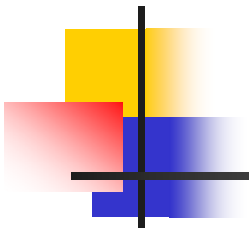


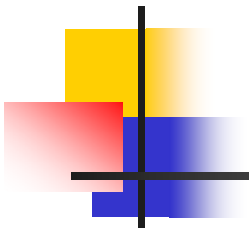
Perspectives on Data, Models, Analysis, and Discipline Differences

Daniel B. Carr
George Mason University



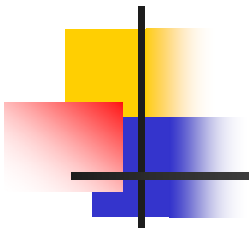
Data, Data Structures, Metadata, Models

- Data
 - Takes many forms
- Data structures for analysis
 - Data tables (data.frames), etc.
 - Cases and variables
 - Increasing complex forms (networks)
- Metadata
 - Ontologies
- Models
 - Distance measures
 - Structure-residual decompositions
 - Deterministic and stochastic parts



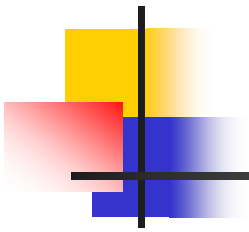
Data Table Structure (data.frames)

	Variable 1	Variable j
Case 1	Value_{11}	Value_{1j}
Case 2	Value_{21}	Value_{2j}
Case i	Value_{i1}	Value_{ij}



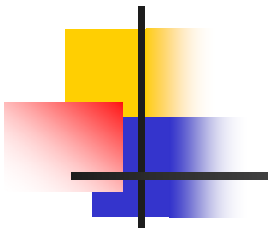
Data Table Metadata

- Provides context
- Answers questions about who, where, what, when, why, how and how well.
 W^5H^2
- The questions are also relevant at different levels: cases, variables and cell entries



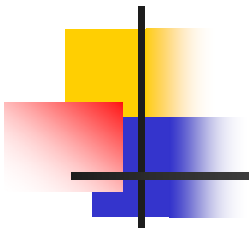
Cases

- Conceptual entities that we want to
 - Describe and compare
 - Model
 - Make inferences about or predict



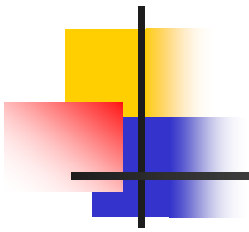
Examples of Cases

- Entities
 - People, cars, purchases, documents
- Geographic entities
 - States, nations, ecoregions, schools, roads
- Processes and procedures
 - Manufacturing processes, algorithms
- Schema: box plots, trees, networks,
- Abstractions: hypotheses and theories



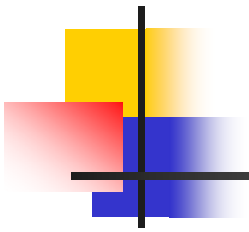
Case Metadata

- Data about the cases that has not made it into the table proper
 - May provide the basis for adding new variables to the table
 - May provide the basis for a
 - Stratified analysis
 - Weighted analysis
 - Eliminates of cases
 - Elimination of all cases



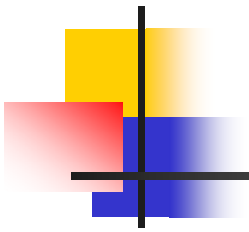
Variables

- Measurements on or descriptors of cases
- Provide a basis for comparison and/or mathematical operations
- Three common classes of variables
 - Spatial coordinates - index
 - Time - index
 - Attributes



Operational Classification of Variables (Attributes)

Class	Meaningful Operations
■ Nominal	$=, \neq$
■ Ordinal	$=, \neq, <, \leq, \geq, >$
■ Interval	above + differences
■ Ratio	above + ratios



Variable Metadata

- Data about measurement or calculation procedures, units of measure
- Federal and pharmaceutical ontologies are increasingly used to resolve naming inconsistencies
- Data about the measurement quality, such as confidence bounds, may appear as separate variables in the data table.

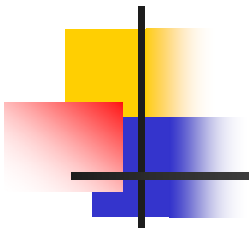
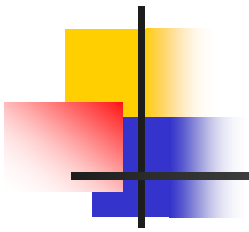


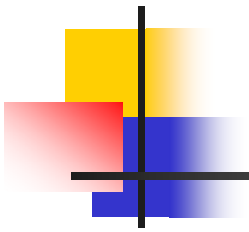
Table Cell Metadata

- Descriptors unique to a table cell
- Sometimes handled in data bases via links.
 - Tables, cases, variables and cells can have their own sets of links to metadata



