CAP4770/5771
Introduction to Data Science
Fall 2015

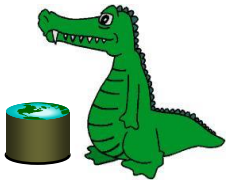
University of Florida, CISE Department
Prof. Daisy Zhe Wang

Based on notes from CS194 at UC Berkeley by Michael Franklin, John Canny, and Jeff Hammerbacher

Data Cleaning and Integration

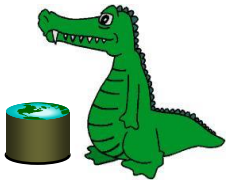
Data Cleaning

Data Integration



Review

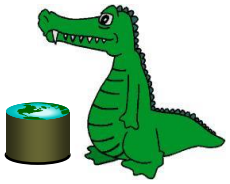
- Data Wrangling
  - The Big Picture (ETL, …)
  - Data types and sources
  - Data models (relational, semi-structured, sparse matrix, unstructured, graph ...)
  - One size does not fit all → noSQL systems
  - Data preparation (Lab 0)
    - Unix – completed? easy/hard?
    - Pandas/Python – completed? easy/hard?
    - Pop quiz – completed? easy/hard?



Schema-on-Read vs. Schema-on-Write

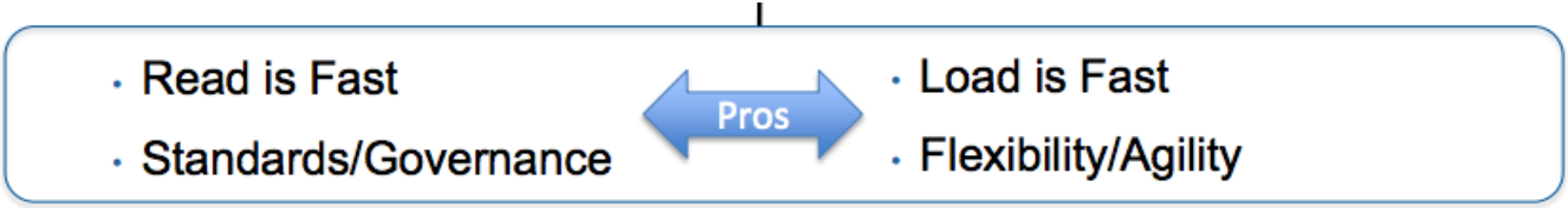
- Schema-on-Write: Traditional data systems require users to create a schema before loading any data into the system.

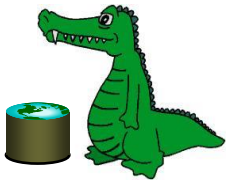
- Schema-on-Read: In Hadoop ecosystem, data can start flowing into the system in its original form, then the schema is parsed at read time (each user can apply their own "data-lens" to interpret the data).



Not "IF" But "WHEN"?

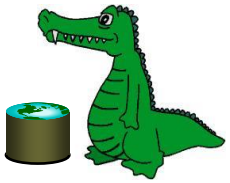
- "Schema on Write"
  - Traditional Approach
- "Schema on Read"
  - Data is simply copied to the file store, no transformation is needed.
  - A SerDe (Serializer/Deserlizer) is applied during read time to extract the required columns (late binding)
  - New data can start flowing anytime and will appear retroactively once the SerDe is updated to parse it.



