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

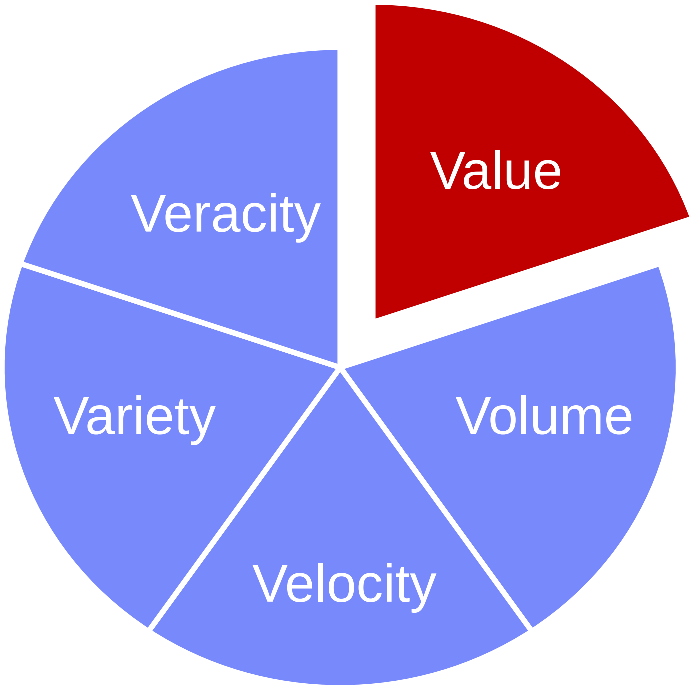
HEALTHCARE DATA

200 Gigabytes

THE SIZE OF A SINGLE SEQUENCED HUMAN GENOME

50 MILLION DEVICES

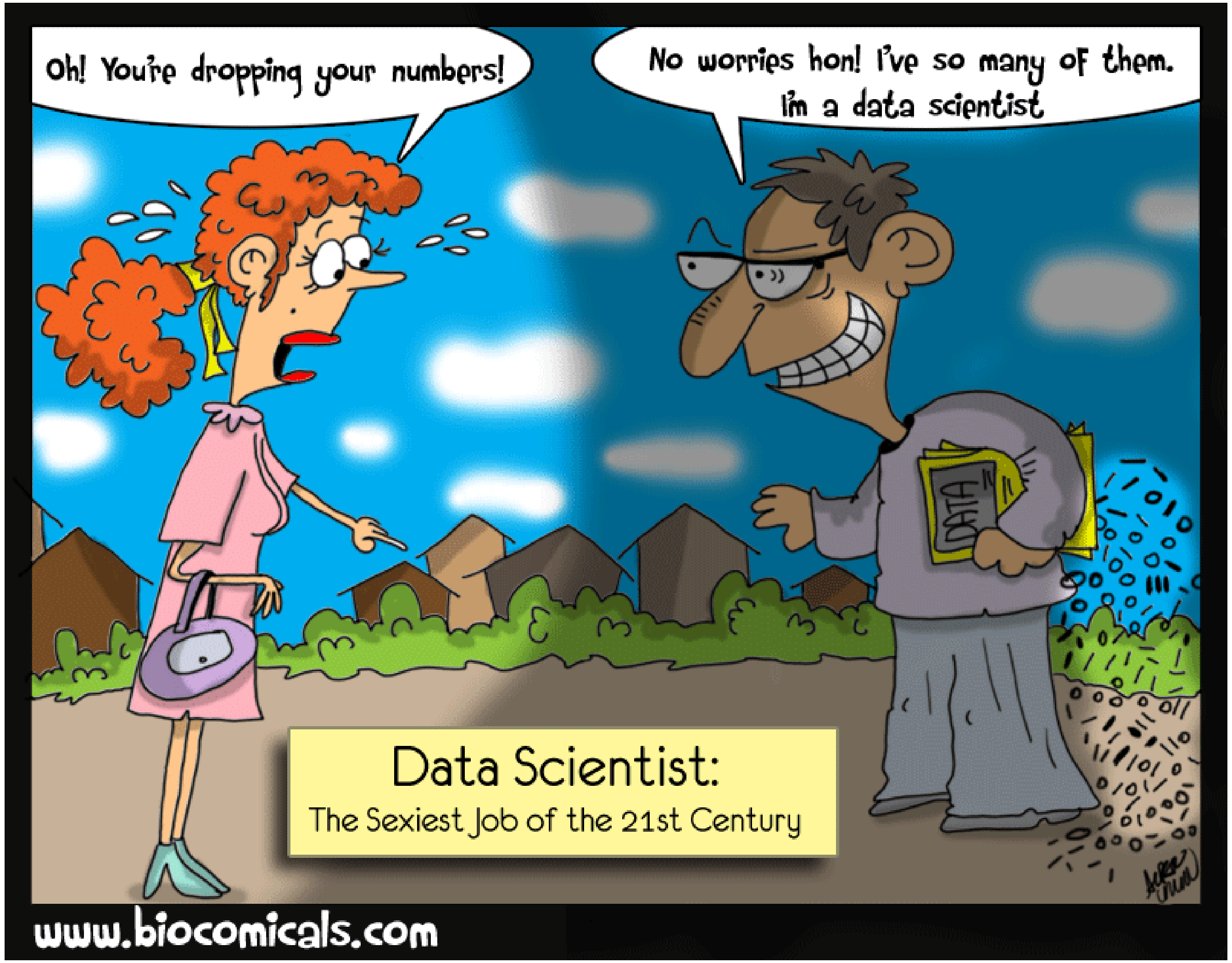
EACH DEVICE GENERATES **156 MB** OF DATA PER **DAY**



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

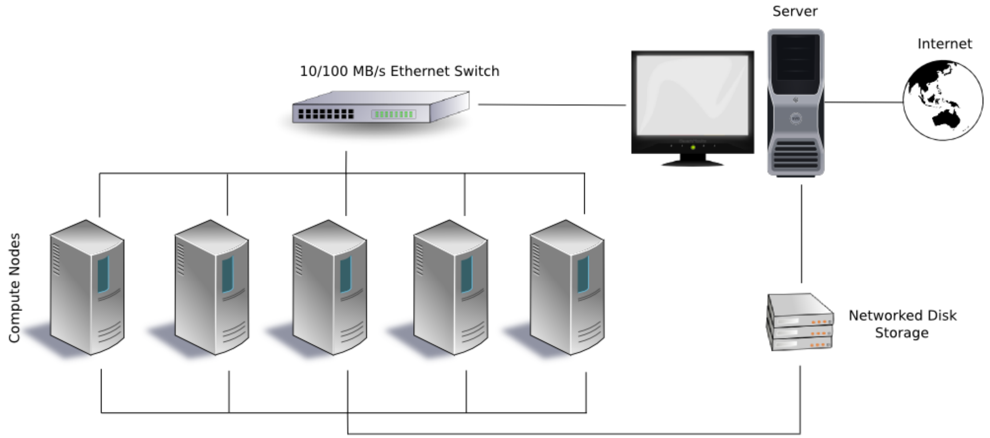
Challenges

- 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

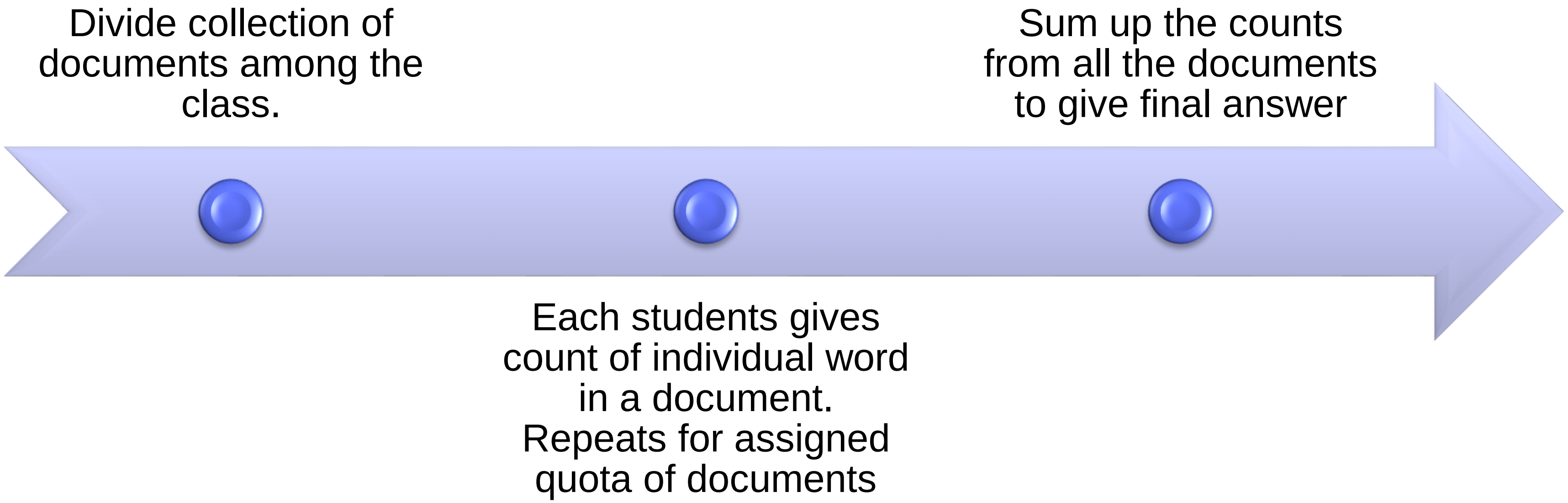


MapReduce

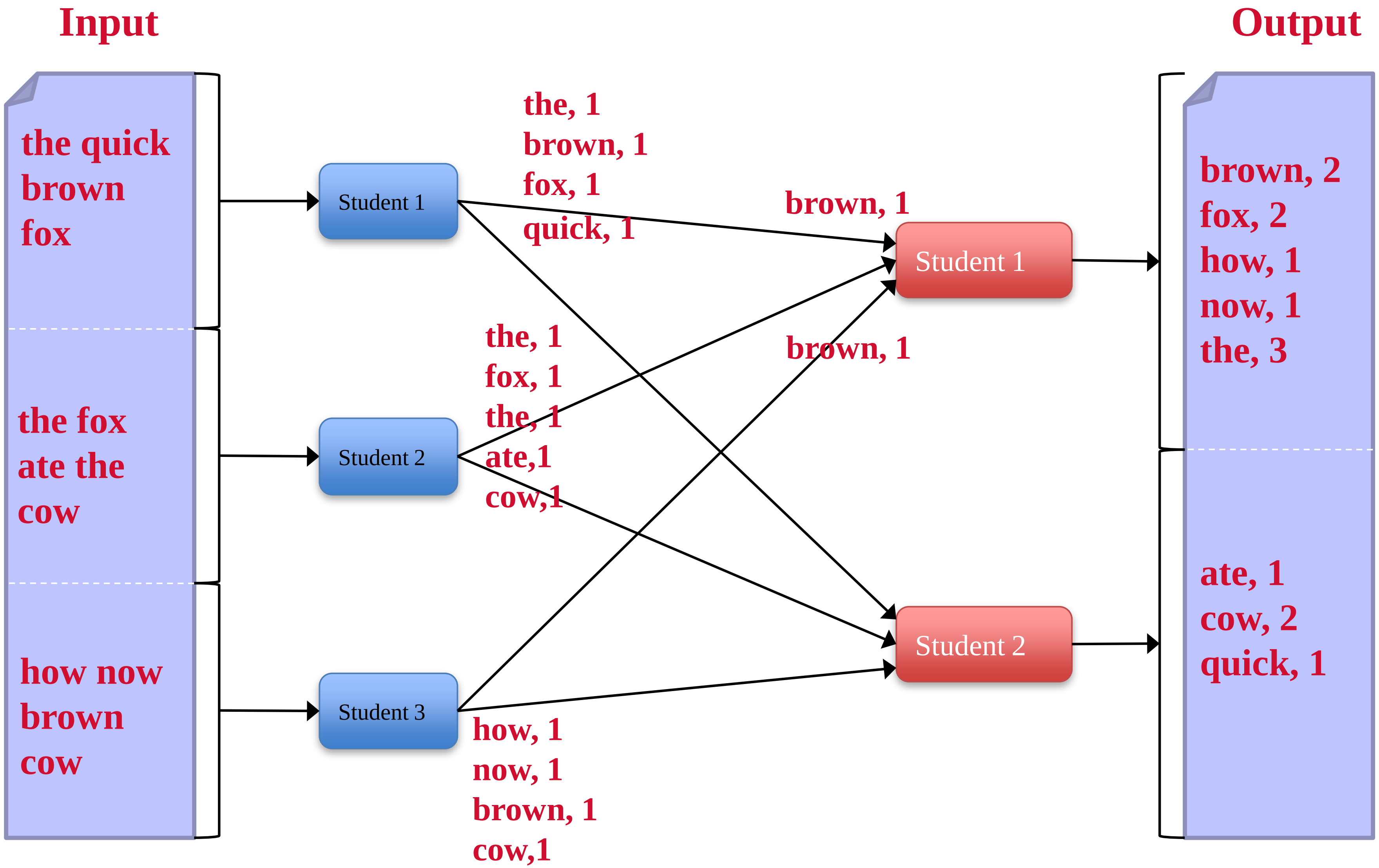
