Software Engineering for Big Data

CS846
Paulo Alencar
David R. Cheriton School of Computer Science
University of Waterloo
Big data technologies describe a new generation of technologies that aim at economically extracting **value** from very large **volumes** of a **variety** of data (structured and unstructured) by enabling high-**velocity** capture, discovery, and analysis.
Big Data Sources

The ability to collect data is increasing exponentially.

Some sources:
- Sensors
- Transactions
- Mobile phones
- Images
- Maps
- Web
- Social networks

Sensor technology and networks
(collecting all sorts of data, e.g., Internet of Things)

Scientific instruments
(measuring all types of data)

Social media and networks
(all of us are generating data)

Mobile devices
(tracking all objects all the time)

[Wang Huan, 2013]
Big Data Trend

*Source: Contents of above graphic created in partnership with Teradata, Inc.*
Big Data and Open Data

Much big data comes from open data sources.

The Intersection:
- Large public government datasets (e.g., weather, census, healthcare)
- Other large datasets (e.g., scientific research, social media)

[Joel Gurin, Open Data Now]
# Big Data Science vs. Databases

<table>
<thead>
<tr>
<th></th>
<th>Databases</th>
<th>BD Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Volume</td>
<td>Modest</td>
<td>Massive</td>
</tr>
<tr>
<td>Examples</td>
<td>Bank records, Personnel records, Census, Medical records</td>
<td>Online clicks, GPS logs, Tweets, Building sensor readings</td>
</tr>
<tr>
<td>Priorities</td>
<td>Consistency, Error recovery, Auditability</td>
<td>Speed, Availability, Query richness (analysis)</td>
</tr>
<tr>
<td>Structured</td>
<td>Strongly (Schema)</td>
<td>Weakly or none (Images)</td>
</tr>
<tr>
<td>Realizations</td>
<td>SQL</td>
<td>NoSQL: Riak, Hbase, Apache River, MongoDB, Cassandra</td>
</tr>
</tbody>
</table>
Data Insights

Data makes everything clearer.
- Study involving seven countries (Ancel Keys, UCB)
- Total # of subjects: 13,000; 5-40 years follow-up

[Ancel Keys, Study from 1958-1970]
Big Data

Why all the excitement?
Example: New models are estimating which cities are most at risk for spread of the Ebola virus.

Mobile phone data from West Africa is being used to map population movements and predict how the Ebola virus might spread.

[BBC News, 2014; CDC]
Big Data

Why all the excitement?

Example: Google Flu Trends.

Using Google search to monitor millions of users’ health tracking behaviors online. Comparison with baseline of influenza activity.

97% accurate initially (2008), but sometimes inaccurate.

Ability to detect outbreaks two weeks ahead of CDC data.
Big Data

Why all the excitement?

Example:
Prediction that Obama would win the 2012 Presidential Election.

Tracking polls and forecasting using historical data and statistical analysis.

“Nate Silver” predicted the winner of all 50 states and the District of Columbia.

[John Canny, 2014]
Big Data Applications

Applications that make sense of big data using, for example, predictive analytics (e.g., classification by clustering or machine learning).

Detect health conditions:

Detect life-threatening conditions using context and wearable or hospital health devices in time to intervene.

[CZ 2014; IBM]
Big Data Applications

Detect extreme weather events (e.g., floods, fires):

Predict weather patterns to detect extreme events (e.g., fires, floods) or plan optimal wind turbine usage.

[CZ 2014; IBM]
Big Data Applications

Analyze transactions:

Make decisions based on real-time business transactional data (e.g., to detect fraud, deal with risks).

[CZ 2014; IBM]
Big Data Applications

Identify threats:

Identify criminals and threats from disparate video, images, audio, and data feeds (e.g., face recognition).

[CZ 2014; IBM]
Big Data Applications

Analyzing sentiment analysis:

Customer sentiment and experience analysis (e.g., to know whether customers like or dislike your products).

[CZ 2014; IBM]
Big Data Applications

Flood prediction using open and big data.

[Cowan, Alencar, McGarry, Palmer, 2015]
Software Engineering (SE) for Big Data (BD)

Software Engineering is “an engineering discipline that is concerned with all aspects of software production.”

