

User Behavior Analysis in Big Data



Morteza(Mori) Zihayat

About Myself

- Mitacs Elevate Postdoctoral Research Fellow
 - ▶ Faculty of Information, University of Toronto
- Big Data Scientist
 - ▶ Spectrum Computing, IBM
 - ▶ Globe and Mail
- Main research interests
 - ▶ User modeling
 - ▶ Big Data Mining and Engineering
 - Finding Meaningful Patterns from Structured and Unstructured Big Data
 - Parallel and Distributed Data Mining
 - ▶ Social Network Analysis

Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion

Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion

Big Data



50 MILLION
TWEETS PER DAY



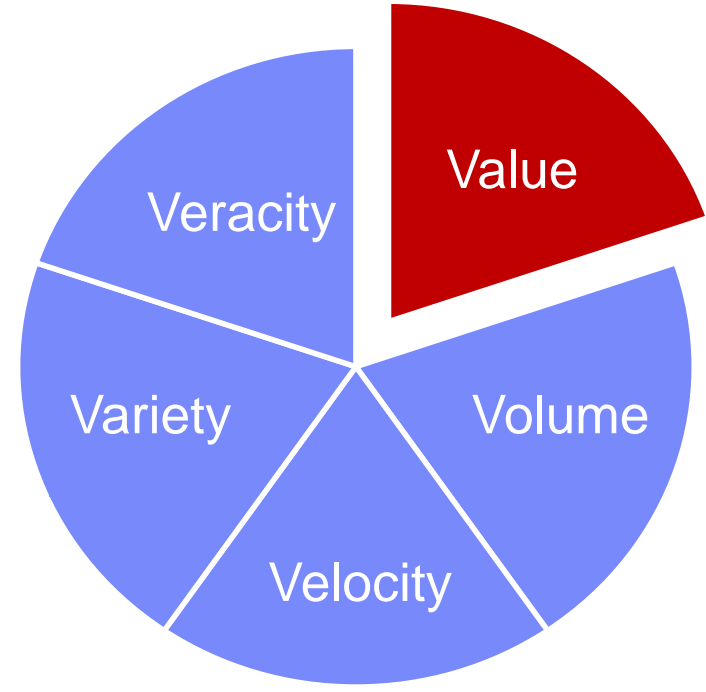
150 Exabytes
HEALTHCARE DATA



200 Gigabytes
THE SIZE OF A SINGLE
SEQUENCED HUMAN GENOME

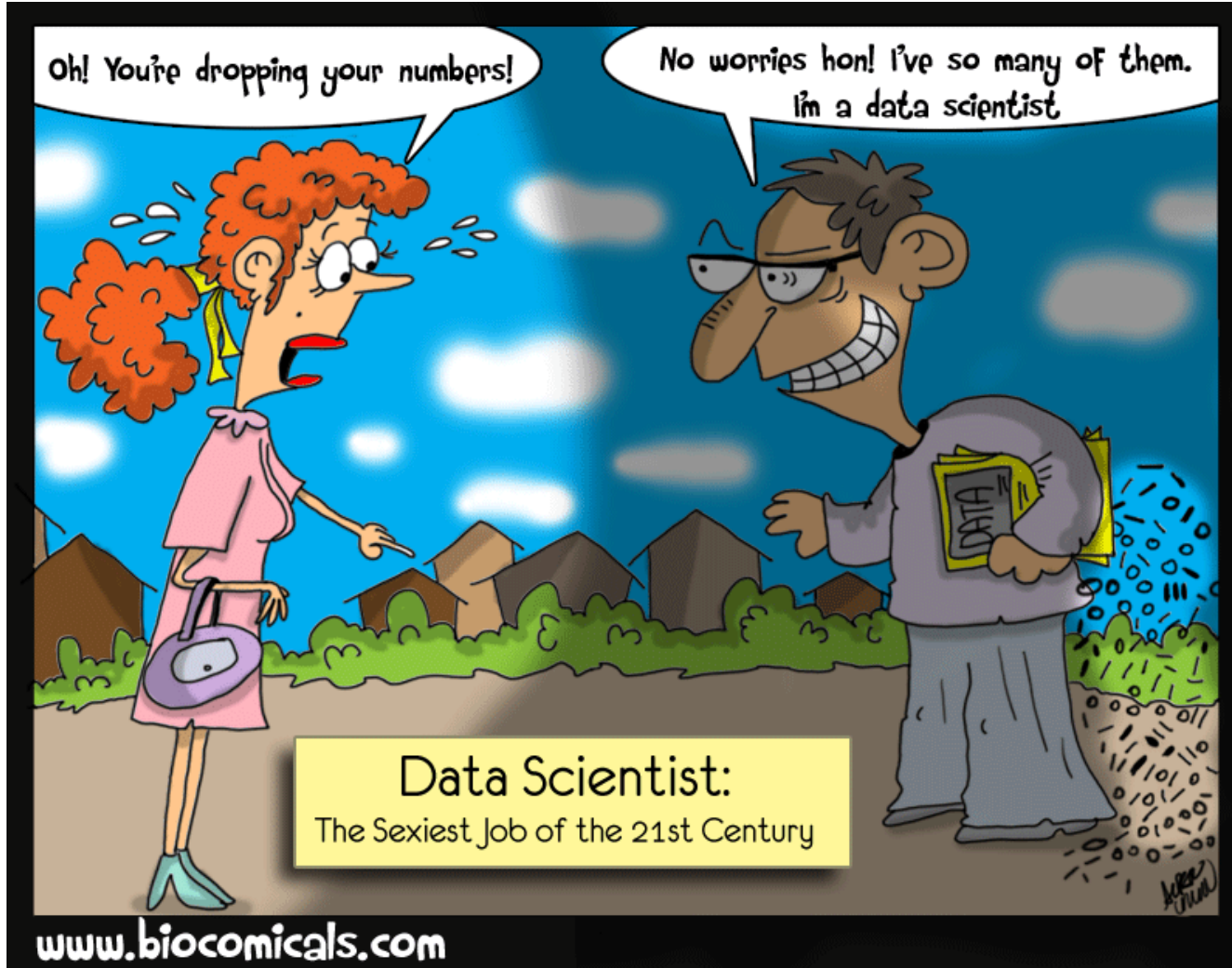


50 MILLION DEVICES
EACH DEVICE GENERATES
156 MB OF DATA PER DAY



0.5% ever analyzed and used
(MIT Technology review)

Big Data to Data Science



Data Scientist: The Sexiest Job of the 21st Century

Harvard
Business Review
Oct. 2012

(c) 2012 Biocomicals
by Dr. Alper Uzon



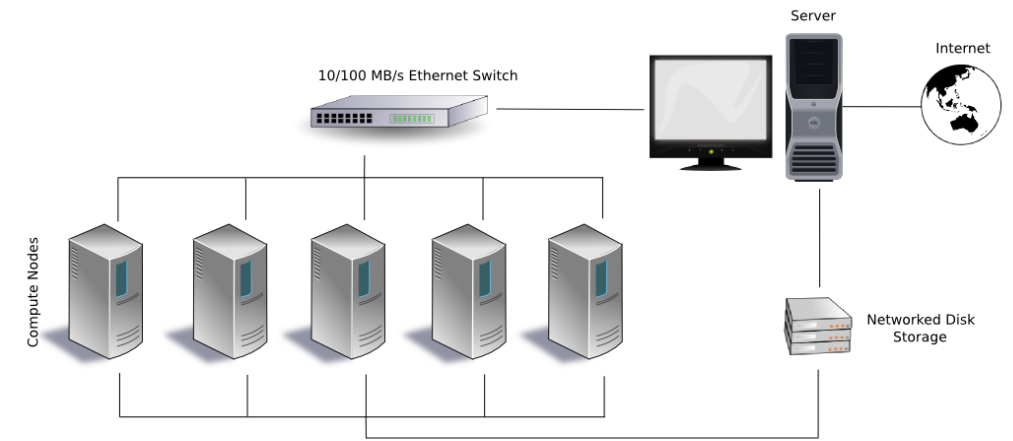
Challenges

- Read/Write to disk is slow
 - ▶ Use multiple disks for **parallel read**
- Hardware failure
 - ▶ Single machine/disk failure
 - ▶ Keep **multiple copies** of data
- How do we **merge** data from different reads
 - ▶ Distributed processing or **Hadoop MapReduce**



Apache Hadoop

- Hadoop is an open-source software framework written in Java
 - ▶ Distributed storage
 - ▶ Computer clusters
- Two main components
 - ▶ HDFS (Hadoop Distributed File System)
 - Provides Distributed Storage
 - ▶ MapReduce (Distributed Data Processing Model)
 - Provides Distributed Processing

