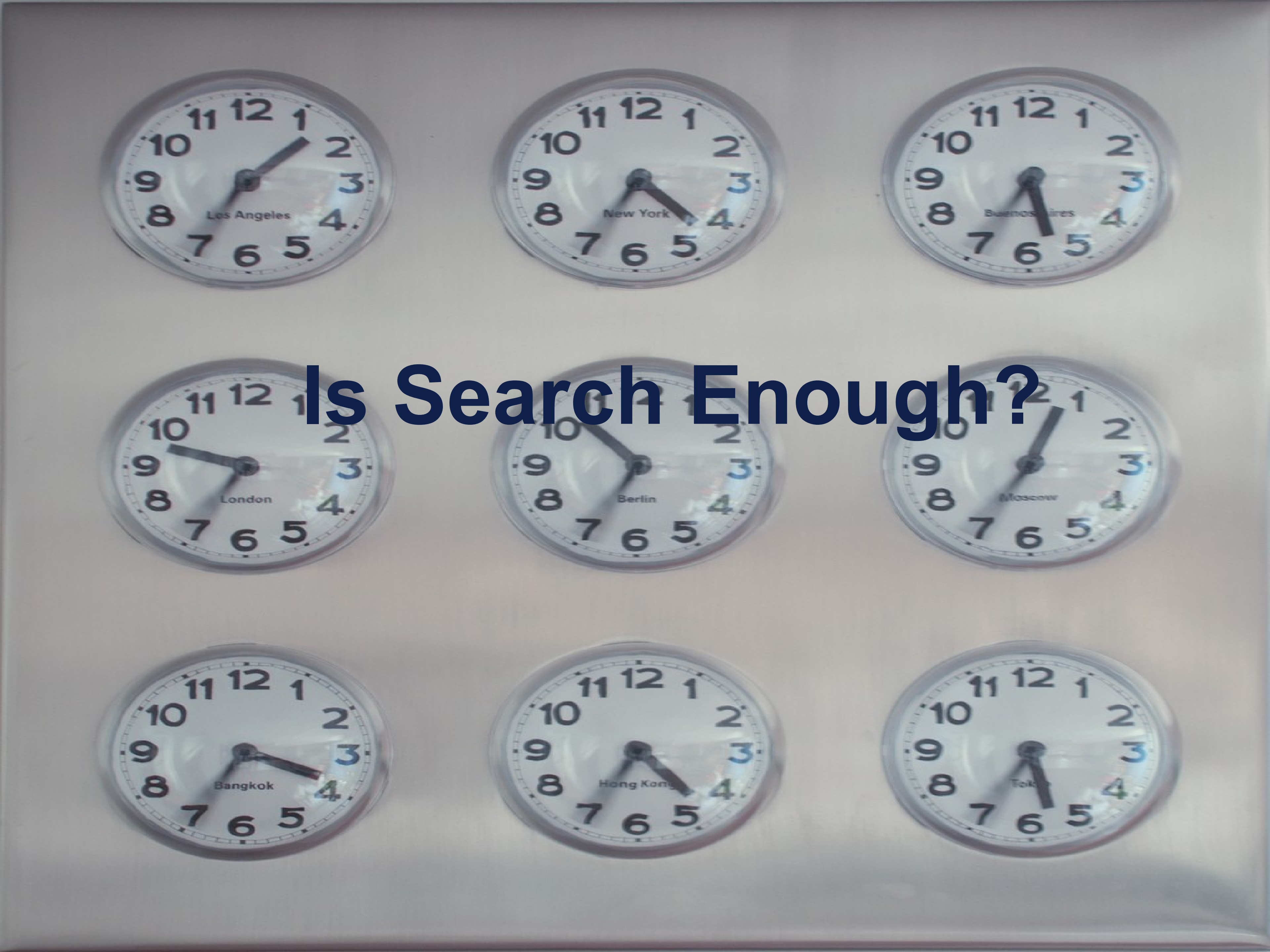
Crowd Sourced Intelligence Built into Search and Hadoop

Grant Ingersoll, LucidWorks, @gsingers

Search | Discover | Analyze



Is Search Enough?

