CSI-777
Principles of Knowledge Mining

Class 1
29 Aug 2018

William G. Kennedy,
PhD, CAPT, USN (Ret.)
Center for Social Complexity
Department of Computational & Data Sciences
College of Science

Introduction

- Questions:
  - first, answers to expected questions
  - second, other questions
- Getting started...

Who are these people?

- Who is the instructor?
  - "Bill"
  - wkennedy@gmu.edu
  - Bio: Navy (BS, MS in CompSci, subs)

    10 yrs at Nuclear Regulatory Commission

    15 yrs at DOE & PhD in IT (AI)

    3yr Postdoc at NRL in cognitive robotics

    11 yrs MASON faculty in CSS (+Psyc)

What is this class?

- an introduction
- graduate level (re grading & motivation)
- a lecture, discussion, hands-on course
- not a programming course, but a data/knowledge mining class

Introduction

- What is expected?
  - Attendance
  - Participation
  - Readings & reviews
  - Exercises
  - Project
  - Grading (point system+)

Introductions

- Undergrad major(s), minor(s), etc.
- Academic interests
- Something else, e.g., favorite movie(?)
- Why this course

(teaming possible for class project)

How will class meetings work?

- Class schedule: Thursdays 4:30-7:10pm
- In class, basic outline:
  - Discussion of previous readings & any exercise
  - Lecture on new material
  - Break(s)?
  - At end of semester, project presentations
- Outside of class (between classes)
  - Some readings & some written reviews (7)
  - Some exercises (5)
  - Contact me with questions, feedback, time of day, etc.

What about syllabus?

What about communications?

- e-mail
  - Mason e-mail

    other
- acknowledgements

Other topics or questions?

Getting Started

Getting Started

- Introduction
  - What knowledge is
  - ”Big Data”
  - Complexity
- Data “preparation”

Introduction: Knowledge

Science Facts

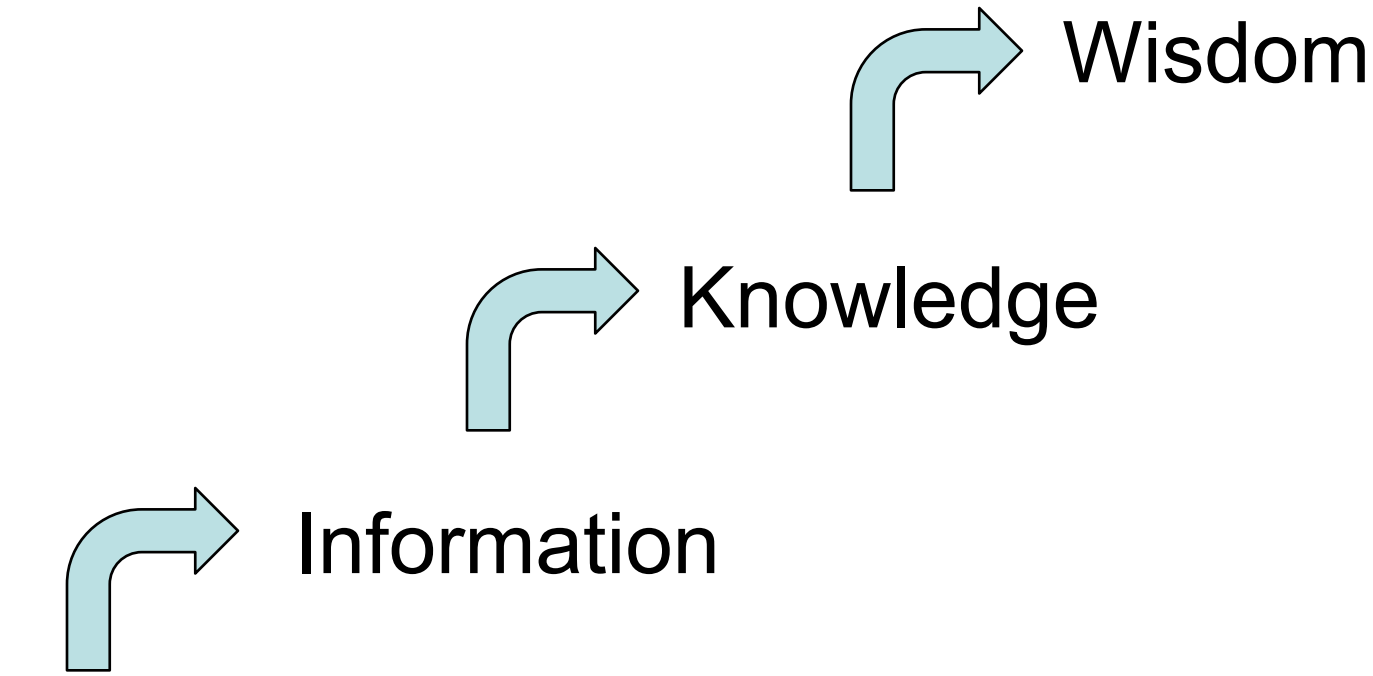
- Noise
- Phenomena (event, condition, …)
- Regularities
  - Nominal
  - Ordinal
  - Interval
  - Ratio

