Musings on Exabyte Scale Principal Component Analysis

Randy Paffenroth

Associate Professor of Mathematical Sciences,
Associate Professor of Computer Science,
Data Science Program,
Worcester Polytechnic Institute

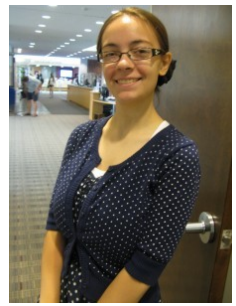
Massachusetts HPC Day
UMass Dartmouth
5-25-17

WPI

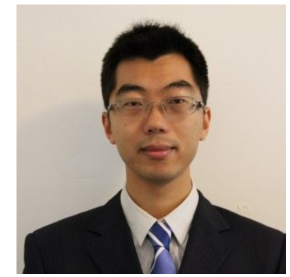
I have the pleasure of working with many great students and collaborators!


Nitish Bahadur


Kelum Gajamannage


Kathleen Kay


Wenjing Li


Haitao Liu


Neehar Mukne


Tyler Reese


Matt Weiss


Lu Zhong


Chong Zhou


Chen Zou


Xiaozhou “Joe” Zou

MITRE


Les Servi

Colorado State University


Sridhar Ramasamy


Vidarshana Bandara


Louis Scharf


Anura Jayasumana

WPI

Large Scale Data Science and classic HPC

By Camelia.boban (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

"Titan render" by Unknown - http://www.olcf.ornl.gov/2012/12/17/titan-trainers-take-road-trip/. Licensed under Public domain via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Titan_render.png#mediaviewer/File:Titan_render.png

"NvidiaTesla" by Mahogny - CameraTransferred from en.wikipedia. Licensed under Public domain via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:NvidiaTesla.jpg#mediaviewer/File:NvidiaTesla.jpg

WPI

Exabyte Scale

On the order of 10^{18} bytes

- 1 Exabyte

  = 1,000 Petabytes

  = 1,000,000 Terabytes

  = 1,000,000,000 Gigabytes

WPI

Getting some intuition for an exabyte

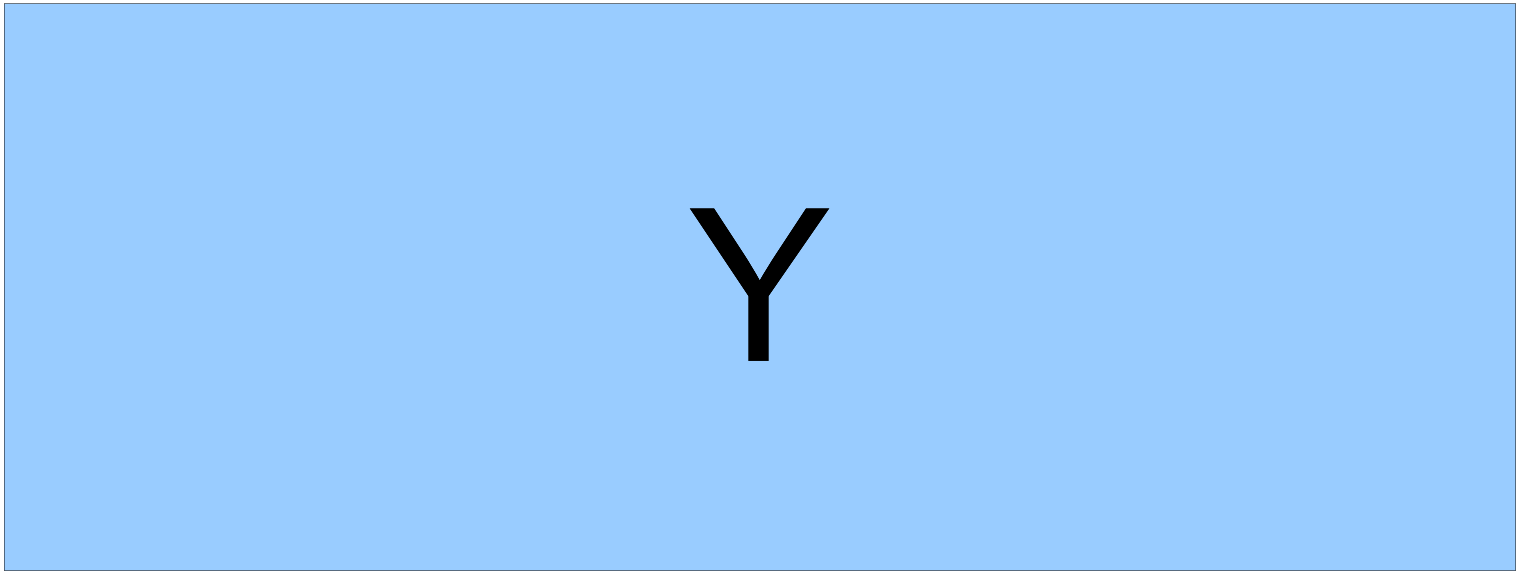
- University of California, Berkeley, estimated that by the end of 1999, the sum of human-produced information (including all audio, video recordings, and text/books) was about 12 exabytes of data.
- According to an International Data Corporation paper sponsored by EMC Corporation, 161 exabytes of data were created in 2006, "3 million times the amount of information contained in all the books ever written".
- In 2004, the global Internet traffic passed 1 exabyte per month for the first time. The global Internet traffic has continued its exponential growth and as of March 2010 it was estimated at 21 exabytes per month.

From: https://en.wikipedia.org/wiki/Exabyte

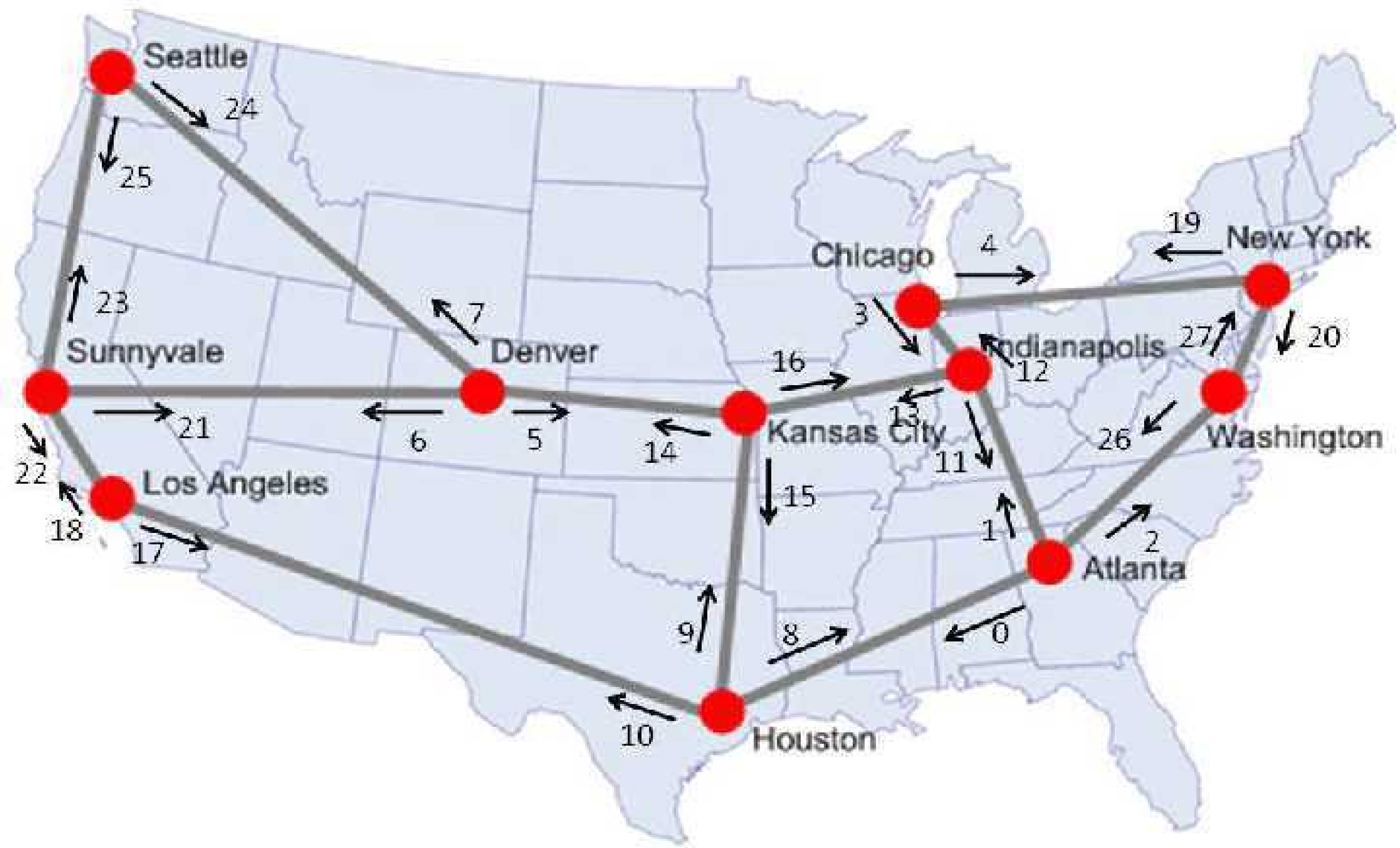
Exabyte Scale Matrix Example

10^{12} bytes from each sensor =
1 terabyte per row

10^{6} sensors =
1,000,000 rows



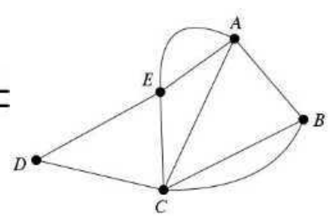
Main example to think about: The Internet

