Data Mining

SU 5050
LECTURE 1
JESSICA L. MCCARTY, PH.D.
AND MIKE BILLMIRE, M.S.

What is Data Mining?

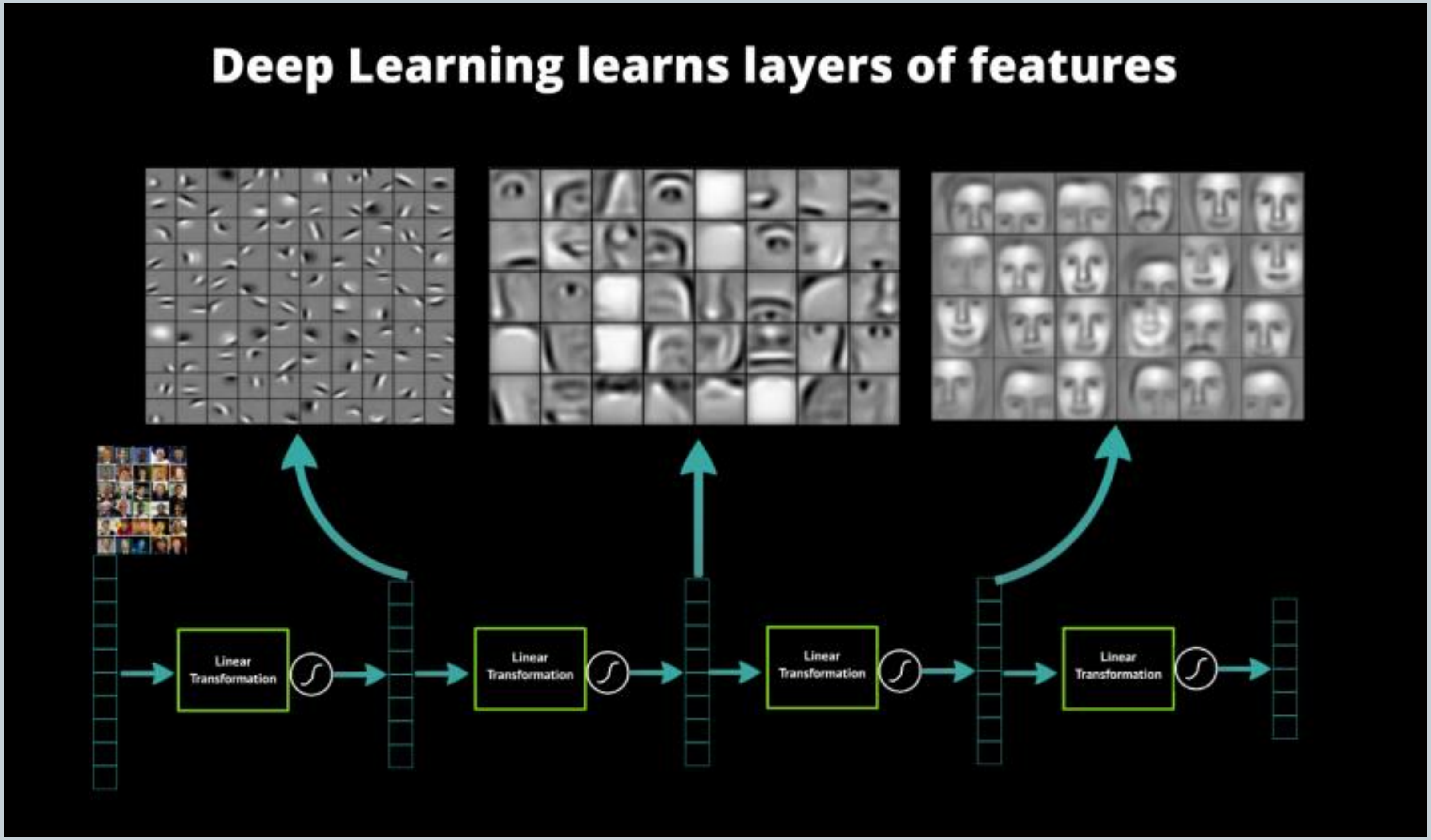
- Recently* coined term for:
  - Confluence of ideas from statistics and computer science
    - Machine learning and database methods
  - Applied to large databases in science, engineering, business

  - *First International Workshop on Knowledge Discovery and Data Mining was in 1995

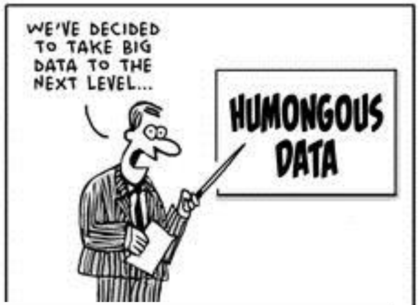
What is Data Mining?