(Stevens, 1946)

Science Facts

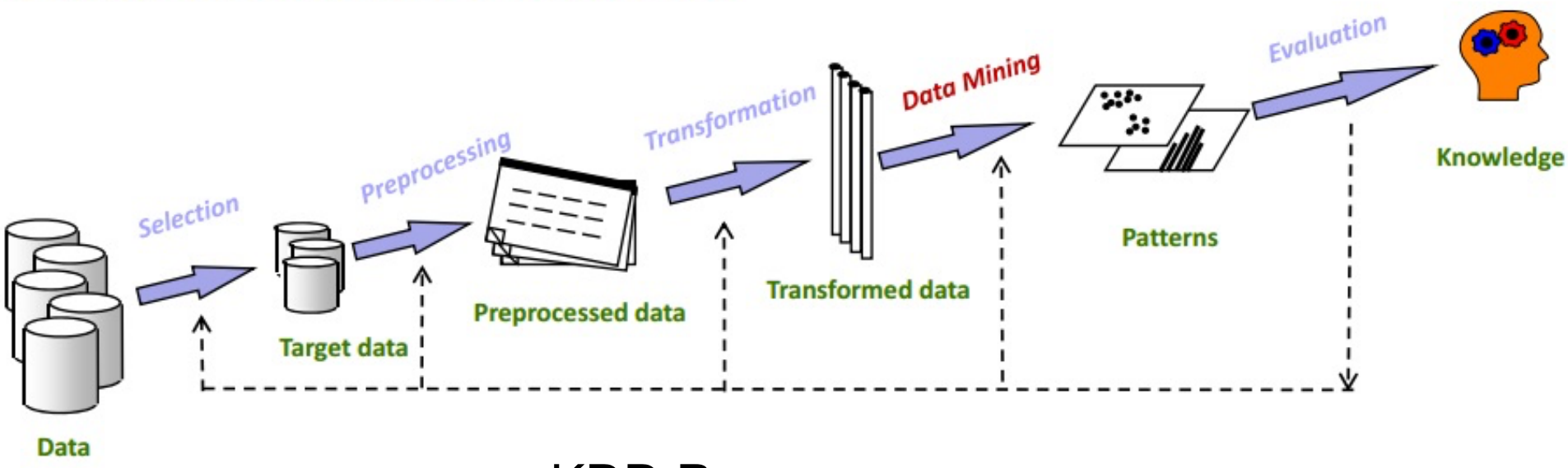
Attribute	Description	Examples	Operations
Nominal	Names (only)	Color(names),sex, IDs, zip codes	Entropy, correlation, shuffle
Ordinal	Ordered objects	Street numbers, mineral hardness, qualitative terms (good, better, best), grades	Sort, percentiles, rank correlation
Interval	Difference meaningful (within range) units	Dates, temperatures (in ^{o}C&F)	Mean, standard deviation, correlation, T&F tests, fixed linear transformations
Ratio	Diff. and ratios meaningful	Temperature (K), Money, counts, age, mass, length, …	Geometric mean, %variation, linear scaling

First, Data…



Knowledge vs. Data Mining

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



KDD Process

(Knowledge Discovery in Databases)

“Big Data”

“Big Data”

Kilobytes	10^3
Megabytes	10^6
Gigabytes	10^9
Terabytes	10^{12}
Petabytes	10^{15}
Exabytes	10^{18}
Zettabytes	10^{21}
Yottabytes	10^{24}

“Big Data”

- Scale steadily increasing
  - Kilobytes, megabytes, gigabytes, terabytes, …
- Earth Observing System of JPL introduced petabytes (10^{15})
- Data collection systems (eg., Large Synoptic Survey Telescope) exabytes (10^{18})

