MO TM

Osama Khalifa, David Corne, Mike Chantler, Fraser Halley

= material liberated from:

David Blei's topic modelling resources

http://www.cs.princeton.edu/~blei/topicmodeling.html

and maybe others

Q: how can I understand 1,000,000 documents without reading them?

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, and search the texts.

A topic is a probability distribution over words

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

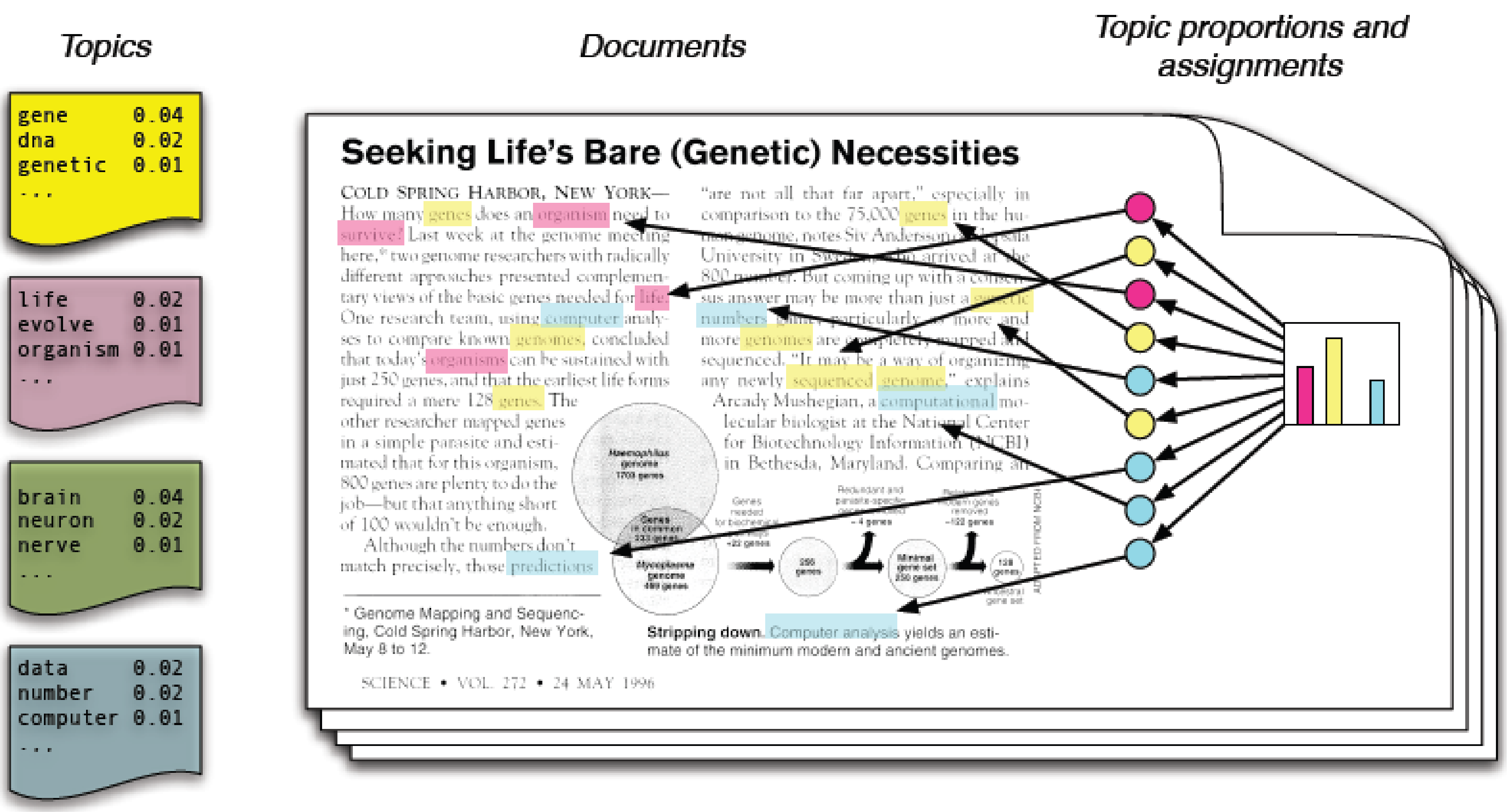
word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

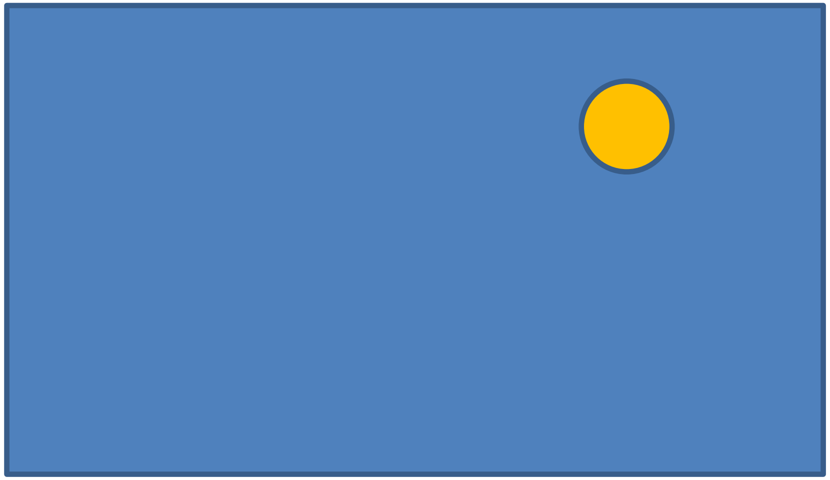
word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

A document (e.g. an EPSRC project summary) is a probability distribution over topics

