Practical Learnings from Real-World Data Mining

Dr. Shailesh Kumar

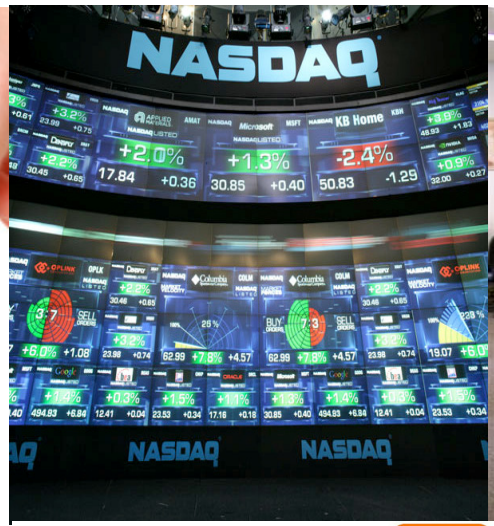
Google, Inc.

In theory, theory and practice are same.
In practice, they are not. *-- Lawrence Peter Berra*

Welcome to the Information Age …

… drowning in data and starving for Knowledge

ATATTAGGTTTTTACCTACC
CAGGAAAAGCCAACCAACC
TCGATCTCTTGTAGATCTG
TTCTCTAAACGAACTTTAAA
ATCTGTGTAGCTGTCGCTC
GGCTGCATGCCTAGTGCA
CCTACGCAGTATAAACAAT
AATAAATTTTACTGTCGTTG
ACAAGAAACGAGTAACTCG
TCCCTCTTCTGCAGACTGC
TTATTACGCGACCGTAAGC



REUTERS Blogger flickr YouTube WIKIPEDIA The Free Encyclopedia

facebook Google+ linkedin

amazon NETFLIX

This data explosion is enabled by…

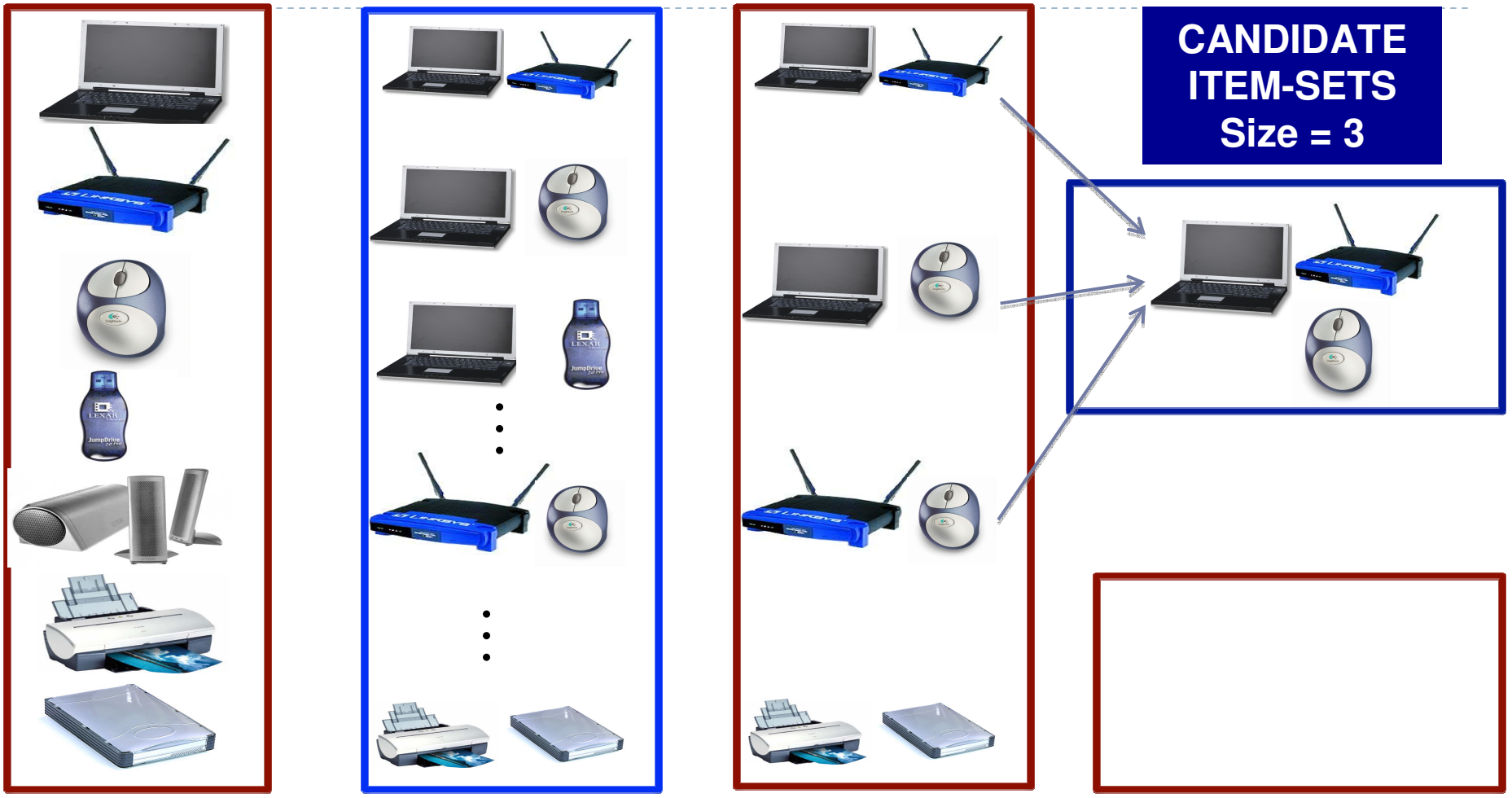
- **Better "Sensors"** – Higher Resolution, More Spectral Bands, Quick Experimental Turnaround, Crowd Sourcing…
- **Higher Bandwidth Communication** – Faster Networks and Routers, Better Compression technologies…
- **Larger Warehouses** – Cheaper Storage, Multi-Level Caching, Scalable Database/Data warehousing technologies…
- **Massive Crunching Power** – Faster Multi-core processors, Parallel Distributed Computing, MapReduce paradigms…
- **Advances in Machine Learning and Data Mining** –

