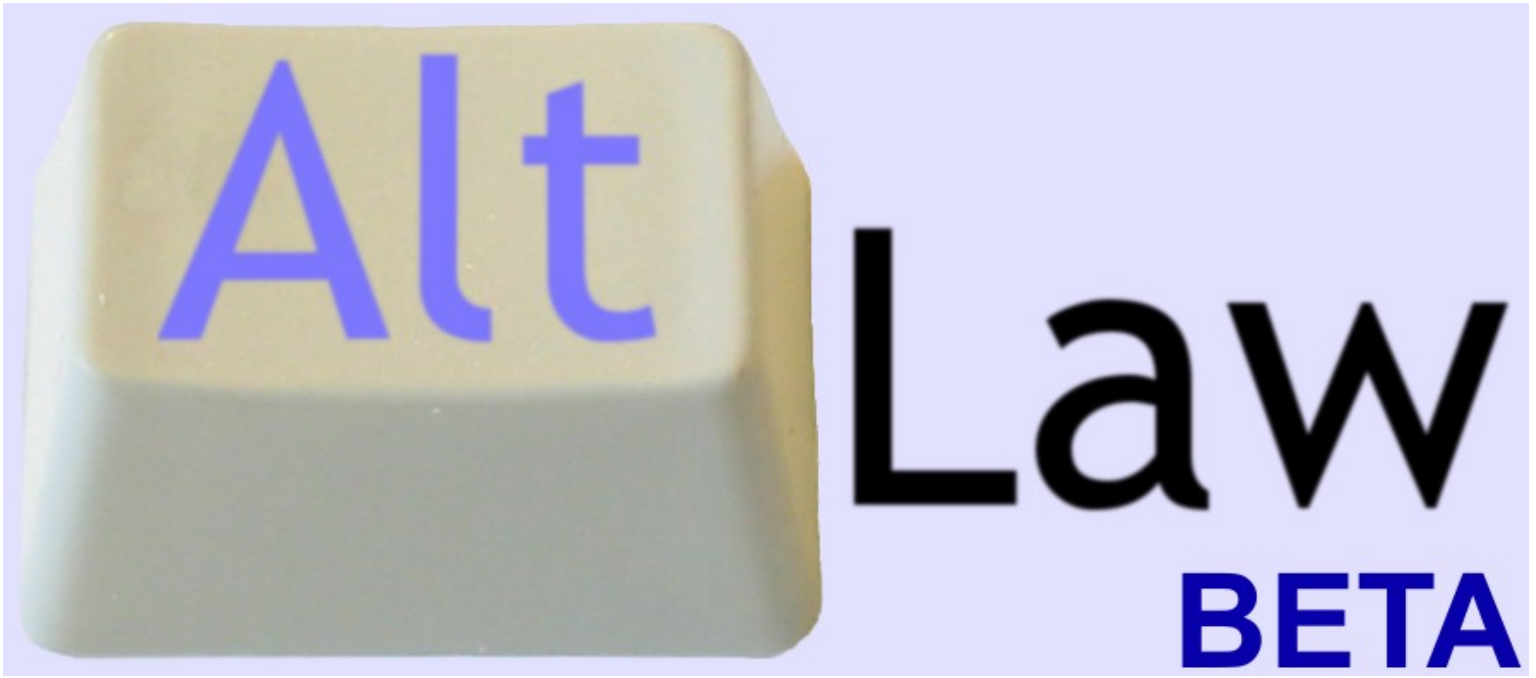
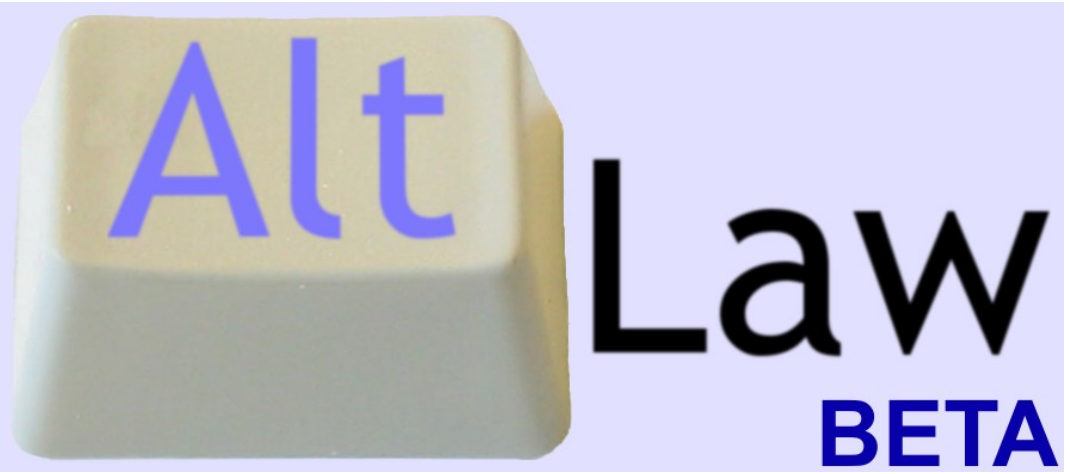


