Reutlingen University

DBKDA Panel 2015, Rome, 27.05.2015

Can we Analyze all the Big Data we Collect?

Moderation:

Fritz Laux, Reutlingen University, Germany

Panelists:

Jerzy Grzymala-Busse, University of Kansas, USA

Yasuhiko Morimoto, Hiroshima University, Japan

Cihan Varol, Sam Houston State University, USA

Sergio De Agostino, Sapienza University of Rome, Italy

Christopher Ireland, The Open University, UK

Reutlingen University

Characterization of Big Data

- Big Data are mostly temporal data and its information is hidden in a sequence of events
  - Examples:
    - click data in a web shop
    - Sensor/machine readings → industry 4.0
    - Digital social network activities

- *Mostly structured data*
- *data rich, but information poor*

- *Usual Big Data definition*
  - Data that cannot be handled by a single system
    - Considering storage and processing power
  - Origin is unclear
    - John R. Mashey (sgi): Big Data ... Presentation on the next technology wave in 1998

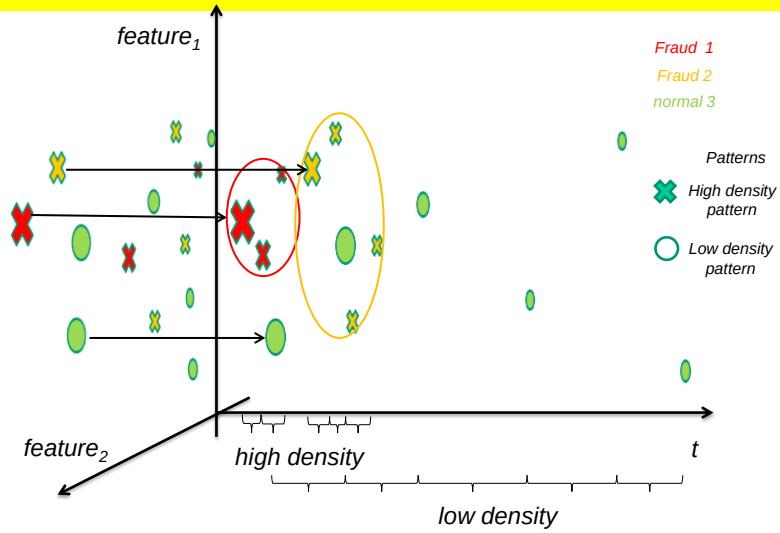
Reutlingen University

Big data analysis

- *If we consider temporal big data only*
- *Information (pattern) is spread over many events within one session or transaction → trend*
- *We want to predict future events or anticipate measures*
  - Timely analysis is needed
  - Most algorithms for (temporal) analysis have a computational complexity of $O(n^2)$
    - sequence pattern search worst case is $O(n^2)$
  - Patterns may change over time
    - Complexity of pattern generation is cruicial, too

Reutlingen University

Example of two temporal patterns (transaction events)



- *No single event discloses a temporal pattern*
  - *Data preparation (restructuring) is possible, but not in realtime and depends on the analysis*

Reutlingen University

Infrequent patterns

- *Strong (frequent) patterns are often not interesting*
- *Looking for infrequent, but important patterns*
  - E.g fraud, hazards, terrorism
  - A timely discovery is even more important
- *How do we know what to look for?*
  - Need for a particular structure and algorithm
- *If we know, how long does it take?*
  - Need efficient algorithms (<= O(n))

5 /6

Reutlingen University

Hypothesis

- *We can not analyze all big data in a timely manner*
- *Some arguments*
  - Big Data are mostly temporal data and its information is hidden in a sequence of events
    - ⇨ At the moment we have no algorithms to find temporal patterns with <= O(n)
  - Infrequent patterns are most interesting
    - ⇨ If data is highly skewed classification tends to produce false positives.
- *Idea*
  - Use parallel processing for data management
  - Research for efficient O(n) parallel algorithms

6 /6

Big Data: Can we analyse it all and so what?

Chris Ireland

The Open University, UK

What's the problem..?

- Volume
- Frequency
- Complexity with respect to interrelationships
- Data type
- Data Mix
- Technology
- Strategies for analysis
- ...Decision Support (Informed, Satisficing [Simon])

Big... or just more?

- How do we define Big Data?
  - Are the definition(s) any use?
  - How do we know we have a Big Data problem?
- Concepts and Technology
  - Data set (complex, large number of data points)
  - Database
    - The number of rows, columns, tables, keys
  - Life of a Queen Bee - 126 billion data points (id, x,y,z,time - ms)

Surely its all relative?

