Clustering:
Probably Approximately Useless?

Rich Caruana
Microsoft Research

JustResearch
David Cohn
Andrew McCallum

CMU
Pedro Artigas
Anna Goldenberg
Anton Likhodedov
Paul Hodor

Cornell
Mohamed Elhawary
Nam Nguyen
Casey Smith

Parts of this work were funded by an NSF Career Award

Clustering:
Probably Approximately Useless?

Rich Caruana
Microsoft Research

And thanks to Dan Bohus (MSR) for help with title!

JustResearch
David Cohn
Andrew McCallum

CMU
Pedro Artigas
Anna Goldenberg
Anton Likhodedov
Paul Hodor

Cornell
Mohamed Elhawary
Nam Nguyen
Casey Smith

Parts of this work were funded by an NSF Career Award

A lot of people TRY clustering,
but few people USE clustering!

Supervised Learning vs. Clustering

Supervised Learning

- Supervisory labels
  - 0/1, multiclass, regression, …
- Crisp performance criteria
  - Accuracy, ROC, log-loss, …
- Well defined goals
  - Minimize loss (on test data)
- Supervised Learning is easy

Unsupervised Clustering

- No labels, just data
  - ???
- Compactness
  - In what metric space?
- Vague goals
  - Groups of "similar" items
- Clustering is harder
  - under-specified
  - yet we expect it to do something useful for our task

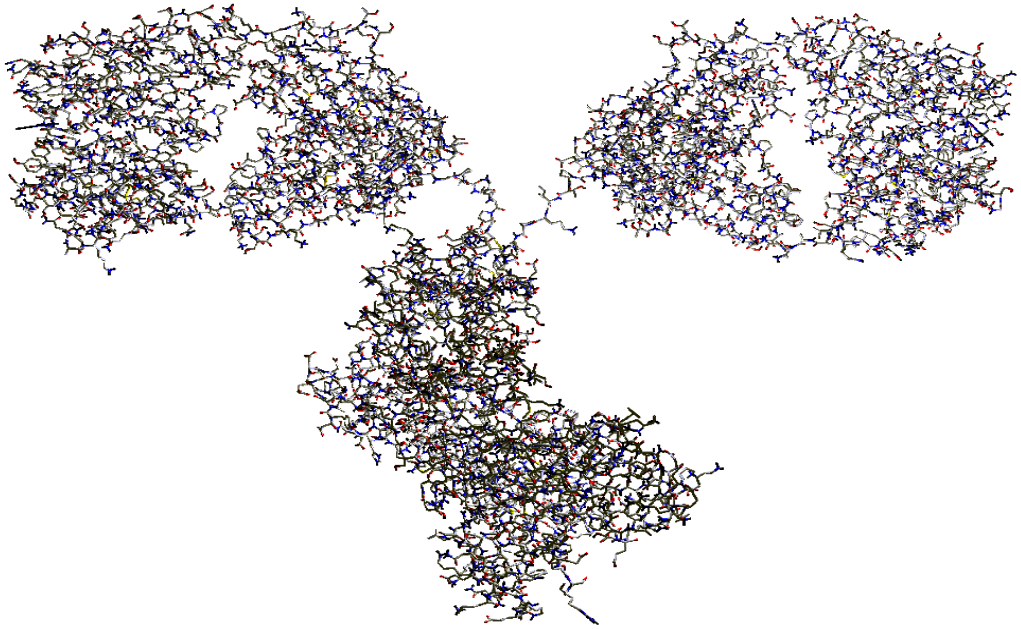
“The problem of … data clustering is generally ill-posed… Performance can vary from one algorithm to another, where each algorithm may be good for some data and bad for other data.”

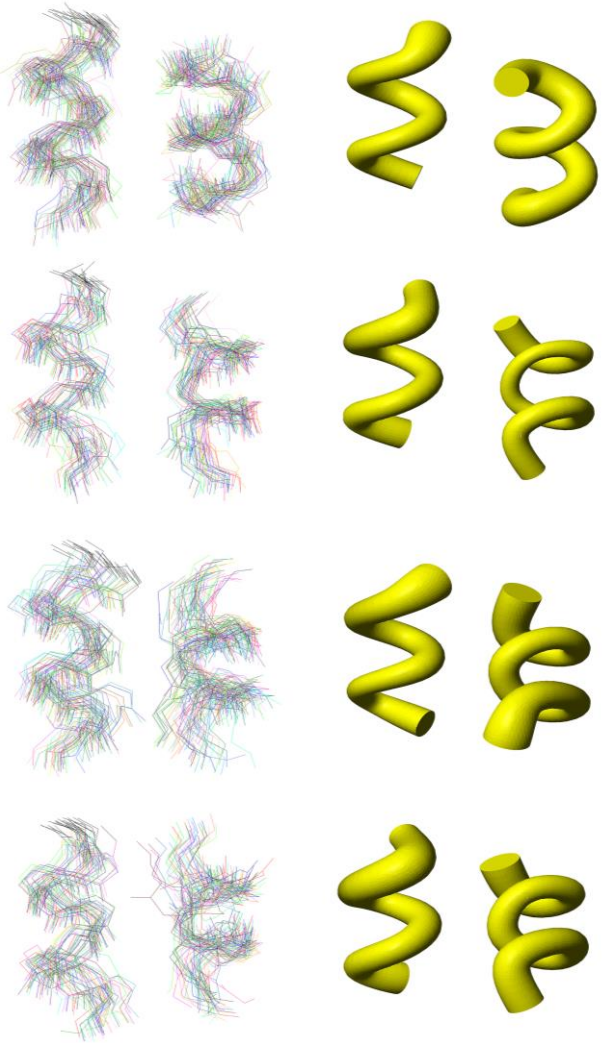
- Tali Tishby, Snowbird2000

I’m a supervised learning guy.

How did I get involved in this?