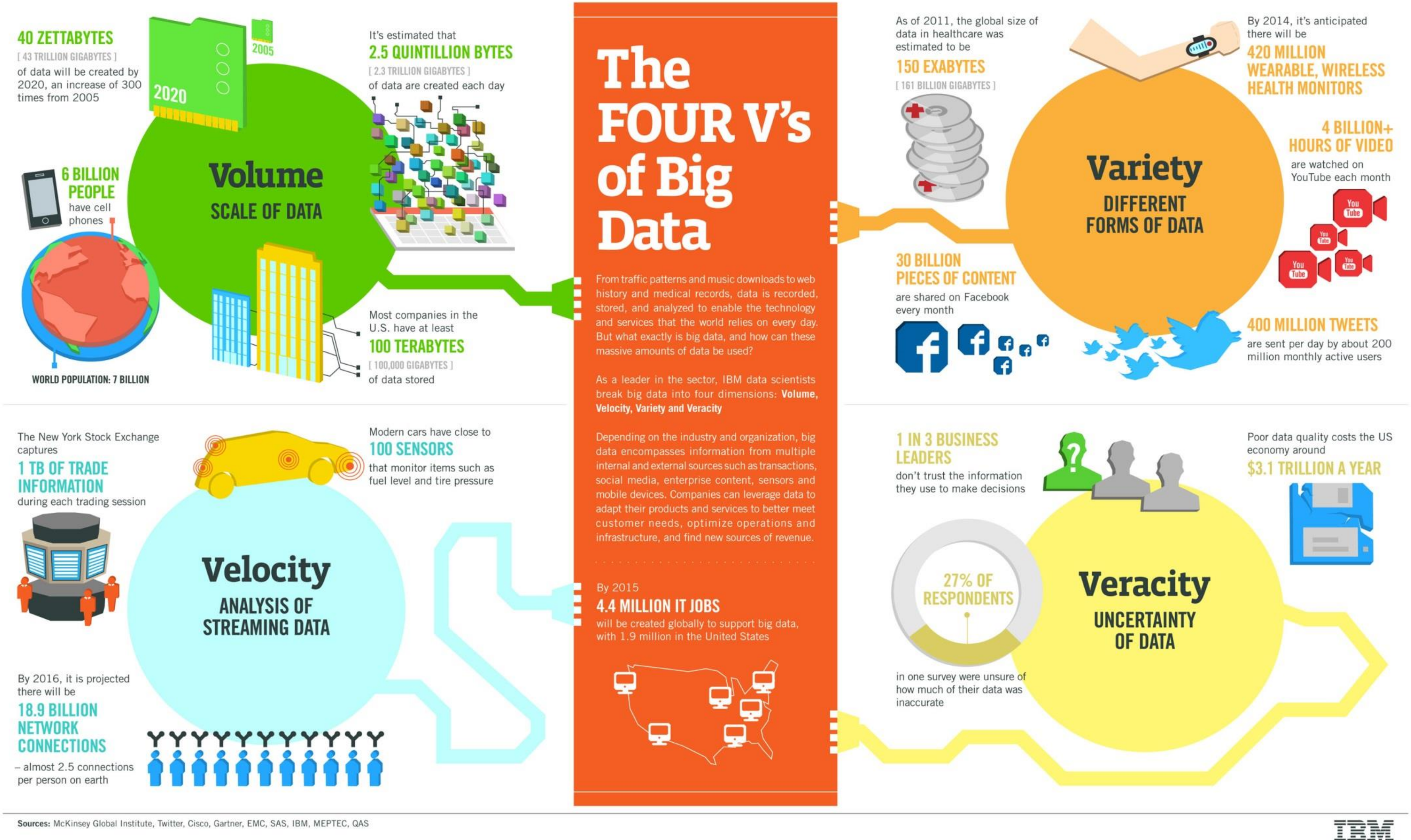
- As less than 20-year-old discipline, is in state of flux
  - Debate over what it is and what it is not
- Terminology is not standard
- Bias, classification, prediction, feature = independent variable
- Target = dependent variable
- Case = exemplar = row

Data Mining's Cousins



Data Mining's Cousins





Data Mining and Deep Learning's Evil Spawn

- Google's Deep Dream Neural Network

- http://googleresearch.blogspot.com/2015/07/deepdream-code-example-for-visualizing.html

Broad and Narrow Definition

- Broad -> traditional statistical methods
- Narrow - > automated and heuristic methods
- Heuristics – exploratory problem-solving techniques that give a non-optimal solution (exhaustive search impractical)
  - Rule of thumb, educated guess, intuitive judgment, stereotyping, common sense
  - Computer Science – technique used when classic methods are too slow; approximate solution after exact solution not found.

