DATA FUSION

TENSOR FACTORIZATIONS BY COMPLEX OPTIMIZATION

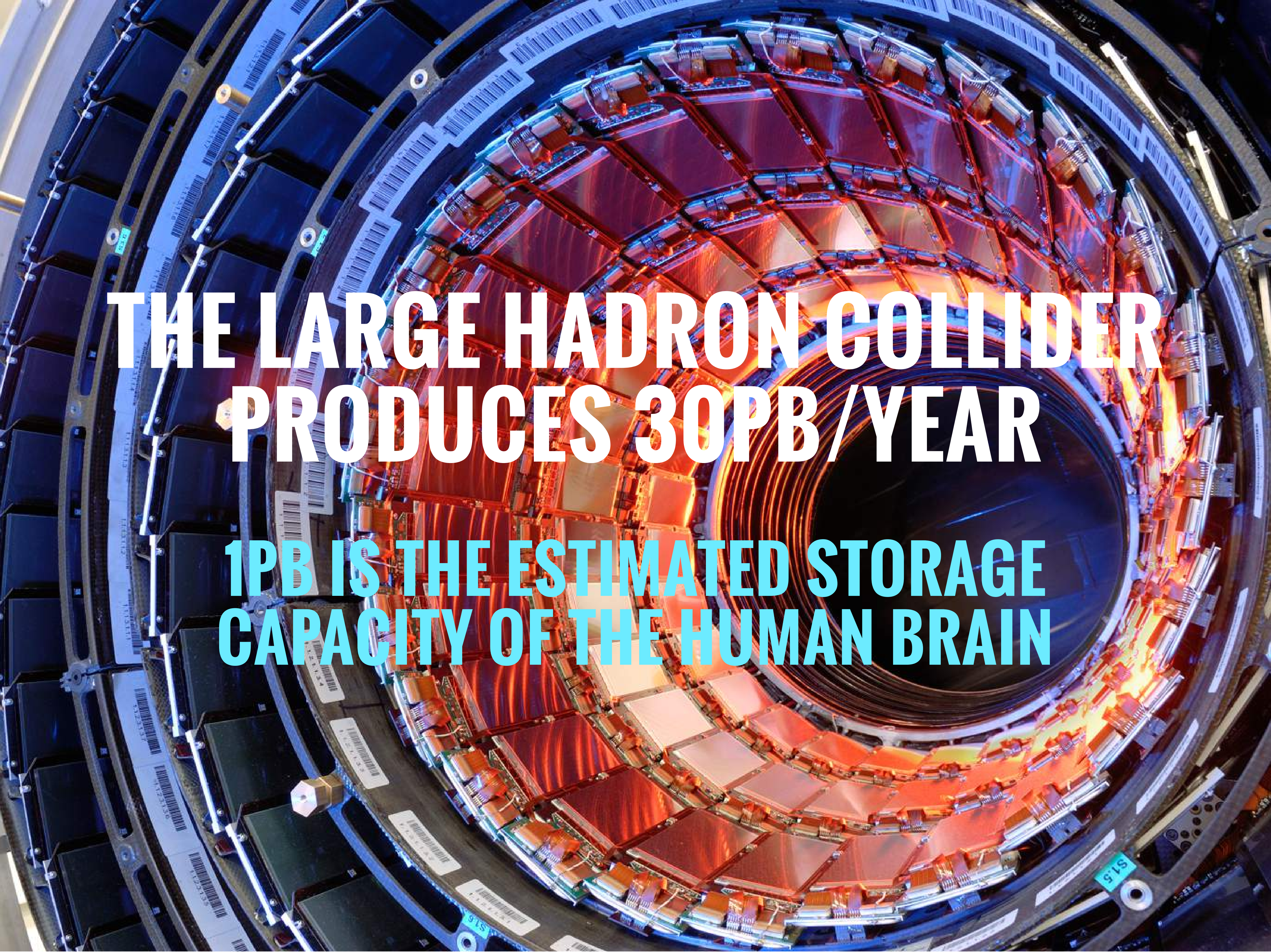
Laurent Sorber

Supervisors:
prof. Lieven De Lathauwer
prof. Marc Van Barel

IWT
KU LEUVEN


SELF-DRIVING CARS
ANALYSE 1GB/S
EQUIVALENT TO WATCHING
400 BLU-RAYS SIMULTANOUSLY
Google
self-driving car


A BOEING 787 GENERATES 500GB/FLIGHT
1000 TREES MADE INTO 1M BOOKS' WORTH OF DATA PER FLIGHT


THE LARGE HADRON COLLIDER PRODUCES 30PB/YEAR
1PB IS THE ESTIMATED STORAGE CAPACITY OF THE HUMAN BRAIN


THIS DOES NOT EVEN INCLUDE
THE MUCH LARGER DEEP WEB

BIG DATA

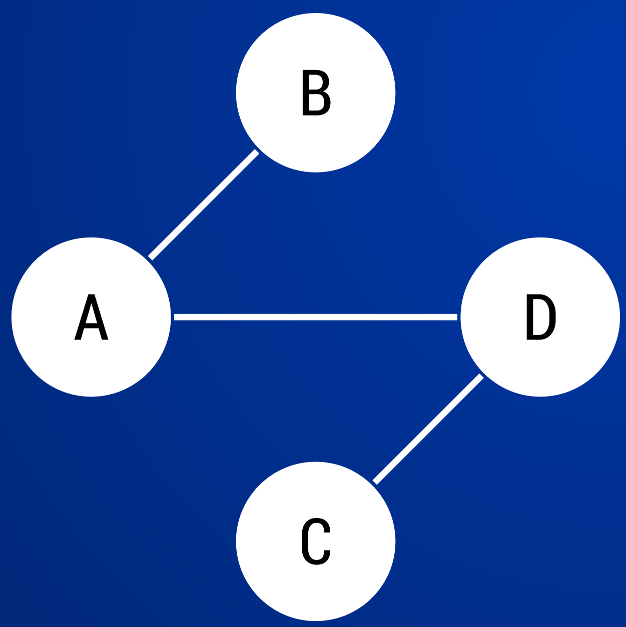
1. Volume
2. Velocity
3. Variety

HOW IS DATA STORED?

GRAPHS

Web graph, Social graph, Electrical circuit,

...



A	B	C	D
0	1	0	1
1	0	0	0
0	0	0	1
1	0	1	0

RELATIONAL DATA

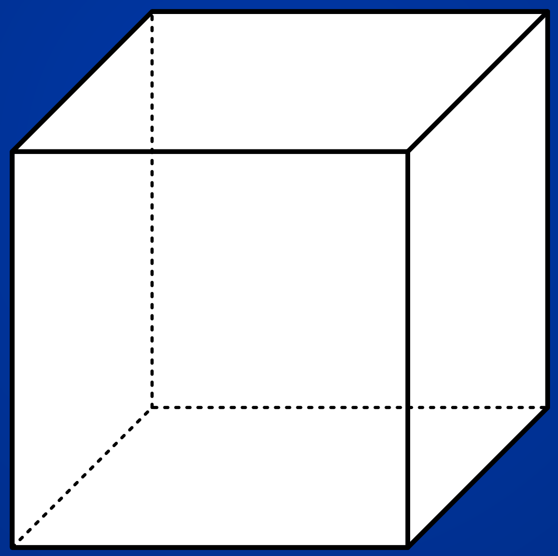
user-rates-movie, student-performs-task,

...

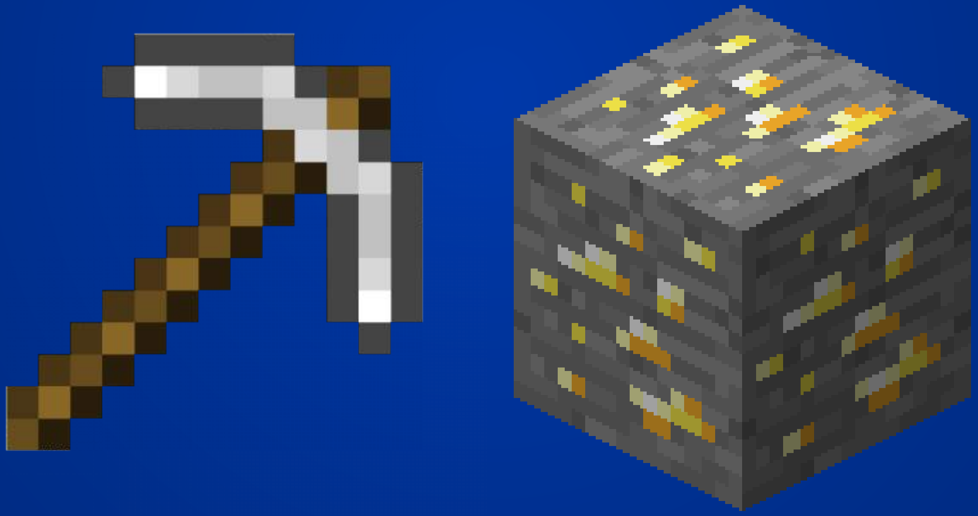
	Dries	Bart	Vicky	Clara
Pretty Woman	3.5	1.5	?	4.5
Titanic	?	2.25	3	4
Star Wars	4.25	?	2	3
The Terminator	4.5	4.5	1.5	?

MATRIX

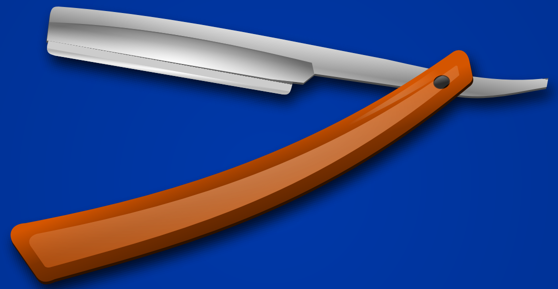
TENSOR



HOW CAN WE MINE THIS NEW NATURAL RESOURCE?



USE OCCAM'S RAZOR!



“Shave the model until you cut into its explanatory power.”

PREDICTING MOVIE RATINGS

	Dries	Bart	Vicky	Clara
Pretty Woman	3.5	1.5	?	4.5
Titanic	?	2.25	3	4
Star Wars	4.25	?	2	3
The Terminator	4.5	4.5	1.5	?