Lessons from Real-World Data Mining

- **Retail: Discovering Needles in a Haystack?**
  - **Is this the right needle to look for?**
  - **Do you understand the nature of your haystack?**
- **Text: Its all in the "Representation"?**
  - What is a term?
  - What does it mean?
  - How much does it matter?
- **Computer Vision: Bridging the "Semantic Gap"**
  - Build semantically deeper features that "mean something"!

The Traditional Market Basket Analysis

Wrong needle in a mysterious haystack!



Lesson: **Know your data (Haystack)**
What process generated the data?

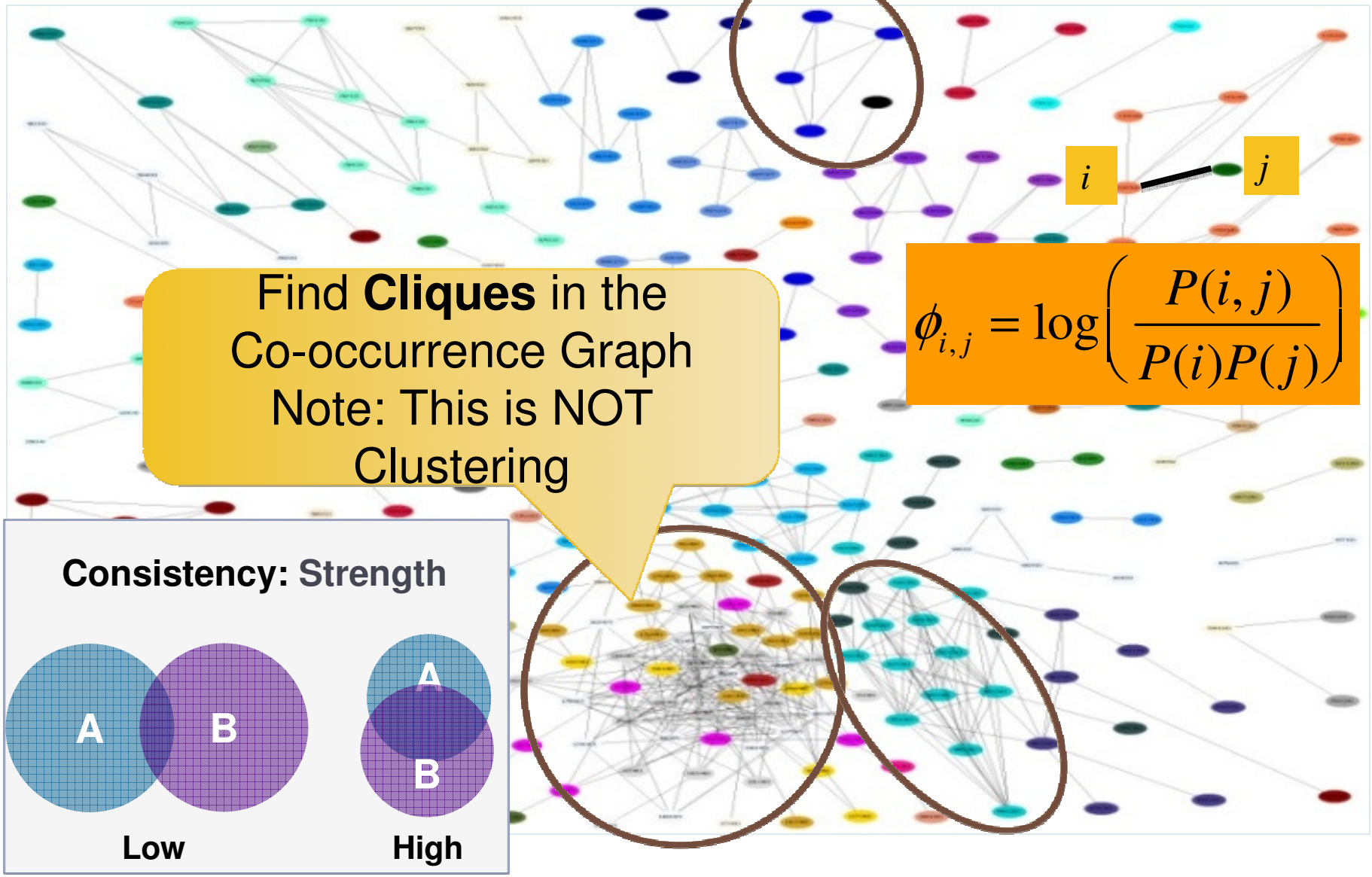
Few buy a complete “logical” product group in the same basket

- already have other products
- buy them from another retailer
- buy them at a different time
- got them as gifts
- ….

mixture of, projections of, latent intentions

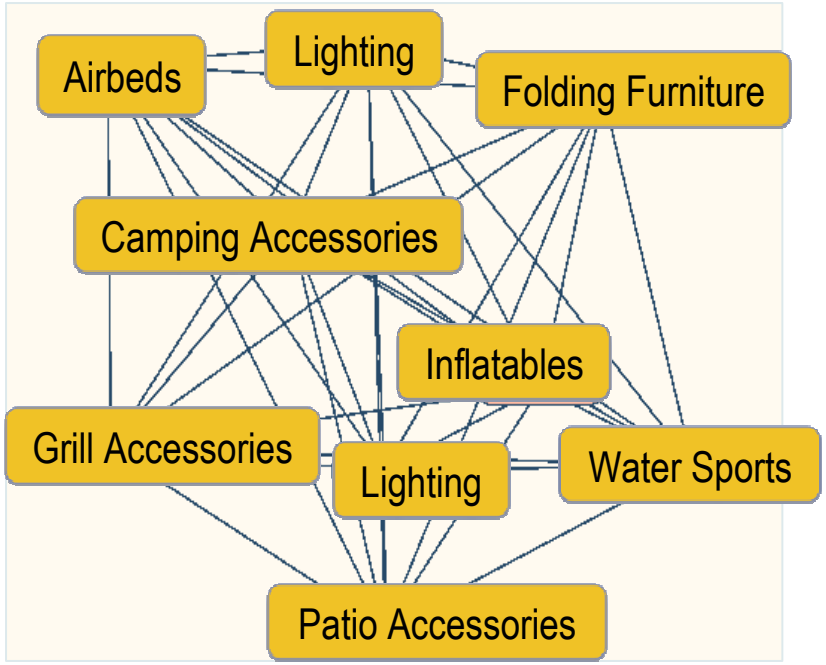
Lesson: **Extract the essence, let go of data**

