文本大数据分析与挖掘：机遇，挑战，及应用前景

Analysis and Mining of Big Text Data: Opportunities, Challenges, and Applications

ChengXiang Zhai (翟成祥)
Department of Computer Science
University of Illinois at Urbana-Champaign
USA

Text data cover all kinds of topics

Topics:
People
Events
Products
Services, …

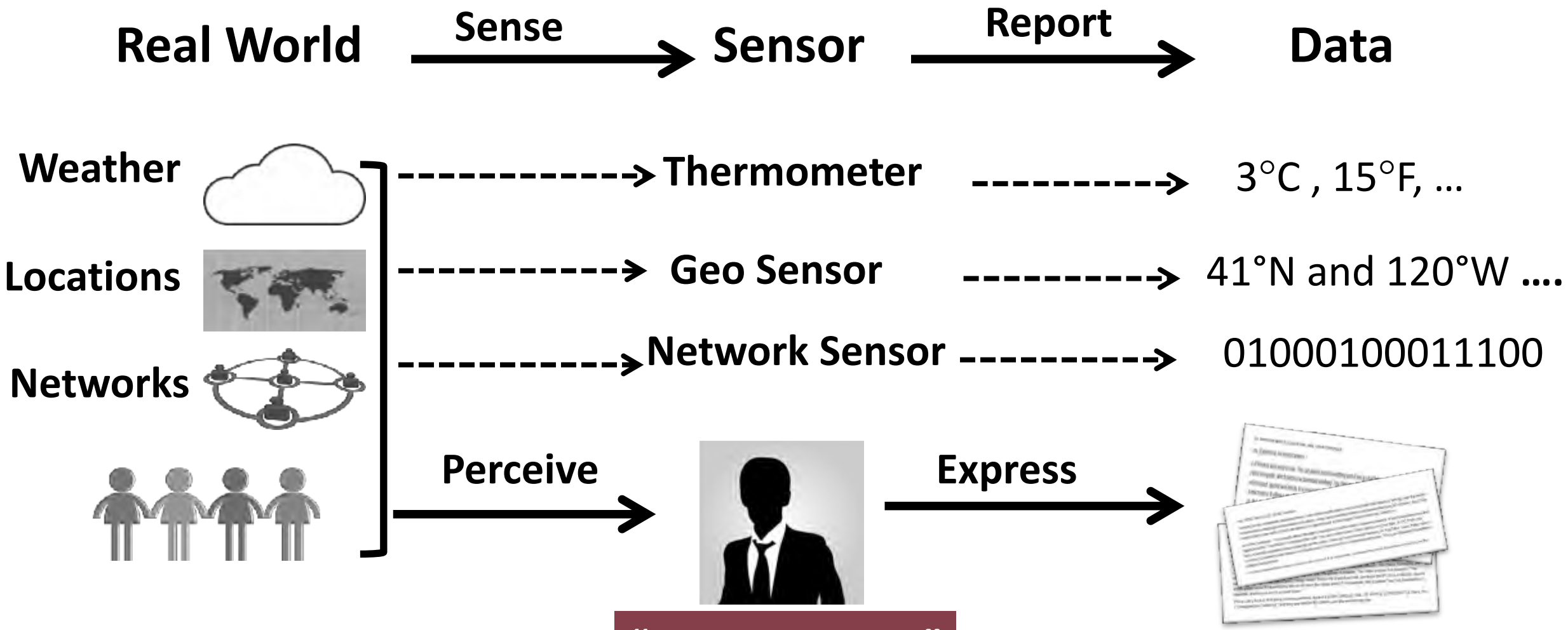
Sources:
Blogs
Microblogs
Forums
Reviews ,…

45M reviews **53M** blogs **1307M** posts **65M** msgs/day **115M** users **10M** groups

tripadvisor blogScope PREVIEW twitter YAHOO! GROUPS …

人= 主观智能“传感器”

Humans as Subjective & Intelligent “Sensors”



文本数据的特殊应用价值
Unique Value of Text Data

- 对所有大数据应用都有应用价值: Useful to all big data applications
- 特别有助于挖掘，利用有关人的行为，心态，观点的知识: Especially useful for mining knowledge about people's behavior, attitude, and opinions
- 直接表达知识；高质量数据（ Directly express knowledge about our world ）➔小文本数据应用 （Small text data are also useful!）

Data ➔ Information ➔ Knowledge

Text Data

Opportunities of Text Mining Applications

+ Non-Text Data

4. Infer other real-world variables (predictive analytics)

2. Mining content of text data

Real World

Observed World

Text Data

+ Context

Perceive

(Perspective)

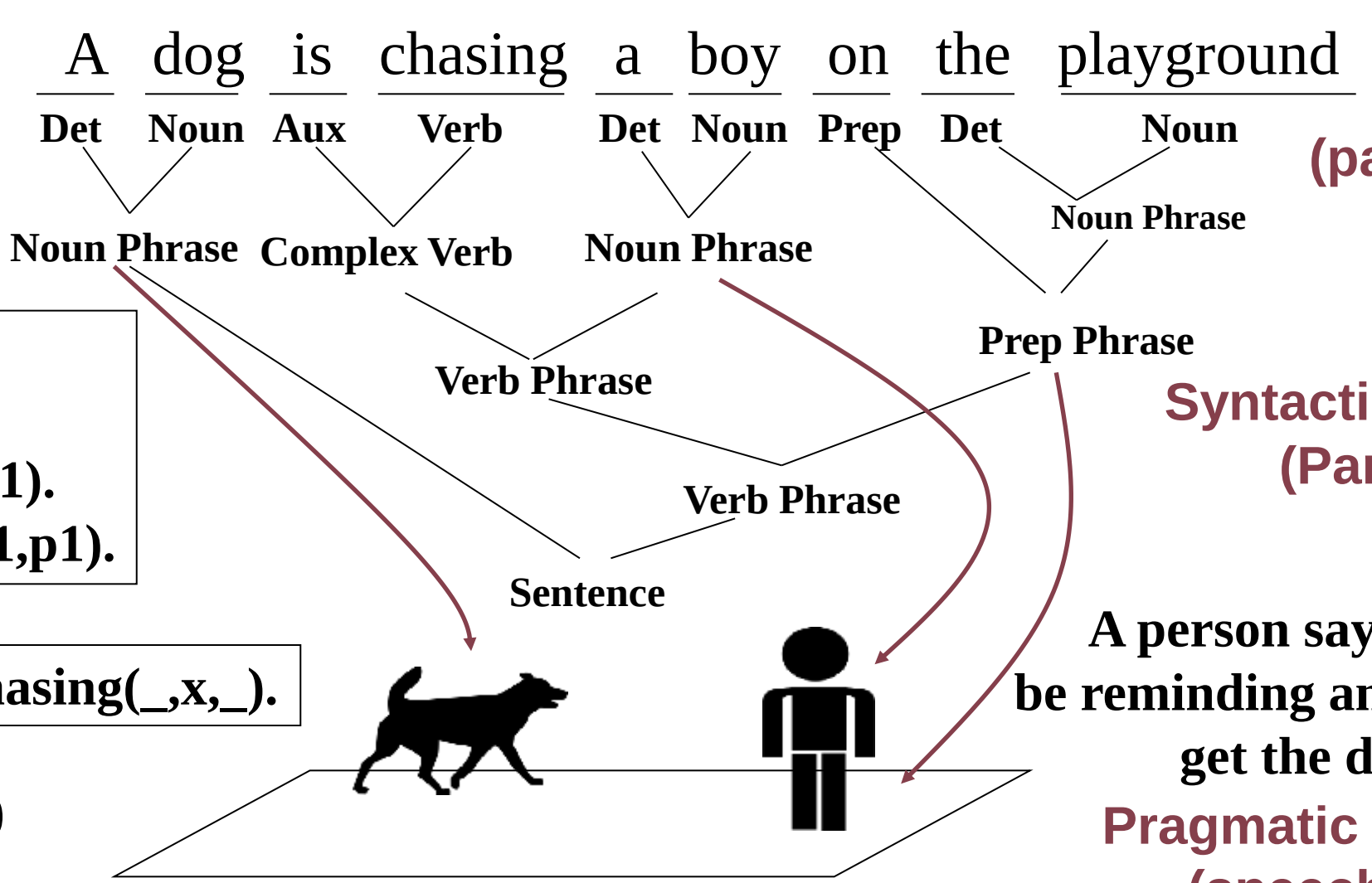
Express

(English)

3. Mining knowledge about the observer

1. Mining knowledge about language

Challenges in Understanding Text Data (NLP)



Lexical analysis (part-of-speech tagging)

Semantic analysis

Syntactic analysis (Parsing)

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)

Inference

A person saying this may be reminding another person to get the dog back.

Pragmatic analysis (speech act)

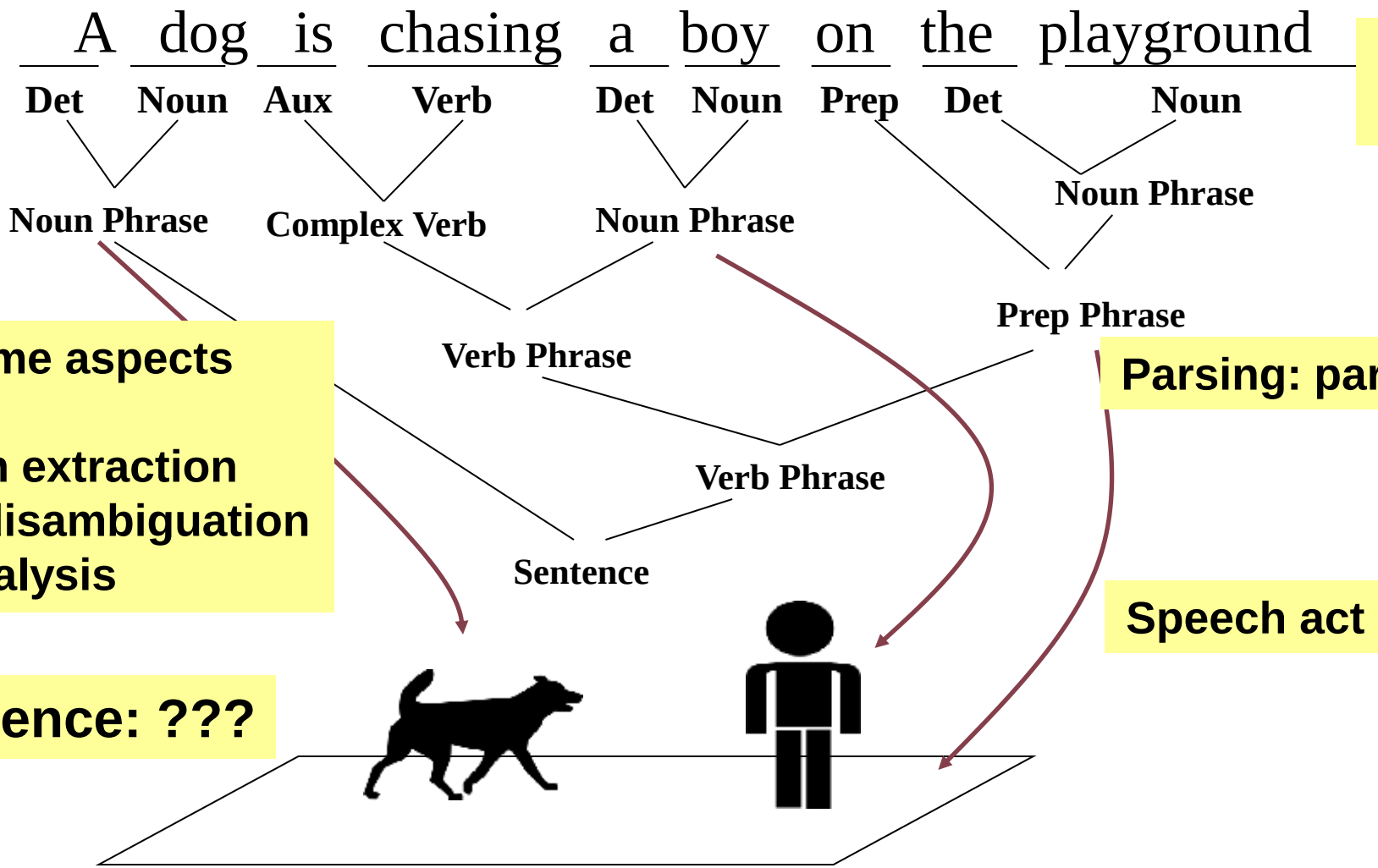
NLP is hard!

