Data Matters

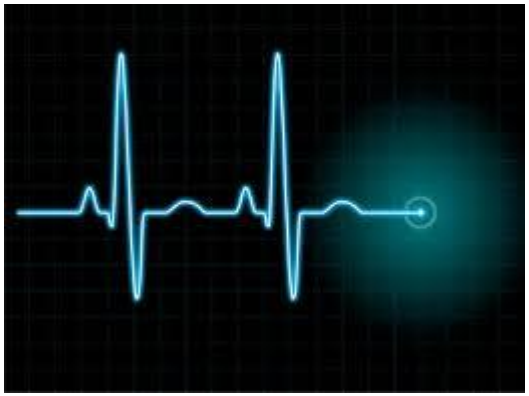
Professor Sally McClean
(si.mcclean@ulster.ac.uk)

Computer Science Research Institute
Coleraine

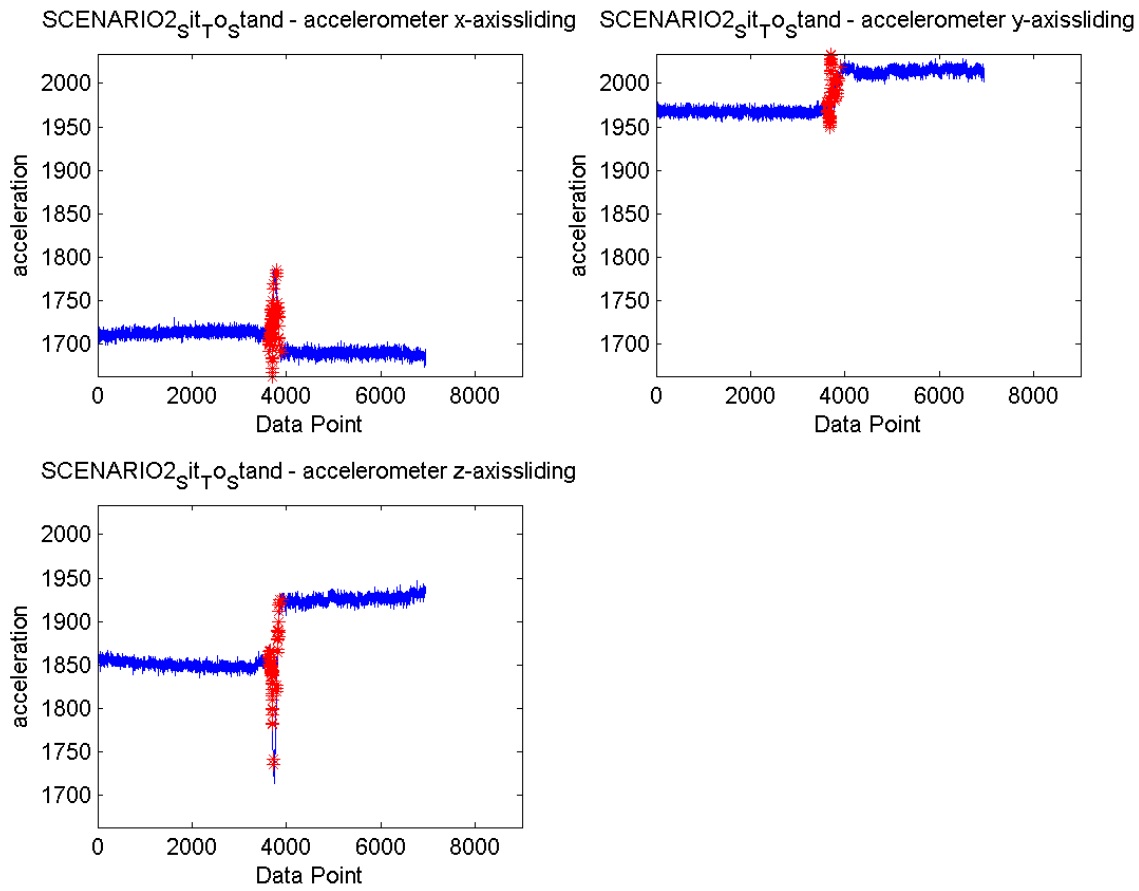
What data do we need?