- MapReduce is a method for distributing a task across multiple nodes
- Consists of two phases
  - Map
  - Reduce
- The data is process in the form of <key,value>
- Each map task processes a discrete portion of the overall data
- After all Maps are complete, the system distributes the intermediate data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

Example: Word counting

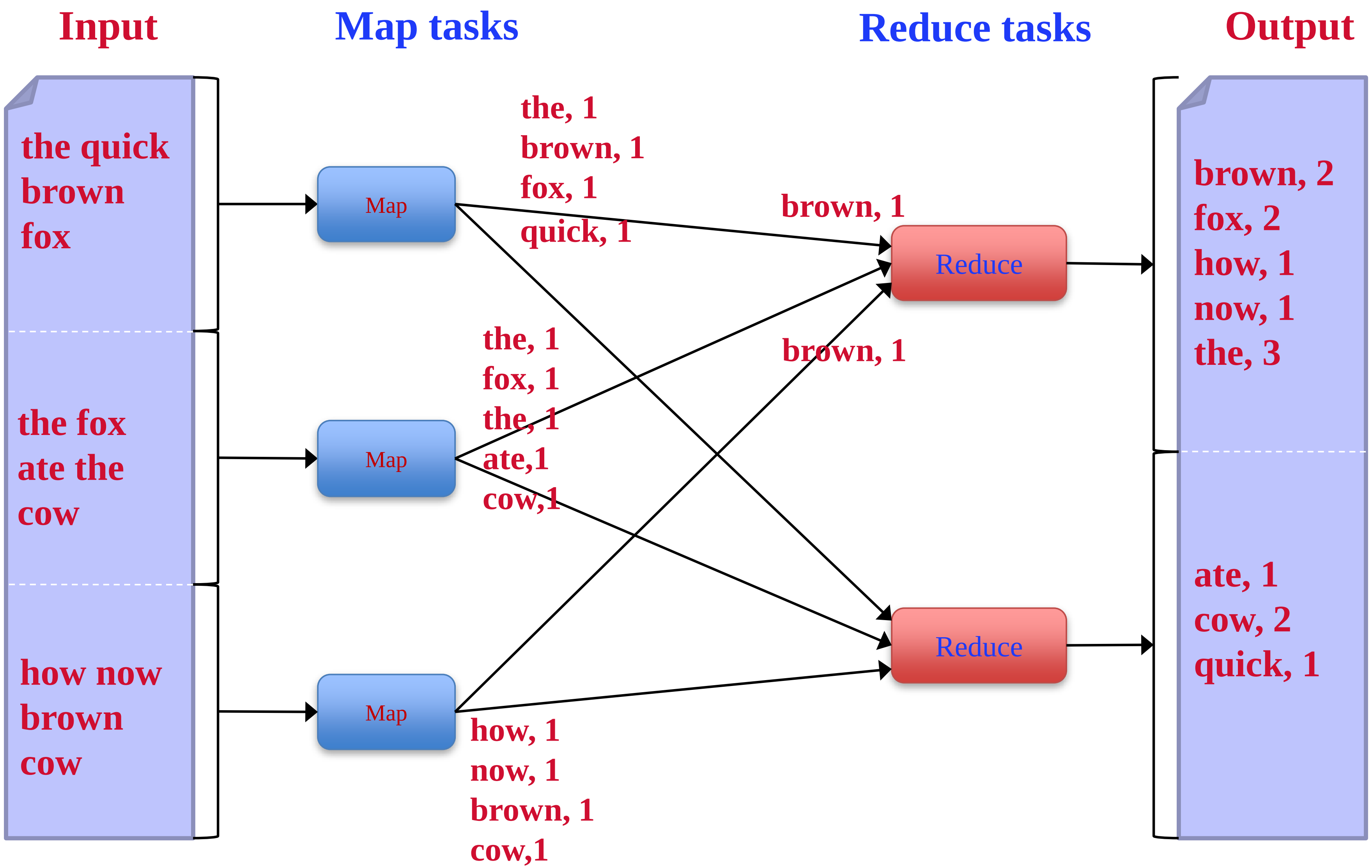
- Counting the number of occurrences of each word in a large collection of documents
    - Input: documents
    - Output: <word,frequency>