	Action content (a)
Pretty Woman	0
Titanic	0.5
Star Wars	1.5
The Terminator	2

	Action preference (b)
Dries	1
Bart	1
Vicky	−0.5
Clara	0

$$r_{m,u} = \mu + a_m \cdot b_u$$

$$4.5 = 2.5 + 2 \cdot 1$$

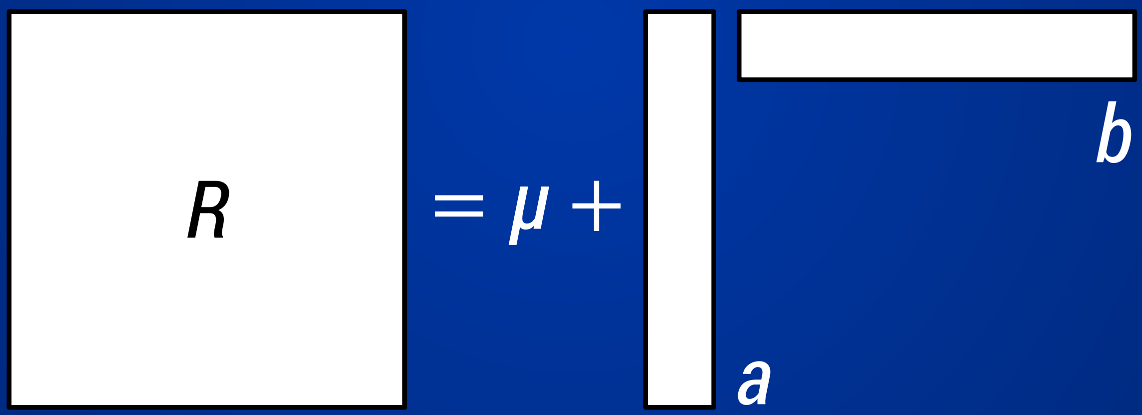
Dries likes The Terminator

$$1.5 = 2.5 + 2 \cdot (-0.5)$$

Vicky does not like The Terminator

WHAT DOES THE MODEL LOOK LIKE?

$$r_{m,u} = \mu + a_m \cdot b_u$$



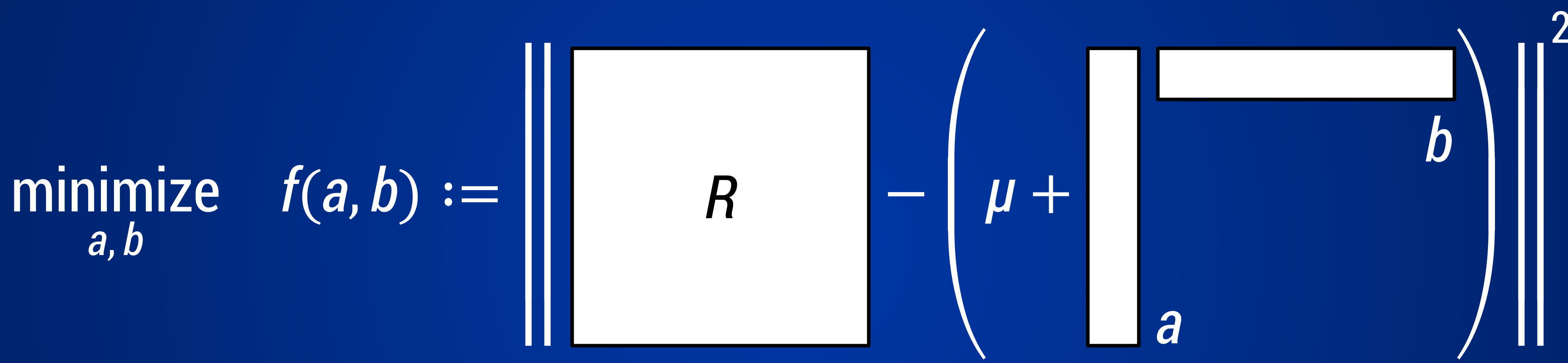
IS THE MODEL ANY GOOD?

	Dries	Bart	Vicky	Clara
Pretty Woman	3.5−2.5	1.5−2.5	?−2.5	4.5−2.5
Titanic	?−3	2.25−3	3−2.25	4−2.5
Star Wars	4.25−4	?−4	2−1.75	3−2.5
The Terminator	4.5−4.5	4.5−4.5	1.5−1.5	?−2.5

$$f(a,b) := \sum_{m,u} \left[r_{m,u} - (\mu + a_m \cdot b_u)\right]^2$$

$$[3.5 - 2.5]^2 + [4.25 - 4]^2 + [4.5 - 4.5]^2 + [1.5 - 2.5]^2 + \ldots$$

MINIMIZE *f* TO FIND THE BEST SOLUTION



However, even the best solution for a and b can not predict all the ratings accurately because the model is too simple!

	Action content (a)	Romance content (c)
Pretty Woman	0	2
Titanic	0.5	1.5
Star Wars	1.5	0.5
The Terminator	2	0

	Action preference (b)	Romance preference (d)
Dries	1	0.5
Bart	1	−0.5
Vicky	−0.5	0.5
Clara	0	1

$$r_{m,u} = \mu + a_m \cdot b_u + c_m \cdot d_u$$

$$2.25 = 2.5 + 0.5 \cdot 1 + 1.5 \cdot (-0.5)$$

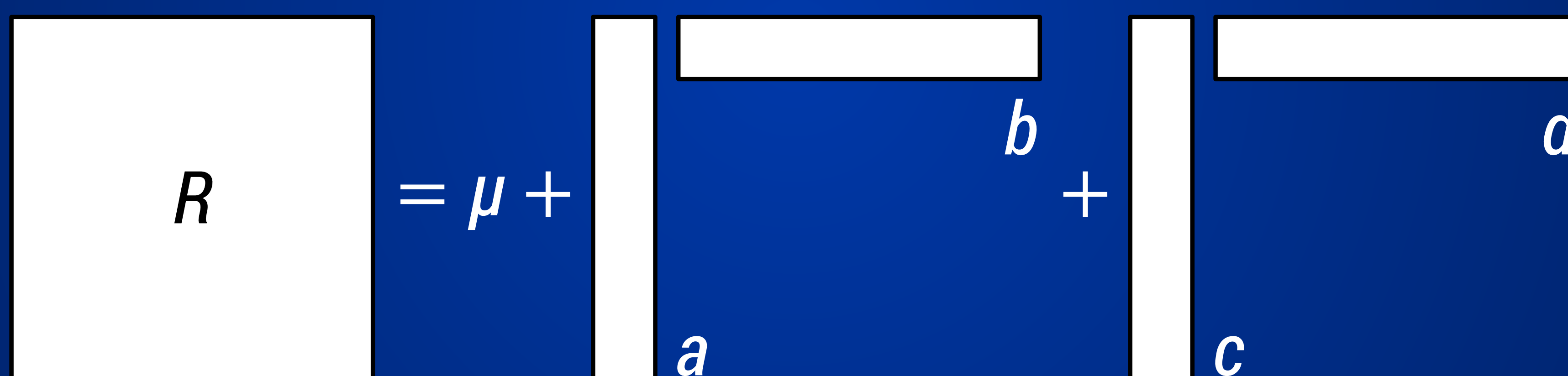
Bart does not like Titanic very much

$$4 = 2.5 + 0.5 \cdot 0 + 1.5 \cdot 1$$

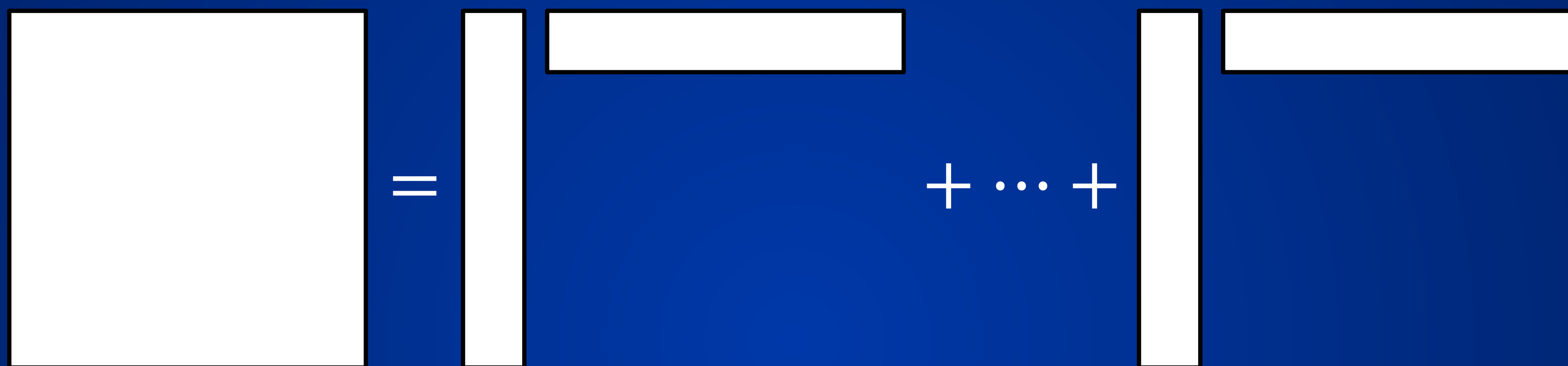
Clara likes Titanic

WHAT DOES THE NEW MODEL LOOK LIKE?