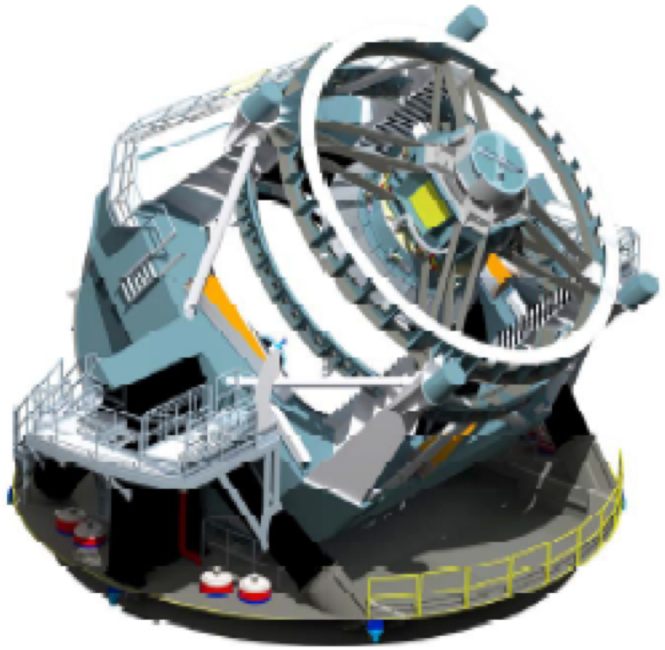
Examples of “Big Data”

- Google processes about 24 petabytes of data each day; it is estimated that Google has 450,000 servers (2009)
- AT&T transfers about 30 petabytes through its network each day (2008)
- German Climate Computing Center stores 60 petabytes of climate data (2009)
- CERN collected 200 petabytes from 800 trillion collisions searching for Higgs boson (2012)
- Facebook’s Hadoop clusters contain more than a 100 petabytes of physical disk space (2012)
- Human brain’s ability to store memories is estimated at about 2.5 petabytes of binary data (2010)
- …

Examples of “Big Data”

Google’s first setup at Stanford circa 1998

Examples of “Big Data”



- Large Scale Synoptic Telescope (LSST) Collects 32 Terabytes nightly or 11.68 Petabytes per Year. An expected 20 year life puts the data collection into the Exabyte range

“Big Data” Issue: ownership

- Lots of Data Available on Internet
  - Facebook – 750,000,000 Unique Monthly Visitors (UMV)
  - Twitter – 250,000,000 UMV
  - LinkedIn – 110,000,000 UMV
  - MySpace – 70,500,000 UMV
  - Google Plus+ - 65,000,000 UMV
  - Others
    - Flickr, Youtube, iTunes, Pinerest
- Commercial loyalty cards
  - Grocery stores
  - Travel websites
  - Airline frequent flier cards
- All collect data on their visitors.

Statistics vs. Data Sciences

- Statistics and Statisticians tend to be defined in terms of Methodology, e.g.
  - Bayesian Methods
  - Nonparametric Methods
  - Sequential Methods
  - Time Series and Stochastic Processes
  - Computational Statistics
  - Biostatistics Methods
- This often leads to a perspective that if the problem doesn't fit into my framework, it is not a statistical problem.

Statistics vs. Data Sciences

- Ed suggests we ought to be Data Centric, i.e. develop methods which are motivated by the data at hand whether it fits standard models or not.

- Wegman, E.J., “On the eve of the 21st century: Statistical science at a crossroads,” *Computational Statistics and Data Analysis*, 32, 239-243, 2000

Statistics vs. Data Sciences

- Most Statistical Methods taught are focused on continuous or discrete numerical data models.
- But Most Useful Applied Methods focus on categorical data.
- Emerging Data Structures
  - Imagery, audio, text data
  - Streaming data
  - Network and graph theoretic data
  - Agent-based simulation data

Statistics vs. Data Sciences

- Traditional Statistics
  - Small to moderate sample sizes
  - Stationary Data
  - Low dimensional
  - Manual computation
  - Mathematically tractable
  - Strong assumptions
  - Well focused questions
  - Closed form algorithms
  - Optimality

- Data Sciences
  - Big data sets – terabytes and more
  - Non-homogeneous
  - High dimensional
  - Computationally intensive
  - Numerically tractable
  - Weak or no assumptions
  - Imprecise questions
  - Iterative algorithms
  - Robustness

