When Trusted Black Boxes Don't Agree: Incentivizing Iterative Improvement and Accountability in Critical Software Systems

Jeanna Neefe Matthews
Graham Northup
Isabella Grasso
Stephen Lorenz
Marzieh Babaeianjelodar
Hunter Bashaw
Sumona Mondal
Clarkson University

Abigail Matthews
University of Wisconsin-Madison

Mariama Njie
Iona College

Jessica Goldthwaite
The Legal Aid Society

AIES, February 7 2020

Decision-Making Processes

- Software increasingly plays a key role in big decisions
  - Regulated areas (housing, hiring, credit) and major public functions (criminal justice, elections)

- Fundamentally changing the landscape of our societal decision-making processes

- Flaws in software AND in the larger socio-technical decision-making systems in which software is developed, deployed and trusted

Criminal Justice System as a Decision-Making System to Secure?

- Decisions of how to deploy police resources, who should get probation vs. jail, forensic analysis of evidence

- Heavy use of software/algorithmic decision making throughout the system
  - Often black boxes for which trade secret protection is claimed to be more important than rights of individual defendants or citizens to understand the decisions
  - IP to reward good ideas vs IP to shield from knowledge of flaws

- Principles to ensure
  - Right to a public trial
  - Rights of defendants to review and confront the evidence against them
  - Better to let an guilty person go free than convict an innocent one?
  - Right to equal justice under the law

A little about Probabilistic Genotyping Software

- Matching evidence samples found at crime scenes to possible suspects
- DNA gold standard vs. Probabilistic
- Cannot manually verify answer
- Many programs that can do this, but little attempt to systematically compare them in case work
  - In fact many hurdles to doing so

Validation Studies

- Developers of the system typically do their own testing and publish results of a validation study
- Validation Studies
  - Unlike in real case work you know the answers (Contributors vs. Non-contributors)
  - But little attempt to match testing space covered by validation study to specific cases in court
  - Peer reviewed = adequately tested for all cases?
- They have a vested interest in demonstrating the system is working, not in finding bugs.

Do sufficient incentives exist for flaws in this software to be identified and fixed?

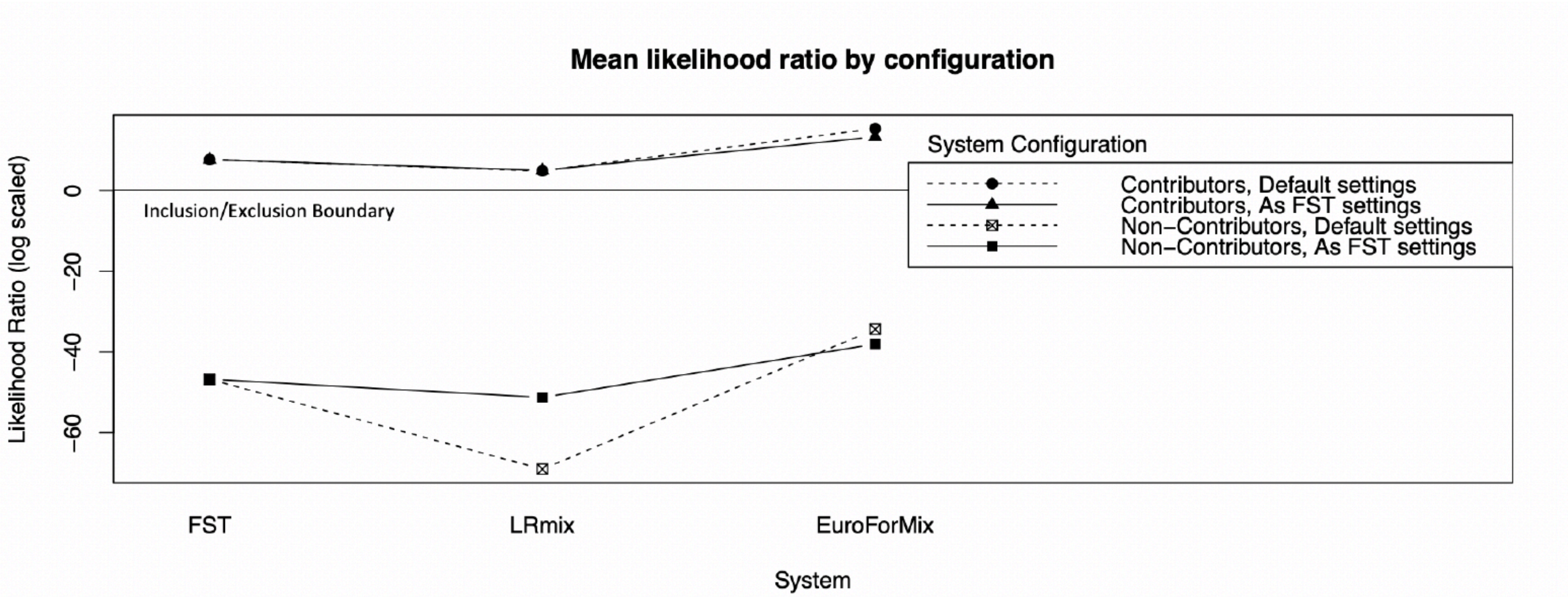
- We are accustomed to market forces incentivizing the costly process of debugging and iterative improvement
  - For many critical software systems, market forces may be utterly insufficient