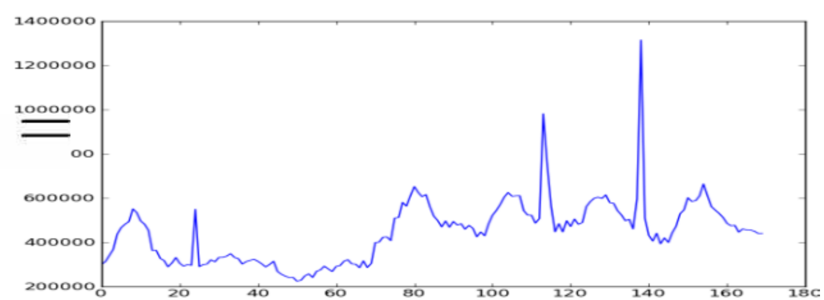


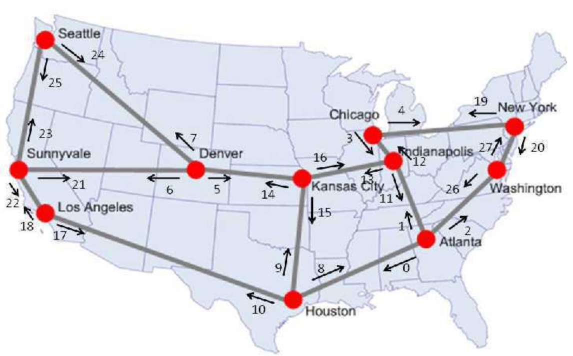
http://www.internet2.edu/products-services/advanced-networking/

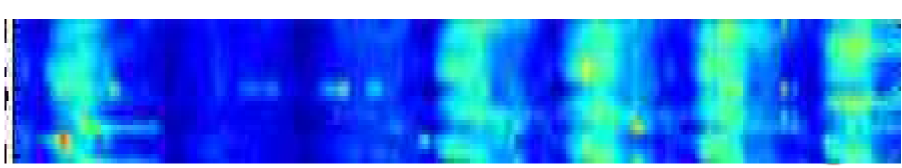
Data matrix

- Given a graph $G = \{E, V\}$ we assign to each vertex $v_i \in V$ a discrete vector valued time series $y_i \in \mathbb{R}^{l_i \times n}$ of dimension l_i and length n.
- We then construct a signal matrix $Y \in \mathbb{R}^{m \times n}$ with $m = \sum_{i=1}^{|V|} l_i$.

$G =$ 

$y_{i,j} =$ 

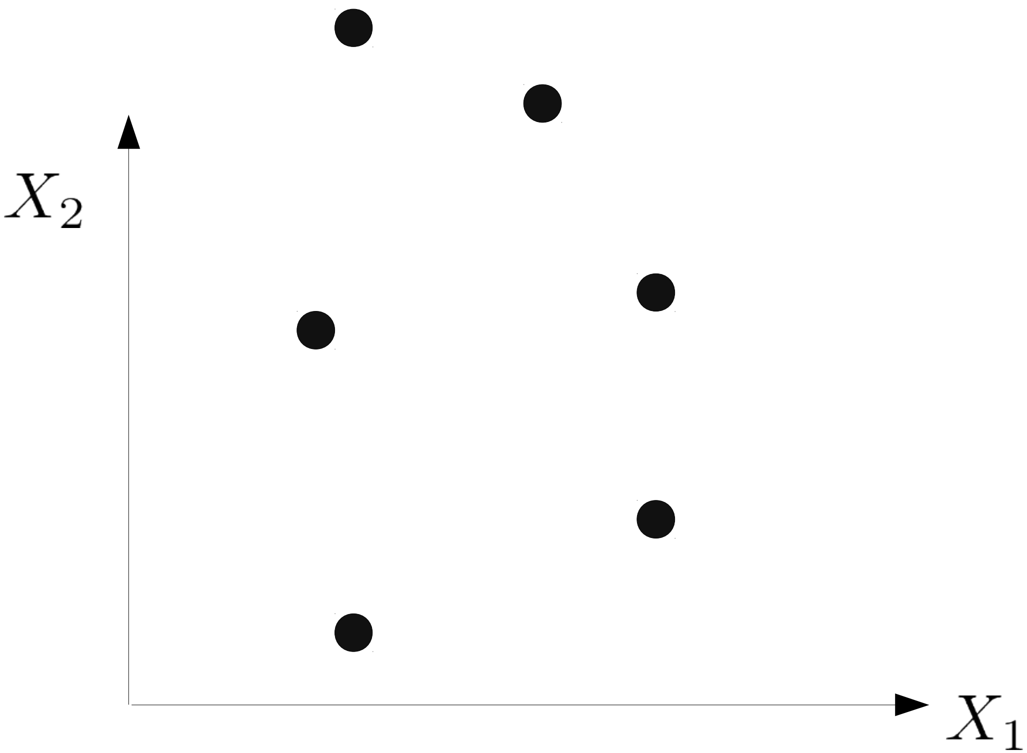


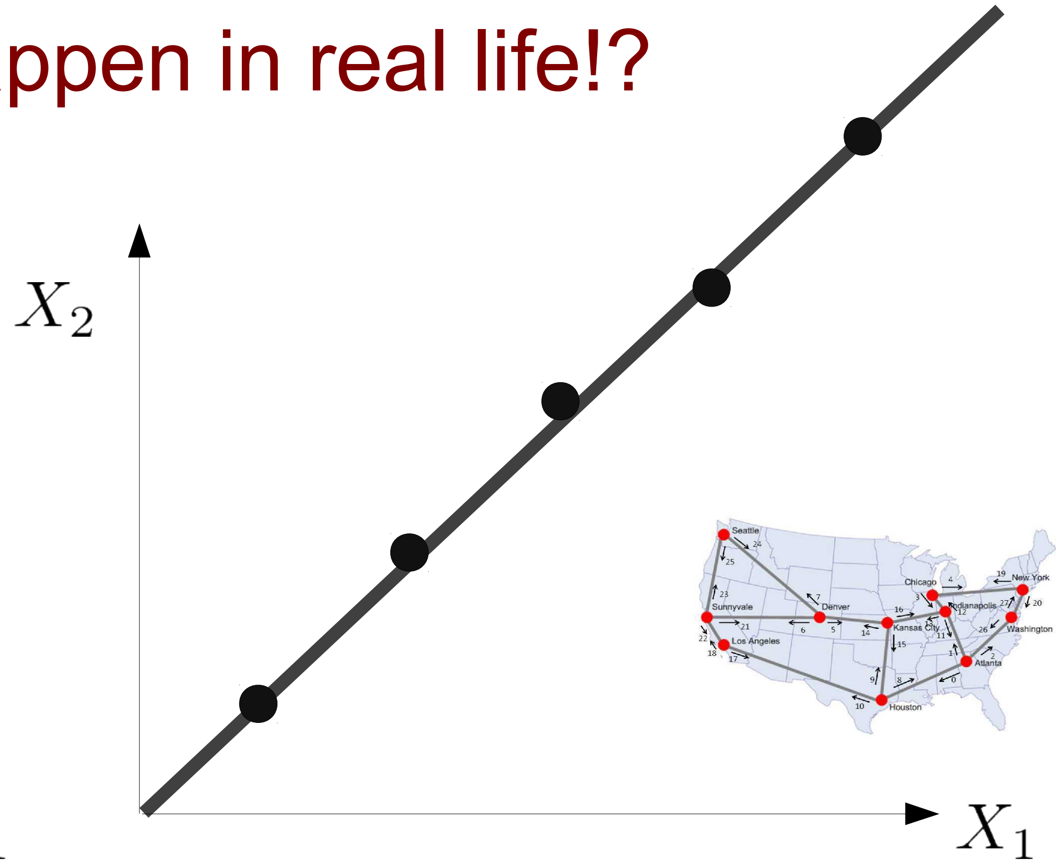
$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{|V|} \end{bmatrix} =$$ 

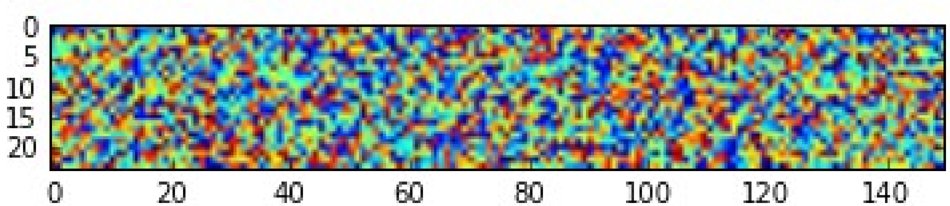
In essence, the data is predictable

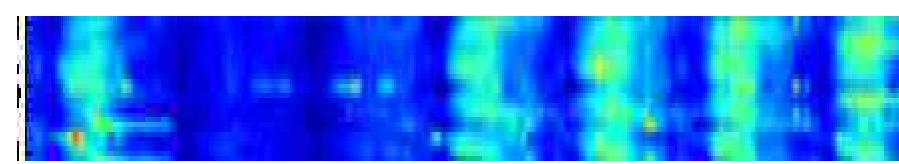
v.s.

Does this happen in real life!?









