Only One Percent Rubbish
A Lesson About Data

Roy Maxion

Dependable Systems Laboratory
Computer Science / Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
Email: maxion@cs.cmu.edu

26 June 2021
IFIP 10.4
Virtual Summer Meeting

Please do not distribute.

First ... thank you to ...

- IFIP Working Group 10.4
  - It's an honor to be here.
- Our sponsors ...
  - National Science Foundation, CERT
- Our teammates ...
  - Vrishab Commuri
  - Harvey Vrsalovic
  - Wangzi He
- Our collaborators ...
  - David Banks, Patricia Loring, and 200+ study participants

 2

This talk is about ...

- This talk is about the sensitivity of ML systems to small irregularities in data.
- We've all heard about that:
  - Autonomous vehicles
  - Face recognition
  - Cat vs dog recognition
  - Bail decisions
  - Parole decisions
  - Loan decisions
  - Etc. – a long list

 3

Irregularities: biased data vs bad data

- Many ML problems are rooted in biased data.
- However, little to no attention has been given to unbiased, but noisy or corrupt, data.
- Nor has a light been shown on how a wee corruption can have dramatic effects on decision-system outcomes.
- Neither has it been shown how to detect such corruption.

 4

In this talk I will ...

- Show how data corruption was discovered in a decision system based on behavioral biometrics (two-factor authentication).
- Show how the cause of the corruption was ascertained and replicated.
- Show the decision-altering effects of only 1% rubbish data in a keystroke-based behavioral-biometric system.

 5

What is keystroke biometrics/dynamics?

- Keystroke dynamics is the process of identifying individual users on the basis of their typing rhythms.
- It's a behavioral biometric ...
  - It's <u>how</u> you do something – your habits; not something you know (e.g., a secret) or something you have (e.g., a fingerprint).
  - Similar to gait – the idiosyncratic way you walk.

 6

How does keystroke dynamics work?

- It is based on the timestamps of key-press and key-release events in the keyboard.
- Basic measures:
  - Key-hold time (average ~ 92 milliseconds)
  - Interkey transition time (d-d average ~ 221 msec)
    - How fast is that? … faster than an eye blink (250-300 ms)
- Everyone is different.
- Users are differentiated from each other on the basis of (dis)similarity in the data -- usually with a machine-learning classifier, such as a Random Forest or a Neural Network.
- No specialized equipment is needed; just a keyboard.

Keystroke biometrics - uses

- Two-factor authentication
  - (1) the password; (2) how you type it
  - Continuous authentication
- Detection
  - Insider threat
  - Deception
  - Neurological conditions
  - Cognitive decline
  - Dementia
  - Stress
  - Emotion
  - Questioned documents

Keystroke dynamics: seriously??

- 21 June 2019
  - European Banking Authority (EBA-Op-2019-06)
  - Approved keystroke dynamics as a method of strong customer authentication.
- 12 May 2021
  - Presidential executive order
  - … agencies shall adopt multi-factor authentication … for data at rest and in transit.

Origin of our undertaking

- We collected two data sets in exactly the same way:
  - Lab – tightly controlled, all participants used the same equipment (e.g., computer, screen, keyboard, mouse, monitoring software, etc.).
  - Field – uncontrolled, participants used whatever equipment they had; monitoring software was the same.

Sketch of experiment

- Conditions
  - Lab: controlled; same apparatus for all participants
  - Field: not controlled; whatever equipment they had
- Participants
  - 100 in each condition
  - Gender and handedness representative of population
- Task
  - Type .tie5Roanl (strong password) 50 times in each of 8 sessions
  - 400 repetitions total
- Data
  - Key-press time
  - Key-release time
  - Key name (e.g., a, b, c, …)
  - 2,480,000 data points
    - 31 features * 200 participants * 400 repetitions
    - 11 * 200 * 400 = 88,000 keystrokes

Research results

- Is there any difference between the lab and field data?
  - Yes
  - Field data contain artifacts that are not seen in lab data
- And if so, does it matter?
  - Yes
  - Decision outcomes can change by nearly 20 percentage points … when only 1-2 percent of the data contains field-type artifacts

