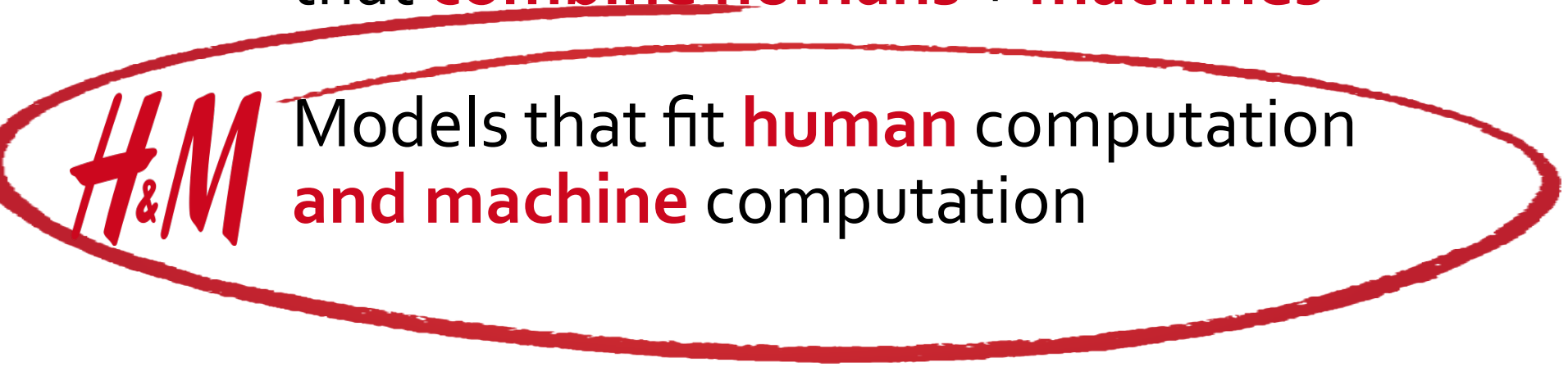
# **Meta-unsupervised-learning:** A model of unsupervised learning applicable to humans and computers

Vikas Garg – MIT

Adam Tauman Kalai – MSR

# Human Computation Theories

H+M Models of computational systems that **combine humans** + **machines**



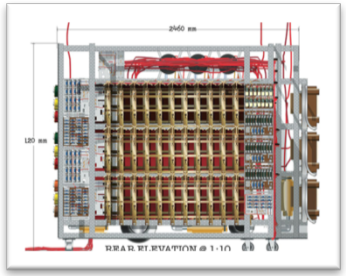

Models that fit **human** computation **and machine** computation

Blum-Vempala HUM (2015)



H&M

Turing Machine (1936)

Common characteristics

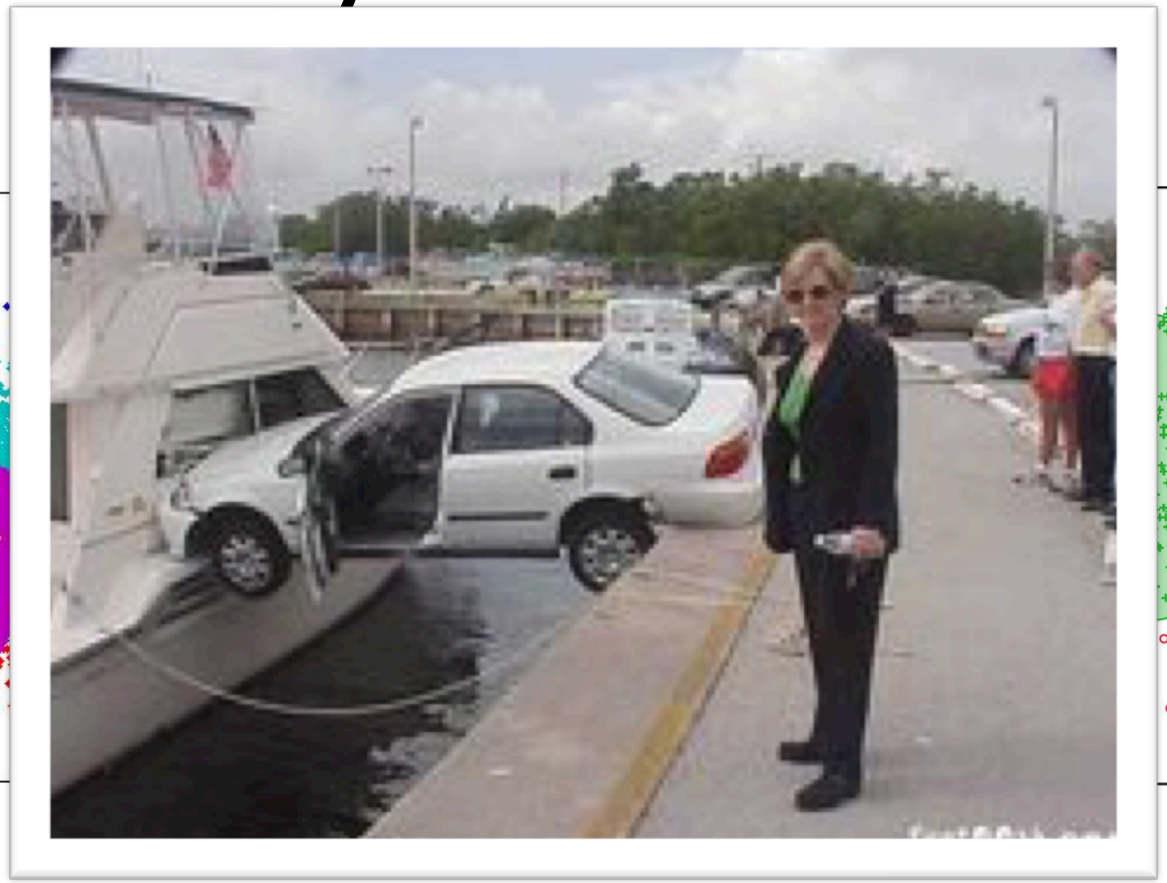
- Time-space-accuracy tradeoffs
- Easier to check than to solve
- Reductions, parallelization, …