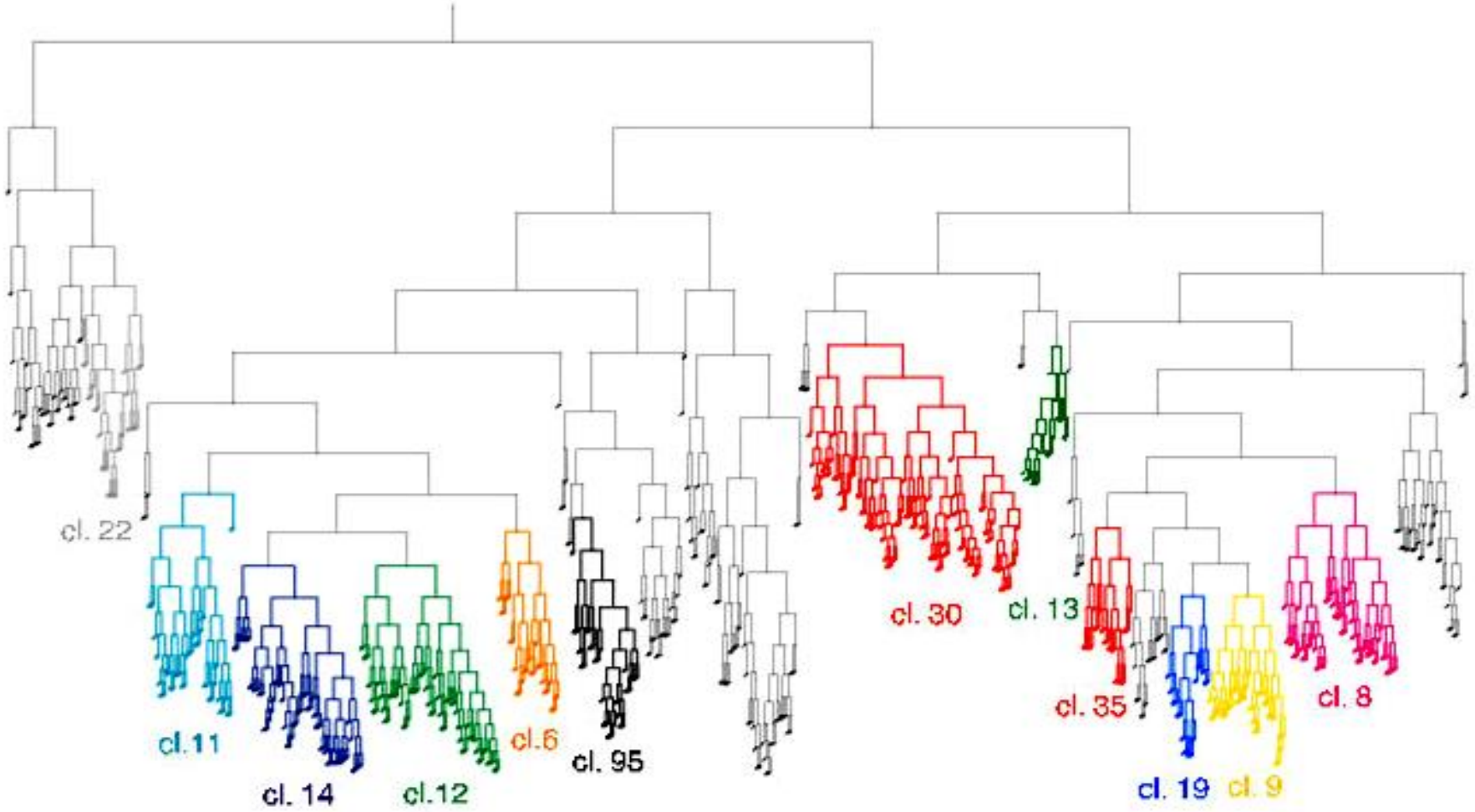
Cluster Proteins in PDB by Structure





Like many of you, I was naïve and thought clustering would be straightforward and useful…

Hierarchical Clustering of Proteins



Took almost a year to generate good clusters!

Why Did it Take Soooooo Long?

Feedback from our Expert

- This pt doesn't belong here
- Move this pt to that cluster
- These two pts shouldn't be in the same cluster
- These two pts should be in the same cluster
- This cluster sucks
- This cluster is good, interesting
- This cluster is too small/large
- Move this cluster near that one
- Move this cluster up/down towards/away from the root
- The clustering is too coarse
- Give me a different clustering
- Give me a clustering half way between these two clusterings
- Cluster using these features
- The label for this pt is "X"

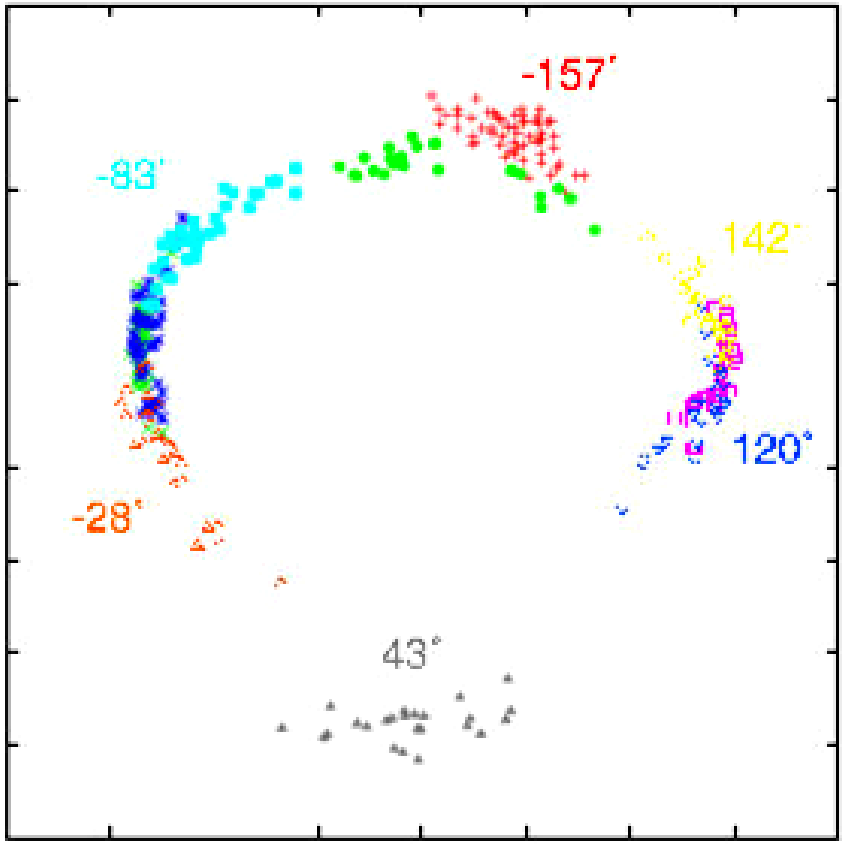
Constraints weaker than labels: don't have to know labels of any pts or even what the labels are, and there are more constraints than labels

Took almost a year to generate good clusters!

Why Did it Take Soooooo Long?

What Did We Learn?

Clustering Eventually Led to Discovery



Important property in scaffolding was helix crossing angle

What Did We Learn About Clustering?

- Clustering doesn't work as described in textbooks
  - Most algs generate one correct clustering (efficiently)
  - If not working, need a better, smarter algorithm
- Textbooks were WRONG!
  - On most data sets, no such thing as "right" clustering
  - Instead, want to find "useful" clusterings for task at hand
- Need to change how we do clustering…

Two Solutions to Clustering Problem

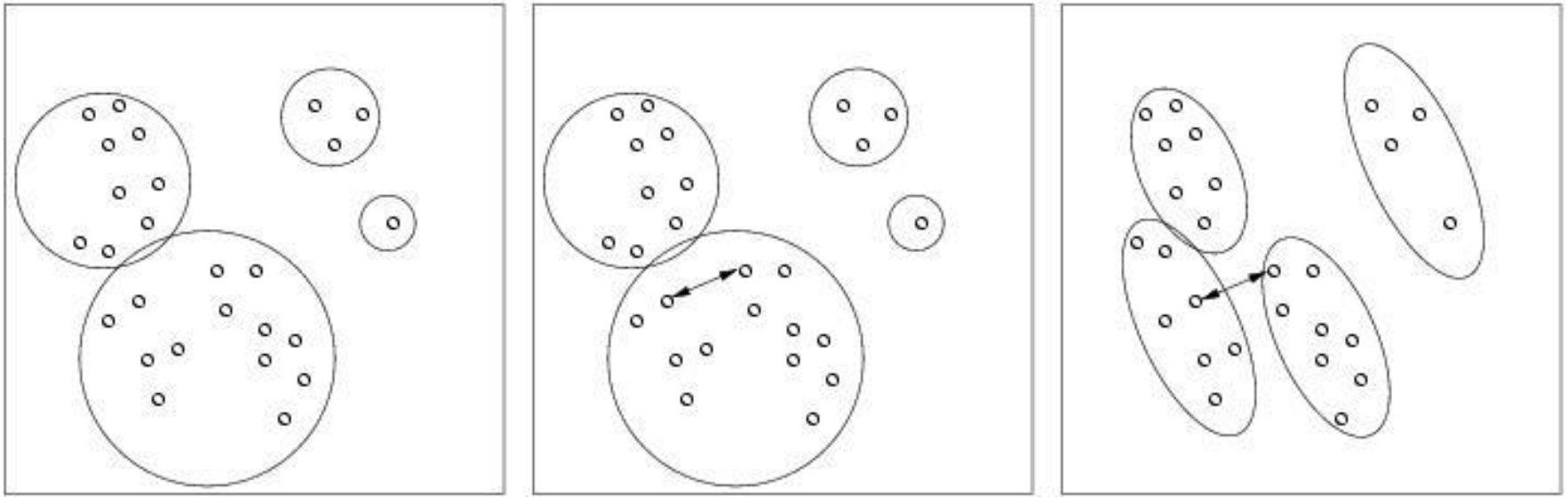
- User in the loop
  - User “steers” the clustering process
  - Iterative refinement
  - User feedback to learn distance function or set of constraints

- User after the loop
  - Use horsepower to generate many diverse high-quality clusterings
  - Human selects useful clustering(s)
  - Organize clusterings to make it easier for user to inspect them

