Mining Health-Related Data

Methods and applications in Research, Public Health, and Patient Care

John H. Holmes, Ph.D.

Center for Clinical Epidemiology and Biostatistics

University of Pennsylvania School of Medicine

CCEB

Why are these guys so happy?

Where we're going today…

- Introduction to databases and warehouses
- Data mining: What is it?
- Output of data mining
- The data mining life cycle
- Data mining applications
- Conclusion

What are we looking for?
The Information Spectrum

Data Information Knowledge Wisdom

160

Data!

160/94

Information

Knowledge

Databases

- Logically coherent collection of data with some inherent meaning
- Databases are designed, built, and populated with data for a specific purpose, for an intended group of users
- Represent some aspect of the real world

“Large” data

- How to define large data
  - » Number of fields
  - » Number of records
  - » Complexity of data model
  - » Breadth of distribution
- Always, the issue is high dimensionality
- Ultimately, large data end up in a centralized resource

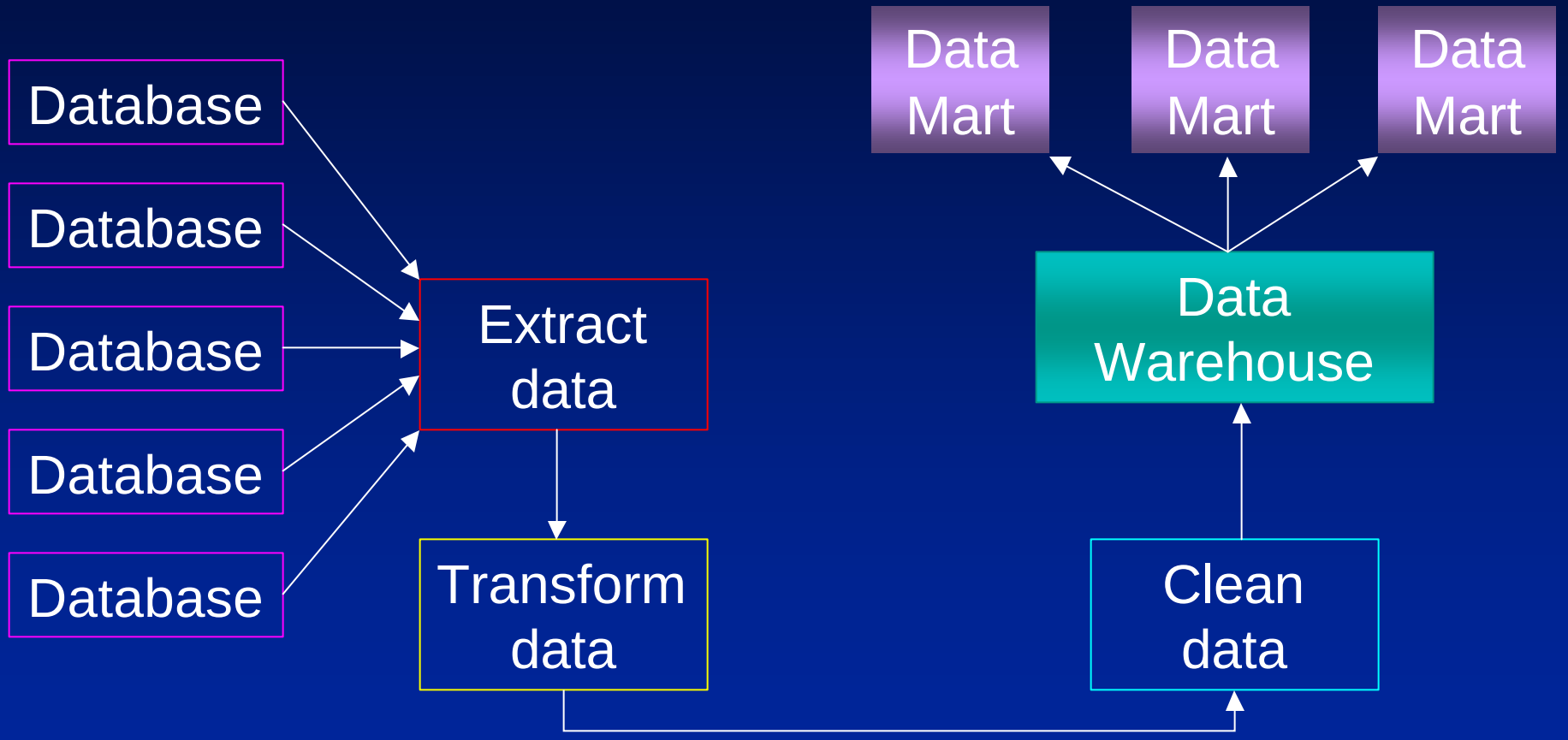
Examples of large data

- CMS Minimum Data Set
- State Medicaid claims databases
- Federally mandated surveillance systems
- Proprietary insurance claims data

Another (generic) example: Data Warehouses

- A centralized resource for long-term data storage
- Support the activities of entire organizations (enterprises)
- Input from distributed databases on scheduled batch basis
- Platform for decision support
- Provide large-scale, temporal data

How does a warehouse work?



Large data gets us into a hole…

- Large number of raw and derived variables renders traditional “manual” methods for discovering patters in data unwieldy
- Hypothesis-driven (biased) analyses may lead to missed associations
- Constantly changing patterns in prospective data require constantly changing analytic approaches that can be informed by data mining

- Introduction to databases and warehouses
- Data mining: What is it?
- Output of data mining
- The data mining life cycle
- Data mining applications
- Conclusion