Stuart Sierra
Program on Law & Technology
Columbia Law School

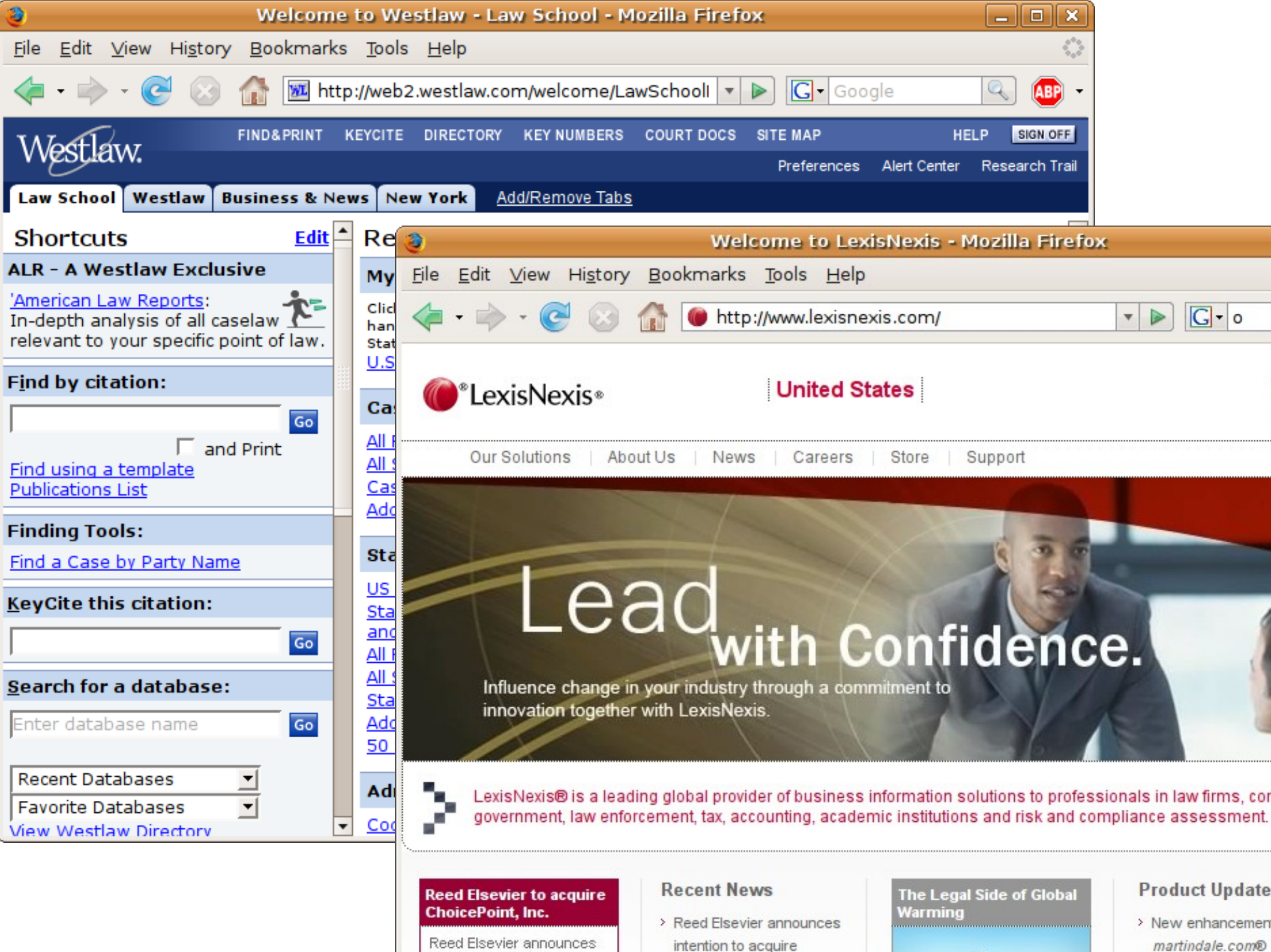
http://altlaw.org/ - the site
http://lawcommons.org/ - wiki & mailing list
http://columbialawtech.org/ - my employer