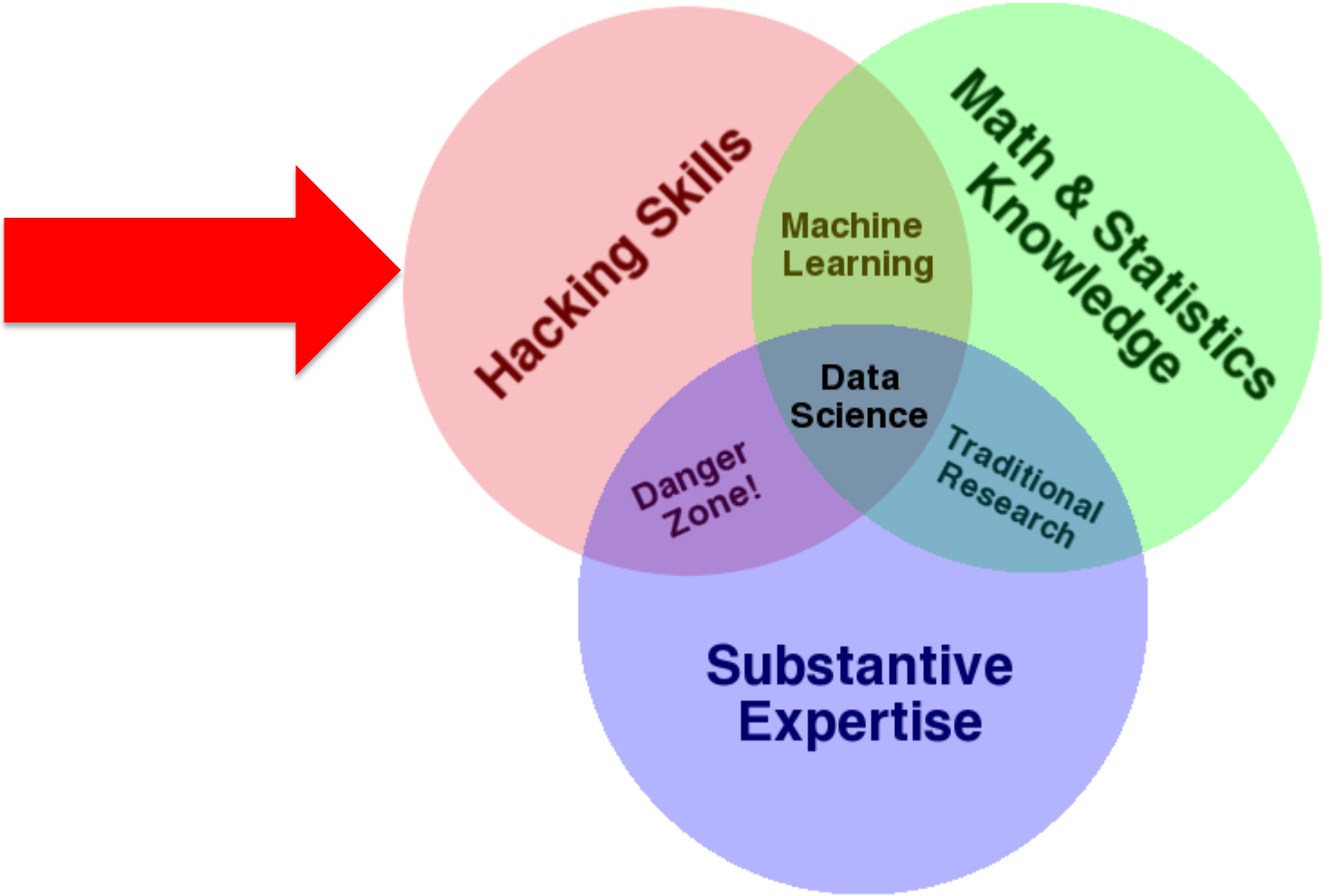


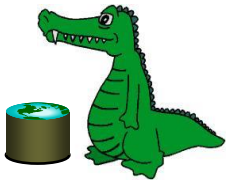
Outline

- Data Cleaning
  - Perspectives on “Dirty Data”
  - Perspectives on Data Quality
  - Some problems and solutions
- Data Integration
  - Item Similarity
  - Schema Matching



