GUIDING USER GROUPINGS

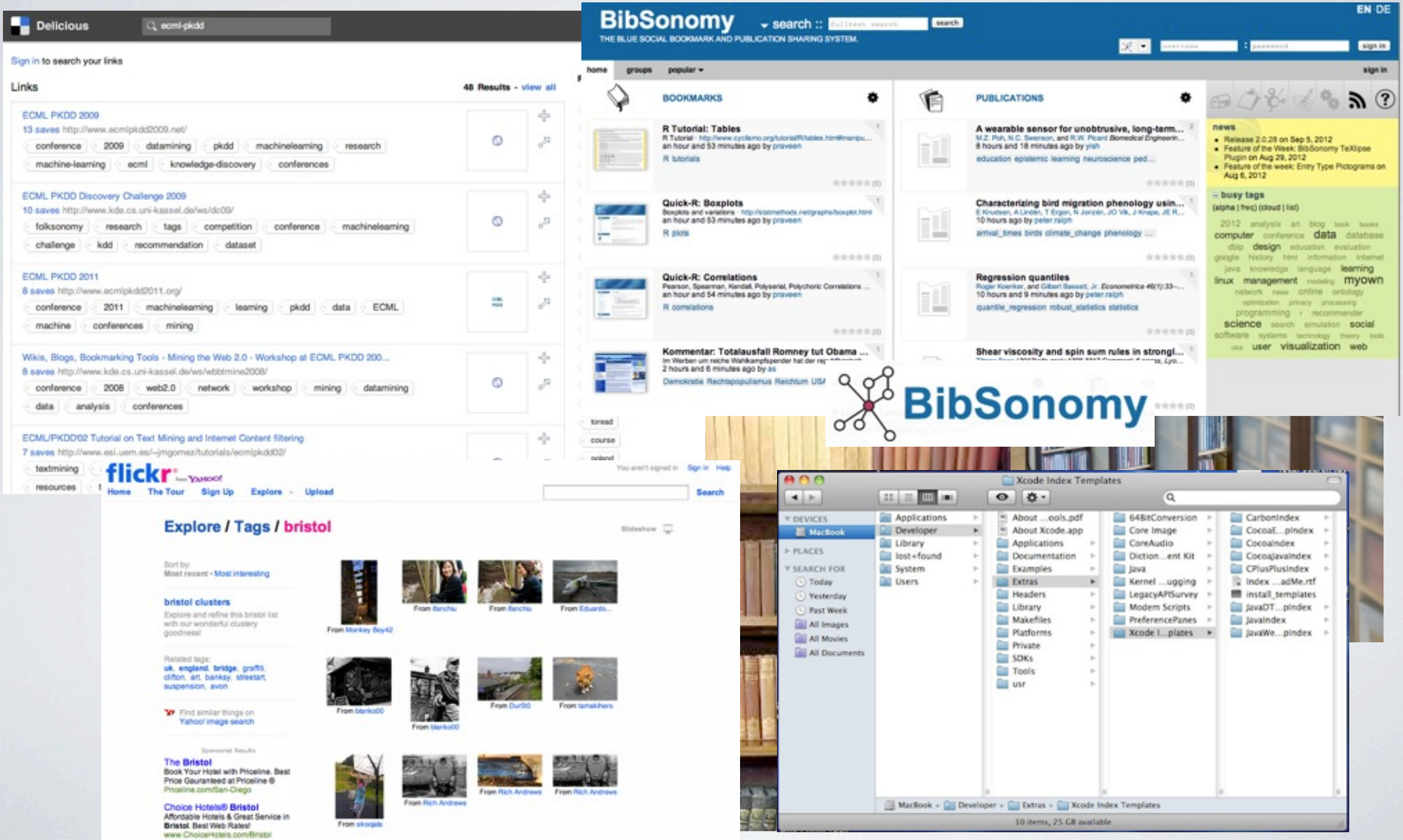
LEARNING AND COMBINING CLASSIFICATION FOR ITEMSET STRUCTURING

Mathias Verbeke, Ilija Subašić and Bettina Berendt

MUSE
September 24, 2012, Bristol

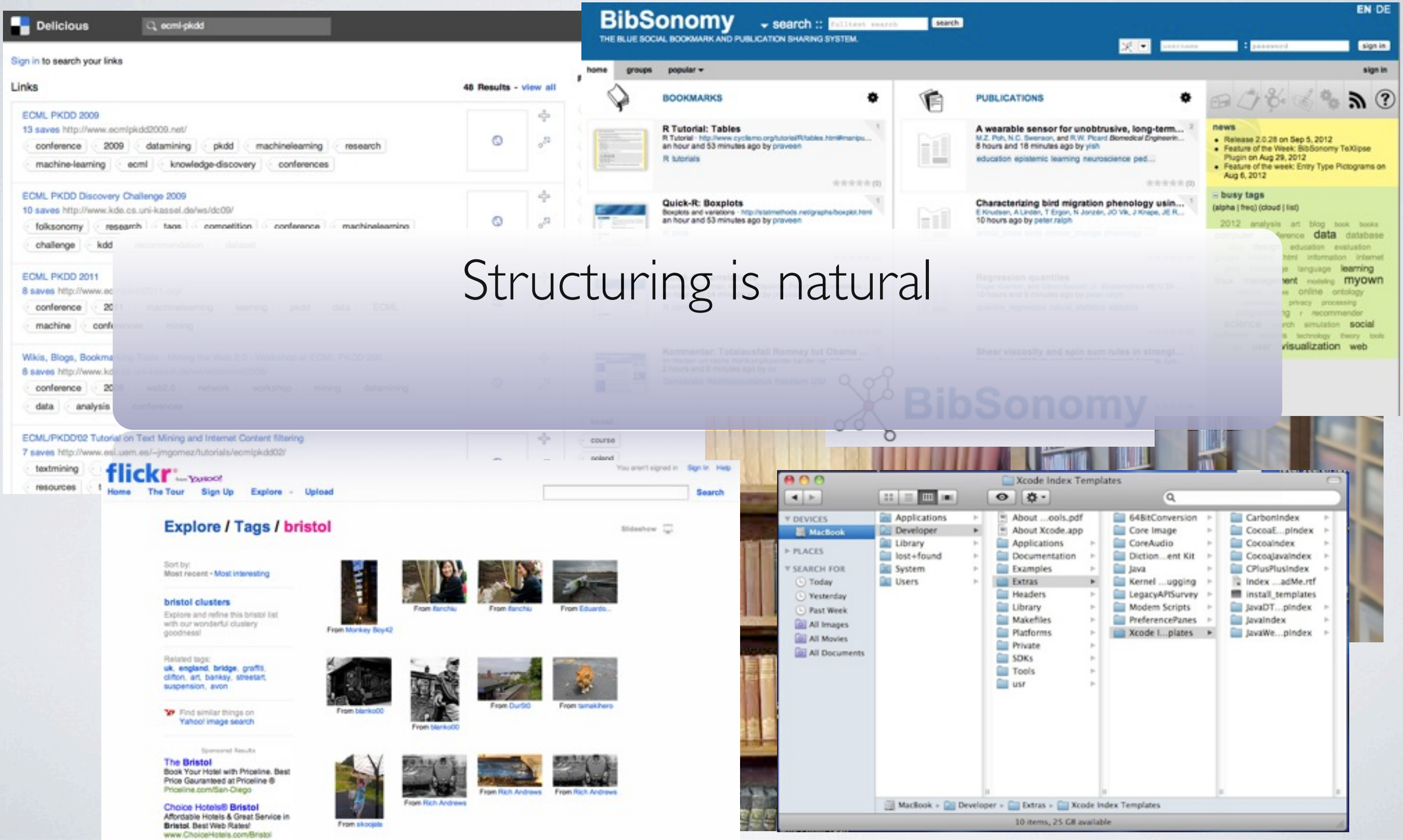
MOTIVATION

MOTIVATION


Delicious
ecml-pkdd
Sign in to search your links
Links
48 Results - view all
ECML PKDD 2009
13 saves http://www.ecmlpkdd2009.net/
conference 2009 datamining pkdd machinelearning research
machine-learning ecml knowledge-discovery conferences
ECML PKDD Discovery Challenge 2009
10 saves http://www.kde.cs.uni-kassel.de/ws/dc09/
folksonomy research tags competition conference machinelearning
challenge kdd recommendation dataset
ECML PKDD 2011
8 saves http://www.ecmlpkdd2011.org/
conference 2011 machinelearning learning pkdd data ECML
machine conferences mining
Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 - Workshop at ECML PKDD 200...
8 saves http://www.kde.cs.uni-kassel.de/ws/wbbtmine2008/
conference 2008 web2.0 network workshop mining datamining
data analysis conferences
ECML/PKDD'02 Tutorial on Text Mining and Internet Content filtering
7 saves http://www.esi.uem.es/~jmgomez/tutorials/ecmlpkdd02/
textmining
resources
BibSonomy
search ::
search
THE BLUE SOCIAL BOOKMARK AND PUBLICATION SHARING SYSTEM.
EN DE
sign in
home
groups
popular
sign in
BOOKMARKS
R Tutorial: Tables
an hour and 53 minutes ago by praveen
R tutorials
Quick-R: Boxplots
an hour and 53 minutes ago by praveen
R plots
Quick-R: Correlations
an hour and 54 minutes ago by praveen
R correlations
Kommentar: Totalausfall Romney tut Obama ...
2 hours and 6 minutes ago by as
Demokratie Rechtspopulismus Reichtum USA
toread
course
poland
PUBLICATIONS
A wearable sensor for unobtrusive, long-term...
8 hours and 18 minutes ago by ylsh
education epistemic learning neuroscience ped...
Characterizing bird migration phenology usin...
10 hours ago by peter.ralph
arrival_times birds climate_change phenology ...
Regression quantiles
10 hours and 9 minutes ago by peter.ralph
quantile_regression robust_statistics statistics
Shear viscosity and spin sum rules in strongl...
news
Release 2.0.28 on Sep 5, 2012
Feature of the Week: BibSonomy TeXlipse Plugin on Aug 29, 2012
Feature of the week: Entry Type Pictograms on Aug 6, 2012
busy tags
(alpha | freq) (cloud | list)
2012 analysis art blog book books computer conference data database dblp design education evaluation google history html information internet java knowledge language learning linux management modeling myown network news online ontology optimization privacy processing programming r recommender science search simulation social software systems technology theory tools usa user visualization web
BibSonomy
flickr from YAHOO!
You aren't signed in Sign In Help
Home The Tour Sign Up Explore Upload
Search
Explore / Tags / bristol
Slideshow
Sort by:
Most recent · Most interesting
bristol clusters
Explore and refine this bristol list with our wonderful clustery goodness!
Related tags:
uk, england, bridge, graffiti, clifton, art, banksy, streetart, suspension, avon
Find similar things on Yahoo! image search
Sponsored Results
The Bristol
Book Your Hotel with Priceline. Best Price Gauranteed at Priceline ®
Priceline.com/San-Diego
Choice Hotels® Bristol
Affordable Hotels & Great Service in Bristol. Best Web Rates!
www.ChoiceHotels.com/Bristol
From Monkey Boy42
From ilanchiu
From ilanchiu
From Eduardo...
From blanko00
From blanko00
From Dur560
From tamakihero
From skoojala
From Rich Andrews
From Rich Andrews
Xcode Index Templates
DEVICES
MacBook
PLACES
SEARCH FOR
Today
Yesterday
Past Week
All Images
All Movies
All Documents
Applications
Developer
Library
lost+found
System
Users
About ...ools.pdf
About Xcode.app
Applications
Documentation
Examples
Extras
Headers
Library
Makefiles
Platforms
Private
SDKs
Tools
usr
64BitConversion
Core Image
CoreAudio
Diction...ent Kit
Java
Kernel ...ugging
LegacyAPISurvey
Modem Scripts
PreferencePanes
Xcode I...plates
CarbonIndex
CocoaE...pIndex
CocoaIndex
CocoaJavaIndex
CPlusPlusIndex
Index ...adMe.rtf
install_templates
JavaDT...pIndex
JavaIndex
JavaWe...pIndex
MacBook » Developer » Extras » Xcode Index Templates
10 items, 25 GB available