Sketch: Our journey of discovery

- Descriptive statistics; n-number summaries
  - Global; across all 100 subjects
  - At the subject level; individual subjects
- Plot the data; histograms
  - Lots of differences; find their origin
- Develop a frequency table reflecting the histograms
- Plot the data; scattergrams
  - Examples from lab and field
- Ask: where is the [experimental] variation?
  - In the apparatus, mainly
- Ascertain and verify the cause
  - Graphs verify the claim

Descriptive statistics

- Start with descriptive statistics
  - Hold times
    - 11 features
    - 50 repetitions
    - 8 sessions
    - 100 subjects
    - 440,000 data points … for each data set
  - Hold times carry most of the information
  - They’re least under the typist’s conscious control

Descriptive stats, n-number summaries

	Lab	Field
Mean	**92.289**	**99.985**
Std	**30.436**	**42.842**
Median	**88.4**	**95.9**
Mode	**80.5**	**80.0**
Kurtosis	**45.388**	**8000.751**
Skewness	**1.693**	**51.278**
Range	**2033.9**	**8687.9**
Max	**2035.3**	**8688.2**
Min	**1.4**	**0.3**
IQR	**38.6**	**37.9**
Upper Quartile	**109.3**	**116.9**
Lower Quartile	**70.7**	**79.0**

!

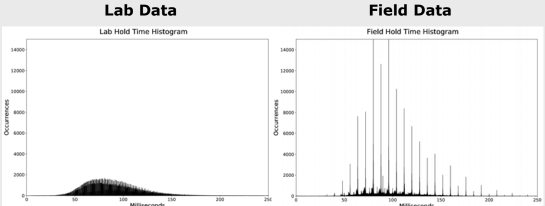
All hold features (h, dd, ud, including <return>) 100 subjects in each of lab and field.
Units: milliseconds

N-number summaries

- Most numbers … not astonishingly different, but
  - Significant difference between the means
  - Two-tailed t-test: (p << .001)
- Kurtoses were quite different, suggesting …
  - The shapes of the distributions must be very different
    - Leads one to wonder *how* they are different
  - Lots of outliers
- So, we plotted some histograms …
  - … and said, “Huh?”

Histograms, lab and field

Lab Data **Field Data**



All hold times for all subjects; 440,000 data points.

100 subjects * 8 sessions * 50 repetitions * 11 keys

Note the spikes in the field data.
Outliers are far to the right, off the graph.

Histograms, hold times only

- These plots lead one to believe that some times are much more prevalent than others
- But why? All field times should be roughly equally distributed … as in the lab data.
  - Should not have giant spikes.
- So, create a frequency table; explore discrepancies between lab/field frequencies

Hold-time frequency counts, L100/F100

Sorted, top 10. Note huge discrepancy in L/F counts.

Lab Hold time	Lab Count	Field Hold time	Field Count
80.5	1656	80.0	15250
81.3	1654	96.0	15094
78.4	1653	88.0	12630
77.6	1627	104.0	10239
75.5	1613	112.0	8371
71.8	1608	72.0	8061
84.2	1600	64.0	7628
76.0	1592	120.0	6669
76.8	1587	128.0	5216
85.5	1585	79.9	4949

Observations on the frequency table

- Hold-time frequencies, sorted, top 10
- Observation: multiples of 8
- Note that a subject-by-subject frequency table would reveal that there are different frequencies, depending on the subject.
- This leads down the path that something unexpected is happening.

Hold-time frequency counts, by subject

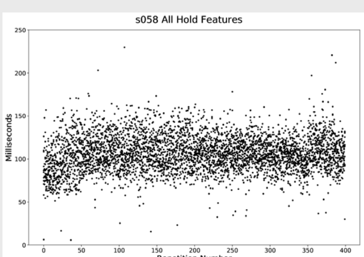
Far more multiples of 8 in field data than in lab data.

s002 (lab) Hold time	s002 (lab) Count	p1156 (field) Hold time	p1156 (field) Count
82.1	33	128.0	276
81.3	31	112.0	272
81.0	31	96.0	208
78.1	30	144.0	178
80.0	30	111.9	172
81.6	29	127.9	164
96.4	29	112.1	143
82.4	29	128.1	131
84.5	28	95.9	127
79.7	28	96.1	103

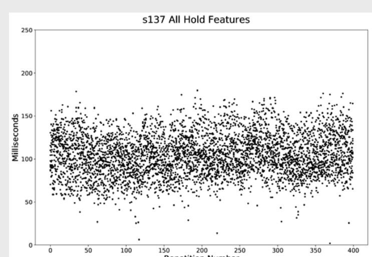
Turn to time-tested approach

- Plot the data … scattergrams to start with
- Several examples of lab data
- Several examples of field data