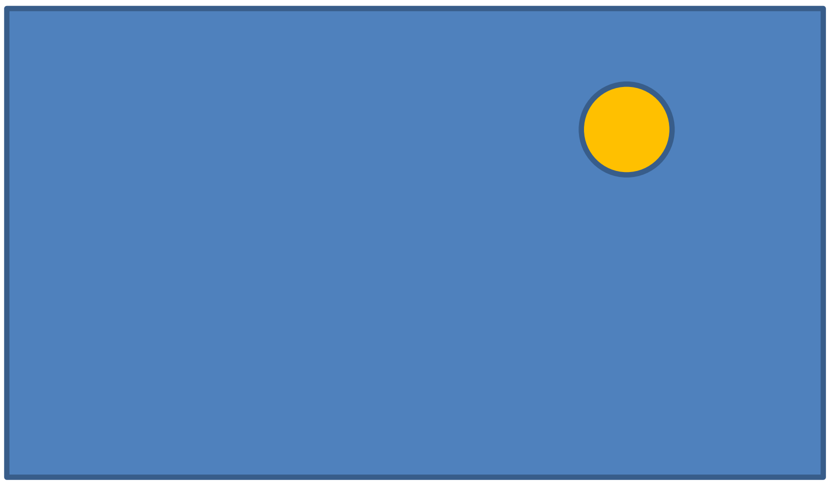


TFIDF / Bag of Words / etc



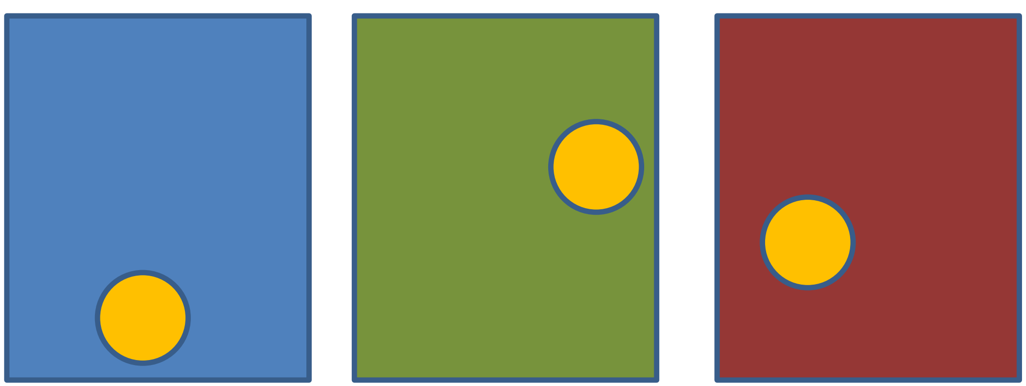
Document is a single point in doc space

TFIDF / Bag of Words / etc



Document is a
single point in
doc space

Topic Modelling



Document is a
vector of points
In small num of
meaningful spaces

a topic model

Is a model of: a document corpus

*It has a COLLECTION T of **TOPICS***

*Each TOPIC is a distribution over **WORDS***

For each DOCUMENT Di there is a

*distribution over **TOPICS** (usually sparse)*

a topic model

Is a model of: a document corpus

*It has a COLLECTION T of **TOPICS***

*Each TOPIC is a distribution over **WORDS***

For each DOCUMENT Di there is a

*distribution over **TOPICS*** (*usually sparse*)

And to do the maths:

For each WORD in a document,

*there is a **TOPIC assignment***

Some of the 600 ICT topics

www.researchperspectives.org/

267 flexible enable required future approach (43)

596 nuclear waste reactor fuel radioactive (43)

042 electron energy electrical exciting charge (43)

566 change research work information technology (43)

010 optical switching transmission wavelength modules (42)

591 statistical models inference data analysis (42)

379 india uk technology rural economy (42)

540 biological systems biology inspired synthetic (42)

442 search retrieval information index query (42)

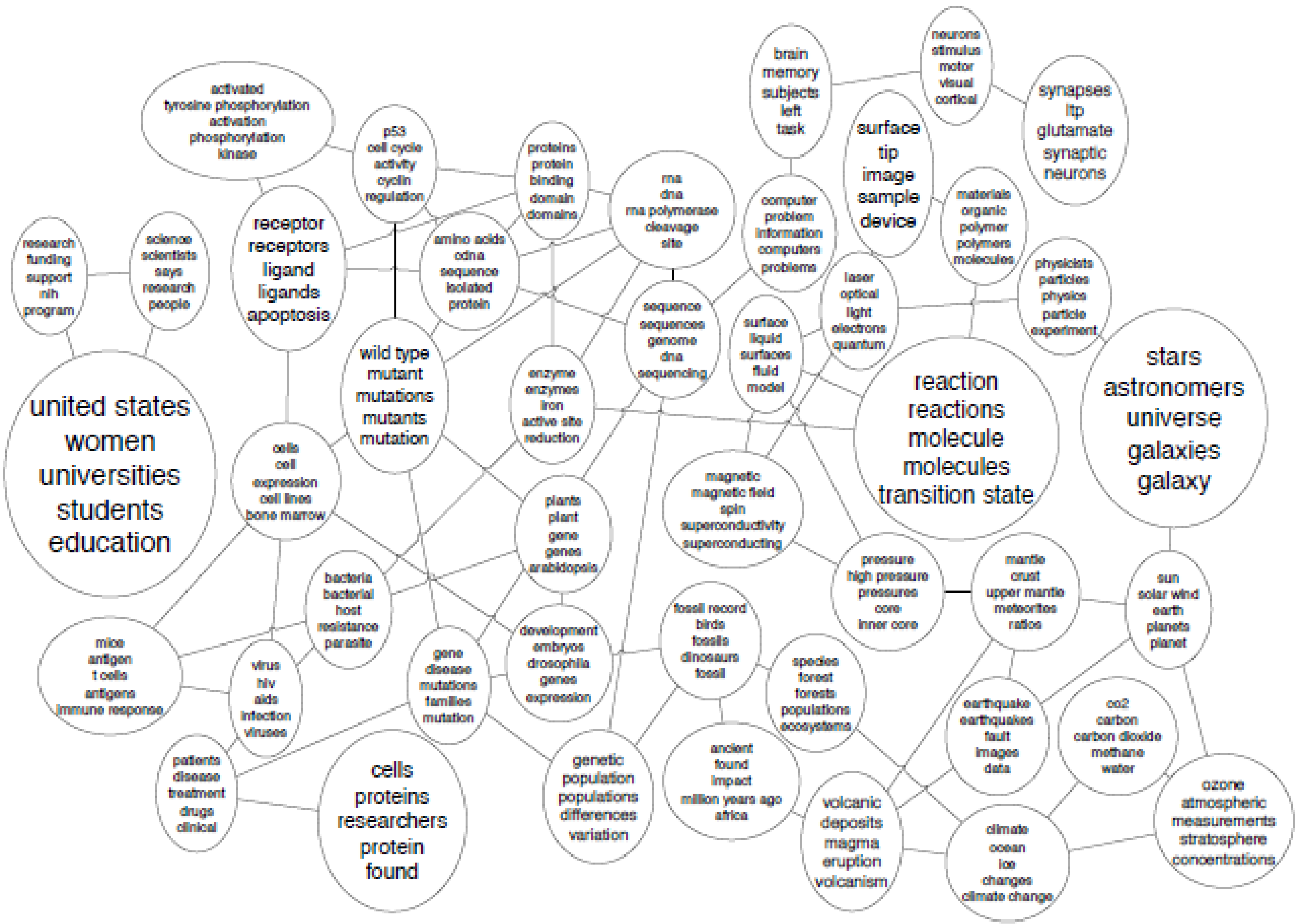
116 security protocols key cryptographic cryptography (41)