MOTIVATION

Structuring is natural



MOTIVATION

Structuring is natural

... also for digital items

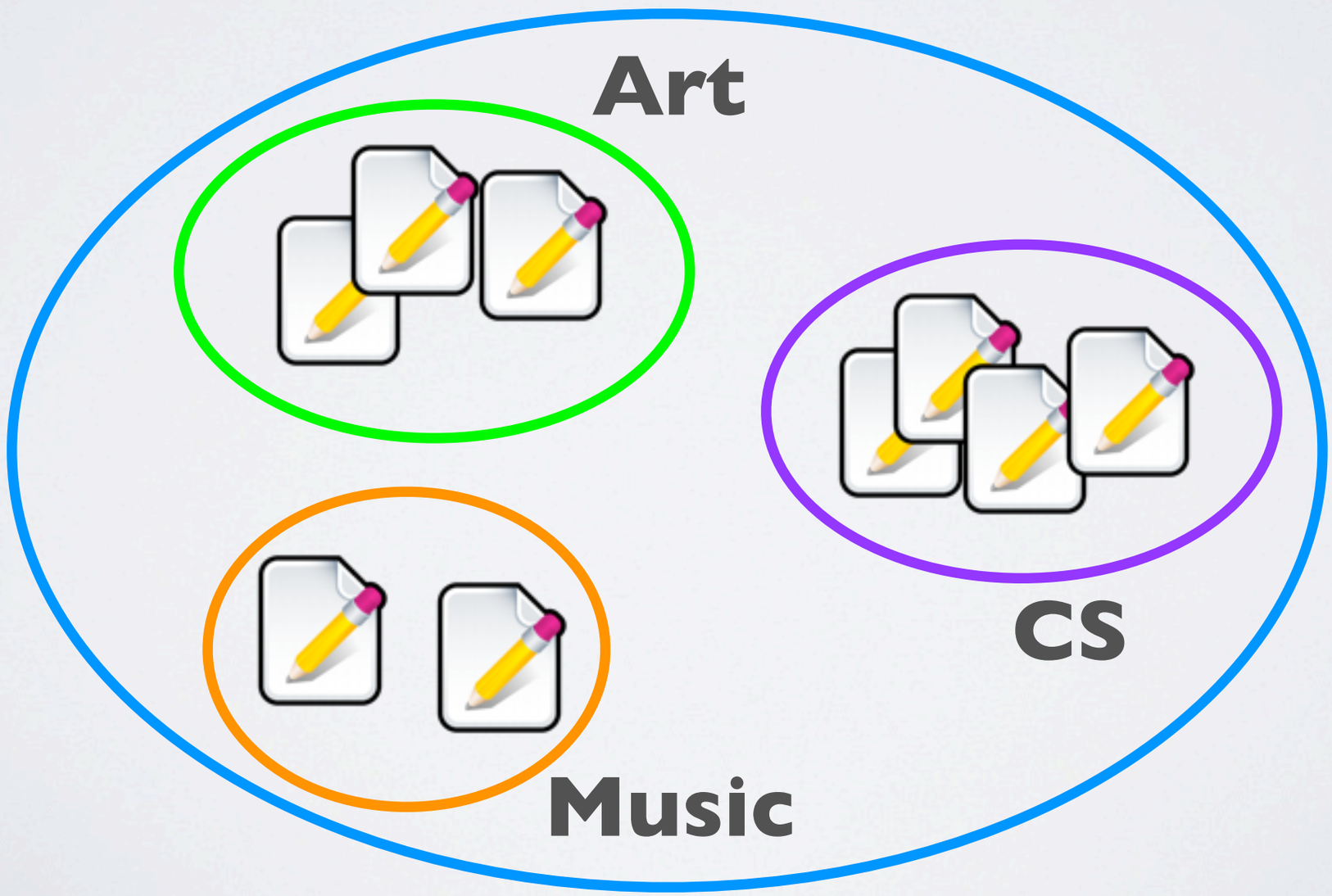
CONTRIBUTIONS

1. A new DM approach that learns an intensional model of user groupings and uses this to group new items
   → Identify structuring dynamics