Complexity

Descriptor	Data Set Size	Storage Mode
• Tiny	10^2	Piece of Paper
• Small	10^4	A Few Pieces of Paper
• Medium	10^6 - 1Mb	USB Memory Stick
• Large	10^8 - 100Mb	USB Memory Stick
• Huge	10^{10} - 10Gb	USB Memory Stick
• Massive	10^{12} - 1Tb	Hard Disk
• Supermassive	10^{15} - petabyte	Distributed Data Archives

The Huber-Wegman Taxonomy of Data Set Sizes

Wegman, E.J. "Huge data sets and the frontiers of computational feasibility," *Journal of Computational and Graphical Statistics*, 4(4), 281-295, 1995

Algorithmetic Complexity

• $O(n^{1/2})$	Plot a scatterplot
• $O(n)$	Calculate means, variances, kernel densities
• $O(n \log(n))$	Calculate fast Fourier transforms
• $O(nc)$	Calculate singular value decomposition of an r by c matrix; solve a multiple linear regression
• $O(n^2)$	Solve many clustering algorithms.
• $O(a^n)$	NP Complete, Traveling Salesman Problem

Algorithmic Complexity

scale	n	n log(n)	n^2
Tiny	10^2	$2x10^2$	10^4
Small	10^4	$4x10^4$	10^8
Medium	10^6	$6x10^6$	10^{12}
Large	10^8	$8x10^8$	10^{16}
Huge	10^{10}	10^{11}	10^{20}
Massive	10^{12}	$1.2x10^{13}$	10^{24}
Super massive	10^{15}	$1.5x10^{16}$	10^{30}

Number of operations for algorithms of various computational complexity and various data set sizes

Algorithmic Complexity

scale	n	n log(n)	n^2
Tiny	<1sec.	<1sec	<1sec.
Small	<1sec.	<1sec.	<1sec.
Medium	<1sec.	<1sec.	6.67 sec.^{2}
Large	<1sec.	<1sec.	18.5 hrs
Huge	<1sec.	0.67 sec.	21.2 yrs
Massive	6.67 sec.	1.33 min	211 millennia
Super massive	1.85 hrs	1.16 days	2.11×10^{12} yrs >>age of universe

Computational Feasibility on a Intel Core i7 Processor (150 Gigaflops).
All times in seconds unless otherwise stated.

#1 (June 2017) PRC's 93 petaflops: 10^6 x faster… still 10^6 years

#1 (June 2018) US's Summit: 143 pentaflops… still 10^6 years

Intro to ”Data Mining”

- Why Data Mining?
- What is Knowledge Discovery in Databases?
- Potential Applications
  - Fraud Detection
  - Manufacturing Processes
  - Targeting Markets
  - Scientific Data Analysis
  - Risk Management
  - Web Intelligence

Intro to ”Data Mining”

Data Mining: On what kind of data?

- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced
  - Object-relational
  - Spatial, Temporal, Spatiotemporal
  - Text, www
  - Heterogeneous, Legacy, Distributed

Intro to ”Data Mining”

Data Mining: Why now?

– Confluence of multiple disciplines
  - Database systems, data warehouses
  - Machine learning
  - Statistical and data analysis methods
  - Visualization
  - Mathematical programming
  - High performance computing

Intro to ”Data Mining”

Why do we need data mining?

- Large number of records (cases) (10^8-10^{12} bytes)
- High dimensional data (variables) (10-10^4 attributes)

How do you explore millions of records, tens or hundreds of fields, and find patterns?

Intro to ”Data Mining”

Why do we need data mining?

- Only a small portion, typically 5% to 10%, of the collected data is ever analyzed.
- Data that may never be explored continues to be collected out of fear that something that may prove important in the future may be missing.
- Magnitude of data precludes most traditional analysis (more on complexity later).

Data Preparation

KDD and data mining have roots in traditional database technology

As database grow, the ability of the decision support process to exploit traditional (I.e. Boolean) query languages is limited.

– Many queries of interest are difficult/impossible to state in traditional query languages
– “Find all cases of fraud in IRS tax returns.”
– “Find all individuals likely to ignore Census questionnaires.”
– “Find all documents relating to this customer’s problem.”

