Basic Issues

- Can existing datasets be effectively used for new research?

- When collecting data for one purpose, is it cost-effective (or even possible) to design it for multiple purposes?

- Can researchers ‘trust’ the data and metadata from existing sources?

- Will the sponsoring organizations allow the use of data from outside sources? IRB, copyrights, MOUs…
Basic Issues

- Is the size and ‘depth’ of databases sufficient?
- Has ground truth been established?
- Have parameters of data collection been sufficiently categorized?
- How to the ‘holes’ in data needs get met?
- Can multiple existing data sources be effectively used instead of collecting new data for each study or grant?
Is there data available?

- Need to ask:
  - What data characteristics are needed to test / prove / disprove a hypothesis?
  - Will the study be sufficiently affected if one or more of those characteristics is not present?

- Database Catalogue as a resource
  - To be covered as a talk in this conference
  - Lists major characteristics of the known databases
Example: Is altered data suitable for research?

- CASIA-IrisV1 dataset

From the website description: “In order to protect our IPR in the design of our iris camera (especially the NIR illumination scheme), the pupil regions of all iris images in CASIA-IrisV1 were automatically detected and replaced with a circular region of constant intensity to mask out the specular reflections from the NIR illuminators. Such editing clearly makes iris boundary detection much easier but has minimal or no effects on other components of an iris recognition system, such as feature extraction and classifier design.”
What about ‘outliers’?

Datasets may be ‘pure’ examples of ‘normal’ conditions or contain unusual cases. For iris, this may be optical diseases, or induced conditions (permanent or temporary) such as tattooed sclera or patterned / colored contact lenses. Images from BBC http://www.bbc.com/news/magazine-30750361
Controlled degraded data

Examples from Q-FIRE (Quality in Face and Iris Research) - Clarkson University

See http://www.clarkson.edu/biosal/pdf/quality.pdf
Training to data

- Databases may become so familiar to researchers that they tune their systems to ensure ‘success’

Note

Problems

On left slap

Image -- from

NIST Special

Database 29
NIST software testing: Use of sequestered data

**Testing Timeline**

1. Collect Data
2. Supply K% of Data to Capable Organizations for Development
3. Retain 100-K% as Sequestered Data
4. Development phase
5. Post vendor-reported Results
6. Host Workshop to present Analysis and Define Next Steps
7. Loop to Step 1 as needed

8. Acquire (software) implementations for formal evaluation
9. Execute SDK-based Independent Evaluation using Sequestered Data
10. Report
Dataset size

- What is the minimum size needed to answer the question?
- What is the amount of data that can be realistically / cost-effectively acquired?
- Will it be possible to extrapolate the analysis to larger dataset sizes? Should that be done?
- What are the assumed characteristics of the dataset as it grows?

Chart from: SUMMARY OF NIST STANDARDS FOR BIOMETRIC ACCURACY, TAMPER RESISTANCE, AND INTEROPERABILITY November 13, 2002
Establishing ground truth

- Was data known at the time of capture (such as who spoke, what the words spoken were to be …) ?

- Was data established by experts (such as when established ‘matches’ between a latent friction ridge print and an exemplar) ?
Changes in features

- Possible? (Tattoo removal or alteration)
- Should these be included in datasets?
- Impossible? (Tooth with root canal becoming a virgin tooth)
- Should these be ‘seeded’ in datasets?

http://www.newlookhouston.com/Tattoo_Removal_Photos.html
Data pedigree

- Has data been (unintentionally) altered? Such as through lossy compression of an image:

- From “Comparison of the WSQ and JPEG 2000 Image Compression Algorithms On 500 ppi Fingerprint Imagery”
Aging of data

- Should samples be taken sequentially, in the same or different sessions?
- Should data (such as photos) be used that were taken in ‘uncontrolled’ circumstances earlier?
- Is the object of the study the impact of aging or is aging a variable to be considered in analysis of the data?
- Can data from earlier in one’s life be compared against current data for the modality under study?

http://www.nist.gov/customcf/get_pdf.cfm?pub_id=915238

CHANGES IN THE VOICE AS WE AGE

Below is a list of commonly reported voice changes as we age:

- Higher pitch voice in men
- Lower pitch voice in women
- Reduced volume and projection of the voice (or thin voice)
- Reduced vocal endurance
- Difficulty being heard in noisy situations
- Tremor or shakiness in the voice

from American Society of Otolaryngology http://www.entnet.org
Metadata about collection characteristics – voice example

Voice analysis factors:
- Recording device type (including microphone)
- A-D conversion?
- Data Compression?
- Frequency clipping?
- Environment

- Redactions /
- Discontinuities
- Vocal content
- Type of speech (coerced, reading, spontaneous…)
- And more

Illustration from www.Bnoack.com (AV Info Non-Profit Informational Website)
# Protect Subject’s Privacy

## Institutional Review Boards (IRBs) in US

- **National Research Act of 1974**
- **Governed by Title 45 of Federal Regulations, Part 46**
- **Exemptions**
  - Research involving analysis of existing data and other materials if they are publicly available, or where the data can be collected such that individual subjects cannot be identified in any way

## Universal Principles

- People should be informed if their personal information is being collected
- Personal information should only be used for the purpose for which it was collected
- Personal information should only be collected for a clearly identified purpose

As stated by the **Office of the Privacy Commissioner of Canada** [www.priv.gc.ca](http://www.priv.gc.ca)
Taking part is voluntary: Taking part in this study is completely voluntary. You may skip any questions that you do not want to answer. If you decide not to take part or to skip some of the questions, it will not affect your current or future relationship with Cornell University. If you decide to take part, you are free to withdraw at any time.

If you have questions: The researchers conducting this study are Randy Jackson and Prof. Simon Cowell. Please ask any questions you have now. If you have any questions later, you may contact Randy Jackson at randy@blabmail.com or at 1-800-555-4365. You can reach Prof. Cowell at meanGuy@abbey.uk or 1-800-555-4365, ext. 1000. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) at 607-255-5138 or access their website at http://www.irb.cornell.edu. You may also report your concerns or complaints anonymously through Ethicspoint (www.hotline.cornell.edu) or by calling toll free at 1-866-293-3077. Ethicspoint is an independent organization that serves as a liaison between the University and the person bringing the complaint so that anonymity can be ensured.

You will be given a copy of this form to keep for your records.

Statement of Consent: I have read the above information, and have received answers to any questions I asked. I consent to take part in the study.

Your Signature ___________________________ Date ___________________________

Your Name (printed) ___________________________

In addition to agreeing to participate, I also consent to having the interview tape-recorded.

Your Signature ___________________________ Date ___________________________

Signature of person obtaining consent ______________ Date ______________

Printed name of person obtaining consent ___________________________

Date ___________________________

This consent form will be kept by the researcher for at least three years beyond the end of the study.

The title of the study should appear at the top of every page.

http://www.irb.cornell.edu/forms/sample.htm
Sample Database Use Form (License Agreement) from Notre Dame University

http://www3.nd.edu/~cvrl/LicenseAgreements/UNDLicenseAgreementF.pdf
Cross-cultural issues

- Capture procedures for some biometrics may touch sensitive cultural issues (such as removing face veils)

- Paternity and familial relationship testing through DNA – how to handle socially sensitive results and how to handle perceived differences in definitions of a family

Cost Considerations

- Standards for how to collect data; amount needed, etc. Such as ISO/IEC 29794 that directly impact costs (number of participants, number of separate visits, etc.)

- Mike Thieme will discuss standards later in this conference

- Collection ‘pristine’ data for each experiment ….
  - Paying participants and/or collection agents

- Spend time and money to verify existing datasets …
  - Signing use agreements
  - Review of data
Impacts upon Teaching

- Should data be collected as part of the teaching process? (Hear talks Stephanie Schuckers and Jeremy Dawson)
- How to design experiments
- How to perform statistical analysis
- How to deal with regulatory requirements
Publishing study results

- Possible limitations based upon the terms of use of the datasets
  - Can the images (or voice samples, etc.) be made publicly available in journals and other publications?
  - How is reproducibility testing by other organizations to be done?
  - Etc.

