Issues in the collection and use of biometric and forensic datasets

Austin Hicklin

26 January 2015

noblis

For the best of reasons

Why to be a cautious collector — and skeptical consumer — of datasets

Examples

Months of analysis conducted before finding that 15% of the images had been removed because a fingerprint examiner "thought they would be difficult to match"

AFIS evaluations where the "ground truth" associations between fingerprints had been made by another AFIS (thereby omitting all data that AFIS couldn't match)

Data collected from a very limited population (e.g. engineering grad students at a specific university), but gender/age/ethnicity information is not retained

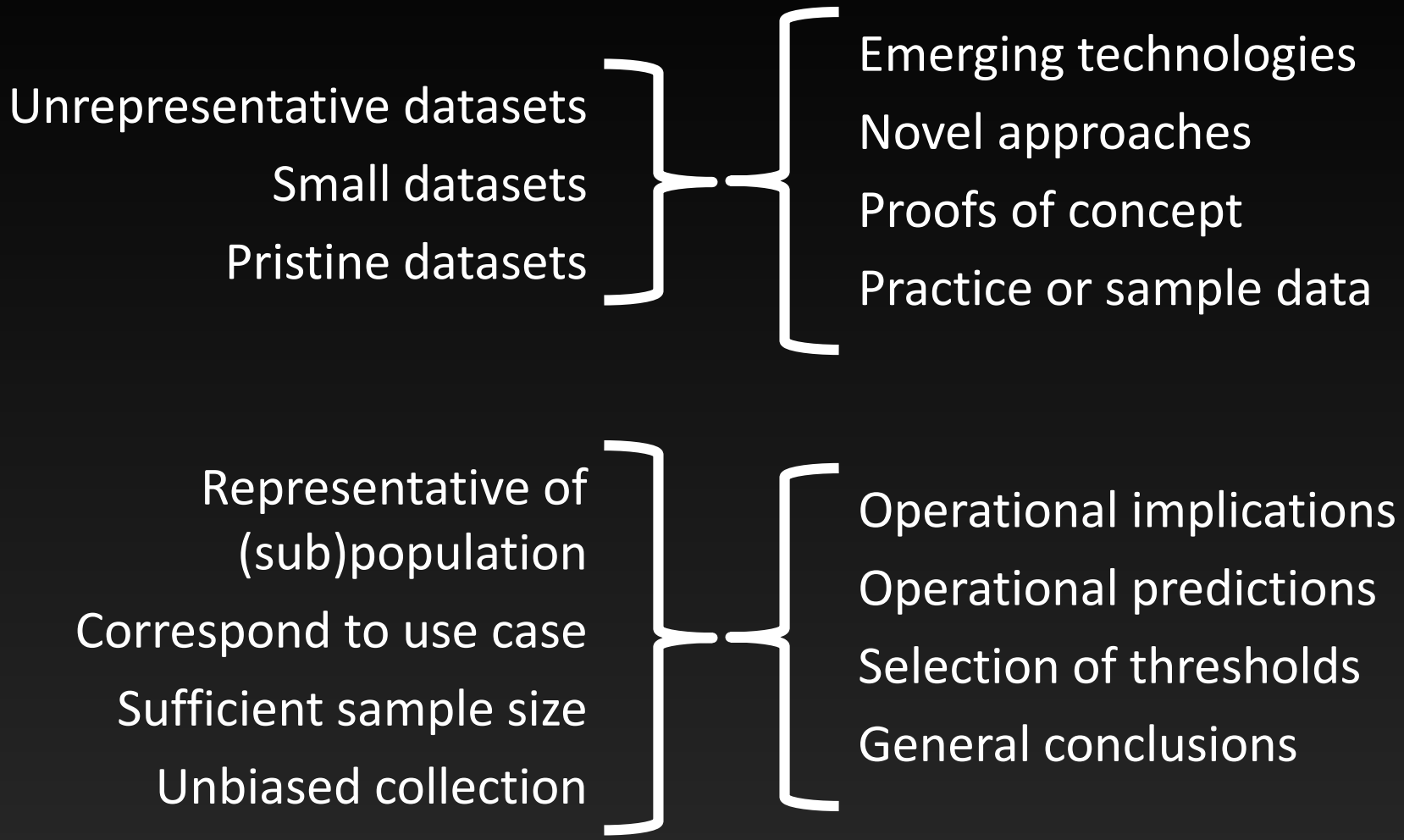
Extensive time spent on statistical measures of confidence for a dataset that was never intended to be representative of anything — and all subsequent dataset collections are wildly outside those bounds

Small, unrepresentative datasets used as the basis for

- Operational technology thresholds
- Policy decisions
- Overstated conclusions in reports and journal publications

Problems in the collection or misuse of datasets can go very wrong

Datasets need to be appropriate for a given purpose



Dataset problems and implications

Collection

Ad hoc data collection
Poor quality control

Dissemination

Inadequate documentation, so the consumer cannot fully understand the data

Usage

Treating arbitrary data as representative

Aspects of Representativeness

- Representative of specific (sub)population
  - Demographics (age, sex, race, etc)
- Representative of use case
  - Data quality
  - Data attributes
    - *collection methods/devices, processing methods, formats, compression methods, etc*
- Unbiased collection
  - (see next)
- Curse of dimensionality
  - Difficult to be representative of many dimensions

Collection biases

- Data collection often perturbs the representativeness of data
- Convenience samples, samples of opportunity, or self-selected samples = explicitly non-representative
- Survivor(ship) bias
  - Many collection processes implicitly or explicitly filter data, e.g.
    - *"ground truth" subject attribution*
    - *Using AFIS to select data*
- Bias often cannot be detected through quality metrics

Availability of data

- Public data is necessary for research, but
  - In some cases, release not possible due to privacy issues or data sensitivity
  - Sequestered data is necessary for evaluations

- Some options in limited access to data
  - E.g. not releasing sensitive data but allowing researchers to submit algorithms & releasing results

Best practices guide

- In progress:
  - *Best practices in the collection and use of biometric and forensic datasets*