Human-guided machine learning and data mining

Ian Davidson
Collegium de Lyon Fellow 2017-2018
University of California - Davis

Merci Tout La Monde!

- Thank you for the invitation to speak.
- Thanks to the OFII for not deporting me yesterday
- Thanks to the Collegium de Lyon for my fellowship

OBJECTIVES OF THE COLLEGIUM

LYON INSTITUTE OF ADVANCED STUDIES

- Offering to **high-level foreign researchers** the possibility to focus fully on their innovative and original research project
- Building, during medium-length stays (5 or 10 months), a **community of fellows** from all scientific fields, with a predominance of human and social sciences (or other sciences in interaction with them).
- Developing **long-term partnerships** between the labs of the University of Lyon (certified as university of excellence-Idex) and the fellows' home scientific institutions

UNIVERSITÉ DE LYON

Big Picture of My Talk

- The traditional machine learning pipeline – when is it good
- Some newer problems that need human involvement
- What role humans can play
  - Some ideas from our group and **limitations**
- A Future Approach? The CP Revolution of DM/ML?

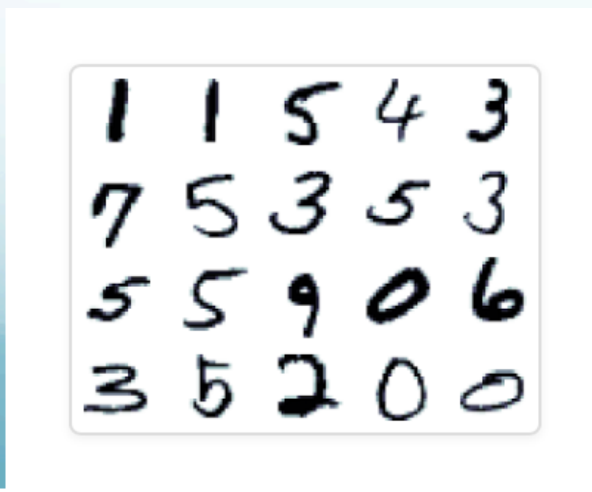
The Typical Learning Pipeline

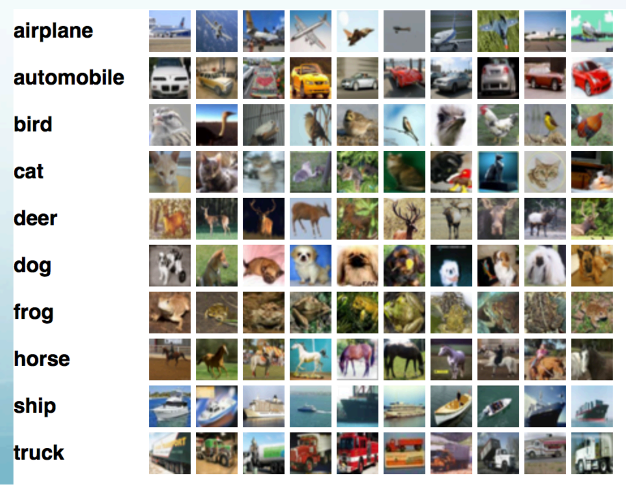
Typical Linear Machine Learning Process

The Typical Learning Pipeline

Typical Linear Machine Learning Process

Data → Learner → Model





Sign Recognition for Google Car

- My student Aubrey Gress spent the summer working at Google so the next driverless car can read signs.

Easy problem
No strong domain knowledge
Easy to annotate instances
Lots of data
Ideal for deep learning

Where The Typical ML Pipeline Does **Not** Work Well

Typical Linear Machine Learning Process

Data → Learner → Model

- Intelligent Tutoring Systems (SoarTech)
- Small Data EHR Records (NMRC)
- Neuroscience (NSF, NMRC, Pennington)
- Trajectory (GPS) (ESI)

Learning in ITS

[With ONR and SoarTech]

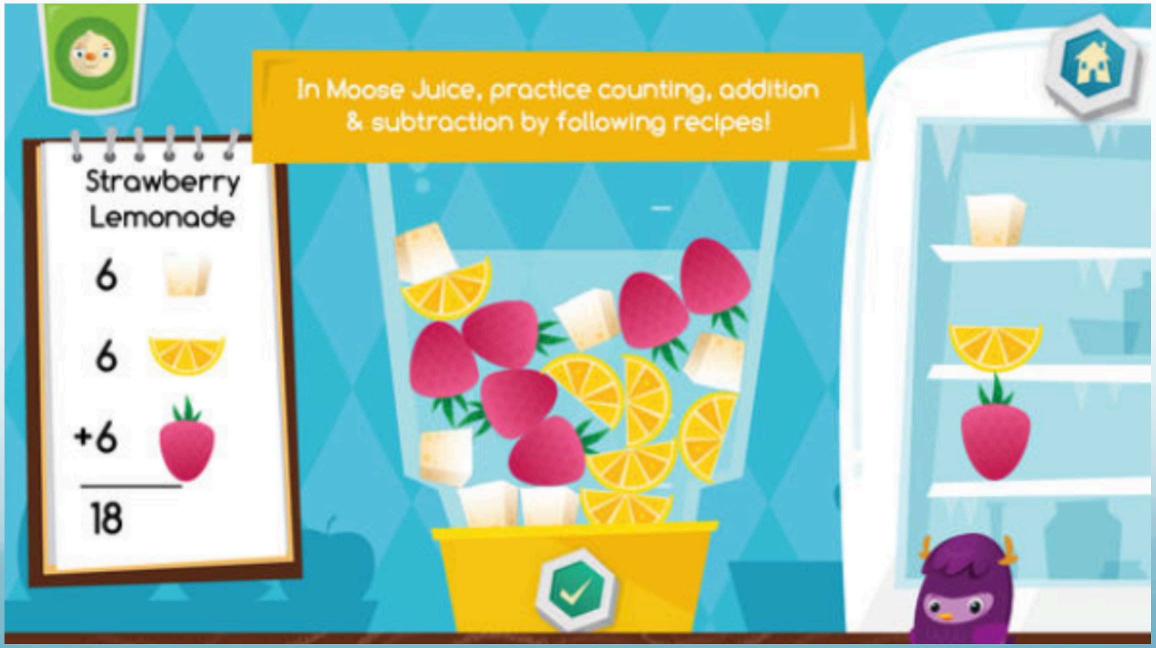
- The future of education?
- Used extensively for small children and DoD!
- Trying to score a person's abilities at many skills

Question 1: There are 3 large marbles, 2 medium marbles and 5 small marbles in a bag. If one of the marbles is chosen randomly, what is the probability that a small marble is chosen?

- ◯ 3/10
- ◯ 1/5
- ◯ 1/3
- ◯ 1/2

Learning in ITS
[With ONR and SoarTech]

- The future of education?
- Used extensively for small children and DoD!
- Trying to score a person's abilities at many skills



Essentially a Big Transfer Learning or Matrix Completion Problem

Skills

Students

	S1	S2	S3	S4	S5	S6	S7	S8	S9	…	S m
St 1											
St 2											
St. .											
St n											

Every time a student answers a question we can fill-in/update a cell
But we will need to then ask lots and lots of questions

But a Domain Expert Can Help Some People Are Smarter!

		S1	S2	S3	S4	S5	S6	S7	S8	S9	…	Sm	
Srini	St 1												$f(\Sigma s_i) = X$
Siegfried	St 2												$f(\Sigma s_i) = Y$
Ian	St. .												$f(\Sigma s_i) = Z$
	St n												

Clearly X >> Z and Y >> Z!

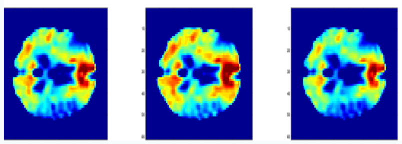
But a Domain Expert Can Help Some People Are Smarter!

	S1	S2	S3	S4	S5	S6	S7	S8	S9	…	S m
St 1											
St 2											
St. .											
St n											

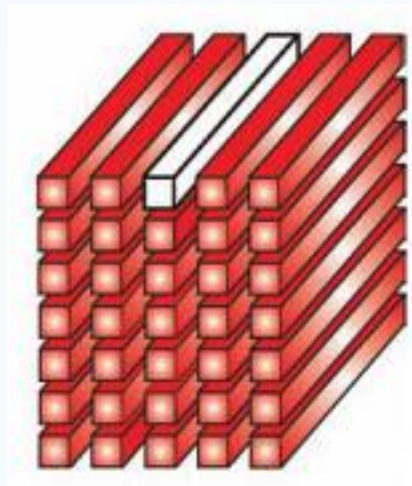
Addition
Multiplication
Long Division

Functional Network Discovery

[With NMRC, Pennington Instiute]

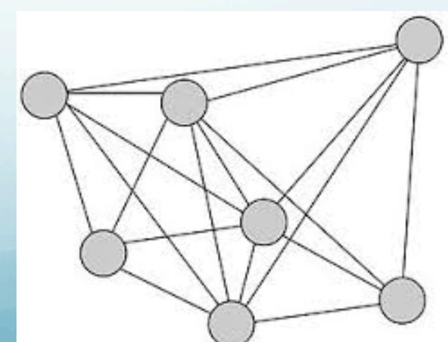


Take functional scans
Co-register with
structural scans



Stack Images Over Time
Each voxel is a time series

Measure correlations over
voxels to construct edge
weights

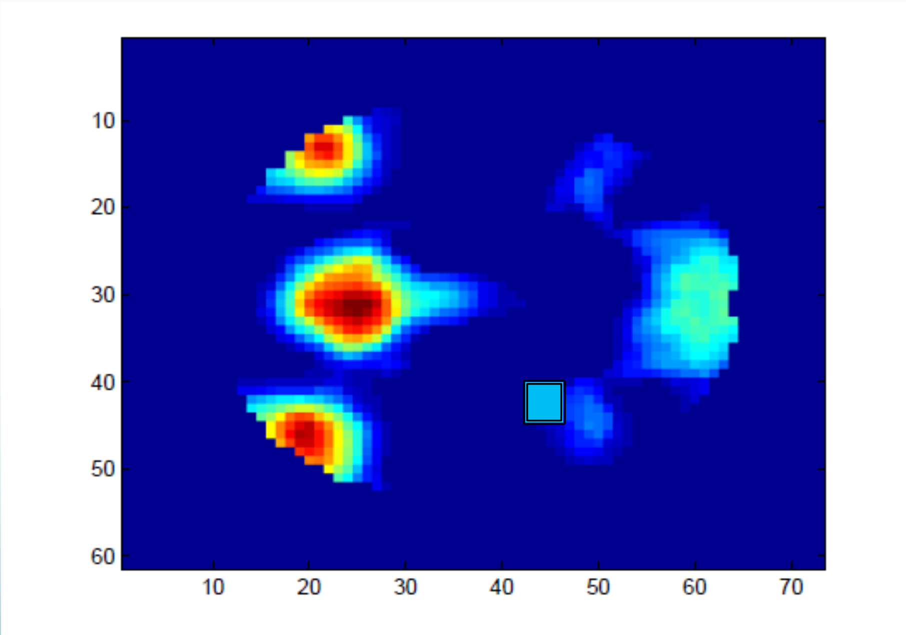


Cutting the graph creates a

i) Foreground community
ii) Background community

Functional Network Discovery

- Synchronized co-activation of spatially separated regions is associated with a **functional network**



Why Human Guidance?

a) Co-activation
b) Lack of wiring
c) Spatial boundaries

Liu et al.: Regional homogeneity, functional connectivity and imaging markers of Alzheimer's disease: A review of resting-state fMRI studies. Neuropsychologia 46, 1648-1656 (2008).
Venkataraman, A. et al.: Exploring Functional Connectivity in fMRI via Clustering. In ICASSP 2009.