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

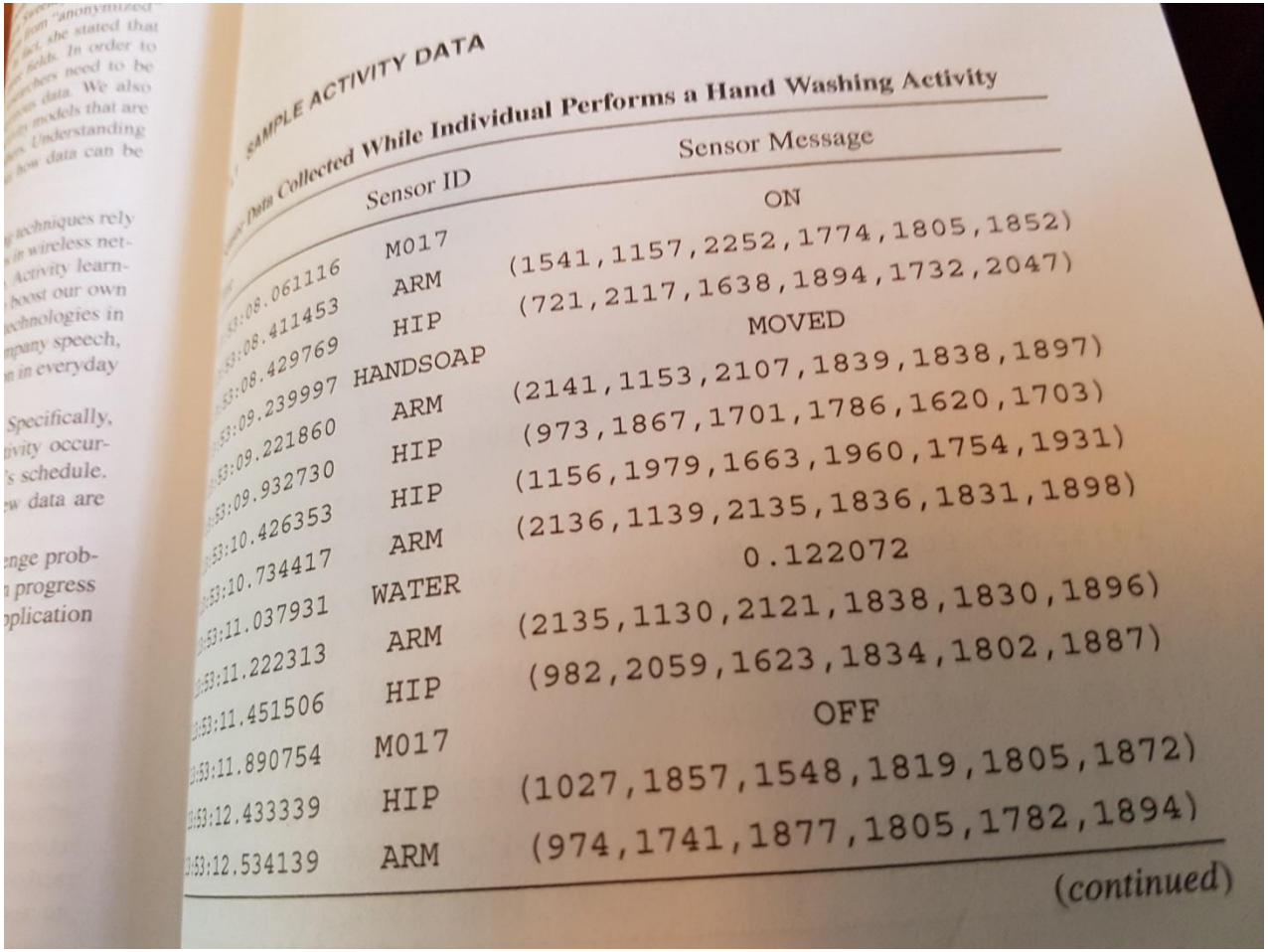
- Raw data
  - Accelerometer data, GPS, door sensor, activity label, gender date of birth, time-stamp., temperature.
- Features (derived data)
  - Age, Activity duration
- Statistical features
  - Mean pulse, variance of accelerometer readings
- Spectral features
  - Signal energy, Signal entropy
- Context features
  - Previous activities, location.
- Image Data



Sit to Stand: accelerometer data



Sample Activity Data (Cook and Krishnan, 2015)



SAMPLE ACTIVITY DATA

Sensor Data Collected While Individual Performs a Hand Washing Activity

	Sensor ID	Sensor Message
53:08.061116	M017	ON
53:08.411453	ARM	(1541,1157,2252,1774,1805,1852)
53:08.429769	HIP	(721,2117,1638,1894,1732,2047)
53:09.239997	HANDSOAP	MOVED
53:09.221860	ARM	(2141,1153,2107,1839,1838,1897)
53:09.932730	HIP	(973,1867,1701,1786,1620,1703)
53:10.426353	HIP	(1156,1979,1663,1960,1754,1931)
53:10.734417	ARM	(2136,1139,2135,1836,1831,1898)
53:11.037931	WATER	0.122072
53:11.222313	ARM	(2135,1130,2121,1838,1830,1896)
53:11.451506	HIP	(982,2059,1623,1834,1802,1887)
53:11.890754	M017	OFF
53:12.433339	HIP	(1027,1857,1548,1819,1805,1872)
53:12.534139	ARM	(974,1741,1877,1805,1782,1894)

(continued)

PAMAP2 Physical Activity Monitoring Data Set

Download: Data Folder, Data Set Description

Abstract: The PAMAP2 Physical Activity Monitoring dataset contains data of 18 different physical activities, performed by 9 subjects wearing 3 inertial measureme monitor.

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	3850505	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	52	**Date Donated**	2012-08-06
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	33884

Source:

Attila Reiss, Department Augmented Vision, DFKI, Germany, attila.reiss '@' dfki.de
Date: August 2012.