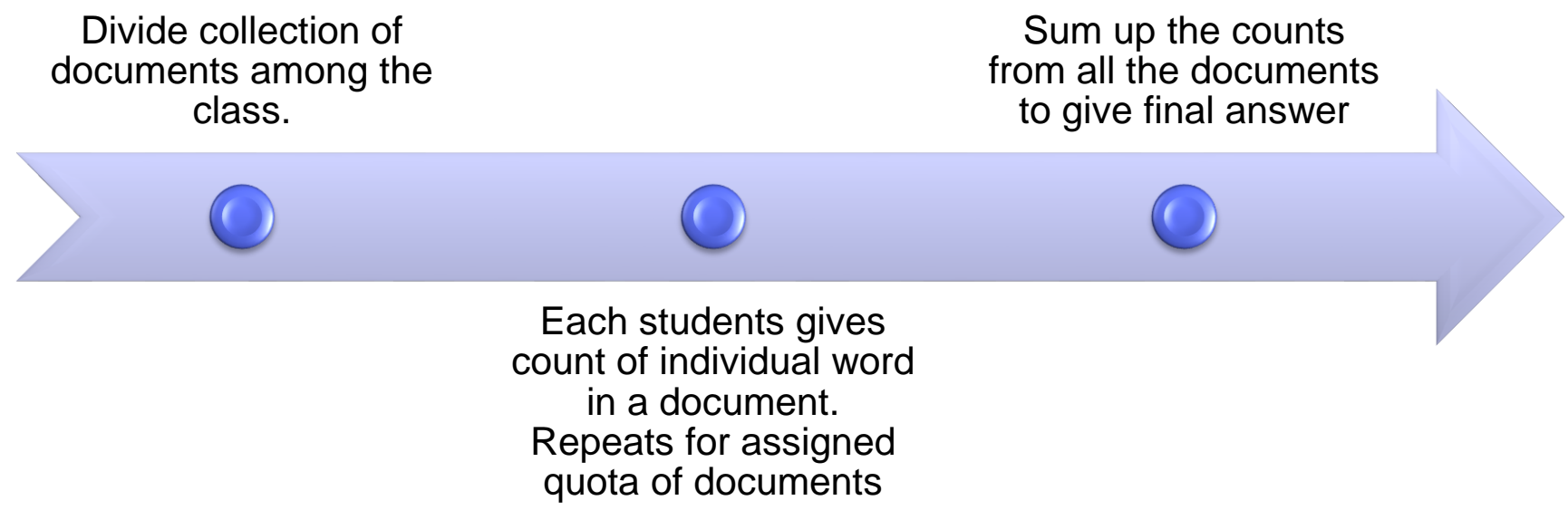


MapReduce

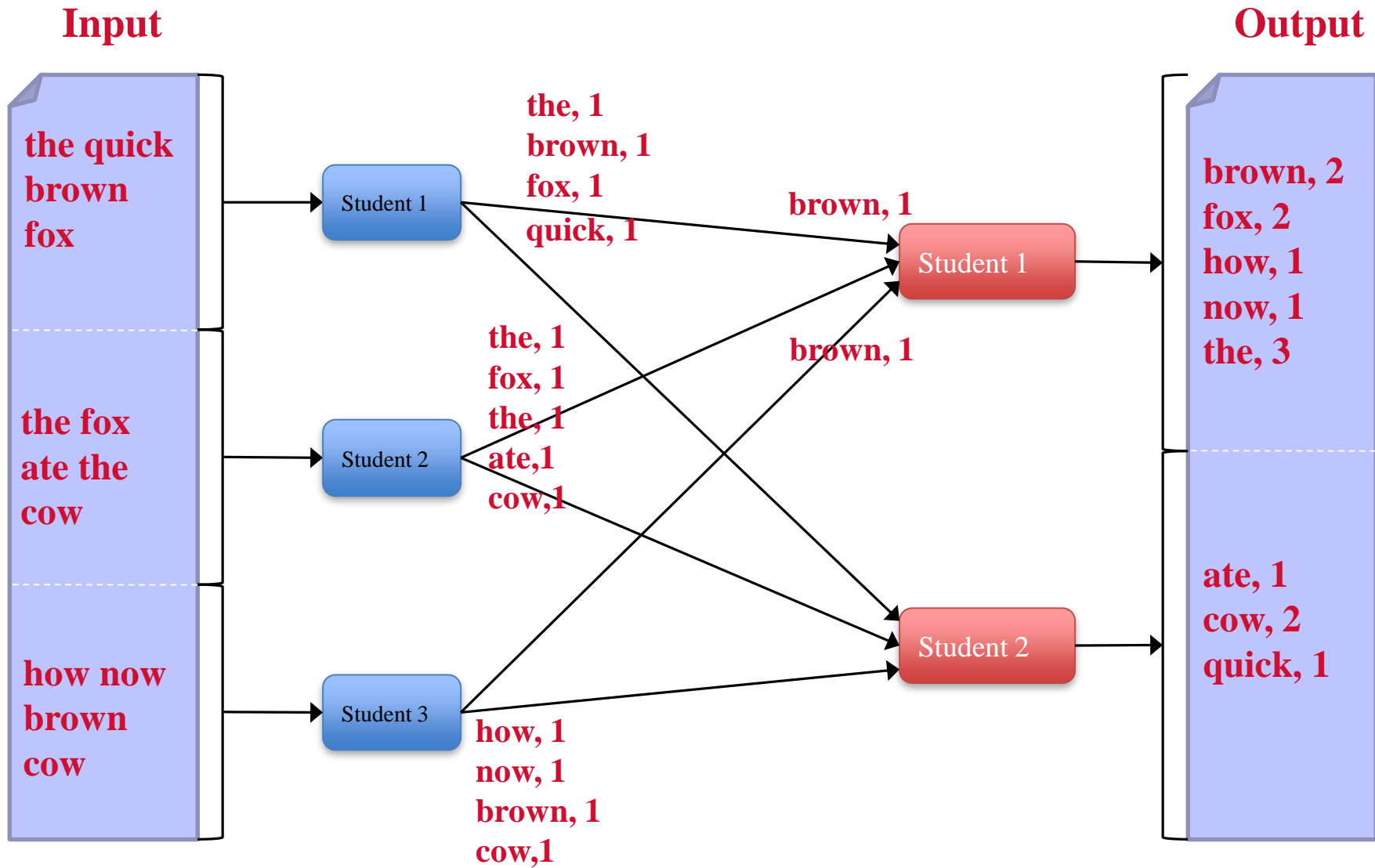
- MapReduce is a method for distributing a task across multiple nodes
- Consists of two phases
 - ▶ Map
 - ▶ Reduce
- The data is process in the form of <key,value>
- Each map task processes a discrete portion of the overall data
- After all Maps are complete, the system distributes the intermediate data to nodes which perform Reduce phase (aggregation)

Example: Word counting

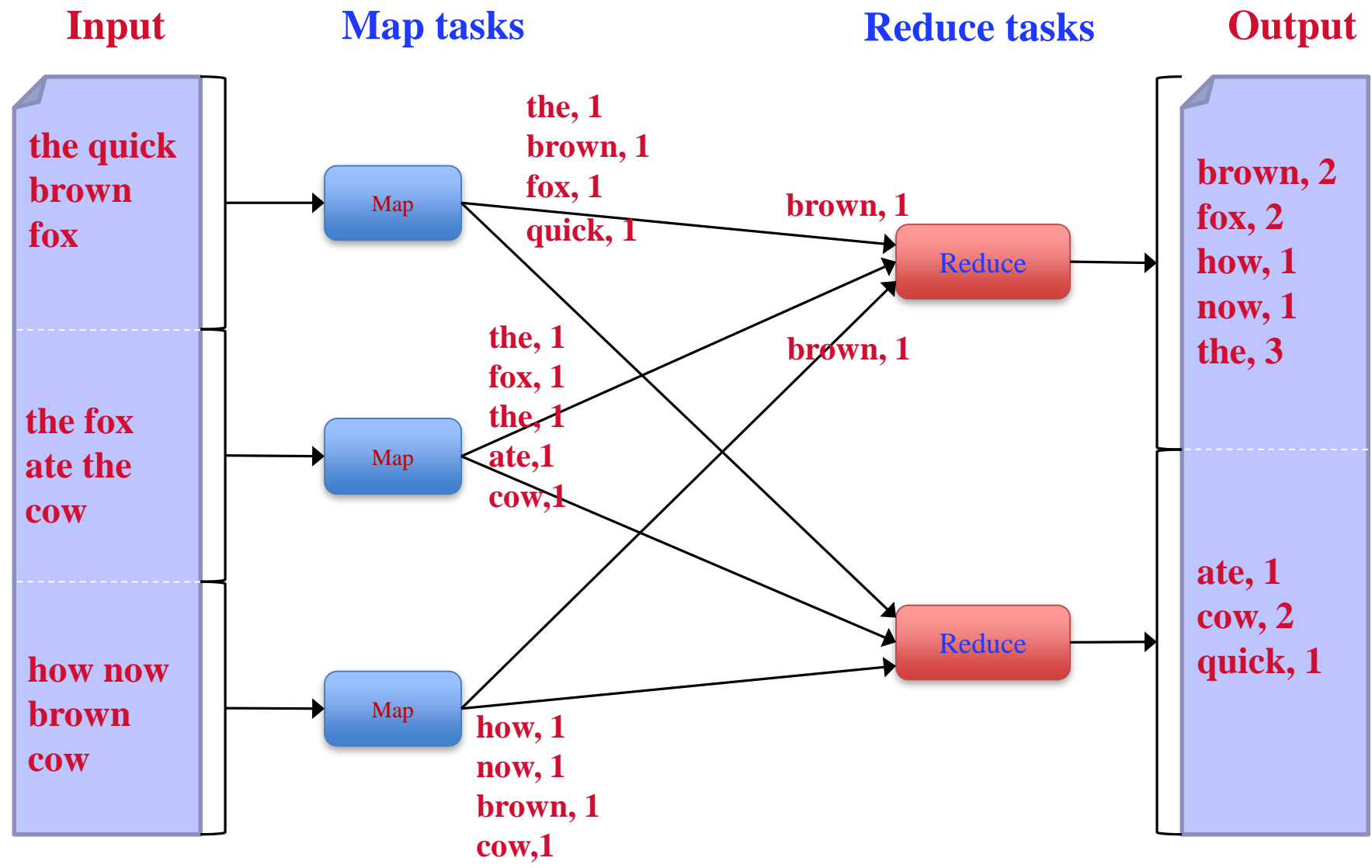
- Counting the number of occurrences of each word in a large collection of documents
 - Input: documents
 - Output: <word,frequency>



