



proudly present:

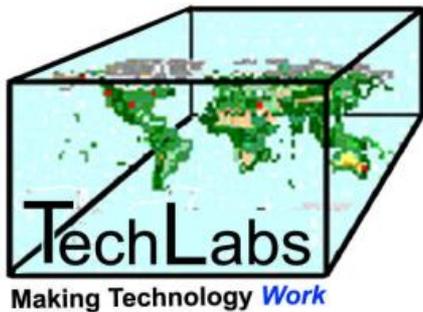


# Mining the Cloud



Mr. Stephen W. Leibholz  
TechLabs, Inc.

Dr. Mitchell C. Kerman  
Systems Engineering Research Center (SERC)  
Stevens Institute of Technology



SYSTEMS ENGINEERING  
Research Center



**STEVENS**  
INSTITUTE of TECHNOLOGY  
THE INNOVATION UNIVERSITY®

# CLOUDDEVELOPERS

Mobile, Big Data & Service Models:  
Critical Take-Aways for Cloud Developers

SUMMIT & EXPO 2014



# Context

**Mining the Cloud:** Searching large, continuously updating *narrative* data collections (measured in gigabytes, terabytes, petabytes, or exabytes) and integrating pertinent information. These data collections include messages, e-mails, documents, webpages, etc.

# The Problem

- Current web and cloud mining techniques are
  - Inefficient
    - Time-consuming
      - Take too long to explore and analyze all of the data
    - Time-late
      - Take too long to convert actionable data into information for user consumption, apprehension, and timely action
  - Ineffective
    - Leave “money” (highly relevant data) on the table
    - Spend a lot of time presenting irrelevant data
  - Inaccurate
    - Not truly based on statistical methods, processes, and techniques
    - Do not take advantage of the newest techniques in artificial intelligence and machine learning
- We want to pose these as questions and challenges to both the development and user communities

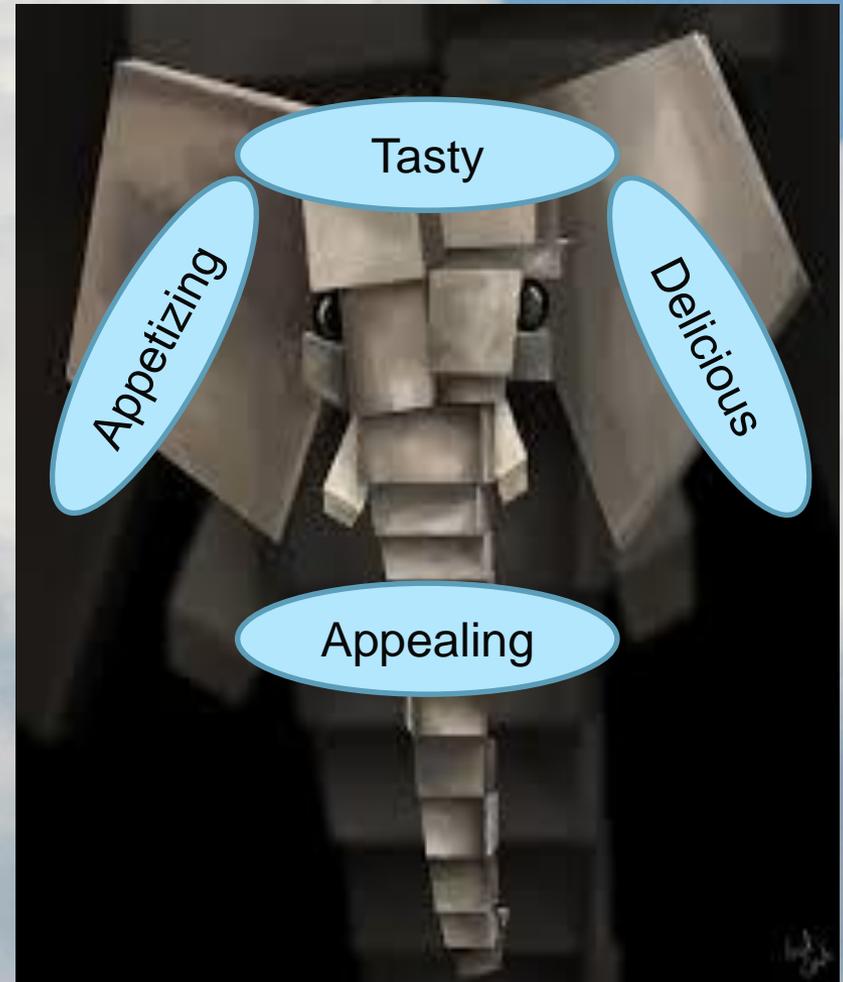
# How do we eat the Digital Elephant?

- One byte at a time?
- Multiple bytes at a time?