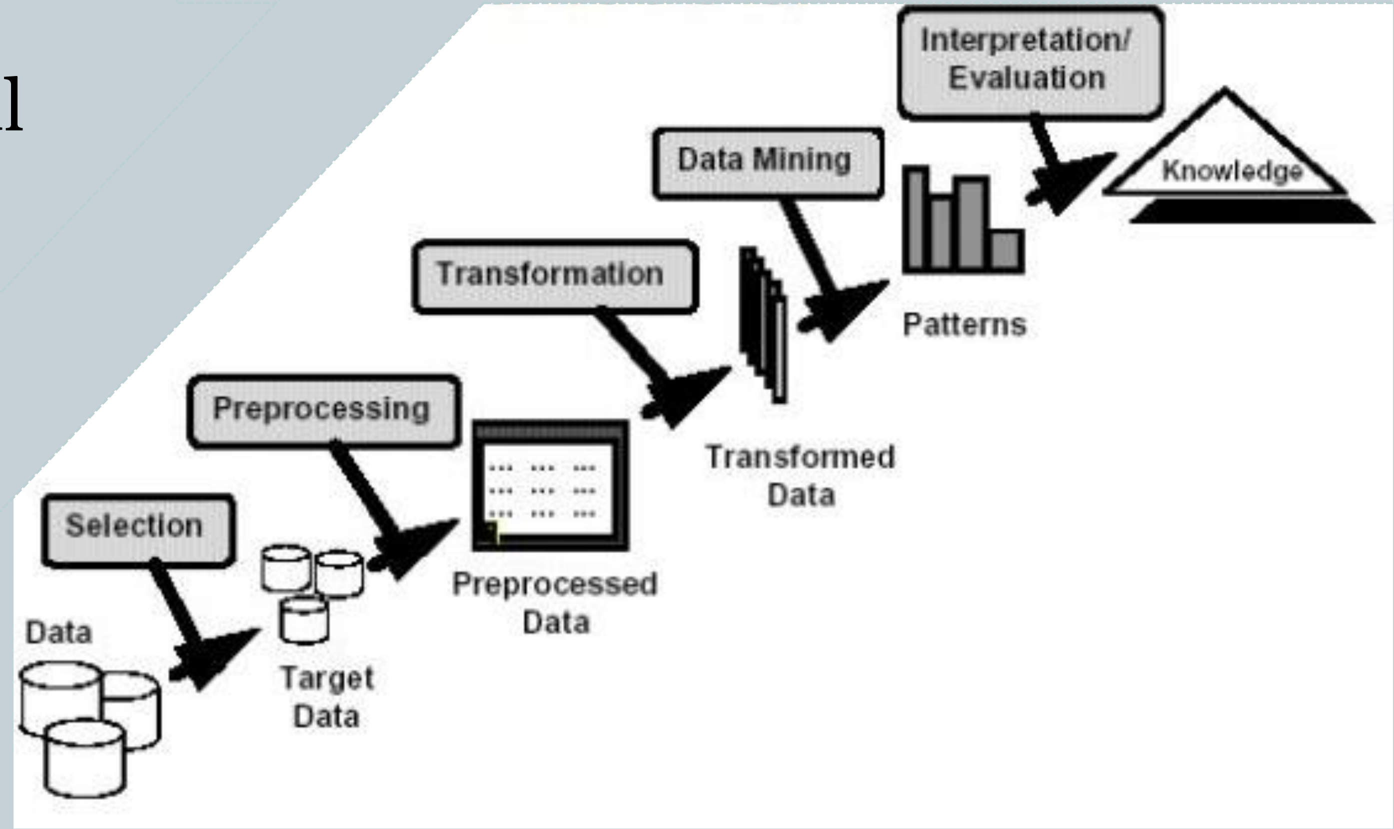
Broad and Narrow Definitions

- Data mining, data dredging, fishing expeditions

- Knowledge Discovery in Databases (KDD)
  - Interactive and iterative
  - Many interactions and feedback loops between steps
  - http://www.usc.edu/dept/ancntr/Paris-in-LA/Analysis/discovery.html