Word Count Execution

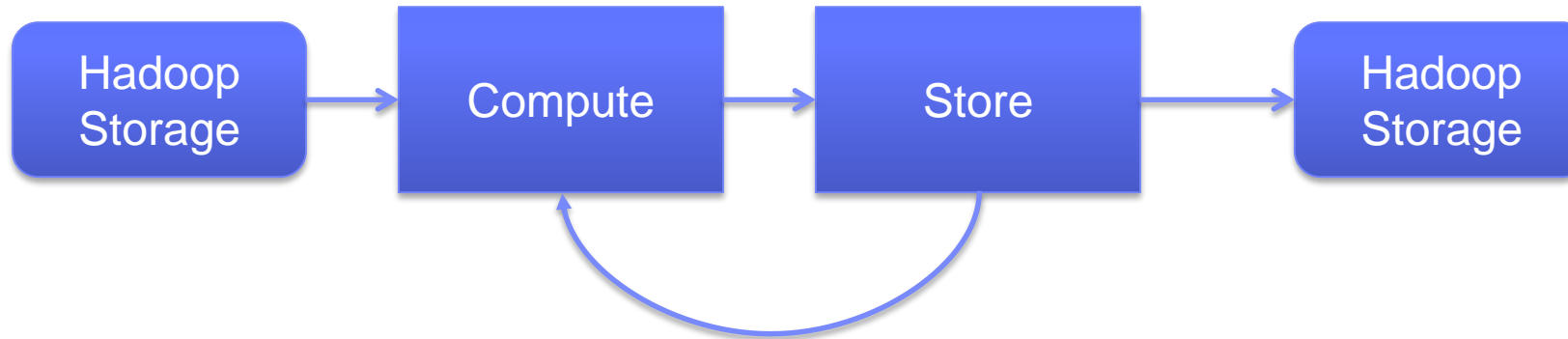


Word Count Execution



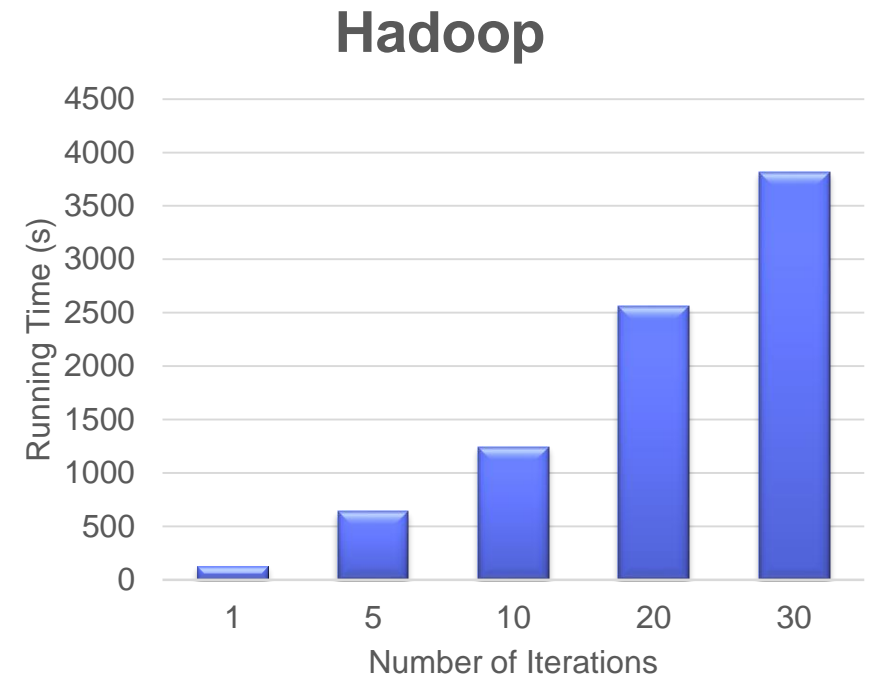
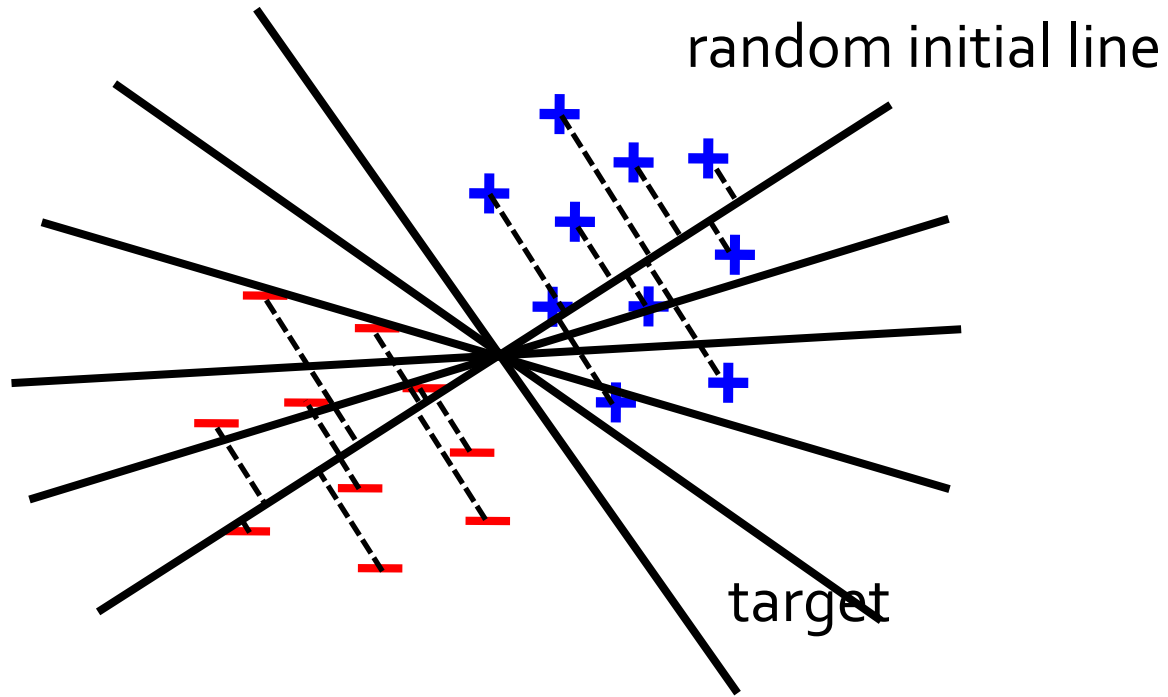
Problem #1

- MapReduce I/O sandbags runtime for advanced analytics.
 - Must persist results after each pass through data
 - Advanced analytics often requires multiple passes through data



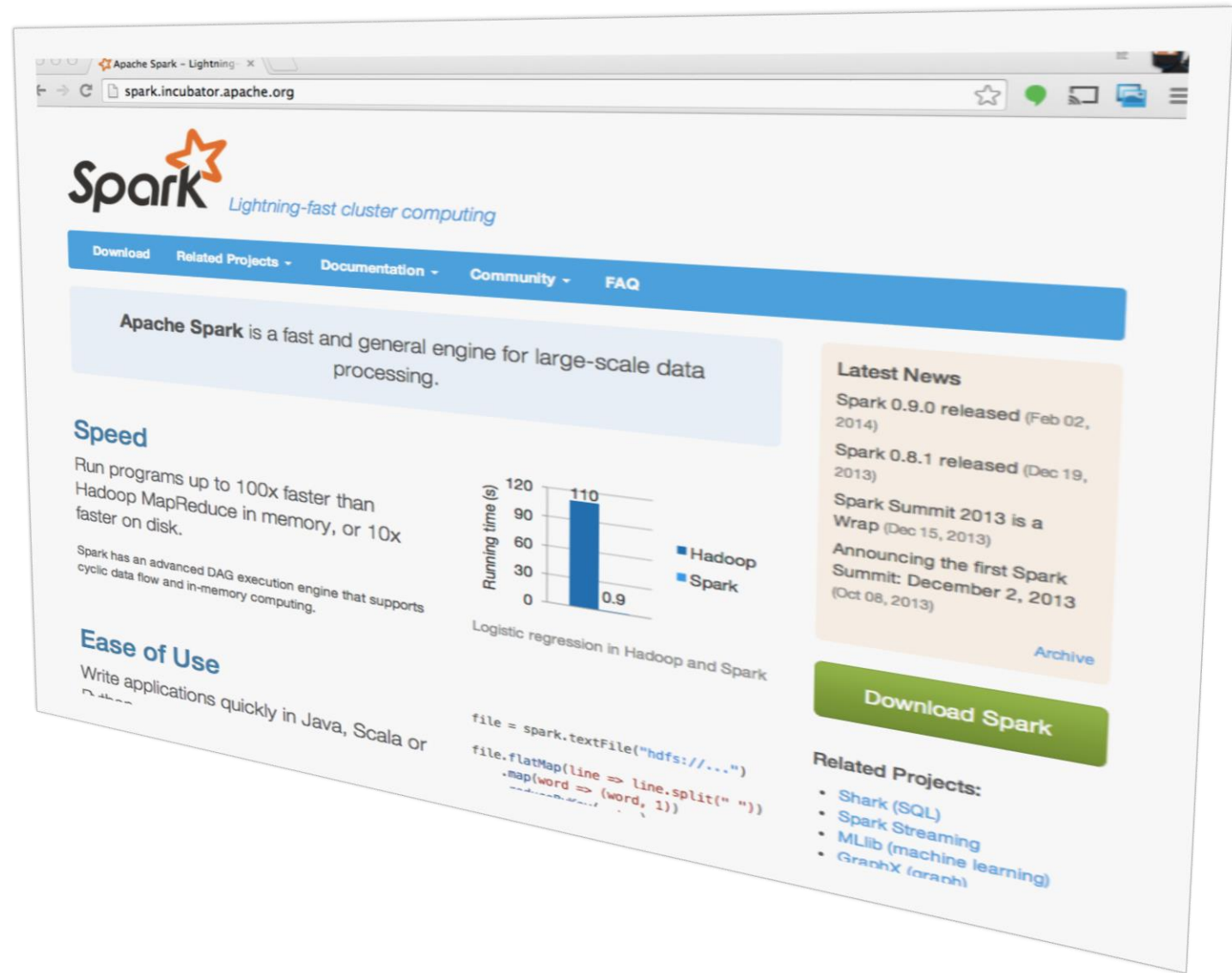
Example

- Goal: find best line separating two sets of points



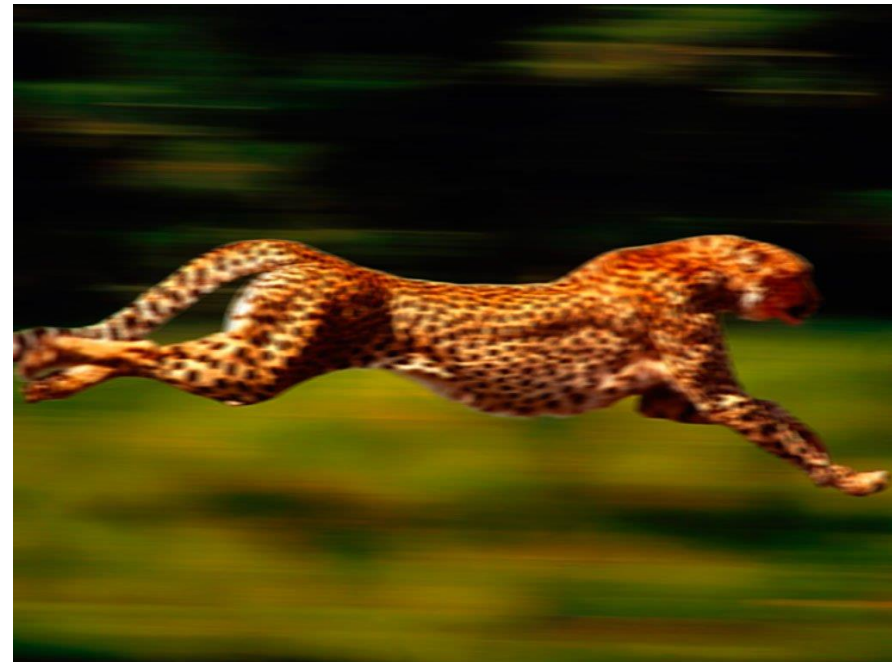
Apache Spark

- A response to limitations in the MapReduce
- UC Berkeley's AMP Lab (2009)
- Fully open sourced in 2010



Spark Performance

- **Machine Learning**
 - ▶ **100x** faster than MapReduce
- **Queries (Shark)**
 - ▶ **100x** faster than Hive
- **Streaming**
 - ▶ **2X** throughput of Storm
- **Graph (GraphX)**
 - ▶ **10X** faster than MapReduce



Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion

The Globe and Mail

www.theglobeandmail.com

GlobalVeyDay In every corner of the globe, there's something to celebrate. Watch the video #GlobalVeyDay

THE GLOBE AND MAIL Enter a term, stock symbol or company name Search Login Register Subscribe Select City Help

Home News Opinion Business Investing Sports Life Arts Tech Drive Real Estate AdChoices

INVESTIGATIONS • CORRESPONDENTS • PUBLIC EDITOR • FEATURED REPORTS • PUZZLES • WATCHLIST • GLOBE UNLIMITED • HOROSCOPES

Ottawa promises to overhaul mental-health services for military
 • INVESTIGATION Remembering 31 Afghan war vets lost to suicide

Trump taps Flynn, Sessions and Pompeo for top positions
 • ALSO Trump reaches out to former foes as he builds team

Teenage Tory wins Ontario by-election, Liberals hold Ottawa-Vanier

Higher down payments needed to battle housing risks: CMHC CEO

Global bonds poised for steepest two-week loss in quarter century

UNLIMITED It pays - big - to be a Canadian CEO

Top military physician skeptical about toxicity of malaria medication

Food prices fall for first time since 2000

JORDAN WESTALL
 People who use opioids are dying every day. Why won't Ottawa talk to us?