487 face expression emotional systems facial (41)

221 data mining analysis large datasets (41)

152 applications development including specific medical (41)

topic models helps us discover structure



it's

HOT

The maths:
inferring a TM from a document corpus

A Generative model for writing a document

A Generative model for writing a document

To generate a document:

A Generative model for writing a document

To generate a document:

SAMPLE from a distribution D_T over the topics T

A Generative model for writing a document

To generate a document:

SAMPLE from a distribution D_T over the topics T

Repeat until document is written:

A Generative model for writing a document

To generate a document:

SAMPLE from a distribution D_T over the topics T

Repeat until document is written:

SAMPLE a topic z from D_T

A Generative model for writing a document

To generate a document:

SAMPLE from a distribution D_T over the topics T

Repeat until document is written:

SAMPLE a topic z from D_T

SAMPLE a word w from z

A Generative model for writing a document

To generate a document:

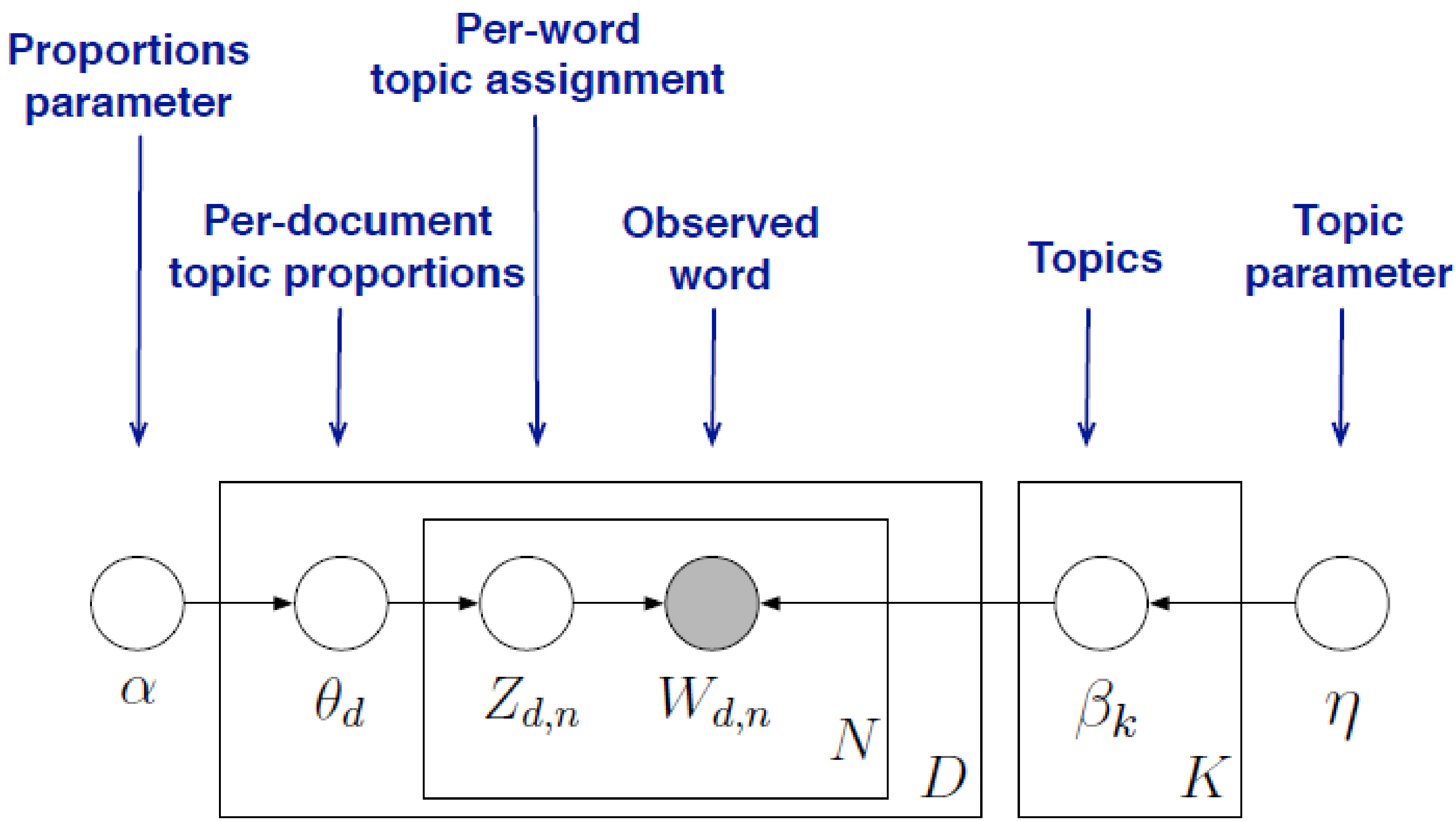
SAMPLE from a distribution D_T over the topics T

Repeat until document is written:

SAMPLE a topic z from D_T

SAMPLE a word w from z

Perhaps only valid for Shakespearean monkeys, but:



$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Joint probability of the topics and D_T s

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

... basically gives likelihood of a given TM of your corpus

Joint probability of the topics and D_T s

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

... basically gives likelihood of a given TM of your corpus

Do some mathematical jiggery-pokery (conjugate your priors, etc...), and ...