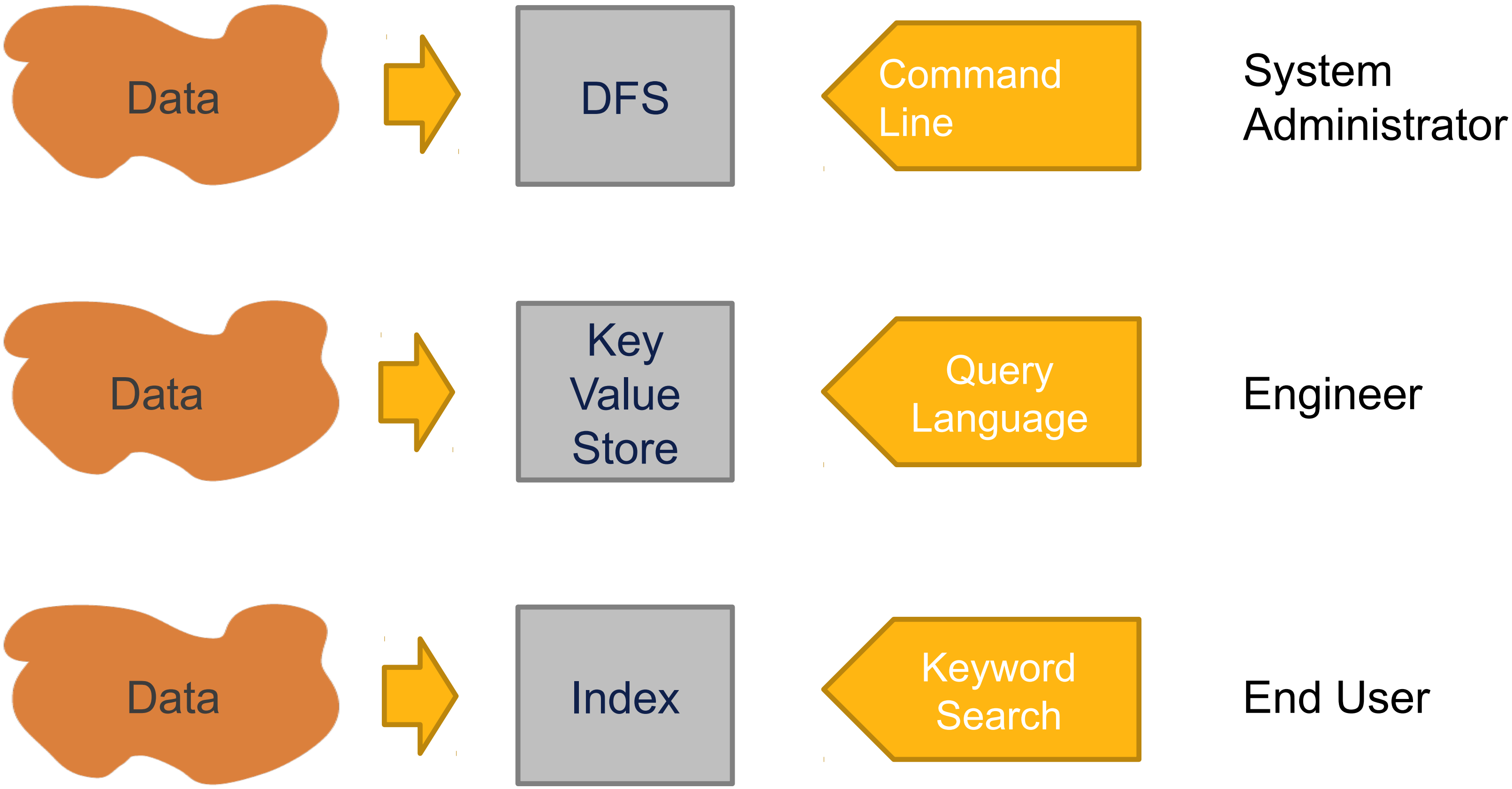
Is Search Enough?