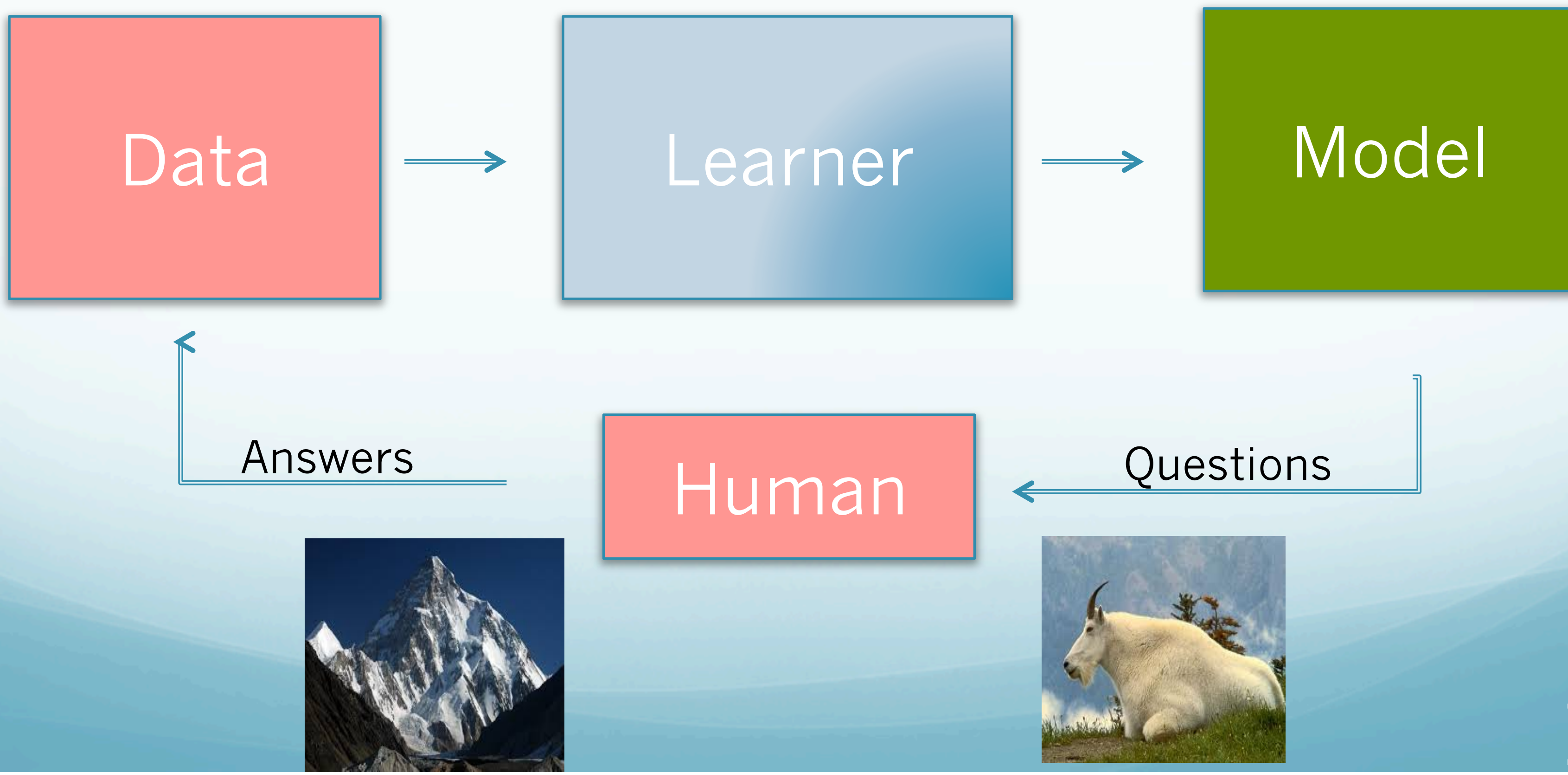
So We Need To Add Human's To ML? How?

Some Typical Ways
1) Human in the Loop Learning

Typically Active Learning

Yahoo! FREP, ONR

ICDM 11, IJCAI 13, SDM 13, ICML 13 …



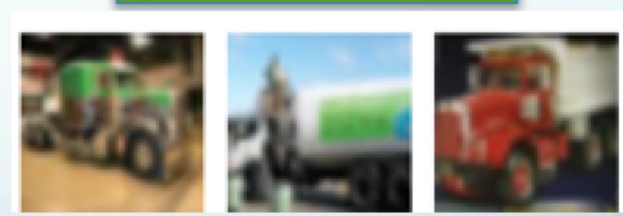
2) Transfer Learning

Transfer in a Related Model

NSF, ONR – Active Transfer Learning Program

ICDM 12, KDD 11/12/13/14 AAAI-15 …

Data → Learner → Target Model



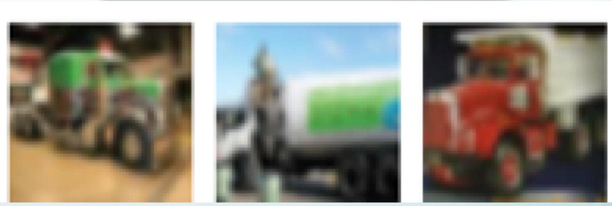
2) Transfer Learning

Transfer in a Related Model

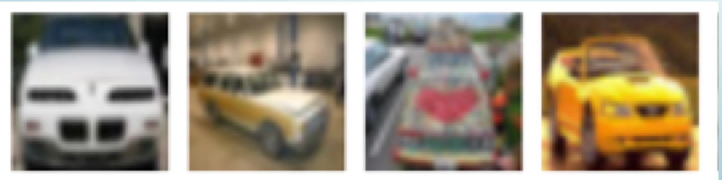
NSF, ONR – Active Transfer Learning Program

ICDM 12, KDD 11/12/13/14 AAAI-15 …

Data → Learner → Target Model



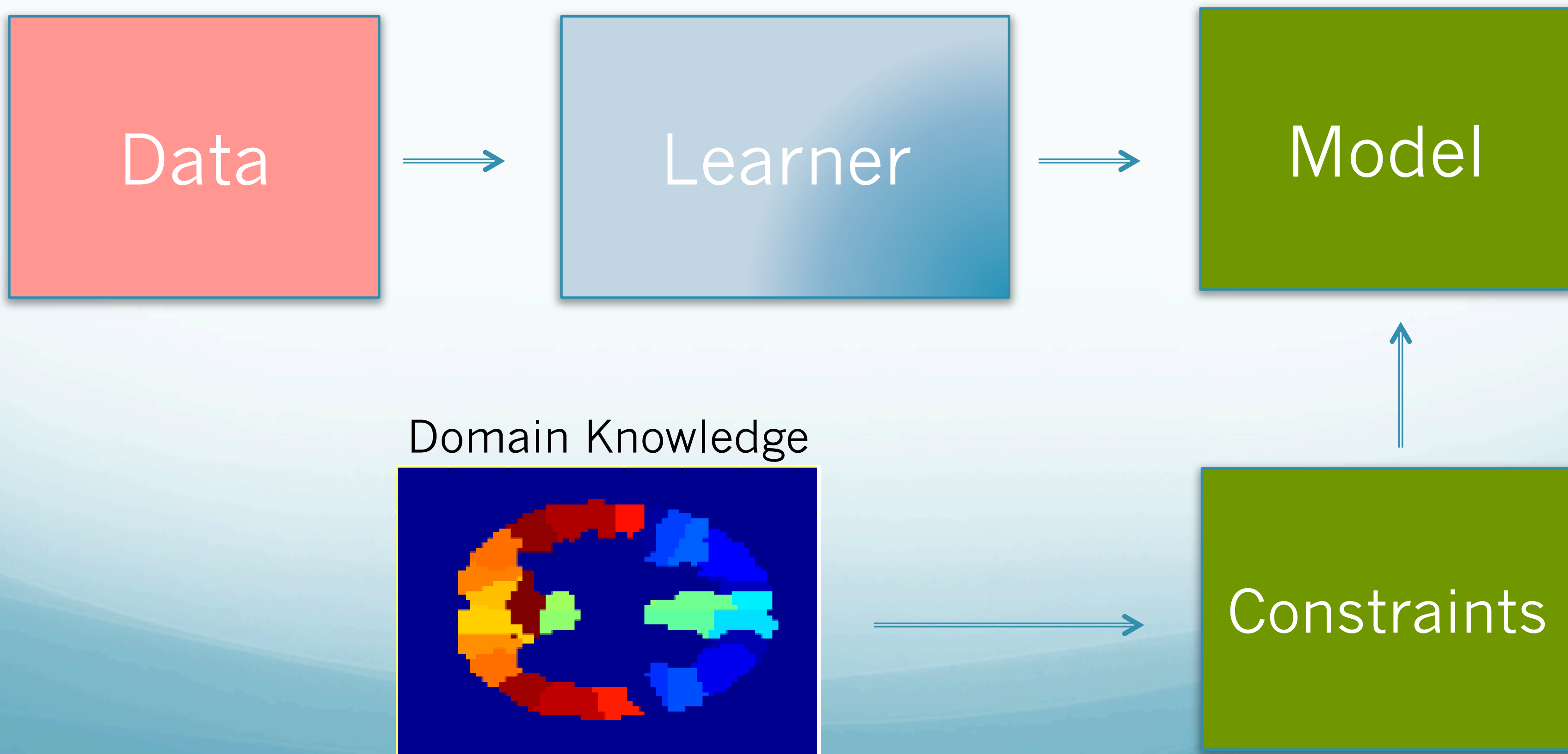
Source Model → Learner



3) Constraining the Model

Adding Human Guidance via Constraints

SDM 05, ECML 07, KDD 10/11/13a-b/15/16/17



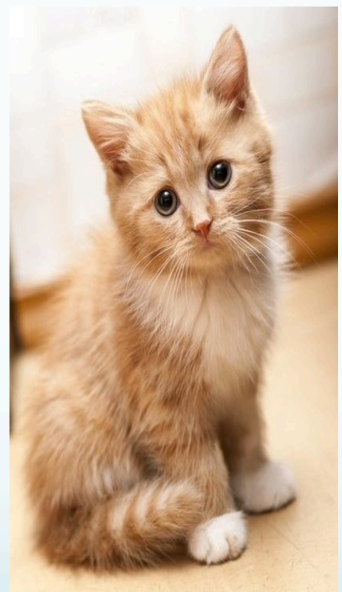
Some Directions of My Group with Limitations

- **Relative guidance – asking humans easier annotations/questions**
  - **IJCAI 13, ICDM 16, AAAI 18**
- Large scale transfer learning – asking humans what tasks are related
  - AAAI 15, ICML 13, AAAI 10, TIPS 16
- **Constrained clustering and block modeling – asking humans what their expectations of clustering should be**
  - **More recently KDD15,17 and ICDM 17**

Supervised Learning and Labeling

- Challenge: Size of output space impacts # annotations
    - Binary classification is simple – only two options

Is this
a cat?



Supervised Learning and Labeling

- Challenge: Size of output space impacts # annotations
    - Multilabel classification is harder

What type of cat is this?

Supervised Learning and Labeling

- Challenge: Size of output space
    - Regression: output could be any real number

How old is this cat?