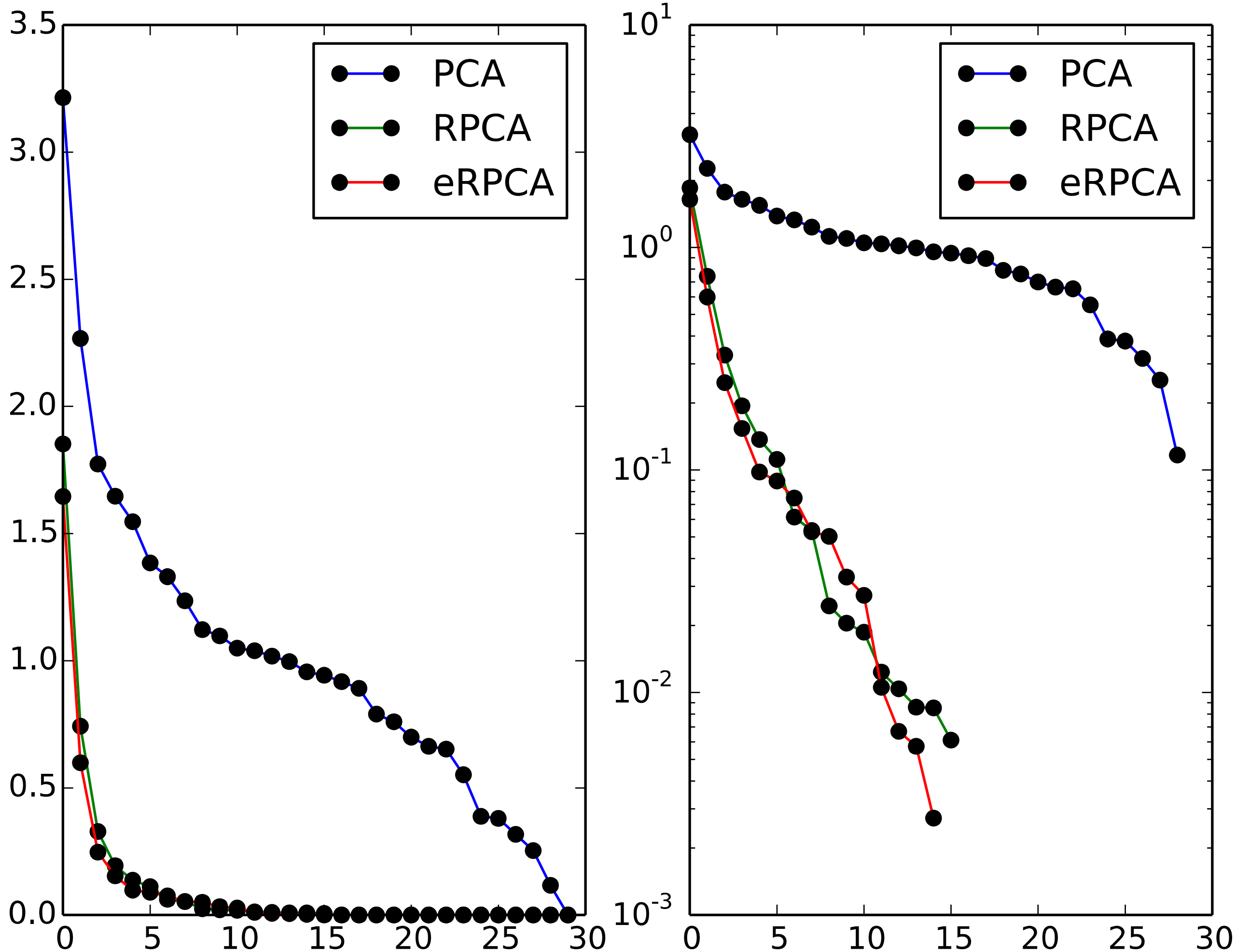
The appropriate structures appear all over the place in real data!

Elisa Rosales

WPI

Singular Values of Matrices

Insurance Satisfaction Surveys



WPI

The appropriate structures appear all over the place in real data!

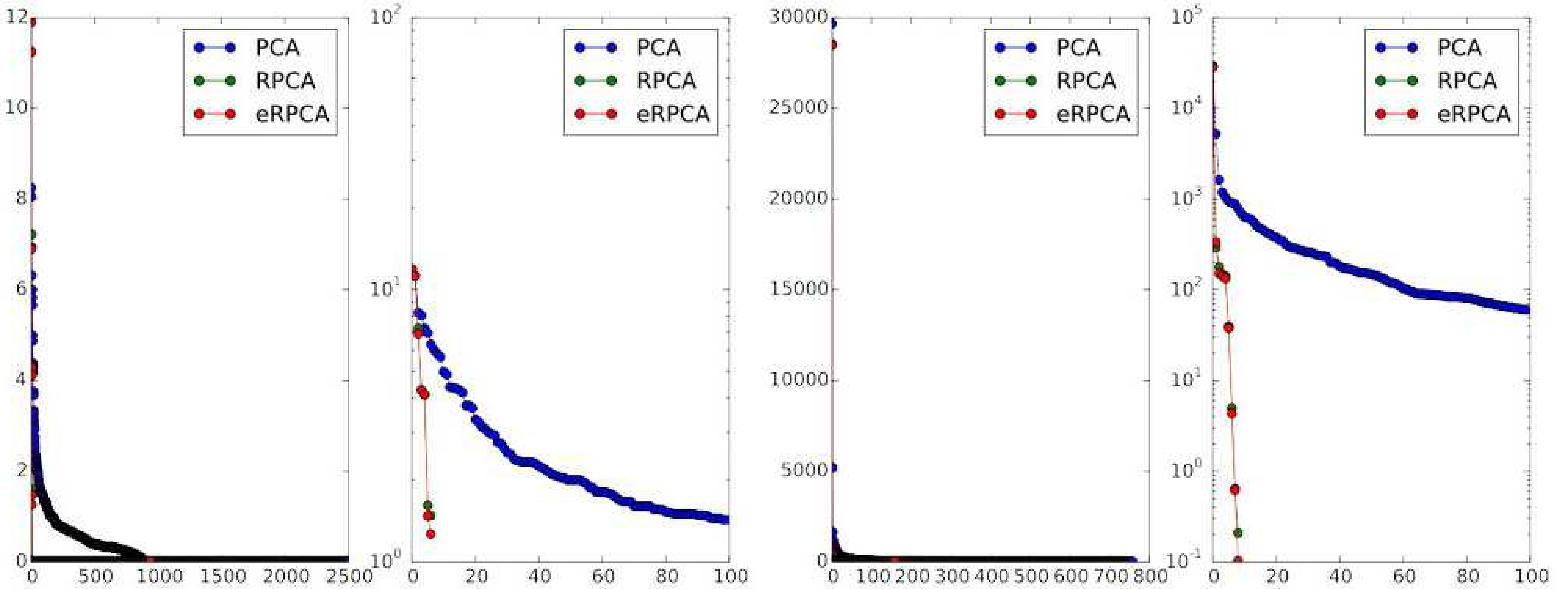
Rakesh Biradar

WPI

Singular Values of Matrices

Amazon product communities

SKAION Internet Attack (e.g., DDoS) simulations



Principal Component Analysis and the Singular Value Decomposition (SVD)

Theorem: Suppose a matrix $Y \in \mathbb{R}^{m \times n}$ (or $\mathbb{C}$), then there exists matrices U, S, and V such that

$$Y = USV^*$$

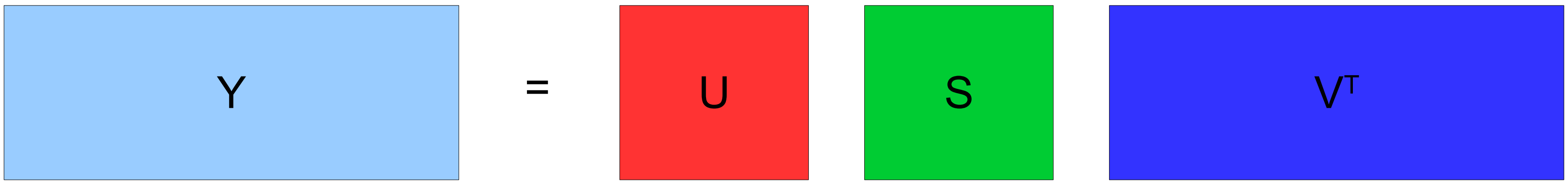
where

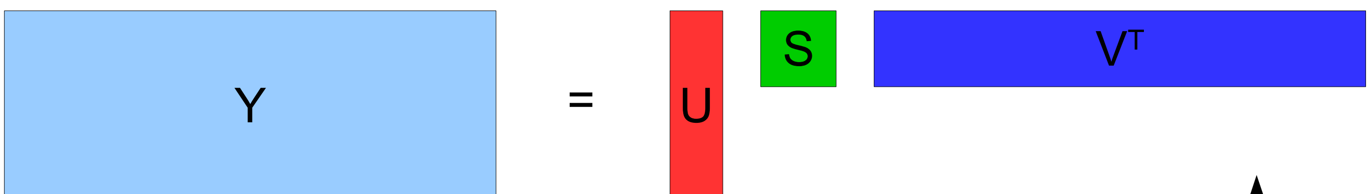
- $U \in \mathbb{R}^{m \times m}$ is unitary,
- $S \in \mathbb{R}^{m \times n}$ is diagonal and all of the diagonal entries are in $\mathbb{R}^+$, and
- $V \in \mathbb{R}^{n \times n}$ is unitary.