Word Count Execution

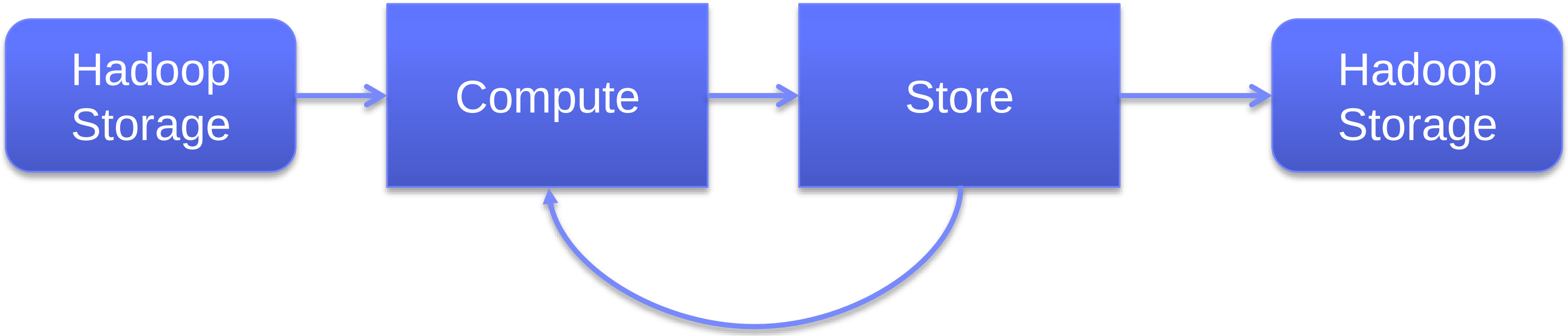


Word Count Execution



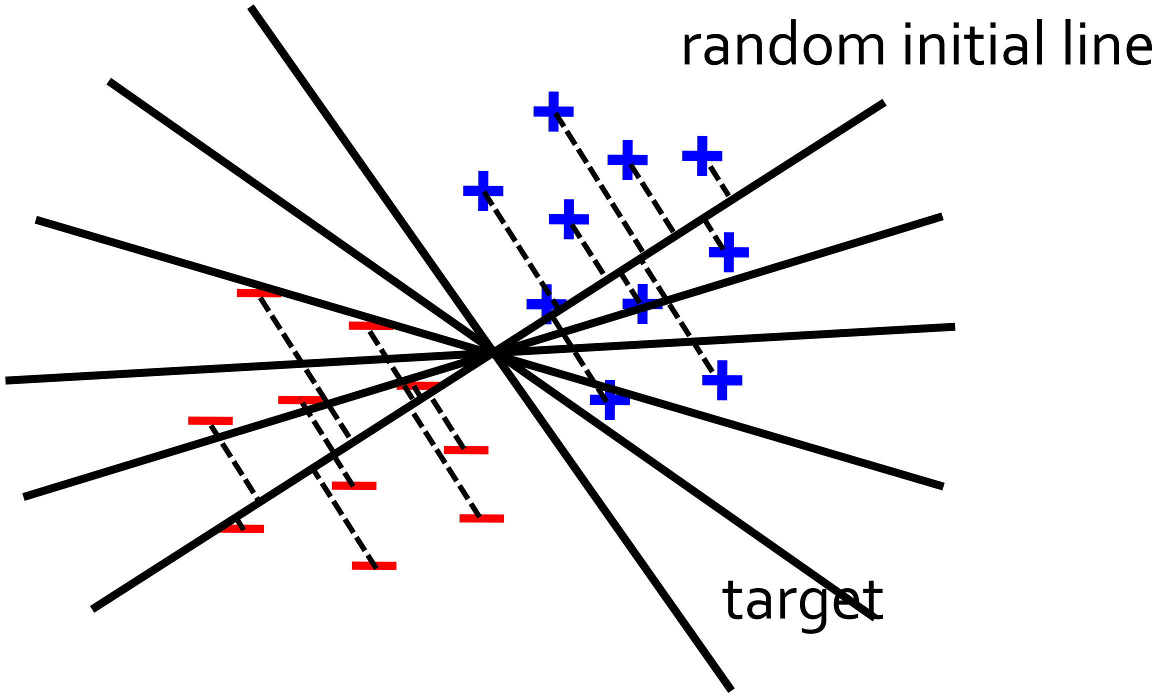
Problem #1

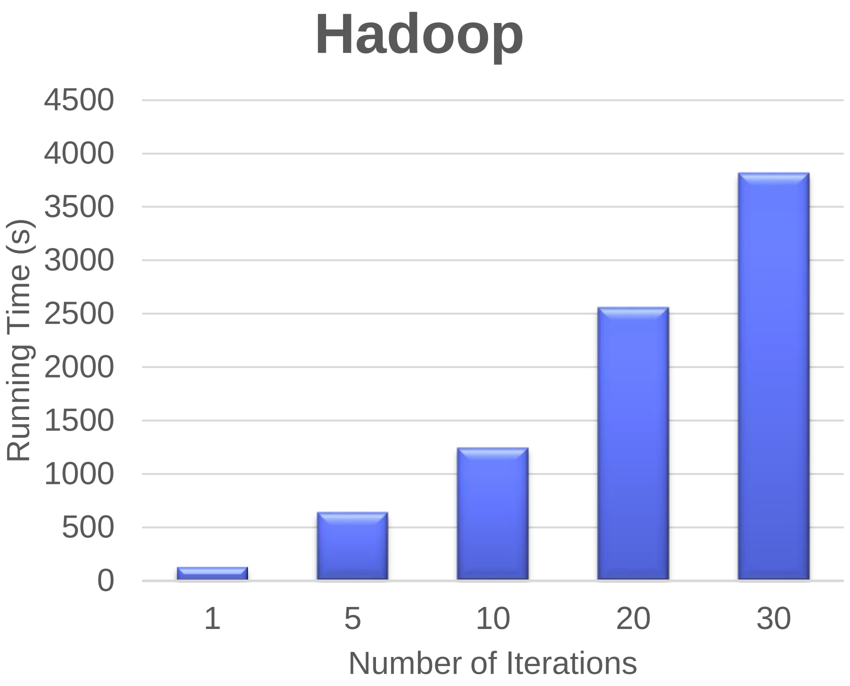
- MapReduce I/O sandbags runtime for advanced analytics.
  - Must persist results after each pass through data
  - Advanced analytics often requires multiple passes through data



Example

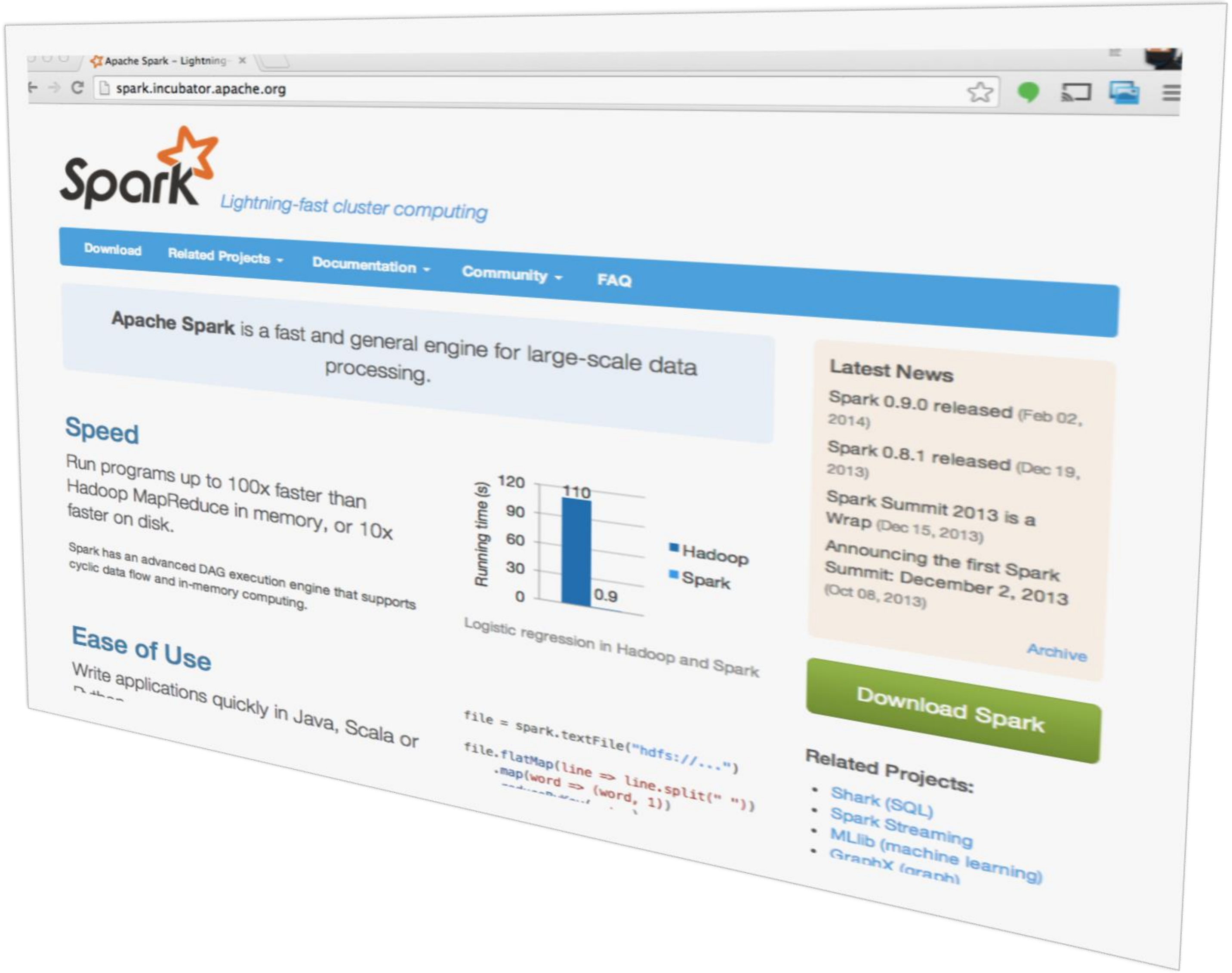
- Goal: find best line separating two sets of points