- In 1983 - Kb - Exam data for an entire school year
- In 1990 - Mb - CASE data for an entire project
- In 2000 - Gb - 114 mins of CD quality audio
- In 2010 - Tb - Monsters v Aliens Movie Development
- In 2020 - Pb? New ways
- In 2030 - Eb? to collect
- In 2040 - Zb? and to
- In 2050 - Yb? classify data.

All about the data..?

- Without data you're just another person with an opinion (Deming)
  - But how much data before it ceases to be an opinion?
- Can we have too much data (how do we access it)?
- How do we know we have the correct data?
- How do we know we have the wrong data?
- How do we analyse the data?
  - What is the question you are trying to answer?
  - Or the answer you are trying to question...

It starts with an idea...

- Retailing (Know your customer)
- Medical (e.g. Genome Analysis, Virus Spread)
- Financial Transactions (market violations?)
- Social Trends (e.g. Tweets)
- Market Analysis (pricing data)
- Understanding the Universe (e.g. CERN)
- ...we are in charge!

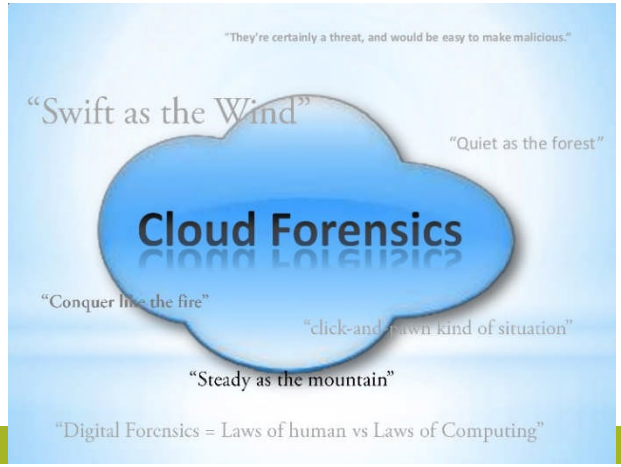
BIG DATA AND DIGITAL FORENSICS

Challenges - Solutions

Cihan Varol
Computer Science Department
Sam Houston State University
Huntsville, Texas, USA
cxv007@shsu.edu