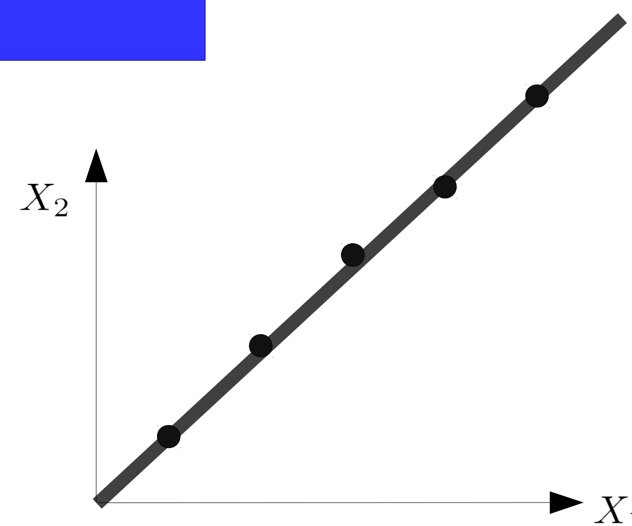
Eckart, C.; Young, G. (1936). "The approximation of one matrix by another of lower rank". Psychometrika 1 (3): 211–8. doi:10.1007/BF02288367.

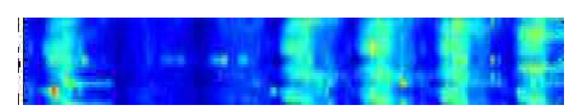
WPI

Low rank matrices

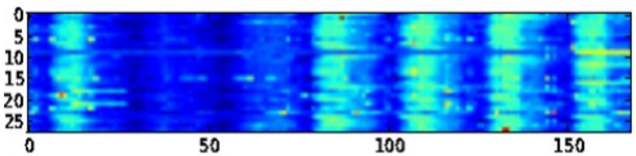




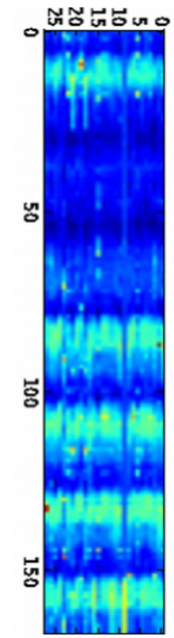




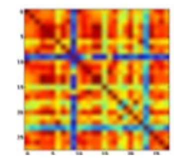
First important (and classic) trick



1 Exabyte of data



=



1 Terabyte of data!

$$Y = USV^T$$

$$\mathrm{PCA}(Y) = US$$

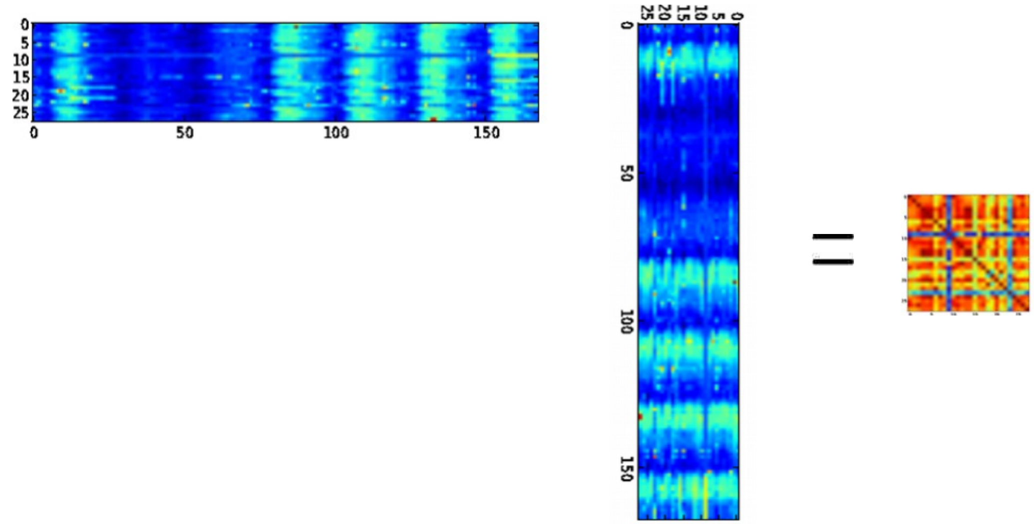
$$\begin{aligned} YY^T &= (USV^T)(USV^T)^T \\ &= USV^T V S^T U^T \\ &= USS^T U^T \\ &= US^2 U^T \end{aligned}$$

Good news: Much smaller
Bad news: Expensive to compute

- Each dot product requires 10^{12} multiplications.
- There are $10^6 * 10^6 = 10^{12}$ total dot products that need to be computed.
- Total cost is $10^{12} * 10^{12} = 10^{24}$ multiplications.

- However, the calculation is embarrassingly parallel, but the data requirements are substantial (1 exabyte in fact).
- Can this be done efficiently?

Yes!



WPI

Standing on the shoulders of giants: Some pretty cool **mathemathematics**

•Over the past 4-5 years there has been a flurry of activity on this problem, much of which we suspect the current audience is aware of.

- Ideas such as matrix completion, robust principal component analysis, and robust matrix completion have generated a lot of interest, including among us!

$$M = U\Sigma V^T$$

Eckart, C.; Young, G. (1936). "The approximation of one matrix by another of lower rank". Psychometrika 1 (3): 211–8.

Matrix completion: The Netflix problem!

$$L = \arg\min_{L_0} \|L_0\|_*$$
$$\text{s.t. } P_\Omega(L_0) = P_\Omega(M)$$

E.Candes and B.Recht, "Exact matrix completion via convex optimization," Foundations of Computational Mathematics, vol. 9, pp. 717–772, December 2009.

$$L = \arg\min_{L_0} \|L_0\|_*$$
$$\text{s.t. } \|P_\Omega(L_0) - P_\Omega(M)\|_F < \delta$$

E. Candes and Y. Plan, "Matrix Completion With Noise," Proceedings of the IEEE, vol.98, no.6, p.11, 2009

Robust principal component analysis

$$L, S = \arg\min_{L_0, S_0} \|L_0\|_* + \lambda\|S_0\|_1$$
$$\text{s.t. } \|M - L_0 - S_0\|_F \le \delta$$

Z. Zhou, X. Li, J. Wright, E. Cande`s, and Y. Ma, "Stable Principal Component Pursuit," ISIT 2010: Proceedings of IEEE International Symposium on Information Technology, 2010.

$$L, S = \arg\min_{L_0, S_0} \|L_0\|_* + \lambda\|S_0\|_1$$
$$\text{s.t. } P_\Omega(L_0 + S_0) = P_\Omega(M)$$

E. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," J. ACM, vol. 58, pp. 11:1–11:37, June 2011.

Our contribution

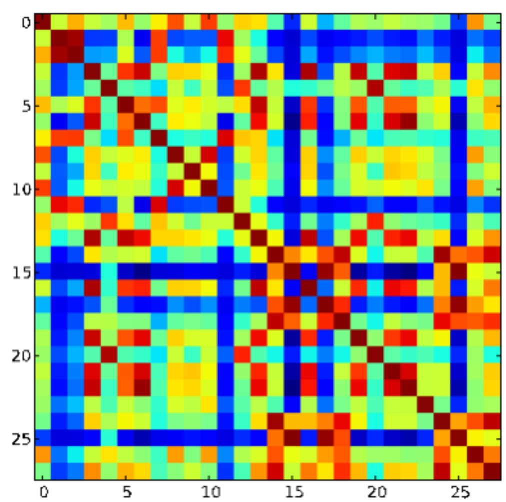
$$L, S = \arg\min_{L_0, S_0} \|L_0\|_* + \lambda\|S_0\|_1$$
$$\text{s.t. } \|P_\Omega(L_0 + S_0) - P_\Omega(M)\| \preceq \epsilon$$

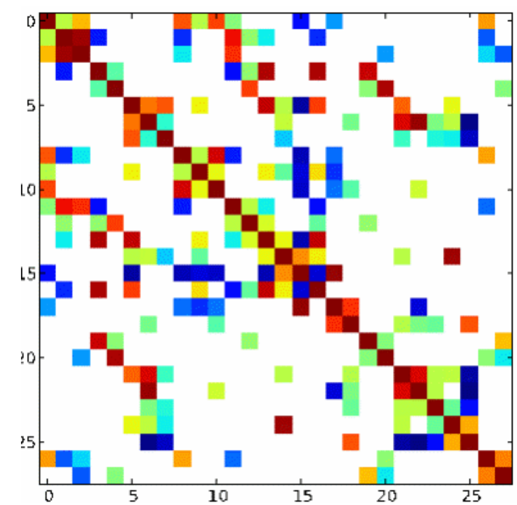
R. Paffenroth, P. Du Toit, R. Nong, L. Scharf, A. Jayasumana and V. Bandara
Space-time signal processing for distributed pattern detection in sensor networks
IEEE Journal of Selected Topics in Signal Processing, Vol. 7, No.1, February 2013
P. Du Toit, R. Paffenroth, R. Nong Stability of Principal Component Pursuit with Point-wise Error Constraints in preparation 2012.