- Natural language is designed to make human communication efficient. As a result,
  - we omit a lot of *common sense* knowledge, which we assume the hearer/reader possesses.
  - we keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve.
- This makes EVERY step in NLP hard
  - Ambiguity is a *killer*!
  - Common sense reasoning is pre-required.

Examples of Challenges

- Word-level ambiguity:
  - "root" has multiple meanings (ambiguous sense)
- Syntactic ambiguity:
  - "natural language processing" (modification)
  - "A man saw a boy *with a telescope*." (PP Attachment)
- Anaphora resolution: "John persuaded Bill to buy a TV for *himself*." (himself = John or Bill?)
- Presupposition: "He has quit smoking" implies that he smoked before.

The State of the Art: Mostly Relying on Machine Learning



POS Tagging: 97%

Parsing: partial >90%(?)

Semantics: some aspects

- **Entity/relation extraction**
- **Word sense disambiguation**
- **Sentiment analysis**

Speech act analysis: ???

Inference: ???

Robust and general NLP tends to be *shallow* while *deep* understanding doesn't scale up.

Grand Challenge:

How can we leverage **imperfect** NLP to build a **perfect** application?

如何将不完善的技术转化为完善的产品？

Answer: Having humans in the loop!

优化人机合作！

文本数据镜拓宽人的感知

TextScope to enhance human perception

TextScope(文本数据镜)

Microscope

Telescope

集信息检索和文本分析挖掘与一体
支持交互式分析，决策支持

TextScope Interface & Major Text Mining Techniques

主题分析 观点分析 预测 推荐

TextScope

Task Panel

Topic Analyzer | Opinion | Prediction | ... | Event Radar

检索 过滤 分类

Search Box

MyFilter1

MyFilter2

...

Select Time

Select Region

My WorkSpace

Project 1

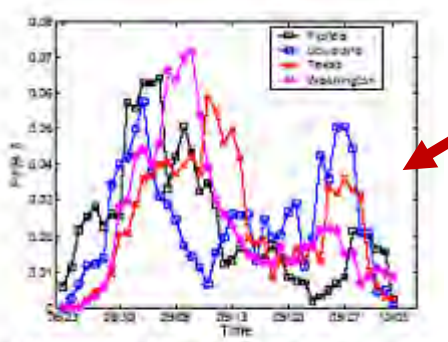
Alert A

Alert B ...

Microsoft (MSFT,) Google, **IBM** (IBM) and other cloud-computing rivals of Amazon Web Services are bracing for an AWS "partnership" announcement with VMware expected to be announced Thursday. …

摘要

RANK	DATE	EXCERPT
1	9/29/2000	Expect earning will be far below
2	12/8/2000	$4 billion cash in company
3	10/19/2000	Disappointing earning report
4	4/19/2001	Dow and Nasdaq soar after rate cut by Federal Reserve
5	7/20/2001	Apple's new retail store
...	...	...

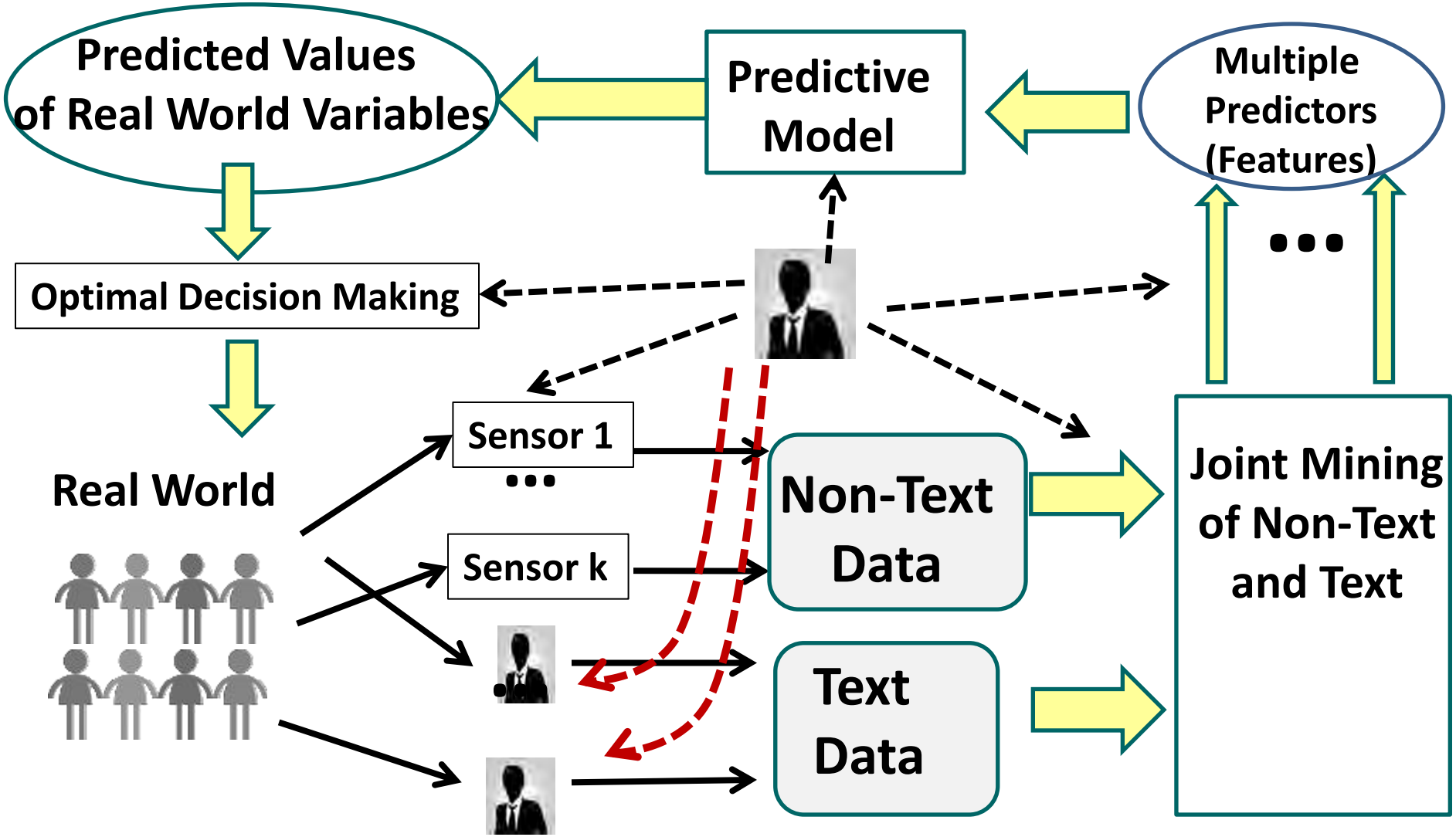


b) Theme "New Orleans" over states (Hurricane Katrina)

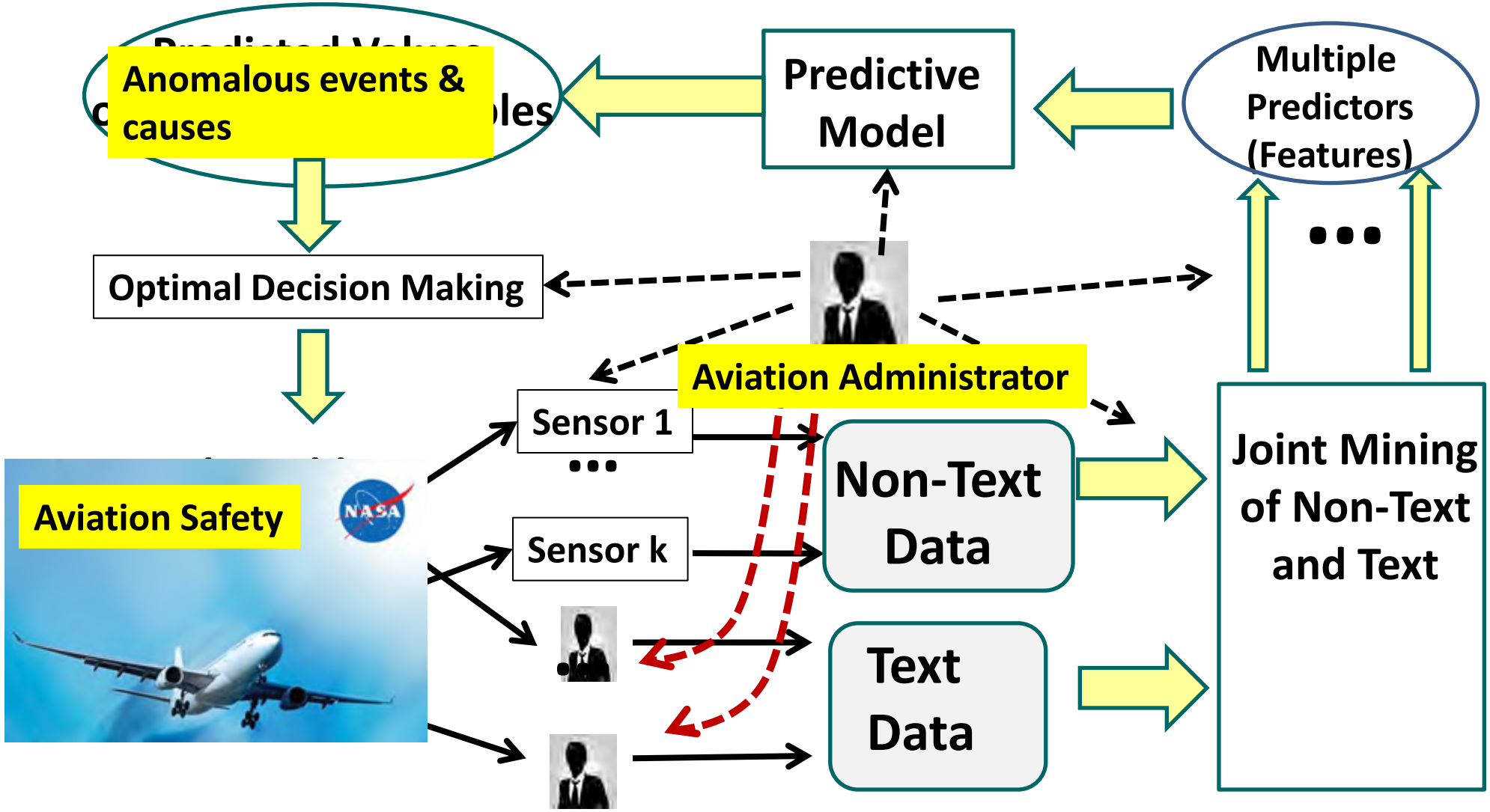
可视化

交互式工作流程管理

TextScope in Action: interactive decision support



Application Example 1: Aviation Safety



Abundance of text data in the aviation domain

Collecting reports since 1976
>860,000 reports as of Dec. 2009

Date & Report Number
- **Report Number** (ACN) was [number]
- **Date of Incident** was between [date] and [date]

Environment
- **Flight Conditions** were [conditions]
- **Lighting** was [conditions]
- **Weather** was [element]

Aircraft
- **Federal Aviation Regs** (FAR) Part was [regulation]
- **Flight Plan** was [type]
- **Flight Phase** was [phase]
- **Make/Model** was [aircraft type]
- **Mission** was [operation]

Place
- **Location** was [identifier]
- **State** was [abbreviation]