LDA /
Gibbs sampling

Latent Dirichlet Allocation

Represent your corpus as a huge vector of topic assignments

[z1, z2, z3, z4, …, zn-1, zn]

Run MCMC to find an assignment that maximises the likelihood

LDA/Gibbs sampling is

popular, standard, faourite

Plenty else going on in the line of more sophisticated models (dynamic topics, topic hierarchies, correlated TMs, etc ..)

... all with their associated Bayesian/statistical inference schemes full of questionable assumptions and simplifications

EVALUATING TMs

Perplexity of a test corpus D_{test}

$$e^{\frac{-\sum_{d \in D_{test}} \log P(w_d | \mathcal{M})}{\sum_{d \in D_{test}} N_d}}$$

Does the test corpus tend to use words that have high probabilities within M?

Pointwise Mutual Information (PMI)

$$Pmi(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}.$$

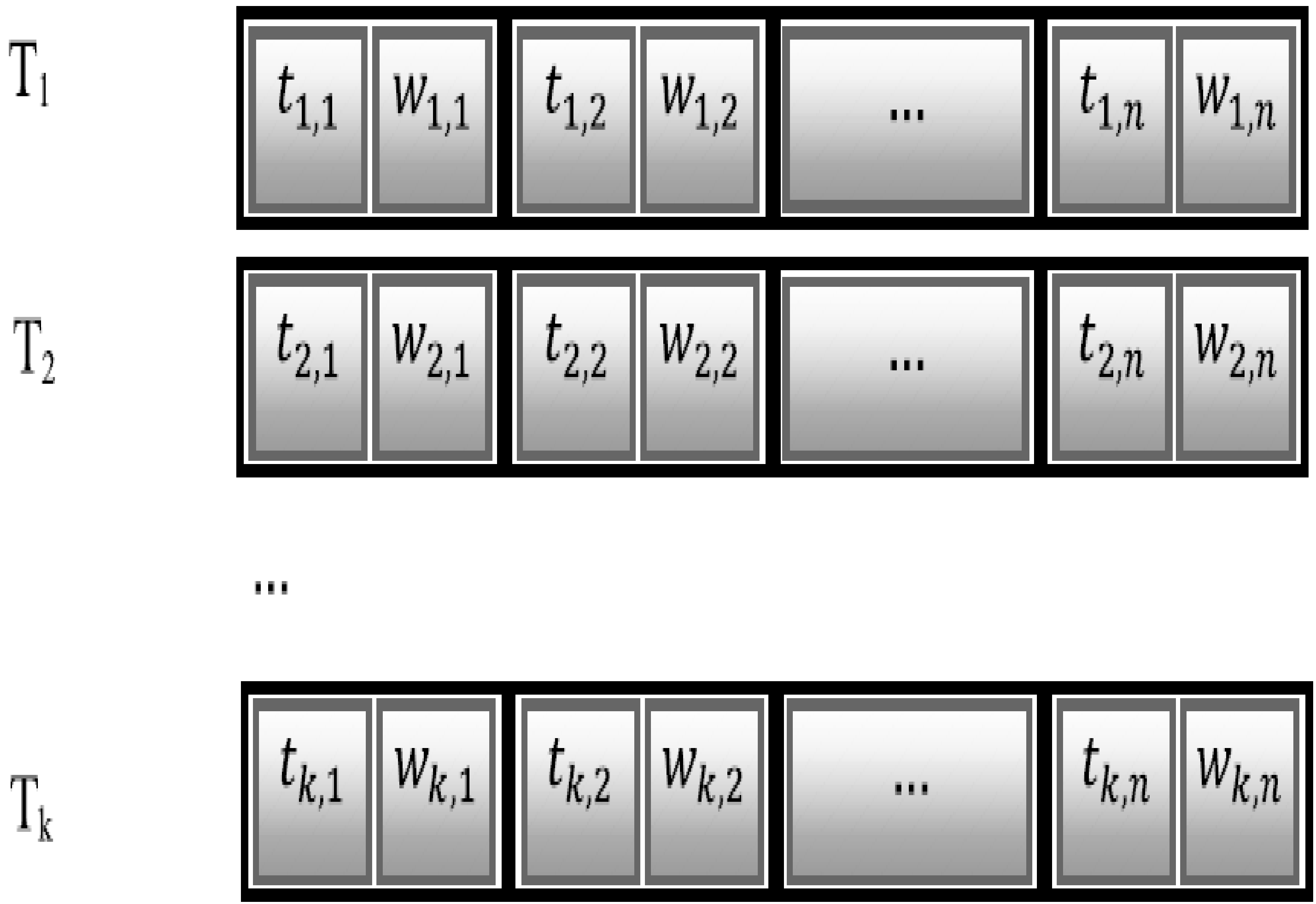
Do documents in the test corpus tend to use words from same topic within a document ?

Coverage

$$Coverage_d = \sqrt{\sum_{w \in d} \left(tf_d(w) - \sum_{i=1}^{K} T_i(w) Prop_d(T_i) \right)^2}$$

Does the TM 'cover' the majority of words in most documents ?

