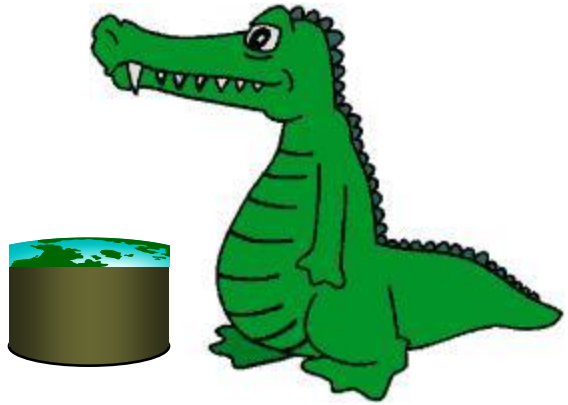
CAP4770/5771
Introduction to Data Science
Fall 2016

University of Florida, CISE Department
Prof. Daisy Zhe Wang



Based on notes from CS194/294 at UC Berkeley by
Michael Franklin, John Canny, and Jeff Hammerbacher

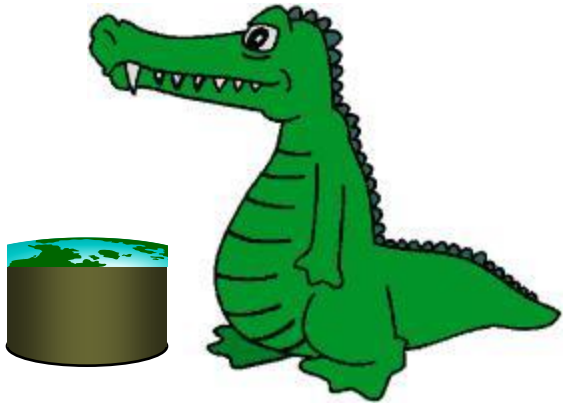
Data Types, Collection and Preparation

Data Types

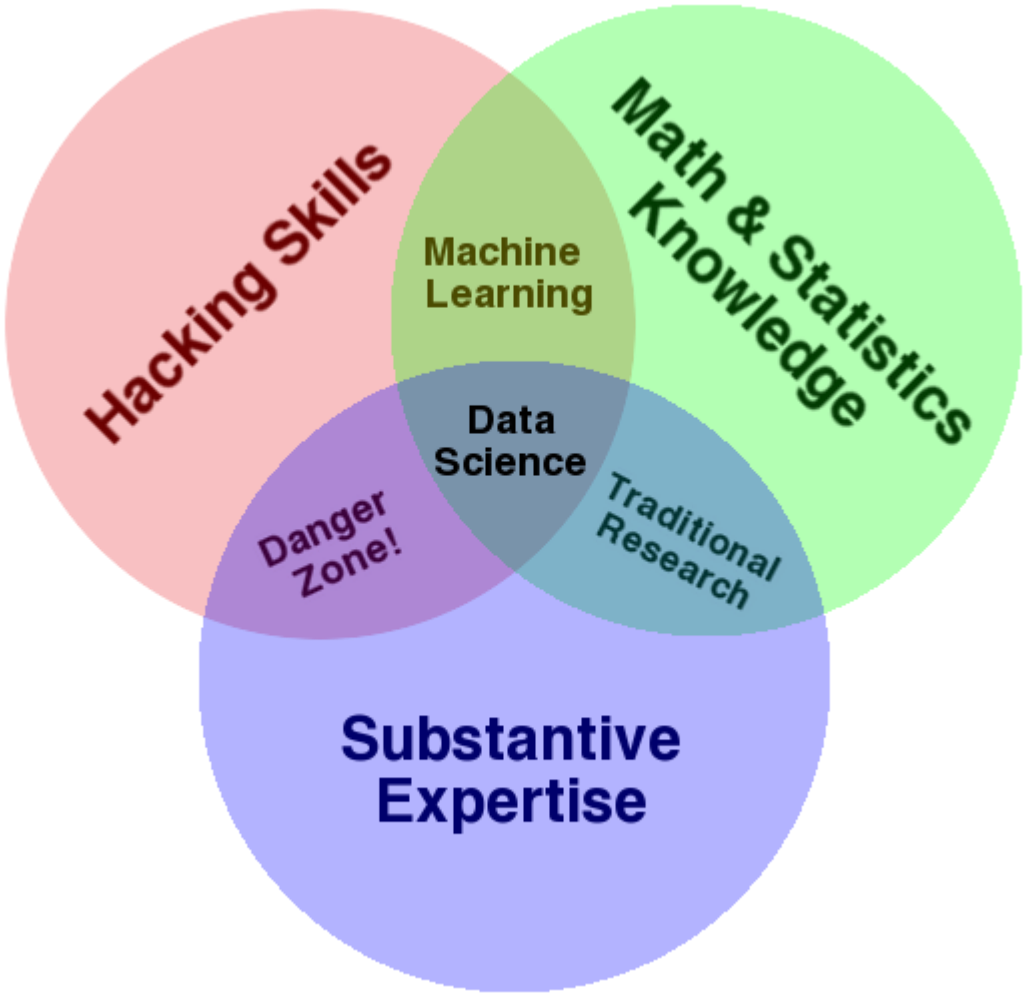
Data Sources

Data Models

Data Preparation



Data Science – A Visual Definition



Kandel et. al. Data Science and Analysis Process Model

- **Discover** data necessary to complete an analysis tasks.
- **Wrangle** data into a desired format.
- **Profile** data to verify its quality and its suitability for the analysis tasks.
- **Model** data for summarization or prediction.
- **Report** procedures and insights to consumers of the analysis.
- Additional Challenge in **Workflow** Management

Challenges in Data Science

- Preparing Data (Noisy, Incomplete, Diverse, Streaming …)
- Analyze Data (Scalable, Accurate, Real-time, Advanced Methods, Probabilities and Uncertainties …)
- Represent Analysis Results (i.e. data product) (Story-telling, Interactive, explainable…)

Outline

- Data Types and Sources
- Data Models
- Data Preparation

Data Sources at Web Companies

- Examples from Facebook
  - Application databases
  - Wikipedia (and other knowledge bases)
  - Web server logs
  - Event logs
  - API server logs
  - Ad server logs
  - Search server logs
  - Advertisement landing page content
  - Images and video

Structured Data

Semi-structured Data

Unstructured Data

From structured to unstructured

The (changing) role of Schema

Schema specify the **structure** and **types** of a data repository, e.g. the types of each column in a table.

They may also specify constraints **within** or **between** data fields.

Traditional databases are **schema-on-write**. You cannot load data into a table without a schema.

Newer (noSQL) data stores are **schema-on-read** or **schemaless**: You can defer applying a schema until you read the data, or avoid schema altogether.

Schema-on-Write

SQL:

```
CREATE SCHEMA Sprockets
  CREATE TABLE NineProngs (source int, cost int,
partnumber int) GO

INSERT INTO NineProngs (source, cost, partnumber)
VALUES (5, 100, 45312453)
```

Schema-on-Read

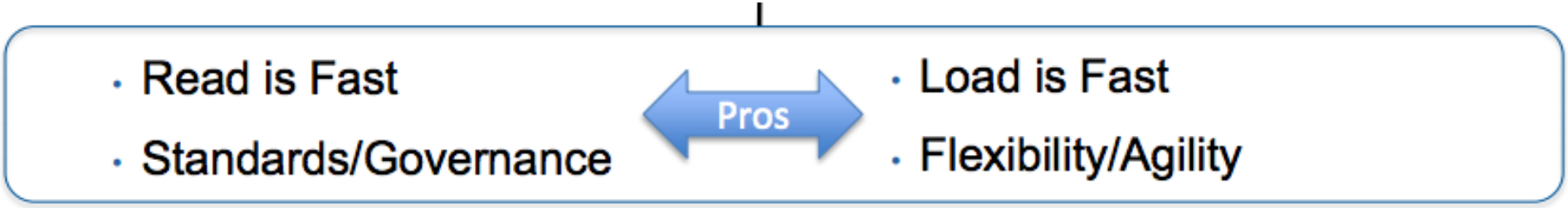
XML: Generalizes HTML and specifies data **structure**. XML schema can be applied later to interpret XML data and specify **data types**. Here is some XML-encoded **data**:

```
<location>
  <latitude>37.78333</latitude>
  <longitude>122.4167</longitude>
</location>
```

When stored without a schema, the numerical data are stored as **strings**.

Schema-on-Read vs. Schema-on-Write

- "Schema on Write"
  - Traditional Approach
- "Schema on Read"
  - Data is simply copied to the file store, no transformation is needed.
  - A SerDe (Serializer/Deserlizer) is applied during read time to extract the required columns (late binding)
  - New data can start flowing anytime and will appear retroactively once the SerDe is updated to parse it.



Tabular Data

- What is a table?
  - A **table** is a collection of **rows** and **columns**
  - Each row has an **index (a.k.a., key)**
  - Each column has a **name**
  - A **cell** is specified by an (index, name) pair
  - A cell may or may not have a **value**

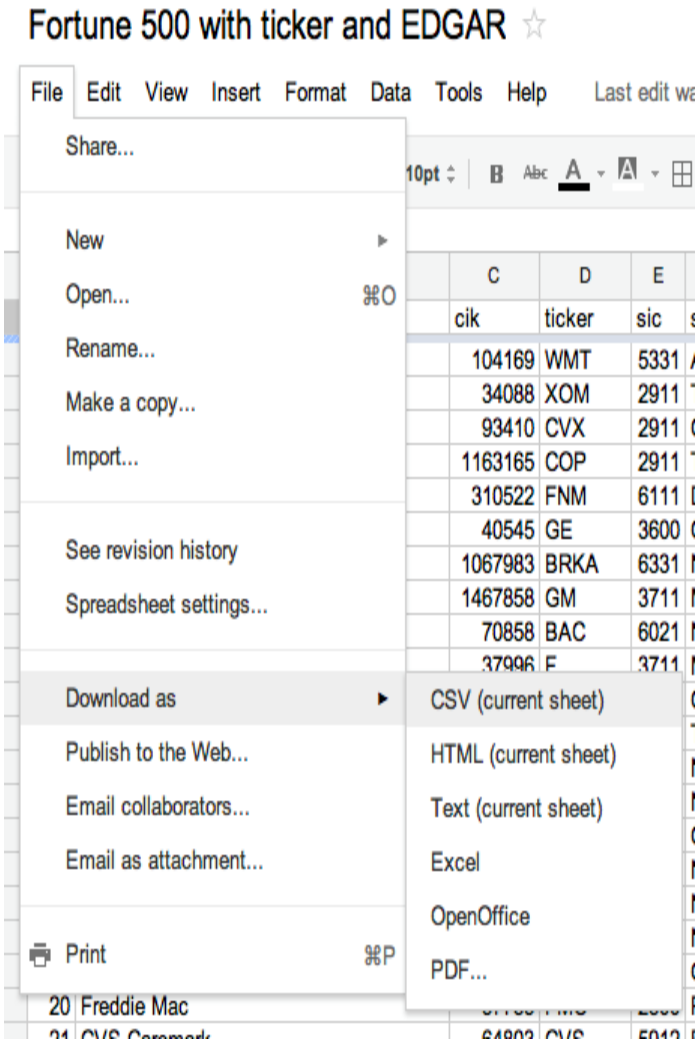
Tabular Data

- Fortune 500

	A	B	C	D	E	F	G	H	I
1	rank	company	cik	ticker	sic	state_location	state_of_incorporation	revenues	profits
2	1	Wal-Mart Stores	104169	WMT	5331	AR	DE	421849	16389
3	2	Exxon Mobil	34088	XOM	2911	TX	NJ	354674	30460
4	3	Chevron	93410	CVX	2911	CA	DE	196337	19024
5	4	ConocoPhillips	1163165	COP	2911	TX	DE	184966	11358
6	5	Fannie Mae	310522	FNM	6111	DC	DC	153825	-14014
7	6	General Electric	40545	GE	3600	CT	NY	151628	11644
8	7	Berkshire Hathaway	1067983	BRKA	6331	NE	DE	136185	12967
9	8	General Motors	1467858	GM	3711	MI	MI	135592	6172
10	9	Bank of America Corp.	70858	BAC	6021	NC	DE	134194	-2238
11	10	Ford Motor	37996	F	3711	MI	DE	128954	6561
12	11	Hewlett-Packard	47217	HPQ	3570	CA	DE	126033	8761
13	12	AT&T	732717	T	4813	TX	DE	124629	19864
14	13	J.P. Morgan Chase & Co.	19617	JPM	6021	NY	DE	115475	17370
15	14	Citigroup	831001	C	6021	NY	DE	111055	10602
16	15	McKesson	927653	MCK	5122	CA	DE	108702	1263
17	16	Verizon Communications	732712	VZ	4813	NY	DE	106565	2549
18	17	American International Group	5272	AIG	6331	NY	DE	104417	7786
19	18	International Business Machines	51143	IBM	3570	NY	NY	99870	14833
20	19	Cardinal Health	721371	CAH	5122	OH	OH	98601.9	642.2
21	20	Freddie Mac	37785	FMC	2800	PA	DE	98368	-14025

Tabular Data

- Fortune 500



Tabular Data –> csv text files

- Fortune 500

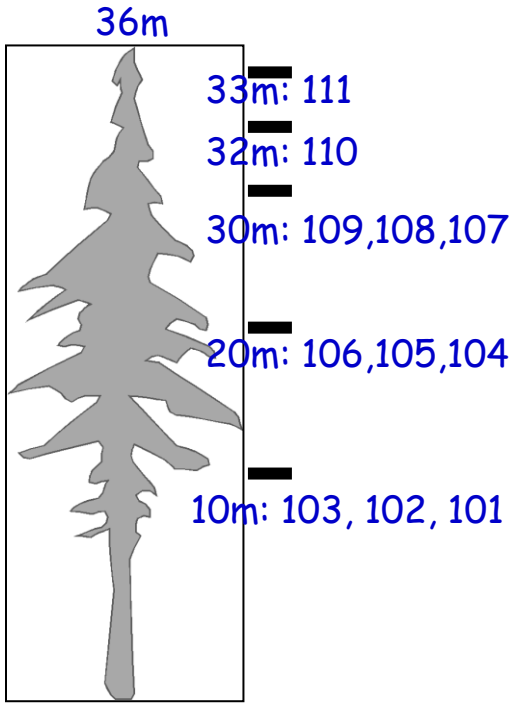
Fortune 500 with ticker and EDGAR – Plus Ticker and EDGAR.txt

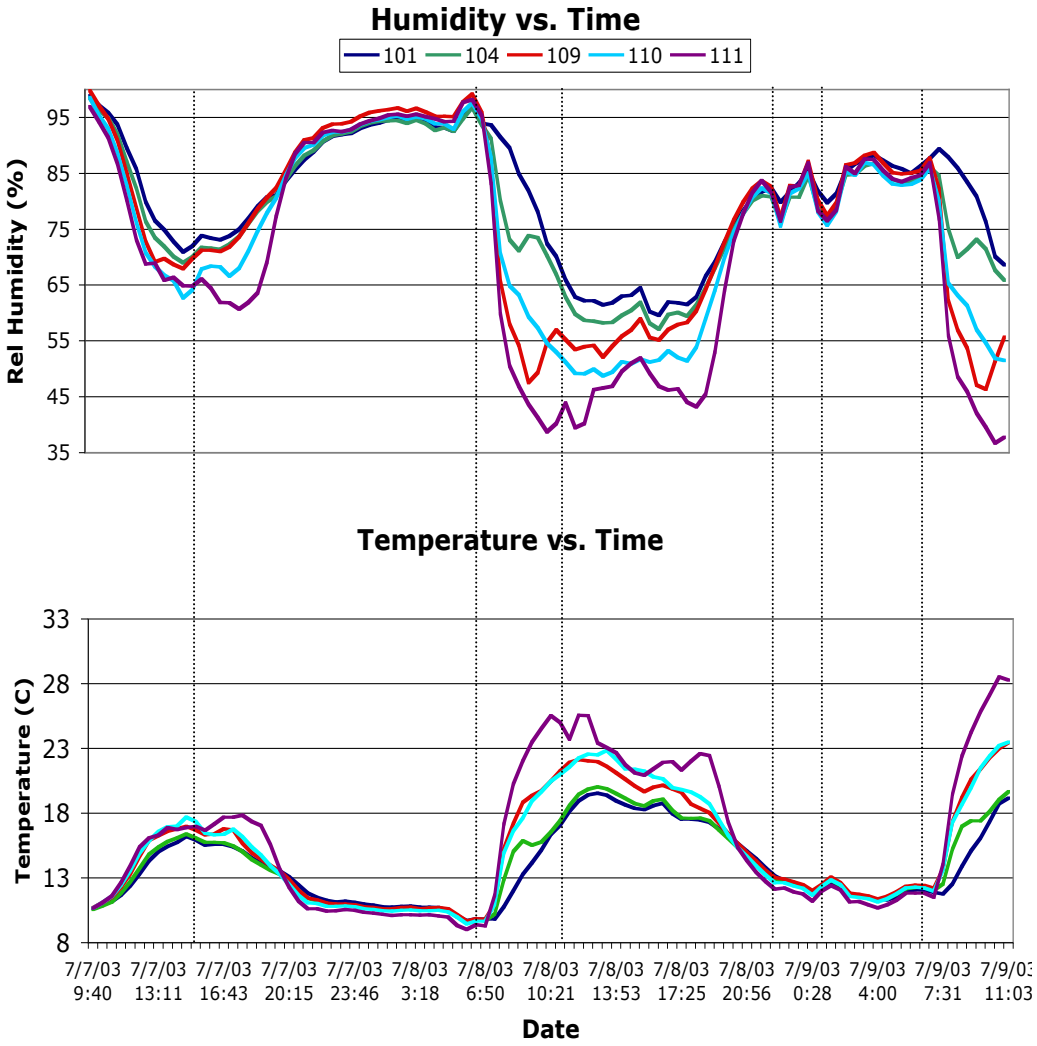
```
rank,company,cik,ticker,sic,state_location,state_of_incorporation,revenues,profits
1,Wal-Mart Stores,104169,WMT,5331,AR,DE,421849,16389
2,Exxon Mobil,34088,XOM,2911,TX,NJ,354674,30460
3,Chevron,93410,CVX,2911,CA,DE,196337,19024
4,ConocoPhillips,1163165,COP,2911,TX,DE,184966,11358
5,Fannie Mae,310522,FNM,6111,DC,DC,153825,-14014
6,General Electric,40545,GE,3600,CT,NY,151628,11644
7,Berkshire Hathaway,1067983,BRKA,6331,NE,DE,136185,12967
8,General Motors,1467858,GM,3711,MI,MI,135592,6172
9,Bank of America Corp.,70858,BAC,6021,NC,DE,134194,-2238
10,Ford Motor,37996,F,3711,MI,DE,128954,6561
11,Hewlett-Packard,47217,HPQ,3570,CA,DE,126033,8761
12,AT&T,732717,T,4813,TX,DE,124629,19864
13,J.P. Morgan Chase & Co.,19617,JPM,6021,NY,DE,115475,17370
14,Citigroup,831001,C,6021,NY,DE,111055,10602
15,McKesson,927653,MCK,5122,CA,DE,108702,1263
16,Verizon Communications,732712,VZ,4813,NY,DE,106565,2549
17,American International Group,5272,AIG,6331,NY,DE,104417,7786
18,International Business Machines,51143,IBM,3570,NY,NY,99870,14833
19,Cardinal Health,721371,CAH,5122,OH,OH,98601.9,642.2
20,Freddie Mac,37785,FMC,2800,PA,DE,98368,-14025
21,CVS Caremark,64803,CVS,5912,RI,DE,96413,3427
22,UnitedHealth Group,731766,UNH,6324,MN,MN,94155,4634
23,Wells Fargo,72971,WFC,6021,CA,DE,93249,12362
24,Valero Energy,1035002,VLO,2911,TX,DE,86034,324
25,Kroger,56873,KR,5411,OH,OH,82189.4,1116.3
26,Procter & Gamble,80424,PG,2840,OH,OH,79689,12736
27,AmerisourceBergen,1140859,ABC,5122,PA,DE,77954,636.7
28,Costco Wholesale,909832,COST,5331,WA,WA,77946,1303
29,Marathon Oil,101778,MRO,2911,TX,DE,68413,2568
30,Home Depot,354950,HD,5211,GA,DE,67997,3338
```

Protein Data Bank – delimited files

```
HEADER    APOPTOSIS                               05-OCT-10   3IZA
TITLE     STRUCTURE OF AN APOPTOSOME-PROCASPASE-9 CARD COMPLEX
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: APOPTOTIC PROTEASE-ACTIVATING FACTOR 1;
COMPND   3 CHAIN: A, B, C, D, E, F, G;
COMPND   4 SYNONYM: APAF-1;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 GENE: APAF1, KIAA0413;
SOURCE   6 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 7108;
SOURCE   8 EXPRESSION_SYSTEM_STRAIN: SF21;
SOURCE   9 EXPRESSION_SYSTEM_VECTOR_TYPE: INSECT VIRUS;
SOURCE  10 EXPRESSION_SYSTEM_PLASMID: PFASTBAC1
KEYWDS    APOPTOSOME, APAF-1, PROCASPASE-9 CARD, APOPTOSIS
EXPDTA    ELECTRON MICROSCOPY
AUTHOR    S.YUAN,X.YU,M.TOPF,S.J.LUDTKE,X.WANG,C.W.AKEY
REVDAT   1   03-NOV-10 3IZA    0
SPRSDE     03-NOV-10 3IZA      3IYT
JRNL       AUTH   S.YUAN,X.YU,M.TOPF,S.J.LUDTKE,X.WANG,C.W.AKEY
JRNL       TITL   STRUCTURE OF AN APOPTOSOME-PROCASPASE-9 CARD COMPLEX
JRNL       REF    STRUCTURE                V.  18   571 2010
```

Internet of Things: Example measurements





In the news: Erika: data collecting sensors are dropped by airplanes

Tabular Data from Sensors

Challenges

- May be many missing fields (a particular sensor may not produce all types of output).
- Device may go offline for a while.
- Device may be damaged (permanently or intermittently).
- Timestamps usually critical but may not be accurate.
- Other meta-data (location, device ID) may have errors.

Log Files – Example Apache Web Log

Processes, usually daemons, create logs
e.g., httpd, mysqld, syslogd

- 66.249.65.107 - - [08/Oct/2007:04:54:20 -0400] "GET /support.html HTTP/1.1" 200 11179 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

- 111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET / HTTP/1.1" 200 10801 "http://www.google.com/search?q=log+analyzer&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7"

- 111.111.111.111 - - [08/Oct/2007:11:17:55 -0400] "GET /style.css HTTP/1.1" 200 3225 ""http://www.loganalyzer.net/" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US; rv:1.8.1.7) Gecko/20070914 Firefox/2.0.0.7"

Syslog – A Standard for System Messages

- Developed by Eric Allman (at Berkeley) as part of the Sendmail project
- Standardized by the IETF in RFC 3164 and RFC 5424
- Listens on port 514 using UDP
- Puts data in /var/log/messages by default
- Enables rich analysis splunk>

Syslog

dhcp-47-129:DataScienceF14> syslog -w 10

Feb 3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting read:]: unexpected field ID 23 with type 8. Skipping.
Feb 3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMUser read:]: unexpected field ID 17 with type 12. Skipping.
Feb 3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAuthenticationResult read:]: unexpected field ID 6 with type 11. Skipping.
Feb 3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAuthenticationResult read:]: unexpected field ID 7 with type 11. Skipping.
Feb 3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting read:]: unexpected field ID 19 with type 8. Skipping.
Feb 3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting read:]: unexpected field ID 23 with type 8. Skipping.
Feb 3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMUser read:]: unexpected field ID 17 with type 12. Skipping.
Feb 3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMSyncState read:]: unexpected field ID 5 with type 10. Skipping.
Feb 3 15:18:49 dhcp-47-129 com.apple.mtmd[47] <Notice>: low priority thinning needed for volume Macintosh HD (/) with 18.9 <= 20.0 pct free space

HTML – a Standard Generalized Markup Language

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">

<head><!-- types/widgets/pages/common/page.tmpl home/index_v3.html generated by index_v3 on Wed 29 Feb 2012 11:04:41 PM PST -->

<title>San Francisco Bay Area — News, Sports, Business, Entertainment, Classifieds: SFGate</title>

<meta http-equiv="content-type" content="text/html; charset=iso-8859-1" />

<meta name="description" content="Find local news & information, updated weather, traffic, classifieds, sports scores, real estate, jobs, cars, food & wine, travel, entertainment, events and more on SFGate.com. Connect to the Bay Area community." />

<meta name="keywords" content="San Francisco, San Francisco Bay Area, news, local events, breaking news, world news, San Francisco Chronicle, SFGate" />

<meta property="fb:page_id" content="105702905593" />

<meta property="fb:admins" content="653226748,658759748" />

<!-- /widgets/sitewide/css/all/inc.html widgets/pages/common/post_write_mtime/css_inc.tmpl -->

<!-- generated by sitewidecss on Thu 16 Feb 2012 10:41:53 AM PST -->

<link rel="stylesheet" type="text/css" title="SFGate" media="all" href="http://imgs.sfgate.com/css1329417713/sitewide/css/sitewide.css" />

<!-- sitewide/css/all/inc.html end css_inc.tmpl -->

Web Crawling – HTML

- What are the tags in HTML used for?
- There's plenty of data, and there are many crawlers for targeted exploration…
  - HTTrack, …
- Common Crawl, about 5 billion web pages, between 0.2-0.5% of Google's web crawl.
  - 60 TB, hosted on Amazon S3, also available for download.
  - Includes link data, page rank.
  - In ARC (Internet Archive) File format.

Web Services

Most large web sites today actively discourage screen-scraping to get their content, and provide Web Service APIs instead.

This is the “right” way to get data from online sources.

Web Services

W3C definition:

a "Web service" as "a software system designed to support interoperable machine-to-machine interaction over a network".

Two kinds:

- XML-based RPC-style messages: WSDL and SOAP
- REST-style stateless interactions, URLs encode state

Application API for Data Aquisition

Twitter: REST API and streaming API with JSON content. Provides sampling, searching and filtering capabilities.

Amazon: has a "product advertising API" in XML with a WSDL spec. Includes product search, reviews etc.

Livejournal: RSS/Atom + custom XML/RPC. Search by keyword, topic, follow friend links.

Netflix: Javascript, Atom and REST interfaces.