[Probabilistic Formulations of Regression with Mixed Guidance, Gress, Davidson, ICDM 2016]

- We assume we have some small set of labeled data $(x_1, y_1), \ldots, (x_n, y_n)$ as well as a set of unlabeled data $x_{n+1}, \ldots, x_m$
- But generating accurate labels for the unlabeled data is too expensive or not possible
- How can we make labeling in regression easier for humans?
- Our idea: ask *easier* questions

Our Work: Bound

- Bound: is $f(x_i) \in [a_i, b_i]$?
  - e.g. "Is this house more than $200,000, but less than $300,000?"
  - Providing an exact price may be too demanding a task, so allowing the user to provide a range of values can better model the user's uncertainty in their prediction.

$\in [\$200{,}000, \$300{,}000]$

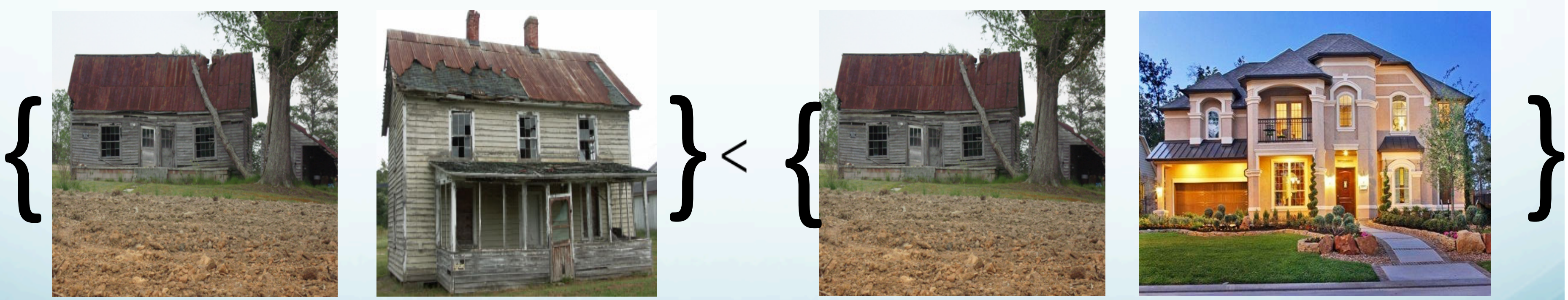
Our Work: Relative

- Relative: is $f(x_i) > f(x_j)$?
  - e.g. “Which of these two houses is more expensive?”
  - Even if the user can’t accurately predict the price of a house, they can probably tell if one house is more valuable than another

<

Our Work: Neighbor

- Neighbor: is $|f(x_i) - f(x_j)| < |f(x_i) - f_k)|$?
    - e.g. “Is house A closer in price to house B or house C?”
    - Given a set of 3 objects, the user can provide which pair of objects’ responses are closest together.



Our Work: Similar

- Similar: is $\left|f(x_i) - f\left(x_j\right)\right| < s$?
  - e.g. "Are the prices of these two houses within $50,000 of each other?"
  - The user may be able to tell if two houses are in roughly the same price range

{   } < [$50,000]

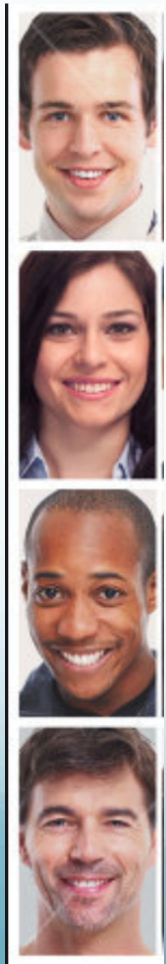
In the End You Have Training Data With Mixed Annotations

Features Annotation - Age

[20-25]

In the End You Have Training Data With Mixed Annotations

Features	Annotation - Age

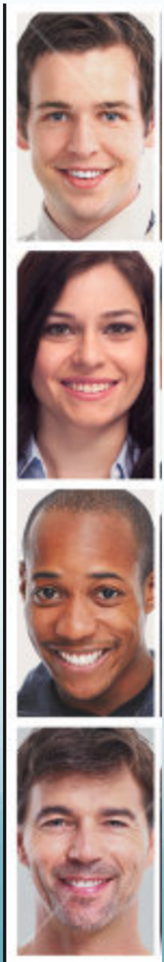


[20-25]

$f(2) \approx f(1)$

In the End You Have Training Data With Mixed Annotations

Features Annotation - Age



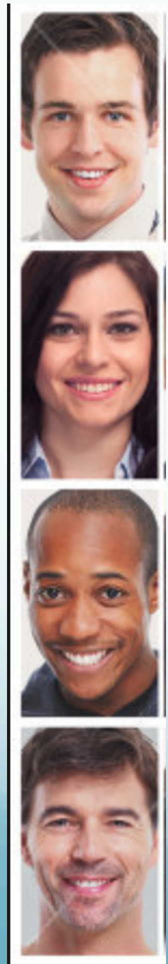
[20-25]

f(2) ≈ f(1)

f(3) > f(1)

In the End You Have Training Data With Mixed Annotations

Features | Annotation - Age



[20-25]

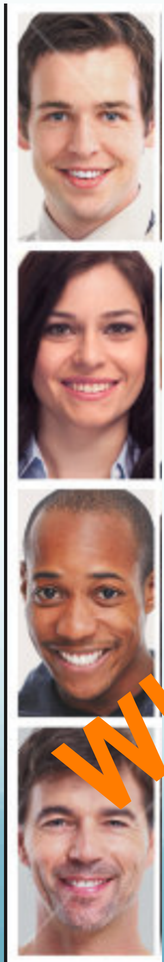
$f(2) \approx f(1)$

$f(3) > f(1)$

$|f(4) - f(2)| < |f(4) - f(3)|$

In the End You Have Training Data With Mixed Annotations

Features   Annotation - Age



[20-25]

f(2) ≈ f(1)

f(3) > f(1)

|f(4) - f(2)| < |f(4) - f(3)|

Which is the most informative type of guidance?

Our Work: Mathematical Formulation

- We derived new loss functions for these four forms of guidance.
- E.g. Relative guidance with the Ridge estimator:
  - $\min_{w} ||Xw - Y||^2 + \lambda_1 \sum_{i,j \in P} \log \sigma\left(\left(x_i - x_j\right)' w\right) + \lambda_2 ||w||^2$
    - σ: The logistic function
    - P: Set of relative pairs
    - λ_1, λ_2: Regularization parameters

Regularizer

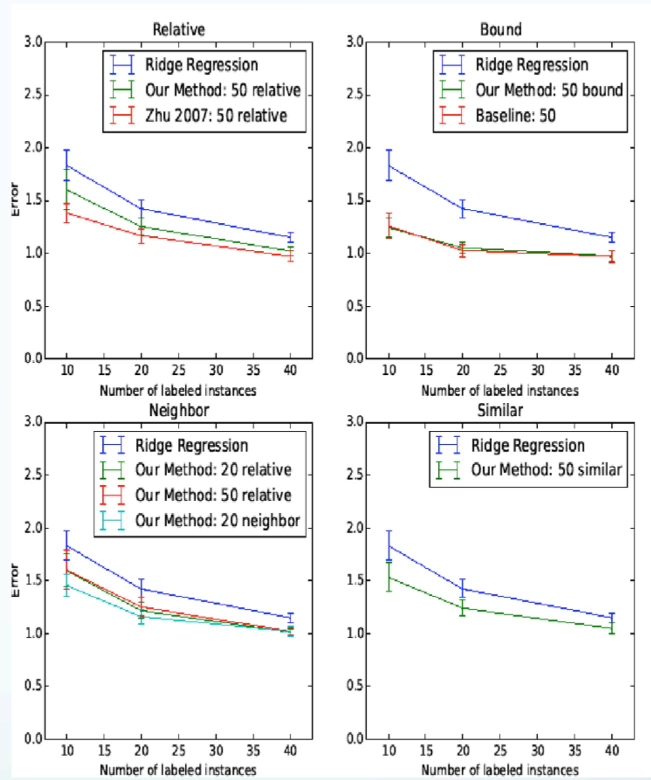
All sorts of tricks: logistic function, logs
Why?
Guarantee convexity and convergence proofs

Our Work: Other Losses

- We derived similar loss functions for the other 3 forms of guidance
    - Bound: $f(x_i) \in [a_i, b_i]$
        - $\sigma\left(b_i - f(x_i)\right) - \sigma\left(a_i - f(x_i)\right)$
    - Similar: $\left|f(x_i) - f\left(x_j\right)\right| \leq s$
        - $\sigma\left(s - \left(f(x_i) + f\left(x_j\right)\right)\right) - \sigma\left(-s - \left(f(x_i) + f\left(x_j\right)\right)\right)$
    - Neighbor: $\left|f(x_i) - f\left(x_j\right)\right| < |f(x_i) - f_k)|$
        - $\min\left\{ \begin{array}{c} 1 - H\left(f(x_k) - f\left(x_j\right)\right), \\ 1 - H\left(f(x_k) + f\left(x_j\right) - 2f(x_i)\right) \end{array} \right\}$
        - H: CDF of exponential distribution (because we used exponential noise)

Our Work: Experimental Results

- Typical results (more in the paper) using our guidance with ridge regression
- Relative and Similar guidance seem to be very valuable
- Neighbor is more valuable than relative
- **Bound worked well on synthetic data, but performed no better than a simple baseline on real data**



Feature Level Guidance

[Gress, Davidson, Human Guided Linear Regression with Feature-level Constraints, AAAI 2018]

- Can we train a regression model with little labeled training data? y = w1x2 + w2x2 + w3x3 .. wmxm
- Our work: leverage ***feature level*** guidance provided by the user
  - "The fuzzy cat's fur has a negative impact on age"
  - "Square footage has a larger positive impact on house price than number of bathrooms"
- Three forms of guidances to constrain w:
  - Sign: "Feature i has a positive impact on the label"
  - Relative: "Feature i has a more positive impact on the label than feature j"
  - Pairwise-Sign: "Features i and j has the same impact (positive or negative) on the label"

Feature Level Guidance

[Gress, Davidson, Human Guided Linear Regression with Feature-level Constraints, AAAI 2018]

- Can we train a regression model with little labeled training data? y = w1x2 + w2x2 + w3x3 .. wmxm
- Our work: leverage ***feature level*** guidance provided by the user
  - "The fuzzy cat's fur has a negative impact on age"
  - "Square footage has a larger positive impact on house price than number of bathrooms"
- Three forms of guidance:
  - Sign: "Feature i has a positive impact on the label"
  - Relative: "Feature i has a more positive impact on the label than feature j"
  - Pairwise-Sign: "Features i and j has the same impact (positive or negative) on the label"

Which is the most informative type of guidance?