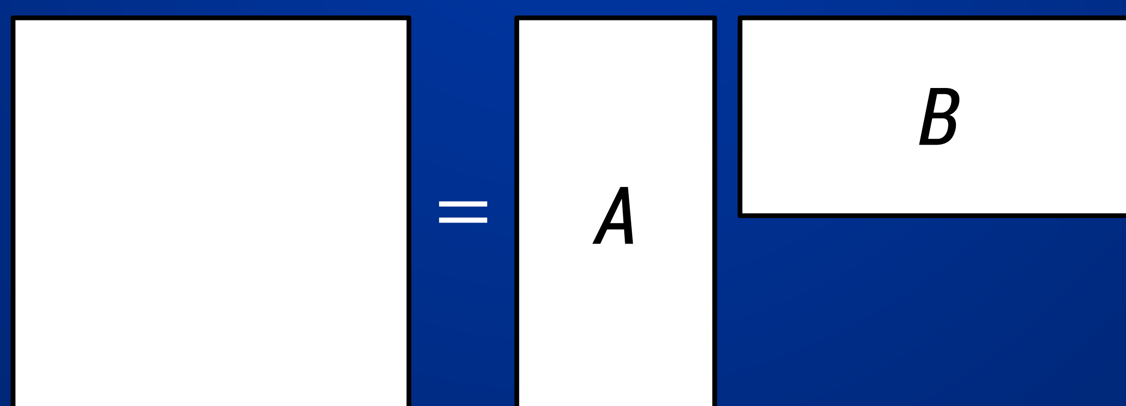
$$r_{m,u} = \mu + a_m \cdot b_u + c_m \cdot d_u$$



MATRIX FACTORIZATION



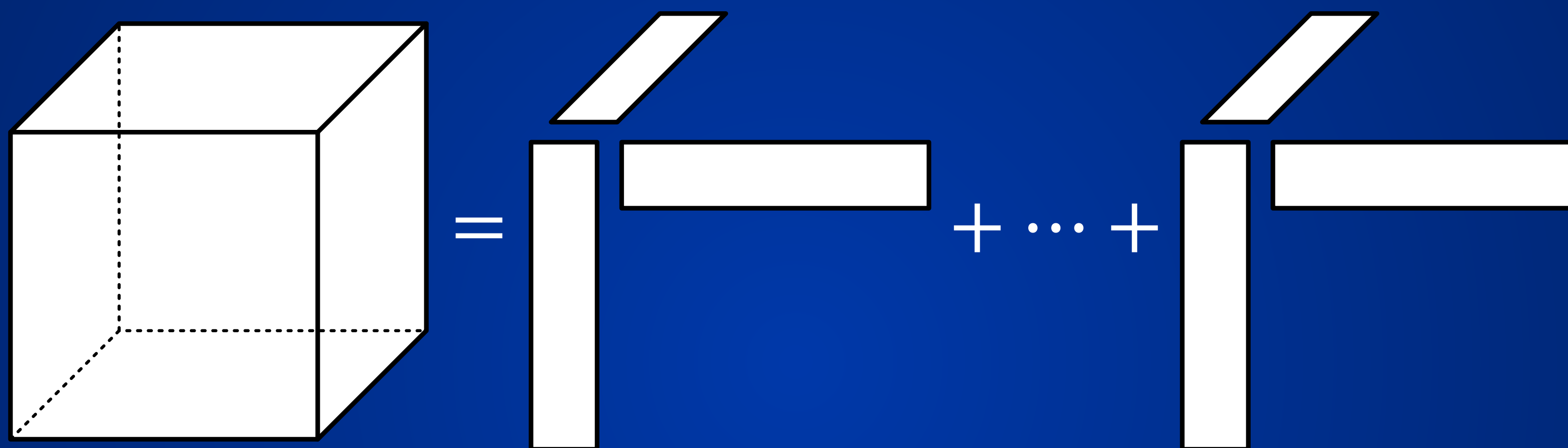
Or equivalently,



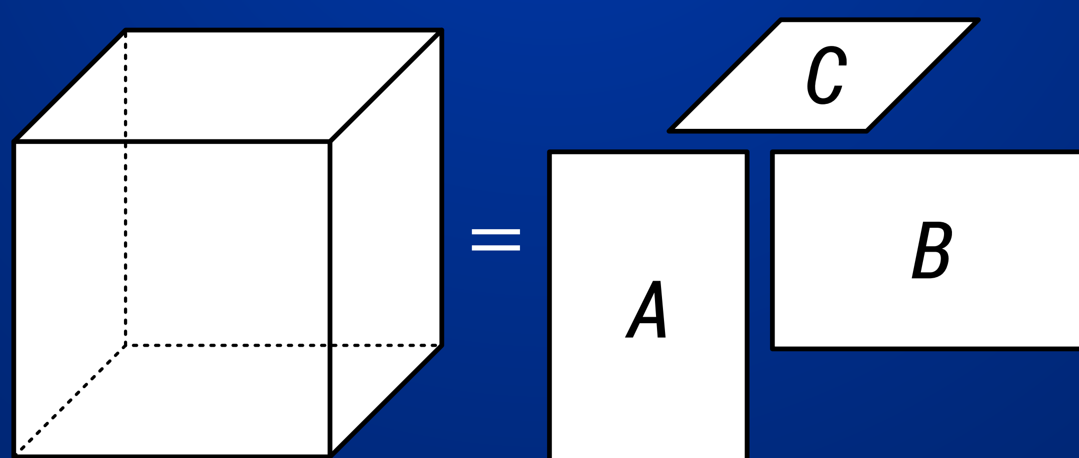
THEN WHAT ARE TENSOR FACTORIZATIONS?

TENSOR FACTORIZATION

Canonical polyadic decomposition (CPD)

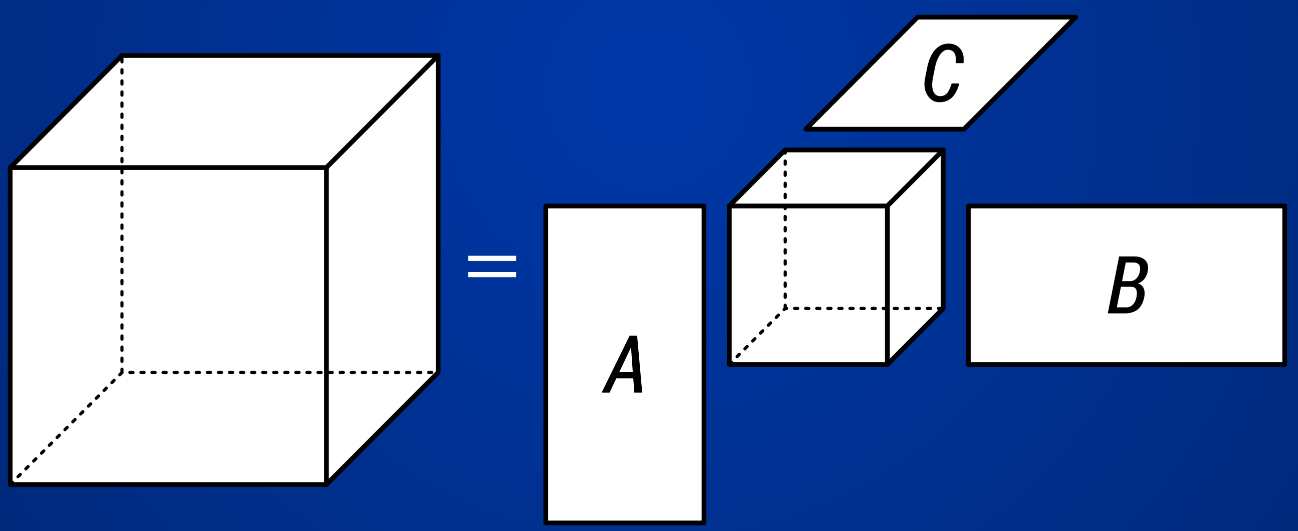


Or equivalently,

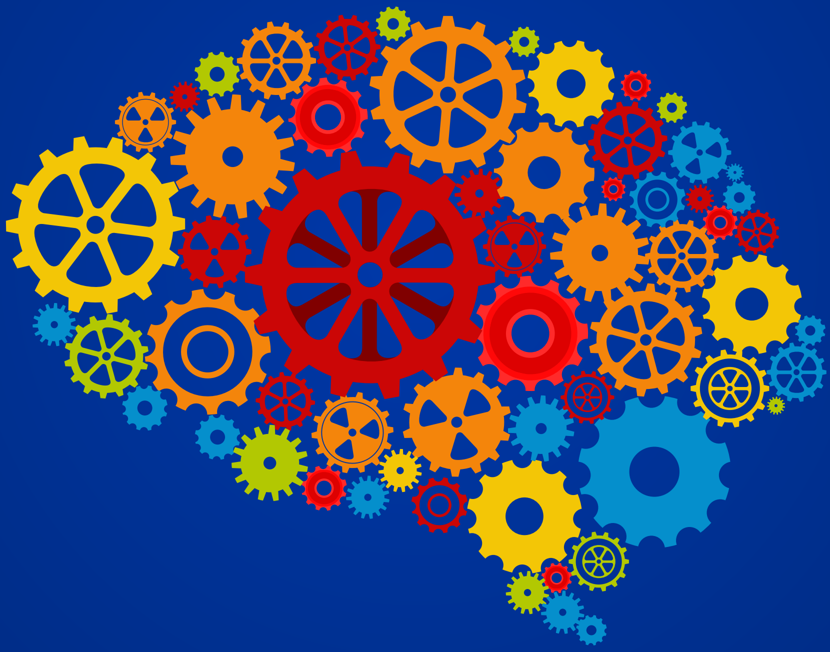


TENSOR FACTORIZATION

Low multilinear rank approximation (LMLRA)

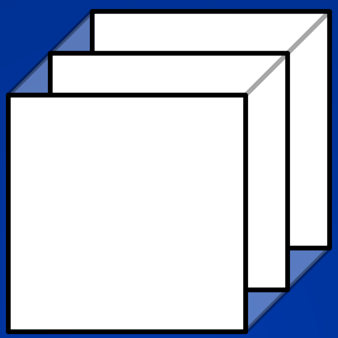


SO WHAT IS DATA FUSION?