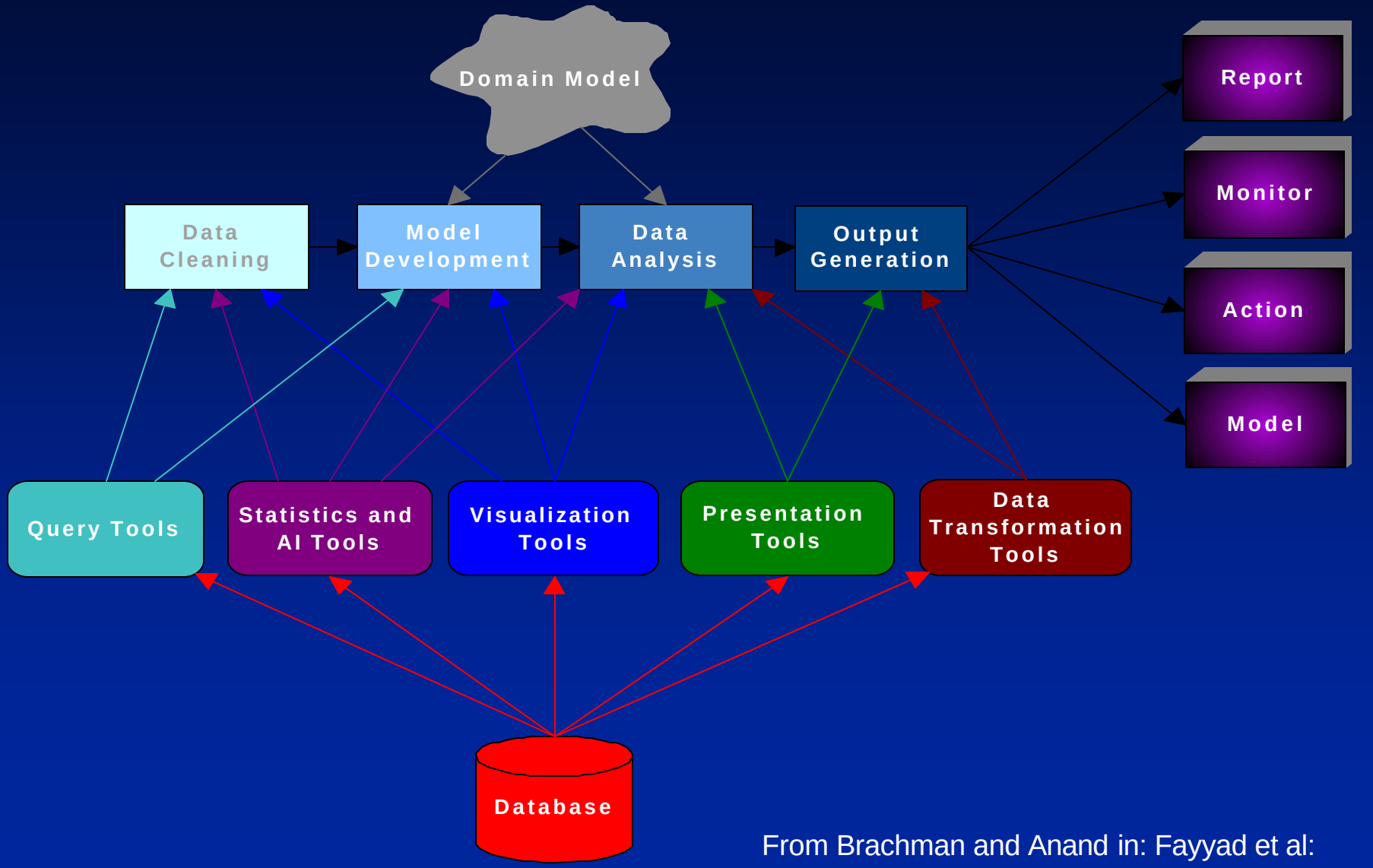
Knowledge Discovery in Databases - KDD

- Data-driven identification of valid, novel, potentially useful, and ultimately meaningful patterns in databases
- Traditionally applied to large-scale enterprise databases (data warehouses)
- Focused on hypothesis generation, not hypothesis testing

Now what are we looking for?
The Information Spectrum revisited

Data Information Knowledge Wisdom

The KDD Process



From Brachman and Anand in: Fayyad et al:
Advances in Knowledge Discovery and Data MIning

Data mining is the application of specialized software tools to the process of *knowledge discovery*

- Introduction to databases and warehouses
- Data mining: What is it?
- Output of data mining
- The data mining life cycle
- Data mining applications
- Data mining resources

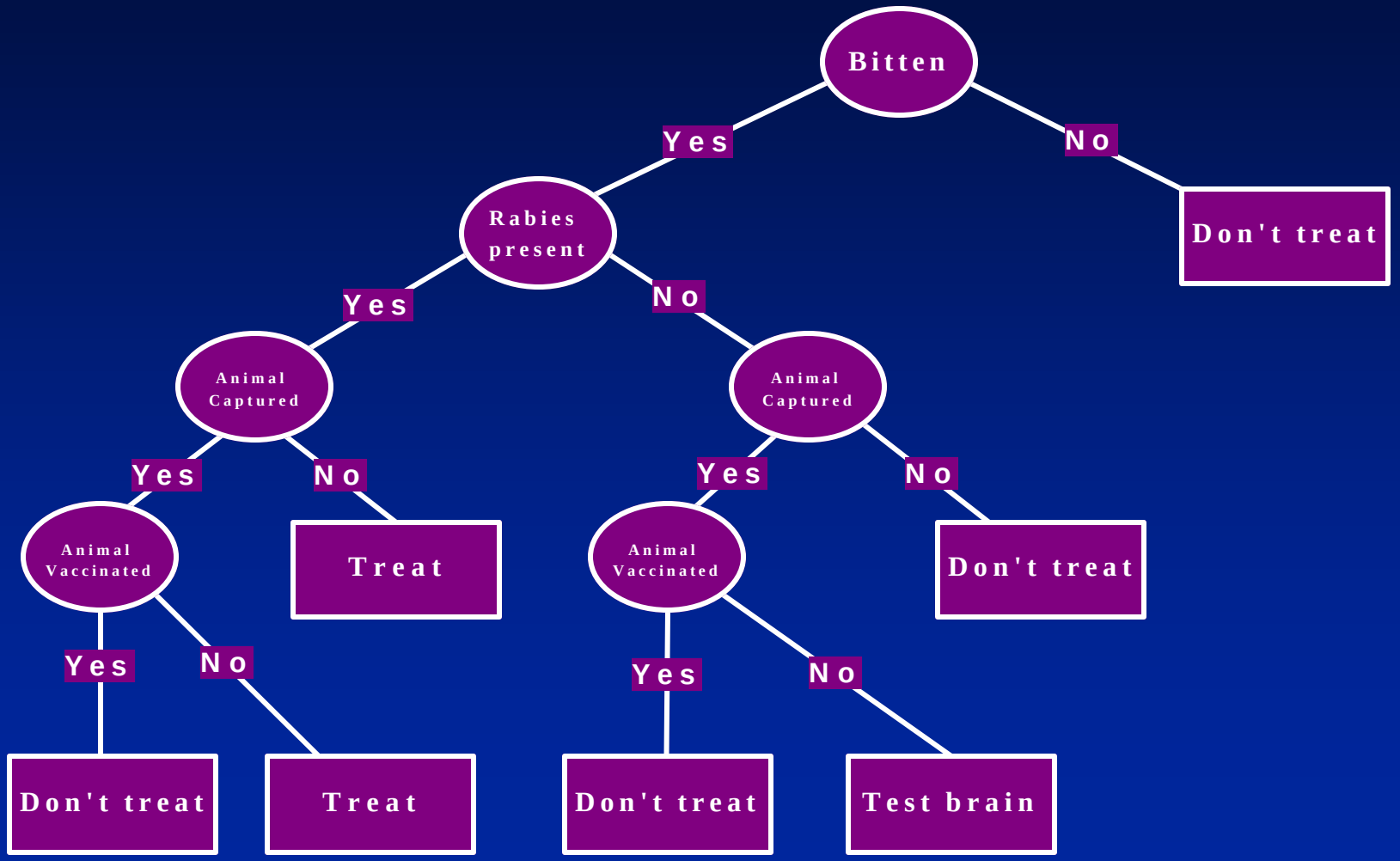
The Ore…
What comes out of the mine?

