# Hidden semantics in text

**Khoat Than**

Hanoi University of Science and Technology

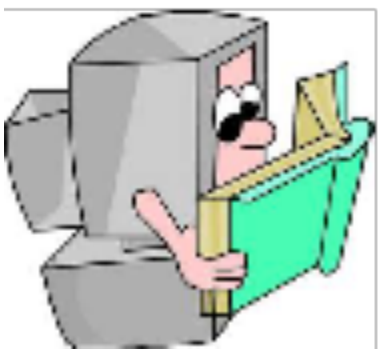
Summer school on Data Mining, Hanoi, 8/2016

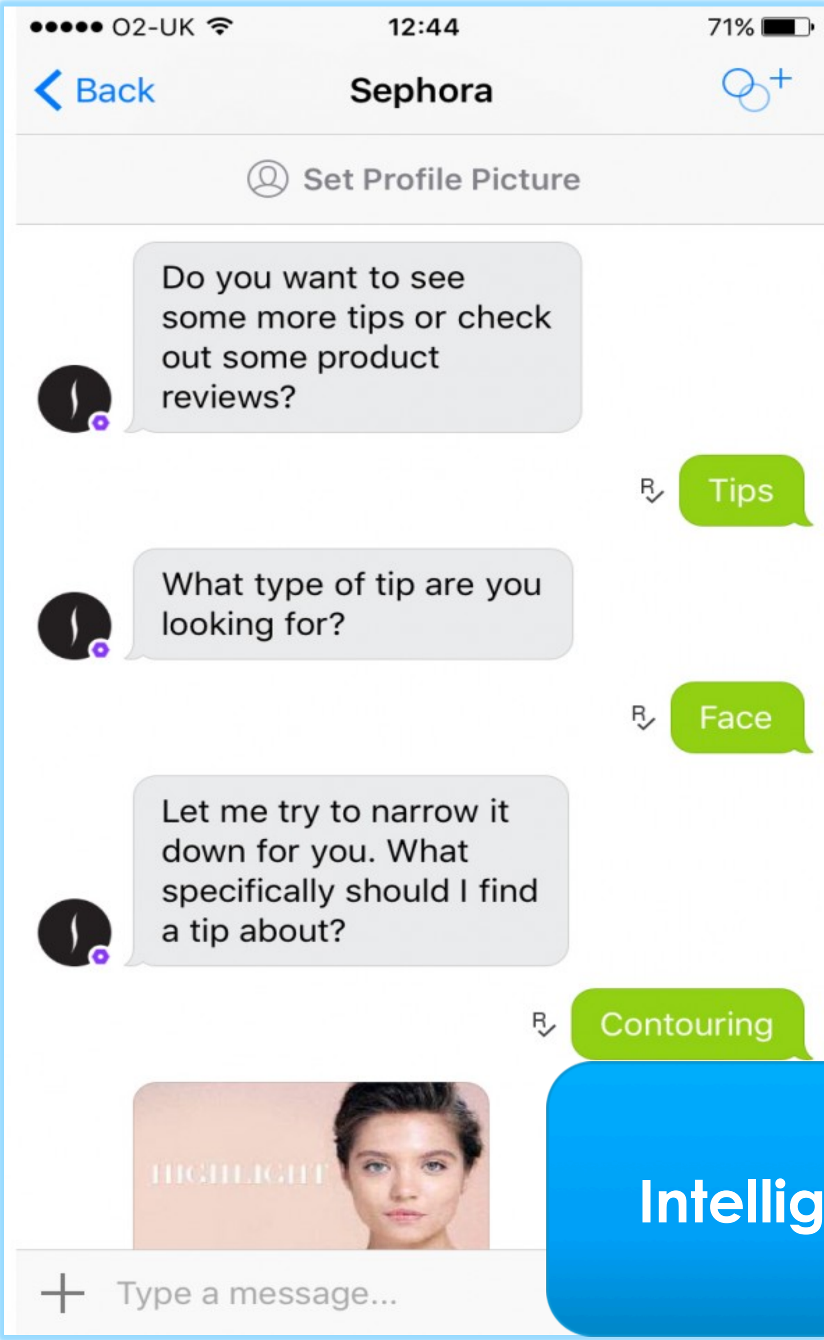
# Contents

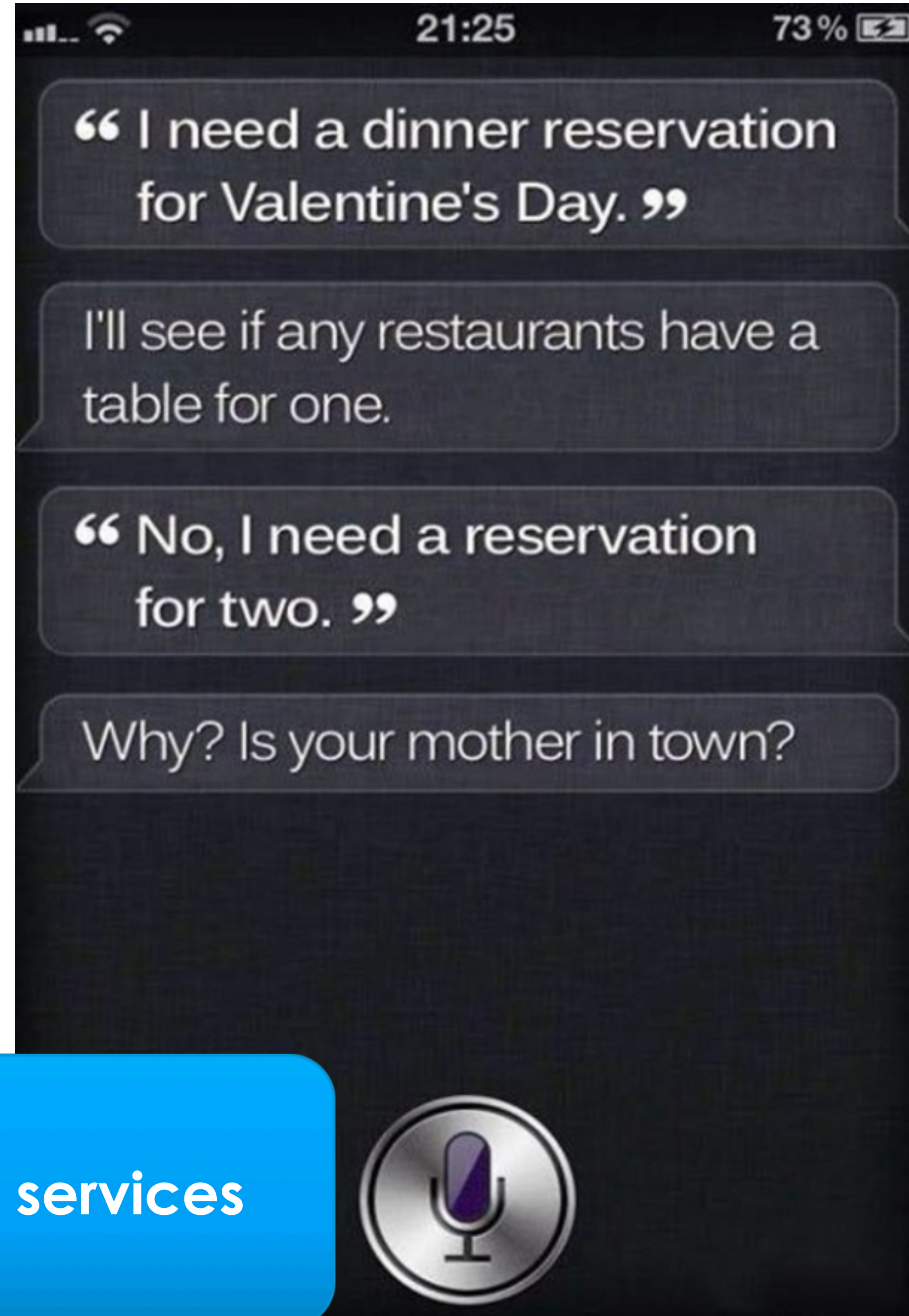
- Open-ended questions
- Hidden semantics: what and why?
- Semantic representation
- Introduction to topic modeling
- Some challenges and lessons learnt

# Open-ended question 1

- Can we help a computer to automatically understand documents and natural languages?






**Intelligent services**

http://www.forbes.com/sites/rachelarthur/2016/03/30/sephora-launches-chatbot-on-messaging-app-kik/

# Open-ended question 2

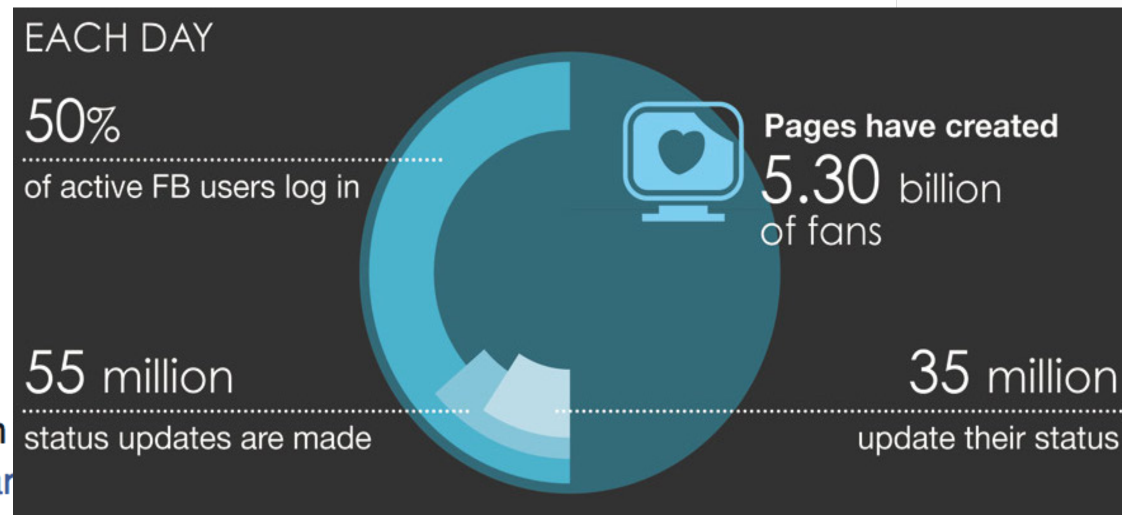
- How to organize, understand, uncover useful knowledge from a huge amount of texts?