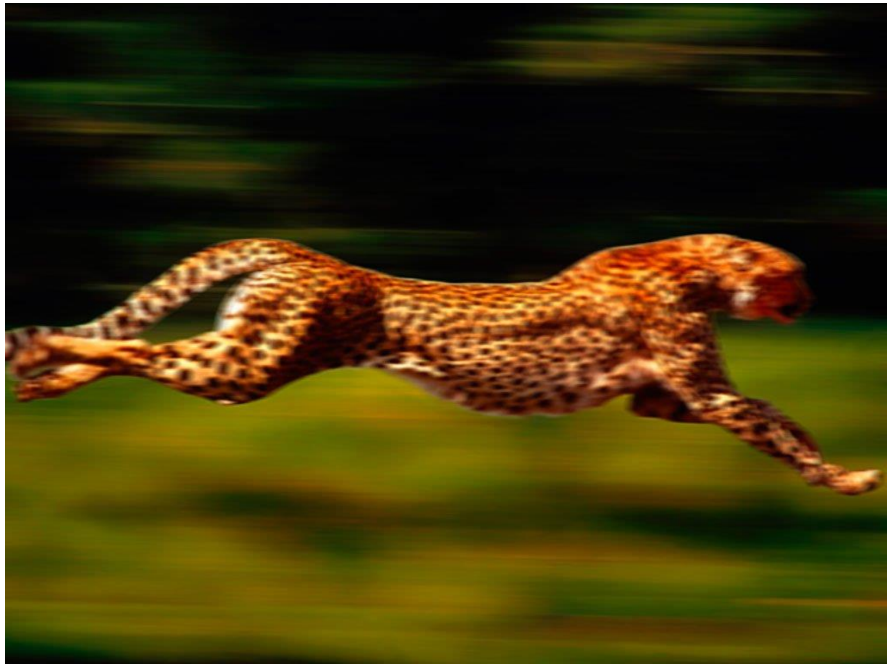
Apache Spark

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



Spark Performance

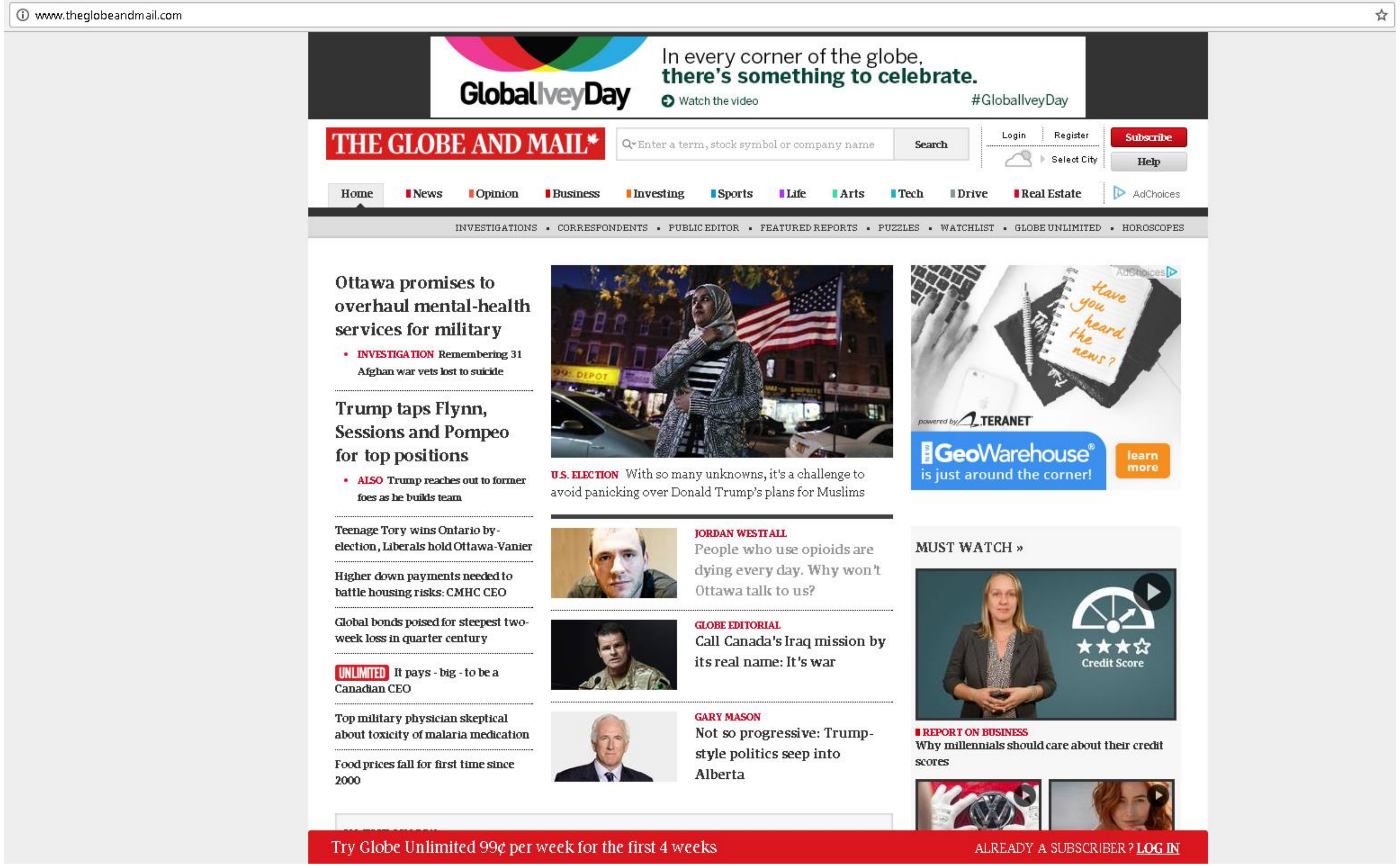
- **Machine Learning**
  - **100x** faster than MapReduce
- **Queries (Shark)**
  - **100x** faster than Hive
- **Streaming**
  - **2X** throughput of Storm
- **Graph (GraphX)**
  - **10X** faster than MapReduce



Outline

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

The Globe and Mail



Digital Media

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

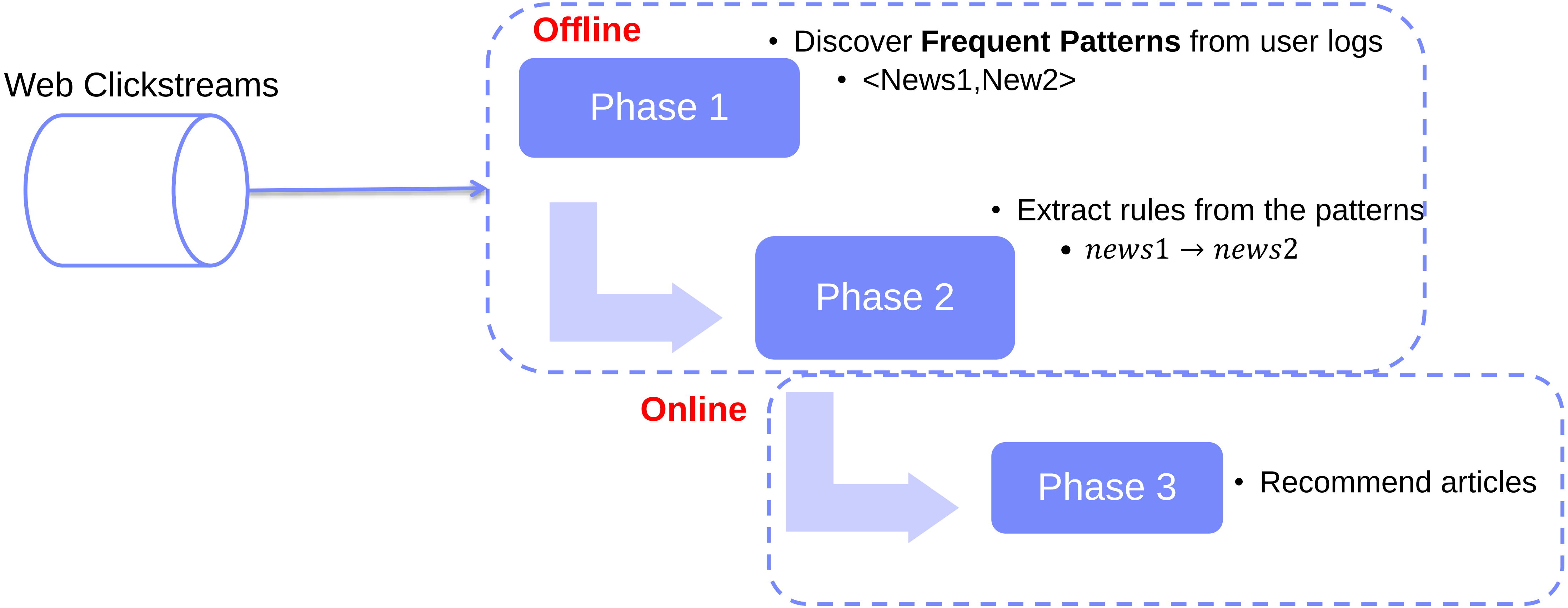
  - Year 2014 (Jan-Jul):
    - Number of Records: 264,735,412
    - Number of Visits: 51,748,518
    - Number of Users: 19,760,853

Preprocessing

- Data Preprocessing and Cleaning
  - Filtering out irrelevant hits
  - Extracting the user types
  - Extracting the event of interest
  - Computing the time spent
  - Roll-up form 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%

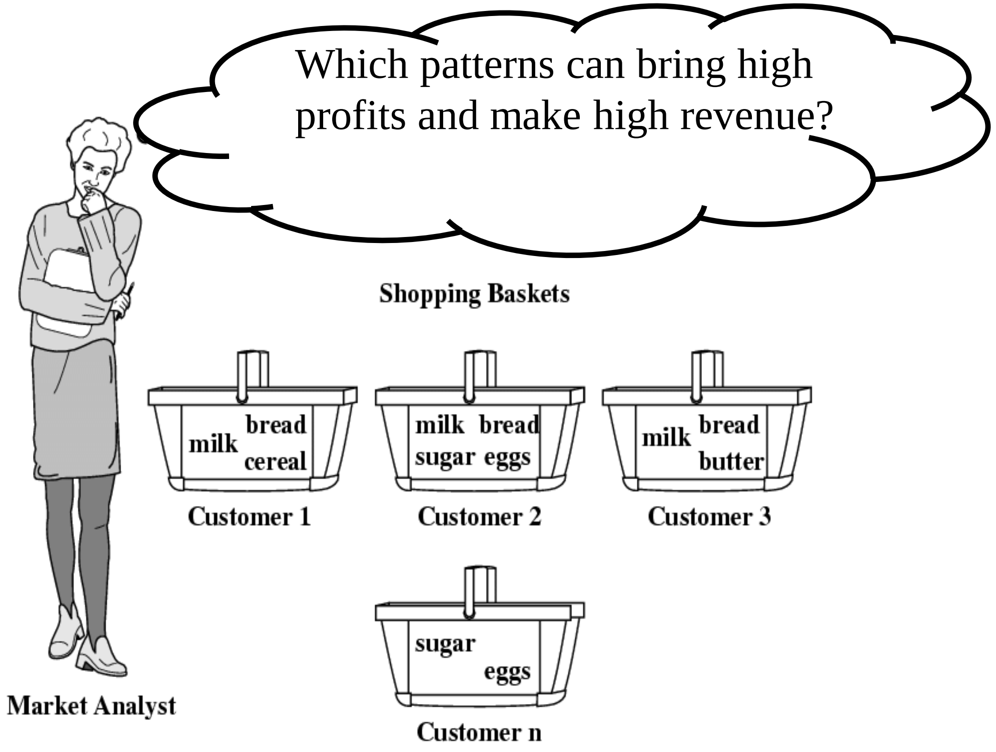
{Bread, Milk} is a frequent itemset

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(Milk,T1) = 3 × 6 = 18
- Utility of itemset in a sequence = sum of utilities of its items:
  - U({Bread, Milk}, C1)= 2 × 1 + 6 × 3 = 20
- Utility of sub-sequence in a sequence = sum of its itemsets' utilities
  - U(<{Milk}{egg}>,C1) = 3 × 6 + 3 × 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