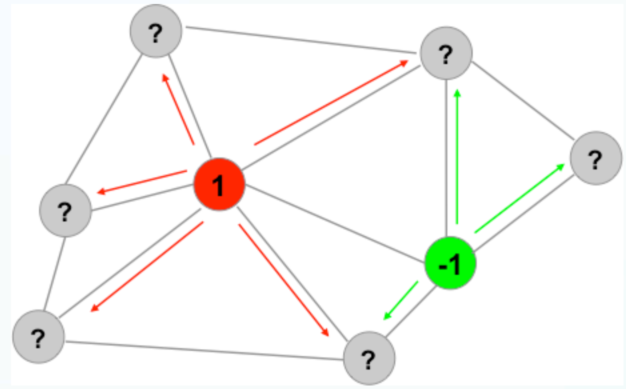
Results

	Nonnegative	Ridge	Lasso	PSRC: p Signs	PRCR: p Pairs	PPSCR: p pairs
Synthetic	0.183(0.030)	0.179(0.029)	0.180(0.029)	**0.141(0.026)**	0.155(0.031)	0.161(0.030)
BH	0.212(0.022)	0.202(0.036)	0.183(0.023)	**0.149(0.018)**	0.176(0.029)	0.163(0.020)
Wine	0.101(0.013)	0.100(0.013)	0.105(0.013)	**0.088(0.010)**	0.093(0.011)	0.091(0.010)
Concrete	0.255(0.022)	0.275(0.020)	0.293(0.018)	**0.220(0.017)**	0.232(0.019)	0.229(0.018)
Housing	0.432(0.041)	0.478(0.043)	0.482(0.049)	0.409(0.038)	0.451(0.042)	**0.399(0.037)**
ITS	0.568(0.092)	0.625(0.095)	0.700(0.118)	**0.525(0.089)**	0.540(0.091)	0.570(0.093)
Heart	2.129(0.220)	2.159(0.220)	2.190(0.187)	**2.007(0.191)**	2.044(0.209)	2.124(0.229)

- Our methods:
  - PSRC: Sign guidance
  - PRCR: Relative Guidance
  - PPSCR: Pairwise-sign Guidance
- **Sign guidance performed best overall**, but all forms of guidance improved accuracy

Non-Expert Active Learning

[Buyue Qian, Xiang Wang, Fei Wang, Hongfei Li, Jieping Ye, Peter Walker, Ian Davidson: Active Learning from Relative Queries. IJCAI 2013]



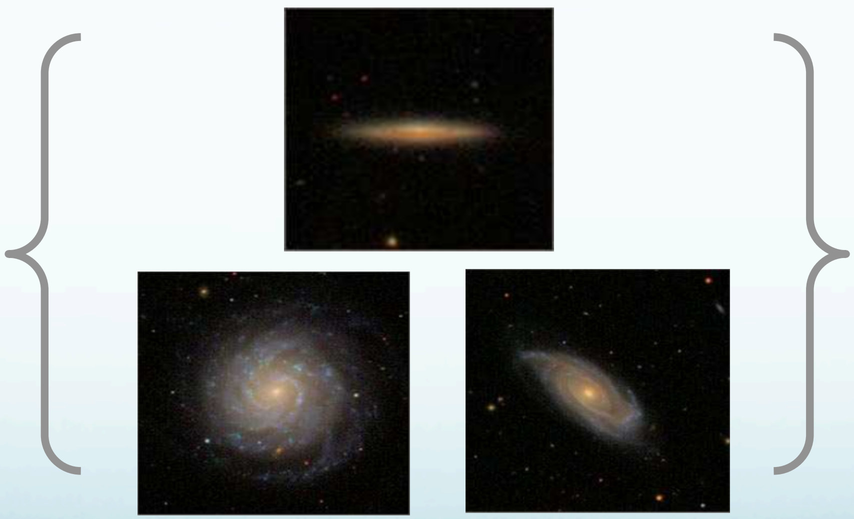
Does **not** generate any new labels
Answerable via crowd-sourcing

Uncertainty in Labeling

Absolute Query Strategy for Relative Queries Don't Work

Main Galaxy

Set of neighbors

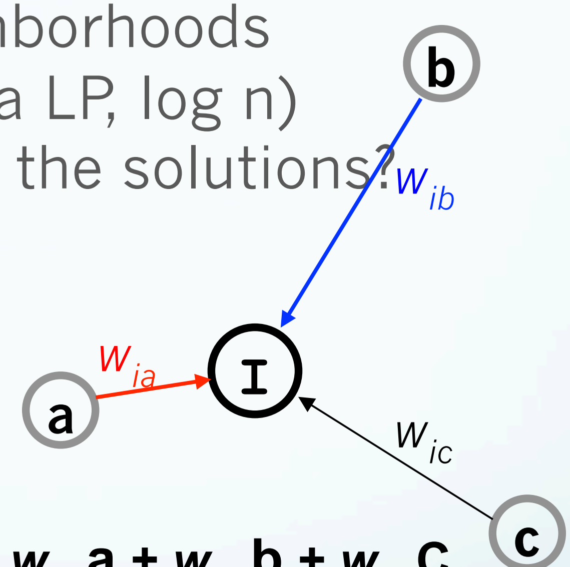


Question: (partially) **order the neighbors** based on their visual similarities to the main Galaxy?

Our Active Scheme

Relative query strategy: which instance to focus on?

X is the set of points, N the set of neighborhoods
Approximate all minimum set covers (via LP, log n)
How many times does a point appear in the solutions?
Weighted set coverage: i.e. connectivity



Query influential point's neighborhood

Instance ***a*** is closer to **I** than ***b***, we have w_{ia} >= w_{ib}

$$\mathbf{w}_i(\mathbf{J}^a - \mathbf{J}^b) \geqslant 0$$

Encode neighborhood guidance: $\mathbf{I} = w_{ia*}a + w_{ib*}b + w_{ic*}c$

$\mathbf{J}^i$ is a single-entry vector whose i-*th* entry is 1 and all other entries 0

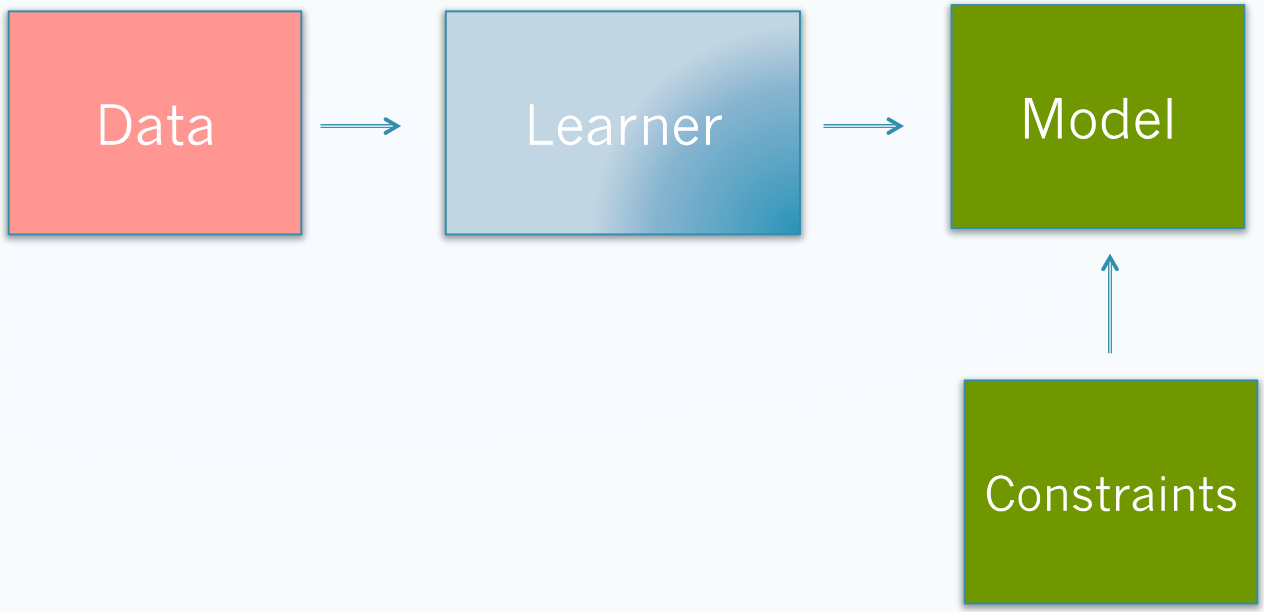
Learning of graph weights:

$$\begin{aligned} \min_{\mathbf{w}_i} \quad & \mathbf{w}_i \mathcal{C}^i \mathbf{w}_i^T \\ s.t. \quad & \mathbf{w}_i \mathbf{1} = 1; \\ & w_{ij} \geqslant 0. \end{aligned}$$

Take Away Message

We can inject human guidance a number of ways

But the underlying solver limits how we can encode their knowledge



Adding Constraints to Clustering

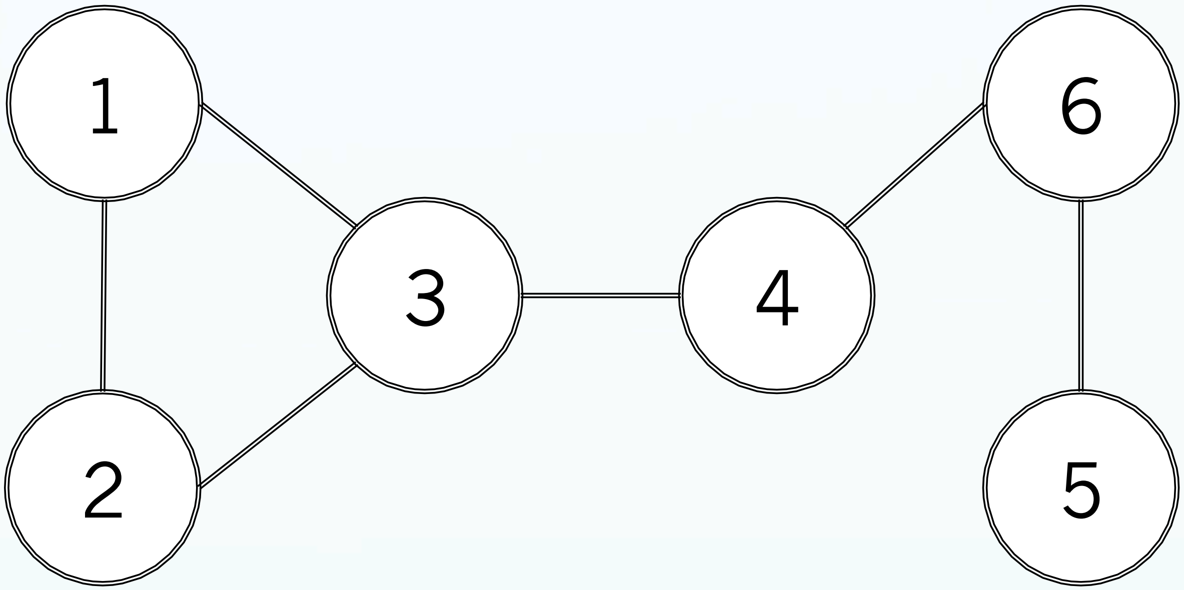
History

- We've been looking at adding constraints to clustering (particularly graphs) for a while
  - KDD 10, AAAI 13, DMKD 14 [Constrained Spectral Methods]
  - SDM 13 [Multi-view Pareto Optimization]
  - ICDM 12, SDM 14 [Active/Self Taught]
  - ICDM 14 [Weighted Spectral Methods]
  - KDD 15 [Contrast and Consensus Formulations]
  - KDD 17 [Constrained Block Models]
  - ICDM 17 [Scaling to huge graphs using RPPM]
- I'll overview the work on graphs.

