Unsupervised Learning of Hierarchical structures in data using **Co-occurrence Analysis**

Dr. Shailesh Kumar

World – a "Hierarchy of Objects"

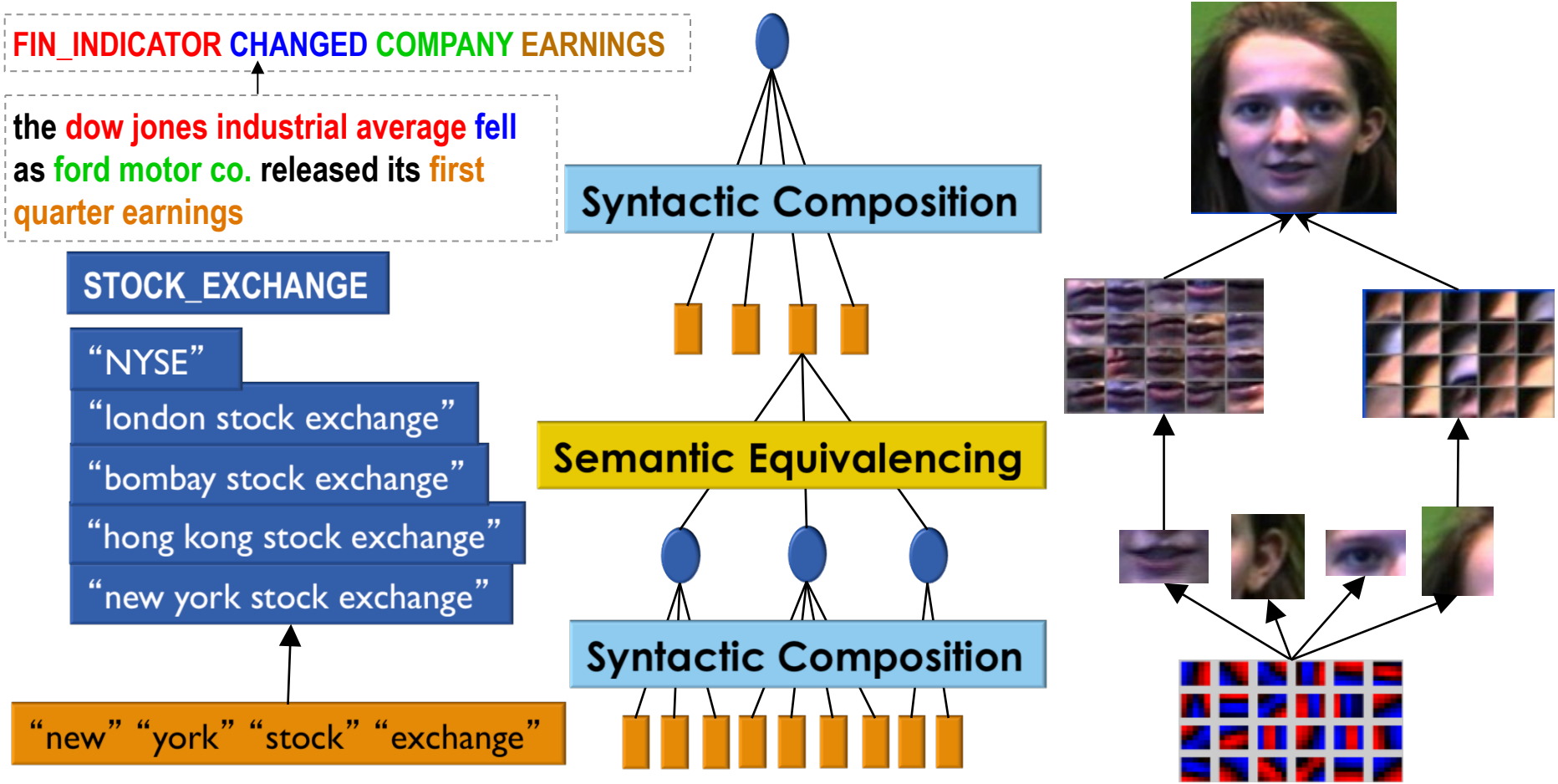
"Entities" at one level of the hierarchy combine in certain "contexts" to form entities at the next level

- **Natural Systems are hierarchical**
  - Atomic Particles → Atoms → Molecules → Crystals → …
  - Cells → Tissues → Organs → Systems → Individuals → Societies
  - Pixels → Edges → Regions → Objects → Images → Videos
- **Man-made systems are also hierarchical**
  - Letters → Words → Phrases → Sentences → Paragraphs → Documents
  - Tokens → Lines → Functions → Modules → Programs → Software
  - Employees → Teams → Groups → Department → Company → Industry

The Quest

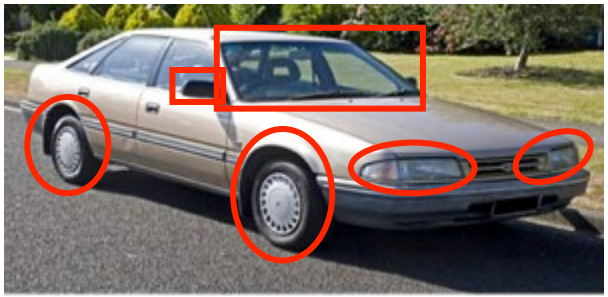
- **Can we learn this hierarchy from lots of data?**
  - Text, Image, Speech, Genetic, Neural, Retail,…
- **What are the "Building blocks" of such a hierarchy?**
  - Domain agnostic, Level agnostic, Data agnostic…
- **How do we adapt these principles to different domains?**
  - Nature of data (sets, sequences, 2-D sequences), etc.
- **Can we discover how the brain builds hierarchies?**
  - Already there is a lot of evidence around this