# How do we eat the Digital Elephant?

- One byte at a time?
- Multiple bytes at a time?
- Change the paradigm
  - Which parts of the elephant are the most appealing, appetizing, tasty, and delicious?
  - In other words, what is most *relevant* to us?



# What is Pertinent / Relevant?

- In Operations and Intelligence scenarios, the Cloud is continuously updated
- Manual search is ineffective
  - Manual scanning is conducted at a rate of 75 documents per hour
- Typical Narrative Data Collections during Operations are very large
  - Emails, Documents, Messages, Memos, and Transcriptions
  - 100,000+ files is the low end
  - 500,000+ files are commonplace
- It's simple math:
  - $(100,000 \text{ documents}) / (75 \text{ documents / hour}) = 1333+ \text{ hours}$
  - $(500,000 \text{ documents}) / (75 \text{ documents / hour}) = 6666+ \text{ hours}$

# Problems with Manual Review

- Cost
  - Gartner estimates the cost of reviewing one gigabyte of ESI at \$18,750.<sup>1</sup>
- Human Error – Reviewers only tag a document the same way 65% of the time.<sup>2</sup>
- Time – manual review is ***SLOW***

1. Logan and Bace, eDiscovery Project Planning and Budgeting: 2008-2011, Feb. 2008
2. Voorhees, Variation in Relevance Judgments and the Measurement of Retrieval Effectiveness, Information Processing & Management, 2000

# Word Search: Summary of NIST Study

- “Although the searchers believed they had found 75% of the relevant documents, their average recall was only 20%. The prior art used word search as a method of automating the correlation process.”  
[\[http://trec.nist.gov/pubs/trec4/t4\\_proceedings.html\]](http://trec.nist.gov/pubs/trec4/t4_proceedings.html)
- *Our studies have verified this and also indicate that manual search only finds about 20% of the relevant data along with about 35% irrelevant selections.*

# ROC Curves

- ROC (Receiver Operating Characteristic) curves stem from experience in WW II.
- Radio operators learned that turning up the signal also turned up the noise.
- Modern signal processing shows us that increased probability of detection means increased false positives.
- The user has to make a tactical decision.

# The Inevitable Tradeoff

Precision (Selectivity)  $\longleftrightarrow$  Recall (Sensitivity)

**Typical Operating-Characteristic  
Tradeoff Benefit Using AI**



# Boolean Operators are Limited

- Results from authors' experience.
- Only documents containing keywords or phrases are selected by keyword search. Connotations are not discovered.
- Synonymy not considered in word search.
- References are known to be variable.
  - e.g., “invoice” possibly missing, but “tender” is present

# Artificial Intelligence Application

- Instead, use statistical approaches to augment word search.
- Iterative supervised learning process is sensitive to anomalous statistics of word usage.
- Numerical data is assumed important.
- Narrative data: Compare frequency of words or phrases in corpus of collected narrative data versus frequency in general (English) language.

# Predictive Coding

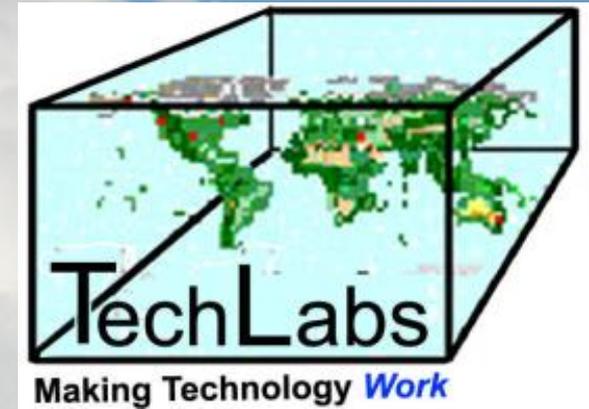
- Derived from the field of Artificial Intelligence
- Methodology used is 40+ years old
- Machine “learns” from data supplied by collectors
- Machine continues to improve over time through iterative learning process and fresh data
- Significant cost savings possible – on the order of 70% to 90%
- General acceptance by courts in regard to legal documents

# Conclusions

- ❖ Word search is ineffective.
- ❖ Manual (linear) review is no longer practical in even modest-size operations due to the exponential growth in data collection and storage.
- ❖ Computers have surpassed humans in performing routine tasks such as document review.
- ❖ Because of the Internet and our engineering business, we have unprecedented access to relevant information in Intelligence.
- ❖ Future technology advances will make further significant changes in legal, medical, financial, and other fields.

# Shameless Plugs

[www.techlabs.com](http://www.techlabs.com)



[www.stevens.edu](http://www.stevens.edu)



**STEVENS**  
INSTITUTE of TECHNOLOGY  
THE INNOVATION UNIVERSITY®

[www.SERCuarc.org](http://www.SERCuarc.org)