GLOBE EDITORIAL
 Call Canada's Iraq mission by its real name: It's war

GARY MASON
 Not so progressive: Trump-style politics seep into Alberta

MUST WATCH

Credit Score

REPORT ON BUSINESS
 Why millennials should care about their credit scores

Try Globe Unlimited 99¢ per week for the first 4 weeks ALREADY A SUBSCRIBER? LOG IN

Digital Media

- Dataset Characteristics:
 - ▶ Attribute: 246
 - ▶ Year: 2014-2015
 - ▶ Size: 2,842 GB

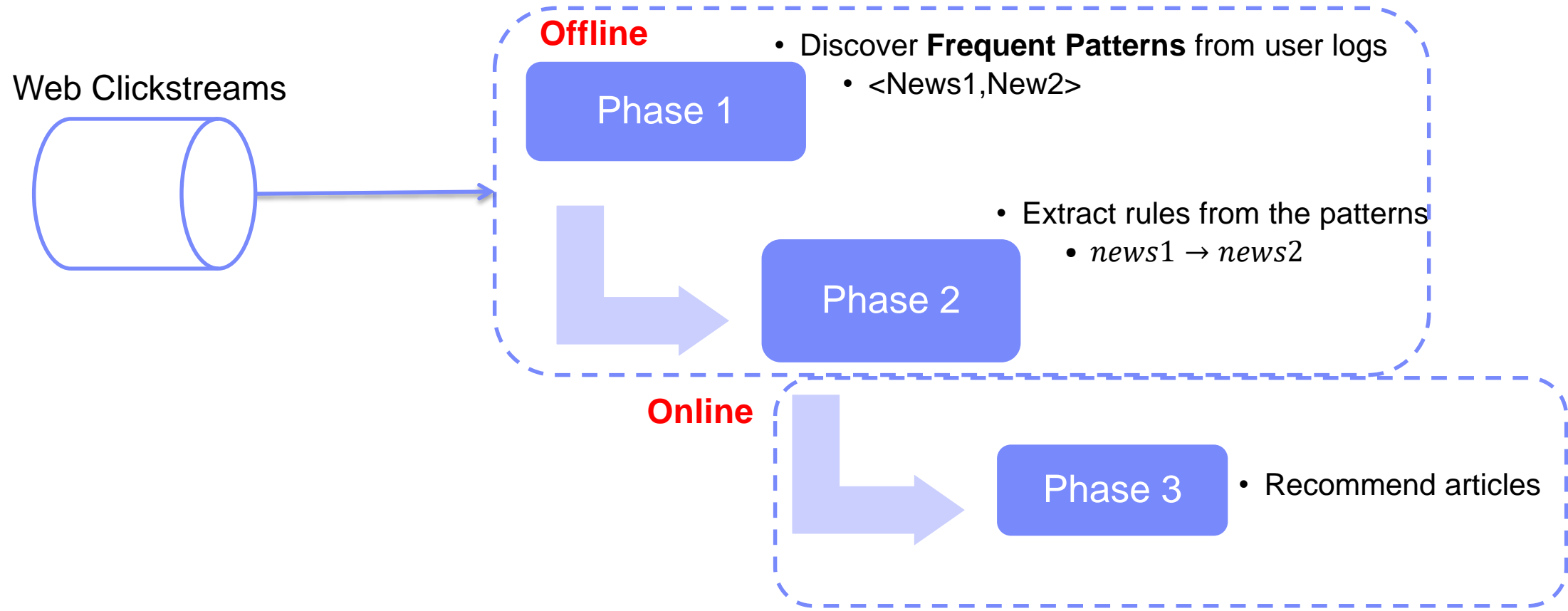
 - ▶ Year 2014 (Jan-Jul):
 - Number of Records: 264,735,412
 - Number of Visits: 51,748,518
 - Number of Users: 19,760,853

Preprocessing

- Data Preprocessing and Cleaning
 - ▶ Filtering out irrelevant hits
 - ▶ Extracting the user types
 - ▶ Extracting the event of interest
 - ▶ Computing the time spent
 - ▶ Roll-up from hit to visit and the user
 - ▶ Converting to expressive format (i.e., json)
 - ▶ ...

Frequency-based News Recommendation

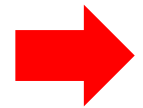
- News articles are not independent
 - ▶ Pattern(behavior): a list of visited articles
 - ▶ Finding sets of co-occurrence news articles



Frequent Pattern Mining

- Frequent Pattern Mining (FPM)
 - ▶ FPM is a fundamental research topic in data mining
 - ▶ Example application
 - Discover sets of items (i.e., itemsets) that are frequently purchased together by customers

TID	Transaction
T_1	{Bread, Milk}
T_2	{Bread, Milk}
T_3	{Bread, Milk, Diaper, Beer}
T_4	{Bread, Milk, Diaper, Beer}
T_5	{Diamond, Necklace}
T_6	{Diamond, Necklace}

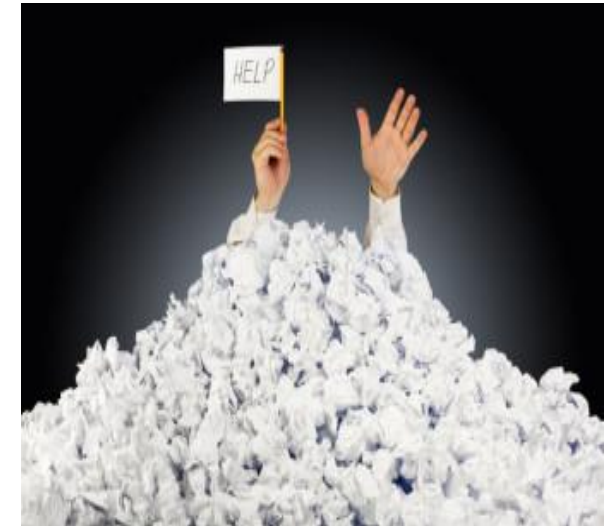


Minimum support threshold: 60%
 $Sup(\{Bread, Milk\}) = 4/6 = 66.6\%$
 {Bread, Milk} is a frequent itemset



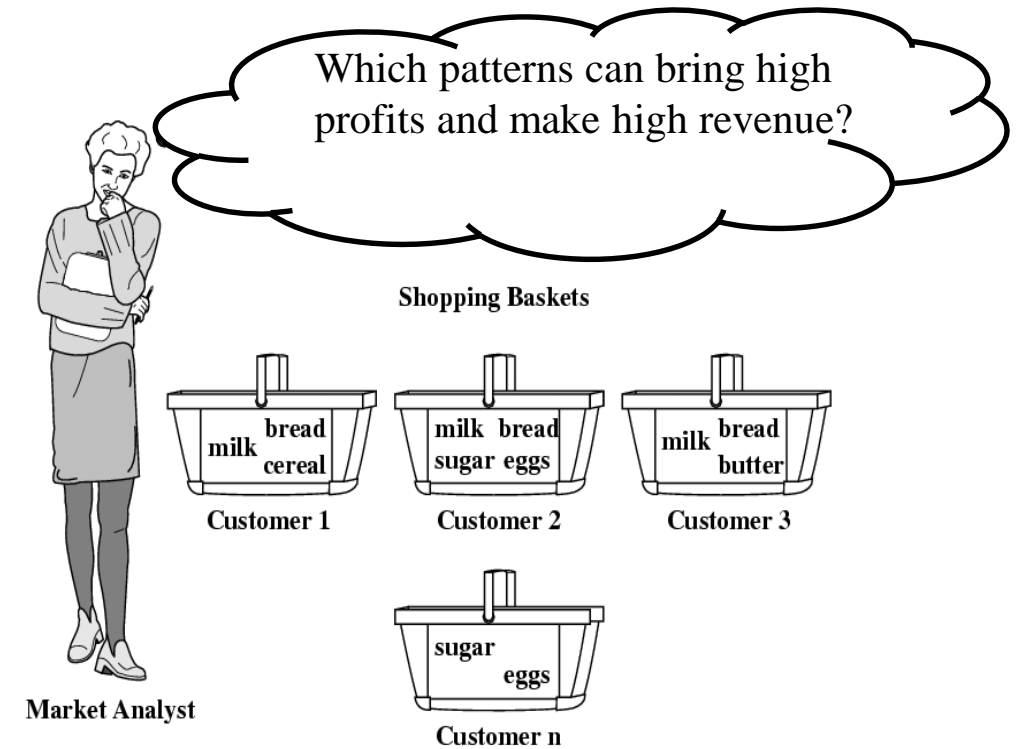
Domain Driven Actionable Knowledge Discovery

- The identified patterns are handed over to business people
- They cannot interpret the patterns for business use
 - ▶ There are many patterns/not informative
 - ▶ Not interested to business needs.
 - ▶ How to interpret the patterns to business actions



Insufficiency of Frequent Pattern Mining

- In Market Analysis
 - ▶ Business objective: **Increase Revenue**
 - ▶ May lose infrequent but valuable patterns
 - ▶ May present too many frequent but unprofitable patterns
 - ▶ Cannot find patterns having high profits



A Motivation Example

TID	Transaction
T_1	{Bread(1), Milk(1)}
T_2	{Bread(1), Milk(1)}
T_3	{Bread(1), Milk(1), Diaper(3), Beer(6)}
T_4	{Bread(1), Milk(1), Diaper(3), Beer(6)}
T_5	{Diamond(1), Necklace(1)}
T_6	{Diamond(1), Necklace(1)}

Item	Unit Profit
Bread	20
Milk	30
Diamond	1,000
Necklace	300
Diaper	300
Beer	70



{Bread, Milk}: \$200



{Diamond, Necklace}: \$2600



{Diaper, Beer}: \$2640



High Utility Sequential Pattern Mining

- Given a set of sequences: find all **sequences** whose **utility** is $>$ a user-specified minimum threshold
 - ▶ Each item has quantity in a transaction
 - ▶ Each item has a value (e.g., price)

Items	Profit
Milk	\$3
Egg	\$2
Birthday Cake	\$20
Birthday Card	\$10
Bread	\$1

CID	TID	
C1	T1	{(Bread,2), (Milk,6)}
C1	T2	{(Birthday Card,2)}
C1	T3	{ (Birthday Cake,2), (egg,3)}
C2	T3	{(Bread,2), (Milk,4), (Yoghurt,3), (Tuna,5)}
C2	T4	{(egg,5), (Pizza,4), (Juice,2)}
C3	T5	{ (Bread,2),(Yoghurt,4), (Milk,3)}
C3	T6	{(Milk,1), (cheese,2)}

What is utility?