# A huge amount of texts

facebook

**Taylor Swift** đã thêm 4 ảnh mới.
4 Tháng 4 lúc 19:52 ·

What an unbelievable run we've had with these memories & all of you. #iHeartAwar




twitter

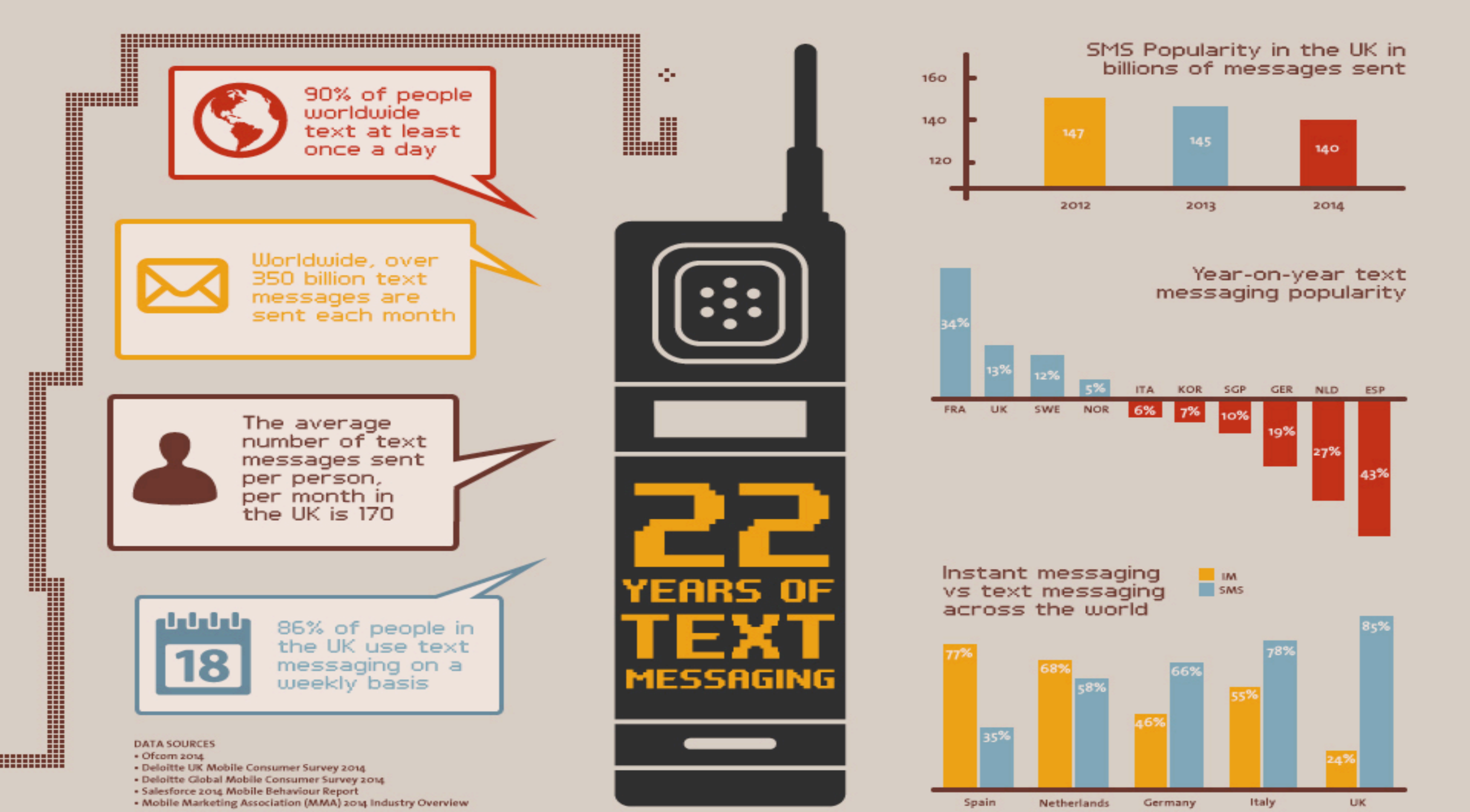
**Basit Alvi** @bpk69 · 6m
Swiss banker whistleblower: CIA behind **Panama Papers** cnb.cx/1WpVjgK
View summary

**Violamagic** @TrautCarol · 6m
Why The **Panama Papers** Scandal Is About Cheating School Children
educationopportunitynetwork.org/why-the-panama...
View summary

**7,174** Tweets sent in 1 second

**862,696** Tweets since opening this page
**0:02:00** seconds ago

# A huge amount of SMS




http://www.openuniversity.edu/news/news/2014-text-messaging-usage-statistics

# One way to answer

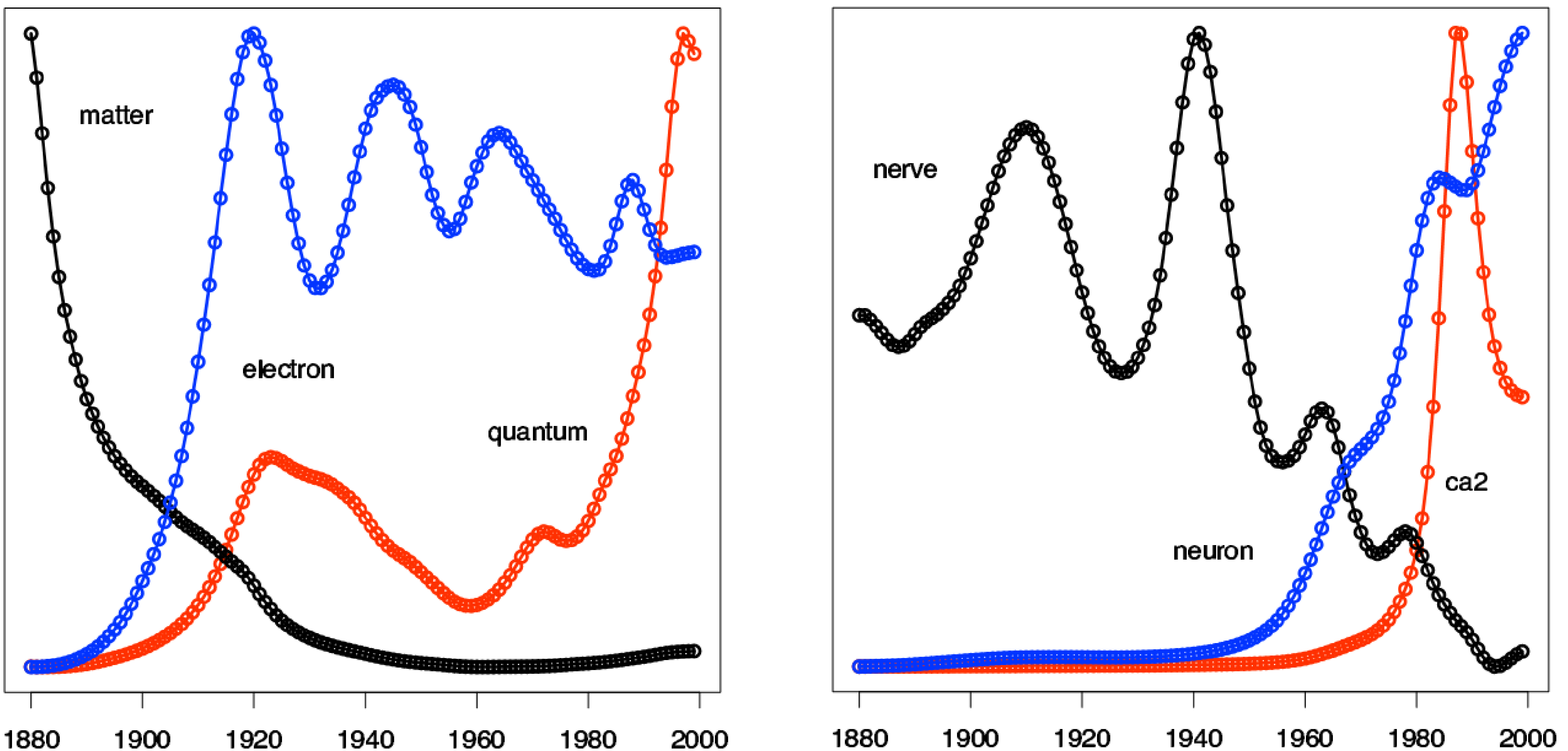
- Help your computer to understand from the very basics
  - Word meanings
  - Word collocation/interaction
  - Sentences, paragraphs
  - …
- to complicated things
  - Themes of paragraphs/documents
  - Opinions, emotions
  - …
- Those are **hidden semantics**

