Extending search-based software testing techniques to big data applications

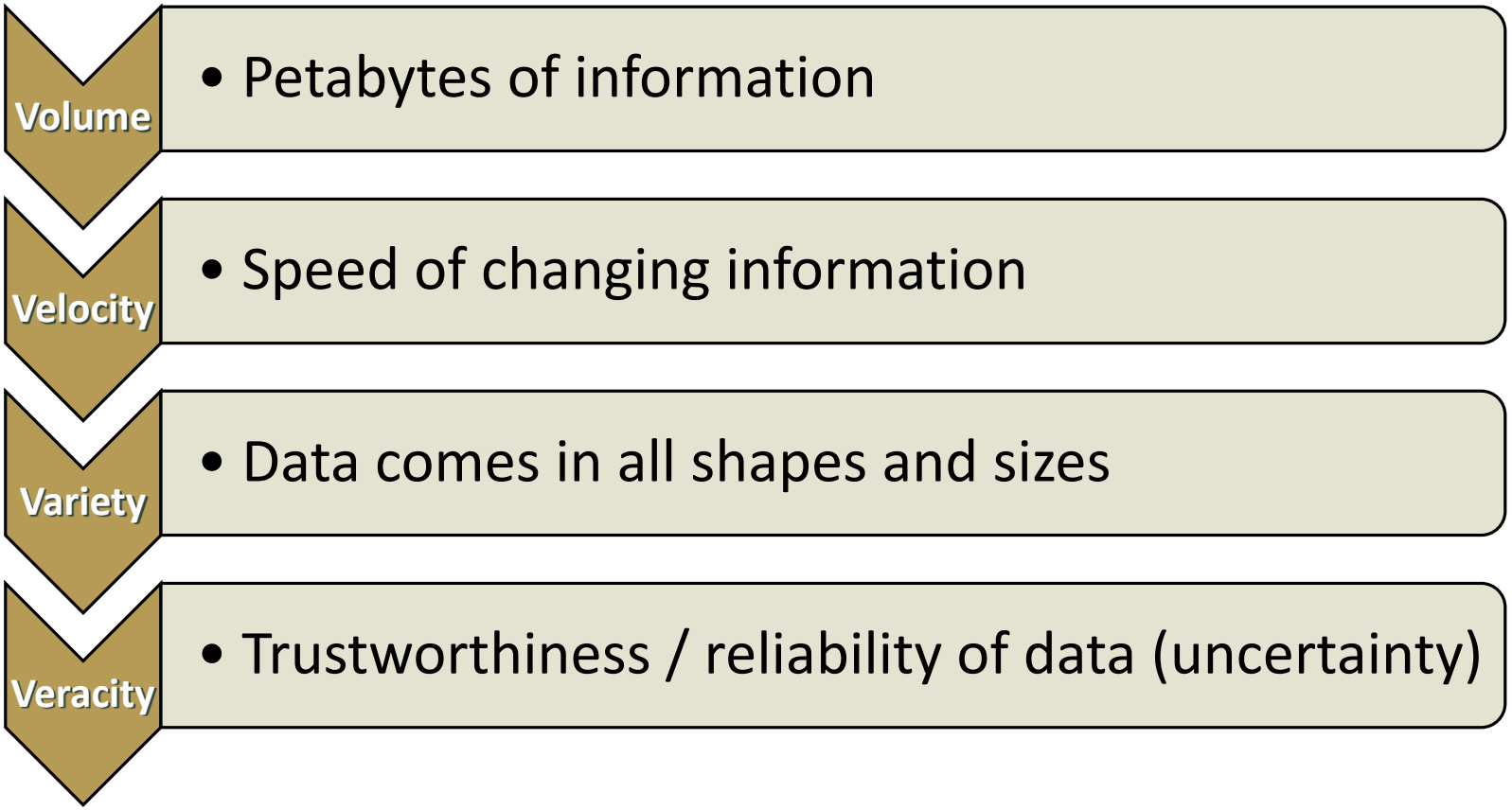
ERIK M. FREDERICKS
REIHANEH H. HARIRI
MAY 17TH, 2016

OAKLAND UNIVERSITY
SCHOOL OF ENGINEERING AND COMPUTER SCIENCE

Big Data?

http://buzzwordpgh.org/wp-content/uploads/2014/08/logo5-300x280.png
http://eduvantis.com/wp-content/uploads/2015/04/big-data.jpg

Big Data?