Massive Data Sets

One Terabyte Dataset

vs

One Million Megabyte Data Sets

Both difficult to analyze

but for different reasons

Massive Data Sets

Data Mining: DM

Massive Data Sets: MD

Data Analysis: DA

Knowledge Discovery in Databases: KDD

$$DM \neq MD$$

$$DM \neq DA$$

$$DA + MD \neq DM$$

Data Mining of Massive Data Sets

Data Mining is a kind of **Exploratory Data Analysis** with **Little or No Human Interaction** using **Computationally Feasible Techniques**, i.e., the Attempt to find Interesting Structure unknown a priori

Massive Data Sets

- Major issues:
  - Complexity
  - Non-homogeneity
- Examples
  - Air Traffic Control: megabyte per min.
  - Highway maintenance: maintenance records over decades, uneven quality & missing data…

Extreme Data

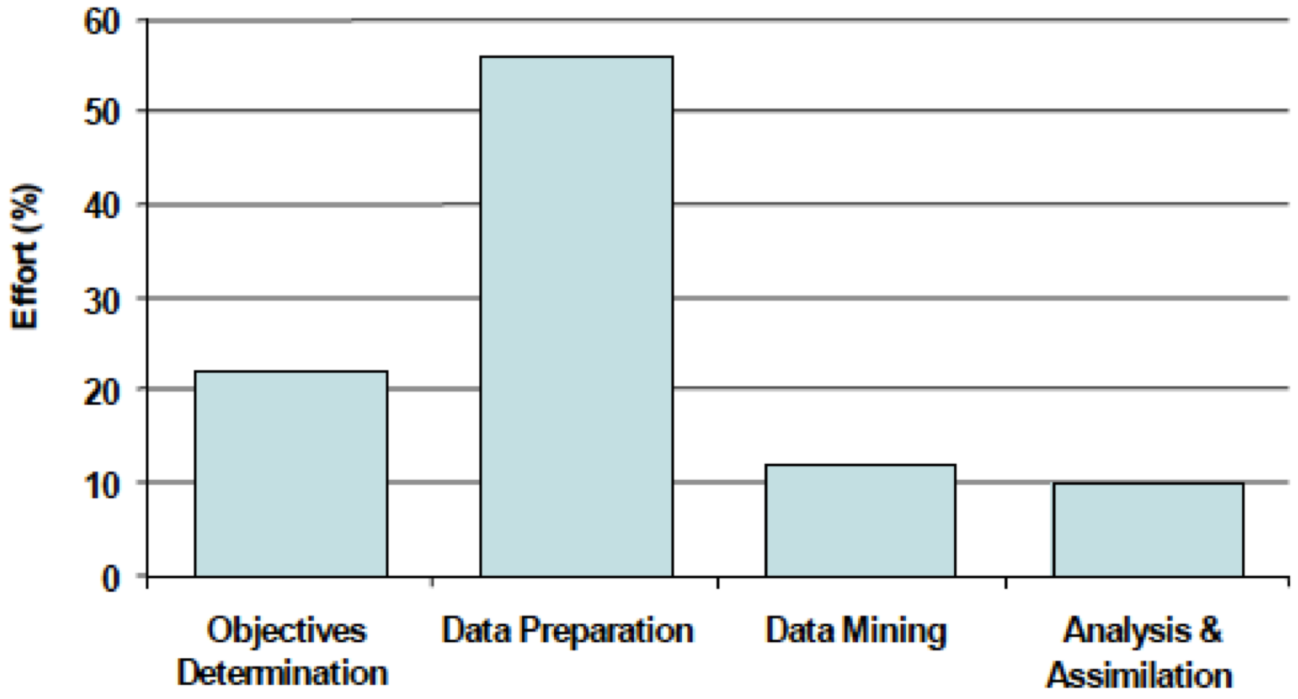
- Personal data on every person in the world:

7,726,791,495

at 4:01pm today

Data Preparation

Data Preparation



Data Preparation

- Data cleaning and quality
- Types of data
- Categorical vs. continuous data
- Problem of missing data
- Problem(?) of outliers
- Dimension reduction, quantization, sampling

Data Quality

- Data may not have any statistically patterns or relationships
- Results may be inconsistent with other data sets
- ”Fake news…”
- Opportunistically collected/biased
- Patterns may be too specific or general

Data Preparation

- Noise
  - Incorrect values?
  - Faulty collection instruments/sensors
  - Data entry errors
  - Technology limitations
  - Naming conventions misused
  - Intentional errors

Data Preparation

- Redundant/stale data
  - Different names in different sources
  - Raw data in one source, processed in another
  - Irrelevant, distracting data costly
  - Changes in variable/collection over time

Data Cleaning

- Remove duplicates (tool-based)
- Missing data filled in (manual, statistical)
- Identify & remove inconsistencies
- Identify & refresh stale data
- Create unique record/case IDs