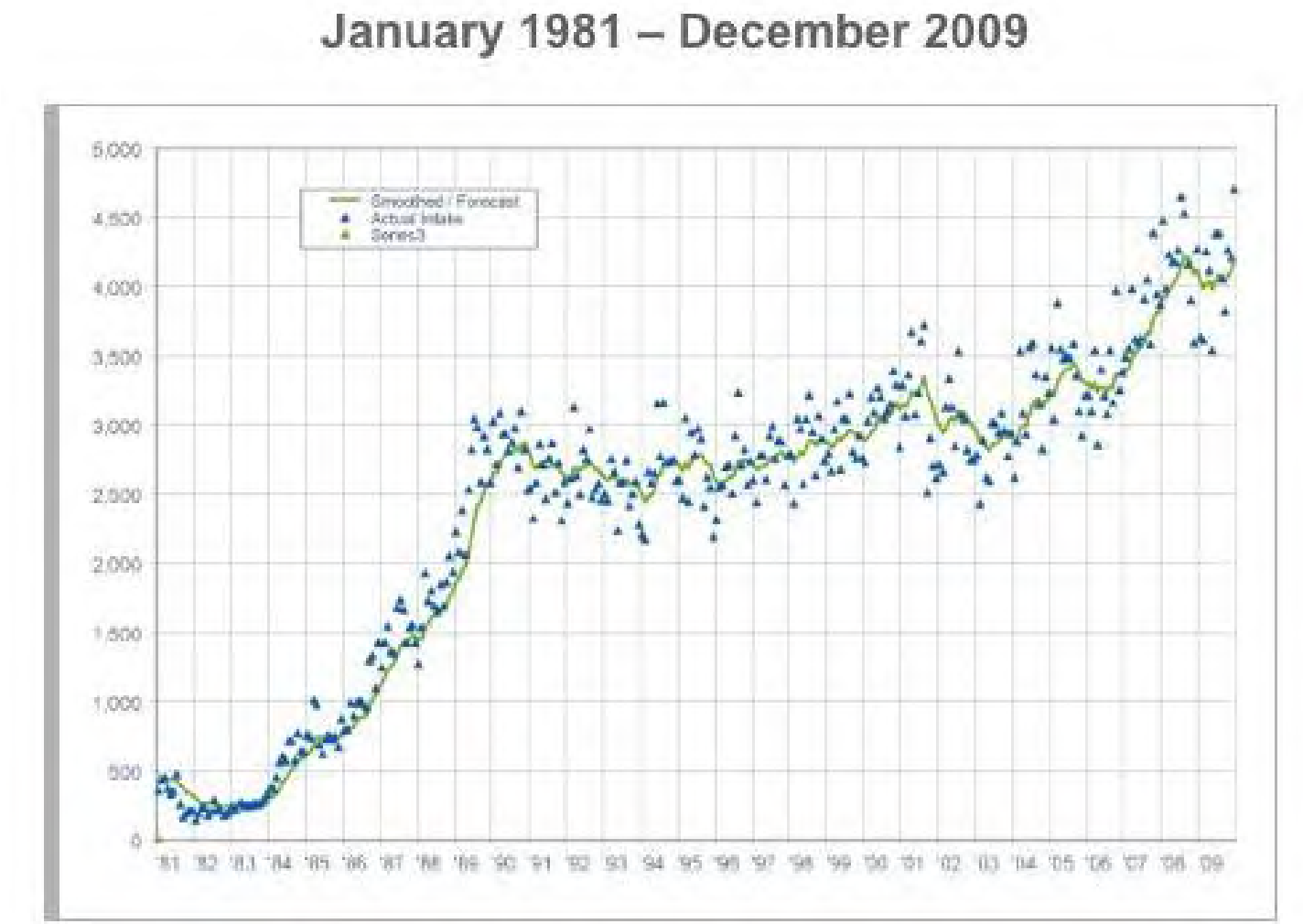
Person
- **Reporter Organization** was [type]
- **Reporter Function** was [position]

Event Assessment
- **Event Type** was [anomaly]
- **Detector** was [equipment/human]
- **Primary Problem** was [most prominent factor]
- **Contributing Factors** were [problem areas]
- **Human Factors** (since 6/09) were [factor]
- **Result** was [consequence]

Text: Narrative / Synopsis
- **Text** contains [words]

Current Search Items:

Monthly intake has been increasing (4k reports/month)



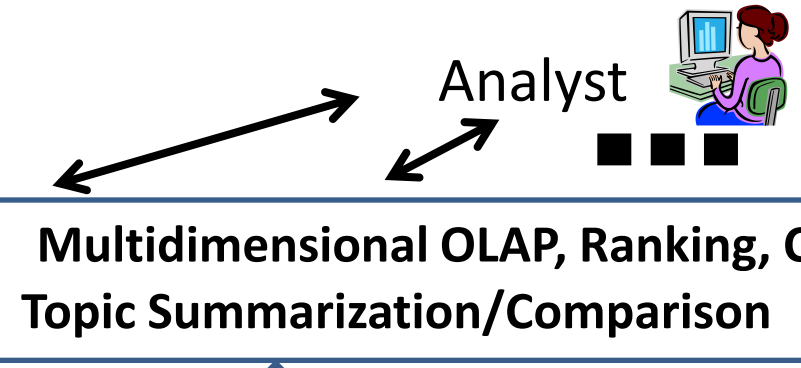
Slide source: http://asrs.arc.nasa.gov/overview/summary.html

Lots of useful knowledge buried in text

ASRS Report ACN: 928983 (Date: 201101, Time: 1801-2400. …)

We were delayed inbound for about 2 hours and 20 minutes. On the approach there was ice that accumulated on the aircraft. … The Captain wrote up … The flight crew [who picked up the plane] the following morning notified us of an **incorrect remark section write up.** I believe a few years ago, there was a different procedure for writing up aborted takeoffs. I think there was some **confusion as to what the proper write-up for the aborted takeoff was**. A **contributing factor** for this incorrect entry into the log may have been **fatigue**. I had personally been awake for about 14 hours and still had another leg to do. …Also **a contributing factor** is that **this event does not happen regularly….** A **more thorough review and adherence to the operations manual section regarding aircraft status would have prevented this**, [as well as], a better recognition of the onset of **fatigue**. The **manual is sometimes so large that finding pertinent data is difficult**. Even after it was determined that the event had occurred, it took me 15 to 20 minutes to find the section regarding aborted takeoffs.

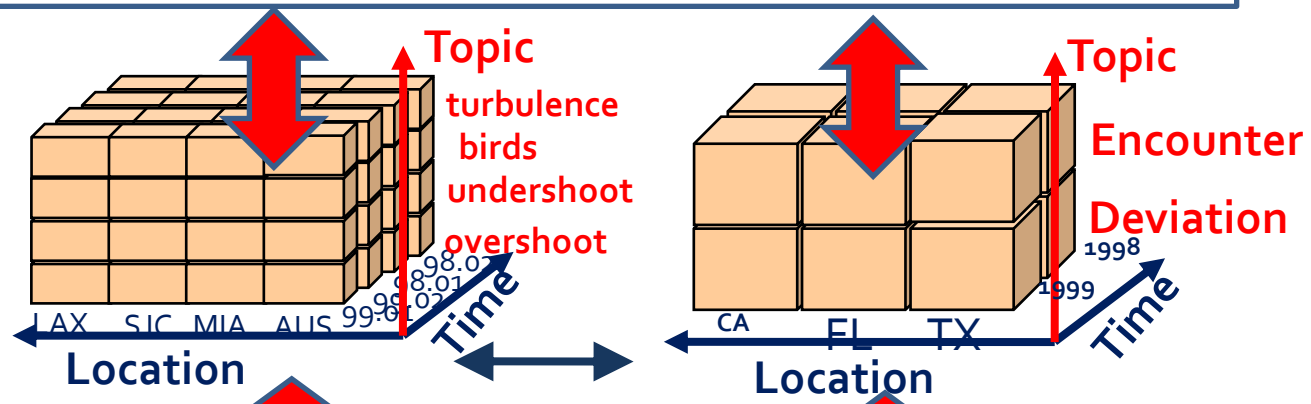
Event Cube



Analysis Support

Multidimensional OLAP, Ranking, Cause Analysis,
Topic Summarization/Comparison ▪▪▪▪▪▪

Event Cube Representation



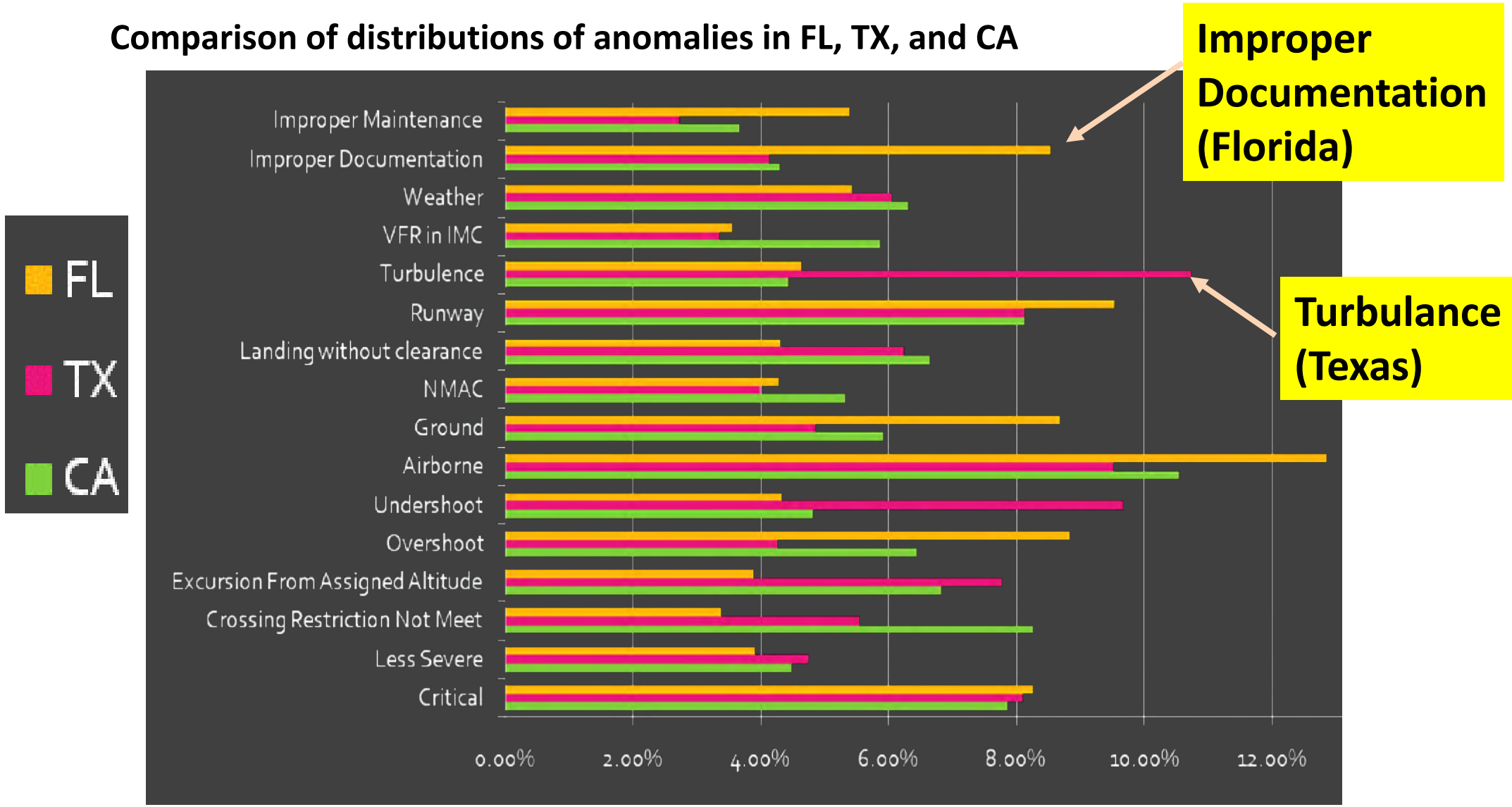
Multidimensonal Text Databasei

Light	Ar...	Phase	Report
Night	equipment : critical	intermediate altitude	...we hit what appeared ... a flock of large white birds. Some of them seem to be ducks..
Daylight	equipment : critical	intermediate altitude	...our flight encountered a small flock of ducks...
Daylight	equipment : less severe	ground: parked	...there were no bird strikes noticed while taxiing..
...	...	...	...