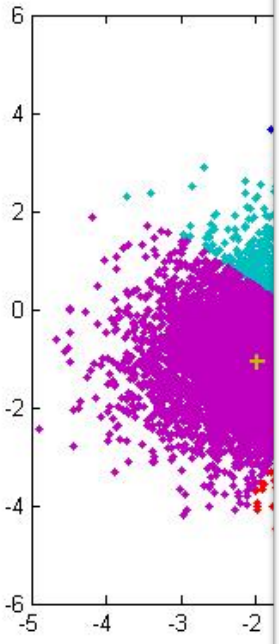
# Unsupervised learning

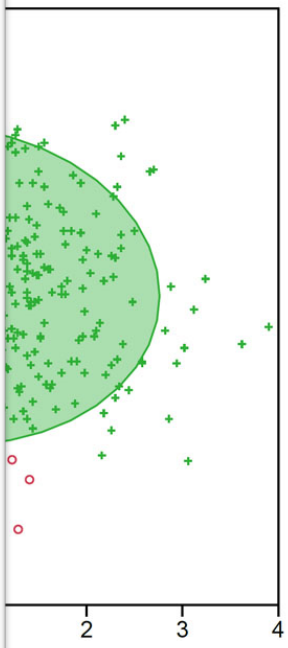
Finding **structure** in **unlabeled data** via:

- Clustering
- Dimensionality reduction
- Dictionary learning
- Topic modeling
- Manifold learning

# Geometry of clustering






$k$-means

Gaussian mixture model

# Human vs. Machine Clustering

The class was great!

Waste of time ADAM I HOPE YOU DIE.

The class was boring.

Best course I have taken in a long time.

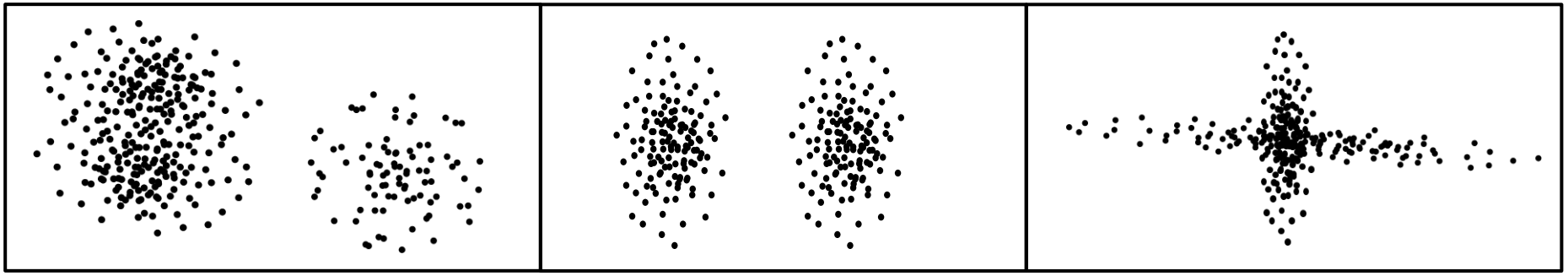
The class was terrible!

I slept the whole time.

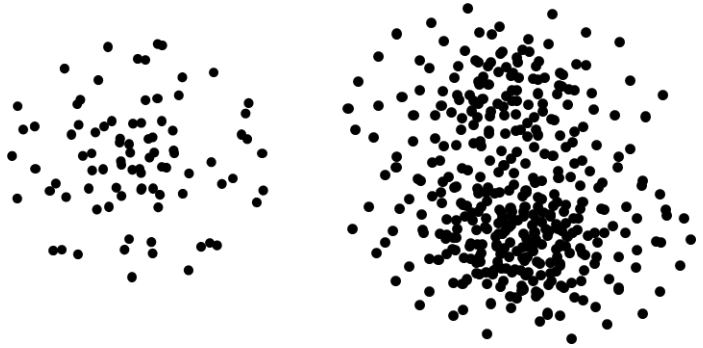
**The class was** great!
**The class was** boring.
**The class was** terrible!

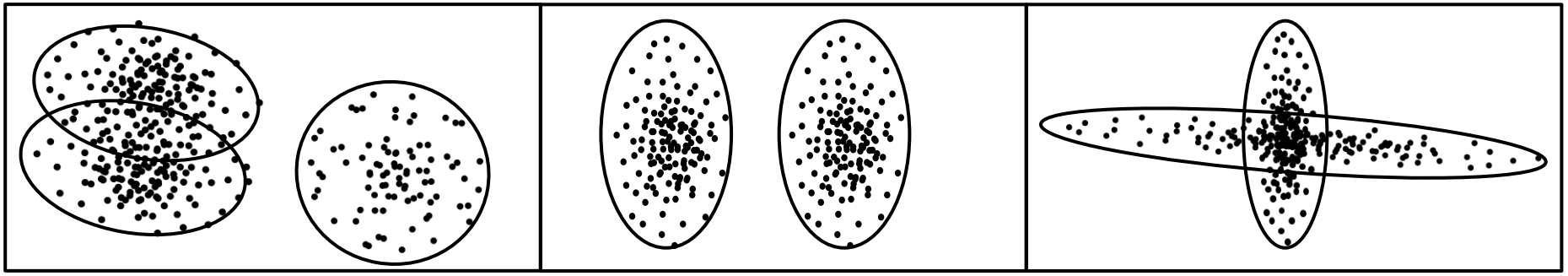
Best course **I** have taken in a long **time**.
Waste of **time** ADAM **I** HOPE YOU DIE.
**I** slept the whole **time**.

