	96,305,090	2370.07
20.09	$< 10^{10}$	86,242,708	2781.11
22.60	$< 10^{9}$	58,157,883	3128.49
24.97	$< 10^{8}$	19,926,459	3455.80
28.62	$< 10^{7}$	747,104	3960.96
31.47	$< 10^{6}$	266,180	4325.60
34.47	$< 10^{5}$	175,981	4693.44
36.84	$< 10^{4}$	163,297	5004.24
39.39	$< 10^{3}$	160,516	5347.71
41.08	$< 10^{2}$	160,138	5577.22
43.88	$< 10^{1}$	160,011	5941.72
45.94	$< 10^{0}$	160,002	6218.82
46.19	$< 10^{-1}$	160,001	6252.20
46.25	$< 10^{-2}$	160,000	6260.20
46.89	$< 10^{-3}$	160,000	6344.31
46.91	$< 10^{-4}$	160,000	6346.99
46.93	$< 10^{-5}$	160,000	6349.69

Tool 5

Parallelism

“Work on random subsets”

The Problem

$$\min_{x \in \mathbb{R}^n} F(x)$$

Size of *x* is BIG

Convex, smooth

Parallel Randomized Coordinate Descent

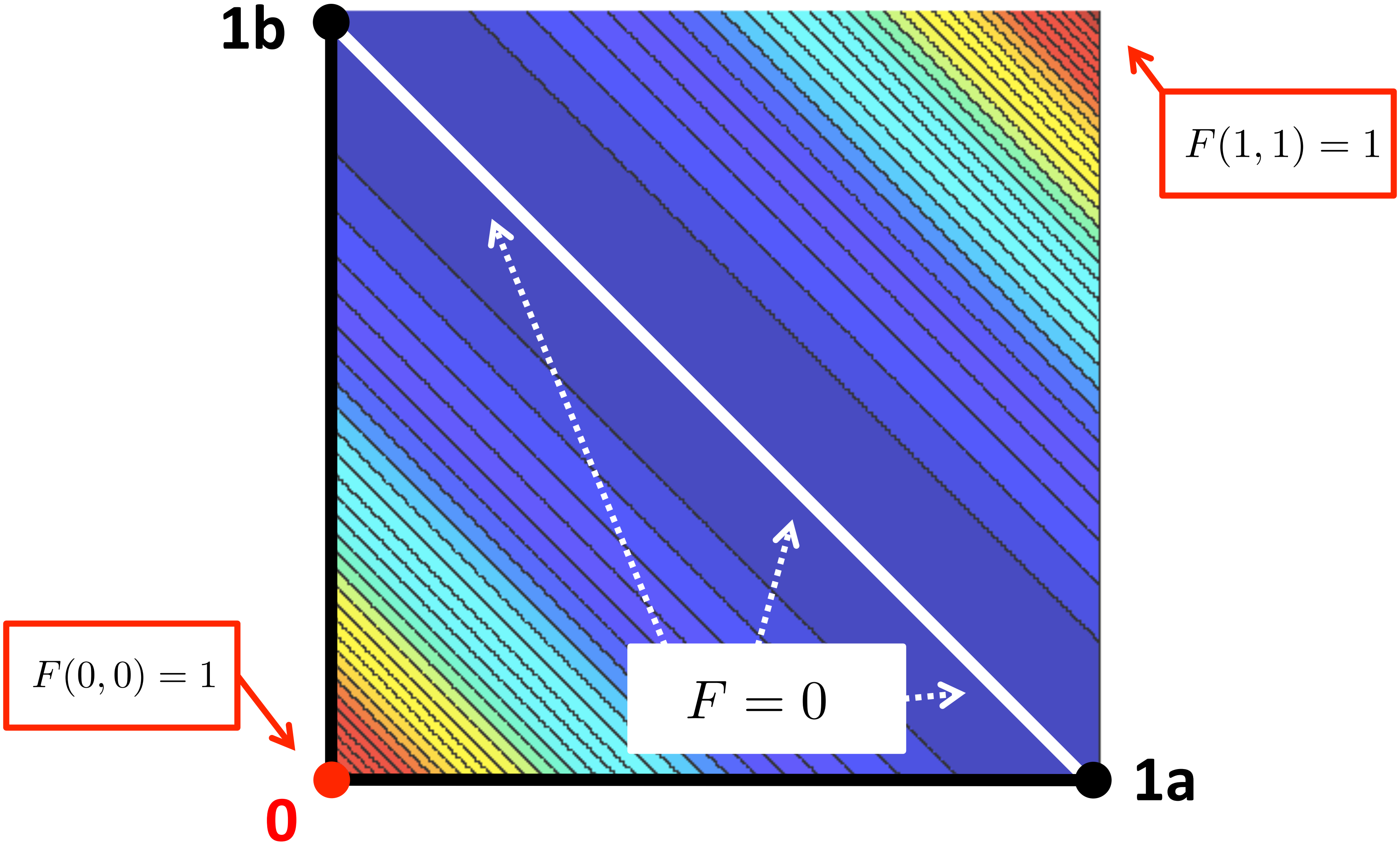
P.R. and Martin Takáč
Parallel Coordinate Descent Methods for Big Data Optimization
Mathematical Programming 156(1), 433-484, 2016