# Hidden semantics

## What and why?

# Hidden semantics: what?

- Evolution/trend of interests over time



# Hidden semantics: what?

- Meanings of pictures

SKY WATER TREE MOUNTAIN PEOPLE

SCOTLAND WATER FLOWER HILLS TREE

SKY WATER BUILDING PEOPLE WATER

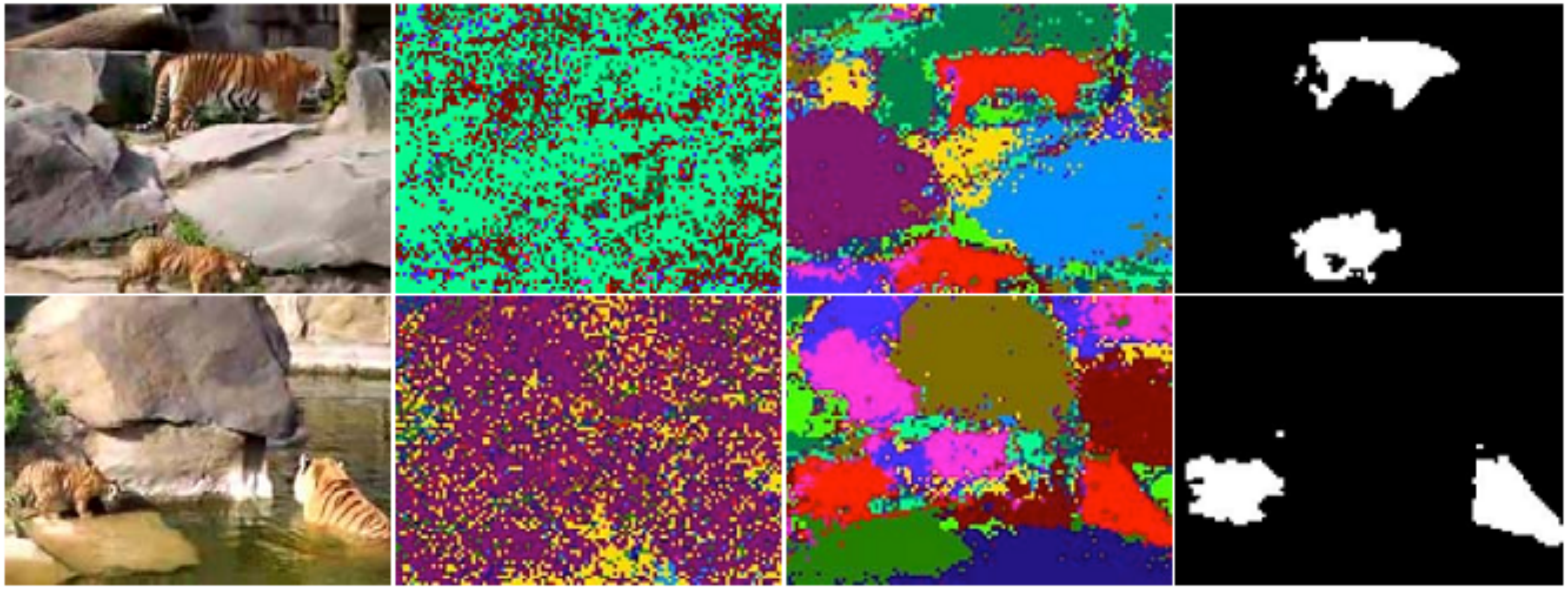
FISH WATER OCEAN TREE CORAL

PEOPLE MARKET PATTERN TEXTILE DISPLAY

BIRDS NEST TREE BRANCH LEAVES

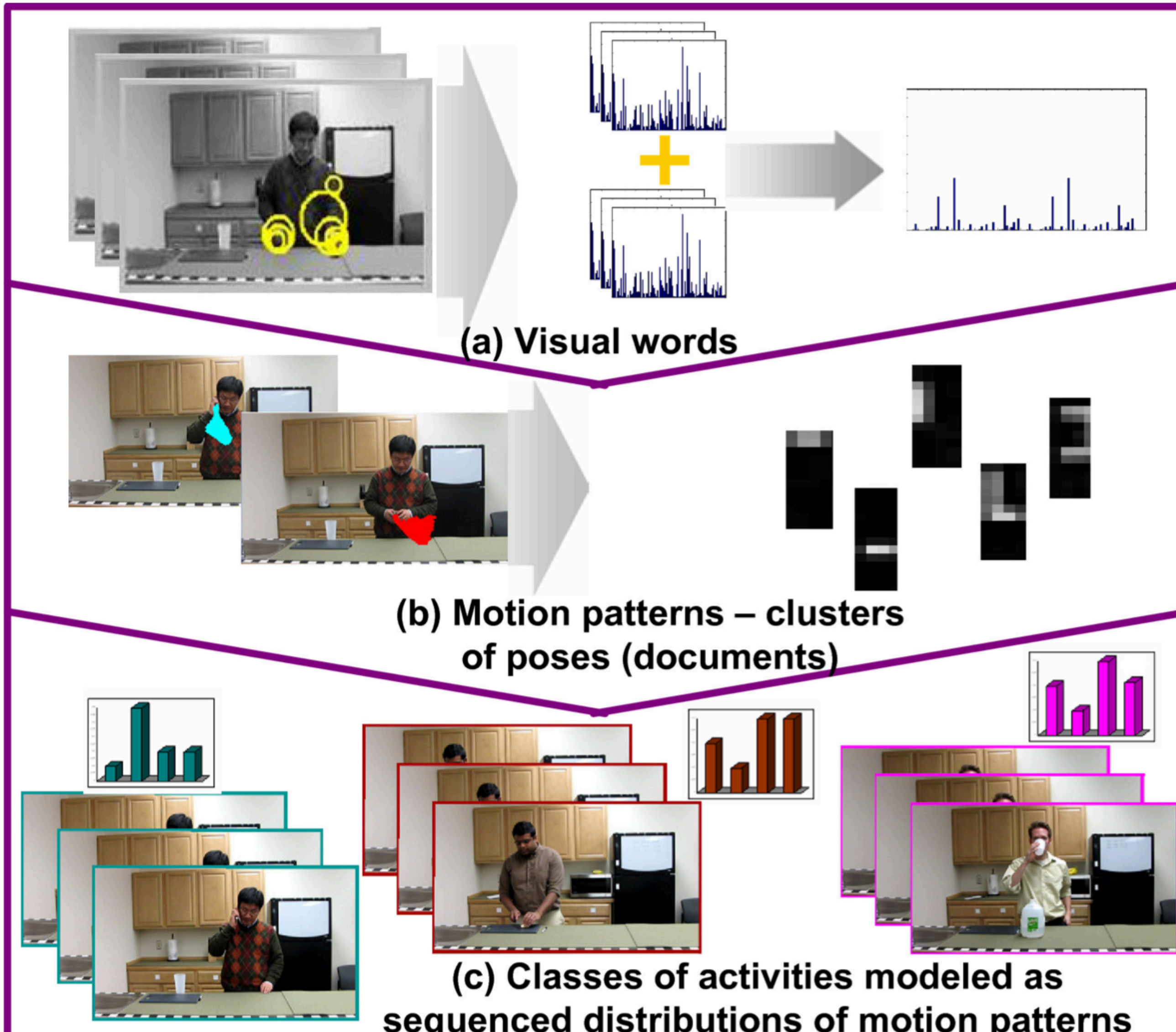
# Hidden semantics: what?

- Objects in pictures



# Hidden semantics: what?

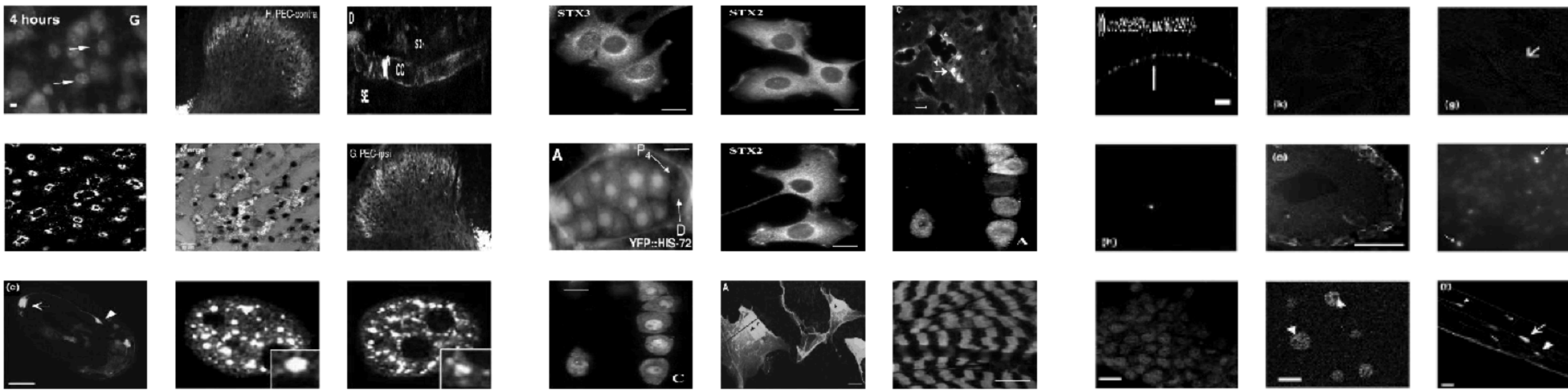
- Activities



# Hidden semantics: what?

- Contents of medical images

EBA Method | Cell structure | Tumors




**EBA Method**

| Top Words | Top proteins |
|---|---|
| mei | Actin |
| sex | VP26 |
| filaments | Some |
| vessels | RT |
| interphase | SCP3 |
| eba | E-H |
| oocytes | 2a |
| injury | PA |
| focal | L4 |
| representation | L1 |

**Cell structure**

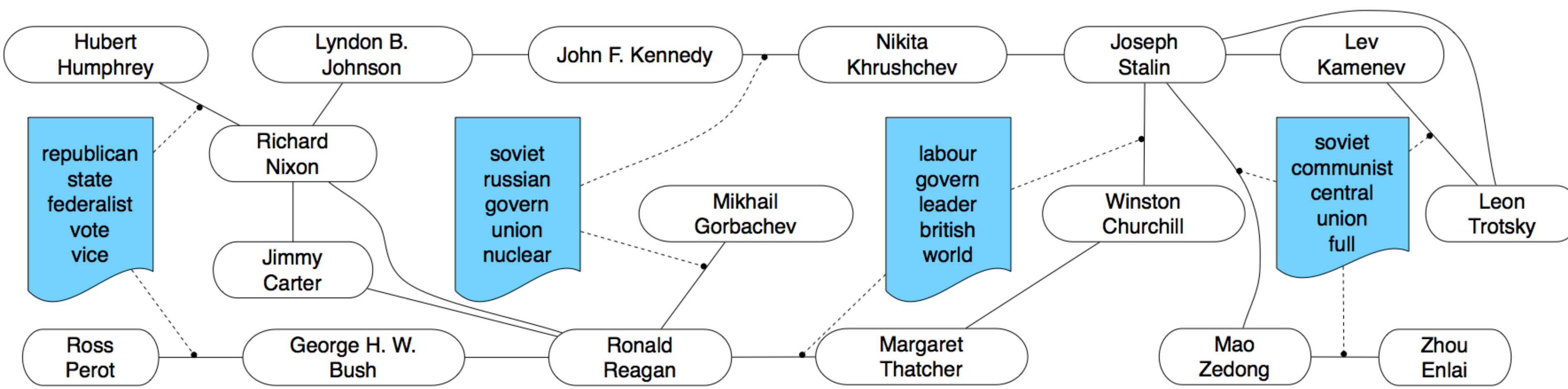
| Top Words | Top proteins |
|---|---|
| embryo | HRP |
| muscle | ZO-1 |
| mitochondrial | Map |
| structure | PC |
| filaments | SSW |
| yfp | P4 |
| epithelial | Df |
| stromule | GFP-MAP4 |
| nestin | Septin2 |
| plastid | SE |

**Tumors**

| Top Words | Top proteins |
|---|---|
| unc | Cx43 |
| dcx | DNA |
| hbl | L1 |
| actin | RNA |
| cbp | caspase-3 |
| optical | E15 |
| tumor | AM |
| murine | Ki-67 |
| signals | DAPI |
| positive | 1 protein |

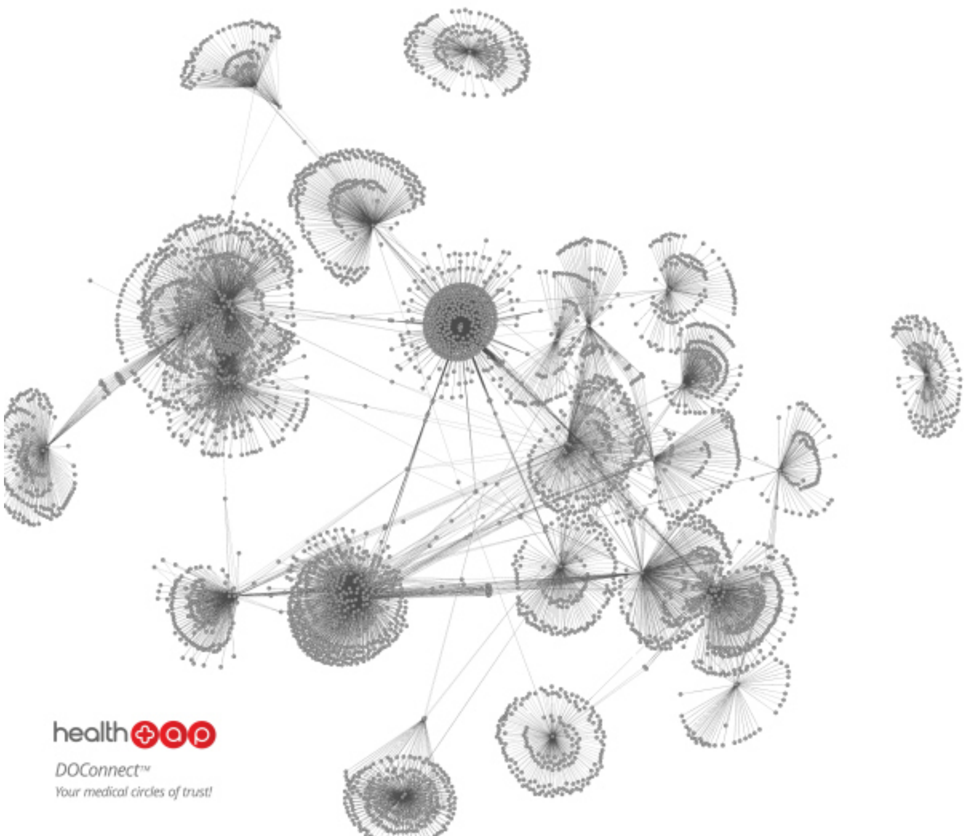
# Hidden semantics: what?