Ω!

$$L = \arg\min_{L_0} \|L_0\|_* \quad \text{s.t. } P_\Omega(L_0) = P_\Omega(M)$$

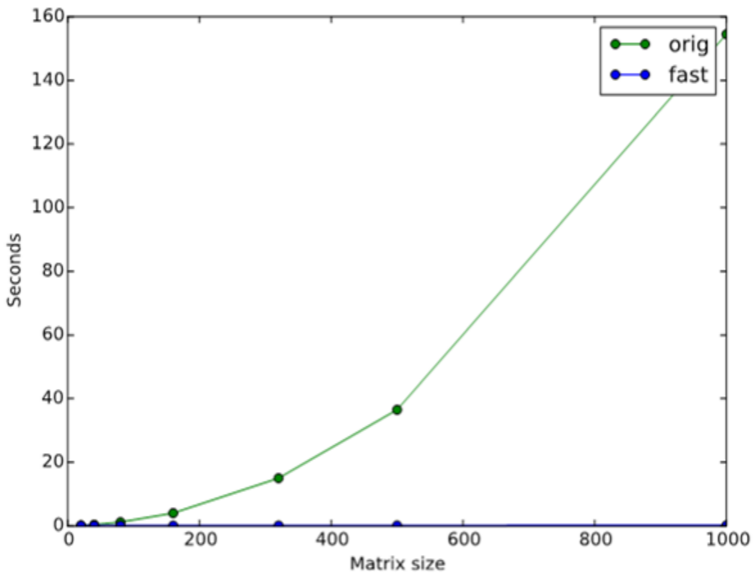
$$L, S = \arg\min_{L_0, S_0} \|L_0\|_* + \lambda \|S_0\|_1 \quad \text{s.t. } \|P_\Omega(L_0 + S_0) - P_\Omega(M)\| \preceq \epsilon$$

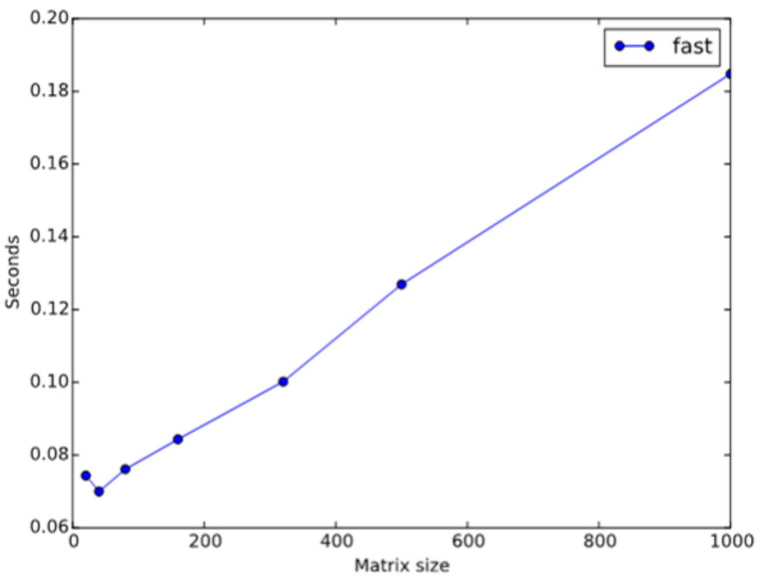




Big Data

Original algorithm. Rank=2, probability of corruption=2%, observations=10 * m
and new algorithm!

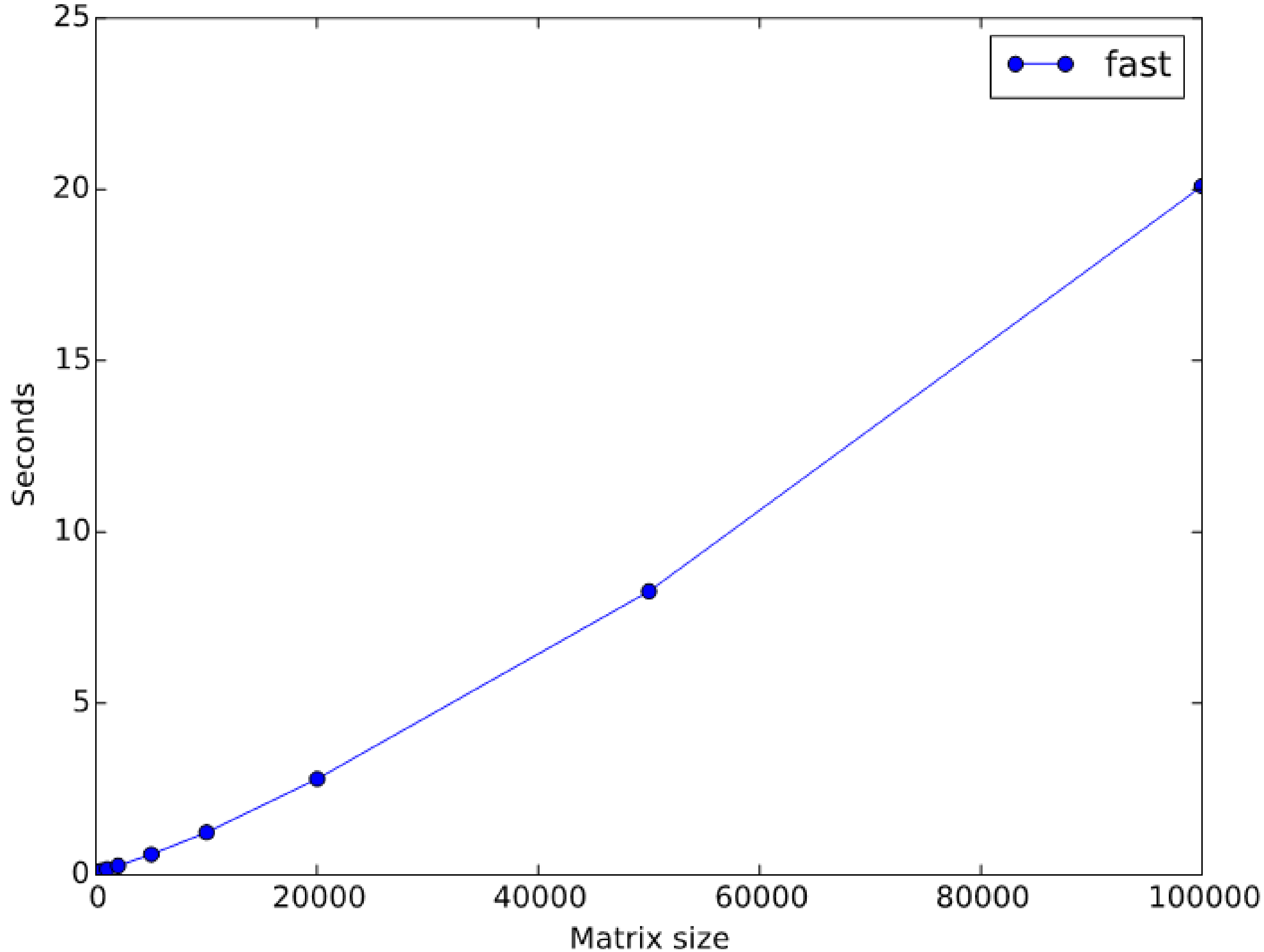




- R. Paffenroth, R. Nong, P. Du Toit, **On covariance structure in noisy, big data. Proceedings** Vol. 8857, Signal and Data Processing of Small Targets, October 2013, Oliver E. Drummond; Richard D. Teichgraeber, Editors.

WPI

Big Data



fast
Seconds
Matrix size
0
5
10
15
20
25
0
20000
40000
60000
80000
100000

Ok, wait a second...

- How can this be?
- I mean, just ***reading*** the full covariance matrix would require $\mathcal{O}(n^2)$ operations!
- Yet, I claim that one can compute the SVD of this same matrix in $\mathcal{O}(n)$ operations!?

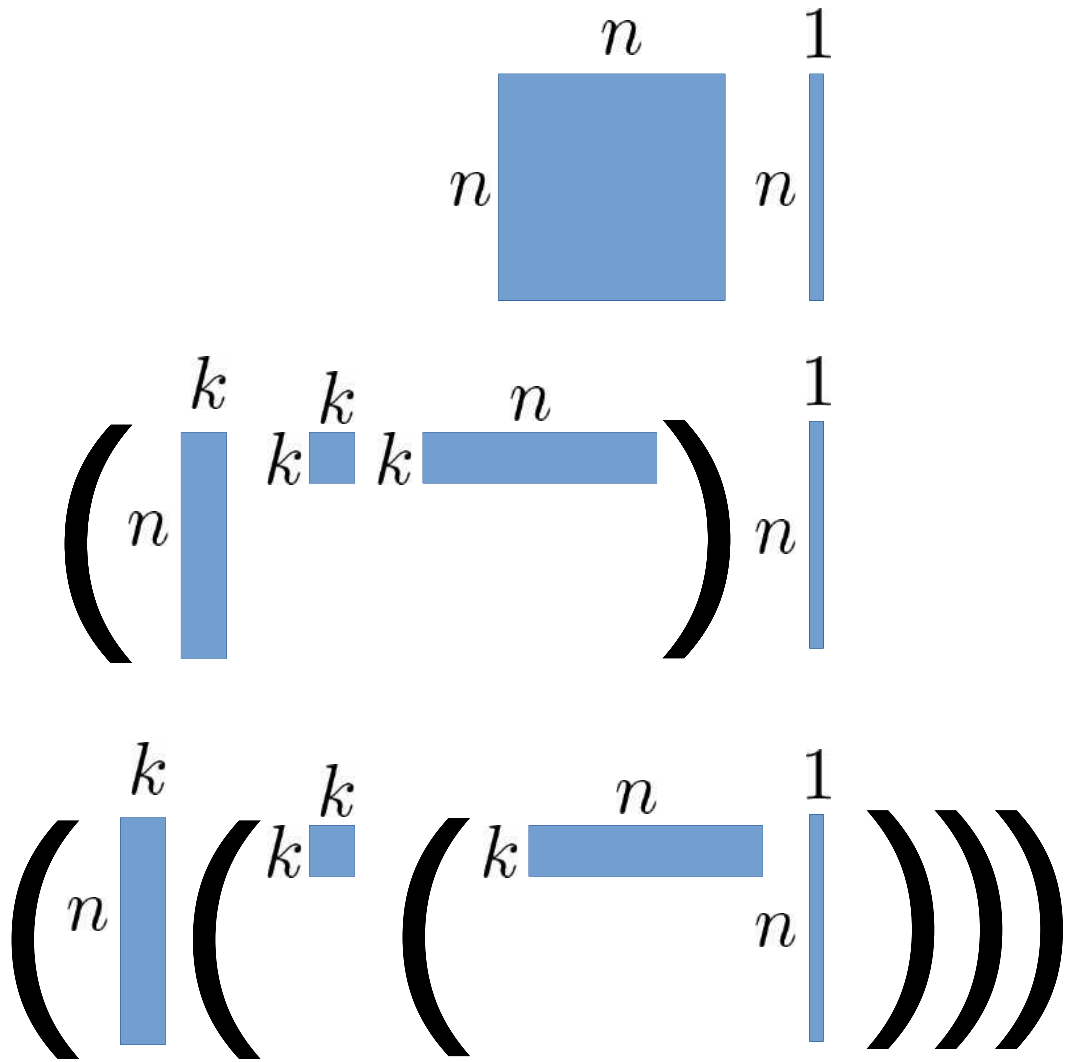
Solution algorithm

- To solve the optimization problem with a fast method that has our desired asymptotic performance properties we use an Augmented Lagrange Multiplier (ALM) formulation which we solve using an Alternating Direction Method of Multipliers (ADMM).
  - The idea is to relax the constraints using a Lagrange multiplier and then split the Lagrangian into two (or more) pieces, each of which is presumed to be easier to minimize than the full Lagrangian.