- Decision tables and trees
- Association rules
- Classification rules
- Prediction rules
- Clusters
- Visualizations

Decision Trees

- Simple, graphical method of representing data attributes and the relationships between them
- Robust data visualization tools
- Nodes (or cells) implicitly test an attribute with a constant or another attribute

A simple decision tree



Rules

- IF {condition} THEN {result} where:
  - condition=*antecedent* (LHS)
  - result=*consequent* (RHS)
- Conditions can be joined by Boolean connectors
  - AND, NOT, OR

Association Rules

- Focus on relationships between *any* attributes
- Most databases have large numbers of association rules that are often trivial (and misleading!)
- Example:
  - IF car-make = Ford
    THEN seat-belts-worn=Yes

Classification rule mining

- Looking for rules that classify cases into one of the known classes

If {VARIABLE}=value
then FATALITY=Yes

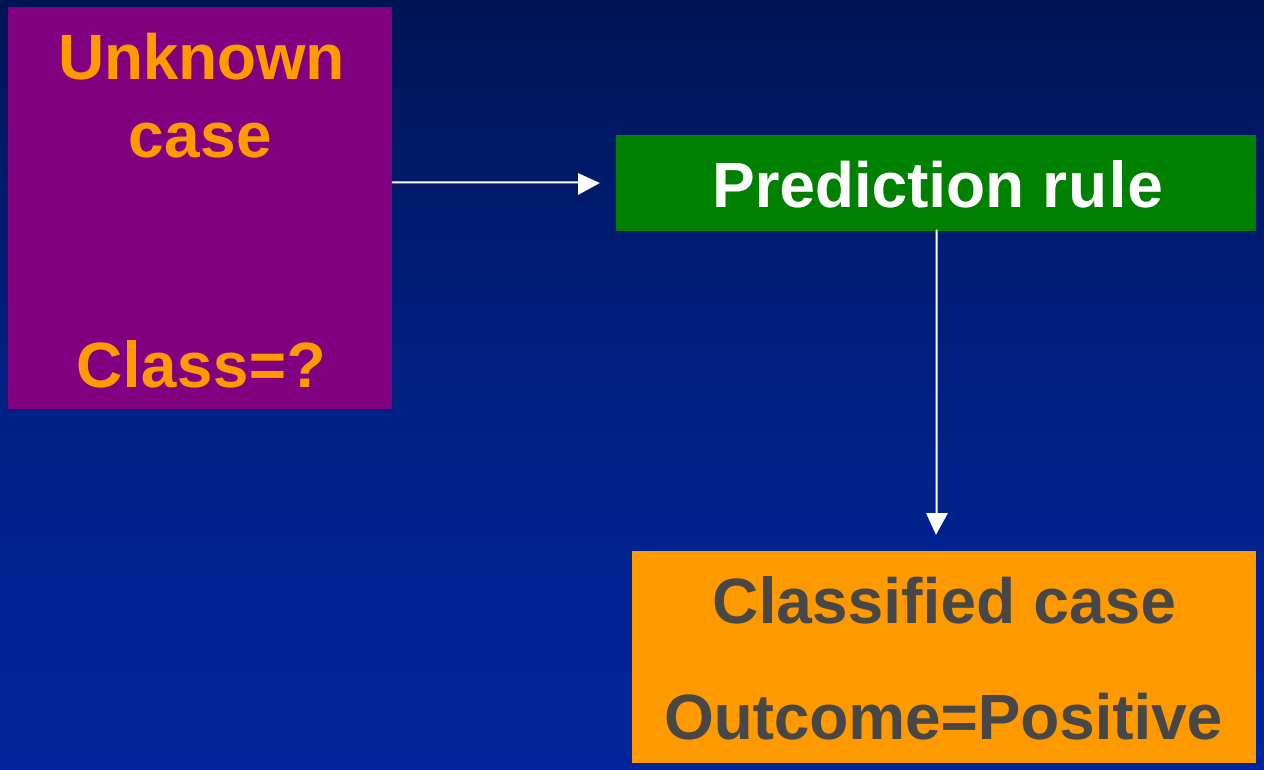
or

If {VARIABLE}=value
then FATALITY=No

Prediction Rules

- Classification rules that are used to predict class membership for objects of unknown class
- May provide simple class membership
- May indicate probability of class membership