H&M idea: use prior data

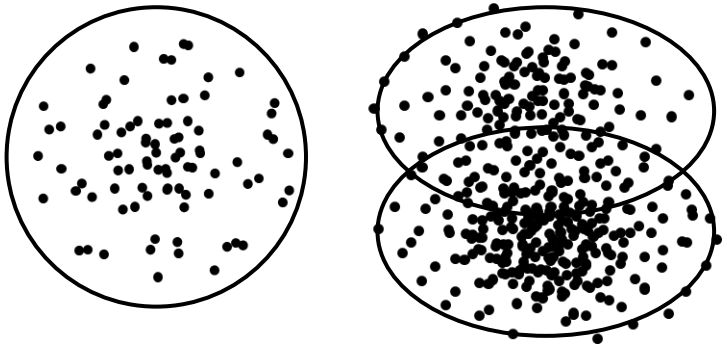


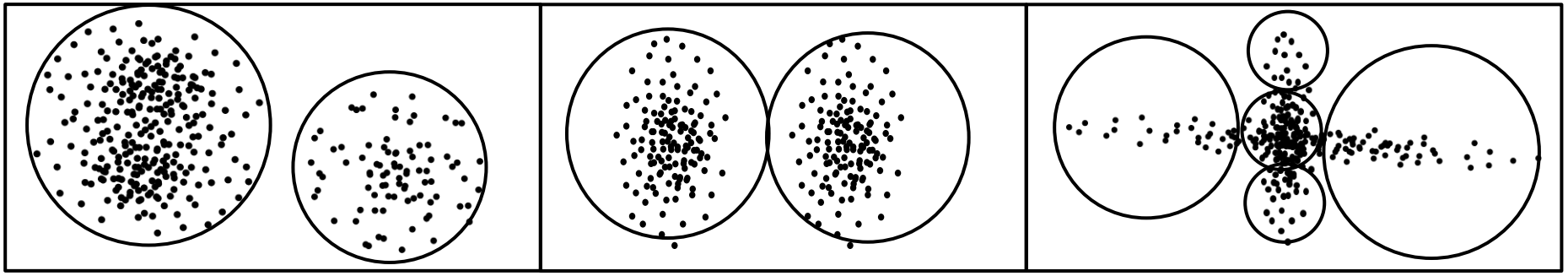
# Clustering with prior labeled data



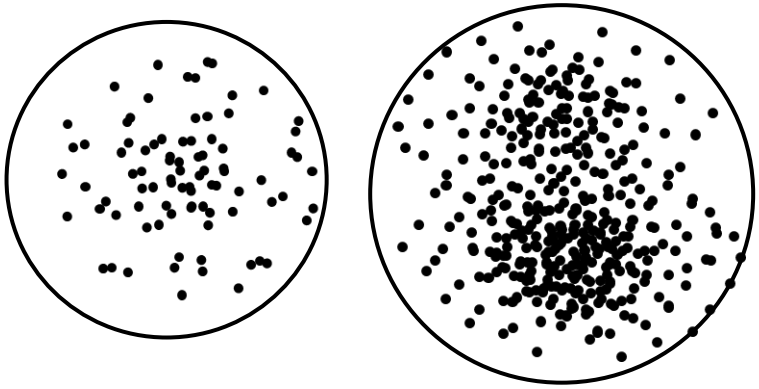


# Clustering with prior labeled data





# Clustering with prior labeled data



# META-MODEL

# Meta-clustering

#dim #data

- Unlabeled data $X \in U d \geq 1, m \geq 1 \uparrow \blacksquare \mathbb{R} \uparrow d \times m$
- Ground truth clustering $Y \in \Pi(X)$
- Predicted clustering $Z \in \Pi(X)$
- Accuracy $\mathrm{acc}(Y,Z) \geq 0$
- Meta-dist. $\mu$ over labeled problems $(X,Y)$
- Training problems $T = (X \downarrow 1, Y \downarrow 1), \ldots, (X \downarrow n, Y \downarrow n) \sim \mu \uparrow n$
- Meta-algorithm $M(T)$ outputs clustering alg. $C$

# Def: Efficient meta-clustering

Fix a class of clustering algorithms $\mathcal{C}$.

$M$ *efficiently meta-clusters as well as* $\mathcal{C}$ *if:*
for any meta-dist. $\mu$ over clustering problems,
for any $\epsilon,\delta>0$, $n \geq poly(1/\epsilon, 1/\delta)$,

$$\Pr_{T\sim\mu^n}\left[\mathrm{acc}_{\downarrow\mu}(M(T)) \geq \max_{C\in\mathcal{C}} \mathrm{acc}_{\downarrow\mu}(C) - \epsilon\right] \geq 1-\delta$$

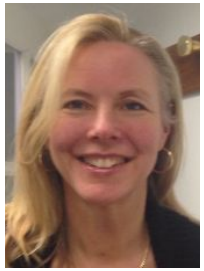
$$\mathbf{E}_{(X,Y)\sim\mu} \mathrm{acc}(Y, C(X))$$

Also, $M$, $M(T)$ must be poly-time.

# Meta-unsupervised learning

Similar to theory of supervised learning:

PAC (Valiant 84)  Agnostic (Kearns, Schapire, Sellie 94)



# Meta-clustering Occam bound

$C\downarrow \mathrm{EMP} \in \mathcal{C}$ maximizes empirical accuracy on training data $T$

$$\Pr_{T\sim\mu^n}\left[\mathrm{acc}_\mu(C\downarrow EMP) \geq \max_{C\in\mathcal{C}} \mathrm{acc}_\mu(C) + \sqrt{2/n \log|\mathcal{C}|/\delta}\right] \geq 1-\delta$$