$News_1$

$News_2$

$News_3$

$News_6$

$News_7$

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

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

Problem 2: News cold-start

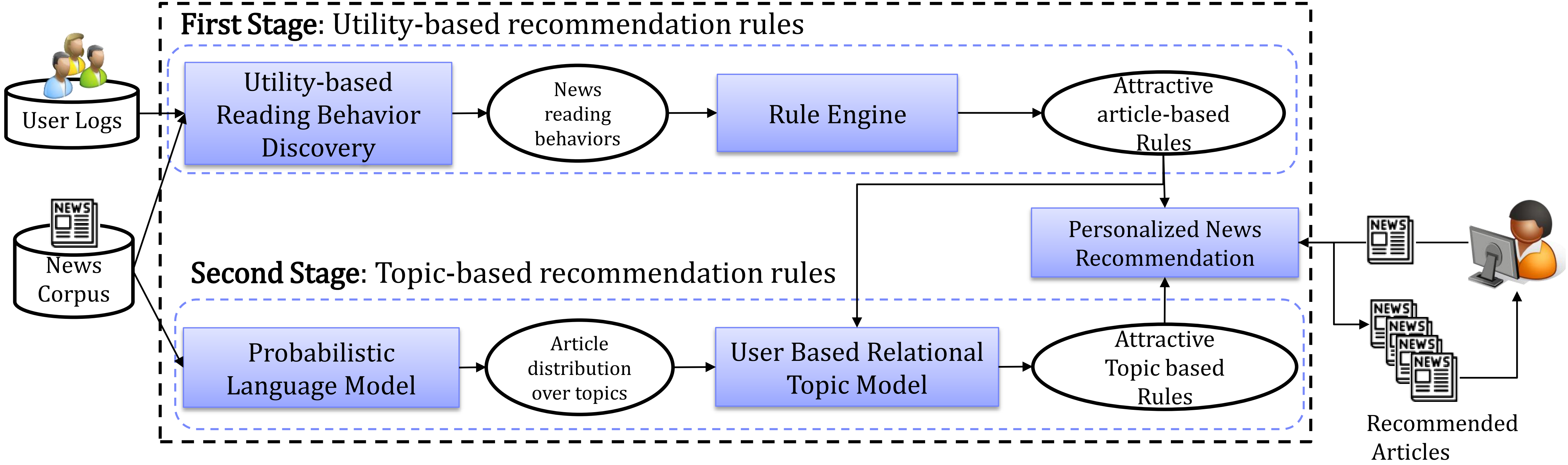
Recent news

Visited articles

Reading behavior

Newly-published articles

PENSYS: The Proposed Framework



FIRST STAGE: NEWS LEVEL

Morteza Zihayat

Stage 1: Utility-based Pattern Mining

Engagement Measures

Time Spent

Cursor movement

Internal Utility Function (F)

External knowledge

Editorial Board

Freshness

External Utility Function (G)

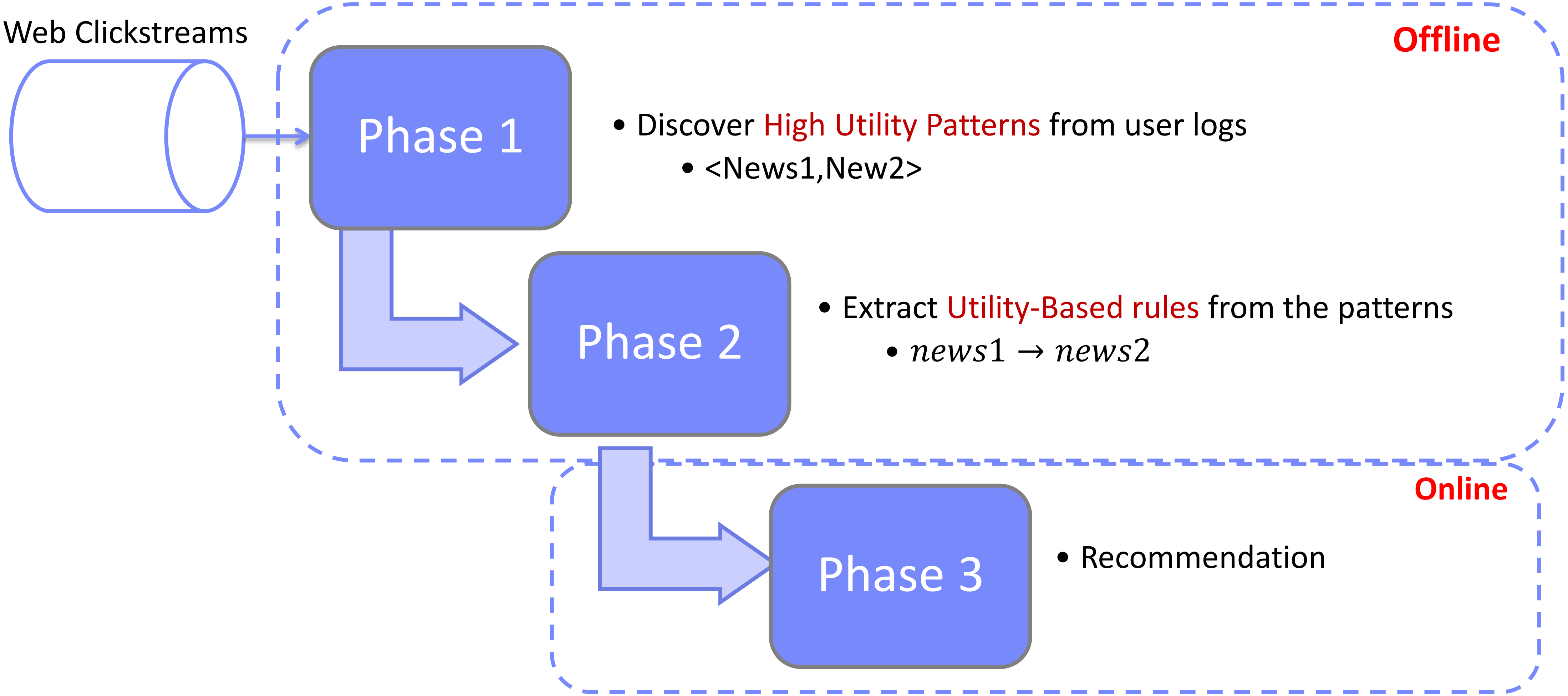
$$Utility = U(F, G)$$

Attractive news reading behavior

Patterns with Utility ≥ **Threshold**

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

Stage 1 Overview



- Discover High Utility Patterns from user logs
  - <News1,New2>
- Extract Utility-Based rules from the patterns
  - $news1 \rightarrow news2$
- Recommendation

Top-2 High Utility Patterns

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

Top-2 Frequent Patterns

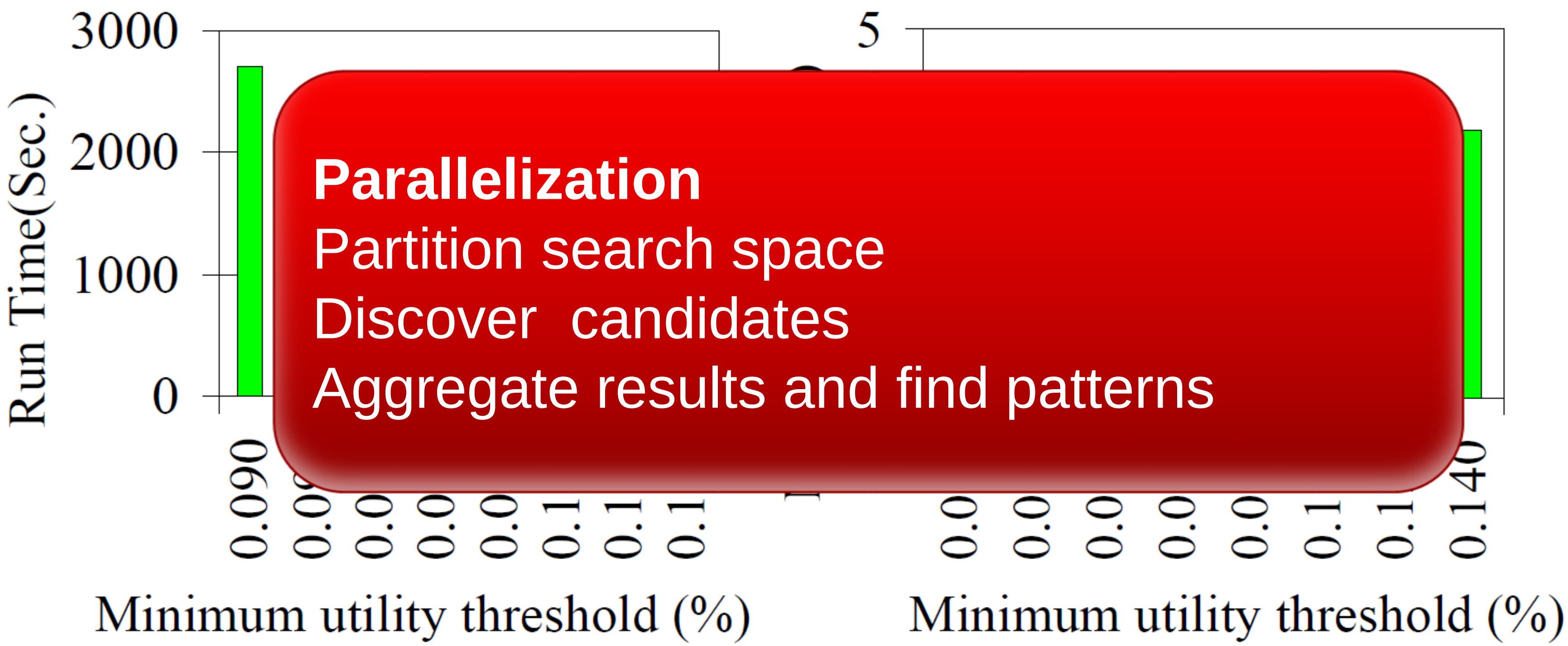
Num.	Set of news	Time (mins)	Support
1.	Target faces calls to withdraw from Canada , Mike Duffy facing 31 charges from Senate expenses scandal, RCMP says	144	254
2.	Florida police say , La Prairie, Quebec mayor dies from wasp stings Canadian professor was killed in targeted attack	105	286