Duo Zhang, ChengXiang Zhai, Jiawei Han, Ashok Srivastava, Nikunj Oza. Topic Modeling for OLAP on Multidimensional Text Databases: Topic Cube and its Applications, *Statistical Analysis and Data Mining*, Vol. 2, pp.378-395, 2009.

Sample Topic Coverage Comparison

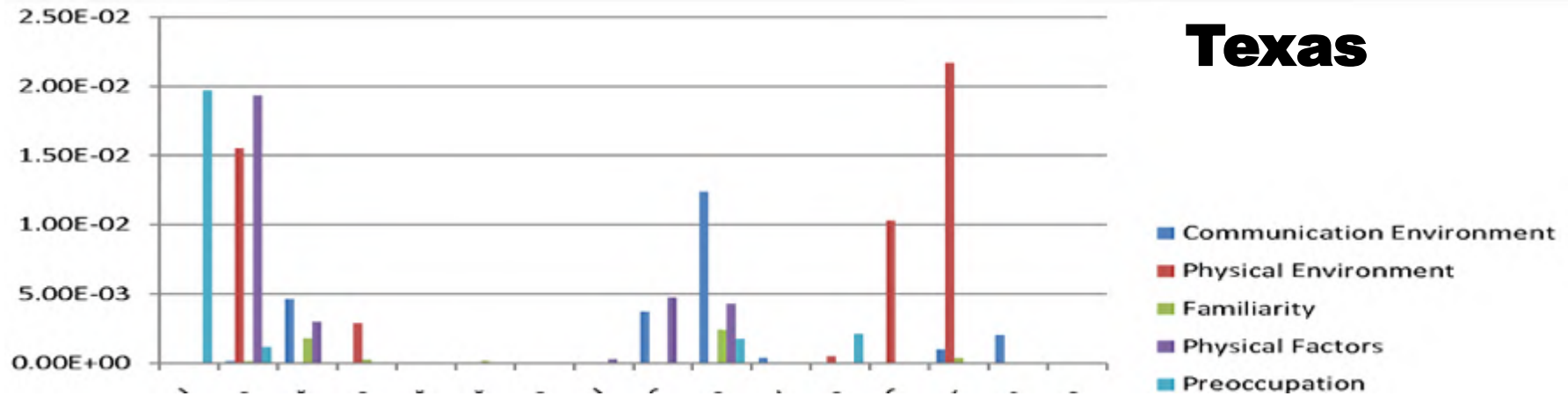
Comparison of distributions of anomalies in FL, TX, and CA



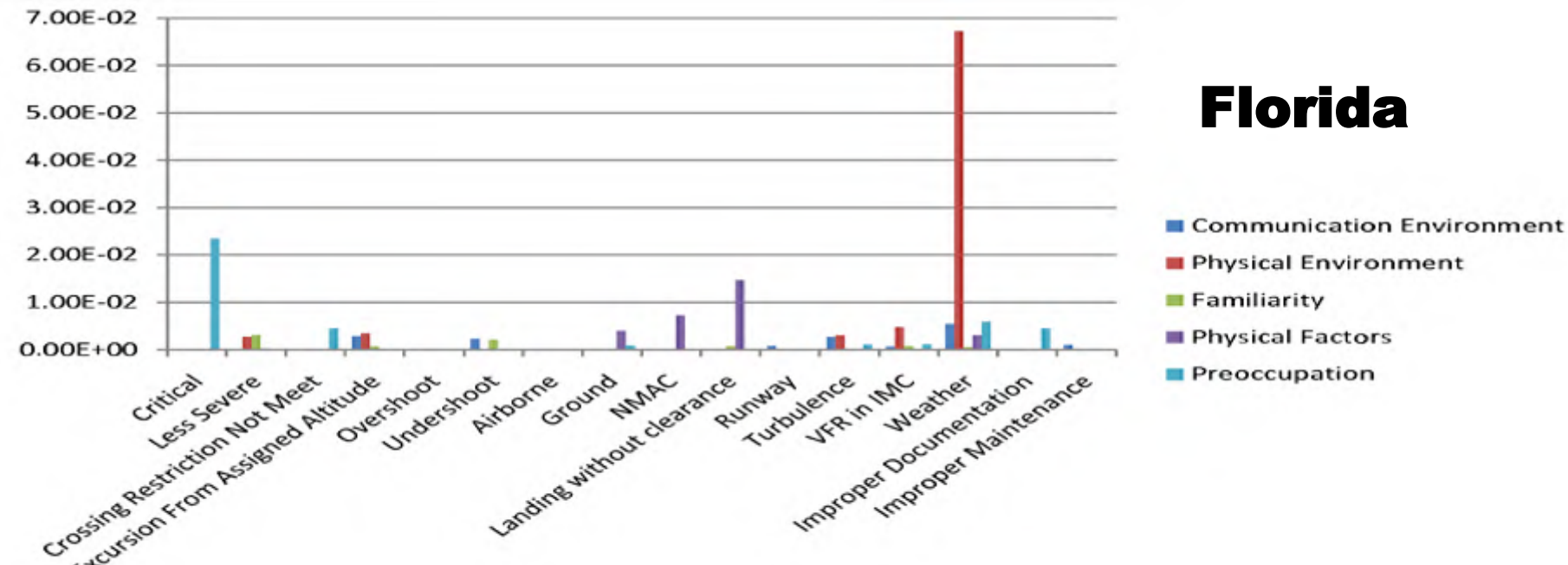
Improper Documentation (Florida)

Turbulance (Texas)

Comparative Analysis of Shaping Factors

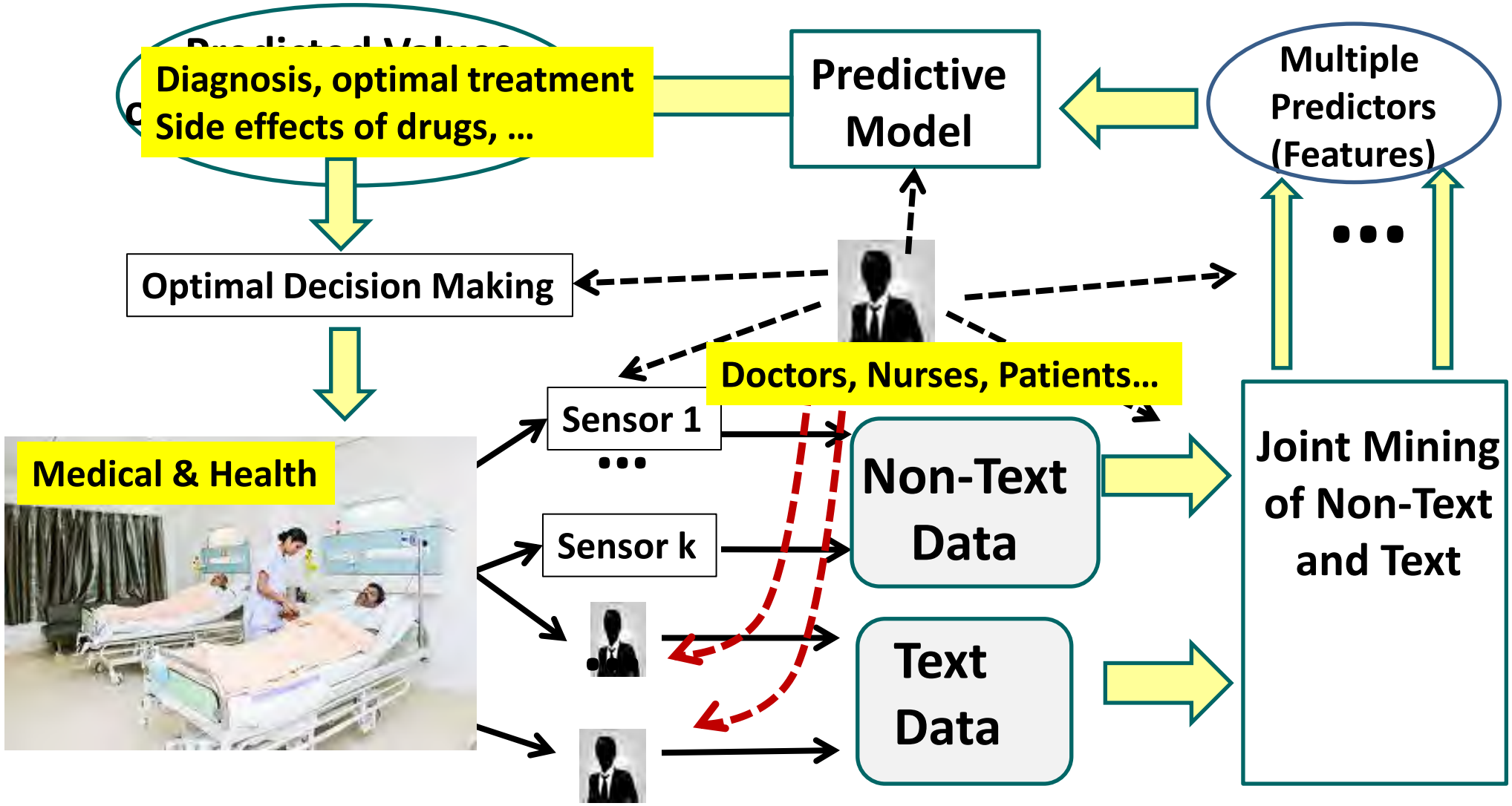


Texas



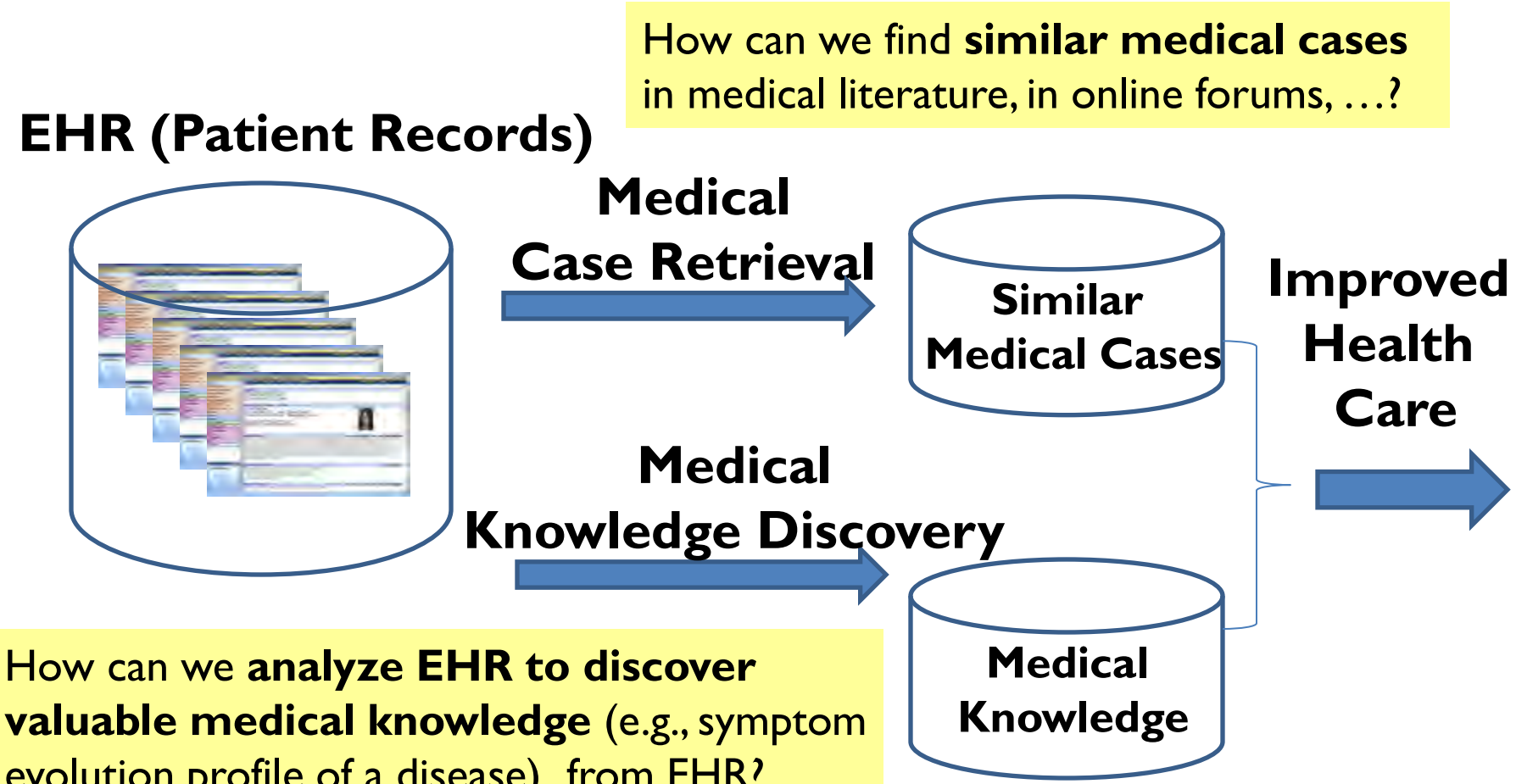
Florida

Application Example 2: Medical & Health



Overview of Text Mining for Medical/Health Applications

How can we find **similar medical cases** in medical literature, in online forums, …?