<table>
<thead>
<tr>
<th>SE topics:</th>
<th>BD techniques:</th>
<th>Domains:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements</td>
<td>Analytics</td>
<td>Environment</td>
</tr>
<tr>
<td>Domain models</td>
<td>Mining</td>
<td>Health</td>
</tr>
<tr>
<td>Architectures</td>
<td>Recommendation</td>
<td>Socio-economic</td>
</tr>
<tr>
<td>Frameworks</td>
<td>Machine learning</td>
<td>Transportation</td>
</tr>
<tr>
<td>Services</td>
<td>Clustering</td>
<td>Agriculture</td>
</tr>
<tr>
<td>Processes</td>
<td>Streams</td>
<td>Smart cities</td>
</tr>
<tr>
<td>Model-driven</td>
<td>Visualization</td>
<td>Energy</td>
</tr>
<tr>
<td>development</td>
<td>Frequent itemsets</td>
<td>Manufacturing</td>
</tr>
<tr>
<td>Testing</td>
<td></td>
<td>e-commerce</td>
</tr>
</tbody>
</table>

[Leskov, Rajaraman, Ullman, 2014; Sommerville, 2011]
Software Engineering (SE) for Big Data (BD)

The focus is on studying how software engineering methods can be used to support the development of applications based on BD techniques.

**SE topics:**
- Requirements
- Domain models
- Architectures
- Frameworks
- Services
- Processes
- Model-driven development
- Testing

**Software Engineering Sources:**

[Leskov, Rajaraman, Ullman, 2014; Sommerville, 2011]
Software Engineering (SE) for Big Data (BD)

The focus is on studying how software engineering methods can be used to support the development of applications based on BD techniques.

**BD techniques:**
- Analytics
- Mining
- Recommendation
- Machine learning
- Clustering
- Streams
- Visualization
- Frequent itemsets

**Big Data Techniques Source:**

[Leskov, Rajaraman, Ullman, 2014; Sommerville, 2011]
Big Data Techniques

Big data techniques include:
- Finding similar items
  - Sets that have a large fraction of their elements in common
  - Finding similar web pages
- Data that arrives in streams
  - The data is lost if you do not do something about it immediately
  - Sensor data, image data, Internet and web traffic
- Finding frequent itemsets ("market basket" model)
  - Finding items that appear together in many baskets
  - Retailers can learn what is commonly bought together
- Clustering
  - Partition a large amount of data into subsets using distance measures
  - Partition the population of consumers into market segments

[Leskov, Rajaraman, Ullman, 2014; Sommerville, 2011]
Big Data Techniques

Big data techniques include:

- **Recommendation systems**
  - Recommending products or services based on what customers liked (e.g., collaborative filtering)
  - Offering customers suggestions about what they might like to buy based on their past history of purchases or product searches
  - Netflix, Amazon

- **Machine learning**
  - Learning a model or classifier from the data, and thus discover something about the data that will be seen in the future
  - Support vector machines, nearest-neighbor models, decision trees
  - Identify spam, sound, heart failure

[Leskov, Rajaraman, Ullman, 2014; Sommerville, 2011]
Software Engineering (SE) for Big Data (BD)

The focus is on studying how software engineering methods can be used to support the development of applications based on BD techniques.

**Domains:**
- Environment
- Health
- Socio-economic
- Transportation
- Agriculture
- Smart cities
- Energy
- Manufacturing
- e-commerce

**Big Data Techniques Applied in Various Domains. Examples are provided in:**
- Short papers
- Regular papers
- Survey papers

[Leskov, Rajaraman, Ullman, 2014; Sommerville, 2011]
Research Questions

• How to model the **requirements** of big data applications?
  – For example, models for the requirements of big data applications using a goal-oriented approach.

• How to use **domain models** for the big data techniques?
  – For example, domain models for recommendation systems applied to web usage mining.

• How can **model-driven development** be applied to big data techniques?
  – For example, model-driven methods that support machine learning application development.
Research Questions

• How can architectural design approaches help support big data software development?
• How can design or architectural patterns support the development of big data systems and applications?
• How to represent the evolution of big data systems?
• How to verify that big data applications behave as expected?
• How to test big data applications?
Example: Model-Driven Development

Case:
How to use model-driven methods to support the development of machine learning applications?