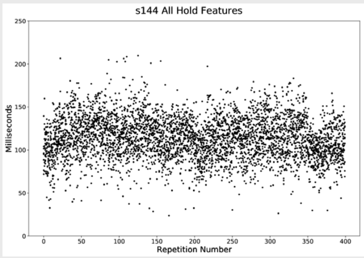
Lab data, typical: s058



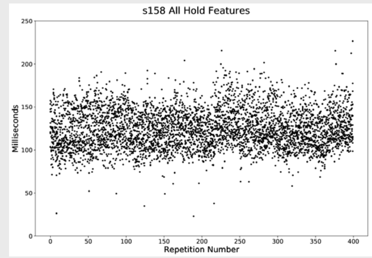
Lab data, typical: s137



Lab data, typical: s144



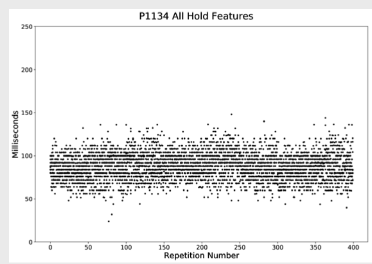
Lab data, typical: s158



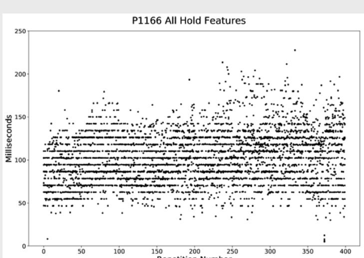
Turn to time-tested approach

- Plot the data … scattergrams to start with
- Several examples of lab data
- Several examples of field data

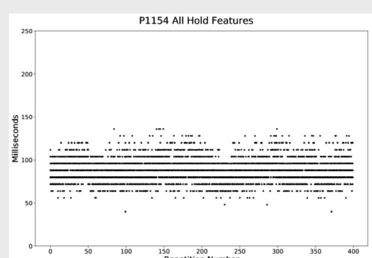
Field data, 4 ms quantization: 1134



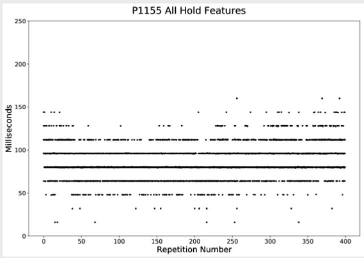
Field data, 8 ms quantization: 1166



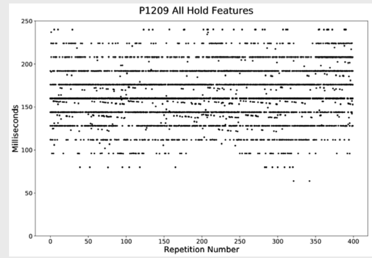
Field data, 8 ms quantization: 1154



Field data, 16 ms quantization: 1155



Field data, 16 ms quantization: 1209



Observation

- If you are going to use hold times in your ML classifier, the striated data we've just seen certainly seems as if it would complicate things.

  (And it does.)

Observation

- The main difference between lab and field conditions was apparatus.
- There were only two variations of apparatus
  - Machines: PCs with Windows OSs of various stripes
  - Keyboards: PS/2, USB, custom
- Examine keyboards and timing protocols

Custom (lab)

- Keyboard – Apple M9034LL/A
- Bypassed keyboard encoder
- Timestamps were captured at the keyboard, not the host
- Keystroke timing resolution, calibrated at 100 microseconds

USB Polling (field)

- We looked at the USB spec (not easy)
- Polling intervals appeared to be powers of 2
  - (But not always; the spec is unclear and often not well implemented.)
- Polling intervals can differ amongst keyboards
- Striations may be artifacts of USB polling
- Hypothesis: USB was involved

The OS talks to a new USB device*

OS: Hello, Stranger. What kind of device are you?

KB: I'm a slow Human Interface Device (HID).

OS: Ok. How often do you want me to request data from you?

KB: *At least* every 16 ms.

OS: Ok. And how much data will you have for me at each 16 ms request/poll?

KB: 12 bytes (with n-key rollover)

OS: Ok, we're set to go. Bus is scheduled; polling starts now.

