Machine Learning for Big Data Systems Acquisition

John Klein

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Motivation

Acquisition Aspiration

- “Choose a modern technology stack” (playbook.cio.gov)

Acquisition Reality

- “The subject matter competencies for successful enterprise IT system acquisition are often missing in government” (GAO)

Trusted knowledge bases are part of the solution

- In FY14 we built knowledge base for NoSQL technology
  - Quality at Scale for Big Data – QuABaseBD
  - http://quabase.sei.cmu.edu
  - Knowledge model – categories, features, allowable values
- Expensive to curate - populate and maintain knowledge as products evolve

Motivation

Acquisition Aspiration

- “Choose a modern technology stack” (playbook.cio.gov)

Acquisition Reality

- “The subject matter competencies for successful enterprise IT system acquisition are often missing in gov

Project Focus

Trusted knowledge bases are part of the soluti

- In FY14 we built knowledge base for NoS technology
  - Quality at Scale for Big Data – QuABas D
  - http://quabase.sei.cmu.edu
  - Knowledge model – categories, features, allowable values
- Expensive to curate - populate and maintain knowledge as products evolve

Problem and Approach

Research Question:

Can we automatically identify relevant document pages that contain the knowledge required for a curator to populate the knowledge base for a product feature?

Approach:

- 2-level supervised machine learning classifier
  - Document model
  - Sentence model
- Train using QuABaseBD contents
- Assess classifier precision, simultaneously extend training set with labeled documents and passages
- Measure classifier improvement

Challenges of Technical Knowledge Curation

Quantity of Information and Diversity of Structure

- Oracle NoSQL – 1000s of fine-grained pages, multiple "volumes"
- Accumulo – single web page with all documentation topics

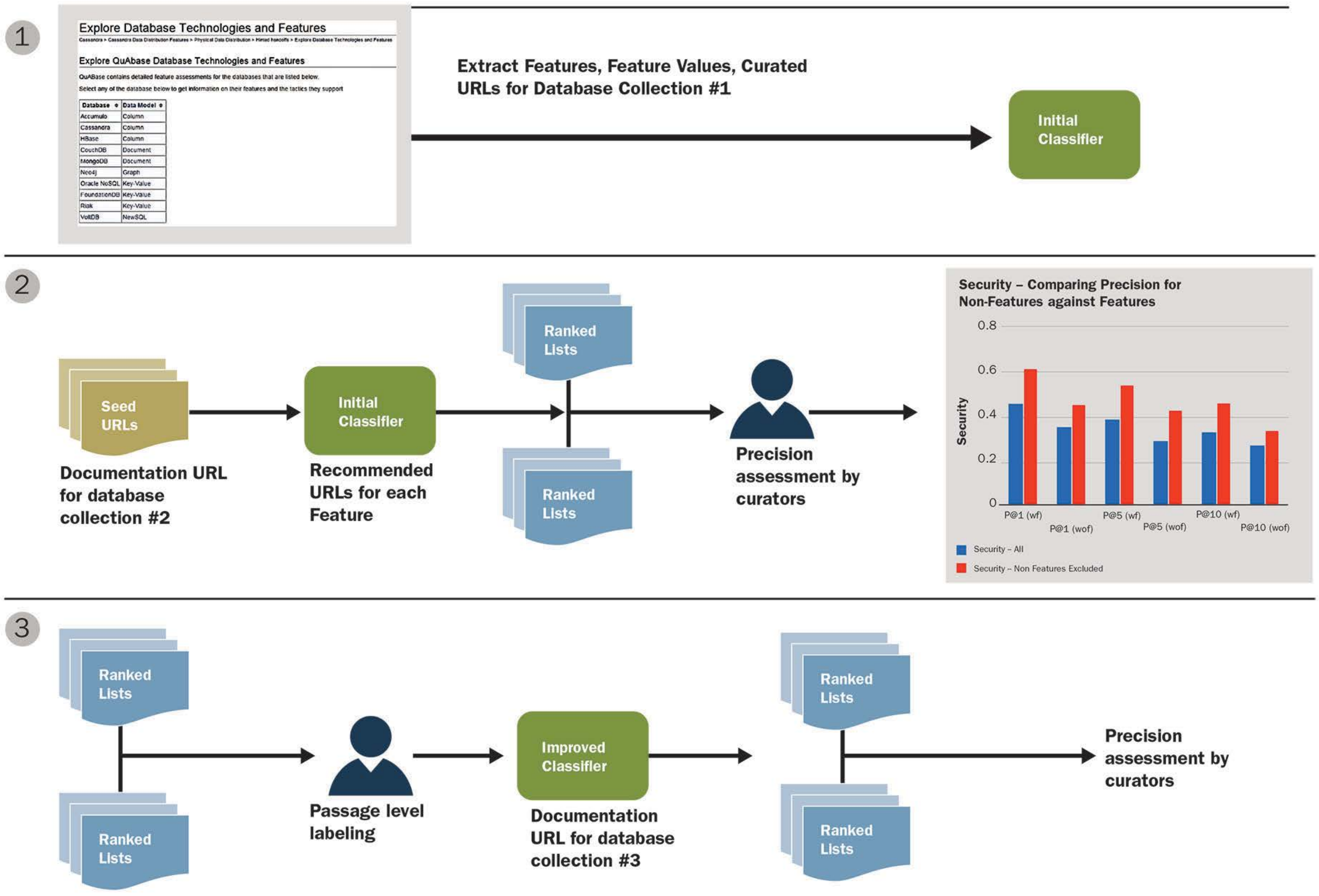
Ambiguous Terminology

- CAP – fundamental quality tradeoff in distributed systems
  - Consistency – Replica or transactional?
  - Availability – System property or semantic dependency ("feature X is available only when configuration flag Y is enabled")
  - Partition – network failure or database shard?