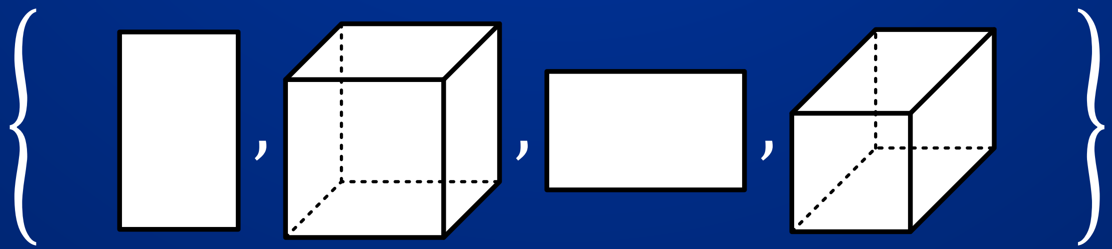
TENSOR FACTORIZATION

Fusion by jointly factorizing a set of matrices
that are homogeneous in size



DATA FUSION

Fusion by jointly factorizing a set of tensors
that are heterogeneous in size

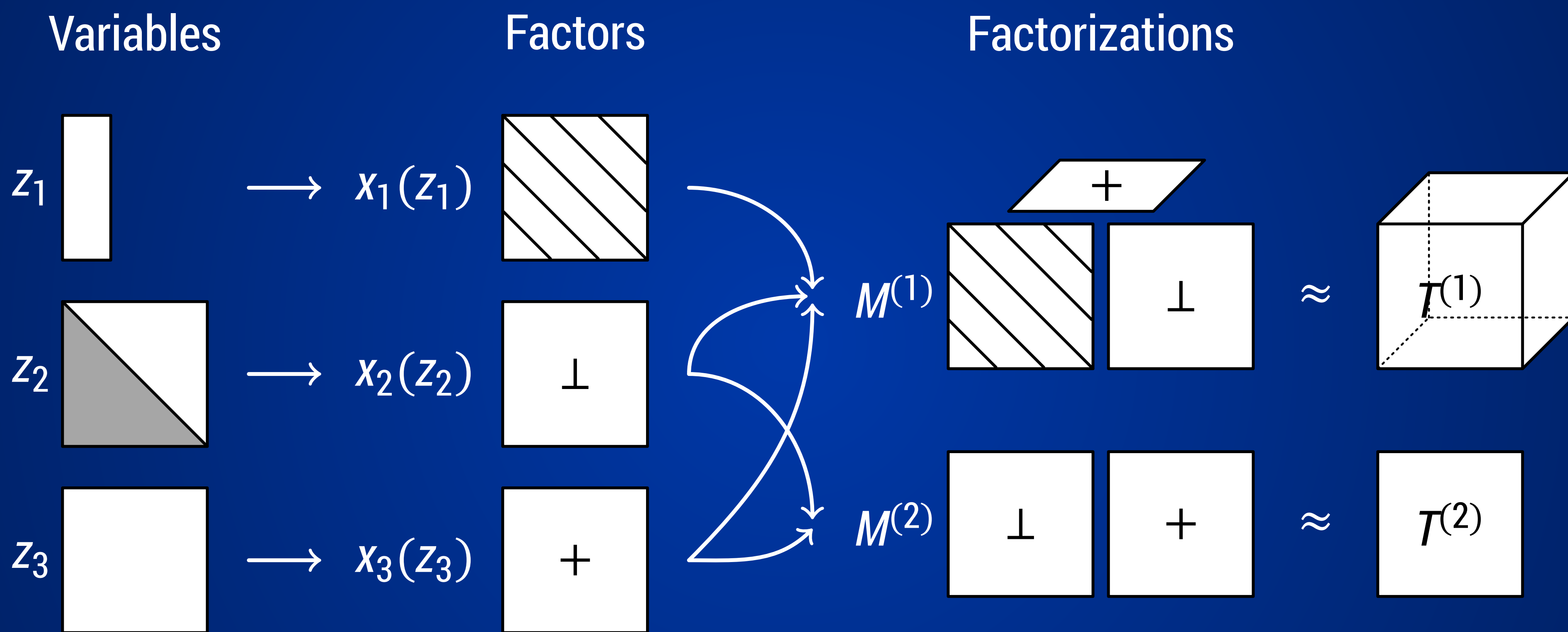


GOAL

For each data set (stored as a tensor):

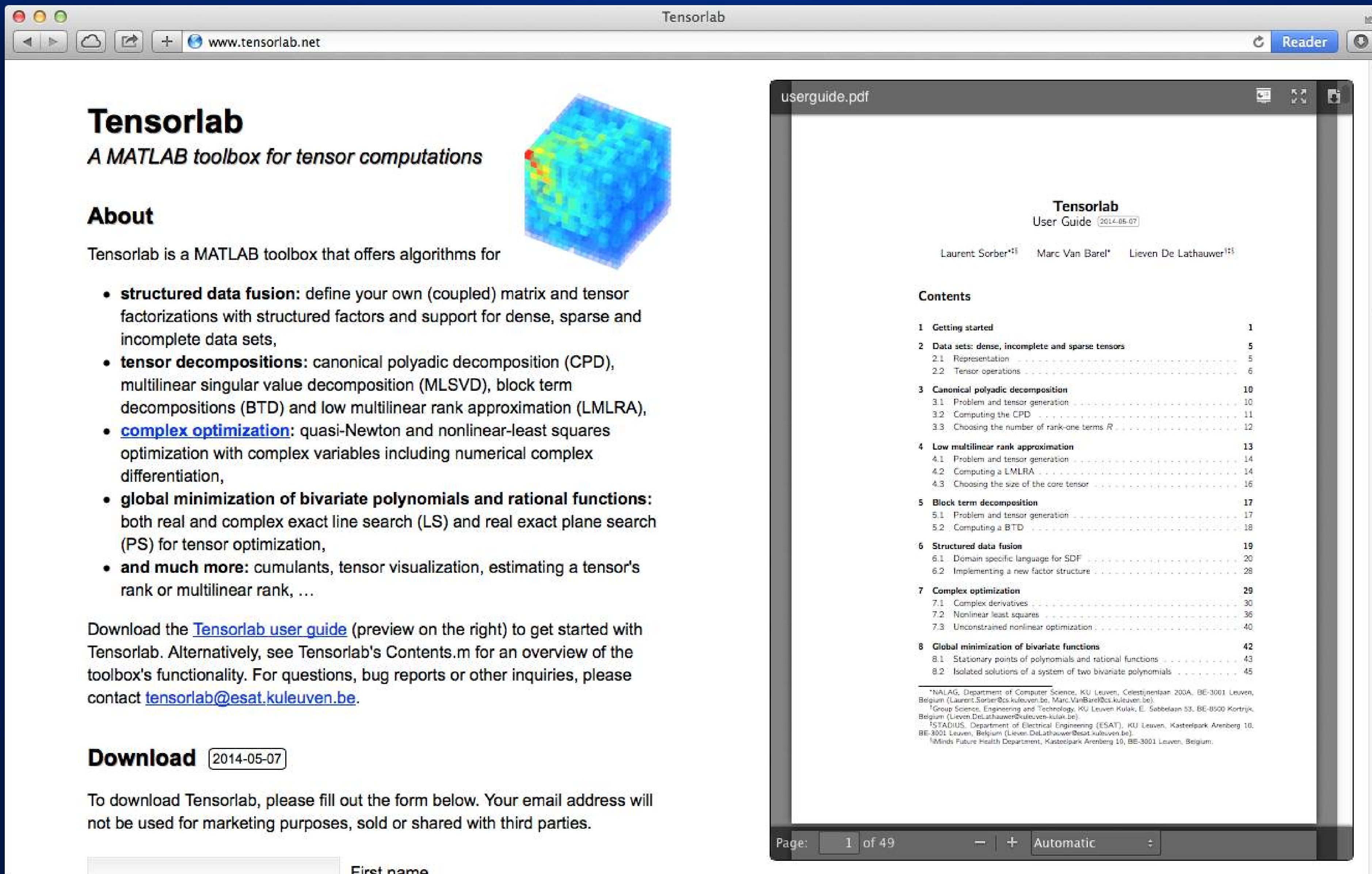
1. Pick a type of tensor decomposition
   CPD, LMLRA, BTD, TT, ...
2. Impose structure on their factors
   Nonnegative, orthogonal, inverse, ...
3. Choose which factors are shared
   1st and 2nd tensor share factor matrix *A*, ...

STRUCTURED DATA FUSION



$$\underset{z}{\text{minimize}} \quad \sum_d \omega_d \left\| M^{(d)}(X(z)) - T^{(d)} \right\|^2$$

WWW.TENSORLAB.NET



ZEISS

BOEING®

D:WAVE
The Quantum Computing Company™

SAMSUNG

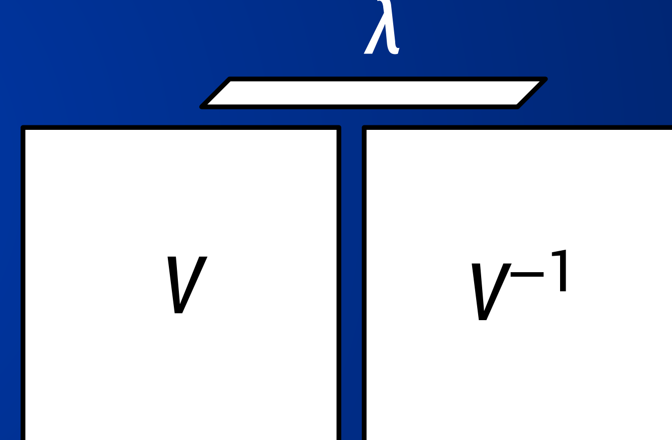
NASA

HITACHI
Inspire the Next

EIGENVALUE DECOMPOSITION

is a CPD with structured factors

$$A = \begin{bmatrix} 0 & 0.5 & & & & \\ 1 & 0 & 0.5 & & & \\ & 0.5 & 0 & 0.5 & & \\ & & 0.5 & 0 & 0.5 & \\ & & & \ddots & \ddots & \ddots \end{bmatrix} =$$



$$\lambda_n = \cos\left(\frac{\pi(2n-1)}{2N}\right) \quad \text{for} \quad n = 1, \ldots, N$$