Data Set Information:

The PAMAP2 Physical Activity Monitoring dataset contains data of 18 different physical activities (such as walking, cycling, playing soccer, etc.), performed by 9 subjects wearing 3 inertial measurement units and a heart rate monitor. The dataset can be used for activity recognition and intensity estimation, while developing and applying algorithms of data processing, segmentation, feature extraction and classification.

http://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring

Attribute Information:

The 54 columns in the data files are organized as follows:

1. timestamp (s)
2. activityID (see across for the mapping to the activities)
3. heart rate (bpm)
4-20. IMU hand
21-37. IMU chest
38-54. IMU ankle

The IMU sensory data contains the following columns:
1. temperature (Â°C)
2-4. 3D-acceleration data (ms-2), scale: Â±16g, resolution: 13-bit
5-7. 3D-acceleration data (ms-2), scale: Â±6g, resolution: 13-bit
8-10. 3D-gyroscope data (rad/s)
11-13. 3D-magnetometer data (Î¼T)
14-17. orientation (invalid in this data collection)

List of activity IDs and activities:
1 lying
2 sitting
3 standing
4 walking
5 running
6 cycling
7 Nordic walking
9 watching TV
10 computer work
11 car driving
12 ascending stairs
13 descending stairs
16 vacuum cleaning
17 ironing
18 folding laundry
19 house cleaning
20 playing soccer
24 rope jumping
0 other (transient activities)

What is Data Quality?

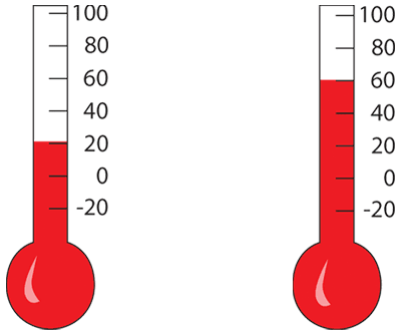
- Generally, you have a problem if the data doesn't mean what you think it does, or should
  - Data not up to spec : garbage in garbage out, glitches, gaps
  - You don't understand the spec : complexity, lack of metadata.
- Poor data quality has many sources and manifestations

Data Quality Problems

- Data quality problems are expensive and pervasive
  - DQ problems cost hundreds of billion $$$ each year.
  - Resolving data quality problems is often the biggest effort in a data analytics study.
- Poor quality data can invalidate the whole study

An Example

16	21/06/2016	run	8.23	34.45



- Can we interpret the data?
  - What do the fields mean?
  - What is the key? The measures?
- Data glitches
  - Typos, multiple formats, missing / default values
- Metadata and domain expertise
  - Field 1 is Temperature. In Fahrenheit or centigrade?
  - Field 5 is Duration. Is it *censored*?
    - How do we handle censored data?
    - How do we handle missing data?

Data Gliches

- Systemic changes to data which are external to the recorded process.
  - Changes in data layout / data types
    - Integer becomes string, fields swap positions, etc.
  - Changes in scale / format
    - Dollars versus euros
  - Temporary reversion to defaults
    - Failure of a processing step
  - Missing and default values
    - Application programs do not handle NULL values well …
  - Gaps in time series

Aspects of Data Quality

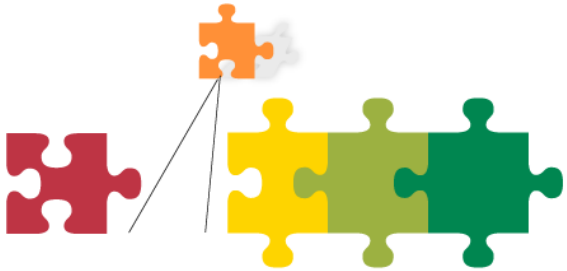
- Accuracy
  - Was the data recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
- Consistency
  - The data agrees with itself.
- Vagueness

What about interpretability, accessibility, metadata, analysis?

Aspects of Data Quality

- Missing, incomplete, ambiguous or damaged data
  - e.g truncated and/or censored data: there are specific methods of analysis to deal with these
  - otherwise: misleading results, bias.
- Highly variable data may be low quality
  - e.g. blurry images
- Vague data is often low quality
  - e.g. concepts are too high level
- Suspicious or abnormal data
  - e.g. outliers
- Departure from models
  - Goodness-of-fit

Missing Data

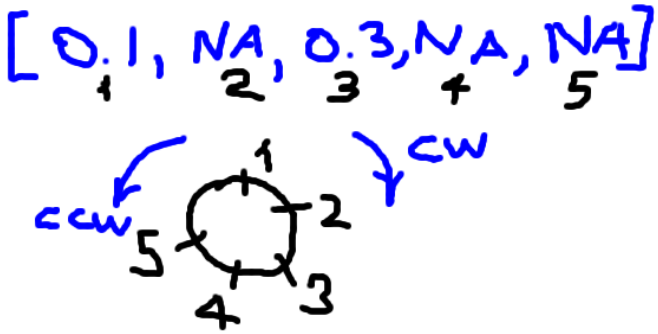


- Values, attributes, entire records, entire sections
- Data truncation and/or censoring – there are specific methods of analysis to deal with these
  - Otherwise: Misleading results, bias.
- Detecting and Analysing Missing Data
- Structured Missing Data