- Interactions of entities



# Hidden semantics: what?

- Communities in social networks

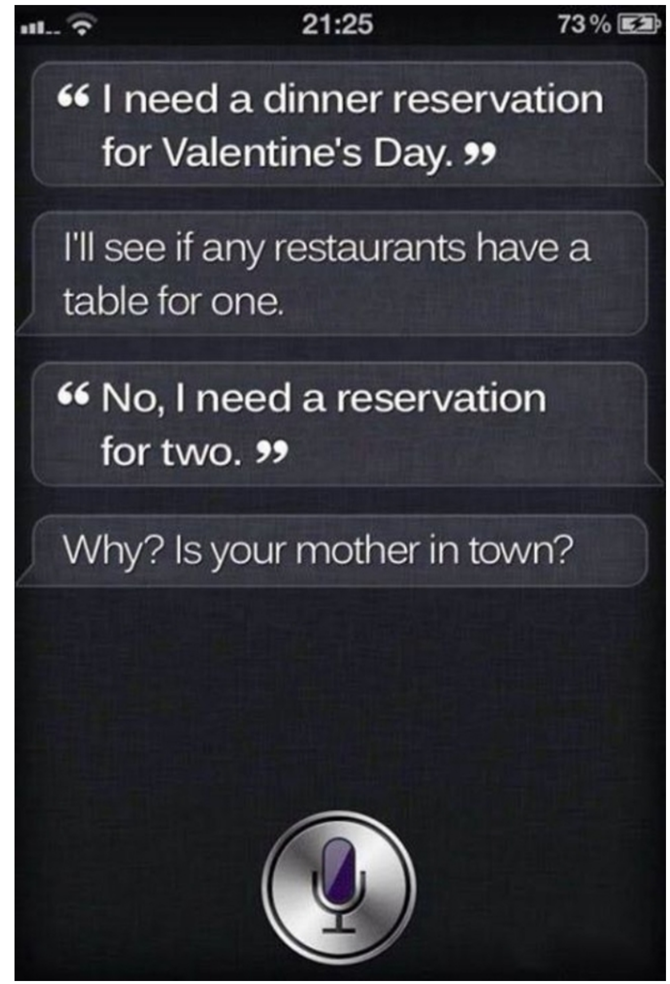


# Hidden semantics: why hard?

- Help your computer understand what is “hard”?
  - Not easy? But what is “easy”?
  - Firm, Solid?
  - Enthusiastic?

→Ambiguity problem
(a word has many different senses)

- Usage styles of languages
  - Slangs, teenage languages
  - Evolvement over time
- Hidden themes are intricately mixed with other structures such as syntax



# Semantics

## Representation & learning

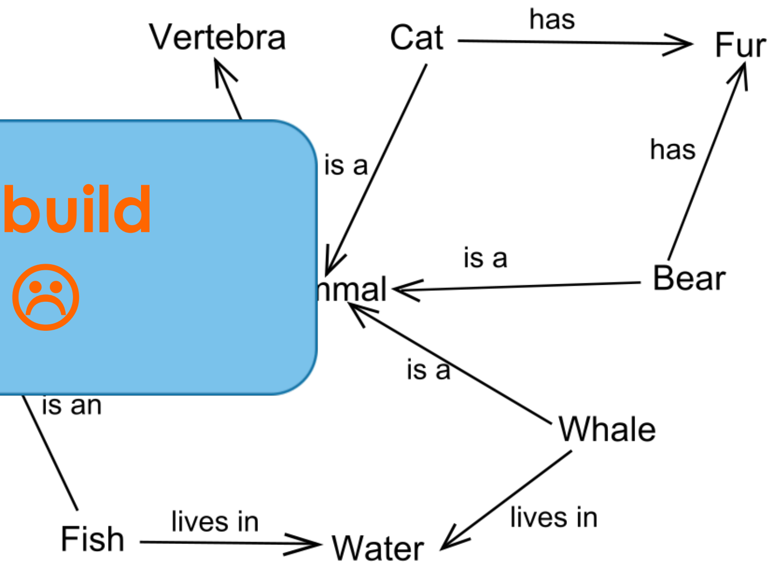
# Semantic representation

- Need a *computational form* to represent knowledge to help a computer to
  - Store knowledge
  - Learn knowledge
  - Make inference

# Some representation approaches

- Classical approaches [Schubert, AAAI 2015]
  - First order logics, Description logics
  - Semantic networks, frames
  - Ontology
  - …

**We have to build manually ☹**




A semantic network (source: wikipedia)

# Some representation approaches

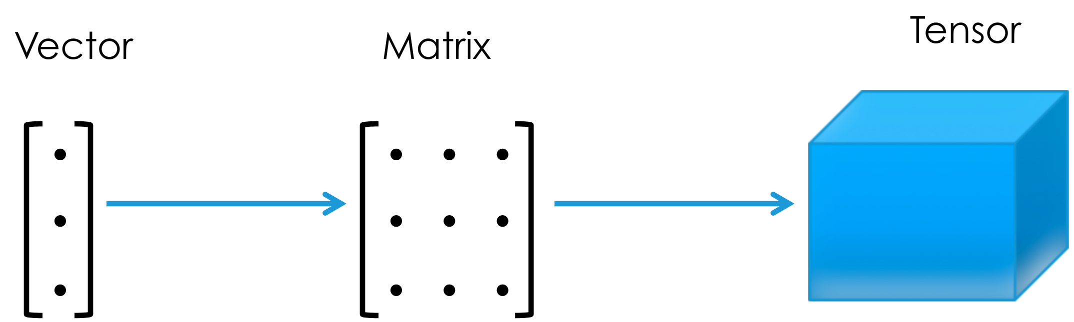
- Machine-learning approaches
  - Topic models [Blei, CACM 2012; Blei et al., JMLR 2003]
  - Deep neural networks
    [LeCun et al., Nature 2015; Collobert et al., JMLR 2011]

**Semantics can be learned automatically from data ☺**

- They tries to learn representation for very basic units, such as words, phrases,…
- Then more complicated forms of semantics can be learned from text collections.

# Learnable representations (1)

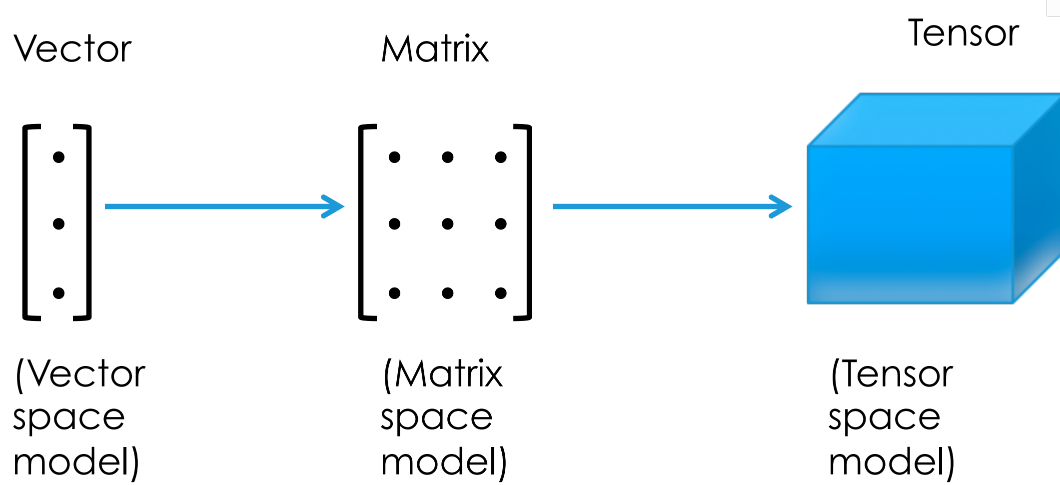
- Different algebraic forms have been used:
  - Vector [Salton et al., CACM 1975]
  - Matrix
  - Tensor



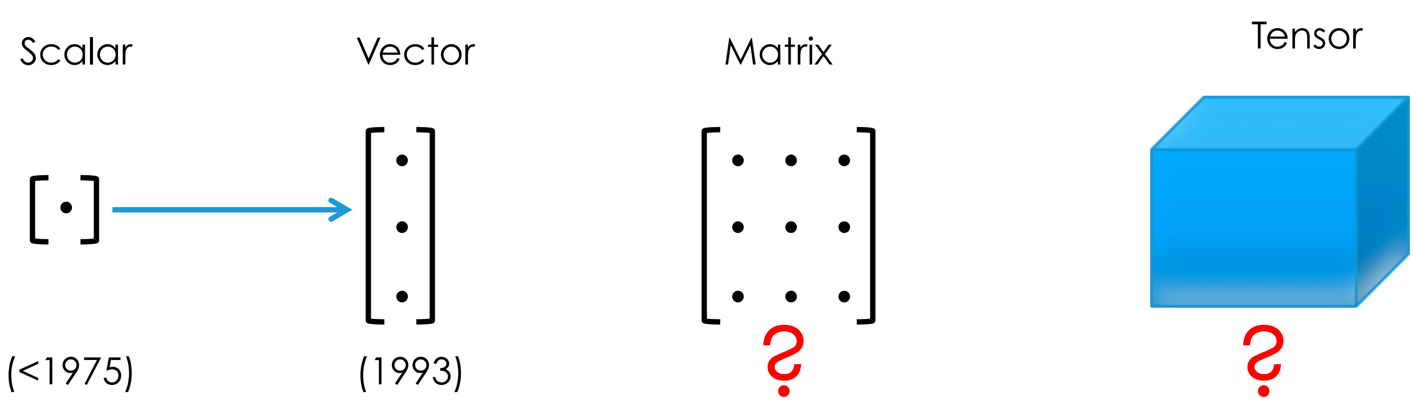

- Finer and finer levels of text are considered
  - A document is represented as a vector [Salton et al., CACM 1975]
  - A paragraph is represented as a vector [Le & Mikolov, ICML 2014]
  - A sentence is represented as a vector [Le & Mikolov, ICML 2014]
  - A phrase is represented as a vector [Mikolov et al., NIPS 2013]
  - A word is represented as a vector [Schütze, NIPS 1993]

# Learnable representations (2)

- More and more complicated tools are used:
  - A document:

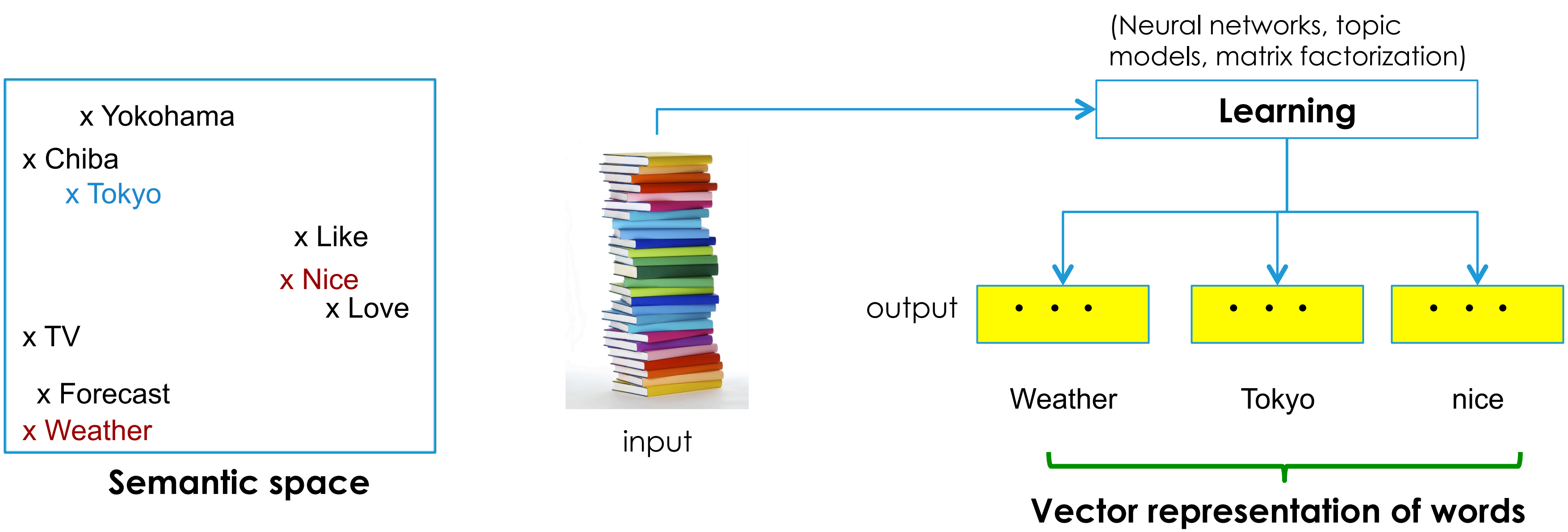



  - A word:



