Descriptive clustering

Christel VRAIN, Thi-Bich-Hanh DAO

LIFO
Université d'Orléans

Workshop on Machine Learning and Explainability

Motivation

- Clustering used extensively in AI applications
- In many domains, data have very good features/attributes to form compact clusters, but
  - features cannot *explain* the clustering well
  - data also described by another set of (potentially *sparse and noisy*) descriptors/tags that are useful for explanation

Setting	Features/attributes	Descriptors/tags
Twitter network	mention/retweet graph	hashtag usage
Images	SIFT features	tags

- Needs to balance compact clusters (w.r.t. to a distance between objects) with
  - their consistency with human expectations
  - their explanations to human
- Aims:
  1. find clusters close to the expert expectations by leveraging knowledge
  2. discover simultaneously explanations during the clustering process

Mainly two frameworks for clustering

- Conceptual Clustering:
  - introduced in the 80's *[Michalski & Stepp, 1983, Fisher, 1985]*
  - presently based on closed patterns (FCA and pattern mining)
  - based on qualitative properties
  - does not take into account quantitative attributes, nor distance between objects (no notion of compactness, e.g. clusters diameter)
- Distance-based clustering:
  - based on dissimilarities between objects
  - appropriate for quantitative data
  - qualitative properties must be encapsulated in a distance

A declarative framework for constrained clustering in CP

Dao, Duong, Vrain, AIJ 2017

- Input: a dataset or a dissimilarity measure between pairs of points
- Clusters are defined by an assignment of points to clusters:
$$G[o] = c, c \in [1, k]$$
- Optimization criterion, e.g. minimizing the maximum diameter
- Constraints are put
  - for representing a partition
  - for breaking symmetries
  - user constraints: size, diameter, split, ...

$G_1 = 1$
$G_i \leq \max_{j \in [1,i-1]}(G_j) + 1$, for $i \in [2, n]$
$\#\{i \mid G_i = k_{min}\} \geq 1$

How to make clustering interpretable?

- Before the clustering process: leverage human knowledge before clustering
  $\rightarrow$ actionable clustering
- After the clustering process $\rightarrow$ explain the cluster :
  - Characterization
  - Generalization
  - Statistics
- During the clustering process. Two assumptions
  - Clustering and explanations are in the same representation space
    $\rightarrow$ conceptual clustering
  - Clustering and explanations are in two different representation spaces.

Actionable clustering

Dao, Vrain, Duong, Davidson, ECAI 2016

Express constraints that makes the clustering useful for a given purpose

Find useful groups each of which you can invite to a different dinner party

- equal number of males and females
- width of a cluster in terms of age at most 10
- each person in a cluster should have at least r other people with the same hobby

Instances $3, 9$ are in the same cluster if $11, 15$ are in different clusters.

$$B_1 \leftrightarrow (G_{11} \neq G_{15})$$
$$B_2 \leftrightarrow (G_3 = G_9)$$
$$B_1 \leq B_2$$

Unifying conceptual and distance clustering

Dao, Lesaint, Vrain, JFPC 2015

- taking into account quantitative and qualitative data
- combining conditions/criteria from both frameworks

- Data:
  - a set $\mathcal{O}$ of objects, a set $\mathcal{I}$ of Boolean properties
  - a dissimilarity measure $d(o, o')$ for any o, o' in $\mathcal{O}$
  - a binary database $\mathcal{D}$: $\mathcal{D}_{op} = 1$, when o satisfies property p
- Clusters are defined by:
  1. assignment of points to clusters: $G[o] = c, c \in [1, k]$
  2. description of clusters: $A[c, p] = 1$ iff p is in the description of cluster c.
- Constraints
  - Constraints of the distance-based model: partition, breaking symmetries
  - Constraints from the conceptual model: an object is in a cluster iff it satisfies all its properties.

$$\forall o \in \mathcal{O}, \forall c \in \mathcal{C} \qquad G[o] = c \Leftrightarrow \sum_{p \in \mathcal{I}} A[c, p](1 - D_{op}) = 0$$

Car Dataset

- 193 objects
- technical properties (22 attributes) :
  - motorization (diesel or not)
  - drive wheels (4, 2 front, 2 rear)
  - power (between 48 and 288)
  - etc.