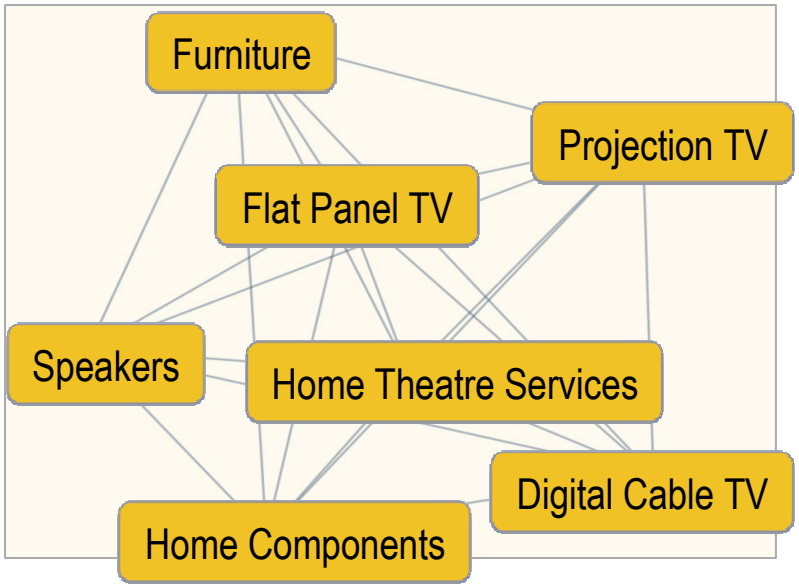
Pair-wise Co-occurrence Statistics



Lesson: **Look for the right Insight**

"Frequent" vs. "Logical" Itemset

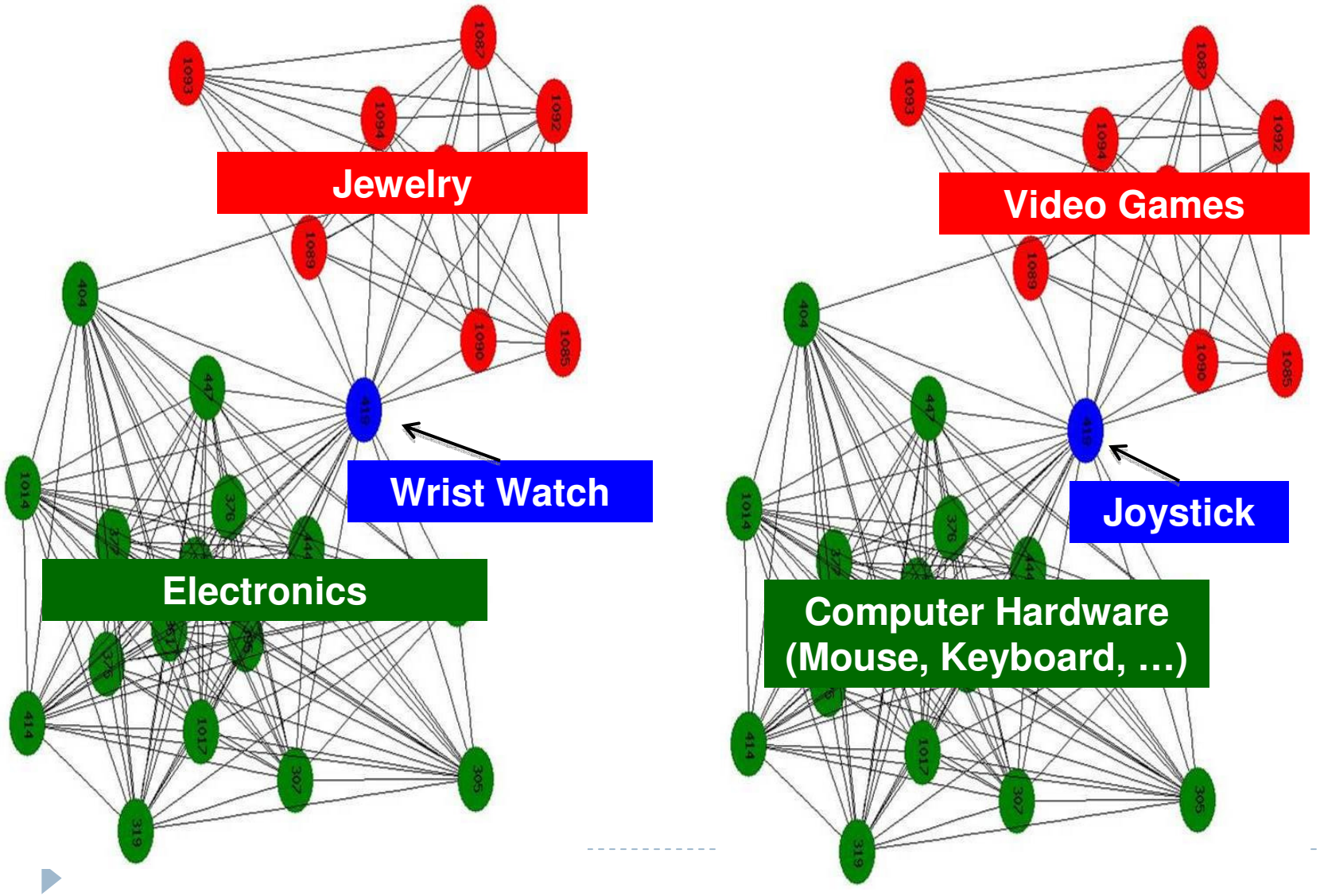




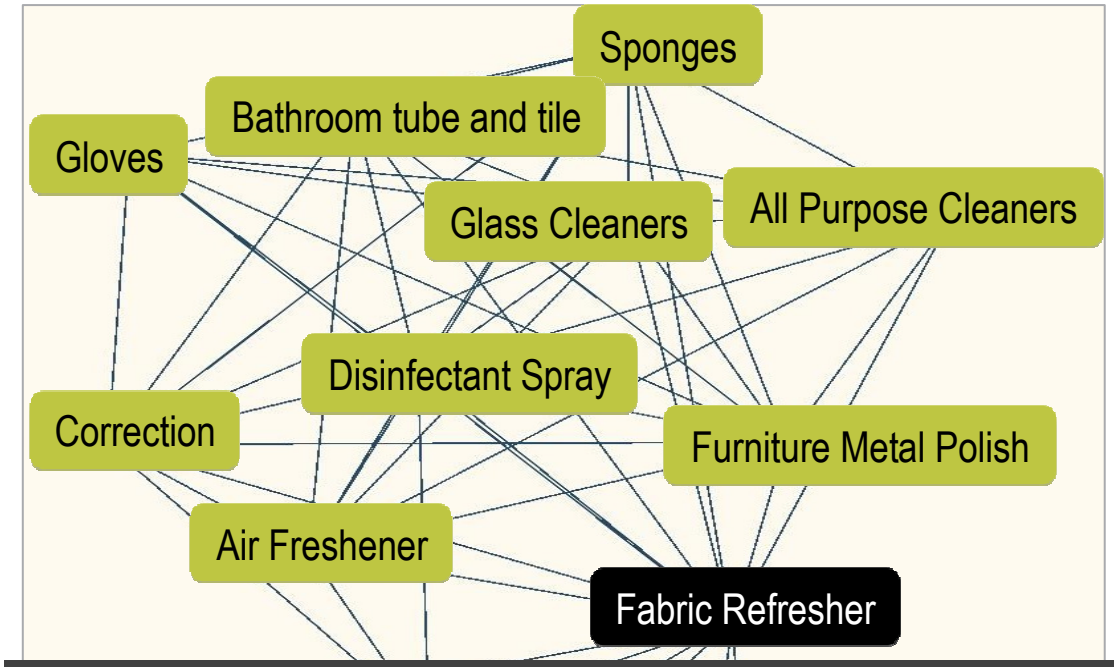
- Novel – Not obvious from the data (support = 0)
- Useful – product bundling, recommendations, layout
- Exhaustive – "No insight left behind!" – however "rare"

Lesson: **Let the data speak for itself!**

Bridge Products – drive traffic across depts.



Lesson: **Serendipity is good!**
I don't know what I don't know – let data help



What Products Does

Fabric Refresher

Bridge To?

The Spirit of “Insight Discovery”!

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

Lessons from Real-World Data Mining

- Retail: Discovering Needles in a Haystack?
  - Is this the right needle to look for?
  - Do you understand the nature of your haystack?
- **Text: Its all in the "Representation"?**
  - **What is a term?**
  - **What does it mean?**
  - **How much does it matter?**
- Computer Vision: Bridging the "Semantic Gap"
  - Build semantically deeper features that "mean something"!