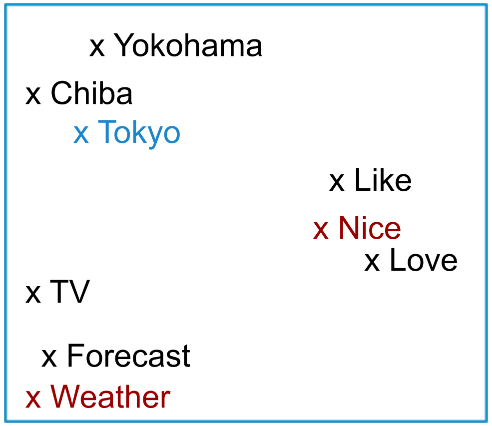
# Word representation

- **Input**: sequences of words (or text collection, or corpus)
  - E.g.: The weather in Tokyo today is nice
- **Output**: k-dimensional vectors, each for a word



# After learning

- Many semantic tasks can be done using *algebraic operations.*




**Semantic space**

- **Semantic similarity**
  - Between words, e.g.,

$$\boldsymbol{V}_{Queen} \approx \boldsymbol{V}_{King} - \boldsymbol{V}_{Man} + \boldsymbol{V}_{Woman}$$

$$Similarity(\boldsymbol{V}_{like}, \boldsymbol{V}_{love}) = cos(\boldsymbol{V}_{like}, \boldsymbol{V}_{love}) = \frac{\boldsymbol{V}_{like} \cdot \boldsymbol{V}_{love}}{||\boldsymbol{V}_{like}|| \cdot ||\boldsymbol{V}_{love}||}$$

  - Between documents, e.g.,

$$Similarity(\boldsymbol{d}_1, \boldsymbol{d}_2) = cos(\boldsymbol{d}_1, \boldsymbol{d}_2) = \frac{\boldsymbol{d}_1 \cdot \boldsymbol{d}_2}{||\boldsymbol{d}_1|| \cdot ||\boldsymbol{d}_2||}$$

- Classification, prediction, inference can be done efficiently

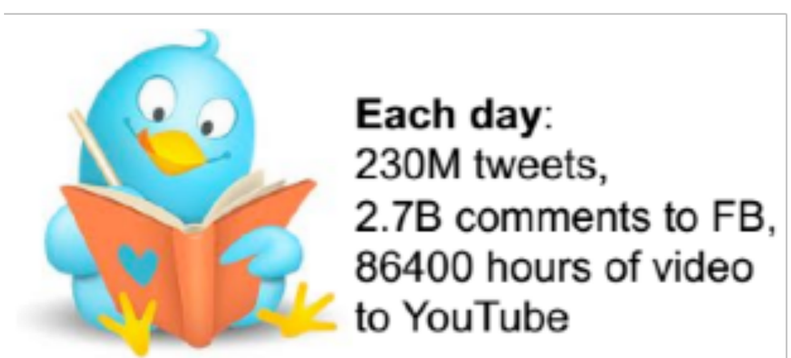
# Fundamental

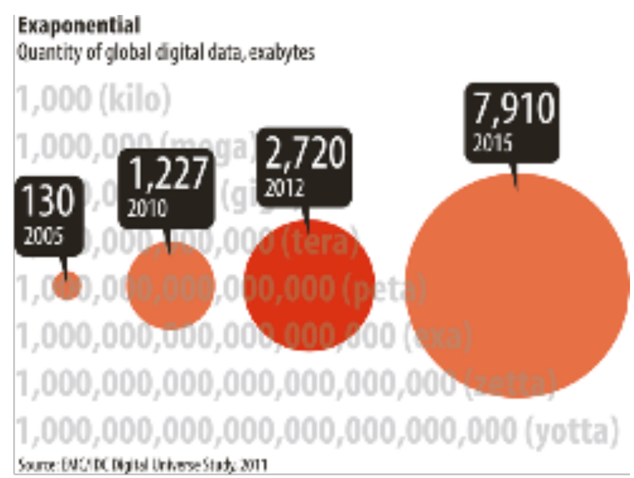
# of

# Topic Modeling

# Topic modeling (1)

- One of the main ways to automatically understand the meanings of text.
- Efficient tools to organize, understand, uncover useful knowledge from a huge amount of data.
- Efficient tools to discover the hidden semantics/structures in data.

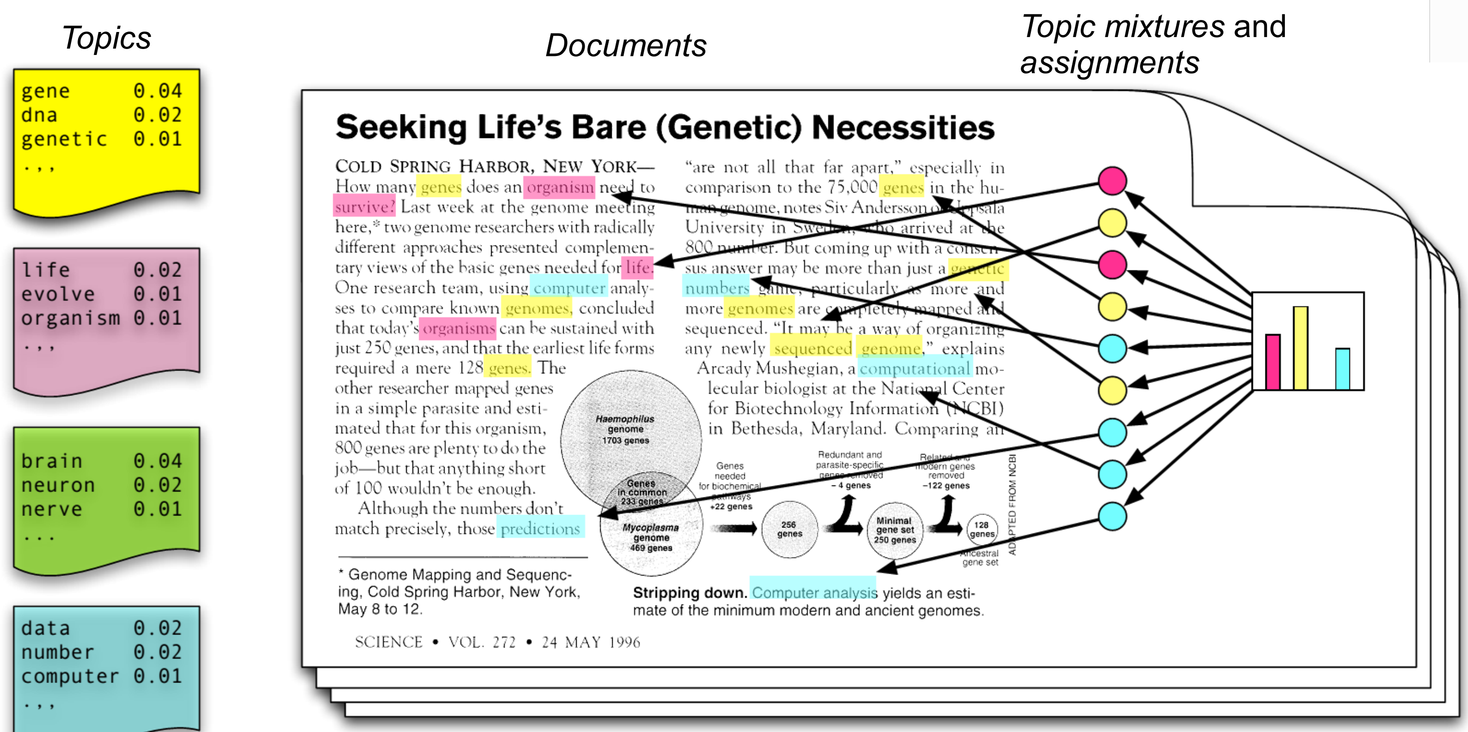


# Topic modeling (2)

- Provides efficient tools for **text analysis** [DiMaggio et al., Poetics, 2013]
  - **Explicit**
    (enable interpretations & exploration of a large text collection, and test hypotheses)
  - **Automated**
    (the algorithms can do with a minimum human intervention)
  - **Inductive**
    (enable researchers to discover the hidden structures of data before imposing their priors on the analysis)
  - **Recognize the rationality of meaning**
    (the meaning of a term might vary across different domains)

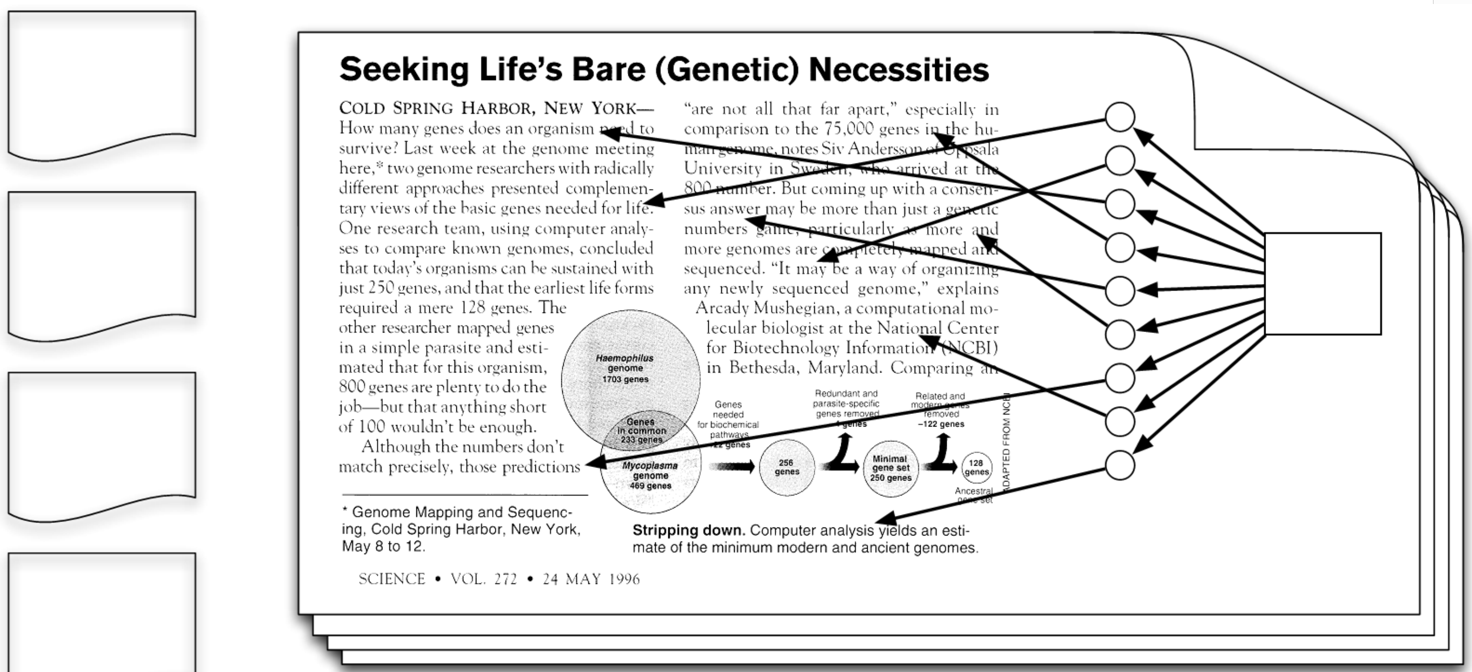
# Topic models: some concepts (1)




David Blei, 2012.