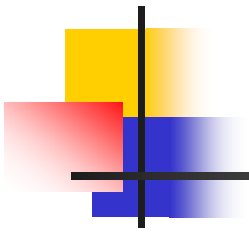
Abstract Duality of Cases and Variables

- Similarity measures for cases and variables
 - Cases:
 - based on selected, appropriately scaled variables
 - Euclidean distance
 - correlation, or other measures
 - Variables:
 - Based on selected cases
 - $1 - \text{abs}(\text{correlation})$ [0 1], $1 - \text{correlation}$ [0, 2]
 - Cosine of angle between vectors or other measures
- Cluster and/or plot using low dimensional layouts
 - Multidimensional scaling, self-organizing maps, etc.



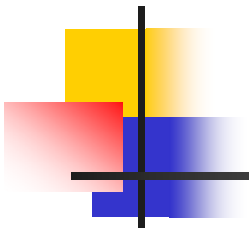
Sources of Data

- Data gathered with modeling in mind
 - Experiments
 - Survey Samples
- Administrative Records
- Monitoring
- ...



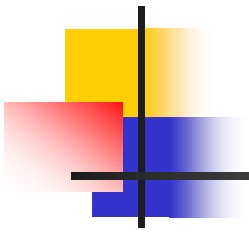
Data Problems

- Data from experiment designs
 - Psychology literature on human bias influencing experiment outcomes
 - Designs limited
 - Uncontrolled variables
 - Too controlled for actual target application
 - Drugs sold to the public
- Administrative records and broadly gathered data
 - Not representative of the target population or circumstance
 - planning for the last war, not the next one
- Problems with variables
 - Confounding, irrelevance, calibration, noise



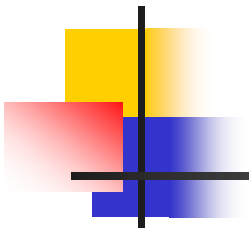
Data Structures

- Analytic methods are often based on simple data tables (n cases $\times$ p variables)
- Frequently **real data** do not take this form. Consider the medical history of people
 - Different events for different people: lobotomies
 - Some people have more repeated events: blood samples
 - The time points of similar events do not match.
 - The measurement procedures for similar events may not be comparable.



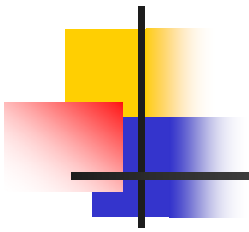
Data Analysis

- Structuring the data to use available methods commonly takes a lot of effort
 - Maybe 80%
- Structuring the analysis for the data is important
 - Views of the data can suggest that data cleaning is needed and the kinds of models that are appropriate
- Major advances in measurement capabilities provide new data that can motivate developing new analysis methods
 - Hyperspectral imagery
 - Microarray data
 - Data from time of flight mass spectrometry



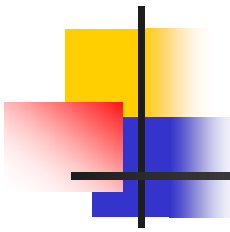
Approaches to Analysis Of New Data Forms

- Transform the data into classic data structures at some sacrifice of information
- Develop new distance measures to compare cases
 - Distance between molecules based on measure applied to their graphs.
 - Distance between documents based on the angle between word counts vectors
 - Consider barriers or penalized paths
 - Consider paths within manifolds, geodesic distance
- Develop a radically different methodology
 - Models for streaming data



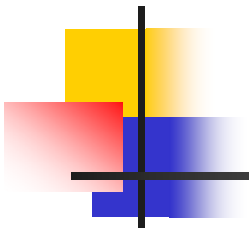
Analysis May Use Increasingly Complex Data Structures

- Networks
 - AT&T daily calling networks
 - Huge, even with aggregation of links between infrequent callers
 - Applies weighted moving average concept from time series to the whole network (daily views)
 - Used in catching fraudulent people
 - Multivariate terrorist networks
- Genomics analysis using metadata



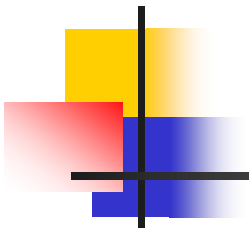
Generalization/Abstraction

- If we include all details everything is unique and in some sense incomparable
 - I assume most of you are unique!
- We have to select details to ignore. To make useful comparisons we need to choose wisely.
- Some have said “There is nothing new under the sun.” This is going too far.
 - Presumably all the people that believe this also believe they are one and the same person.



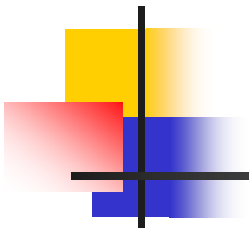
Scientific and Economic Focus

- Scientific modeling depends knowledge that helps us focus on what is important.
 - This knowledge evolves
- Economic interests actively avoid focus on some areas
 - Companies choose not to assess the health impact of their products and garbage: Toxic releases, diesel particulates, etc.
- Science is dependent upon economic and political interests



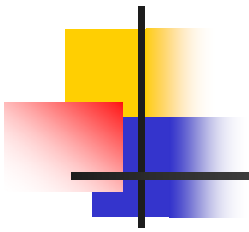
Statistics View

- We analyze data
 - Everything can be data at all levels of abstraction
- Good data is precious.
- Collection methodology is important.
Designing procedures to obtain crucial data can be the key part of the analysis process.
- Data relevance, quality and estimate uncertainty should be considered in the analysis process



Human Computer Interface View

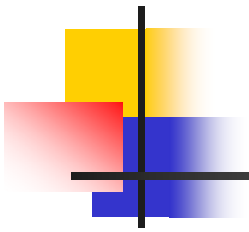
- Data is available
- Navigation and access to data are important
- Views of data are important
 - Overviews
 - Access to details
- Usability Studies
- Little emphasis on quantitative modeling



Statistical Learning

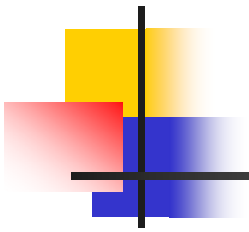
Theories/Machine Learning View

- Increasing interdisciplinary
 - Harder to characterize
- Data is stored large data bases
 - Fast algorithms and scalability are important
 - See Andrew Moore's work
 - <http://www-2.cs.cmu.edu/~awm/>
 - Models need to work, but not necessarily be parsimonious for human understanding
 - Partition and modeled locally
 - The data can be the model (nearest neighbor methods)



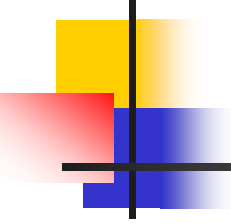
An Inclusive View?

- Common barriers include language and discipline blinders.
- Each discipline would like to claim priority.
- Progress is being made at least in the data mining community
- In general each discipline has important contributions to make. Other important sources include
 - Cartography, psychology, education, etc.



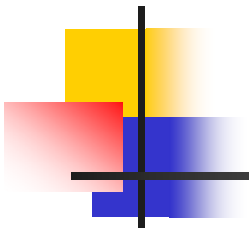
Visualization Process

- Design collection ⇒
- Transform data ⇒
- Construct models and visual mappings ⇒
- Manipulate views ⇒
- Perform tasks ⇒
- Annotate and Archive =>
- Raw data
- Data tables
- Models, graphic objects and views
- More views
- Understand, analyze, critique models, iterate, report
- Replicable analysis, highlights, insight into the analysis process



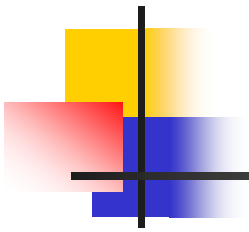
Convenient Data For Visualization Is Often Processed Data

- Computer processed monitoring data
 - Satellite image processing
 - Internal calibration, etc.
 - NASA images, level 3 processing
 - Sophisticated scientists often use less processed data
 - Captured results from processed burst or streaming data
 - High energy physic experiments
- Human and computer processed data
 - Talking to people who collected the data can be important
 - Known problems – Bad altitude sensor example
 - See the data in early stages can be instructive
 - Nuclear fuel production example



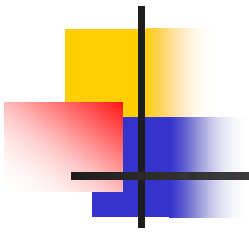
Visualizing Models And Their Parameters

- The meaning of data is limited without models
 - Ideally models connect to scientific theory
 - Empirical models can be useful for prediction
- Many scientists think it is more important to focus on models and model parameters than the data, especially when the model form is fairly stable over time.
- Models classes
 - Seriously wrong
 - All are wrong at some level of detail
 - Many can be excellent for practical purposes



Model Selection and Data

- Model selection
 - Scientific equations
 - Empirical models, deterministic and stochastic components
 - Computational convenience
 - Historically limited by computational feasibility
 - Gaussian distribution and least squares
 - Personal motivations
 - gaining experience, discipline fads
 - Involves assumptions
- Data
 - Can suggest more appropriate models at the start
 - Can provide a basis for model criticism and refinement
 - Can suggest the need for more or better data



Visualizing Concepts, Underlying Truth, Conjectured Worlds

- Some consider models less important than
 - The concepts they instantiate: an underlying reality or process
- Some evaluations utilize conjectured futures
 - What would be a person's medical costs if they didn't smoke?
 - Does reduced life time lower costs?