Deep-Parsing Natural Languages

深度语法分析是自然语言应用的核武器

ArchSummit 全球架构师峰会 2017 (北京)

李维（京东硅谷研究院）

12/08/2017 (liweinlp.com)

主办方 Geekbang 极客邦科技 InfoQ

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14 北京国际会议中心

扫描关注大会官网

Outline

- 人工智能的历史和现状简介：从感知到认知

  此消彼长的经验主义和理性主义钟摆

- 深度解析（Deep Parsing）是什么?

- NLP 架构纵览

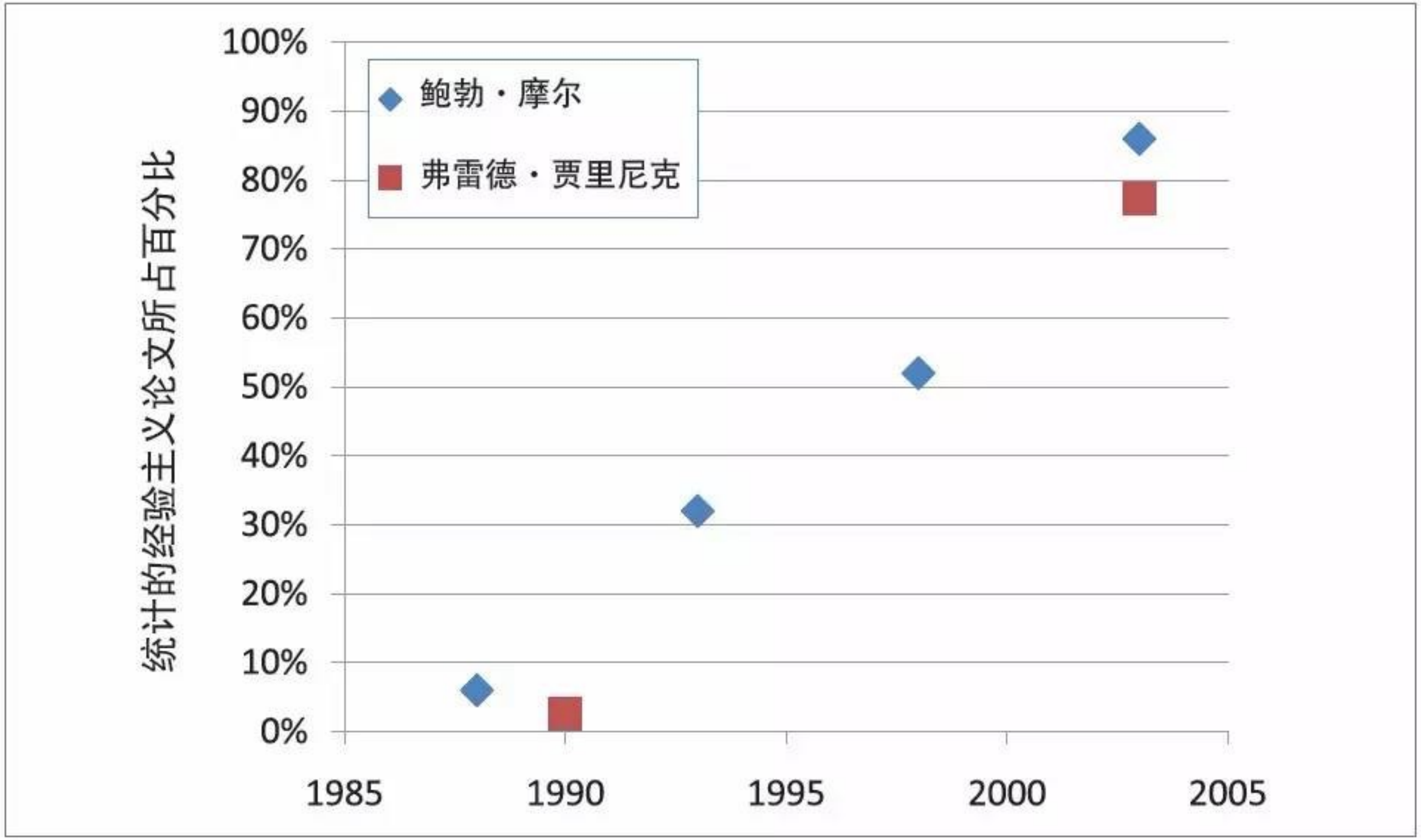
- 核武器应用举例

Outline: AI History

- 人工智能的历史和现状简介：从感知到认知

  此消彼长的经验主义和理性主义钟摆

NLP Mainstream Since 1990s



Courtesy of Prof. Church: **“A Pendulum Swung Too Far”**
http://blog.sciencenet.cn/blog-362400-988692.html

NLP频道
liweinlp.com

Two Basic Approaches to NLP

Approach	Pros	Cons
Statistical Learning (based on keywords)	• Good for document-level • High recall • Robust • Easy to scale • Fast development (if data available)	• Requires large annotation • Coarse-grained • Difficult to debug • Fail in short messages • Only shallow NLP • No understanding
Grammar Engineering (based on sentence structure)	• Good for sentence level • Handles short messages well • High precision • Fine-grained insights • Easy to debug • Parsing and understanding	• Requires deep skills • Requires scale up skills • Requires robustness skills • Moderate recall (coverage) • Parser development slow

- Complementary rather than competing
- Hybrid: Best of both worlds
- Balance and configurability between precision and recall

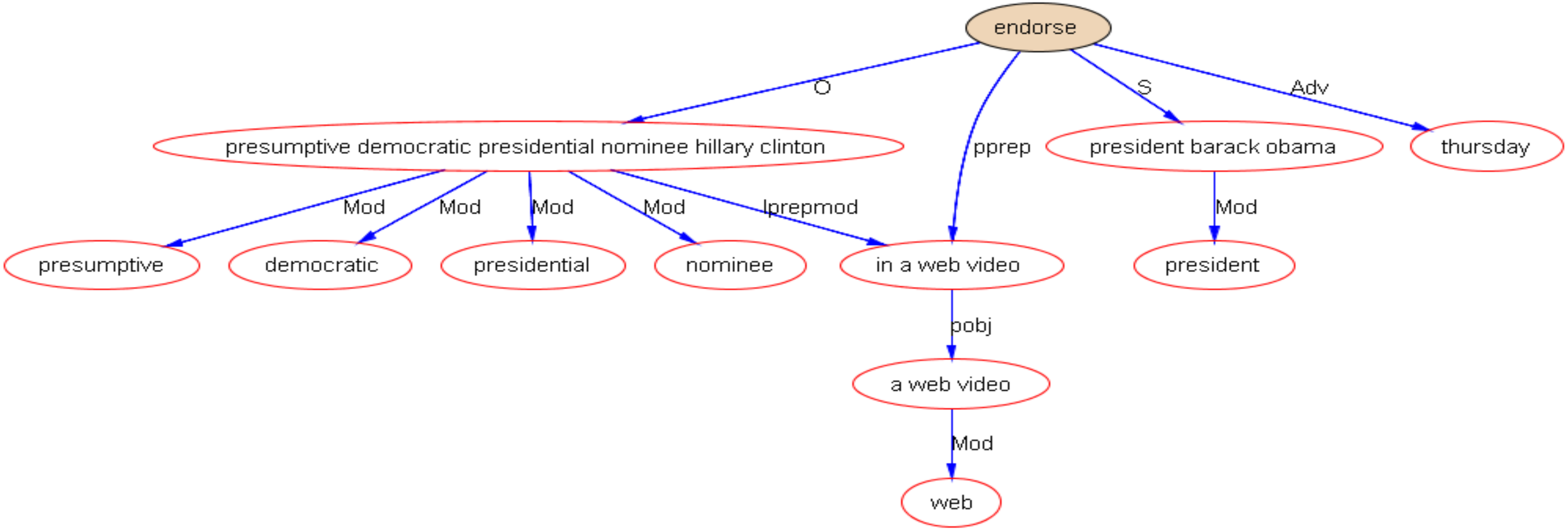
Outline: What Is Deep Parsing

- 人工智能的历史和现状简介：从感知到认知

  此消彼长的经验主义和理性主义钟摆

- 深度解析（Deep Parsing）是什么?

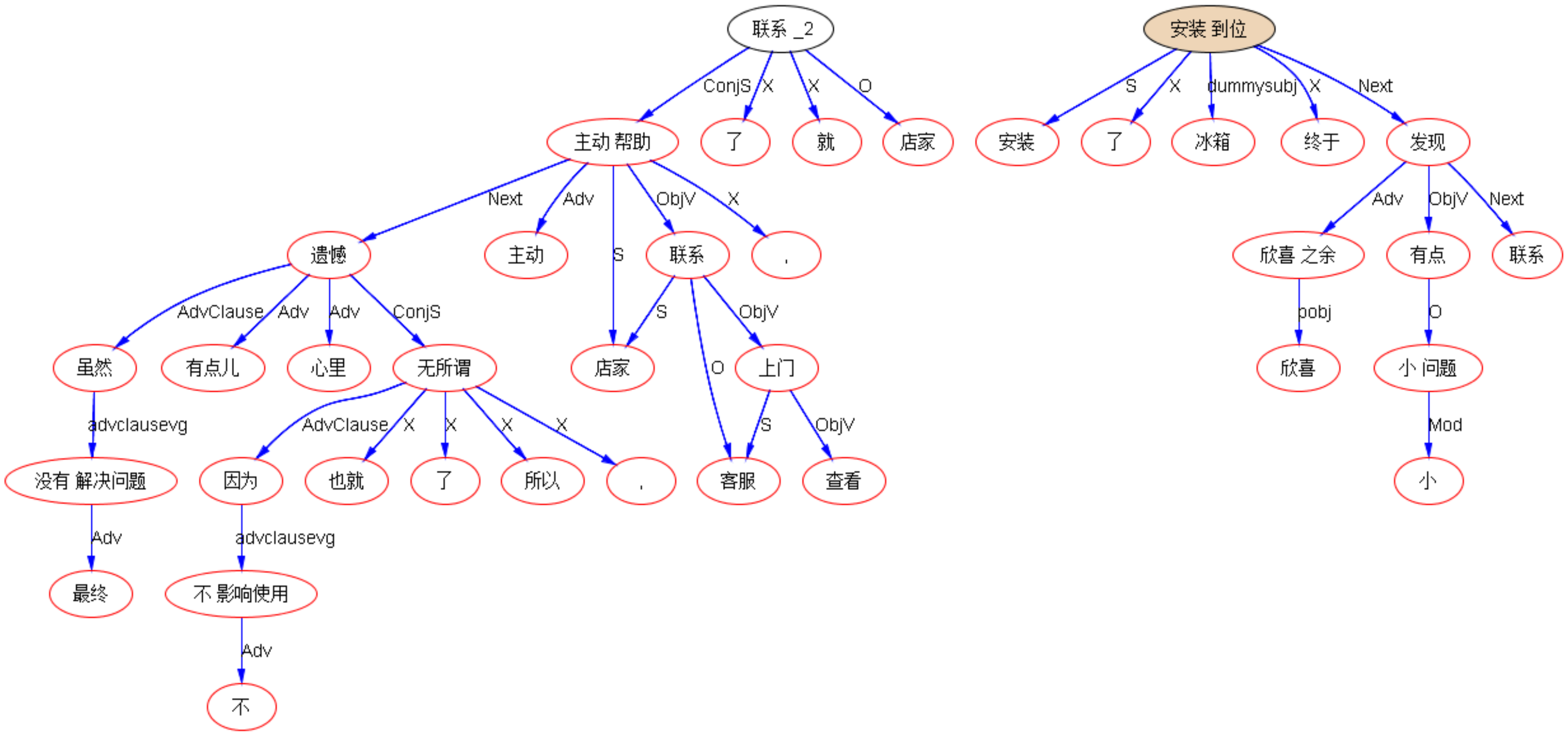
Deep Parsing: Unstructured to Structures



1. Input: President Barack Obama endorsed presumptive Democratic presidential nominee Hillary Clinton in a web video Thursday .