MATLAB

```
[V,L] = eig(A);
```

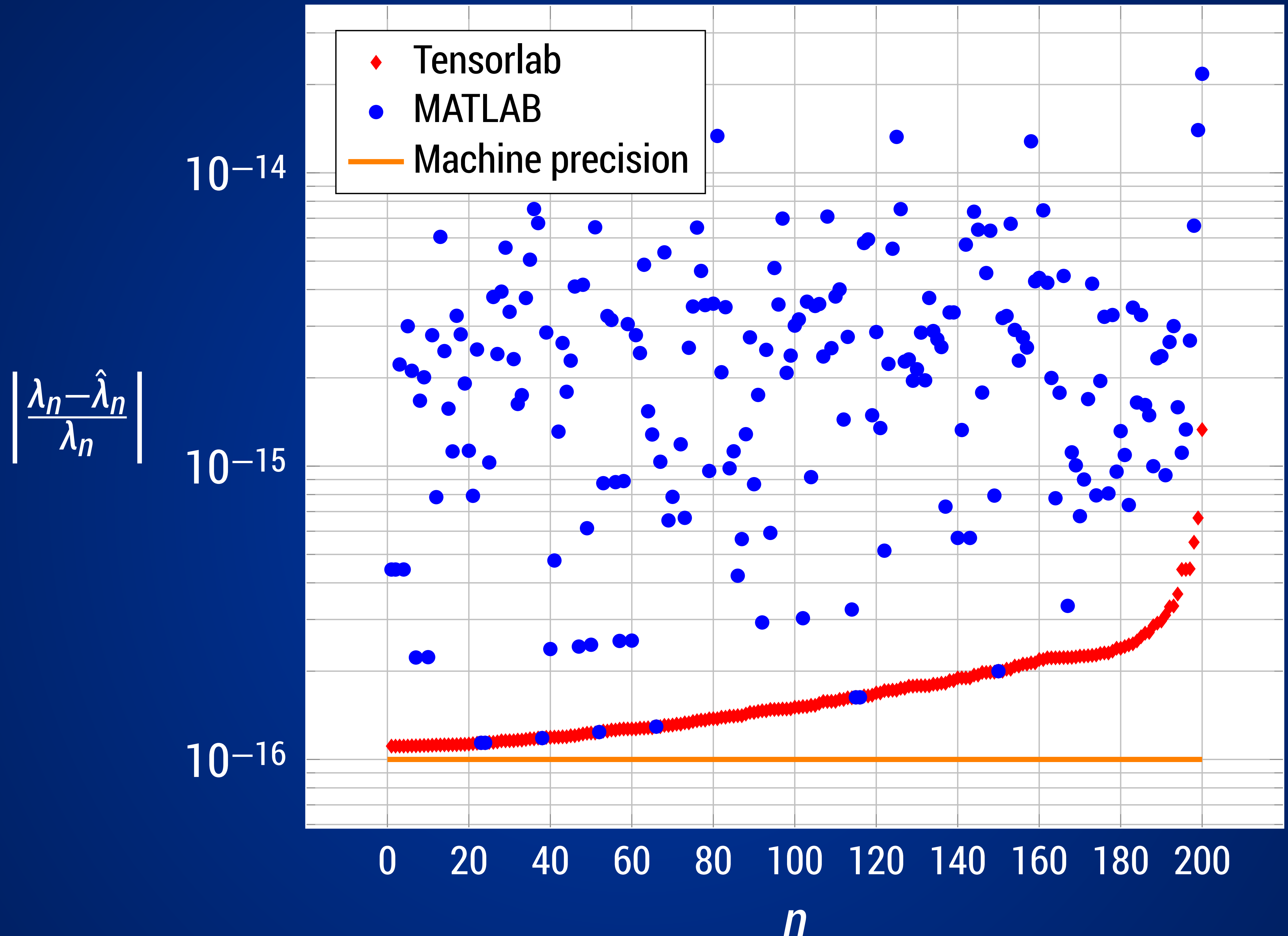
TENSORLAB

```
model.variables.v = randn(N,N); % N-by-N matrix
model.variables.l = randn(1,N); % 1-by-N vector
```

```
model.factors.V    =  'v';
model.factors.Vinv = {'v',@struct_invtransp}; % One of 32 s
model.factors.L    =  'l';
```

```
model.factorizations.evd.data = A;
model.factorizations.evd.cpd  = {'V','Vinv','L'};
```

```
sol = sdf_nls(model); % sol.factors, sol.variables
```



Tensorlab
MATLAB
Machine precision
10^{-14}
10^{-15}
10^{-16}
$\left|\frac{\lambda_n-\hat{\lambda}_n}{\lambda_n}\right|$
0
20
40
60
80
100
120
140
160
180
200
n

ALGORITHMS

$$\underset{z}{\text{minimize}} \quad \sum_d \omega_d \left\| M^{(d)}(X(z)) - T^{(d)} \right\|^2$$

- User's choice of underlying solver
  Quasi-Newton, nonlinear least squares, ...
- Solver exploits the structure in the factors
  Nonnegative, orthogonal, inverse, ...
- Solver exploits the structure of the decomposition
  CPD, LMLRA, BTD, TT, ...
- Based on complex optimization
  Solve complex-valued problems with the same code

LINE AND PLANE SEARCH

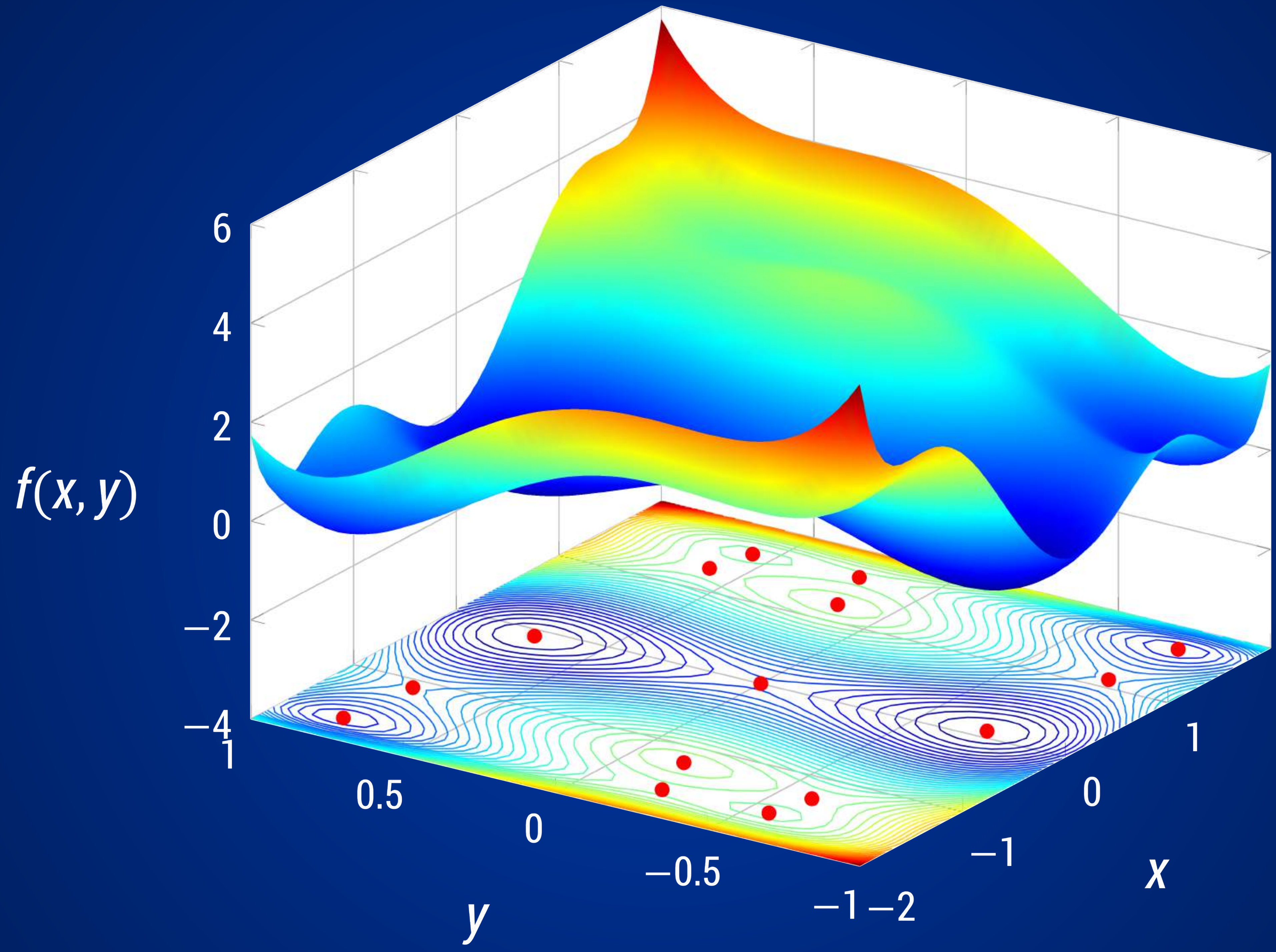
$$\underset{z}{\text{minimize}} \quad f(z)$$

Line search subproblem

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad f(z_k + x \cdot p)$$

Plane search subproblem

$$\underset{x,y \in \mathbb{R}}{\text{minimize}} \quad f(z_k + x \cdot p + y \cdot q)$$



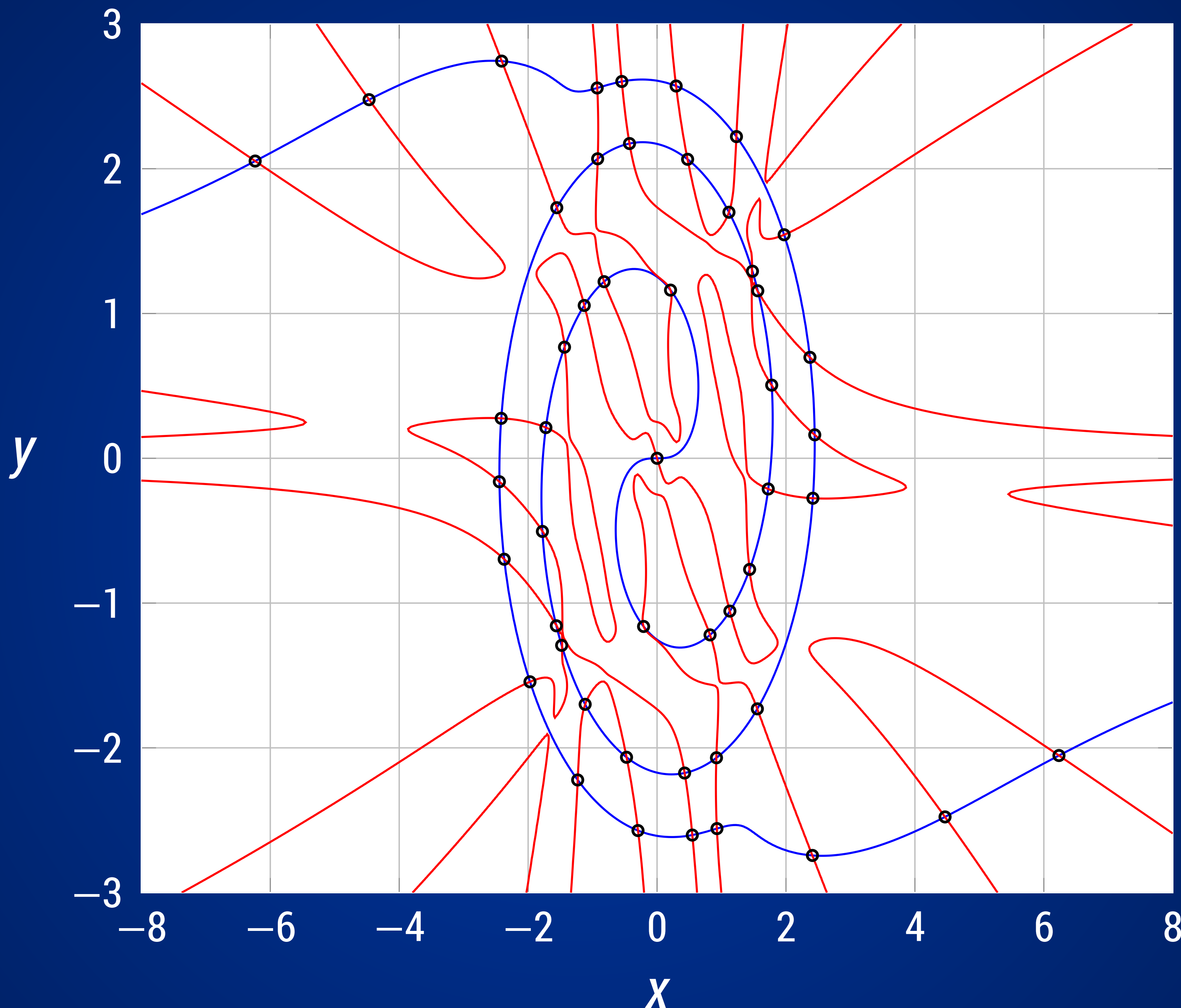
6
4
2
0
−2
−4
f(x, y)
1
0.5
0
−0.5
−1
y
−2
−1
0
1
2
x