LDA vs MOEA-TM



T1
$t_{1,1}$ $w_{1,1}$ $t_{1,2}$ $w_{1,2}$ ... $t_{1,n}$ $w_{1,n}$
T2
$t_{2,1}$ $w_{2,1}$ $t_{2,2}$ $w_{2,2}$ ... $t_{2,n}$ $w_{2,n}$
...
Tk
$t_{k,1}$ $w_{k,1}$ $t_{k,2}$ $w_{k,2}$ ... $t_{k,n}$ $w_{k,n}$

(an important aside)

The ML/AI topic modelling community seem to be focussed on **scale**

bigger and bigger corpuses,

clever tricks to infer TMs fast,

(an important aside)

We are focussed on **quality**

- reasonably sized corpuses
- TMs that are coherent, and make sense; and who cares if it takes a few hours to run?

LDA vs MOEA-TM

‘Standalone MOEA-TM’

- **Evolves TMs using 2 objectives:**
- **coverage and PMI**
- **uses small *no. of words per topic***

LDA-initialized MOEA-TM

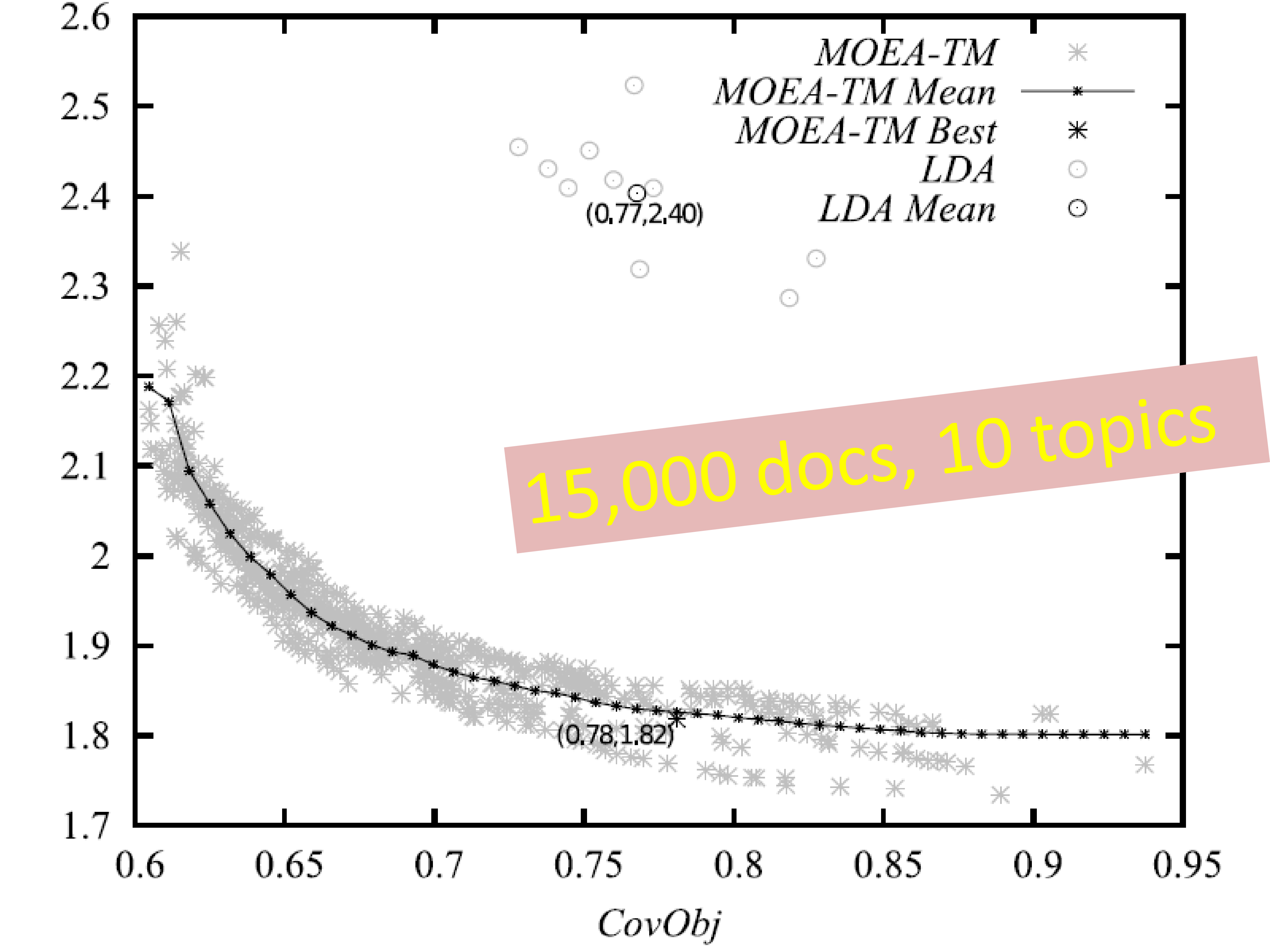
- **coverage, PMI, and perplexity**
- **compare with LDA run for equivalent t**

Everything is here

is.gd/MOEATM

LDA

http://mallet.cs.umass.edu



MOEA-TM
MOEA-TM Mean
MOEA-TM Best
LDA
LDA Mean
(0.77,2.40)
(0.78,1.82)
15,000 docs, 10 topics
2.6
2.5
2.4
2.3
2.2
2.1
2
1.9
1.8
1.7
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
CovObj