#bits

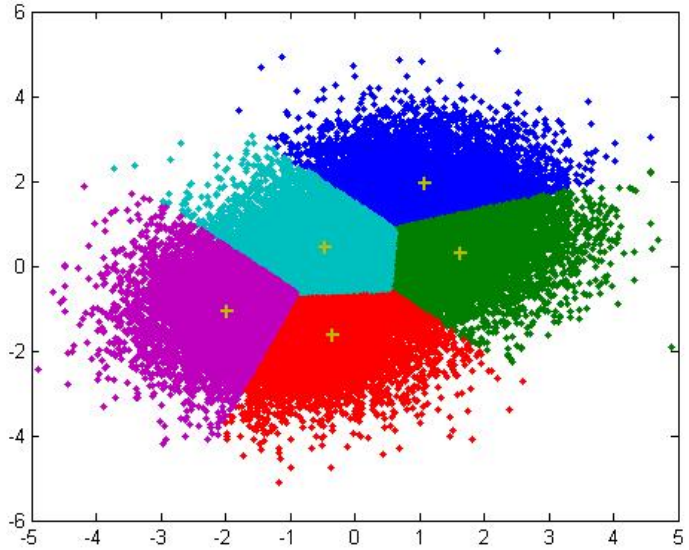
If $M$ efficiently finds $C\downarrow \mathrm{EMP}$,
then $M$ efficiently meta-clusters as well as $\mathcal{C}$
using #problems $\approx$ #bits to represent $\mathcal{C}$

# EXPERIMENTAL EVALUATION

Silhouette heuristic (intrinsic clustering quality)

Rand Index (accuracy metric w.r.t. ground-truth/gold-standard)

# Silhouette clustering quality heuristic




How much closer, on average, are points to other points in their same cluster than to points in the nearest other cluster? (normalized to be in $[-1, 1]$)

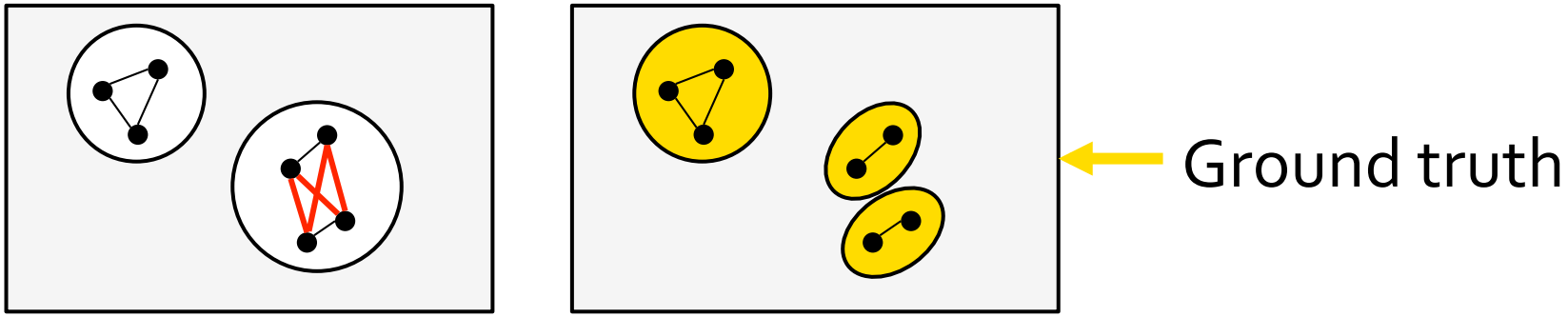
$$a = \text{Avg}_{x' \in C(x)} |x - x'|$$

$$b = \min_{C' \neq C(x)} \text{Avg}_{x' \in C'} |x - x'|$$

$$\text{Silh} = \text{Avg}_x \, b(x) - a(x) / \max\{a(x), b(x)\}$$

Often used to pick #clusters $k$

# Rand Index: ground truth *accuracy*




- Fraction of edge/non-edge agreement pairs
- Adjusted Rand Index (ARI)
  - Corrects for random chance
  - Common empirical clustering accuracy metric


International Federation of Classification Societies

# Cluster Benchmark Data Repository

Search ...

- Welcome to Repository
- Data Sets
- Philosophy
- Citation Policy
- Disclaimer
- Contribute a Data Set
- Licenses
- Contact
- IFCS Homepage

## Welcome to the IFCS Cluster Benchmark Data Repository

The aim of this Repository is to stimulate better practice in benchmarking (performance comparison of methods) for cluster analysis by providing a variety of well documented high quality datasets and simulation routines for use in practical benchmarking.

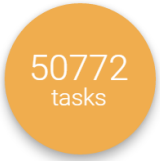
The repository collects datasets with and without given "true" clusterings. A particular feature of the repository is that every dataset comes with a comprehensive documentation, including information on the specific nature of the clustering problem in this dataset and the characteristics that useful clusters should fulfill, with scientific justification.

**Currently < 10 datasets**

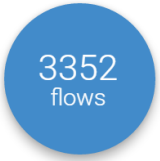
OpenML beta

Exploring machine learning better, together


Find or add **data** to analyse




Download or create scientific **tasks**




Find or add data analysis **flows**

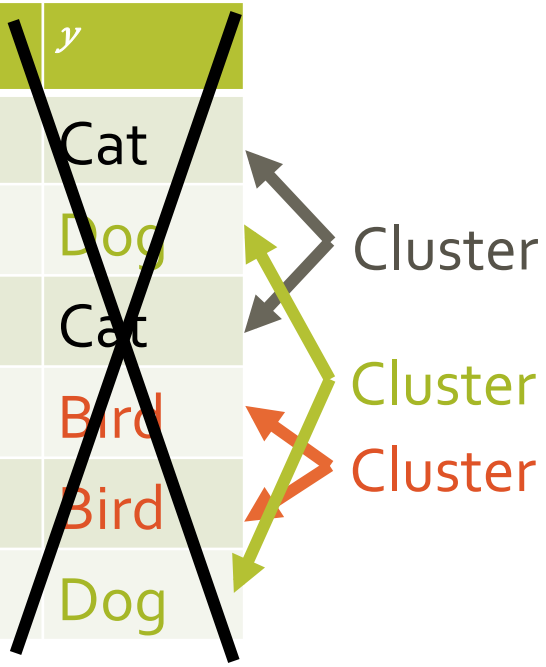

Upload and explore all **results** online.