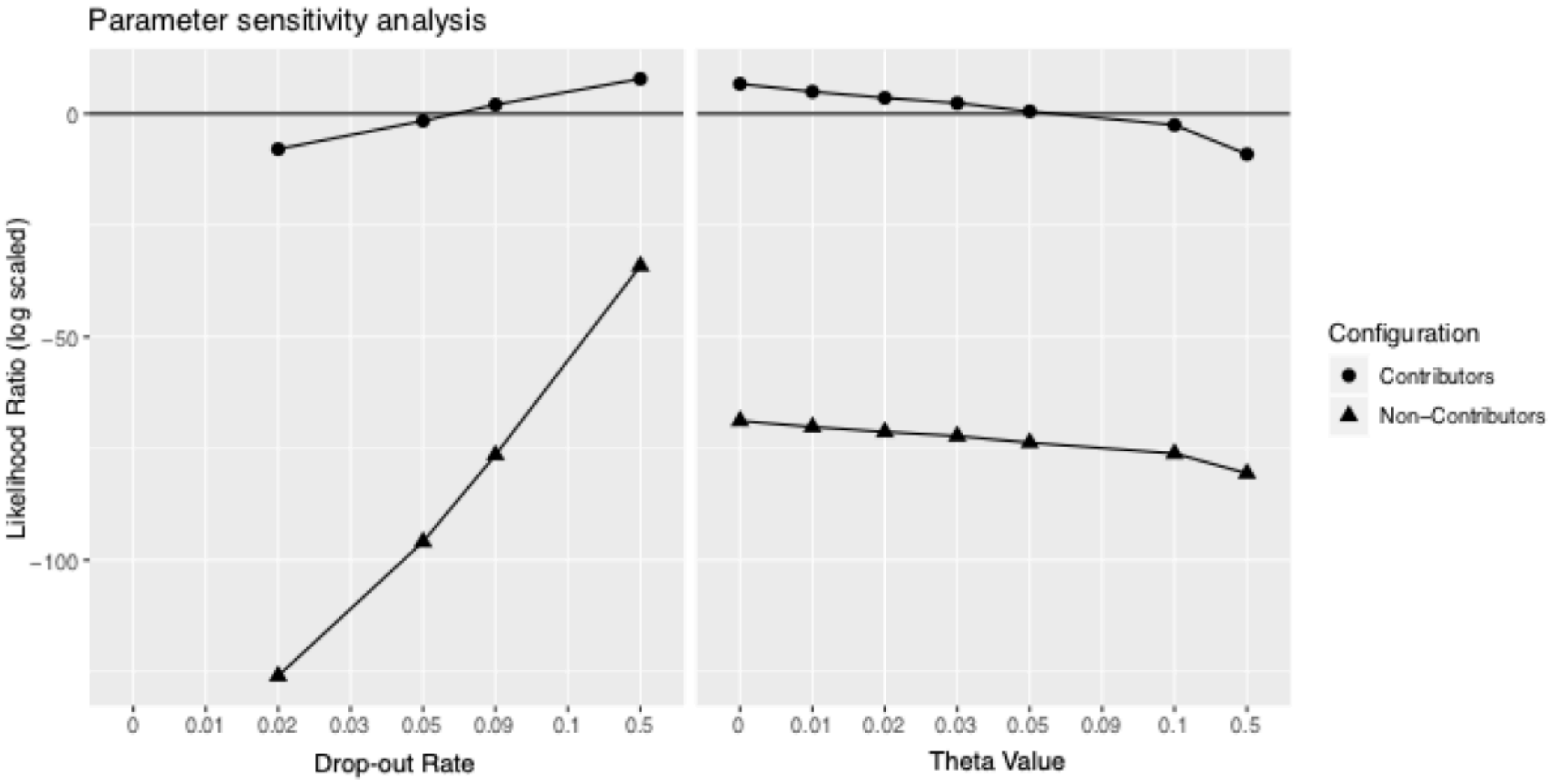
- Forensic software – trade secret protection, little oversight, expensive, hurdles in terms of service to testing, history of covering up bugs found post-deployment, inability to do manual checking, then…
  - You think the software is incorrect in your real case?
  - You are just complaining because you are guilty!