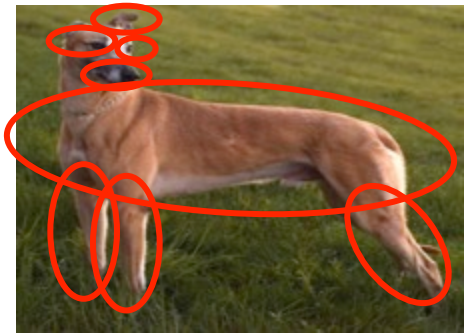
Two building blocks of Learning



Two building blocks of learning

- **Syntactic Composition/Conjunction** – **"is a PART of"**





- **Semantic Equivalence/Disjunction** – **"is a TYPE of"**

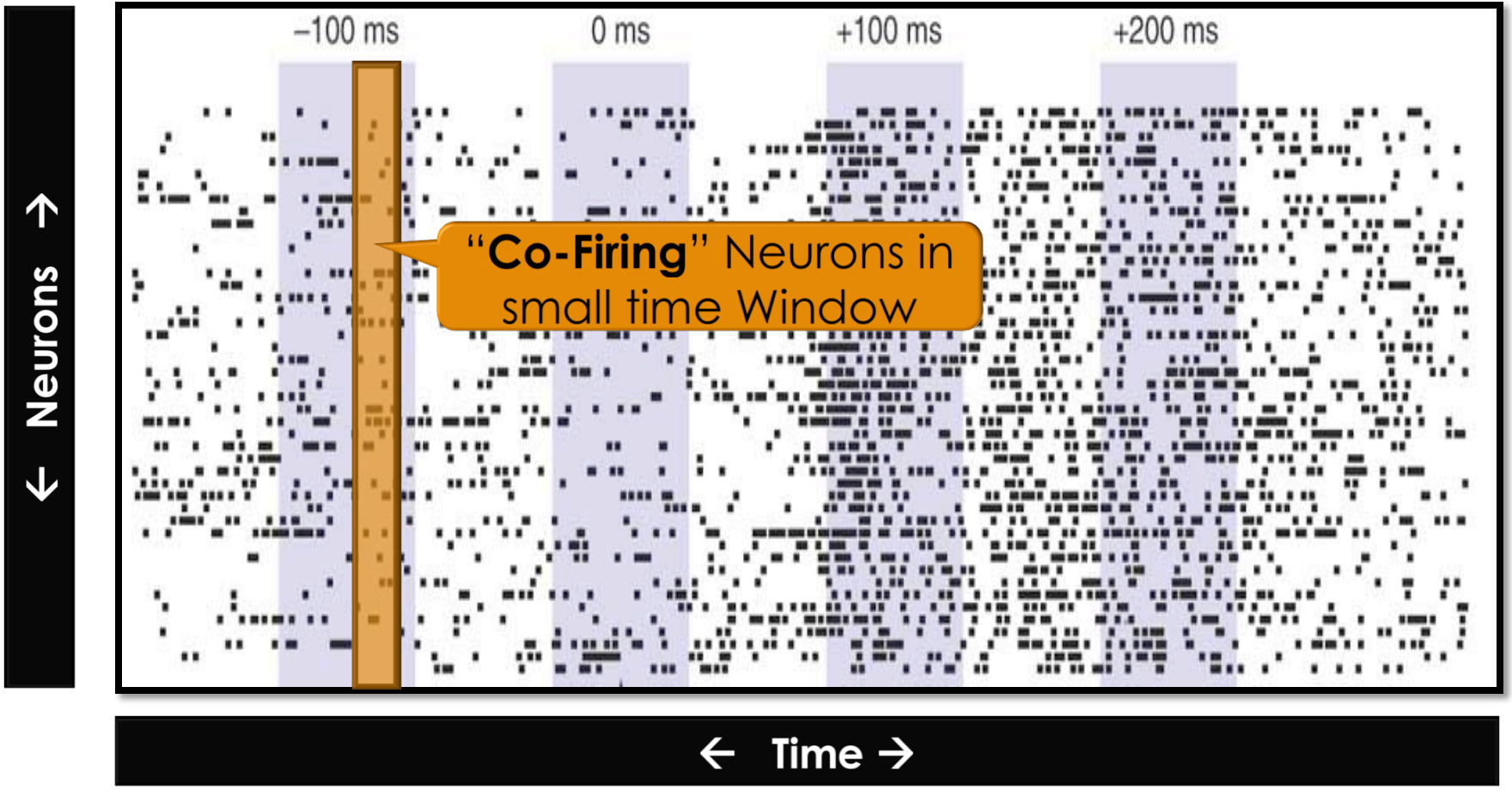
Co-occurrence Analytics

A scalable, unsupervised, hierarchical framework that

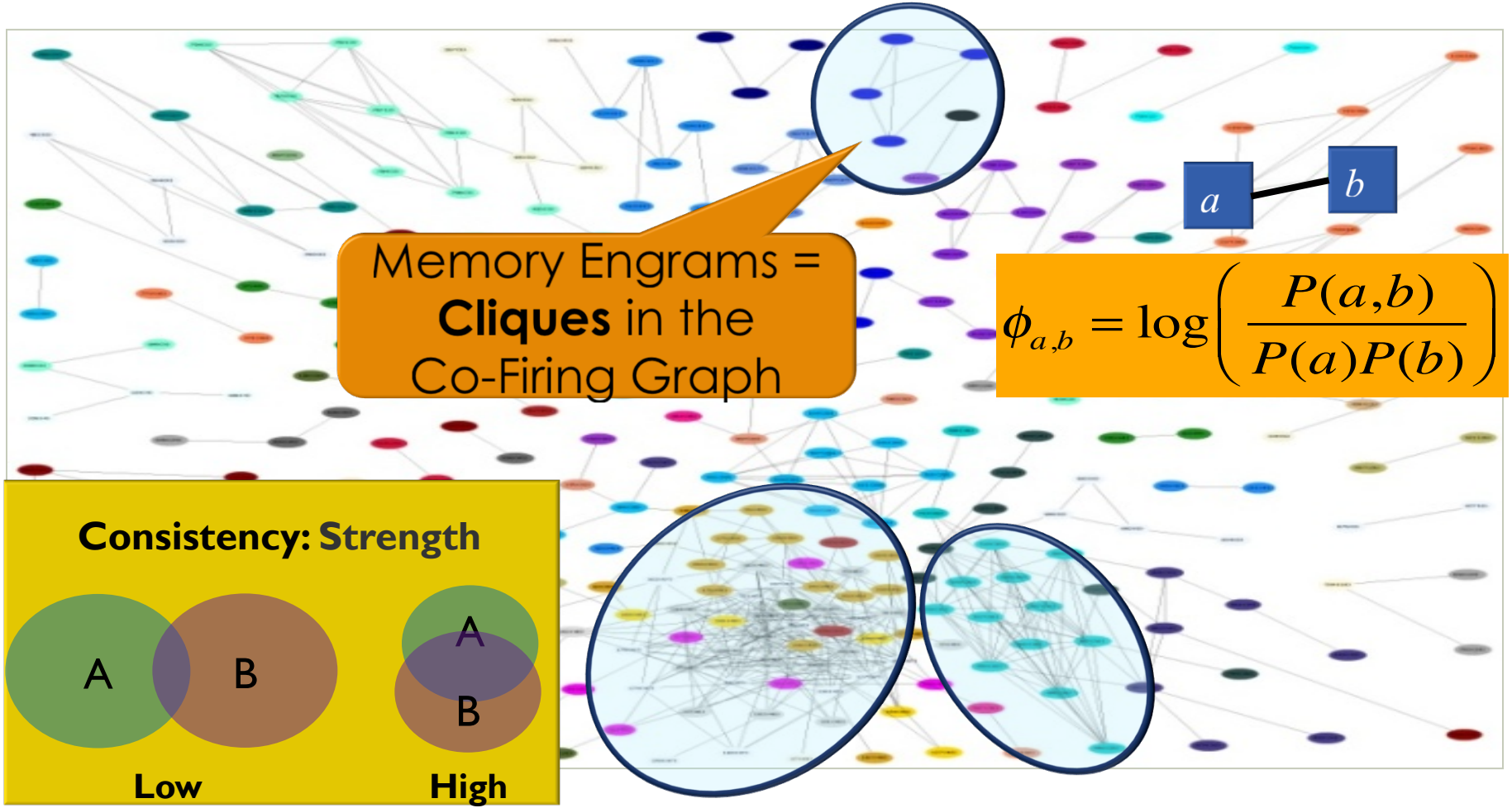
- Analyzes pair-wise **relationships** among **entities**
- Co-occurring in various **contexts**
- To build a **Co-occurrence Graph(s)** in which
- It discovesr **coherent higher order structures**

Discovering Memory Engrams

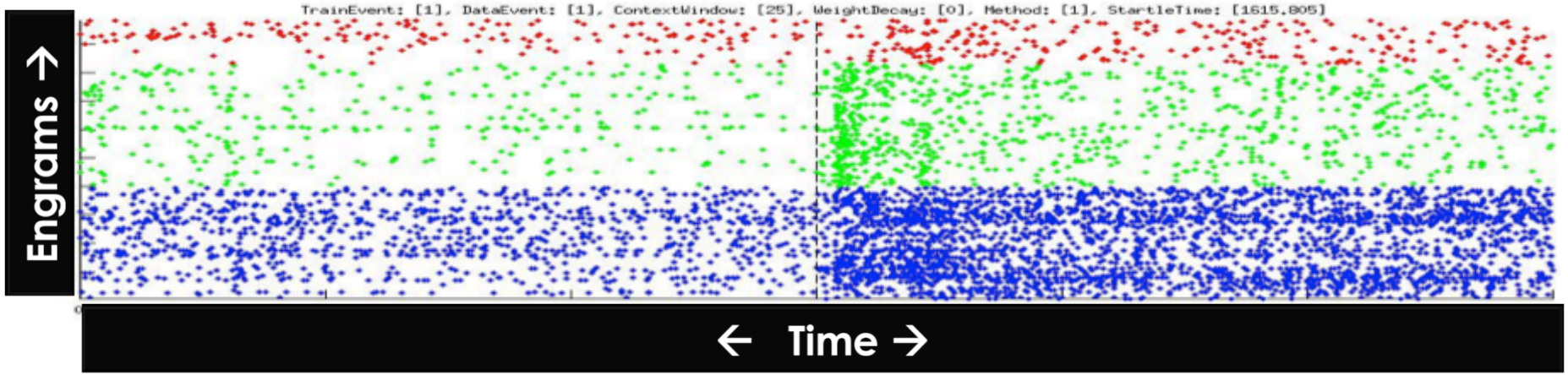
Spiking Neurons – A "Crazy Haystack!"



"Co-Firing" Consistency Graph



Memory engrams = Co-firing neurons



Co-occurrence Analytics - steps

- **Context** – Nature of Co-occurrence
  - E.g. Window of time
- **Co-occurrence** – Definition of Co-occurrence
  - E.g. Co-firing
- **Consistency** – Strength of Co-occurrence
  - E.g. Point-wise Mutual Information
- **Coherence** – Tightness of a group of entities
  - E.g. Modularity or Cliqueness score
- **Community** – Locally optimal groups of entities
  - E.g. Memory engrams