Volume
• Petabytes of information
Velocity
• Speed of changing information
Variety
• Data comes in all shapes and sizes
Veracity
• Trustworthiness / reliability of data (uncertainty)

Techniques for Managing Big Data

Hadoop / MapReduce

Apache Spark

NOSQL, BigTable, etc.

Position

SBST techniques can enhance testing techniques for big data applications.

- Focus on automated test suite generation
- Reduce enormous search space generated by big data

- Isn't reducing the search space the entire point of SBST?
  - Of course!
  - Big data is simply the next obstacle to be overcome using SBST!
    - Extend our techniques to this new paradigm

http://www.spring.org.uk/images/cynical.jpg

Issues and Possible Solutions

Nearly all facets of software testing can be impacted by big data!

Issues that concern the SBST community…

- **Test suite generation**
- Combinatorial testing
- Mutation testing
- etc.

Test Suite Generation

Test suite

- Typically comprise a set of test cases
- Generally concerned with validating a particular **operating context**
  - Combination of parameters that specify system and environmental configuration
- Well-studied problem in SBST community [Fraser.2011]

However…

- Big data adds a new wrinkle!
- How can we possibly generate enough test suites to adequately cover the 4 V's?

Impact of Big Data

Test suites provide measure of coverage for **known operating contexts**

Consider a nation-wide **medical records network (MRN)**

- Patient data recorded in Detroit, MI
- Immediately available in Austin, TX
- Patients, doctors, nurses, etc. all interface using heterogeneous devices
  - Network supported by heterogeneous devices
- Data such as patient records, medical imaging, video, etc. ALL available

Deriving test suites to cover entire application becomes quickly non-trivial!

- More reasonable to focus on subsets of application
- E.g., Android/iOS/WinPhone application that interfaces with network

Applications of SBST

SBST techniques now needed more than ever!