Data Preparation

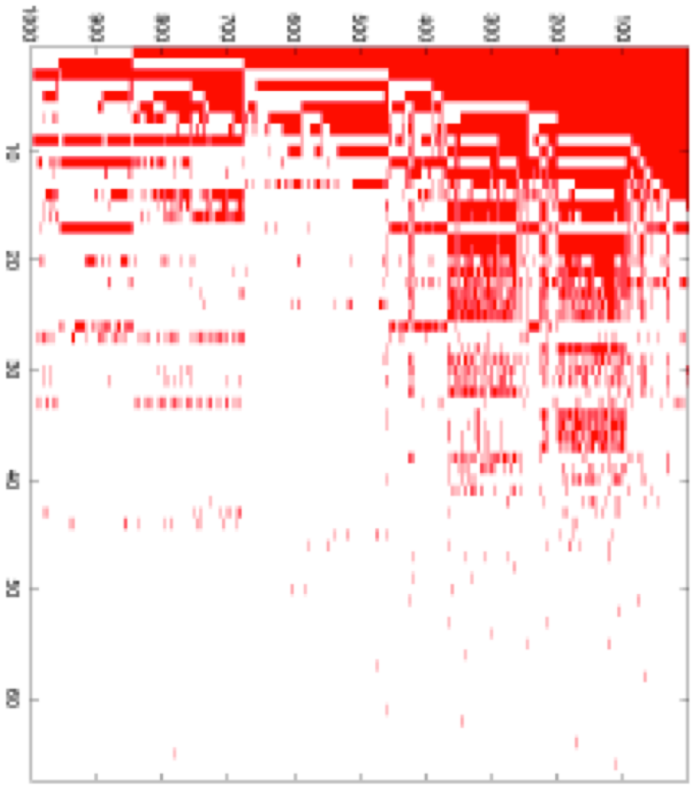
- Categorical vs. continuous data
  - Most statistical theory, many graphics tools for continuous data (only)
  - Data in databases usually categorical
  - Often useful to convert continuous into bins (low, medium, high)

Missing Data

- May or may not be a problem
  - Missing data may be irrelevant to use
  - Filling in data presumes a model
  - Plot missing data…

Missing Data

- Missing Value Plot
  - A plot of variables by cases
  - Missing values colored red
  - Special case of "color histogram" with binary data
  - "Color histogram" also known as "data image"
  - This example is 67 dimensions by 1000 cases



Problem of Outliers

- Easy to detect in low dimensional data
- High dimensional outlier may not show up in low dimensional projections
- Algorithms (min covariance determinant (MCD) or min. volume ellipsoid (MVE)) exponentially complex (expensive)

Data Compression

- Sampling
- Quantization

Sampling

- May be practical rather than exhaustive processing
- Tools can help select nec'y data
- To work, data must satisfy certain conditions (avoid biases)
- Sampling a DBMS can be more expensive than sequential full processing

Data Quantization

- Thinning vs. Binning
  - First thoughts usually statistical sampling
  - Quantization in engineering success story
  - Binning is statistician's quantization

Data Quantization

- Images quantized to 8-24 bits (256-16M)
- Audio quantized to 16 bits (65,536 levels)
- How many?
  - Statistician: few 100
  - Computer scientist: 3
  - Human perception: 7 ± 2