Talking Points

- AltLaw
  - History, motivation
  - Data sources
  - Back-end



- Semantic Web
  - What I've done
  - What I want
  - Problems I see


Welcome to Westlaw - Law School - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://web2.westlaw.com/welcome/LawSchooll
Google
Westlaw.
FIND&PRINT KEYCITE DIRECTORY KEY NUMBERS COURT DOCS SITE MAP HELP SIGN OFF
Preferences Alert Center Research Trail
Law School Westlaw Business & News New York Add/Remove Tabs
Shortcuts Edit
ALR - A Westlaw Exclusive
'American Law Reports:
In-depth analysis of all caselaw relevant to your specific point of law.
Find by citation:
Go
and Print
Find using a template
Publications List
Finding Tools:
Find a Case by Party Name
KeyCite this citation:
Go
Search for a database:
Enter database name
Go
Recent Databases
Favorite Databases
View Westlaw Directory
Welcome to LexisNexis - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.lexisnexis.com/
LexisNexis®
United States
Our Solutions | About Us | News | Careers | Store | Support
Lead with Confidence.
Influence change in your industry through a commitment to innovation together with LexisNexis.
LexisNexis® is a leading global provider of business information solutions to professionals in law firms, cor
government, law enforcement, tax, accounting, academic institutions and risk and compliance assessment.
Reed Elsevier to acquire ChoicePoint, Inc.
Reed Elsevier announces
Recent News
Reed Elsevier announces intention to acquire
The Legal Side of Global Warming
Product Update
New enhancemen
martindale.com®

AltLaw

BETA

The free legal search engine — over 700,000 documents.

Enter a case name, citation, or key words and phrases:

search cases | search codes

About AltLaw | Advanced Search | Coverage

Browse Cases | Browse U.S. Code

Front-end

Data Sources – Large Corpora

- Paul Ohm's corpus, http://bulk.altlaw.org/
  - 7 GB, 200,000+ files harvested from court web sites
- Cornell U.S. Code
  - 748 MB of XML
- http://bulk.resource.org/courts.gov/c/
  - 2 GB, 700,000+ federal cases, XHTML
- http://pacer.resource.org/
  - 736 GB, 2.7 million PDFs, 1.8 million HTML files