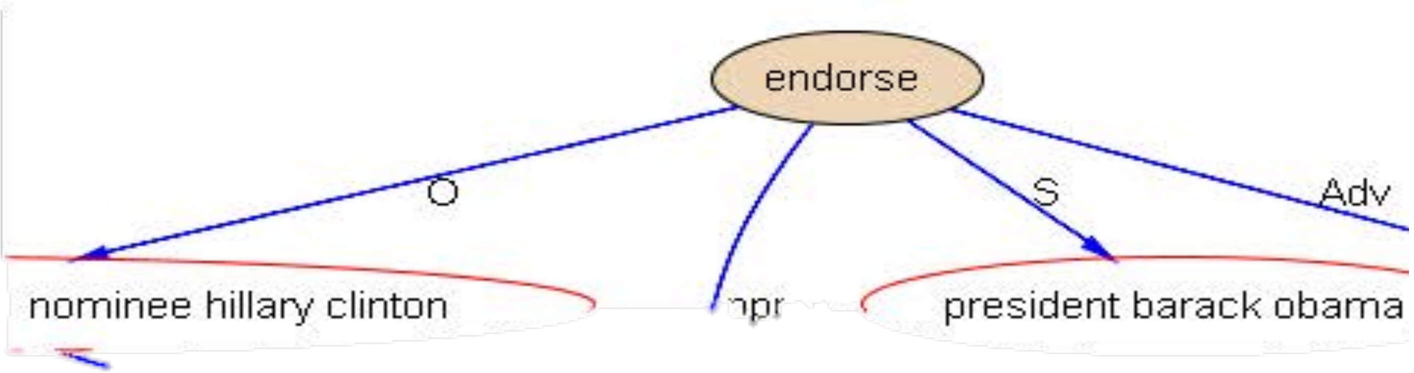
NLP频道
liweinlp.com

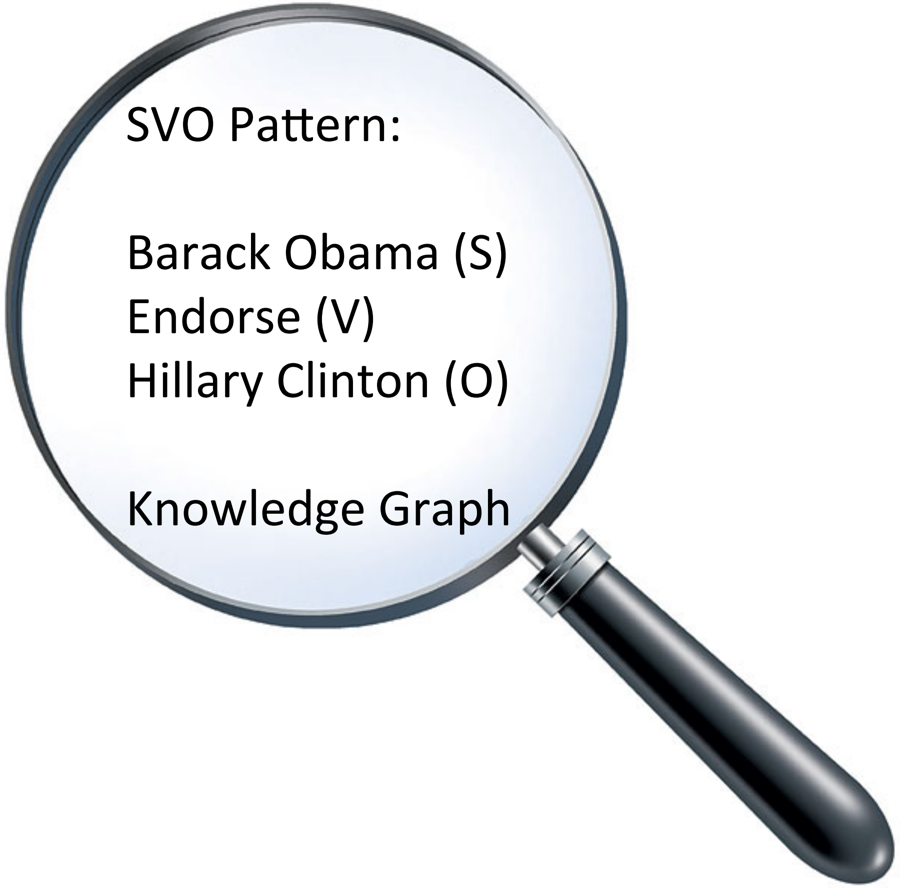
Why parsing? Limited Patterns



终于冰箱安装到位了，欣喜之余发现有点儿小问题，就联系了店家，店家主动帮助联系客服上门查看，虽然最终没有解决问题，心里有点儿遗憾，但是因为不影响使用，所以也就无所谓了.

Subtree Pattern: Data to Intelligence



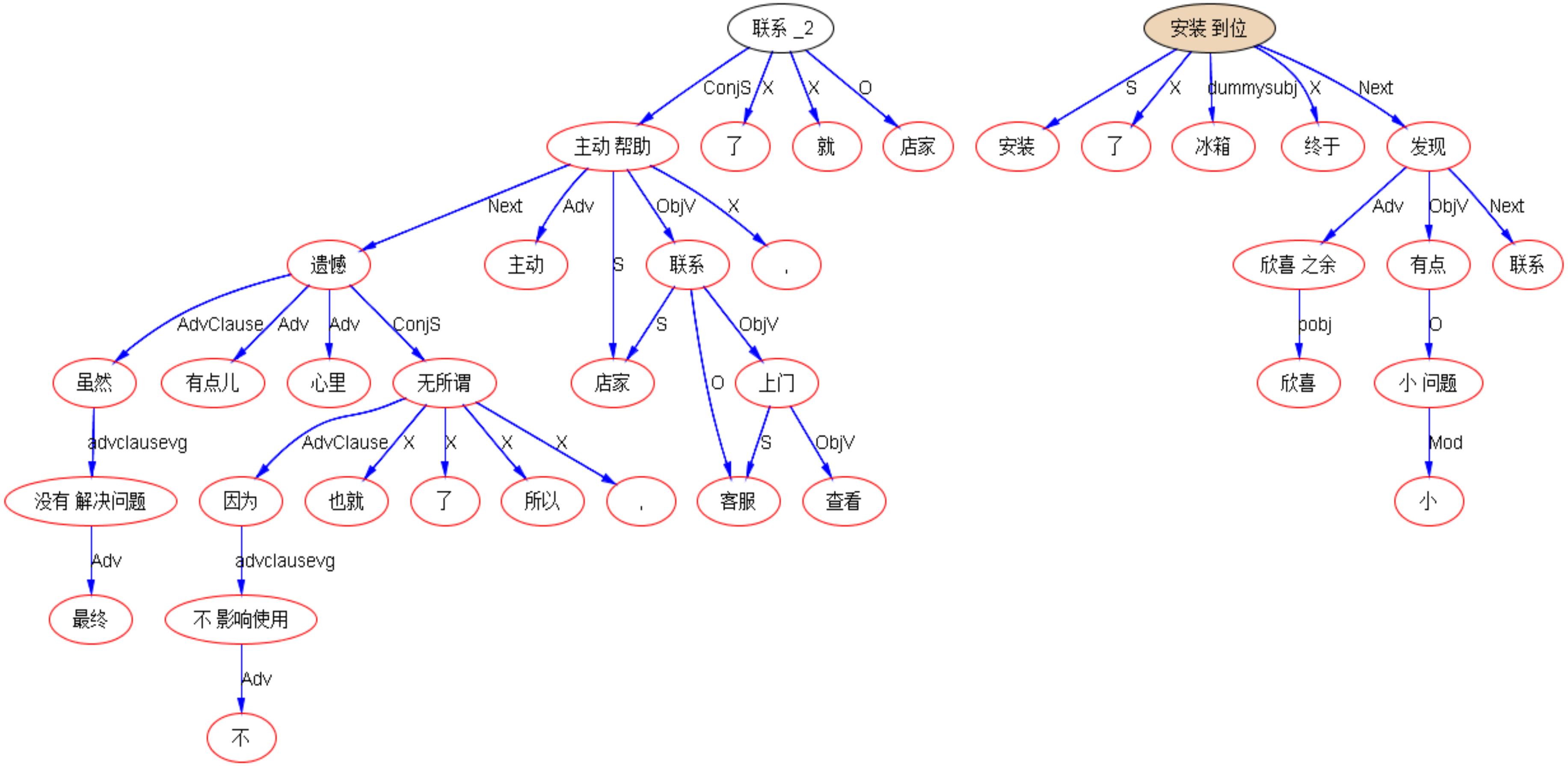


SVO Pattern:

Barack Obama (S)
Endorse (V)
Hillary Clinton (O)

Knowledge Graph

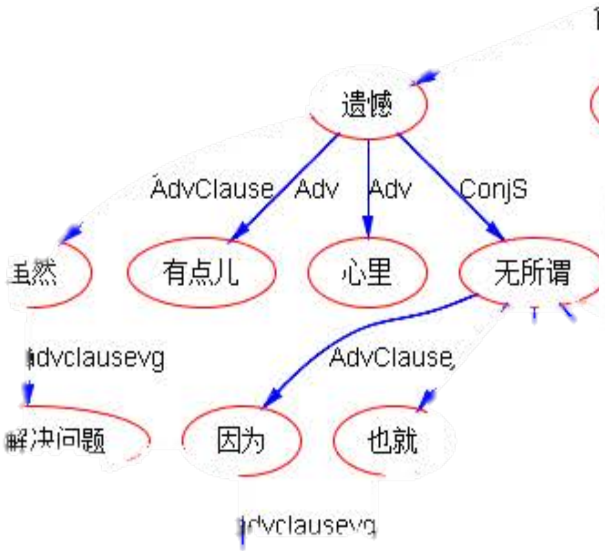
Deep Parsing: Unstructured to Structures