16th IMA Leslie Fox Prize (2nd), 2013

2014 OR Society Doctoral Prize

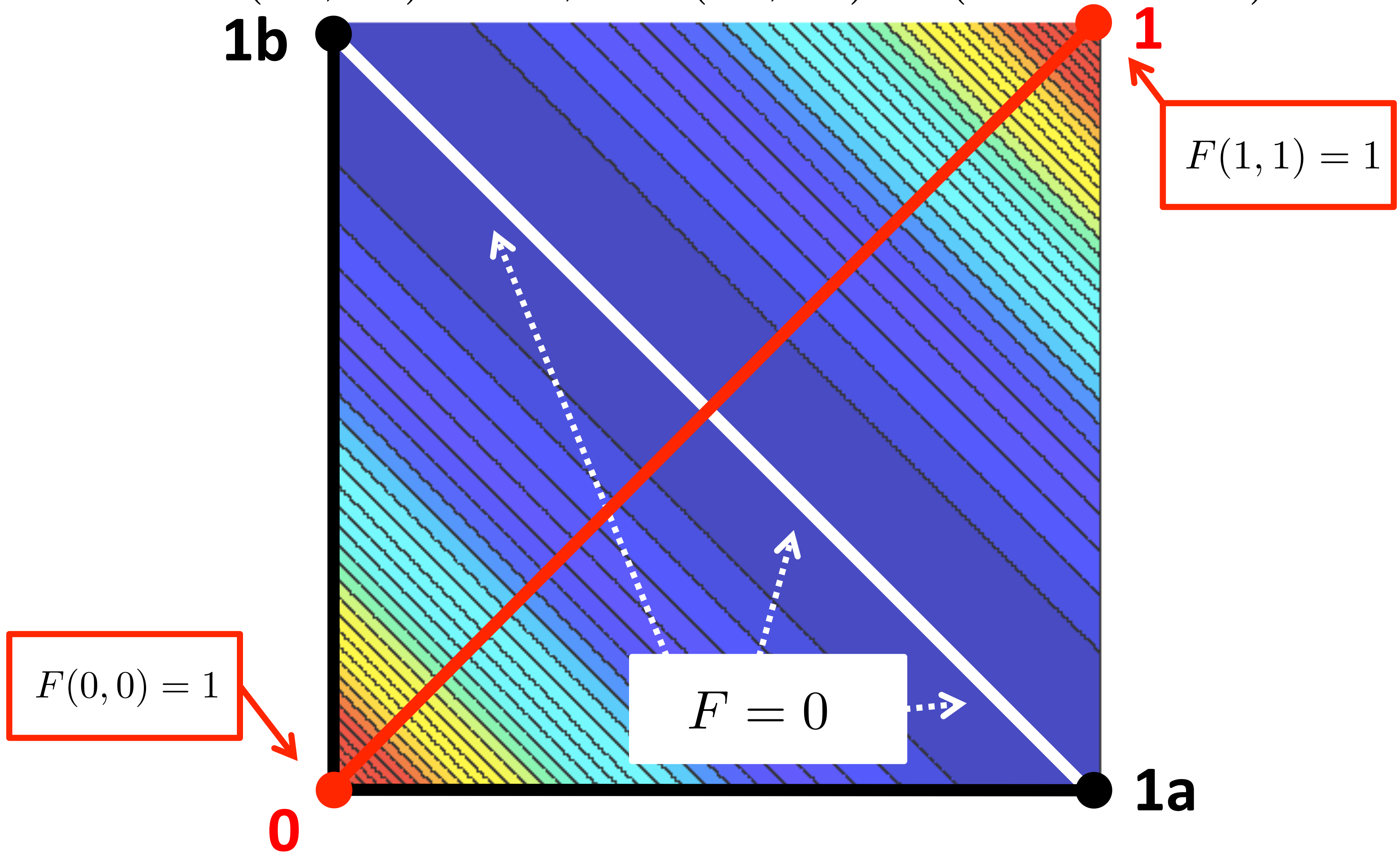
Additive Strategy

$$x = (x^1, x^2) \in \mathbb{R}^2, \quad F(x^1, x^2) = (x^1 + x^2 - 1)^2$$



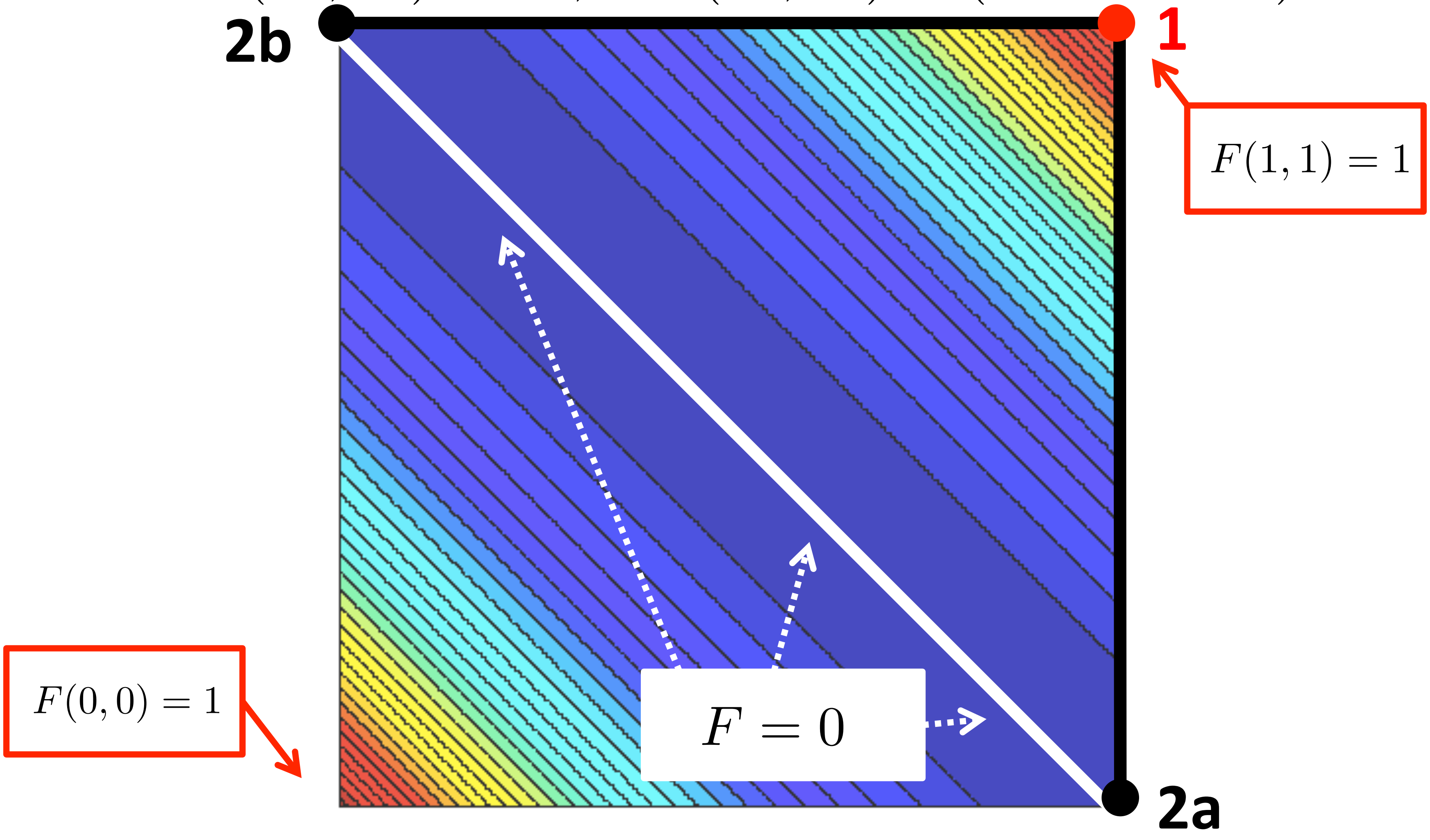
Additive Strategy

$$x = (x^1, x^2) \in \mathbb{R}^2, \quad F(x^1, x^2) = (x^1 + x^2 - 1)^2$$



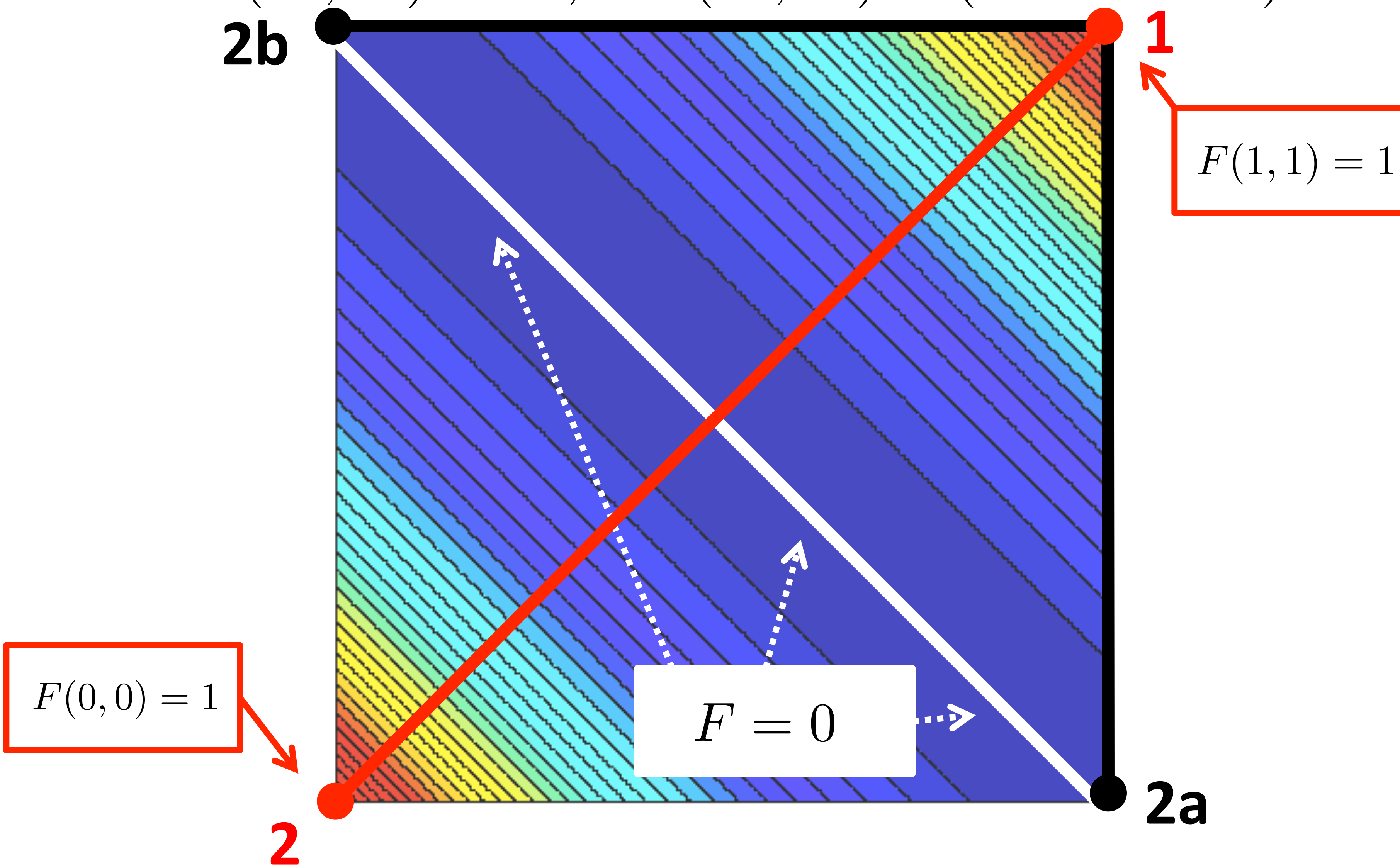