- Keyword search is a commodity
- Holistic view of the data and the user interactions with that data are critical
- Search, Discovery and Analytics are the key to unlocking this view of users and data

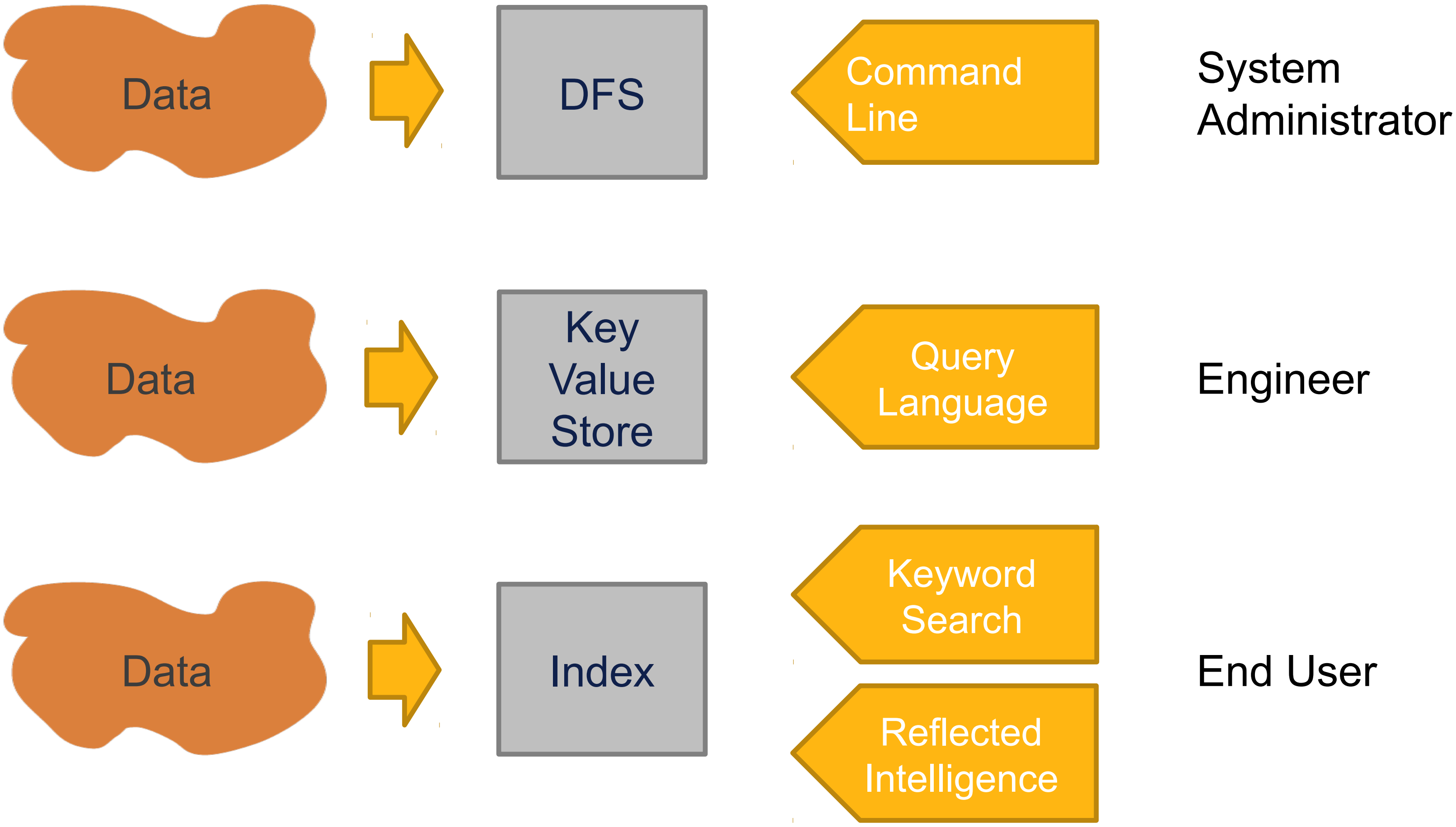
Agenda

- Intro
- Search (R)evolution
- Reflected Intelligence Use Cases
- Building a Next Generation Search and Discovery Platform
  - LucidWorks
- Easy Technical Wins
- 1+1=3