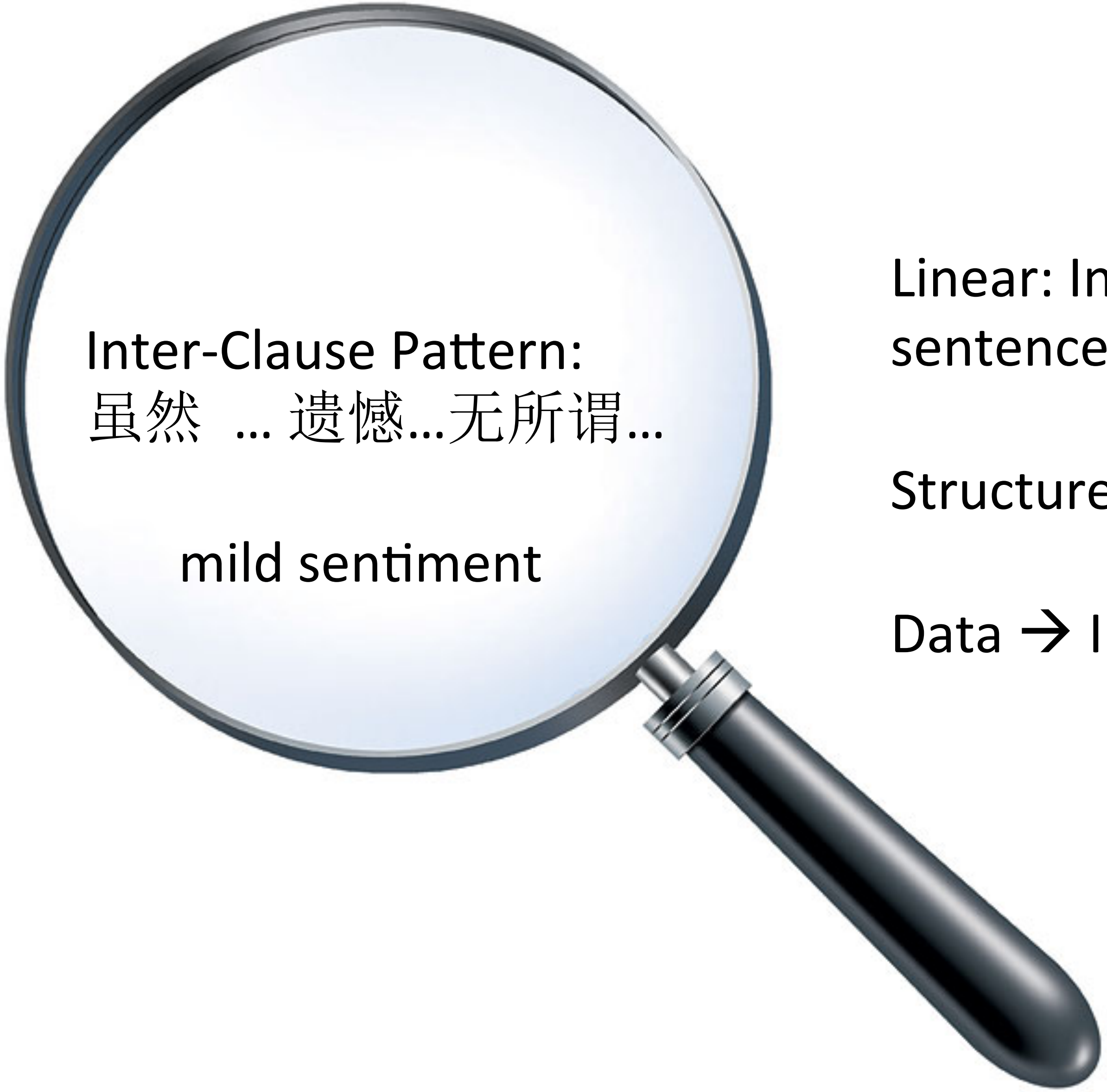


终于冰箱安装到位了，欣喜之余发现有点儿小问题，就联系了店家，店家主动帮助联系客服上门查看，虽然最终没有解决问题，心里有点儿遗憾，但是因为不影响使用，所以也就无所谓了.

NLP频道
liweinlp.com

Subtree Pattern: Data to Intelligence





Inter-Clause Pattern:
虽然 … 遗憾…无所谓…

mild sentiment

Linear: Infinite number of sentences

Structure: Limited patterns

Data → Intelligence

NLP频道
liweinlp.com

Outline: NLP Architectures

- 人工智能的历史和现状简介：从感知到认知

  此消彼长的经验主义和理性主义钟摆

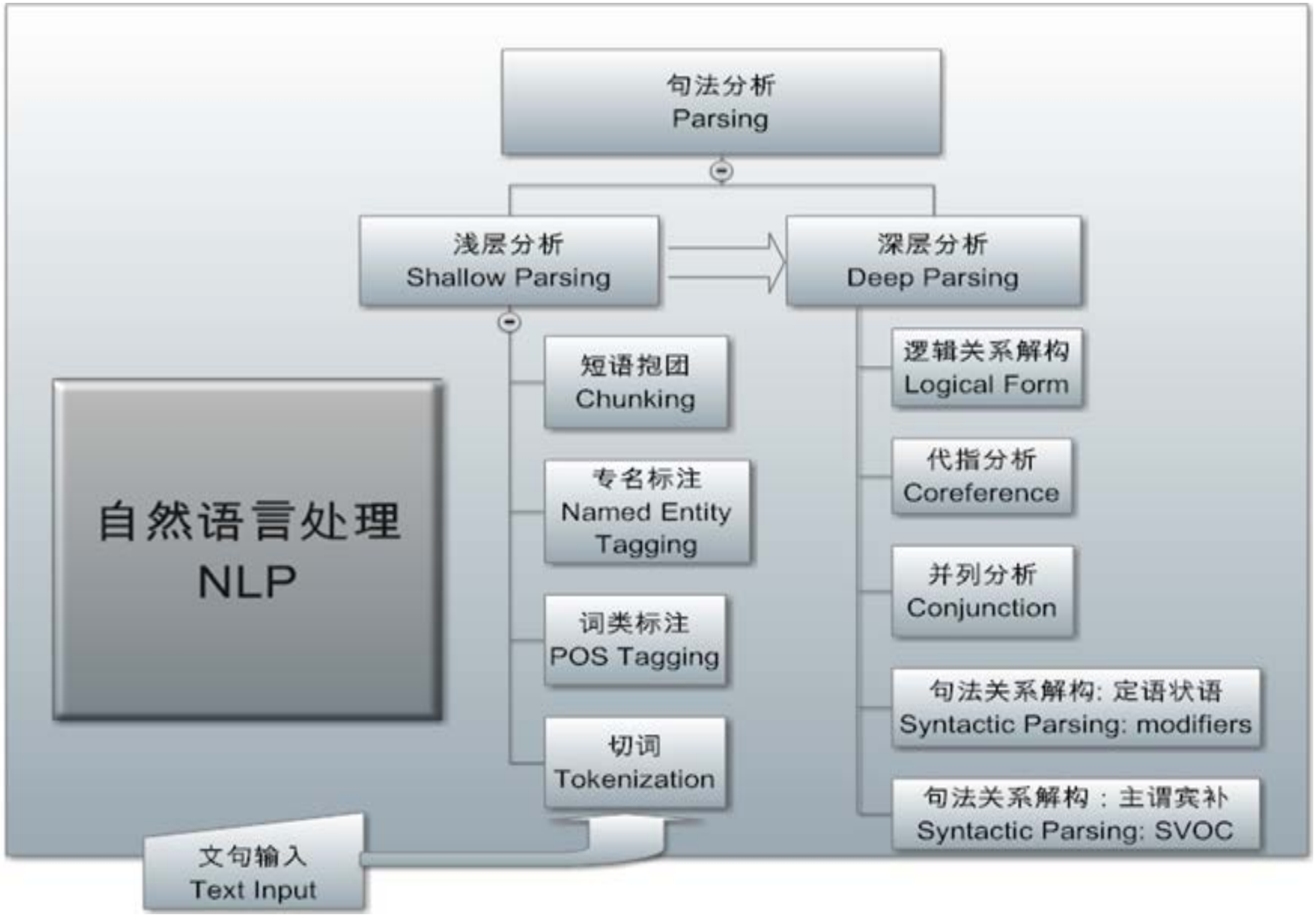
- 深度解析（Deep Parsing）是什么?

- NLP 架构纵览

NLP Architecture 1: Deep Parser as Core

Cascaded FSAs break through Chomsky’s hierarchy walls

Robust, linear, F-measure: scale up to big data



句法分析
Parsing

浅层分析
Shallow Parsing

深层分析
Deep Parsing

短语抱团
Chunking

专名标注
Named Entity Tagging

词类标注
POS Tagging

切词
Tokenization

逻辑关系解构
Logical Form

代指分析
Coreference

并列分析
Conjunction

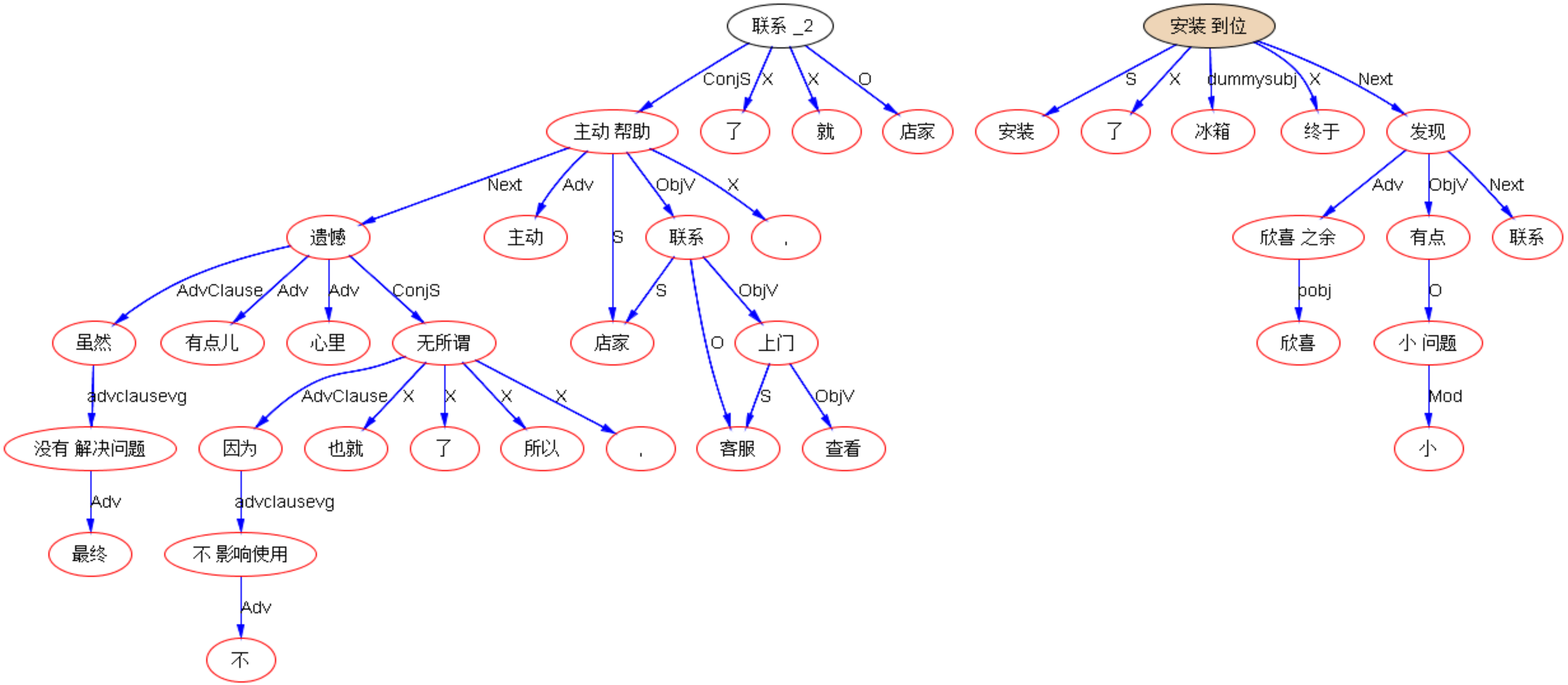
句法关系解构：定语状语
Syntactic Parsing: modifiers