Schema for Data Mining

- Search for meaningful patterns



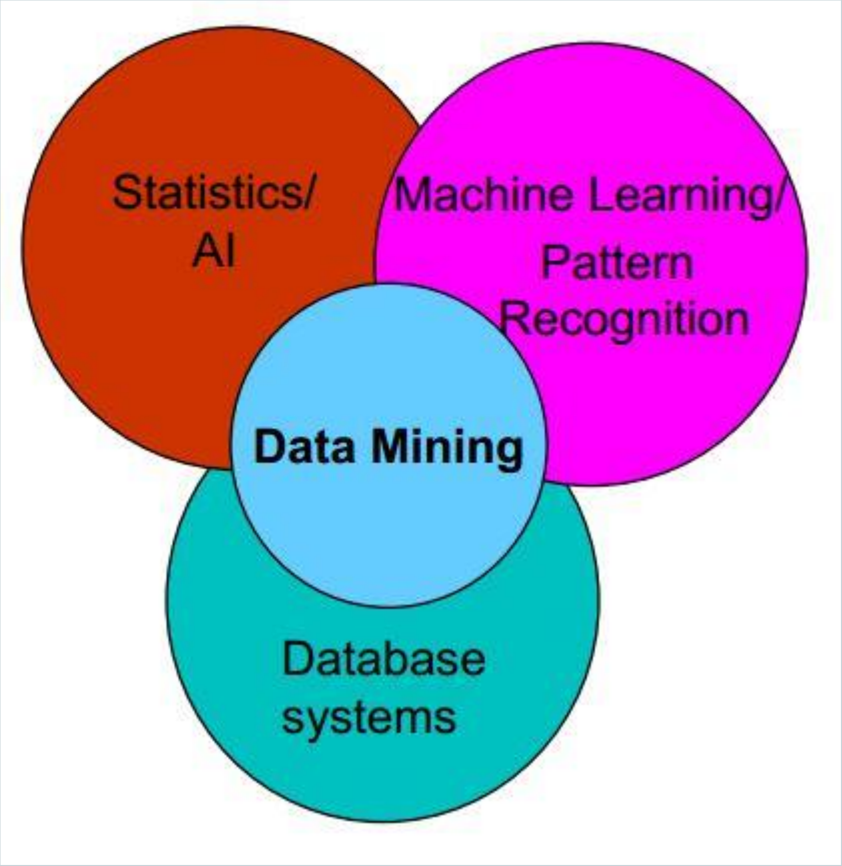
Drivers

- Market -> From focus on products/service to focus on customers
- IT -> From focus on up-to-date balance to focus on patterns in transactions (Data warehouses, cloud)
- Automatic Data Capture of Transactions (bar codes, POS devices, mouse clicks, GPS/locational data)
- Internet -> personalized interactions, longitudinal data

Core Disciplines

- Statistics: Visualization (Descriptive Stats) & Regression, Cluster Analysis (Models)
- Machine Learning: Neural Nets
- Database Retrievals: Association Rules
- Parallel developments: Decision trees, k means, nearest neighbors, Online Analytical Processing (OLAP) Exploratory Data Analysis (EDA)

Core Disciplines

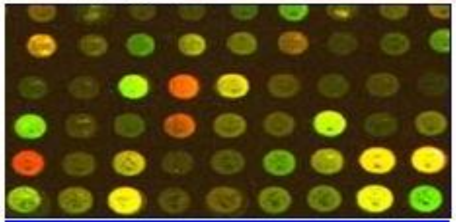
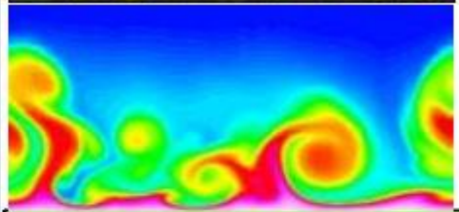
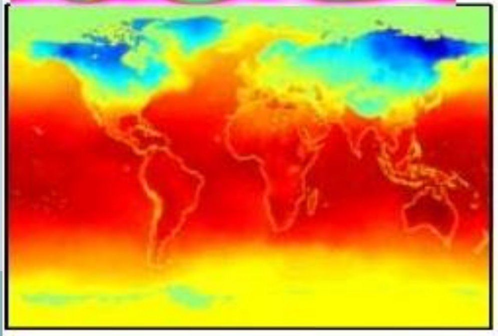