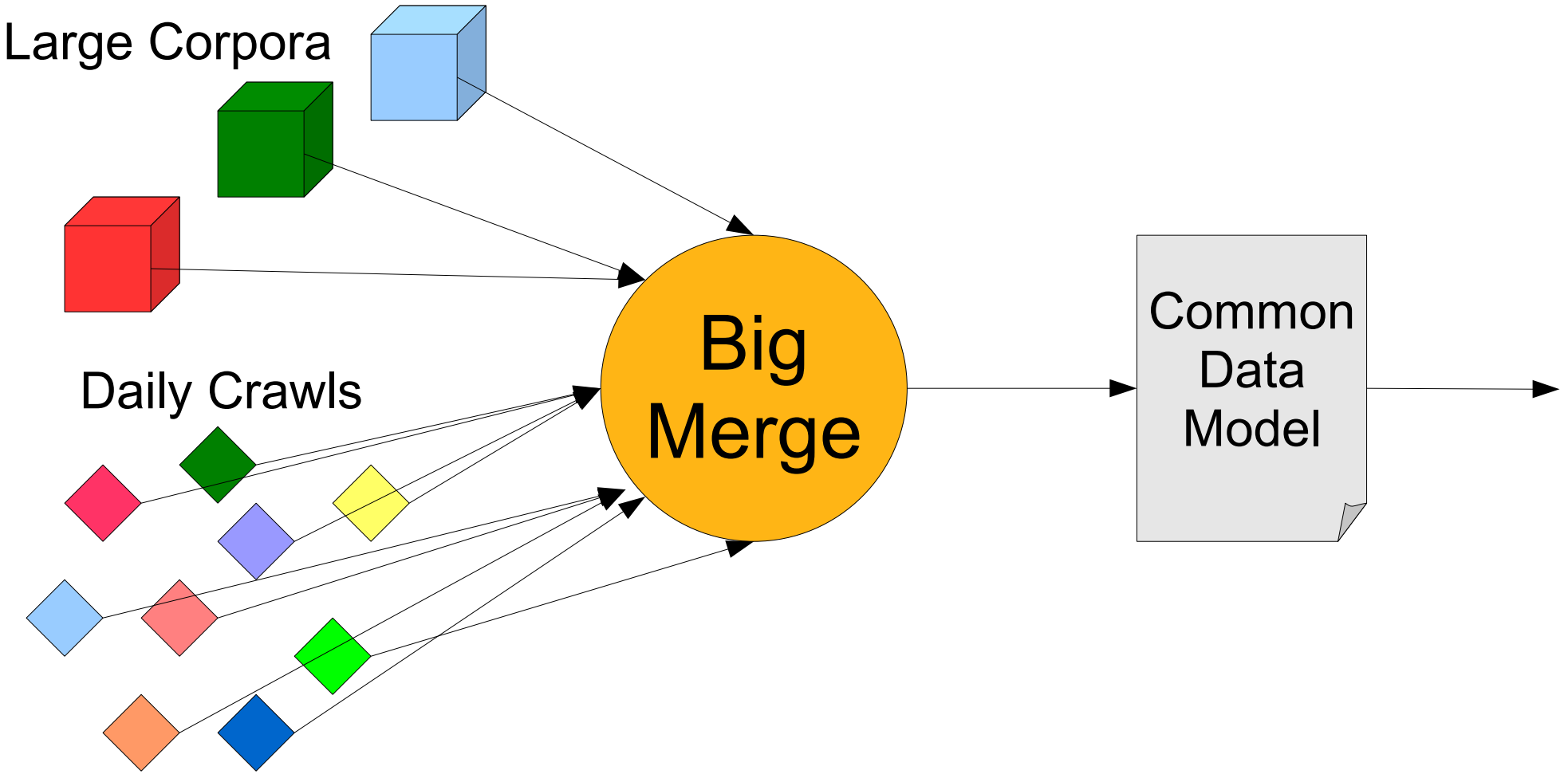
Data Sources – Court Web Sites

www.supremecourtus.gov
www.ca1.uscourts.gov
www.ca2.uscourts.gov
www.ca3.uscourts.gov
www.ca4.uscourts.gov
www.ca5.uscourts.gov
www.ca6.uscourts.gov
. . .

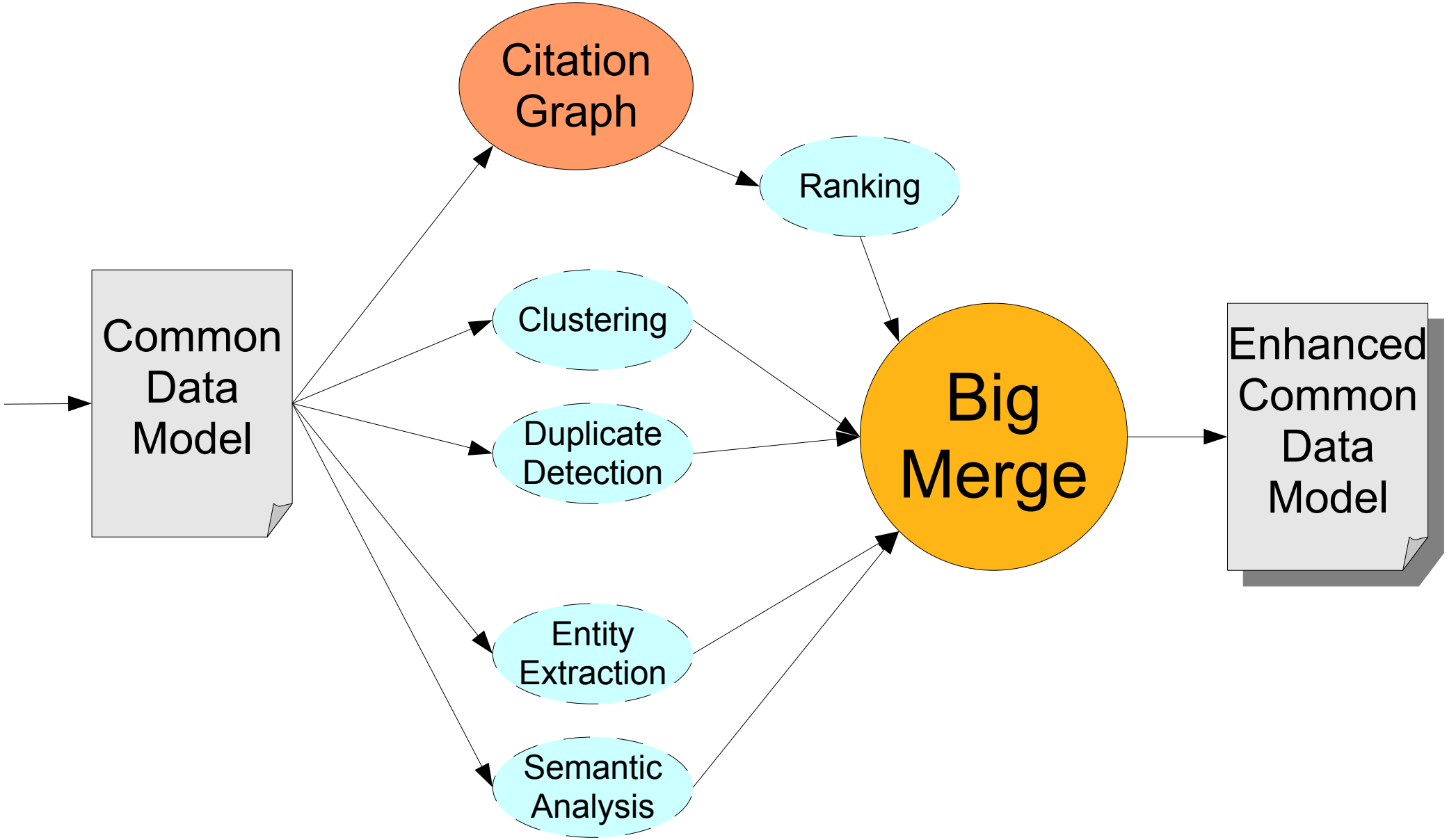
14 appeals courts total
94 district courts
?? state courts
?? local/other courts