Recommendation Rules

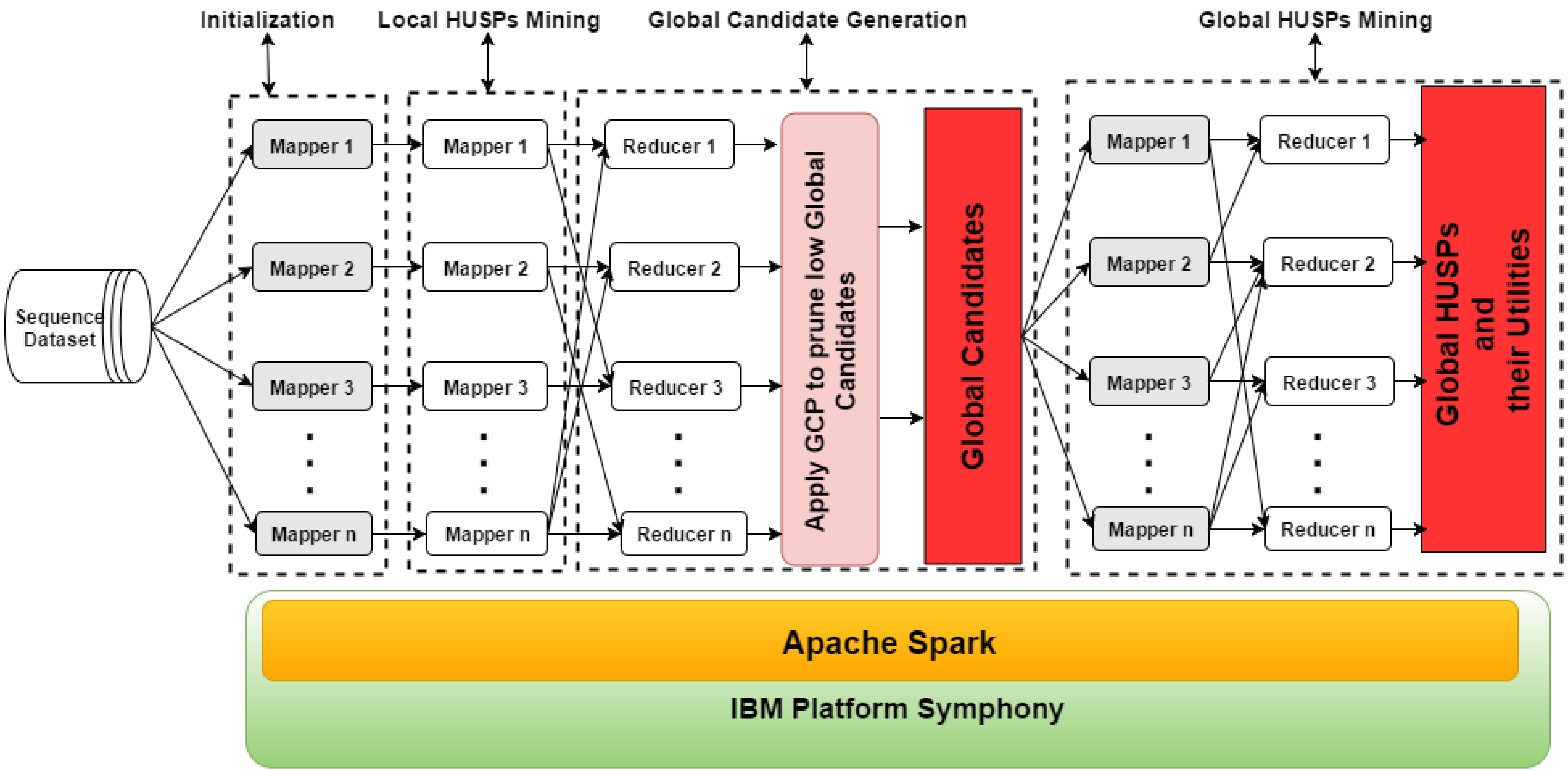
Top 3 of our Utility-Based Association Rules, sorted by uconf, time spent in descending order	Time Spent (minute)
[Researchers find dozens of genetic links to schizophrenia] ==> [Anti-Semitic graffiti found in Thornhill, days after Islamic centre defaced] [Video: Fearless 93-year-old tackles CN Tower 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]	239.98
[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]	239.93
[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]	239.8