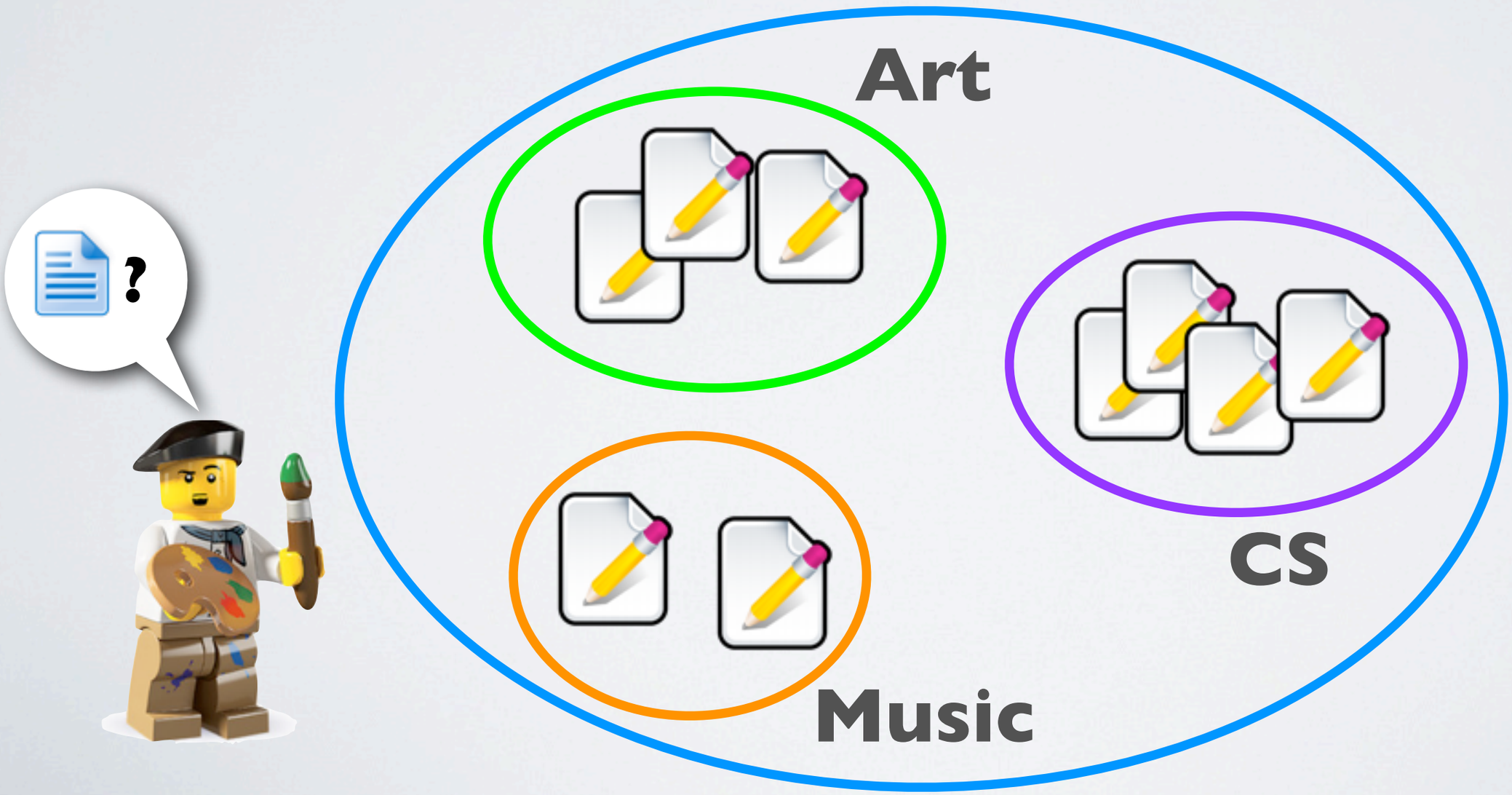
2. New divergence measure

3. A study of grouping behaviour in a social bookmarking system

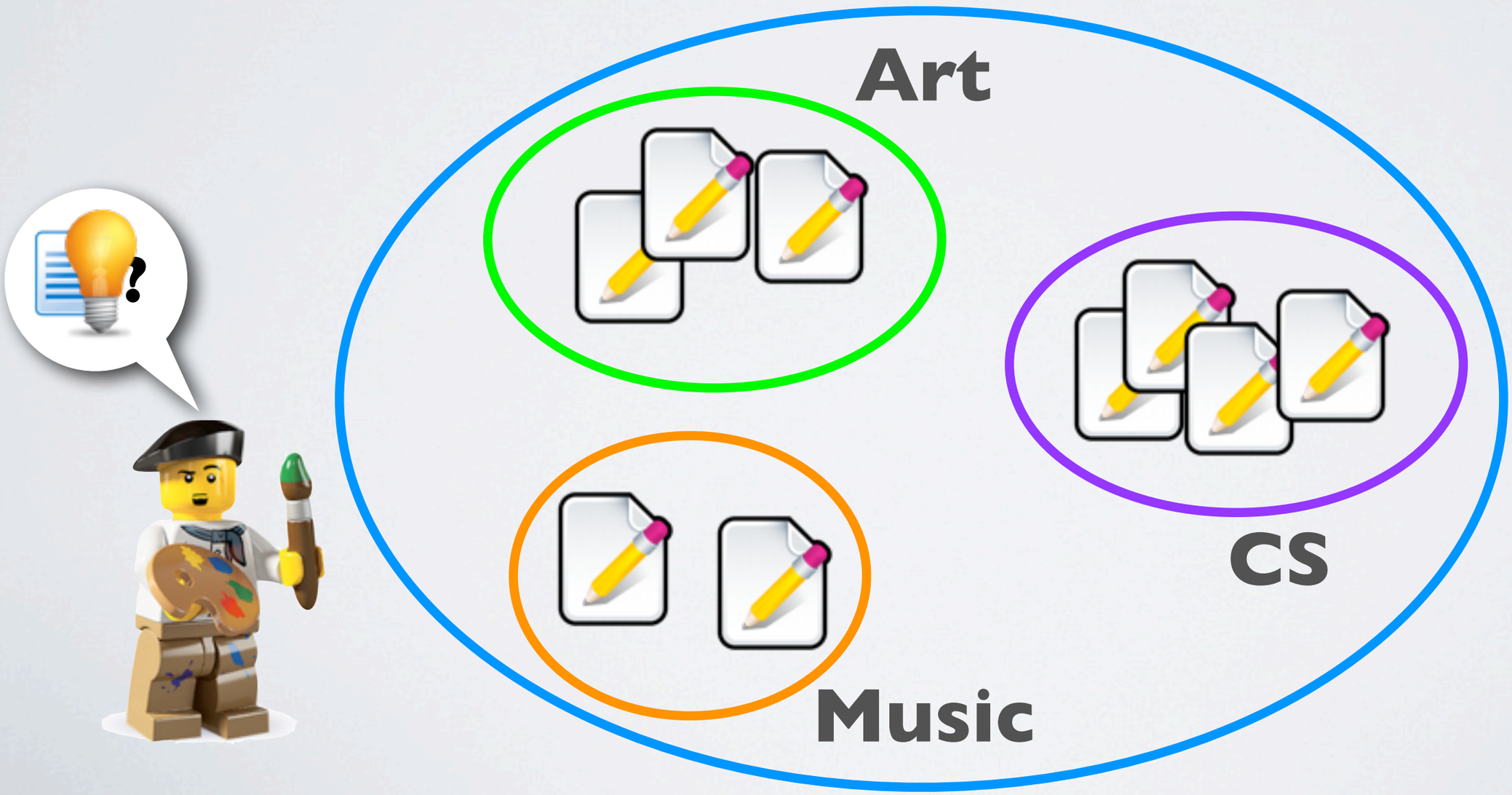
STRUCTURING



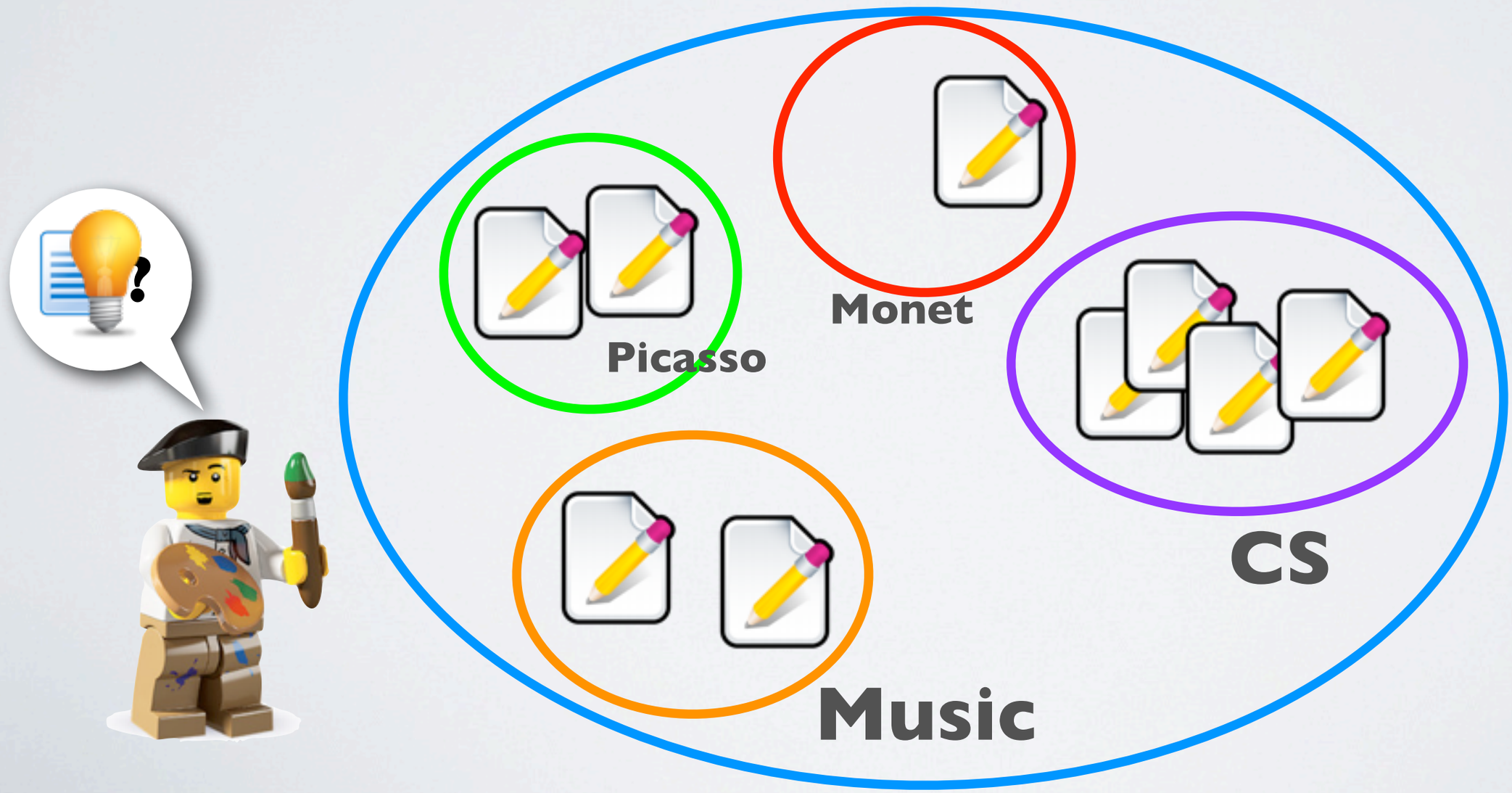
STRUCTURING



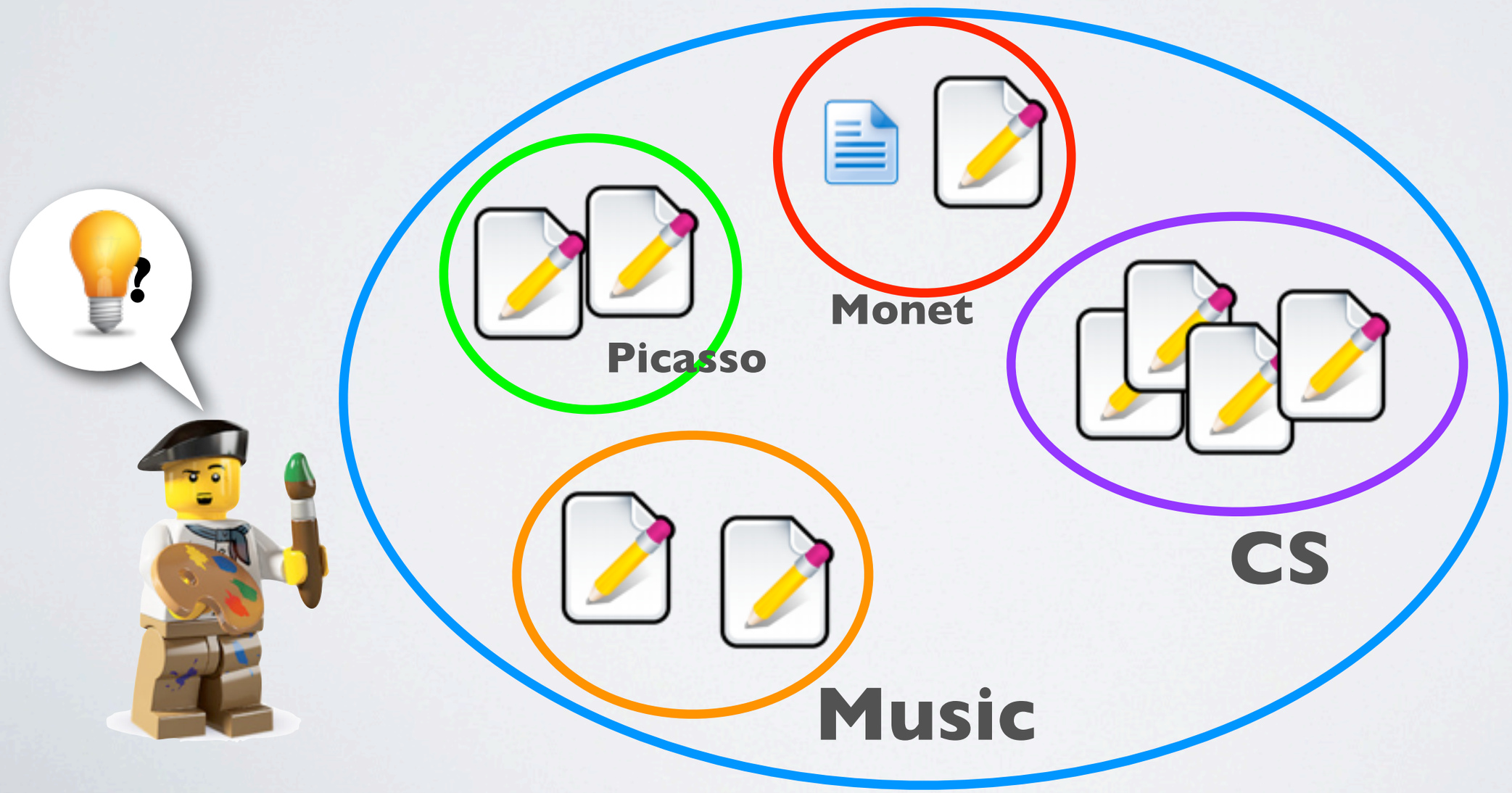
STRUCTURING



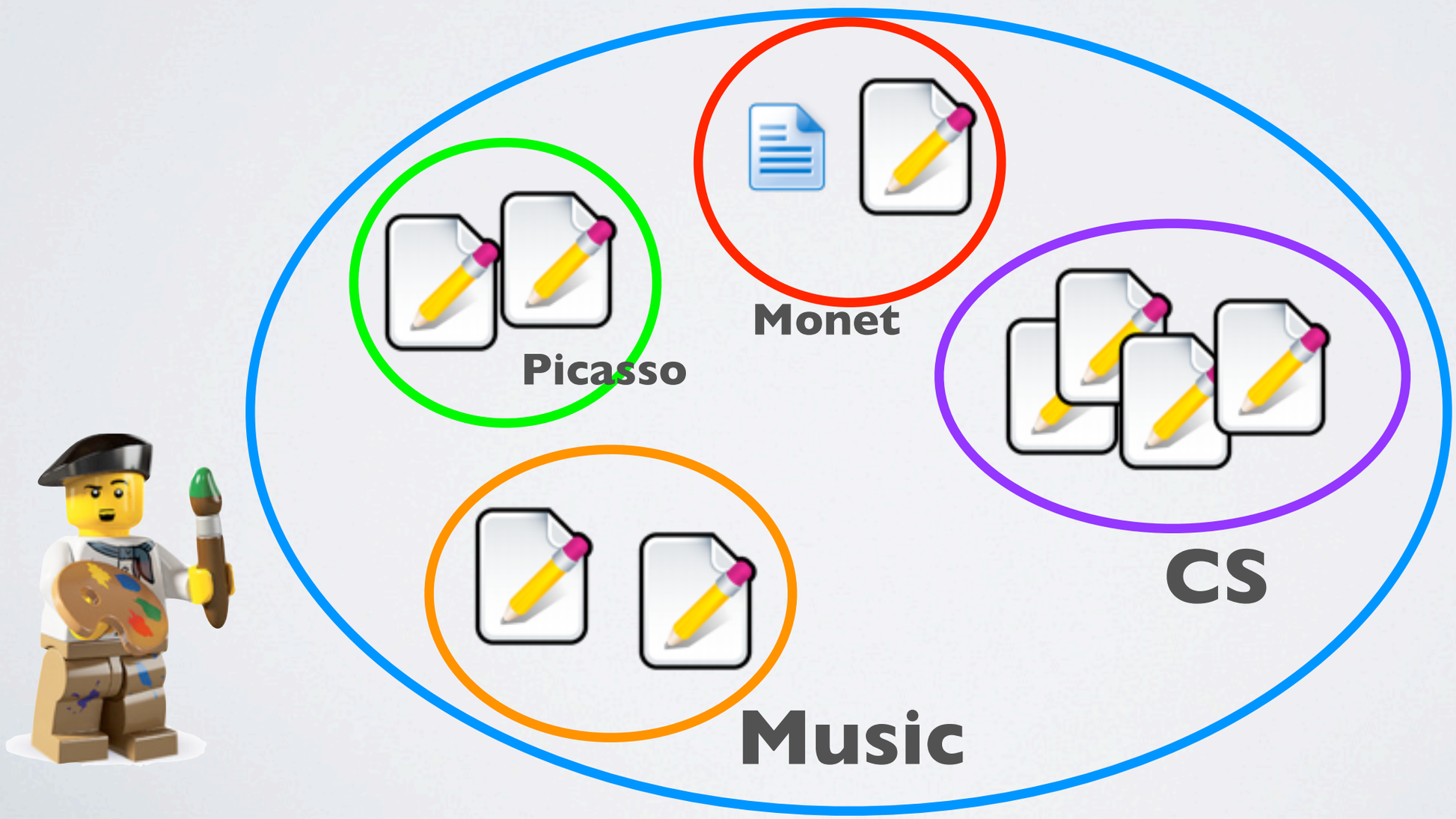
STRUCTURING



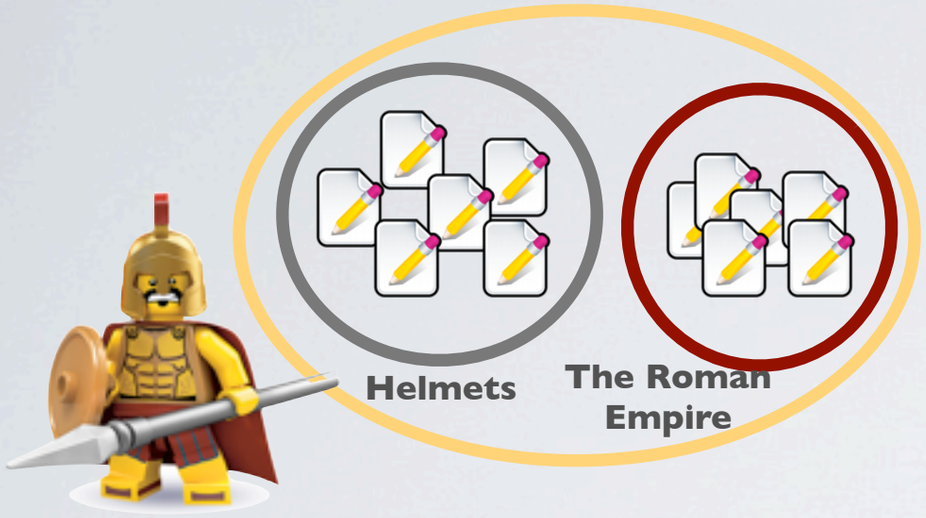
STRUCTURING

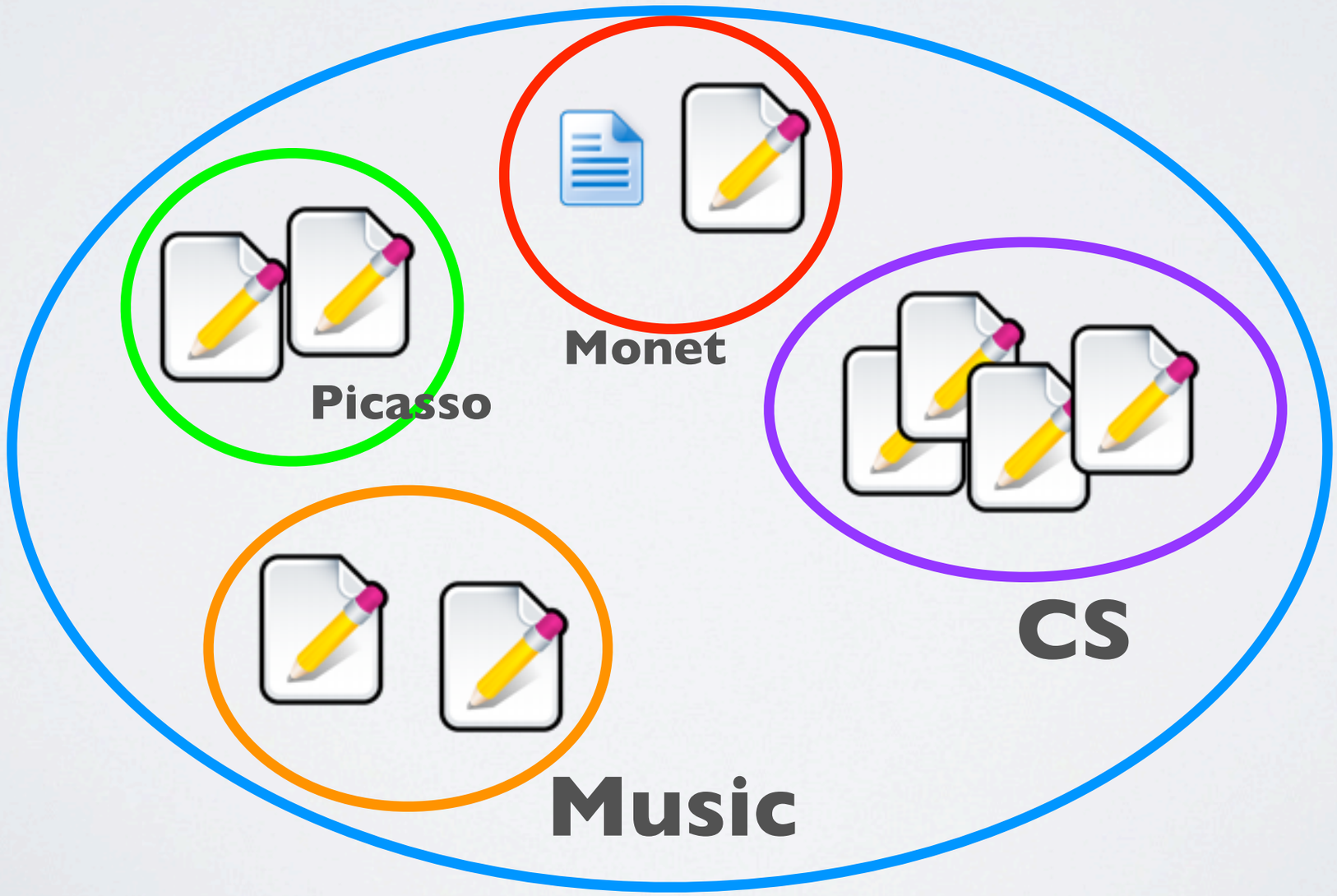


STRUCTURING

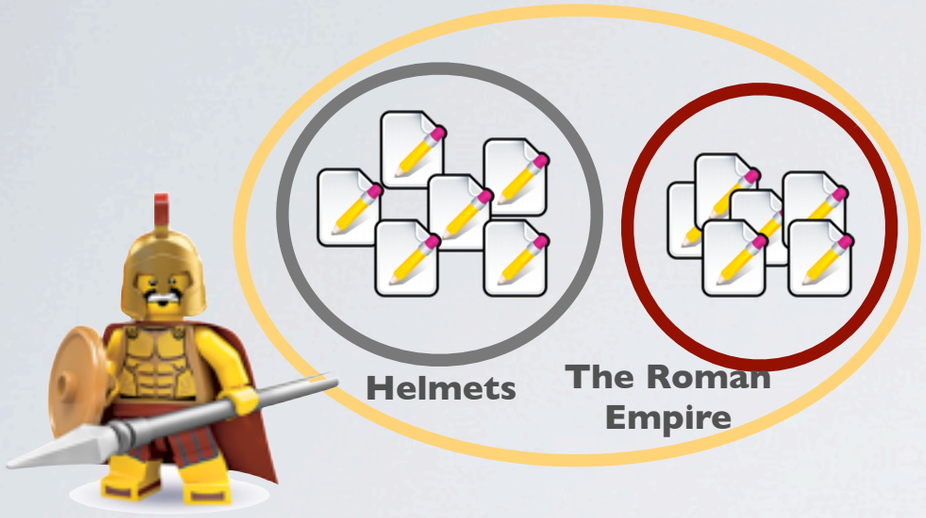


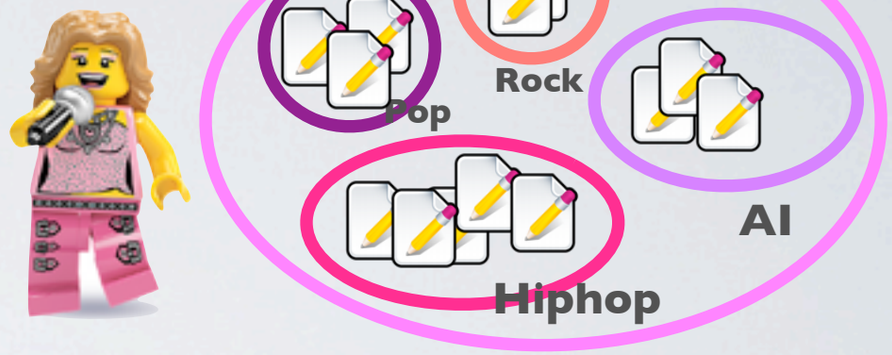
STRUCTURING

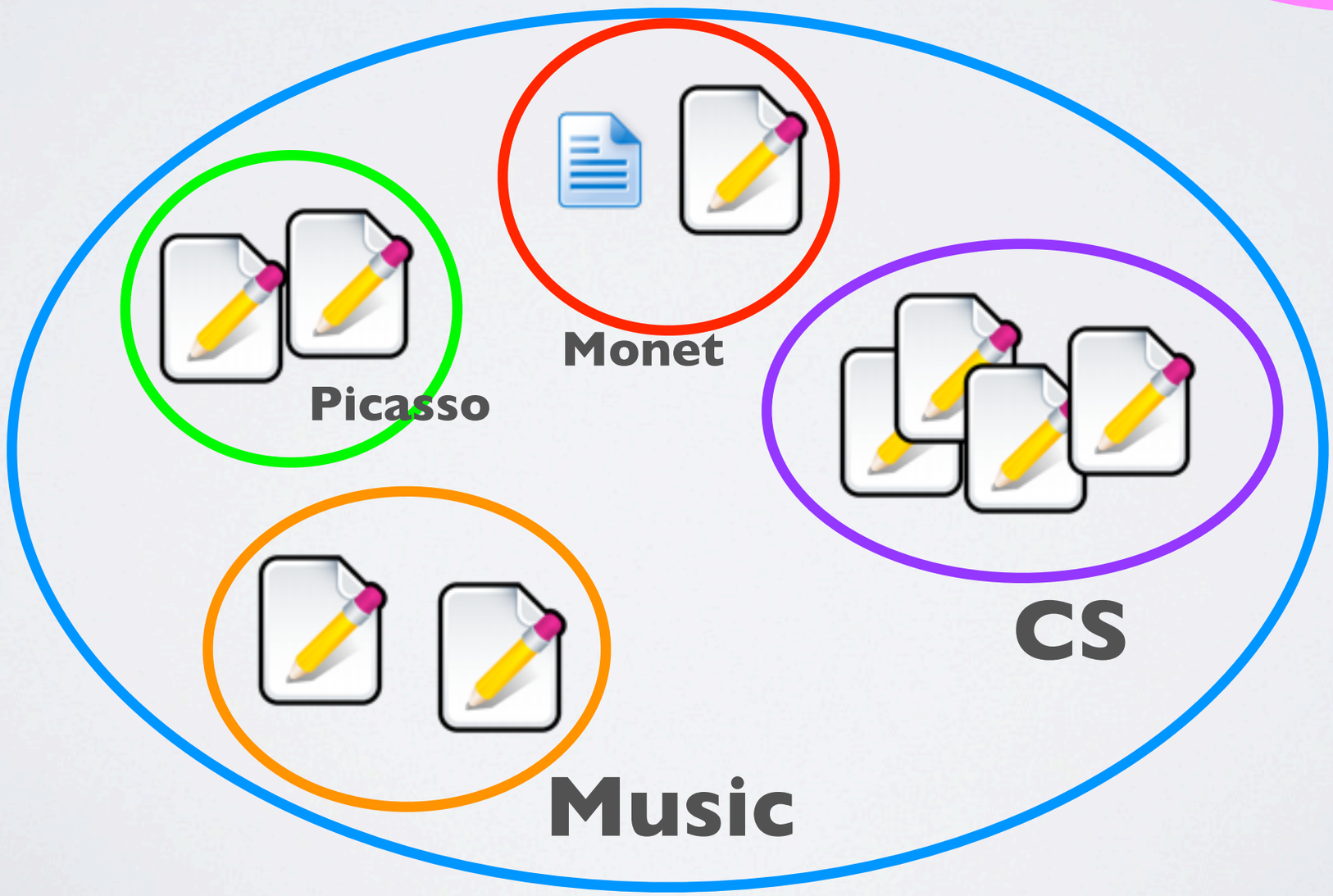