- Should data be aggregated for presentation to avoid privacy issues? Etc.

From:
IREX III Supplement I
Failure Analysis

http://www.nist.gov/customcf/get_pdf.cfm?pub_id=910385

Figure 9: On the left are examples of highly compressed images. The graph on the right shows the distribution of file sizes for the overall dataset.
Company-sponsored data

- Who does the research benefit?
- Why should people participate?
- Who ‘owns’ rights --- DNA lines, etc.
- Can / should company-sponsored data collection be used by researchers?

From: The Hand Vein Pattern Used as a Biometric Feature  
Master Literature Thesis by Annemarie Nadort, vreije Universiteit Amsterdam

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Test Images</th>
<th>Match Attempts</th>
<th>FAR</th>
<th>FRR</th>
<th>FTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross and Smith - 1995 [54]</td>
<td>20</td>
<td>2</td>
<td>40 genuine attempts</td>
<td>760 impostor attempts</td>
<td>0%</td>
</tr>
<tr>
<td>Wang and LeeRam - 2005 [1]</td>
<td>12</td>
<td>6</td>
<td>72 genuine attempts</td>
<td>792 impostor attempts</td>
<td>0%</td>
</tr>
<tr>
<td>Tsinghua University [112]</td>
<td>13</td>
<td>5 (and reference image)</td>
<td>260 genuine attempts</td>
<td>3120 impostor attempts</td>
<td>0%</td>
</tr>
<tr>
<td>Harbin University [112]</td>
<td>48</td>
<td>5 (and reference image)</td>
<td>960 genuine attempts</td>
<td>45120 impostor attempts</td>
<td>0%</td>
</tr>
<tr>
<td>Watanabe - 2005 [102]</td>
<td>70,000</td>
<td>1</td>
<td>14,000 genuine attempts</td>
<td>1,960 impostor attempts</td>
<td>0.0008%</td>
</tr>
<tr>
<td>Kondo - 2002 [83]</td>
<td>678</td>
<td>1</td>
<td>678 genuine attempts</td>
<td>45906 impostor attempts</td>
<td>0.04%</td>
</tr>
<tr>
<td>Miura - 2006 [81]</td>
<td>678</td>
<td>1</td>
<td>678 genuine attempts</td>
<td>45906 impostor attempts</td>
<td>1%</td>
</tr>
<tr>
<td>Miura - 2004 [82]</td>
<td>678</td>
<td>1</td>
<td>678 genuine attempts</td>
<td>45906 impostor attempts</td>
<td>0.145%</td>
</tr>
<tr>
<td>Lin and Fan - 2004 [67]</td>
<td>32</td>
<td>15</td>
<td>480 genuine attempts</td>
<td>149000 impostor attempts</td>
<td>3.5%</td>
</tr>
</tbody>
</table>

Table 1: Overview of the performance experiments performed for some of the methods described in section 3.1 and 3.2.

\[ FRR = \frac{NFR}{NEV} \]  
\[ FAR = \frac{NFA}{NIVA} \]  

Where:
- FRR is the false rejection rate  
- NFR is the number of false rejections  
- NEVA is the number of enrollee verification attempts
- FAR is the false acceptance rate  
- NFA is the number of false acceptances  
- NIVA is the number of impostor verification attempts

From: The Hand Vein Pattern Used as a Biometric Feature  
Master Literature Thesis by Annemarie Nadort, vreije Universiteit Amsterdam
Other factors

- When collecting data
  - ergonomics
  - accessibility
  - universality
  - understanding of instructions

- Formatting of data
  - APIs for access, etc.

Those wishing to submit software for MINEX testing shall be required to provide NIST with an SDK (Software Development Kit) library which complies with the API (Application Programmer Interface) specified in this document. At a minimum, the SDK submitted must provide functionality to create MINEX compliant templates based on individual fingerprint images. Support for matching pairs of MINEX compliant templates is encouraged, but optional.


http://zing.ncsl.nist.gov/biousa/docs/HFES%20instructions.ppsx