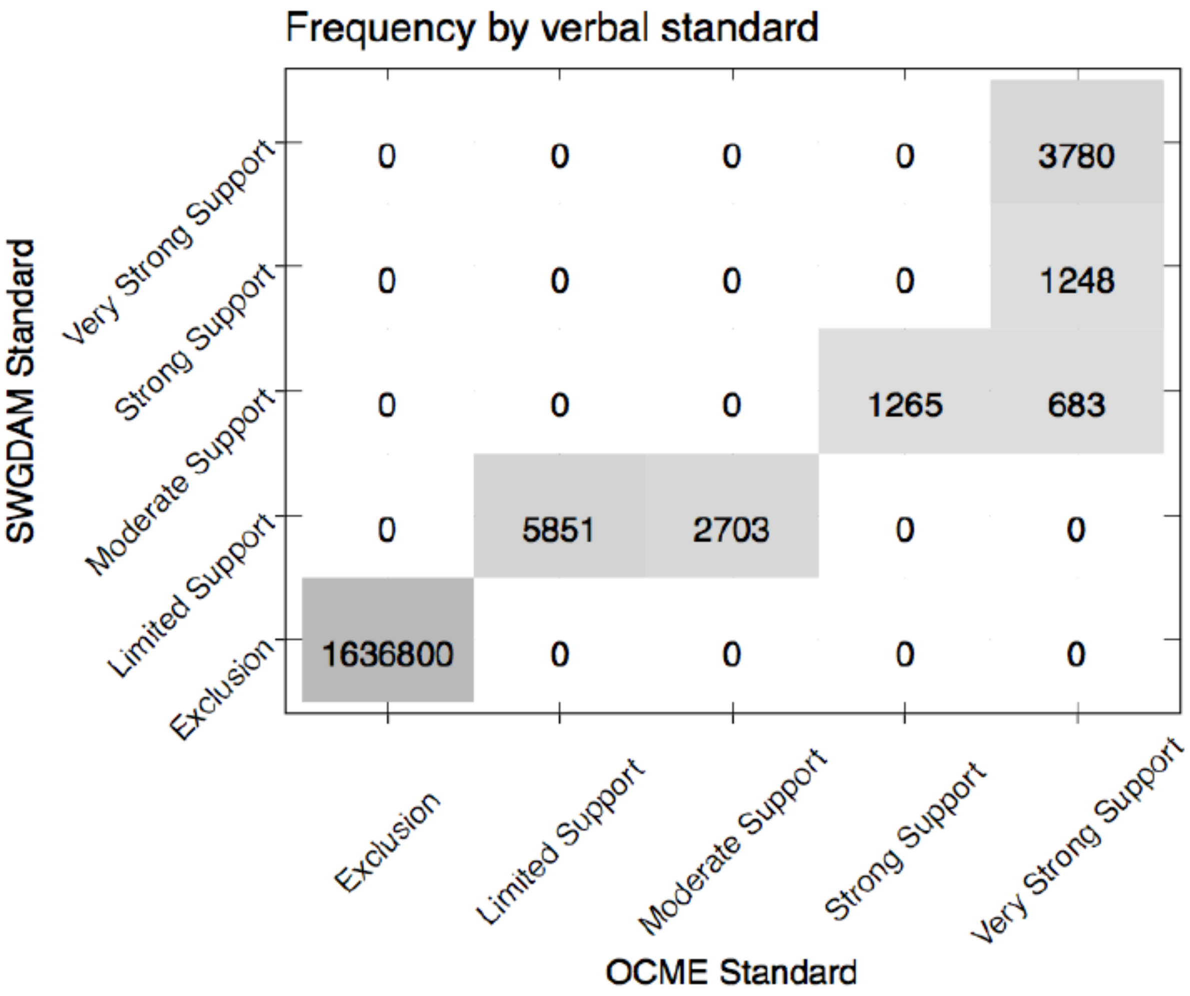
- Interests of developers vs. deciders vs. those decided about
  - Rare bugs matter to individuals
  - Developers: finding and fixing is expensive! Isn't it good enough?
  - Customers/deciders: Decision-making more efficient/ minimize risk