Characteristics:
- Design classes of applications
- Applications built from models
- Specific features are selected
- Reusable components

Benefits:
- Minimize programming
- Increase productivity
- Decrease time to market
Example: Domain-Models

Case: How to use domain-models to support modeling the requirements of web usage mining applications?

Characteristics:
- Requirements representation
- Domain models
- Domain-specific abstractions
- Domain-specific languages
- Web usage mining

Benefits:
- Improve productivity
- Language independent
- Improve reuse
Example: Metamodels

Case:
How to use metamodels to design and implement risk-aware recommendation systems?

Characteristics:
- Design representation
- Recommendation systems
- Risk-aware features

Benefits:
- Improve reusability
- Code generation
Example: Architecture

Case:
How to use models to represent the architectural design of big data context-aware applications involving sensors?

Characteristics:
- Architectural design
- Sensor-based applications
- Context-aware approach

Benefits:
- Reference architectures
- Improve productivity
Course Structure

Seminar-based course covering current research papers:
- Overview (short papers)
- Regular papers

Course Activities:
- Selecting short and regular papers
- Presenting papers and leading discussions
- Developing a research project:
  - Focus on a research topic involving some papers
  - Conduct a literature review
  - Conduct a case study
Course Structure

Marking:

- Presentations and leading discussions (50%)
  Short/Regular papers (35%)
  Participation/Paper discussions (15%)

- Project (50%)
  Presentation (20%)
  Report (30%)
Course Structure

What is next?

Short paper presentations:
- Three or four classes: short papers
- Special guest(s)

Regular paper presentations:
- Two presentations per class

Project presentations:
- TBA
Short Papers

Two or three-page overview papers. More than 70 papers are provided. Start selecting your papers!

Examples:

• Make Earth Observations Open Access
• Facial Analytics: From Big Data to Law Enforcement
• Lessons for Big Data Projects
• The Parable of the Google Flu
• New Smart Cities
• Managing your Digital Life
• Big Data meets Public Health
• The Science of Managing Data Science
• Turning Data into Knowledge
• Big Data for Modern Industry
Short Papers

Presenting the papers and leading discussions:
- We will use about 7-10 slides for each presentation (pdf)
- We will need about 7 minutes for the presentation and 3 minutes for discussion
- Participation is important!
Regular Papers

These are research-oriented papers covering aspects of software engineering applied to big data applications. More than 80 papers will be provided.

Examples:

• Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems
• Guiding the Introduction of Big Data in Organizations
• CAP: Community Activity Prediction Based on Big Data Analysis
• The Evolvement of Big Data Systems
• Managing a Big Data Project
• Mapping the Data Shadows of Hurricane Sandy
Regular Papers

More examples:

• Development of an Intelligent Environmental Knowledge System for Sustainable Agricultural Decision Support
• iCARE: A Framework for Big Data-based Banking Analytics
• SemantEco: A Semantically Powered Modular Architecture for Integrating Distributed Environmental and Ecological Data
• Intelligent Services for Big Data Science
• Towards Model-Driven Engineering of Big Data Analytics
• Assembling Cloud-based Geographic Information Systems
• A Domain Model of Web Recommender Systems
• An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things
• Big Data for Building Energy Performance
Survey Papers

These are research-oriented papers presenting surveys of specific relevant areas. They can be used to support research and as general references. More than 30 survey papers will be provided.

Examples:

• Data-intensive applications, challenges, techniques, and technologies: A Survey on Big Data
• Beyond the Hype: Big Data concepts, methods, and analytics
• Geospatial Big Data: Challenges and Opportunities
• Big Data for Health
• Towards Scalable Systems for Big Data Analytics
• Big Data for Open Digital Collaboration – A Research Roadmap
Project

The students are encouraged to select project topics in line with the course main objectives (e.g., related to their thesis theme, expertise or interest).

Some projects:
- Software framework for face recognition
- Designing risk-aware recommendation systems
- Designing context-aware IoT applications
- Literature review: software engineering for big data applications
Questions

Please contact:

Paulo Alencar
palencar@cs.uwaterloo.ca

Course website:
cs.uwaterloo.ca/~palencar/cs846/