Motivation for User-in-Loop Clustering

- 1M - 100M documents
  - Group them into classes or a hierarchy so docs can be browsed easily
  - Don’t know in advance what classes or hierarchy to use
  - Have criteria in mind, but can’t verbalize them
- This problem is ubiquitous
- No such thing as the right clustering, but you’ll know good/bad clusters when you see them

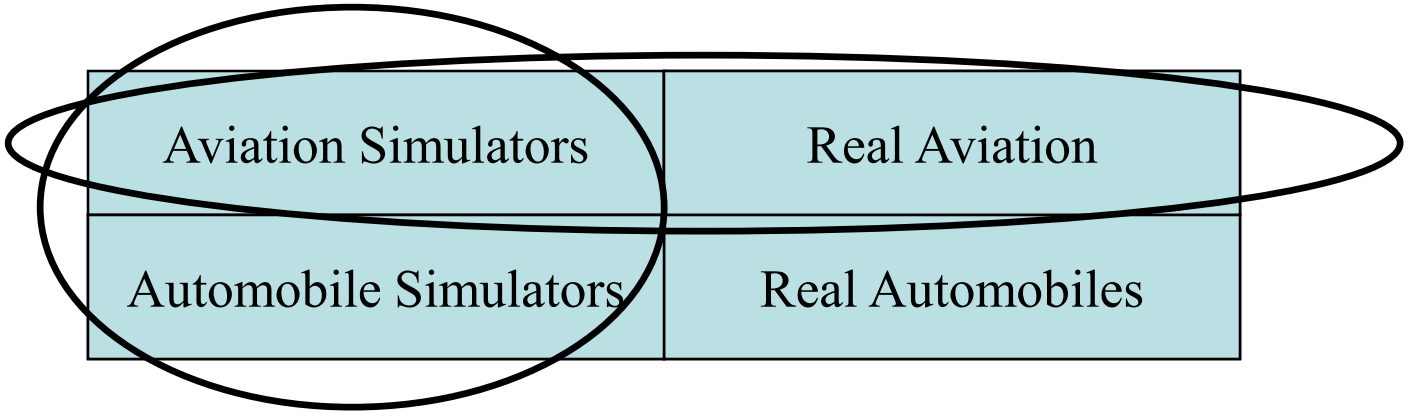
It's Easier to Criticize Than to Create



- Do initial clustering
- Like art, know good/bad clusters when you see them
- Critique the clustering (provide feedback)
- Re-cluster with algorithm sensitive to your critique
- Repeat until happy with final clustering

Proof-of-Concept Experiment

- 20,000 USENET articles from four subjects
- Different users give different pairwise constraints:



- "Accuracy" increases from 50% to 80% with 10 pairwise constraints (labels hidden from clustering)

Naive Bayes Bags-of-Words Model

Generative Model:

$$p(d) = \prod_{w_j \in V} p(w_j \mid \theta)^{N(w_j, d)}$$

Natural Distance Metric:

$$KLD_M(d_1 \| d_2) = |d_1| KLD(d_1 \| M_{12}) + |d_2| KLD(d_2 \| M_{12})$$

Learning a Distance Metric to Implement Constraints

Modify KL-Divergence:

$$KLD' = \prod_{w_j \in V} \lambda_j \cdot p(w_j \mid \theta_{d1}) \log\left(\frac{p(w_j \mid \theta_{d2})}{p(w_j \mid \theta_{d1})}\right)$$

where the lambda weights control the warping of the distance metric for each word.

Adjusting the Weights

Given constraint that documents d_1 and d_2 should be in different clusters:

$$\frac{\partial \, KLD'_M(d_1 \| d_2)}{\partial \, \lambda_j} = |d_1| p(w_j | \theta_{d1}) \log\left(\frac{p(w_j | \theta_{d1d2})}{p(w_j | \theta_{d1})}\right) +$$

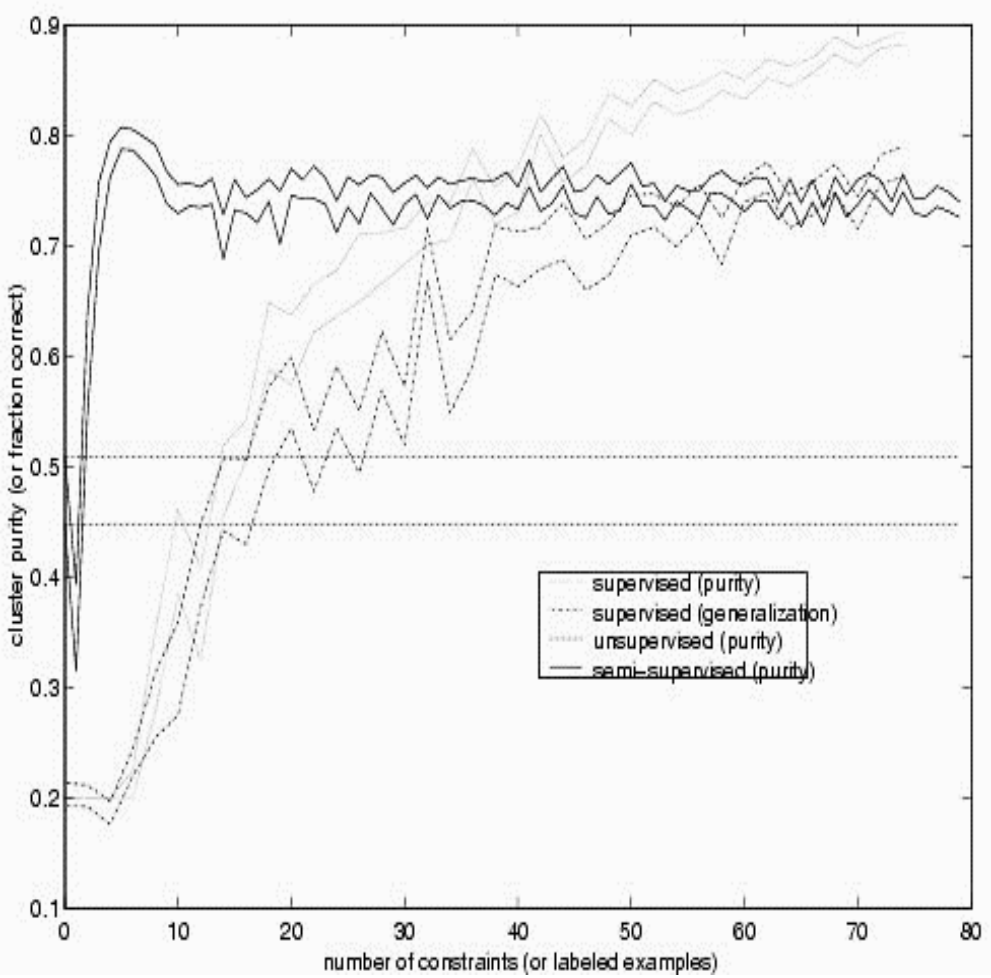
$$|d_2| p(w_j | \theta_{d2}) \log\left(\frac{p(w_j | \theta_{d1d2})}{p(w_j | \theta_{d2})}\right)$$

Text Clustering Experiment

- 5 Reuters topics:
  - *Business*
  - *Health*
  - *Politics*
  - *Sports*
  - *Tech*

- EM-based soft-clustering: compute p(c|d) from p(d|c)

- Add constraints one-at-a-time that docs with different labels should not be in same cluster



Constraints (User Feedback)

- Don’t have algorithms that can handle all kinds of constraints at same time
- Not known how to balance multiple, conflicting, incomparable constraints?
- Soft vs. hard constraints?
  - Often not possible to satisfy all constraints
- Re-cluster from scratch, or incrementally?
- Serious UI/HCI issues…

Some users want a clustering system that can be driven like a car, with a set of controls, a dashboard full of indicators, and a windshield to see the current clustering.