Performance

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



Big Data Framework



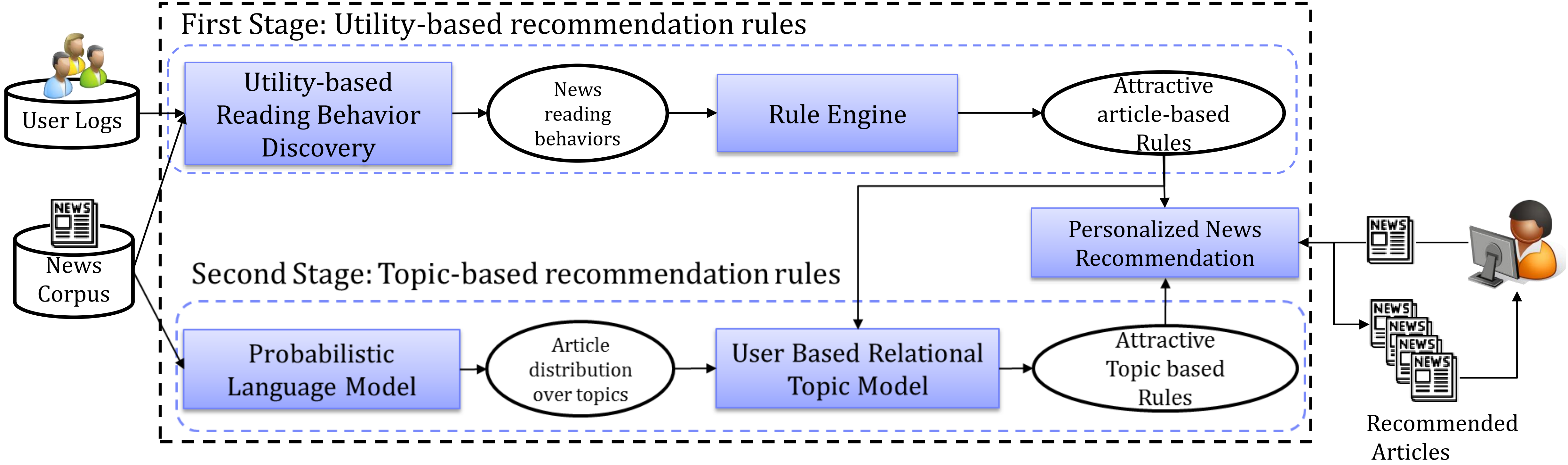
Results

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

SECOND STAGE: TOPIC LEVEL

Morteza Zihayat

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

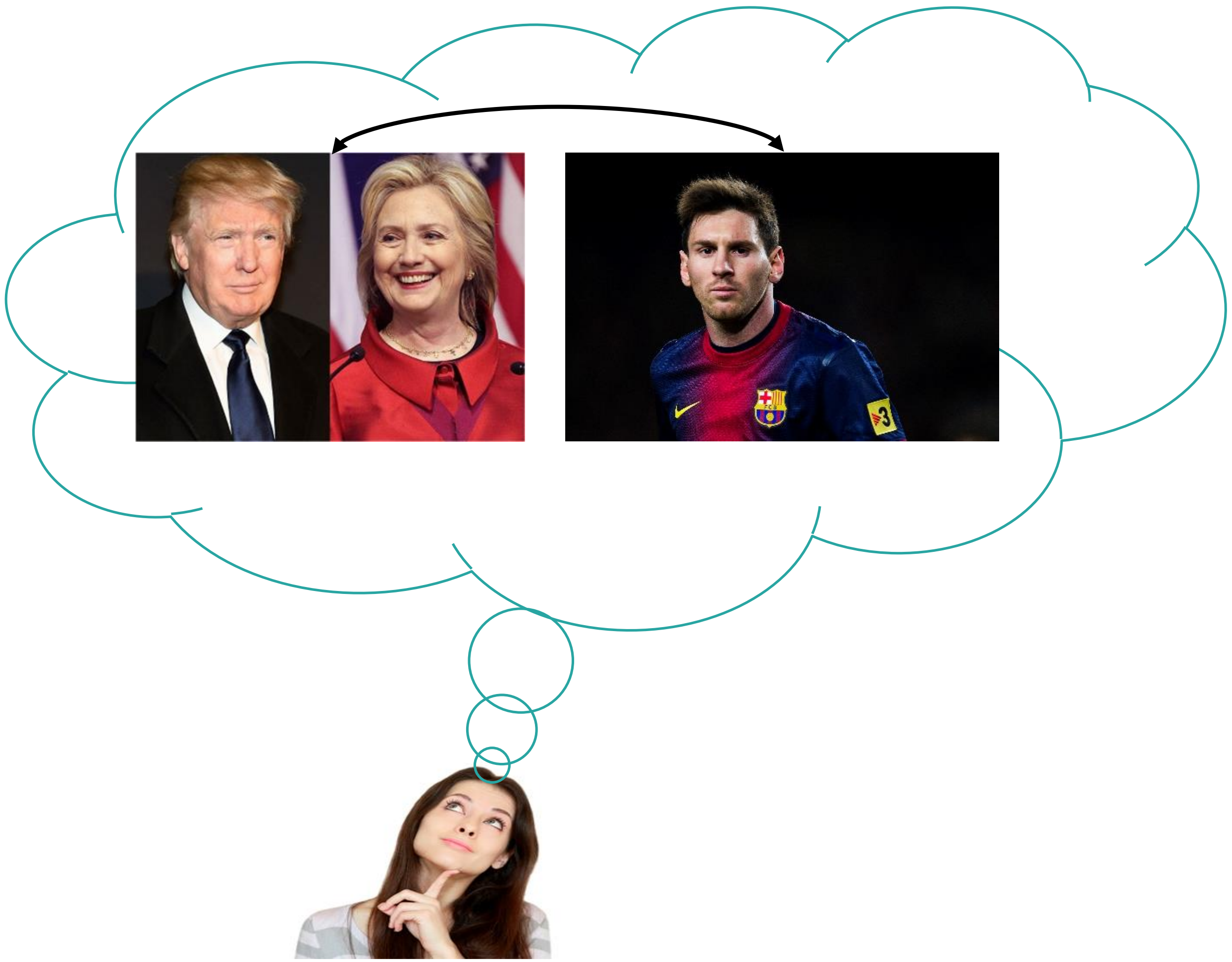
(a)

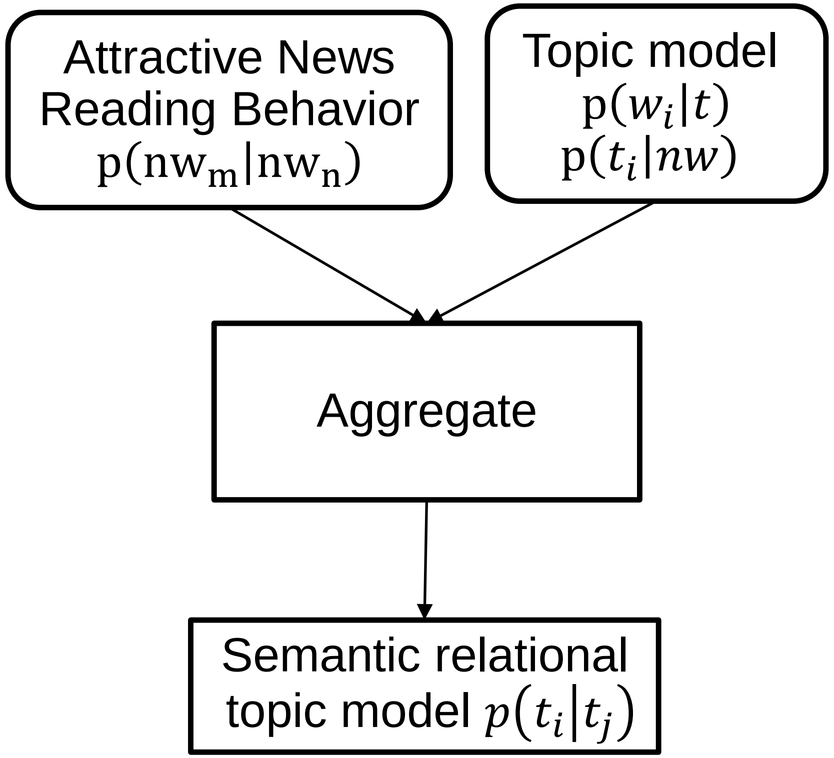
Topic 1: **Health**	Topic 2: **Debt**	Topic 3: **Travel**	Topic 4: **Sport**	Topic 5: **Politics**
$w, p(w\|Topic1)$	$w, p(w\|Topic2)$	$w, p(w\|Topic3)$	$w, p(w\|Topic4)$	$w, p(w\|Topic5)$
Drug, 0.5	Debt, 0.5	Hotel, 0.5	Soccer, 0.5	Obama, 0.5
Health, 0.4	Bond, 0.4	Travel, 0.4	Sport, 0.4	Politics, 0.4
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)

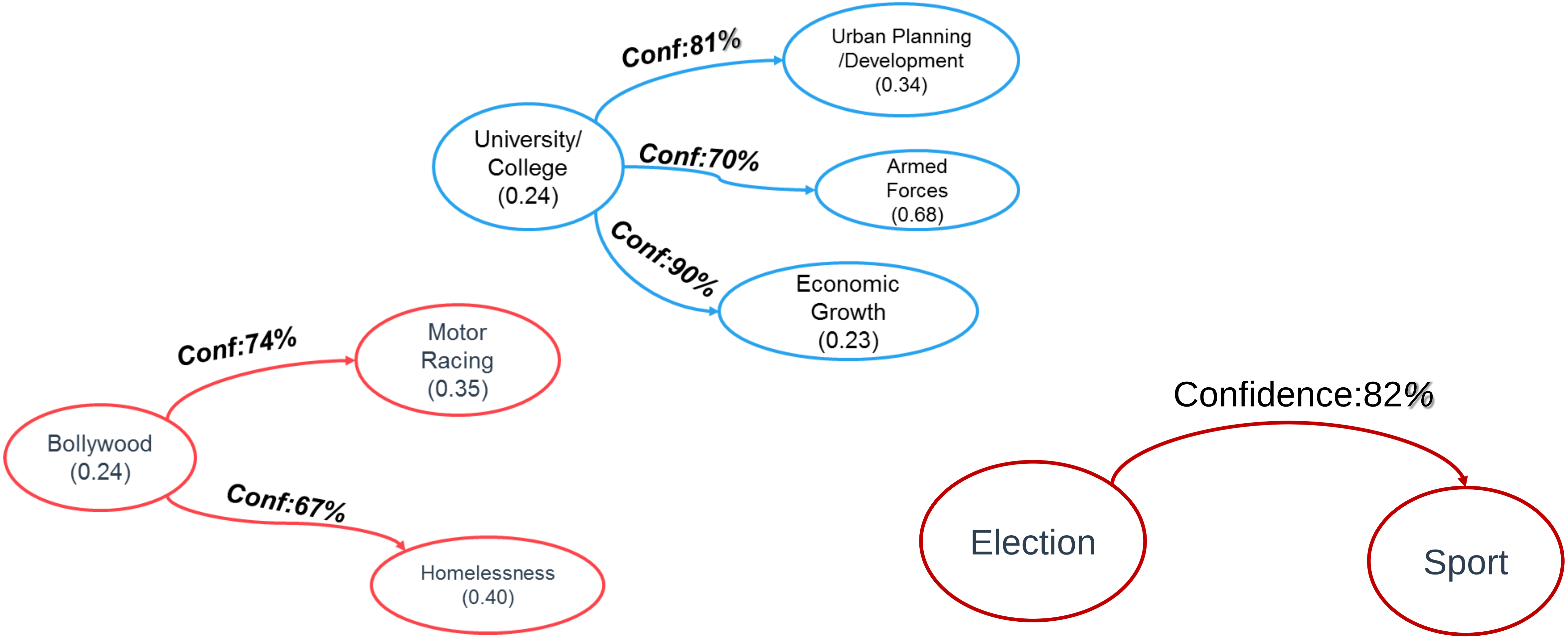
nw_1	nw_2	nw_3	nw_4	nw_5
$t, p(t\|nw_1)$	$t, p(t\|nw_2)$	$t, p(t\|nw_3)$	$t, p(t\|nw_4)$	$t, p(t\|nw_5)$
Topic 1, 0.6	Topic 2, 0.7	Topic 3, 0.6	Topic 4 0.8	Topic 5, 0.6
Topic 3, 0.2	Topic 3, 0.4	Topic 1, 0.4	Topic 5, 0.4	Topic 3, 0.4
....	....	....	....	....