# Classification data → Clustering data

| $x$ | | | | | | | $y$ |
|---|---|---|---|---|---|---|---|
| 1 | 2.2 | 2 | 1.3 | 1.7 | 4 | -1 | Cat |
| 1 | 1.7 | 1 | 2.4 | 3.9 | 5 | -1 | Dog |
| 0 | 3.9 | 4 | 17 | 2.8 | -1 | 1 | Cat |
| 1 | 2.8 | 5 | 17 | 1.1 | 2 | -1 | Bird |
| 0 | 1.1 | -1 | 17 | 3.8 | 0 | -1 | Bird |
| 1 | 4.5 | 2 | 17 | 1.1 | 8 | 1 | Dog |

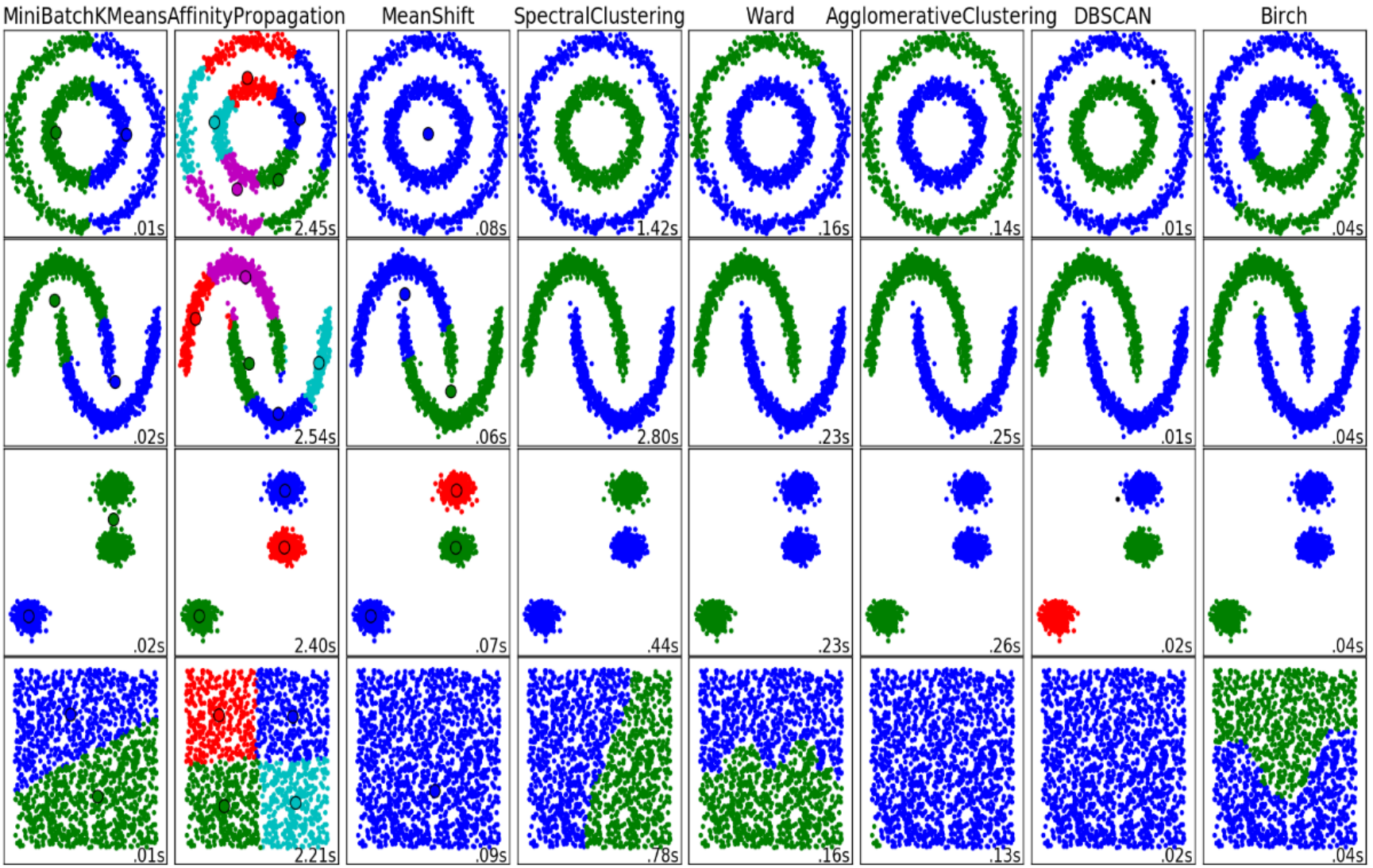
# Classification data → Clustering data

| *x* | | | | | | | *y* |
|---|---|---|---|---|---|---|---|
| 1 | 2.2 | 2 | 1.3 | 1.7 | 4 | -1 | Cat |
| 1 | 1.7 | 1 | 2.4 | 3.9 | 5 | -1 | Dog |
| 0 | 3.9 | 4 | 17 | 2.8 | -1 | 1 | Cat |
| 1 | 2.8 | 5 | 17 | 1.1 | 2 | -1 | Bird |
| 0 | 1.1 | -1 | 17 | 3.8 | 0 | -1 | Bird |
| 1 | 4.5 | 2 | 17 | 1.1 | 8 | 1 | Dog |