How can we **analyze EHR to discover valuable medical knowledge** (e.g., symptom evolution profile of a disease) from EHR?

Medical Case Retrieval

Query: "Female patient, 25 years old, with fatigue and a swallowing disorder (dysphagia worsening during a meal). The frontal chest X-ray shows opacity with clear contours in contact with the right heart border. Right hilar structures are visible through the mass. The lateral X-ray confirms the presence of a mass in the anterior mediastinum. On CT images, the mass has a relatively homogeneous tissue density."

Find all medical literature articles discussing a similar case

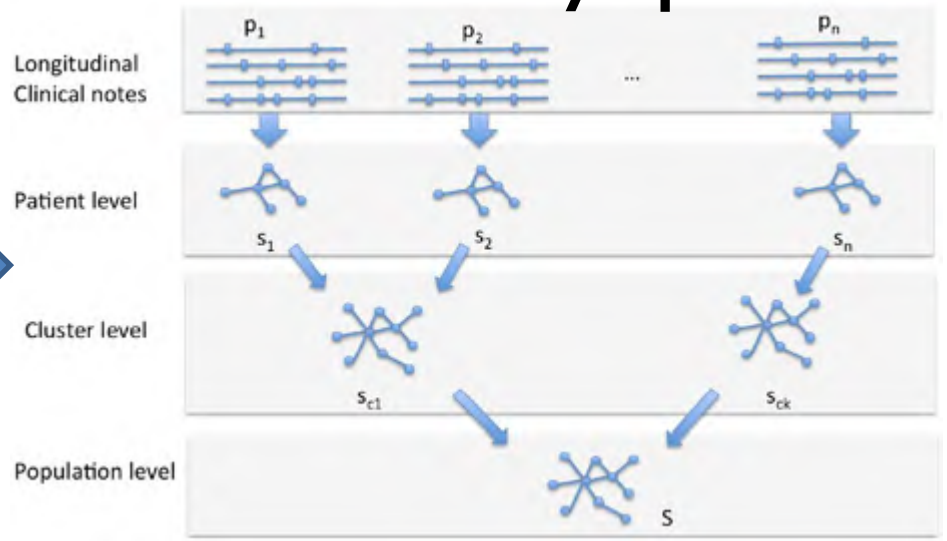
We developed techniques to leverage medical ontology and Feedback to improve accuracy. **The UIUC-IBM team was ranked #1 in ImageCLEF 2010 evaluation**.

Parikshit Sondhi, Jimeng Sun, ChengXiang Zhai, Robert Sorrentino and Martin S. Kohn. Leveraging Medical Thesauri and Physician Feedback for Improving Medical Literature Retrieval for Case Queries, Journal of the American Medical Informatics Association , 19(5), 851-858 (2012). doi:10.1136/amiajnl-2011-000293.

Extraction of Symptom Graphs from EHR

EHR (Patient Records)

Multi-Level Symptom Graphs



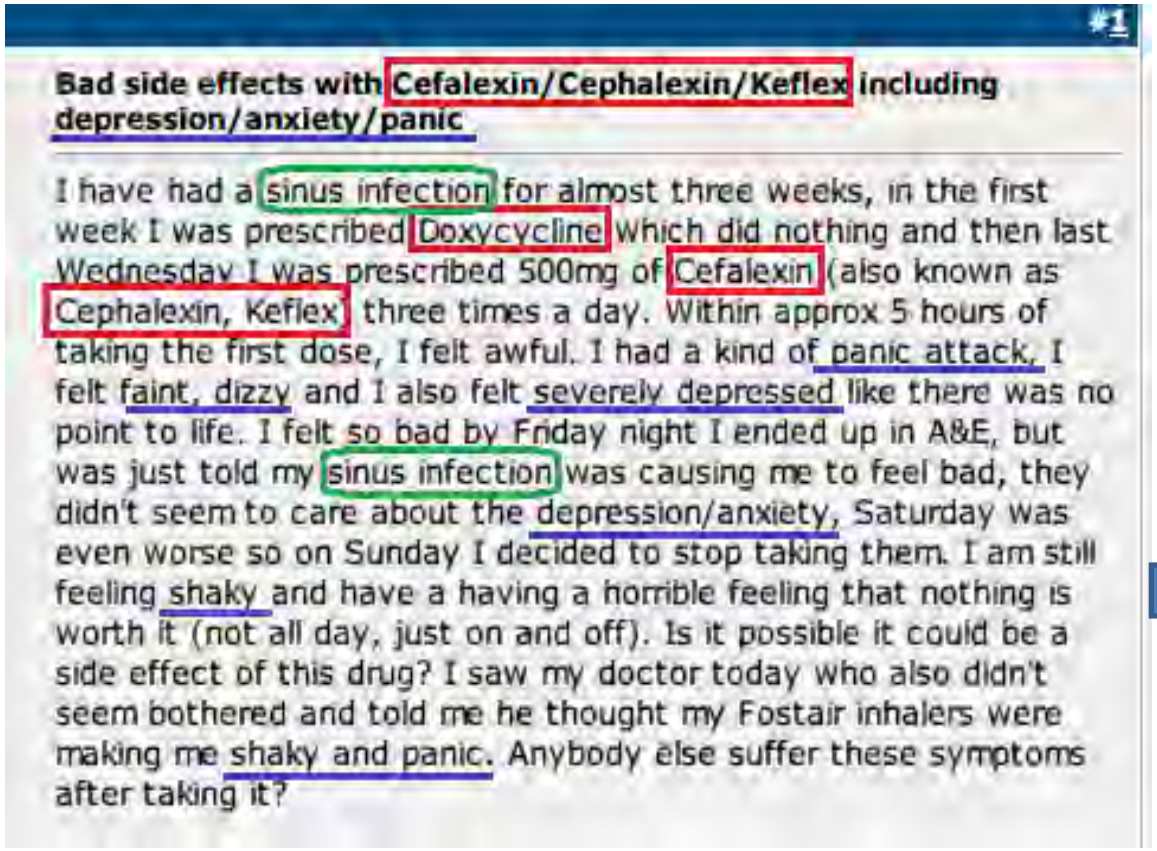
Predict the future onset of a disease (e.g., Congestive Heart Failure) for a patient

Discovery of symptom profiles of diseases

Discovered symptoms improves accuracy of prediction by +10%

Parikshit Sondhi, Jimeng Sun, Hanghang Tong, ChengXiang Zhai. SympGraph: A Mining Framework of Clinical Notes through Symptom Relation Graphs, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 12), pp. 1167-1175, 2012.

Discovery of Adverse Drug Reactions from Forums



Bad side effects with Cefalexin/Cephalexin/Keflex including depression/anxiety/panic

I have had a sinus infection for almost three weeks, in the first week I was prescribed Doxycycline which did nothing and then last Wednesday I was prescribed 500mg of Cefalexin (also known as Cephalexin, Keflex) three times a day. Within approx 5 hours of taking the first dose, I felt awful. I had a kind of panic attack, I felt faint, dizzy and I also felt severely depressed like there was no point to life. I felt so bad by Friday night I ended up in A&E, but was just told my sinus infection was causing me to feel bad, they didn't seem to care about the depression/anxiety, Saturday was even worse so on Sunday I decided to stop taking them. I am still feeling shaky and have a having a horrible feeling that nothing is worth it (not all day, just on and off). Is it possible it could be a side effect of this drug? I saw my doctor today who also didn't seem bothered and told me he thought my Fostair inhalers were making me shaky and panic. Anybody else suffer these symptoms after taking it?

Green: Disease symptoms
Blue: Side effect symptoms
Red: Drug

Drug: Cefalexin
ADR:
panic attack
faint
....

Sheng Wang et al. 2014. SideEffectPTM: an unsupervised topic model to mine adverse drug reactions from health forums. In *ACM BCB 2014*.

Sample ADRs Discovered

Drug(Freq)	Drug Use	Symptoms in Descending Order
Zoloft (84)	antidepressant	**weigh gain, weight,** depression, side effects, mgs, **gain weight**, anxiety, **nausea**, head, brain, **pregnancy, pregnant, headaches**, depressed, **tired**
Ativan (33)	anxiety disorders	Ativan, **sleep**, Seroquel, doc prescribed seroqual, **raising blood sugar levels**, anti-psychotic drug, **diabetic, constipation, diabetes**, 10mg, benzo, **addicted**
Topamax (20)	anticonvulsant	Topmax, **liver**, side effects, **migraines, headaches, weight**, Topamax, pdoc, neurologist, supplement, **sleep, fatigue, seizures, liver problems, kidney stones**
Ephedrine (2)	stimulant	**dizziness, stomach**, Benadryl, **dizzy, tired, lethargic**, tapering, **tremors, panic attach, head**