History

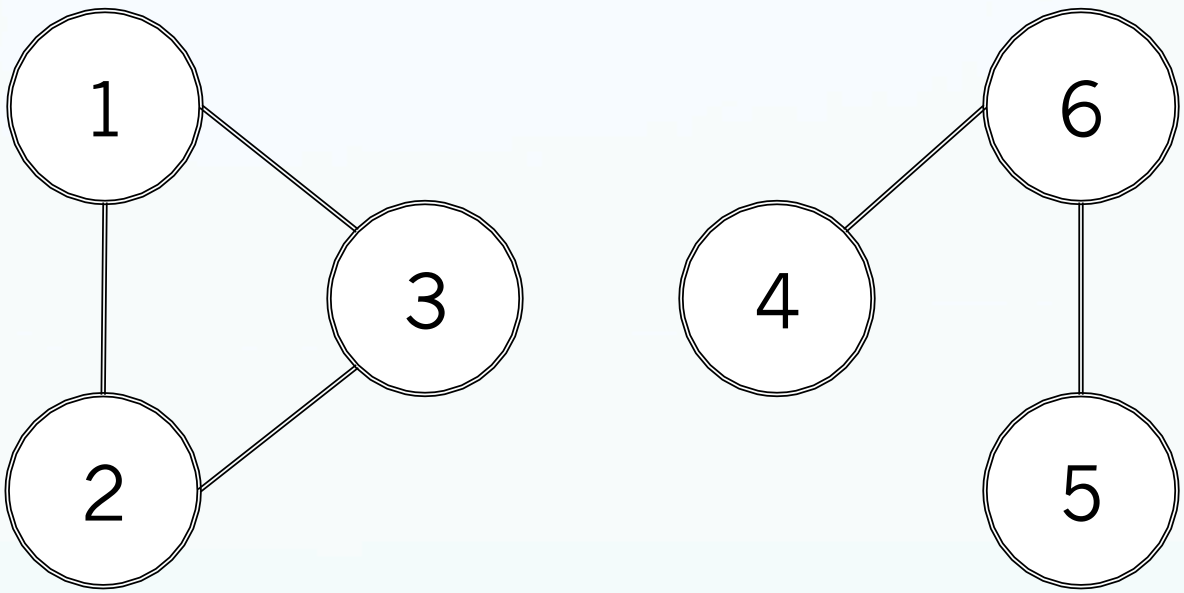
- We've been looking at adding constraints to clustering (particularly graphs) for a while
  - **KDD 10, AAAI 13, DMKD 14 [Constrained Spectral Methods]**
  - SDM 13 [Multi-view Pareto Optimization]
  - ICDM 12, SDM 14 [Active/Self Taught]
  - ICDM 14 [Weighted Spectral Methods]
  - KDD 15 [Contrast and Consensus Formulations]
  - **KDD 17 [Constrained Block Models]**
  - **ICDM 17 [Scaling to huge graphs using RPPM]**
- I'll overview the work on graphs.

Intuition behind Segmenting a Graph – Think of a Social Network

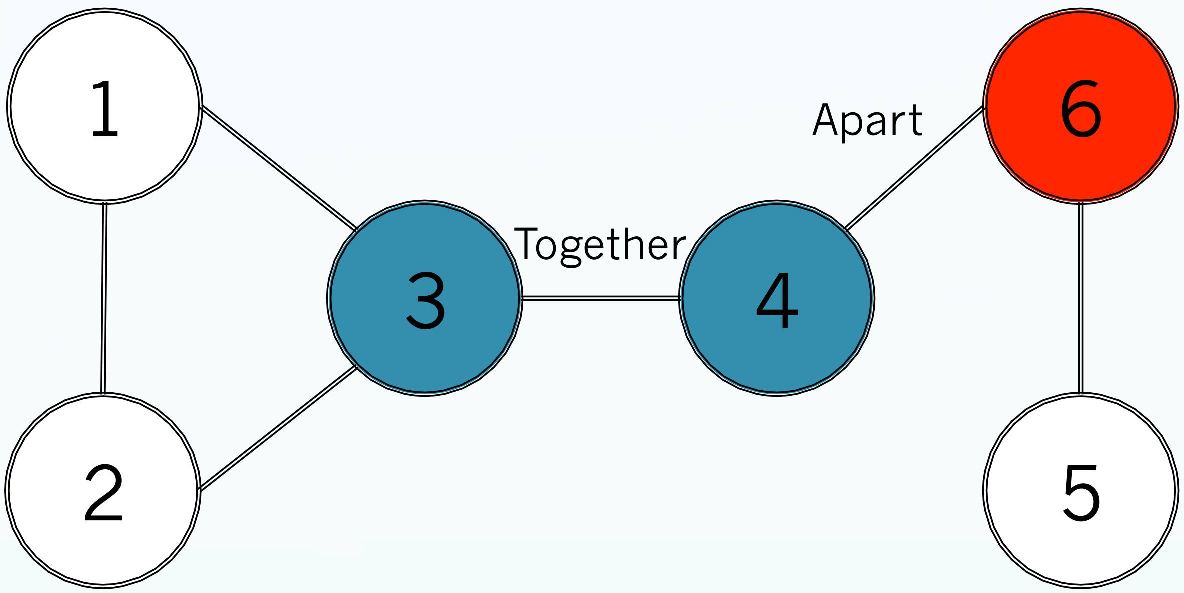


Imagine this is your ego network in Facebook. Want to Create two dinner parties

Intuition behind Segmenting a Graph



Intuition behind Constrained Graph Segmentation

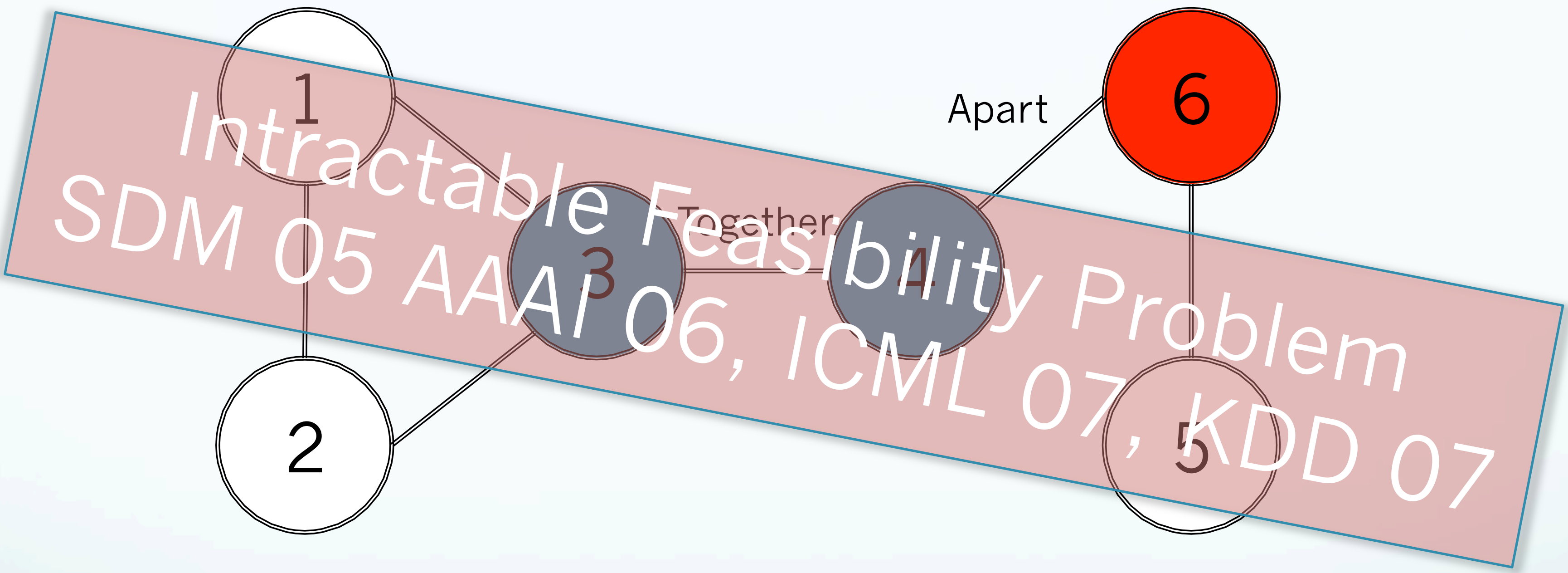


Maybe 6 divorced 4 because they were having an affair with 3

Find a *constrained cut* that:

- Minimizes the cost (friendships links broken)
- Satisfies these constraints

Intuition behind Constrained Graph Segmentation



Maybe 6 divorced 4 because they were having an affair with 3

Find a *constrained cut* that:

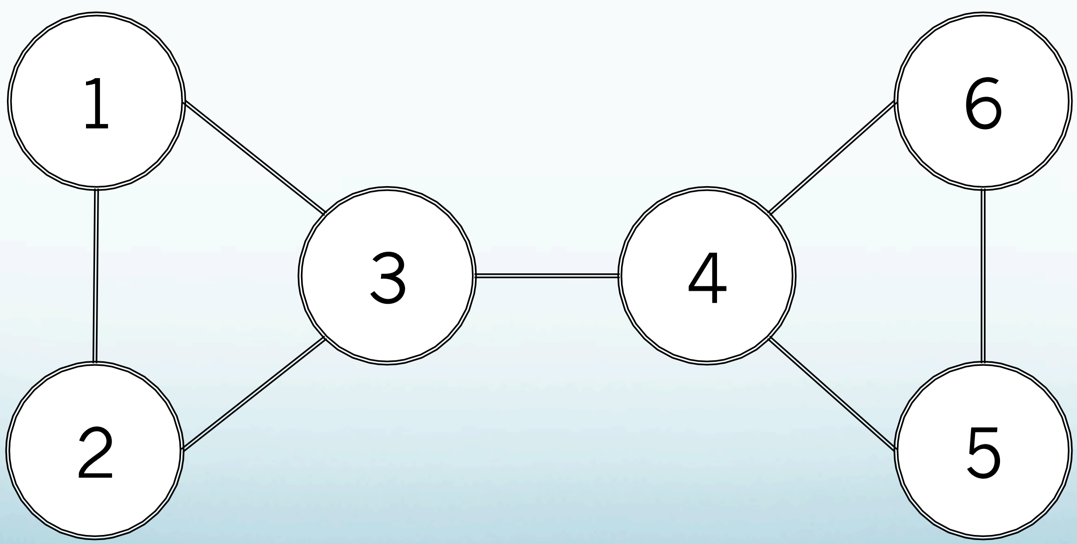
- Minimizes the cost (friendships links broken)
- Satisfies these constraints

Relaxing the Problem
Spectral Clustering

Objective for spectral clustering
(Shi and Malik, 2000)

$$\underset{\mathbf{v} \in \mathbb{R}^N}{\operatorname{argmin}} \quad \mathbf{v}^T \bar{L} \mathbf{v},$$
$$\text{s.t.} \quad \mathbf{v}^T \mathbf{v} = 1, \ \mathbf{v} \perp D^{1/2} \mathbf{1}.$$

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



J. Shi and J. Malik. Normalized cuts and image segmentation. PAMI, 22(8):888–905, 2000.
X. Wang, I. Davidson. Flexible constrained spectral clustering. In KDD 2010, pp. 563-572.
X. Wang, B. Qian, I. Davidson. On constrained spectral clustering and its applications. DMKD, 2014.

Constrained Spectral Clustering

Objective for spectral clustering
(Shi and Malik, 2000)

$$\boxed{\begin{aligned} &\underset{\mathbf{v}\in\mathbb{R}^N}{\operatorname{argmin}} \quad \mathbf{v}^T \bar{L} \mathbf{v}, \\ &\text{s.t.} \quad \mathbf{v}^T\mathbf{v} = 1, \; \mathbf{v} \perp D^{1/2}\mathbf{1}. \end{aligned}}$$

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad Q = \begin{bmatrix} +1 & +1 & +1 & +1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 \\ -1 & -1 & -1 & -1 & +1 & +1 \\ -1 & -1 & -1 & -1 & +1 & +1 \end{bmatrix}$$

Finding normalized min-cut

Lower-bounding how well the constraints in Q are satisfied

Normalizing v

Ruling out the trivial solution

$$\boxed{\underset{\mathbf{v}\in\mathbb{R}^N}{\arg\min}\, \mathbf{v}^T \bar{L} \mathbf{v},} \; \text{s.t.} \; \boxed{\mathbf{v}^T \bar{Q} \mathbf{v} \geq \alpha,} \; \boxed{\mathbf{v}^T\mathbf{v} = \mathrm{vol}(\mathcal{G}),} \; \boxed{\mathbf{v} \neq D^{1/2}\mathbf{1}.}$$

J. Shi and J. Malik. Normalized cuts and image segmentation. PAMI, 22(8):888–905, 2000.
X. Wang, I. Davidson. Flexible constrained spectral clustering. In KDD 2010, pp. 563-572.
X. Wang, B. Qian, I. Davidson. On constrained spectral clustering and its applications. DMKD, 2014.