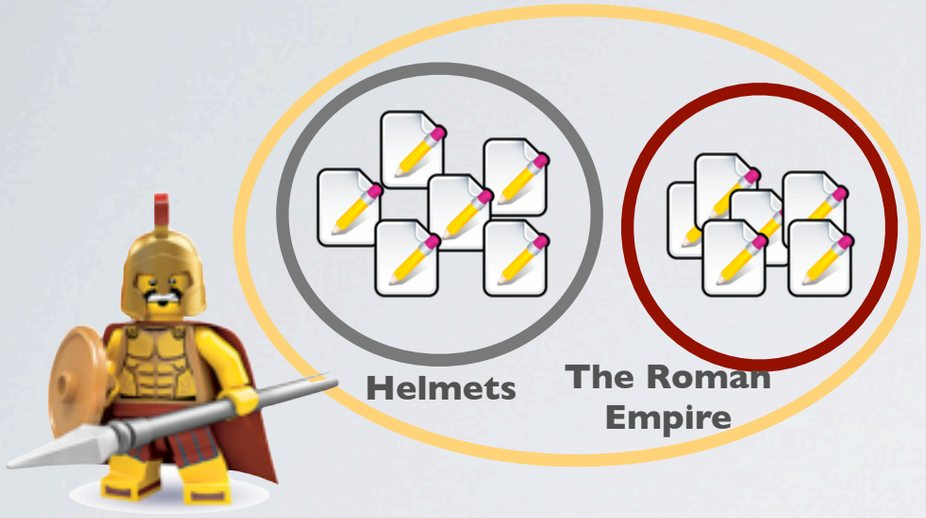
STRUCTURING

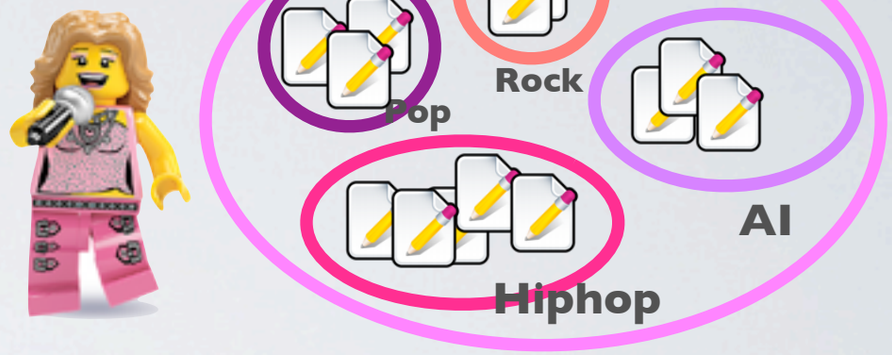


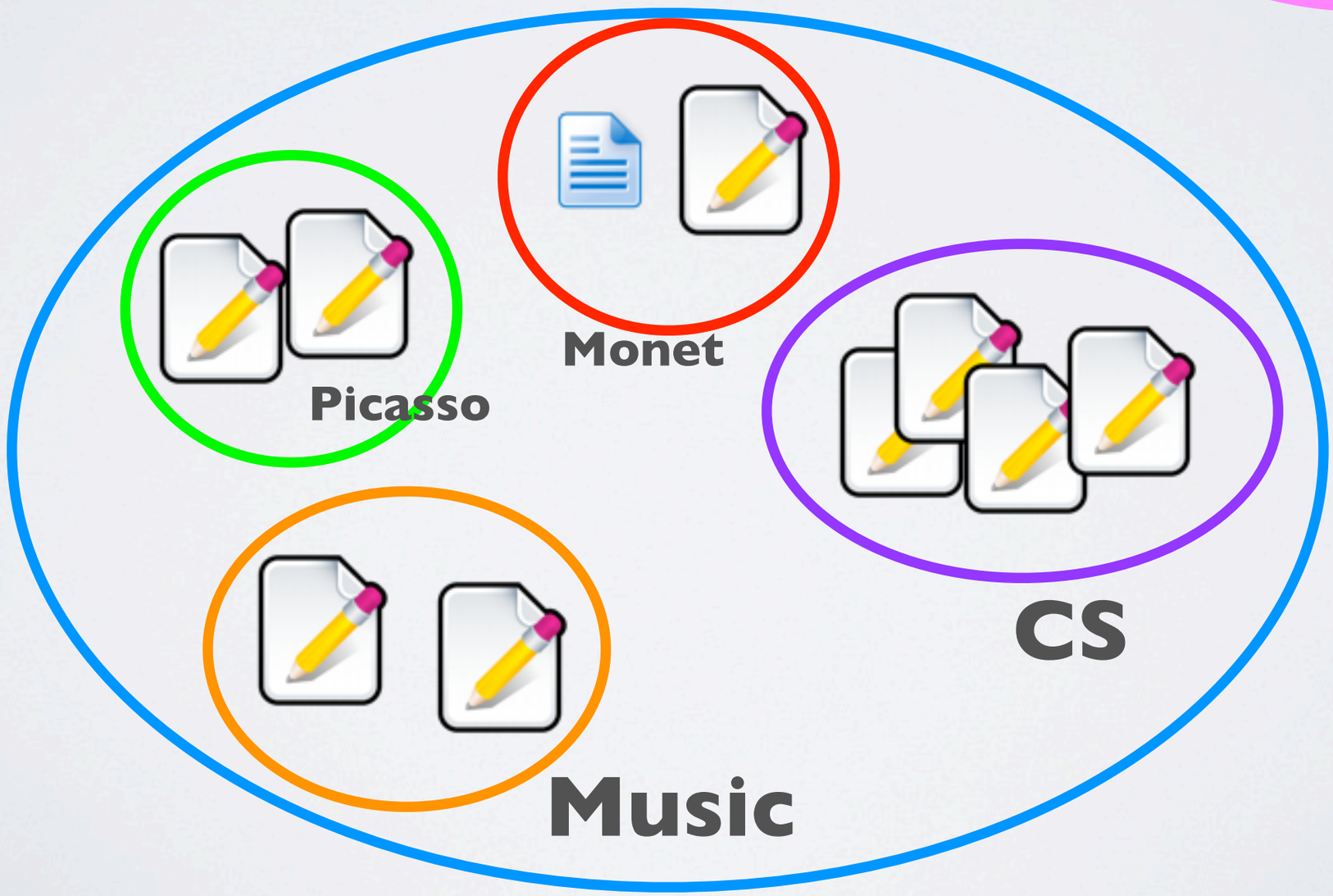




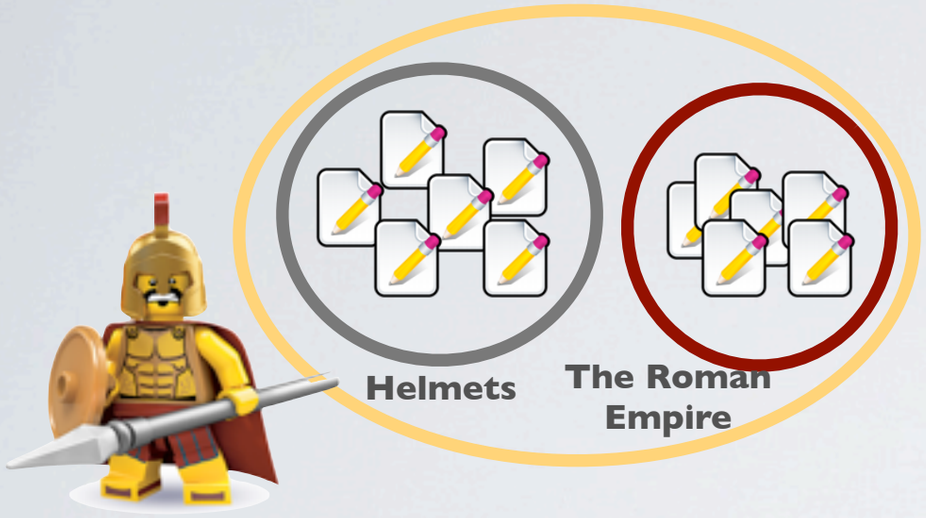
STRUCTURING

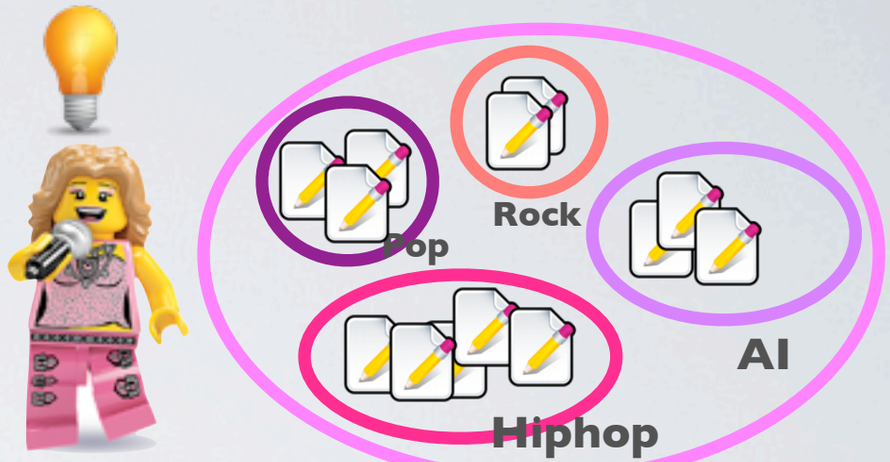


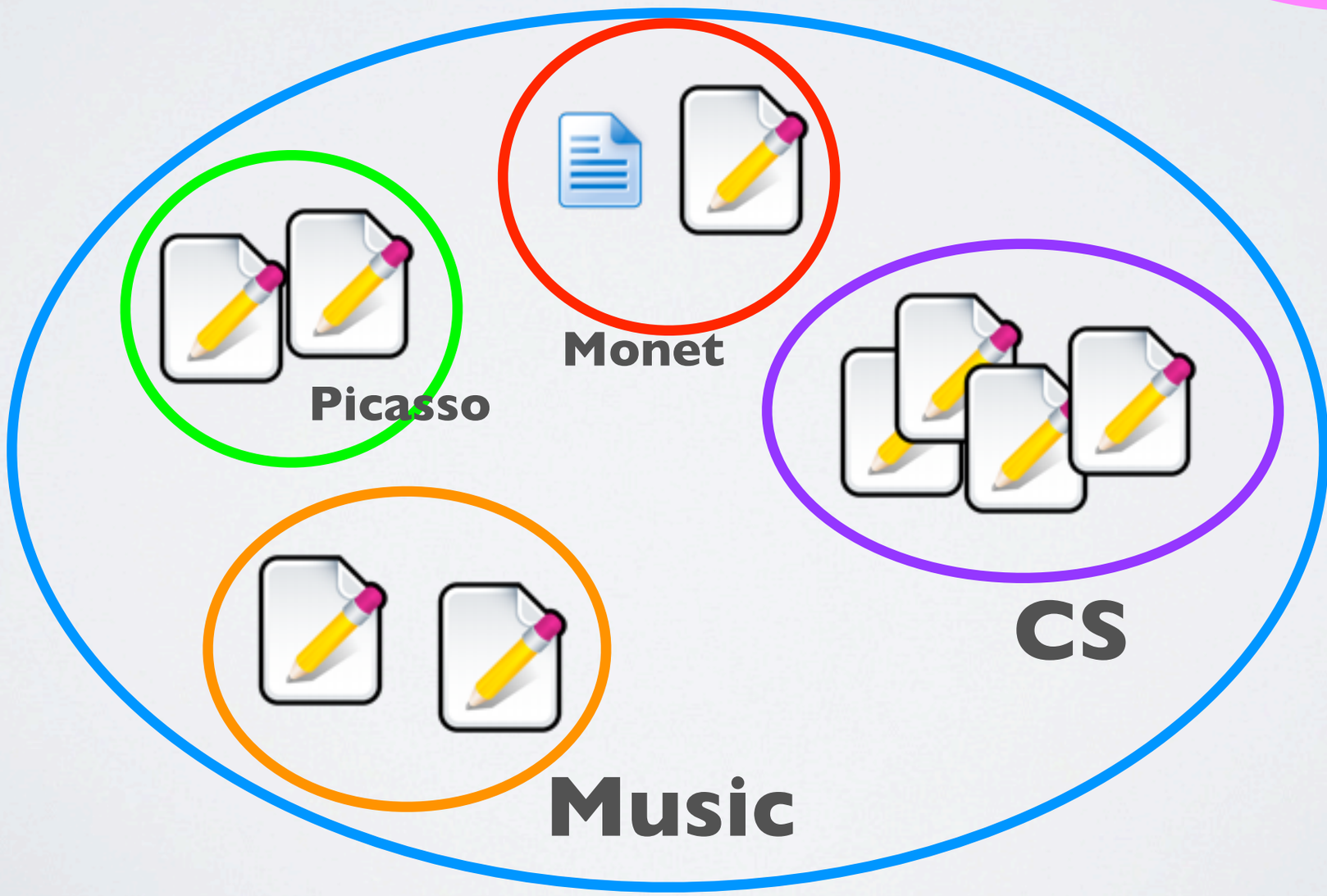




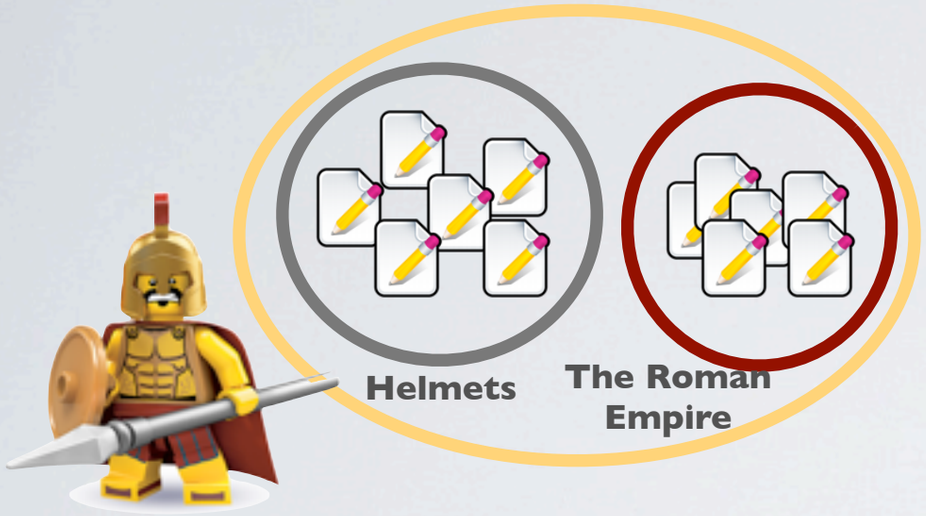
STRUCTURING

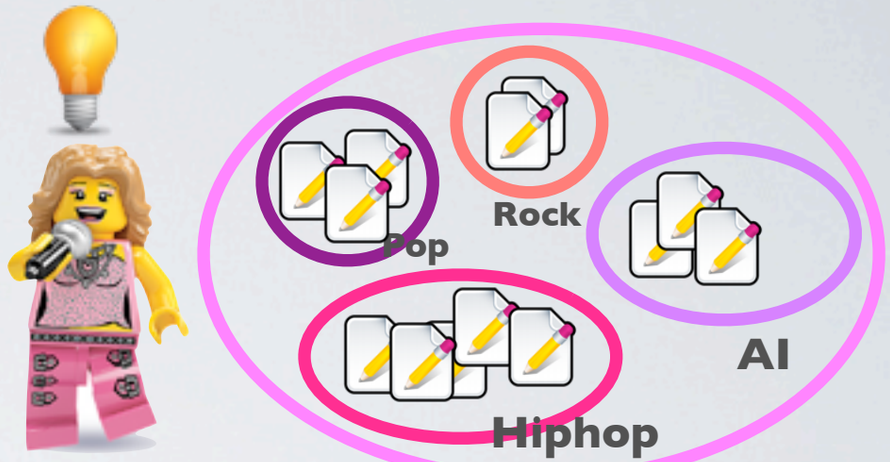


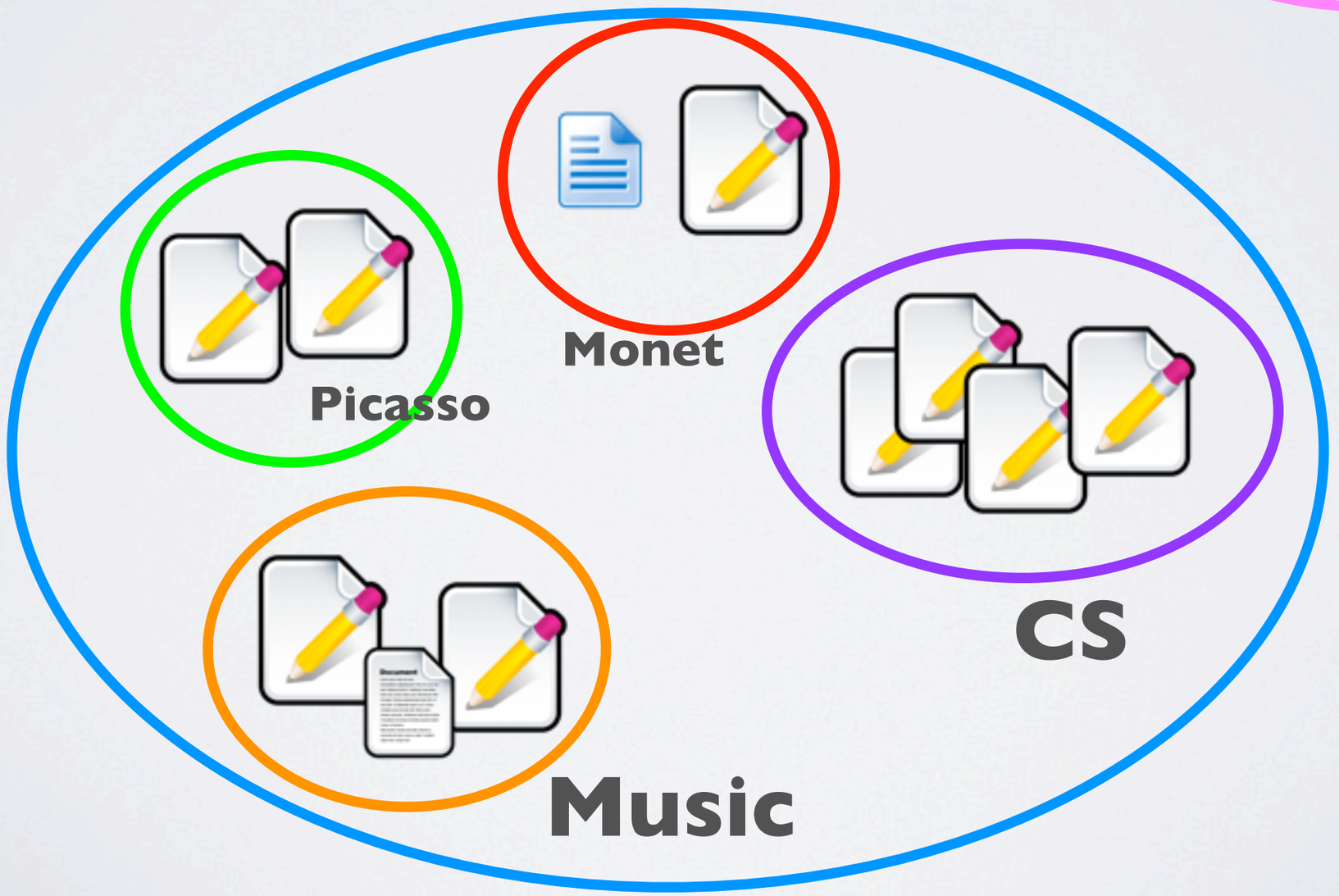




STRUCTURING







... AND ITS DYNAMICS

- Goal: insight in structuring dynamics
  → dynamic conceptual clustering that simulates the intellectual structuring process

- Two types of *guides*:
  1. own prior structuring
  2. structuring of peers

AT A GLANCE

Combination of 2 data mining tasks:

1. Learn model of structuring (classification)
   = intension: set of conditions for an object to belong to a certain class
   (vs. extension: list of objects in class)

2. Use intension or extension to structure new items

   A. based on own structuring

   B. based on k peers

GROUPING GUIDANCE

BASIC NOTATION

- U: the set of all users (used symbols: u, v, w)
- T: the set of all time points $\{0, 1, ..., tmax\}$, where $tmax$ represents the time at which the last item arrives
- D: the set of all items (used symbol: d)

  $D_u^t \subseteq D$: the set of all $d \in D$ already considered by $u \in U$ at $t \in T$

  $d_u^t \in (D \setminus D_u^t)$: the item assigned to the structure by user u at t

GROUPING GUIDANCE

GROUPINGS AND CLASSIFIERS

- G: (machine-induced) groupings for each user's items
- C: classifiers (i.e. intensions) learned for these groupings

  - OG: Observed Grouping
  - GS: Simulated Grouping, guided by self
  - Gn: Simulated Grouping, guided by *n* peers

INITIAL CLASSIFIER LEARNING

Goal: determine intensional definitions for
the user-generated groupings

→ Each group is regarded as a class for which
a definition needs to be calculated