Using Data Imputation

- Imputation is a strategy for coping with Missing Data
- In federated data, between 30%-70% of the data points may have at least one missing attribute - data wastage if we ignore all records with a missing value
- The remaining data is likely to be seriously biased
- This leads to a lack of quality in the data and a lack of confidence in the results
- Understanding the pattern of missing data can unearth data integrity issues

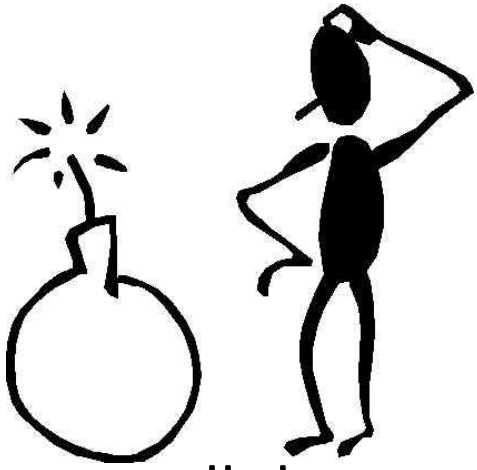
Mechanisms for Data Imputation



- Standalone imputation
  - Mean, median, other point estimates
    - Assumes the distribution of the missing values is the same as the non-missing values.
    - Does not take into account inter-relationships
    - Introduces bias
    - Convenient, easy to implement

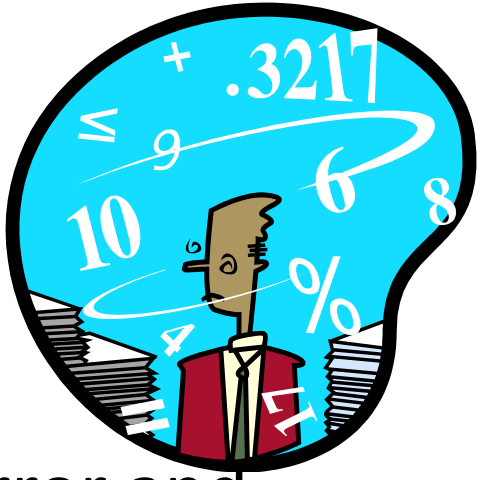
- Better imputation - use attribute relationships

Checking for Problem Data



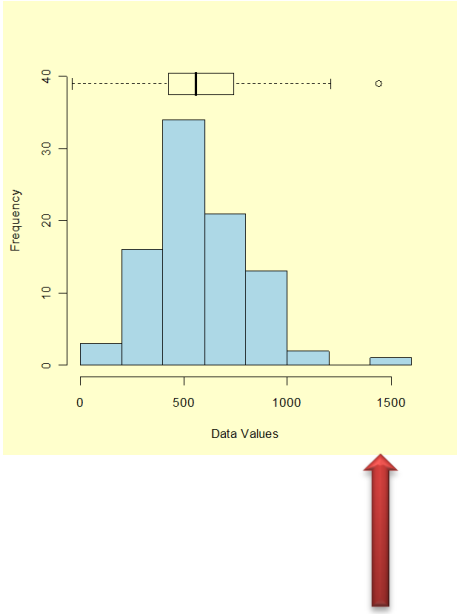
- Match data specifications against the data - are all the attributes present?
- Scan individual records - are there gaps?
- Rough checks : number of files, file sizes, number of records, number of duplicates
- Compare estimates (averages, frequencies, medians) with “expected” values and bounds; check at various levels of granularity since aggregates can be misleading.
- Values are truncated or censored - check for spikes and dips in distributions and histograms