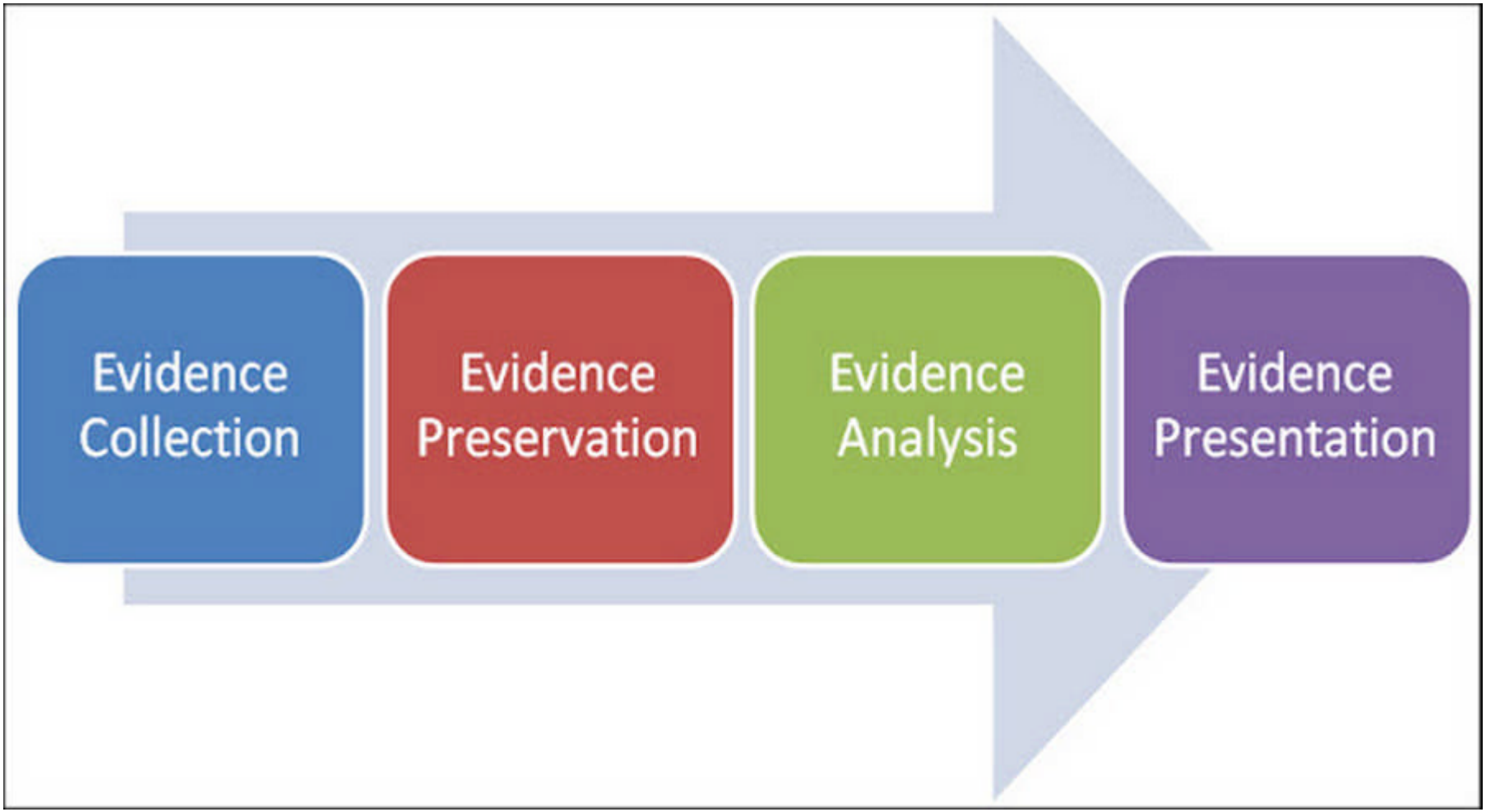
DBKDA ‘15
05/27/2015

Digital Forensics Branches



. . .

Digital Forensics Process



Challenges Faced by Digital Forensics Experts

- Storage capacity that needs to be analyzed
- Data backup
- Protecting data from threats

Rethinking Digital Forensics

- A new workflow?
- Evidence Collection and Acquisition
  - Complete collection could be impossible
  - Prioritization (triage) right at the start
  - From copy-all, analyze later
  - To acquire what is necessary, evaluate "on the scene"

Rethinking Digital Forensics

- Rethinking the workflow
- Preservation
  - Big data storage
- Evidence Analysis
  - Intelligent Digital Forensics (NIJ guideline, social-network analysis, artificial intelligence)
- Evidence Presentation
  - Log details of examinations, precise algorithms applied
  - Explain validity of machine learning techniques

Tools (to cope)

- Hadoop
- Sleuth Kit Hadoop Framework
- …

Conclusion

- Forensic analysts need to
  - Adapt methods to the new scenery
  - Relax some requirements
  - Add some new tools to the arsenal

Mining Big Data

Jerzy W. Grzymala-Busse'"

' University of Kansas, Lawrence, KS 66045, USA

" Department of Expert Systems and Artificial Intelligence,

University of Information Technology and Management, 35-225 Rzeszow, Poland

DIKW Paradigm

Wisdom

Knowledge

Information

Data

Data

Big Data

- analytics-ready structured data form only a tiny subset of big data
- unstructured data - audio, image, video and unstructured text

Mining Big Data

- Business as usual,
- Sampling, dimensionality reduction,
- Hadoop etc.

Big Data Systems

- Hadoop - batch oriented data processing system (Google),
  - Map Reduce
- High Performance Computing Clusters (LexisNexis Risk Solutions),
  - Thor cluster (destination for loaded data),
  - Roxie (data processing),
  - Enterprise Control Language,
- Storm - real-time data processing system.

Conclusion

Everything must change

so that everything can stay the same

From The Leopard by Giuseppe Tomasi di Lampedusa

Applying Compression and Decompression to Big Data in Parallel

Sergio De Agostino
Sapienza University di Rome

Big Data Compression

Compression is considered advantageous and actionable by the Big Data Community for the following reason:

> it is possible to compress data so that many predicates can be evaluated without having to decompress.

I would like to discuss another potentially good reason for big data compression.

Locality Principle

The locality principle makes big data compression and, more importantly, decompression highly parallelizable.

In the future, using the computational resources to speed up compression and decompression could be more practical than employing sophisticated techniques to query compressed data (as, for example, compressed pattern matching).

Zipf's Law

The frequency of an item in a realistic data set is, more or less, inversely proportional to its rank in the frequency table.

Occurences of the most frequent item= c

Occurences of *i*-th most frequent one= c / i

No need of a global view of data to make compression (local reference).