- *Caricature: most info is transmitted via the keyboard USB endpoint descriptor.
- It's a conversation about allocating resources on the bus.

PS/2 vs USB

- PS/2 is much faster.
- The PS/2 controller can generate an interrupt as soon as the keyboard has clocked in the 11 (8 data + 3 frame) bits.
- The USB controller will send interrupts at a maximum of 1 every ~8 ms (+/- some jitter) or 16 ms , or 32 ms , depending on the bInterval polling value.

Measuring polling intervals in field data

- Built a polling discovery tool comprising …
  - Dictionary: a look-up test of polling rate
  - Spike: inter-spike latencies
  - DBSCAN: Density-based spatial clustering of applications with noise
  - Visual-line (tool of last resort)

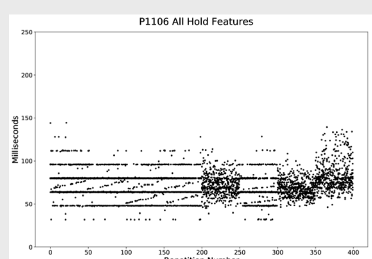
Polling rates (millisec) 100 field subjs

Polling Rate	Count	
2.0	1	
2.5	1	…….. Some were fractional; need USB protocol analyzer.
2.7	1	
2.75	1	
2.8	2	Possibly low-quality keyboards.
3.5	1	
4.0	7	
5.0	2	
5.2	2	
5.3	2	
5.35	2	
5.5	1	
7.8	1	
8.0	54	most prevalent
10.0	3	
16.0	7	
22.0	1	
PS/2	11	PS/2
	100	Total

Things can get even more strange

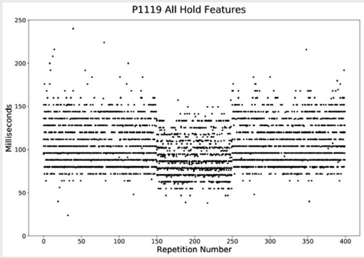
- So … USB polling explains the striations in the data.
- What you've seen so far has been more or less well-behaved field data.
- Not everything was that "good".
- Here are some examples of weird data from the field – things that you'd never expect.

1106 changed keyboard



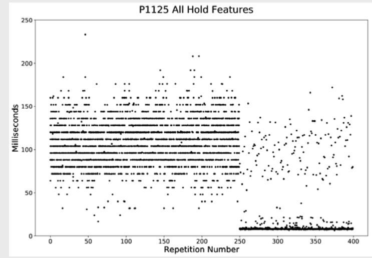
Changed keyboard; wireless at home, internal PS/2 on the road.

1119 changed keyboard



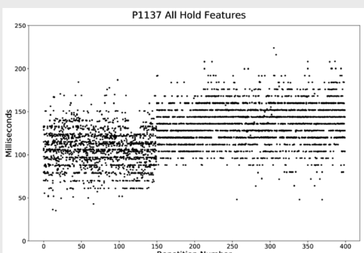
Changed keyboard, office vs home

1125 keyboard macro?



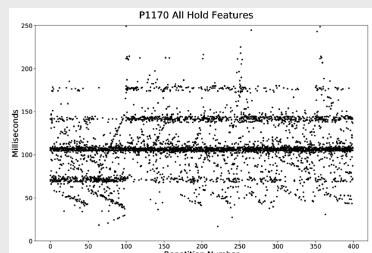
Extremely short times. Keyboard macro?

1137 mean shifted



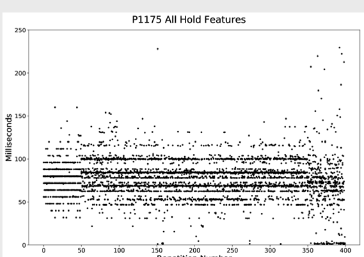
Mean shifted/jitter; reason unknown

1170 ramps, 37 ms polling



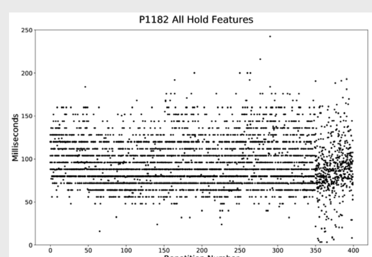
Ramps, 37 ms polling, unexplained general strangeness

1175 changed polling; bursts



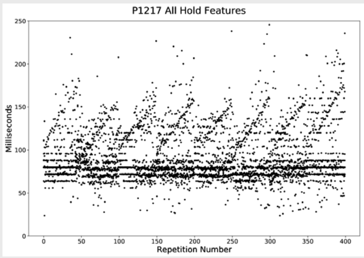
Changed polling; bursts in last session

1182 changed keyboard, remote login



Changed keyboard to PS/2; same machine, remote login

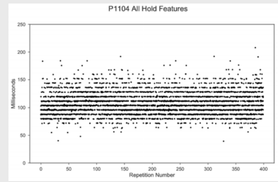
1217 unexplained ramps



Unexplained ramps; insufficient memory/resources?

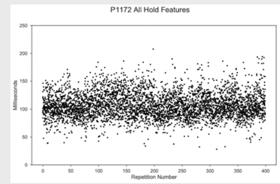
 49

Field data, compare: 1104 vs 1172



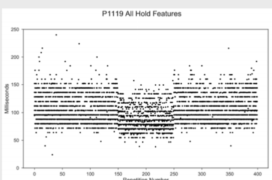
Left: Personal workstation
Right: Dell laptop

Same typist, 2 keyboards



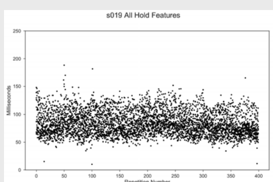
 50

Field data vs lab data; same person



Field 1119

Same typist; field/lab



Lab 019

 51

These were dropped from the study

- 1120: egregious quantization (37ms) and jitter
- 1125: sessions 6-8 very short hold times akin to bursts, except for return key; 1193 bursts – the most bursts of any subject
- 1170: large polling interval (34ms), extreme jitter, ramps
- 1172: duplicate subject (1104)
- 1175: too many changes across sessions, possible keyboard switch, high jitter, last session contained 85 bursts

 52

Other people's data; or was it just us?

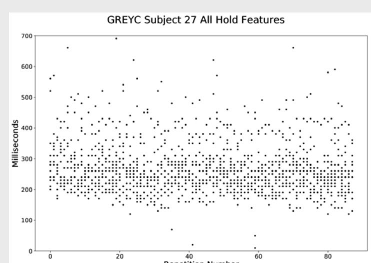
- We examined 10 other well-known data sets.
- Found the same quantization phenomena.

1. A review on the public benchmark databases for static keystroke dynamics (2015). Giot, Dorizzi & Rosenberger. Computers & Security, 55, pp 46–61.
   - **GREYC, WEBGRAYC, BIOCHAVES, KEYSTROKE100, GREYC-NISLAB, CMU**
2. Observations on typing from 136 million keystrokes (2018). Dhakal, Feit, Kristensson & Oulasvirta. CHI proceedings.
3. Coursera
4. Study on the BeiHang keystroke dynamics database (2011). Li, Zhang, Cao, Zhao, Gao & Liu. International Joint Conference on Biometrics, pp 1-5.
5. User authentication through keystroke dynamics (2002). Bergadano, Gunetti & Picardi. ACM Trans. Inf. Syst. Secur. Vol 5, No 4, November, pp 367-397.

- Our own data will be made available.

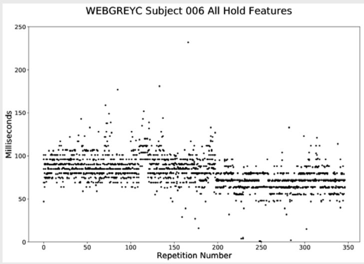
 53

GREYC: subject 27, all hold features



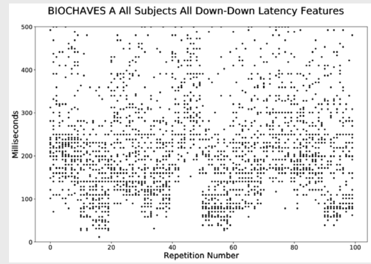
 54

Web GREYC: Subj 006 all hold features



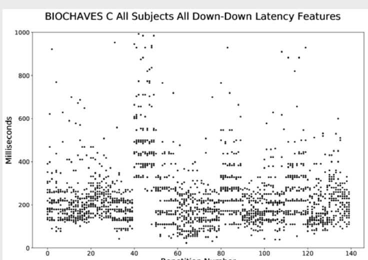
 55

BioChaves-A: all subjs, all dd latency features



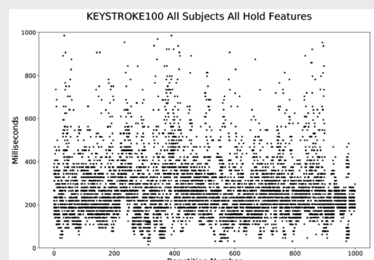
 56

BioChaves-C: all subjs, all dd latency features



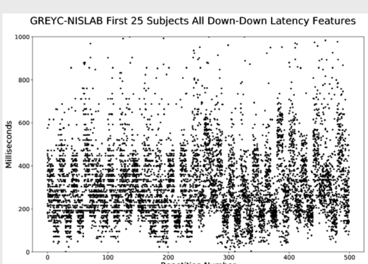
 57

KEYSTROKE100: all subjs, all hold features



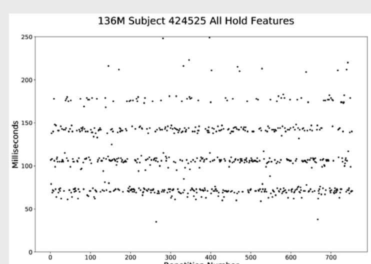
 58

GREYC-NISLAB: 1st 25 subjs, all dd features



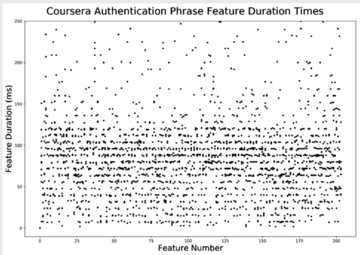
 59

136m, subject 424525, all hold features



 60

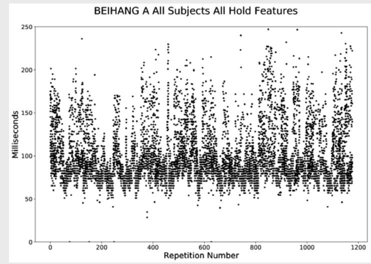
Coursera: holds & ud latency features



One subject: P-OD

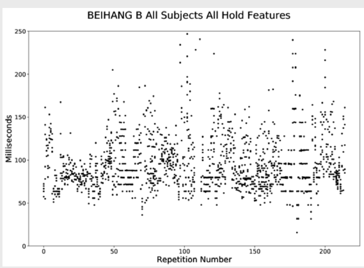
 61

Beihang-A: all subjs, all hold features



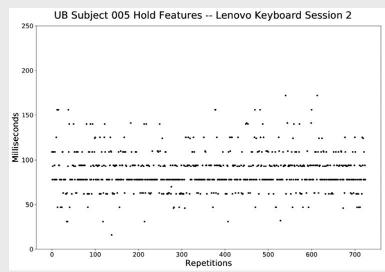
 62

Beihang-B: all subjs, all hold features



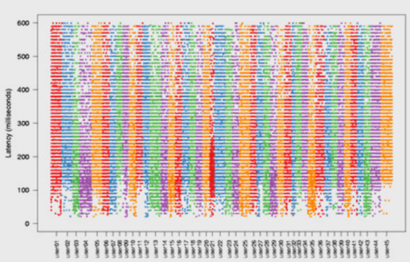
 63

Univ of Buffalo



 64

Bergadano, Gunetti & Picardi - 2002



 65

Conclusion regarding field data

- USB keyboards are injecting polling artifacts into the typing data
- It wasn't just our data
  - Many known data sets share USB characteristics
- Next question: does it matter?
  - Classification / machine-learning experiments

 66

Classification: the rubber meets the road

- Are classifier results (random forest) different when using quantized data?
- Four experiments: lab vs field-similar
  1. Quantize the lab data so that it looks like field data
     - 8 ms and 16 ms
  2. Quantize just one subject (one-shot)
     - 8 ms and 16 ms
  3. Quantize two subjects: 16 ms: s029 and s077
  4. Quantize in proportion to field-data polling rates
- What to look for …
  - Overall classification accuracy
  - Changes in the misclassification matrix

Classifier regime

- Random forest w/ seeds held constant
  - No nondeterminism; fully repeatable
- Training
  - 25 repetitions (out of 50) from each session drawn at random: training set contains 200 repetitions per subject; 20,000 vectors for entire data set
- Testing
  - Use the remaining repetitions not used in training: 200 reps/subject; 20,000 total
- Control: repeat random draw 5 times
  - Maximum accuracy variation due to random draws: .005%

Quantization method

- Rule: Round half to even

  (1) Divide observed keystroke feature duration by desired polling rate

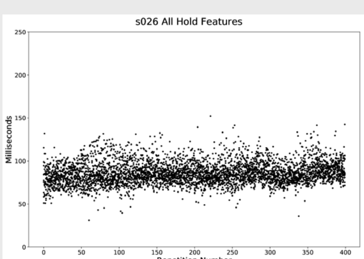
  (2) Round to nearest integer
  - If rounding is ambiguous (e.g., xxx.5) then round to nearest even integer

  - Quantized value = rounded value * desired polling rate

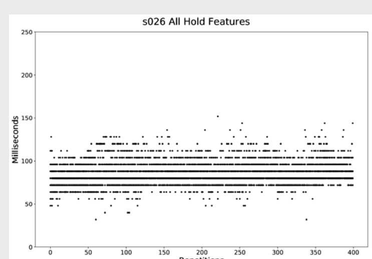
Examples of quantized lab data

- Original lab data
- Quantized to 8 ms
- Quantized to 16 ms

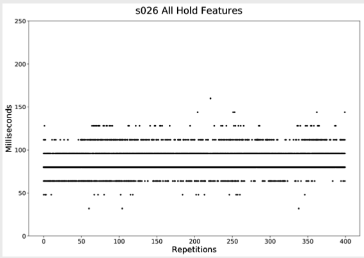
s026 original lab data



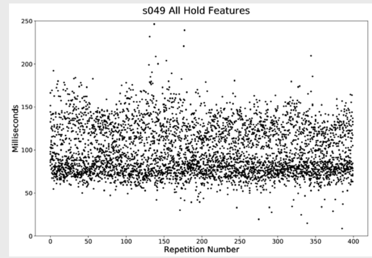
s026 quantized lab data, 8 ms



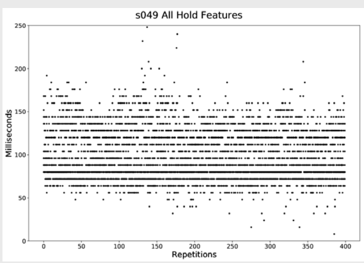
s026 quantized lab data, 16 ms



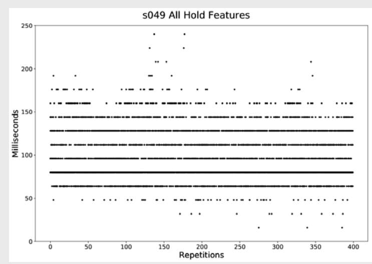
s049 original lab data



s049 quantized lab data, 8 ms



s049 quantized lab data, 16 ms



General classification

- Overall classification accuracy
  - Lab: 90.55%
  - Field: 92.56%
- These cannot be compared, because sample frames and subject pools were dissimilar.
  - When the subjects are different, and they all type differently, producing different data sets, they cannot be compared.
  - This is one reason why many keystroke biometric results cannot be taken seriously.
- So we compare lab data with quantized lab data (i.e., field-similar data).

Lab, no quantization, RF, h, dd, ud

	s077	**s078**	**s079**	**s080**	**s081**
s077	0.825	0.015	0.000	0.000	0.025
s078	0.000	0.835	0.000	0.010	0.005
s079	0.000	0.000	0.785	0.000	0.000
s080	0.000	0.000	0.000	0.945	0.000
s081	0.000	0.000	0.000	0.005	0.885

Misclassification matrix excerpt.
Accuracy: 90.55%

All subjs 8ms quantization, RF, h, dd, ud

	s077	s078	s079	s080	s081
s077	0.860	0.000	0.000	0.005	0.030
s078	0.000	0.840	0.000	0.005	0.010
s079	0.000	0.000	0.830	0.000	0.000
s080	0.000	0.000	0.000	0.955	0.000
s081	0.000	0.000	0.000	0.000	0.905

s079: shift from .785 to .830.
4.5 percentage point difference.
1073 cells changed value.
Accuracy: 90.37%

 79

Lab, no quantization, RF, h, dd, ud

	s029	s030	s032	s034	s035
s029	0.805	0	0	0	0
s030	0	0.975	0	0	0
s032	0	0.005	0.62	0.005	0
s034	0	0	0	0.825	0
s035	0	0	0	0	0.895

Misclassification matrix excerpt.
Accuracy: 90.55%

 80

s032, one-shot, 16 ms quantization

	s029	s030	s032	s034	s035
s029	0.805	0	0	0	0
s030	0	0.965	0	0	0
s032	0	0.01	0.78	0.01	0
s034	0	0	0	0.81	0
s035	0	0	0	0	0.895

s032: shift from .62 to .78.
16 percentage point difference.
1030 cells changed value.
Accuracy: 90.85%

 81

Classification: results

Quantize all lab data into field-similar data, n ms

Quant level	Classification accuracy	Cells changed	Most egregious	Diagonal Change	Percentage Points
All data					
0	90.55	--	--	---	---
8	90.37	1073	s079	.785/.830	4.5
16	89.46	1190	s034	.825/.765	6
One-shot					
8	90.69	1038	s079	.785/.875	9
16	90.85	1030	s032	.620/.780	16
Two-shot (8ms: s007 & s111 / 16ms: s018 & s032)					
8	90.52	1038	s007	.755/.870	11.5
16	90.87	1058	s032	.620/.810	19
Proportional (averaged over 10 runs)					
Various	91.48	1084	Various	.802/.883	8.15

Quantizing just one subject changed the diagonal by 16 percentage points!
Quantizing two subjects … by 19 points!

 82

Other classifiers; it wasn't just ours

SVM

Unquantized lab data:	**74.06**
8 ms quantized lab data:	**73.97**
16 ms quantized lab data:	**73.69**

Manhattan KNN*

Unquantized lab data:	**77.86**
8 ms quantized lab data:	**77.62**
16 ms quantized lab data:	**77.53**

Random Forest

Unquantized lab data:	**90.55**
8 ms quantized lab data:	**90.37**
16 ms quantized lab data:	**89.45**

*** Used Manhattan distance**

- **Misclassification matrices had similar outcomes**

 83

Observations

- USB quantization changes outcomes.
- Quantizing just a single subject can induce changes in misclassification matrices.
- I.e., if just one subject out of 100 (1%) uses a keyboard that injects artifacts into the data, the misclassification matrix can change.

 84

What do we make of this?

- Context-sensitive: in some contexts these shifts can invert a verdict of guilty vs innocent.
- In other contexts, they can result in a user being incorrectly labeled as fraudulent instead of legitimate, or vice versa.
- If you are Google and are trying to classify someone as willing or not willing to click on an ad, a 1% change on the diagonal is a lot of money.
- The significance of the magnitude of the change depends upon the application, the context and the costs of decisions, risks and errors.

 85

The law looks at things differently

- Different types of crimes have different standards of evidence.
  - In tort law, more than 50% belief is needed to levy fines. Consider a matrix shift from .49 to .51.
  - In criminal cases, reasonable doubt can turn on 1% or 2% evidentiary change, but it is never defined … leaving things open to interpretation, which is sometimes not good.
- And a company can fire someone without any due process whatsoever; they can interpret classification outcomes in any way they wish.

 86

Now what?

- It would appear that future keystroke studies will be hampered (at best) by USB artifacts.
- These artifacts are difficult/impossible to control.
- Maybe this ends keystroke dynamics.
- Remedy: gaming keyboards for $150.
  - Ok for companies/governments, but not for most users.
  - Or, modify keyboard HID descriptors for 1 ms polling
- Future: uncertain

 87

Summarizing … what we have seen

- Two data sets; lab and field
  - Same task, slightly different apparatus
- Discovery
  - Field: USB keyboards injected timing artifacts
  - And it wasn't just in our data
- Impact
  - Consequential changes in misclassification matrices
- Upshot
  - Context-dependent; depends on cost of error
  - Significant doubt in credence of classification results
  - Could mean serious differences in adjudications
  - Uncertain future

 88

General conclusion

- Artifacts injected into data, whether through chance or through malice, and irrespective of their source (in our case, keyboards), can have unwanted and injurious effects on classification outcomes.
- Caution is warranted, as is careful screening of collected data.

 89