Additive Strategy

$$x = (x^1, x^2) \in \mathbb{R}^2, \quad F(x^1, x^2) = (x^1 + x^2 - 1)^2$$



Additive Strategy

$$x = (x^1, x^2) \in \mathbb{R}^2, \quad F(x^1, x^2) = (x^1 + x^2 - 1)^2$$



Additive Strategy

$$x = (x^1, x^2) \in \mathbb{R}^2, \quad F(x^1, x^2) = (x^1 + x^2 - 1)^2$$

$F(1,1) = 1$

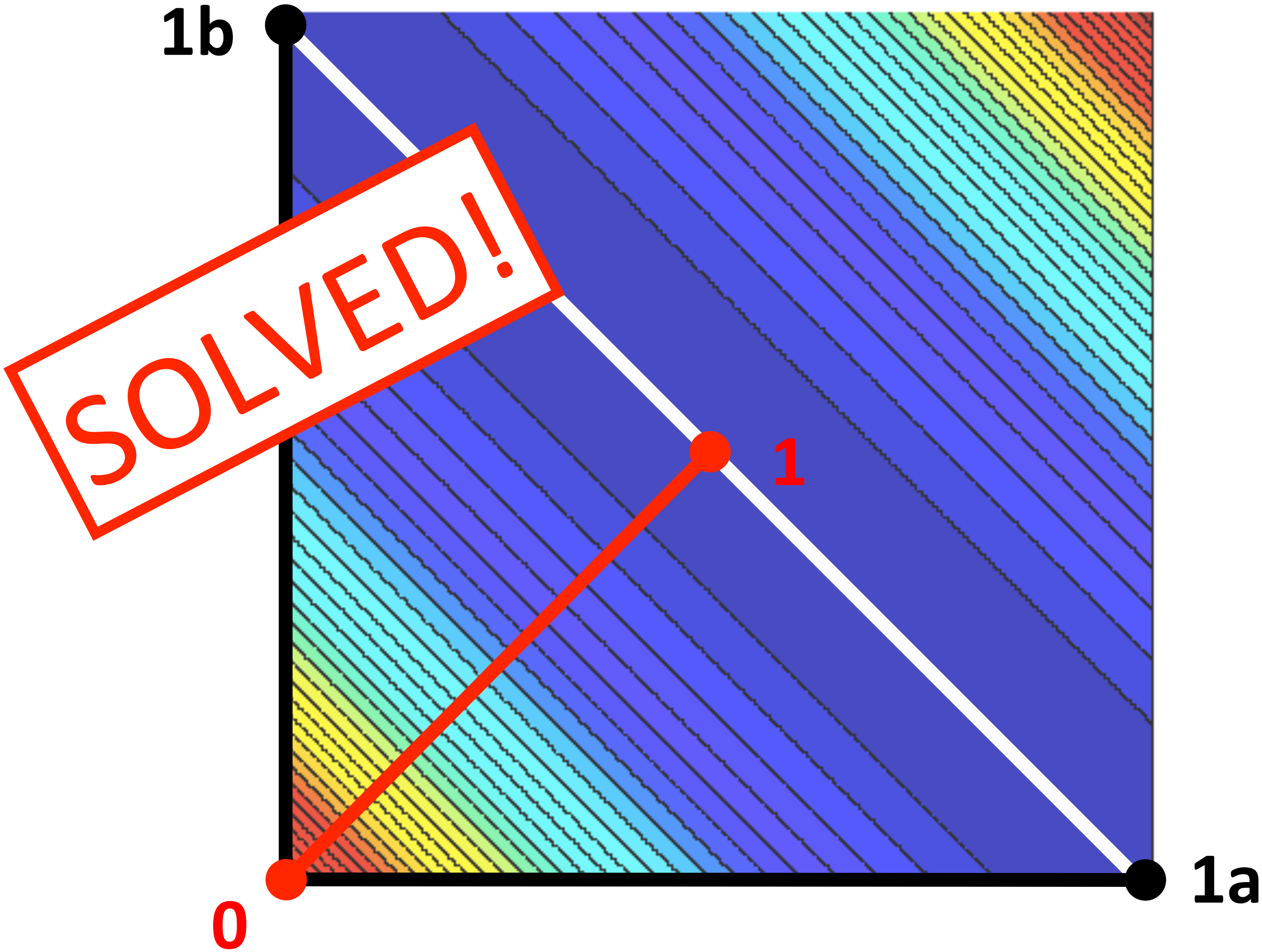
OOPS!