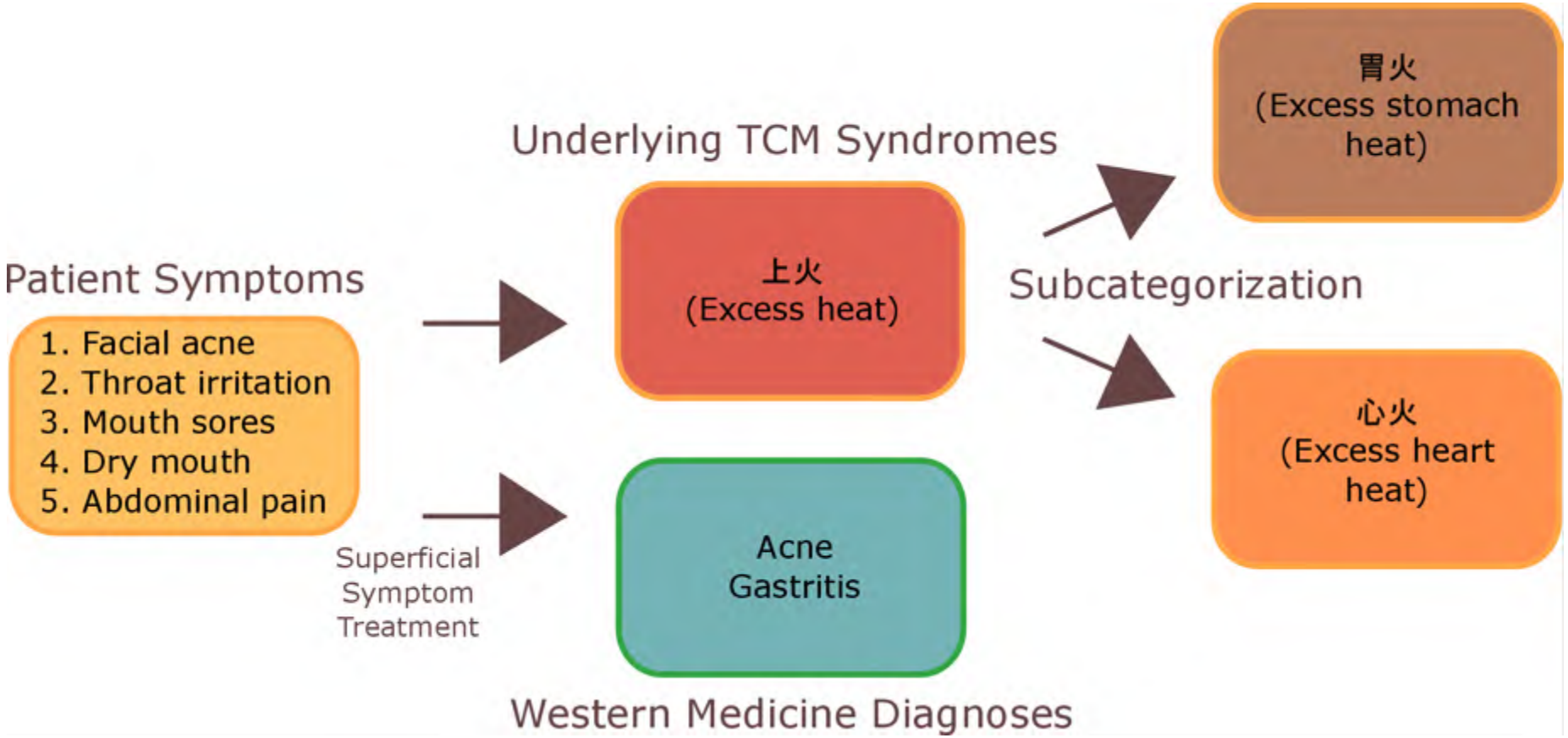
Unreported to FDA

Mining Traditional Chinese Medicine Patient Records

- Collaboration with Beijing TCM Data Center
  - Clinical warehouse since 2007
  - More than 300,000 clinical cases from six hospitals
  - Each hospital has ~ 3 million patient visits
- Two lines of work
  - Subcategorization of patient records
  - TCM knowledge discovery

Beijing TCM Data Center

Subcategorization of Patient Records



Edward W Huang, Sheng Wang, Runshun Zhang, Baoyan Liu, Xuezhong Zhou, and ChengXiang Zhai. **PaReCat: Patient Record Subcategorization for Precision Traditional Chinese Medicine.** ACM BCB, Oct. 2016.

TCM Knowledge Discovery

- 10,907 patients TCM records in digestive system treatment
- 3,000 symptoms, 97 diseases and 652 herbs
- Most frequently occurring disease: chronic gastritis
- Most frequently occurring symptoms: abdominal pain and chills

- Ground truth: 27,285 manually curated herb-symptom relationship.

Sheng Wang, Edward Huang, Runshun Zhang, Xiaoping Zhang, Baoyan Liu, Xuezhong Zhou, and ChengXiang Zhai , A Conditional Probabilistic Model for Joint Analysis of Symptoms, Diagnoses, and Herbs in Traditional Chinese Medicine Patient Records" , IEEE BIBM 2016.

Top 10 herb-symptoms relationships

Herb	Symptom	Rank
five-flavor berry (五味子)	palpitations (心悸)	11
Japanese lady bell (南沙参)	dry mouth (口干)	13
Chinese cucumber (瓜蒌皮)	chest tightness (胸闷)	20
Chinese white olive (青果)	coughing (咳嗽)	29
crow-dipper (半夏)	abdominal swelling (脘痞)	30
ginger-lily (草豆蔻)	bloating (腹胀)	43
broad-leaf privet (女贞子)	tinnitus (耳鸣)	49
Chinese buckeye seed (娑罗子)	bloating (腹胀)	57
midnight horror (木蝴蝶)	coughing (咳嗽)	71

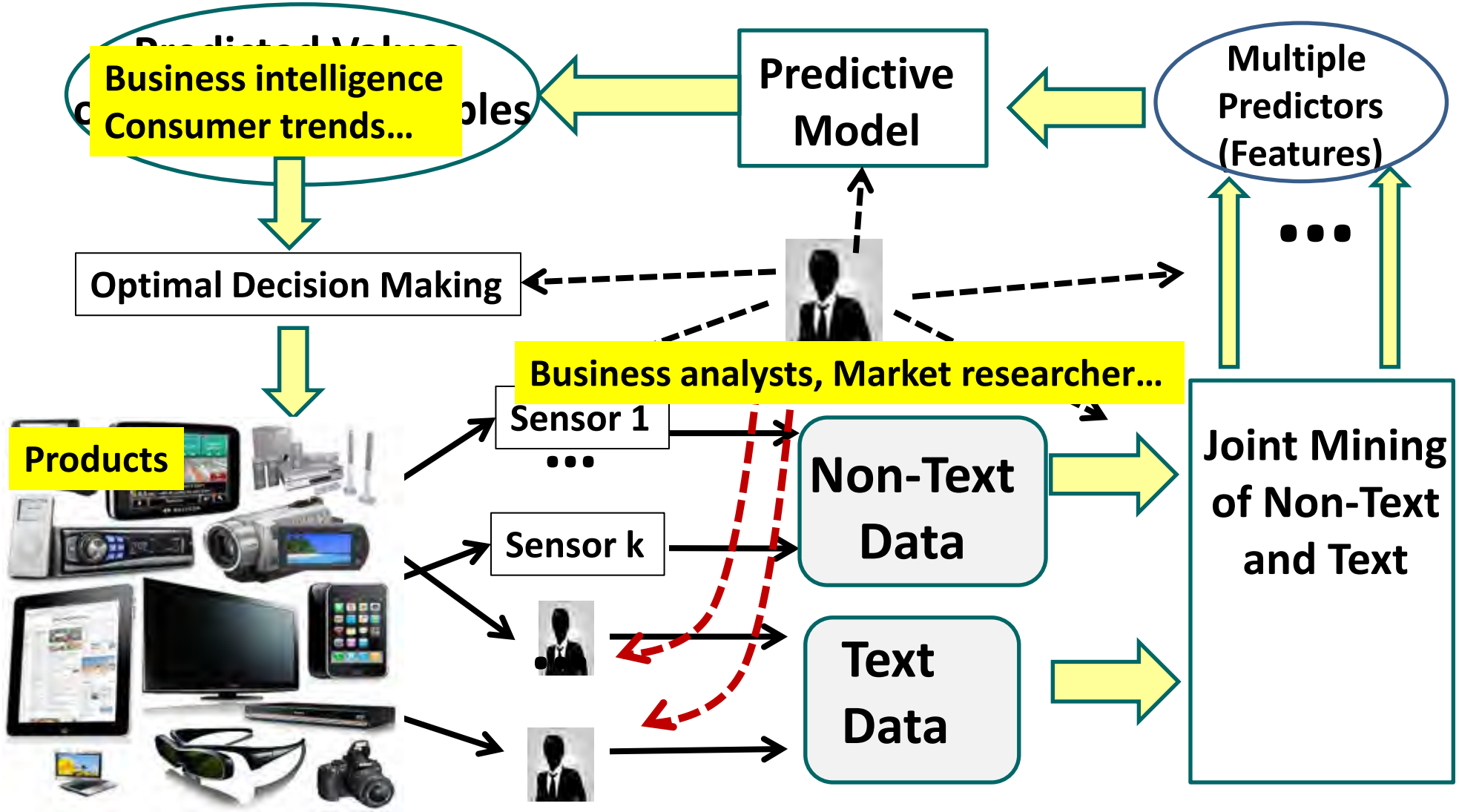
Typical Symptoms of three Diseases

chronic gastritis	constipation	reflux esophagitis
胃胀 epigastric distention	排便困难 difficult bowel movements	痞满 upper abdominal distention and fullness
胃脘畏寒 epigastric chills	腹胀 abdominal distension	反酸 acid reflux
痞满 upper abdominal distention and fullness	大便干燥 dry, hard stools	烧心 heartburn
烧心 heartburn	乏力 hypodynamia	嗳气 belching
反酸 acid reflux	腹部畏寒 abdominal chills	咽部异物感 paresthesia pharynges

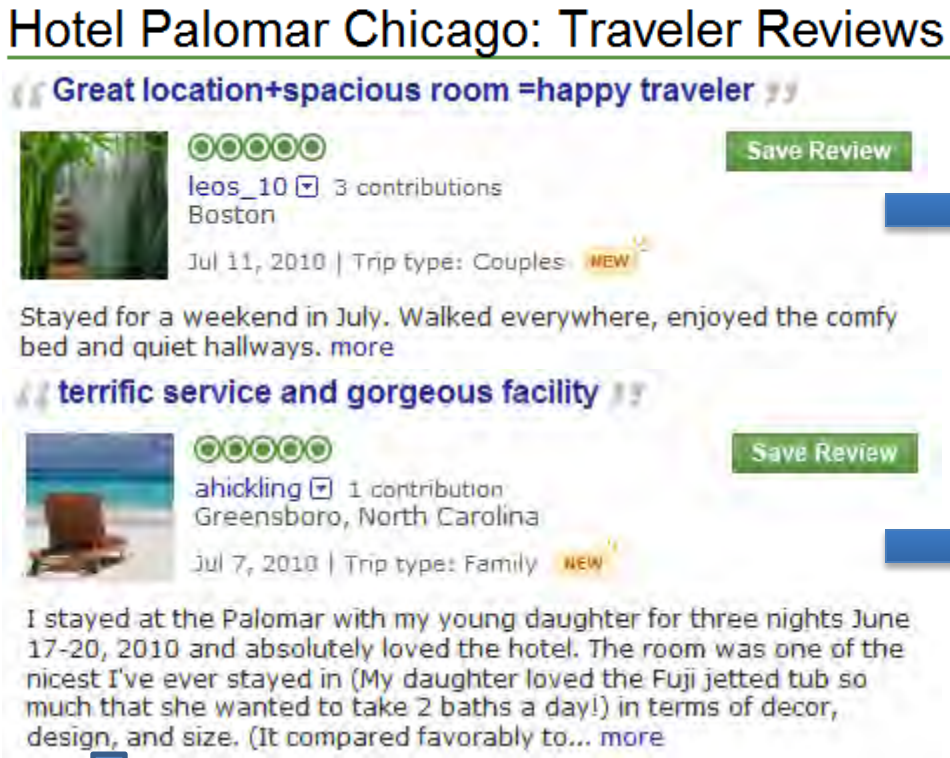
Typical Herbs for three Diseases

hyperlipidemia	hepatic steatosis	chronic pharyngitis
荷叶 lotus leaf	荷叶 lotus leaf	桔梗 Chinese bellflower root
生山楂 raw hawthorn	茵陈 Oriental wormwood	紫苏叶 perilla root
生地黄 *R. glutinosa* root	牡丹皮 peony root bark	连翘 weeping forsythia fruit
醋莪术 turmeric vinegar	虎杖 Japanese knotweed	炒牛蒡子 stir-fried burdock
虎杖 Japanese knotweed	栀子 gardenia fruit	玄参 figwort root

Application Example 3: Business intelligence

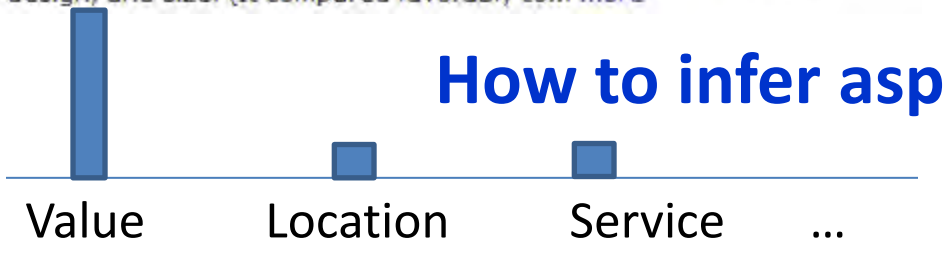


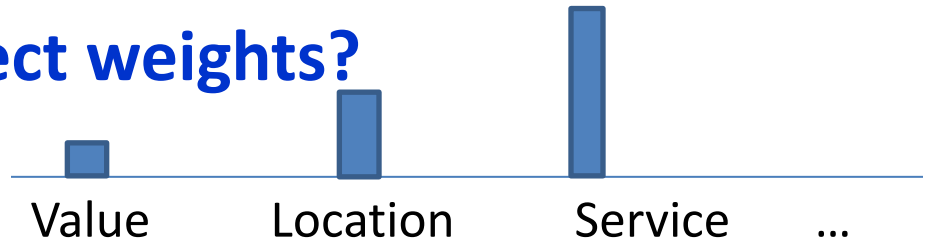
Motivation



How to infer aspect ratings?