- Utility of item in a transaction = internal utility (quantity of items in the transaction) x external utility (profit of the item).
 - ▶ $U(\text{Milk}, T1) = 3 \times 6 = 18$
- Utility of itemset in a sequence = sum of utilities of its items:
 - ▶ $U(\{\text{Bread, Milk}\}, C1) = 2 \times 1 + 6 \times 3 = 20$
- Utility of sub-sequence in a sequence = sum of its itemsets' utilities
 - ▶ $U(\langle \{\text{Milk}\} \{\text{egg}\} \rangle, C1) = 3 \times 6 + 3 \times 2 = 24$
 - ▶ If more than one occurrence, then maximum value among occurrences

Items	Profit
Milk	\$3
Egg	\$2
Birthday Cake	\$20
Birthday Card	\$10
Bread	\$1

CID	TID	
C1	T1	{(Bread,2), (Milk,6)}
C1	T2	{(Birthday Card,2)}
C1	T3	{(Birthday Cake,2), (egg,3)}
C2	T3	{(Bread,2), (Milk,4), (Yoghurt,3), (Tuna,5)}
C2	T4	{(egg,5), (Pizza,4), (Juice,2)}
C3	T5	{(Bread,2), (Yoghurt,4), (Milk,3)}
C3	T6	{(Milk,1), (cheese,2)}

Problem 1: Actionability



→ Actionability → Business objective



News₂



News₃



News₆



News₇



- **Frequent patterns**
 - **75%** frequency
 - **2** minutes

- **Actionable patterns**
 - **30%** frequency
 - **12** minutes

Problem 2: News cold-start

Recent news



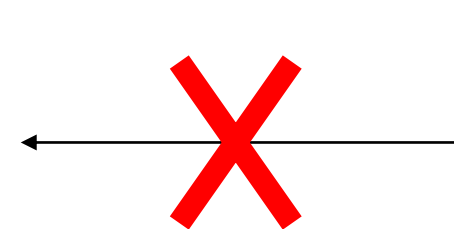
Visited articles



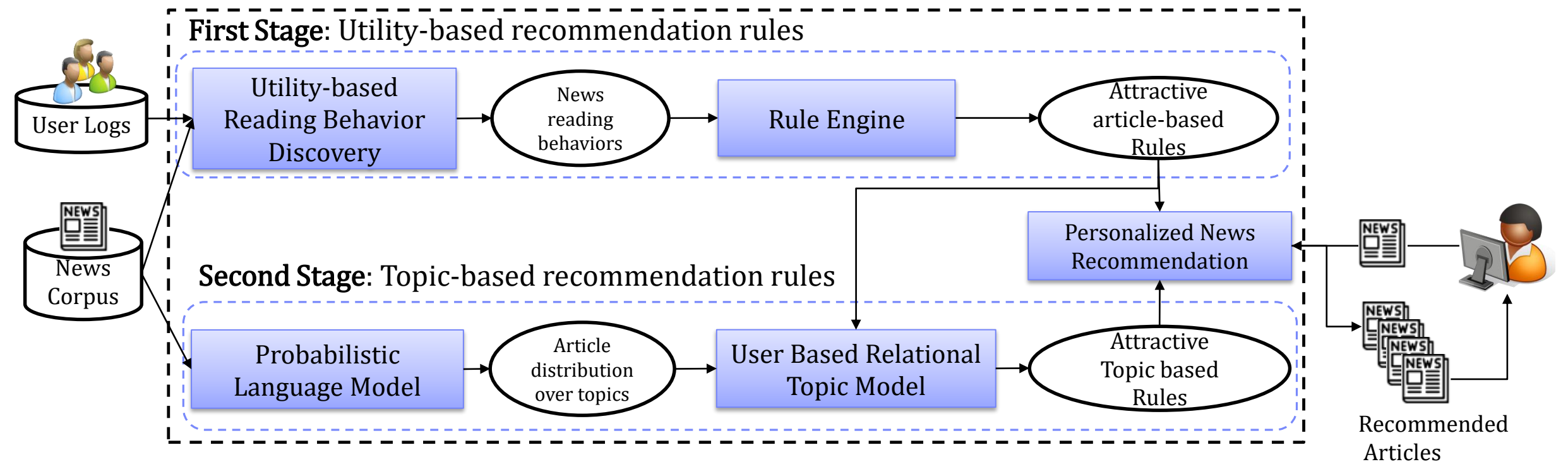
Reading behavior



Newly-published articles



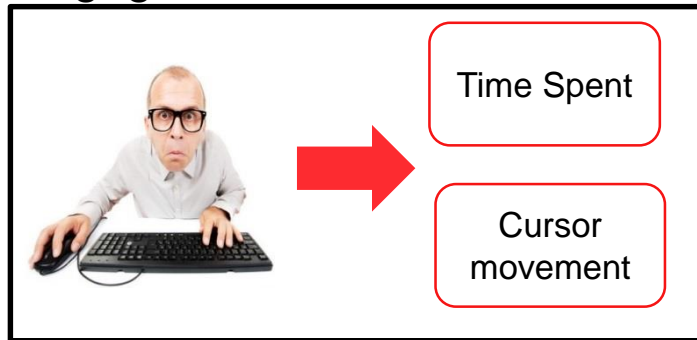
PENSYS: The Proposed Framework



FIRST STAGE: NEWS LEVEL

Stage 1: Utility-based Pattern Mining

Engagement Measures



Internal Utility Function (F)

External knowledge



External Utility Function (G)

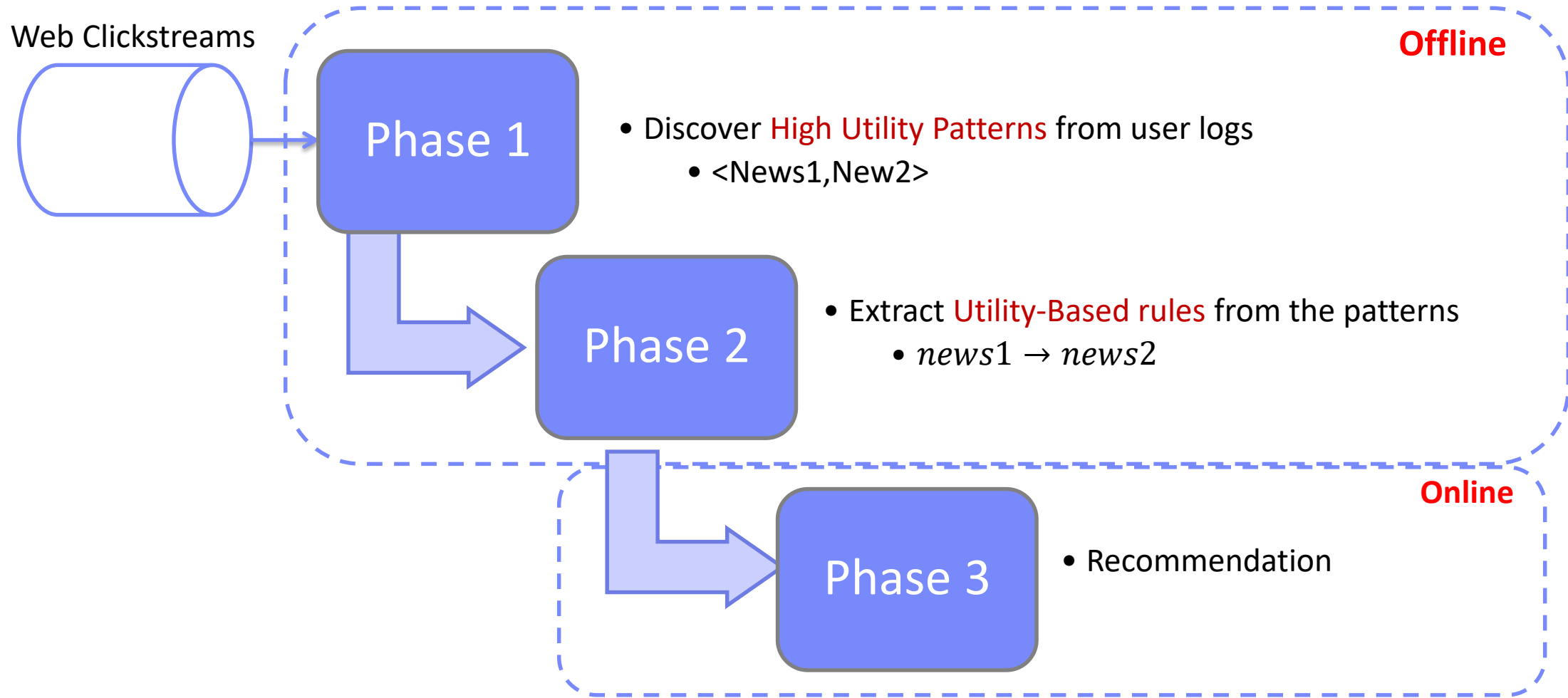
$$Utility = U(F, G)$$

Attractive news reading behavior

Patterns with Utility \geq **Threshold**

$$Utility(news) = Time\ spent \times Freshness$$

Stage 1 Overview



Top-2 High Utility Patterns

Num.	Set of news	Time (mins)	Support
1.	Vigil held for daughter of Conservative Party president	2537	107
	MH17: Disaster ratchets up Russia-Ukraine tensions		
2.	Retiree, 60, wonders how long her money will last ,	1473	102
	Which is better, a RRIF or an annuity You may be surprised		

Top-2 Frequent Patterns

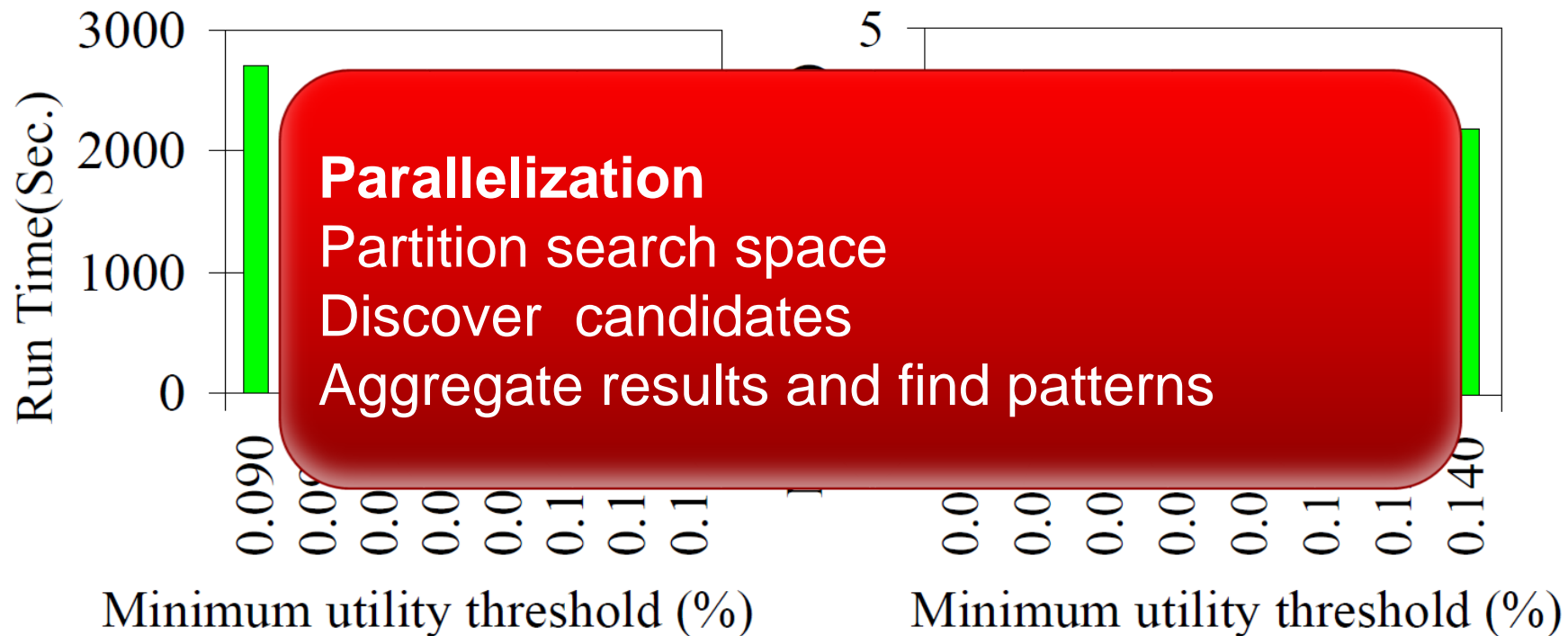
Num.	Set of news	Time (mins)	Support
1.	Target faces calls to withdraw from Canada ,	144	254
	Mike Duffy facing 31 charges from Senate expenses scandal, RCMP says		
2.	Florida police say , La Prairie, Quebec mayor dies from wasp stings	105	286
	Canadian professor was killed in targeted attack		