$F(0,0) = 1$

2

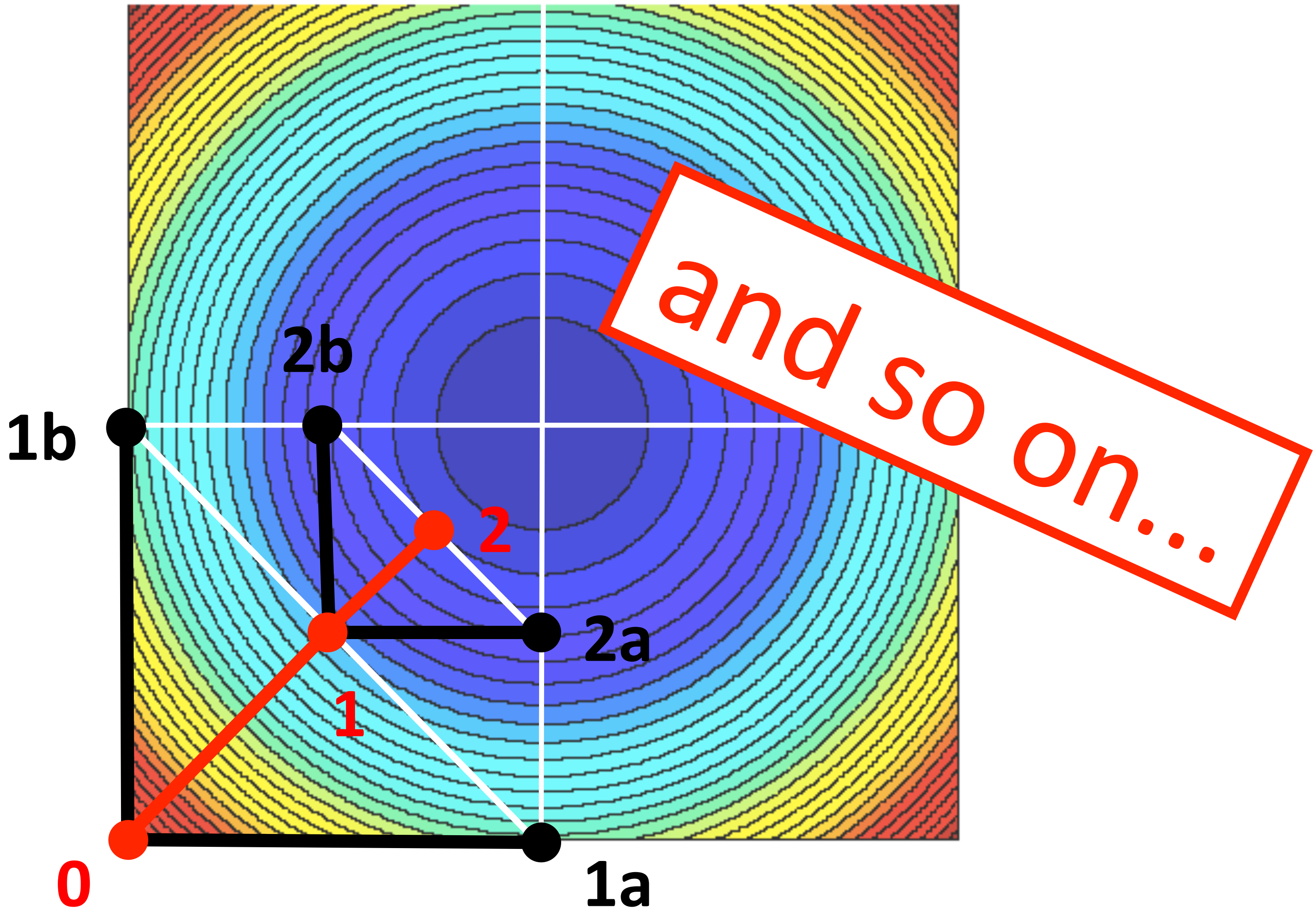
Averaging Strategy

$$x = (x^1, x^2) \in \mathbb{R}^2, \quad F(x^1, x^2) = (x^1 + x^2 - 1)^2$$



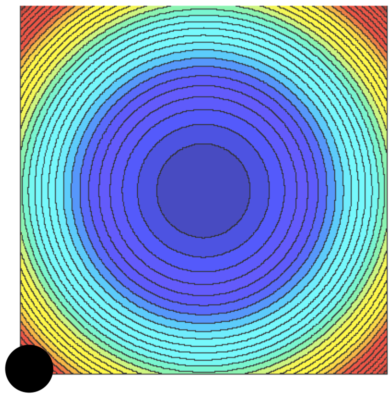
Averaging Can Be Bad, Too!

$$x = (x^1, x^2) \in \mathbb{R}^2, \quad F(x^1, x^2) = (x^1 - 1)^2 + (x^2 - 1)^2$$



Actually, Averaging Can Be Very Bad!

$$F(x) = (x^1 - 1)^2 + (x^2 - 1)^2 + \cdots + (x^n - 1)^2$$



$$x_0 = 0 \in \mathbb{R}^n \;\Rightarrow\; F(x_0) = n$$

BAD!!!

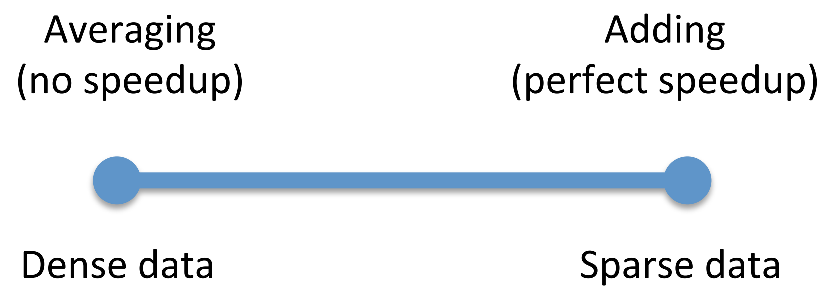
$$k \geq \frac{n}{2} \log\left(\frac{n}{\epsilon}\right)$$

$$F(x_k) = n\left(1 - \frac{1}{n}\right)^{2k} \leq \epsilon$$

WANT

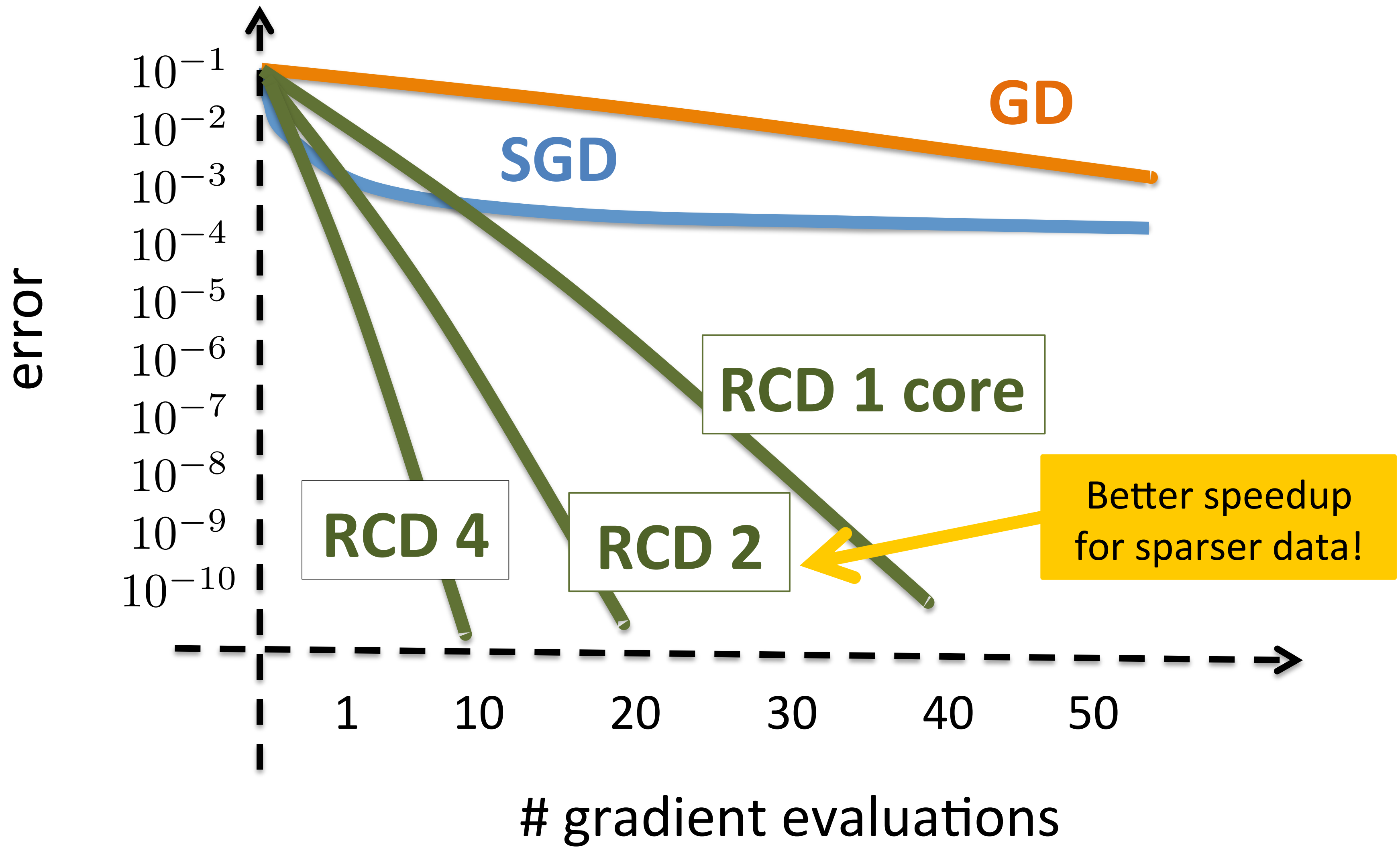
How to Combine the Updates?

