Some standard disclaimers:

"Certain equipment, instruments, procedures, and/or materials are identified in this presentation in order adequately to specify the experimental procedure. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials, procedures, or equipment identified are necessarily the best available for the purpose."

"The opinions expressed in this presentation are those of the author, and do not reflect the opinion of NIST."

- NIST's roles
  - Standards
    - Establishing standards
    - Building / identifying consensus
    - Suggesting when consensus is absent
  - Technology
    - Evaluating state-of-the-art
    - Suggesting when state-of-the-art is nebulous
  - Other roles as assigned by Congress, by Executive Branch, by statute [e.g., Patriot Act]
  - Measurement

- On measurement
  - Accuracy
    - Not the same as precision
  - Reliability / Repeatability
    - Confidence intervals
    - Probability
    - Functions
      - Analytic [e.g., Gaussian, Poisson, Weibull]
      - Empirical
  - Relevance
    - Measuring that which is relevant to the system
    - Not making measurements just because they are easy

Topics – about five minutes each

- Context
- Determination
- Limits
- Implications of limits

Topics

- **Context**
- Determination
- Limits
- Implications of limits

Contexts of image-based biometric ground truth

- Identity
  - Fundamental question
  - Primary business process
  - Certainty difficult/impossible in operational env’t
  - Certainty feasible, not guaranteed, in lab env’t
- Attributes of image
- Attributes of subject
- Test environment
- Operational scenario

Contexts of image-based biometric ground truth

- Identity
- Attributes of image
  - Intrinsic attributes certain
    - e.g., height, width, pixel depth
  - Extrinsic attributes ‘not so much’
    - e.g., impression type, scanner
- Attributes of subject
- Test environment
- Operational scenario

Contexts of image-based biometric ground truth

- Identity
- Attributes of image
- Attributes of subject
  - e.g., date of birth, place of residence
  - Secondary business process
  - hit [usually] or miss
- Test environment
- Operational scenario

Contexts of image-based biometric ground truth

- Identity
- Attributes of image
- Attributes of subject
- Test environment
  - NIST test environment
  - NITB
    - CMF
    - IQMI
- Operational scenario

Contexts of image-based biometric ground truth: NIST’s test environment

- Test data
  - Database or repository
  - Probe / query set
  - Test conditions and parameters
- AFIS testing - relevant measurements
  - FMR [FAR]
  - FNMR [1 – TAR]
  - FMR, FNMR definable in verification mode [see next slide]
  - Confidence intervals for FMR, FNMR
  - Performance, in this context, is FMR & FNMR
  - Speed [or throughput] is operationally important
  - Speed measured, but not included in performance

Contexts of image-based biometric ground truth: NIST's test environment

- FMR: p(A(i), B(j)) {i ≠ j} => M
  - probability that subject A(i), when tested against identity B(j), will be incorrectly reported as a 'match'
  - **not** the same as the probability that subject A(i) will be reported as a match against either B(j) or B(k) or B(n) or ...
  - equivalent to **some** definitions of FAR
- FNMR: p(A(i), B(j)) {i = j} => NM
  - probability that subject A(i), when tested against identity B(j), will be incorrectly reported as a 'non-match'
  - equivalent to **most** common definitions of [1 – TAR]
- FMR, FNMR require **knowledge of identity**

Contexts of image-based biometric ground truth NIST's test environment

- CMF extract
  - 1.68M tenprint records, 1.68M **subjects**
  - FD-249 image data => 10 rolled, 4 flat, AFVs for rolled
  - Type-2 [bio/demographic] data largely censored
- IQMI [Image Quality Multiple Instance
  - 285K tenprint records, 51K subjects
  - 6 [generally], 5, or 4 records per **subject**
  - FD-249 image data => 10 rolled, 4 flat, AFVs for rolled
  - Some type-2 data consistently present

- NIST’s context
- CMF extract
  - Duplicated **identity** [consolidation]
  - Perfection not required
    - If we never see adverse effects of imperfections in our measurements, then the imperfections have caused no problem [no harm, no foul]
    - CMF extract mostly used to model operational matching; since it is a snapshot of part of the real CMF, perhaps it should replicate its warts