  discretization : 64 qualitative attributes
- price (quantitative attribute)

Car dataset

- Conceptual setting

→ (e) concepts + maximizing min. size of clusters

→ (f) concepts + maximizing min. size of concepts

Price distribution not convincing

- Distance-based setting

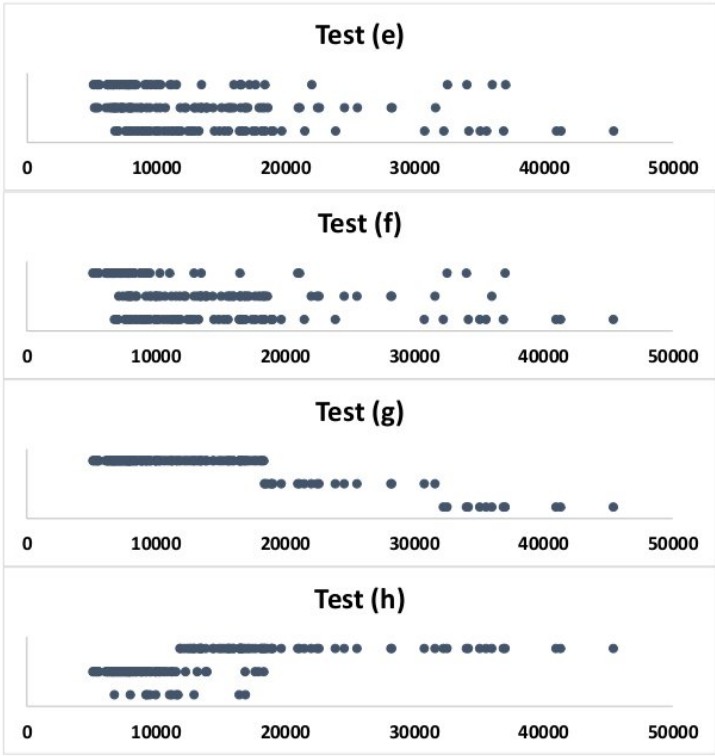
→ (g) minimizing max diameter

No convincing concepts

- Unified framework

→ (h) concepts + minimizing max diameter

A better modeling of the 3 car ranges with concepts based on size, engine power, fuel consumption, . . .



Descriptive clustering formulation

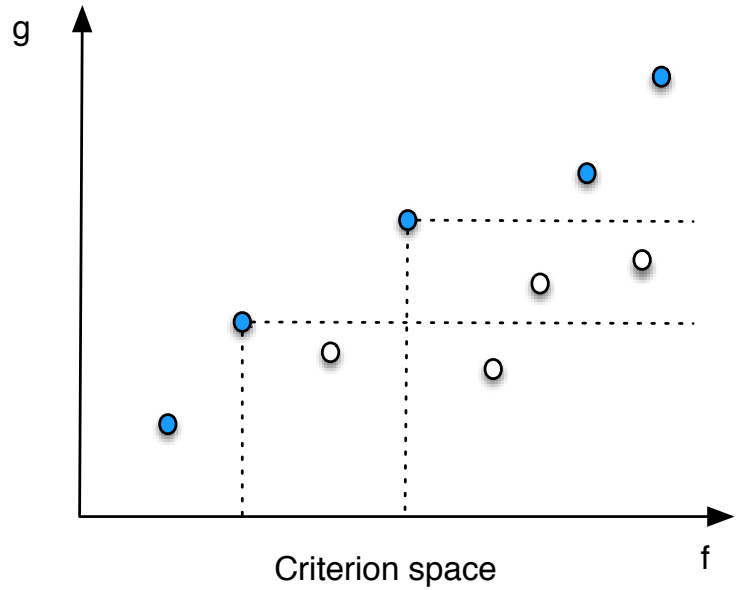
Dao, Kuo, Ravi, Vrain, Davidson, IJCAI 2018

- Data: n data instances described by *numerical features* X and interpretable *boolean descriptors/tags* D
- Aims: Simultaneously look for clusters which are both
  - good/compact in one modality (e.g. SIFT features for images or graph distance)
  - useful/descriptive in another modality (e.g. tags)
- The objectives are not compatible

$\rightarrow$ computation of a Pareto front corresponding to Pareto optimal solutions, allowing to model a trade-off with both objectives