The global minimizer of $f(x,y)$ is a solution of a bivariate polynomial system, e.g.,

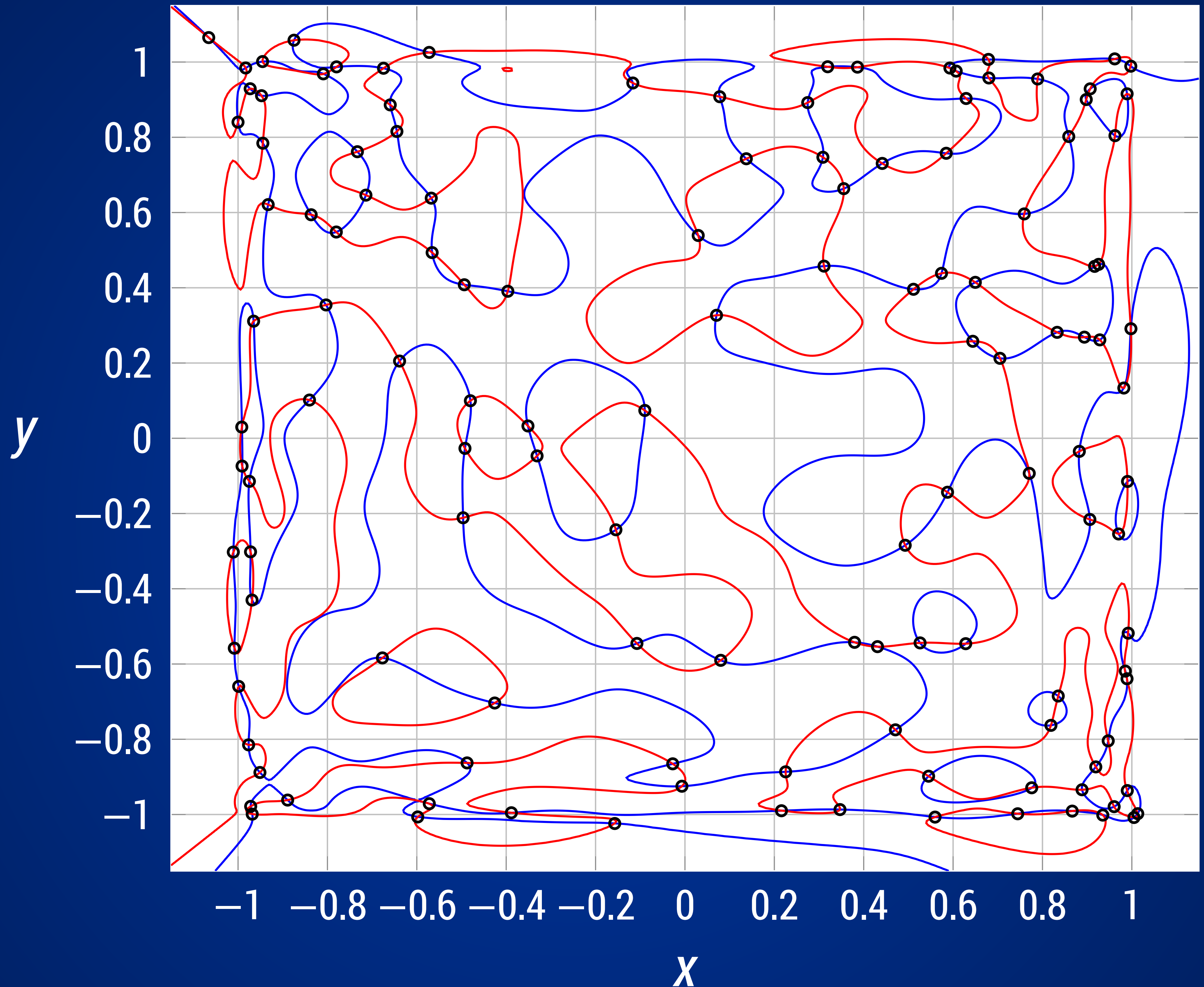
$$\begin{cases} 2x^2 - xy + y^4 - 8 = 0 \\ 5y^3 + 7xy^2 + 3 = 0 \end{cases}$$

Developed an algorithm that computes all solutions simultaneously by solving an eigenvalue decomposition