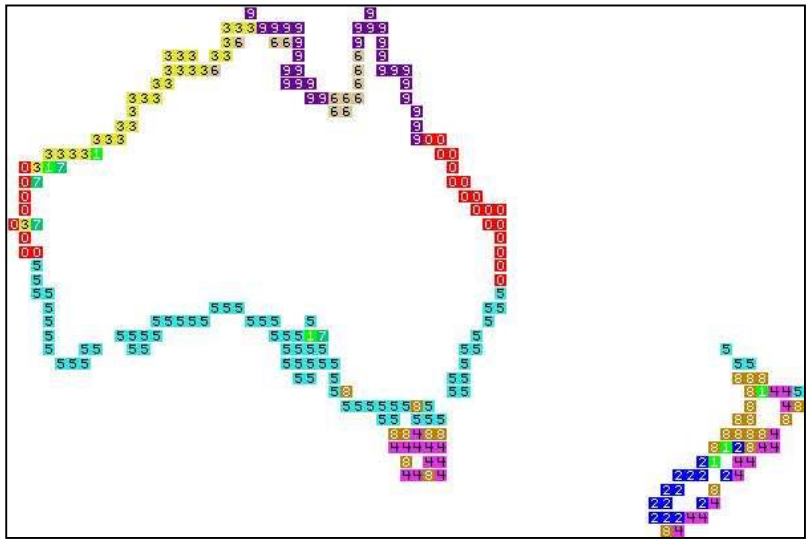
Other users want to take the bus.

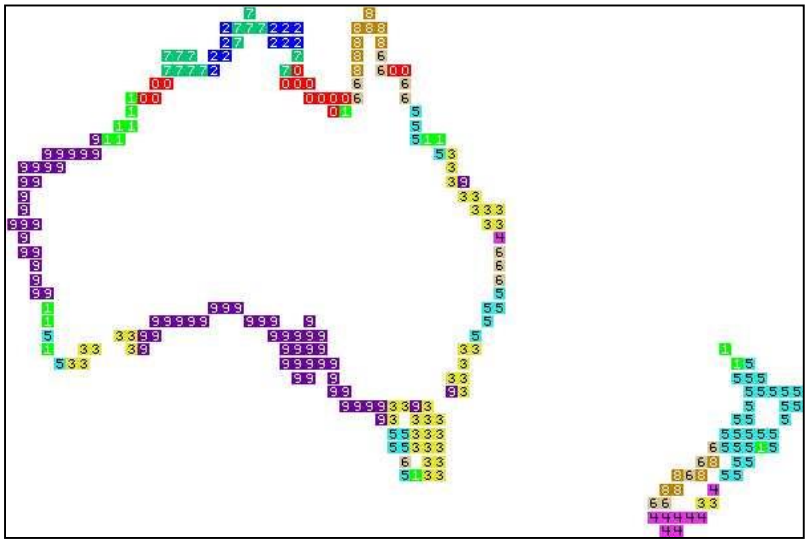
User After the Loop: Meta Clustering

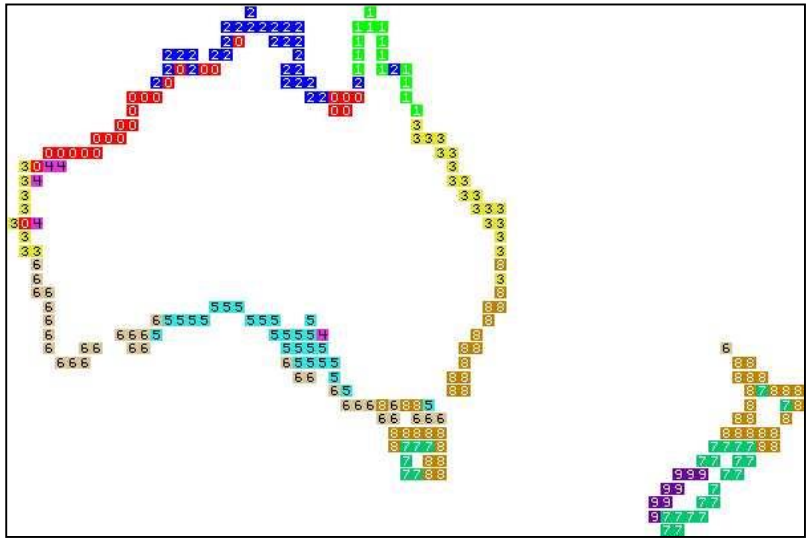
- Automatically generate many different clusterings
- Cluster the *clusterings* to organize results
- Show user clusters of clusterings (MetaClusters)

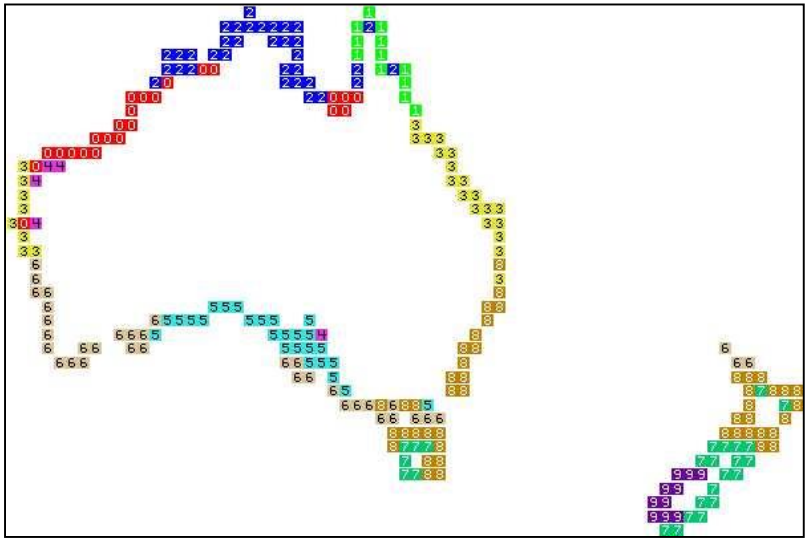
- Human out of main loop: just select best clustering
- No need to "learn" distance metric
  - Returns distance metric after user selects best clustering

4 Different Clusterings of Australia Coastline









Research Questions

- How to generate different clusterings?
  - EM/k-means gets stuck in local minima
  - Stochastically re-weight distance metric
- How to measure distance between clusterings?
  - Hop distance (hierarchical clustering methods)
  - Rand Score: pair-wise overlap
- How to organize clusterings for user?
  - Hierarchical Clustering
  - MDS?
- How to combine/merge clusterings?
  - Consensus clustering

Data Sets

- 20 NewsGroups
- Australia Coastline
- Bergmark Web Searches
- Blogs (Buzzmetrics)
  - Sample blog data by author, by blog, by time
  - 9 Prolific Authors (but not blog admins)
  - 30 Blogs
  - 5000 documents
  - > 100 documents per author
- Cover Type (satellite tree ground cover)
- Letter Recognition
- Phonemes

How to Evaluate Clusterings?

- 20 NewsGroups: classify into newsgroups
- Australia Coastline: expert only
- Bergmark Web Searches: keywords used for web crawls
- Blogs (Buzzmetrics)
  - Measure accuracy by author, by blog, by time
  - 9 Prolific Authors (but not blog admins)
  - 30 Blogs
  - 5000 documents
  - > 100 documents per author
- Cover Type (satellite tree ground cover): true ground cover
- Letter Recognition: letter
- Phonemes: speaker or phoneme label