- f: feature-focused objective to minimize compactness
- g: descriptor-focused objective to maximize interpretability

Pareto optimal solutions and Pareto front



- Partition P' dominates P iff better in one criterion and not worse in the other
- P is a Pareto optimal solution iff there is no P' which dominates P
- Pareto front = $\{(f(P), g(P)) \mid \text{P is a Pareto optimal solution}\}$

Compute the complete Pareto front

$\mathcal{P} \leftarrow \emptyset$;
$s_1^f \leftarrow$ minimize f subject to $\mathcal{C}$;
$i \leftarrow 1$;
while $s_i^f \neq NULL$ **do**
    $s_i^g \leftarrow$ maximize g subject to $\mathcal{C} \cup \{f \leq f(s_i^f)\}$;
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{s_i^g\}$;
    $i \leftarrow i + 1$;
    $s_i^f \leftarrow$ minimize f subject to $\mathcal{C} \cup \{g > g(s_{i-1}^g)\}$;
return $\mathcal{P}$;

Data and variables

- Data:
  - X: $n \times f$ matrix of n data instances with f numerical features
  - D: $n \times r$ matrix of the same n instances with r tag indicators
- Variables:
  - cluster indication matrix Z: $n \times k$ boolean matrix
    $Z_{ic} = 1$ indicates the i-th instance is in the c-cluster
  - cluster description matrix S: $k \times r$ boolean matrix
    $S_{cp} = 1$ means the p-th tag is included in the description of the c-th cluster

Partitioning constraints

- Each instance is in one cluster
- Each cluster has at least one element
- Breaking symmetries between clusters

$$\begin{array}{rl} \forall i = 1, \ldots, n, & \sum_{c=1}^{k} Z_{ic} = 1 \\ \forall c = 1, \ldots, k, & \sum_{i=1}^{n} Z_{ic} \geq 1 \end{array}$$

$$\begin{array}{rl} & Z_{11} = 1 \\ \forall i = 2, \ldots, n, \forall c = 2, \ldots, k, & \sum_{j=1}^{i-1} Z_{jc-1} \geq Z_{ic} \end{array}$$

- Each cluster description has at least one tag

$$\forall c = 1, \ldots, k, \quad \sum_{i=1}^{n} S_{cp} \geq 1$$

Cluster description constraints

- Each cluster is described by a non empty subset of tags
- An instance in a cluster must satisfy *most* of its descriptions (up to α exceptions):

$$\forall c = 1, \ldots, k, \ \forall i = 1, \ldots, n,$$

$$Z_{ic} = 1 \implies \sum_{p=1}^{r} S_{cp}(1 - D_{ip}) \leq \alpha$$

- A tag is included in a cluster description if and only if *most* of the instances in the cluster (up to β exceptions) possess it:

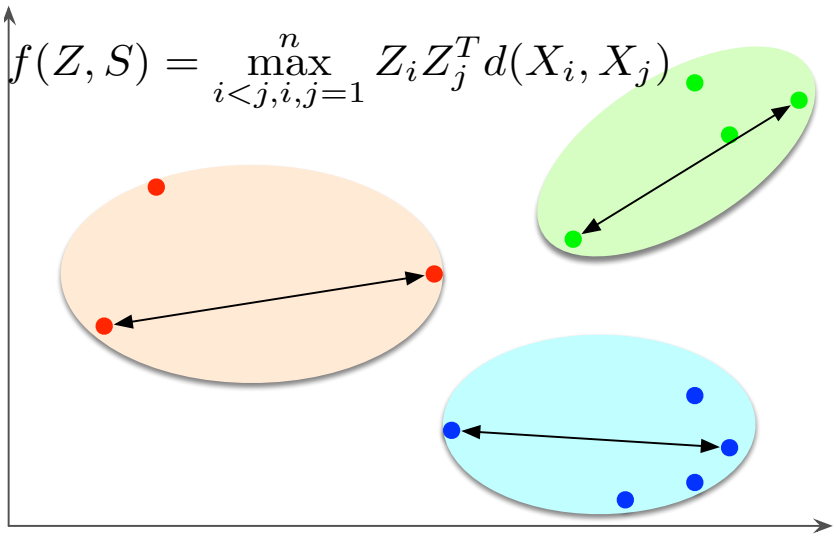
$$\forall c = 1, \ldots, k, \ \forall p = 1, \ldots, r,$$