- **Topic**: is a set of semantically related words
- **Document:** is a mixture of few topics [Blei et al., JMLR 2003]
- **Topic mixture:** shows proportions of topics in a document

# Topic models: some concepts (2)

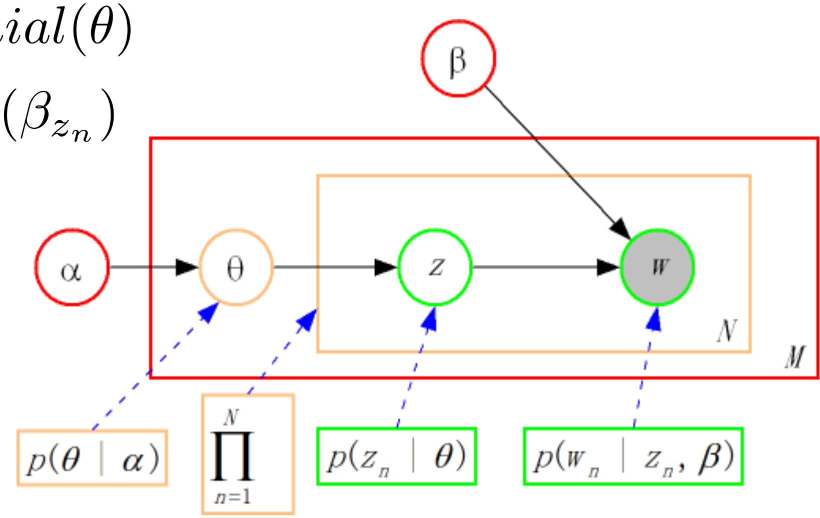



David Blei, 2012.

- In reality, we only **observe the documents**.
- The other structures (topics, mixtures, ...) are **hidden**.
- Those structures compose a **Topic Model**.

# Topic models: LDA

- Latent Dirichlet allocation (LDA) [Blei et al., JMLR 2003] is the most famous topic model.
  - LDA assumes a corpus to be composed from K topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$
- Each document is generated by
  - First choose a topic mixture $\theta \sim Dirichlet(\alpha)$
  - For the n^th word in the document
    - Choose topic index $z_n \sim Multinomial(\theta)$
    - Generate word $w_n \sim Multinomial(\beta_{z_n})$



# LDA

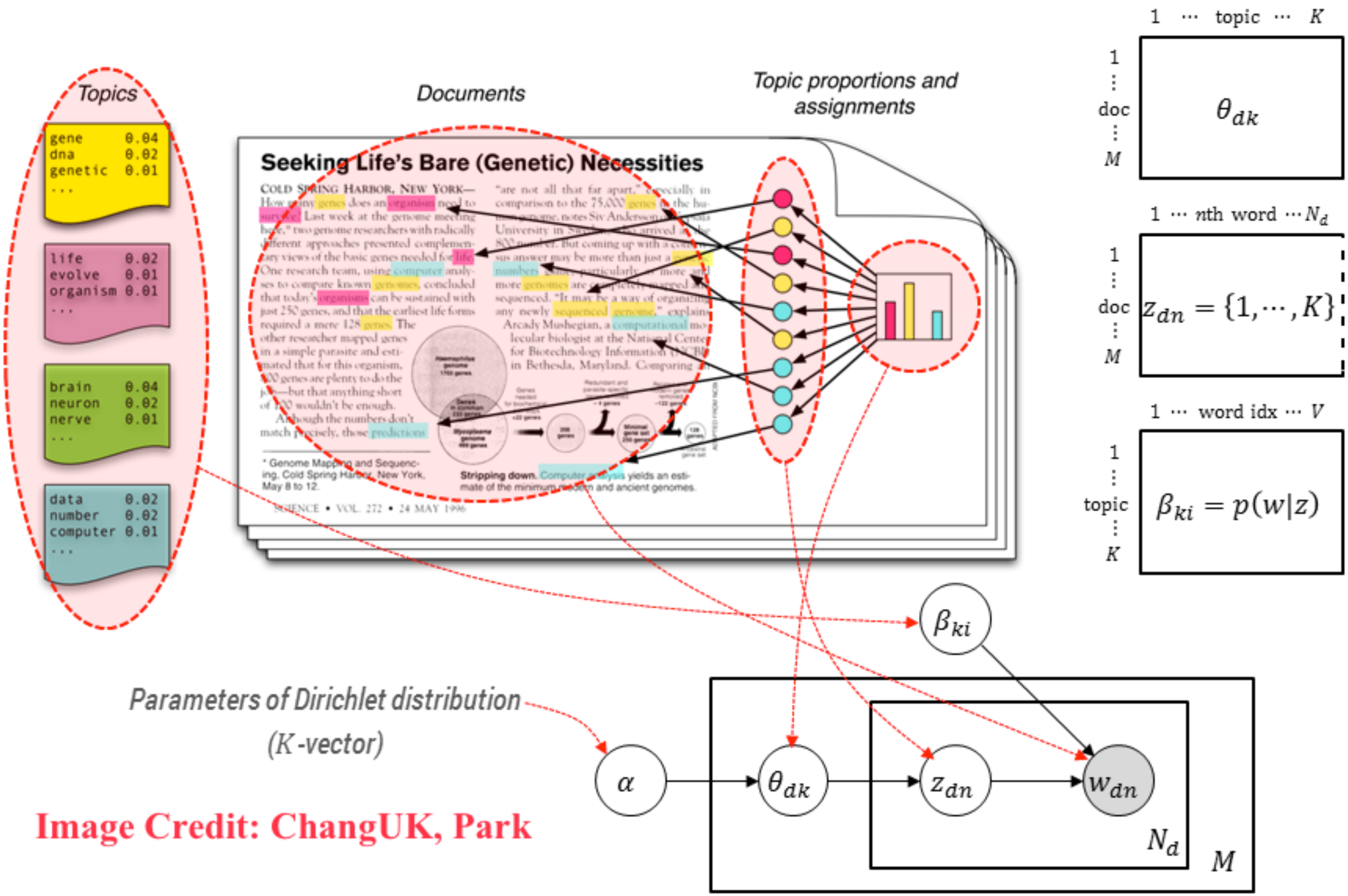
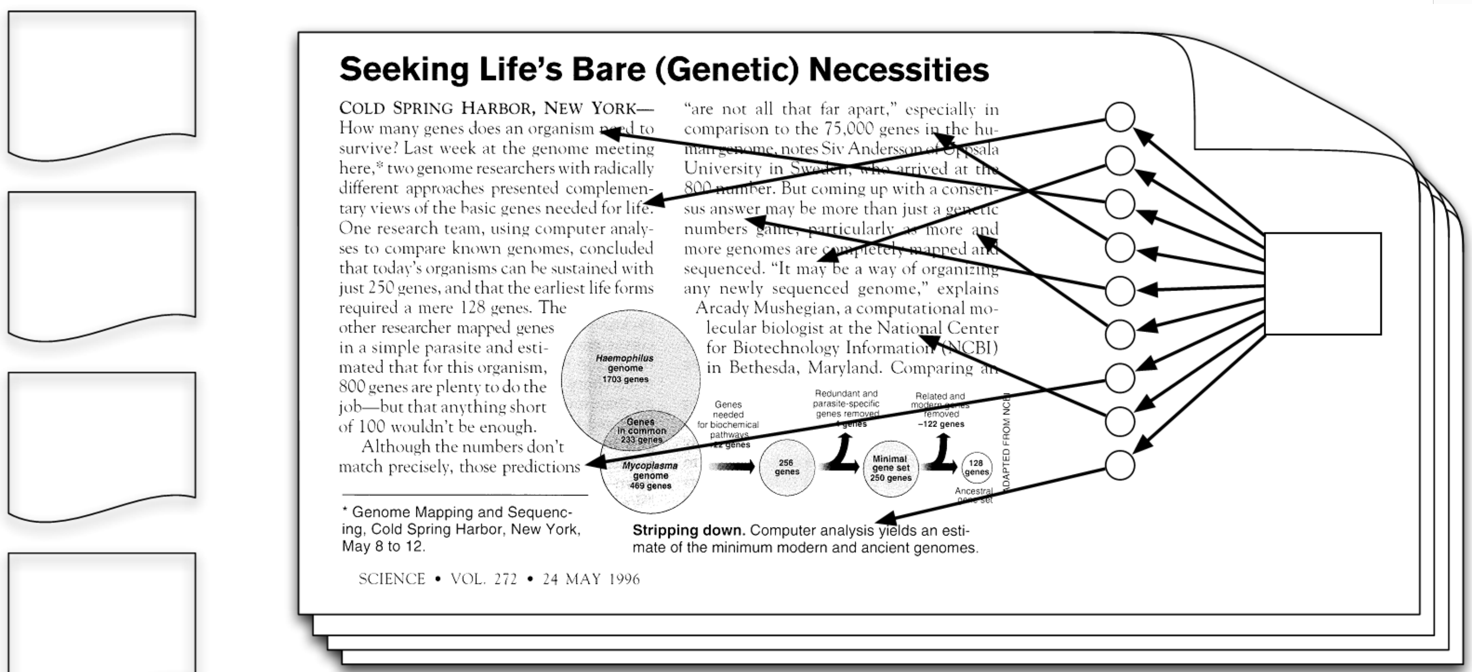



**Image Credit: ChangUK, Park**

# Topic models: learning

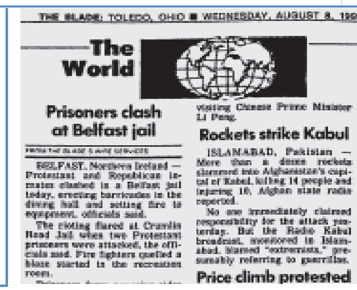



David Blei, 2012.

- Given a corpus, our aim is to *infer the hidden variables*,
- e.g., topics, relations, interactions, ... $P(\boldsymbol{\beta}, \theta, z | corpus)?$

# Topic models: posterior inference

**Rockets strike Kabul** -- AP, August 8, 1990.
More than a dozen rockets slammed into Afghanistan's capital of Kabul today, killing 14 people and injuring 10, Afghan state radio reported. No one immediately claimed repsonsibility for the attack. But the Radio Kabul broadcast, monitored in Islamabad, blamed ``extremists,'' presumably referring to U.S.-backed guerrillas headquartered in Pakistan. Moslem insurgents have been fighting for more than a decade to topple Afghanistan's Communist-style government. In the past year, hundreds of people have died and thousands more injured in rocket assaults on the Afghan capital.


THE BLADE: TOLEDO, OHIO ■ WEDNESDAY, AUGUST 8, 1990

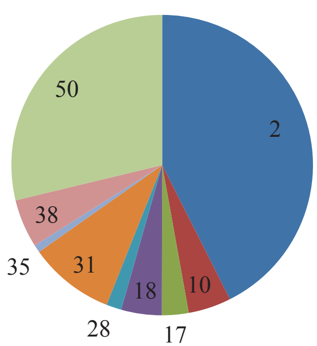
The World

Prisoners clash at Belfast jail

Rockets strike Kabul

Price climb protested

How much topics contribute to the news?