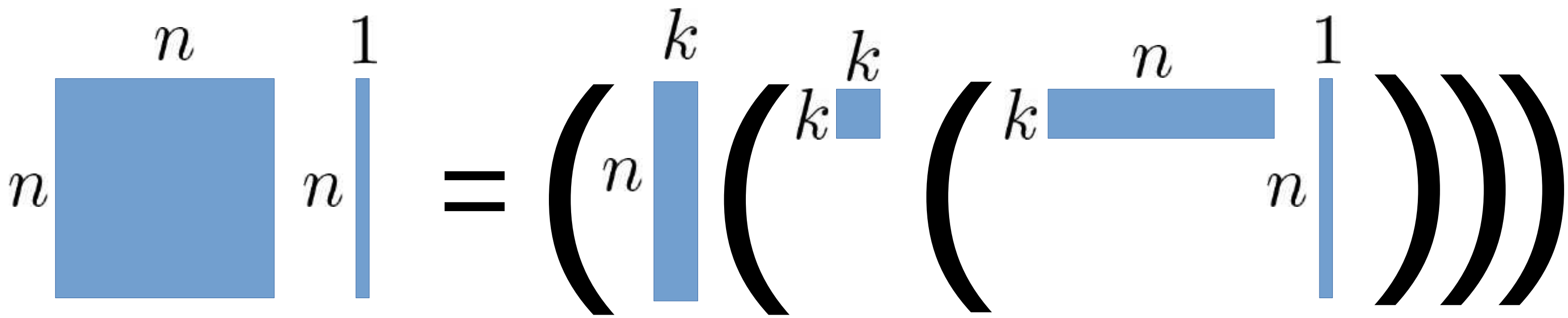
- S. Boyd, “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” Foundations and Trends in Machine Learning, vol. 3, no. 1, pp. 1–122, 2010.
- E. J. Candés, X. Li, Y. Ma, J. Wright, and E. J. Candes, “Robust Principal Component Analysis?,” Journal of the ACM, vol. 58, no. 3, 2011.

WPI

How can this be? Low rank helps...



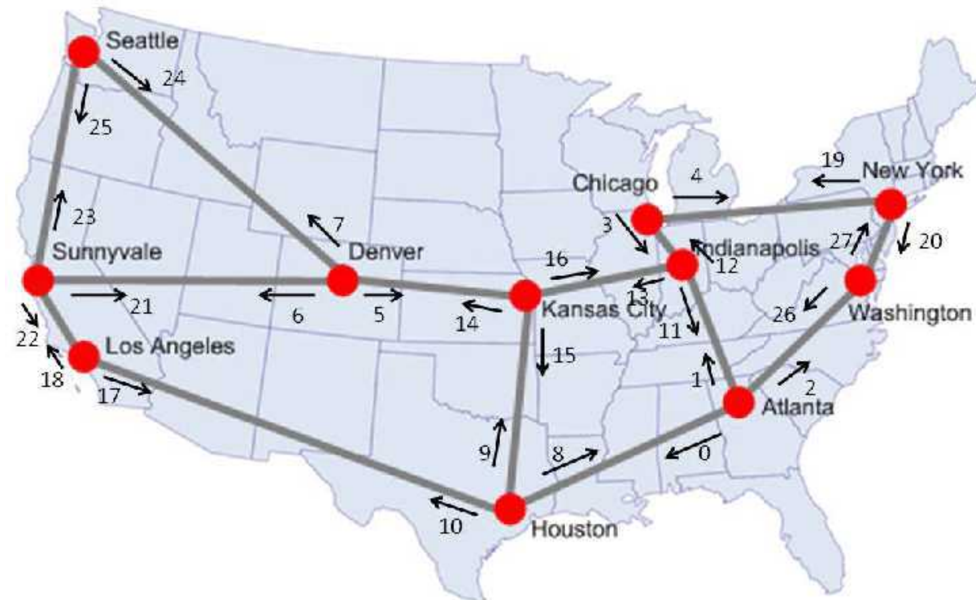
How can this be? Math helps...

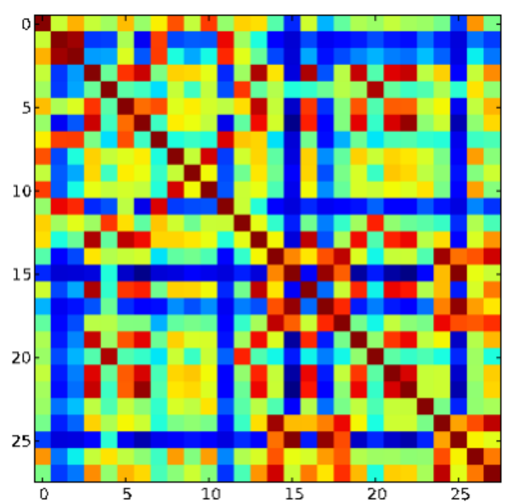


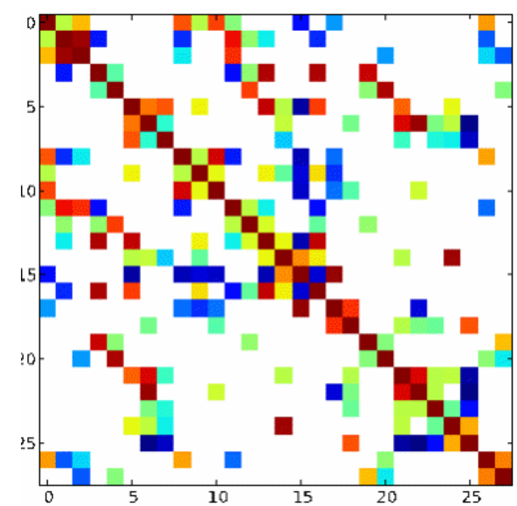
- Matrix free methods
  - Quite popular in the HPC community
- SVD
  - Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions." SIAM review 53.2 (2011): 217-288.

WPI

How can this be? Implementation helps...