↓

Definitions used to assign new items to these gorups

CLASSIFIER SELECTION

= selection of peer guides

CLASSIFIER SELECTION

= selection of peer guides

→ Requires divergence measure between groupings of non-identical item sets

CLASSIFIER SELECTION

= selection of peer guides

→ Requires divergence measure between groupings of non-identical item sets

... but existing measures require large overlap between sets

CLASSIFIER SELECTION

= selection of peer guides

→ Requires divergence measure between groupings of non-identical item sets

... but existing measures require large overlap between sets

↓

Inter-guide measure of diversity:

$$udiv(u,v) = \frac{1}{2}(1/|G_u^t| \sum_{x \in G_u^t} min_{y \in G_v^t} gdiv(x,y) + 1/|G_v^t| \sum_{y \in G_v^t} min_{x \in G_u^t} gdiv(y,x))$$

CLASSIFICATION

Selected classifiers are used to classify the item under consideration

Two cases:

- Self-guided classification
- Peer-guided classification

SELF-GUIDED

U

SELF-GUIDED

U

SELF-GUIDED

U

t_u^1: item a reaches u

t_u^2: item b reaches u

OG_u^0: observed initial grouping, learned: intensional description via classifier learning from extension

apply classifier OC_u^0 to item a; learn the new classifier CS_u^1

apply classifier CS_u^1 to item b; learn the new classifier CS_u^2

OG_u^{tmax}	$G20_u^{tmax}$	$G10_u^{tmax}$	$G5_u^{tmax}$	$G1_u^{tmax}$	GS_u^0
(observed)	(simulated, 20 peers)	(simulated, 10 peers)	(simulated, 5 peers)	(simulated, 1 peer)	(simulated, self-guided)

SELF-GUIDED

U

OG_u^0: observed initial grouping, learned: intensional description via classifier learning from extension

apply classifier OC_u^0 to item a; learn the new classifier CS_u^1

apply classifier CS_u^1 to item b; learn the new classifier CS_u^2

apply classifier CS_u^2 to item c; learn the new classifier CS_u^3

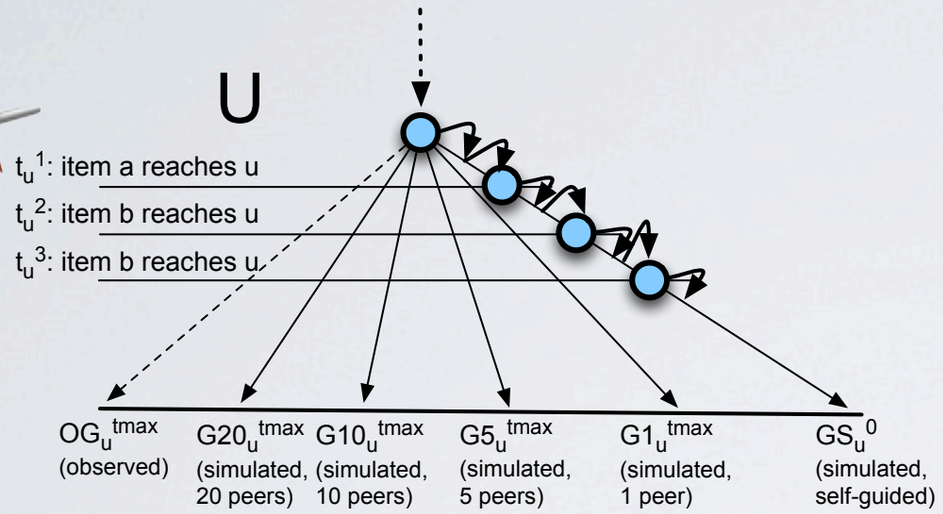
t_u^1: item a reaches u

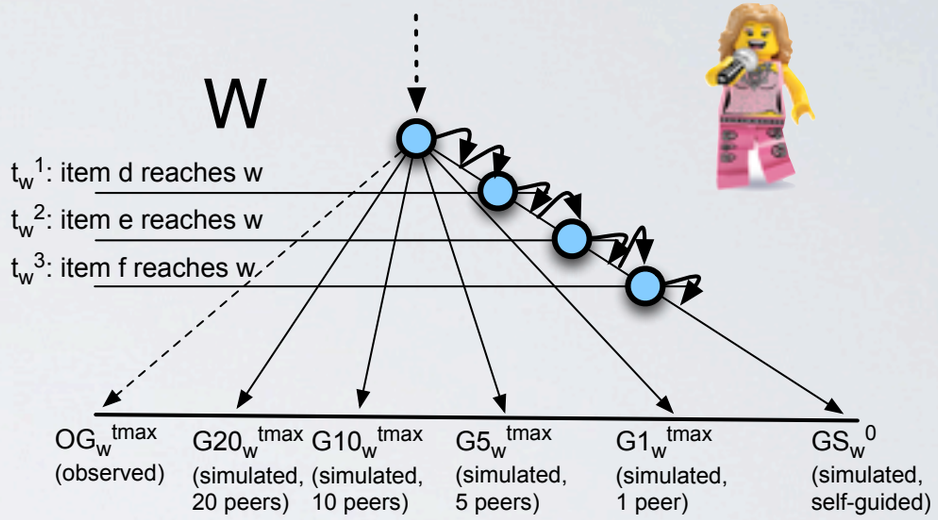
t_u^2: item b reaches u

t_u^3: item c reaches u

OG_u^{tmax}	$G20_u^{tmax}$	$G10_u^{tmax}$	$G5_u^{tmax}$	$G1_u^{tmax}$	GS_u^0
(observed)	(simulated, 20 peers)	(simulated, 10 peers)	(simulated, 5 peers)	(simulated, 1 peer)	(simulated, self-guided)

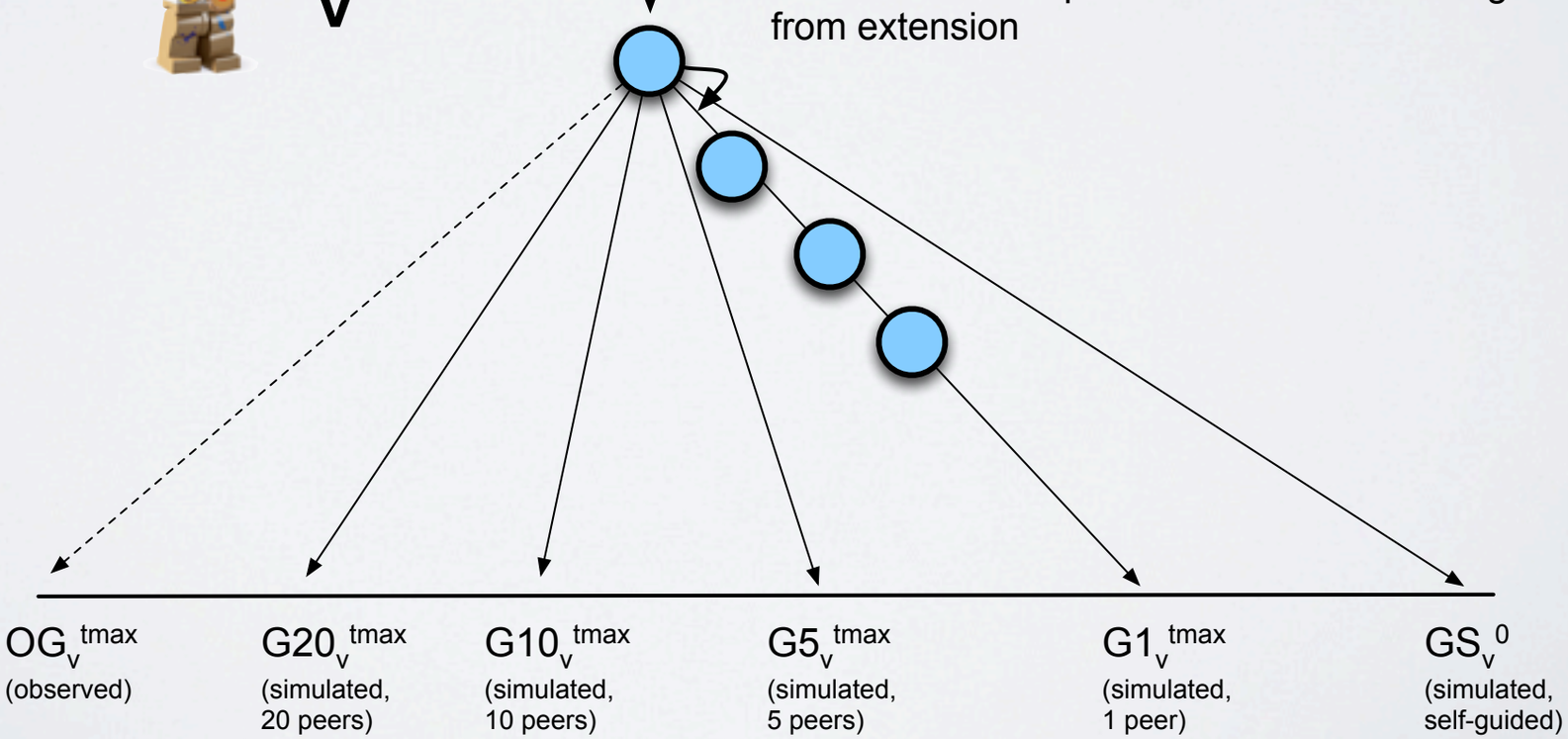
PEER-GUIDED





V

OG_v^0: observed initial grouping, learned: intensional description via classifier learning from extension



PEER-GUIDED

U

t_u^1: item a reaches u

t_u^2: item b reaches u

t_u^3: item b reaches u

OG_u^{tmax} (observed)

$G20_u^{tmax}$ (simulated, 20 peers)

$G10_u^{tmax}$ (simulated, 10 peers)

$G5_u^{tmax}$ (simulated, 5 peers)

$G1_u^{tmax}$ (simulated, 1 peer)

GS_u^0 (simulated, self-guided)

W

t_w^1: item d reaches w

t_w^2: item e reaches w

t_w^3: item f reaches w

OG_w^{tmax} (observed)

$G20_w^{tmax}$ (simulated, 20 peers)

$G10_w^{tmax}$ (simulated, 10 peers)

$G5_w^{tmax}$ (simulated, 5 peers)

$G1_w^{tmax}$ (simulated, 1 peer)

GS_w^0 (simulated, self-guided)

V

OG_v^0: observed initial grouping, learned: intensional description via classifier learning from extension

apply classifier OC_u^0 to item g; learn the new classifier $C1_v^1$

t_v^1: item g reaches v

OG_v^{tmax} (observed)

$G20_v^{tmax}$ (simulated, 20 peers)

$G10_v^{tmax}$ (simulated, 10 peers)

$G5_v^{tmax}$ (simulated, 5 peers)

$G1_v^{tmax}$ (simulated, 1 peer)

GS_v^0 (simulated, self-guided)

PEER-GUIDED

U

t_u^1: item a reaches u

t_u^2: item b reaches u

t_u^3: item b reaches u

OG_u^{tmax} (observed)

$G20_u^{tmax}$ (simulated, 20 peers)

$G10_u^{tmax}$ (simulated, 10 peers)

$G5_u^{tmax}$ (simulated, 5 peers)

$G1_u^{tmax}$ (simulated, 1 peer)

GS_u^0 (simulated, self-guided)

W

t_w^1: item d reaches w

t_w^2: item e reaches w

t_w^3: item f reaches w

OG_w^{tmax} (observed)

$G20_w^{tmax}$ (simulated, 20 peers)

$G10_w^{tmax}$ (simulated, 10 peers)

$G5_w^{tmax}$ (simulated, 5 peers)

$G1_w^{tmax}$ (simulated, 1 peer)

GS_w^0 (simulated, self-guided)