$$S_{cp} = 1 \iff \sum_{i=1}^{n} Z_{ic}(1 - D_{ip}) \leq \beta$$

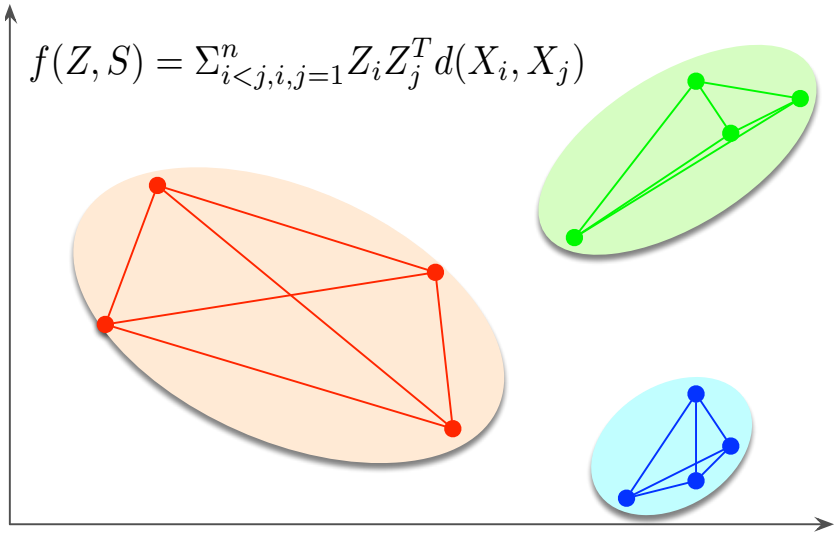
- With dense tags dataset, stronger version with $\alpha = \beta = 0$

Feature-focused optimization criteria

Finding compact clusters, based on their distance $d(\cdot,\cdot)$ defined over pairs of instances



Diameter $\arg\min f(Z,S)$



Sum of within-cluster distances

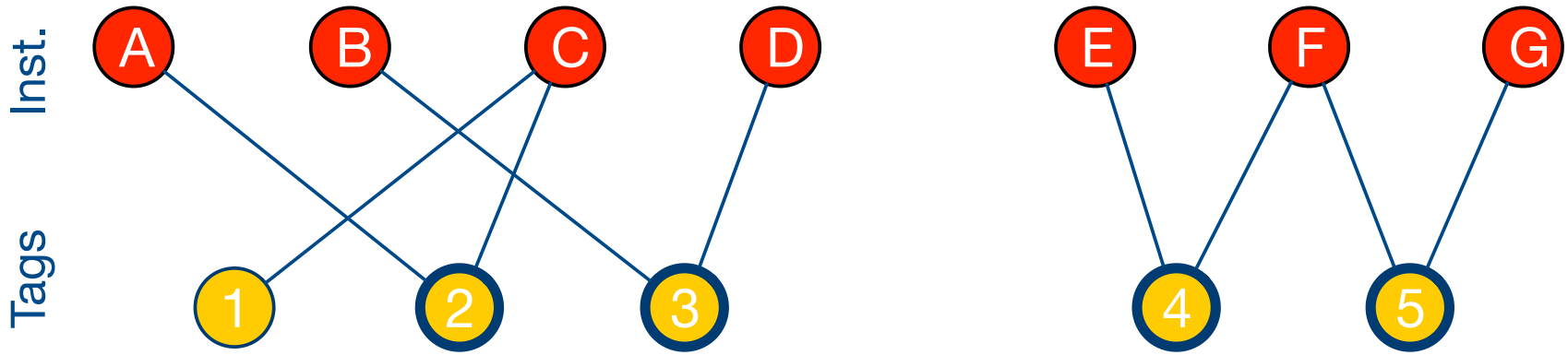
Descriptor-focused optimization criteria

Tags may be dense, sparse or noisy $\Rightarrow$ different objective functions

1. Minimize tag disagreement (MTD)
   - minimize $\alpha + \beta$
   - useful when the tags contain noise
2. Max-min complete tag agreement (MMCTA):
   - the tags of a cluster are shared by *all* its instances ($\alpha = \beta = 0$)
   - maximize the tag set of each cluster (size of the smallest)
   - Use: tags well populated with little noise
3. Max-min neighborhood agreement (MMNA):
   - each pair of instances in a same cluster must share *at least q* tags
   - maximize *q*
   - Use: tags are sparse