Recommendation Rules

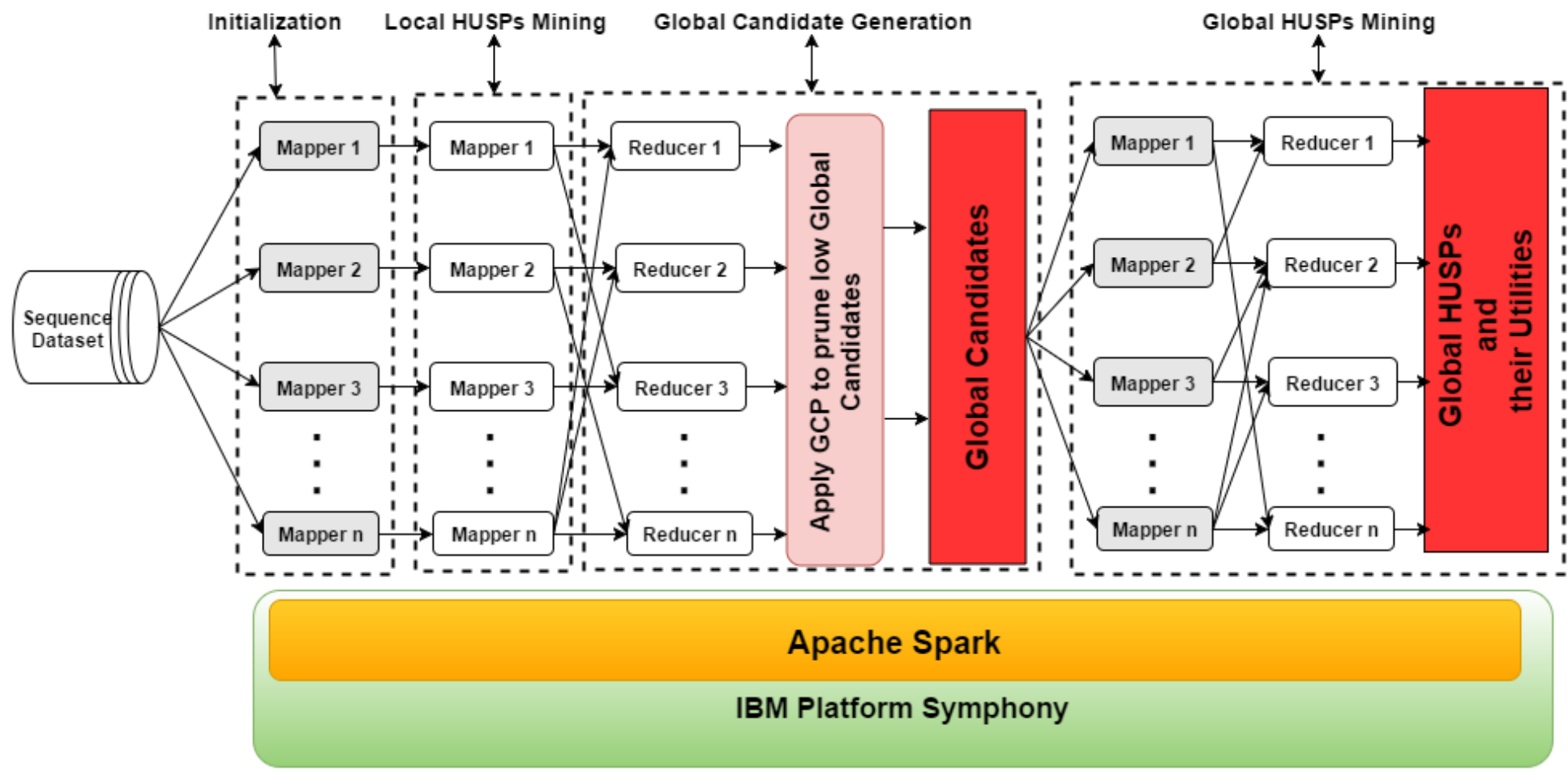
Top 3 of our Utility-Based Association Rules, sorted by uconf, time spent in descending order	Time Spent (minute)
<p>[Researchers find dozens of genetic links to schizophrenia] ==> [Anti-Semitic graffiti found in Thornhill, days after Islamic centre defaced] [Video: Fearless 93-year-old tackles CN Tower EdgeWalk] [Teen use of human growth hormone on the rise] [Canadian poised to become terror tourist given 10-year sentence] [Ontario Liberals waste no time playing hardball with opposition] [The Gaza war has done terrible things to Israeli society]</p>	239.98
<p>[Researchers find dozens of genetic links to schizophrenia] ==> [Anti-Semitic graffiti found in Thornhill, days after Islamic centre defaced] [Teen use of human growth hormone on the rise] [Canadian poised to become terror tourist given 10-year sentence] [Ontario Liberals waste no time playing hardball with opposition] [The Gaza war has done terrible things to Israeli society]</p>	239.93
<p>[Researchers find dozens of genetic links to schizophrenia] ==> [Anti-Semitic graffiti found in Thornhill, days after Islamic centre defaced] [Video: Fearless 93-year-old tackles CN Tower EdgeWalk] [Canadian poised to become terror tourist given 10-year sentence] [Ontario Liberals waste no time playing hardball with opposition] [The Gaza war has done terrible things to Israeli society]</p>	239.8

Performance

- Only 116000 records
 - ▶ Over 4 GB memory usage in average
 - ▶ Around 45 mins run time



Big Data Framework

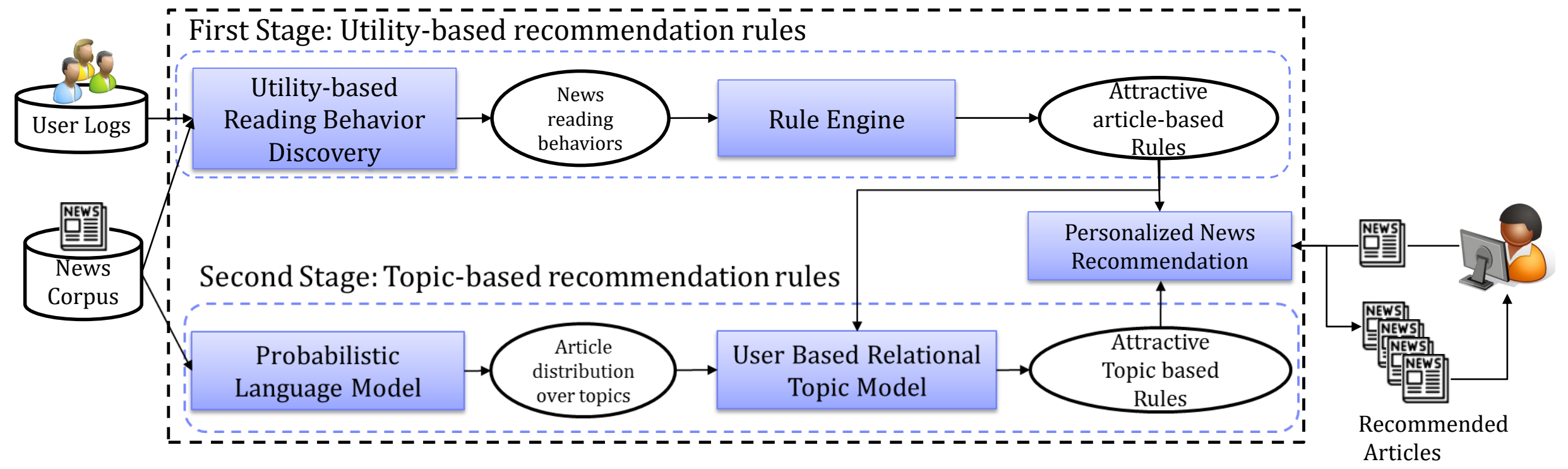


Results

<i>m = Minutes, h = Hours</i>				
<i>Dataset</i>	δ (%)	<i>BigHUSP</i>	<i>BigHUSP_{Basic}</i>	<i>BigHUSP_{SA}</i>
<i>Globe</i>	0.09	1.6 <i>m</i>	3.6 <i>m</i>	0.99 <i>h</i>
	0.08	2.3 <i>m</i>	4.4 <i>m</i>	1.4 <i>h</i>
	0.07	3.1 <i>m</i>	6.6 <i>m</i>	2.2 <i>h</i>
	0.06	5.0 <i>m</i>	11.0 <i>m</i>	3.3 <i>h</i>
	0.05	9.2 <i>m</i>	20.7 <i>m</i>	4.5 <i>h</i>
<i>synthDS1</i>	0.05	3.0 <i>m</i>	10.0 <i>m</i>	1.1 <i>h</i>
	0.04	4.26 <i>m</i>	14.4 <i>m</i>	1.2 <i>h</i>
	0.03	6.23 <i>m</i>	17.9 <i>m</i>	1.6 <i>h</i>
	0.02	9.9 <i>m</i>	27.2 <i>m</i>	1.9 <i>h</i>
	0.01	14.3 <i>m</i>	29.4 <i>m</i>	3.2 <i>h</i>
	0.009	37.6 <i>m</i>	76.5 <i>m</i>	7.8 <i>h</i>
<i>ChainStore</i>	0.09	15.0 <i>m</i>	33.4 <i>m</i>	6.4 <i>h</i>
	0.08	19.9 <i>m</i>	56.2 <i>m</i>	12.0 <i>h</i>
	0.07	25.0 <i>m</i>	77.0 <i>m</i>	13.4 <i>h</i>
	0.06	34.8 <i>m</i>	107.7 <i>m</i>	14.6 <i>h</i>
	0.05	38.7 <i>m</i>	159.8 <i>m</i>	17.4 <i>h</i>
<i>synthDS2</i>	0.09	13.1 <i>m</i>	26.3 <i>m</i>	7.7 <i>h</i>
	0.08	16.3 <i>m</i>	34.5 <i>m</i>	9.3 <i>h</i>
	0.07	20.6 <i>m</i>	47.2 <i>m</i>	15.7 <i>h</i>
	0.06	23.8 <i>m</i>	51.8 <i>m</i>	17.8 <i>h</i>
	0.05	32.2 <i>m</i>	85.3 <i>m</i>	21.4 <i>h</i>

SECOND STAGE: TOPIC LEVEL

PENSYS: The Proposed Framework



Stage 2: A Semantic Relational Topic Model

- News cold start problem
 - ▶ Content based recommendation systems
 - ▶ Similar topics