V

OG_v^0: observed initial grouping, learned: intensional description via classifier learning from extension

apply classifier OC_u^0 to item g;
learn the new classifier $C1_v^1$

apply classifier CS_u^1 to item h;
learn the new classifier $C1_v^2$

t_v^1: item g reaches v

t_v^2: item h reaches v

OG_v^{tmax} (observed)

$G20_v^{tmax}$ (simulated, 20 peers)

$G10_v^{tmax}$ (simulated, 10 peers)

$G5_v^{tmax}$ (simulated, 5 peers)

$G1_v^{tmax}$ (simulated, 1 peer)

GS_v^0 (simulated, self-guided)

PEER-GUIDED

U

t_u^1: item a reaches u

t_u^2: item b reaches u

t_u^3: item b reaches u

OG_u^{tmax} (observed)

$G20_u^{tmax}$ (simulated, 20 peers)

$G10_u^{tmax}$ (simulated, 10 peers)

$G5_u^{tmax}$ (simulated, 5 peers)

$G1_u^{tmax}$ (simulated, 1 peer)

GS_u^0 (simulated, self-guided)

W

t_w^1: item d reaches w

t_w^2: item e reaches w

t_w^3: item f reaches w

OG_w^{tmax} (observed)

$G20_w^{tmax}$ (simulated, 20 peers)

$G10_w^{tmax}$ (simulated, 10 peers)

$G5_w^{tmax}$ (simulated, 5 peers)

$G1_w^{tmax}$ (simulated, 1 peer)

GS_w^0 (simulated, self-guided)

V

OG_v^0: observed initial grouping, learned: intensional description via classifier learning from extension

apply classifier OC_u^0 to item g;
learn the new classifier $C1_v^1$

apply classifier CS_u^1 to item h;
learn the new classifier $C1_v^2$

apply classifier CS_w^2 to item i;
learn the new classifier $C1_u^3$

t_v^1: item g reaches v

t_v^2: item h reaches v

t_v^3: item i reaches v

OG_v^{tmax} (observed)

$G20_v^{tmax}$ (simulated, 20 peers)

$G10_v^{tmax}$ (simulated, 10 peers)

$G5_v^{tmax}$ (simulated, 5 peers)

$G1_v^{tmax}$ (simulated, 1 peer)

GS_v^0 (simulated, self-guided)

DATASET

CiteULike dataset

→ sampled with p-core subgraphs to overcome sparsity

# users	377
# documents	11,400
# tags	12,982
timeframe	01/2009 - 02/2010

INITIAL GROUPING

Tagging as implicit structuring

- First 7 months to learn initial grouping
- Modularity clustering

INITIAL GROUPING

Tagging as implicit structuring

- First 7 months to learn initial grouping
- Modularity clustering

Initial classifier learning

- High dimensional input space (BoW of abstracts)

→ Naive bayes

SIMULATING GROUPINGS

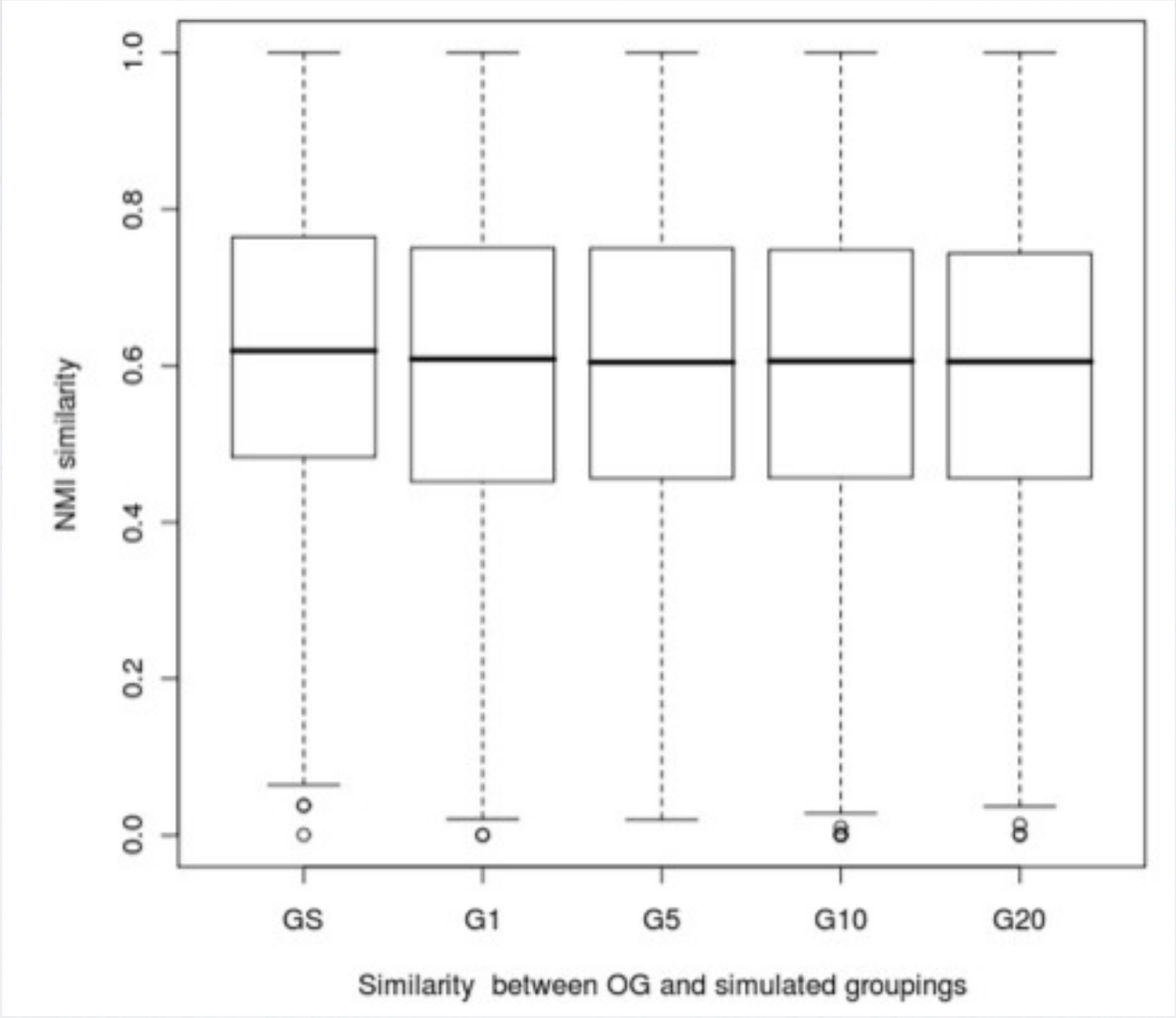
- Groups represented by language models
- Jensen-Shannon divergence as inter-group divergence

$$JS(\Theta_x, \Theta_y) = \frac{1}{2} KL(\Theta_x, \Theta_z) + \frac{1}{2} KL(\Theta_y, \Theta_z)$$

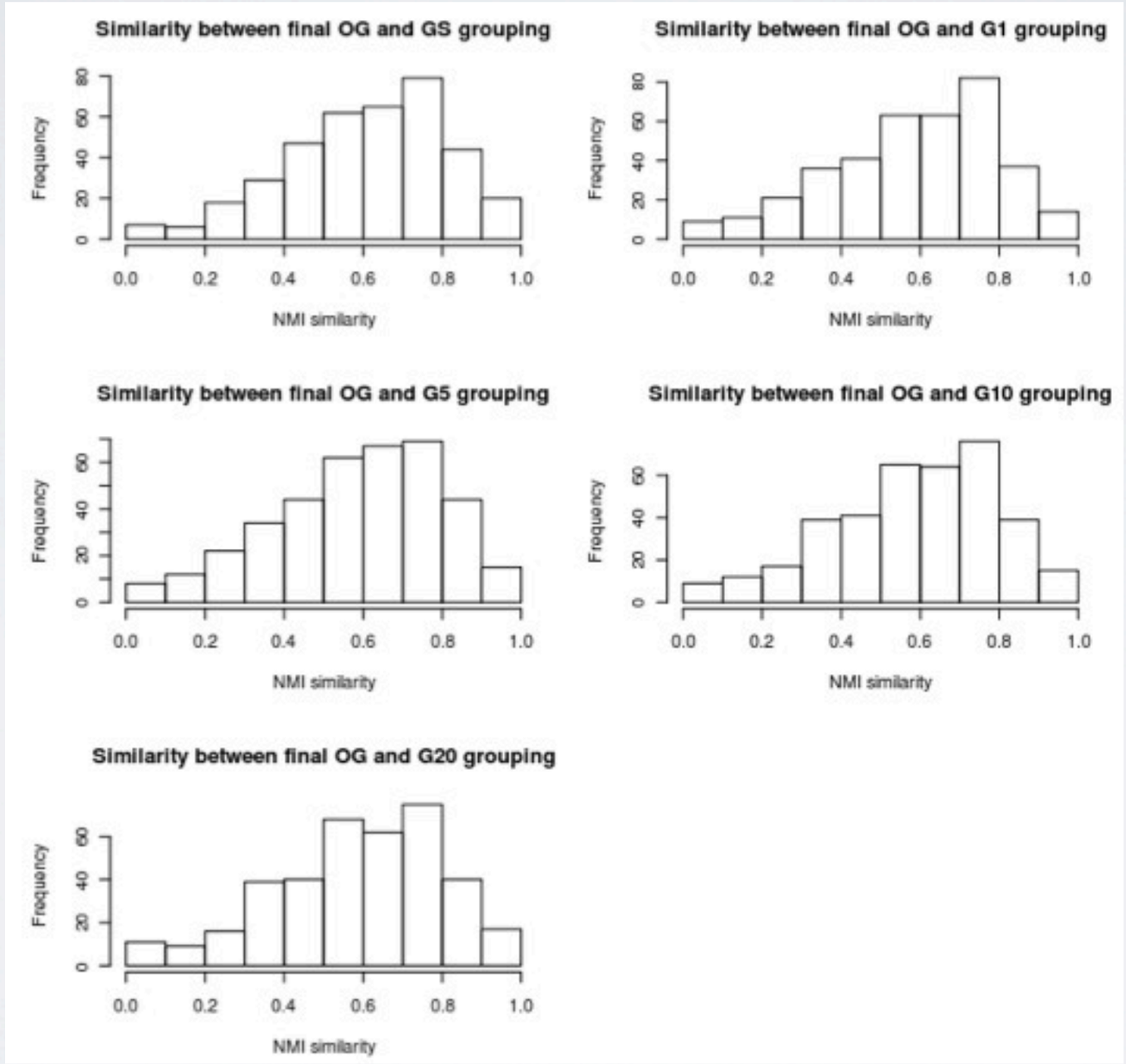
- Normalized Mutual Information to compare groupigns

$$NMI(G, G') = H(G) + H(G') - \frac{H(G, G')}{\sqrt{H(G)H(G')}}$$

RESULTS


NMI similarity
0.0
0.2
0.4
0.6
0.8
1.0
GS
G1
G5
G10
G20
Similarity between OG and simulated groupings

SIMILARITY DISTRIBUTION



CONCLUSIONS

- Investigate and simulate collaborative structuring

→ Learning and combining classifiers for itemset structuring

- New divergence measure

- Tested on social-bookmarking platform for literature management

CONCLUSIONS

LIMITATIONS

- Observed groupings based on tag assignements
- Simple classifier

... but provides initial insights into grouping behaviour and behaviour of users in social bookmarking systems

FUTURE WORK

- Applications: (tag) recommendation and social search

→ Adds new level to individual and social measures

- Regrouping based on peers

- Hybrid measure for itemset structuring

THANKS!

QUESTIONS?