User Interactions With Big Data



User Interactions With Big Data



Search (R)evolution

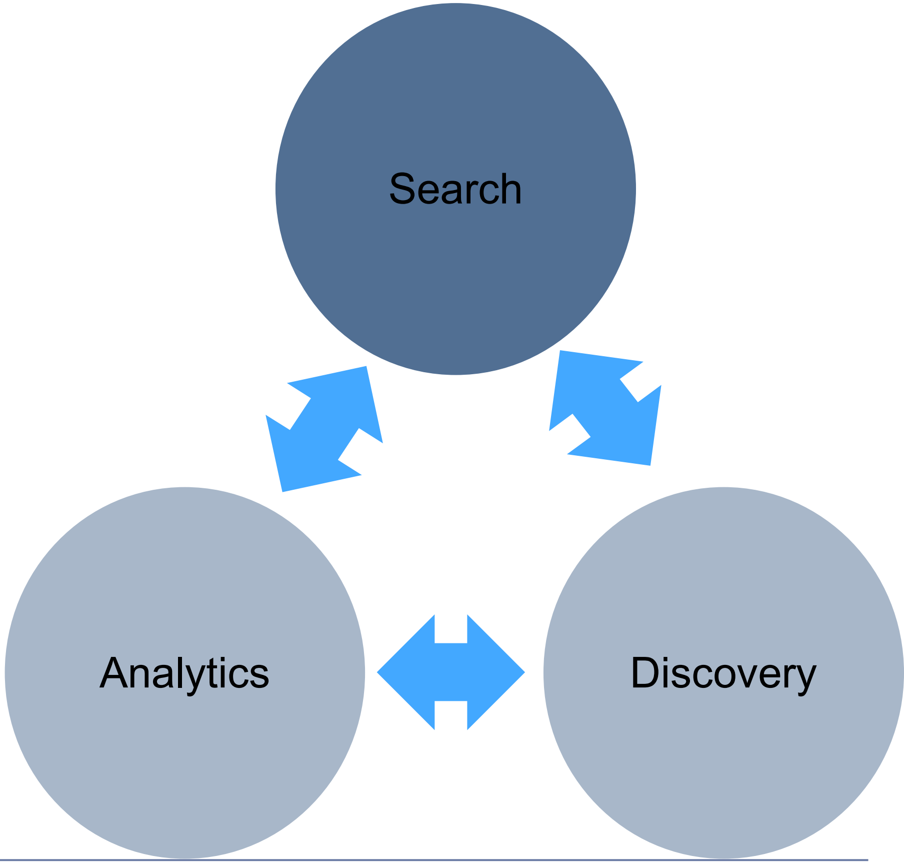
- Search use leads to search abuse
  - denormalization frees your mind
  - scoring is just a sparse matrix multiply
- Lucene/Solr evolution
  - non free text usages abound
  - many DB-like features
  - noSQL before NoSQL was cool
  - flexible indexing
  - finite State Transducers FTW!
- Scale
- “This ain’t your father’s relevance anymore”

Search, Discovery and Analytics

- Large-scale analysis is key to reflected intelligence
  - correlation analysis
  - based on queries, clicks, mouse tracks,
    - even explicit feedback
  - produce clusters, trends, topics, SIP's
  - start with engineered knowledge,

    refine with user feedback
- Large-scale discovery features

  encourage experimentation
- Always test, always enrich!



Social Media Analysis in Telecom

- Goal
  - Detect flash-mob traffic events
  - Provision additional resources before failures
- Method: Correlate mobile traffic analysis with social media analysis
  - events cause traffic micro-bursts
  - participants tweet the events ahead of time
  - tweet locations converge on burst location
- Deploy operations faster to predict outages and better handle emergency situations
  - high cost bandwidth augmentation can be marshaled as the traffic appears
  - anticipation beats reaction

Provenance is 80% of Value

- Problem
  - Broadcasters don’t know what audiences really like at a micro level
- Method:
  - Analysis of social media to determine advertising reach and response
  - Time resolution of social traffic can provide detailed response metrics
- Results:
  - In one case the untargeted advertising was worth 5x more if with supporting response data

Claims Analysis

- Goal
  - Insurance claims processing and analysis
  - fraud analysis
- Method
  - Combine free text search with metadata analysis to identify high risk activities across the country
  - Integrate with corporate workflows to detect and fix outliers in customer relations
- Results
  - Questions that took 24-48 hours now take seconds to answer