Zip Compressors

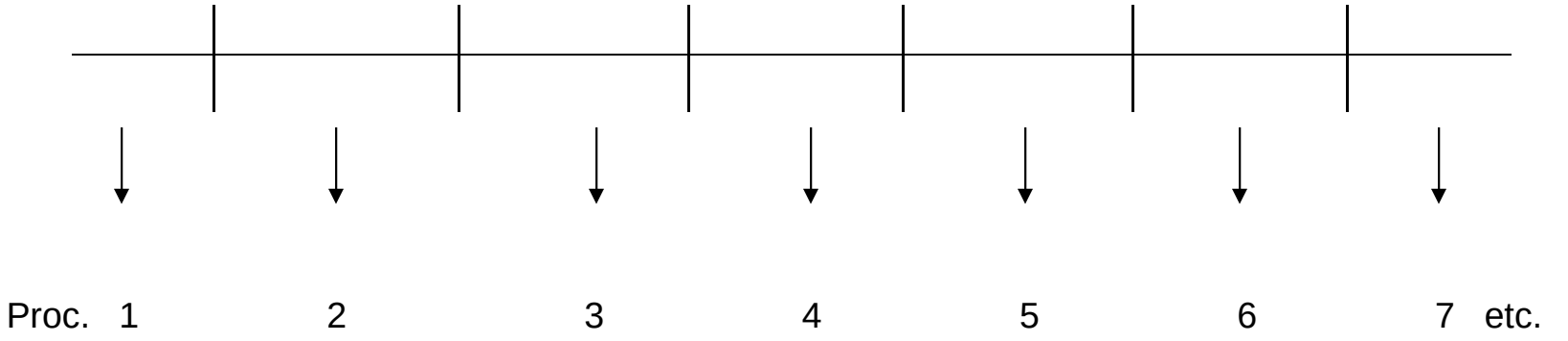
Most Zip compressors store 32 or 64 KB of history to average about a 50 percent gain (local reference).

Bzip2 stores between 100 KB and 900 KB of history and gains just 2 percent with respect to the other compressors.

How long must be a data block to be compressed independently by a node in a network with no relevant loss?

Scalability and Robustness

A block length of a few hundreds kilobytes guarantees scalability and robustness (one order of magnitude more than the history to achieve compression).



Conclusion

Decompressing a few hundreds kilobytes is extremely fast.

Such parallel approach is scalable and robust for any compression method.

On the other hand, evaluating predicates on compressed data needs ad-hoc techniques which might be less effective (lack of robustness).

Thank you

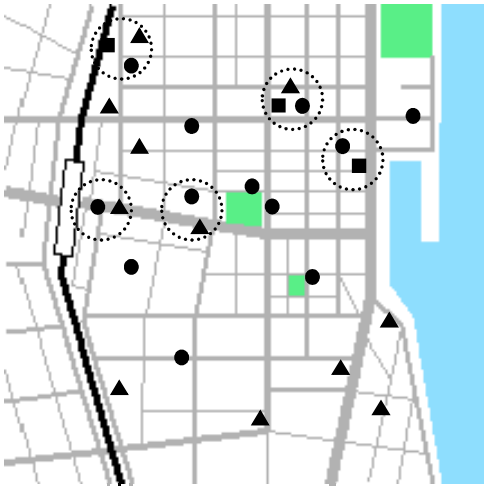
Yasuhiko Morimoto, Hiroshima University, Japan

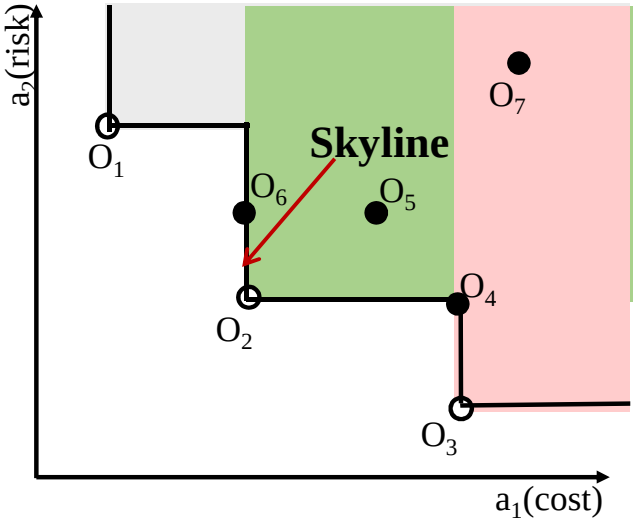
Data Mining for numerical data

Decision Tree / Regression Tree for Huge Correlated Numerical data

Spatio-Temporal Data Mining

Privacy-Aware Information Retrieval (Skyline Query)





Y. Morimoto, DBKDA 2015 Panel

We (database researchers) have been working on “Big Data” for 40 years.

“Big Data” problem is not a new one but a problem that
we have always been working on.

> “Big Data” was data that could not be stored in “floppy disk” in 1980s.

> Ask youself.
> Could we analyze “Big Data” could not be stored in “floppy disk” in 1980s?

Y. Morimoto, DBKDA 2015 Panel

Can we analyze all the “Big Data” we collect?

In all “3V”, (Volume, Velocity, Variety)

YES, but … we have to solve

Hardware issues (storage etc.)

Algorithmic issues (MapReduce computing etc.)

Ethical issues (privacy etc.)

Y. Morimoto, DBKDA 2015 Panel