
Chris A. Mattmann, NASA JPL,
USC & the ASF
@chrismattmann
mattmann@apache.org
APACHECON
NORTH AMERICA

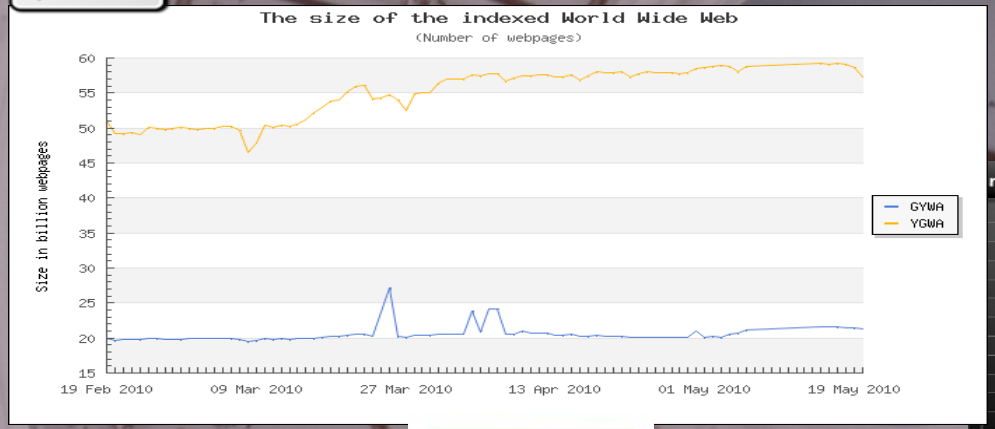
Content Extraction from Images and Video in Tika

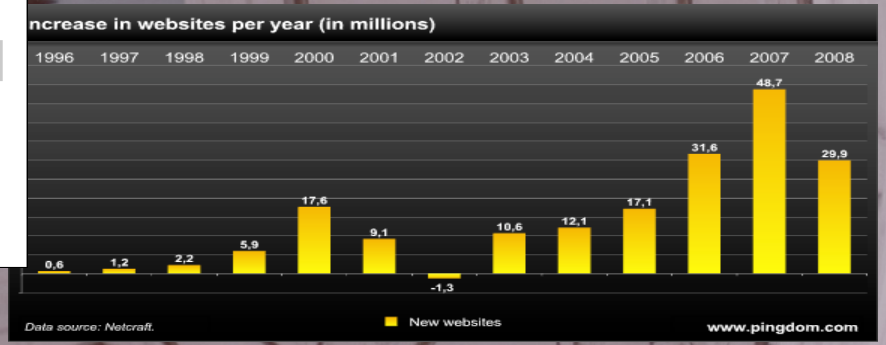

Background: Apache Tika
APACHECON
NORTH AMERICA

Outline

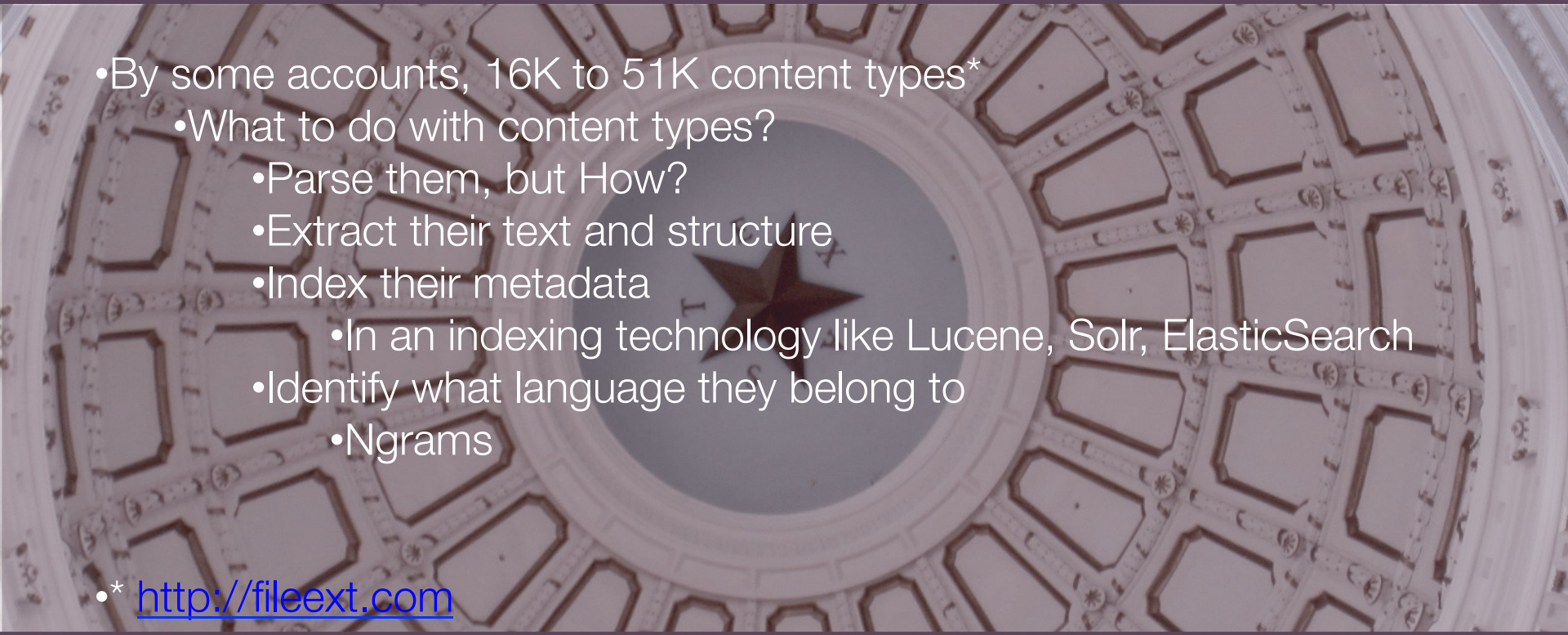
- Text
- The Information Landscape
- The Importance of Content Detection and Analysis
- Intro to Apache Tika