Minimize tag disagreement (MTD)

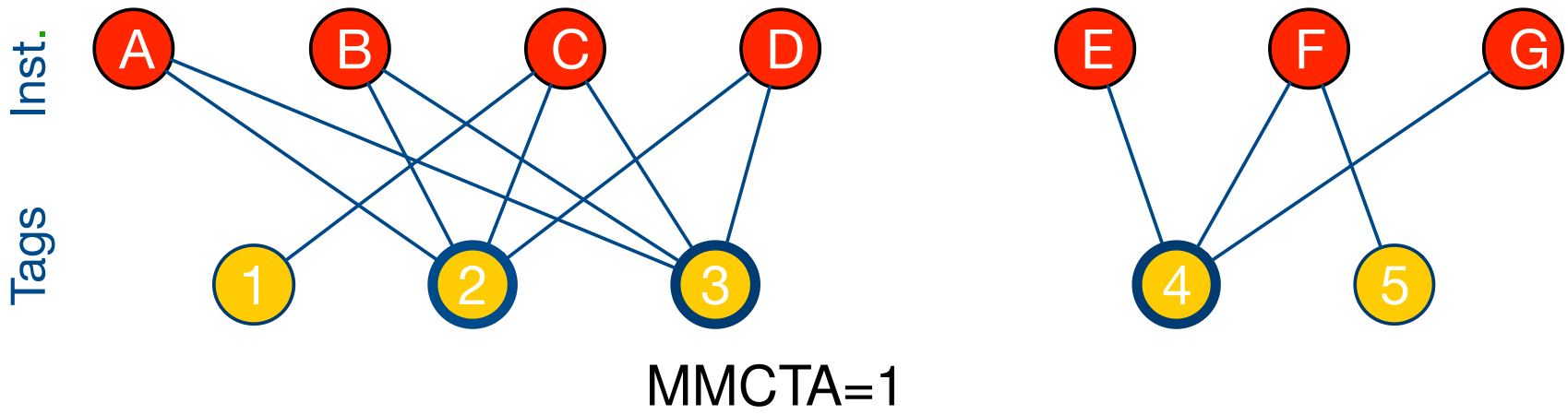
- allows disagreements
- α number of tags an instance may not posses
- β number of instances a tag may not cover
- useful when tags are sparse and/or noisy



$$\alpha = 1, \beta = 2$$

Max-min complete tag agreement (MMCTA)

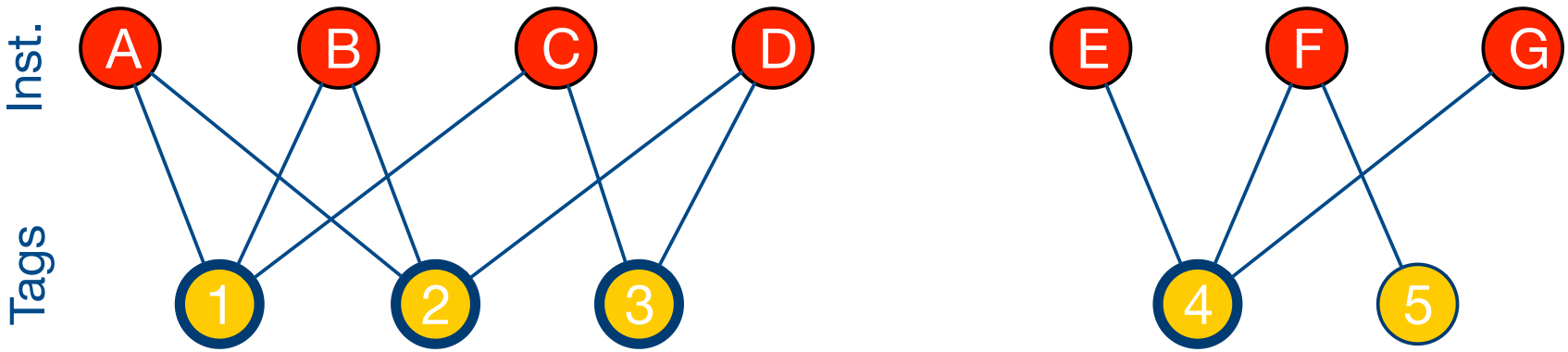
- the tags of a cluster are shared by *all* its instances ($\alpha = \beta = 0$)
- useful when the tags are well populated with little noise



