Northeastern University

CS6020: Collecting, Storing, and Retrieving Information Systems

Introduction & Essential Concepts

Lesson 1

INTRODUCTION

Lesson Objectives

- After completing this lesson, you are able to:
  - specify the role of the data scientist and data science in decision making
  - express the overarching principles of collecting, storing, and retrieving data
  - explain the course structure

The Value of Data

- Data drives decision making in most organizations, *e.g.*,
  - where to locate a new franchise,
  - what customers to target in marketing,
  - where bottlenecks exist in a process,
  - how customers feel about a product,
  - and so forth.

The Role of the Data Scientist

Data Scientists turn data into actionable information.

Data for Analytics

- Data needs to be in a format that allows for qualitative, quantitative, and statistical analysis.
- In an ideal world, data looks like this:

A	B	C	D	E	F
REGION	MARKET	STORE	IN BALANCE DATE	FISCAL PERIOD	MODEL
North	Great Lakes	65061011	01/03/03	200205	4055T
North	Shenandoah Valley	62067017	01/03/03	200205	2500P
North	Shenandoah Valley	32139049	01/03/03	200205	2500C
North	New England	2004014	01/03/03	200205	4055T
North	New England	72074014	01/03/03	200205	4500C
North	New England	12011011	01/03/03	200205	3002P
North	New England	2105015	01/03/03	200205	2500P
North	New England	22022012	01/03/03	200205	4055T
North	New England	22022012	01/03/03	200205	3002C
North	New York South	12118068	01/03/03	200205	4500C

But Data Looks Often Like This

- Data is often unformatted, formatted in a way that is not conducive to analysis, or is missing critical pieces…

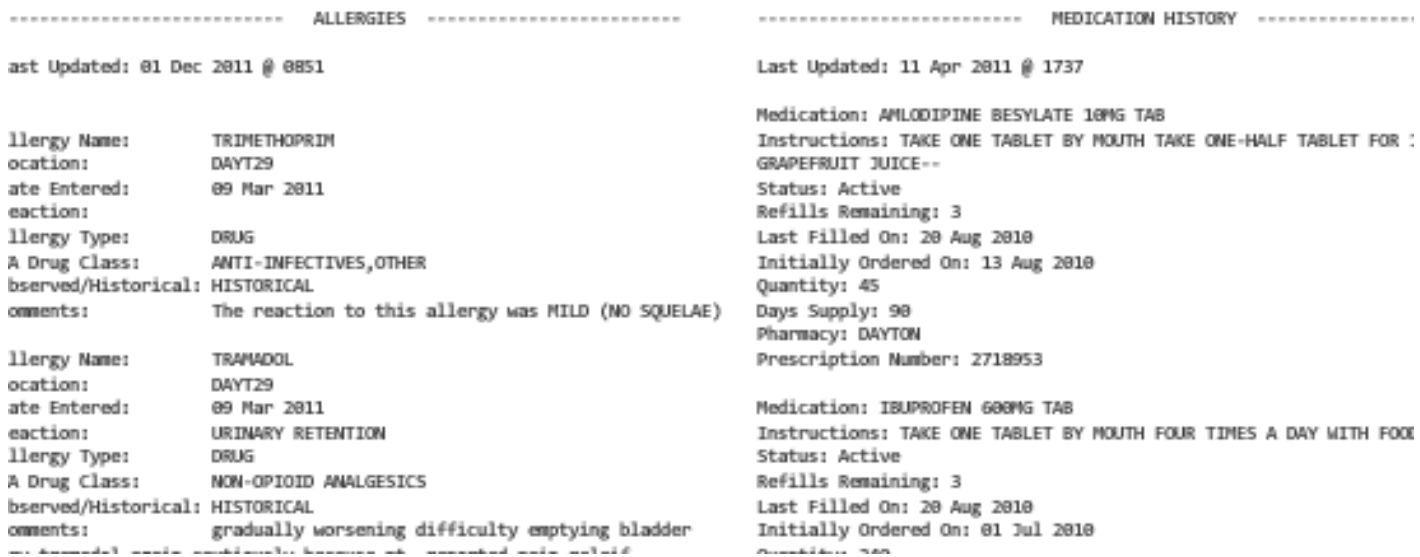

```
  } ],
  "metadata" : {
    "custom_fields" : {
      "TEST" : {
        "CFPB1" : ""
      }
    },
    "renderTypeConfig" : {
      "visible" : {
        "table" : true
      }
    },
    "availableDisplayTypes" : [ "table", "fatrow", "page" ],
    "rdfSubject" : "0",
    "rowIdentifier" : 53173967
  },
  "owner" : {
    "id" : "dfzt-mv86",
    "displayName" : "CFPB Administrator",
    "roleName" : "publisher",
    "screenName" : "CFPB Administrator",
    "rights" : [ "create_datasets", "edit_others_datasets", "e
"view_domain", "view_others_datasets", "create_pages", "edit_p
  },
```



```
  </owner>
▼<rights>
    <rights>read</rights>
  </rights>
▼<tableAuthor id="54a3-qyun" displayName="
  ▼<rights>
      <item>create_datasets</item>
      <item>edit_others_datasets</item>
      <item>edit_nominations</item>
      <item>approve_nominations</item>
      <item>moderate_comments</item>
      <item>manage_stories</item>
      <item>feature_items</item>
      <item>change_configurations</item>
      <item>view_domain</item>
      <item>view_others_datasets</item>
      <item>create_pages</item>
```