The Information Landscape





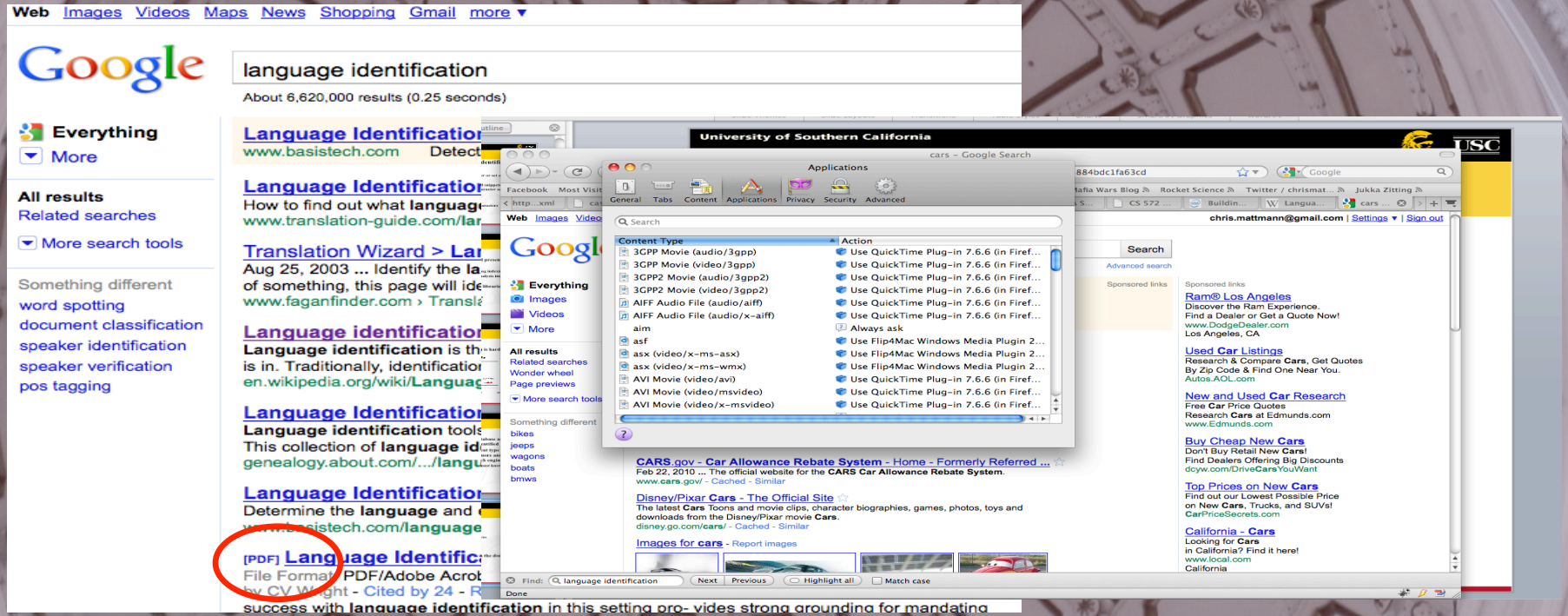
Proliferation of Content Types



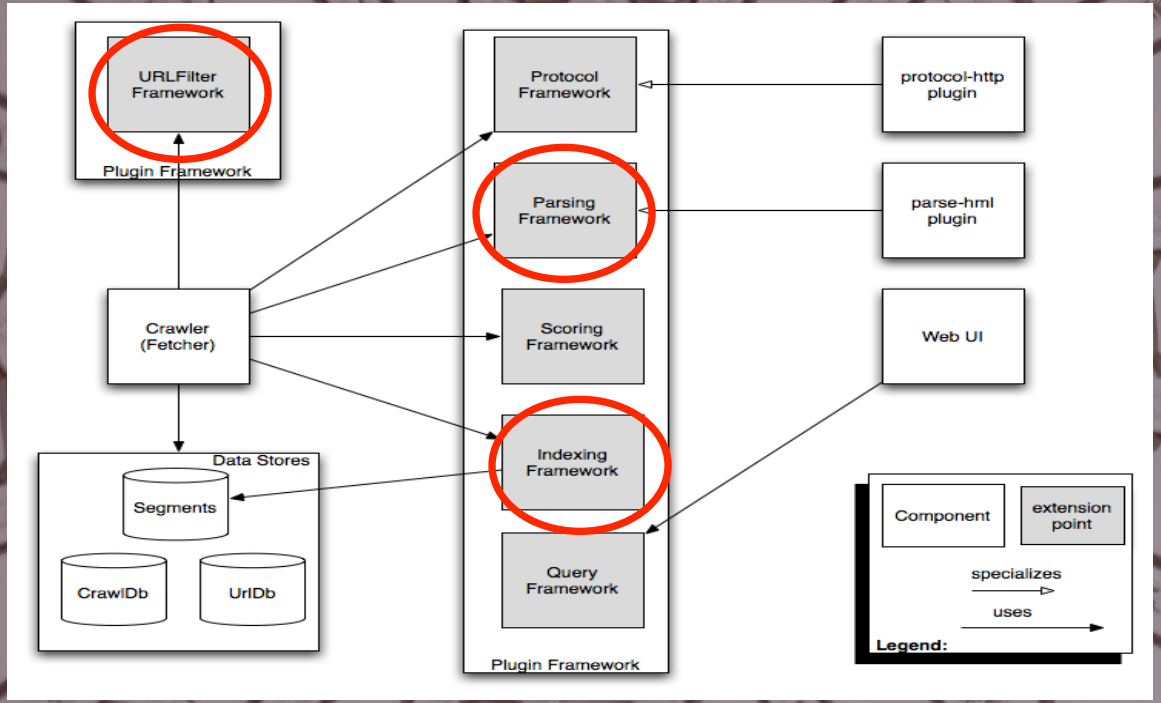
- By some accounts, 16K to 51K content types*
  - What to do with content types?
    - Parse them, but How?
    - Extract their text and structure
    - Index their metadata
      - In an indexing technology like Lucene, Solr, ElasticSearch
    - Identify what language they belong to
      - Ngrams

- * http://fileext.com

Importance: Content Types

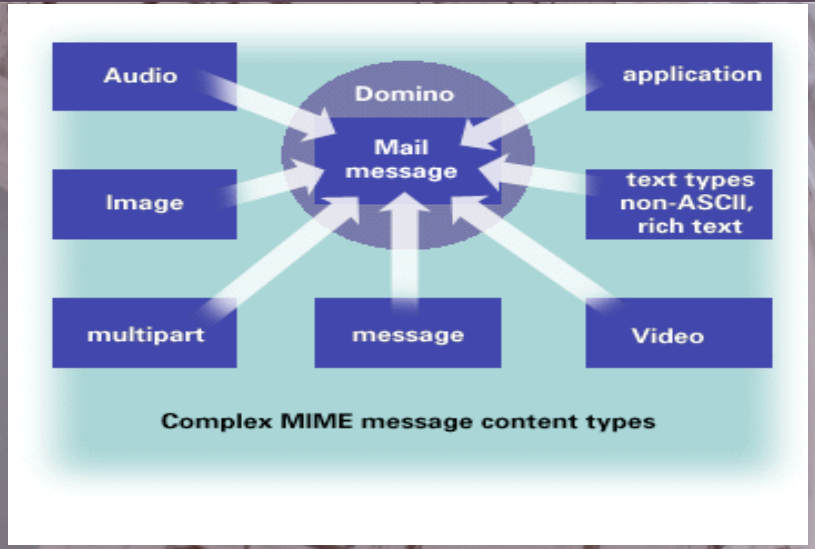


Importance: Content Types



IANA MIME Registry

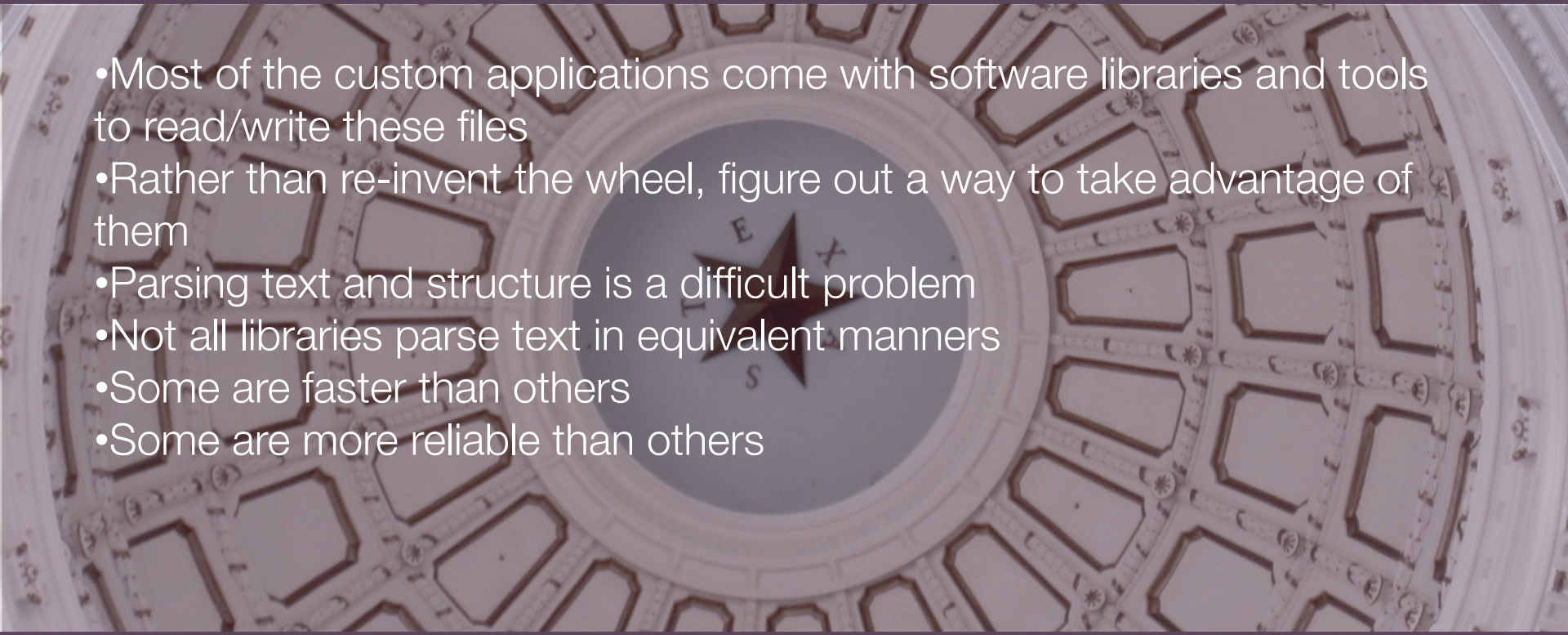
- Identify and classify file types
  - MIME detection
  - Glob pattern
    - *.txt
    - *.pdf
  - URL
    - http://…pdf
    - ftp://myfile.txt
  - Magic bytes
  - Combination of the above means
- Classification means reaction can be targeted