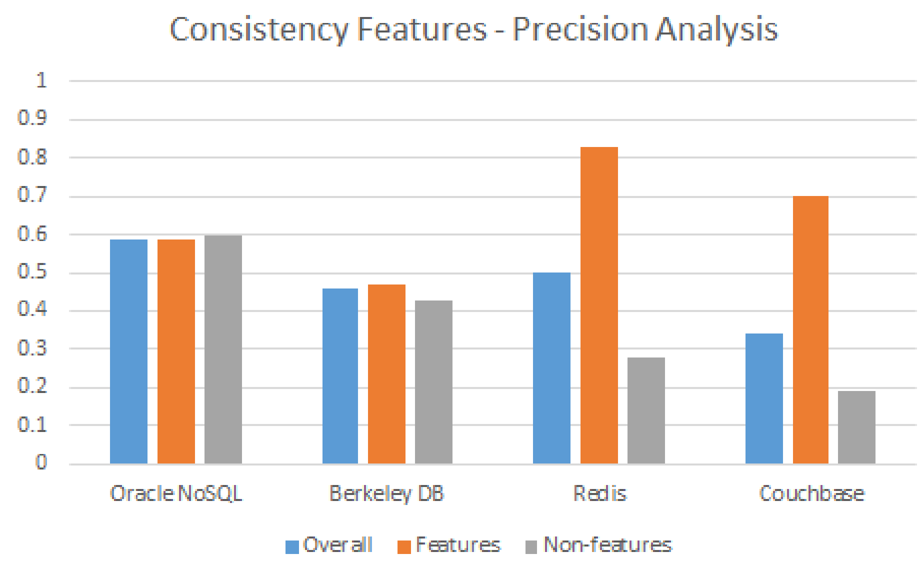
Unsupported Features

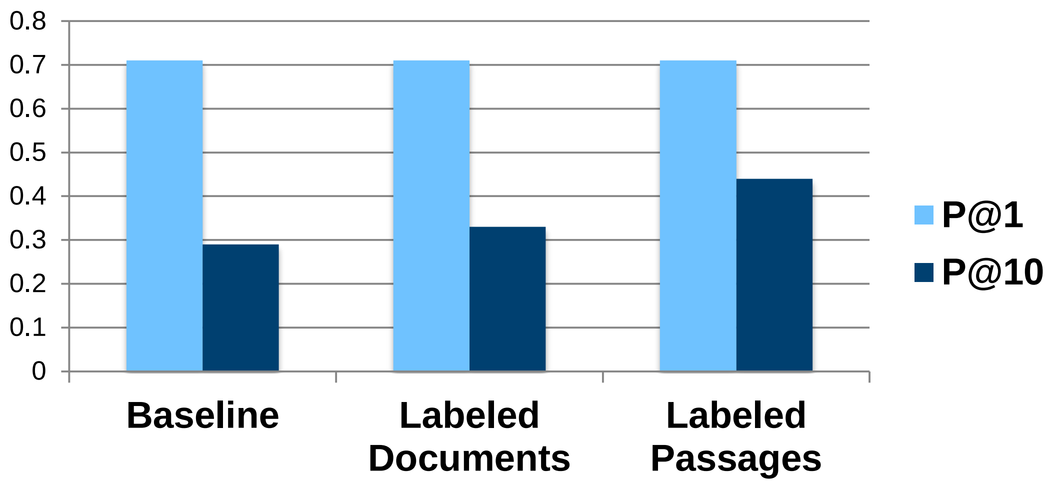
- Rare to find explicit statement that a feature is *not* supported
- Closed-world assumption requires rich feature dependency model

Experiment Approach



Results Towards Automation





Precision better for supported features (Orange bars) (p=0.03)

Sensitive to

- Documentation structure
- Product feature-richness

Classifier performance improved as training set was extended

Future Work

Classifier performance was limited by available training data

- Extend training sets and identify limit of classifier performance
- Assess classifier performance on other knowledge base feature categories

Systematically investigate performance sensitivities to develop confidence measures

- Quantify differences in document structure and writing style, product feature-richness, other heuristics

Assess classifier performance on new versions of product/documentation

- Knowledge base evolution/maintenance scenario may be more automate-able

Research Team

Principal Investigator: Prof. Ian Gorton (Northeastern U., ex-SEI)

Classifier Development: Prof. Yiming Yang

(CMU Language Technology Institute)

Domain Experts: Soumya Simanta

(SEI) John Klein

Contact Information

John Klein
Senior Member of Technical Staff
Software Solutions Division
Telephone: +1 617-283-2170
Email: jklein@sei.cmu.edu

U.S. Mail
Software Engineering Institute
Customer Relations
4500 Fifth Avenue
Pittsburgh, PA 15213-2612
USA

Web
www.sei.cmu.edu
www.sei.cmu.edu/contact.cfm

Customer Relations
Email: info@sei.cmu.edu
Telephone: +1 412-268-5800
SEI Phone: +1 412-268-5800
SEI Fax: +1 412-268-6257