MMCTA=1

Max-min neighborhood agreement (MMNA)

- Every pair of instances in a cluster must share at least q tags
- useful when tags are sparse



MMNA=1

Two methods

- Integer linear programming (ILP): all the constraints and objectives can be transformed into linear form
- Constraint programming (CP): using global constraints, reified constraints and restart search

CP formulation

- Supplementary variables $G_i \in \{1, \ldots, k\}$ for $i = 1, .., n$
  $G_i = c$ means i-th instance is in c-cluster
- Channeling constraints:

$$Z_{ic} = 1 \Longleftrightarrow G_i = c$$

- Global constraints:

$$precede(G, [1, .., k])$$
$$atleast(G, 1, k)$$
$$diameter(G, f, d)$$

- Reified constraints to express description constraints
- For MMCTA, MMNA, new global constraints to enforce:

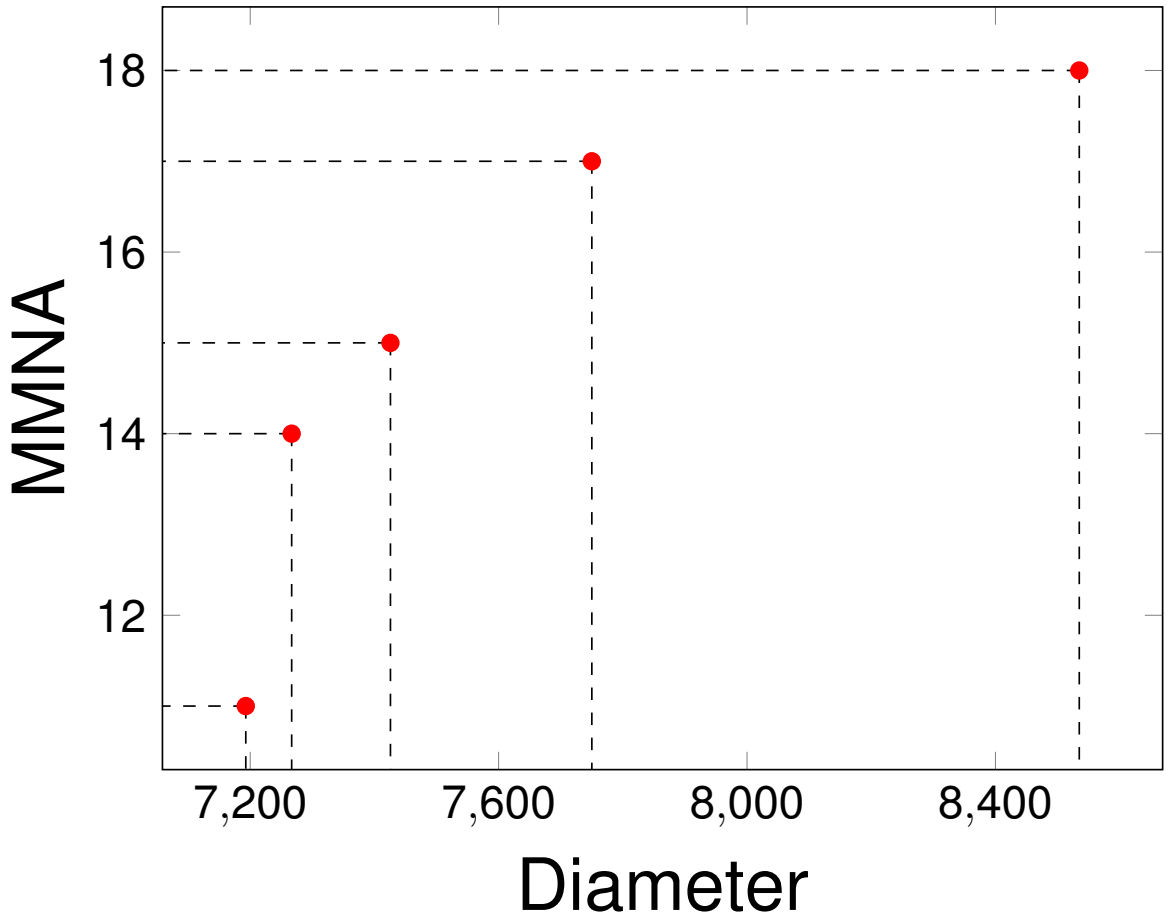
$$\forall i, j = 1, .., n,\ G_i = G_j \Longrightarrow \sum_{p=1}^{r} D_{ip} D_{jp} \geq q$$