800 docs, 10 topics

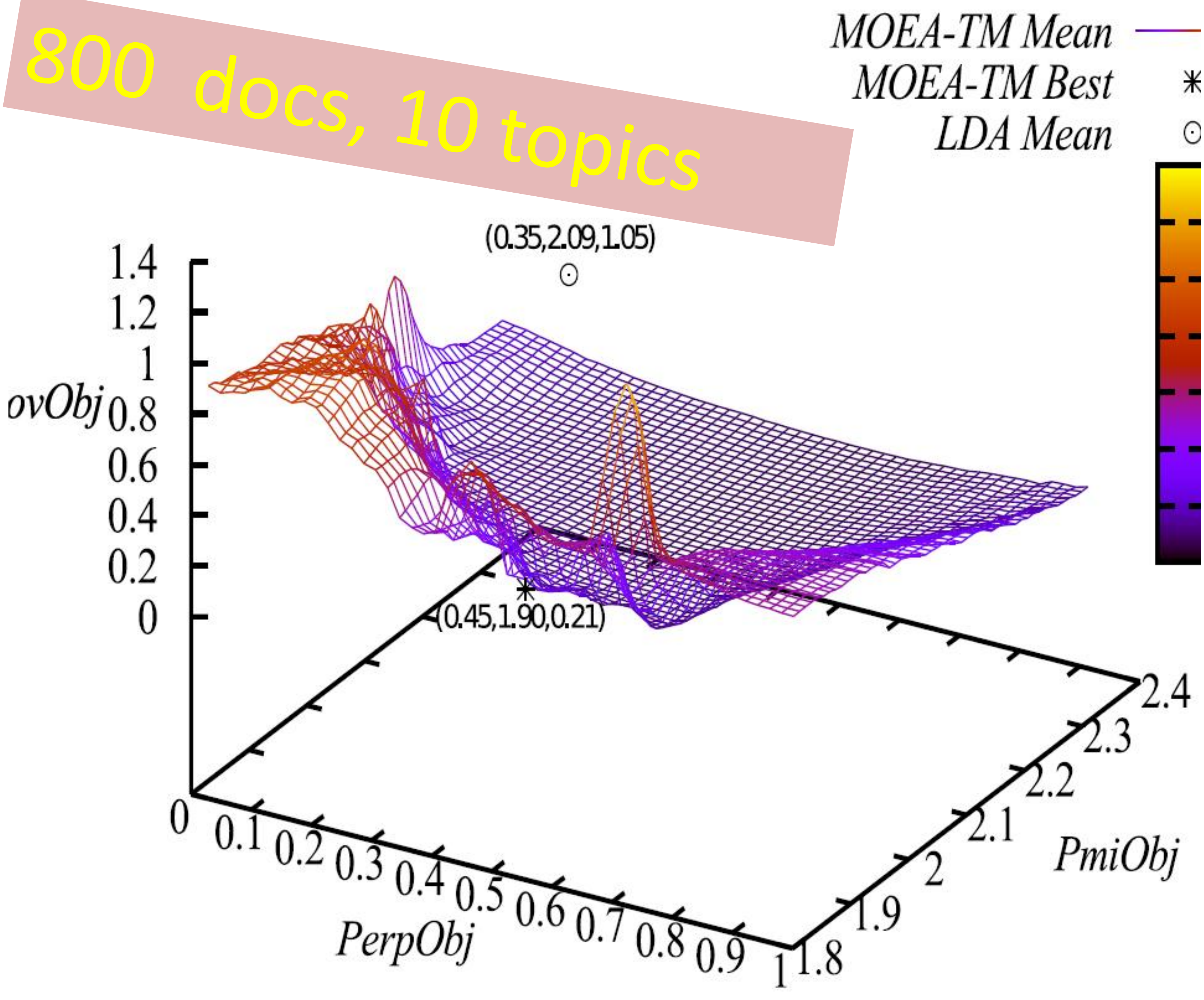
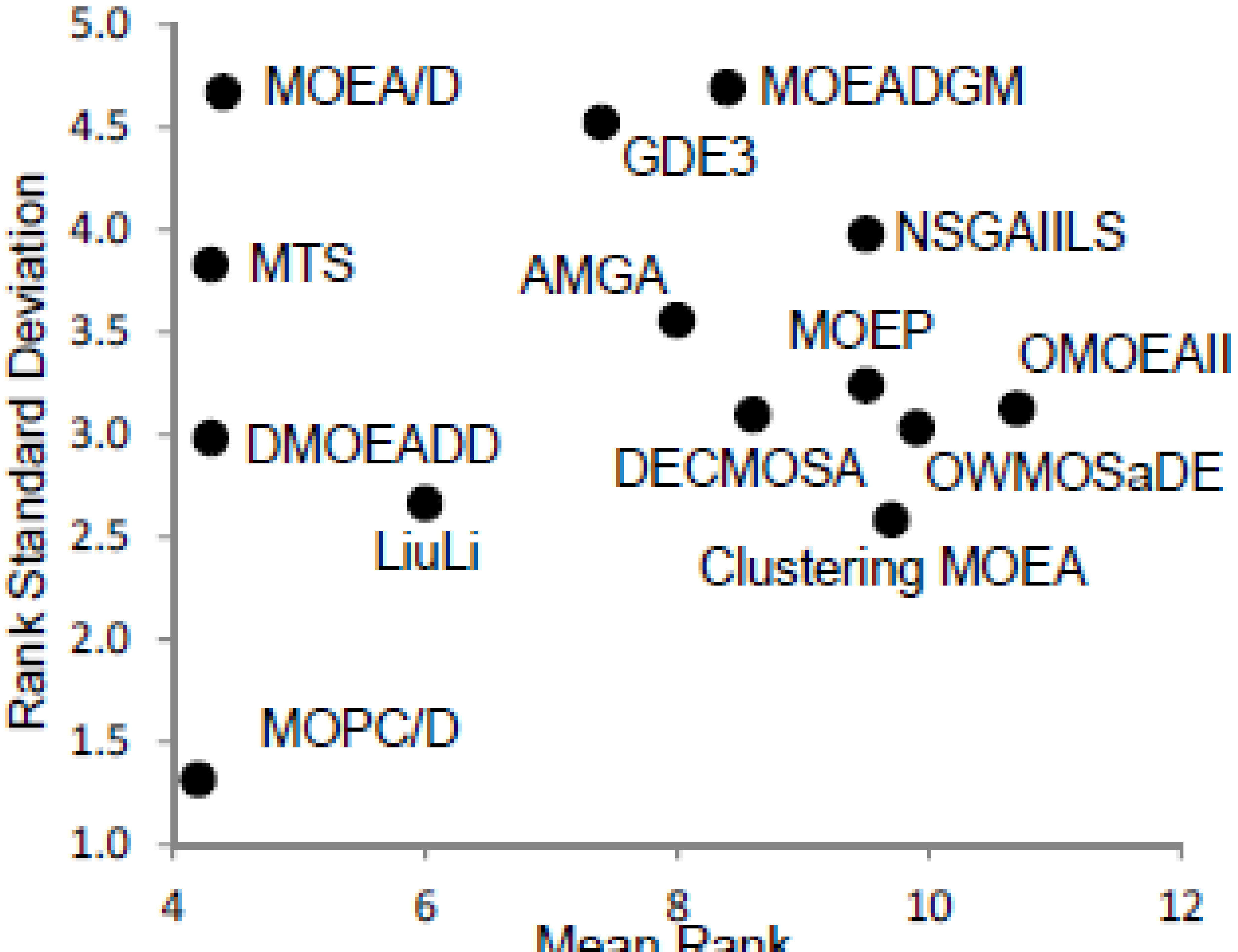