# WHICH ALGORITHM?

# 2.3.1. Overview of clustering methods




A comparison of the clustering algorithms in scikit-learn

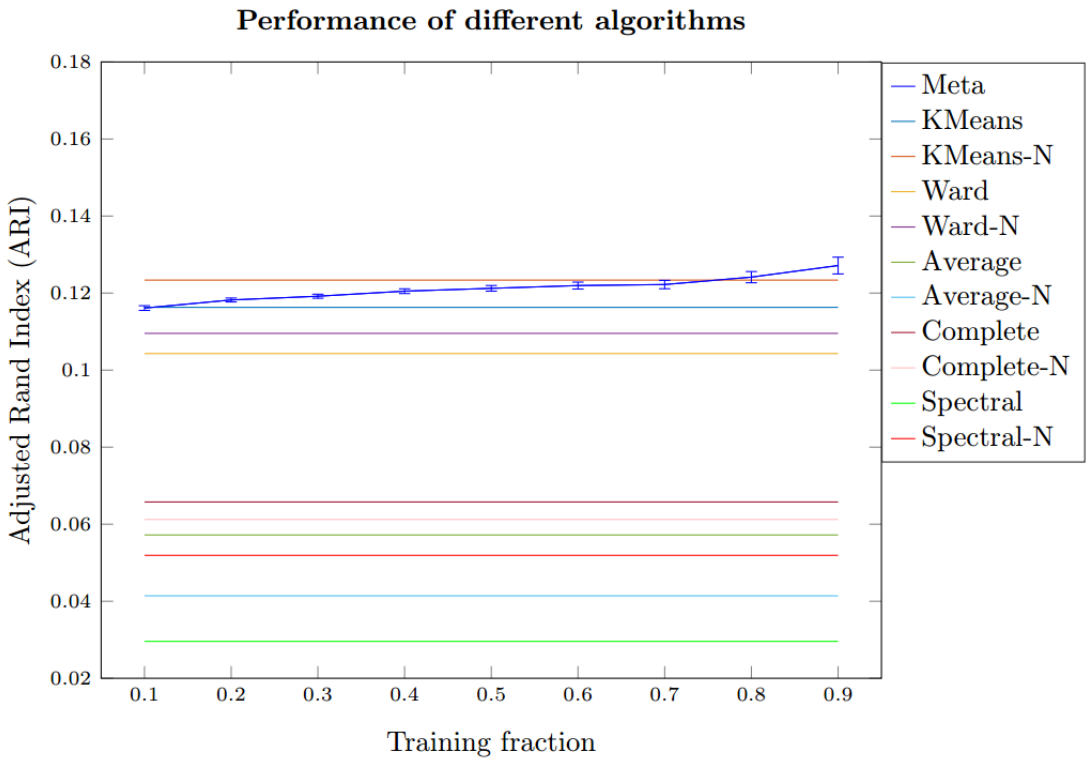
| Method name | Parameters | Scalability | Usecase | Geometry (metric used) |
|---|---|---|---|---|
| K-Means | number of clusters | Very large `n_samples`, medium `n_clusters` with MiniBatch code | General-purpose, even cluster size, flat geometry, not too many clusters | Distances between points |
| Affinity propagation | damping, sample preference | Not scalable with n_samples | Many clusters, uneven cluster size, non-flat geometry | Graph distance (e.g. nearest-neighbor graph) |
| Mean-shift | bandwidth | Not scalable with `n_samples` | Many clusters, uneven cluster size, non-flat geometry | Distances between points |
| Spectral clustering | number of clusters | Medium `n_samples`, small `n_clusters` | Few clusters, even cluster size, non-flat geometry | Graph distance (e.g. nearest-neighbor graph) |
| Ward hierarchical clustering | number of clusters | Large `n_samples` and `n_clusters` | Many clusters, possibly connectivity constraints | Distances between points |
| Agglomerative clustering | number of clusters, linkage type, distance | Large `n_samples` and `n_clusters` | Many clusters, possibly connectivity constraints, non Euclidean distances | Any pairwise distance |
| DBSCAN | neighborhood size | Very large `n_samples`, medium `n_clusters` | Non-flat geometry, uneven cluster sizes | Distances between nearest points |
| Gaussian mixtures | many | Not scalable | Flat geometry, good for density estimation | Mahalanobis distances to centers |
| Birch | branching factor, threshold, optional global clusterer. | Large `n_clusters` and `n_samples` | Large dataset, outlier removal, data reduction. | Euclidean distance between points |

