- **Data Science Roles**
- Business Person
- The Programmer
- The Enterprise
- The Web Company
- **Data Cleaning and Quality**
- Data Cleaning
  - Missing data
  - Entity resolution
  - unit mismatch
- Dealing with Dirty Data — Statistician view
  - process that produces data
    - distortion - some samples are corrupted by a process
    - Selection Bias - likelihood of a sample depends on its value
    - Left and Right Censorship - users come and go from our scrutiny
    - Dependence - samples are supported to be independent, but are not (ex: social networks)
  - Add new models for each type of imperfection
    - cannot model everything
    - what's the best trade-off between accuracy and simplicity
- Dirty Data — Database View
  - results are absolute (relational model)
  - improving the quality of the values in the dataset
- Dirty Data — Expert View
  - The data doesn't look right
  - Domain experts have implicit model of the data that they can test again…
- Dirty Data — Data Scientist
  - combination of all of the above
- Data Quality Problems
  - (Source) Data is dirty on its own
  - transformation corrupt data
  - clean datasets screwed up by integration
  - "Rare" errors can become frequent after transformation/integration
  - Clean datasets can suffer "bit rot": data loses value/accuracy over time
  - Any combination of the above
- Where does Dirty Data Come from?
  - extract transform load process
- Dirty Data Problems
  - Parsing text into fields
  - Naming conventions
  - Missing required field
  - Primary key violation
  - Licensing/Privacy issues prevent use of the data as you would like
  - Different representations
  - Fields too long
  - Redundant Records
  - Formatting issues — especially dates
- The meaning of Data Quality
  - Data Interpretation

        - data is useless unless we know all the rules behind it
    - Data Suitability
        - can you get answer from available data
            - use of proxy data
            - relevant data is missing
- Data Quality Continuum
    - Data and information are not static
    - Flows in a data collection and usage process
        - Data gathering
        - data delivery
        - data storage
        - data integration
        - data retrieval
        - data mining/analysis
- **Data Acquisition and Usage**
- Data Gathering
    - Experimentation, observation, and collection
    - Sources of problem
        - manual entry
        - approximations, surrogates — SW/HW constraints
        - No uniform standards for content and formats
        - Parallel data entry (duplicates)
        - Measurement or sensor errors
- Data Gather — Potential Solutions
    - Preemptive:
        - Process architecture (build in integrity checks)
        - Process management (reward accurate data entry, sharing, stewards)
    - Retrospective
        - cleaning focus (duplicate removal, merge/purge, name/addr matching, field value standardization)
        - Diagnostic focus (automated detection of glitches)
- Data Delivery
    - Destroying/mutilating information by bad pre-processing
        - inappropriate aggregation
        - NULLs converted to default values
    - Loss of data:
        - Buffer overflows
        - Transmission problems
        - No checks
- Data Delivery — Potential Solutions
    - Build reliable transmission protocols: use a relay server
    - Verification: checksums, verification parser
        - Do the uploaded files fit an expected pattern?
    - Relationships
        - Dependencies between data streams and processing steps?
    - Interface agreements
        - Data quality commitment from data supplier
- Data Storage
    - physical storage

        - potential issue but storage is cheap
    - problems in logic
        - Poor metadata:
            - Data feeds derived from programs or legacy sources — what does it mean?
        - Inappropriate data models
            - Missing timestamps, incorrect normalization, etc
        - Ad-hoc modifications
            - Structure the data to fit the GUI
        - Hardware / software constraints
            - Data transmission via Excel spreadsheets, Y2K
- Data Storage — Potential Solutions
    - Metadata: document and publish data specifications
    - Planning: assume that everything bad will happen
        - can be very difficult to anticipate all problems
    - Dat exploration
        - use data browsing and data mining tools to examine the data
            - does it meet the specifications you assumed?
            - has something changed?
- Data Retrieval
    - Exported data sets are often a view of the actual data
        - problems occur because:
            - source data or need for derived data not properly understood
            - just plain mistakes: inner join vs. outer join, not understanding NULL values
    - Computational constraints: Full history too expensive
        - supply limited snapshot instead
- Data Mining and Analysis
    - Problems in the analysis
        - Scale and performance
        - Confidence bounds?
        - Black boxes and dart boards
        - Attachment to models
        - Insufficient domain expertise
        - Casual empiricism (use arbitrary number to support a pre-conception)
- Retrieval and Mining — Potential Solutions
    - Data exploration
        - Determine which models and techniques are appropriate
        - Find data bugs
        - Develop domain expertise
    - Continuous analysis
        - are the results stable? How do they change?
    - Accountability
        - make the analysis part of the feedback loop
- **Data Quality Constraints and Data Integration**
- Data quality constraints
    - Capture many data quality problems using schema's static constraints
        - NULLs not allowed, field domains, foreign key constraints, etc
    - Many other quality problems are due to problems in workflow
        - Can be captured by dynamic constraints
        - E.g. orders above 200 are processed by biller 2

- The constraints follow an 80-20 rule
  - a few constraints capture most cases
  - thousands of constraints to capture the last few cases
- Constraints are measurable — data quality metrics?

- Data Quality Metrics
  - We want a measurable quantity
    - indicates what is wrong and how to improve
    - Realize that DQ is messy problem, no set of numbers will be perfect
  - Metrics should be directionally correct with improvement in data use
  - Types of metrics
    - static vs. dynamic constraints
    - operational vs. diagnostic
- Examples of Data Quality Metrics
  - Conformance to schema: evaluate constraints on a snapshot
  - Conformance to business rules: evaluate constraints on DB changes
  - Accuracy: perform expensive inventory or track complaints (proxy)
    - audit samples
  - accessibility
  - interpretability
  - glitches in analysis
  - successful completion of end-to-end process
- Technical approaches
  - multi-disciplinary approach to attack data quality problems
    - no one approach solves all the problems
  - process management: ensure proper procedures
  - statistics: focus on analysis — find an repair anomalies in data
  - database: focus on relationships — ensure consistency
  - metadata / domain expertise
    - what does the data mean? how to interpret?
- Data Integration
  - combine data sets (acquisitions, across departments)
  - common source of problems
    - Heterogeneous data: no common key, different field formats
      - approximate matching
    - Different definitions: what is a customer — account, individual, family?
    - Time synchronization
      - does the data relate to the same time periods?
      - are the time windows compatible?
    - Legacy data: spreadsheets, ad-hoc structures
- Duplicate Record Detection
  - Resolve multiple different entries:
    - entity resolution, reference reconciliation, object ID/consolidation
  - Remove duplicates: Merge/purge
  - Record Linking (across data sources)
  - Approximate Match (accept fizziness)
  - House holding (special case)
    - different people in same house?
- Processing/Standardization
  - convert to canonical form

    - example: mailing addresses
- More Sophisticated Techniques
    - Use evidence from multiple fields
        - Positive and Negative instances are possible
    - Use evidence from linkage pattern with other records
    - clustering-based approaches
- Lots of Additional problems
    - Examples
        - address vs number, street …
        - units
        - differing constraints
        - multiple versions and schema evolution
        - other metadata
- Data Integration — Solutions
    - Commercial Tools
        - Significant body of research in data integration
        - many tools for address matching, schema mapping are available
    - Data browsing and exploration
        - many hidden problems and message meanings: must extract metadata
        - view before and after results
            - did the integration go the way you thought?
- **Estimation**
- Estimation
    - Statistical Inference
        - making conclusion based on data in random samples
        - example
            - use data to guess the value of an unknown number
        - create an estimate of the unknown quantity
            - depends on the random sample taken
- Assumptions
    - example
        - estimate the number of planes
        - see a plane with the number 44
    - The main assumption
        - The serial numbers of planes we see are uniform random sample drawn with replacement from 1,2,3,…N
- Estimation
    - If you saw the serial numbers 1 23 48 57 92
    - 92 is N
- The Largest Number Observed
    - Is it likely to be close to N?
        - How likely?
        - How close?
    - Some options:
        - Could try to calculate probabilities and draw a probability histogram
        - Could simulate and draw an empirical histogram
- Verdict on the Estimate
    - The largest serial number observed is likely to be close to N
    - But, it is also likely to underestimate N

  - Another idea for an estimate
    - average of the serial numbers observed ~ N/2
  - New estimate: 2 times the average of seen
- Bias
  - Biased estimate
    - on average across all possible samples, the estimate is either too high or too low
  - Bias creates a systematic error in one direction
  - Good estimates typically have low bias
- Variability
  - The value of an estimate varies from one sample to another
  - High variability makes it hard to estimate accurately
  - Good estimates typically have low variability
- Bias-Variance Tradeoff
  - The max has low variability, but it is biased
    - It under estimates the number of planes
  - 2*average has little bias, but it is highly variable
    - It over estimates the number of planes
  - Which one to choose?
    - Pick based on your utility?
- **Statistics**
- Normal Distributions, Mean, Variance
  - The mean of a set of values is the average of the values
  - variance is a measure of the width of a distribution
  - standard deviation is the square root of variance
  - normal distribution is characterized by mean and variance
- Properties of the mean
  - Balance point of the histogram
    - Not the “half-way point” of the data (median)
  - If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail
- Defining Variability
  - measure variability around the mean
- Standard deviation
  - measures roughly how far off the values are from the average
- Central Limit Theorem
  - If the samples are
    - A large set
    - drawn at random with replacement
  - Then no matter what the distribution of the population
    - probability distribution of the sample average is roughly bell-shaped (normal distribution)
  - The distribution of sum (or mean) of n identically-distributed random variables X_i approaches a normal distribution as n -> infinity
  - Common parametric statistical tests assume normally-distributed data, but depend on sample mean and variance
  - Tests work reasonably well for data that are not normally distributed as long as the samples are not too small
- Bounds and Normal Approximations
  - Chebychev’s Inequality

    - no matter what the shape of the distribution, the proportion of values in the range "average +- z SDs" is at least 1 - 1/z^2
- Correcting Distributions
  - many statistical tools assume data are normally distributed
- Other Import Distributions
  - Poisson: distribution of counts that occur at a certain "rate"
    - observed frequency of a given term in a corpus
    - number of visits to a web site in a fixed time interval
    - number of web site clicks in an hour
  - Exponential: interval between two such events
  - Zipf/Pareto/Yule distributions
    - govern frequencies of different terms in a document, or web site visits
  - Binomial/Multinomial
    - number of counts of events
    - example: 6 die tosses out of n trials
  - Understand your data's distribution before apply any model