- We should do data-dependent combination of the results obtained in parallel
- There is rich theory for this now



Zheng Qu and P.R.
Coordinate Descent with Arbitrary Sampling II: Expected Separable Overapproximation
Optimization Methods and Software 31(5), 858-884, 2016

Performance



Problem with 1 Billion Variables

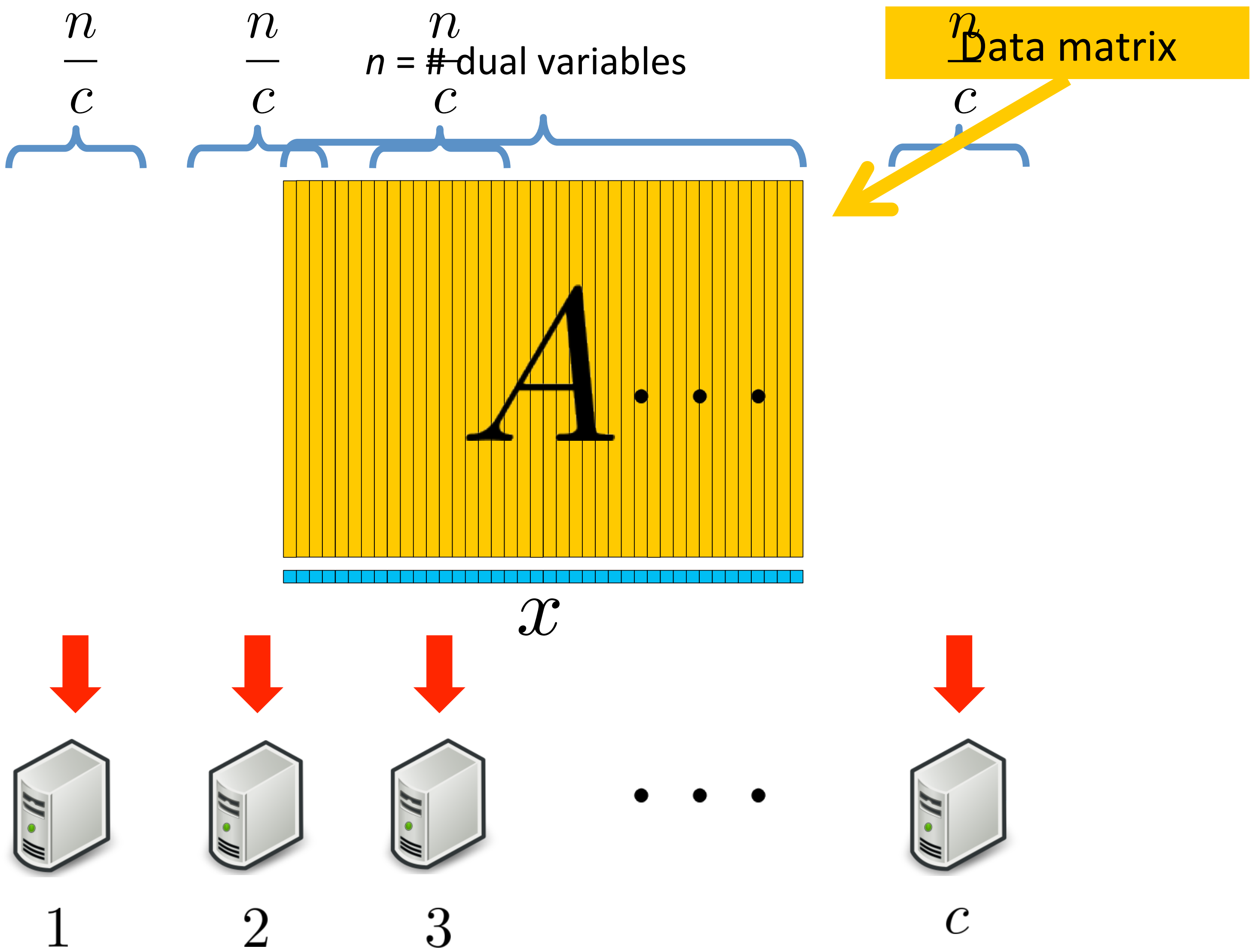
$(k \cdot \tau)/n$	$F(x_k) - F^*$ 1 core	$F(x_k) - F^*$ 8 cores	$F(x_k) - F^*$ 16 cores	Elapsed Time 1 core	Elapsed Time 8 cores	Elapsed Time 16 cores
0	6.27e+22	6.27e+22	6.27e+22	0.00	0.00	0.00
1	2.24e+22	2.24e+22	2.24e+22	0.89	0.11	0.06
2	2.25e+22	3.64e+19	2.24e+22	1.97	0.27	0.14
3	1.15e+20	1.94e+19	1.37e+20	3.20	0.43	0.21
4	5.25e+19	1.42e+18	8.19e+19	4.28	0.58	0.29
5	1.59e+19	1.05e+17	3.37e+19	5.37	0.73	0.37
6	1.97e+18	1.17e+16	1.33e+19	6.64	0.89	0.45
7	2.40e+16	3.18e+15	8.39e+17	7.87	1.04	0.53
⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	3.49e+02	4.11e+01	3.68e+03	31.71	3.99	2.02
27	1.92e+02	5.70e+00	7.77e+02	33.00	4.14	2.10
28	1.07e+02	2.14e+00	6.69e+02	34.23	4.30	2.17
29	6.18e+00	2.35e-01	3.64e+01	35.31	4.45	2.25
30	4.31e+00	4.03e-02	2.74e+00	36.60	4.60	2.33
31	6.17e-01	3.50e-02	6.20e-01	37.90	4.75	2.41
32	1.83e-02	2.41e-03	2.34e-01	39.17	4.91	2.48
33	3.80e-03	1.63e-03	1.57e-02	40.39	5.06	2.56
34	7.28e-14	7.46e-14	1.20e-02	41.47	5.21	2.64
35	-	-	1.23e-03	-	-	2.72
36	-	-	3.99e-04	-	-	2.80
37	-	-	7.46e-14	-	-	2.87

Tool 6

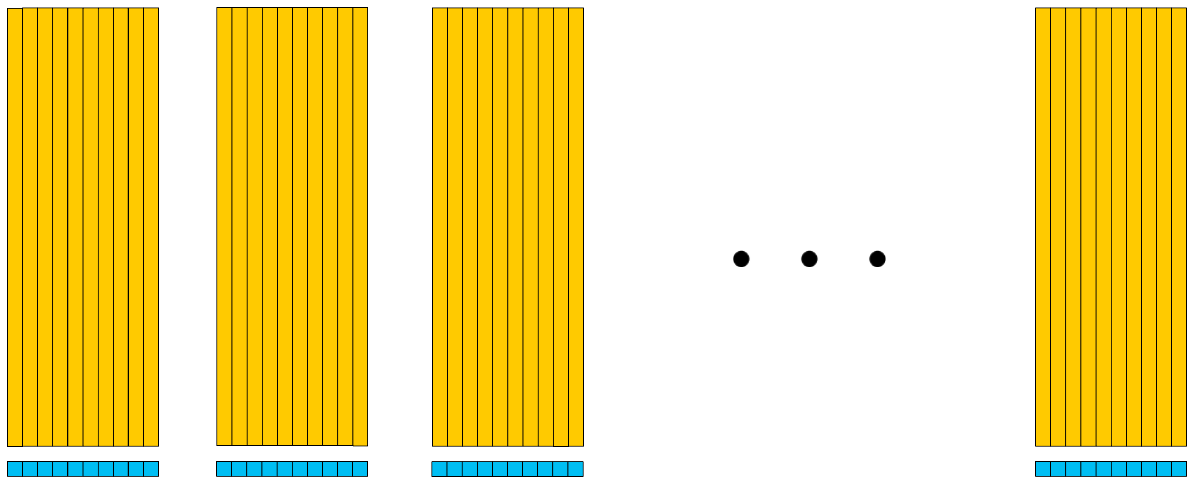
Distributed Computation

“Communication hurts”