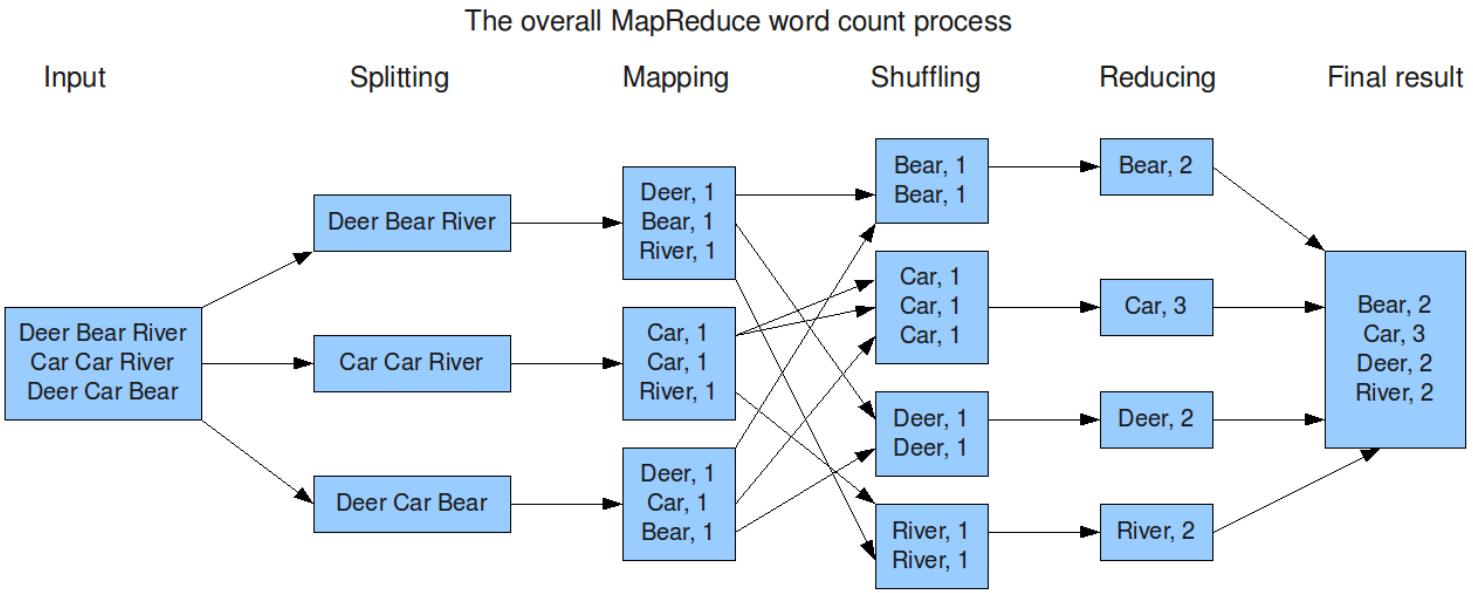
Explore a **massive** solution space

Augment existing big data approaches to support SBST

Applications of SBST

Hadoop/MapReduce, for example

- Comprises, at its core, Map and Reduce functions



http://www.cs.uml.edu/~jlu1/doc/source/report/img/MapReduceExample.png

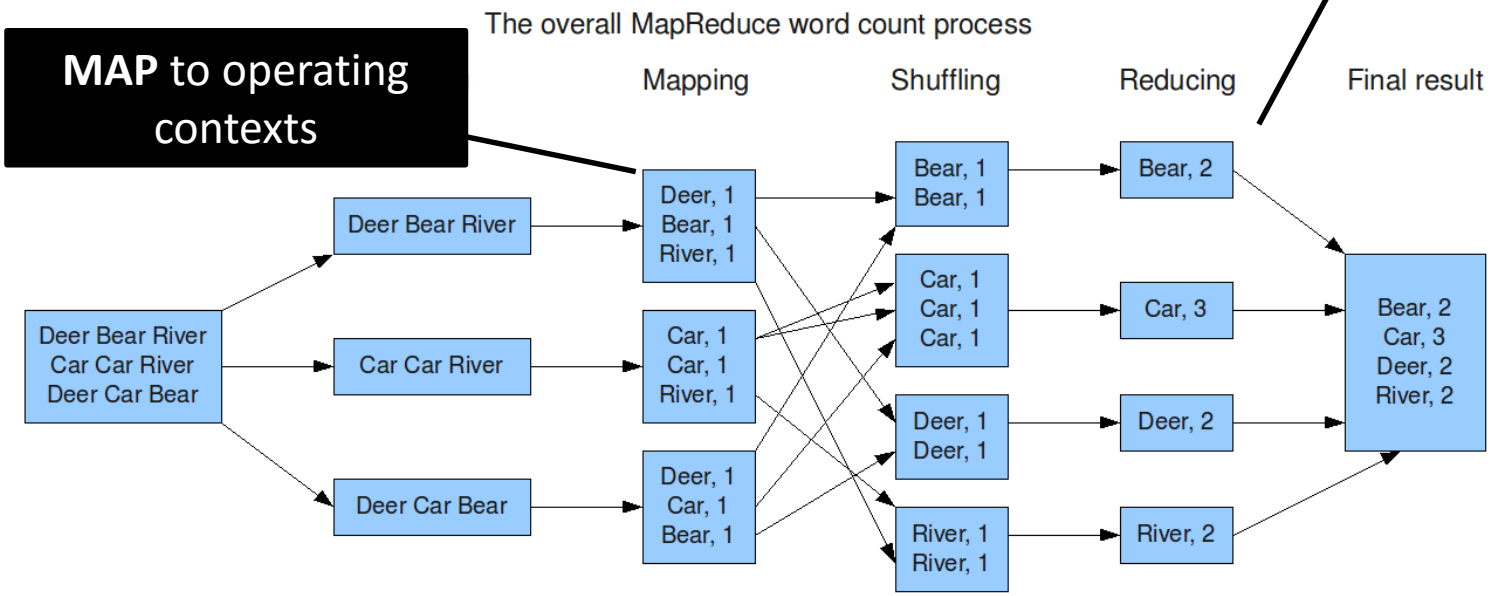
Applications of SBST

Hadoop/MapReduce, for example

- Comprises, at its core, Map and Reduce functions

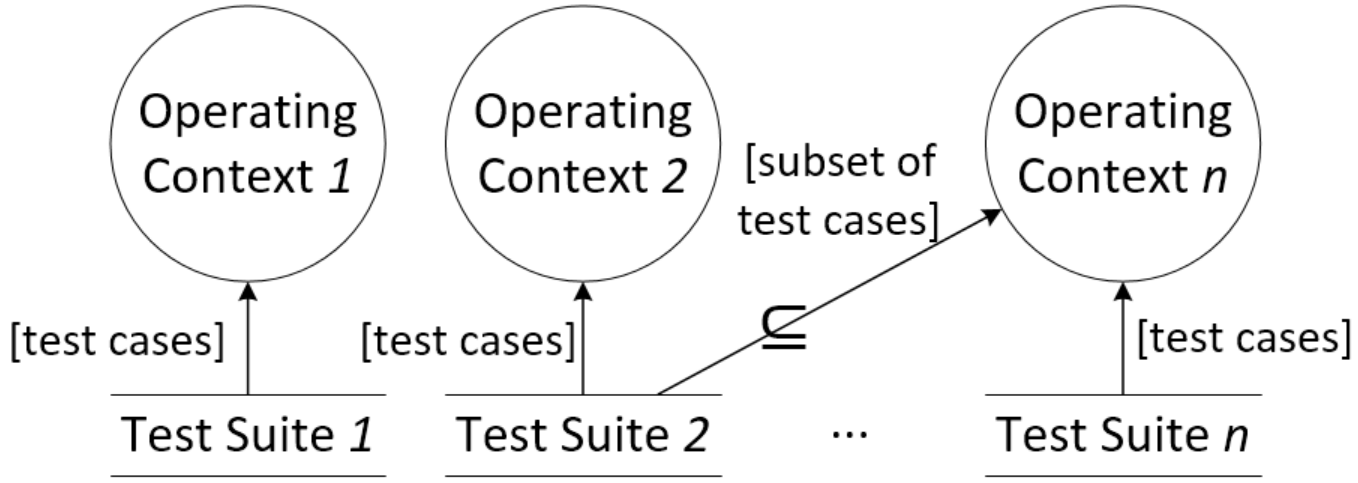
REDUCE to minimal coverage criteria

MAP to operating contexts



http://www.cs.uml.edu/~jlu1/doc/source/report/img/MapReduceExample.png

Applications of SBST



1 → BLOB data

2 → Network reliability

n → Video playback

Applications of SBST

Parallelized genetic algorithm (GA) for generating test suites with Hadoop [Geronimo.2012]

- Each GA generation is a MapReduce job
  - Fitness evaluation performed by *Mappers*
  - *Reducer* collects results and performs evolutionary operations
  - Extend paradigm to manage big data – mappers concerned with operating contexts

Automated test generation using relational databases [McMinn.2015]

- Testing integrity constraints on relational database schema
  - Constraint and column coverage
- Augmented random search and alternating variable method
  - Generate test suites
- Highly-relevant to big data, as big data is typically schema-less!

Acknowledgements

The authors would like to thank **Oakland University** for supporting this work.

Discussion

Testing applications that **interface** with big data

Dealing with **unstructured** data

Extending search-based techniques to the big data (testing) domain

http://www.macrobusiness.com.au/wp-content/uploads/2011/12/cartoon-round-table-discussion.jpg

References

[Fraser.2011] G. Fraser and A. Arcuri. Evosuite: automatic test suite generation for object-oriented software. In Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering, ESEC/FSE ’11, pages 416–419, Szeged, Hungary, 2011. ACM.

[Geronimo.2012] L. Di Geronimo, F. Ferrucci, A. Murolo, and F. Sarro. A parallel genetic algorithm based on hadoop mapreduce for the automatic generation of junit test suites. In Proceedings of the 2012 IEEE Fifth International Conference on Software Testing, Verification and Validation, ICST ’12, pages 785–793, 2012

[McMinn.2015] P. McMinn, C. J. Wright, and G. M. Kapfhammer. The effectiveness of test coverage criteria for relational database schema integrity constraints. ACM Transactions on Software Engineering and Methodology, 25(1):8:1–8:49, 2015.