Data Science – One Definition





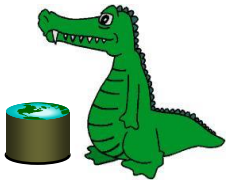
DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406Bn
Microsoft	Redmond, WA	$392Bn
Intl. Business Machines	Armonk, NY	$194Bn

```
SELECT Market_Cap
From Companies
Where Company_Name = “Apple”
```

```
Number of Rows: 0
```

```
Problem:
Missing Data
```



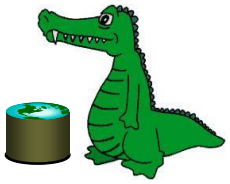
DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406Bn
Microsoft	Redmond, WA	$392Bn
Intl. Business Machines	Armonk, NY	$194Bn

```
SELECT Market_Cap
From Companies
Where Company_Name = “IBM”
```

```
Number of Rows: 0
```

```
Problem:
Entity Resolution
```



DB-hard Queries

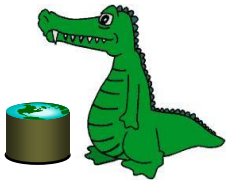
Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406
Microsoft	Redmond, WA	$392
Intl. Business Machines	Armonk, NY	$194
Sally's Lemonade Stand	Alameda,CA	$460

```
SELECT MAX(Market_Cap)
From Companies
```

```
Number of Rows: 1
```

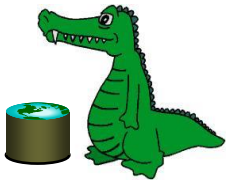
```
Problem:
Unit Mismatch
```

WHO'S CALLING WHO'S DATA DIRTY?



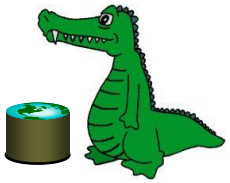
Dirty Data

- The Statistics View:
  - There is a process that produces data
  - We want to model ideal samples of that process, but in practice we have non-ideal samples:
    - **Distortion** – some samples are corrupted by a process
    - **Selection Bias** - likelihood of a sample depends on its value.. Examples?
    - **Left and right censorship** - users come and go from our scrutiny
    - **Dependence** – samples are supposed to be independent, but are not.. Examples? in social networks?
  - You can add new models for each type of imperfection, but you can't model everything.
  - What's the best trade-off between accuracy and simplicity?



Dirty Data

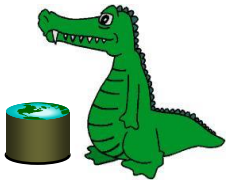
- The Database View:
  - I got my hands on this data set
  - Some of the values are missing, corrupted, wrong, duplicated
  - Results are absolute (relational model)
  - You get a better answer by improving the quality of the values in your dataset



Dirty Data

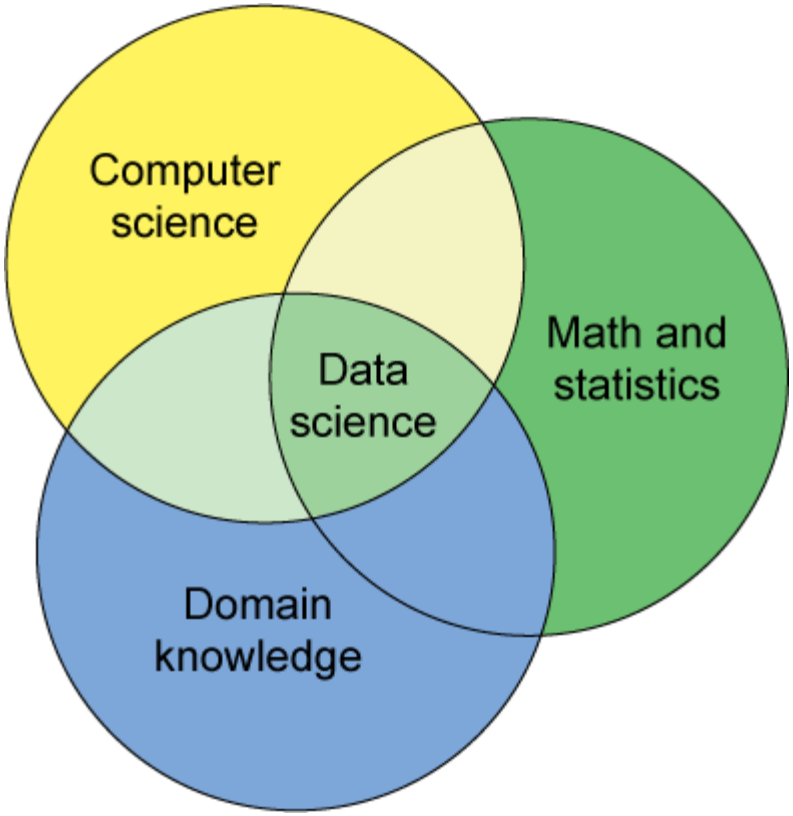
- The Domain Expert's View:
  - This Data Doesn't look right
  - This Answer Doesn't look right
  - What happened?