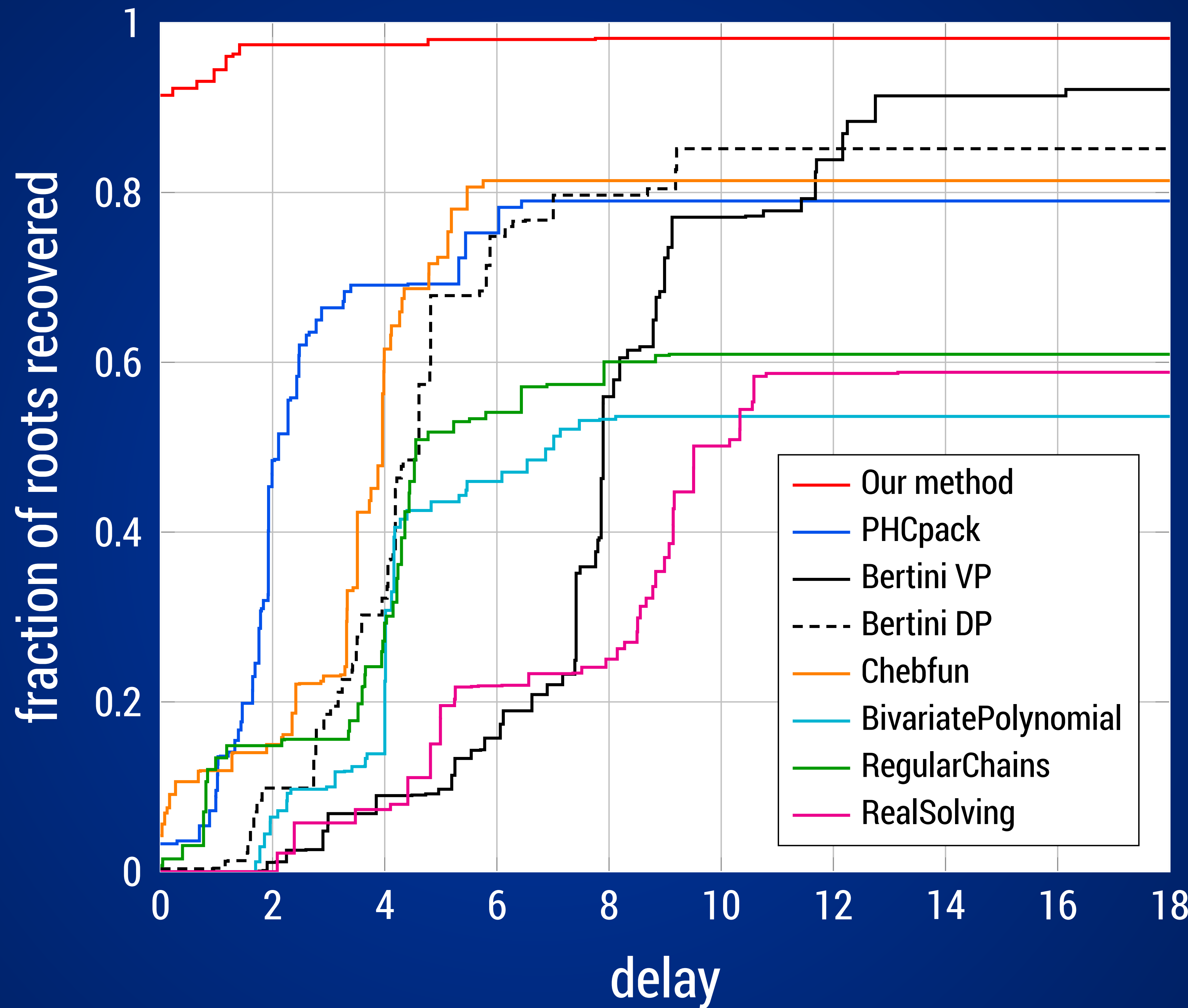
$$A \cdot x = \lambda \cdot B \cdot x$$



3
2
1
0
−1
−2
−3
y
−8
−6
−4
−2
0
2
4
6
8
x



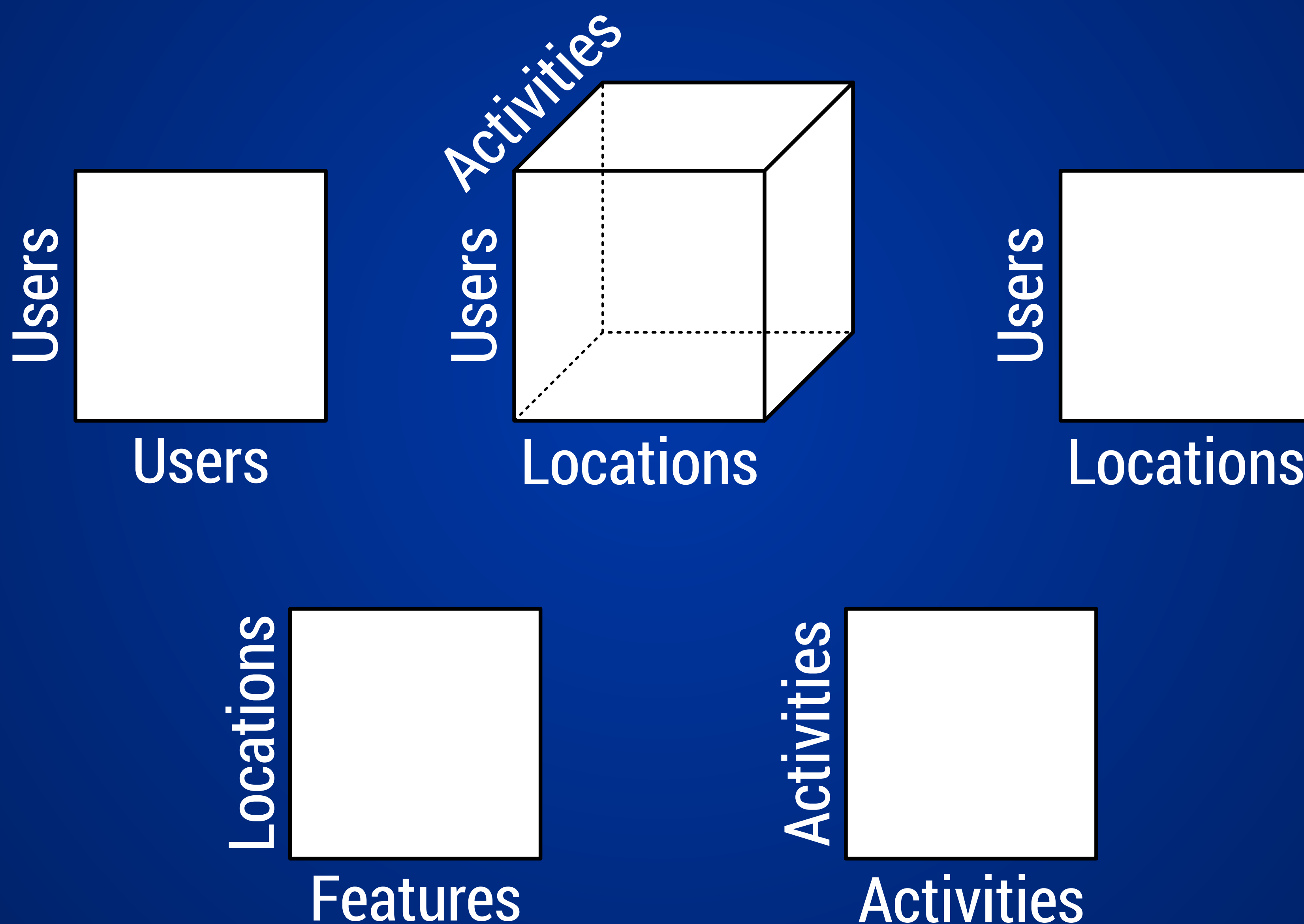
y
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
x

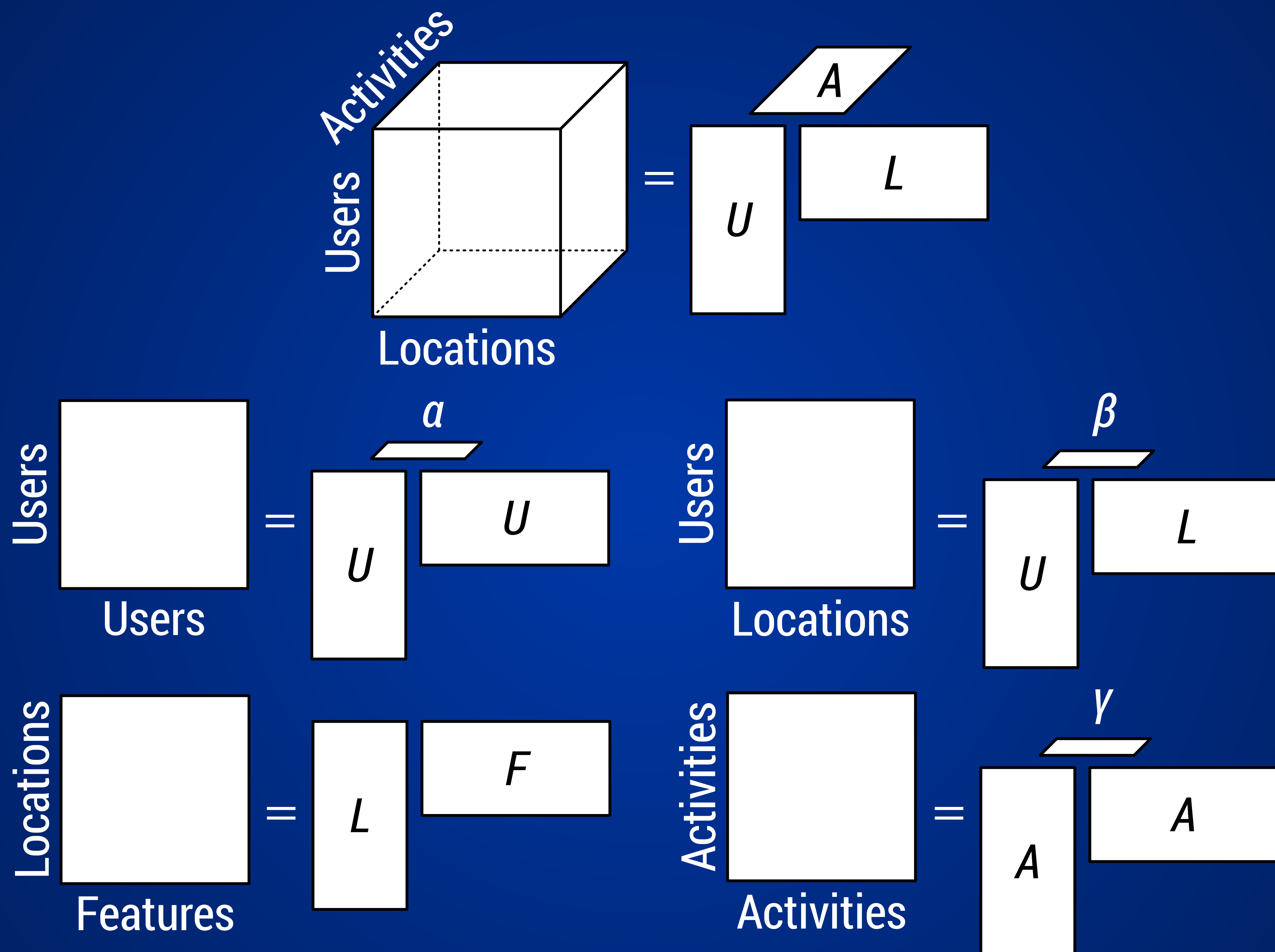


fraction of roots recovered
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
12
14
16
18
delay
Our method
PHCpack
Bertini VP
Bertini DP
Chebfun
BivariatePolynomial
RegularChains
RealSolving

Microsoft Research tracked 164 users' GPS coordinates for 2.5 years

Let's apply data fusion to predict if users took part in activies!

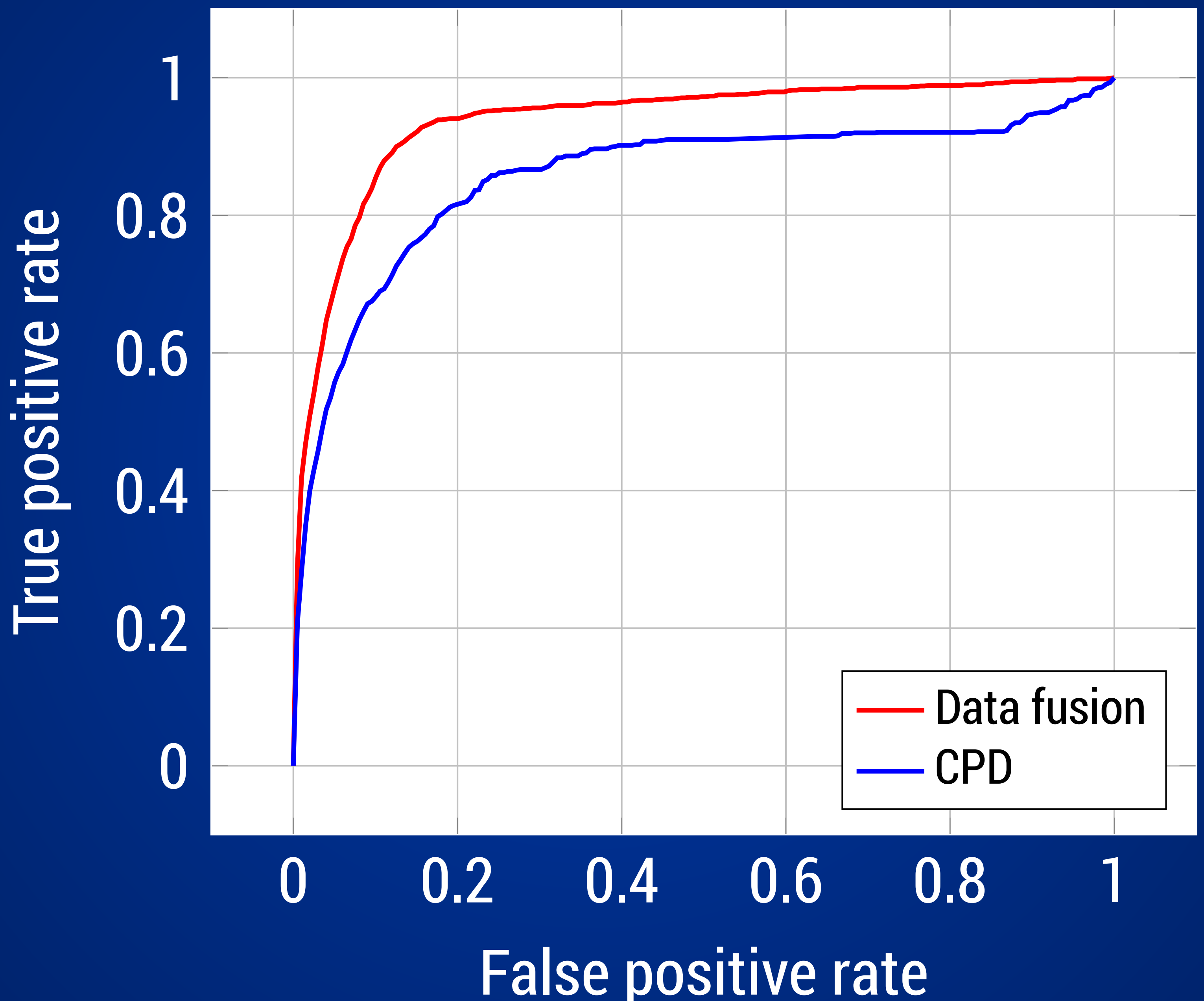



Activities
Users
Locations
A
U
L
Users
Users
α
U
U
Users
Locations
β
U
L
Locations
Features
L
F
Activities
Activities
γ
A
A