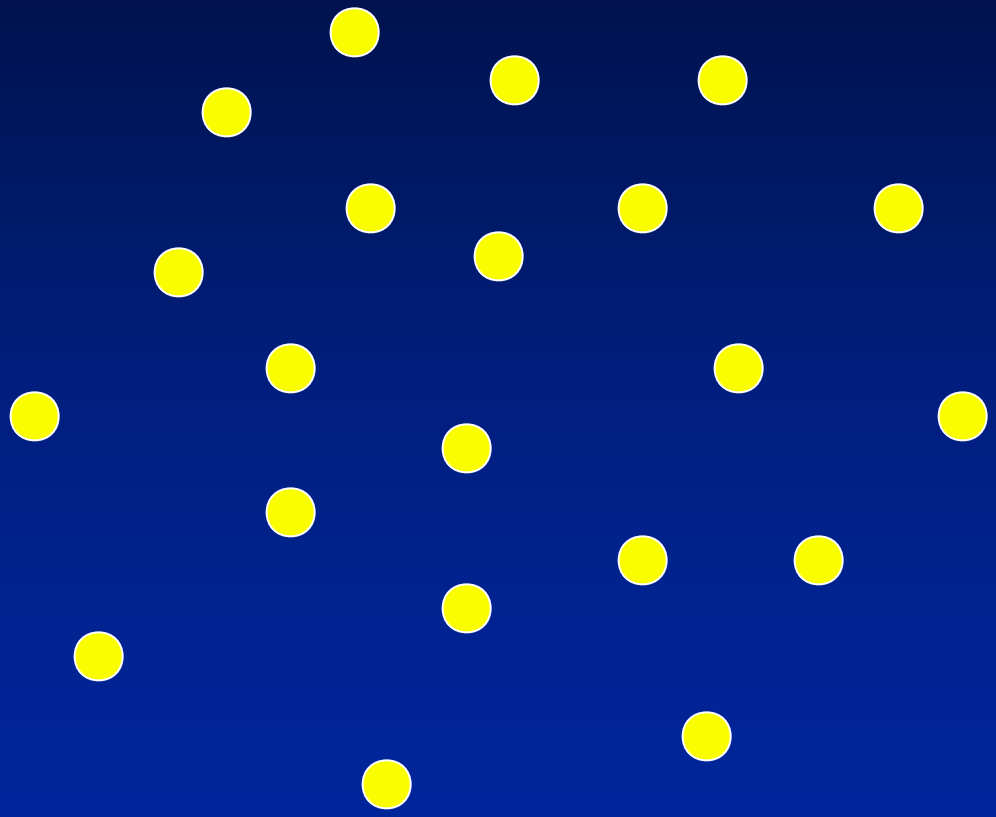
How a prediction rule works



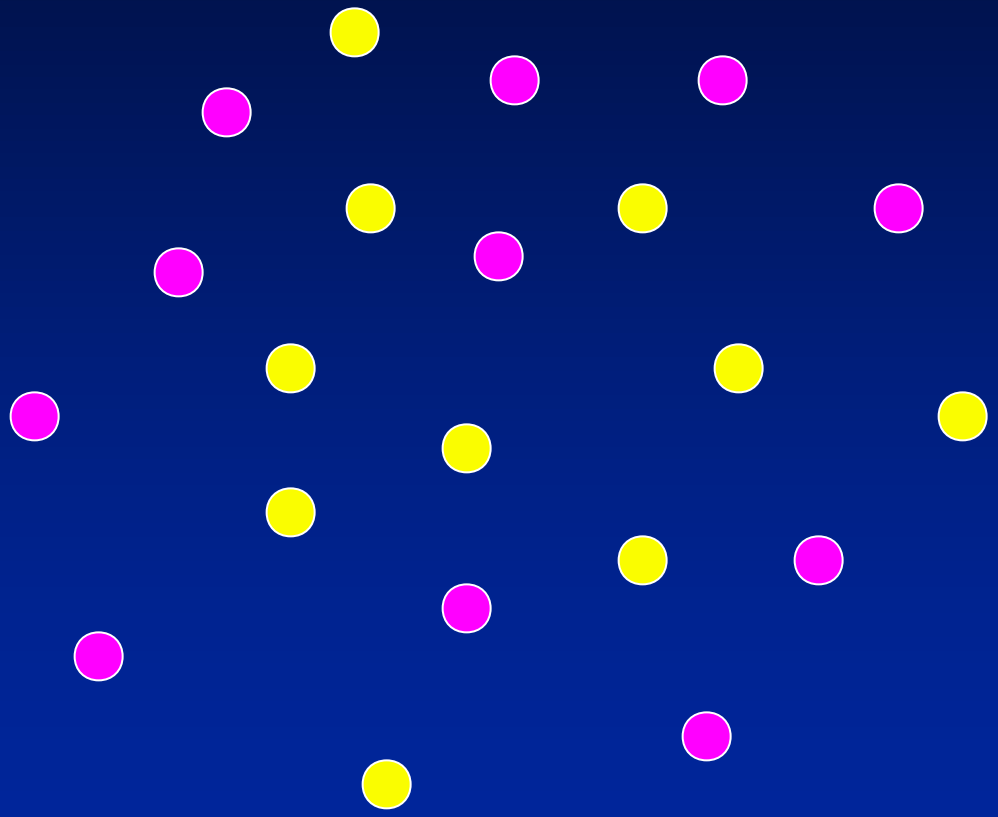
Clustering

- Grouping of objects into sets defined by some type of similarity
- Clusters are constructed by maximizing intraclass similarity and minimizing interclass similarity
  - » Objects in a cluster are similar to others in the same cluster
  - » Objects in one cluster are dissimilar to those in another cluster