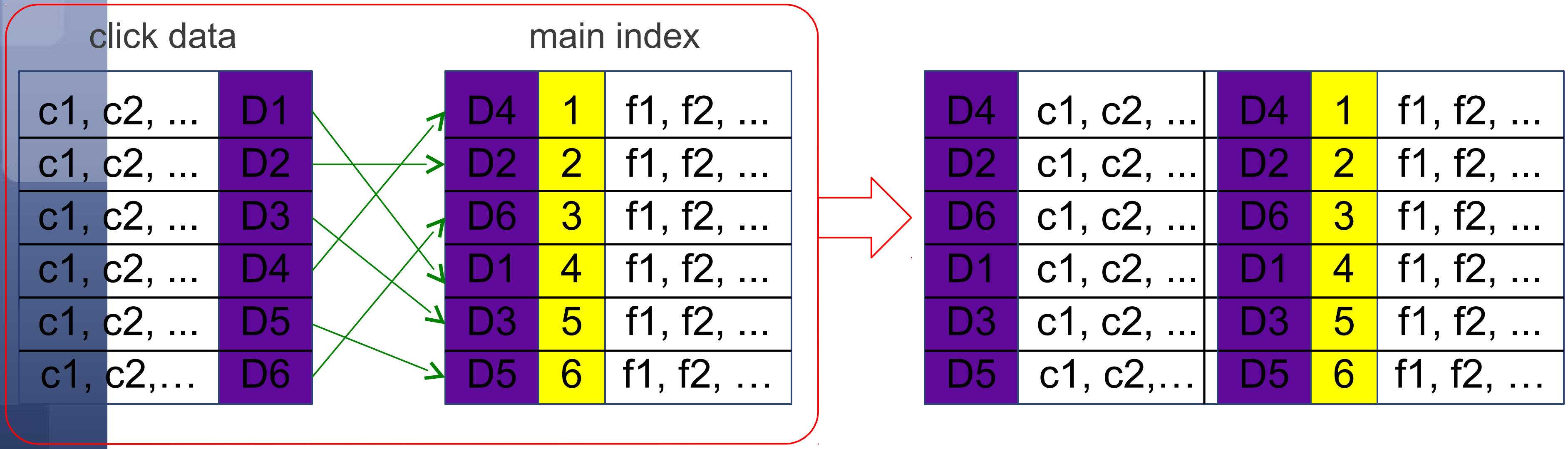
Can Search Catch the Bad Guys?

- Online Drug Counterfeit detection
- Identify commonly used language indicating counterfeits
  - you know it when you see it
  - and you know you have seen it
- Leverage:
  - Statistically Interesting Phrases
  - Clustering
  - Other Analysts
- Feed to analyst via search-driven application
  - enrich based on analysts feedback

Learn to Rank

- Go beyond TF/IDF by leveraging user votes
- Log all clicks per query
- Periodically process the logs to determine most popular items per query
- “Update” Lucene index underneath the hood with query X boost factors
  - Alternatively: train a classifier to learn rankings
  - Beware of self-fulfilling results!
- Profit!

Via ParallelReader



click data

c1, c2, ...	D1
c1, c2, ...	D2
c1, c2, ...	D3
c1, c2, ...	D4
c1, c2, ...	D5
c1, c2,…	D6

main index

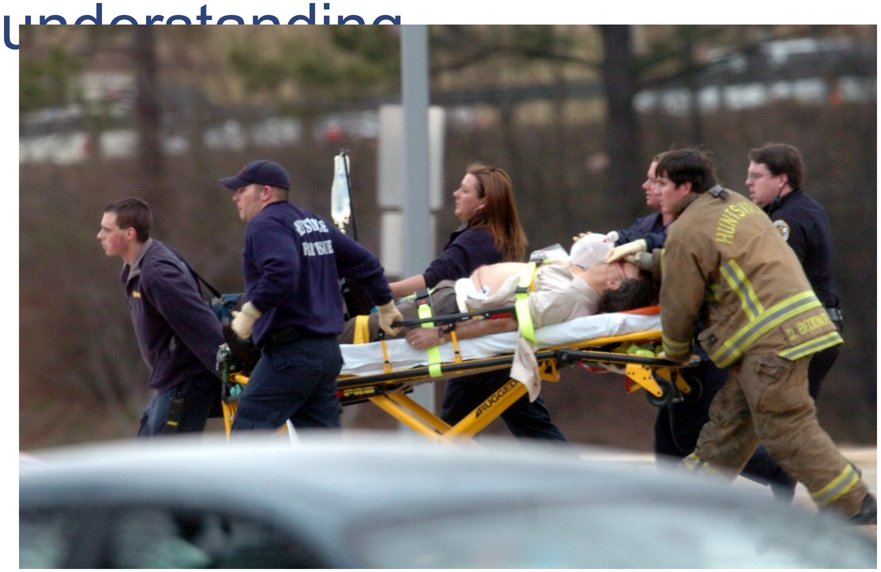
D4	1	f1, f2, ...
D2	2	f1, f2, ...
D6	3	f1, f2, ...
D1	4	f1, f2, ...
D3	5	f1, f2, ...
D5	6	f1, f2, …