Table 2. PMI for standalone MOEA-TM and LDA for, for three corpora / ten topics.

2D	MOEA TM		LDA	
	Mean PMI	St. Deviation	Mean PMI	St. Deviation
Wiki Corpus	0.3483	0.0078	0.2158	0.0163
EPSRC Corpus	0.4264	0.0080	0.3371	0.0106
News Corpus	0.3913	0.0077	0.2448	0.0216

Table 4. PMI scores for LDA-Initialized MOEA TM and Pure LDA for the three corpora with ten topics.

3D	MOEA TM				LDA			
	PMI	St. Dev	-LL	st. Dev	PMI	St. Dev	-LL	st. Dev
Wiki Corpus	0.3105	0.0135	8.0716	0.0294	0.2013	0.0194	8.0822	0.0262
EPSRC Corpus	0.3889	0.0085	15.034	0.1005	0.3404	0.0101	15.096	0.0960
News Corpus	0.3428	0.0159	51.990	0.5377	0.2445	0.0208	53.261	0.6977

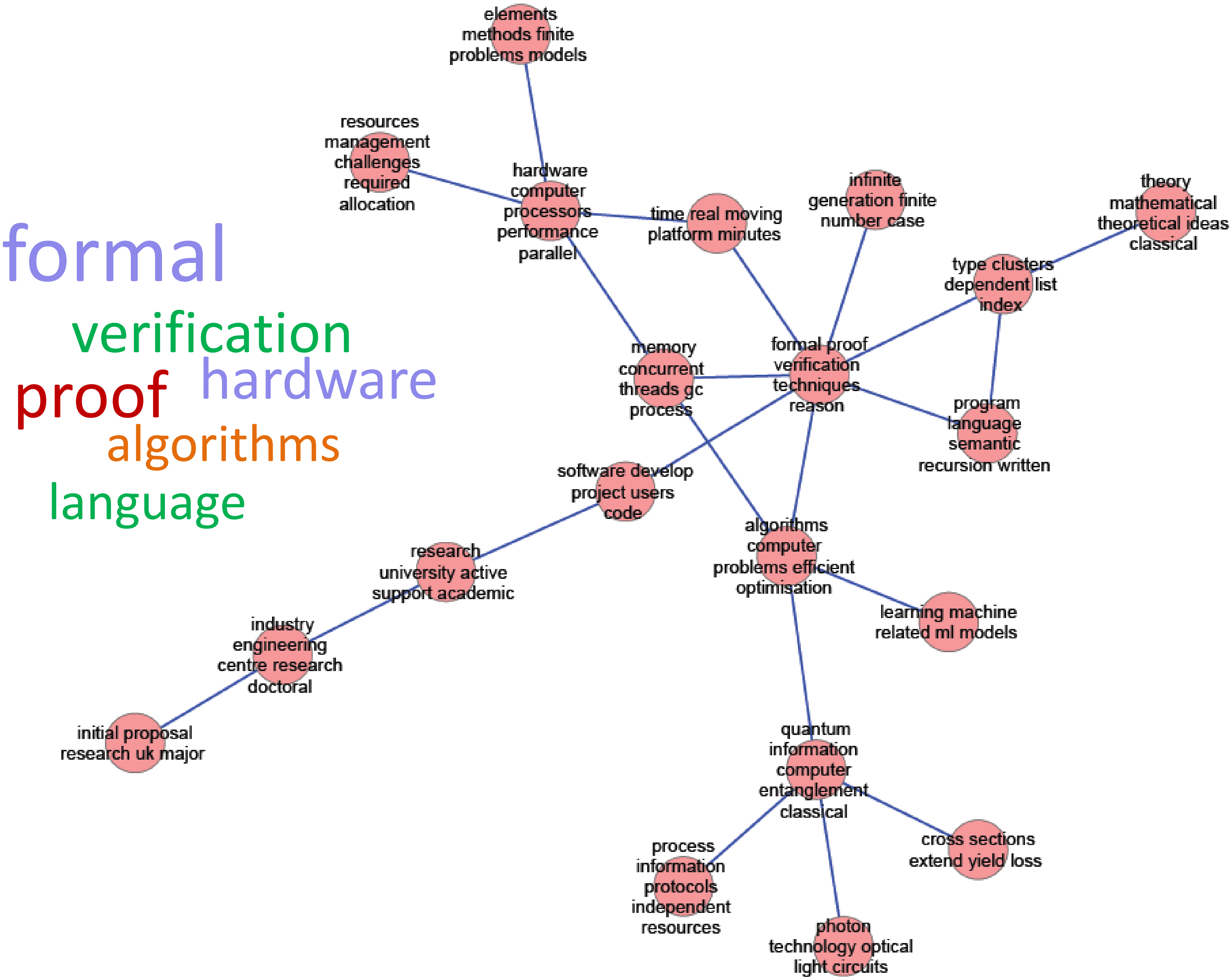


Rank Standard Deviation
Mean Rank
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4
6
8
10
12
MOEA/D
MOEADGM
GDE3
NSGAIILS
MTS
AMGA
MOEP
OMOEAII
DMOEADD
DECMOSA
OWMOSaDE
LiuLi
Clustering MOEA
MOPC/D