Generating Diverse Clusterings

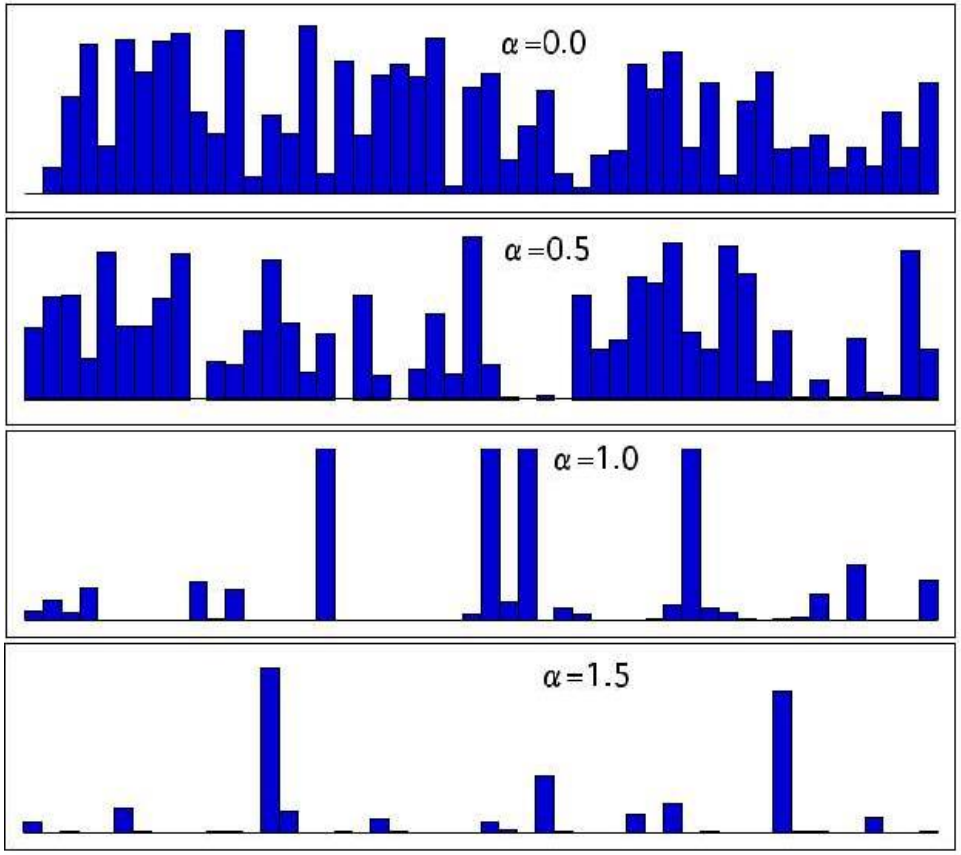
- Collect K-Means local minima from random restarts
  - effective on some problems, not on others

- Random weighting of features in distance function

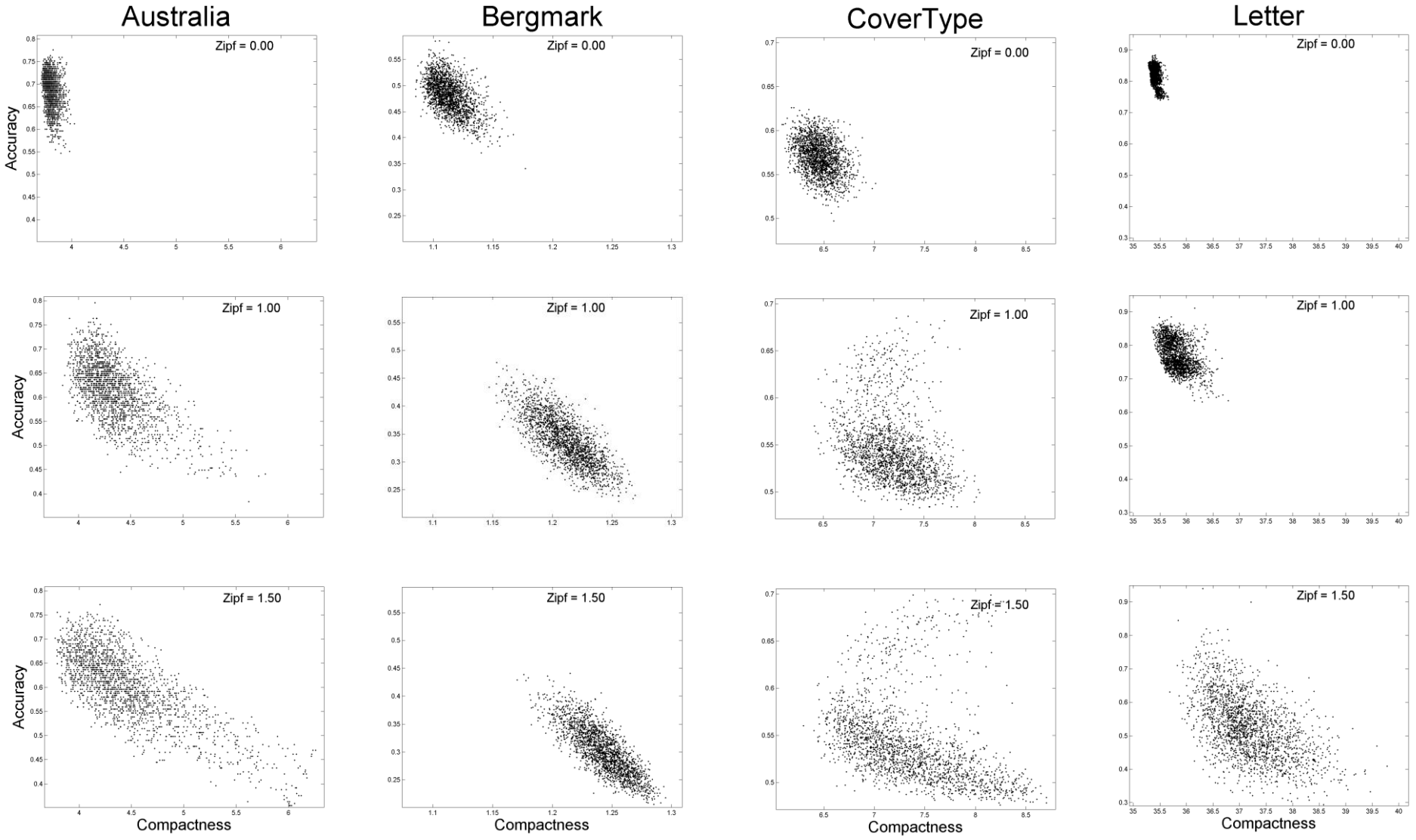
$$Dist(O_i, O_j) = \sum_l w_l (x_{i,l} - x_{j,l})^2$$

  - uniform random weights?
  - power-law random weights
    - some evidence that feature importance often is power-law distributed
    - we'll use well-known Zipf distribution: $P(i) \propto \frac{1}{i^{\alpha}}$