- Objective:

$$\arg\min_{\mathbf{v}\in\mathbb{R}^N} \mathbf{v}^T \bar{L}\mathbf{v}, \text{ s.t. } \mathbf{v}^T\bar{Q}\mathbf{v} \geq \alpha, \ \mathbf{v}^T\mathbf{v} = \mathrm{vol}(\mathcal{G}),$$

- Introducing Karush-Kuhn-Tucker (KKT)

$$\begin{aligned}
\text{(Stationarity)} \quad & \bar{L}\mathbf{v} - \lambda\bar{Q}\mathbf{v} - \mu\mathbf{v} = 0, \\
\text{(Primal feasibility)} \quad & \mathbf{v}^T\bar{Q}\mathbf{v} \geq \alpha, \mathbf{v}^T\mathbf{v} = \mathrm{vol}(\mathcal{G}), \\
\text{(Dual feasibility)} \quad & \lambda \geq 0, \\
\text{(Complementary slackness)} \quad & \lambda(\mathbf{v}^T\bar{Q}\mathbf{v} - \alpha) = 0.
\end{aligned}$$

- Let $\beta = -\frac{\mu}{\lambda}\mathrm{vol}(\mathcal{G})$
- The problem becomes

$$\bar{L}\mathbf{v} = \lambda(\bar{Q} - \frac{\beta}{\mathrm{vol}(\mathcal{G})}I)\mathbf{v}$$

$$\mathbf{v}^T\mathbf{v} = \mathrm{vol}(\mathcal{G}),$$

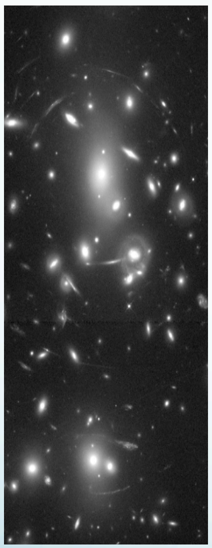
This is a generalized eigenvalue problem.

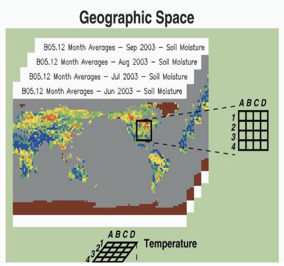
Some Take Aways

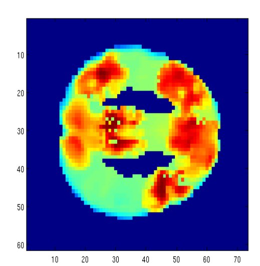
- We relaxed a discrete optimization problem
  - No guarantees of optimality after the rounding
- We were limited to conjunctions of constraints
- We were limited to binary relationship constraints
- We were limited to making one objective a constraint
  - We did a Pareto optimization formulation [SDM 13] but the code is challenging difficult to implement

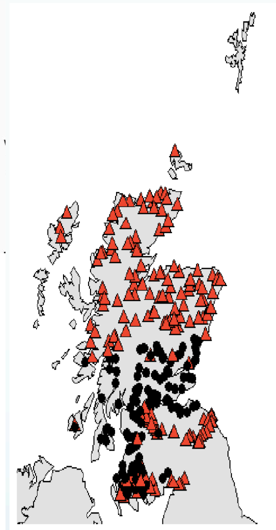
Network Discovery in Spatial Temporal Data

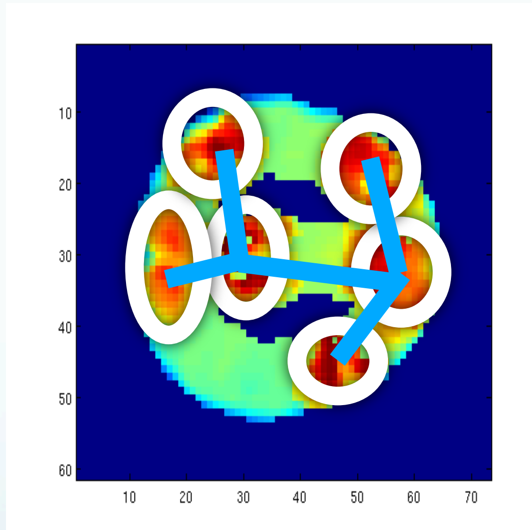
[Bai, Davidson, Unsupervised Network Discovery, KDD 17]

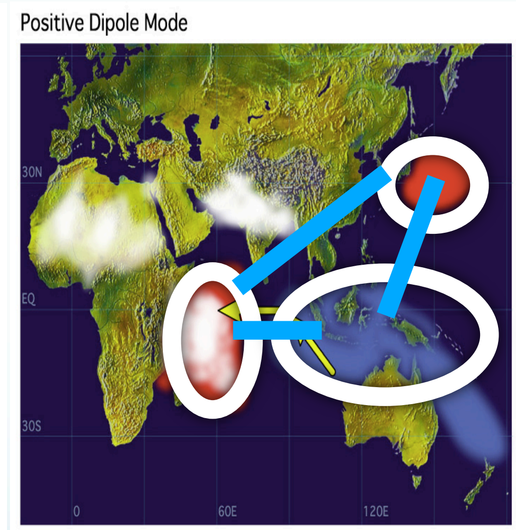






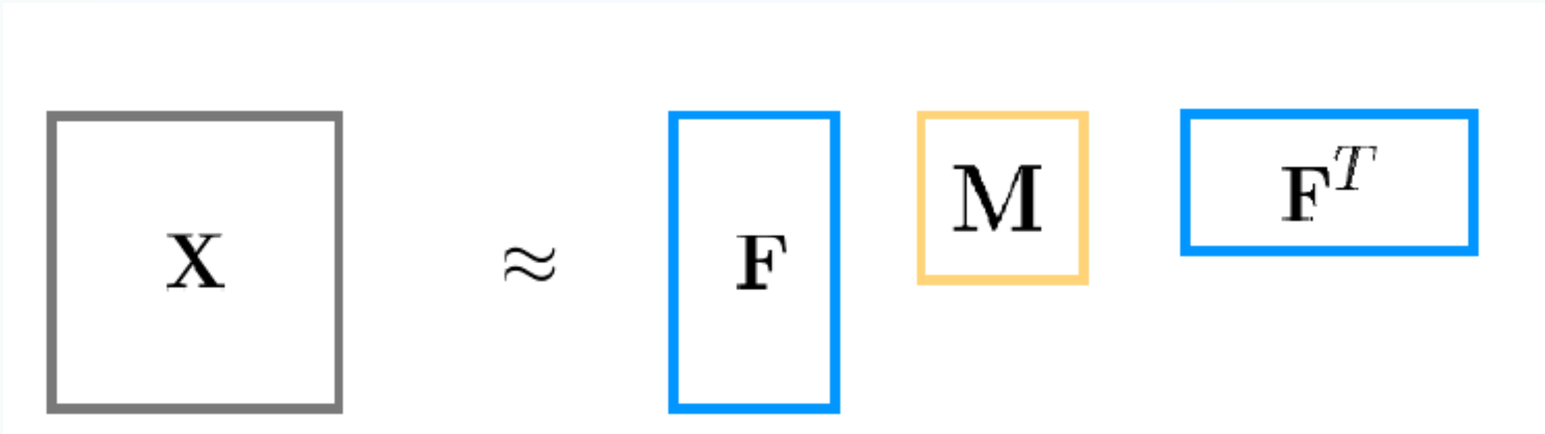






Many observations over time of the same locations
We can convert them into a graph as shown before

Regular Block Modelling



Input:

X: n x n weighted graph, edge weights are correlations

Output:

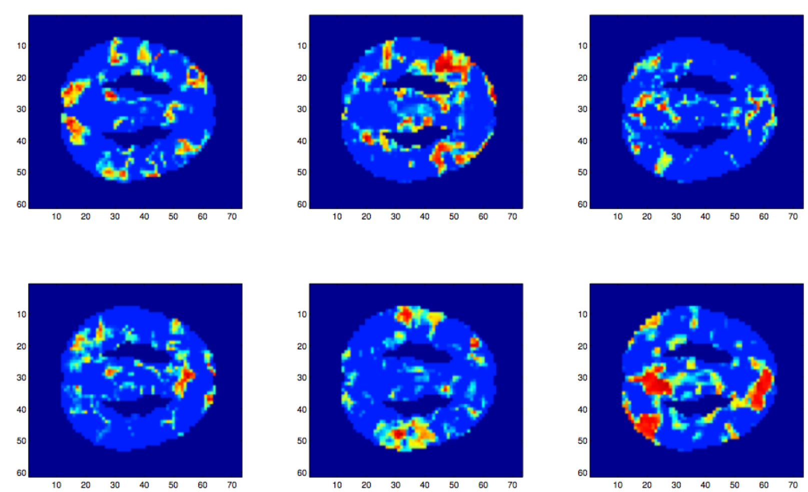
F: n x k block indicator matrix

Each block's indicator matrix is stored column wise

M: k x k interaction matrix

Regular Block Modeling on Spatial Data

$$\mathbf{X} \approx \mathbf{F} \mathbf{M} \mathbf{F}^T$$



Adding in Side Information

- “Affinity” matrix
- Absolute Correlations
- Graph: ***N*** by ***N***

Kernel/Graph Regularization

$$\min_{\mathbf{F}\geq 0, \mathbf{M}\geq 0} \|\mathbf{X} - \mathbf{F}\mathbf{M}\mathbf{F}^T\|_F^2 + \beta tr(\mathbf{F}^T \mathbf{\Theta} \mathbf{F})$$

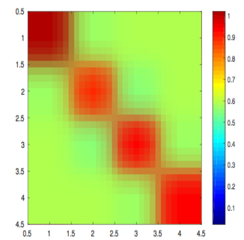
$$s.t. \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}$$

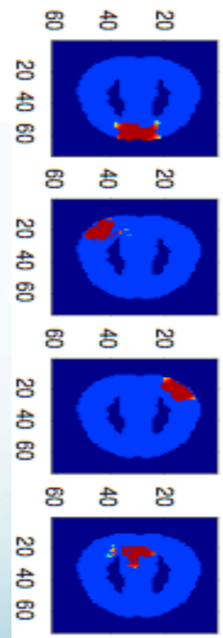
- Cluster indicator matrix **(Nodes)**
- ***N*** by ***k***
- [0,1]
- Column-wise orthogonal

- Mixing matrix **(Edges)**
- ***k*** by ***k***
- Nonnegative
- Associations between clusters

Requiring the Blocks to Be Spatially Contiguous

$$\mathbf{X} \approx \mathbf{F} \quad \mathbf{M} \quad \mathbf{F}^T$$





$$\underset{\mathbf{F}\geq 0, \mathbf{M}\geq 0}{Minimize} \|\mathbf{X}-\mathbf{F}\mathbf{M}\mathbf{F}^T\|_F^2 + \beta tr(\mathbf{F}^T \boldsymbol{\Theta} \mathbf{F})$$

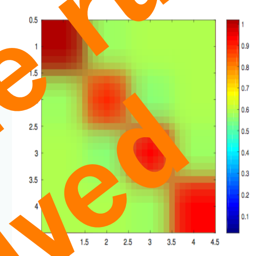
$$s.t. \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}$$

Requiring the Blocks to Be Spatially Contiguous

Can't put constraints on M

Multiplicate update rules can't be derived

$$\mathbf{X} \approx \mathbf{F} \; \mathbf{M} \; \mathbf{F}^T$$



$$\underset{\mathbf{F}\geq 0, \mathbf{M}\geq 0}{Minimize} \|\mathbf{X} - \mathbf{F}\mathbf{M}\mathbf{F}^T\|_F^2 + \lambda tr(\mathbf{F}^T \Theta \mathbf{F})$$

$$s.t. \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}$$

Recap

- Relative guidance – asking humans easier annotations/questions
  - IJCAI 13, ICDM 16, AAAI 18
- Large scale transfer learning – asking humans what tasks are related
  - AAAI 15, ICML 13, AAAI 10, TIPS 16
- Constrained block models and clustering – asking humans what their expectations of clustering should be
  - More recently KDD15,17 and ICDM 17

Some Directions of My Groups with **Limitations**

- **Relative guidance – asking humans easier annotations/questions**
  - **IJCAI 13, ICDM 16, AAAI 18**
- The results didn't match out intuition. Why?
  - Tricks to ensure convexity and convergence proofs meant we couldn't model the human knowledge as we wanted to
  - We kept on adding in regularizers and having to tune the hyper-parameters

Some Directions of My Groups with Limitations

- **Constrained clustering – asking humans what their expectations of clustering should be**
  - **More recently KDD15,17 and ICDM 17**

- The constraints had to be in a very specific form
- Some constraints we couldn't even model (i.e. constraints on M in block models)

A Solution?