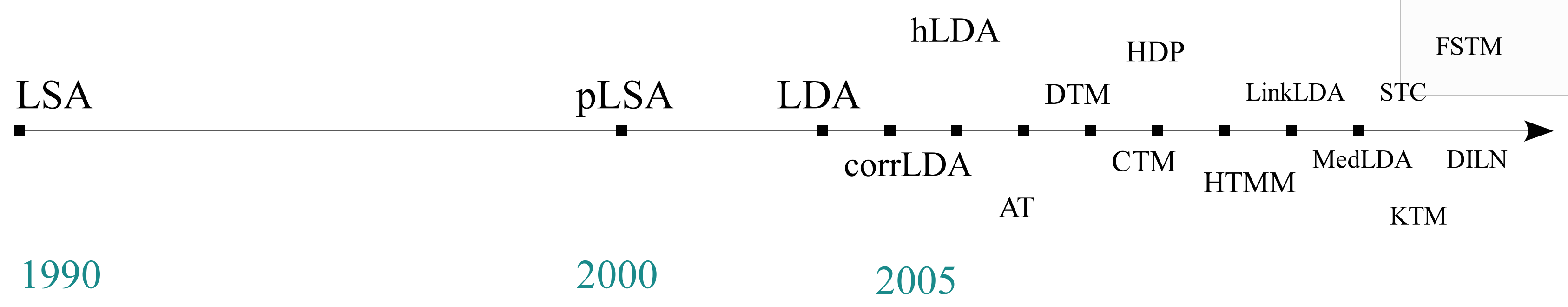



Some topics previously learned from a collection of news

| 2 | 10 | 18 | 31 | 38 | 40 | 42 | 50 |
|---|---|---|---|---|---|---|---|
| police | sinhalese | contra | fire | shuttle | gorbachev | index | beirut |
| students | tamil | sandinistas | winds | nasa | soviet | stock | hezbollah |
| palestinians | iranian | chamorro | firefighters | space | republics | yen | lebanon |
| curfew | dam | ortega | mph | launch | politburo | points | aoun |
| sikh | khomeini | rebels | blaze | magellan | yeltsin | market | syrian |
| gaza | sri | sandinista | brush | mars | moscow | shares | militia |
| rangoon | cemetery | aid | homes | spacecraft | tass | trading | lebanese |
| moslem | accord | nicaragua | acres | telescope | party | dow | amal |
| israeli | wppss | managua | water | venus | treaty | unchanged | troops |
| militants | guerrillas | ceasefire | weather | astronauts | grigoryants | volume | wounded |

- Infer the hidden variables for a given document, e.g.,
  - What topics/objects appear in? $P(\theta, z | w, \beta)$?
  - What are their contributions? $P(\theta | w, \beta)$? $P(z | w, \beta)$?

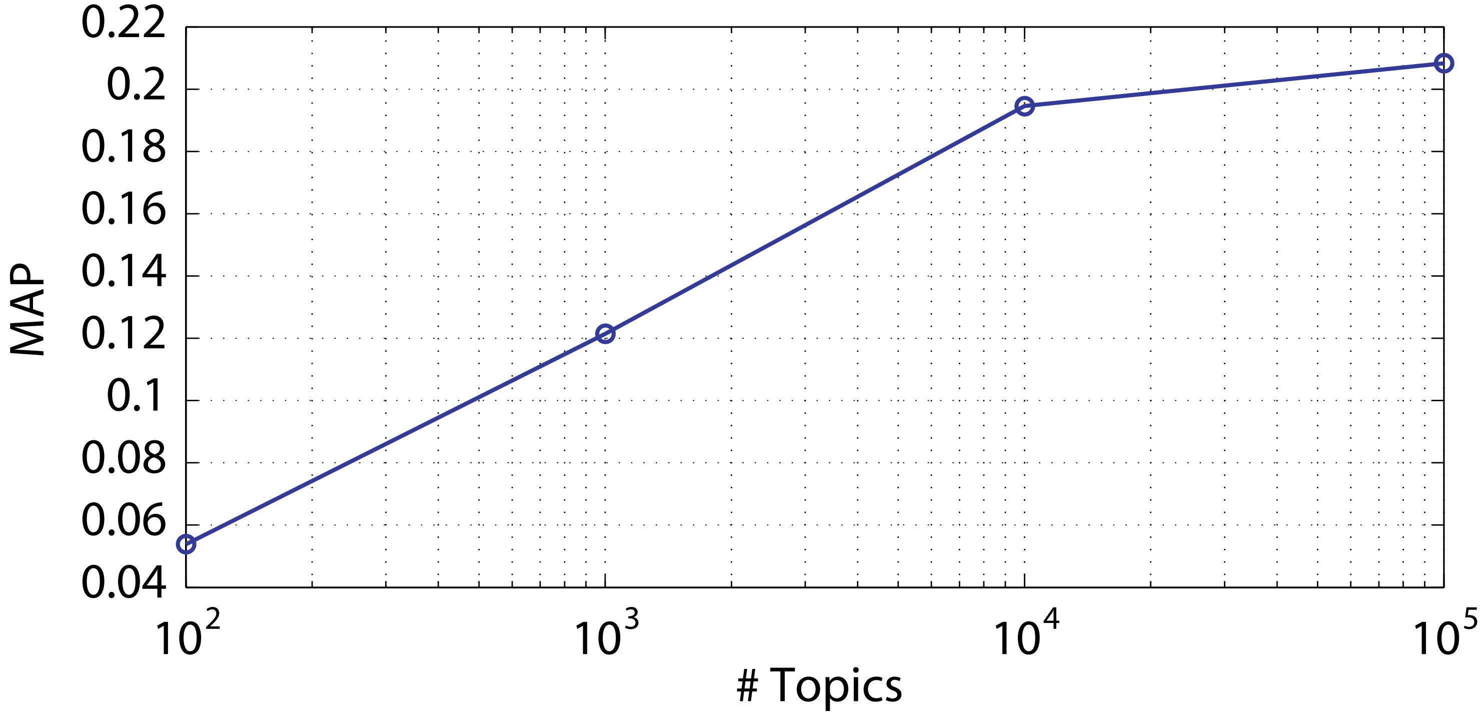
# Recent trends in topic modeling




- *Large scale learning:* learn models from huge corpora (e.g., 100 millions of documents).
- *Sparse modeling:* respect the sparseness nature of texts.
- *Nonparametric models:* automatically grow the model size.
- *Theoretical foundation:* provide guarantees for learning and posterior inference.
- *Incorporating meta-data:* encode meta-data into a model.

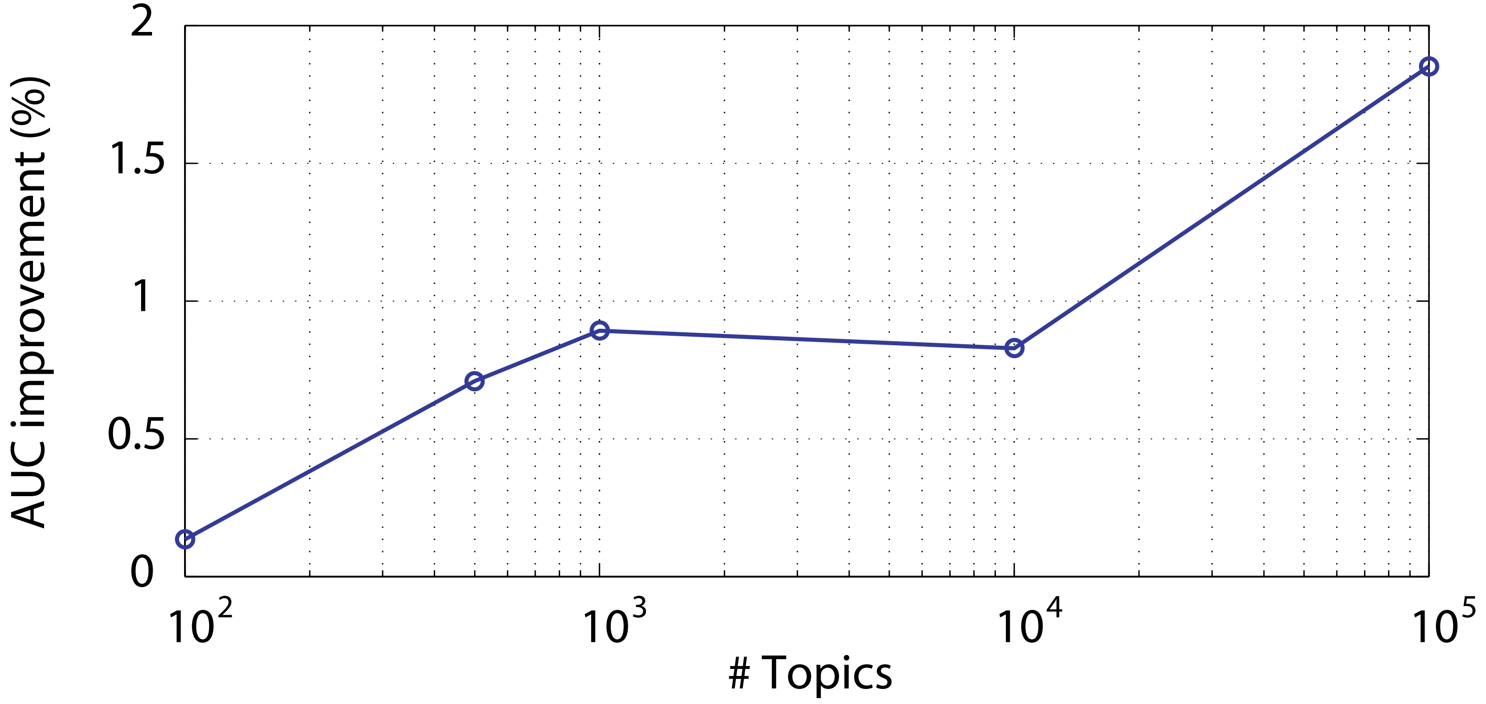
# Recent applications (1)

- Boosting performance of Search engines over the baseline [Wang et al., ACM TIST 2014]



# Recent applications (2)

- Boosting performance of Online advertisement over the baseline [Wang et al., ACM TIST 2014]



# Some challenges

## Lessons learnt and Our solutions

# Challenges: first

- *Can we develop a **fast inference method** that has provably **theoretical guarantees** on quality?*
- Inference on each data instance:
  - What topics appear in a document?
  - What are they talking about?
  - What animals appear in a picture?
- Vital role in many probabilistic models:
  - Enable us to design fast algorithms for massive/stream data.
  - Ensure high confidence and reliability when using topic models in practices
- But: inference is often intractable (NP-hard) [Sontag & Roy, NIPS 2011]

# Challenges: second

- *How can we learn a **big topic model** from big data?*
- Big model:
  - billions of variables/parameters
  - Which might not fit in the memory of a supercomputer
- Many applications lead to this problem:
  - Exploration of a century of literature
  - Exploration of online forums/networks
  - Analyzing political opinions
  - Tracking objects in videos
- But largely unexplored in the literature.

# Challenges: third

- *Can we develop **methods with provable guarantees** on quality for handling **streaming/dynamic** text collections?*
- Many practical applications:
  - Analyzing political opinions in online forums
  - Analyzing behaviors & interests of online users
  - Identifying entities and temporal structures from news.
- But: existing methods often lack a theoretical guarantee on quality.

# Lessons: learnability

- In theory:
  - A model can be recovered exactly if the number of documents is sufficiently large ☺
    [Anandkumar et al., NIPS 2012; Arora et al., FOCS 2012; Tang et al., ICML 2014]
  - It is impossible to guarantee learnability of a model when having few documents ☹
  - A model cannot be learned from very short texts ☹
    [Arora et al., ICML 2016; Tang et al., ICML 2014]
- In practice: [Tang et al., ICML 2014]
  - Once there are sufficently many documents, further increasing the number may not significantly improve the performance.
  - The document length should be long, but need not too long.
  - A model performs well when the topics are well separated.