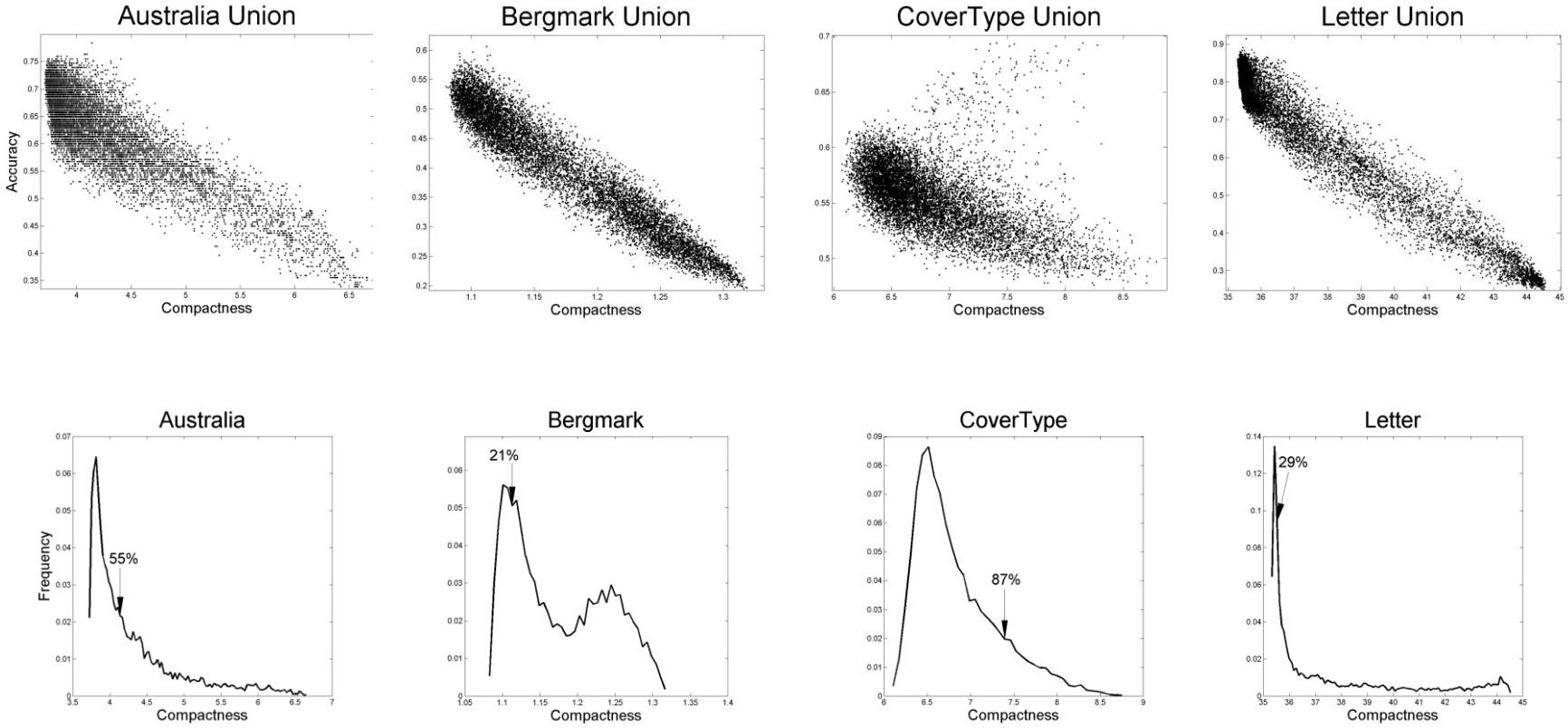
Zipf Distribution Weights



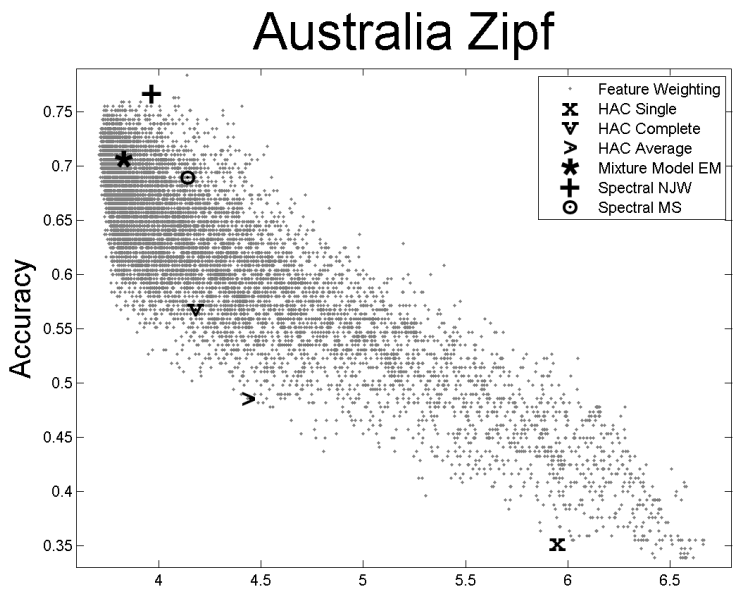
Zipf Shape Parameter Affects Diversity

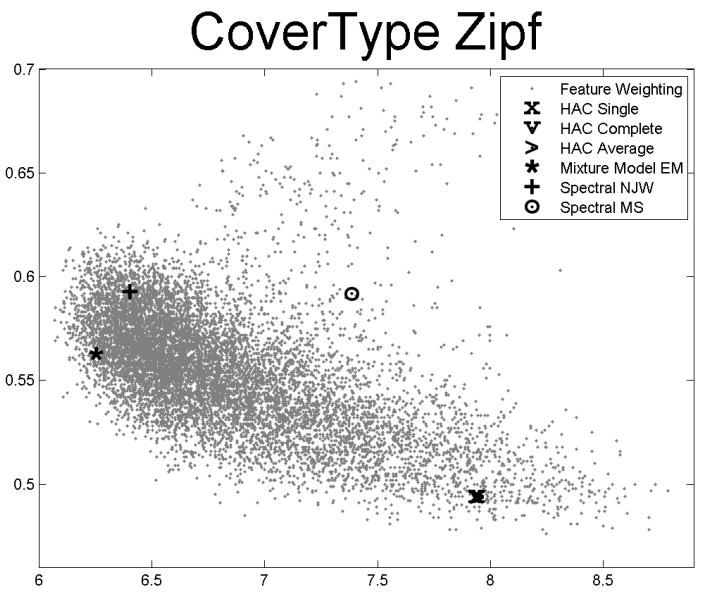


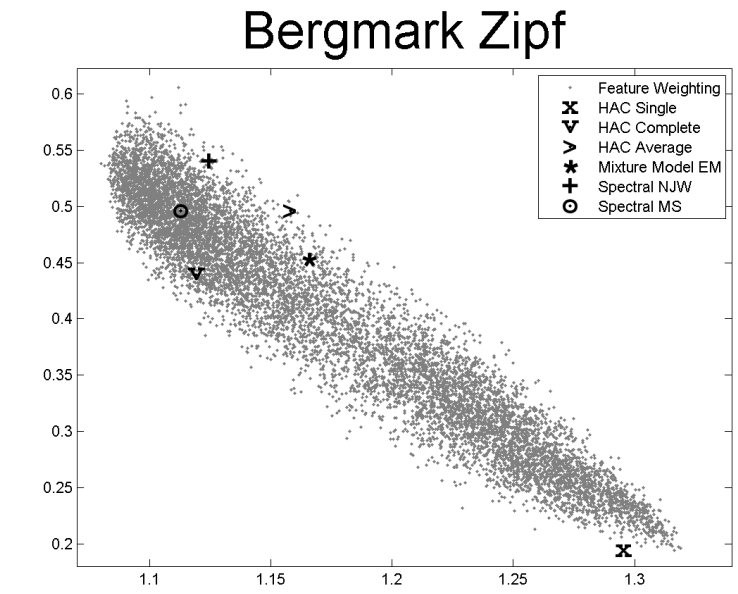
Compactness ≠ Accuracy

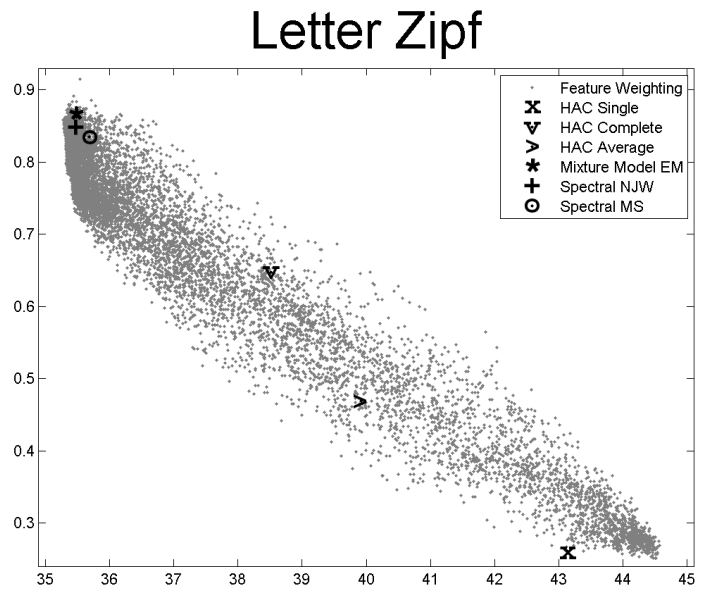


Clustering Methods Other Than K-Means?