Outlier Analysis

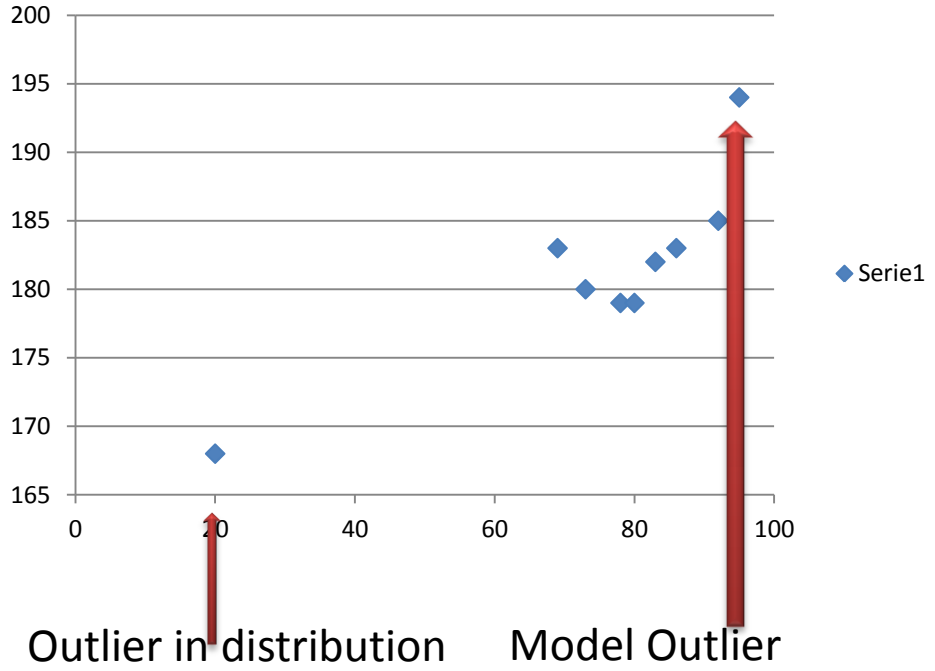


- We may decide that an outlier is an error and eliminate it from the study.
- However, outliers are often the most interesting part of the data- exception mining.
- Robust statistics and robust analysis circumvent the outliers
  - Median, semi inter-quartile range (SIQR) are robust metrics (not sensitive to outliers).
  - Mean, variance, range are not robust
  - Robust regression, robust ANOVA etc. are techniques that are not influenced by outliers.

Example Outliers



Outliers in distribution



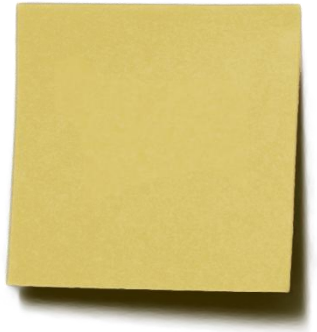
Outlier in distribution

Model Outlier

Data and Metadata

- Supplementary information about the Data is often stored alongside.
- This may include details of the data collection methodology, units, codes, meanings, missing data types, names, relationships, domain knowledge etc.
- Poor metadata is associated with poor quality data.

Annotation



- **Annotation** is a type of metadata _(e.g. a comment, label or explanation) attached to text, image, or other data. Often, annotations make reference to a specific part of the original data.
- Markup languages like XML can annotate in a way that is syntactically distinguishable from the data.
- Annotation can be used to add information about an image, or machine-readable semantic information, as in the semantic web.
- Automated annotation for Activity Learning is a current research topic e.g.
  - S. Szewcyzk, K. Dwan, B. Minor, B.Swedlove, and D. Cook Annotating Smart Environment Sensor Data for Activity Learning. Technology and Health Care 17.3 (2009): 161-169.
  - eecs.wsu.edu/~cook/pubs/th09.pdf

Data (and Metadata) Storage

- On a sensorised device e.g. a smart phone (in a compressed form)
- On a local server
- In the Cloud – data centre, data farm.

- Openstack- Python based cloud management system.
- Openstack Trove: openstack database-as-a-service (DaaS)– SQL and NoSQL.

Cloud Computing?

1. Web-scale problems
2. Distributed, heterogeneous and parallel.
3. Large data centers
4. Different models of computing
5. Highly-interactive Web applications

Key technologoes

- Virtualization- orchestration engine
- Hadoop includes:
  - Distributed File System - distributes data
  - Map/Reduce - distributes application
  - Elasticity
  - Batch-oriented processing of large datasets
- Ajax: the “front-end” of cloud computing
  - Highly-interactive Web-based applications

Characteristics

- Cloud computing is an umbrella term used to refer to Internet based development and services.
- A number of characteristics define cloud data, applications services and infrastructure:
  - Remotely hosted: Services or data are hosted on remote infrastructure.
  - Ubiquitous: Services or data are available from anywhere.
  - Commodified: The result is a utility computing model similar to gas or electricity - you pay for what you get!
- Cloud Computing is very data-intensive

  -May also be processing intensive

Semantic Models for the Cloud

- The cloud landscape is diverse and heterogeneous, making interoperability a major challenge. Adding semantics can help address such issues.
- Heterogeneity arises from different sources, and may be vertical or horizontal, potentially involving different abstraction levels, different services, different software and different types of data.
- Such heterogeneity may be bridged via common underlying semantics to improve inter-operability, integration and reasoning.

Modelling of semantic concepts

- The retrieval of data and metadata based on semantic content is essential to handle huge amounts of data.
- Many application domains require data mining and data analytics systems that can operate efficiently and handle heterogeneous data
- Tools are therefore required to facilitate the handling of huge volumes of data by semantically linking and managing them.
- The modelling of semantic concepts in data and metadata represents an attempt to overcome the semantic gap between humans and machines.
- This gap that can be bridged by the use of software agents that can automatically recognise certain semantic concepts.