句法关系解构：主谓宾补
Syntactic Parsing: SVOC

自然语言处理
NLP

文句输入
Text Input

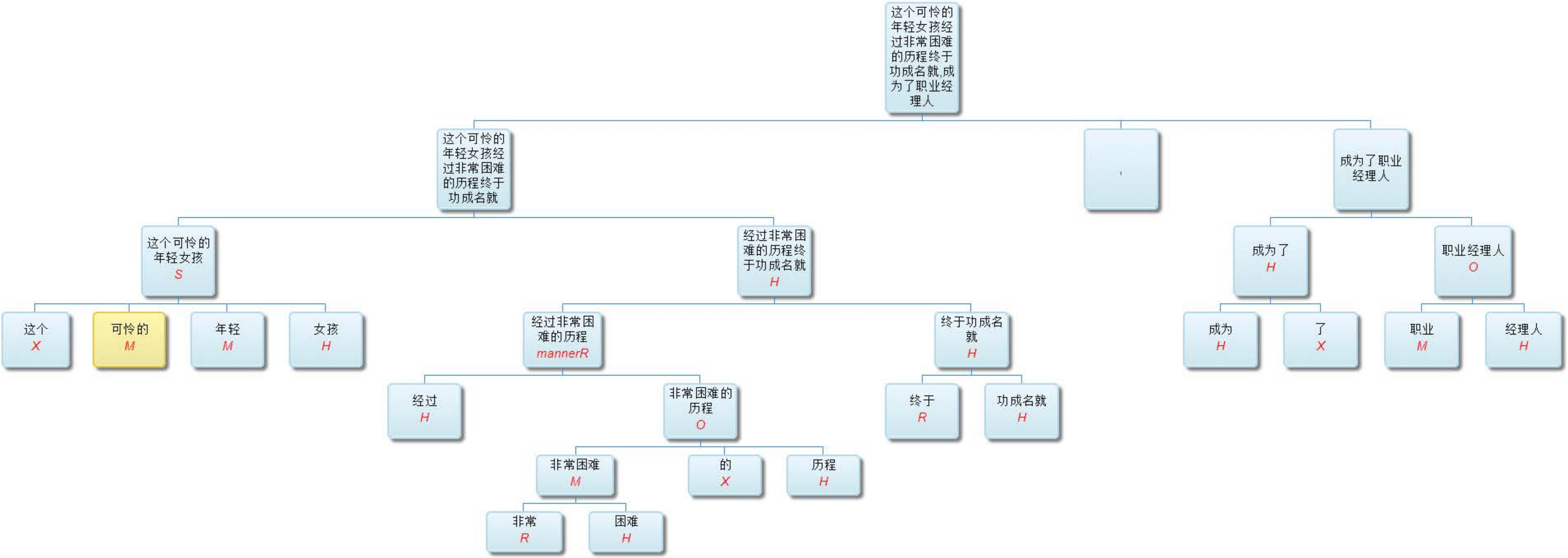
Sample Deep Parse Tree (dependency)



终于冰箱安装到位了，欣喜之余发现有点儿小问题，就联系了店家，店家主动帮助联系客服上门查看，虽然最终没有解决问题，心里有点儿遗憾，但是因为不影响使用，所以也就无所谓了.

NLP频道
liweinlp.com

Sample Deep Parse Tree (PS flavor)