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

Diagnosed by

Screening questionnaire

Problems:

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

Depression Acuity Detection

Mental state

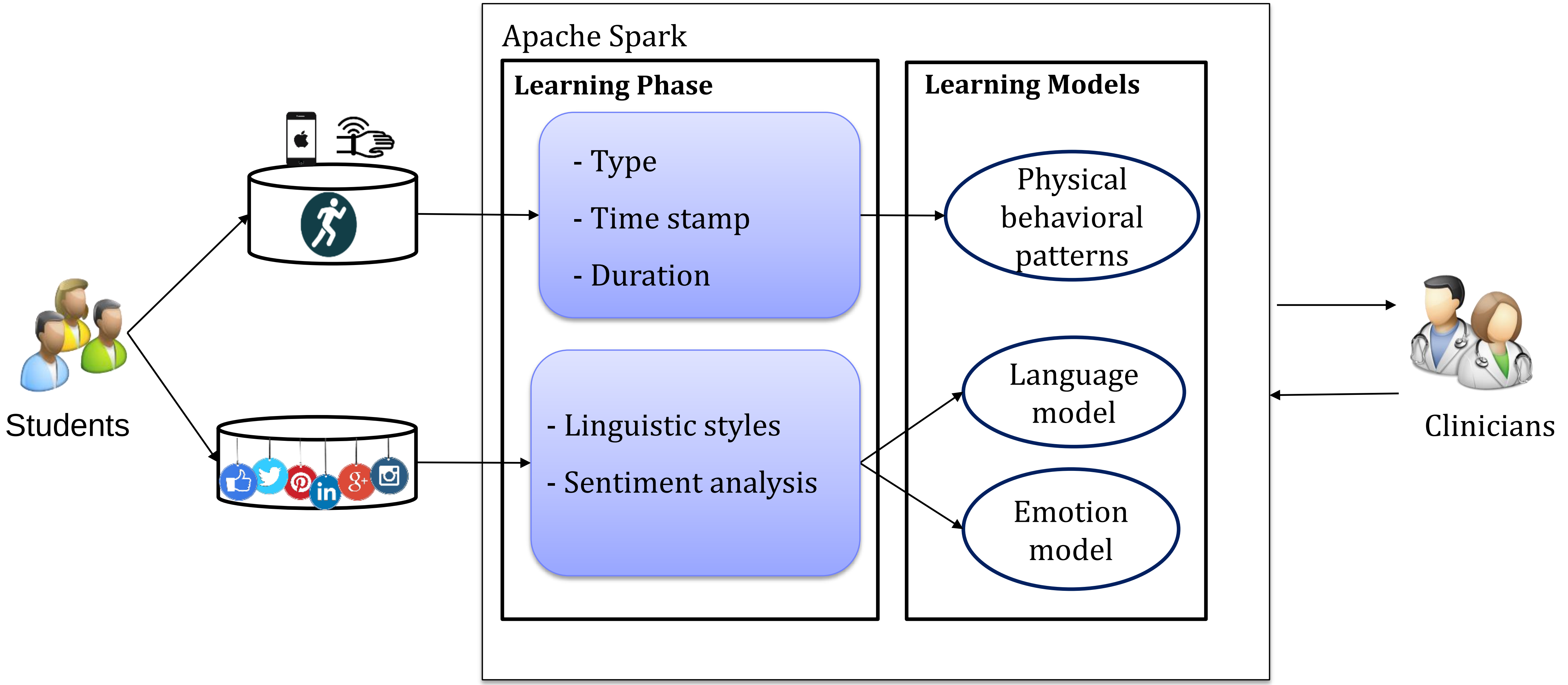
Physical wellness

Learning models

Sooner

More Accurate

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

UNIVERSITY OF TORONTO

THE GLOBE AND MAIL

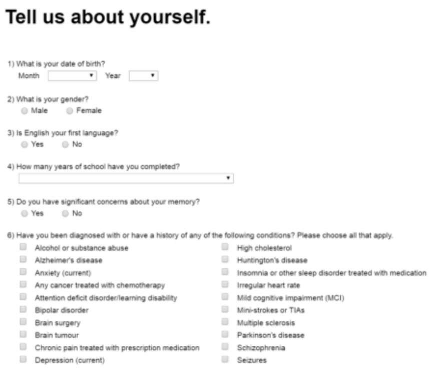
Brain Health

- 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

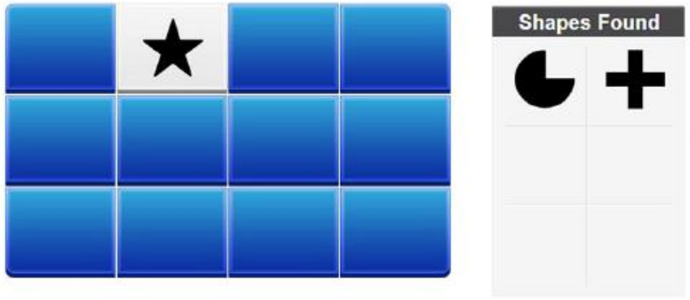
cogniciti Baycrest dapasoft

Assessments

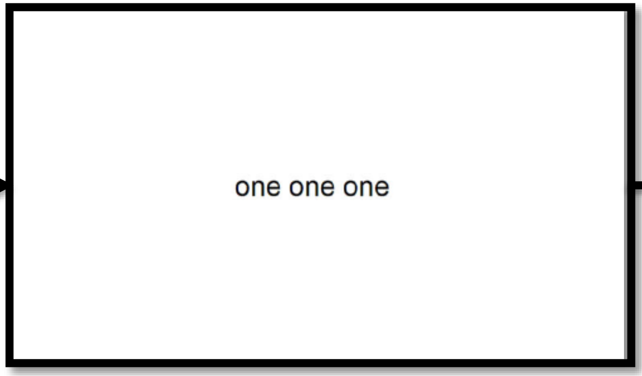
Demographic Test



Shape match test

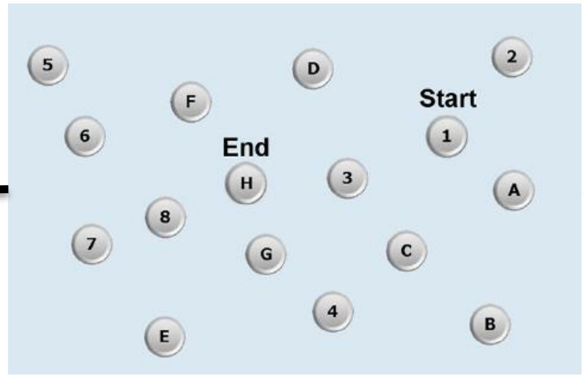


Stroop interference

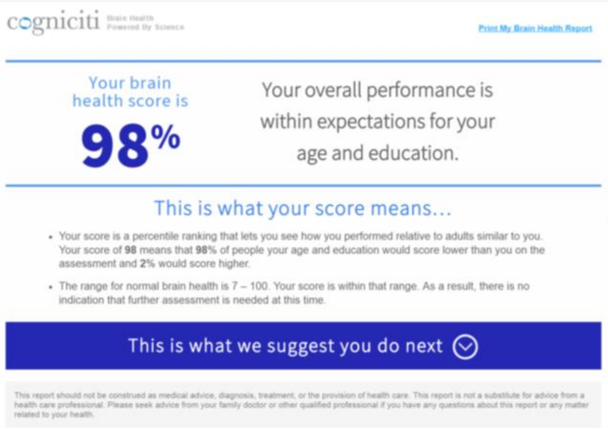


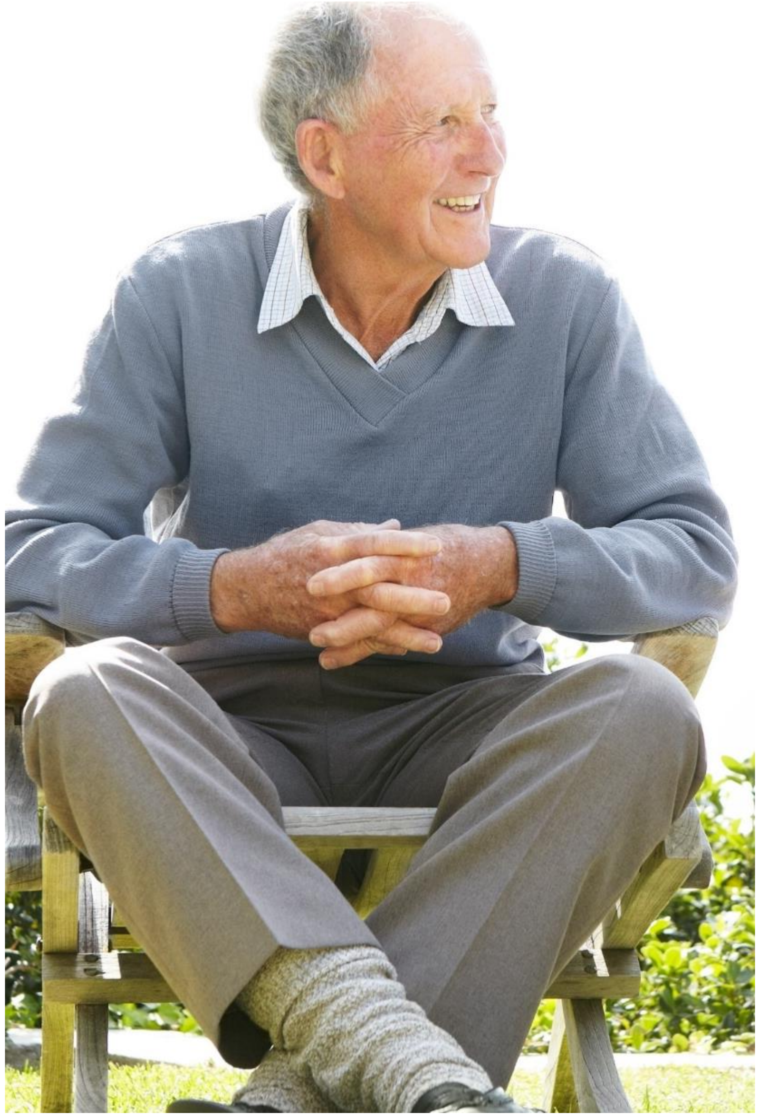
Face-name association

Letter-number alternation



Unified Score





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!

UNIVERSITY OF TORONTO

THE GLOBE AND MAIL

User Behavior Analysis in Big Data

mori.zihayatkermani@utoronto.ca

Morteza Zihayat