D4	c1, c2, ...	D4	1	f1, f2, ...
D2	c1, c2, ...	D2	2	f1, f2, ...
D6	c1, c2, ...	D6	3	f1, f2, ...
D1	c1, c2, ...	D1	4	f1, f2, ...
D3	c1, c2, ...	D3	5	f1, f2, ...
D5	c1, c2,…	D5	6	f1, f2, …

- Pros:
  - All click data (e.g. searchable labels) can be added
- Cons:
  - Complicated and fragile (rebuild on every update)
  - Though only the click index needs a rebuild
  - No tools to manage this parallel index in Solr

Virginia Tech - Help the World

- Grab data around crisis
  - Crowd sourced from Twitter, etc.
- Search immediately
- Large-scale analysis enriches data to find ways to improve responses and understanding
- http://www.ctrnet.net



Veoh - Cross Recommendations

- Cross recommendation as search
  - with search used to build cross recommendation!
- Recommend content to people who exhibit certain behaviors (clicks, query terms, other)
- (Ab)use of a search engine
  - but not as a search engine for content
  - more like a search engine for behavior

Recommendation Basics

- See Ted’s talk from this morning on Multi-modal Recommendation Algorithms
- Go get Mahout/Myrrix or just do it in y(our) search engine

Search Engine for Reflected Intelligence

- Map-reduce “big data” part
  - Logs record user + item occurrence
  - Group by user to get rows of occurrence matrix
  - Self-join to get co-occurrence
  - Log-likelihood test to find anomalies