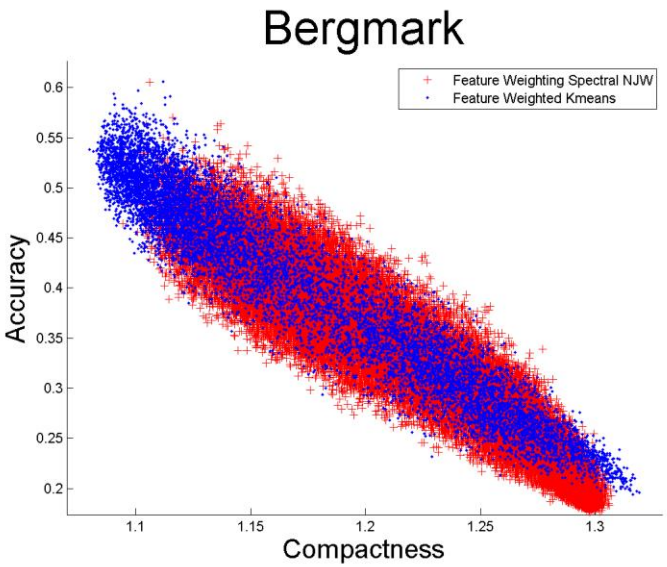


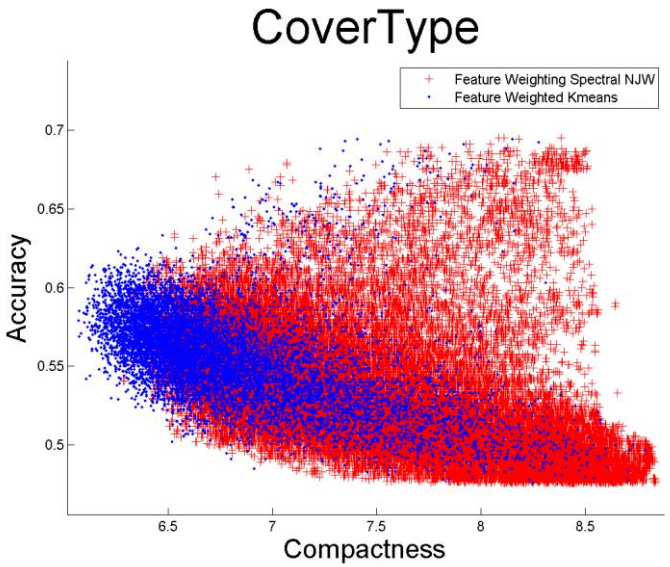






K-Means & Spectral Qualitatively Different

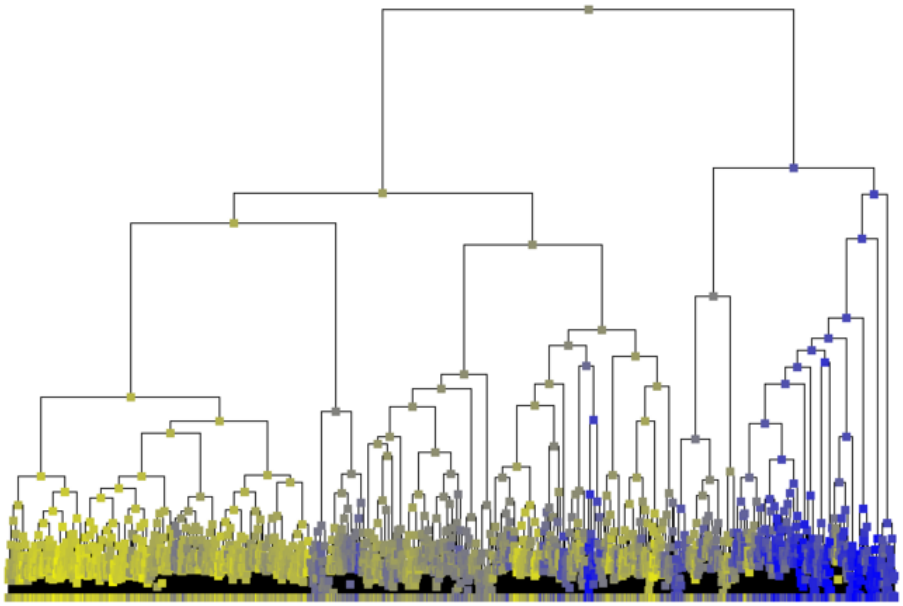




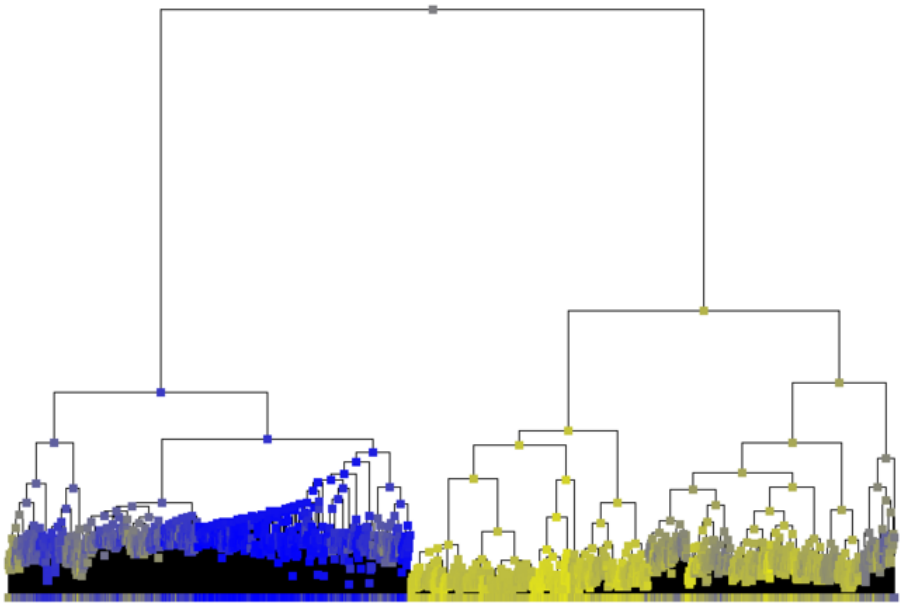
So What Do We Show To Users?

- Now we can generate 100k different clusterings
- Won't have accuracy on real clustering problems
  - If you have a computable criterion like accuracy for selecting the best clustering, if possible optimize the clustering process to it (not always feasible)
- Can't show users 100k clusterings --- hard to look at one clustering, let alone pick best of 100k
- Can only show users a few clusterings
- Cluster the clusterings: MetaClustering
  - need a distance metric over clusterings
  - and a clustering algorithm for the meta-level
- Show users representative (consensus) clusterings or allow users to browse hierarchy of clusterings