Many Custom Applications

- You need these apps to parse these files
  - …and that's what Tika exploits

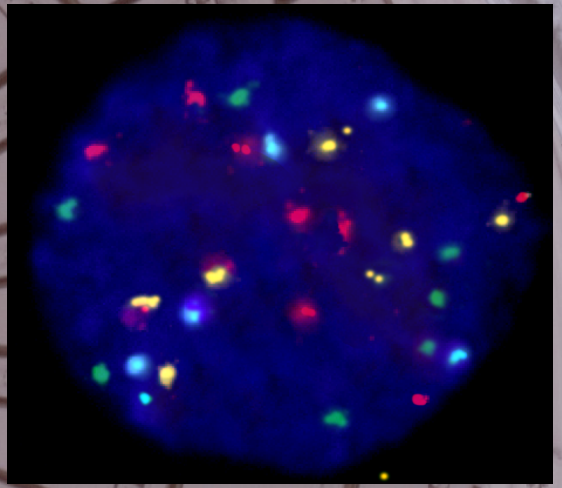
Third Party Parsing Libraries



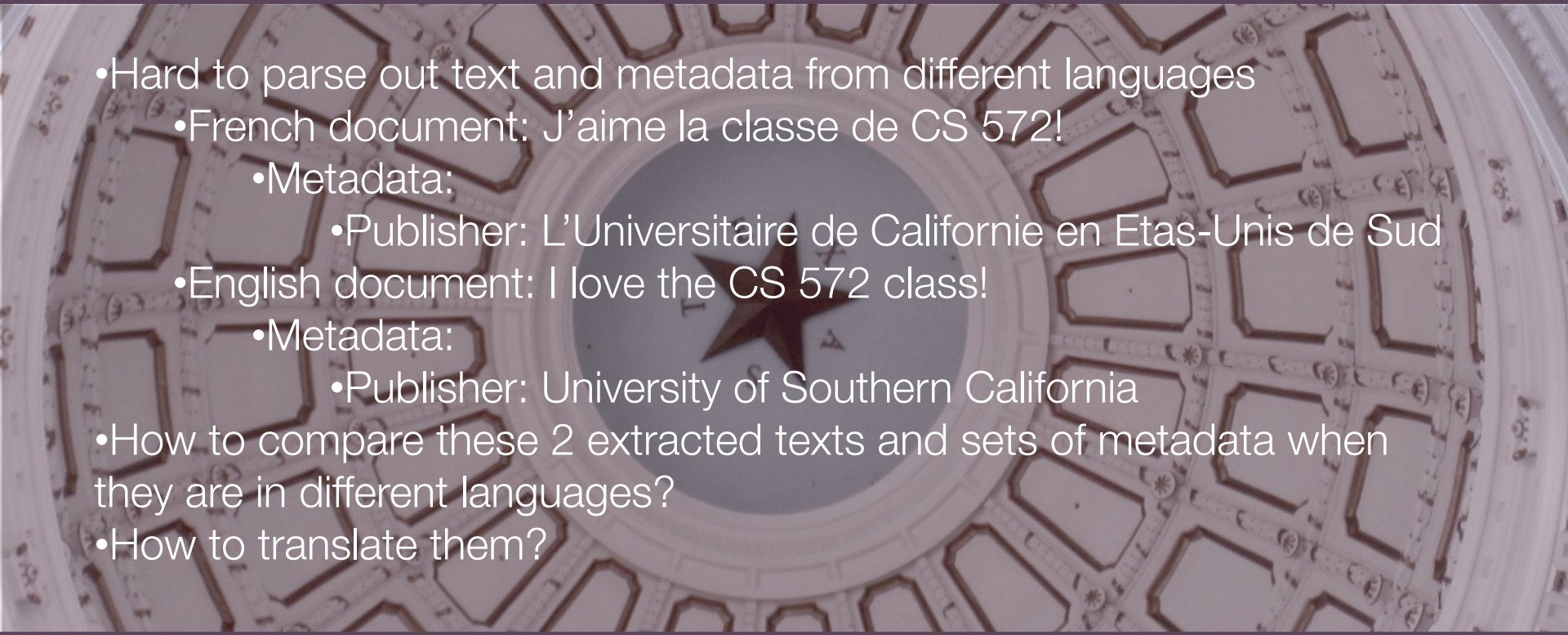
- Most of the custom applications come with software libraries and tools to read/write these files
- Rather than re-invent the wheel, figure out a way to take advantage of them
- Parsing text and structure is a difficult problem
- Not all libraries parse text in equivalent manners
- Some are faster than others
- Some are more reliable than others

Extraction of Metadata

- Important to follow common Metadata models
  - Dublin Core
  - Word Metadata
  - XMP
  - *EXIF*
  - Lots of standards and models out there
    - The use and extraction of common models allows for content intercomparison
  - All standardizes mechanisms for searching
  - You always know for X file type that field Y is there and of type String or Int or Date



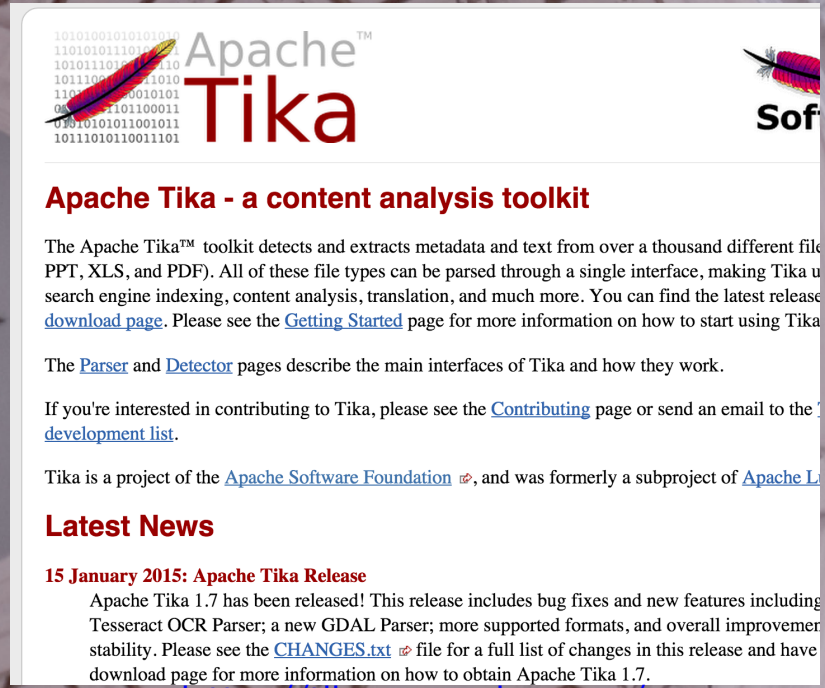
Lang. Identification/Translation



- Hard to parse out text and metadata from different languages
  - French document: J'aime la classe de CS 572!
    - Metadata:
      - Publisher: L'Universitaire de Californie en Etas-Unis de Sud
  - English document: I love the CS 572 class!
    - Metadata:
      - Publisher: University of Southern California
- How to compare these 2 extracted texts and sets of metadata when they are in different languages?
- How to translate them?

Apache Tika

- A content analysis and detection toolkit
- A set of Java APIs providing MIME type detection, language identification, integration of various parsing libraries
- A rich Metadata API for representing different Metadata models
- A command line interface to the underlying Java code
- A GUI interface to the Java code
- Translation API
- REST server
- Ports to NodeJS, Python, PHP, etc.



http://tika.apache.org/

Tika's History

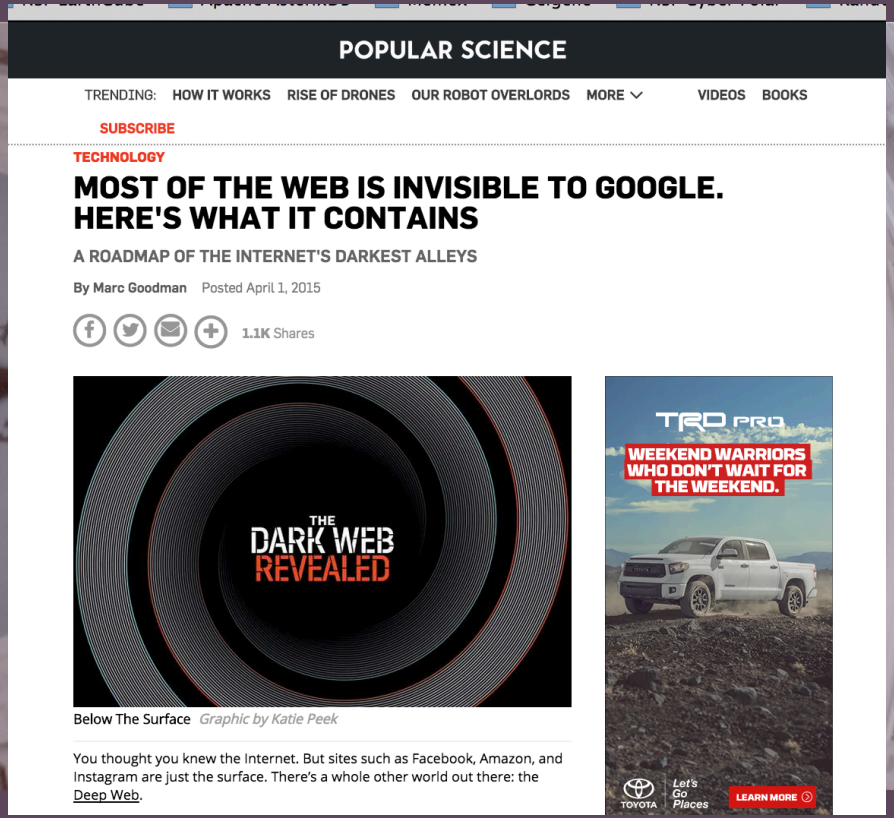
- Original idea for Tika came from Chris Mattmann and Jerome Charron in 2006
- Proposed as Lucene sub-project
- Others interested, didn't gain much traction
- Went the Incubator route in 2007 when Jukka Zitting found that there was a need for Tika capabilities in Apache Jackrabbit
- A Content Management System
- Graduated from the Incubator to Lucene sub-project in 2008
- Graduated to Apache TLP in 2010
- Many releases since then, currently VOTE'ing on 1.8


Images and Video
APACHECON
NORTH AMERICA

The Dark Web

- The web behind forms
- The web behind Ajax/Javascript
- The web behind heterogeneous content types

- Examples
  - Human and Arms Trafficking
    - Tor Network
  - Polar Sciences
    - Cryosphere data in archives
- DARPA Memex / NSF Polar Cyber Infrastructure



POPULAR SCIENCE

TRENDING: HOW IT WORKS RISE OF DRONES OUR ROBOT OVERLORDS MORE VIDEOS BOOKS

SUBSCRIBE

TECHNOLOGY

MOST OF THE WEB IS INVISIBLE TO GOOGLE. HERE'S WHAT IT CONTAINS

A ROADMAP OF THE INTERNET'S DARKEST ALLEYS

By Marc Goodman Posted April 1, 2015

1.1K Shares

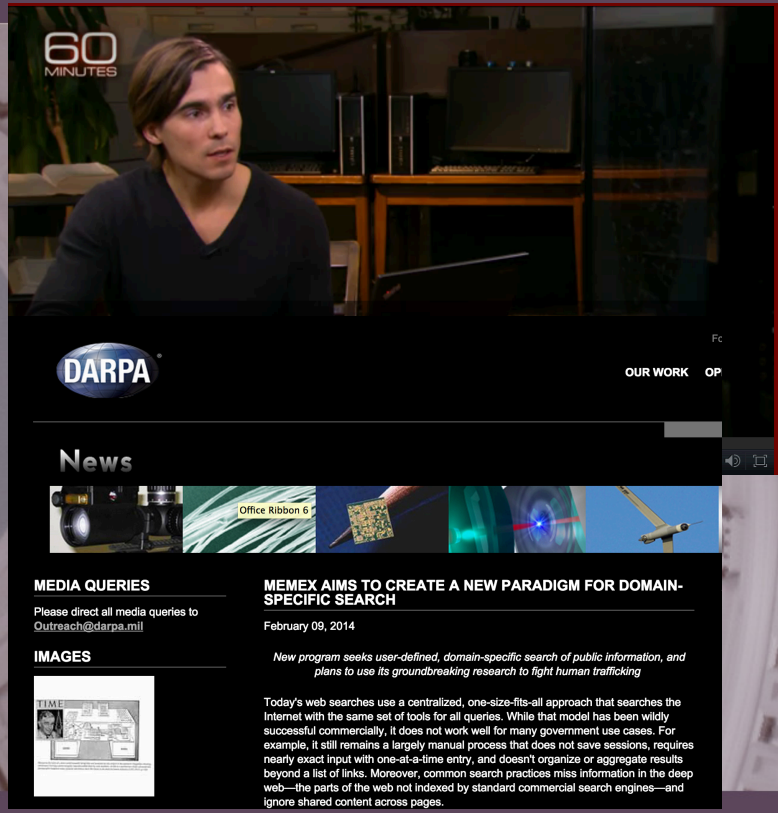
Below The Surface *Graphic by Katie Peek*

