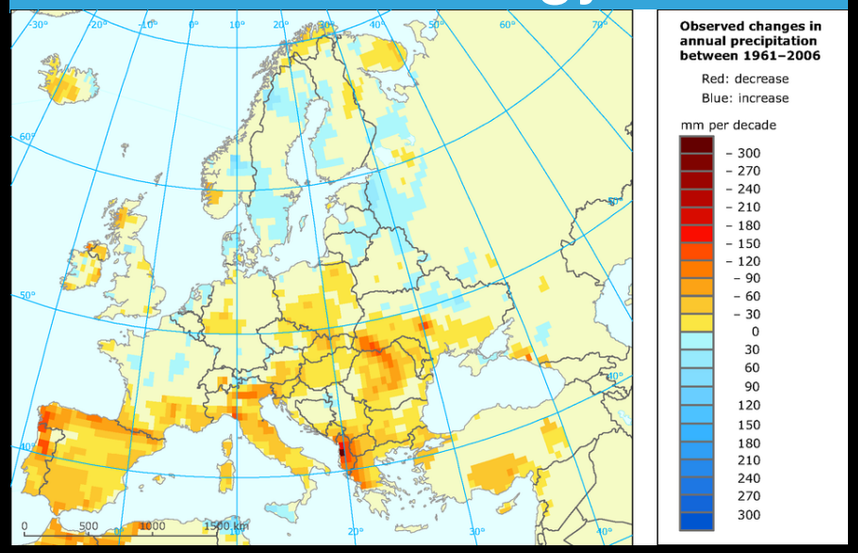
# Subspace Clustering with Missing and Corrupted Data

Zachary Charles (UW-Madison)

Joint with Amin Jalali and Rebecca Willett (UW-Madison)

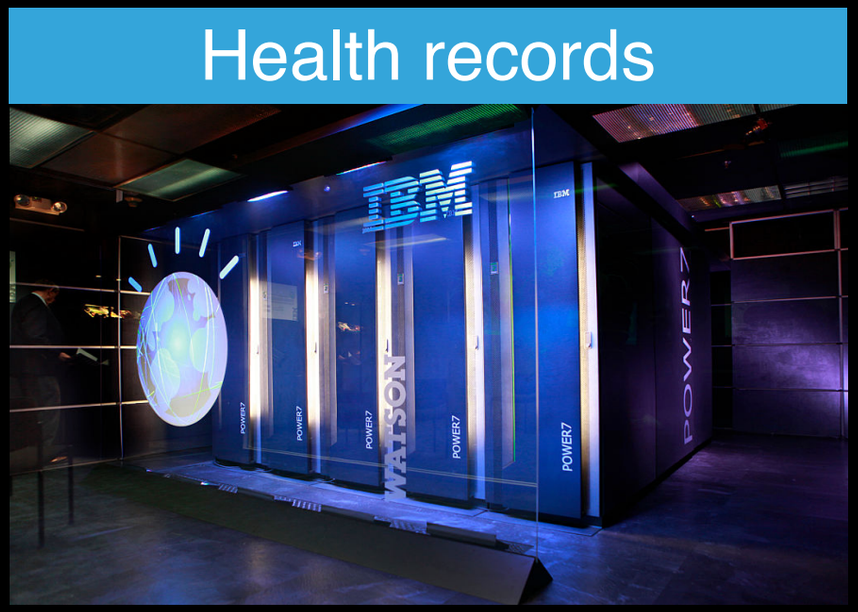
# Real data is messy (or missing)

## Climatology




## Genomics

## Health records



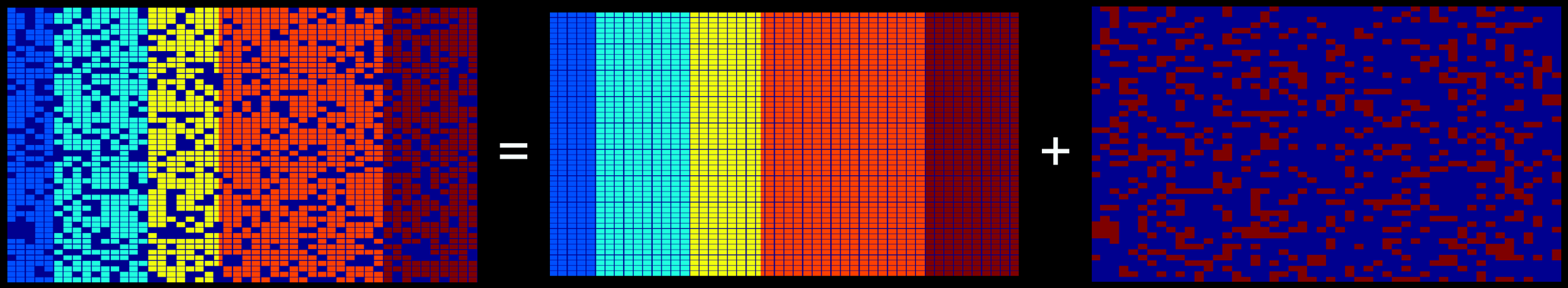

## Recommender systems

amazon

How do we remove noise and fill in missing values?