A growing interest in using Constraint Solvers (CP/SAT/MIP) in ML and DM

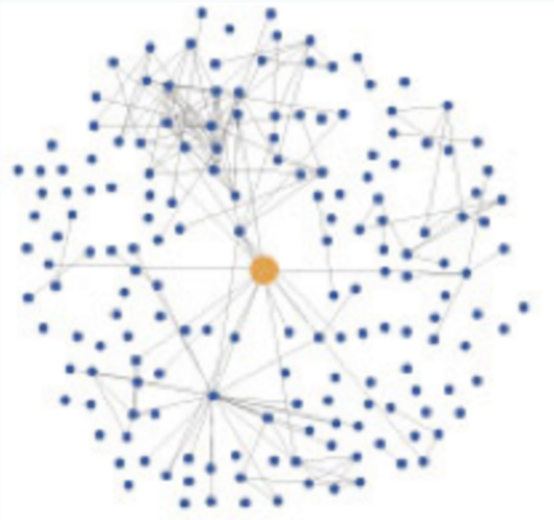
- Meetings
  - Dagstuhl 11201 (2011): Constraint Programming meets Machine Learning and Data Mining
  - Dagstuhl 14411 (2014): Constraints, Optimization and Data
- Workshops
  - CoCoMile 2012 ECAI, CoCoMile 2013 AAAI
- Journal Special Issue
  - AIJ - Combining Constraint Solving, Mining & Learning 2017
- CP 2017
  - Special track on ML/DM + CP
- Two dozen+ papers in the last five years on the topic at:
  - IJCAI, AAAI, KDD, ICDM, SDM etc.
- IJCAI 2017 tutorial (Siegfried, Tias and myself)
- https://sites.uclouvain.be/cp4dm/tutorial/ijcai17/

Benefits and Uses of CP For ML/DM

- Because the search algorithm is branch and bound no mandated restrictions on objective function and constraints form
  - But clever filtering algorithms are needed for scalability
- Three main uses explored so far
  - A) Use constraints as a dialog mechanism to allow complex feedback
  - B) Using CP to model novel problems
  - C) CP as a post-processor to DM/ML

Use #A1 - New Constraints

[Duong, Vrain and Davidson, ECAI16]



- New types of constraints
  - My dinner party problem: 'Segment by ego network into k groups (k dinner parties) may yield poor results
    - So require each group has:
      - 1) #Males = #Females,
      - 2) Diameter wrt age < 10 and
      - 3) Everyone has at least q people at the party with at least r common interests
  - These are very different cardinality, density etc.
  - Definitely not linear or encodable in a matrix.

Use #A2 – For Block Modeling

[Work with Peter Stuckey's group at U.Melb]

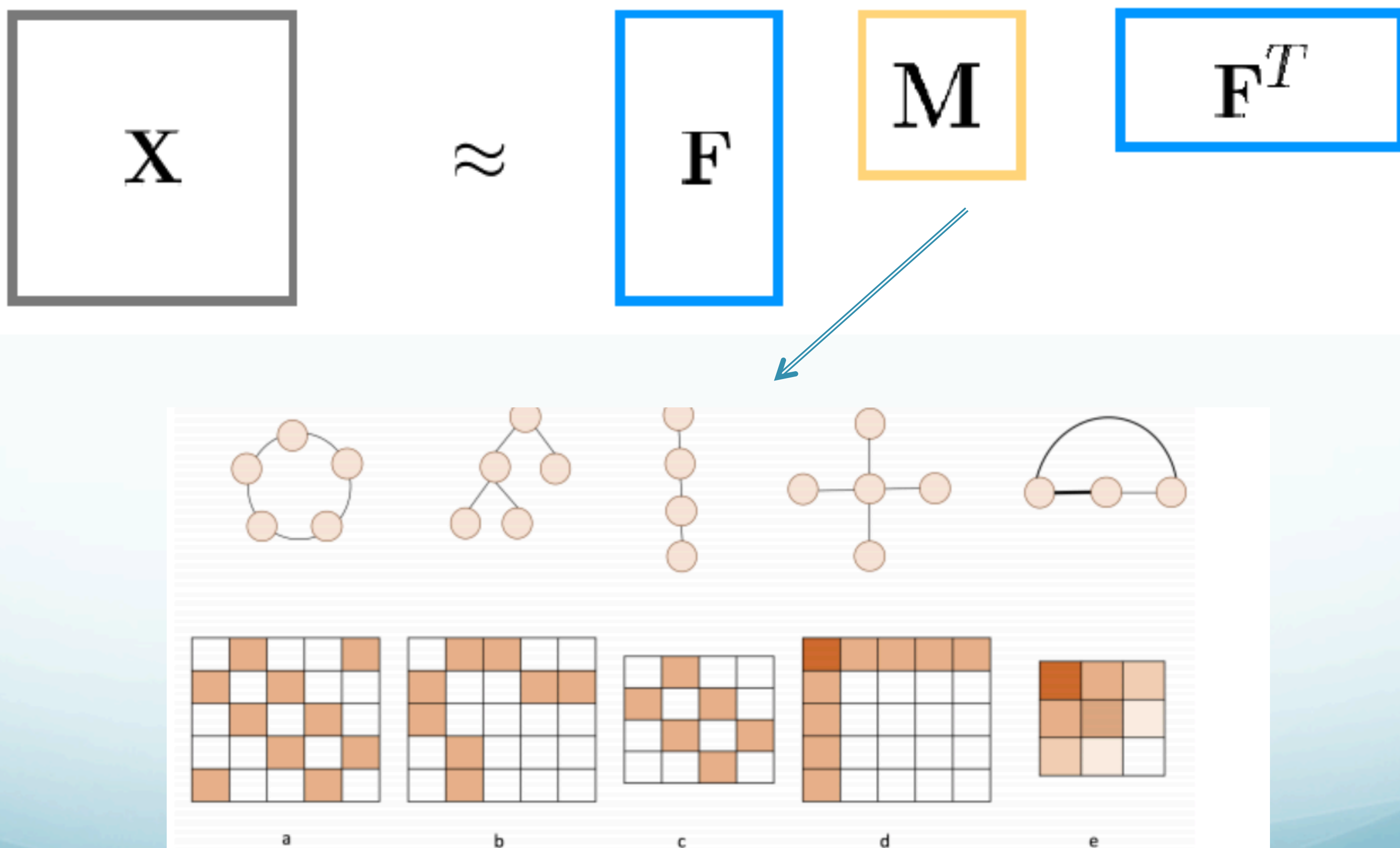


Figure 2: Some examples of image graph structures: a) ring, b) hierarchy, c) stick, d) star, e) core-periphery

Use #B3 – Outlier Description Problem

- Two sets of points: normal, abnormal
  - Question: What make the normal set normal
  - Example: Represent a car by vector describing part locations
    - What common properties do the non-lemons have that the outliers do not.
  - Related to
    - **Discovering Outlying Aspects in Large Datasets**. Nguyen Xuan Vinh, Jeffrey Chan, Simone Romano, James Bailey, Christopher Leckie, Kotagiri Ramamohanarao and Jian Pei. *Data Mining and Knowledge Discovery*, 30(6), pp. 1520-155, 2016

The Benefits of CP

Kuo, Chia-Tung, and Ian Davidson. "A Framework for Outlier Description Using Constraint Programming." *AAAI*. 2016.

Lower and upper bound on # NN for normal and outliers

Objective Maximize $k_N - k_O$

Variables $F = [f_1, f_2, \ldots, f_{|S|}] \in \{0,1\}^{|S|}$

Projection vector

$k_{min} \leq k_O \leq k_N \leq k_{max}$

NN distance $0 \leq r \leq r_{max}$

Constraints $\forall x \in N,\ |\mathcal{N}_F(x,r)| \geq k_N$

$\forall y \in O,\ |\mathcal{N}_F(y,r)| < k_O$

Multi-criteria optimization over k, r and F

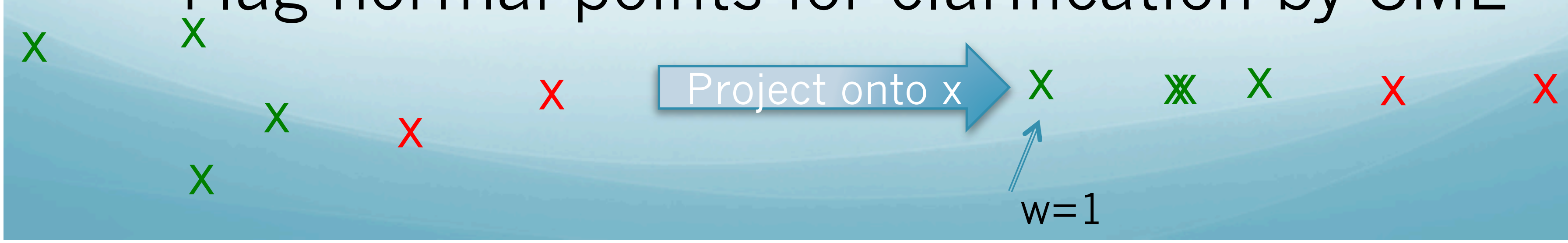
Human in Loop Extension

Kuo, Chia-Tung, and Ian Davidson. "A Framework for Outlier Description Using Constraint Programming." *AAAI*. 2016.