# Lessons: practical effectiveness

- Collapsed Gibbs sampling (CGS):
  - Most efficient
  - Better than VB and BP in large-scale applications [Wang et al., TIST 2014]
- Belief propagation (BP):
  - Memory-intensive
- Variational Bayes (VB): [Jiang et al., PAKDD 2015]
  - Often slow
  - And inaccurate
- Collapsed variational Bayes (CVB0): [Foulds et al., KDD 2013]
  - Most efficient and accurate

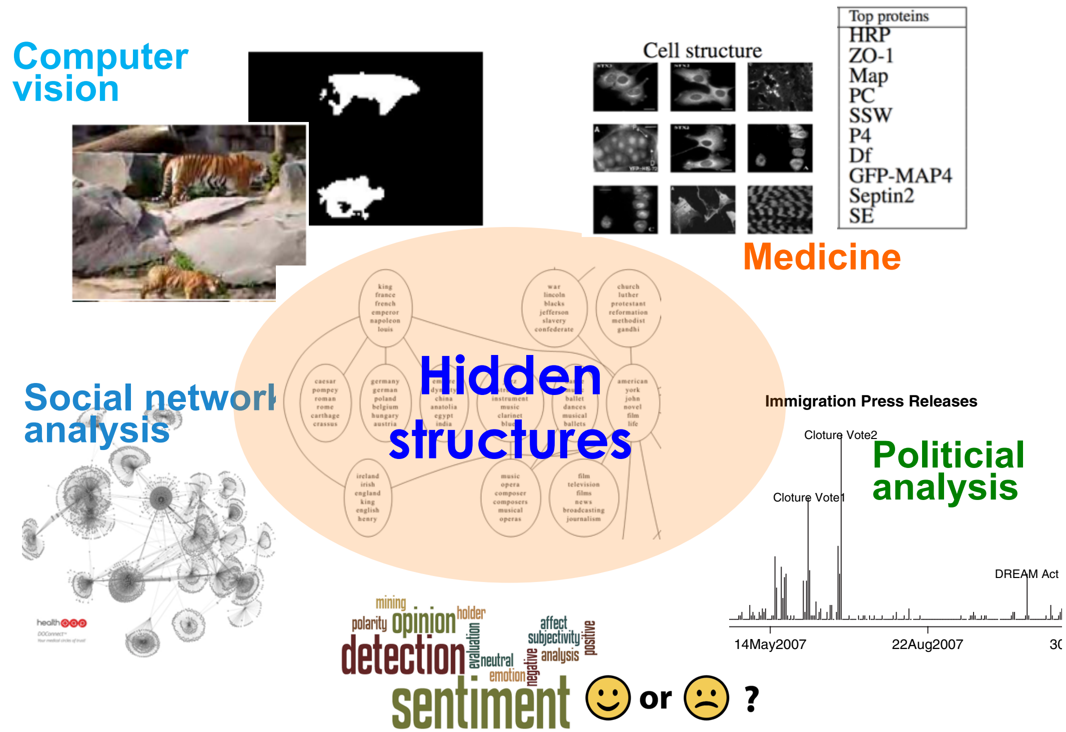
# Lessons: posterior inference

- Inference for individual texts:
  - *Variational method (VB)* [Blei et al., JMLR 2003]
  - *Collapsed VB (CVB)* [Teh et al., NIPS 2007]
  - *CVB0* [Asuncion et al., UAI 2009]
  - *Gibbs sampling* [Griffiths & Steyver, PNAS 2004]
  - *OPE* [Than & Doan, 2015]
- It is often intractable in theory [Sontag & Roy, NIPS 2011].
- But it might be tractable in practice
  [Than & Doan, ACML 2014; Arora et al., ICML 2016]
- OPE is a fast algorithm that has provable guarantees on quality.

# Our works

- Develop models & methods that help us to infer hidden structures from big/streaming data

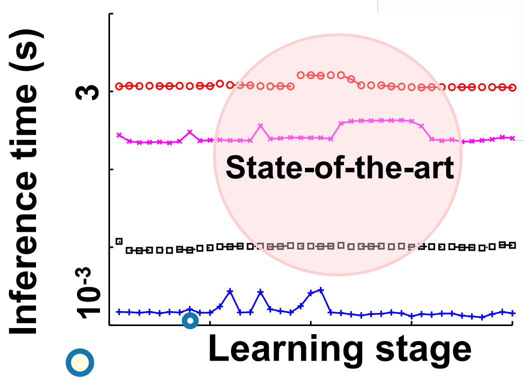
- Many applications



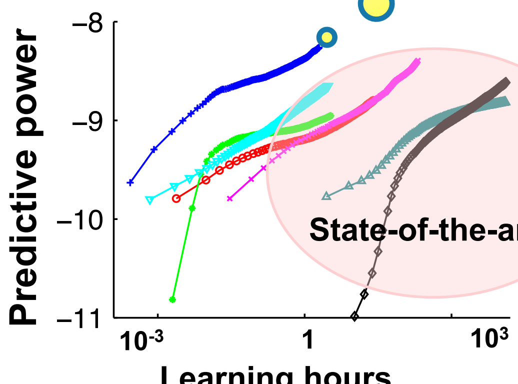

- Our related projects: NAFOSTED (VN), AFOSR (US)

# Some recent results

- Some achievements
  - Inference for individual texts with a theoretical guarantee of fast convergence
    → 5-100 times faster
  - Stochastic learning for streams with far less training documents, yet much better performance
    → better predictiveness, 20-1000 times faster



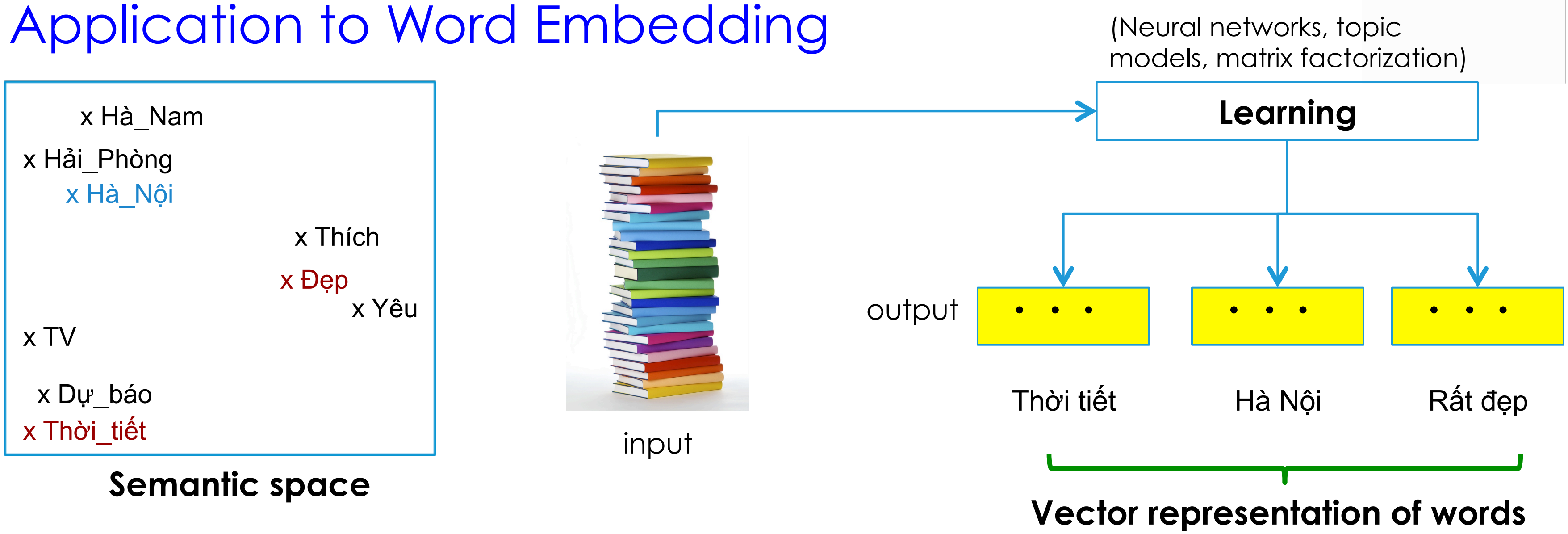

Orders of magnitude faster




Orders of magnitude faster

# Some recent results

- Application to Word Embedding





- *5-15% improvement in classification accuracy,* by combination of
  - Manifold learning
  - Sparse codings
  - Topic models

# References


- Anandkumar, Anima, et al. "A spectral algorithm for latent dirichlet allocation." In *NIPS*. 2012.
- Arora, Sanjeev, Rong Ge, and Ankur Moitra. "Learning topic models--going beyond SVD." In *FOCS, 2012*.
- Asuncion A., P. Smyth, and Max Welling. Asynchronous distributed estimation of topic models for document analysis. *Statistical Methodology*, 8(1):3–17, 2011.
- Blei D., Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3(3):993–1022, 2003.
- Broderick T., Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael Jordan. Streaming variational bayes. In *NIPS*, pages 1727–1735, 2013.
- J. Foulds, L. Boyles, C. DuBois, P. Smyth, and Max Welling. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In *KDD*, pages 446–454. ACM, 2013.
- Griffiths T.L. and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228, 2004.
- Hoffman M., David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- Mimno D. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage*, 5(1):3, 2012.
- Smola A. and Shravan Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.
- Sontag D. and Daniel M. Roy. Complexity of inference in latent dirichlet allocation. In *NIPS*, 2011.
- Tang J., Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *ICML*, pages 190–198, 2014.
- Teh Y.W., D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, volume 19, page 1353, 2007.
- WANG, Y., ZHAO, X., SUN, Z., YAN, H., WANG, L., JIN, Z., ... & ZENG, J. Peacock: Learning Long-Tail Topic Features for Industrial Applications. ACM Transactions on Intelligent Systems and Technology, Vol. 9, No. 4, Article 39, 2014.

# References


- Arora, Sanjeev, et al. "Provable algorithms for inference in topic models." *ICML* (2016).
- Bengio, Yoshua, et al. "A neural probabilistic language model." Journal of Machine Learning Research 3.Feb (2003): 1137-1155.
- Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.
- Collobert, Ronan, et al. "Natural language processing (almost) from scratch." Journal of Machine Learning Research 12.Aug (2011): 2493-2537.
- Deerwester, Scott, et al. "Indexing by latent semantic analysis." *Journal of the American Society for Information Science* 41.6 (1990): 391.
- DiMaggio et al., "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding", Poetics 41 (2013): 570-606.
- Harris, Z. "Distributional structure". *Word* 10 (1954): 146–162.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *Nature* 521.7553 (2015): 436-444.
- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.
- Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." *Communications of the ACM* 18.11 (1975): 613-620.
- Than, Khoat, and Tung Doan. "Guaranteed algorithms for inference in topic models." *arXiv preprint arXiv:1512.03308* (2015).
- Than, Khoat, and Tung Doan. "Dual online inference for latent Dirichlet allocation." *ACML*. 2014.
- Schubert, Lenhart K. "Semantic Representation." *AAAI*. 2015.
- Schütze, Hinrich. "Word Space". *Advances in Neural Information Processing Systems 5* (1993). pp. 895–902
- Zeng et al. "A Comparative Study on Parallel LDA Algorithms in MapReduce Framework". In *PAKDD*, 2015.

Thank you