Ebay: Many APIs for searching, buying and posting. WSDL descriptions, client code in Java and .NET

Flickr: Comprehensive API set, free for non-commercial use. REST, XML-RPC, SOAP, with client code in many languages.

vBulletin: REST interface, most actions supported

XML files

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- bookstore.xml -->
<bookstore>
   <book ISBN="0123456001">
      <title>Java For Dummies</title>
      <author>Tan Ah Teck</author>
      <category>Programming</category>
      <year>2009</year>
      <edition>7</edition>
      <price>19.99</price>
   </book>
```

XML

- XML = *Extensible Markup Language*.
- While HTML uses tags for formatting (e.g., "italic"), XML uses tags for semantics (e.g., "this is an address").
- Key idea: create tag sets for a domain (e.g., genomics), and translate all data into properly tagged XML documents.

JSON

JSON (Javascript Object Notation) by contrast is a schemaless data description language (Schema support was added later):

```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
       "streetAddress": "21 2nd Street",
       "city": "New York",
       "state": "NY",
       "postalCode": "10021-3100" },
  "phoneNumbers": [
      { "type": "home",
        "number": "212 555-1234" },
      { "type": "office",
        "number": "646 555-4567" } ],
  "children": [],
  "spouse": null
}
```

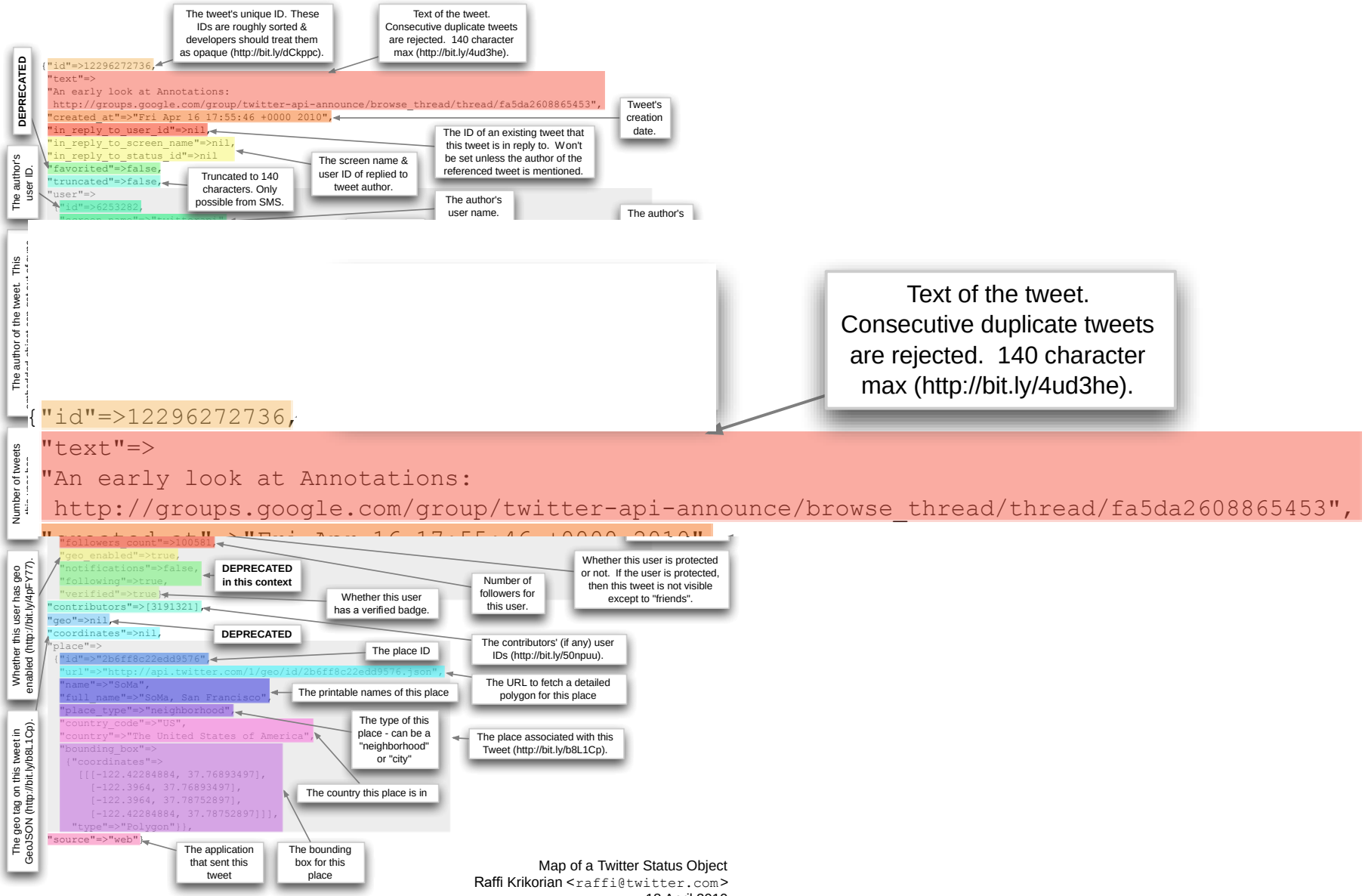
JSON

JSON is typically used to represent hierarchical data structures directly in the target language (Javascript or Java).

Transformations on the data are **procedural** in the target language (not declarative in a language as in XML/Xquery).

Easier for some tasks, but painful for e.g. schema changes.

Twitter Data in JSON Format



Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

Multimodal Data Sources on World Cup Common

Twitter
~5M tweets with 244,000 associating images

News & Blogs
776 blog posts and associated pictures

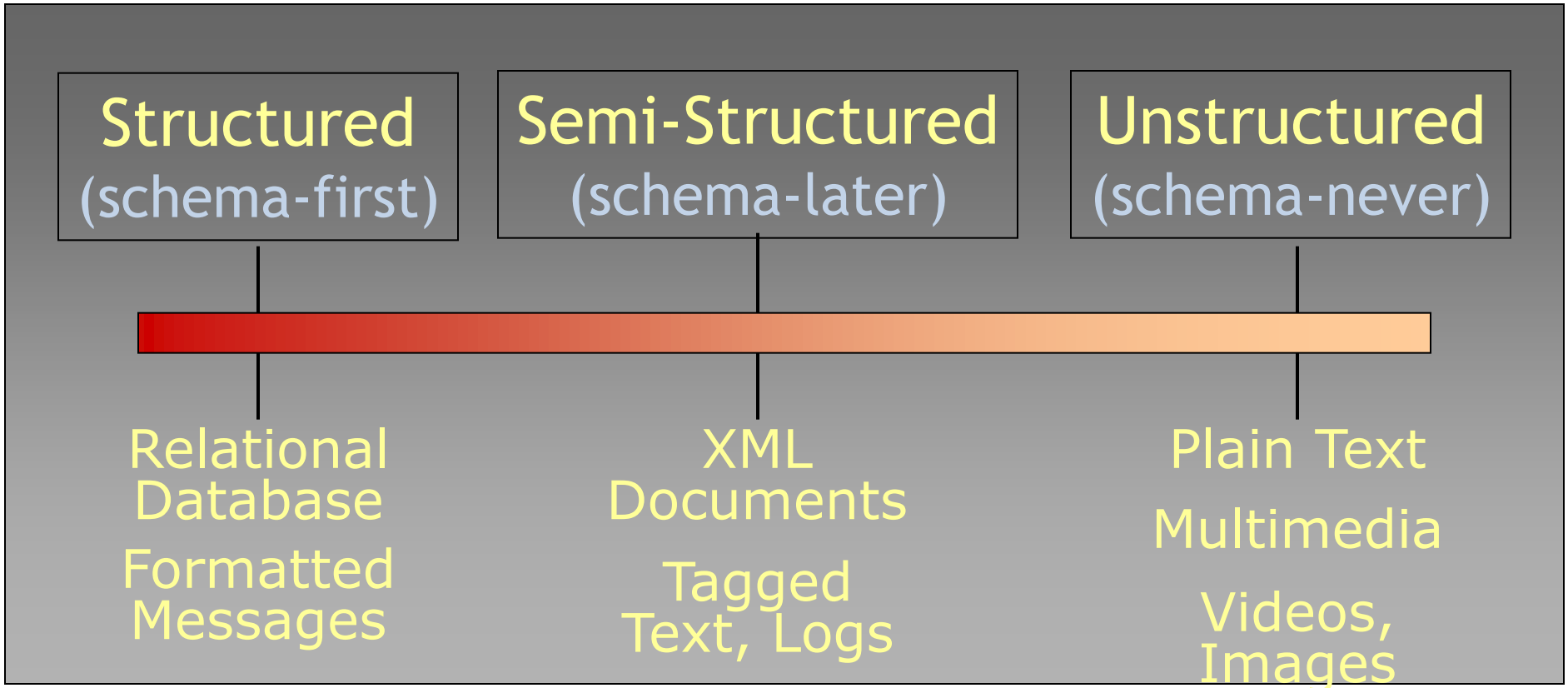
Facebook
250 posts and associated images

Flickr
913 images and associated captions

Google Images
1,920 images and associated captions

Youtube
460 minutes of Videos and Audio

The Structure Spectrum



Other special types of data

- Graph data
- High dimensional data
- Hyperspectral/Lidar data
- Spatial & temporal data

- Streaming vs. static data vs. dynamic

One Size Does not Fit All

– Dr. Michael Stonebraker, Turing Award 2015

Graph Data

Lots of interesting data has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

Some of these graphs can get quite large (e.g., Facebook* user graph)