Clustering tagged images

[Lampert C.H, Nickisch H., Harmeling S., CVPR 2009]

- 30000 images from 50 classes of animals
- Each image described by 2000 SIFT features and variable number of tags (black, fast, timid, etc.)
- Animal names are not given to algorithm
- Randomly sample 100 images from 10 first animal classes

Trade off compactness vs. useful description



First Pareto point: Diameter minimized. MMCTA=4. MMNA=11

Cl#	Composition by animals	Description by tags
C1	1 grizzly bear, 2 dalmatian, 1 horse, 2 blue whale	big, fast, strong, muscle, newworld, smart
C2	5 antelope, 2 grizzly bear, 2 beaver, 5 dalmatian, 5 persian cat, 5 horse, 6 german shepherd, 3 siamese cat	furry, chewteeth, fast, quadrapedal, newworld, ground
C3	69 beaver, 64 dalmatian, 42 persian cat, 29 blue whale, 42 siamese cat	tail, fast, newworld, timid, smart, solitary
C4	100 killer whale, 69 blue whale, 1 siamese cat	tail, fast, fish, smart
C5	95 antelope, 97 grizzly bear, 29 beaver, 29 dalmatian, 53 persian cat, 94 horse, 94 german shepherd, 54 siamese cat	furry, chewteeth, fast, quadrapedal, newworld, ground