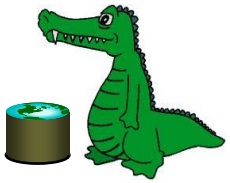
- Domain experts have an implicit model of the data that they can test against…



Dirty Data

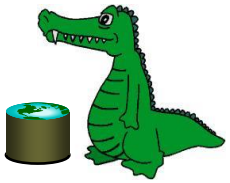
- The Data Scientist's View:
  - Some Combination of all of the above



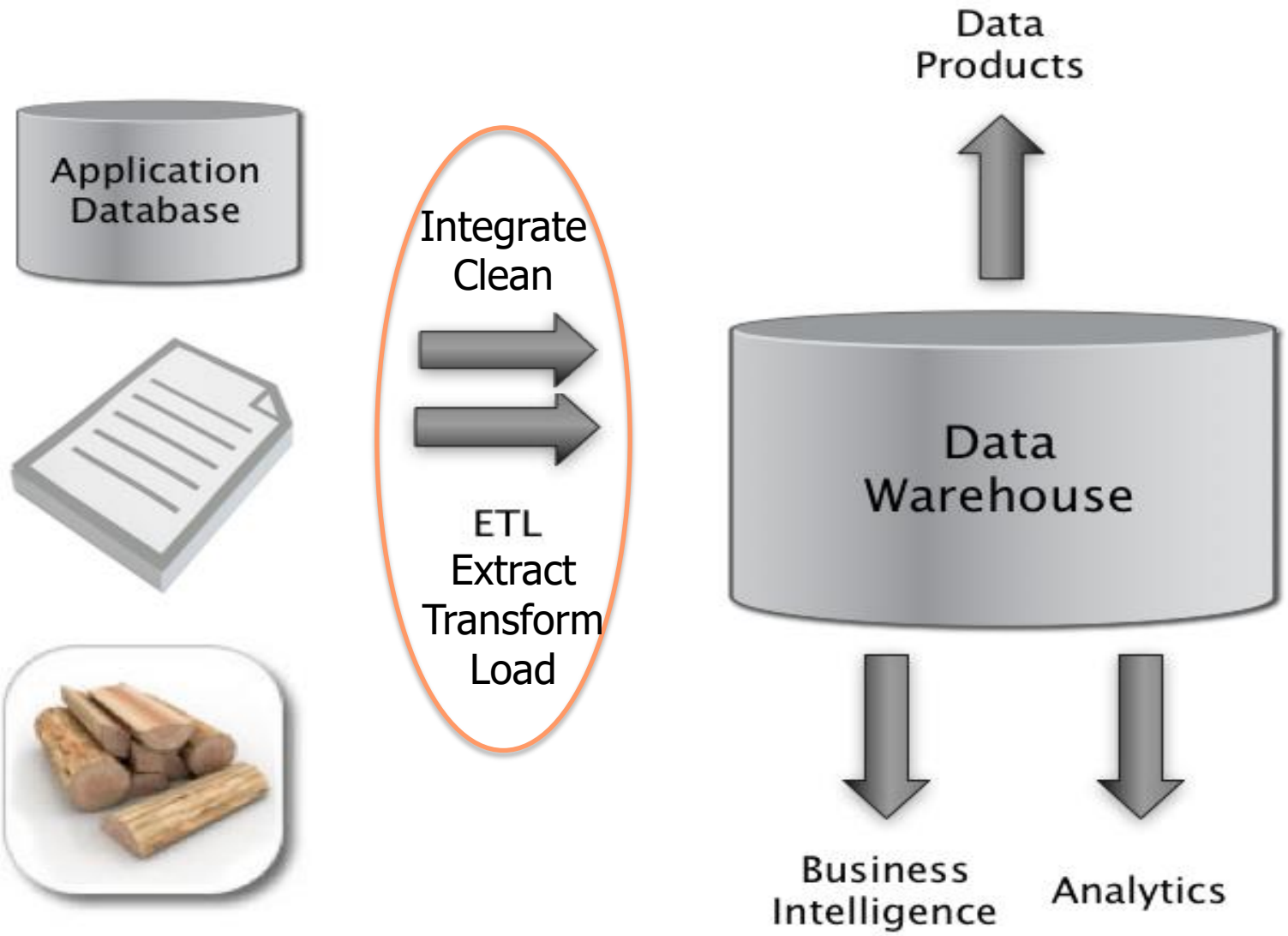


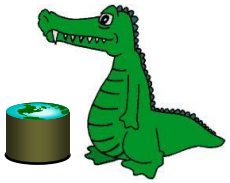
Data Quality Problems

- (Source) Data is dirty on its own.
- Transformations corrupt the data (complexity of software pipelines).