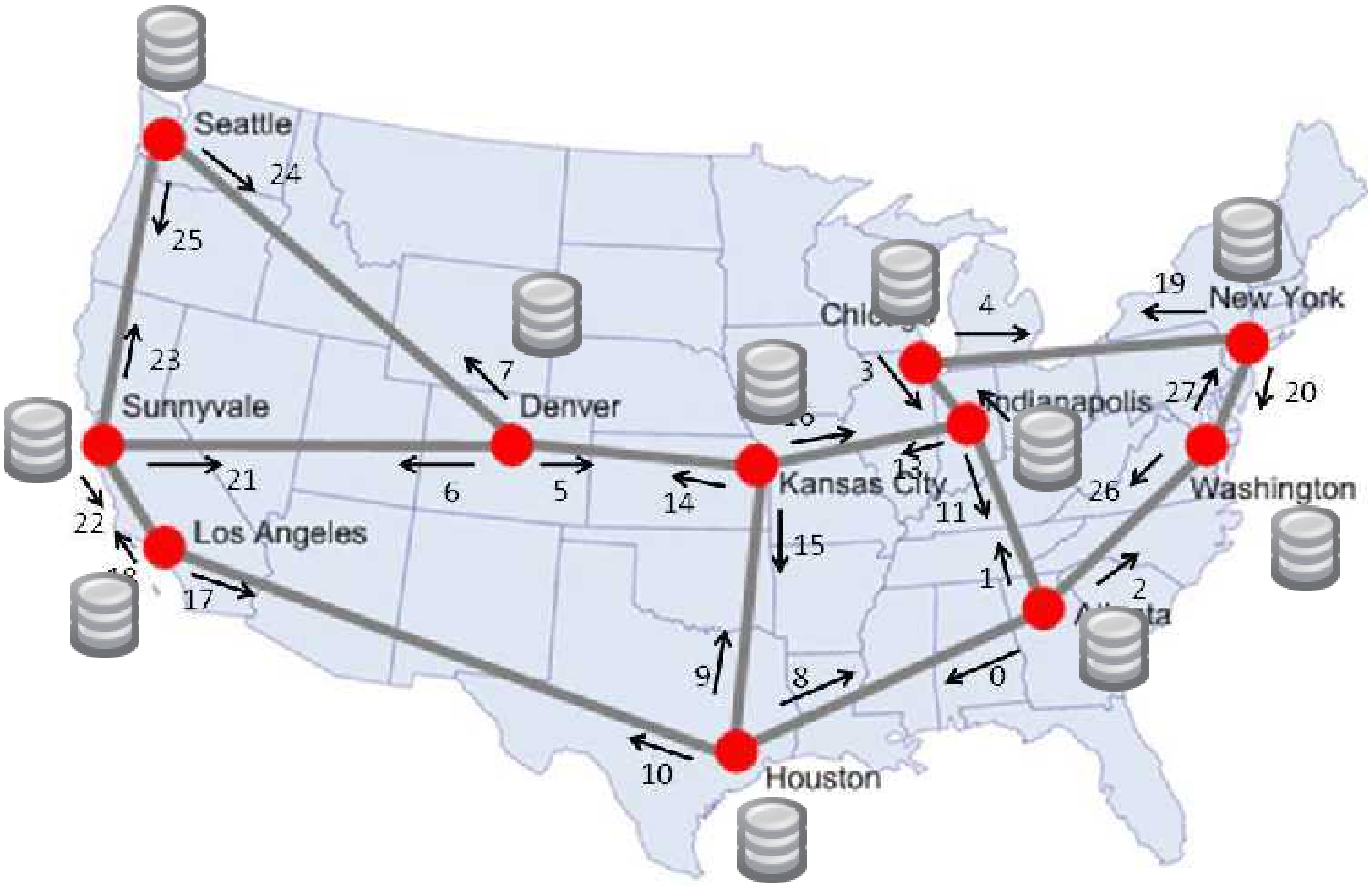




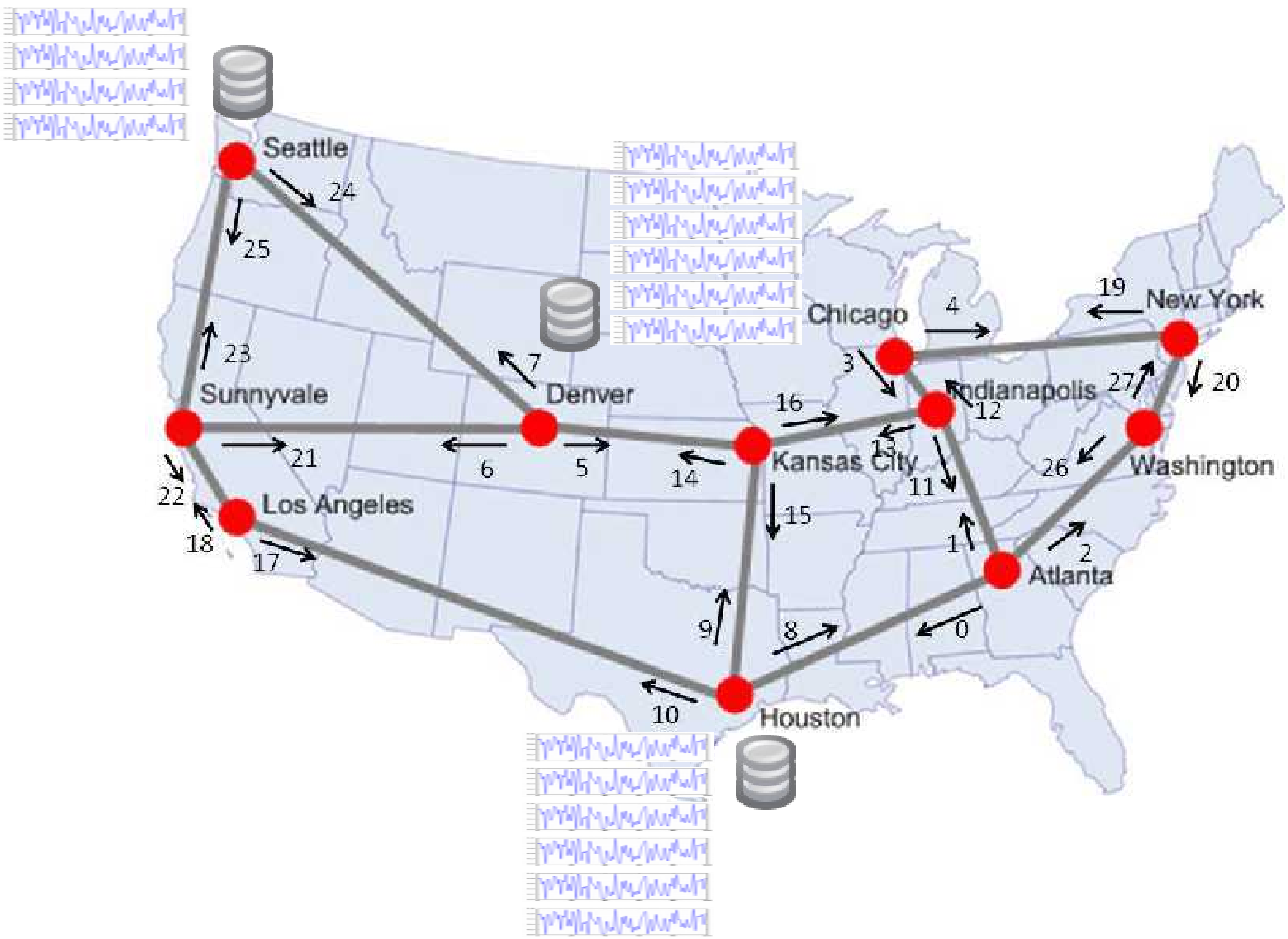


Think about as distributed databases.

Distributed databases.

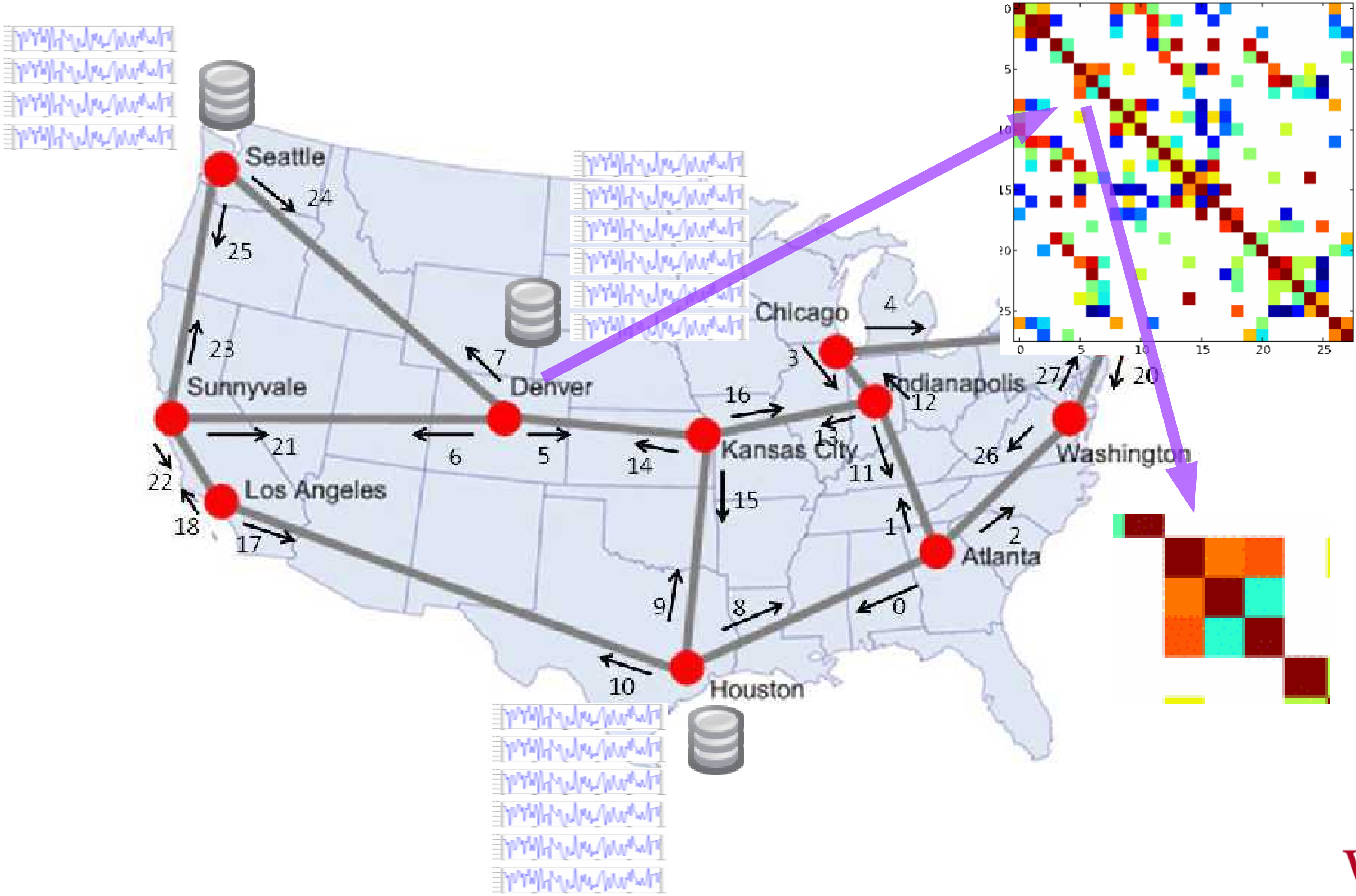


Distributed databases.