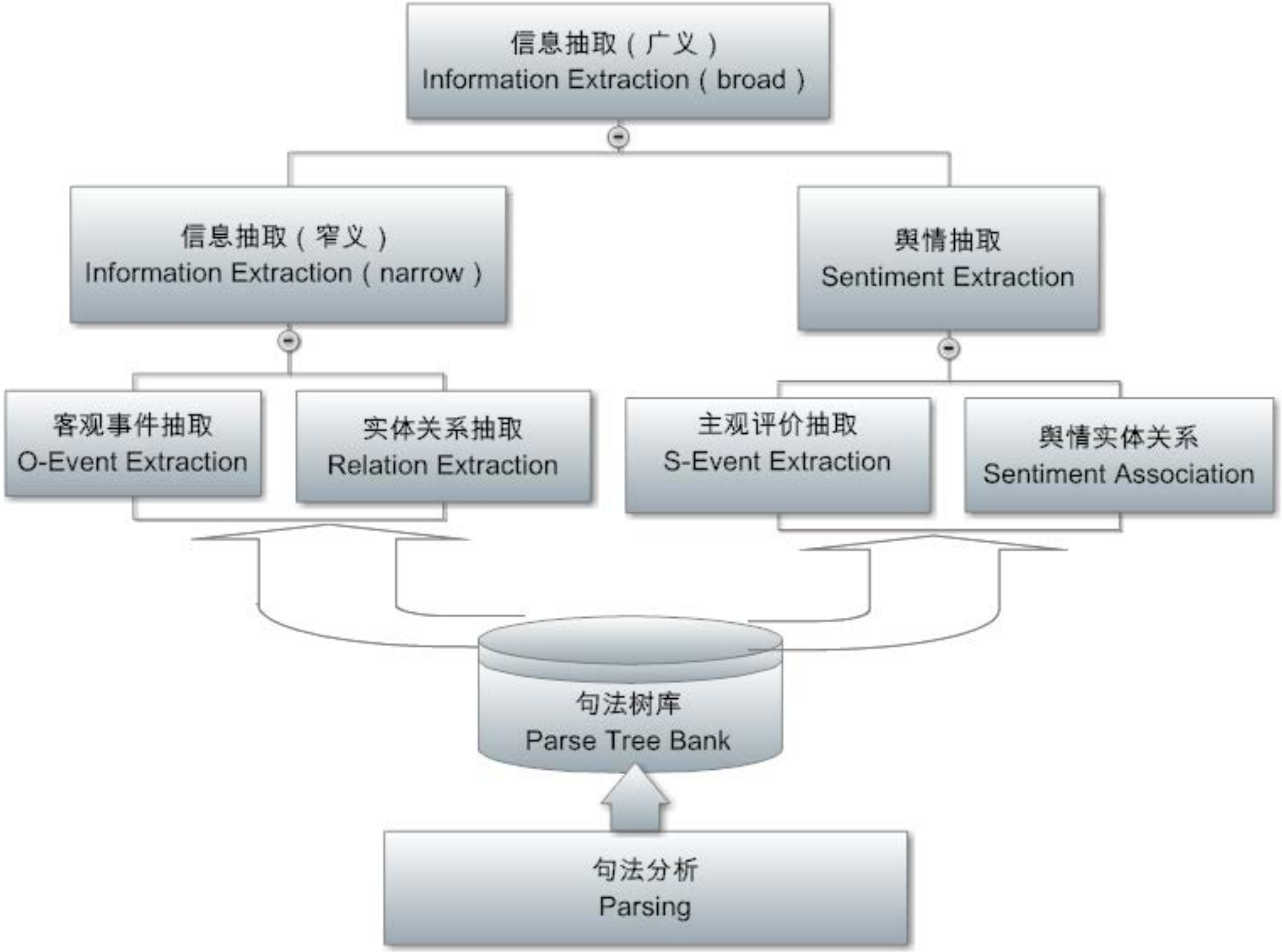


NLP频道
liweinlp.com

NLP Architecture 2: Information Extraction

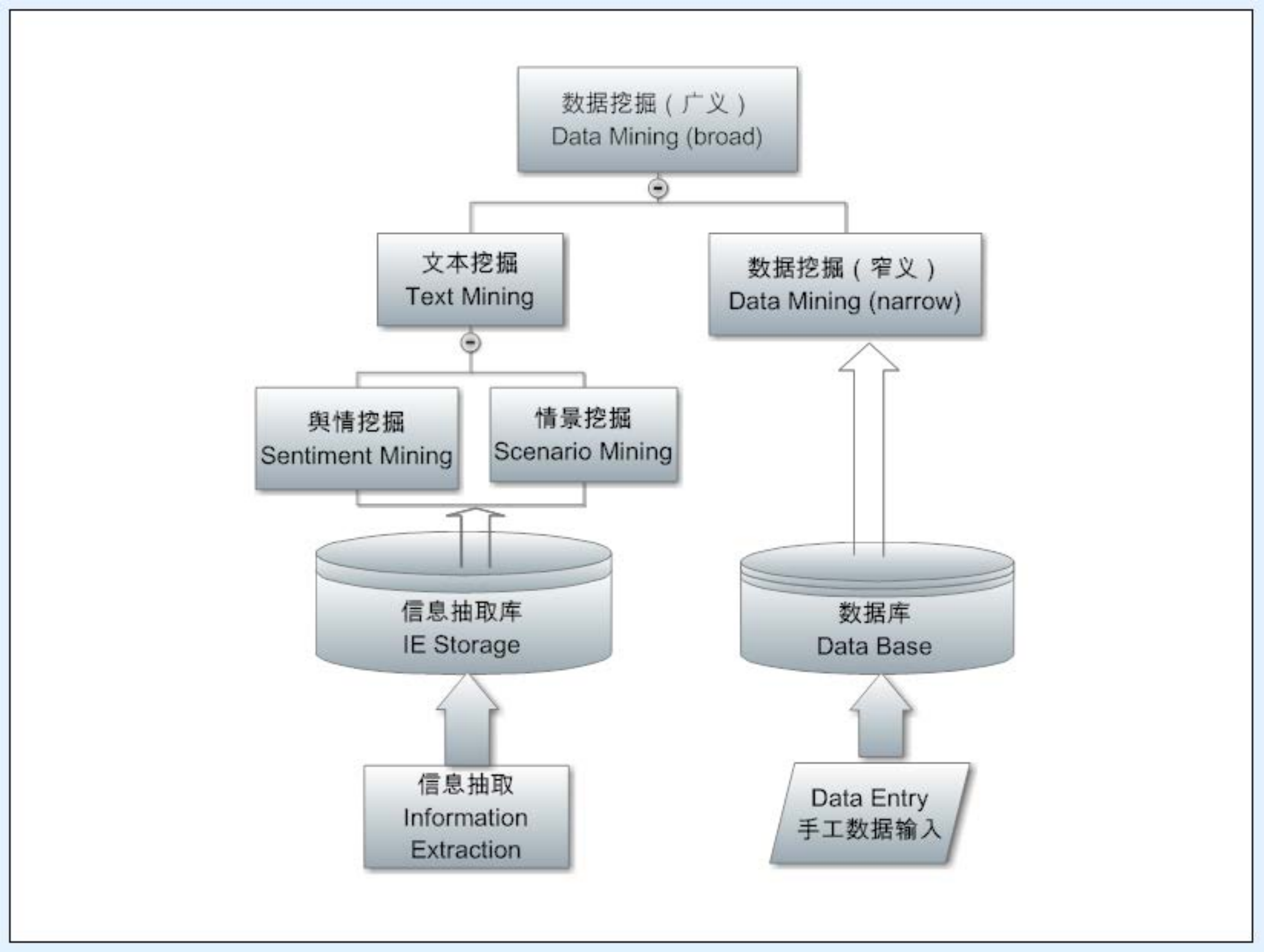
京东

Including sentiment analysis （on subjective language）



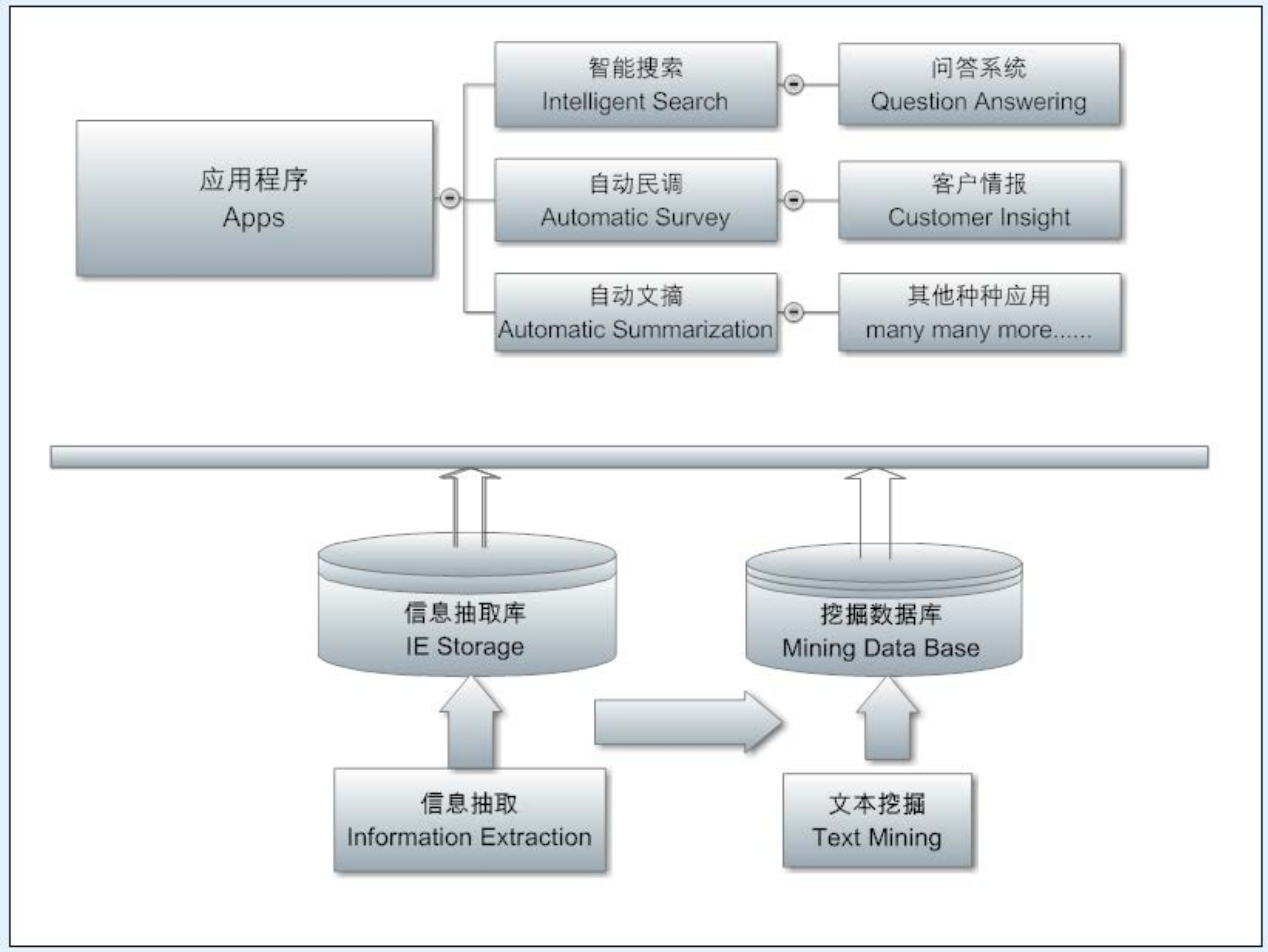
NLP频道
liweinlp.com

NLP Architecture 3: Text Mining



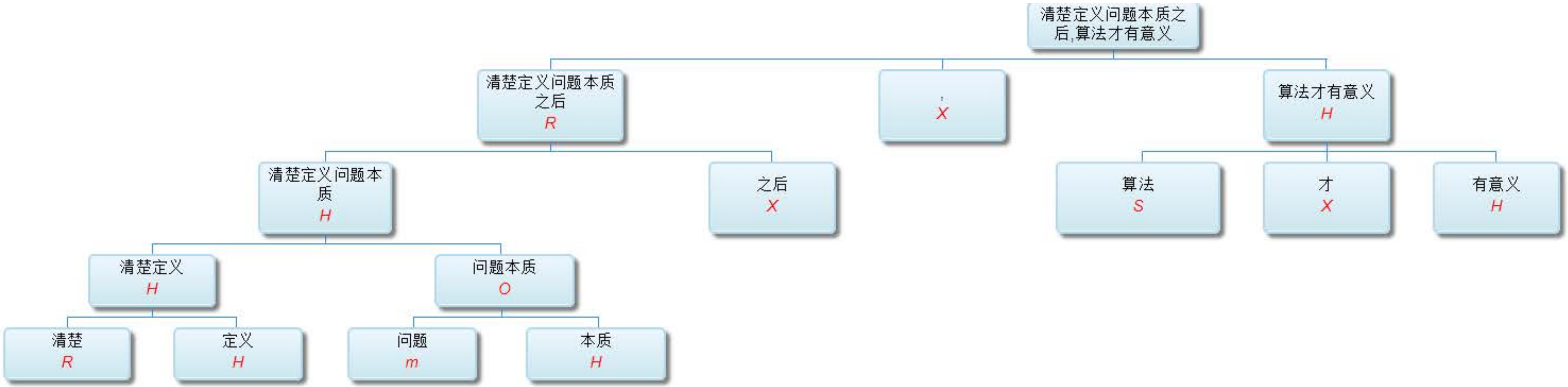
NLP频道
liweinlp.com

NLP Architecture 4: Landing on Applications



NLP频道
liweinlp.com

Sample Deep Parse Tree



清楚定义问题本质之后,算法才有意义

清楚定义问题本质之后 R

, X

算法才有意义 H

清楚定义问题本质 H

之后 X

算法 S

才 X

有意义 H

清楚定义 H

问题本质 O

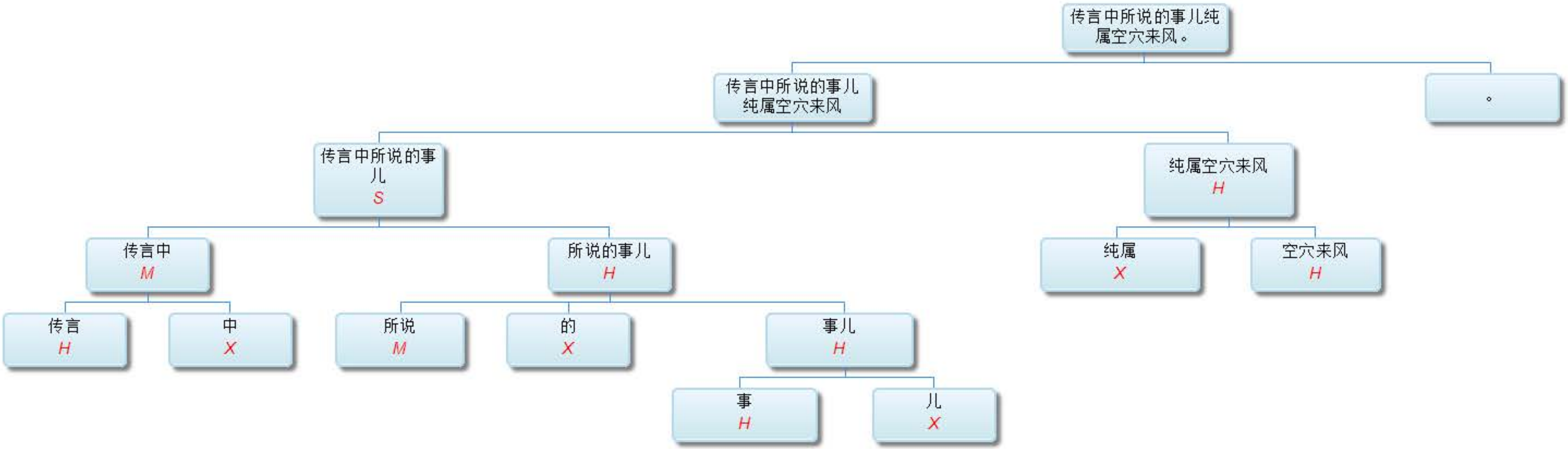
清楚 R

定义 H

问题 m

本质 H

Sample Deep Parse Tree



NLP频道
liweinlp.com

Outline: NLP Applications

- 人工智能的历史和现状简介：从感知到认知

  此消彼长的经验主义和理性主义钟摆

- 深度解析（Deep Parsing）是什么?