Questions

- If the court would accept results from a number of different forensics system, what does it say when their results differ?
- What does it say when the results are highly sensitive to changes in parameters that for which it would be difficult to determine the correct setting?
- What will incentivize iterative improvement/ finding and fixing bugs found post-deployment rather than using trade secret to shield from disclosure?


Mean likelihood ratio by configuration
Likelihood Ratio (log scaled)
0
−20
−40
−60
Inclusion/Exclusion Boundary
System Configuration
Contributors, Default settings
Contributors, As FST settings
Non−Contributors, Default settings
Non−Contributors, As FST settings
FST
LRmix
EuroForMix
System


Parameter sensitivity analysis
Likelihood Ratio (log scaled)
0
−50
−100
0
0.01
0.02
0.03
0.05
0.09
0.1
0.5
Drop-out Rate
0
0.01
0.02
0.03
0.05
0.09
0.1
0.5
Theta Value
Configuration
● Contributors
▲ Non–Contributors



Frequency by verbal standard

SWGDAM Standard \ OCME Standard	Exclusion	Limited Support	Moderate Support	Strong Support	Very Strong Support
Very Strong Support	0	0	0	0	3780
Strong Support	0	0	0	0	1248
Moderate Support	0	0	0	1265	683
Limited Support	0	5851	2703	0	0
Exclusion	1636800	0	0	0	0

Procurement Phase Wishlist

- When public money used for critical software systems (e.g. criminal justice software), require! or at least give credit for:
    - Software artifacts: bug reports, internal testing plans and results, software requirements and specifications, risk assessments, design documents, etc.
    - Source code
    - No clauses preventing third party review or publishing of defects found
    - Access to executables for third party testing
    - Testing against diverse sub-population benchmarks
    - Bug bounties
- Require scriptable interfaces!
- Design software systems to be compared! And regularly compare them. Standards from NIST
- Requirements for validation studies to clearly specify range of testing – not all or nothing
- Reward/Fund/Incentivize non-profit third party entities to do independent testing and find problems!

Conclusion

- We must add the right incentives to make critical software responsive to the needs of individuals, to society and to the law
  - Not just to needs of customers, deciders and developers
- Flaws in the larger socio-technical decision-making processes in which critical software is developed, deployed and trusted
- We should not be deploying critical software systems in an environment that does not incentivize iterative improvement and debugging

Thank you!

jnm@clarkson.edu

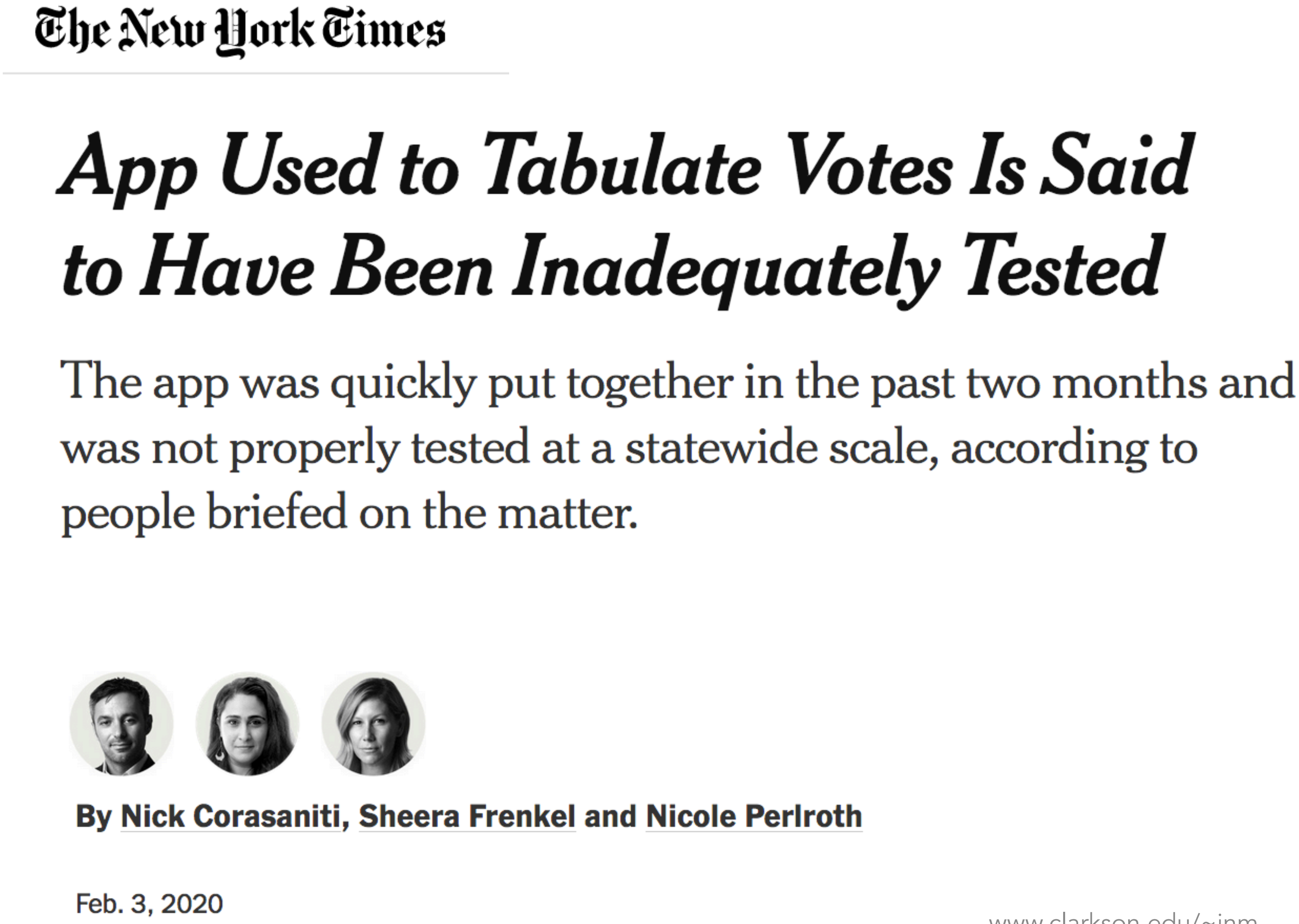
http://www.clarkson.edu/~jnm

@jeanna_matthews

US v. Daniel Gissantaner 1:17cr130

Five specific questions from the Court:

- Has the system been adequately validated?
- Has the system been adequately peer reviewed?
- Have error rates been determined?
- Is the system generally accepted?
- Has it been applied correctly in this case?



The New York Times

App Used to Tabulate Votes Is Said to Have Been Inadequately Tested

The app was quickly put together in the past two months and was not properly tested at a statewide scale, according to people briefed on the matter.

By Nick Corasaniti, Sheera Frenkel and Nicole Perlroth

Feb. 3, 2020