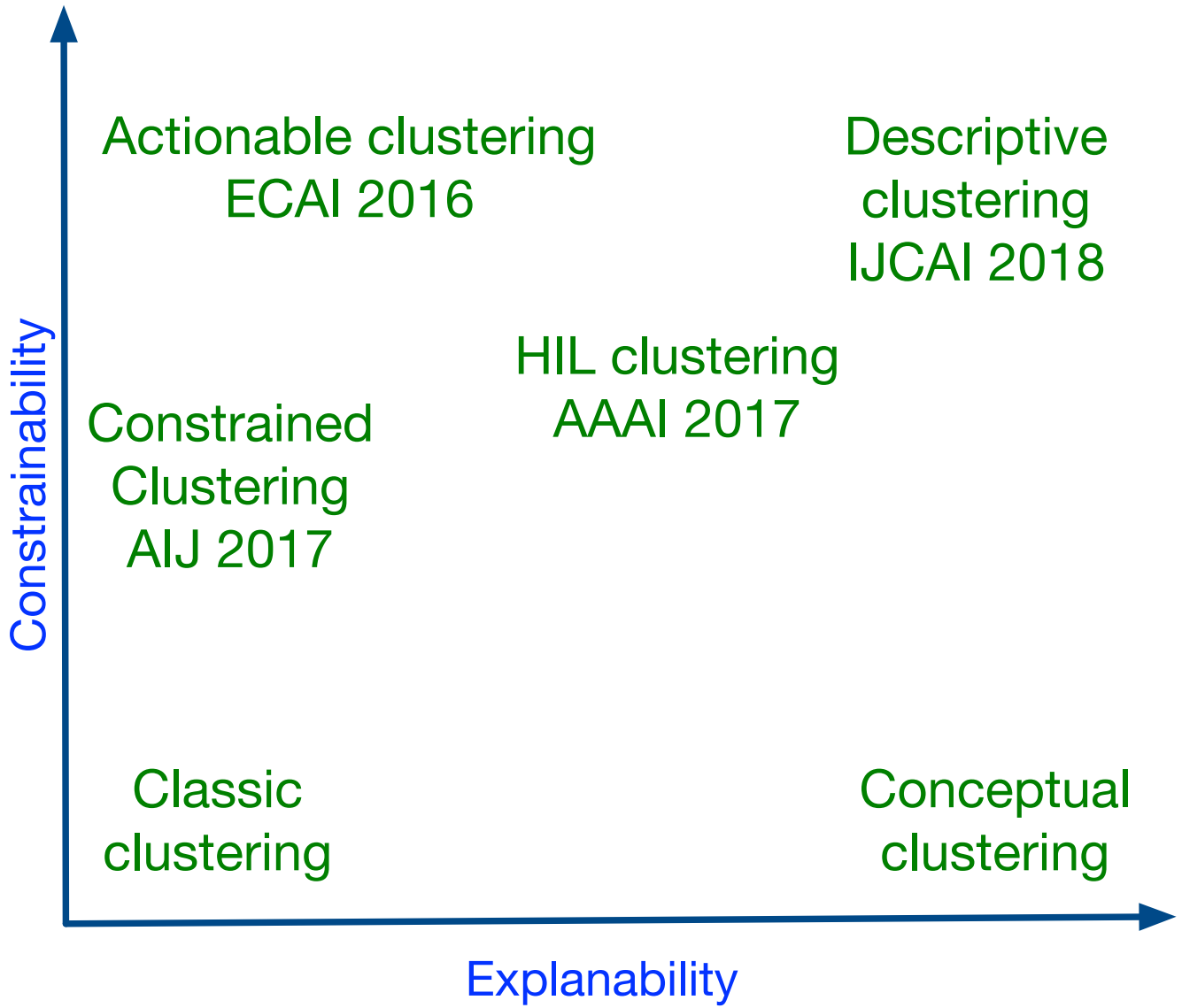
Third Pareto point. MMCTA=9, MMNA=15

Cl#	Composition by animals	Description by tags
C1	2 antelope, 4 dalmatian, 2 horse, 3 german shepherd, 4 siamese cat	furry, lean, longleg, tail, chewteeth, walks, fast, muscle, quadrapedal, active, agility, newworld, oldworld, ground
C2	2 beaver, 1 persian cat, 1 horse, 1 german shepherd	furry, tail, chewteeth, fast, quadrapedal, agility, newworld, ground, smart
C3	100 grizzly bear, 98 beaver, 99 persian cat, 1 siamese cat	furry, paws, chewteeth, claws, fast, quadrapedal, fish, newworld, ground, smart, solitary
C4	100 killer whale, 100 blue whale	spots, hairless, toughskin, big, bulbous, flippers, tail, strainteeth, swims, fast, strong, fish, plankton, arctic, ocean, water, smart, group
C5	98 antelope, 96 dalmatian, 97 horse, 96 german shepherd, 95 siamese cat	furry, lean, longleg, tail, chewteeth, walks, fast, muscle, quadrapedal, active, agility, newworld, oldworld, ground

Fifth Pareto point: MMNA maximized. MMCTA=15, MMNA=18

Cl#	Composition by animals	Description by tags
C1	100 antelope, 100 dalmatian	furry, big, lean, longleg, tail, chewteeth, walks, fast, strong, muscle, quadrapedal, active, agility, newworld, oldworld, ground, timid, group
C2	100 horse, 99 german shepherd, 98 siamese cat	black, brown, gray, patches, furry, lean, longleg, tail, chewteeth, walks, fast, muscle, quadrapedal, active, agility, newworld, oldworld, ground, smart, domestic
C3	100 grizzly bear, 100 beaver, 1 siamese cat	brown, furry, paws, chewteeth, claws, fast, muscle, quadrapedal, active, nocturnal, fish, newworld, ground, smart, solitary
C4	100 killer whale, 100 blue whale	spots, hairless, toughskin, big, bulbous, flippers, tail, strainteeth, swims, fast, strong, fish, plankton, arctic, ocean, water, smart, group
C5	100 persian cat, 1 german shepherd, 1 siamese cat	gray, furry, pads, paws, tail, chewteeth, meatteeth, claws, walks, fast, quadrapedal, agility, meat, newworld, oldworld, ground, smart, solitary, domestic

Towards actionable/explainable clustering



Future work

- Framework is extendable to other types of compactness and description
- Scalability
  - Smart data
  - Sampling ??
  - Relaxing the search for an optimal solution
- Learning constraints and preferences
- Leveraging knowledge by means of constraints in other frameworks