Heterogeneous Data (and Metadata) Integration

- Many tools for address matching, schema mapping are available.
- Many hidden problems and meanings : must extract metadata.
  - Understanding NULL values
- Computational constraints
  - e.g., too expensive to give a full history, we'll provide a segment.
- Incompatibility
- Time synchronization

- **WE MAY UNEARTH NEW KNOWLEDGE IN THE PROCESS**

Two sensor data fusion

- Each sensors take a measurement: z_1 and z_2 e.g. temperature.
- The fused weighted estimator is:

  $z=k.z_1+(1-k).z_2$ where $0<k<1$

- The weighted variance estimator is:

  $z=(\sigma_1^{-2}/(\sigma_1^{-2}+\sigma_2^{-2})\, z_1 + (\sigma_2^{-2}/(\sigma_1^{-2}+\sigma_2^{-2})\, z_2$

  where σ_1^2 and σ_2^2 are the respective variances.

- k can take other values and we can have more than two sensors.
- NB if k=0.5 our fused value is just the mean.
- Here a high variance (low quality) gets a low weight and vice versa.

Data Fusion

D	E	F
	TEMP A	TEMP B
	35.125	33.3125
	35.1875	33.375
	35.25	33.4375
	35.3125	33.5
	36.4375	34.5625
	36.625	34.75
	36.6875	34.75
	37.1875	35.0625
	37.25	34.5625
	37.5	35.4375
mean	36.25625	34.275
variance	0.89796	0.625174

The Mechanisms

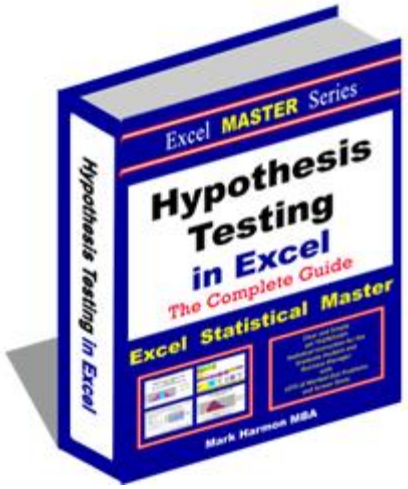
- The "Global As View" (GAV) approach provides a mediated mapping from the mediated schema to the original sources,
- The "Local As View" (LAV) approach provides a mapping from the original sources to the mediated schema This requires more sophisticated inferences to resolve a query on the mediated schema, but makes it easier to add new data sources to a (stable) mediated schema.
- Semantic integration addresses the structuring of the architecture of the integration, and resolves semantic conflicts between heterogeneous data sources.
- A common strategy involves the use of ontologies to explicitly map the schema.

Evaluation

- Evaluation means making observations of all aspects of our research.
- We need to clarify why our research is important, how it improves on the state-of the art, that it is correctly implemented and how we might improve on it.
- Without evaluation we cannot replicate results.
- Hypothesis testing is fundamental to evaluation.

Cohen and Howe (1988). How Evaluation Guides AI Research

Hypothesis Testing



To do this, in each case we have a:

Null Hypothesis (H_o): which is a statement of null effect, e.g. 'there is no association between x and y', and an:

Alternative Hypothesis (H_1): which is a statement of effect, e.g. 'there is an association between x and y'.

Tests of Significance

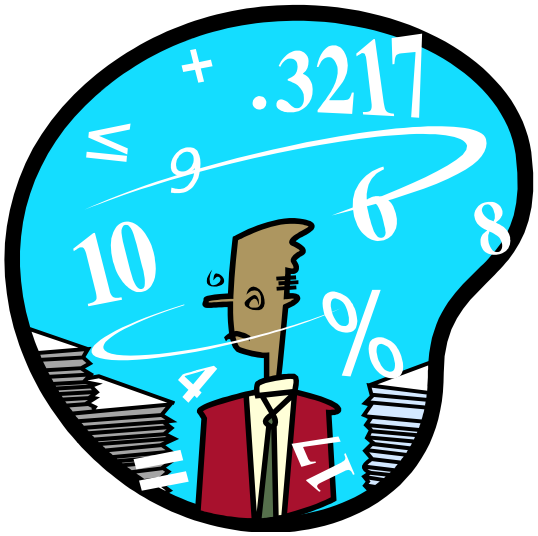
- The test of significance allows us to decide which of the two hypotheses (H_o or H_1) we should accept.

- We say that a result is significant at the 5% level if the probability that the discrepancy between the actual data and what is expected assuming the null hypothesis is true has probability less that 0.05 of occurring.

The Chi-Squared Test

A common type of hypothesis test for categorical data is a test of association between different categorical variables. This is called a chi-squared test.

The null hypothesis being tested here is
H_o: no association
against the alternative:
H_1: there is an association between two categorical variables.