Distribution of Data

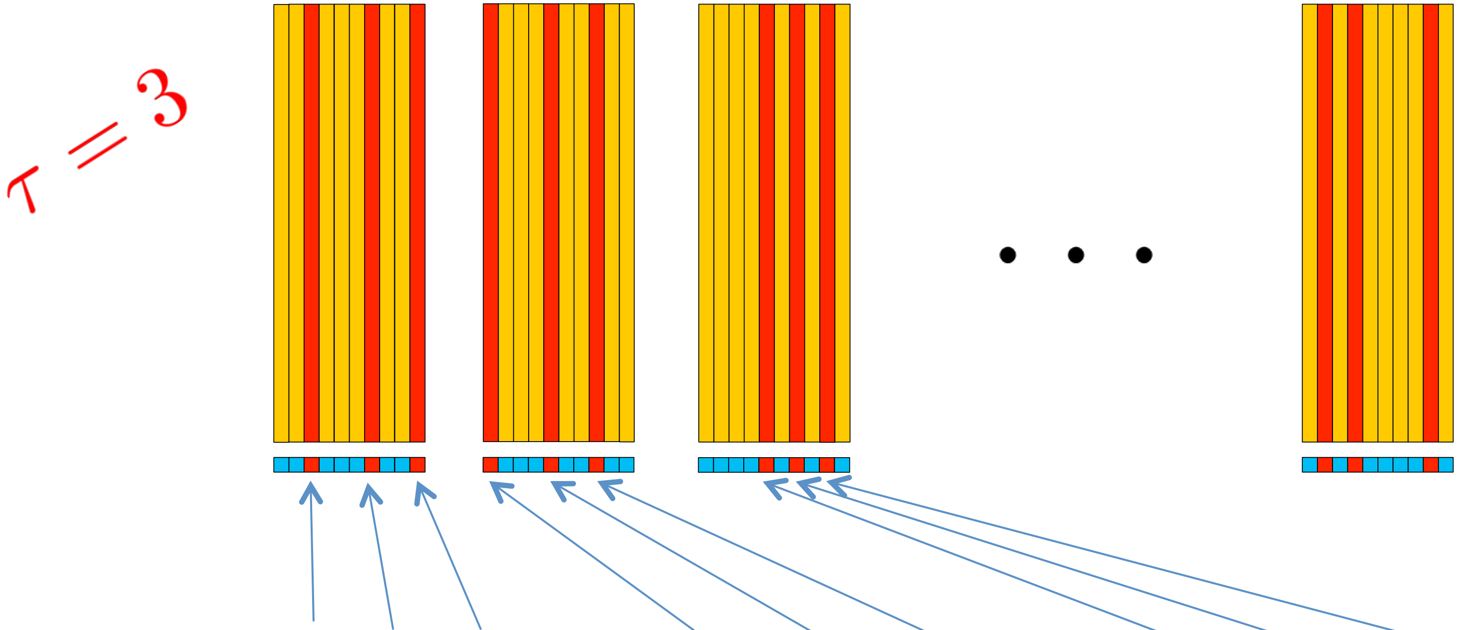


Distributed sampling



Distributed sampling

Each computer (node) independently pick τ variables from those it owns, uniformly at random



$$S_k = \{3, 7, 10\} \cup \{11, 15, 18\} \cup \{25, 27, 29\} \cup \cdots$$

Random set of variables

Also see: CoCoA+ [Ma, Smith, Jaggi et al 15]

There is Theory for this…

Key: Get the right stepsize parameters *v*

The leading term in the complexity bound then is:

$$\max_i \left(\frac{1}{p_i} + \frac{v_i}{p_i \lambda \gamma n} \right)$$

$$||$$

$$\frac{n}{c\tau} + \frac{\text{Something that looks complicated}}{\lambda \gamma c \tau}$$

$$||$$

$$\frac{n}{c\tau} + \max_i \frac{\lambda_{\max}\left(\sum_{j=1}^{d}\left(1 + \frac{(\tau-1)(\omega_j-1)}{\max\{n/c-1,1\}} + \left(\frac{\tau c}{n} - \frac{\tau-1}{\max\{n/c-1,1\}}\right)\frac{\omega_j'-1}{\omega_j'}\omega_j\right) A_{ji}^\top A_{ji}\right)}{\lambda \gamma c \tau}$$

Experiment

Machine: 128 nodes of Hector Supercomputer (4096 cores)

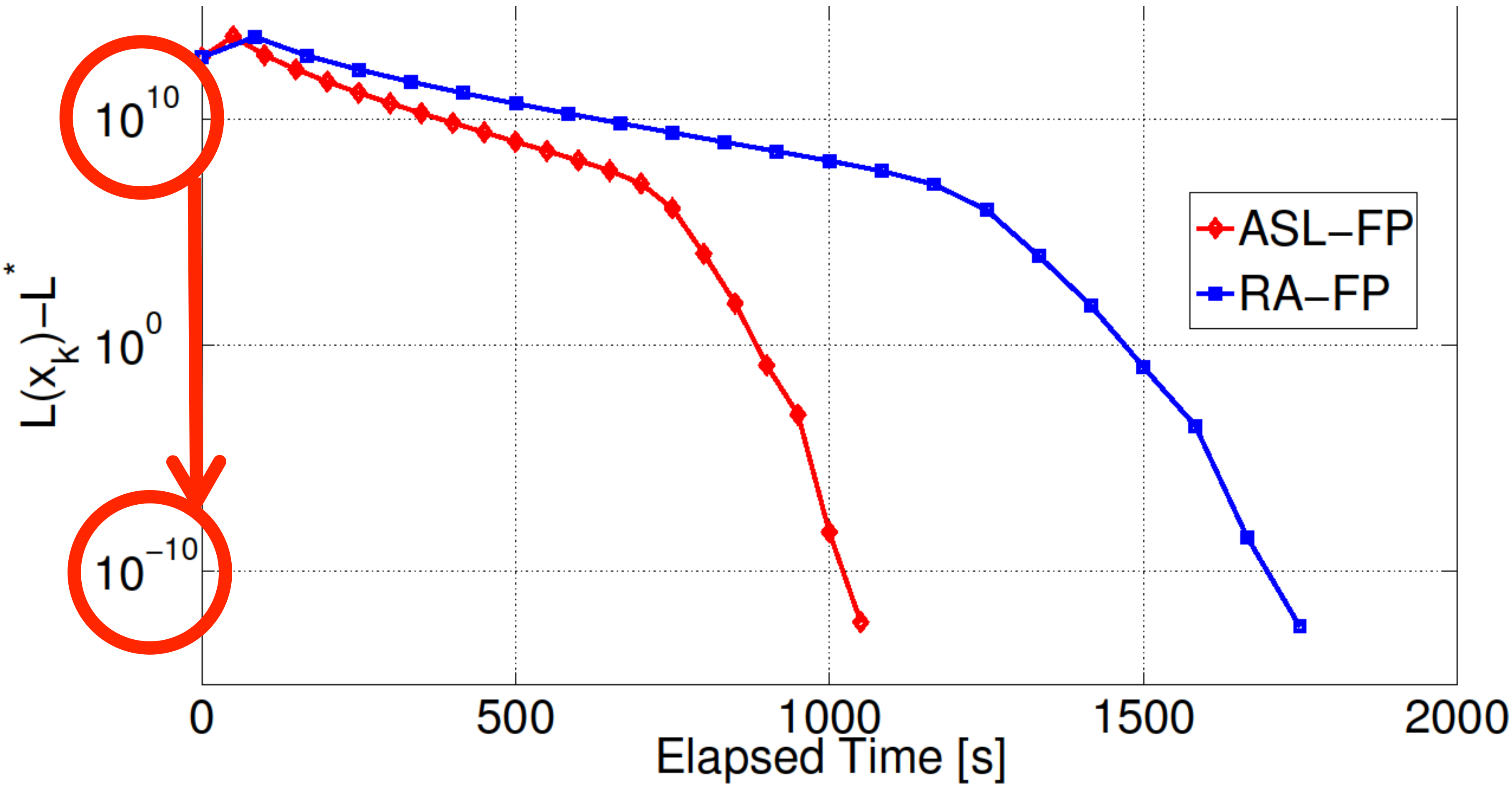
Problem: LASSO, n = 1 billion, d = 0.5 billion, 3 TB