- NLP 架构纵览

- 核武器应用举例

  社媒舆情分析，大数据挖掘，智能搜索，对话系统 …………

Sentiment Analysis

Why deep parsing, not deep learning?

Learning without parsing does not work for social media sentiment analysis

- *Social media is dominated by short messages*
- *Statistical learning breaks in short messages: no sufficient data points*
- *Deep parsing enables linguistic analysis for best precision*
- *Deep parsing enables insights mining 2 magnitudes more efficient*
  - *parsing-supported rule has power of about 100 ngram rules*
- ***Deep learning is a great algorithm but still delinked from parsing***
  - *Parsers trained by deep learning are all research systems*
    - *difficult to adapt to real life text of social media (or other genres)*
    - *knowledge bottleneck: domains where labeled data are insufficient*
  - *Real life deep learning systems are mostly end-to-end, still no structures*

Sentiment Analysis: Bag of Words vs. Parsing 京东

KEYWORD CHALLENGE

The iPhone has never been good.

The iPhone has never been this good.

ASSOCIATION CHALLENGE

Another reason to switch from Visa to MasterCard

I prefer MasterCard over Visa.

MasterCard is way better than Visa.

CLASSIFICATION CHALLENGE

I had a wonderful day today. Even my instant coffee tastes great. However my Dell laptop doesn't boot again. Maybe I should check out the MacBook. It [MacBook] seems so easy to use.

Coarse-grained Classification thumbs-up and down: overall tone positive (3 vs 1)

Fine-grained Analysis uncovers "why" behind sentiments:

(1) Instant coffee / tastes great
(2) Dell Laptop / does not boot
(3) Macbook / easy to use

Deep Parsing Supports Deep Sentiments

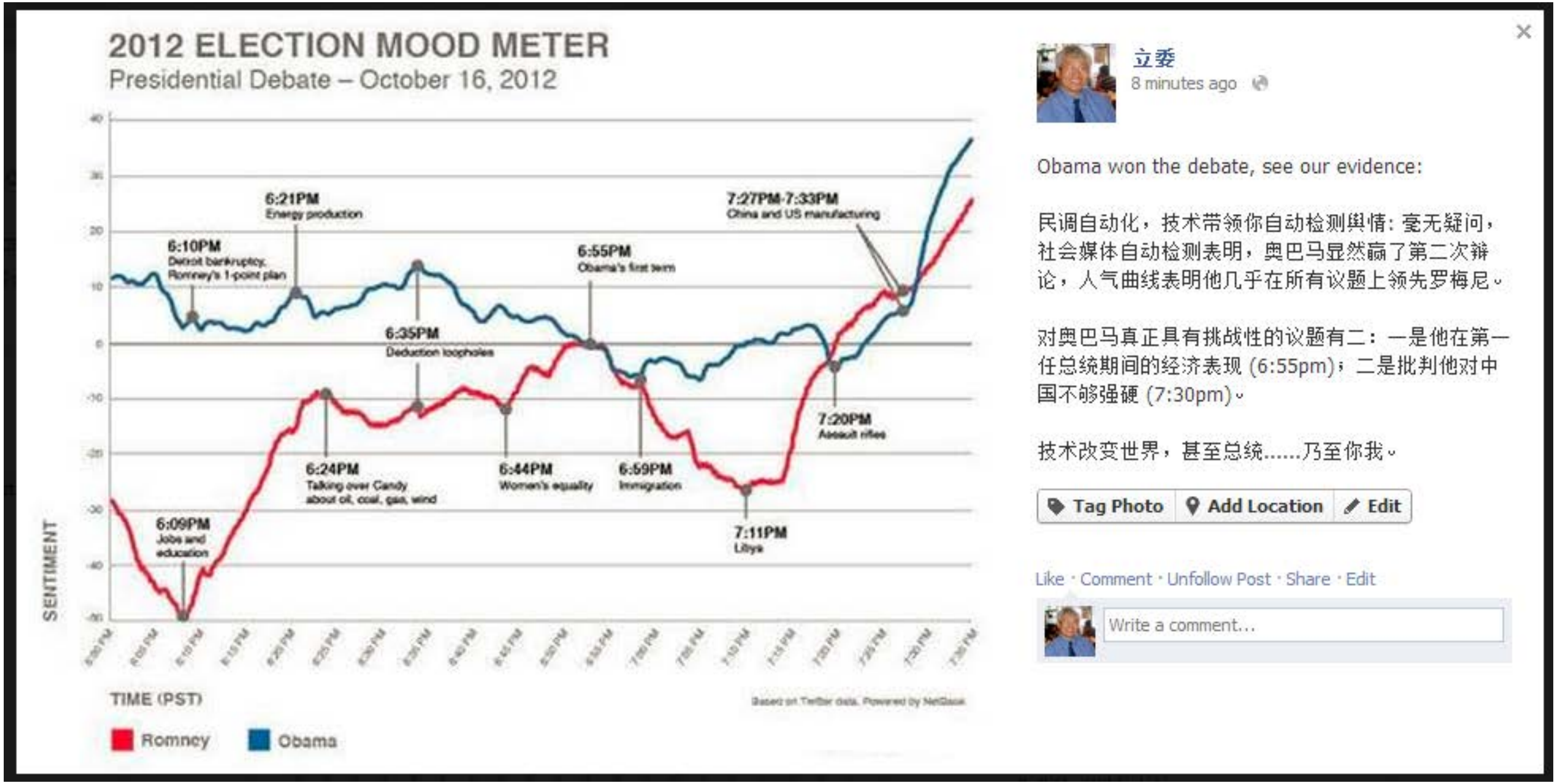
京东

Sentiment analysis has different layers

1. sentiment classification: thumbs-up and down (or neutral)
2. sentiment association: to associate a sentiment with a topic or brand as its object
3. deep sentiment insights:
   - (i) who has the sentiment?
   - (ii) how intense?
   - (iii) why?
   - (iv) Evaluations, comparisons and contrasts;
   - (v) needs and wish-list;
   - (vi) positive/negative actions (e.g. adopt / abandon);
   - (vii) purchase intent;
   - (viii) pros and cons

Most learning systems stop at 1 and sometimes at 2. All 3 can be done via deep parsing.

Illustration: Real-time Polls



Obama won the debate, see our evidence:

民调自动化，技术带领你自动检测舆情：毫无疑问，社会媒体自动检测表明，奥巴马显然赢了第二次辩论，人气曲线表明他几乎在所有议题上领先罗梅尼。

对奥巴马真正具有挑战性的议题有二：一是他在第一任总统期间的经济表现 (6:55pm)；二是批判他对中国不够强硬 (7:30pm)。

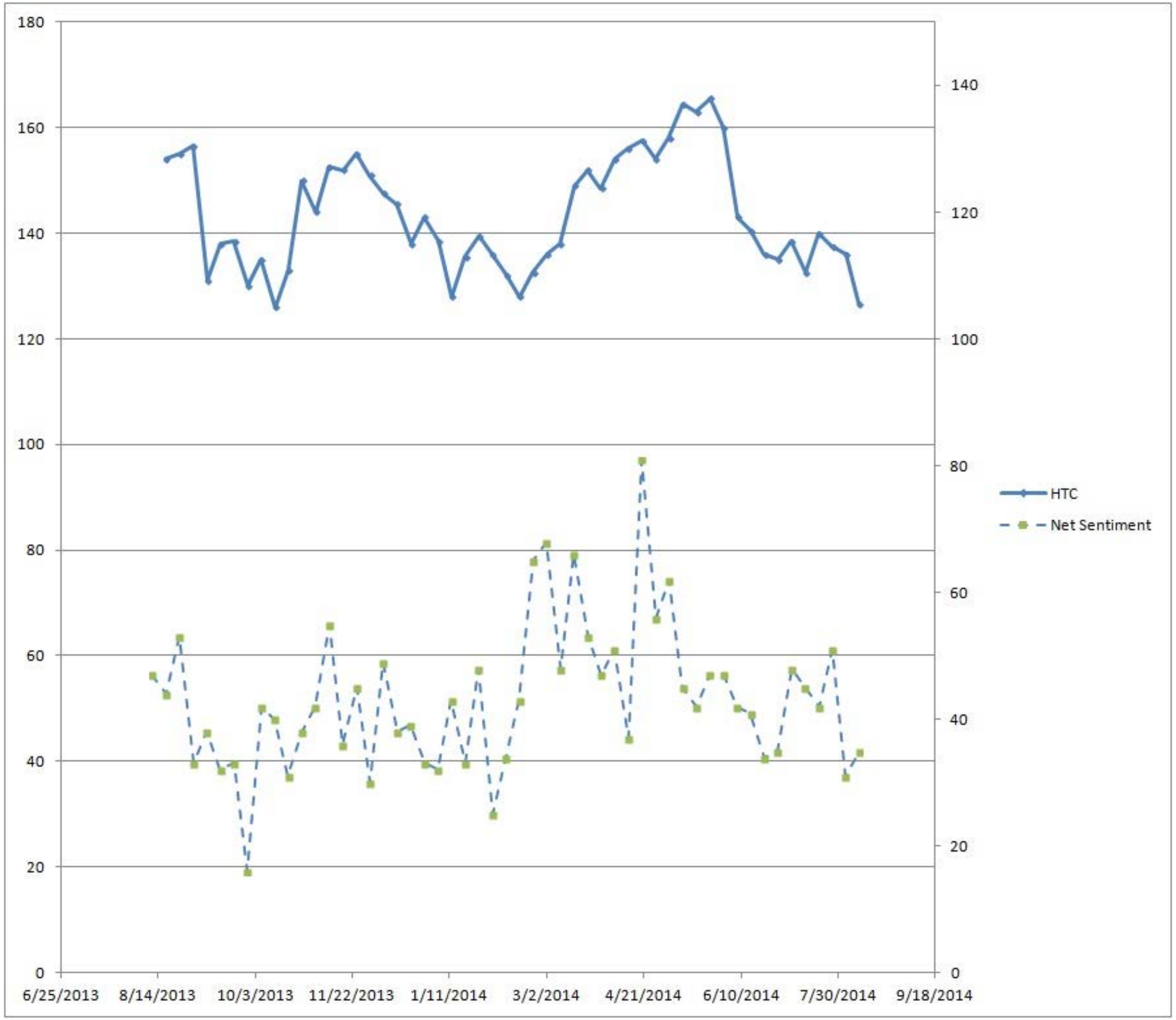
技术改变世界，甚至总统......乃至你我。

Challenges observed:

economy topic
at 6:55pm;

China topic
at 7:30pm

Illustration: Stock Market Trends



Topic: HTC

Data 1: Stock Market Performance

Data 2: Chinese social media (Weibo, Tianya, Facebook, Twitter…)

Time range: 2013/08 – 2014/08

Strong correlation observed

NLP频道
liweinlp.com

Big Data Mining: Who benefits?

