CSCI568

Discussion 4: Data Fundamentals

Organize Your Mind

Two general purposes

predictive

descriptive

~~Four~~ families of tasks

cluster analysis

classification

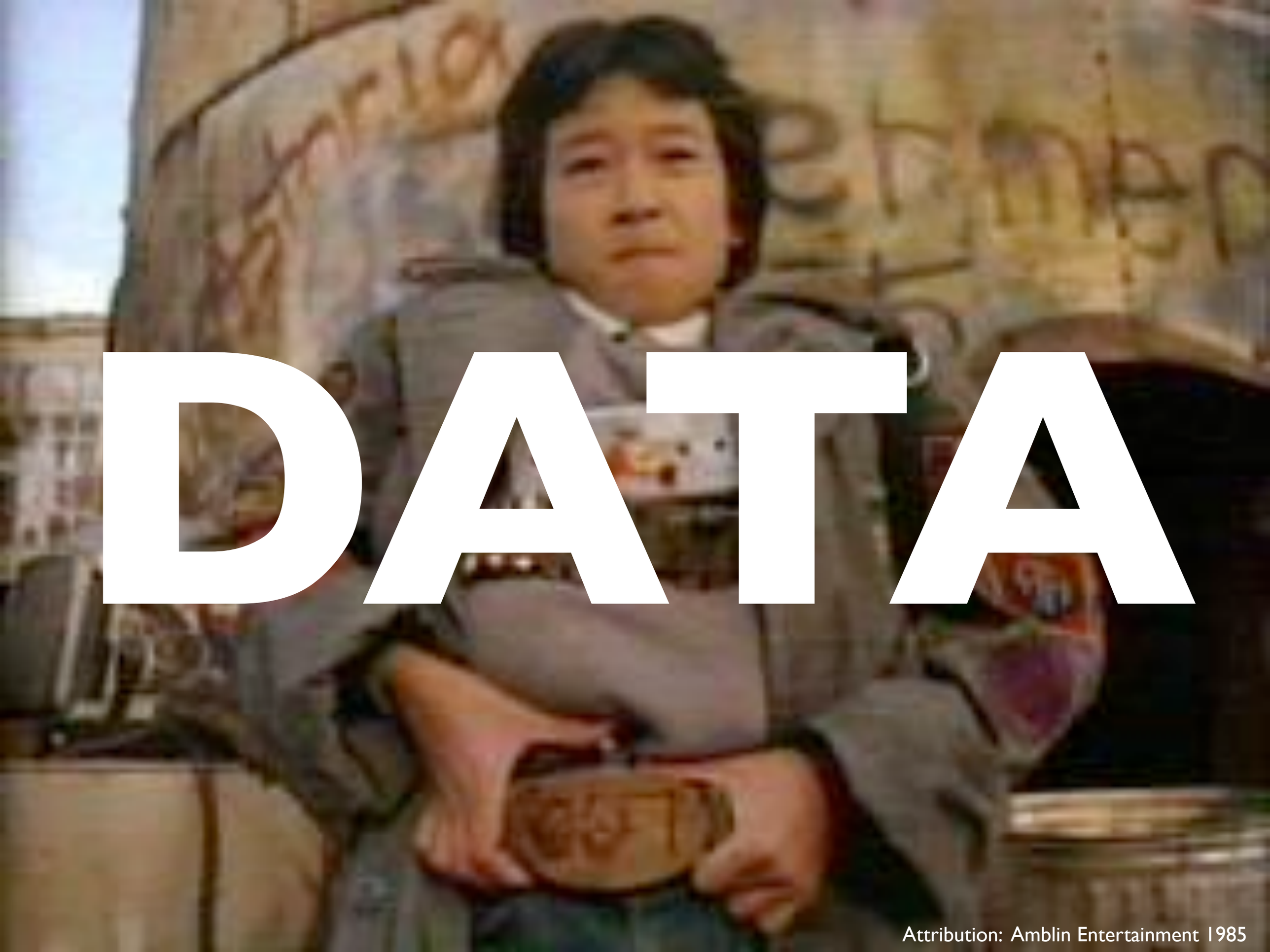
exploration

association analysis

anomaly detection

visualization & interaction

DATA



Attribution: Amblin Entertainment 1985

Common “Formats?”

Data Mining Starts With Data! (duh)

- type
- quality
- preprocessing
- existing/obvious relationships

Types

Categorical
(qualitative)

Nominal
zip codes
student ID
eye color
gender

Ordinal
hardness
grades
street numbers

Numeric
(quantitative)

Interval
dates
temperature (in C/F)

Ratio
temp in K
money
age
lengths

Data Transformations

(changes that don't affect the *meaning* of an attribute)

Categorical
(qualitative)

Numeric
(quantitative)

Nominal
preserve 1:1 mapping

Ordinal
preserve order

Interval
consistent scaling

Ratio
preserve ratio, but units can change

Bottom line: *statistical* operations should yield same results whether or not data is transformed.

(DM 27)

of possible values

- finite ... discrete
- infinite.. continuous (limited precision)

A "Special" Type: Asymmetric Attributes

- We only care about presence (or non-presence)

- Influences the meaning of "similarity"

Data Sets
(collections of objects)

- Dimensionality (# of attributes)
- Sparsity (how often a value exists)
- Resolution
    - too fine? pattern buried in noise
    - too coarse? pattern not evident

Data Sets Have Types Too

DM 29 - 36

Record-Based

- General (most common)
- Transaction / Market Basket Data
    - One attribute includes a set of values
- Data Matrix
    - sparse data matrices
        - document-term matrices

Graph-Based

- Relationships among objects
  - data objects are mapped to nodes
  - relationships are represented as links
- Object attributes might be graphs

Ordered Data

- Temporal data
    - "candy sales peak before Halloween
        - purchase history, predictions
        - subject to temporal autocorrelation
- Sequence data (position, no timestamps)
- Spatial data (eg, weather data by location)
    - subject to spatial autocorrelation

Unstructured Data

- non-record data can become record data
- but you might not capture certain aspects of the data

Data Quality Issues

- collected for other purposes
- collected without specified application
- errors
- *often cannot be addressed at the source*

Data Mining Includes...

- detection and correction of problems
- use of algorithms that tolerate poor quality

Data Is Rarely Perfect (damn!)

- human error
- limitation of measuring tools
- flaws in collection process
- missing values
- missing records
- duplicate records

Measurement Error

- recorded value differs from true value
- measure the measurement error (aka “error”) of continuous attributes

Collection Error

- eg, omitting data objects or attributes
- inappropriately including a data object
- collection process problems
- keyboard errors

Noise & Artifacts

- NOISE = random component of measurement error
    - distortion or added spurious objects
    - most often with temporal / spatial datasets
    - elimination is often difficult
        - so we use algorithms that are robust against noise
- ARTIFACTS = deterministic errors
    - think: “a repeated streak on a set of photos”

Precision, Bias, Accuracy (simple, but important!)

- Precision: closeness of repeated measurements
    - std. dev. of a set of values
- Bias: systematic imprecision
    - diff between mean of values and real value
- Accuracy: closeness to true value

Outliers

- *data objects* that have characteristics that are very different from most of the other objects
- *value*s of an attribute that are very different from typical values of that attribute
- outliers are not noise per se
    - are often legitimate data objects/values

Missing Values (and what to do about it)

- Eliminate data object or attribute
- Estimate missing values
- Ignore missing values during analysis

DM 40, 41

Inconsistent Values

- think: "incorrect city for a zip code"
- often easy to detect
  - eg, human height shouldn't be negative
- can use other valid data to resolve inconsistencies

DM 41, 42

Duplicate Data

- duplicates or “almost-duplicates”
- eg, John Smith vs John Q. Smith
    - Same object but slightly different attributes
- Values should be resolved/merged
    - aka “deduplication”

DM 42, 43

What Defines Quality Data?

Data is of "high quality" if it is suitable for its intended use.

Timeliness
Relevance
Documentation / Knowledge

Exploring Data

Common Summary Stats

- frequency, mode
- percentiles (esp. ordered data)
- mean, median, trimmed mean
- min, max, range, variance/std-dev
    - absolute average deviation (AAD)
    - median absolute deviation (MAD)
    - interquartile range (IQR)

Multivariate Summary Statistics

- Covariance
  - "to what degree do the values of two attributes vary together?"
- Correlation
  - "how strongly are two attributes linearly related?"

Visualization

Simple: the display of information in a graphical format

Why Visualizations?

- Humans have a great ability for analyzing large amounts of information visually
  - Can detect general patterns/trends
  - Can detect outliers / unusual patterns

Homework

- Reading 4
- Post one cool data visualization on Piazza