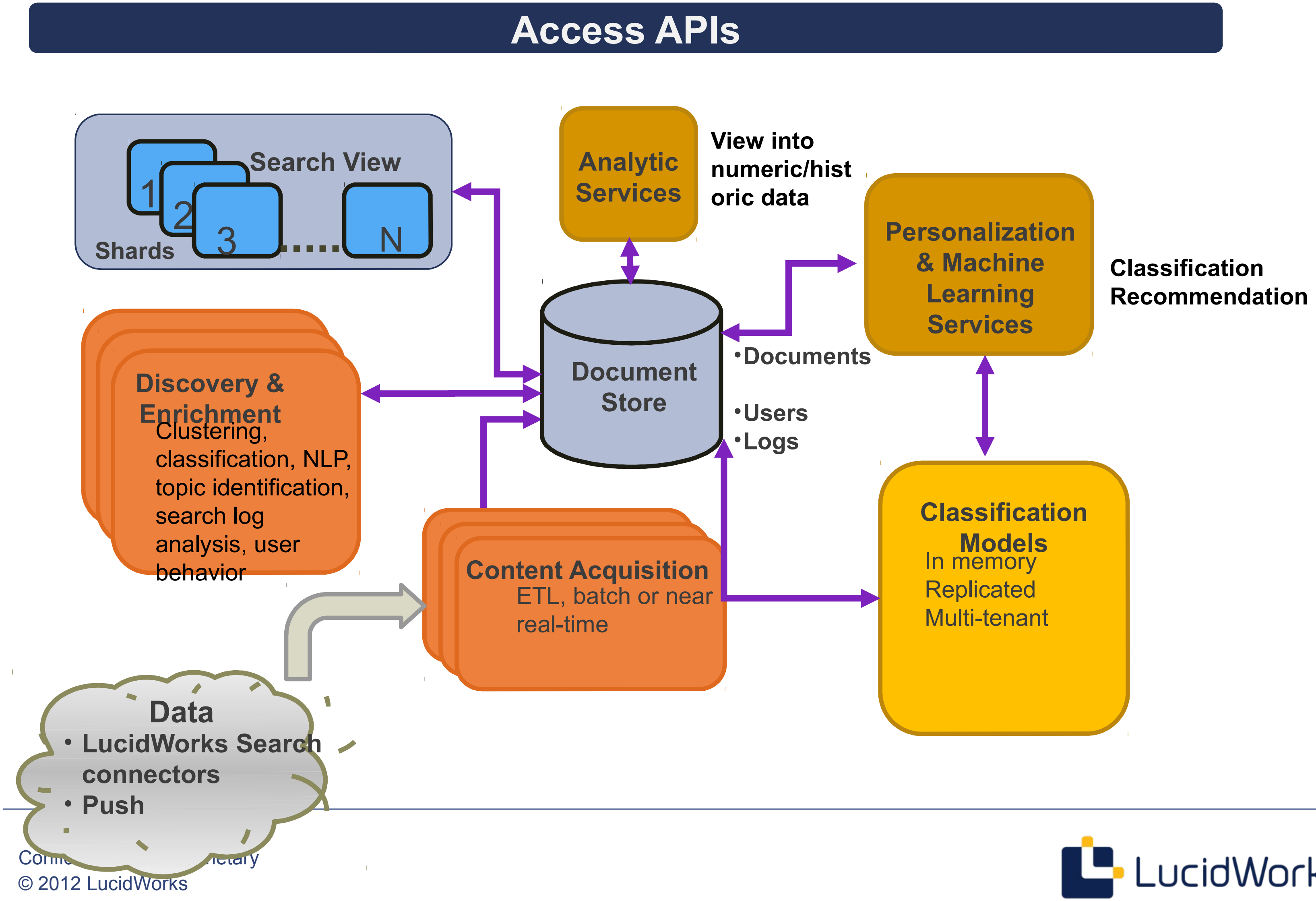
- Search part
  - Anomalous cooccurrences are indicators
    - (or use statistical scores to provide fancy boosts)
  - Indicator fields and other meta-data are indexed
  - Recommendation implemented using a single search

  - Boosts, functions, similarity also can reflect learned behavior

What Platform Do You Need?

- Fast, efficient, scalable search
  - bulk and near real-time indexing
  - handle billions of records with sub-second search and faceting
- Large scale, cost effective storage and processing capabilities
- NLP and machine learning tools that scale to enhance discovery and analysis
- Integrated log analysis workflows that close the loop between the raw data and user interactions
- Easy API access with support for programming language of their choice
- Content acquisition across a variety of enterprise, Internet and social connectors

Reference Architecture



Access APIs

Search View

Shards

1

2

3

N

Analytic Services

View into numeric/historic data

Personalization & Machine Learning Services

Classification Recommendation

Document Store

•Documents

•Users

•Logs

Discovery & Enrichment

Clustering, classification, NLP, topic identification, search log analysis, user behavior