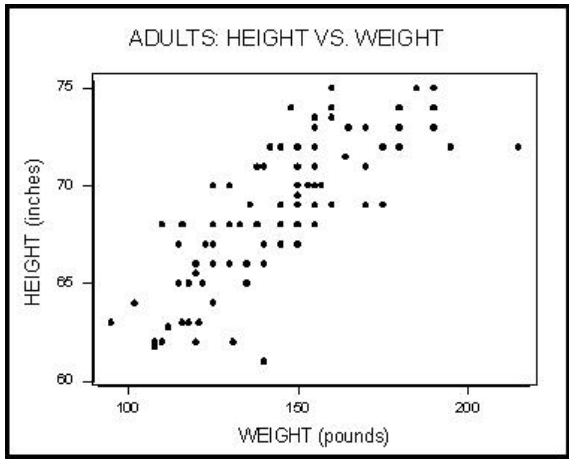
Example:- smoking habit and gender

	Smoker	Non-smoke r	ALL
Male	20	37	57
Female	8	27	35
ALL	28	64	92

CHI-SQUARE = 1.532

The critical chi-square value (5% significance level) is 3.84. Since the calculated value (1.532) is less than the critical value, we favour the null hypothesis that there is no association between smoking habit and gender.

Associations for Continuous Data



- For continuous data, we measure association between two variables using the **correlation coefficient** (**r**) where the correlation coefficient always lies between -1 and +1.
- If it is -1 then we have a perfect negative correlation, zero means no correlation and +1 means perfect positive correlation.
- So, for example, a correlation of 0.8 between weight and height would mean that there is a strong positive correlation between weight and height i.e. tall people tend to be heavier than short people.
- As for categorical data, we use hypothesis tests to test significance of the correlation coefficient i.e. H_0: r=0 V $H_{1:}$ r≠0.

Height and Weight for PAMAP2 participants

Height	Weight
182	83
169	78
187	92
194	95
180	73
183	69
173	86
179	87
168	65

Is there a significant correlation between height and weight?

$$t = \frac{r}{\mathrm{sqrt}\left[(1-r^2)/(N-2)\right]}$$

Degrees of freedom=N-2

Difference in Means

- We can use ***(Student's) t-test*** to test whether a difference in means is statistically significant.
- If the data are paired we use a ***paired*** t-test, which can determine differences to be significant when the training sets are the same for both systems.
- If the data are not paired we use an unpaired (independent sample) t-test.
- Alternative statistical tests have been proposed (if the data are not normally distributed), such as McNemar's test for equality of medians.
- Although no test is perfect when data is limited and independent trials are not practical, some statistical test that accounts for variance is highly desirable.

Suaave.org

Sensing, **U**nmanned, **A**utonomous **A**erial **VE**hicles

An Example
SUAAVE:
Combining Aerial Robots and Wireless Networking

UCL

University of ULSTER

UNIVERSITY OF OXFORD

EPSRC Engineering and Physical Sciences Research Council

Safe Landing Zones- SLZ (Dr. Timothy Patterson)

- Need to land safely in event of failure
- Potential sites assessed in terms of
  - Attainability
  - Distance from man made structures
  - Terrain classification
    - Grass, Gorse, Rock, Trees, Water
  - Roughness
    - Smooth, Rough, Very Rough
- Image processing techniques used for terrain classification
  - Roughness based on variation in greyscale intensity
  - Edge detection used to segment image
  - Maximum likelihood classifier for terrain classification

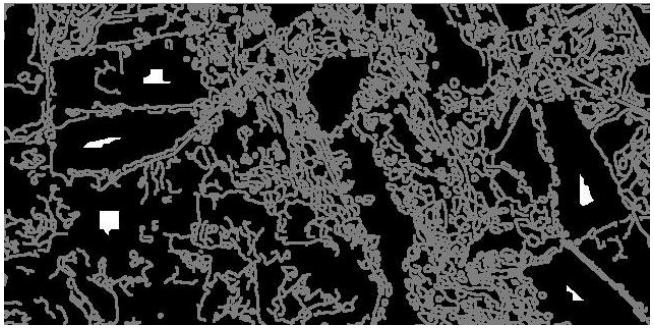
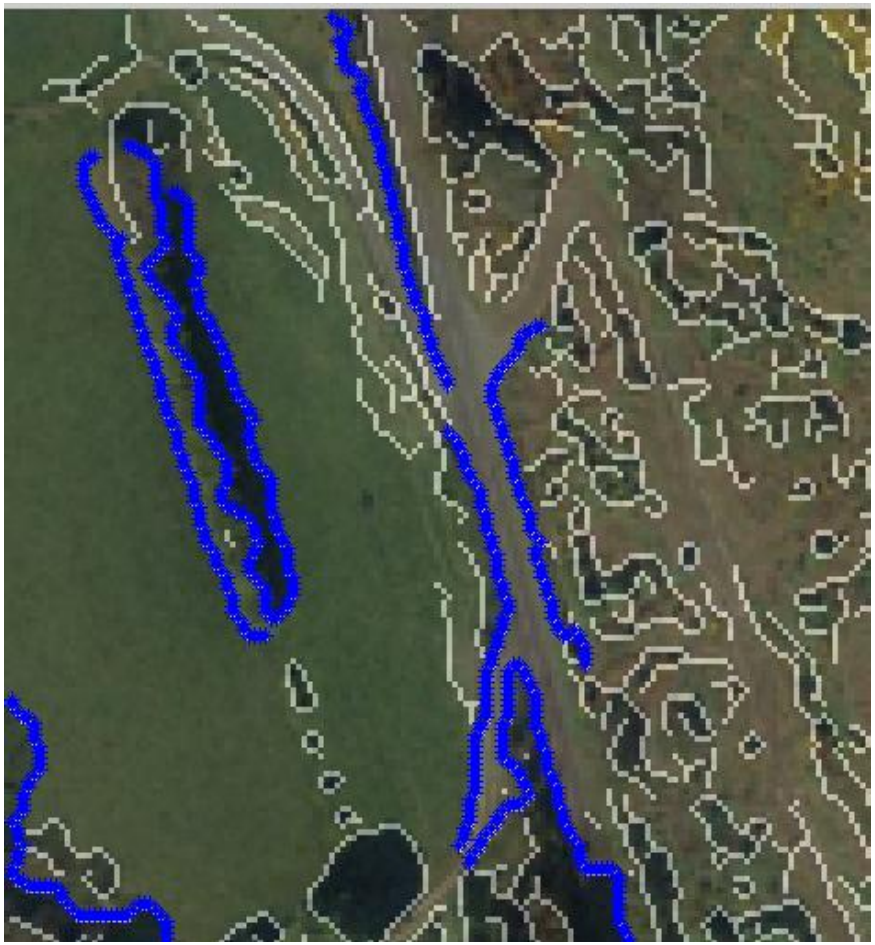