MetaClustering: Clustering the Clusterings



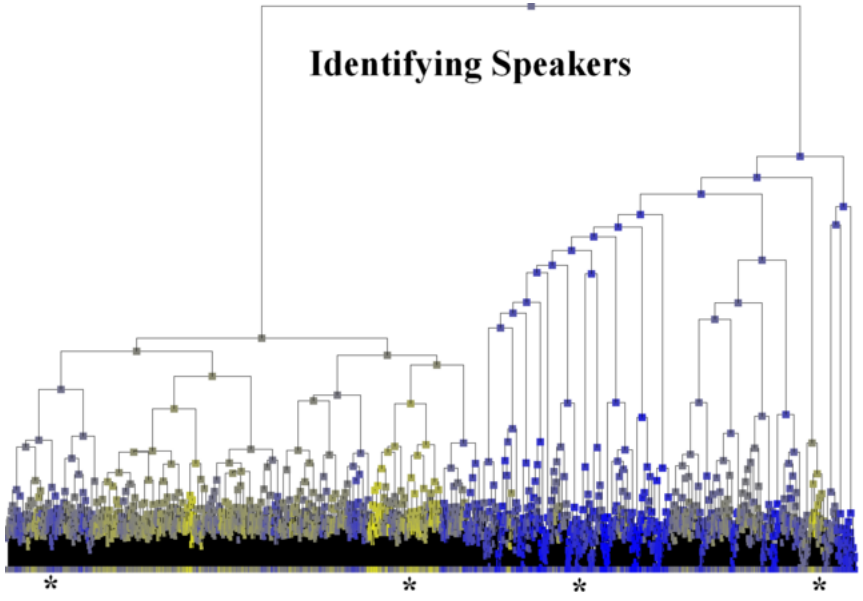
Australia

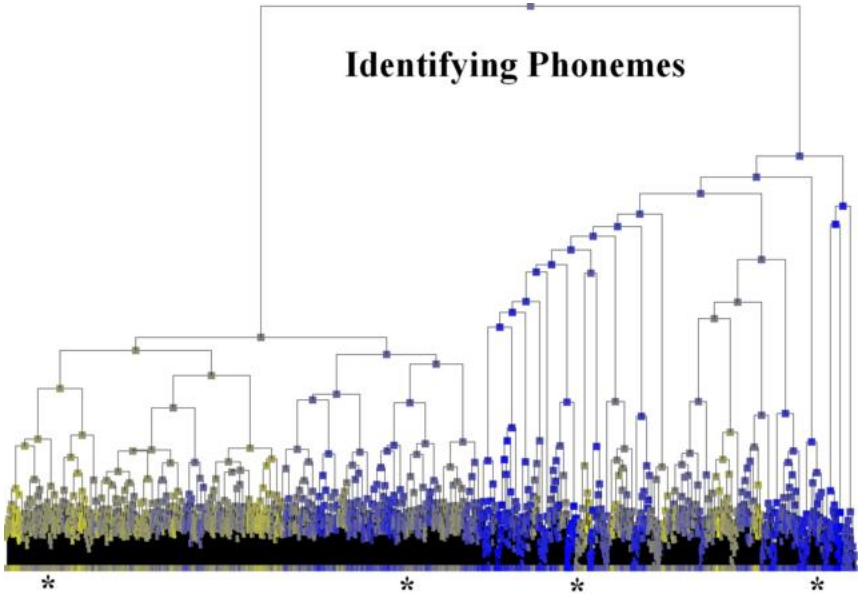


Letter

Clusterings with high quality tend to be close in the clustering tree.

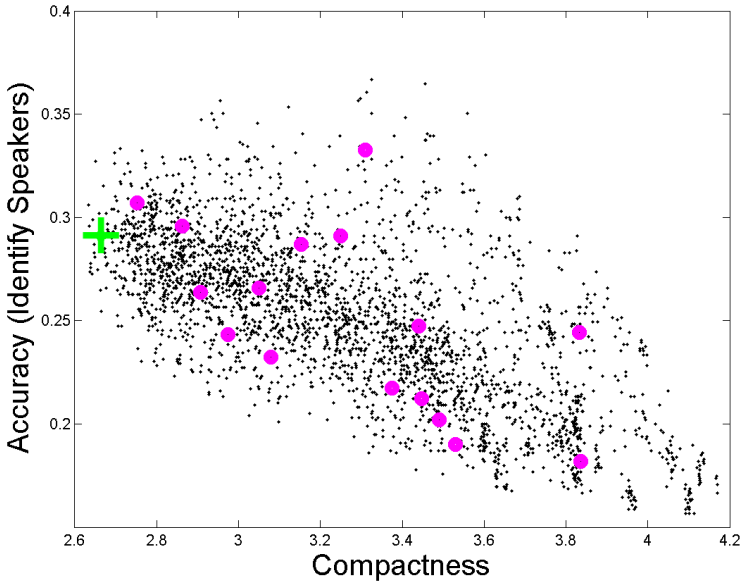
Clustering the Clusterings (Phoneme Data)

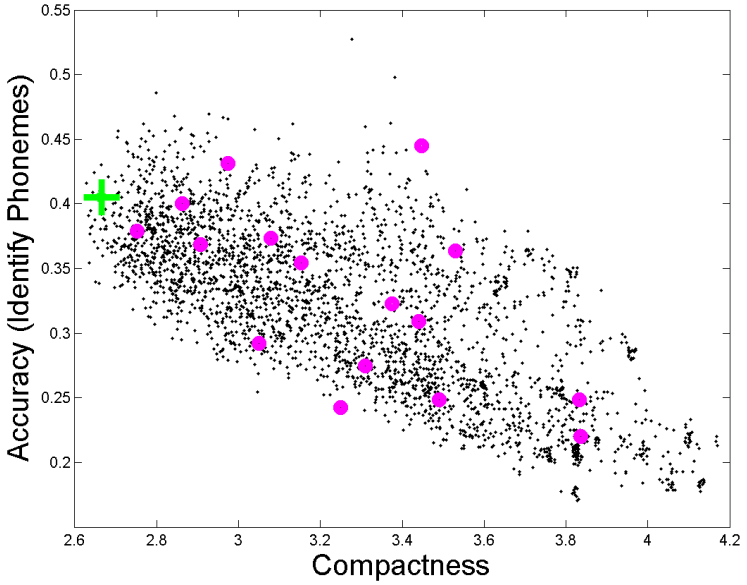


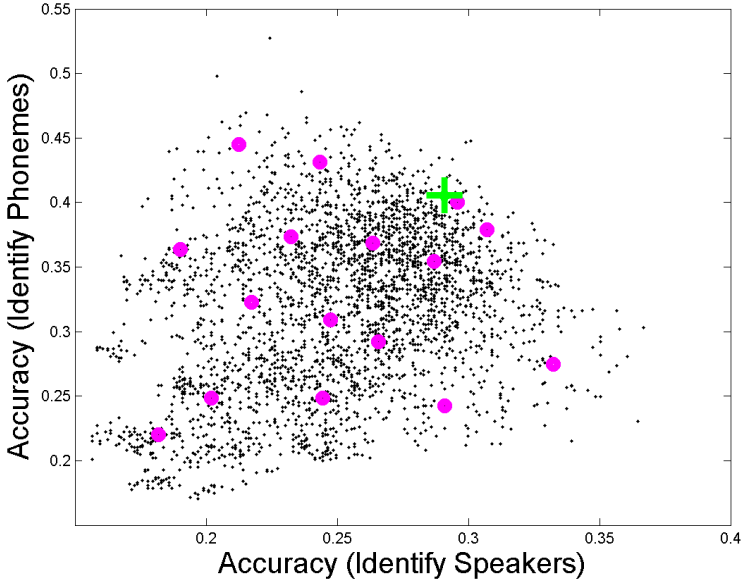


"paint" meta clustering tree with any info you have

Consensus Clusterings at Meta-level (Phoneme Data)







- Multiple Clusterings
- Aggregation of All Clusterings
- Aggregation of Meta Clusterings

Take Home Messages

- There are many different clusterings of complex datasets
  - Unlikely to find clustering you want on 1st try
  - Most "compact" clustering is rarely the "right" clustering
  - Will never be able to fully automate clustering
  - Going to need to roll up your sleeves and dive in
- Use side-information to guide clustering
- Allow human to drive clustering
- Generate Diverse Clusterings, MetaCluster, then Select
- Need more research: both algorithms and user interfaces
- I now prefer Meta-Clustering where feasible because diversity can take you places you didn't expect to go
- In rare cases there is a unique answer: phylogenetic trees

A Few References

- Semi-Supervised Clustering with User Feedback (1999-2003)
  - David Cohn, Rich Caruana, Andrew McCallum
- An Impossibility Theorem for Clustering (2002)
  - Jon Kleinberg
- Meta Clustering (2006)
  - Rich Caruana, Mohamed Elhawary, Nam Nguyen, Casey Smith
- Non-redundant Multi-view Clustering via Orthogonalization (2007)
  - Y Cui, X. Fern, J. Dy
- Simultaneous Unsupervised Learning of Disparate Clusterings (2008)
  - P. Jain, R. Meka, I. Dhillon
- Clustering: Art or Science? (2009)
  - Isabelle Guyon, Ulrike von Luxburg, Robert Williamson
- Multiple Non-Redundant Spectral Clustering Views (2010)
  - Donglin Niu, Jennifer Dy, Michael Jordan

Workshops

- NIPS 2009: Clustering: Science or Art?
- KDD 2010: MultiClust: Discovering, Summarizing, and Using Multiple Clusterings
- PKDD 2011: MultiClust II
- SIAM 2012: MultiClust III
- KDD 2013: MultiClust IV

Thank You!