Images from MRI

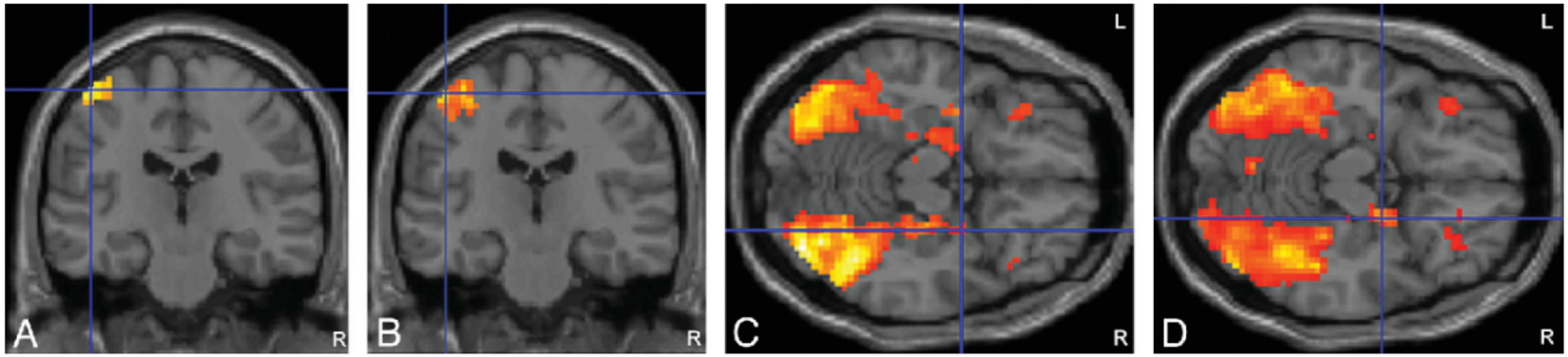
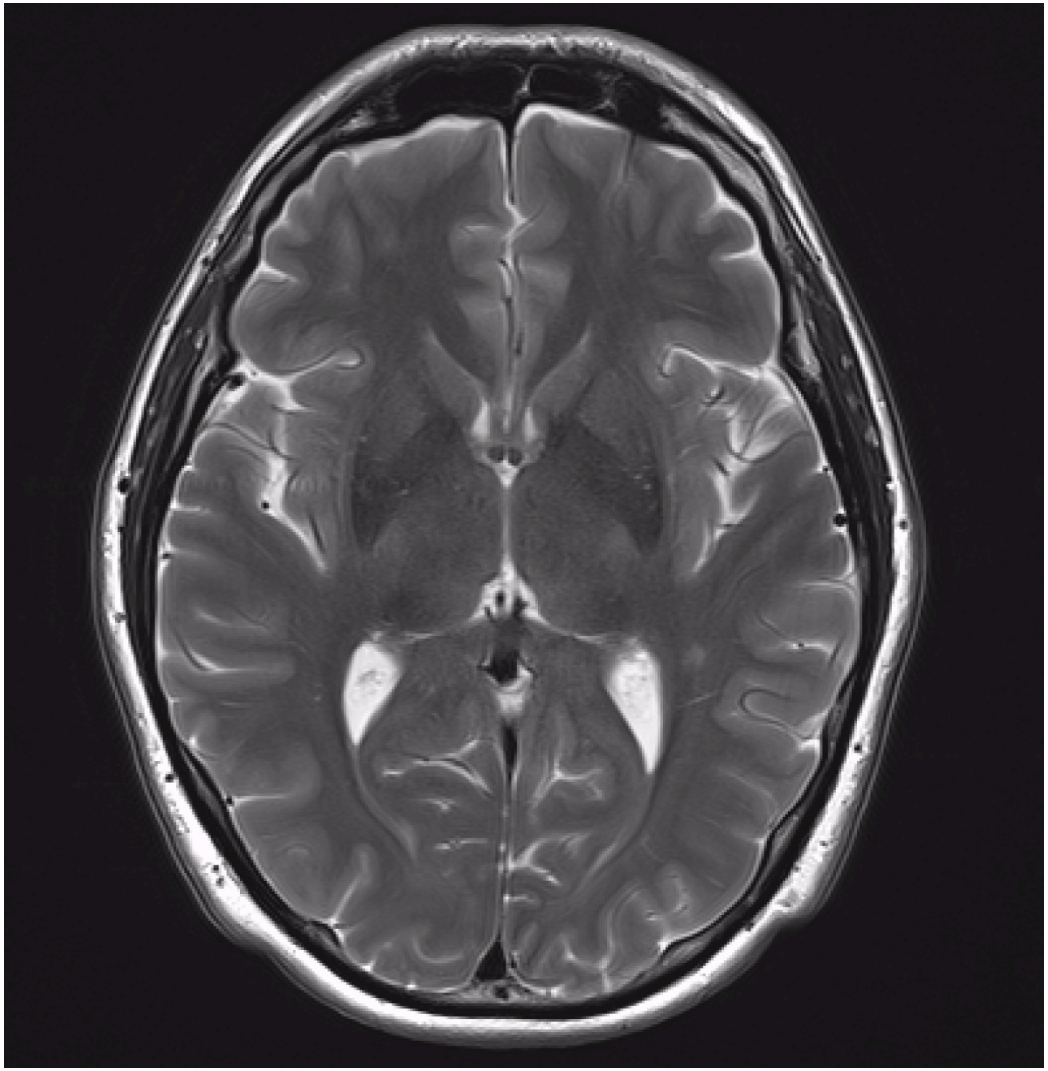


Fig 2. A and B, Examples of left motor cortex activation (crosshairs) in a group analysis for the open 1T (TE 70 ms, A) and the 3T scanner (TE 40 ms, B), (P .05, family-wise error corrected). On both scanners, the activation is well-detected. Notice the larger size of the activation area on the 3T scanner (B). C and D, Examples of right amygdala activation (crosshairs) in a group analysis for the open 1T (TE 70 ms, C) and 3T scanners (TE 20 ms, D), (P .001, uncorrected). Right amygdala activation is well-detected on both scanners. Similarly, extensions of the visual cortex activation and OFC activations can be seen for both scanners. Notice the larger activation areas on the 3T scanner (D). Images are in neurologic orientation.