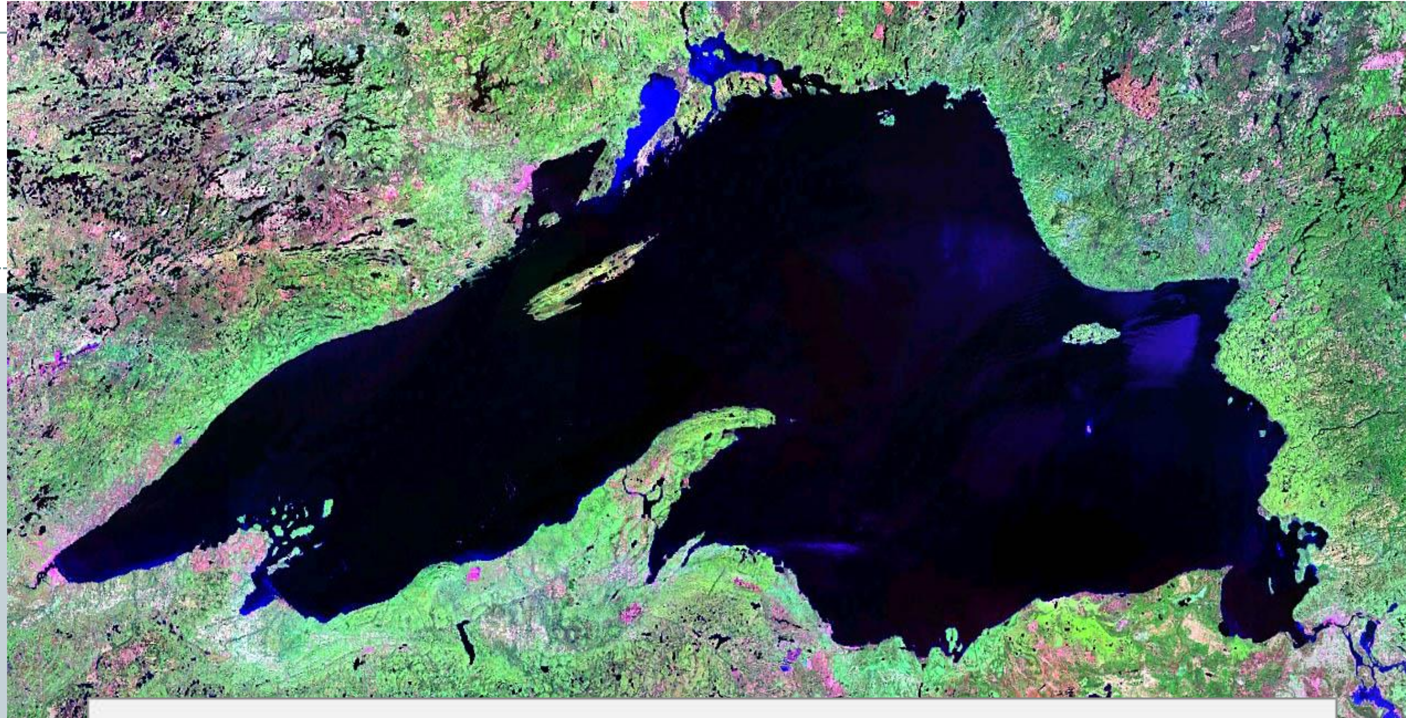
Why Mine Data? Commercial Viewpoint

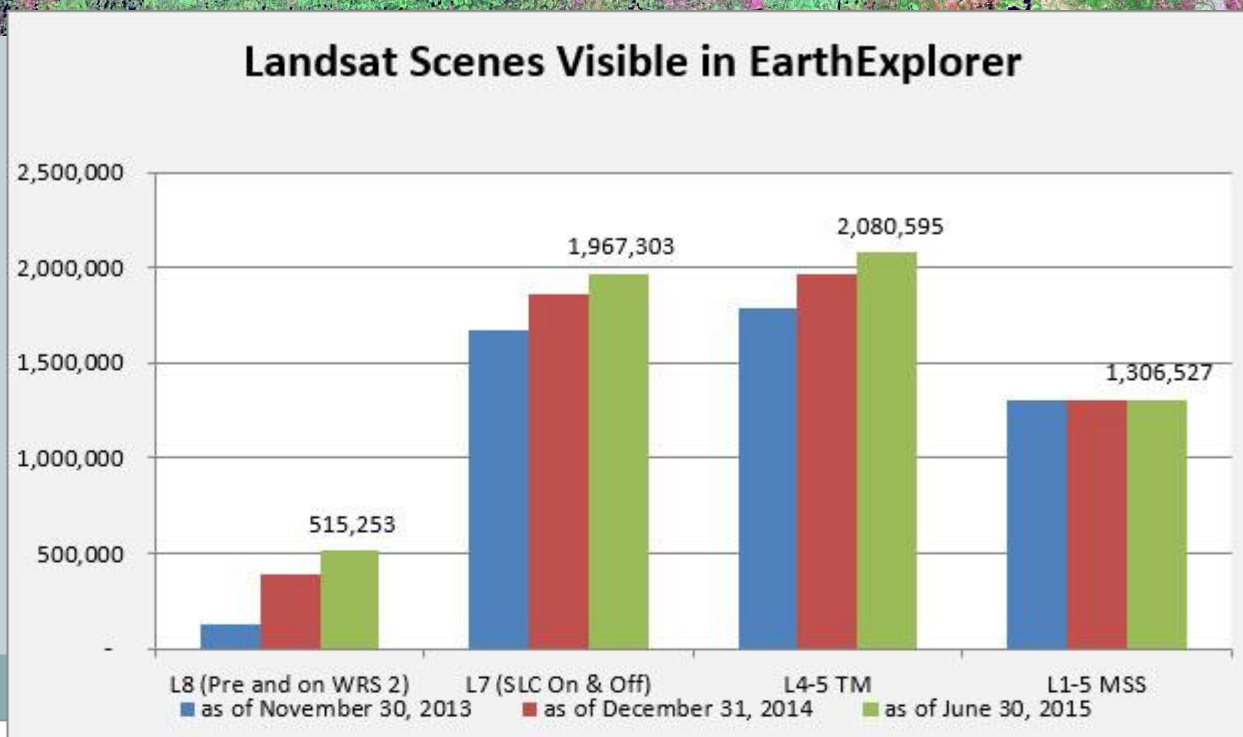
- Loads of data collected and stored
- Computers are cheaper and more powerful
- Competitive pressure is strong
    - Provide better, customized service for an **edge**