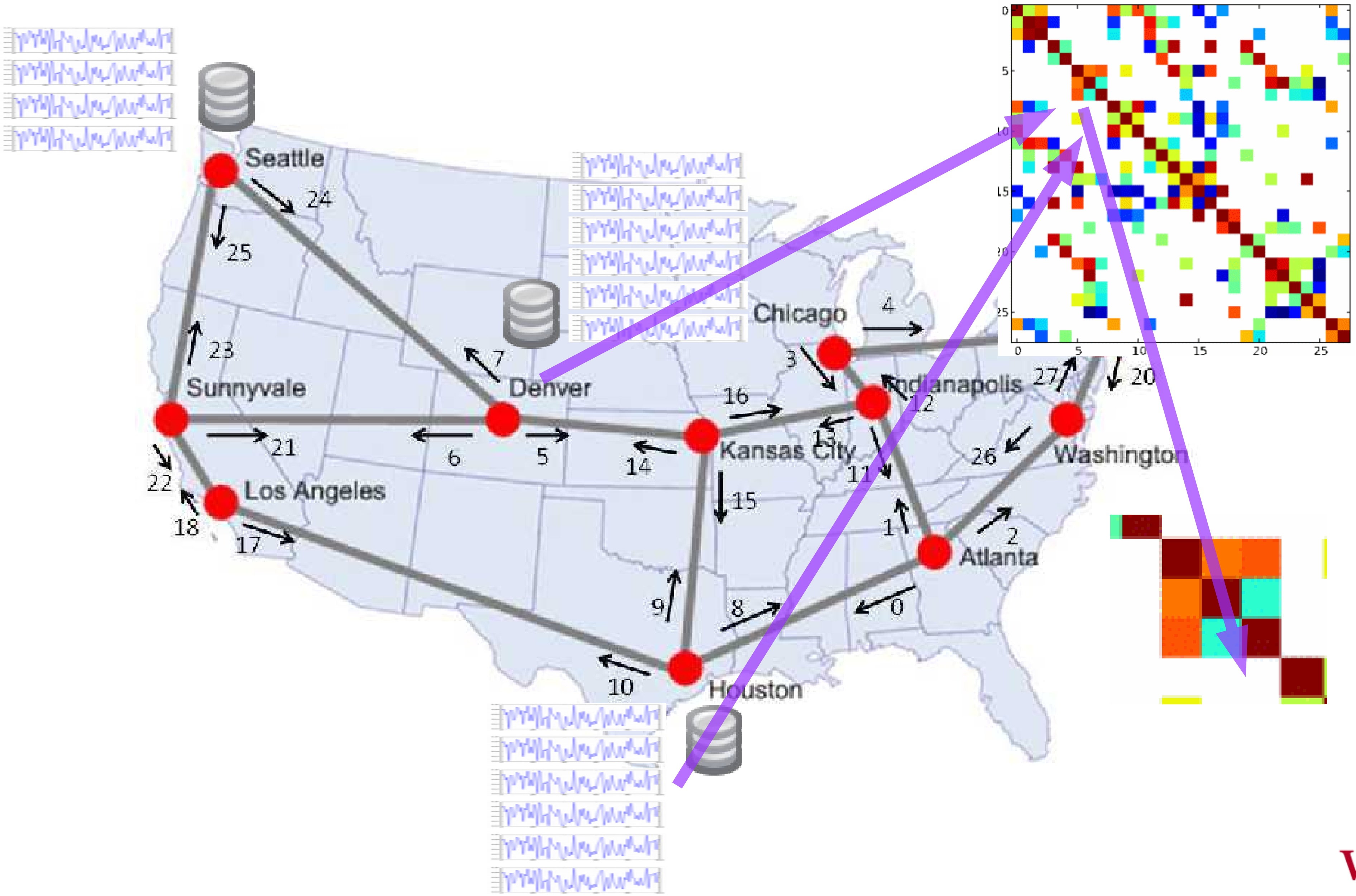
WPI

Distributed databases.

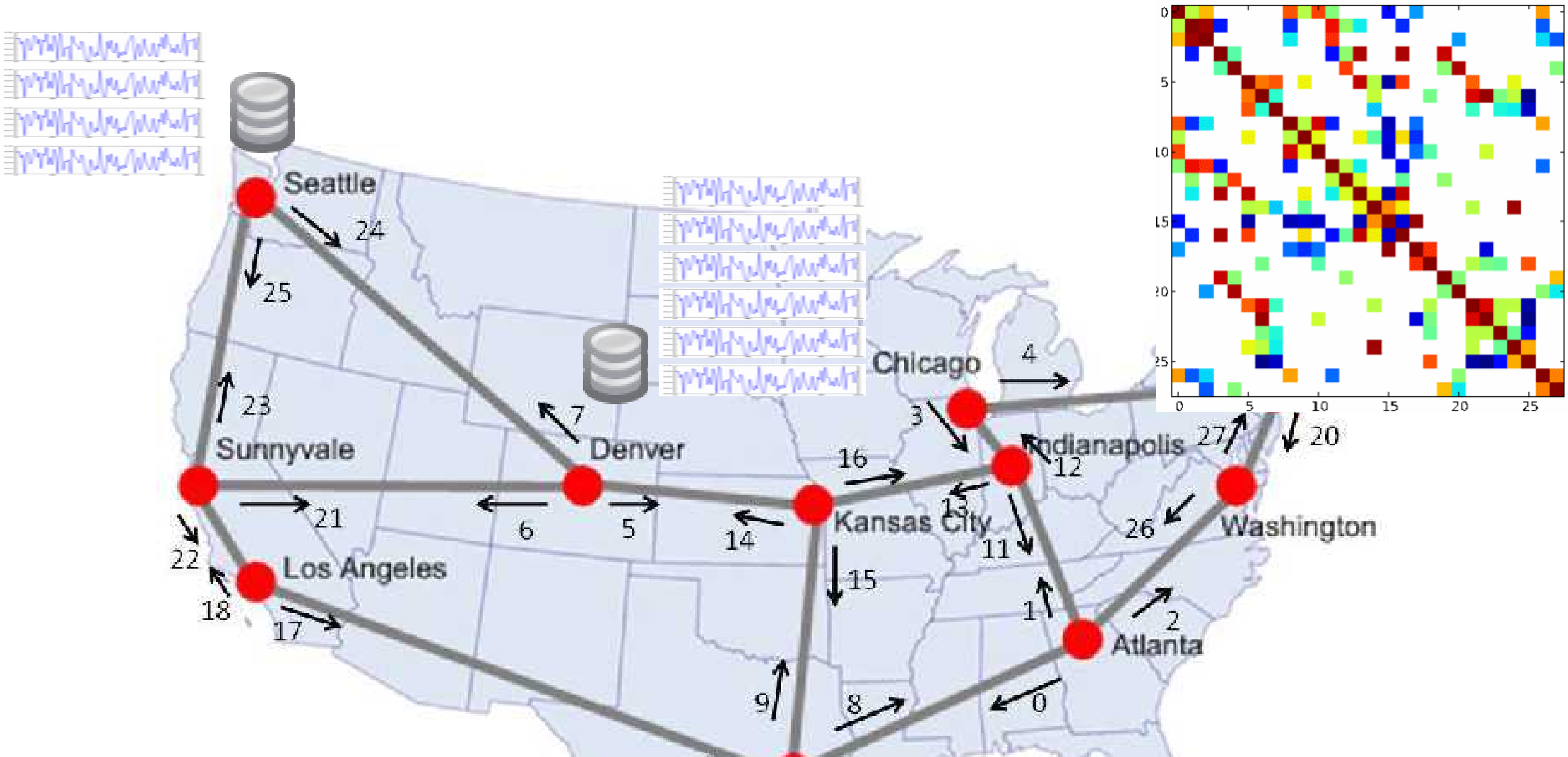


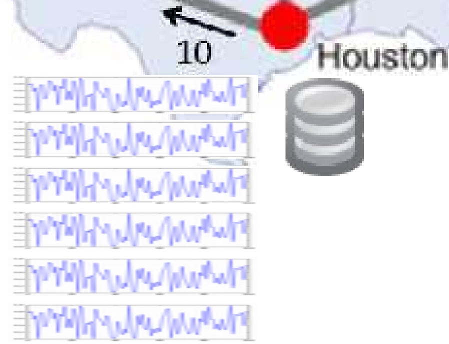
WPI

Distributed databases.



Distributed databases.





No terabyte time series need to be transmitted, just scalars!

WPI

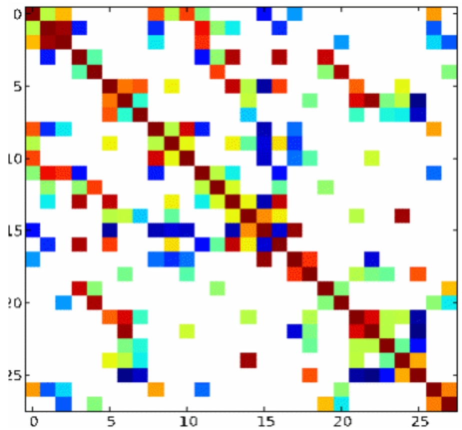
Large Scale Data Science

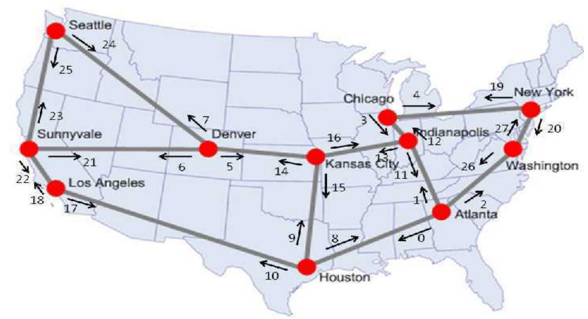
Computer Science **Math**

By Holger Motzkau 2010, Wikipedia/Wikimedia Commons (cc-by-sa-3.0), CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=11115505

WPI

Questions?





$$\min_{L,S} \|L\|_* + \lambda\|S\|_1$$

$$\text{s.t. } |P_\Omega(M) - P_\Omega(L+S)| \preceq \bar{\epsilon}$$