Logical Item-Set Mining

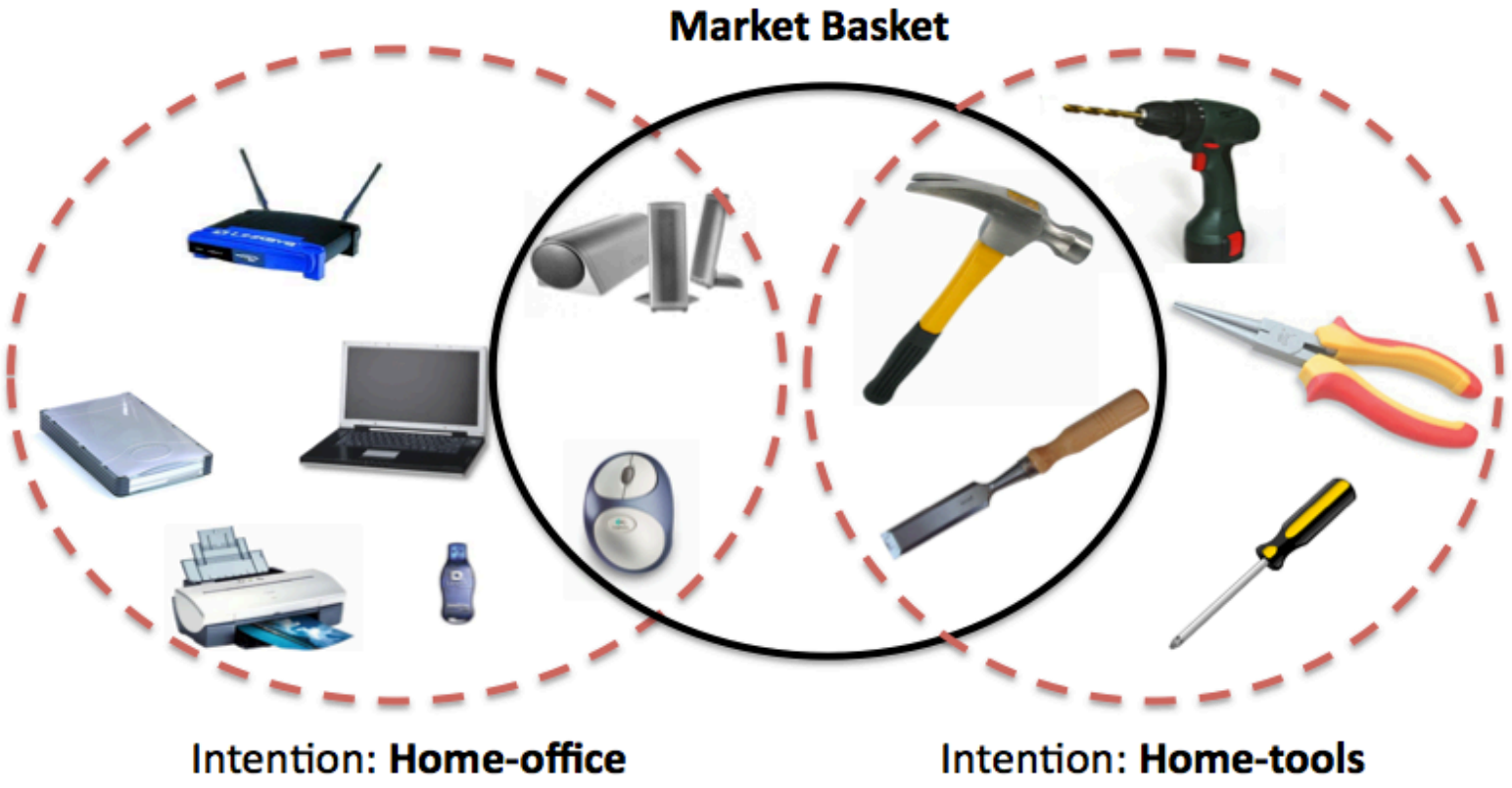
Why **Market Basket Analysis** failed?

Few buy a complete "logical" item-set in same basket

- already have other products
- buy them from another retailer
- buy them at a different time
- got them as gifts
- didn't know they needed it

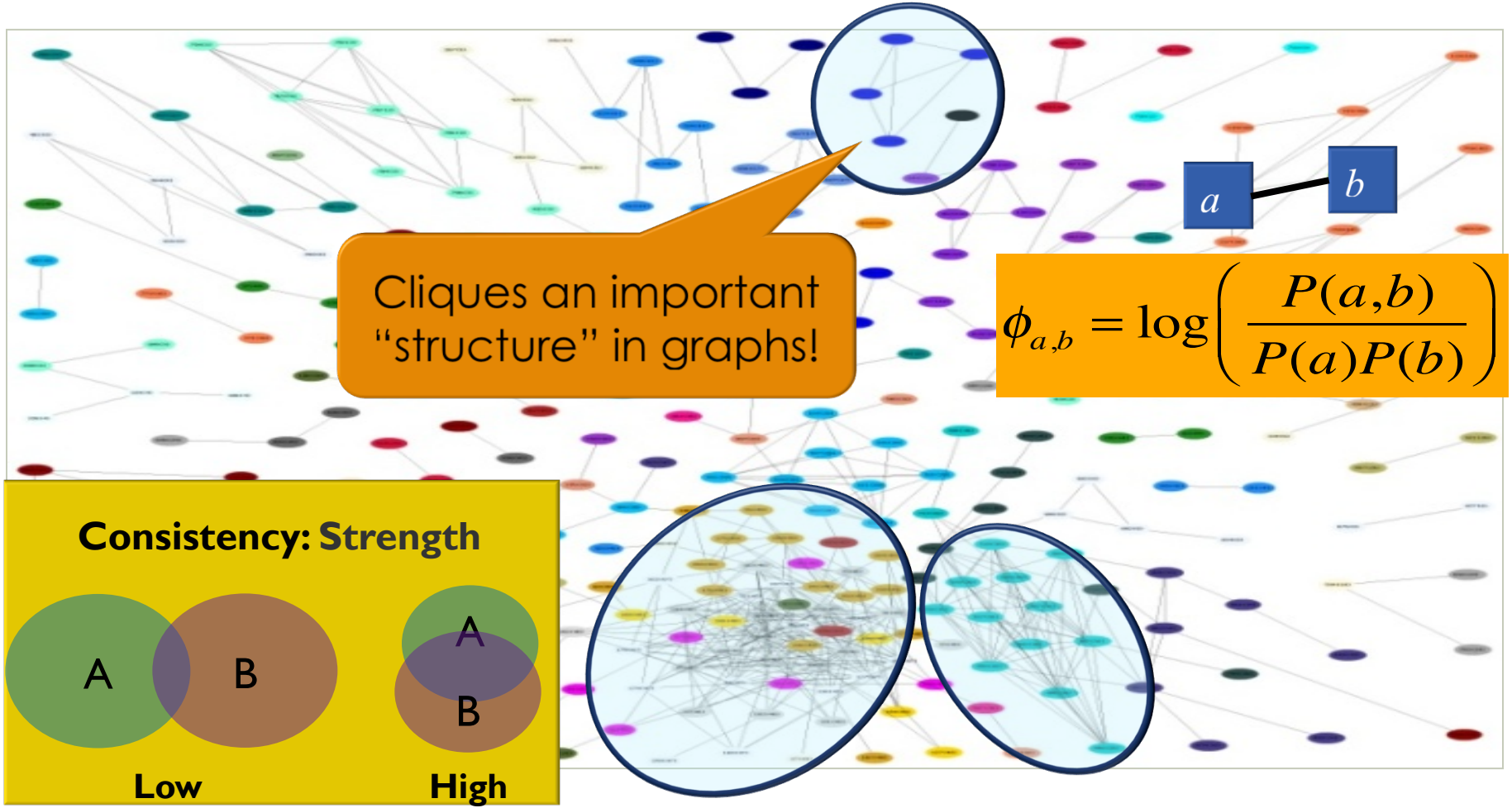
Projection of **latent** customer **intentions**

The **right** needle in the **wrong** haystack?