MO is better at TM than LDA

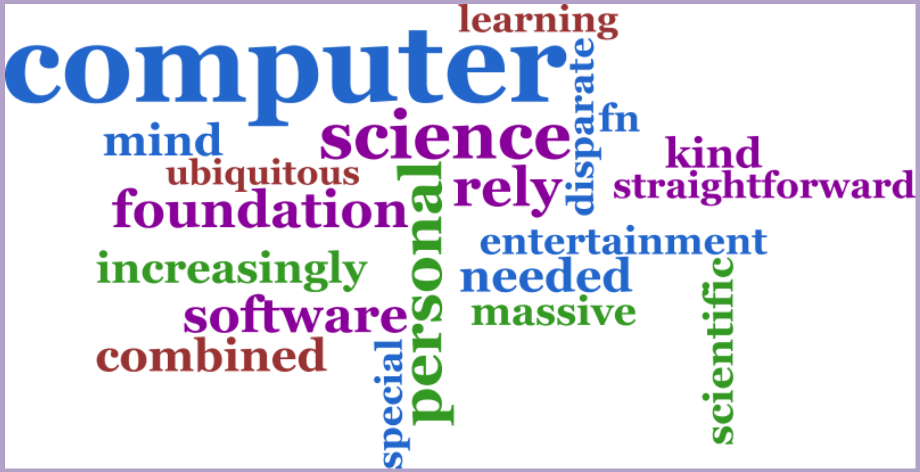
but, faster would be nice

currently we are using many more topics, and optimizing PMI, using fast (sampled) approximations


formal
verification
proof
hardware
algorithms
language
elements methods finite problems models
resources management challenges required allocation
hardware computer processors performance parallel
time real moving platform minutes
infinite generation finite number case
theory mathematical theoretical ideas classical
type clusters dependent list index
memory concurrent threads gc process
formal proof verification techniques reason
program language semantic recursion written
software develop project users code
algorithms computer problems efficient optimisation
research university active support academic
learning machine related ml models
industry engineering centre research doctoral
initial proposal research uk major
quantum information computer entanglement classical
process information protocols independent resources
cross sections extend yield loss
photon technology optical light circuits

600 topics model / ICT grants only / predictions of future funded topic interaction based on Adamic/Adar scores for currently unconnected topics

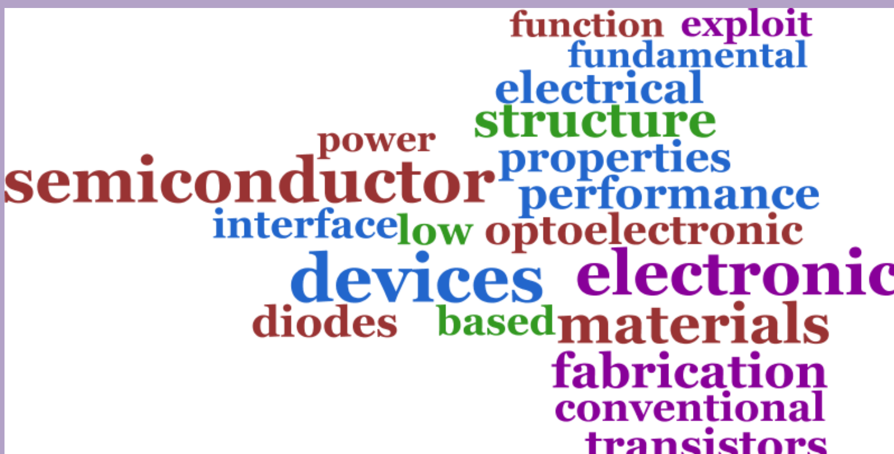
Highest rated

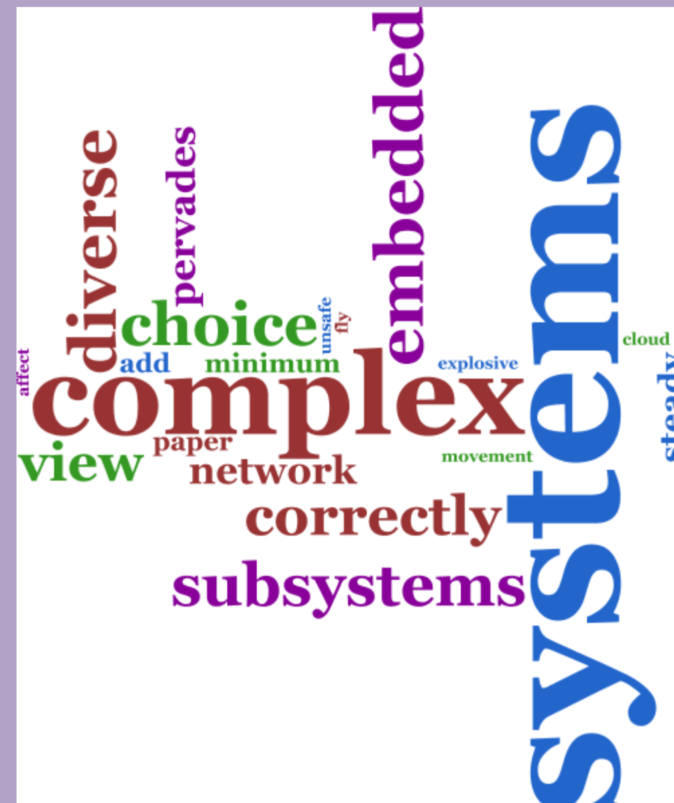


data analytics in personal / entertainment / ubiquitous computing?

600 topics model / ICT grants only / predictions of future funded topic interaction based on Adamic/Adar scores for currently unconnected topics

5th Highest rated





Complex systems research in the semiconductor/optoelectronics systems ?

stop now