Test Image from new MRI



Data Quantization

- Binning, at small scale
- Conventions:

  d = dimension

  k = number of bins

  n = sample size

  Typically k<<n

Data Quantization

- Want evaluation of sample to equal evaluation of whole dataset
- “self-consistent”
- Eg., if transformation is convex WRT eval,

  eval(sample) ≤ eval(dataset)

Data Quantization

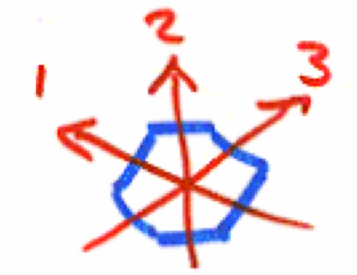
- Distortion is error due to quantization
- Distortion minimized when quantization regions are most like a (hyper) sphere

Geometry-based Quantization

- Need space-filling tessellations
- Need congruent tiles
- Need as spherical as possible

Geometry-based Quantization

- (polytope = shape with flat sides)
- 1D data: only possible polytope is line segment
- 2D data: only polytopes: equilateral triangle, square, and hexagons

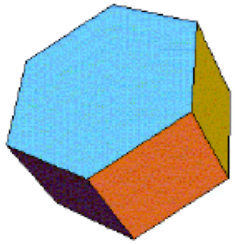


Geometry-based Quantization

- 3D: tetrahedron, cube, hexagonal prism, rhombic dodecahedron, truncated octahedron

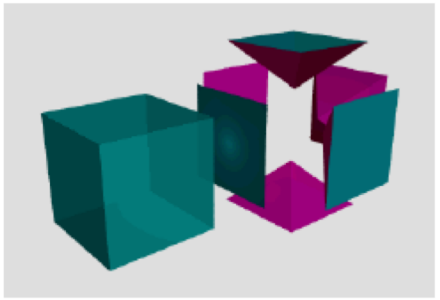
Tetrahedron

Cube



Hexagonal Prism

Truncated Octahedron



Rhombic Dodecahedron

Quantization Strategies

- Optimally (minimizing distortion), want to use roundest polytope in d-dimensions
- Computational complexity always O(n)
- Storage complexity is 3k
- #tiles grows exponentially (the “curse”)

Quantization Strategies

- Summary:
  - Geometric approach good for 4-5 D
  - Adaptive tilings may improve growth rate but may increase distortion
  - Good for large n, but weaker for large d

Quantization Strategies

- Another approach: distance based clustering (in EE, ”vector” quantization)
- Form bins via clustering $O(n^2)$
  - Poor for >>n
  - Not too sensitive to dimension, d
  - Clusters may not be round...
- Bottom line: good for large d, not large n, not particularly useful for “massive” data

Quantization Strategies

- Third approach: Density-based clustering
- Density estimation O(n)
- Not distance-based, not $O(n^2)$
- Roundness may be a problem

Data Quantization

- Analysis on finite subset has theoretical advantages:
  - Analysis less delicate (multiple forms of convergence equivalent)
  - Analysis often more natural (already quantized or categorical)
  - Graphical analysis not much changed because human vision limit ~million pixels

Summary

Data > Info > Knowledge > Wisdom

Big Data

Data preparation

What’s next?

Course Overview (tentative)

- Data preparation
- Databases
- Classification or supervised learning
- Supervised v. Unsupervised learning
- Density estimation
- Color
- Visual data mining
- Mining data streams
- Text data mining

Next Week

- Read and write a review of Rowley (2006)
- Install software

Next Week

Reviews

1. I, too, have read the material. The review isn't "for" me.
2. Make it personal
3. Connect/integrate the material with your previous knowledge
4. To have an informed opinion you have to have read and understand the material and compare it with previous knowledge and maybe facts.
5. Has a minimum length to get something valuable.
6. Has a maximum length to focus on the most important points.
7. Format (single/double space) is not the focus. The content is.

Your Questions?

My Questions

- What you expected?
- How does it serve your interests?
- At right level?

Backup