HYDRA

2014 OR Society Doctoral Prize

P.R. and Martin Takáč
Distributed Coordinate Descent for Learning with Big Data
Journal of Machine Learning Research 17:1-25, 2016

LASSO: 3TB data + 128 nodes



Experiment

Machine: 128 nodes of Archer Supercomputer

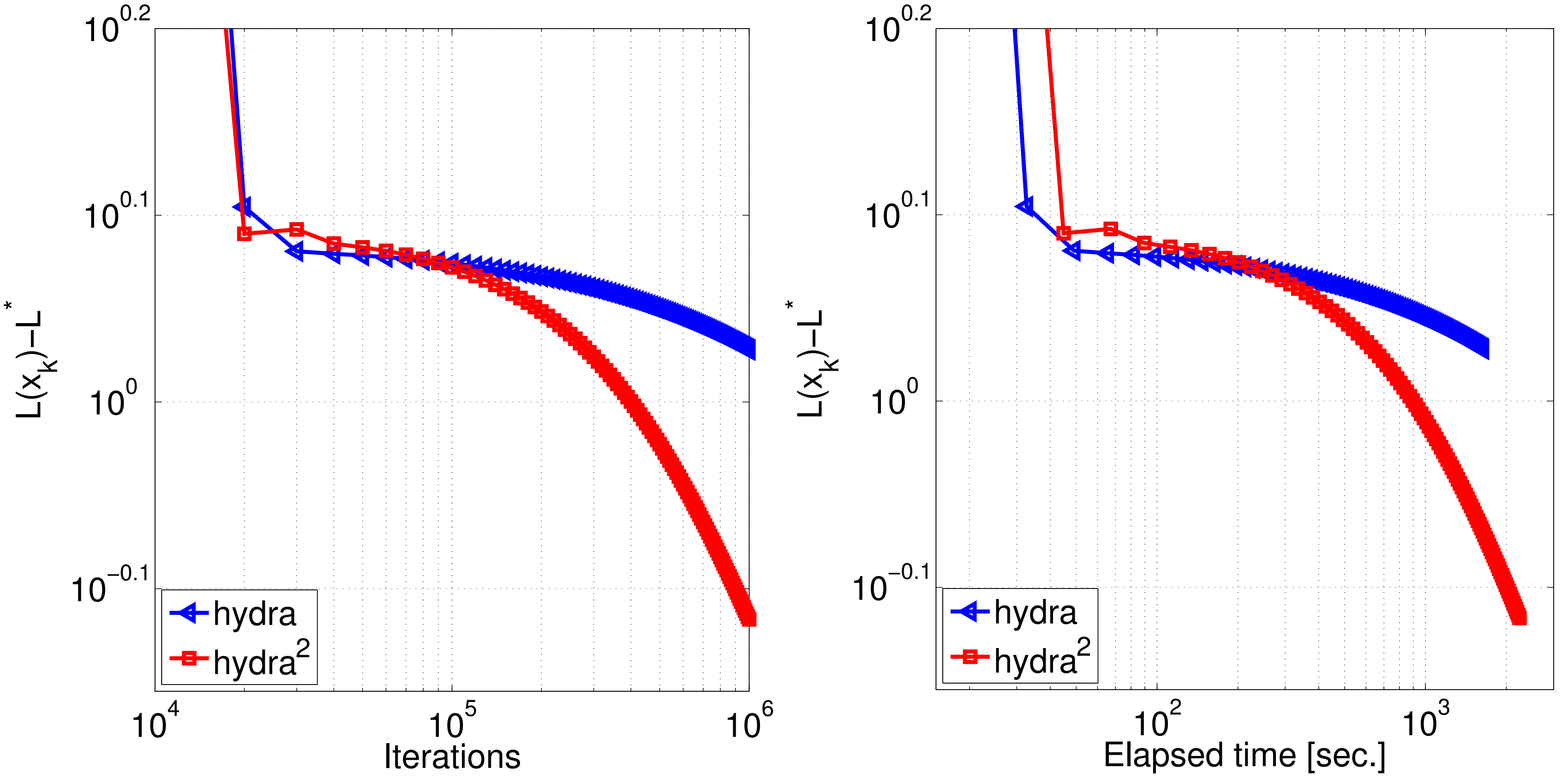
Problem: LASSO, n = 5 million, d = 50 billion, 5 TB (60,000 nnz per row of A)

HYDRA²

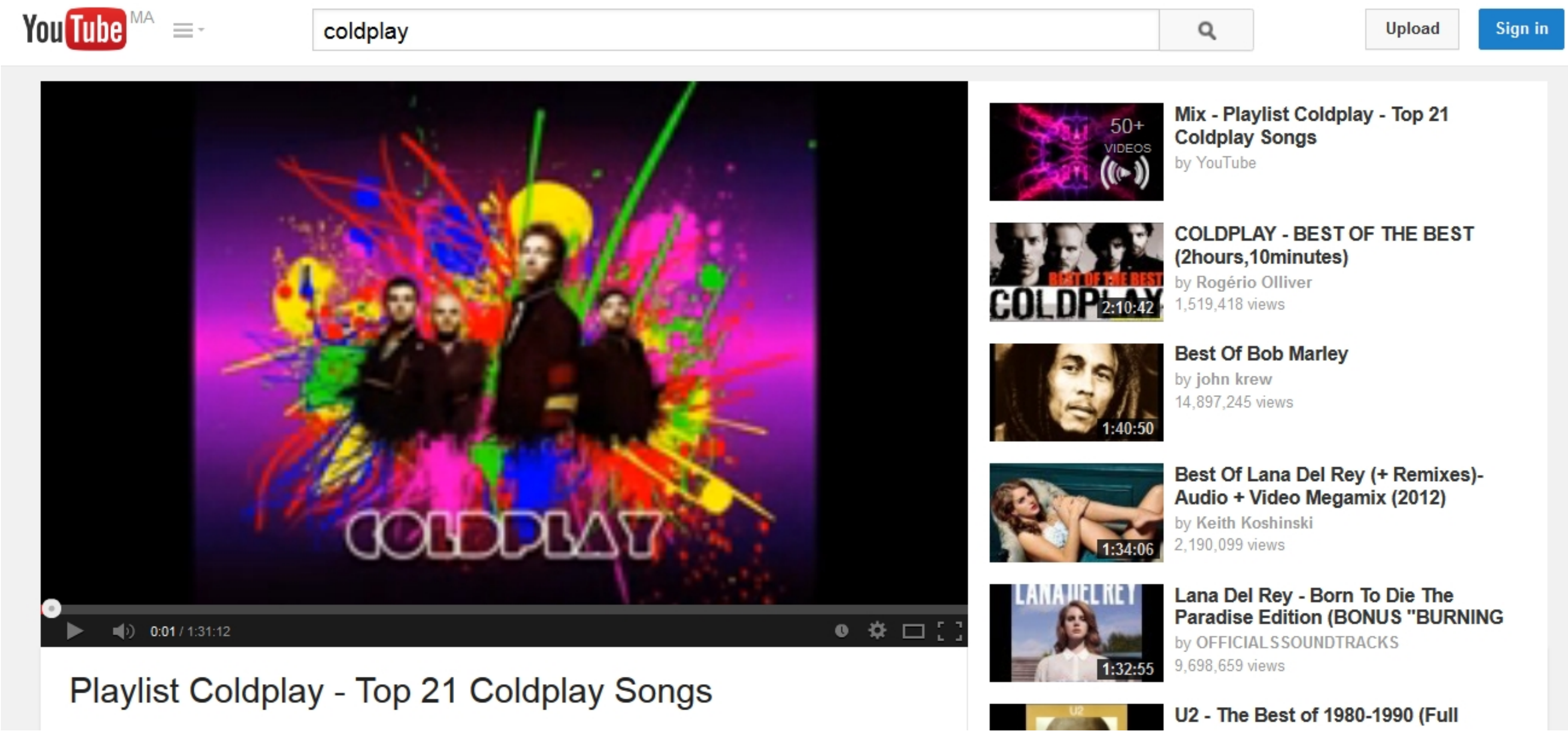
PDF

Olivier Fercoq, Zheng Qu, P.R. and Martin Takáč. **Fast Distributed Coordinate Descent for Minimizing Non-strongly Convex Losses** *IEEE Int. Workshop on Machine Learning for Signal Processing*, 2014