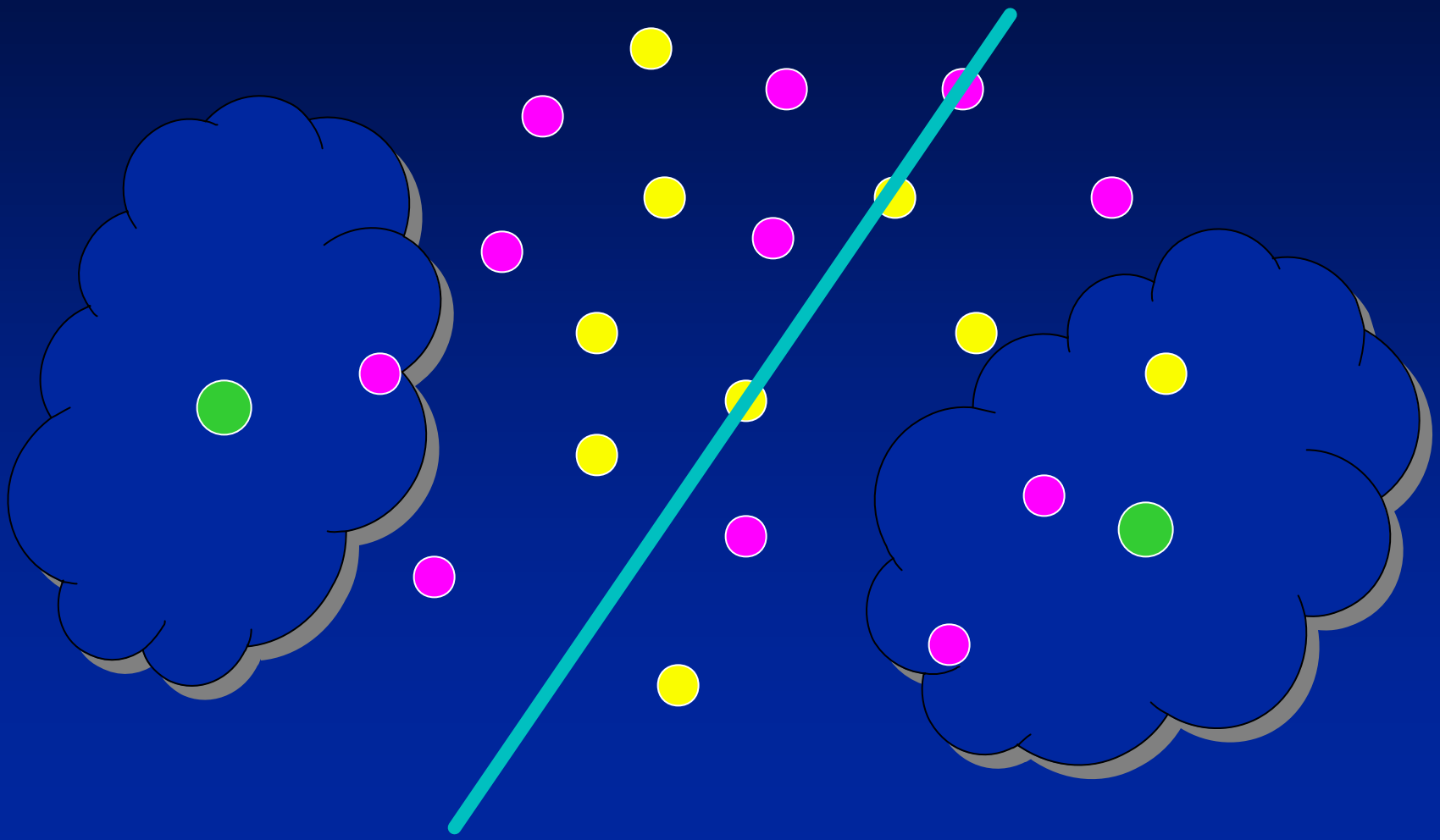
An Example: Raw Data



Applying a clusterer:
Identifying similarities and dissimilarities



Applying a clusterer:
Identifying similarities and dissimilarities



Online Analytical Processing (OLAP)

- Analysis techniques applied to data warehouses
  - » Summarization
  - » Consolidation
  - » Aggregation
  - » “Rotational” analysis
- Support multidimensional databases
- Predecessor of modern data mining

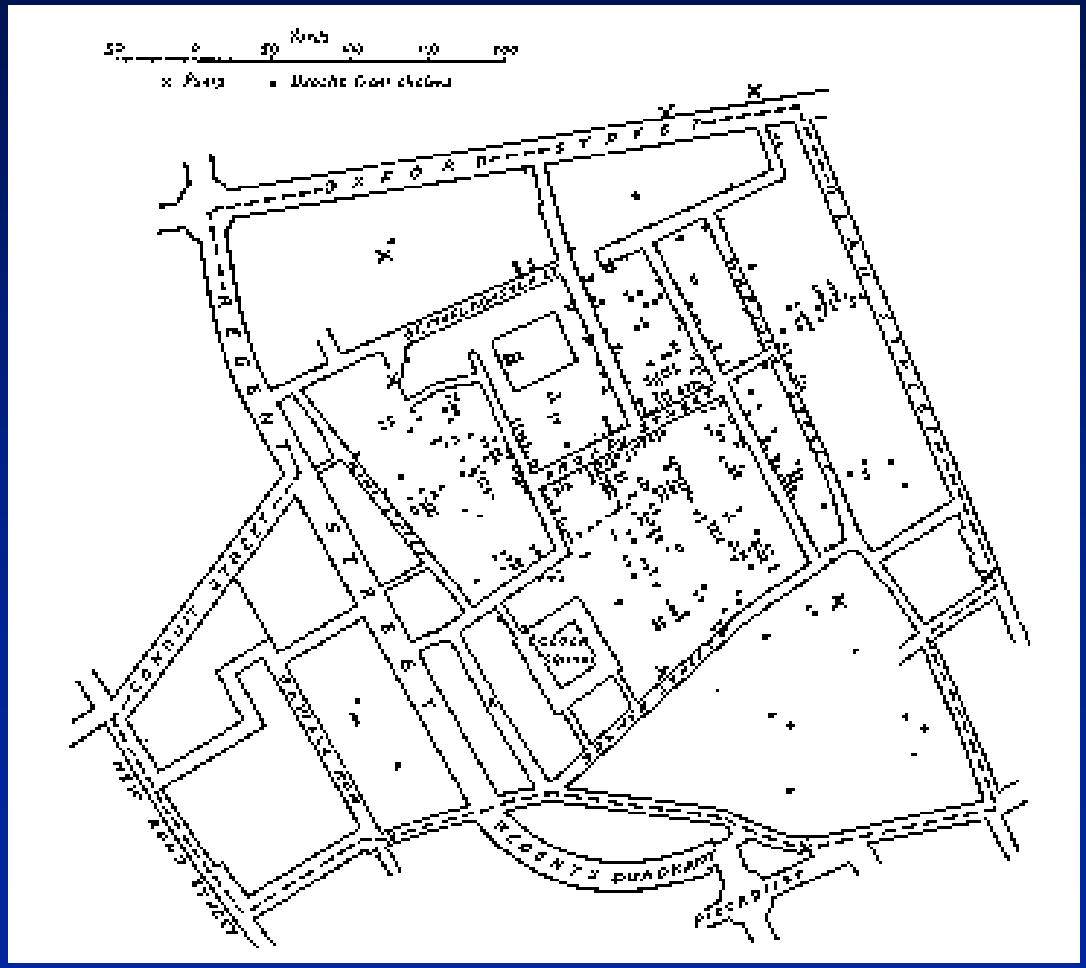
Data Visualization

- Used when data are not in an organized form
- Often a good first step to data reduction and transformation
- Focuses on graphical techniques