For businesses: social listening

- Consumer in:sights: sentiments and why
- Brand image: trends
- Competitive research: where do we stand

For consumers

- Purchase decision
- Personalized service

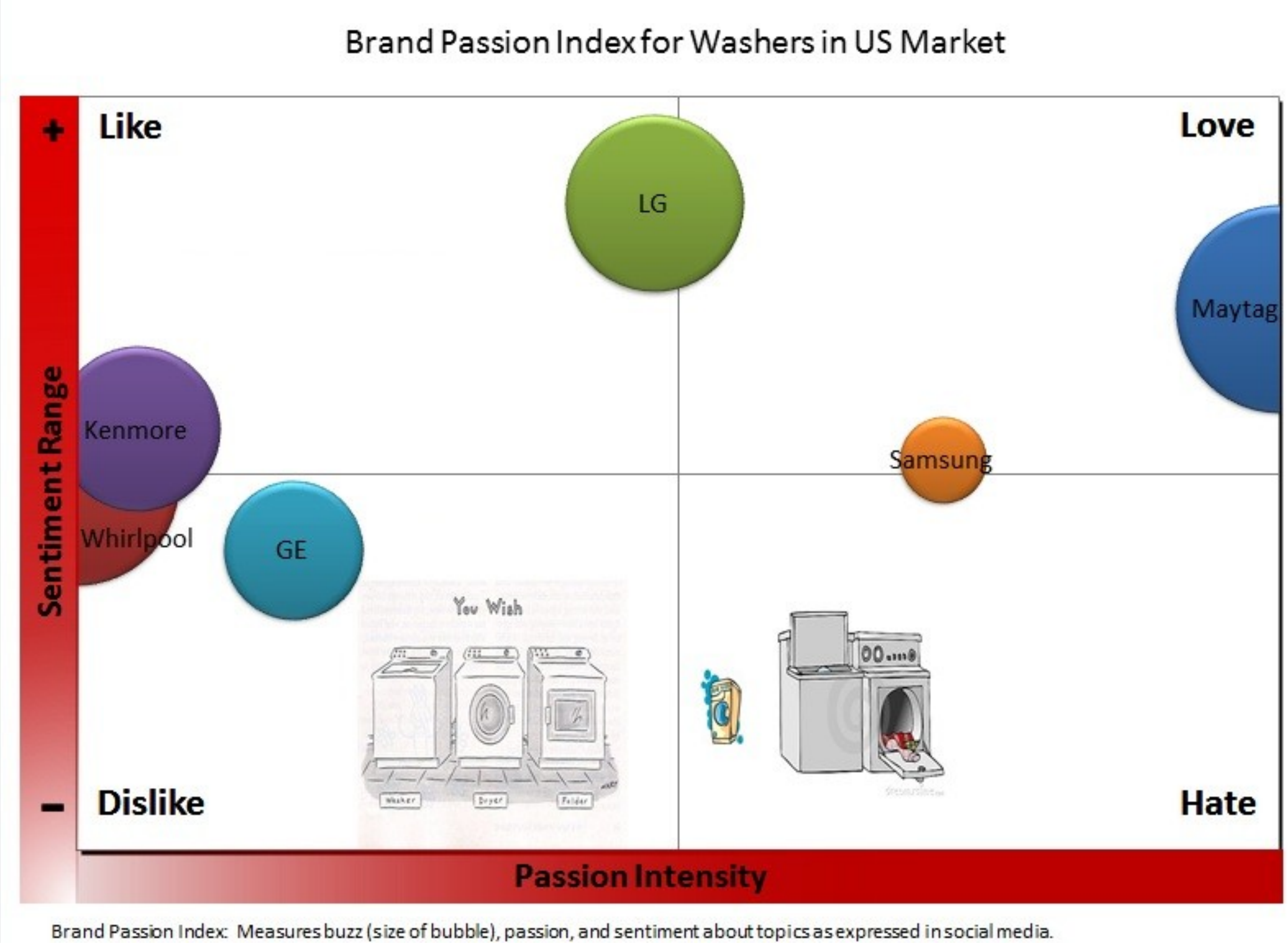
For government

- Election campaign
- Public opinions on policies and social topics

Others?

- Hot topics or anywhere public opinions are involved
- Stock market trends correlation

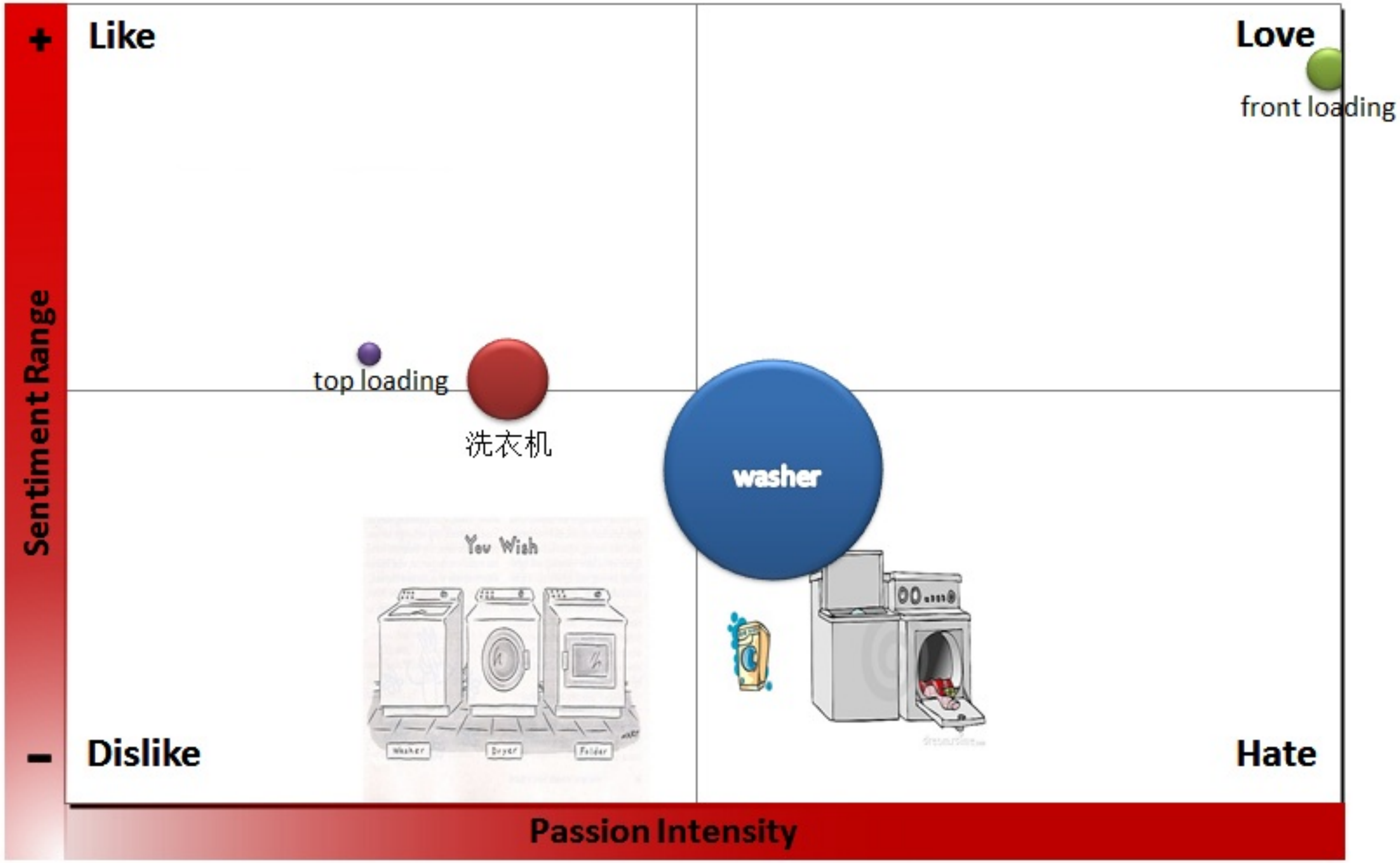
For consumers: Purchase Decision



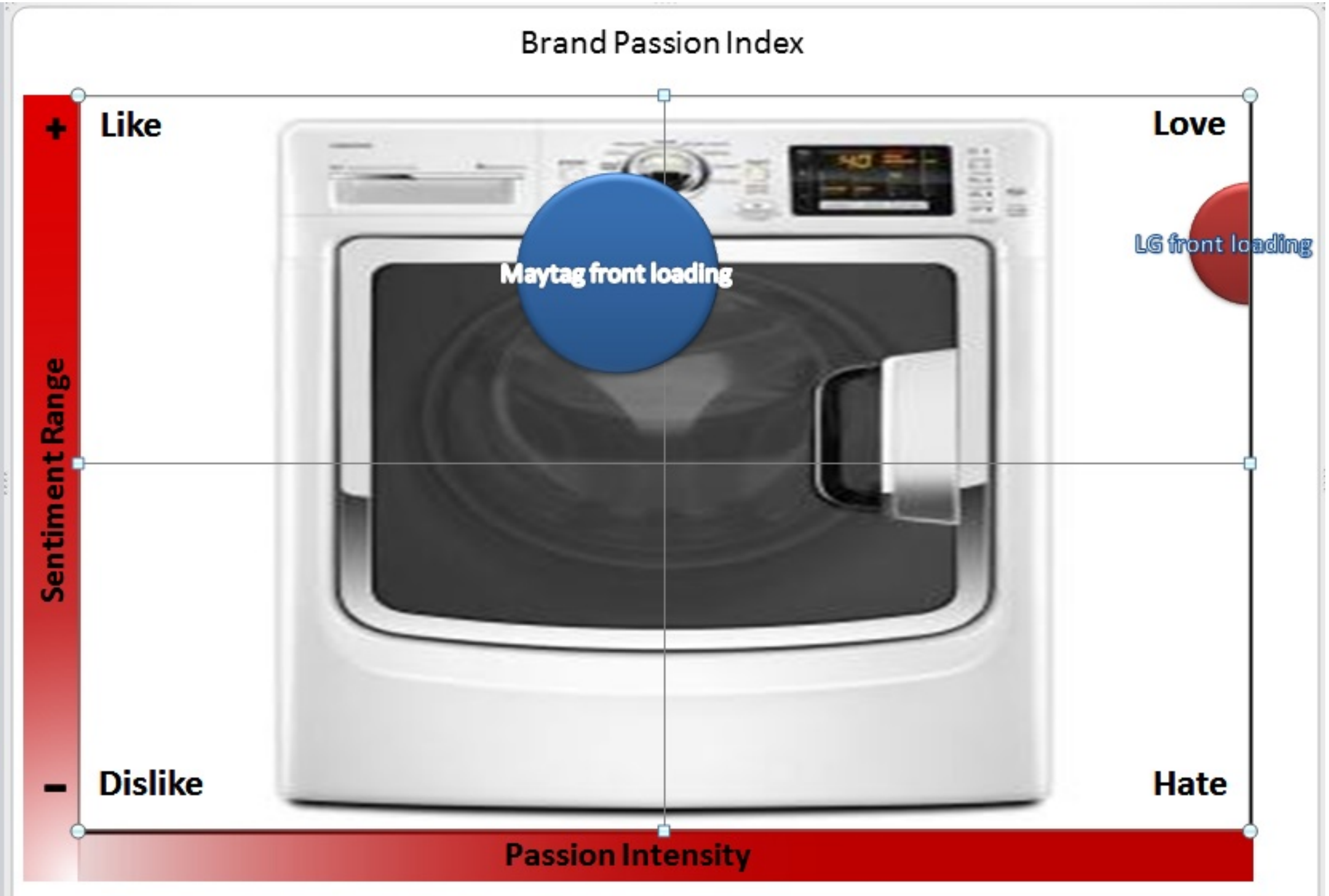
NLP频道
liweinlp.com

For consumers: Purchase Decision

Brand Passion Index for Two Types of Washers



For consumers: Purchase Decision



NLP频道
liweinlp.com

Intelligent Search and Chatbots

Three types of Chatbots:

1. Domain knowledge QA: e.g. customer service;
2. Open domain knowledge QA: e.g. Who won Nobel Prize in 2015?
3. Interactive chatting: e.g. just for fun (killing time); in time, for comfort (senior people); for mental health counselling

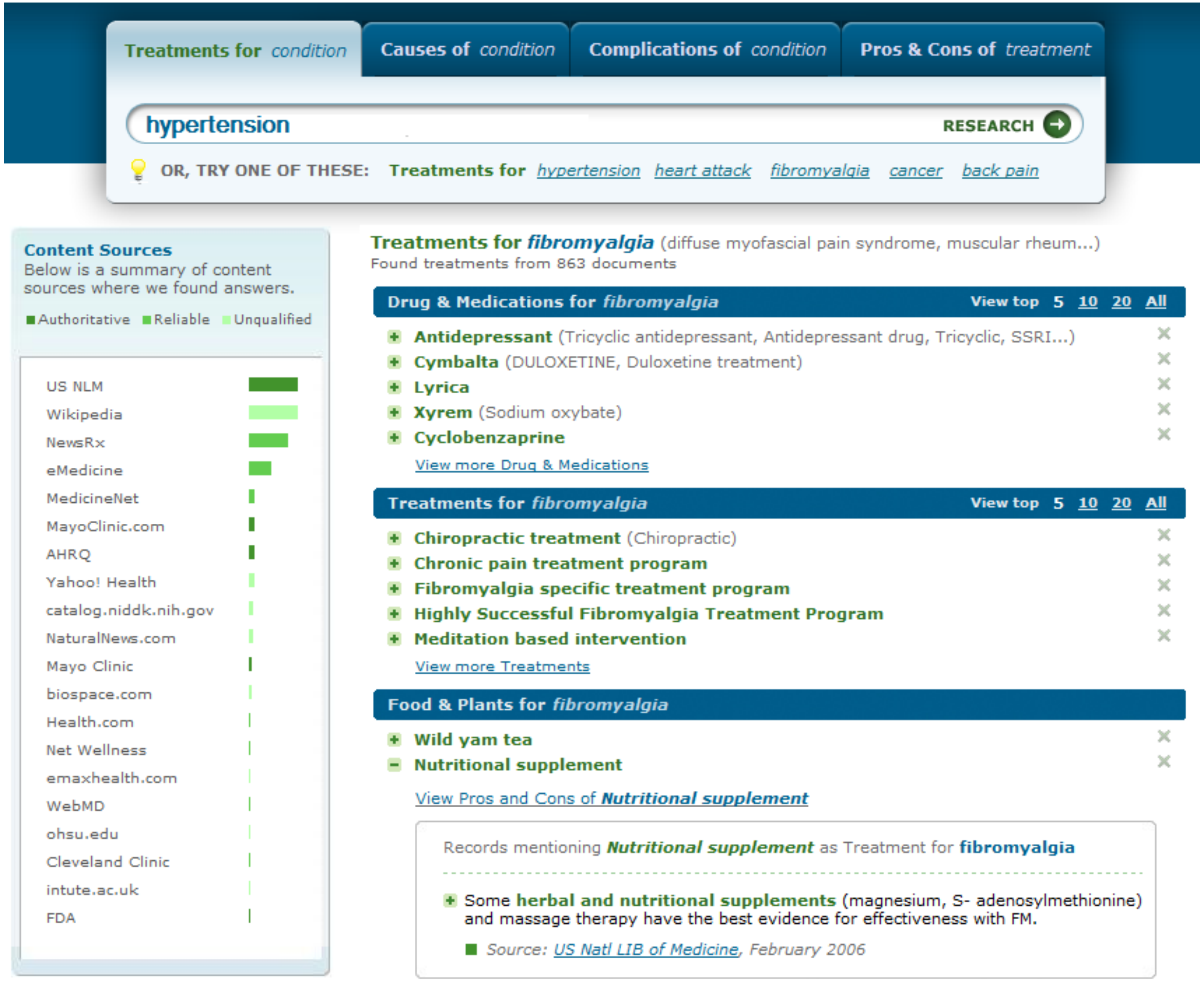
Q: questions are a subset of language, tractable
for decoding intent, asking point, and hidden slots

A: 1 and 2 can be based on Knowledge Graph enabled by deep parser;
3 can be enabled by learning from human chats plus parsing

A mixture/convergence of 3 is possible

Apply NLP to Verticals: Medicine Domain

京东



Treatments for *condition* | Causes of *condition* | Complications of *condition* | Pros & Cons of *treatment*

hypertension RESEARCH

OR, TRY ONE OF THESE: Treatments for hypertension heart attack fibromyalgia cancer back pain

Content Sources
Below is a summary of content sources where we found answers.

■Authoritative ■Reliable ■Unqualified

US NLM
Wikipedia
NewsRx
eMedicine
MedicineNet
MayoClinic.com
AHRQ
Yahoo! Health
catalog.niddk.nih.gov
NaturalNews.com
Mayo Clinic
biospace.com
Health.com
Net Wellness
emaxhealth.com
WebMD
ohsu.edu
Cleveland Clinic
intute.ac.uk
FDA

Treatments for *fibromyalgia* (diffuse myofascial pain syndrome, muscular rheum...)
Found treatments from 863 documents

Drug & Medications for *fibromyalgia* View top 5 10 20 All

- **Antidepressant** (Tricyclic antidepressant, Antidepressant drug, Tricyclic, SSRI...)
- **Cymbalta** (DULOXETINE, Duloxetine treatment)
- **Lyrica**
- **Xyrem** (Sodium oxybate)
- **Cyclobenzaprine**

View more Drug & Medications

Treatments for *fibromyalgia* View top 5 10 20 All

- **Chiropractic treatment** (Chiropractic)
- **Chronic pain treatment program**
- **Fibromyalgia specific treatment program**
- **Highly Successful Fibromyalgia Treatment Program**
- **Meditation based intervention**

View more Treatments

Food & Plants for *fibromyalgia*

- **Wild yam tea**
- **Nutritional supplement**

View Pros and Cons of ***Nutritional supplement***

Records mentioning ***Nutritional supplement*** as Treatment for **fibromyalgia**

Some **herbal and nutritional supplements** (magnesium, S- adenosylmethionine) and massage therapy have the best evidence for effectiveness with FM.

Source: US Natl LIB of Medicine, February 2006

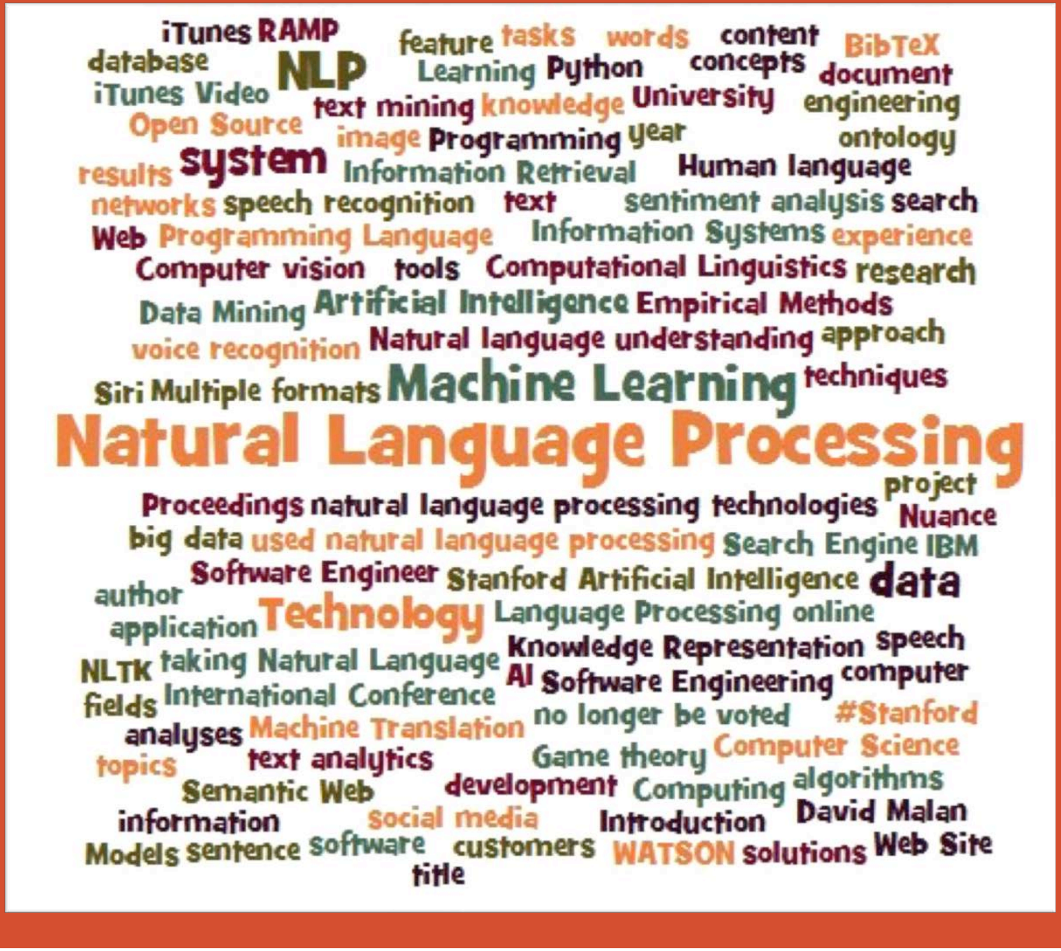
Some Big Data Verticals:

1. News
2. Social Media

3. Medicine
4. Legal
5. Education
6. Financing

7. Multilingual

NLP频道
liweinlp.com



And we are hiring!

At Beijing & Silicon Valley

NLP频道:
liweinlp.com
www.linkedin.com/in/liwei4nlp

京东