- User behavior?
 - ▶ Hybrid approach

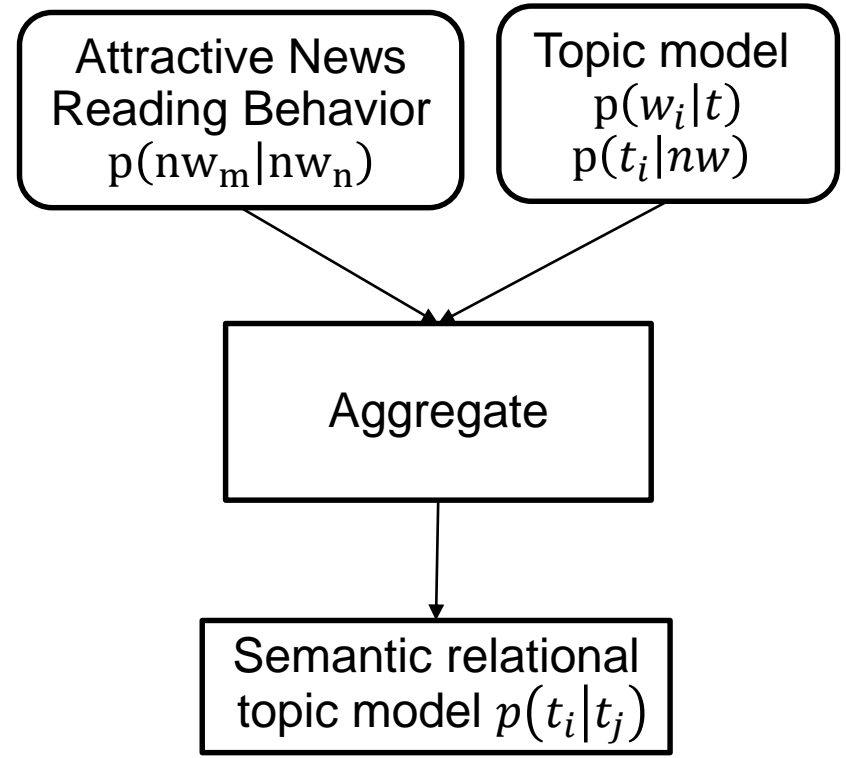
(a)

Topic 1: Health	Topic 2: Debt	Topic 3: Travel	Topic 4: Sport	Topic 5: Politics
$w, p(w Topic1)$	$w, p(w Topic2)$	$w, p(w Topic3)$	$w, p(w Topic4)$	$w, p(w Topic5)$
Drug, 0.5	Debt, 0.5	Hotel, 0.5	Soccer, 0.5	Obama, 0.5
Health, 0.4	Bond, 0.4	Travel, 0.4	Sport, 0.4	Politics, 0.4
People, 0.25	Credit, 0.25	Park, 0.25	NBA, 0.25	Election, 0.25
Disease, 0.15	Investors, 0.15	Mountain, 0.15	Ronaldo, 0.15	War, 0.15
....

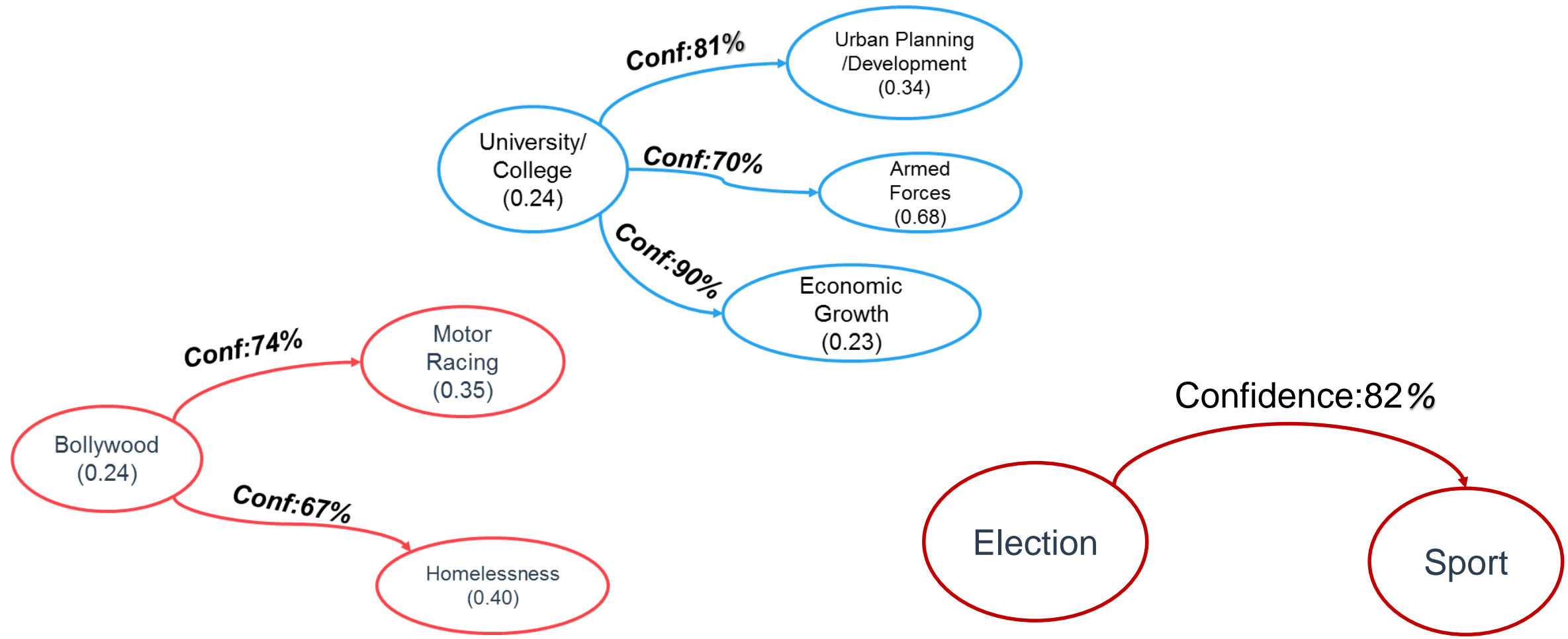
(b)

nw_1	nw_2	nw_3	nw_4	nw_5
$t, p(t nw_1)$	$t, p(t nw_2)$	$t, p(t nw_3)$	$t, p(t nw_4)$	$t, p(t nw_5)$
Topic 1, 0.6	Topic 2, 0.7	Topic 3, 0.6	Topic 4, 0.8	Topic 5, 0.6
Topic 3, 0.2	Topic 3, 0.4	Topic 1, 0.4	Topic 5, 0.4	Topic 3, 0.4
....