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds and quantities (TB/hour globally)
  - Remote sensors on a satellite
  - High powered telescopes
  - Microarrays replicating genome
  - Scientific simulations
- Classifying data
- Hypothesis formation
- Visualizations
- http://usdaapps.devpost.com/

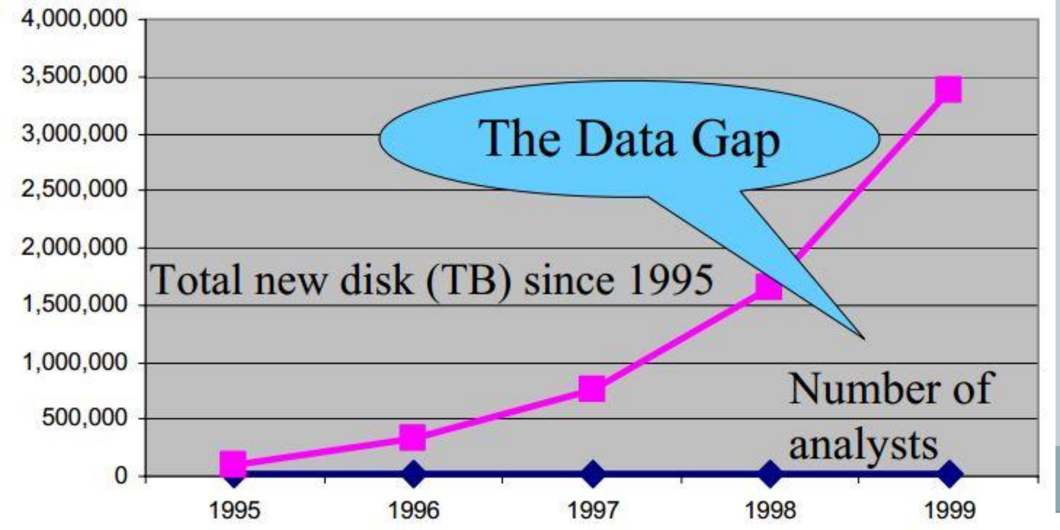








Landsat Scenes Visible in EarthExplorer
2,500,000
2,000,000
1,500,000
1,000,000
500,000
-
515,253
1,967,303
2,080,595
1,306,527
L8 (Pre and on WRS 2)
L7 (SLC On & Off)
L4-5 TM
L1-5 MSS
as of November 30, 2013
as of December 31, 2014
as of June 30, 2015

Motivation for Mining Large Data Sets

- Often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Scope of analyst-based methods different
- Much of the data never analyzed at all



What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Process

1. Develop understanding of application, goals
2. Create dataset
3. Data cleaning and preprocessing
4. Data reduction and projection
5. Choose data mining task
6. Choose data mining algorithms
7. Use algorithms to perform task
8. Interpret and iterate thru 1-7 *if necessary*
9. Deploy: integrate into/create new operational system

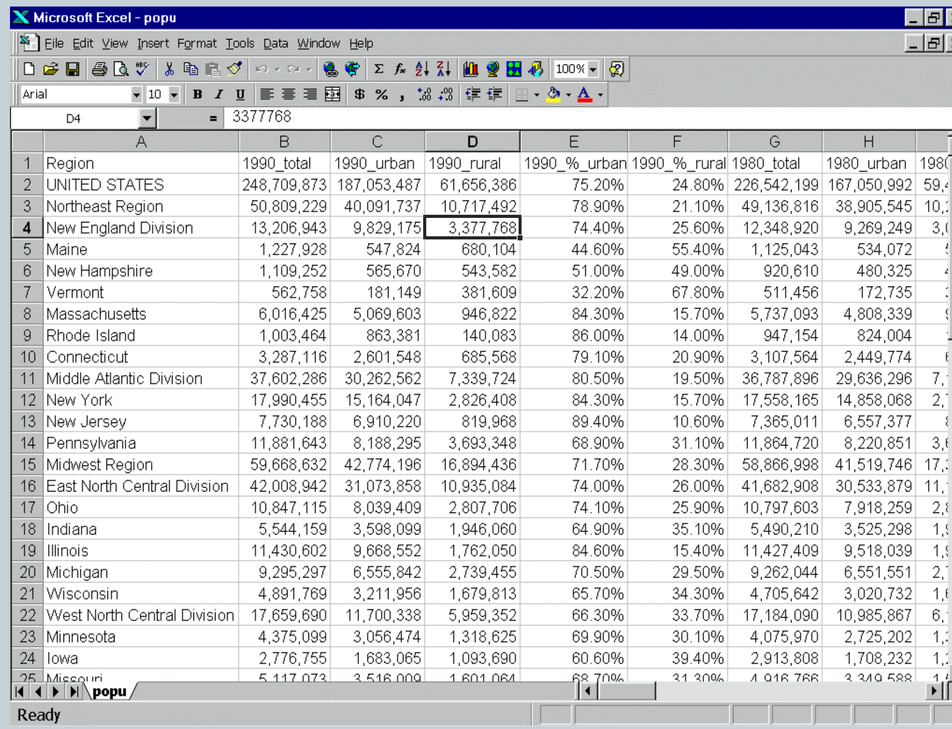
Data Mining

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Typical characteristics of mining data