You thought you knew the Internet. But sites such as Facebook, Amazon, and Instagram are just the surface. There's a whole other world out there: the Deep Web.

http://www.popsci.com/dark-web-revealed

DARPA Memex Project

- Crawl, analyze, reason, and decide about the dark web
- 17+ performers
- JPL is a performer based on the Apache stack of Search Engines technologies
  - Apache Tika, Nutch Solr



DARPA Memex Project

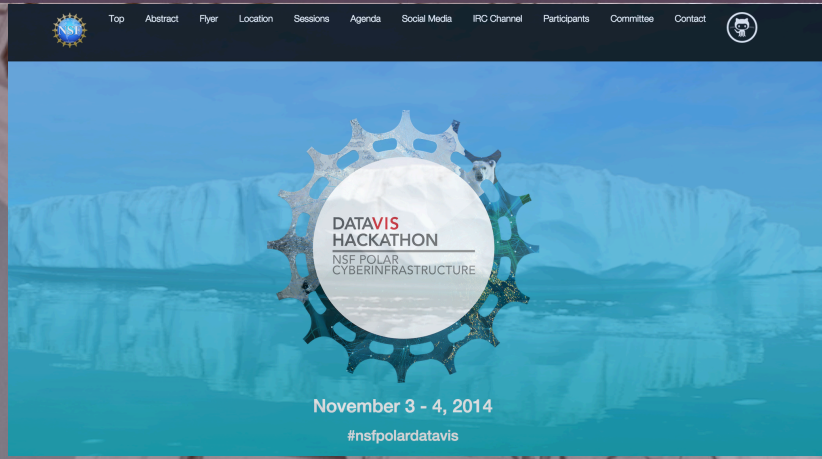
APACHECON NORTH AMERICA

- 60 Minutes (February 8, 2015)
  - DARPA: Nobody's Safe On The Internet News:
    - http://www.cbsnews.com/news/darpa-dan-kaufman-internet-security-60-minutes/
    - http://www.cbsnews.com/videos/darpa-nobodys-safe-on-the-internet
  - 60 Minutes Overtime (February 8, 2015)
    - New Search Engine Exposes The "Dark Web"
      - http://www.cbsnews.com/news/darpa-dan-kaufman-internet-security-60-minutes/
      - http://www.cbsnews.com/videos/new-search-engine-exposes-the-dark-web
- Scientific American (February 8, 2015)
  - Human Traffickers Caught on Hidden Internethttp://www.scientificamerican.com/article/human-traffickers-caught-on-hidden-internet/
  - Scientific American Exclusive: DARPA Memex Data Maps
  - http://www.scientificamerican.com/slideshow/scientific-american-exclusive-darpa-memex-data-maps/

NSF Polar CyberInfrastructure

APACHECON NORTH AMERICA

- 2 specific projects
  - http://www.nsf.gov/awardsearch/showAward?AWD_ID=1348450&HistoricalAwards=false
  - http://www.nsf.gov/awardsearch/showAward?AWD_ID=1445624&HistoricalAwards=false
- I call this my "Polar Memex"
  - Crawling NSF ACADIS, Arctic Data Explorer and NASA AMD
  - Exposing geospatial and temporal content types (ISO 19115; GCMD DIF; GeoTopic Identification; GDAL)
  - Exposing Images and Video

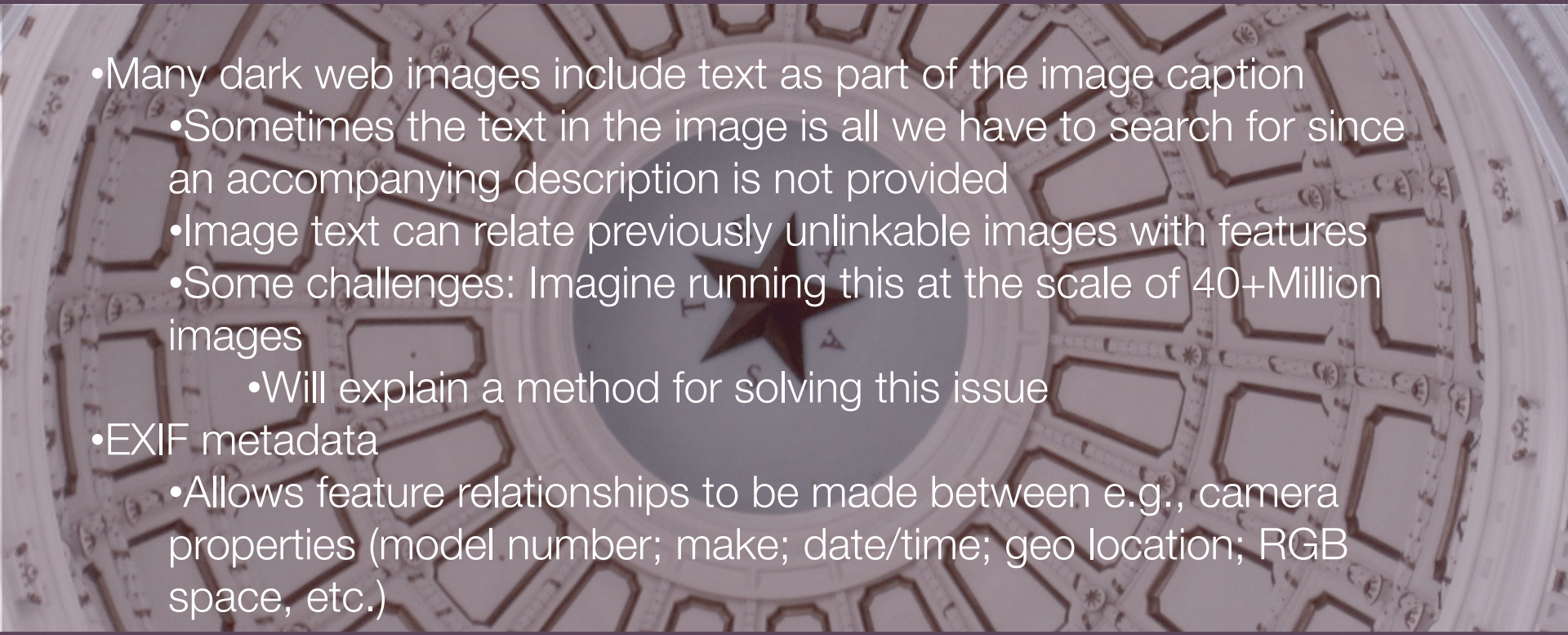


- http://nsf-polar-cyberinfrastructure.github.io/datavis-hackathon/

Specific improvements

- Tika doesn't natively handle images and video even though it's used in crawling the web
  - Improve two specific areas
  - Optical Character Recognition (OCR)
  - EXIF metadata extraction
    - Why are these important for images and video?
  - Geospatial parsing
    - Geo reference data that isn't geo referenced (will talk about this later)

OCR and EXIF



- Many dark web images include text as part of the image caption
  - Sometimes the text in the image is all we have to search for since an accompanying description is not provided
  - Image text can relate previously unlinkable images with features
  - Some challenges: Imagine running this at the scale of 40+Million images
    - Will explain a method for solving this issue
- EXIF metadata
  - Allows feature relationships to be made between e.g., camera properties (model number; make; date/time; geo location; RGB space, etc.)

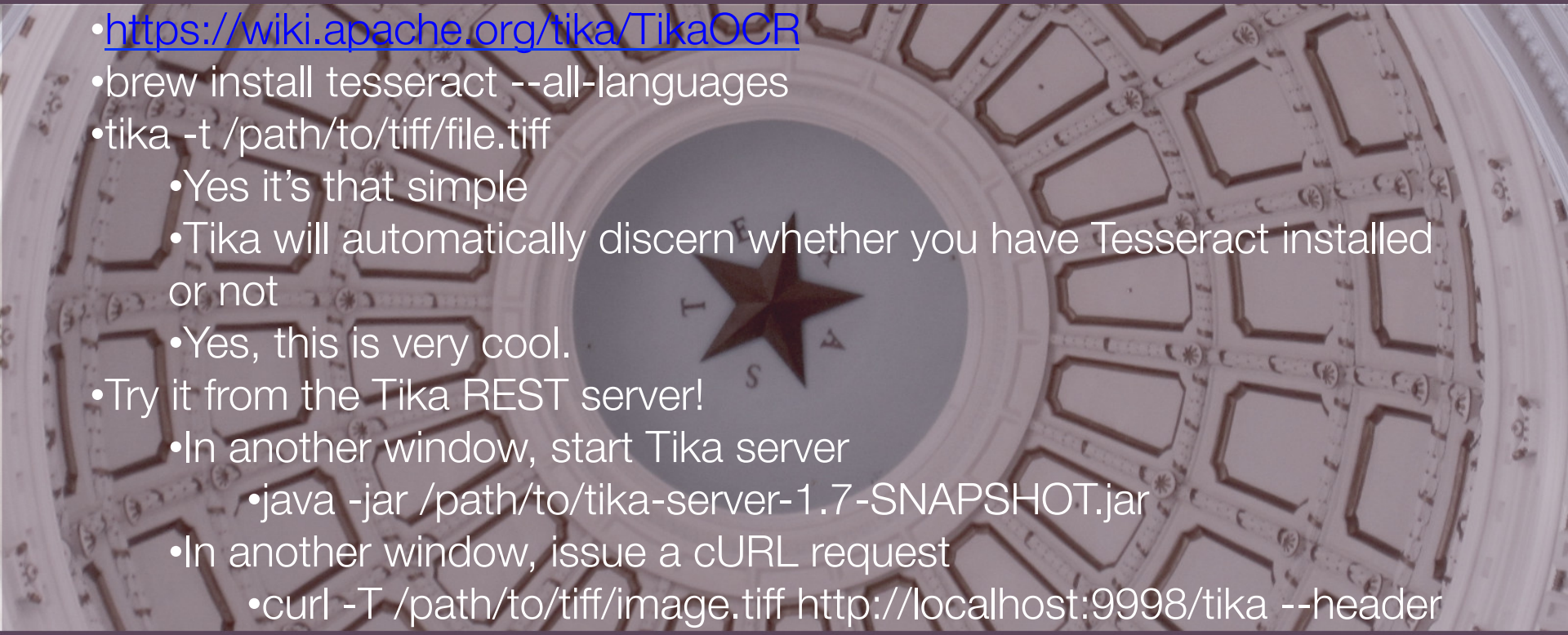
Enter Tesseract

- https://code.google.com/p/tesseract-ocr/

- Great and Accurate Toolkit, Apache License, version 2 (“ALv2”)
- Many recent improvements by Google and Support for Multiple Languages

- Integrate this with Tika!
  - http://issues.apache.org/jira/browse/TIKA-93
  - Thank you to Grant Ingersoll (original patch) and Tyler Palsulich for taking the work the rest of the way to get it contributed

Tika + Tesseract In Action



- https://wiki.apache.org/tika/TikaOCR
- brew install tesseract --all-languages
- tika -t /path/to/tiff/file.tiff
  - Yes it's that simple
  - Tika will automatically discern whether you have Tesseract installed or not
  - Yes, this is very cool.
- Try it from the Tika REST server!
  - In another window, start Tika server
    - java -jar /path/to/tika-server-1.7-SNAPSHOT.jar
  - In another window, issue a cURL request
    - curl -T /path/to/tiff/image.tiff http://localhost:9998/tika --header "Content-type: image/tiff"

Tesseract – Try it out

EXIF metadata

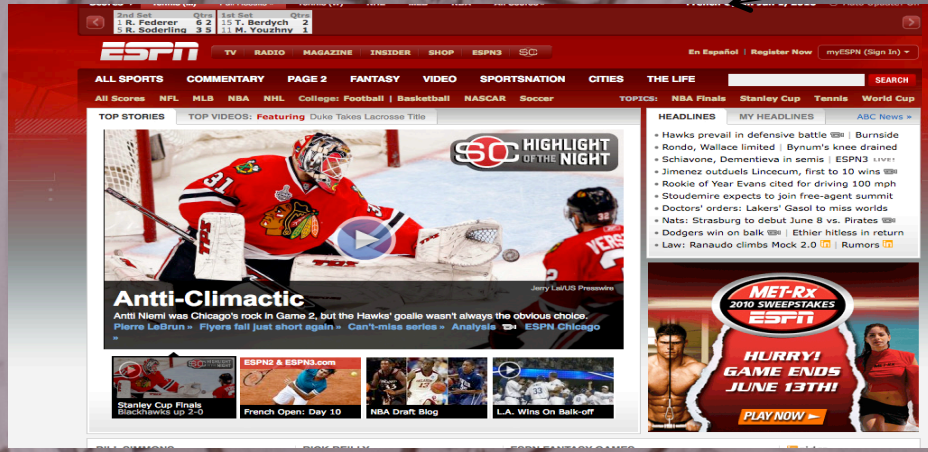
- Example EXIF metadata
  - Camera Settings; Scene Capture Type; White Balance Mode; Flash; Fnumber (Fstop); File Source; Exposure Mode; Xresolution; Yresolution; Recommended EXIF interoperability Rules, Thumbnail compression; Image Height; Image Width; Flash Output; AF Area Height; Model; Model Serial Number; Shooting Mode; Exposure Compensation..
    - AND MANY MORE
- These represent a “feature space” that can be used to relate images, *even without looking directly at the image*
  - Will speak about this over the next few slides

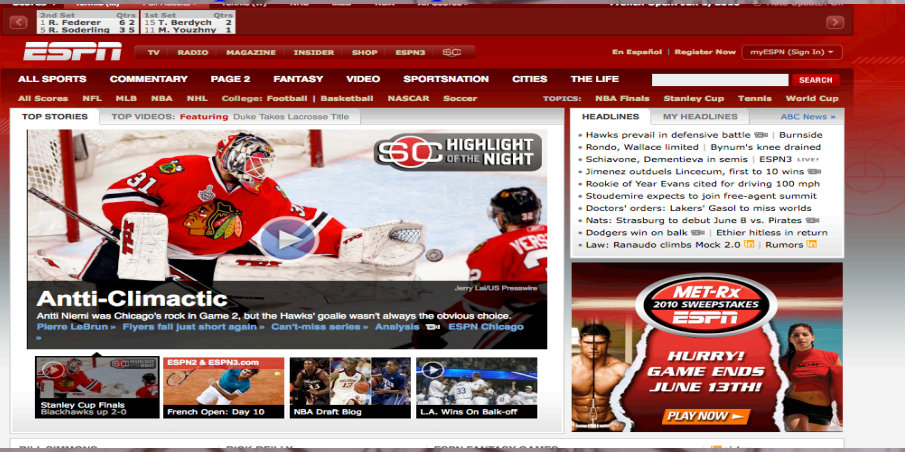
What are web duplicates?

- One example is the same page, referenced by different URLs

/ http://espn.go.com

http://www.espn.com

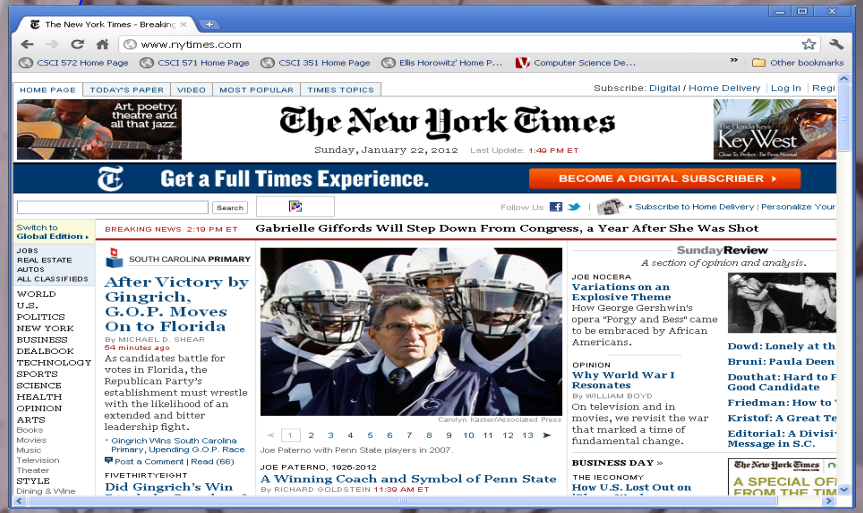


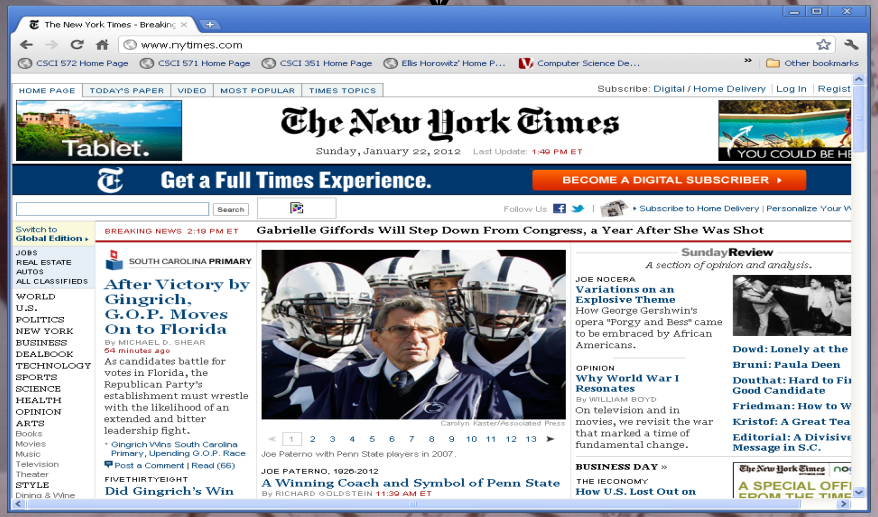


- How can two URLs differ yet still point to the same page?
  - the URL's host name can be distinct (virtual hosts),
  - the URL's protocol can be distinct (http, https),
  - the URL's path and/or page name can be distinct

What are web duplicates?

- Another example is two web pages whose content differs slightly





- Two copies of www.nytimes.com snapshot within a few seconds of each other;
- The pages are essentially identical except for the ads to the left and right of the banner line that says The New York Times;

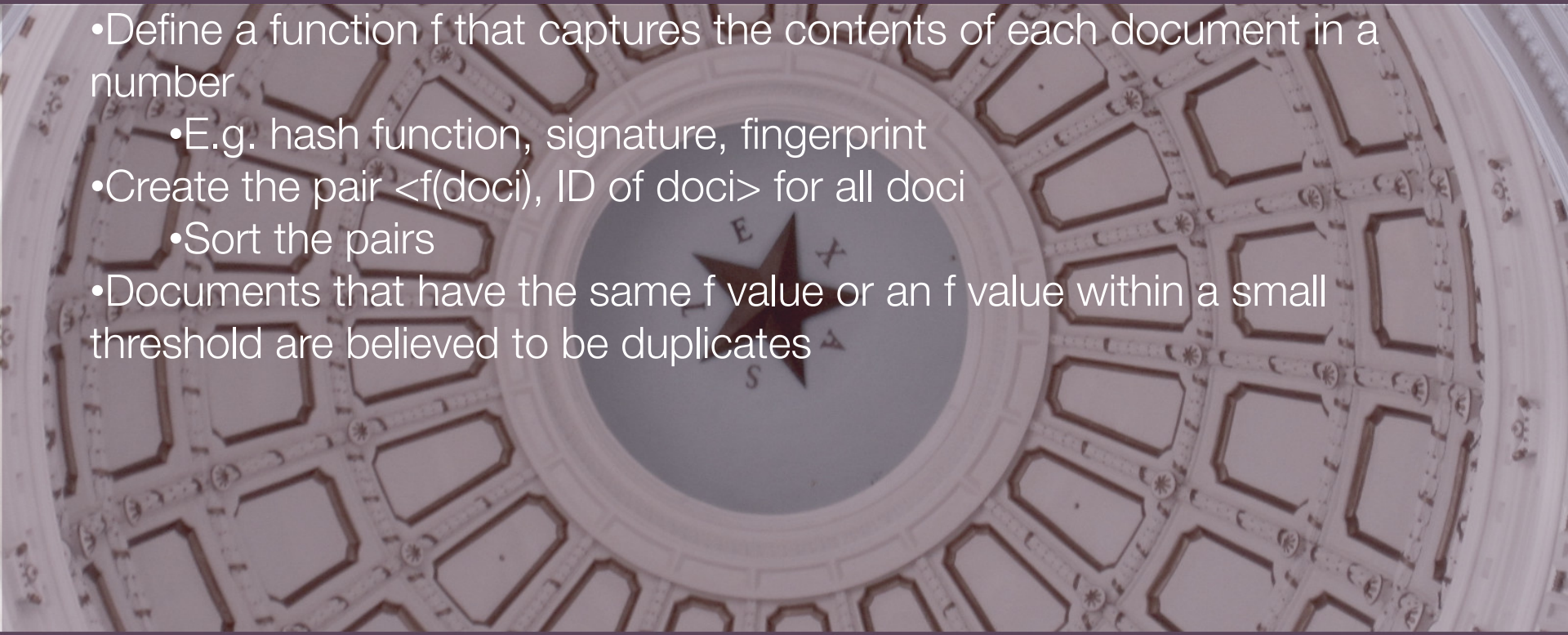
Solving (near) Duplicates

- Duplicate: Exact match;
  - Solution: compute fingerprints or use cryptographic hashing
  - SHA-1 and MD5 are the two most popular cryptographic hashing methods
- Near-Duplicate: Approximate match
  - Solution: compute the syntactic similarity with an edit-distance measure, and
  - Use a similarity threshold to detect near-duplicates
  - e.g., Similarity > 80% => Documents are “near duplicates”

Identifying Identical Documents

- Compare character by character two documents to see if they are identical
  - However, this could be very time consuming if we must test every possible pair
- We might hash just the first few characters and compare only those documents that hash to the same bucket
  - But what about web pages where every page begins with <HTML>
- Another approach would be to use a hash function that examines the entire document
  - But this requires lots of buckets
- A better approach is to pick some fixed random positions for all documents and make the hash function depend only on these;
  - This avoids the problem of a common prefix for all or most documents, yet we need not examine entire documents unless they fall into a bucket with another document
  - But we still need a lot of buckets

General Paradigm: Similarity



- Define a function f that captures the contents of each document in a number
  - E.g. hash function, signature, fingerprint
- Create the pair <f(doci), ID of doci> for all doci
  - Sort the pairs
- Documents that have the same f value or an f value within a small threshold are believed to be duplicates

Distance Measures

- Distance measure must satisfy 4 properties
  - No negative distances
  - D(x,y) = 0 iff x=y
  - D(x,y) = d(y,x) symmetric
  - D(x,y) <= d(x,z) + d(z,y) triangle inequality
- There are several distance measures that can play a role in locating duplicate and near-duplicate documents
  - Euclidean distance – d([x1…xn], [y1,…,yn]) = sqrt(Sum(xi-yi)^2) i=1…n
  - **Jaccard distance – d(x,y) = 1 – SIM(x,y) or 1 minus the ratio of the sizes of the intersection and union of sets x and y**
  - Cosine distance – the cosine distance between two points (two n element vectors) is the angle that the vectors to those points make; in the range 0 to 180 degrees
  - Edit distance – the distance between two strings is the smallest number of insertions and deletions of single characters that will convert one string into the other
  - Hamming distance – between two vectors is the number of components in which they differ (usually used on boolean vectors)

Jaccard Similarity

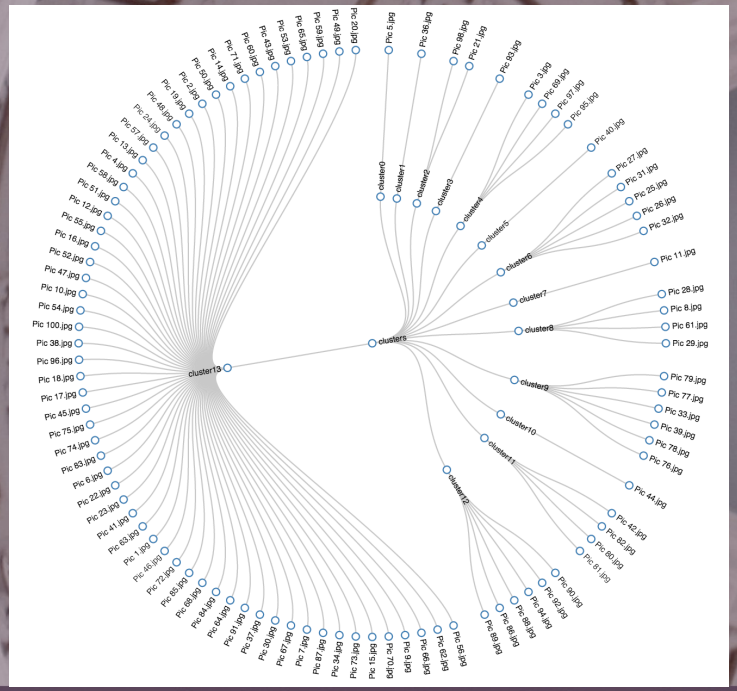
- Similarity Measures
- Resemblance(A,B) is defined as
  - size of (S(A,w) intersect S(B,w)) / size of (S(A,w) union S(B,w))
- Containment(A,B) is defined as
  - size of (S(A,w) intersect S(B,w)) / size of (S(A,w))
- 0 <= Resemblance <= 1
- 0 <= Containment <= 1
- EXIF metadata can be treated as “FEATURES” that you can compute containment and resemblance.

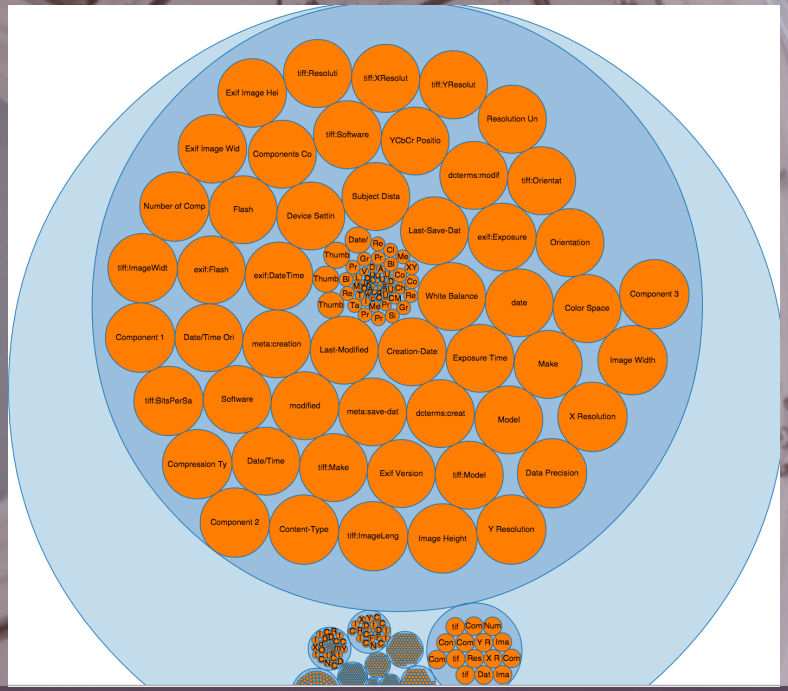
Tika Image Similarity

- http://github.com/chrismattmann/tika-img-similarity/
- First pass it a directory e.g., of Images
  - For each file (image) in the directory
    - Run Tika, extract EXIF features
    - Add all unique features to “golden feature set”
  - Loop again
    - Use extracted EXIF metadata for file, compute size of feature set, and names, compute containment which is a “distance” of each document to the golden feature set
  - Set a threshold on distance, then you have clusters

Tika Image Similarity

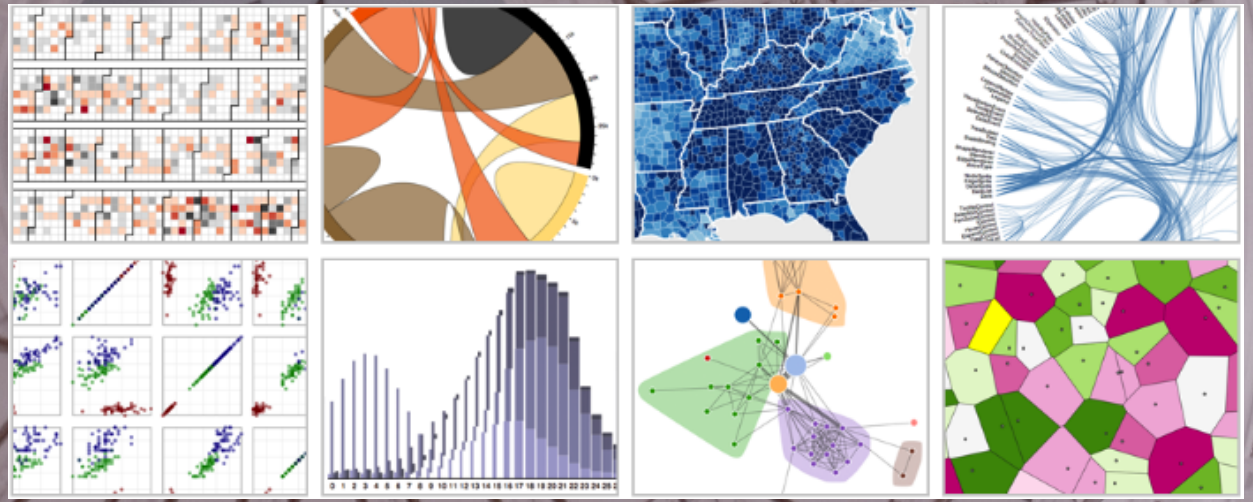
- Results are extremely promising





Wait, DataViz??!

- http://d3js.org/
- Invented by Mike Bostock and Vadim Ogievetsky and Jeff Heer http://vis.stanford.edu/papers/d3



Wait, DataViz??!

- Creates SVG tied to DOM aspects of the page
- Page loads e.g., data, (JSON or other), controls access via DOM
- Manipulate DOM and bind DOM to SVG elements
- Tons of examples

- https://github.com/mbostock/d3/wiki/Gallery

Demo

- Tika Image Similiarity

Image Similarity Viz

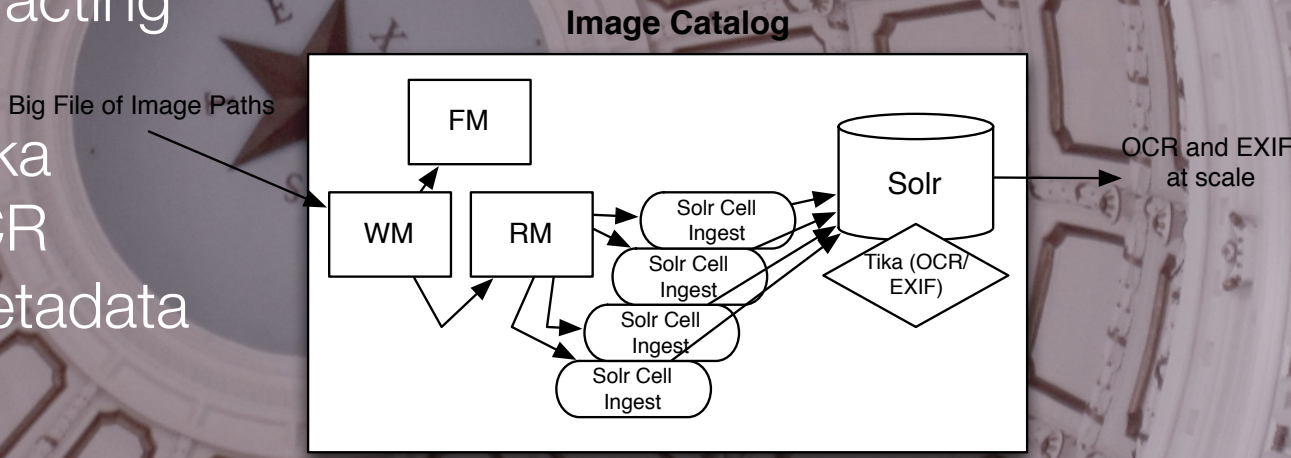
- Dendogram, flare dendogram
  - Excellent for showing cluster relationships as generated by tika-img-similarity
- Circle packing
  - What metadata features distinguish each cluster?
- Dynamic versions of each allow for interaction
- Future work
  - Integrating into Nutch administration GUI and allowing for Tika-based similarity and clustering
- The power of this approach: doesn't require Computer Vision

Image Catalog (“ImageCat”)

- OCR and EXIF metadata around images
  - Can handle similarity measures
  - Can allow for search of features in images based on text
  - Can relate images based on EXIF properties (all taken with flash on; all taken in same geographic region, etc.)
- How do you do this at the scale of the Internet
  - “Deep Web” as defined by DARPA in domain of e.g., human trafficking ~60M web pages, 40M images
- You use ImageCatalog, of course! ☺

Image Catalog (“ImageCat”)

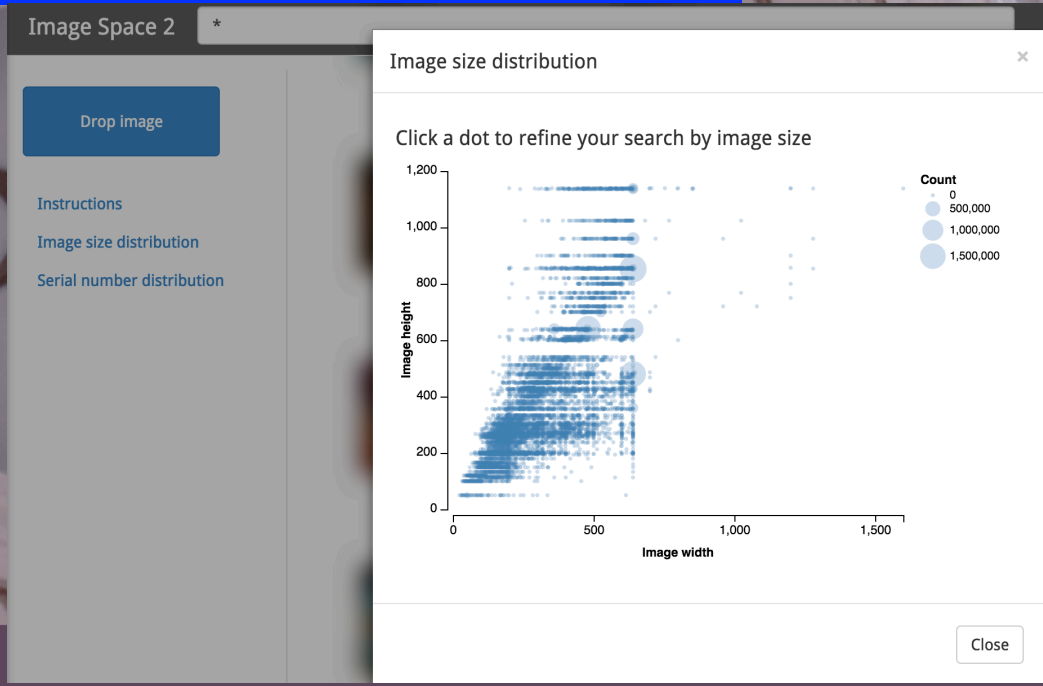
- Apache OODT – ETL, Map Reduce over LONG list of files
  - Partition files into 50k chunks
  - Ingest into Solr Extracting RequestHandler
- Apache Solr / Extracting RequestHandler
  - Augmented Tika + Tesseract OCR
  - Tika + EXIF metadata



- https://github.com/chrismattmann/imagecat/

ImageSpace

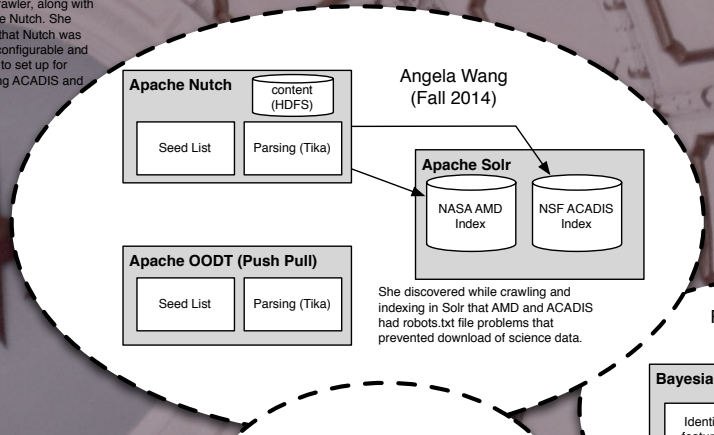
- With ImageCat you can build…Image Space
  - https://github.com/memex-explorer/image_space/
- Connect to ImageCat
  - Search on similar images
  - EXIF, Jaccard, computer vision based approaches
- Continuum Analytics + Kitware, Inc. + JPL
- Funded by DARPA Memex



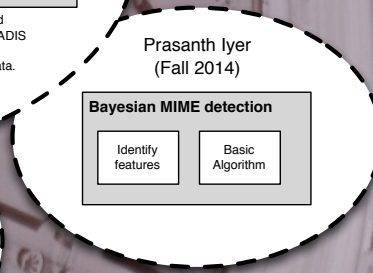
NSF Polar Work: Fall 2014

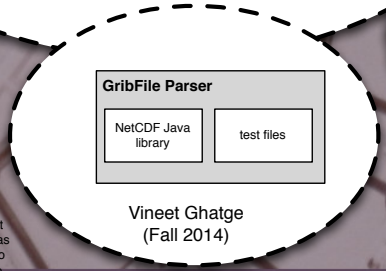
- So far, two semesters of projects
- Fall 2014 and Spring 2015
- Fall 2014
  - Crawl NASA AMD, NSF ACADIS and NSIDC ADE
  - Bayesian MIME detection
  - Gridded Binary Image Parser

Angela started out exploring Apache OODT and Push Pull as a crawler, along with Apache Nutch. She found that Nutch was more configurable and easier to set up for crawling ACADIS and AMD.



Prasanth wanted to examine the MIME detector in Tika - he was wondering if the issue with parsing science data was that the MIME detector was incorrectly detecting it.





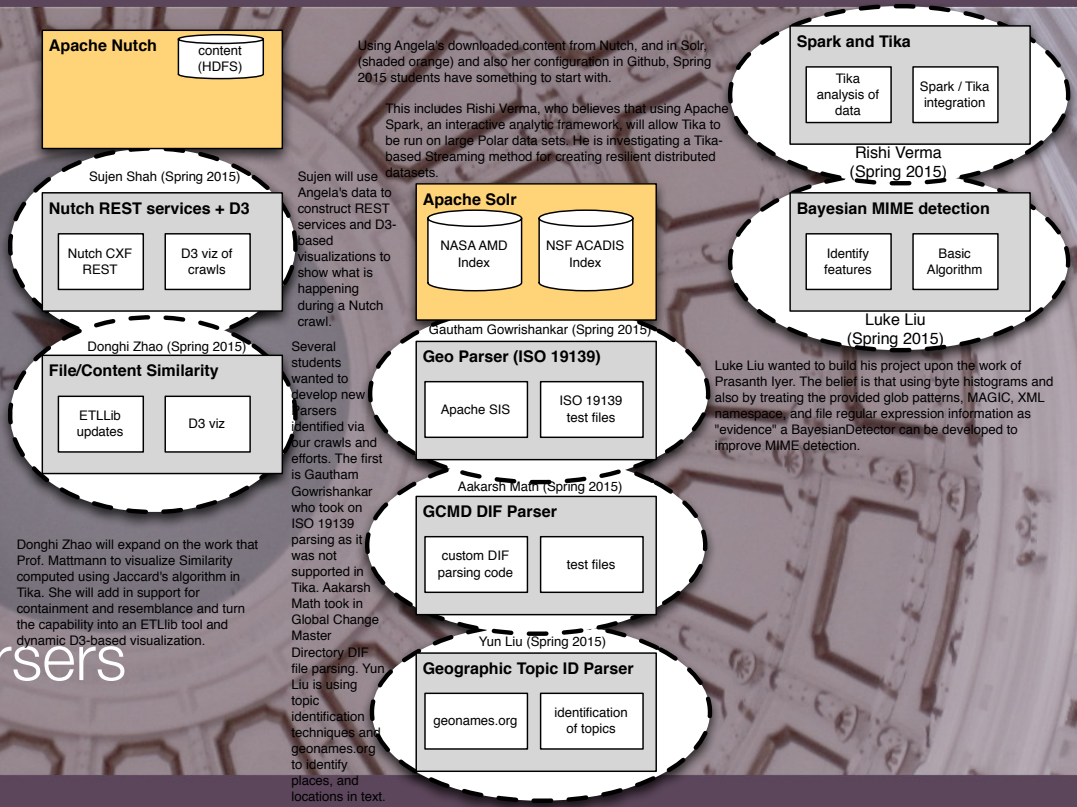
One of the science data files present in AMD that Tika didn't support and wouldn't parse was Grib (Gridded Binary) Files. So Vineet's project was a GribFile parser in Tika.

Prasanth wanted to treat the information provided from the glob file pattern; MIME magic; file name regular expression, and XML root chars as "evidence" in a Bayesian learning algorithm. He came up with the basic algorithm idea, and did some preliminary data gathering.

NSF Polar Work: Spring 2015

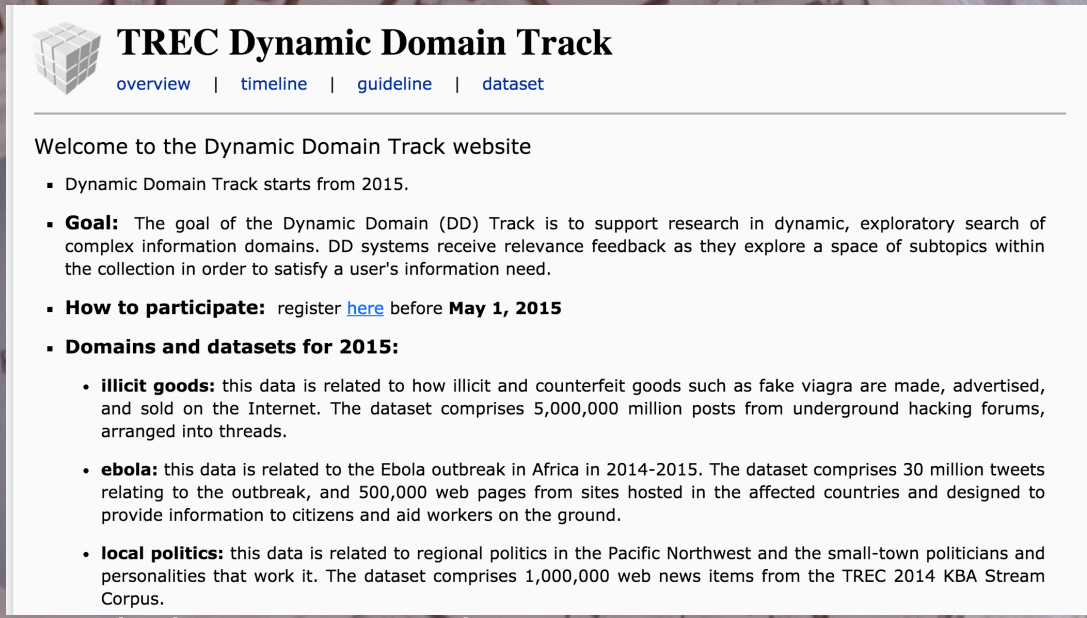
APACHECON NORTH AMERICA

- Nutch REST API
  - Drives crawling and eventually dataviz
- DataViz in D3 for image similarity
- Geo Topic Parser
  - ML and NLP with geonames
- GCMD DIF, ISO19115 parsers
- Bayesian Detector
- Spark and Tika



NIST Text Retrieval Conf (TREC)

- DARPA Memex started a new TREC "track" in Dynamic Domains
  - http://trec-dd.org/
- Memex contributions + polar contributions
- Polar
  - 1.7m URLs
  - 158Gb, lots of images and data to work on
- http://github.com/chrismattmann/trec-dd-polar/



TREC Dynamic Domain Track

overview | timeline | guideline | dataset

Welcome to the Dynamic Domain Track website

- Dynamic Domain Track starts from 2015.
- **Goal:** The goal of the Dynamic Domain (DD) Track is to support research in dynamic, exploratory search of complex information domains. DD systems receive relevance feedback as they explore a space of subtopics within the collection in order to satisfy a user's information need.
- **How to participate:** register here before **May 1, 2015**
- **Domains and datasets for 2015:**
  - **illicit goods:** this data is related to how illicit and counterfeit goods such as fake viagra are made, advertised, and sold on the Internet. The dataset comprises 5,000,000 million posts from underground hacking forums, arranged into threads.
  - **ebola:** this data is related to the Ebola outbreak in Africa in 2014-2015. The dataset comprises 30 million tweets relating to the outbreak, and 500,000 web pages from sites hosted in the affected countries and designed to provide information to citizens and aid workers on the ground.
  - **local politics:** this data is related to regional politics in the Pacific Northwest and the small-town politicians and personalities that work it. The dataset comprises 1,000,000 web news items from the TREC 2014 KBA Stream Corpus.

Cop out: What about videos?

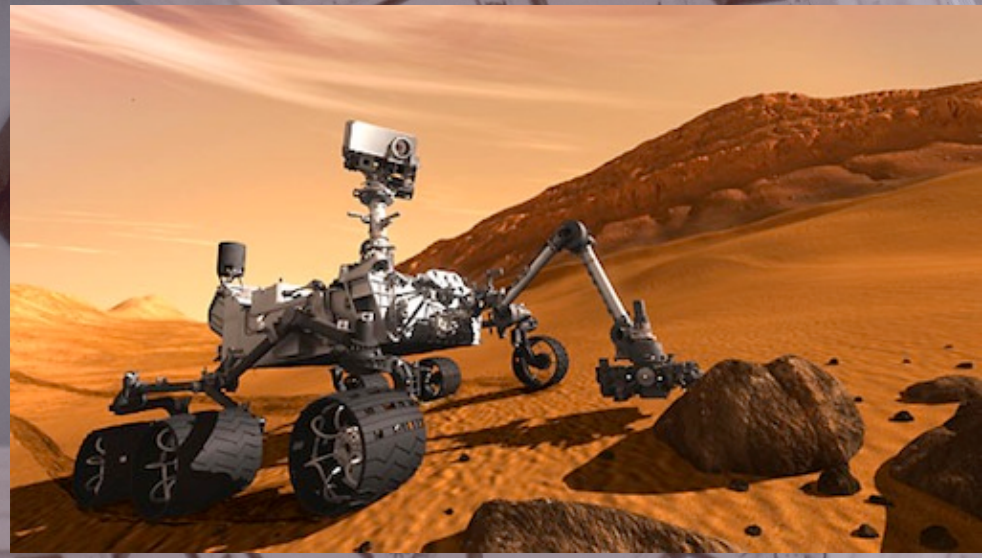
- I know I know
- Working on FFMPEG and Tika support for metadata
- https://issues.apache.org/jira/browse/TIKA-1510
- Have someone working on similarity and deduplication methods for video (Michael Ryoo)
  - Pooled Motion for First Person Videos http://arxiv.org/abs/1412.6505
- Streaming video parser
  - https://issues.apache.org/jira/browse/TIKA-1598

Tie back to NASA / JPL

- Transition into Physical Oceanographic Distributed Active Archive Center (PO.DAAC) for images from satellites

- PolarCyberInfrastructure community

- Science images and videos for Mars



Thank you!

- Chris Mattmann
- @chrismattmann
- mattmann@apache.org

- http://memex.jpl.nasa.gov/
- http://trec-dd.org/

- http://nsf-polar-cyberinfrastructure.github.io/datavis-hackathon/


Chris A. Mattmann, NASA JPL,
USC & the ASF
@chrismattmann
mattmann@apache.org
APACHECON
NORTH AMERICA