Big Data in Healthcare

- *D+collab* posted a challenge on *GitHub* to reimagine the Patient Record, so it is more analyzable than this typical data format:



```
--------------------------  ALLERGIES  -------------------------

ast Updated: 01 Dec 2011 @ 0851


llergy Name:        TRIMETHOPRIM
ocation:            DAYT29
ate Entered:        09 Mar 2011
eaction:
llergy Type:        DRUG
A Drug Class:       ANTI-INFECTIVES,OTHER
bserved/Historical: HISTORICAL
omments:            The reaction to this allergy was MILD (NO SQUELAE)

llergy Name:        TRAMADOL
ocation:            DAYT29
ate Entered:        09 Mar 2011
eaction:            URINARY RETENTION
llergy Type:        DRUG
A Drug Class:       NON-OPIOID ANALGESICS
bserved/Historical: HISTORICAL
omments:            gradually worsening difficulty emptying bladder
```

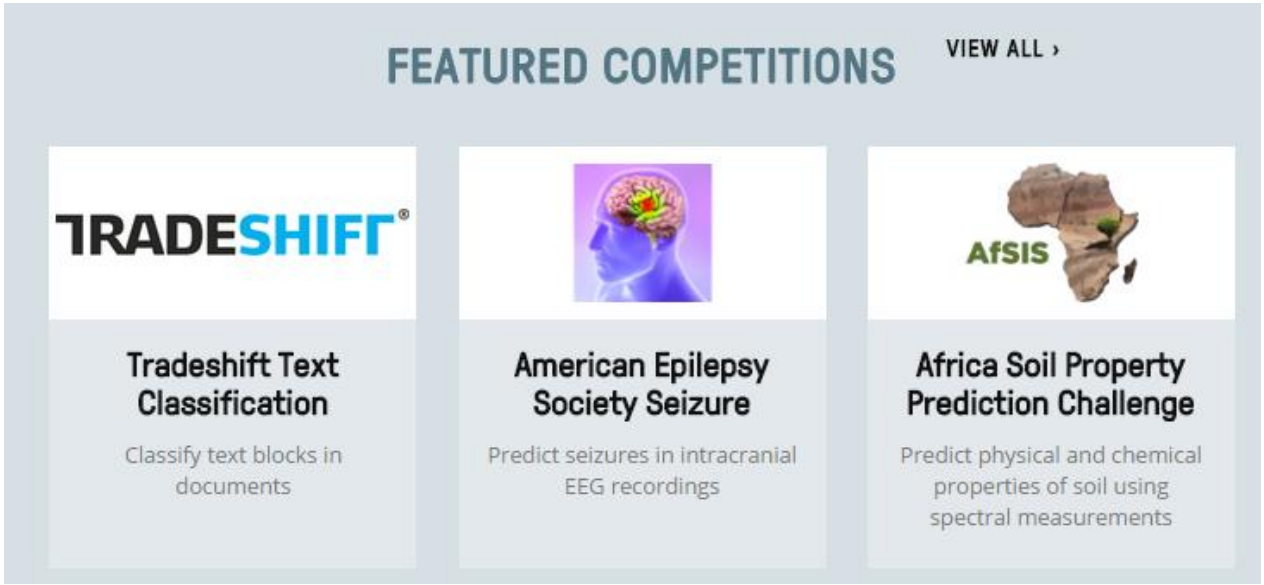
```
--------------------------  MEDICATION HISTORY  ----------------

Last Updated: 11 Apr 2011 @ 1737

Medication: AMLODIPINE BESYLATE 10MG TAB
Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR 
GRAPEFRUIT JUICE--
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 13 Aug 2010
Quantity: 45
Days Supply: 90
Pharmacy: DAYTON
Prescription Number: 2718953

Medication: IBUPROFEN 600MG TAB
Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOO
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 01 Jul 2010
```

Big Data Challenges



- There are numerous challenging "big data" problems.
- *Kaggle.com* runs competitions in which data scientists from all over the world participate.