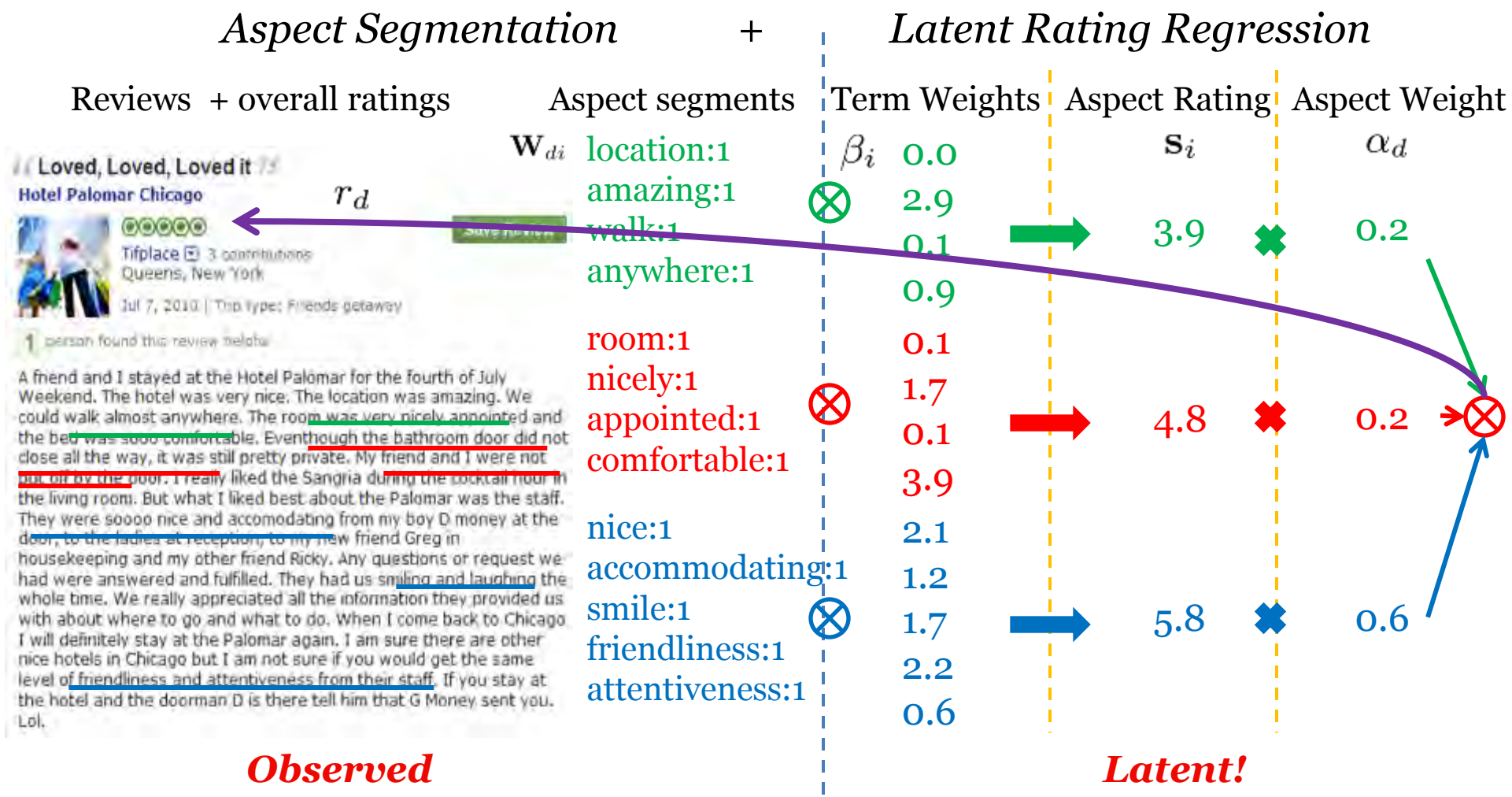
How to infer aspect weights?





Hongning Wang, Yue Lu, ChengXiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD'10), pages 115-124, 2010.

Solving LARA in two stages: Aspect Segmentation + Rating Regression



Sample Result 1: Rating Decomposition

- Hotels with the same overall rating but different aspect ratings

(All 5 Stars hotels, ground-truth in parenthesis.)

Hotel	*Value*	*Room*	*Location*	*Cleanliness*
Grand Mirage Resort	4.2(4.7)	3.8(3.1)	4.0(4.2)	4.1(4.2)
Gold Coast Hotel	4.3(4.0)	3.9(3.3)	3.7(3.1)	4.2(4.7)
Eurostars Grand Marina Hotel	3.7(3.8)	4.4(3.8)	4.1(4.9)	4.5(4.8)

- Reveal detailed opinions at the aspect level

Sample Result 2: Comparison of reviewers

- Reviewer-level Hotel A
  - Different reviewers' rati
  - Reveal differences in op

Reviewer	Value
Mr.Saturday	3.7(4.0)
Salsrug	5.0(5.0)

"Good Price for what we got"

Riu Palace Punta Cana

Save Review

salsrug 13 contributions
Marylander

Oct 27, 2008 | Trip type: Family

1 person found this review helpful

We stayed for six days, five nights. Overall, we had a very good time. The was pretty good and the staff was very friendly. They definitely do not skimp on the free alcohol. The room was a little smelly, which we had read on trip advisor so we bought a candle with us - no problem. They only thing I had an issue with was the little bugs. They were like gnats or fleas but they weren't either. I had some candy and popcorn which we brought from the States to munch on. I left it out on the table and within 40 minutes, the bag was infested. DO NOT keep any open food in your room. Also we ended up having to wash all of our clothes (clean and dirty) and airing out our luggage when we got home because we could still smell the room on them. For the price we paid, we really did have an excellent time besides those small things. The pool was awesome and the beach was spectacular. Out of the nearby resorts that we saw, Riu Palace Punta Cana was the best (it was also the nicest out of the other Riu's on Punta Cana). We went on the 1/2 day Outback Safari and had a great time. We got coffee and souvenirs cheapers than other places and the hotel. General - not good or bad just things that we noticed - There were a lot of topless sunbathers. The crowd is middle aged (35 - 55) so we were on the younger side and the majority of the people were European or Brazilian. It helps to know some spanish but it's not a necessity.

Liked — The beach was excellent.

Disliked — Room smell and little bugs.

My ratings for this hotel

Value
Rooms
Location
Cleanliness
Check in / front desk
Service
Business service (e.g., internet access)

Sample Result 3:Aspect-Specific Sentiment Lexicon

Value	*Rooms*	*Location*	*Cleanliness*
resort 22.80	view 28.05	restaurant 24.47	clean 55.35
value 19.64	comfortable 23.15	walk 18.89	smell 14.38
excellent 19.54	modern 15.82	bus 14.32	linen 14.25
worth 19.20	quiet 15.37	beach 14.11	maintain 13.51
bad -24.09	*carpet -9.88*	*wall -11.70*	*smelly -0.53*
money -11.02	*smell -8.83*	*bad -5.40*	*urine -0.43*
terrible -10.01	*dirty -7.85*	*road -2.90*	*filthy -0.42*
overprice -9.06	*stain -5.85*	*website -1.67*	*dingy -0.38*

Uncover sentimental information directly from the data

Application 1: Discover consumer preferences

- Amazon reviews: no guidance

Table 2: Topical Aspects Learned on MP3 Reviews

Low Overall Ratings			High Overall Ratings		
unit	jack	service	files	player	vision
usb	headphone	charge	format	music	video
battery	warranty	problem	included	download	player
charger	replacement	support	easy	headphones	quality
reset	problem	hours	convert	button	great
time	player	months	mp3	set	product
hours	back	weeks	videos	hours	sound
work	months	back	file	buds	radio
thing	buy	customer	wall	volume	accessory
wall	amazon	time	hours	ear	fm

battery life accessory service file format volume video

Application 2: User Rating Behavior Analysis

	Expensive Hotel		*Cheap Hotel*	
	5 Stars	*3 Stars*	*5 Stars*	*1 Star*
Value	0.134	0.148	**0.171**	0.093
Room	0.098	0.162	0.126	0.121
Location	0.171	0.074	0.161	0.082
Cleanliness	0.081	**0.163**	0.116	**0.294**
Service	**0.251**	0.101	0.101	0.049

People like expensive hotels because of good service

People like cheap hotels because of good value

Application 3:
Personalized Recommendation of Entities

Query: 0.9 value
0.1 others

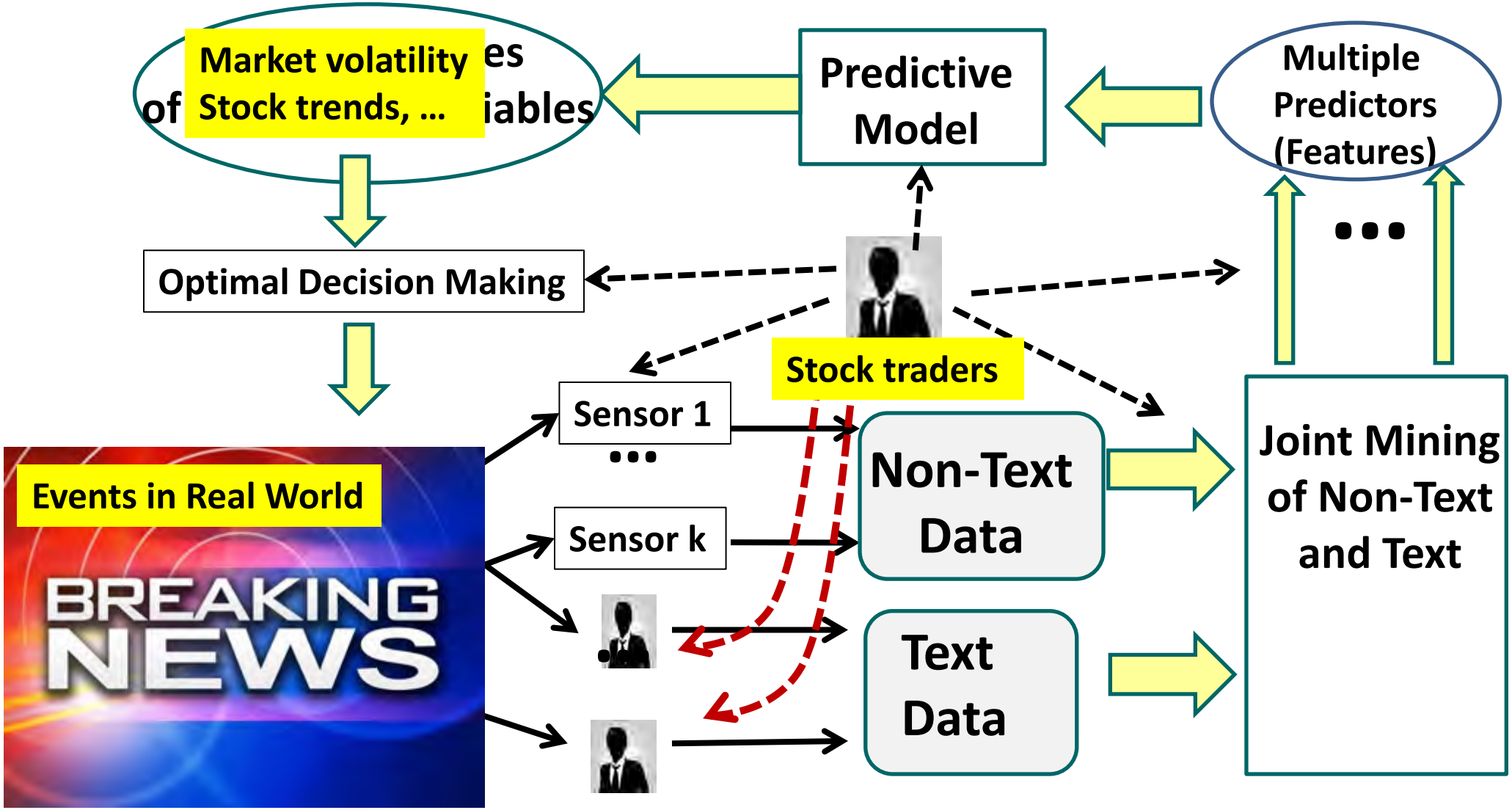
Non-Personalized

Personalized

Table 10: Personalized Hotel Ranking

Hotel	Overall Rating	Price	Location
Majestic Colonial	5.0	339	Punta Cana
Agua Resort	5.0	753	Punta Cana
Majestic Elegance	5.0	537	Punta Cana
Grand Palladium	5.0	277	Punta Cana
Iberostar	5.0	157	Punta Cana
Elan Hotel Modern	5.0	216	Los Angeles
Marriott San Juan Resort	4.0	354	San Juan
Punta Cana Club	5.0	409	Punta Cana
Comfort Inn	5.0	155	Boston
Hotel Commonwealth	4.5	313	Boston

Application Example 4: Prediction of Stock Market



Text Mining for Understanding Time Series

What might have caused the stock market crash?