- "Standard format is spreadsheet
  - Row = observation unit
  - Column = variable



	A	B	C	D	E	F	G	H
1	Region	1990_total	1990_urban	1990_rural	1990_%_urban	1990_%_rural	1980_total	1980_urban
2	UNITED STATES	248,709,873	187,053,487	61,656,386	75.20%	24.80%	226,542,199	167,050,992
3	Northeast Region	50,809,229	40,091,737	10,717,492	78.90%	21.10%	49,136,816	38,905,545
4	New England Division	13,206,943	9,829,175	3,377,768	74.40%	25.60%	12,348,920	9,269,249
5	Maine	1,227,928	547,824	680,104	44.60%	55.40%	1,125,043	534,072
6	New Hampshire	1,109,252	565,670	543,582	51.00%	49.00%	920,610	480,325
7	Vermont	562,758	181,149	381,609	32.20%	67.80%	511,456	172,735
8	Massachusetts	6,016,425	5,069,603	946,822	84.30%	15.70%	5,737,093	4,808,339
9	Rhode Island	1,003,464	863,381	140,083	86.00%	14.00%	947,154	824,004
10	Connecticut	3,287,116	2,601,548	685,568	79.10%	20.90%	3,107,564	2,449,774
11	Middle Atlantic Division	37,602,286	30,262,562	7,339,724	80.50%	19.50%	36,787,896	29,636,296
12	New York	17,990,455	15,164,047	2,826,408	84.30%	15.70%	17,558,165	14,858,068
13	New Jersey	7,730,188	6,910,220	819,968	89.40%	10.60%	7,365,011	6,557,377
14	Pennsylvania	11,881,643	8,188,295	3,693,348	68.90%	31.10%	11,864,720	8,220,851
15	Midwest Region	59,668,632	42,774,196	16,894,436	71.70%	28.30%	58,866,998	41,519,746
16	East North Central Division	42,008,942	31,073,858	10,935,084	74.00%	26.00%	41,682,908	30,533,879
17	Ohio	10,847,115	8,039,409	2,807,706	74.10%	25.90%	10,797,603	7,918,259
18	Indiana	5,544,159	3,598,099	1,946,060	64.90%	35.10%	5,490,210	3,525,298
19	Illinois	11,430,602	9,668,552	1,762,050	84.60%	15.40%	11,427,409	9,518,039
20	Michigan	9,295,297	6,555,842	2,739,455	70.50%	29.50%	9,262,044	6,551,551
21	Wisconsin	4,891,769	3,211,956	1,679,813	65.70%	34.30%	4,705,642	3,020,732
22	West North Central Division	17,659,690	11,700,338	5,959,352	66.30%	33.70%	17,184,090	10,985,867
23	Minnesota	4,375,099	3,056,474	1,318,625	69.90%	30.10%	4,075,970	2,725,202
24	Iowa	2,776,755	1,683,065	1,093,690	60.60%	39.40%	2,913,808	1,708,232
25	Missouri	5,117,073	3,516,009	1,601,064	68.70%	31.30%	4,916,766	3,349,588

- Many rows, many columns
- Many rows, few columns
- Few rows, many columns
- Opportunistic data collect