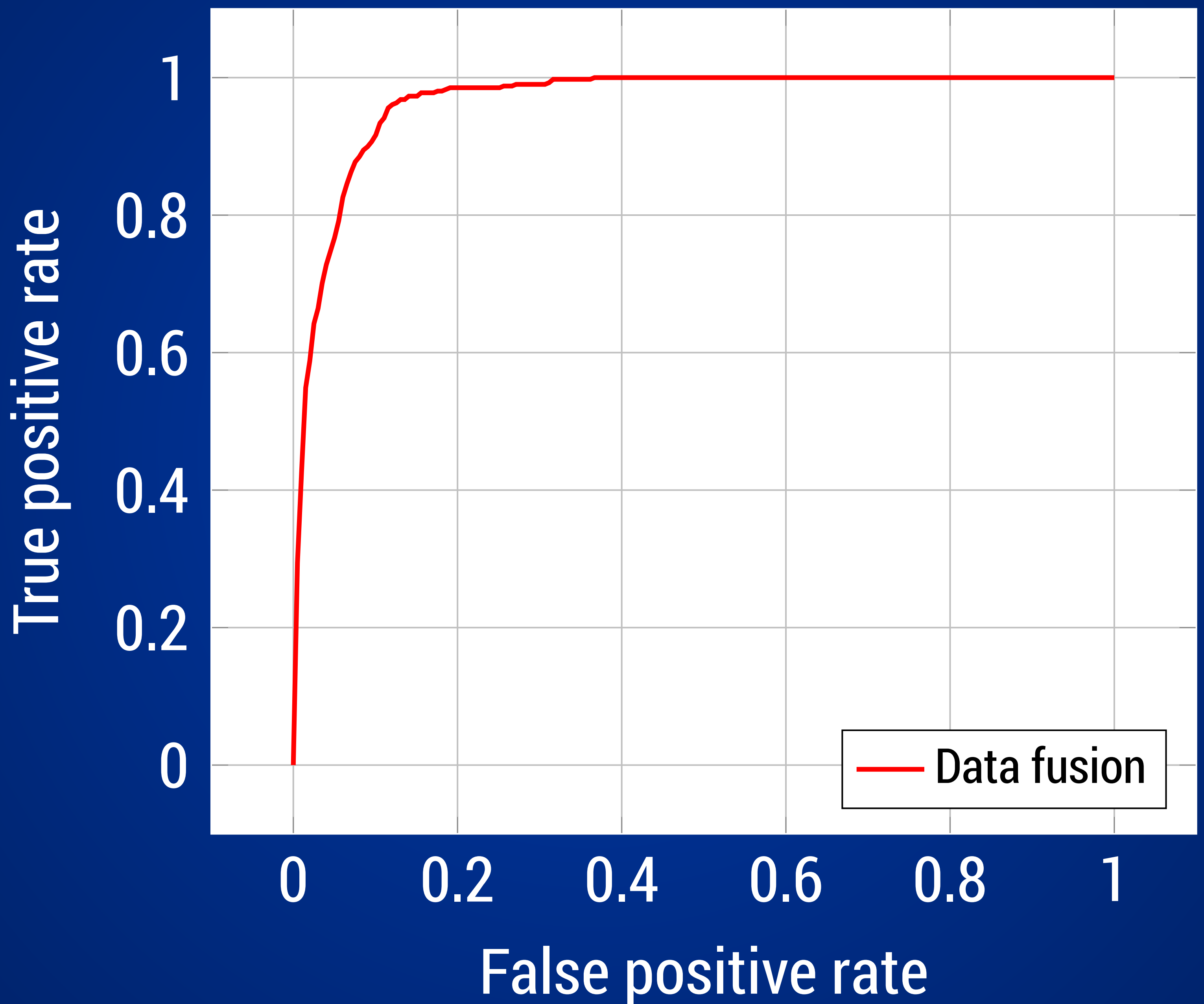
TENSORLAB

```
model.factorizations.UserLocAct.data = UserLocAct;
model.factorizations.UserLocAct.cpd  = {'U','L','A'};

model.factorizations.UserUser.data = UserUser;
model.factorizations.UserUser.cpd  = {'U','U','alpha'};

model.factorizations.LocFea.data = LocFea;
model.factorizations.LocFea.cpd  = {'L','F'};

model.factorizations.ActAct.data = ActAct;
model.factorizations.ActAct.cpd  = {'A','A','beta'};

model.factorizations.UserLoc.data = UserLoc;
model.factorizations.UserLoc.cpd  = {'U','L','gamma'};
```

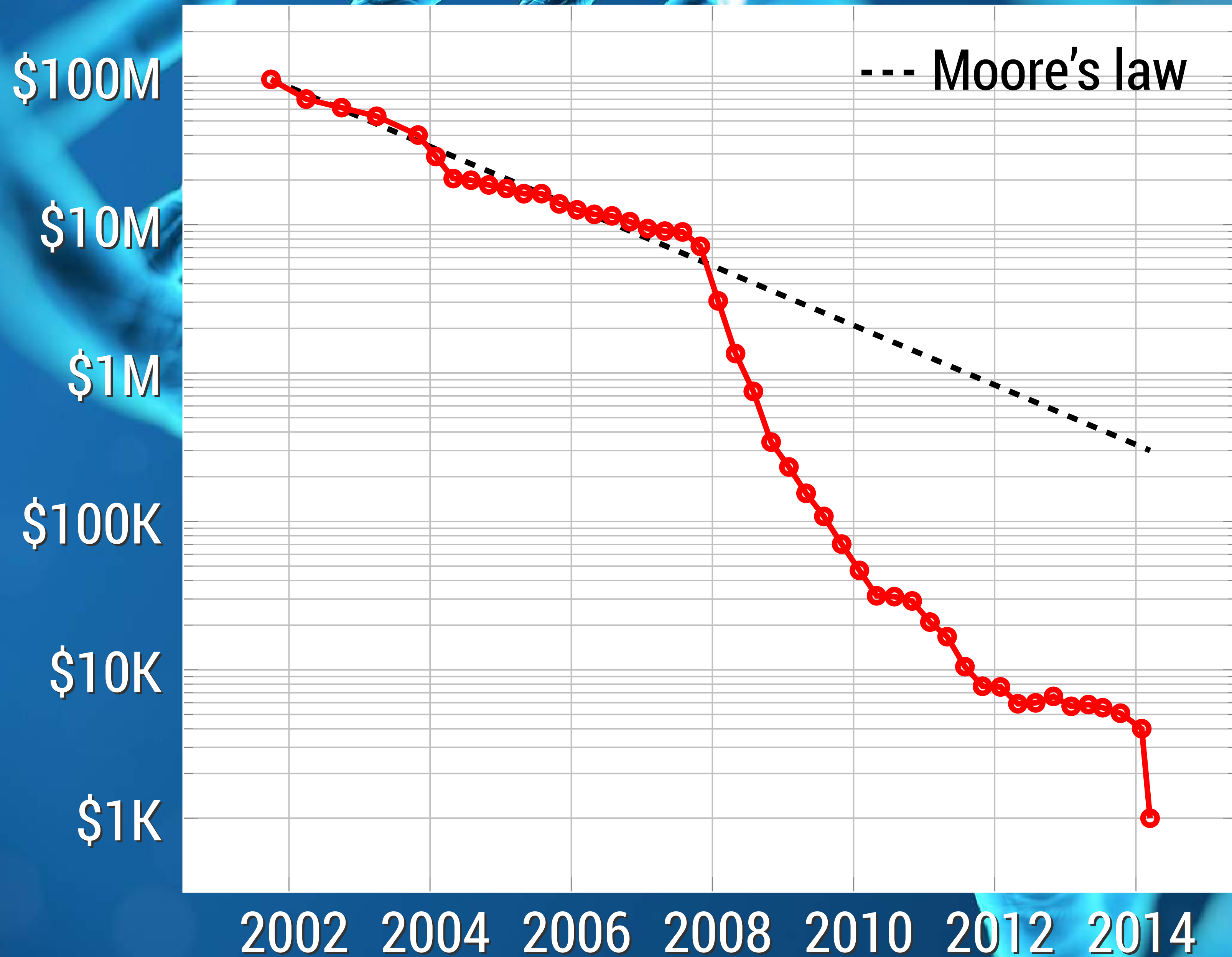
80% OF USER-LOCATION-ACTIVITY MISSING



50 USERS MISSING



COST OF SEQUENCING A HUMAN GENOME



THE END?

- Big data is upon us!
- Structured data fusion is a paradigm shift in our capabilities to extract knowledge from this exciting new resource

THANK YOU!

REFERENCES

- Sorber, L., Van Barel, M., De Lathauwer, L. Structured data fusion. ESAT-STADIUS IR 13-177, KU Leuven. Submitted.
- Sorber, L., Van Barel, M., De Lathauwer, L. Numerical solution of bivariate and polyanalytic polynomial systems. *SIAM Journal on Numerical Analysis*. Accepted.
- Sorber, L., Domanov, I., Van Barel, M., De Lathauwer, L. Exact line and plane search for tensor optimization. ESAT-STADIUS IR 13-02, KU Leuven. Submitted.
- Sorber, L., Van Barel, M., De Lathauwer, L. Optimization-based algorithms for tensor decompositions. *SIAM Journal on Optimization 23*, 2 (2013), 695—720.
- Sorber, L., Van Barel, M., De Lathauwer, L. Unconstrained optimization of real functions in complex variables. *SIAM Journal on Optimization 22*, 3 (2012), 879—898.