Sept 11 attack!

Any clues in the companion news stream?

Dow Jones Industrial Average [Source: Yahoo Finance]

Stock-Correlated Topics in New York Times: June 2000 ~ Dec. 2011

AAMRQ (American Airlines)	AAPL (Apple)
russia russian putin	paid notice st
europe european	russia russian europe
germany	olympic games olympics
bush gore presidential	she her ms
police court judge	oil ford prices
airlines airport air	black fashion blacks
united trade terrorism	**computer technology software**
food foods cheese	**internet com web**
nets scott basketball	football giants jets
tennis williams open	japan japanese plane
awards gay boy	
moss minnesota chechnya	

Topics are biased toward each time series

Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas A. Rietz, Daniel Diermeier. Mining causal topics in text data: iterative topic modeling with time series feedback, Proceedings of the 22nd ACM international conference on Information
and knowledge management (CIKM ' 13), pp. 885-890, 2013.

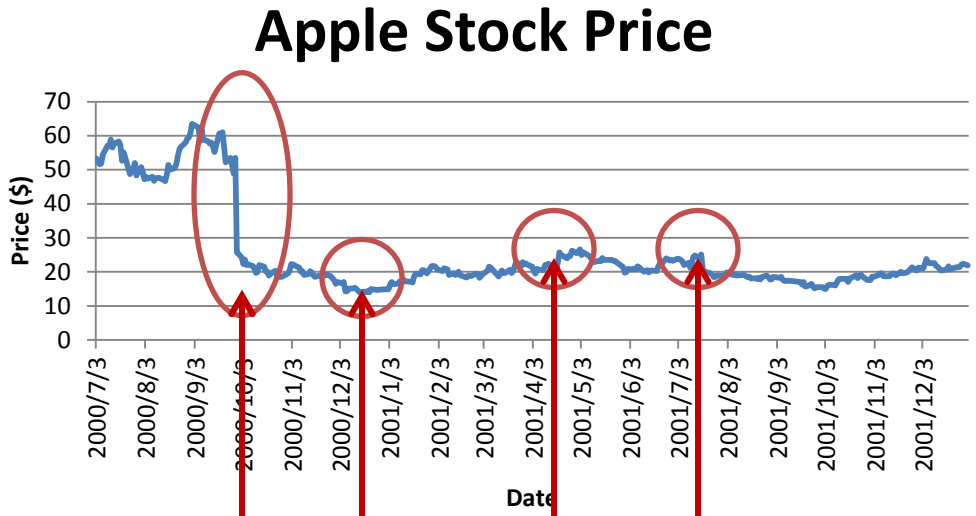
“Causal Topics” in 2000 Presidential Election

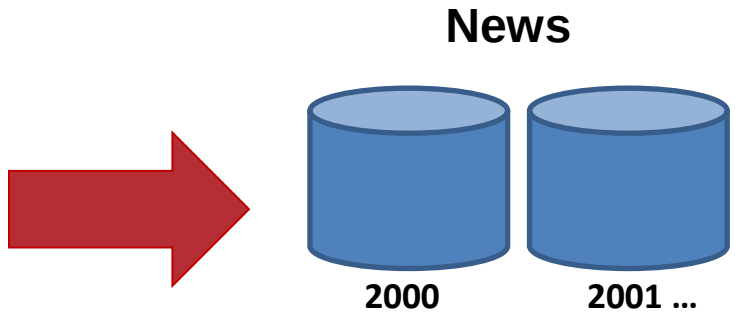
Top Three Words in Significant Topics from NY Times
tax cut 1
screen pataki guiliani
enthusiasm door symbolic
oil energy prices
news w top
pres al vice
love tucker presented
partial **abortion** privatization
court supreme **abortion**
gun control nra

Text: NY Times (May 2000 - Oct. 2000)
Time Series: Iowa Electronic Market
http://tippie.uiowa.edu/iem/

Issues known to be important in the 2000 presidential election

Information Retrieval with Time Series Query





RANK	DATE	EXCERPT
1	9/29/2000	Expect earning will be far below
2	12/8/2000	$4 billion cash in company
3	10/19/2000	Disappointing earning report
4	4/19/2001	Dow and Nasdaq soar after rate cut by Federal Reserve
5	7/20/2001	Apple's new retail store
…	…	…

Hyun Duk Kim, Danila Nikitin, ChengXiang Zhai, Malu Castellanos, and Meichun Hsu. 2013. Information Retrieval with Time Series Query. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval* (ICTIR '13),

Top ranked documents by American Airlines stock price

Rank	Date	Excerpt
1	10/22/2001	Fleeing the war
2	12/11/2001	Us and anti-Taliban forces in **Afghanistan**
3	11/18/2001	Fate of **Taliban** Soldiers Under Discussion
4	11/12/2001	Tally and dead and missing in **Sep 11 terrorist attacks**
5	9/25/2001	Soldiers in **Afghanistan** …
6	11/19/2001	Recover operation at **World Trade Center**
7	11/3/2001	4343 died or missing as a result of the attacks on **Sep 11**
8	11/17/2001	Dead and missing report of **Sep 11 attack**

All top ranked documents are related to September 11, terrorist attack

Top ranked 'relevant' documents by Apple stock price

Rank	Date	Excerpt
1	9/29/2000	Fourth-**quarter earning** far below estimates
2	12/8/2000	$4 billion reserve, not $11 billion
3	10/19/2000	Announced **earnings report**
4	4/29/2001	Dow and Nasdaq soar after rate cur by Federal Reserve
5	7/20/2001	Apple's **new retail stores**
6	12/6/2000	Apple warns it will record **quarterly loss**
7	3/24/2001	Stocks perk up, with Nasdaq posing gain
8	8/10/2000	Mixing Mac and Windows

- **Retrieved relevant events: Disappointing earning report, store open, etc.**

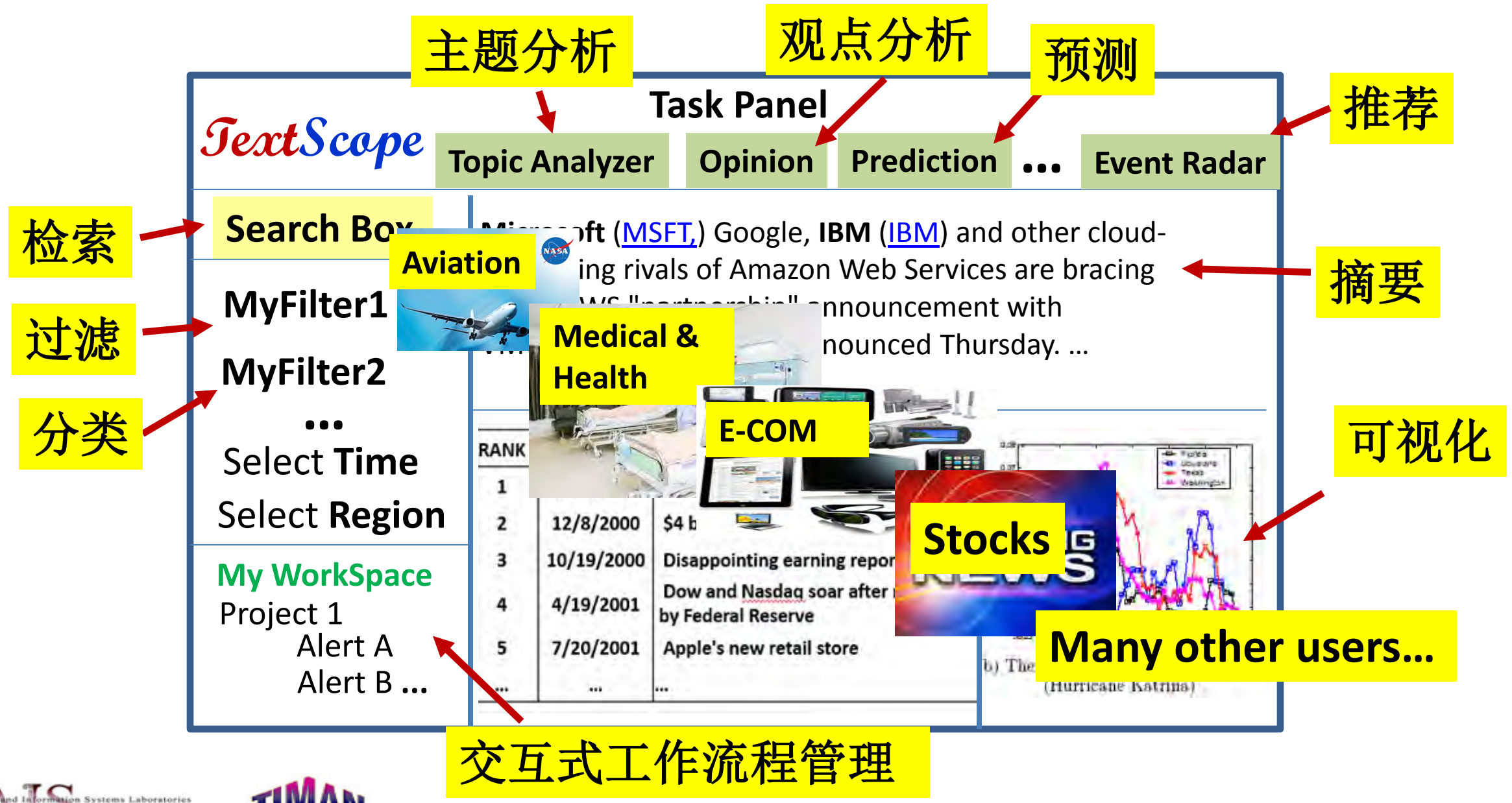
总结 (Summary)

Human as subjective, intelligent sensor

- **人= 主观智能“传感器”：** 文本数据的广泛特殊应用价值
  - 对所有大数据应用都有应用价值
  - 特别有助于挖掘，利用有关人的行为，心态，观点的知识
  - 直接表达知识（高质量数据）：小文本数据应用
- **文本数据理解困难：** 必须优化人机合作
  - 用计算机所长，统计方法，机器学习
  - 将不完善的技术转化为有用的产品
- **文本数据镜：** TextScope
  - 集信息检索和文本分析挖掘与一体
  - 支持交互式分析，决策支持
  - 应用实例：飞行安全，医疗卫生，智能商务，金融市场分析

Maximization of combined intelligence of humans and computers

前景与技术挑战：支持多种应用的通用文本数据镜



Thank You!

Questions/Comments?

Looking forward to opportunities to collaborate!