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

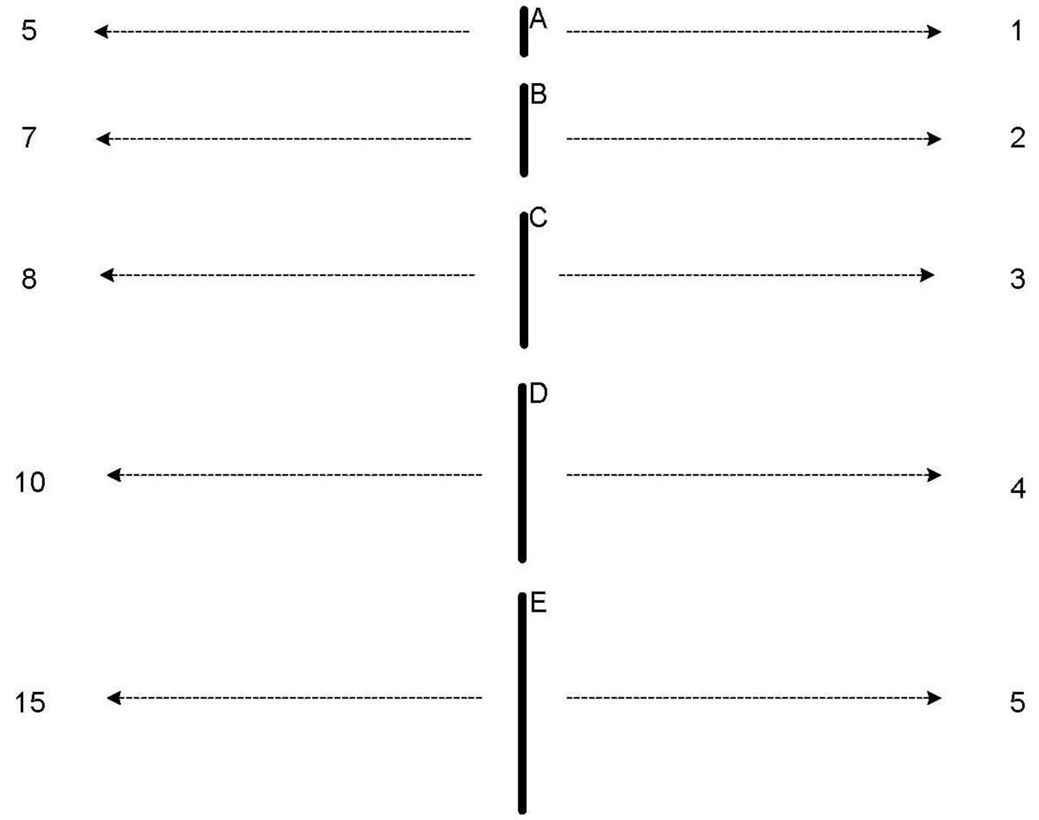
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Measurement of Length

- **The way you measure an attribute is somewhat may not match the attributes properties.**



Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $= \ \neq$
  - Order: $< \ >$
  - Addition: $+ \ -$
  - Multiplication: $* \ /$

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Types of Data Sets

- **Record**
  - Data Matrix
  - Document Data
  - Transaction Data
- **Graph**
  - World Wide Web
  - Molecular Structures
- **Ordered**
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Syllabus

- Class Goals
- Assignments
- Expectations
- Schedule
- Canvas Site

"Drowning in data, yet starving for knowledge."
-- **Anonymous**

"Where is the knowledge we have lost in information?" -- **T.S. Eliot**

Data Mining
SU 5050

Instructors	Jessica L. McCarty, PhD Adjunct Faculty, School of Technology, MTU Research Scientist, Michigan Tech Research Institute mtri.org	Michael Billmire, MS and CMS-GIS/LIS Research Scientist, Michigan Tech Research Institute
Contact	jmccarty@mtu.edu Cell: 502.415.1628 Work: 734.994.7236	mgbillmi@mtu.edu Work: 734.913.6853
Office Hours	Thurs 10 am – 12 pm, Online via Adobe Connect, via Google Hangout, Email or Phone (work, then cell) **Any communication will answered immediately.*	* Please discuss lab issues with Billmire during lab times; email only if you have tried 10 times and can NOT make it work.
Class Meets	Online Lecture: Mondays and Wednesdays 12:00 to 12:55 pm via http://mtu.adobeconnect.com/datamining/ Online Lab Instruction: Fridays 12:00 to 12:55 pm via http://mtu.adobeconnect.com/datamining/	
Canvas	The Canvas site will be used to distribute pdf copies of lecture slides and lab assignments, online midterm and final, and for online submissions. Lectures will be made available after each class, including links to video recordings (requires Adobe and Flash Player).	
Objectives	This course will be taught in three modules: 1. Overview of current techniques, including theory and applications of data mining and big data for geospatial techniques; 2. Application focuses on open source programming and library development (Python); 3. Writing a research plan suitable for research submission and proof-of-concept study.	
Prerequisites	This course is a lot of work. Lab assignments usually require work outside of class and lab times. The course is designed so that students without a programming and geospatial background can succeed, but previous experience will no doubt be helpful. Although **not required for successful completion** of this course, courses in the following areas can be a helpful background: computer programming, statistics, surveying, remote sensing/GIS.	
Required Readings	1. **Textbook** – Russell, M.A.. 2013. *Mining the Social Web*. Second Edition. Sebastapol, CA: O'Reilly Media, Inc. Available at as Ebook, Print & Ebook, Print: http://shop.oreilly.com/product/0636920030195.do . Instructors have a copy of the Ebook. NOTE: You can save money on your online purchase: http://www.retailmenot.com/view/oreilly.com?c=5659596 2. **Miscellaneous Readings** - from various sources will be available on Canvas.	

Who is this Prof McCarty Person?

- PhD in Geography, University of Maryland (2009)
- Research Scientists and Adjunct Professor at Michigan Tech
- mtri.org
- @jmccarty_geo
- Climate and Carbon, Fire, Air Quality, Land Cover/Land Use Change, Food Security, Regional & Natural Planning, Data Mining, Remote Sensing, GIS

On a personal note

- Native of Eastern Kentucky (Appalachia)
- Izzy Dawg
- Recently moved to Houghton