- Data sets are clean but integration (i.e., combining them) screws them up.
- "Rare" errors can become frequent after transformation or integration. Examples?
- Data sets are clean but suffer "bit rot"
  - Old data loses its value/accuracy over time
- Any combination of the above



Big Picture: Where can Dirty Data Arise?

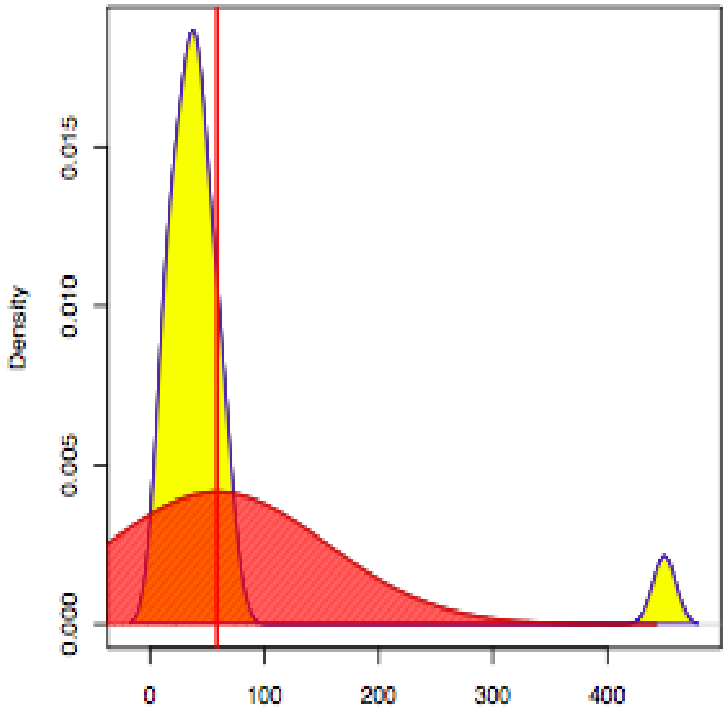




Numeric Outliers: Dirty data or Interesting event?