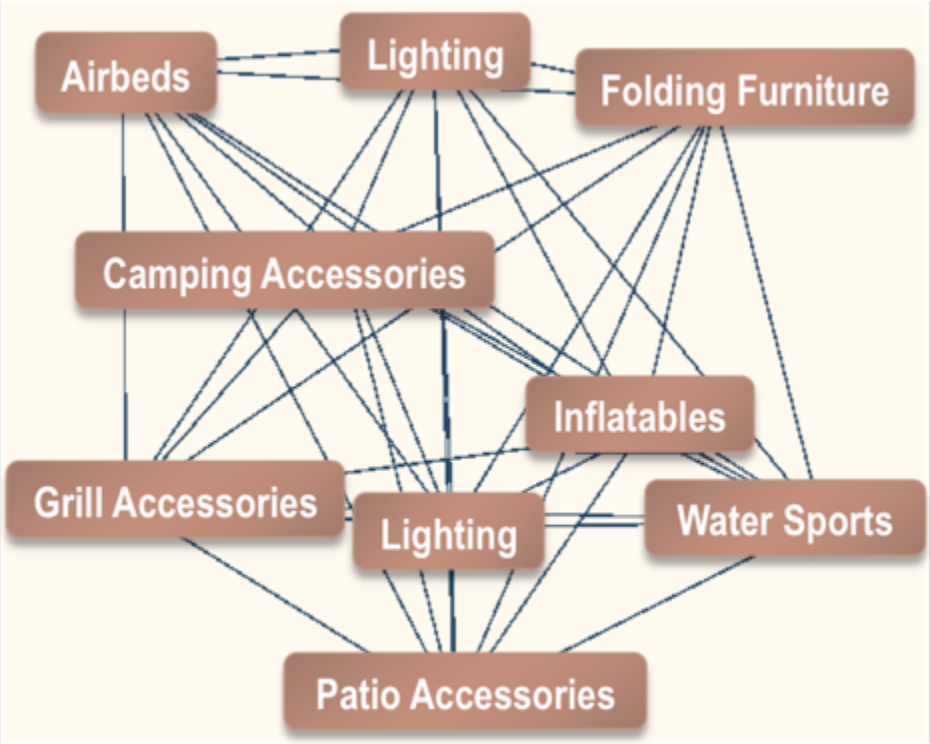


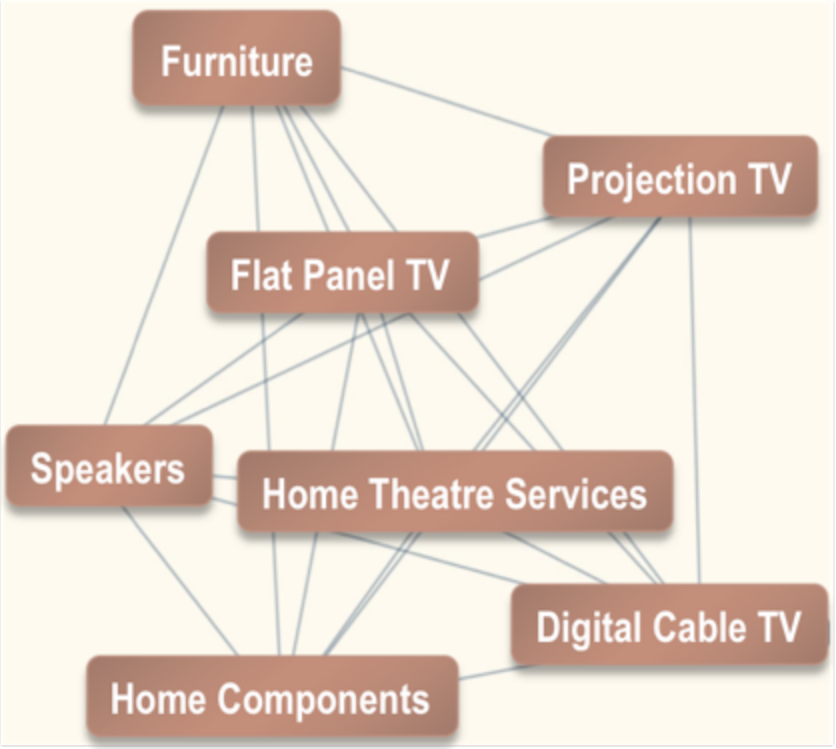
Mixture of **Projections** of **latent** customer **intentions**

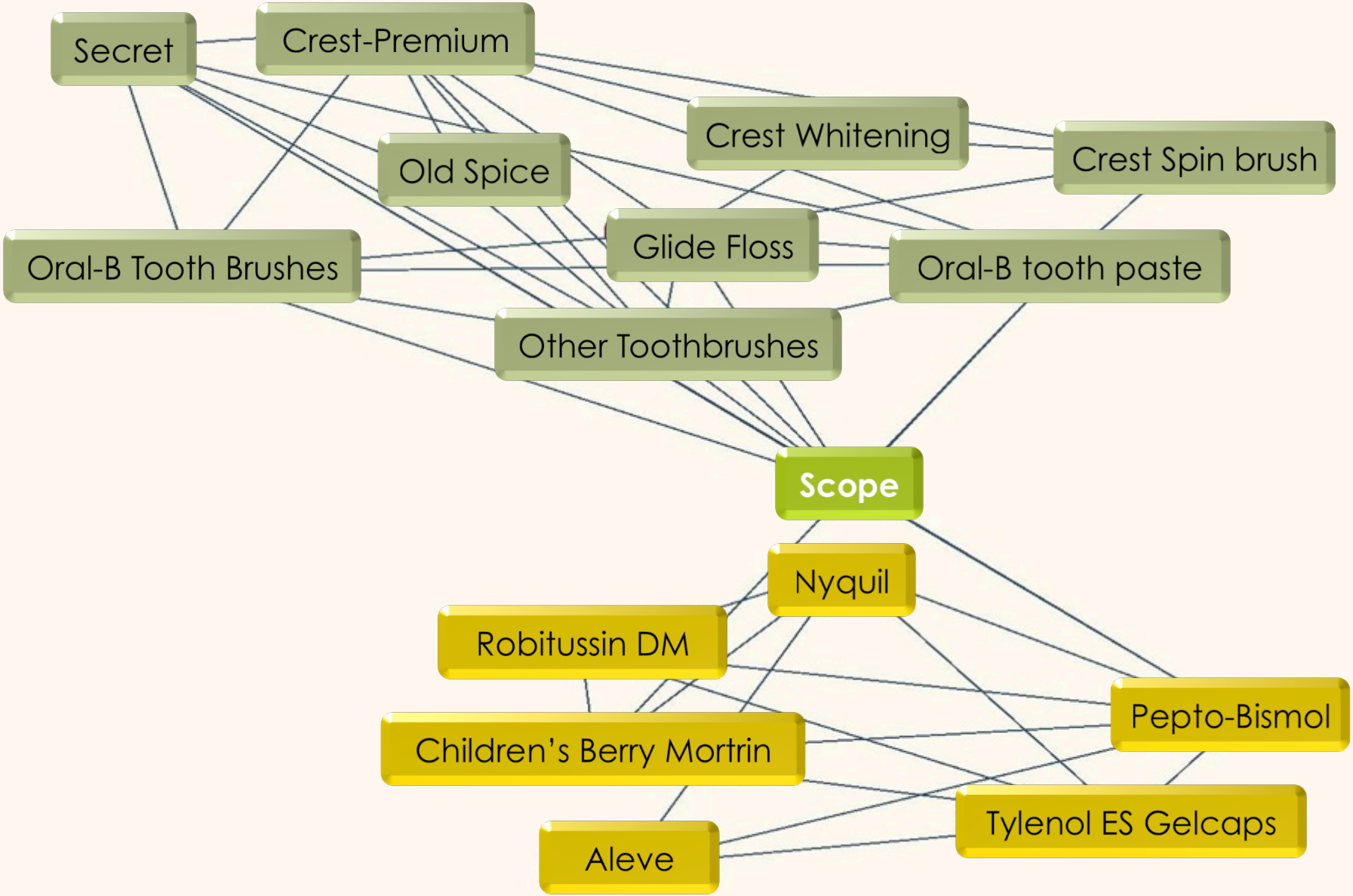
Extract **Knowledge**, throw away **Data**



Reconstruct needles, don't **find** them!






Secret
Crest-Premium
Crest Whitening
Crest Spin brush
Old Spice
Glide Floss
Oral-B Tooth Brushes
Oral-B tooth paste
Other Toothbrushes
Scope
Nyquil
Robitussin DM
Pepto-Bismol
Children's Berry Mortrin
Tylenol ES Gelcaps
Aleve

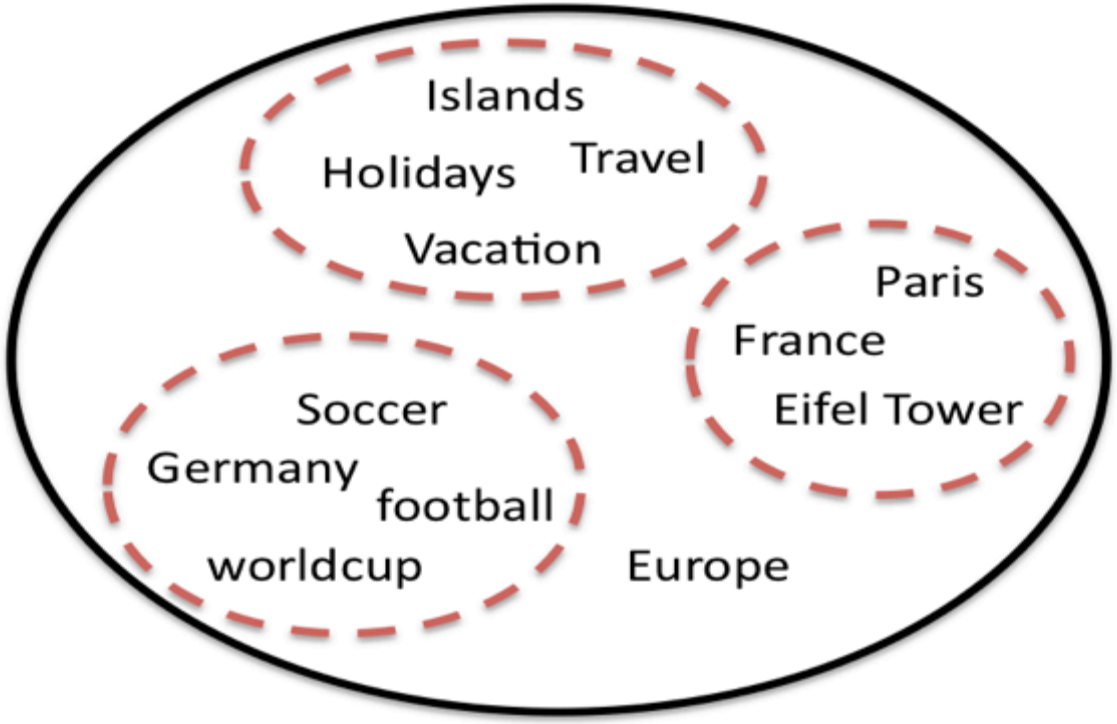
"**Data Mining** is like a box of chocolates. You never know what (**Insights**) you're gonna get."!!!

Discovering Tag Communities

Community Detection in Tagsets

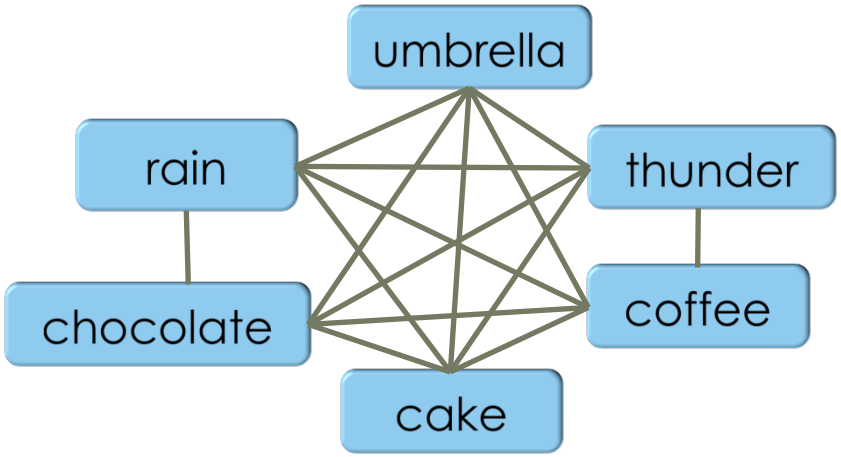
- **Tagset data…**
  - **Flickr** – tags describing **images**
  - **YouTube** – tags describing **videos**
  - **AdWords** – tags describing **advertisements**
  - **IMDB** – tags describing **movies**
  - **Keywords** – tags describing **scientific publications**
- **Key Challenges**
  - **Noisy Tag-sets** – create robust graphs!
  - **Weighted graphs** – communities in weighted graphs
  - **Overlapping communities** – no clustering!

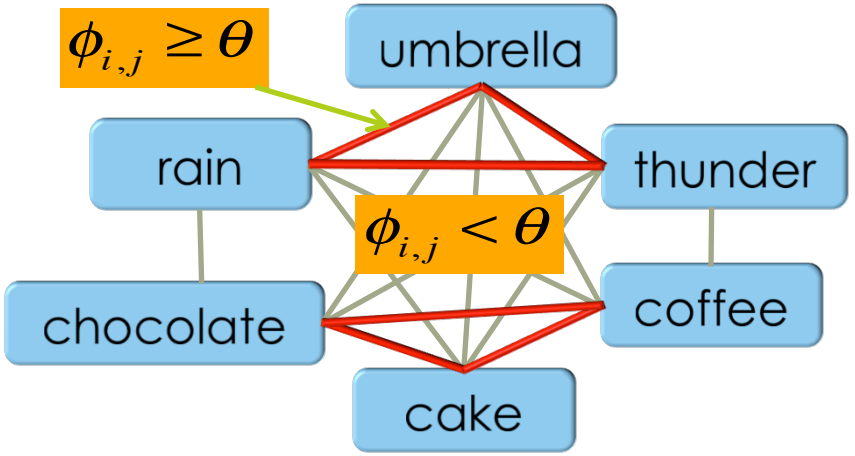
Tagsets – a "Crazy Haystack"!

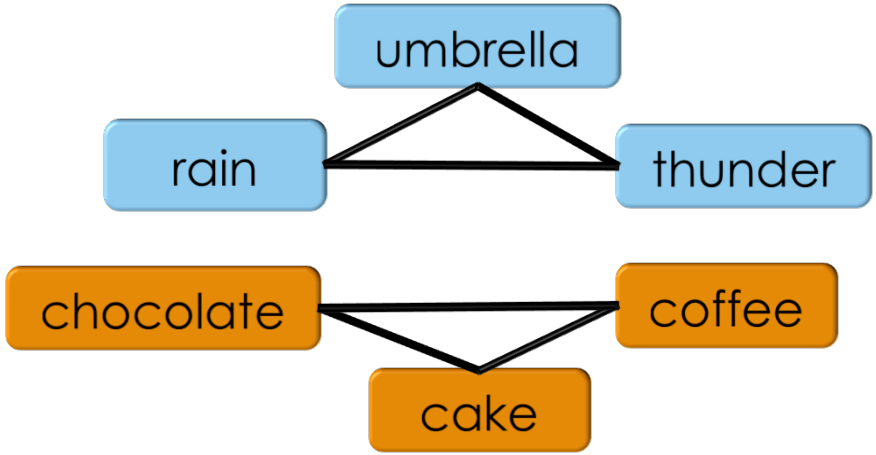


Mixture of **Projections** of **latent** **Concepts**

Creating Robust Co-oc Graph







$$\phi_{a,b} = \log\left(\frac{P(a,b)}{P(a)P(b)}\right)$$

De-noising – for better graphs

Co-occurrence of Tags with tag "**wedding**"

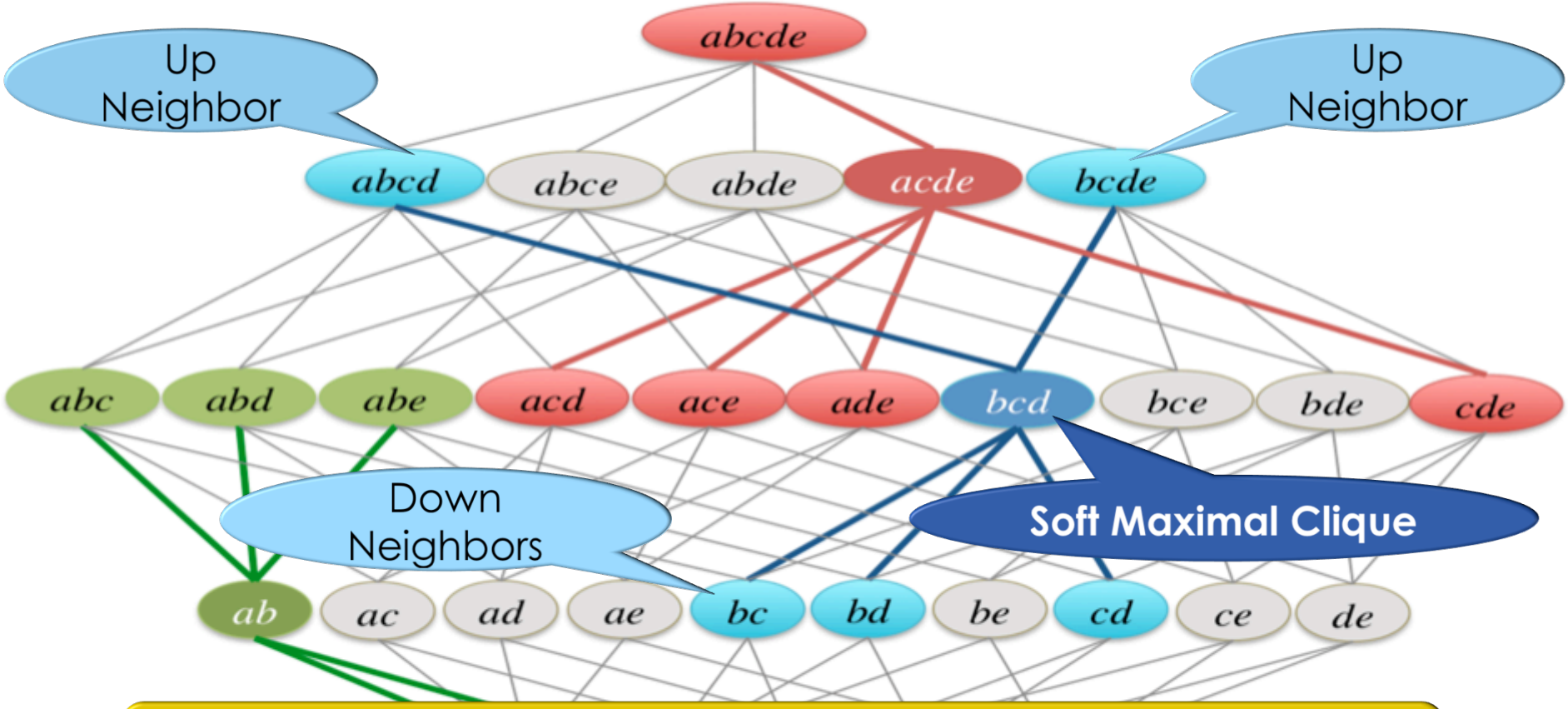
Tag	Before Denoising	After Denoising
bride	0.3257	0.5750
reception	0.3720	0.5728
marriage	0.3195	0.5658
cake	0.1699	0.3629
love	0.0148	0.2449
honeymoon	0.0183	0.2262
jason	0.2081	0
chris	0.1461	0

Consistency + Denoising

Tag	Most Consistent Tags in IMDB dataset
food	lifestyle, money, restaurant, drinking, cooking
road	truck, motorcycle, car, road-trip, bus
singer	singing, song, dancing, dancer, musician
suicide	suicide-attempt, hanging, depression, mental-illness, drowning
hospital	doctor, nurse, wheelchair, ambulance, car-accident

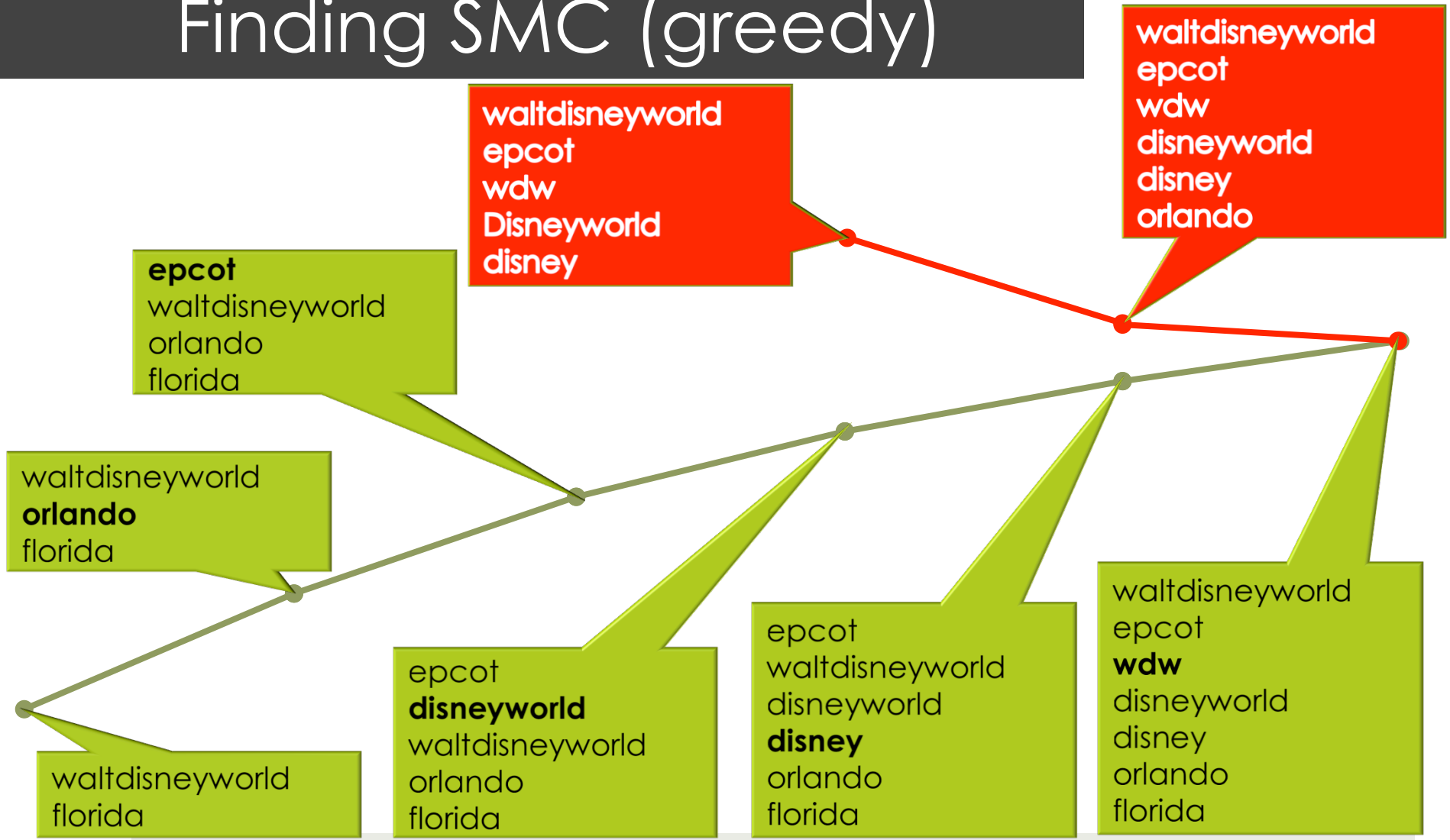
Tag	Most Consistent Tags in FLICKR dataset
art	painting, gallery, paintings, sculpture, artist
france	paris, french, eiffeltower, tower, europe
island	tropical, islands, newzealand, thailand, sand
animals	zoo, pets, wild, cats, animal
airplane	flying, airshow, fly, military, aviation

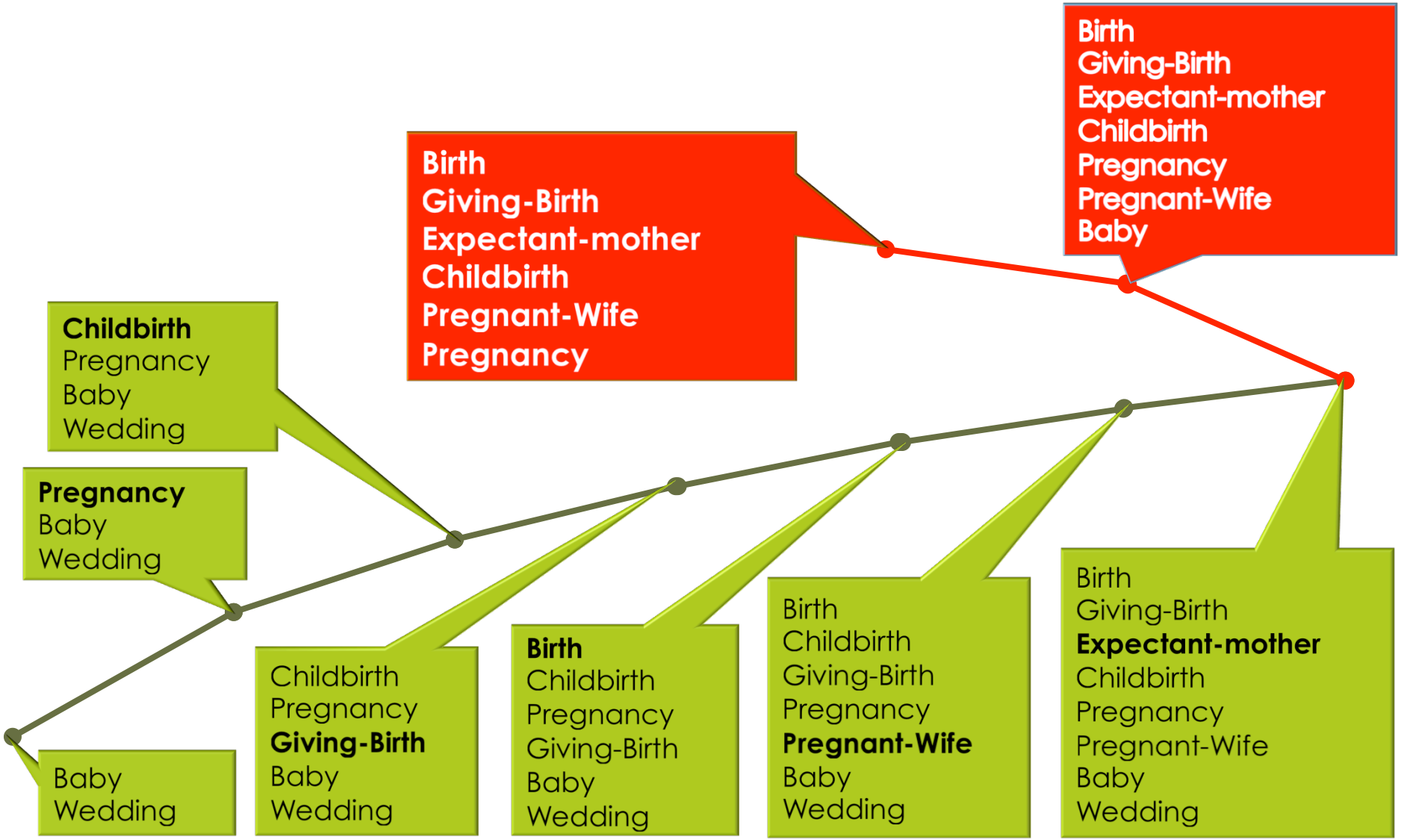
Soft Maximal Clique (SMC)



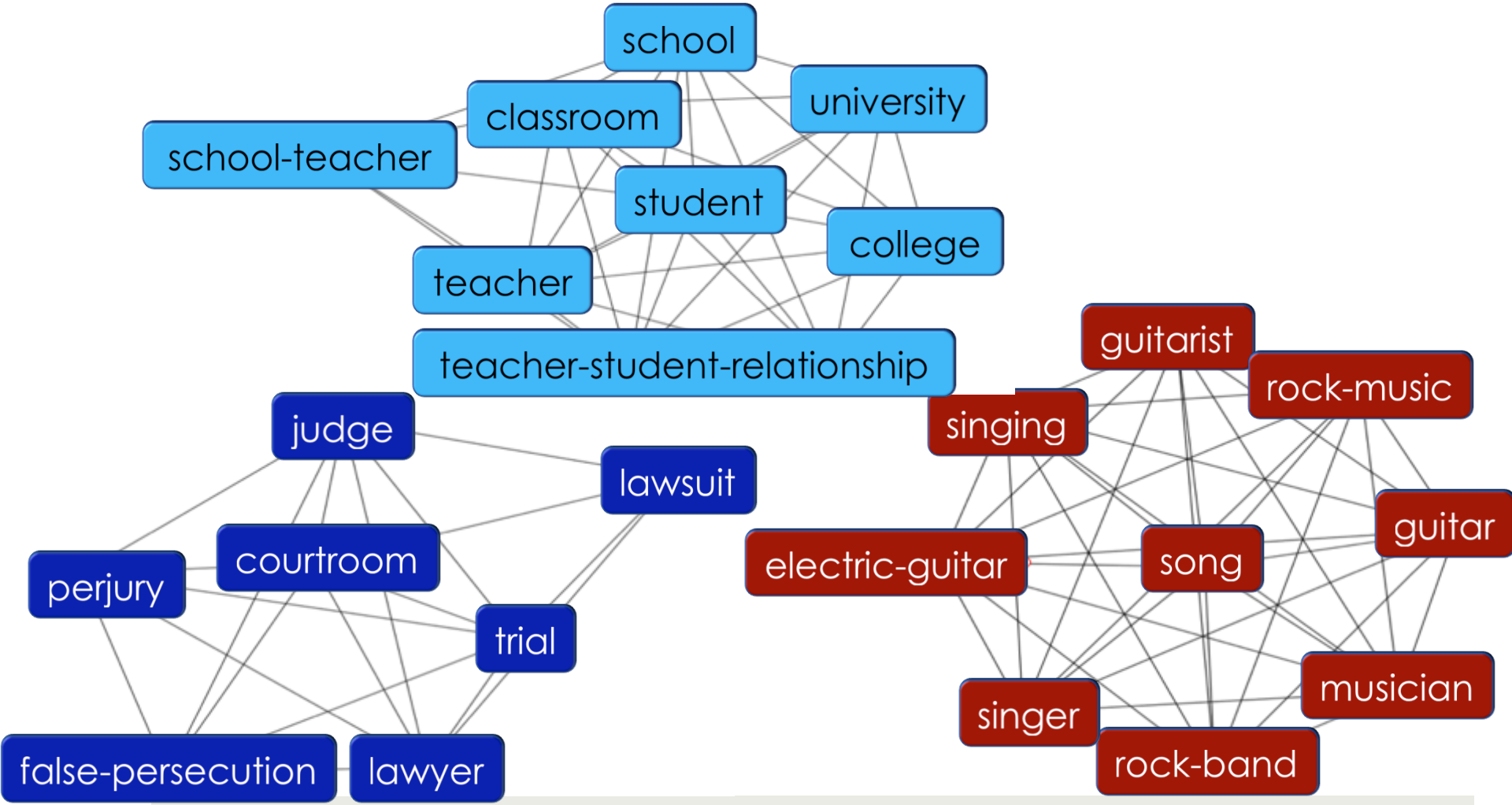
Coherence of a **Soft Maximal Clique** is higher than the coherence of all of its **Up** as well as **Down** neighbors

Finding SMC (greedy)




Baby
Wedding
Childbirth
Pregnancy
Giving-Birth
Baby
Wedding
Birth
Childbirth
Pregnancy
Giving-Birth
Baby
Wedding
Birth
Childbirth
Giving-Birth
Pregnancy
Pregnant-Wife
Baby
Wedding
Birth
Giving-Birth
Expectant-mother
Childbirth
Pregnancy
Pregnant-Wife
Baby
Wedding
Pregnancy
Baby
Wedding
Childbirth
Pregnancy
Baby
Wedding
Birth
Giving-Birth
Expectant-mother
Childbirth
Pregnancy
Pregnant-Wife
Baby

Discovered Communities (IMDB)



"You shall know a word by the company it keeps"

– *John Rupert Firth, 1957*

Finding Phrases in Text

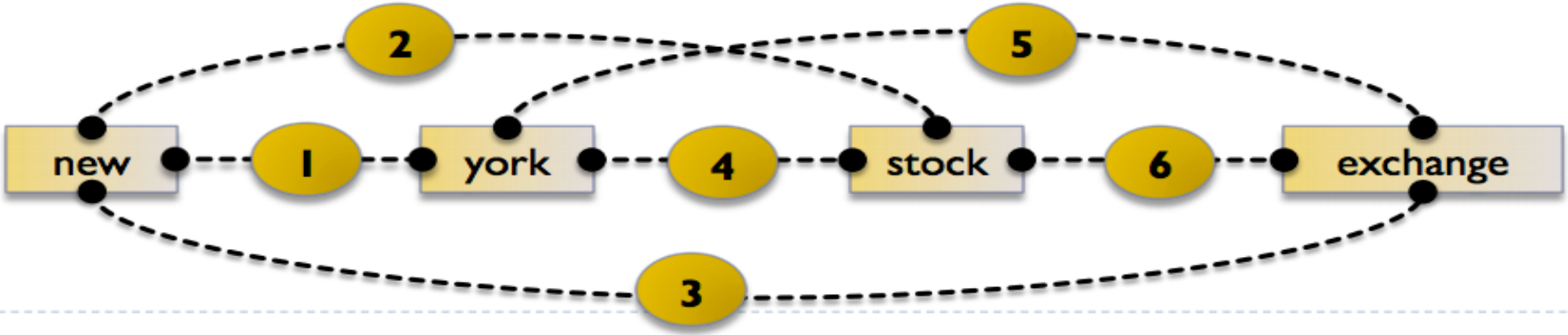
Unsupervised Text "**Enrichment**"

- **Syntactic Tokenization**
  - He distributes **Time** magazine in **new york**
  - Today **new york times** reported **new** rise in crime
- **Semantic Tokenization / Disambiguation**
  - I was **right** to avoid a **suit** against **apple**
  - Man in red **suit** on my **right** was drinking **apple** juice
- **Conceptual Tokenization**
  - **filed** a **suit charging orange** of **illegal behavior**
  - **submitted** a **case accusing apple** of **unauthorized conduct**
- **Contextual Weighting**
  - rain, thunder, umbrella, lightening, **chocolate**
  - kids, birthday, candies, **chocolate**, cake, candles

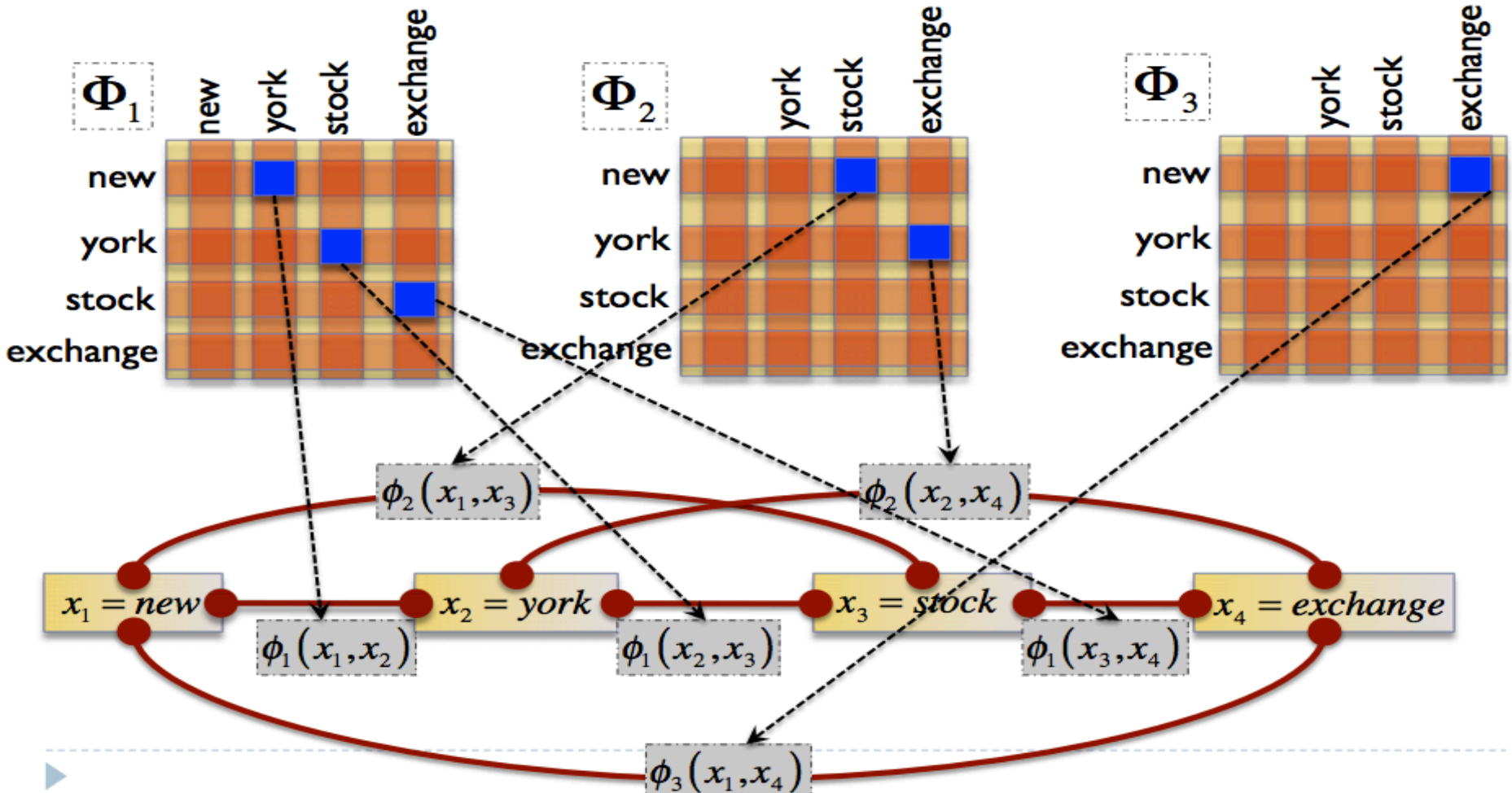
Positional Bigrams $O(nV^2)$

Positional Bi-grams = Co-occurrence of a pair of ***words*** in a ***specific relative position*** w.r.t. each other.