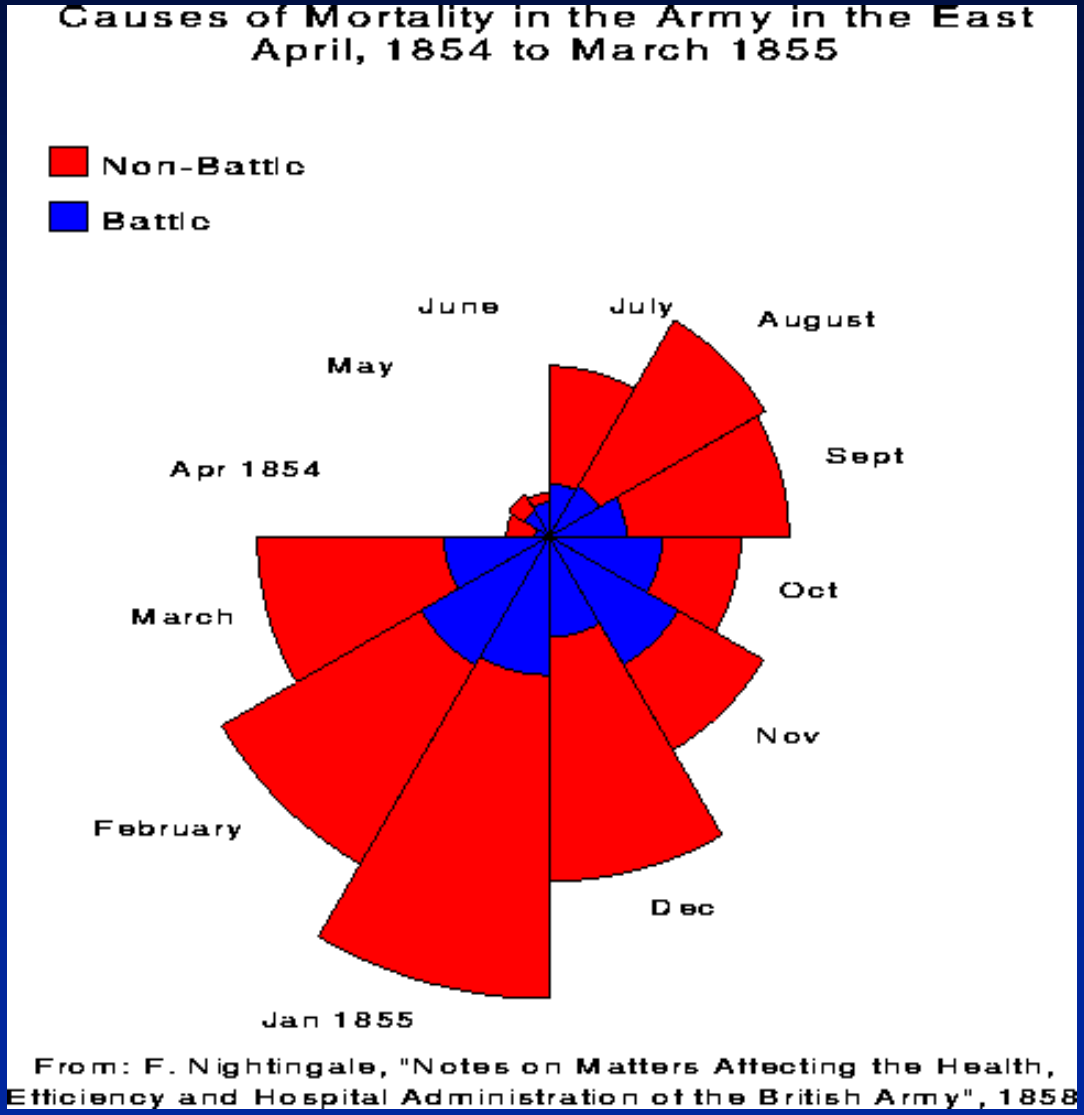
Visualization Tools

- Bar charts
- Pie charts
- Line graphs
- X-Y plots
- Maps
- Density plots

Visualization tools of historical interest: Snow's Cholera Map



Visualization tools of historical interest: Nightingale's Coxcomb Plot



From: F. Nightingale, "Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army", 1858

- Introduction to databases and warehouses
- Data mining: What is it?
- Output of data mining
- The data mining life cycle
- Data mining applications
- Data mining resources

The Data Mining Life Cycle

- Data preparation
- Data reduction
- Data modeling and prediction
- Evaluation

1. Data Preparation

- Standardization of coding from attribute to attribute of like concept
- Applying attribute transformations
- Applying attribute normalizations
- Database denormalization
- Discretization
- Missing data

Coding standardization

- Make sure all variables of like meaning are coded the same way

- Example
  - » Height should be in same units from record to record in the database

Attribute transformations

- Some techniques require data to be normally distributed or otherwise “smoothed”

- Example
  - » WBC (heavily skewed to the right)
  - » Solution: log transform

Attribute normalization

- Involves normalizing range to a specific normalization scheme, usually between 0 and 1, or -1 and +1
- Required by clustering methods for polytomous categorical data

Database denormalization

- Required if the software can’t handle normalized databases
- Example
  - Large claims databases, where patients have many claims records
  - Solution: do the needed joins to create a flat file and mine that

Discretization

- Some data mining software can't handle continuously valued attributes
  - » Example: Older evolutionary computation methods
  - » Solution: discretize
- Approaches:
  - » Histogram analysis
  - » Statistical binning
  - » Machine learning methods
    - – Entropy-based discretization

Handling missing data

- Special coding regimes
  - » Use negative or very large numbers to code numerical missing data
- Imputation
  - » Estimate what the missing value should be, using statistical methods
- Ignore it
  - » Some software can do this!

2. Data Reduction

- Why?
  - » Data mining methods are not open-ended as to capacity
  - » Trivia and minutiae are noise
  - » Noise may overwhelm software
  - » Output may overwhelm users

Methods for Data Reduction

- Segmentation
- Deletion
- Sampling
- Feature selection

Segmentation

- Divide the database up into manageable chunks, while analyzing all of the data

- Methods
  - » Kohonen maps
  - » Clustering

Deletion

- Rows
  - » Restrict the exploration to selected records in the database
  - » Problem: you could end up missing a rare cancer!
- Columns
  - » Restrict the exploration to selected fields in the database
  - » Problem: you could end up missing important risk factors!

Feature selection

- Heuristic approaches
  - » Cognitive domain model
  - » Expert panel
- Statistical methods
  - » Univariate and bivariate analysis
  - » Regression
  - » Nearest neighbors
  - » Clustering

3. Data Modeling and Prediction

- Use of mined information to create or augment knowledge
- Application of mined information for classification and prediction
- Employs the Output of data mining (we'll come back to that!)

Two families of data mining tools

- Statistical/Probabilistic

- Machine learning/Artificial intelligence

Statistical and probabilistic data mining tools

- Univariate
- Multivariate
- Bayesian classifiers
- Statistical classifiers

Machine learning tools

- Neural networks
- Decision tree induction
- Evolutionary computation

Decision Tree Induction

- Decision trees
  - » A node represents a test on an attribute
  - » A branch represents the test's outcome
  - » Leaf nodes represent decisions
- Created (induced) from data by means of an entropy based metric, information gain
  - » Used to select recursively the attribute that best separates the data into separate classes at a given node

Evolutionary Computation

- Framework and algorithms based on genetics metaphor
- Each unique combination of responses to a set of variables mapped to a unique rule or “chromosome”
- Each rule or “chromosome” is mapped to a possible outcome or “phenotype”
- Genetic operators mimic Darwinism (survival of the fittest rule)

Mining Complex Data

- Spatial
- Time-series
- Text
- Web

Spatial Data

- Contain topological information
- Organized according to a complex multidimensional indexing structure
- Require spatial reasoning and representation
- Examples
  - » Maps
  - » Imaging data

An example:
MRI of an abdominal aortic aneurysm

Mining time-series data

- Trend analysis
  - » Trend movements
    - – General direction a time-series graph move over time
    - – Cyclic variations
      - Long term oscillations of a trend line over time
    - – Seasonal variation
      - Cyclic variations tied to recurring points in time
    - – Random variation
      - Variations in movement of trend-line that are not cyclic

Text mining

- Text databases
  - » Collections of text-based documents
  - » Data semi- or unstructured
- Methods
  - » Keyword and similarity retrieval
  - » Latent semantic indexing

Keyword and similarity retrieval

- Keyword retrieval
  - » Document represented by a string containing one or more keywords
  - » Query formulated using keyword vectors with Boolean operators
- Similarity-based retrieval
  - » Similar to keyword retrieval, but retrieval is based on degree of similarity between keywords in document and in query vector

Web mining

- Issues
  - Size
  - Complexity
  - Dynamic nature
  - Broad coverage
  - Lots of chaff

Identifying Web usage patterns

- Uses Web logs to discover patterns of access to pages
  - » URL, IP of accessing user, and timestamp
  - » Lots of simple data, but often confusing patterns!
- Applications
  - » Marketing
  - » Web site design

4. Evaluation

- Heuristics
  - » Does this make good clinical sense?
- Multi-method comparisons
- Statistical methods
  - » Tests of association
  - » Sensitivity, specificity, predictive values, ROC curves

- Introduction to databases and warehouses
- Data mining: What is it?
- Output of data mining
- The data mining life cycle
- Data mining applications
- Data mining resources

Three illustrative applications of data mining

- Epidemiologic surveillance
  - » Motor vehicle-associated fatalities
- Patient safety
  - » Features associated with medication errors
- Research
  - » Intelligent data analysis

Fatality Analysis Reporting System (FARS)

- Prospective surveillance database of all fatal vehicle accidents occurring in the US
- Available at http://www-fars.nhtsa.dot.gov/
- Person-, vehicle-, and crash-level data

The FARS data model

Some possible questions about FARS

- What variables are associated with fatality?
- What variables predict fatality?
- Why not just use logistic regression?
- How to go about mining this database?

Some characteristics of FARS

- Denormalized
  - The Person File contains pertinent data from the Vehicle and Crash Files
- Large
  - 100,968 person records
  - 72 candidate variables
- Unbalanced
  - 42,116 deaths (41.7%)
- Many missing values
  - Bicycles don’t have airbags!
- Some variables are continuous
  - Require discretization for some DM tools
- Interactions
  - Passenger airbag deployment vs. year of vehicle
- Prospective
  - Even within a given year, new patterns emerge over time

What we're left with

Candidate predictors

- Age
- Sex
- Roadway function
- Manner of collision
- Model year
- Body type
- Rollover
- Emergency use vehicle
- Impact type
- Fire and/or explosion
- Person type
- Seating position
- Location in vehicle
- Ejection from vehicle
- Alcohol use
- Drug use
- Work-related injury
- Restraint use
- Weather conditions
- Surface conditions

Outcome: Fatality (Yes/No)

Sample decision tree output

```
:...ejection = Yes:
        :...hospital = No : Fatality (2968/36)
        :   hospital = Yes :
        :   :...driver = Yes : Fatality (1337/388)
        :       driver = No :
        :       :...rollover = Yes : No Fatality (1120/379)
        :           rollover = No :
        :           :...collision = No : Fatality (12/1)
        :               collision = Yes :
        :               :...urban = No : No Fatality (275/99)
        :                   urban = Yes :
        :                   :...fire_exp = Yes : No Fatality (10/2)
        :                       fire_exp = No :
        :                       :...air_bag = Yes : Fatality (18/6)
        :                           air_bag = No :
        :                           :...rest_use = Yes : No Fatality (10/2)
        :                               rest_use = No :
        :                               :...sex = No : Fatality (70/31)
        :                                   sex = Male:
        :                                   :...drinking = No : No Fatality (107/42)
        :                                       drinking = Yes : Fatality (5/1)
```

Sample rules from a decision rule inducer

If DRIVER=No
And EJECTION=No
And DRINKING=No
And DRUGS=No
Then FATAL=No

If HOSPITAL=Yes
Then FATAL=No

If PEDESTRIAN=No
Then FATAL=No

If EJECTION=Yes
And HOSPITAL=No
Then FATAL=Yes

If DRIVER=Yes
And ROAD=Rural
And RESTRAINT=No
Then FATAL=Yes

If WORK_INJ=Yes
Then FATAL=Yes

Three illustrative applications of data mining

- Epidemiologic surveillance
  - » Motor vehicle-associated fatalities
- Patient safety
  - » Features associated with medication errors
- Research
  - » Intelligent data analysis

Mining a large dataset for diagnoses associated with medication errors

- Nationwide Inpatient Sample
  - » Healthcare Cost and Utilization Project of U.S. Agency for Health Research and Quality
- Details
  - » Seven million hospital discharges in 1997
  - » Data from 22 states

Predictor variables

Demographics

- Age in years
- Died in hospital
- Discharge disposition
- Length of stay
- Primary insurance
- Race
- Sex
- Income

Discharge Diagnoses

- Cancer
- Circulation
- Congenital
- Dematologic
- Endocrine
- Gastrointestinal
- Genitourinary
- Hematologic
- Musculoskeletal
- Neurologic
- Obstetrical
- Psychiatric
- Pulmonary
- Others

Class distribution

Medication Error	**52,491**	**(0.73%)**
No Medication Error	**7,095,929**	**(99.27%)**
Total	**7,148,420**	**(100.0%)**

Classification variable

- Presence of a diagnosis suggesting medication error, defined as:
  - Incorrect dosage
  - Incorrect route of administration
  - Incorrect drug
  - Incorrect time or frequency
  - Problem associated with medication administration that could lead to an ADE

Data mining approach

- Decision tree induction (See5)
- Evolutionary computation (EpiCS)
- Logistic regression

Results: Sample negative rules

If AGE<=22
LOS<=2
INCOME>$50K
Then No error present

If LOS>2 and LOS<=6
INCOME >$25K
DIAGNOSIS=Dermatologic
Then No error present

If AGE>37 and AGE<=45
LOS=5
INSURER=Private
DIAGNOSIS<>Gastrointestinal
Then No error present

Results: Sample positive rules

If LOS=2
DIAGNOSIS=Psychiatric
Then Error present

If LOS>13 and LOS<=16
INSURER=Medicaid
INCOME <$25K
DIAGNOSIS<>Obstretric
Then Error present

If DISPOSITION=Transferred to another hospital
LOS=1
SEX=Male
DIAGNOSIS=Psychiatric
Then Error present

Three illustrative applications of data mining

- Epidemiologic surveillance
  - » Motor vehicle-associated fatalities
- Patient safety
  - » Features associated with medication errors
- Research
  - » Intelligent data analysis

Using data mining to inform statistical analysis

- Data
  - » FARS 2001 person file
- Data mining methods
  - » EpiXCS
  - » See5 (decision tree inducer)
- Primary statistical analysis method
  - » Logistic regression

The problem…

- The dataset is too large to analyze via traditional statistical methods
  - » >100K cases
  - » >100 variables, plus interactions
- Variable selection for logistic regression via bivariate methods too cumbersome
- How can we build a robust logistic model using these data?

Alternatives

- Bootstrapping
- Mine the data to identify candidate predictors and interactions
- Or, both!

Here's how the methods compared on variable selection

Significant Predictors			
	EpiXCS	LR	See5
Cyclists	X	X	X
Motircyclists	X	X	X
Ejection	X	X	X
Fire/Explosion	X	X	X
Driver's side impact	X	X	X
Not hospitalized	X	-	X
Vehicle rollover	X	-	X
Pickup trucks	X	-	X
Age >55	X	-	-
Rear-end impact	X	-	-
Inappropriate restraint	X	-	-

And on classification…

Classification Performance on Testing Set			
	EpiXCS	LR	See5
AUC	0.85	0.86	0.80
PPV	0.84	0.77	0.84

- Introduction to databases and warehouses
- Data mining: What is it?
- Output of data mining
- The data mining life cycle
- Data mining applications
- Conclusion

What kinds of questions can (and can't) be answered by data mining?

- Can…
  - Are there attributes that seem to be associated with others?
  - Are there any attributes that may be associated with an outcome that I might not be considering?
  - What variables should be included in a regression model?
- Can't…
  - Is an observed association statistically significant?
  - Can I always rely on what the miner is telling me?

Where is data mining appropriate?

- Large data
  - » Data warehouses
  - » Data marts
  - » Temporal databases
- Small data
  - » Specialized registries
  - » Ad hoc clinical research databases
- When the data are complex enough that you’re not sure that just looking at them will give you the answers you need

What is left after you strip away the hype?

- Data mining is a computer science discipline in development
  - » New techniques and software appear frequently, many of them untested
  - » Old techniques have been prematurely rejected
  - » Data mining is not a panacea…

BE CIRCUMSPECT!

Ethical concerns

- Data mining can’t be used to make policy, at least not by itself!
- Data mining results should not be reported in the literature, unless it’s a data mining article
- There is no substitute for the intellectual enterprise, of which data mining is a small part

Some data mining resources

- One-stop shopping web site
  - » KD nuggets www.kdnuggets.com
- Software
  - » Weka www.cs.waikato.ac.nz/ml/weka/
  - » MLC++ www.sgi.com/tech/mlc/
  - » IBM Intelligent Miner www-3.ibm.com/software/data/iminer/
  - » SPSS Clementine www.spss.com