Content Acquisition

ETL, batch or near real-time

Classification Models

In memory

Replicated

Multi-tenant

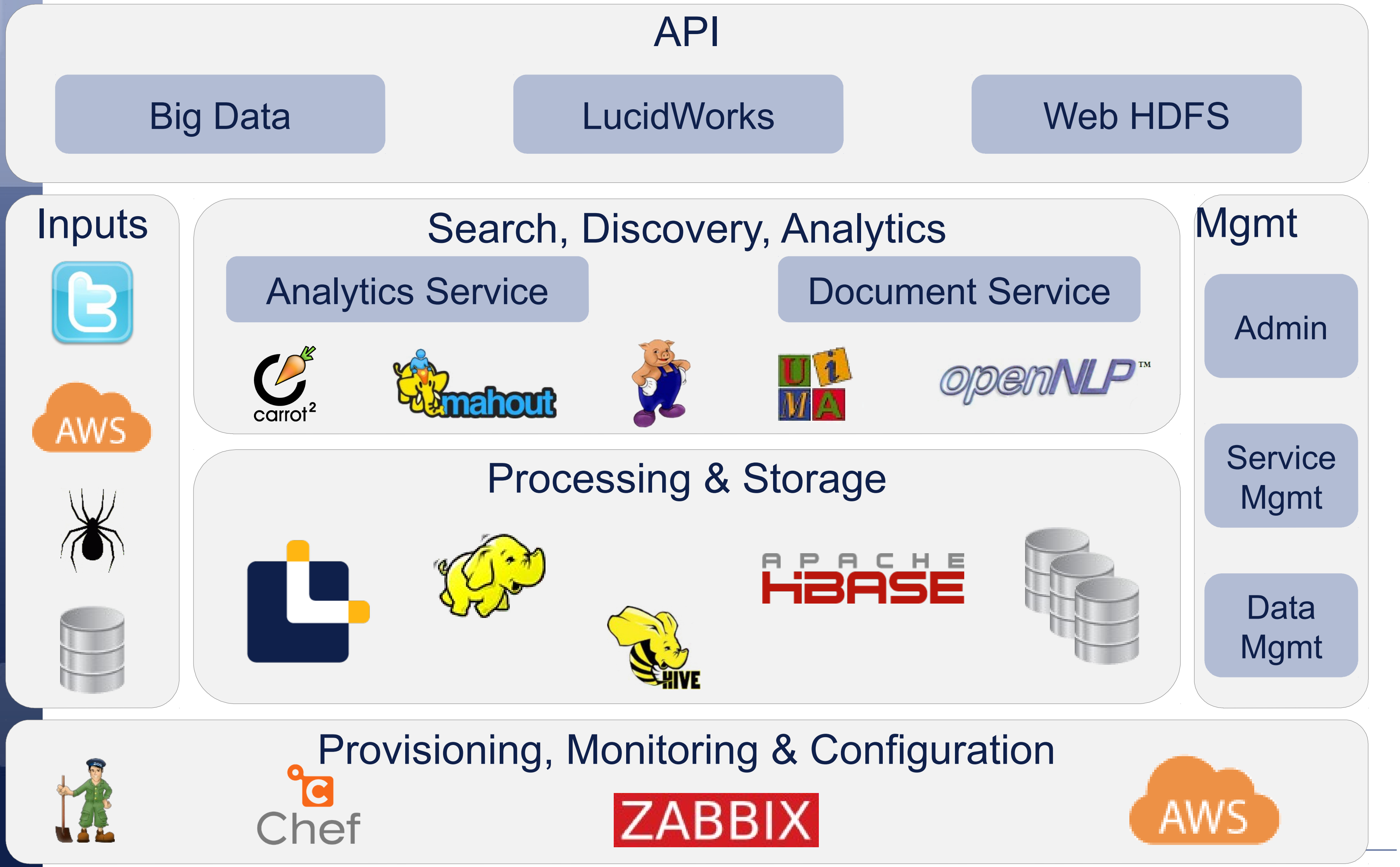
Data

- LucidWorks Search connectors
- Push

LucidWorks

- LucidWorks provides the leading packaging of Apache Lucene and Solr
  - build your own, we support
  - founded by the many prominent Lucene/Solr experts
- LucidWorks Search
  - “Solr++”
  - UI, REST API, MapR connectors, relevance tools, much more
- LucidWorks Big Data
  - Big Data as a Service
  - Integrated LucidWorks Search, Hadoop, machine learning with prebuilt workflows for many of these tasks

LucidWorks Big Data



API

Big Data

LucidWorks

Web HDFS

Inputs

Search, Discovery, Analytics

Analytics Service

Document Service

Processing & Storage

Mgmt

Admin

Service Mgmt

Data Mgmt

Provisioning, Monitoring & Configuration

Easy Technical Wins

- Analyze logs from application stored in Hadoop/MapR
- Seamlessly store search indexes in Hadoop/MapR
  - and feed to Pig, Mahout and others
  - use mirrors + NFS to directly deploy indexes
- LucidWorks 2.5 easily connects with Hadoop/MapR
  - Click ranking, other log analysis built in
  - Classification as service
  - Offline Enrichment

1 + 1 = 3

Learn More

- Talk to Grant

  @gsingers

  grant@lucidworks.com

- LucidWorks

  http://www.lucidworks.com

  Hash Tags

  #lucene #solr #lucidworks