Stage 2: A Semantic Relational Topic Model



Topic-based Rules Example



Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion

Depression

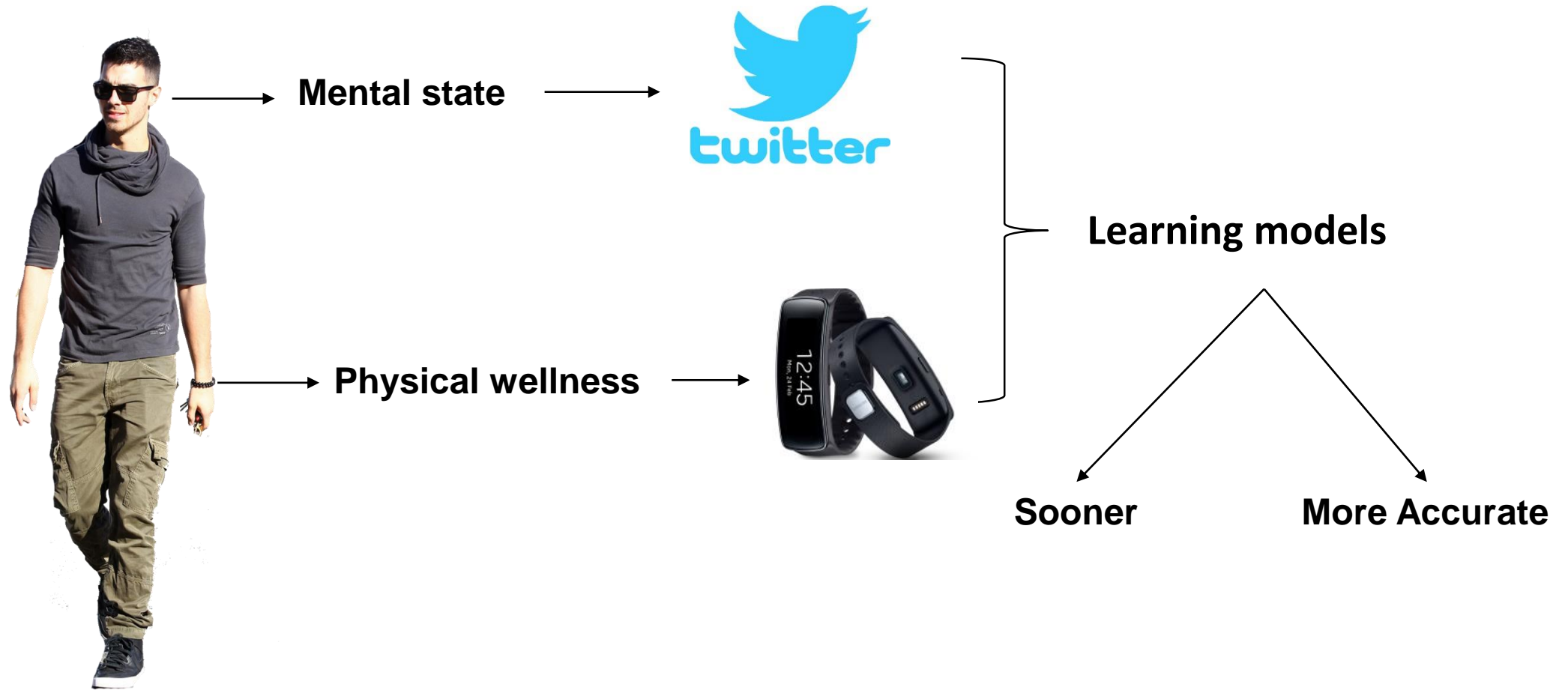


Screening questionnaire

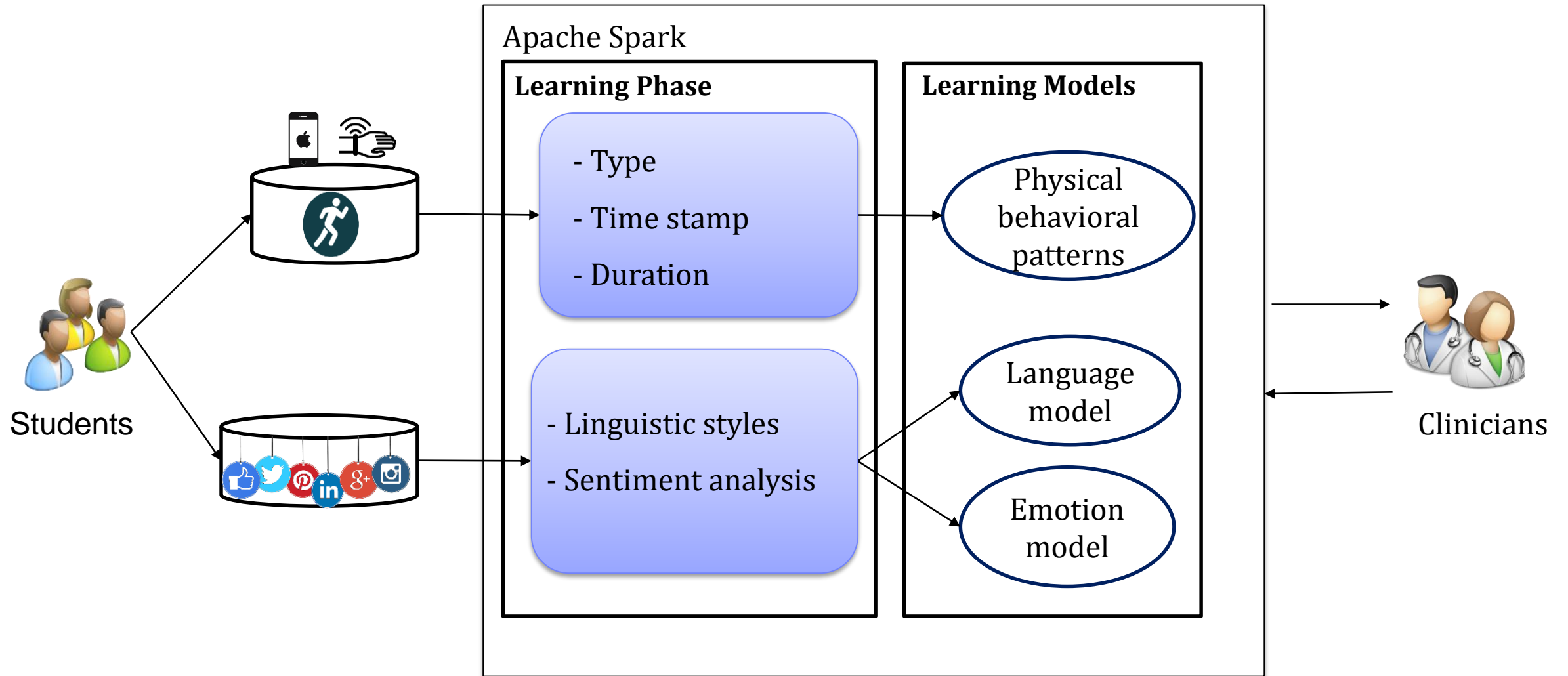
Problems:

- Recalling events
- Affected by current mental state
- Low quality of data

Depression Acuity Detection



Proposed Framework



Results

1. Depression terms

- ▶ 65% higher

2. Mood classification

- ▶ 69% accuracy

3. Physical wellness indicators

- ▶ Time
- ▶ Duration
- ▶ Sequential order of events matters



Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- **Cogniciti: An Online Brain Health Assessment**
- Conclusion

Brain Health

- Motivation
 - ▶ 47 million people suffer from dementia (Alzheimer's Society of Canada)
 - ▶ Early detection of dementia: \$219 billion saving
- Solution
 - ▶ An early warning test for brain health



cogniciti

Baycrest

dapasoft

Assessments

Demographic Test

Tell us about yourself.

1) What is your date of birth?
 Month Year

2) What is your gender?
 Male Female

3) Is English your first language?
 Yes No

4) How many years of school have you completed?

5) Do you have significant concerns about your memory?
 Yes No

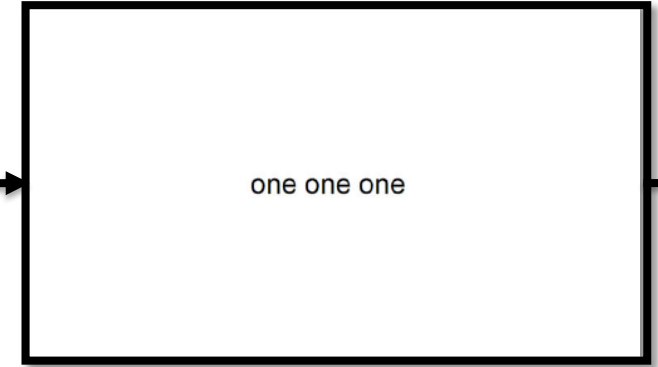
6) Have you been diagnosed with or have a history of any of the following conditions? Please choose all that apply.

<input type="checkbox"/> Alcohol or substance abuse	<input type="checkbox"/> High cholesterol
<input type="checkbox"/> Alzheimer's disease	<input type="checkbox"/> Huntington's disease
<input type="checkbox"/> Anxiety (current)	<input type="checkbox"/> Insomnia or other sleep disorder treated with medication
<input type="checkbox"/> Any cancer treated with chemotherapy	<input type="checkbox"/> Irregular heart rate
<input type="checkbox"/> Attention deficit disorder/learning disability	<input type="checkbox"/> Mild cognitive impairment (MCI)
<input type="checkbox"/> Bipolar disorder	<input type="checkbox"/> Mini-strokes or TIA
<input type="checkbox"/> Brain surgery	<input type="checkbox"/> Multiple sclerosis
<input type="checkbox"/> Brain tumour	<input type="checkbox"/> Parkinson's disease
<input type="checkbox"/> Chronic pain treated with prescription medication	<input type="checkbox"/> Schizophrenia
<input type="checkbox"/> Depression (current)	<input type="checkbox"/> Seizures

Shape match test



Stroop interference



Unified Score

Your brain health score is 98%

Your overall performance is within expectations for your age and education.

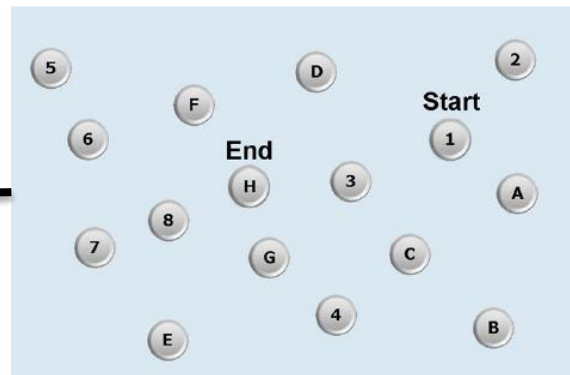
This is what your score means...

- Your score is a percentile ranking that lets you see how you performed relative to adults similar to you. Your score of 98 means that 98% of people your age and education would score lower than you on the assessment and 2% would score higher.
- The range for normal brain health is 7 – 100. Your score is within that range. As a result, there is no indication that further assessment is needed at this time.

This is what we suggest you do next

This report should not be construed as medical advice, diagnosis, treatment, or the provision of health care. This report is not a substitute for advice from a health care professional. Please seek advice from your family doctor or other qualified professional if you have any questions about this report or any matter related to your health.

Letter-number alternation



Face-name association





Strong consumer response from one Canadian media release

Test re-test
reliability

72%

Site Visits

153,000

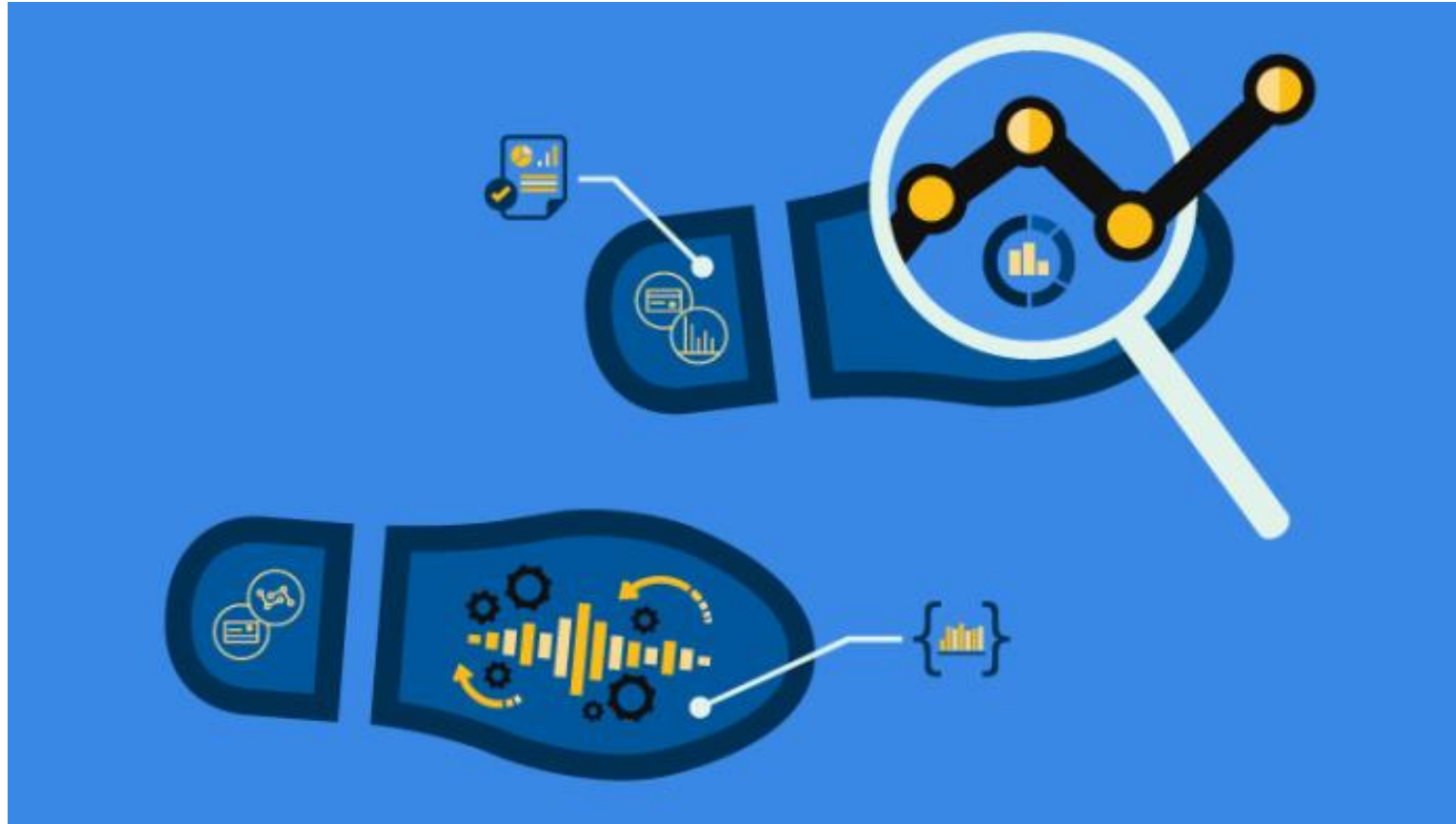
Completed
Assessments

41,000

Conclusions

- User behavior analysis in big data is an important area of research
- We have done some (hopefully) interesting work in this area
 - ▶ Utility-based pattern discovery in big data streams
 - ▶ Depression acuity detection
 - ▶ Online Brain Health Assessment
- A lot more research needs to be done!





User Behavior Analysis in Big Data
mori.zihayatkermani@utoronto.ca

Morteza Zihayat