Lesson: **Things are not what they appear**
What is a word in "Bag-of-Words"?

- **Segmentation:** What is a term?
  - New York Stock Exchange → 4 words?
  - "New York" "Stock Exchange" → 2 phrases?
  - "New York Stock Exchange" → 1 phrase?
- **Disambiguation & Equivalencing:** What does it mean?
  - '**rock** band', '**rock** climbing', '**rock**ing chair', 'the **rock**'
  - **orange** juice, **orange** flag, **orange** blog,
  - **apple** store, **apple** pie, the big **apple**
- **Weighting**
  - lightening,, thunder, rain, **road**, **umbrella**, **chocolate**

Lesson: **Supervision is best delayed**
Unsupervised, Statistical > NLP

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)

Lesson: **Stay Language Independent**
Chinese Giga Word Corpus

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

Lesson: **Does it mean what it says?**
Disambiguation and Equivalence

Equivalencing SIMILARITY = 0.995

- we filed a suit charging dell of illegal behavior
- they submitted a case accusing apple of unauthorized conduct

Disambiguation SIMILARITY = 0.171

- i was right to avoid a suit against apple
- on my right was a man in a suit drinking apple juice

You shall know a word by the company it keeps

-- Firth, J. R. 1957:11

Lesson: **What Independence assumption?**
How much does a term weigh?

- IDF Weights: Global weights irrespective of “context”
- Contextual Weights: Local weights based on other words

idf	
10.2765	hinduism
8.6589	hindu
8.6259	finger
7.8524	kerala
7.3432	mother
6.7895	smile
6.6507	child
6.576	women
6.5535	point
6.4512	happy
6.0312	orange
5.2129	india
4.312	family

ctx	
8.8989	child
8.8033	smile
8.338	happy
7.982	mother
6.0989	women
4.8763	family
4.208	india
2.9307	hinduism
2.8871	hindu
2.8318	orange
1.4355	kerala
0.2292	point
0	finger