# Matrix completion

- We assume the data is inherently low rank.
- Allows us to impute missing data with convex optimization. [Candès, Tao, 2009].
- Extensions allow for missing data and noise. [Voltchinskii, 2011] [Klopp, 2012]



http://perception.csl.illinois.edu/matrix-rank/home.html

Issue: Most data is not low rank.

# Union of subspaces

- What if the data comes from a union of low-dimensional subspaces?

Image segmentation

# Union of subspaces model

- Data matrix $X = Y + Z$.
- $Z =$ corruption matrix.
- $Y$ comes from a union of d-dimensional subspaces:

$$S_1 \cup S_2 \cup \cdots \cup S_L$$

- Cluster $X$ by the subspaces, reduce to low-rank matrix completion.

New problem: Subspace clustering

# Subspace clustering

- Key idea: self-expressivity [Elhamifar, Vidal, 2009].
- Each column is a *sparse* linear combination of other columns from the same subspace.

- Optimization formulation:

$$\min_{c} \|c\|_0 \text{ s.t. } x_i = Xc,\ c_i = 0$$

# Sparse subspace clustering (SSC):

- SSC [Elhamifar, Vidal, 2009]:

$$\min_{c} \|c\|_1 \text{ s.t. } x_i = Xc, \ c_i = 0$$

- LS-SSC [Soltanolkotabi et al., 2014] [Wang, Xu, 2016]:

$$\min_{c} \|c\|_1 + \frac{\lambda}{2}\|Xc - x_i\|_2^2$$

- Prior work focused on the amount of noise this can tolerate.
- Success criteria: no false positives.

# Main results

- $L$ $d-$dimensional randomly selected subspaces from $\mathbb{R}^n$.
- $\Omega(d)$ samples are drawn randomly on the unit sphere from each subspace.
- $d = O(n/\#\text{ of samples})$.

Additive noise of norm $\delta$ OR $M$ missing entries per sample

| | $\delta$ | $M$ |
|---|---|---|
| Wang, Xu | $O(1/d)$ | $O(n/d^2)^*$ |
| C., Jalali, Willett | $O(1/\sqrt{d})$ | $O(n/d)$ |

# Contributions

- We guarantee success for additive noise bounded by:
  - The alignment of the subspaces (subspace incoherence).
  - The distribution of points in a subspace (inradius).
- Novel subspace incoherence definition leads to better bounds.

- Extend additive noise case to missing data.
  - Randomly zeroing out entries $\equiv$ projecting on to random axis-aligned subspaces.
  - Apply Johnson-Lindenstrauss style results.

# Subspace incoherence

Primal: $$\min_{c} \|c\|_1 + \frac{\lambda}{2}\|Ac - x\|_2^2$$

Dual: $$\max_{\nu} \langle x, \nu \rangle - \frac{1}{2\lambda}\|\nu\|_2^2 \; s.t. \; \|A^T\nu\|_\infty \leq 1$$

$$\nu / \|\nu\|_2 = \text{Dual direction}$$

Our subspace incoherence: Maximum inner product of the dual vectors and the uncorrupted samples.

Prior subspace incoherence: Maximum inner product of the projected dual vectors and the uncorrupted samples.

By avoiding projection, we can better measure the affinity between the corrupted and the true subspaces.

# Contributions

- LS-SSC can be used location agnostically.

$$\min_{c} \|c\|_1 + \frac{\lambda}{2}\|Xc - x_i\|_2^2$$

- Assume missing entries are set to zero.
- Zeros may be a observed zeros or missing.
- Allows use in presence-only data settings:
  - Population sampling.
  - Disease screening.

# Conclusion

- Low rank assumptions may not hold true in general.
- Union of subspaces model can explain full rank data.
- Convex analysis, high-dimensional statistics can guarantee subspace clustering methods succeed.

## Open problems:

- How do we guarantee clustering accuracy?
- Information-theoretic limits?
- What about unions of low-dimensional non-linear spaces?

# Fin.