NIST's context

- IQMI
  - Duplicated **identity** [consolidation]
  - Accuracy of those **correlate** data elements [biographic/demographic] which we use
  - Differentiation by data types [some are clean, some not so clean]
  - Differentiation by individual records [ditto]
  - Perfection required
  - Perfection: perfect knowledge, not perfect data

NIST’s context

- Common problem: consolidation
- DB-specific problem [IQMI]: correlate data

Contexts of image-based biometric ground truth

- Identity
- Attributes of image
- Attributes of subject
- Test environment
- Operational scenario
  - "when were you born?"
    - Who is asking? what questions will be answered?
    - Maryland DNR [year in which you turned 65?]
    - Maryland DMV [what goes on operator's license?]
    - US TSA [are you are who your ID says you are?]
  - "have you ever been arrested?"
    - Legal question [rights, privileges] => no
    - Security investigation [candor, trust] => yes*

Topics

- Context
- **Determination**
- Limits
- Implications of limits

NIST's determination of GT [consolidation]:

- Match scores underlie all analyses on this system
- Each match score is independent of all others
- Scoring codes designed to **censor** [ignore] results from erroneous records
- Scoring codes read a list of subject IDs of interest: scores pertaining to other IDs are ignored
- Scoring codes read a list of identities [true mates]
- Problematic records can remain in repository without penalty

NIST’s strategy – consolidation:

- Maintain record [master list] of consolidations
- Apply transitivity to build equivalence classes:

  A=B & B=C => A=C, and thus

  {A,B,C} share the same identity
- Conduct ten-print match of all against all, turning off filtering to the extent that time permits
- Visually validate all unexpected results
  - **Unexpected** matches
  - **Unexpected** failures to match

NIST's strategy – consolidation:

- Build tools to facilitate visual validation of **unexpected** results
- Rank cases by rough cost-benefit criteria:
  - Extremely easy to decide [high-scoring 'non-mates', low-scoring 'mates']; sort low-to-high
  - Less easy to decide, but with relatively high probability of changing our equivalence classes [moderate-scoring 'non-mates']; sort high-to-low
  - Less easy to decide, and with relatively low probability of changing our equivalence classes [low-scoring 'non-mates']; sort high-to-low

NIST’s strategy – consolidation:

- Visual validation tool [triage]
- Reads next record number, tells analyst which finger-pairs are available [in both records]
  - Analyst responds with finger number
  - Tool presents finger images side-by-side
  - Analyst responds:
    - # [number of next finger-pair to review]
    - I [Ident]
    - N [Non-ident]
    - Q [Questionable – flag to review later]
    - X [eXit – time for a coffee break]
  - Tool keeps running log of results, marking Automatic


```
images for finger position          5  require display under control of the window manager
for probe     30007091  and gallery     30007431  choose finger (1-14) to examine / e[X]it / [S]kip
 / [I]dent / [N]on-ident / [Q]uestionable
```

Operational Ground-Truth IBPC 2012-03-08

```
==> results file <==
3000709l 3000743l 02000 N
3000743l 3000787l 03500 I
3000787l 3000984l 05000 Q
3000984l 3000709l 09000 A

==> score.file <==
3000709l 3000743l 02000
3000743l 3000787l 03500
3000787l 3000984l 05000
3000984l 3000709l 09000
```

NIST's strategy – consolidation:

- Learn from adjudicating cases:
  - Keep running tabs to establish high threshold beyond which no changes are expected
  - Keep running tabs to establish low threshold beyond which no changes are expected
- Apply different procedures as context requires
  - CMF extract could tolerate a few missed consolidations because anomalous results would be checked retrospectively [modest filtering allowed]
  - IQMI could tolerate no consolidation errors, but then again, it was only 1/6th the size [no filtering allowed]

NIST's strategy – consolidation:

- Process the no-brainers internally
- Leave everything else to FEs
- NIST provided complete package of score files, image records, and software to Fes
- Records entrusted to NIST without authority to delegate trust were processed on site
- Records coming from FBI were processed at NIST or at CJIS by contract FEs