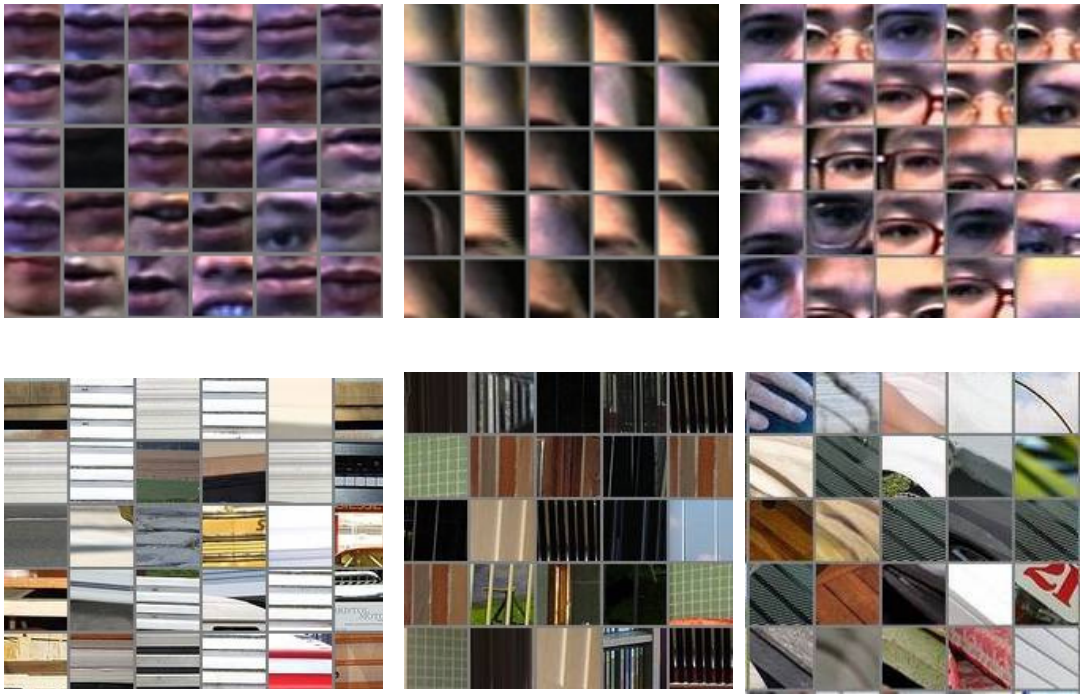
Lessons from Real-World Data Mining

- Retail: Discovering Needles in a Haystack?
  - Is this the right needle to look for?
  - Do you understand the nature of your haystack?
- Text: Its all in the “Representation”?
  - What is a term?
  - What does it mean?
  - How much does it matter?
- **Computer Vision: Bridging the “Semantic Gap”**
  - **Build features that “mean something”!**

Lesson: **Use Semantically deep features** Lines → Objects

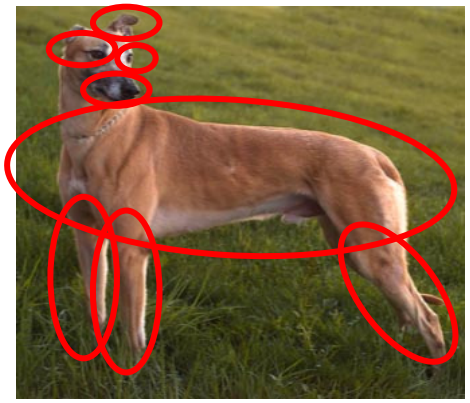
SIFT Bag-Of-Words

Semantically Deep features



Learning a Hierarchy-of-Objects
The "Building Blocks of Learning"

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



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

Retrieving Similar Images















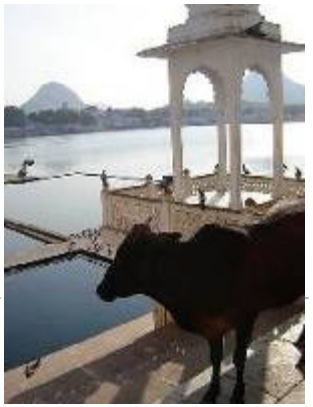
Retrieving Similar Images

Retrieving Similar Images

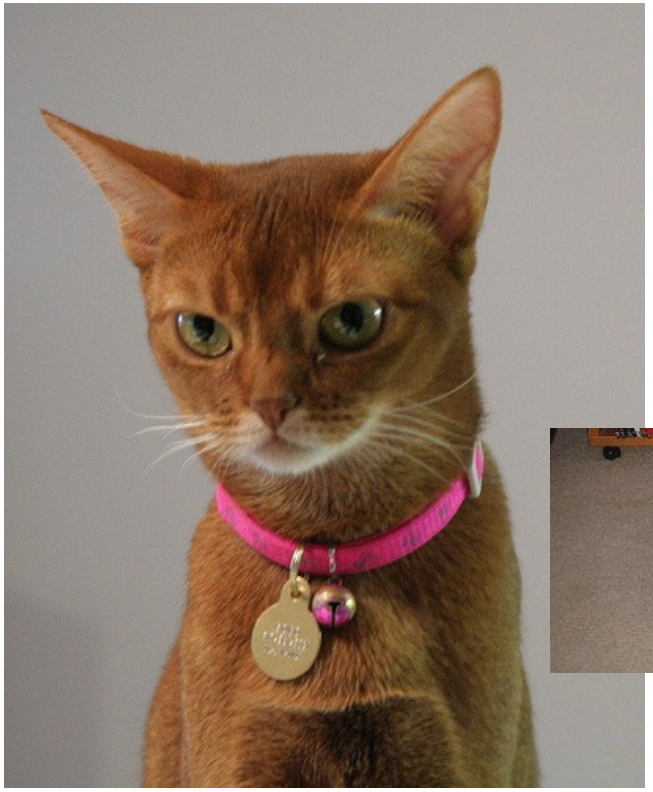




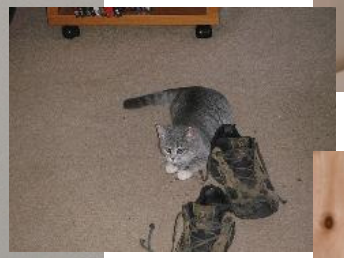



Content based Image Retrieval






ERROR: undefined
OFFENDING COMMAND: f‘~

STACK: