Getting and Cleaning Data

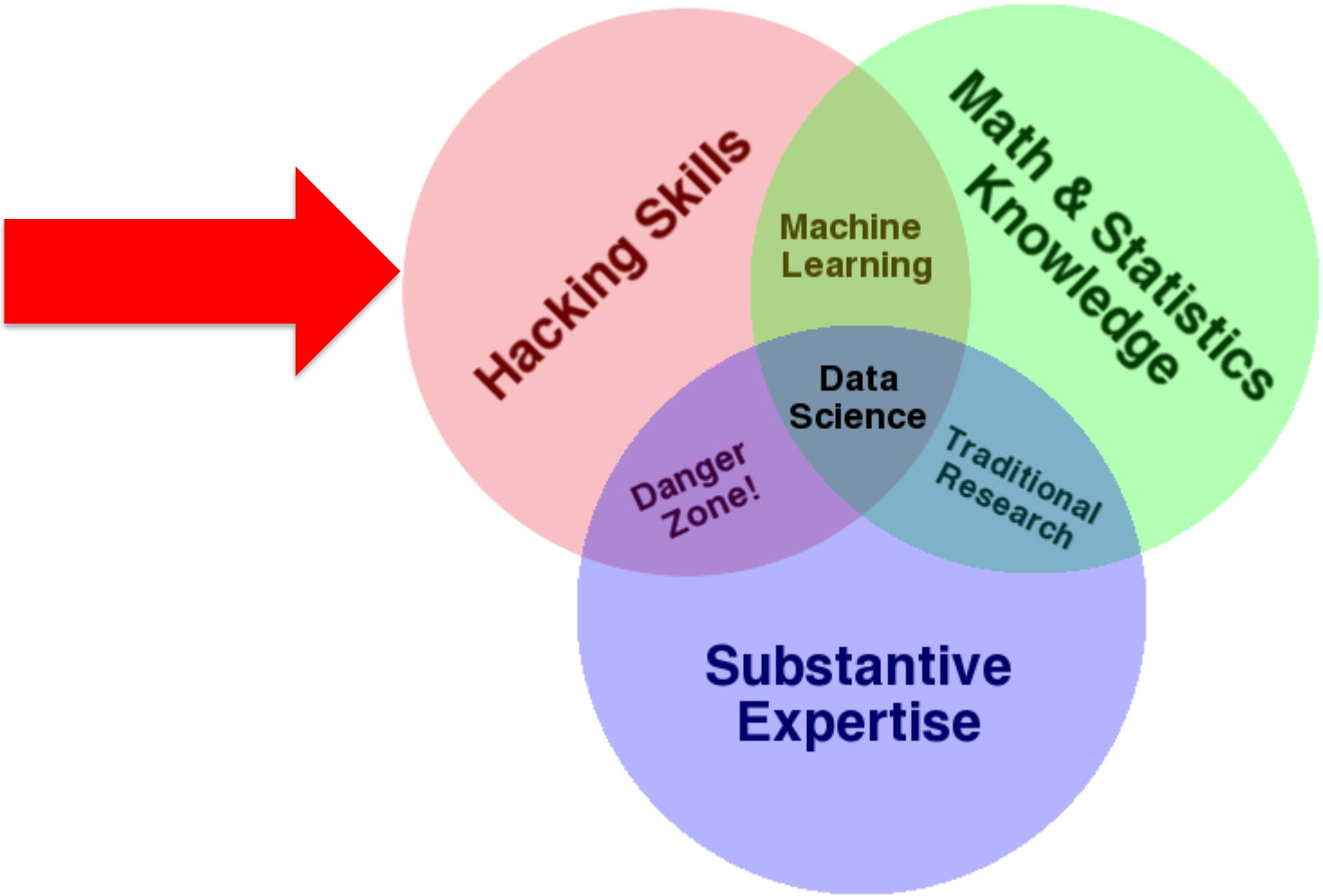
Dr Muhammad Atif Tahir

Including notes from John Canny, Xing Su

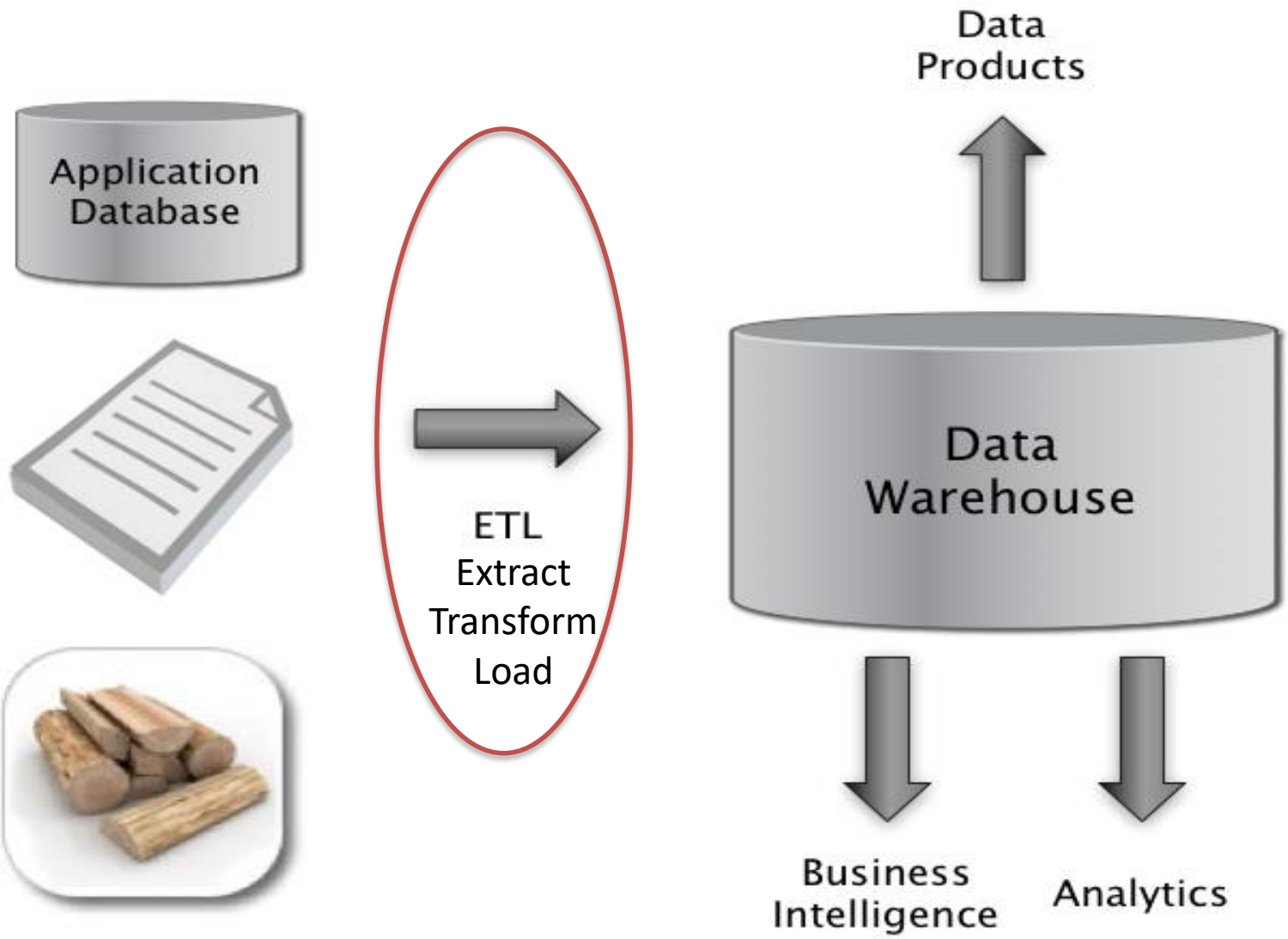
Outline

- Data Science – Why all the excitement?
  - examples
- Where does data come from
- So what is Data Science
- Doing Data Science
- About the course
  - what we'll cover
  - requirements, workload etc.

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, web site, HDFS…
  - Sinks: Python, R, SQLite, RDBMS, NoSQL store, files, HDFS…

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** (for files or network)

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, Matlab

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

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**

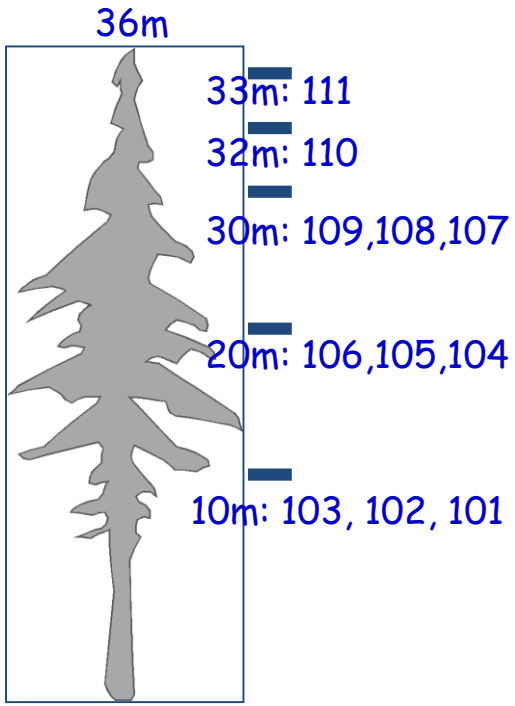
Tabular Data

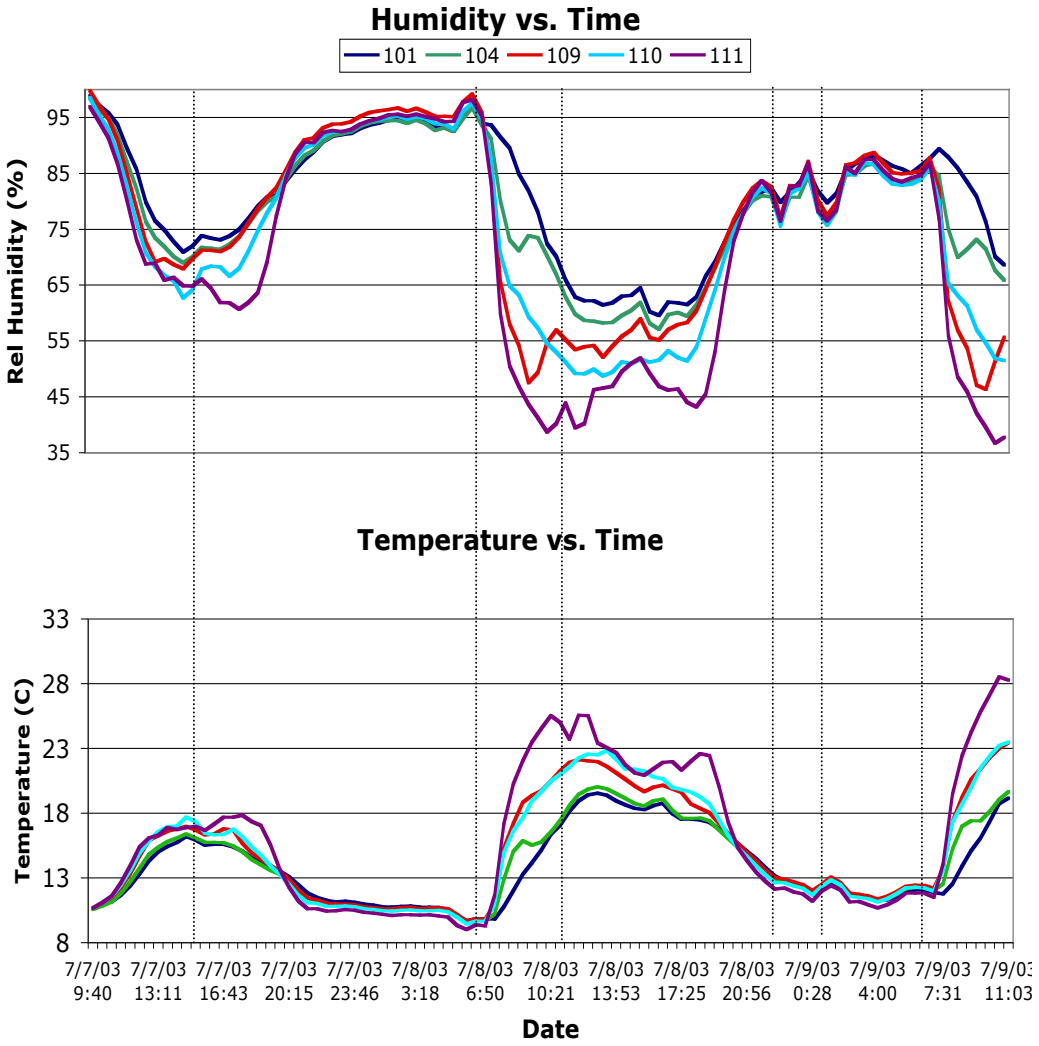
	A	B	C	D	E	F	G	H	I
1	rank	company	cik	ticker	sic	state_location	state_of_incorporation	revenues	profits
2	1	Wal-Mart Stores	104169	WMT	5331	AR	DE	421849	16389
3	2	Exxon Mobil	34088	XOM	2911	TX	NJ	354674	30460
4	3	Chevron	93410	CVX	2911	CA	DE	196337	19024
5	4	ConocoPhillips	1163165	COP	2911	TX	DE	184966	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

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;
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





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"`

Rest from Introducing Data Science by Davy Cielen Arno D. B. Meysman Mohamed Ali

Chapter 2