NIST's strategy – biographic/demographic

- Exploration of temporal and geographic effects upon matchability
  - DAT [1.05] in this case, not useful
  - DOB [2.022] shouldn't conflict with DOA, DPR
  - DPR [2.038] what is really wanted
  - DOA [2.045] should agree with DPR
  - ORI [1.08] less specific than CRI
  - RES [2.041] might be useful; must parse
  - CRI [2.073] what is really wanted

NIST’s strategy – geographic data

- ORI
  - Related to creation of derivative record
  - Not useful
- RES
  - Not always present
  - Not always credible
  - Not easy to parse
  - Not useful
- CRI
  - Not always credible
  - Not useful

NIST’s strategy – temporal data

- DAT
  - Referred to date of derivative record [c.f. ORI]
- DOB
  - Useful for corroboration
- DPR
  - Desired data
- DOA
  - Useful for corroboration

NIST's processes – temporal data

- Convert all dates to days since 1900-01-01 [there were no dates prior to 1900]
- Ignore DAT [contained nothing of value]
- Compute days from DOB to DOA
  - Flag unreasonably low age at time of arrest
- Compute days from DOA to DPR
  - Flag negative interval [DPR **before** DOA]
  - Flag lengthy interval [a week is reasonable; three months is questionable]
- Modify criteria as experience with data increases

NIST’s processes – temporal data [continued]

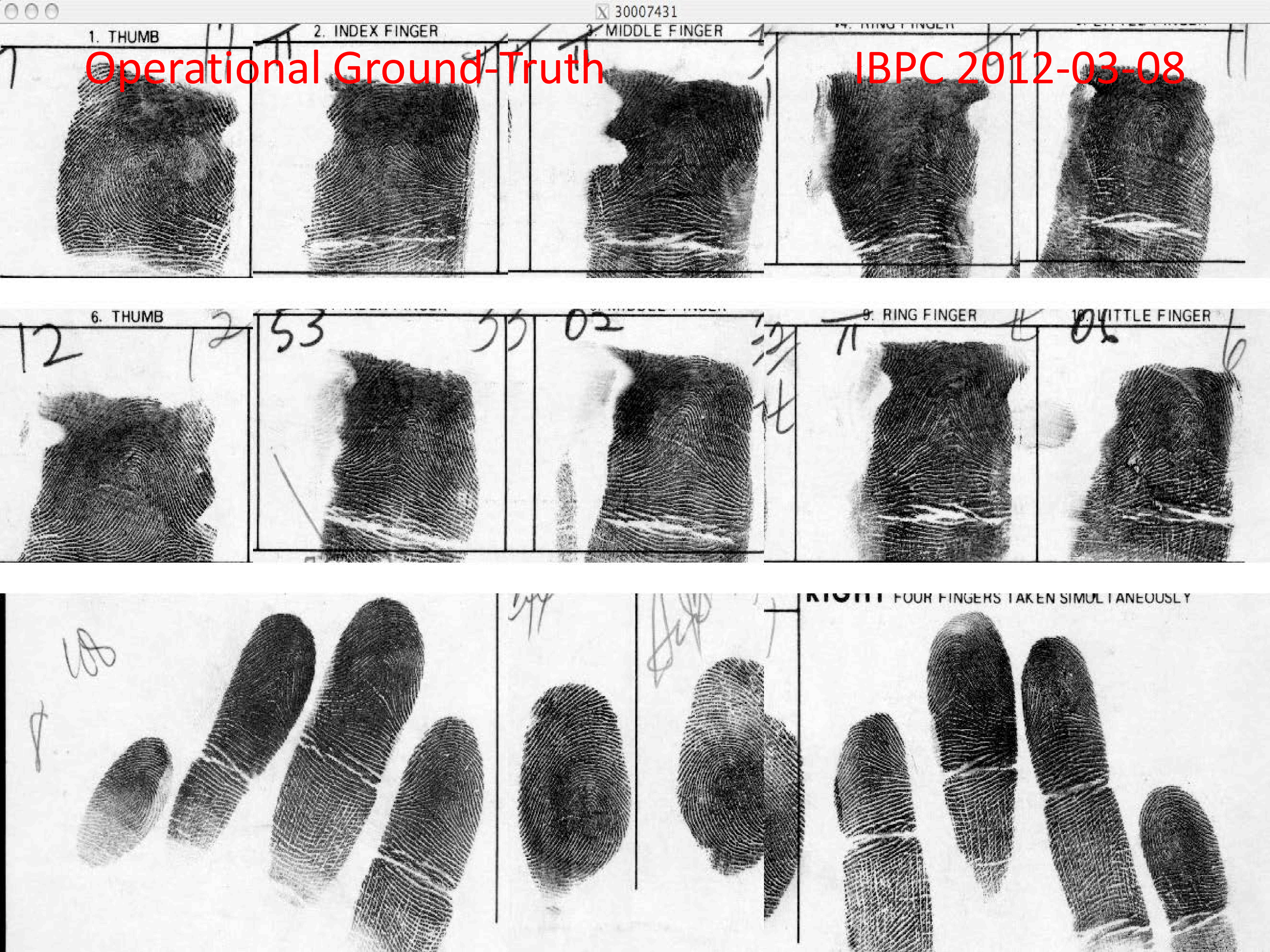
- Examine each date field [original and elapsed] collectively:
  - Sort
  - Count
- Find **sensible** explanation for anomalies
  - Cluster of dates on 1900-01-01
    - an EDP default beginning date
  - Cluster of dates on 1970-01-01
    - a mini-computer & UNIX default beginning date
  - Assume many/most errors have a reasonable basis
    - e.g., DOB used for DOA

NIST's processes – temporal data [continued]

- Develop a feel for what is probably right and what is probably questionable
  - DOA & DPR before 1970 almost surely wrong
  - DOA & DPR after 1995 raises no flags
  - DOA & DPR before 1988 presumptively wrong, but accepted if there was corroboration
  - DOA & DPR on or after 1988 presumptively correct, but record inspected for anomalies
- Reduce the questionable cases to a manageable amount and manually inspect
- Developed tool to reconstruct virtual FD-249

Operational Ground-Truth

IBPC 2012-03-08



Lessons learned:

- Immediately run internal consistency check
  - Record contents into database: finger images, other images, type-2 fields
  - Simple, automated tools [sort, count, sequence check]
  - Manually inspect records
- Immediately perform rapid consolidation check using normal operating mode [i.e., with filtering]
- During downtime, perform thorough consolidation check [i.e., without filtering]
- Use anomalies to trigger closer inspection of data
- Look for patterns in anomalies

More lessons learned:

- Trust data essential to the business process of the entity creating or recording it, but distrust data not essential: for example, trust 01-10, but not 11-14
- One knows more about one's own sampling from a database than about another's extraction process used to create that database
  - Randomness and bias of former easy to assess
  - Randomness and bias of latter difficult to assess

What we achieved

- Large operational database[s] useful for measuring extremely low FMRs
- Ability to correlate matchability with temporal data, with a high degree of confidence
- Techniques to correlate matchability with intrinsic and derived image data, but **not** biographical data, with a high degree of confidence [IAI-IEC 2010 presentation]
- Methodology for replicating this work with other large sets of biometric data

Topics

- Context
- Determination
- **Limits**
- Implications of limits

- NIST's observations – consolidation:
  - There was exactly one consolidation of subject IDs within the 50,855 subjects in IQMI [0.00002]
  - There were a non-negligible [i.e., > 3K] number of consolidations within the 1.68M subjects in the CMF extract [~0.002]
  - There were a significant number of consolidations among AZ, LAC, TXDPS, and CMF extract [~0.01]

- NIST's observations – non-identity:
  - **Systemic** image errors [~0.1] in one DB
    - Differing tenprint card formats
    - Scan coordinates for format A, cards in format B
  - **Systemic** metadata errors [0.1 to 1.0] in some DBs
    - Censoring
    - IT system [e.g., default dates]
    - Individual enroller quirks [e.g., DOB used for DOA]
  - **Non-systemic** metadata errors difficult to quantify [~0.001 to ~0.1]
    - Enrollee-induced error
    - Enroller error

Topics

- Context
- Determination
- Limits
- **Implications of limits**

Implications of limits

- On FMR
- On FNMR
- On correlation of bio/demographic data & match score

Implications of limits of GT on FMR:

FMR = probability that a decision D that would correctly have been classified D_{NM} will instead be classified D_M; call such a decision D_{XM}

$|D| = |P| * |G|$, or

number of decisions = [size of probe] * [size of gallery]

$|D_M|$ = Summation over p in P of $|g(p)|$:

$$|D_M| = \sum_{p \in P} |g(p)|$$

$|D_M| = |P| *$ **mntm** [mean number true mates]

$|D_{NM}|=|D|-|D_M|=|P|*|G|-|P|*$ **mntm** $=|P|*[|G|-$ **mntm**]

thus: limit {as **mntm** / |G| approaches 0} $(|D_{NM}|) = |P|*|G|$

$FMR = |D_{XM}| / |D_{NM}| \cong |D_{XM}| / |G| * |P|$

Implications of limits of GT on FMR:

For large operational databases, the increasing the number of true mates will have negligible impact on FMR

However, increasing the number of unreported true mates can cause a dramatic increase in the reported FMR, because with a good matcher, almost every unreported true mate of the probe set will result in an **apparent** false match

Such **apparent** false matches can easily dominate the FMR

Implications of limits

Postulate a gallery of 2M whose consolidation has been effected by matcher whose FNMR is 0.002 and whose real FMR is 0.000001, tested by a probe set of 1M [and an orthogonality factor of 90%]; also assume that 1% of subjects in gallery had falsely identified themselves

There would have been 20K claims of non-identity, of which all but 40 would have been detected; of these 40 undetected consolidations, half would not be in play; of the remaining 20, 90% would remain unmatched [no harm, no foul] when probed with a new image from the same subject, but 10% [or 2 subjects] would be apparent false matches, elevating the apparent FMR 3-fold, from 0.000001 to 0.000003

Implications of limits of GT on FNMR:

FNMR = probability that a decision D that would correctly have been classified D_M will instead be classified D_{NM}; call such a decision D_{XNM}

$|D_M|$ = Summation over p in P of |g(p)|:

$$|D_M| = \sum_{p \in P} |g(p)|$$

$|D_M| = |P|$ * **mntm** [mean number true mates]

FNMR = $|D_{XNM}| / |D_M| = |D_{XNM}| / |P|$ * **mntm**

Note that gallery size |G| is not relevant

Implications of limits

Postulate a gallery of 2M whose consolidation has been effected by matcher whose FNMR is 0.002 and whose real FMR is 0.000001, tested by a probe set of 1M, each with one mate in the gallery [**mntm** = 1.0]; also assume that 1% of subjects in gallery had falsely identified themselves

The effect on measured FNMR is undetectable: in this case there would have been 2*10^12 decisions, of 1*10^6 nominally should have been match decisions; however, we expect about 2*10^3 failures, and in fact observe 2*10^3 failures; any matches [or failures to match] with undetected duplicates will not be noted

Implications of limits of GT on correlation of match score with bio/demographic data

Observation: everything in the **real** [vs **ideal**] world is random [non-deterministic]

Question: “how random?”

- Deceit by subject
- Systemic error
- Memory error
- Transcription error [noise]
- Systematic extraction

Implications of limits of GT on correlation of match score with bio/demographic data

- Deceit by subject
  - Identity [name, SSN, military ID #]
  - Attributes [age, DOB]
- Systemic error
  - Overlaying data
  - Swapping data
- Memory error
  - Enrollee's memory
  - Enroller's memory
- Transcription error [noise]
  - Typos
- Systematic extraction
  - Every 10^{th} record vs every 7^{th} day vs. every nnn01 zip code

Implications of limits

Aside from temporal data, identifying GT too difficult to permit much analysis: certainty, or even quantification of uncertainty, was lacking; when looking for subtle effects, one must be able to trust one's data

This does not apply to the images themselves; claims of height and width can be tested, although in reality we ignored the claims and measured the images directly

- Contact information:

Stephen S Wood

National Institute of Standards and Technology

100 Bureau Drive, Mail Stop 8940

Gaithersburg, Maryland 20899-8940

301-975-4722

swood@nist.gov