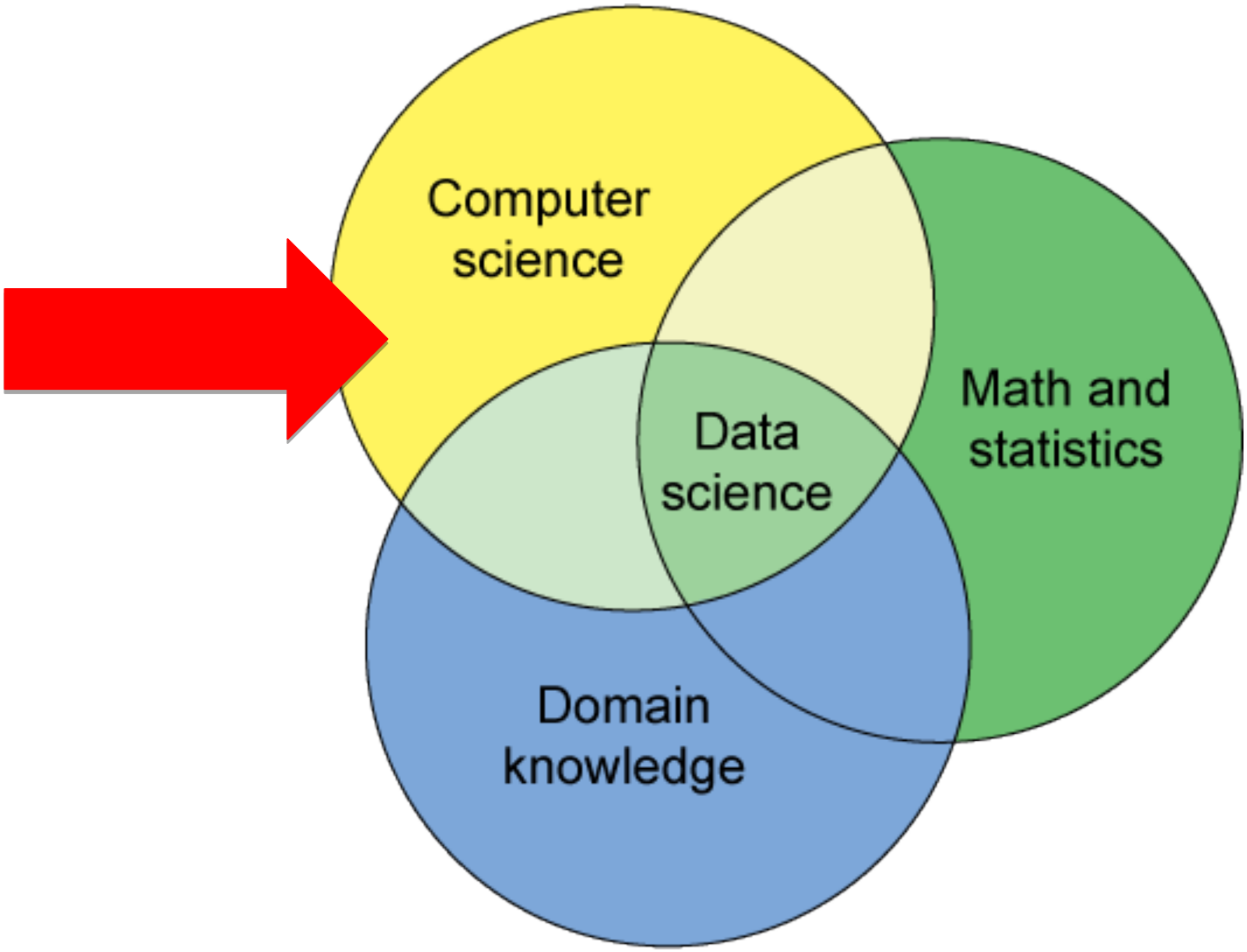
Introduction to Data Science
Lecture 2
Data Preparation (part 1)

CS 439 Fall 2023
Farhan Khan
GIK Institute

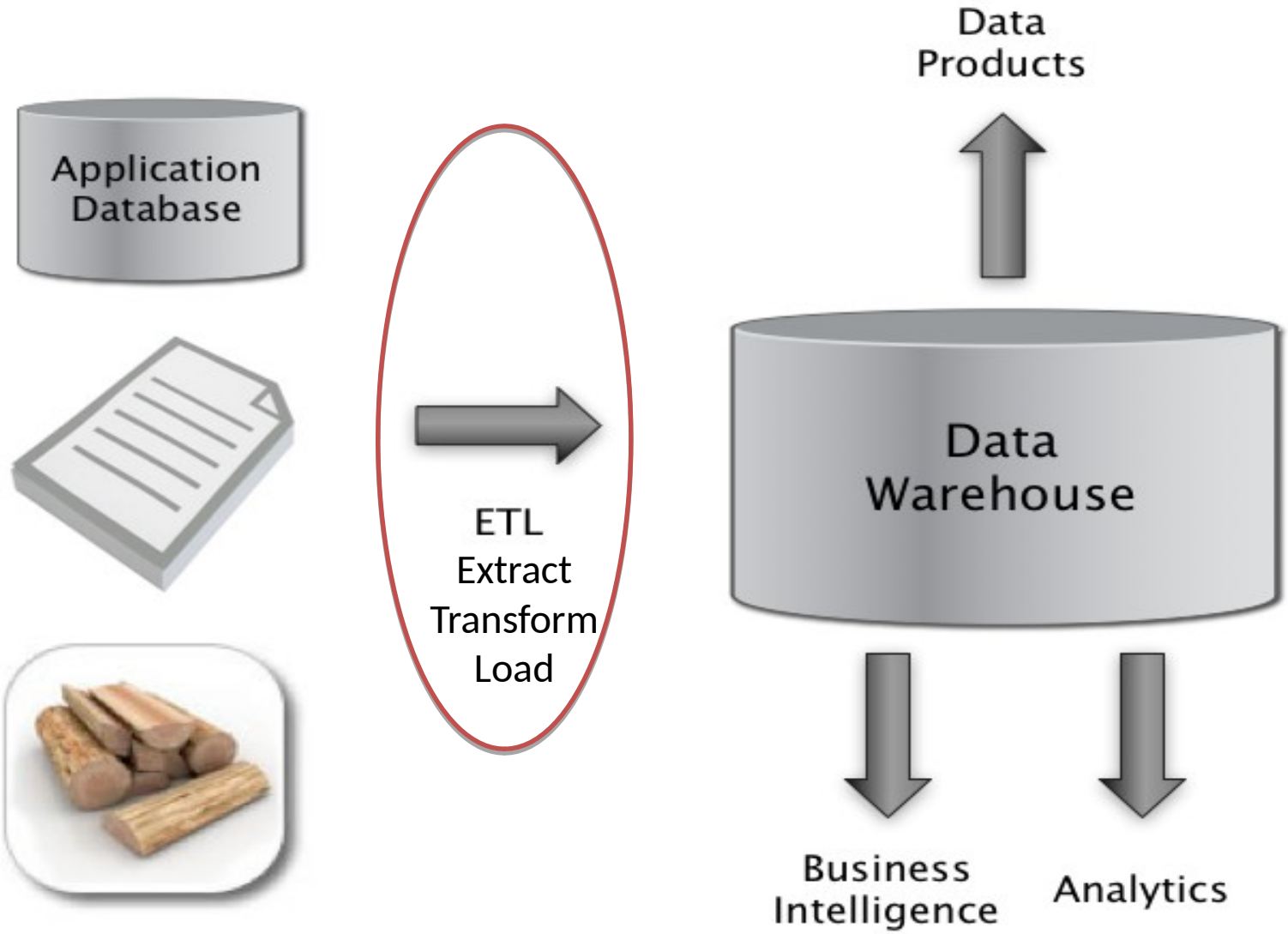
Outline

- Discuss Kandel et al. Paper
- Lecture – Data Prep, and Manipulation
- Exercise – Unix text Utilities
- Review of exercise

Data Science – One Definition



The Big Picture



Data Preparation overview

- ETL
  - We need to **extract** data from the **source(s)**
  - We need to **load** data into the **sink**
  - We need to **transform** data at the source, sink, or in a **staging area**

  - Sources: file, database, event log, …
  - Sinks: Python, R, SQLite, RDBMS, Data Warehouse

Data Preparation overview

- Process model
  - The construction of a new data preparation process is done in many phases
    - Data **characterization**
    - Data **cleaning**
    - Data **integration**
  - We must efficiently move data around in space and time
    - Data **transfer**
    - Data **serialization** and **deserialization**

Data Preparation overview

- Workflow
  - The transformation **pipeline** or **workflow** often consists of many steps
    - For example: Unix pipes and filters
    - $ cat data_science.txt | wc | mail -s "word count" hammer@example.com
  - If the workflow is to be used more than once, it can be **scheduled**
    - Scheduling can be time-based or event-based
  - Recording the execution of a workflow is known as capturing **lineage** or **provenance**

The Businessperson

- Data Sources
  - Web pages
  - Excel
- ETL
  - Copy and paste
- Data Warehouse
  - Excel
- Business Intelligence and Analytics
  - Excel functions
  - Excel charts
  - Visual Basic?!

The Programmer

- Data Sources
  - Web scraping, web services API
  - Excel spreadsheet exported as CSV
  - Database queries
- ETL
  - wget, curl, Beautiful Soup, lxml
- Data Warehouse
  - Flat files
- Business Intelligence and Analytics
  - Numpy, Matplotlib, R

The Enterprise

- Data Sources
  - Application databases
  - Intranet files
  - Application server log files
- ETL
  - Informatica, IBM DataStage, Ab Initio, Talend
- Data Warehouse
  - Teradata, Oracle, IBM DB2, Microsoft SQL Server
- Business Intelligence and Analytics
  - Business Objects, Cognos, Microstrategy
  - SAS, SPSS, R

The Web Company

- Data Sources
  - Application databases
  - Logs from the services tier
  - Web crawl data
- ETL
  - Flume, Sqoop, Pig, Crunch, Oozie
- Data Warehouse
  - Hadoop/Hive, Spark/Shark
- Business Intelligence and Analytics
  - Custom dashboards: Argus, BirdBrain
  - R

Compare to Categories in the Kandel et al. Paper?

- Hackers?
- Scripters
- Application Users

- The process:
- Discover, Wrangle, Profile, Model, Report
  - Today we'll start discussing "Wrangling"
  - integrating, cleaning and transforming

Impediments to Collaboration

- Diversity of tools and PLs makes it hard to share
- Finding a script or computed result is harder than just writing the program from scratch!
  - Q: How could we fix this?
- View that much of the analysis work is “throw away”

Data Sources at Web Companies

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

Tabular Data

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

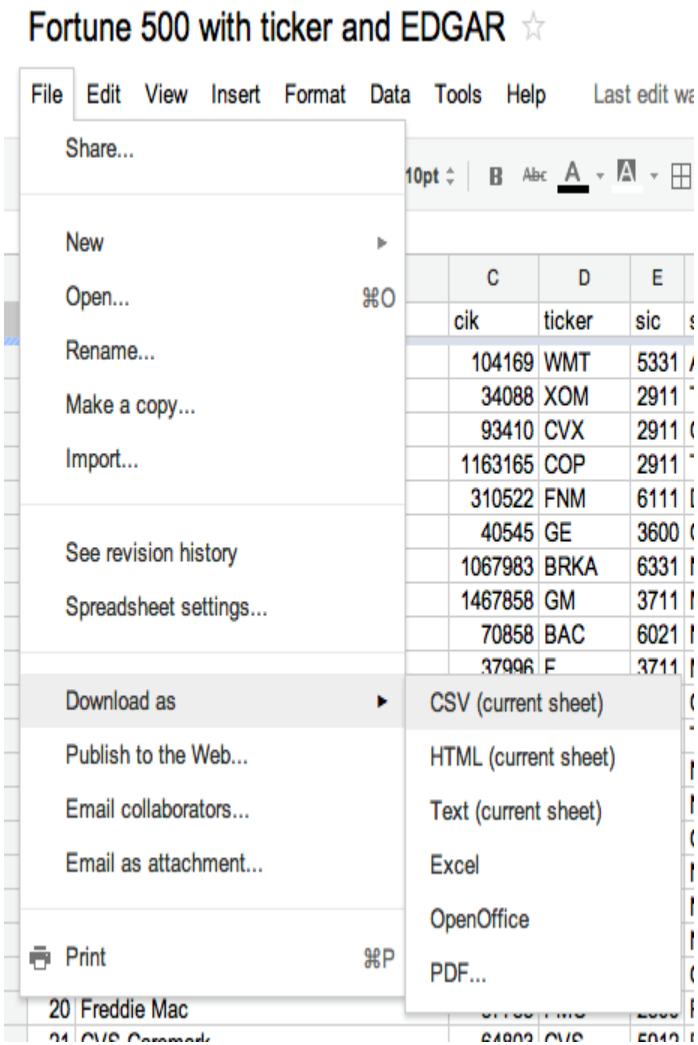
Tabular Data

- 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)

- 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
```

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

Log Files - Syslog

- 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
- Functionality extended by syslog-ng and rsyslog
  - More complex message formatting
  - Content-based filtering
  - TCP as a transport

Syslog

dhcp-47-129:DataScienceS14 Michael$ 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

Syslog – XML Format (Yikes!)

dhcp-47-129:DataScienceS14 Michael$ syslog -w 1 -F xml


```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple Computer//DTD PLIST 1.0//EN" "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<array>
	<dict>
		<key>ASLMessageID</key>
		<string>3280612</string>
		<key>Time</key>
		<string>Feb  3 15:18:49</string>
		<key>TimeNanoSec</key>
		<string>608197000</string>
		<key>Level</key>
		<string>5</string>
		<key>PID</key>
		<string>47</string>
		<key>UID</key>
		<string>0</string>
		<key>GID</key>
		<string>0</string>
		<key>ReadGID</key>
		<string>80</string>
		<key>Host</key>
		<string>dhcp-47-129</string>
		<key>Sender</key>
		<string>com.apple.mtmd</string>
		<key>Facility</key>
		<string>daemon</string>
		<key>Message</key>
		<string>low priority thinning needed for volume Macintosh HD (/) with 18.9 &lt;= 20.0 pct free space </string>
	</dict>
</array>
</plist>
```

Some Questions

1) How Many Characters are there in a Tweet?

2) How Many Bytes are there in a Tweet?

Tweet JSON Format

```
{"id"=>12296272736,
 "text"=>
 "An early look at Annotations:
  http://groups.google.com/group/twitter-api-announce/browse_thread/thread/fa5da2608865453",
 "created_at"=>"Fri Apr 16 17:55:46 +0000 2010",
 "in_reply_to_user_id"=>nil,
 "in_reply_to_screen_name"=>nil,
 "in_reply_to_status_id"=>nil
 "favorited"=>false,
 "truncated"=>false,
 "user"=>
  {"id"=>6253282,
   "screen_name"=>"twitterapi",
   "followers_count"=>100581,
   "geo_enabled"=>true,
   "notifications"=>false,
   "following"=>true,
   "verified"=>true},
 "contributors"=>[3191321],
 "geo"=>nil,
 "coordinates"=>nil,
 "place"=>
  {"id"=>"2b6ff8c22edd9576",
   "url"=>"http://api.twitter.com/1/geo/id/2b6ff8c22edd9576.json",
   "name"=>"SoMa",
   "full_name"=>"SoMa, San Francisco",
   "place_type"=>"neighborhood",
   "country_code"=>"US",
   "country"=>"The United States of America",
   "bounding_box"=>
    {"coordinates"=>
      [[[-122.42284884, 37.76893497],
        [-122.3964, 37.76893497],
        [-122.3964, 37.78752897],
        [-122.42284884, 37.78752897]]],
     "type"=>"Polygon"}},
 "source"=>"web"}
```

The tweet's unique ID. These IDs are roughly sorted & developers should treat them as opaque (http://bit.ly/dCkppc).

Text of the tweet. Consecutive duplicate tweets are rejected. 140 character max (http://bit.ly/4ud3he).

DEPRECATED

Tweet's creation date.

The ID of an existing tweet that this tweet is in reply to. Won't be set unless the author of the referenced tweet is mentioned.

The author's user ID.

The screen name & user ID of replied to tweet author.

Truncated to 140 characters. Only possible from SMS.

The author's user name.

The author's

The author of the tweet. This

Text of the tweet. Consecutive duplicate tweets are rejected. 140 character max (http://bit.ly/4ud3he).

Number of tweets

Whether this user has geo enabled (http://bit.ly/4pFY77).

DEPRECATED in this context

Whether this user has a verified badge.

Number of followers for this user.

Whether this user is protected or not. If the user is protected, then this tweet is not visible except to "friends".

DEPRECATED

The contributors' (if any) user IDs (http://bit.ly/50npuu).

The place ID

The URL to fetch a detailed polygon for this place

The printable names of this place

The type of this place - can be a "neighborhood" or "city"

The place associated with this Tweet (http://bit.ly/b8L1Cp).

The geo tag on this tweet in GeoJSON (http://bit.ly/b8L1Cp).

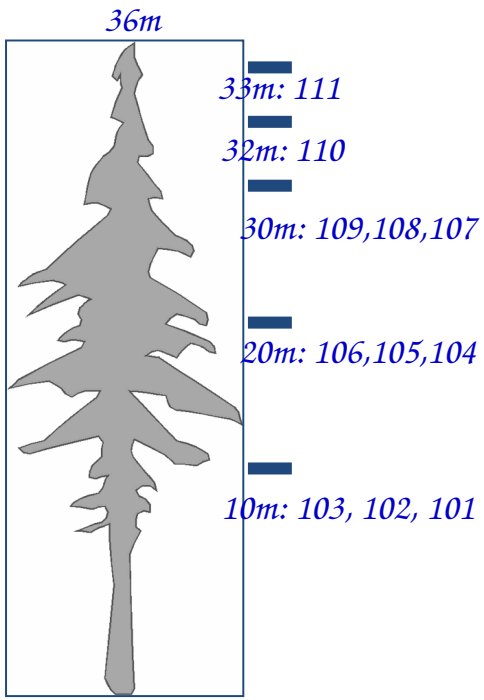
The country this place is in

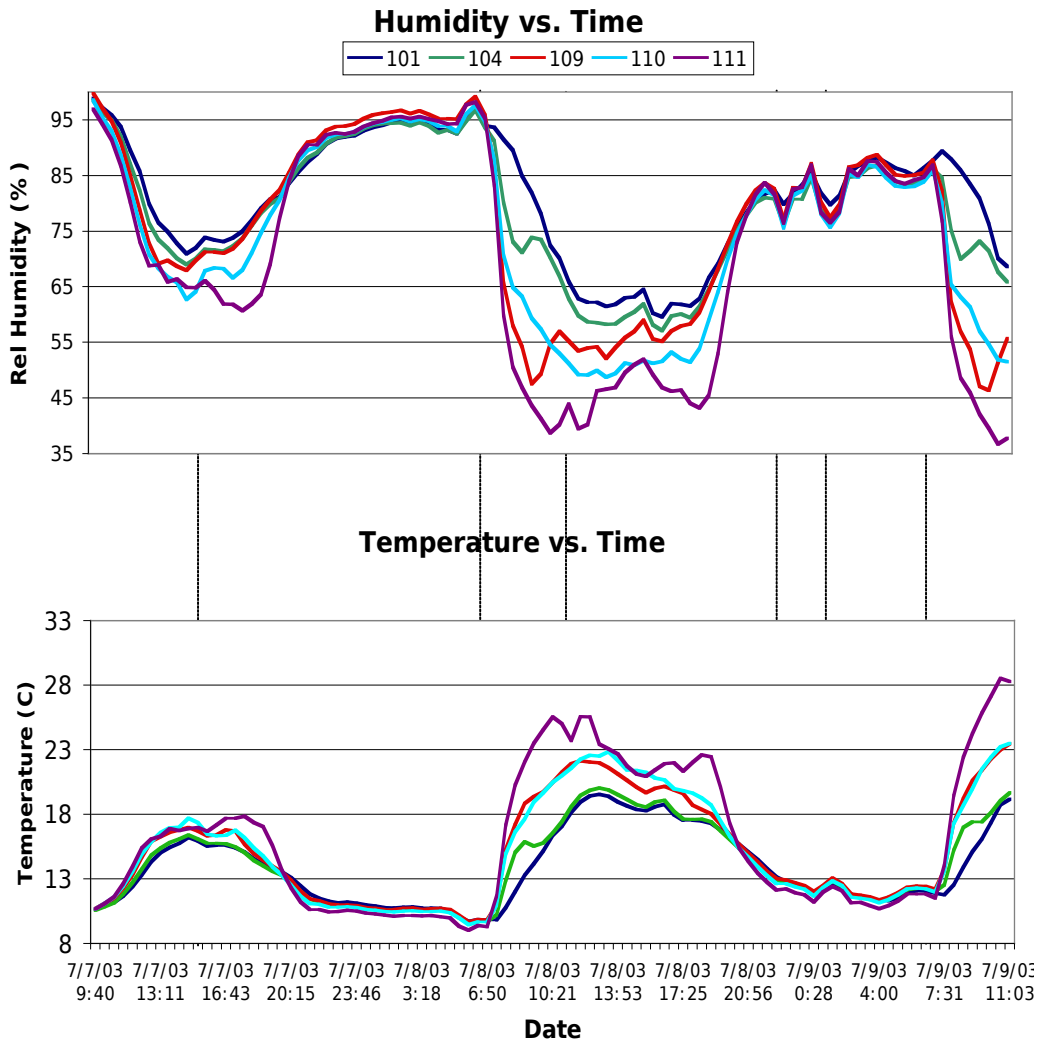
The application that sent this tweet

The bounding box for this place

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

Internet of Things:
Example measurements

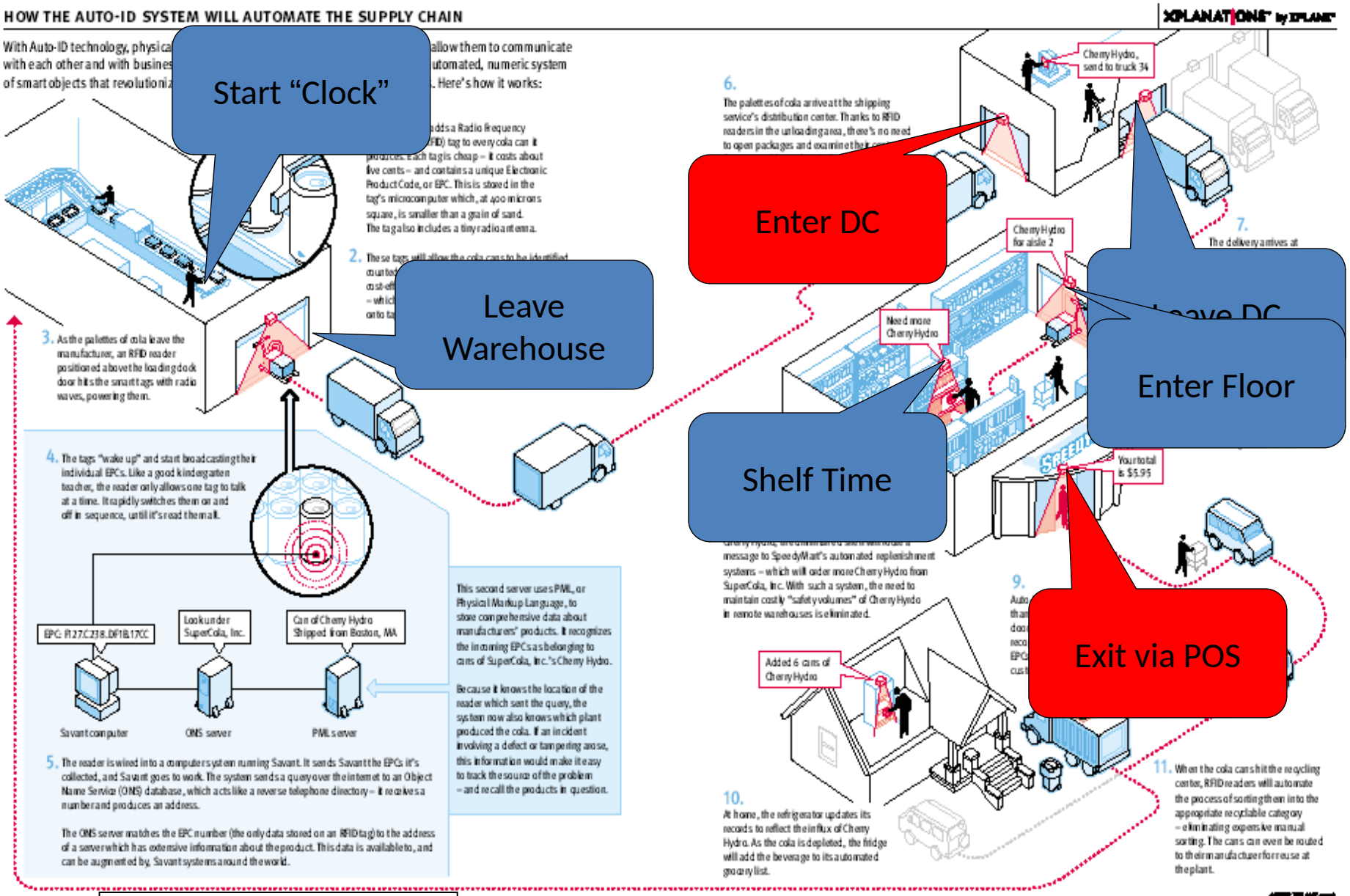




Internet of Things
(RFID tags)

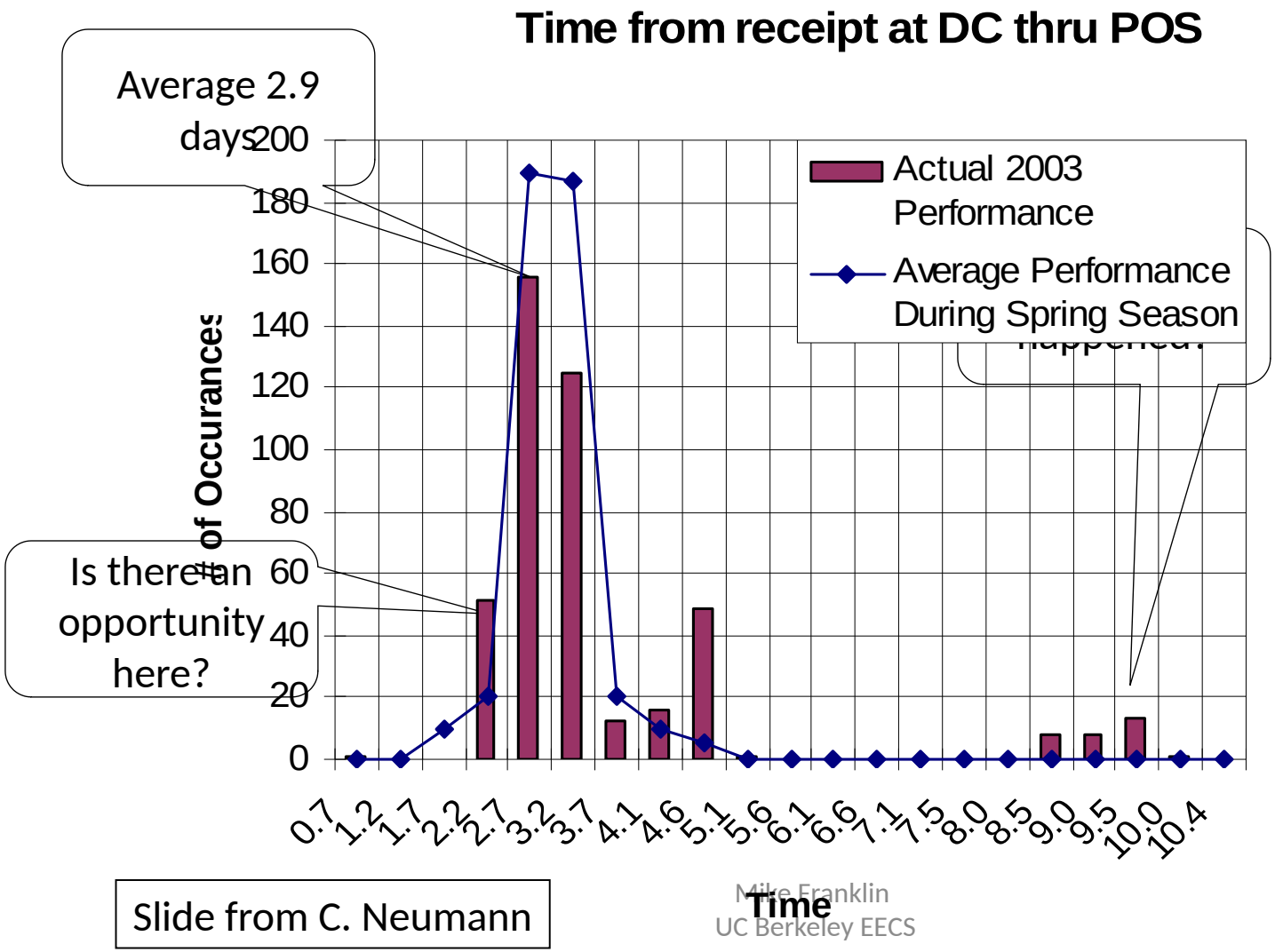
Tag ID	Responses	Timestamp
8576 2387 2345 8678	9	11:07:05
8576 4577 3467 2357	1	11:07:05
8576 3246 3267 5685	7	11:07:06

Note: # responses can be used to estimate signal strength



Slide from C. Neumann

Example: Velocity thru Retail Supply Chain (from Oat Sys)



Slide from C. Neumann

Protein Data Bank

```
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;
```

Files

- What is a file?
  - A **file** is a named sequence of **bytes**
    - Typically stored as a collection of pages (or blocks)
  - A **filesystem** is a collection of files organized within an hierarchical namespace
    - Responsible for laying out those bytes on physical media
    - Stores file metadata
    - Provides an API for interaction with files
  - Standard operations
    - open()/close()
    - seek()
    - read()/write()

Files

- Hierarchical namespace
  - / is known as the root of a filesystem
  - On Linux, the Filesystem Hierarchy Standard specifies which files live where
    - System executables in /usr/bin
    - Log files in /var/log
  - Permissions can be applied to all files beneath a directory
  - Files are not always arranged in a hierarchical namespace
    - Content-addressable storage (CAS)
    - Often used for large multimedia collections

Files

- Byte sequence

$ cat tobits.py

```
import sys

with open(sys.argv[1], "rb") as f:
    byte = f.read(1)
    while byte:
        sys.stdout.write(bin(ord(byte))[2:].zfill(8))
        byte = f.read(1)
```

#To run

python tobits.py <file_name>

Files

- Bit sequence

```
dhcp-47-129:DataScienceS14 Michael$ python tobits.py tobits.py
01101001011011010111000001101111011100100111010000100000011100110111100101110011000010100000101000
11101110110100101110100011010000010000001101111011100000110010101101110001010000111001101111001010
11001100101110011000010111001001100111011101100101101100110001010111010010110000100000001000100111
10010011000100010001000101001001000000110000101110011001000000110011000111010000010100000100101100
00100111100101110100011001010010000000111101001000000110011000101110011100100110010101100001011000
10000101000001100010010100100001010000010010111011101101000011010010110110001100101001000000110000
10011110010111010001100101001110100000101000001001000010010111001101111001011100110010111001110011
10111010001100100011011110111010101110100001011100111011101110010011010010111010001100101001010000
01100010011010010110111000101000011011110111001001100100001010000110001001111001011101000110010100
01010010010100101011011001100100011101001011101001011100111101001100110011010010110110001101100000
10100000111000001010010010100100001010000010010000100101100010011110010111010001100101001000000001
11101001000000110011000101110011100100110010101100001011001000010100000110001001010010000101000000
10010111001101111001011100110010111001110011011101000110010001101111011101010111010000101110011110
11101110010011010010111010001100101001010000010001001011100011011100010001000101001000010101
```

Files

- Byte sequence

```
dhcp-47-129:DataScienceS14 Michael$ ls -l tobits.py
-rw-r--r--  1 Michael  admin  169 Feb  3 13:30 tobits.py
dhcp-47-129:DataScienceS14 Michael$ wc tobits.py
       8      17     169 tobits.py
```

File Formats

- Considerations for a file format
  - Data model: tabular, hierarchical, array
  - Physical layout
  - Field units and validation
  - Metadata: header, side file, specification, other?
  - Plain text or binary
  - Encoding: ASCII, UTF-8, other?
  - Delimiters and escaping
  - Compression, encryption, checksums?
  - Schema evolution