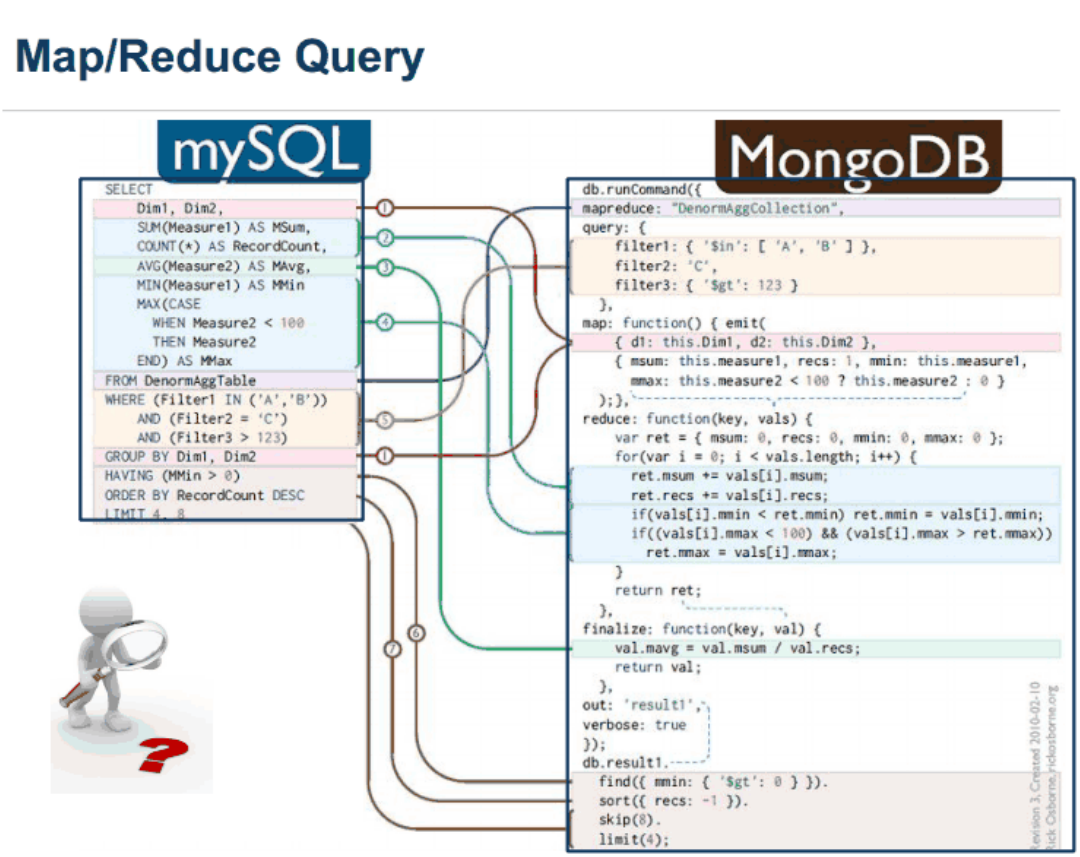
Data Repositories

- Data is stored in a variety for formats and repositories:
  - Simple text files, *e.g.*, CSV
  - Structured text files, *e.g.*, XML, JSON
  - Relational databases, *e.g.*, MySQL, Oracle, SQL Server, JavaDB
  - Non-Relational databases, *e.g.*, CouchDB, MongoDB, Hadoop, Redis, Cassandra
  - Embedded data, *e.g.*, HTML