1. Word ***new*** occurs ***one position before*** word ***york***
2. Word ***new*** occurs ***two positions before*** word ***stock***
3. Word ***new*** occurs ***three positions before*** word ***exchange***
4. Word ***york*** occurs ***one position before*** word ***stock***
5. Word ***york*** occurs ***two positions before*** word ***exchange***
6. Word ***stock*** occurs ***one position before*** word ***exchange***

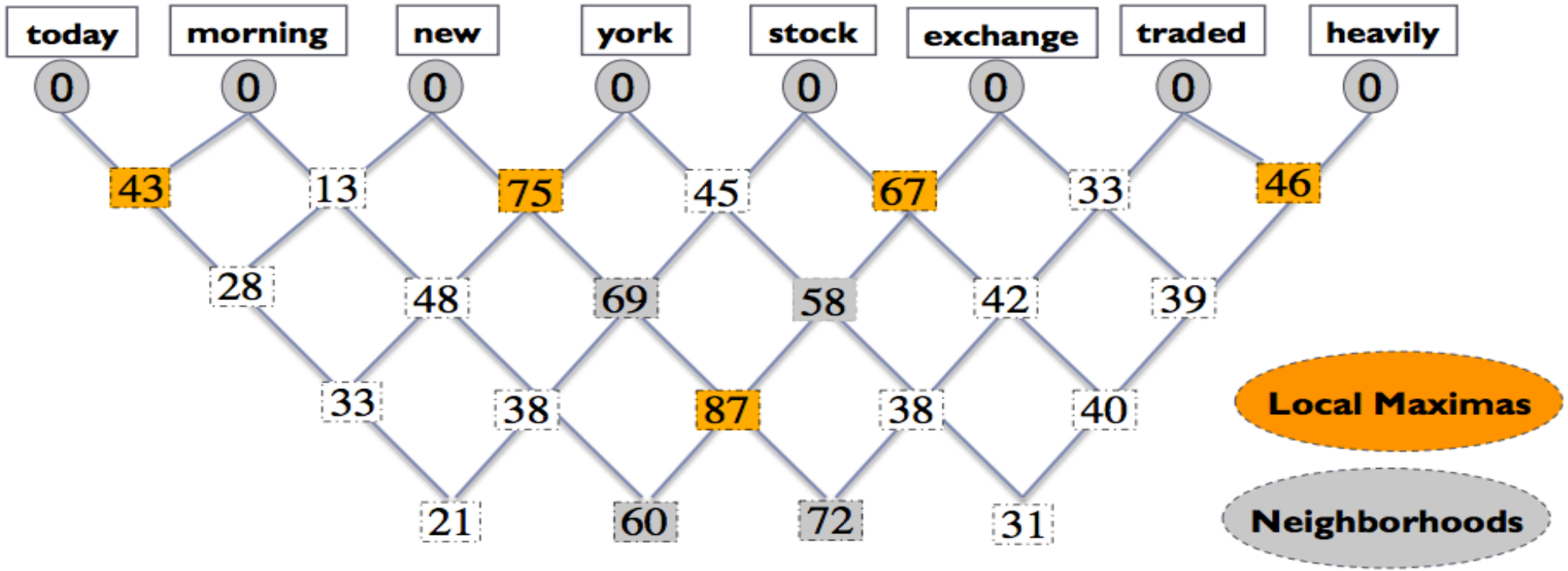


Using Positional Bigrams



Phrase: Soft Maximal Sequence

Local maxima in "**Sequence coherence lattice**"



Phrases in Wall Street Journal data

High Cohesiveness Low Frequency	Medium Cohesiveness Medium Frequency	Low Cohesiveness High Frequency
russian president boris yeltsin	new york stock exchange	he would not elaborate
prime minister viktor chernomyrdin	dow jones industrial average	sometime next year
prime minister ryutaro hashimoto	nasdaq composite index	few weeks before
palestinian leader yasser arafat	new york mercantile exchange	nearly two years ago
serbian president slobodan milosevic	automated teller machines	made no comment
syrian president hafez al-assad	international monetary fund	gain market share
british foreign secretary douglas hurd	u . n . officials	tons of cocaine
iraqi deputy prime minister tariq aziz	u . n . peacekeepers	had no chance
secretary general kofi annan	u . n . spokesman	made no comment
senate majority leader bob dole	an international tribunal	materials contained herein
lieutenant general raoul cedras	intensive care unit	people have been killed
nba commissioner david stern	highway traffic safety administration	as soon as possible
bulls coach phil jackson	mitsubishi heavy industries ltd	dollar rose as high as
federal reserve policy makers	rio de janeiro	threatened to pull out
	freshly ground black pepper	forced to give up
		as far as possible
		as early as tomorrow
		kilometers (<num> miles) north
		kilos (<num> pounds)

And… It's language agnostic

世 界 贸 易 组 织	World Trade Organization	10.504958
总 统 克 林 顿	President Clinton	9.011669
亚 洲 金 融 危 机	Asian Financial Crisis (1997)	9.702676
巴 塞 罗 那 奥 运	Barcelona Olympics (1992)	10.333340
总 统 布 什	President Bush	7.339422
俄 罗 斯 总 统 叶 利 钦	Russian President Yeltsin	11.698954
糖 尿 病	diabetes	11.583491
艾 滋 病 毒	Human immunodeficiency virus	14.140515
络 绎 不 绝	in endless stream	8.637969
紧 锣 密 鼓	an intense publicity campaign	11.915520
兴 致 勃 勃	be highly interested in	11.415492

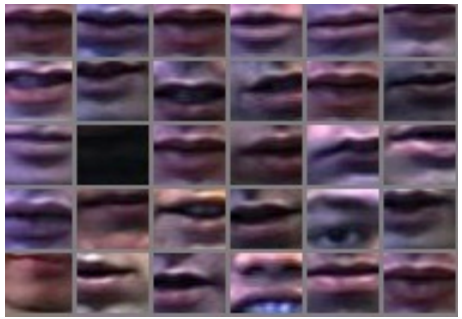
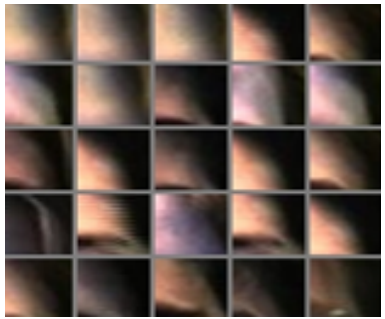
"You shall know a PIXEL by the company it keeps"

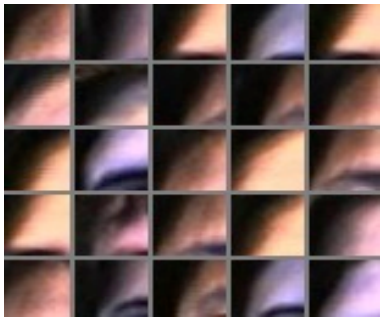
Recognizing "objects" in Images

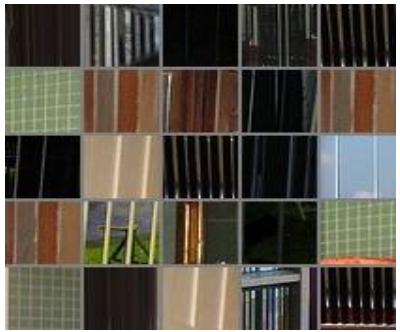
Images – the "Craziest Haystack!"

- Image Understanding (the way humans do) is very difficult
- Images have far too many degrees of freedom!
  - Translation, Scale, Rotation, Illumination, Color, Shadows, …
- Even for Human brain its hard!
  - 30-40% of brain processes vision!
- Most computer vision systems:
  - Use Low level features
  - Complex models to compensate
- SIFT features →
  - Still don' t capture semantics

Phrases and Concepts in Images





Content Based Image Retrieval














Content Based Image Retrieval

Content Based Image Retrieval











Content Based Image Retrieval

Content Based Image Retrieval


















Content Based Image Retrieval















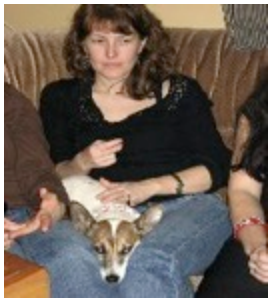




Content Based Image Retrieval

Content Based Image Retrieval









Content Based Image Retrieval












Content Based Image Retrieval













Conclusions

- **World is a hierarchy of objects**
  - Raw data is the lowest level objects
- **Discovering the "grammar of the data"**
  - Syntactic Composition
  - Semantic Equivalencing
- **Co-occurrence Analysis Framework**
  - Unsupervised Hierarchical Structure Discovery
  - Language and Domain agnostic