

# Machine Learning for Big Data Systems Acquisition

John Klein

Software Engineering Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213

Copyright 2015 Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

This material has been approved for public release and unlimited distribution except as restricted below.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at [permission@sei.cmu.edu](mailto:permission@sei.cmu.edu).

DM-0002843



# 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 solution

- In FY14 we built knowledge base for NoSQL technology
  - Quality at Scale for Big Data – QuABase
  - <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

1

Explore Database Technologies and Features

CaseWare • Cassandra Data Distribution Features • Physical Data Distribution • HBase Security • Explore Database Technologies and Features

Explore QuBase Database Technologies and Features

QuBase contains detailed feature assessments for the databases that are listed below. Select any of the database below to get information on their features and the tactics they support

Database	Data Model
Accumulo	Column
Cassandra	Column
HBase	Column
CouchDB	Document
MongoDB	Document
Neo4j	Graph
Oracle NoSQL	Key-Value
FoundationDB	Key-Value
Riak	Key-Value
VolDB	NewSQL

Extract Features, Feature Values, Curated URLs for Database Collection #1



2



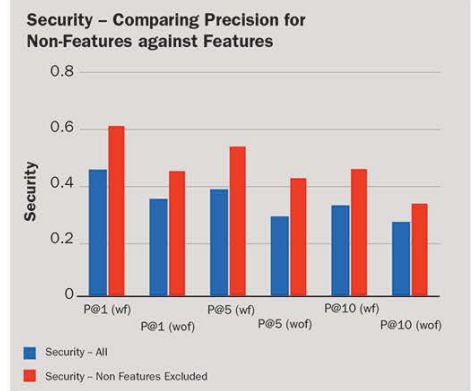
Documentation URL for database collection #2



Recommended URLs for each Feature



Precision assessment by curators



3



Passage level labeling



Documentation URL for database collection #3

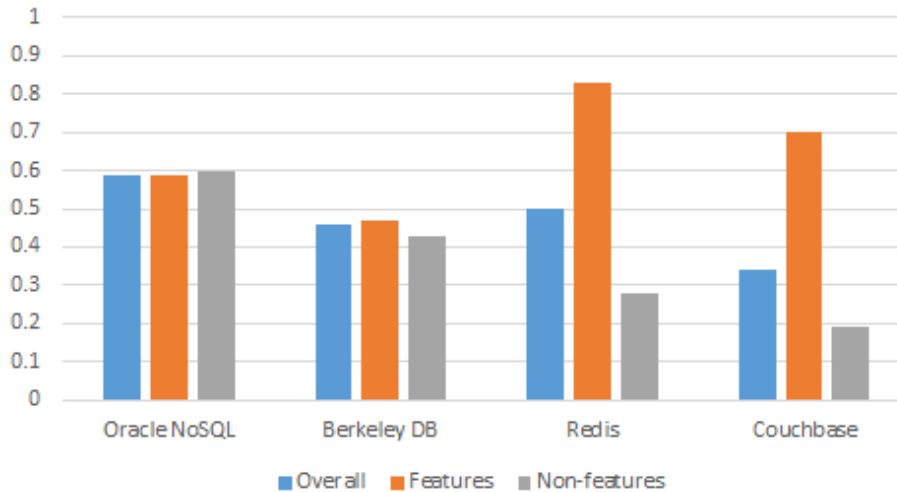


Precision assessment by curators



# Results Towards Automation

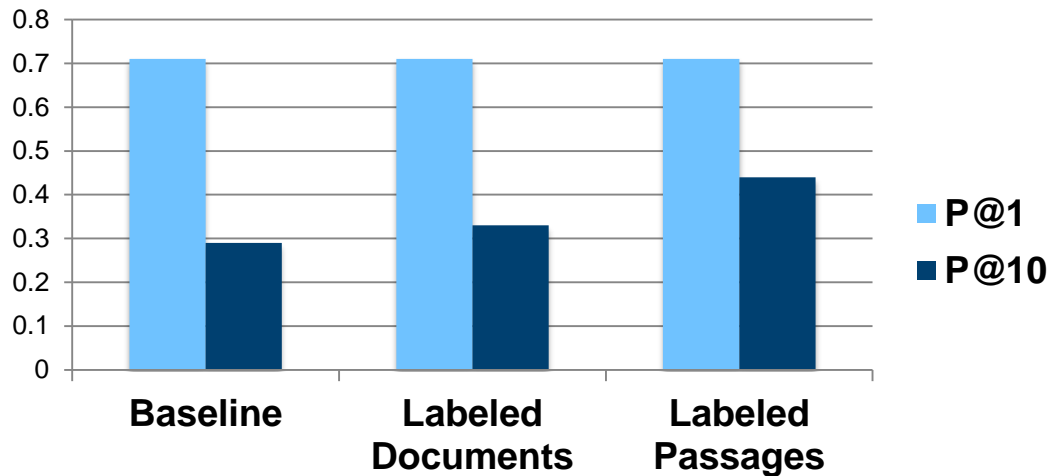
Consistency Features - Precision Analysis



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](mailto: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](http://www.sei.cmu.edu)  
[www.sei.cmu.edu/contact.cfm](http://www.sei.cmu.edu/contact.cfm)

## Customer Relations

Email: [info@sei.cmu.edu](mailto:info@sei.cmu.edu)  
Telephone: +1 412-268-5800  
**SEI Phone:** +1 412-268-5800  
**SEI Fax:** +1 412-268-6257