- 20-40 new cases daily
- PDF, WordPerfect, HTML, plain text

Back-end (1)

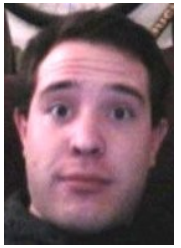


Back-end (2)


Citation Graph
Ranking
Common Data Model
Clustering
Duplicate Detection
Big Merge
Enhanced Common Data Model
Entity Extraction
Semantic Analysis

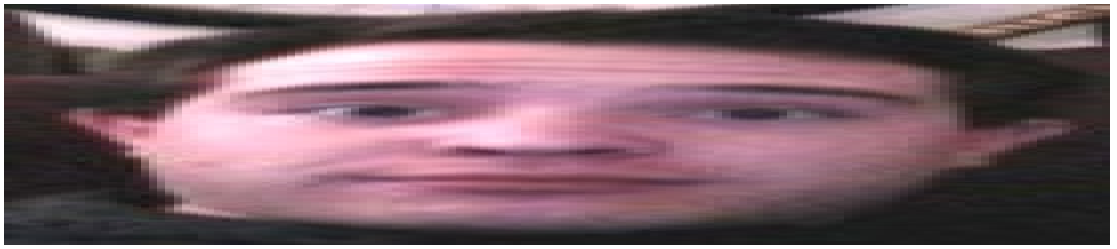
Scaling Stuart

- Java



-
- Ruby

-
- Clojure



The Grand Unified Data Model

- Key-value pairs? (files, Berkeley DB)
- Documents? (Solr/Lucene, CouchDB)
- Trees? (XML, JSON, Objects)
- Graphs? (RDF)
- Tables? (SQL)

- “Disk is the new tape.”
  - NO random access
  - NO disk seeks
  - Run at full disk transfer rate, not seek rate
- Data must be splittable
- Process each record in isolation

Secret Weapons

- Hadoop – open-source MapReduce
- Amazon EC2 – cluster by the hour
- Clojure – Lisp on the JVM
- Solr – full-text search +
  document storage;
  no SQL database!
- Ruby on Rails

The Grand Unified Data Model

- Key-value pairs? (files, Berkeley DB)
- Documents? (Solr/Lucene, CouchDB)
- Trees? (XML, JSON, Objects)
- Graphs? (RDF)
- Tables? (SQL)

Mismatch

- Hadoop
  - Disk is the new tape
  - Flat key/value files
  - Isolated records
- Solr / Lucene
  - Denormalized
  - Flat documents

- RDF
  - Normalized
  - Random access
  - Graph structure
  - Linked records

Semantic Web – What I Want

- Publish linked data for others
- Accept new data without writing new parsers/scrapers
- Richer internal data model
- Inference over multiple data sources

AltLaw on the Semantic Web

- Persistent URIs for federal courts
  - e.g. http://id.altlaw.org/courts/us/fed/app/3
  - 303 redirects to HTML/RDF
- Beginnings of an ontology
  - http://github.com/lawcommons/altlaw-vocab
  - Extension of Dublin Core & Bibliontology
- Semantic web crawler
  - Output uses “HTTP Vocabulary in RDF”

Questions

- What's in it for you?
  - How do you want my data?
    - Bulk RDF/XML downloads
    - RDFa embedded in HTML
    - SPARQL endpoint
  - What would you do with it?
- What's in it for me?
  - Universal data model
  - Less data transformation