LASSO: 5TB data (d = 50 billion) 128 nodes



Used in YouTube



Tool 7

Importance Sampling

“Sample more important data more often”

P.R. and Martin Takáč

On Optimal Probabilities in Stochastic Coordinate Descent Methods

Optimization Letters 10(6), 1233-1243, 2015

2014 OR Society Doctoral Prize

The Problem

$$\min_{x \in \mathbb{R}^n} F(x)$$

Smooth and strongly convex

Really, really large

SYNC

Arbitrary Sampling:

$S_k \subseteq \{1, 2, \ldots, n\}$

Choose a random set S_k of coordinates

For $i \in S_k$ do

$$x_{k+1}^i \leftarrow x_k^i - \frac{1}{v_i} \nabla_i F(x_k)$$

For $i \notin S_k$ do

$$x_{k+1}^i \leftarrow x_k^i$$

Partial derivative

Stepsize parameter

Complexity Theorem

$$k \geq \left(\max_i \frac{v_i}{p_i \mu} \right) \log \left(\frac{F(x_0) - F(x_*)}{\epsilon \rho} \right)$$

$p_i = \mathbb{P}(i \in S_k)$

strong convexity constant of F

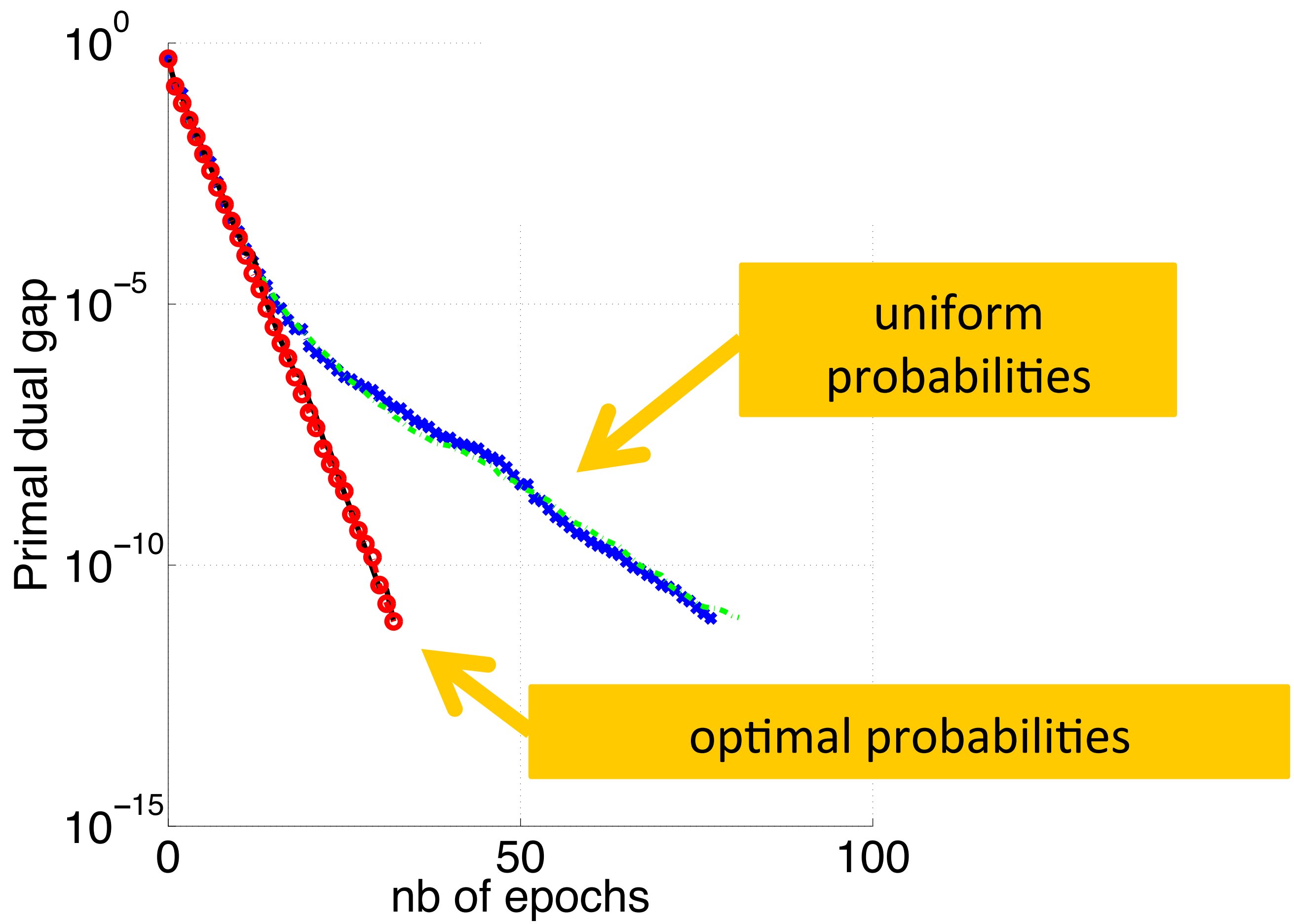
$$\mathbb{P}\left(F(x_k) - F(x_*) \leq \epsilon\right) \geq 1 - \rho$$

Uniform vs Optimal Sampling

$$p_i = \frac{1}{n} \quad \Rightarrow \quad \max_i \frac{v_i}{p_i \mu} = \frac{n \max_i v_i}{\mu}$$

$$p_i = \frac{v_i}{\sum_i v_i} \quad \Rightarrow \quad \max_i \frac{v_i}{p_i \mu} = \frac{\sum_i v_i}{\mu}$$

Uniform vs Optimal Sampling



$$\text{Data} = \texttt{cov1}, \quad n = 522{,}911, \quad \mu = 10^{-6}$$

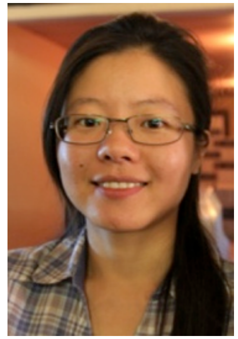
Part 4
Conclusion

Conclusion

- Data, data science, machine learning, ATI
- Data science applications
  - structure of the objective (simple, data-defined)
  - imaging, empirical risk minimization, truss topology design, spam filtering, …
- Outlined a few key tools/tricks developed for big data optimization

Martin Takáč
(Lehigh)

Jakub Mareček
(IBM)



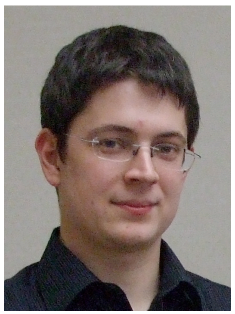
Zheng Qu
(Hong Kong)

Olivier Fercoq
(Telecom ParisTech)

Rachael Tappenden
(Johns Hopkins)

Robert M Gower
(Edinburgh)

Virginia Smith
(Berkeley)



Jakub Konečný
(Edinburgh)

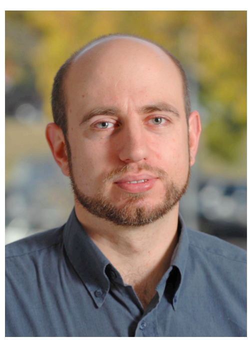
Jie Liu
(Lehigh)

Michael Jordan
(Berkeley)

Dominik Csba
(Edinburgh)

Tong Zhang
(Rutgers & Baidu)

Zeyuan Allen-Zhu
(Princeton)



Nati Srebro
(TTI Chicago)

Donald Goldfarb
(Columbia)

Chenxin Ma
(Lehigh)

Martin Jaggi
(ETH Zurich)