# Which clustering algorithm?

Fix $k$=2 clusters

250 binary classification problems from openml.org

5 algorithms * 2 normalizations




**Meta** takes max ARI estimate from:

- # dimensions
- # examples
- **Silhouette score**
- max/min singular value

# HOW MANY CLUSTERS?

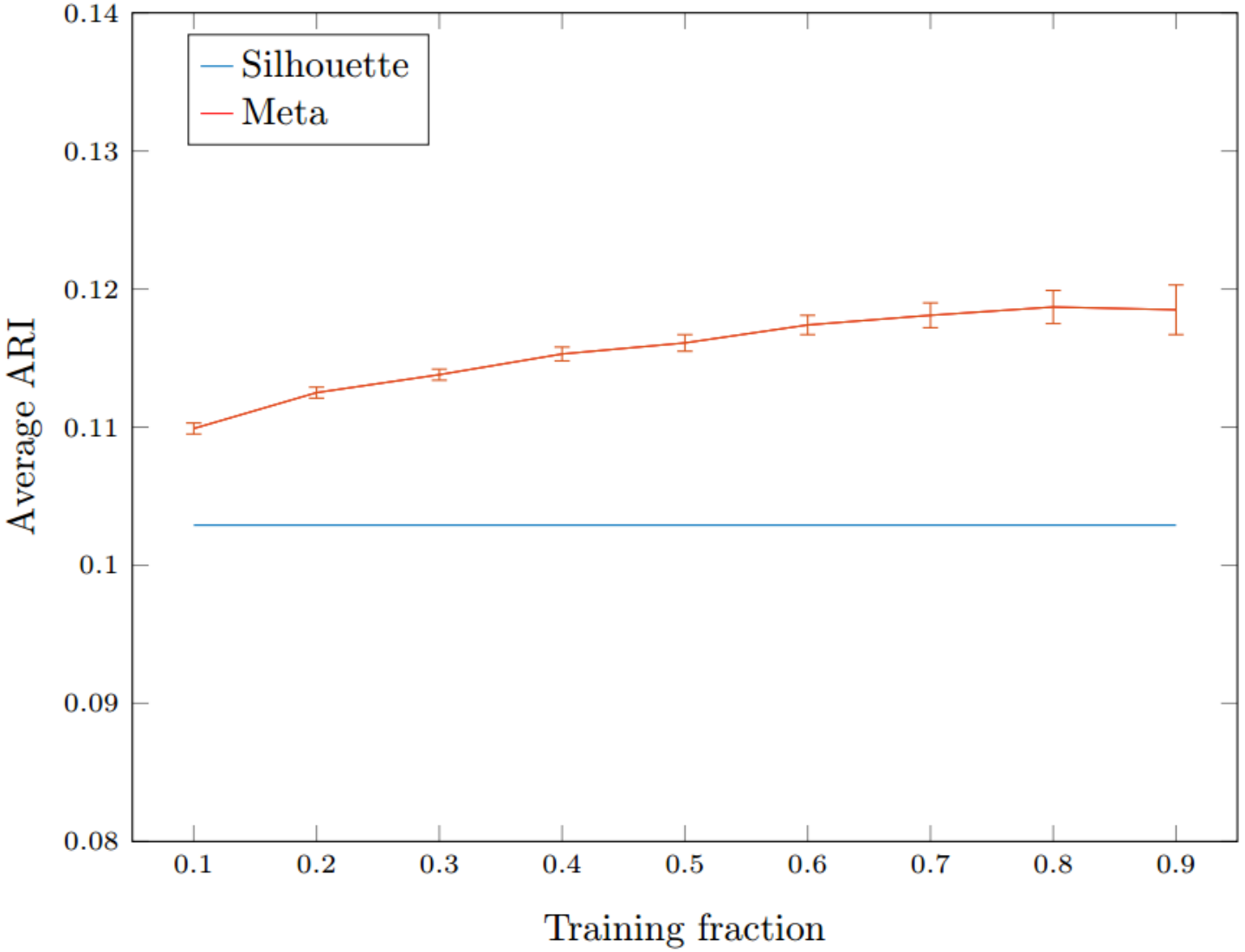
# Std. Silh approach to pick $k$

- Given new clustering problem, run $k$-means for $k=2,\ldots,10$ and compute $\text{Silh}\downarrow 2$ , …,$\text{Silh}\downarrow 10$
- Take $k$ to maximize $\text{Silh}\downarrow k$

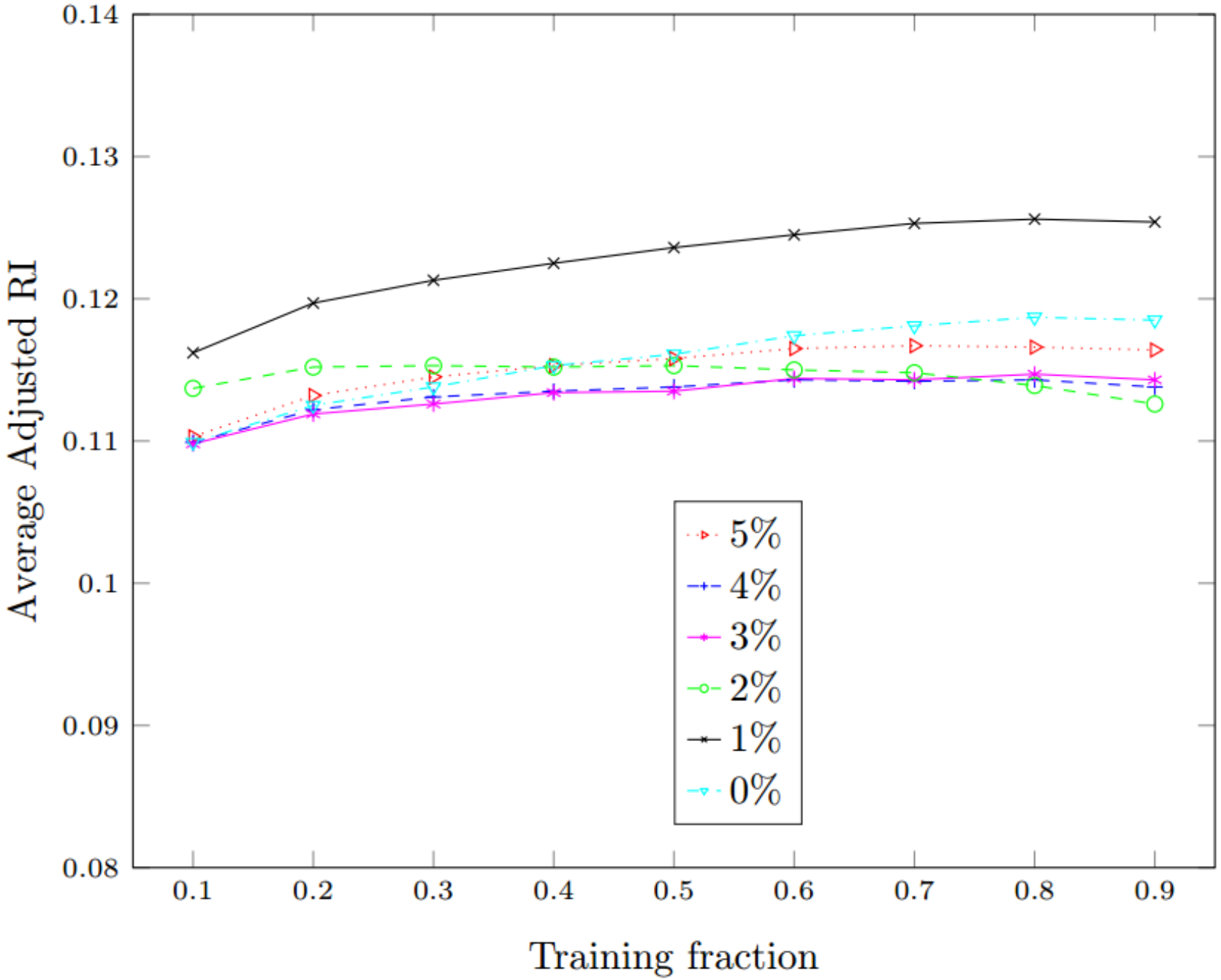
# Meta-$k$

- For each $k=2,\ldots,10$, fit $\mathrm{ARI}_k = \alpha_k + \beta_k \mathrm{Silh}_k$
  - Use least-squares linear regression
  - Data is ground-truth problems from openml.org
- Given new clustering problem, compute $\mathrm{Silh}_2, \ldots, \mathrm{Silh}_{10}$ and estimate $\mathrm{ARI}_2, \ldots, \mathrm{ARI}_{10}$
- Take $k$ to maximize $\mathrm{ARI}_k$