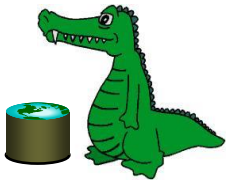
12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)



- median 37
- mean 58.52632
- variance 9252.041

Adapted from Joe Hellerstein's 2012 CS 194 Guest Lecture

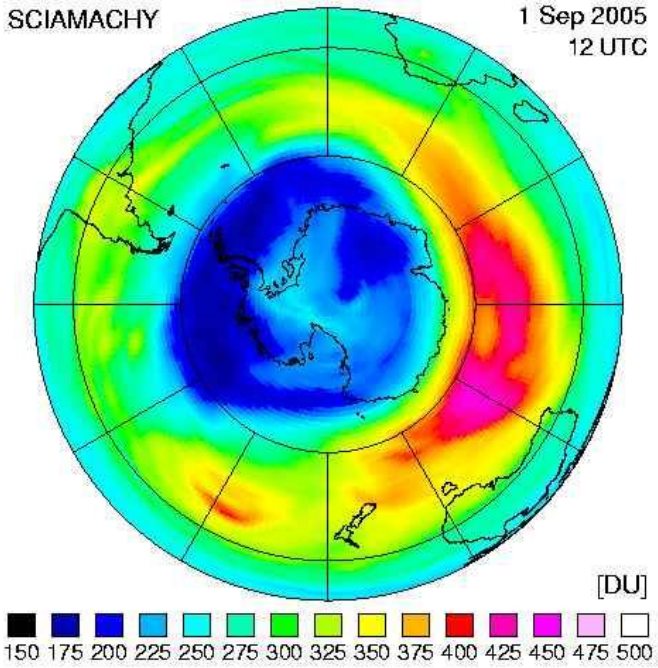


Data Cleaning Makes Everything Okay?

The appearance of a hole in the earth's ozone layer over Antarctica

First detected in 1976, was so unexpected that scientists at the National Center for Atmospheric Research didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning.

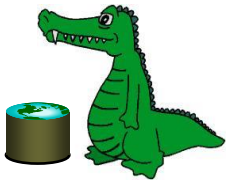
How to differentiate unexpected data vs. dirty data?



In fact, the data were rejected as unreasonable by data quality control algorithms

Dirty Data Problems

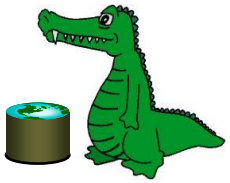
- From Stanford Data Integration Course:
  1) parsing text into fields (separator issues)
  2) Naming conventions: ER: NYC vs New York
  3) Missing required field (e.g. key field)
  4) Different representations (2 vs Two)
  5) Fields too long (get truncated)
  6) Primary key violation (from un- to structured or during integration
  7) Redundant Records (exact match or other)
  8) Formatting issues – especially dates
  9) Licensing issues/Privacy/ keep you from using the data as you would like?



Conventional Definition of Data Quality

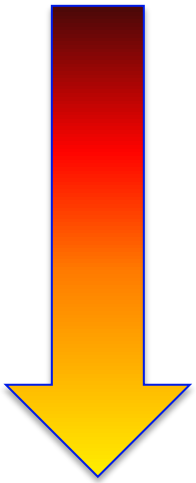
- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
- Consistency
  - The data agrees with itself.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

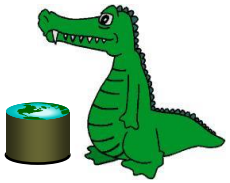


The Data Quality Continuum

- Data and information is not static, it flows in a data collection and usage process
  - Data gathering
  - Data delivery
  - Data storage
  - Data integration
  - Data retrieval
  - Data mining/analysis



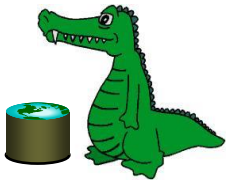
Adapted from Ted Johnson's SIGMOD 2003 Tutorial