Objective	Maximize $k_N - k_O$
Variables	$F = [f_1, f_2, \ldots, f_{\vert S\vert}] \in \{0,1\}^{\vert S\vert}$
	$k_{min} \leq k_O \leq k_N \leq k_{max}$
	$0 \leq r \leq r_{max}$
Constraints	$\forall x_i \in N,\ \vert\mathcal{N}_F(x_i, r)\vert \geq (1 - w_i)k_N$
	$\sum_{i=1}^{n} w_i \leq w_{max}$
	$\forall y \in O,\ \vert\mathcal{N}_F(y, r)\vert < k_O$

I can ignore some points
From the NN constraint

Multi-criteria optimization over k, r, F and w
Flag normal points for clarification by SME



Two Sub Space Explantion

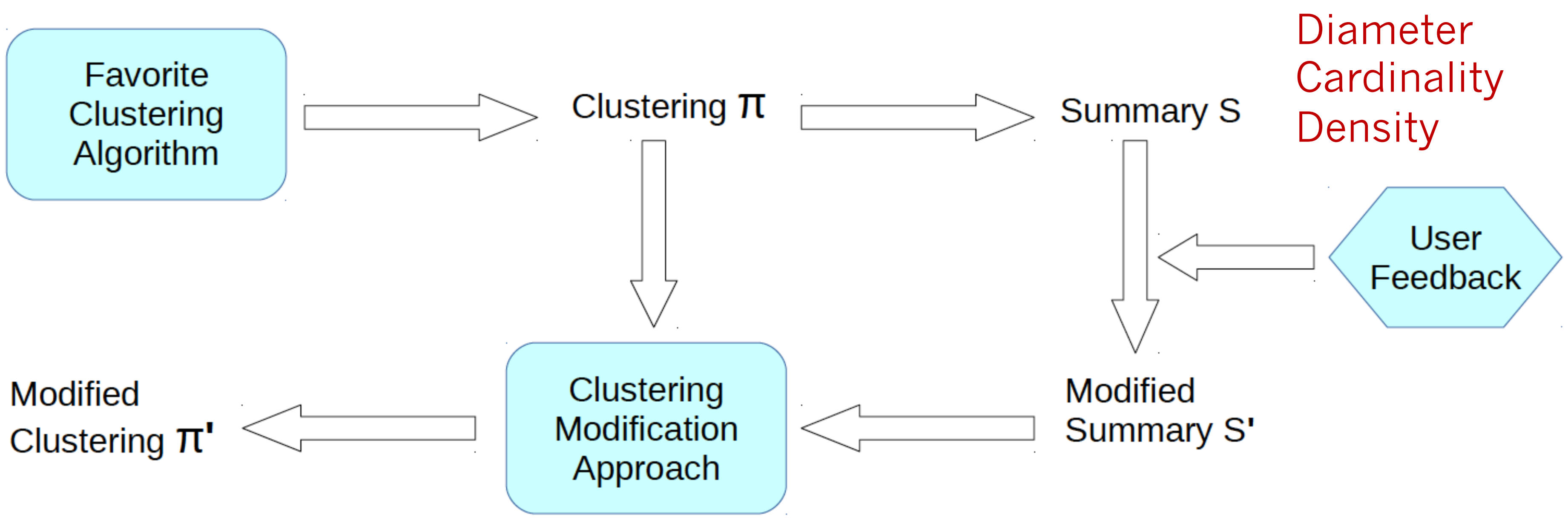
Kuo, Chia-Tung, and Ian Davidson. "A Framework for Outlier Description Using Constraint Programming." *AAAI*. 2016.

Objective	Maximize $k_N - k_O$
Variables	$F = [f_1, \ldots, f_{\|S\|}], G = [g_1, \ldots, g_{\|S\|}] \in \{0,1\}^{\|S\|}$
	$k_{min} \leq k_O \leq k_N \leq k_{max}$
	$0 \leq r_F, r_G \leq r_{max}$
Constraints	$\forall x \in N,\ \|\mathcal{N}_F(x, r_F)\| \geq k_N$ AND $\|\mathcal{N}_G(x, r_G)\| \geq k_N$
	$\forall y \in O,\ \|\mathcal{N}_F(y, r_F)\| < k_O$ OR $\|\mathcal{N}_G(y, r_G)\| < k_O$

Multi-criteria optimization over k, r, F and G

Use #C4 – HIL Clustering

A Framework for Minimal Clustering Modification via Constraint Programming, Tom Kuo et. al. AAAI 17



Minimally modify Π to obtain Π' to satisfy S'

$$\begin{array}{ll} \underset{\Pi'}{\textit{minimize}} & d(\Pi, \Pi') \\ \textit{subject to} & \Pi' \textit{ satisfies } S' \end{array}$$

Constraints are elicited from \pi and updated to obtain \pi'

Intractable Problem

A Framework for Minimal Clustering Modification via Constraint Programming, Tom Kuo et. al. AAAI 17

Theorem (1)

The reclustering problem where $\ell = 2$ is NP-complete.

Proof idea: reduction to Covering Points by Unit Squares.

Even for very limited settings

Theorem (2)

Suppose the number of dimensions along which the maximum diameter must be reduced is a variable ℓ. The reclustering problem is NP-complete for any $k \geq 3$.

Proof idea: similarly reduction to Covering Points by Unit Hypercubes.

Formulation

A Framework for Minimal Clustering Modification via Constraint Programming, Tom Kuo et. al. AAAI 17

$$\underset{z,C,L,H,\Pi'}{\text{minimize}} \quad \sum_{i=1}^{n} z[i]$$

Number of modifications

subject to

$$\forall c = 1,\ldots,k,\ \forall i = 1,\ldots,n,\ C[c,i] = \mathbb{I}[\Pi'[i] = c]$$

$$\forall i = 1,\ldots,n,\ z[i] = \mathbb{I}[\Pi'[i] \neq \Pi[i]]$$

$$\forall c = 1,\ldots,k,\ \ \forall t = 1,\ldots,f,$$

$$L[c,t] = \min_{i=1,\ldots,n} \{C[c,i](X[i,t] - M_u[t])\} + M_u[t]$$

$$H[c,t] = \max_{i=1,\ldots,n} \{C[c,i](X[i,t] - M_l[t])\} + M_l[t]$$

Smallest/largest values for t^th fe

$$H[c,t] - L[c,t] \leq \mathcal{D}'[c,t]$$

User given constant

Results

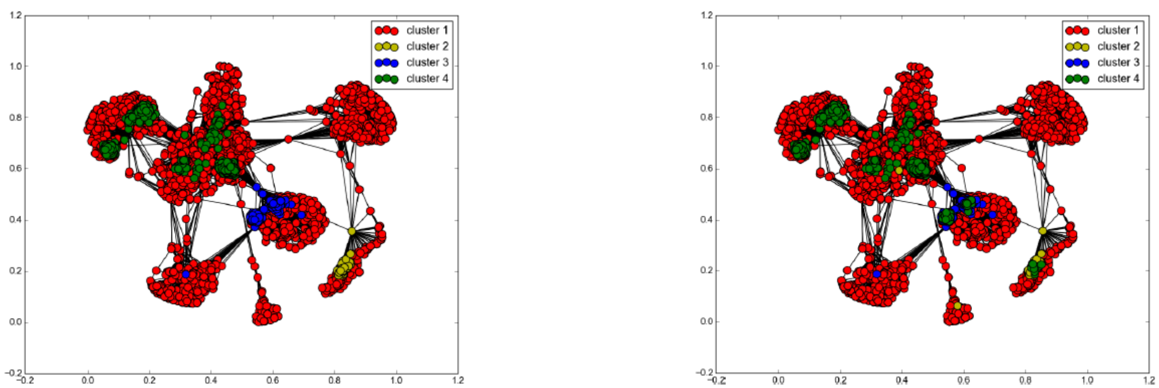
A Framework for Minimal Clustering Modification via Constraint Programming, Tom Kuo et. al. AAAI 17

Data: Facebook egonets[1]

Initial clustering: 4-way clustering from spectral clustering on *friendship graph*

Modification: balance (i.e. bounds diameters) two features/dimensions, gender and some language

Results:



(a) Initial (b) Modified

Figure: Visualization of clusterings on Facebook egonets graph.

Conclusion

- Many problems require human involvement as:
  - Data limitations (size and annotations)
  - Strong domain expertise
  - Challenging problem
- We covered several directions
  - Easier human annotation
  - Transfer learning
  - Constraints
- But formulations in procedural and MP formulations are limited
- CP is a potential solution?
  - For those in Lyon – I'm giving a shortened version of the IJCAI tutorial @ Lyon 1 on December 8th?

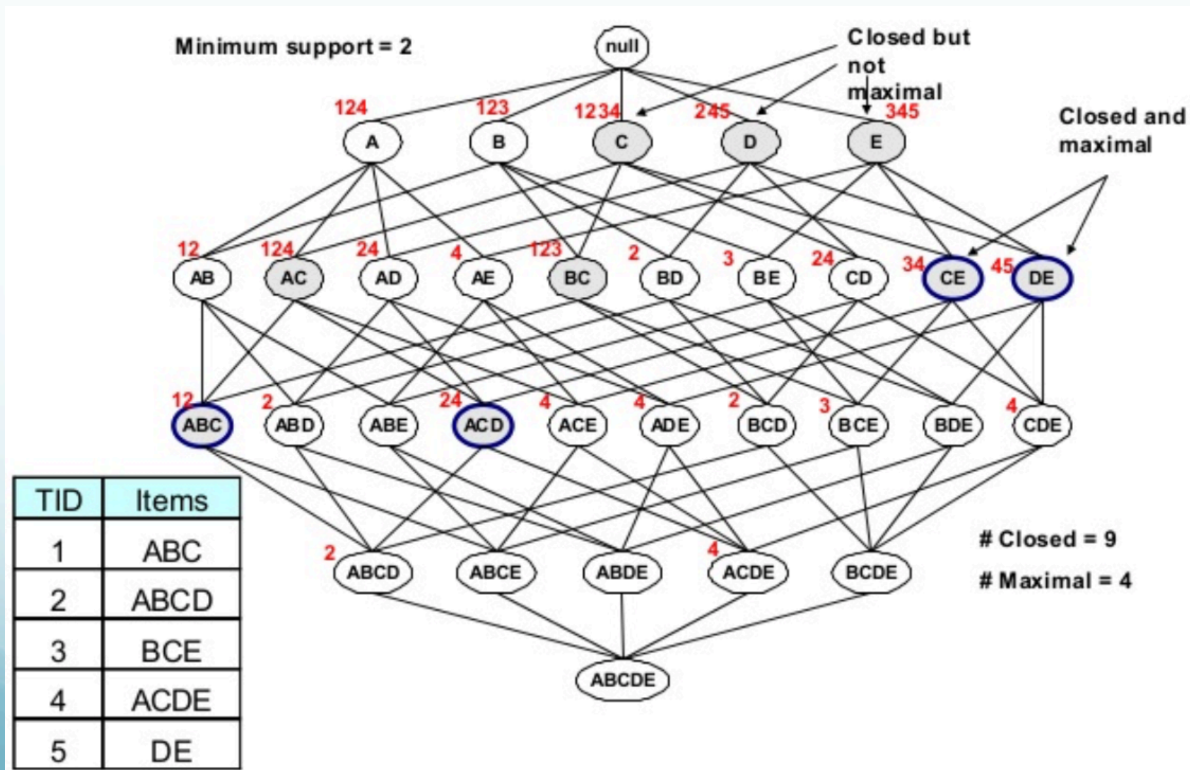
Merci and Questions

davidson@cs.ucdavis.edu
www.cs.ucdavis.edu/~davidson

Combinations

Mueller, Marianne, and Stefan Kramer. "Integer Linear Programming Models for Constrained Clustering." Discovery science 2010. Vol. 6332. 2010.

First to use the idea of PM as a pre-processor?



TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

Each pattern is a potential cluster

Compute some distance over the instances covered by it

Let this be referred to w.

Combinations

Mueller, Marianne, and Stefan Kramer. "Integer Linear Programming Models for Constrained Clustering." Discovery science 2010. Vol. 6332. 2010.

Non-overlapping formulation

instance i covered by at most 1 pattern j

Some measure of cluster quality

maximize	$\frac{1}{k}(w_{max} - w)^T x$	
subject to	(i) $Ax \leq \mathbf{1}$	
	(ii) $Ax \geq y$	(v) $x \in \{0,1\}^n$
	(iii) $\mathbf{1}^T x = k$	(vi) $y \in \{0,1\}^m$
	(iv) $\mathbf{1}^T y \geq m \cdot minCompl$	

x is a binary indicator vector for patterns

y set to 1 means an instance is covered by a cluster

Cover at least minCompl % of instances but not all of them