# Choosing # clusters $k$

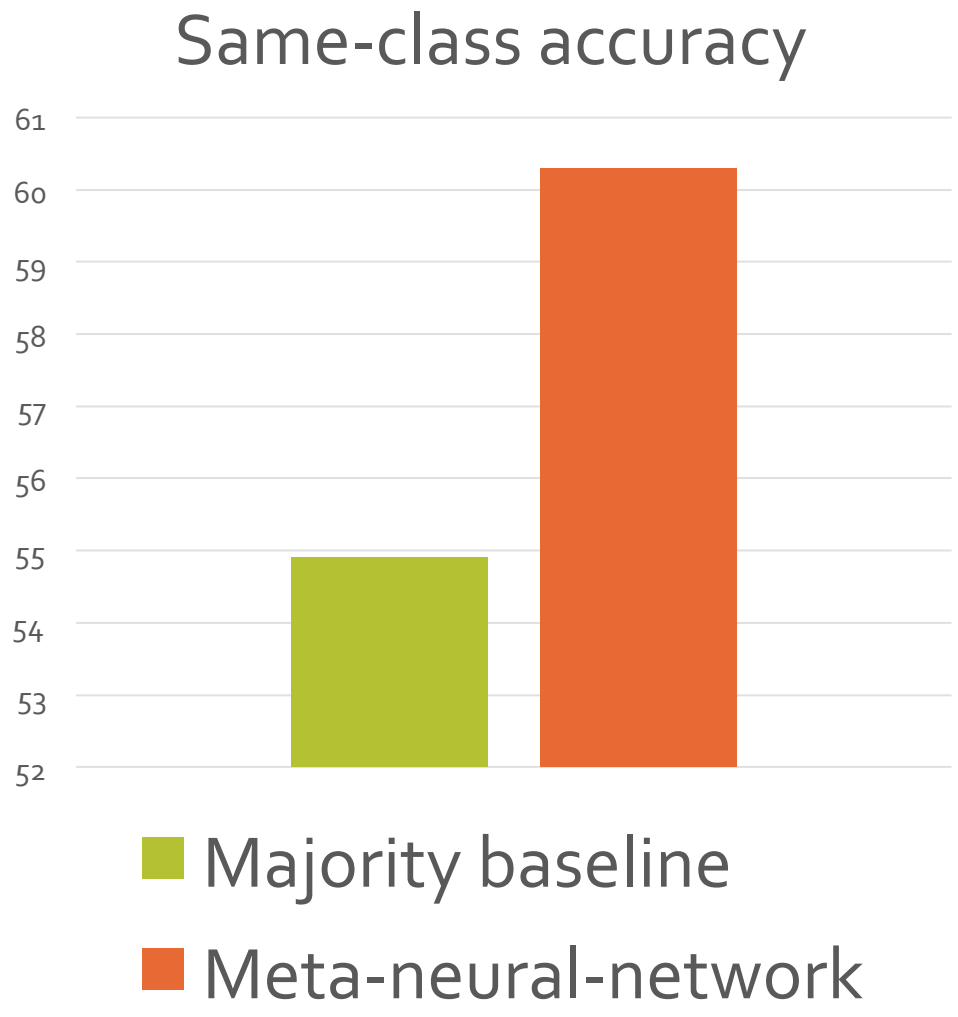


# META-OUTLIER-REMOVAL


0.14
0.13
0.12
0.11
0.1
0.09
0.08
Average Adjusted RI
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training fraction
5%
4%
3%
2%
1%
0%

# Meta-**neural-network**-unsupervised-learning

Lots of little data makes big data



# HUMAN EXPERIMENTS?

# Cluster this

# Cluster this

# Cluster this

***The class was*** *great!*

*Waste of time ADAM I HOPE YOU DIE.*

***The class was*** *boring.*

*Best course I have taken in a long time.*

***The class was*** *terrible!*

*I slept the whole time.*

# Summary

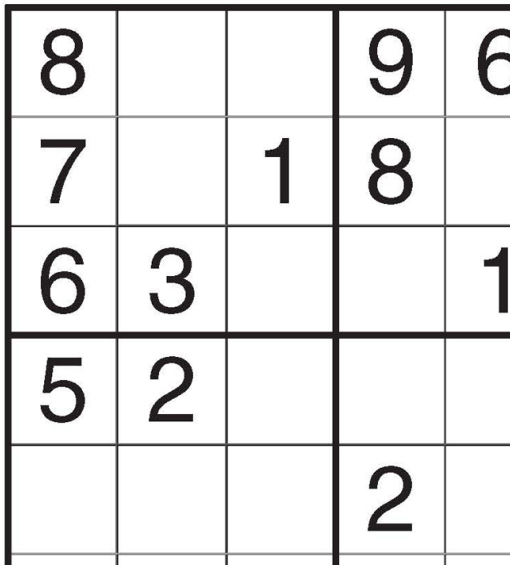
Don't miss the meta-boat

- Meta-clustering
- Meta-topic-modeling
- Meta-feature-selection
- **Meta-classification**
- **Meta-algorithms**


H&M theories that apply to Humans & Machines

# An open meta-problem

Humans solve problems, write programs, come up with and improve strategies




How can computers do the same?

# Meta-thank-you