From blog.sqlauthority.com

MySQL vs MongoDB

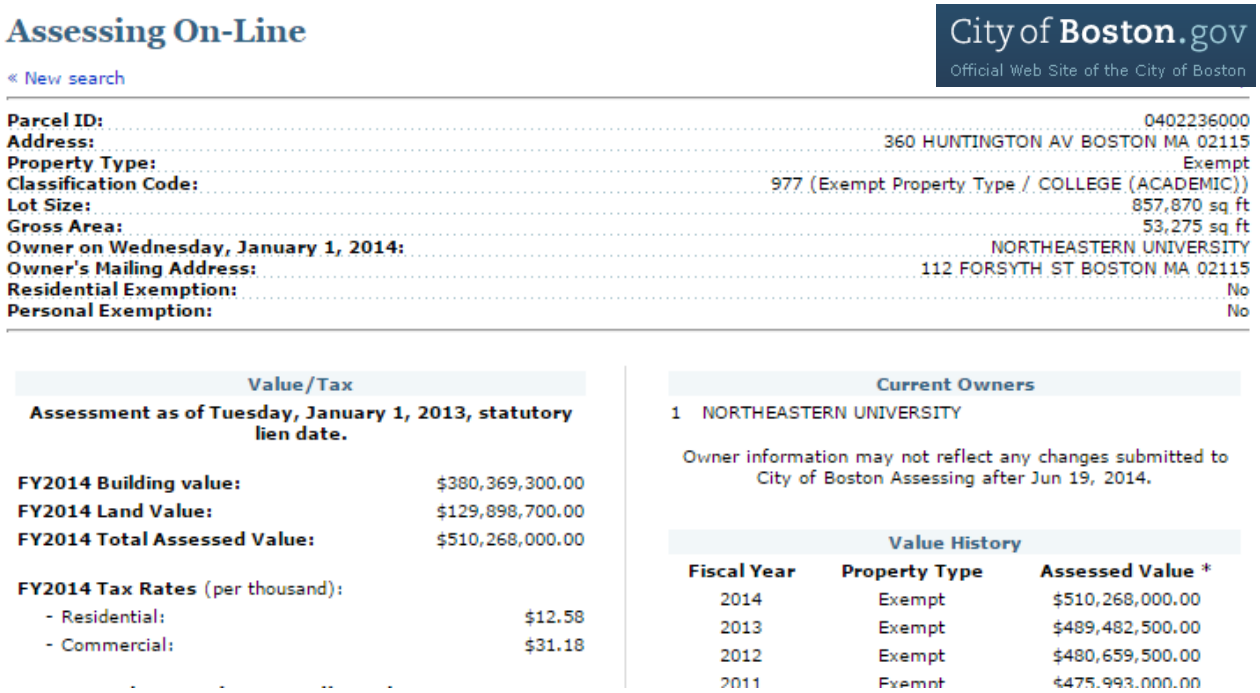
- Here's an example of **MongoDB** and **MySQL** and how the data is stored differently.



http://www.pinaldave.com/bimg/scalebase/scalebase14.png

Embedded Data

- Often data is embedded in documents or on web pages and it must be “scraped”:



Assessing On-Line

« New search

City of Boston.gov
Official Web Site of the City of Boston

Parcel ID:	0402236000
Address:	360 HUNTINGTON AV BOSTON MA 02115
Property Type:	Exempt
Classification Code:	977 (Exempt Property Type / COLLEGE (ACADEMIC))
Lot Size:	857,870 sq ft
Gross Area:	53,275 sq ft
Owner on Wednesday, January 1, 2014:	NORTHEASTERN UNIVERSITY
Owner's Mailing Address:	112 FORSYTH ST BOSTON MA 02115
Residential Exemption:	No
Personal Exemption:	No

Value/Tax

Assessment as of Tuesday, January 1, 2013, statutory lien date.

FY2014 Building value:	$380,369,300.00
FY2014 Land Value:	$129,898,700.00
FY2014 Total Assessed Value:	$510,268,000.00
FY2014 Tax Rates (per thousand):	
- Residential:	$12.58
- Commercial:	$31.18

Current Owners

1 NORTHEASTERN UNIVERSITY

Owner information may not reflect any changes submitted to City of Boston Assessing after Jun 19, 2014.

Value History

Fiscal Year	Property Type	Assessed Value *
2014	Exempt	$510,268,000.00
2013	Exempt	$489,482,500.00
2012	Exempt	$480,659,500.00
2011	Exempt	$475,993,000.00

Map of this Course

- CSV
- JSON
- SQL
- XML
- HTML
- Text

Summary

- In this lesson, you learned that:
  - Data Scientists turn data from various sources into actionable information
  - data must often be "*munged*" and "*wrangled*" to be useful for analysis and visualization
  - there are numerous sources and formats for data
  - different databases store data in different formats
  - this course addresses the challenges of collecting, cleaning, and storing data

Summary, Review, & Questions…