Image pre-processing



Pixel Classification

- The probability provided by the Maximum Likelihood Classifier $p(\mathbf{x}|C_k)$ can be combined with prior knowledge, $p(C_k)$ (e.g. information derived from a map) using Bayes' rule

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})}$$

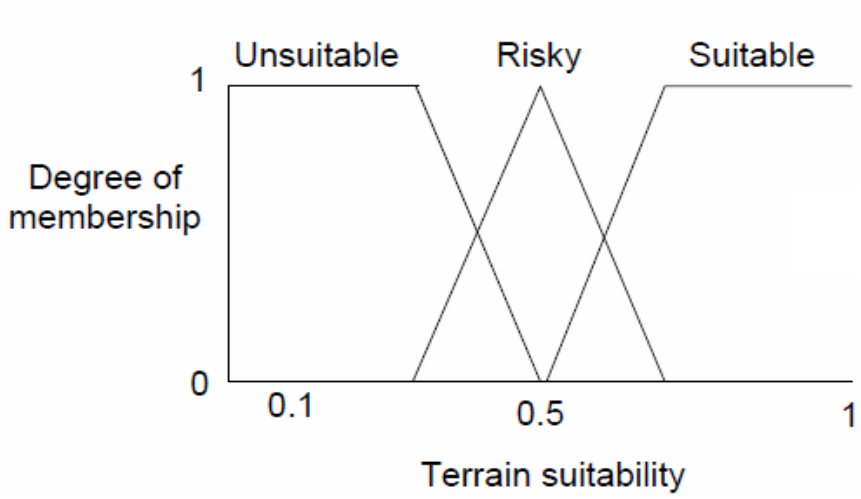
- Only pixels which can be relatively accurately classified and have a consistent corresponding OS feature code are included in the output. These are feature codes {1090, 1097, 1131}, i.e. {road, path, water}
- The result of aerial image pre-processing is a matrix containing feature codes of the same dimensions as the rescaled input aerial image

Multi-Resolution Terrain Classification

- The Multi-Resolution Expectation Maximization (MREM) algorithm fuses historic knowledge in the form of OS and training data with multiple images of a scene (from different heights or different UAVs) to improve terrain classification.
- Enable the efficient weighting of an image's contribution to the classification of an area
  - Altitude of capture
  - Time since capture
  - Sensor type
- Apportion an image's contribution to the classification of an area according to its coverage of that area

SLZ detection – Terrain Suitability

A numerical suitability measure in the range [0..1] is assigned to each terrain type by a human expert familiar with the operational area. This suitability measure is used to assign a fuzzy classification of unsuitable, risky, or suitable to each terrain type



Terrain type	Suitability
Grass	0.9
Gorse	0.7
Rock	0.5
Trees	0.3
Water	0.1

Experimental results/Evaluation

For the purposes of evaluating SLZ detection accuracy, SLZs for100 aerial images were manually validated by a human expert.

A true positive (TP) is a correctly identified SLZ with a high safety weighting, a true negative (TN) is a correctly identified SLZ with low safety weighting. Similarly for false positive (FP) and false negative (FN)

Table 4: Validated results based on SLZs detected within 100 images.

	TP	TN	FP	FN	Total SLZs
Incl. Knowledge	688 (76%)	169 (18.7%)	0	47 (5.3(%)	904
No Knowledge	807 (78%)	86 (8.3%)	104 (10.1%)	37 (3.6%)	1034

Overall it was found that incorporating knowledge provided a more reliable method of SLZ detection with 94.7% of potentially suitable SLZs assigned a correct safety score. In contrast when knowledge was not included there were a significant amount of false positives (10.1%) with 86.3% of potential SLZs being assigned a correct safety score.


The end