Data Gathering

- How does the data enter the system?
- Sources of problems:
  - Manual entry
  - No uniform standards for content and formats
  - Parallel data entry (duplicates)
  - Approximations, surrogates – SW/HW constraints
  - Measurement or sensor errors.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial



Data Gathering - Solutions

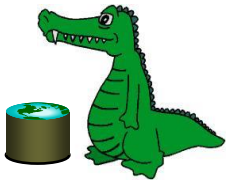
- Potential Solutions:
  - Preemptive:
    - Process architecture (build in integrity checks)
    - Process management (reward accurate data entry, data sharing, data stewards)
  - Retrospective:
    - Cleaning focus (duplicate removal, merge/purge, name & address matching, field value standardization)
    - Diagnostic focus (automated detection of glitches).

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Delivery

- Destroying or mutilating information by inappropriate pre-processing
  - Inappropriate aggregation
  - Nulls converted to default values
- Loss of data:
  - Buffer overflows
  - Transmission problems
  - No checks

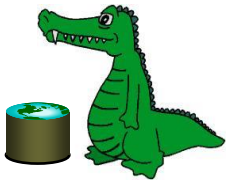
Adapted from Ted Johnson's SIGMOD 2003 Tutorial



Data Delivery - Solutions

- Build reliable transmission protocols
  - Use a relay server
- Verification
  - Checksums, verification parser
  - Do the uploaded files fit an expected pattern?
- Relationships
  - Are there dependencies between data streams and processing steps
- Interface agreements
  - Data quality commitment from the data stream supplier.

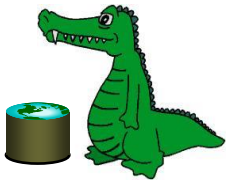
Adapted from Ted Johnson's SIGMOD 2003 Tutorial



Data Storage

- You get a data set. What do you do with it?
- Problems in physical storage
  - Can be an issue, but terabytes are cheap.
- Problems in logical storage
  - Poor metadata.
    - Data feeds are often derived from application programs or legacy data 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

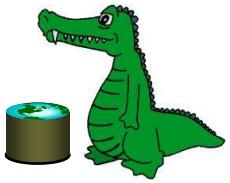
Adapted from Ted Johnson's SIGMOD 2003 Tutorial



Data Storage - Solutions

- Metadata
  - Document and publish data specifications.
- Planning
  - Assume that everything bad will happen.
  - Can be very difficult.
- Data exploration
  - Use data browsing and data mining tools to examine the data.
    - Does it meet the specifications you assumed?
    - Has something changed?

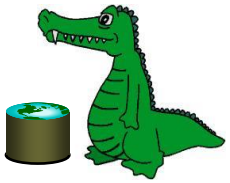
Adapted from Ted Johnson's SIGMOD 2003 Tutorial



Data Retrieval

- Exported data sets are often a view of the actual data. Problems occur because:
  - Source data not properly understood.
  - Need for derived data not understood.
  - Just plain mistakes.
    - Inner join vs. outer join
    - Understanding NULL values
- Computational constraints
  - E.g., too expensive to give a full history, we'll supply a snapshot.
- Incompatibility
  - Ebcdic? Unicode?

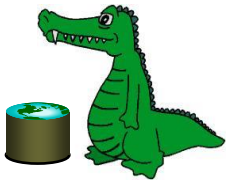
Adapted from Ted Johnson's SIGMOD 2003 Tutorial



Data Mining and Analysis

- What are you doing with all this data anyway?
- Problems in the analysis.
  - Scale and performance
  - Confidence bounds?
  - Black boxes and dashboards
  - Attachment to models
  - Insufficient domain expertise
  - Casual empiricism

Adapted from Ted Johnson's SIGMOD 2003 Tutorial



Retrieval and Mining - 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.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial