Max
Planck
Institute
for
Software Systems

aircloak

High-Quality Analytics and Strong Anonymization through Cloaking

Sebastian Probst Eide
Aircloak

Mar. 2015

Aircloak and MPI-SWS

- Max Planck Institute for Software Systems (MPI-SWS)
  - German computer science institute
  - Basic research
  - 100% government funded

- Aircloak
  - 2-yr old startup
  - 7 people
  - Seed funding German government (EXIST)

aircloak

Max Planck Institute for Software Systems

Conventional anonymization

- Perturb data, rendering it safe (statistical microdata publishing)
  - Many techniques
  - K-anonymization, data swapping, randomization, ….
- Many problems
  - Data specific, complex to work with
    - What techniques to use?
    - How to set parameters?
    - Complex and fiddly trade-off between utility and privacy
  - Can’t combine datasets
  - Generally doesn’t deal well with dynamic data
  - Complex data simply cannot be effectively anonymized without destroying utility

aircloak

Max Planck Institute for Software Systems

Approach taken by medical research

- Anonymize as much as possible ***without compromising*** analytic quality
  - Typically means pseudonymization only

- Share data only within an appropriate *trust framework*
  - Careful vetting of data recipients
  - Contractual oversight

- Inform users, gain consent

aircloak

Max Planck Institute for Software Systems

Approach taken by medical research

- Anonymize as much as possible ***without compromising*** analytic quality
  - Typically means pseudonymization only

If this was perfect…..

- Share data only within an appropriate *trust framework*
  - Careful vetting of data recipients
  - Contractual oversight

- Inform users, gain consent

aircloak

Max Planck Institute for Software Systems

Approach taken by medical research

- Anonymize as much as possible ***without compromising*** analytic quality
  - Typically means pseudonymization only

If this was perfect…..

- Share data only within an appropriate *trust framework*
  - Careful vetting of data recipients
  - Contractual oversight
- Inform users, gain consent

This wouldn't be necessary

aircloak

Max Planck Institute for Software Systems

Our system goals

- *Substantial* improvements in anonymization with minor loss in analytic quality
  - “Legally” anonymous
- One-size-fits-all anonymization
  - No tuning parameters
  - Independent of type of data
- Reduction or (in many cases) elimination of contractual oversight
  - Increased data sharing at lower cost
- Elimination of informed user consent

aircloak

Max Planck Institute for Software Systems

Fundamentals of our approach

- Retain the raw (or pseudonymized) data
- Protect data in database
  - Zero-access, zero-password
- Run analytics over raw data
  - Eliminates decisions of how to manipulate the data
  - Retains full data fidelity
- Anonymize the answers, not the data
  - Beyond differential-privacy
  - Active anonymization: history of queries and answers is examined

aircloak

Max Planck Institute for Software Systems

Cloak: Zero-access, zero-password black box

- Data stored in cloaks, in raw or pseudonymized form
  - But overall system is anonymous!
- Cloaks are closed, hardened black boxes
  - No direct access to data inside cloaks
  - Operates without a password
- Data encrypted while on disk
  - But there is no password to decrypt it

aircloak

Max Planck Institute for Software Systems

Password-less: how is that possible?

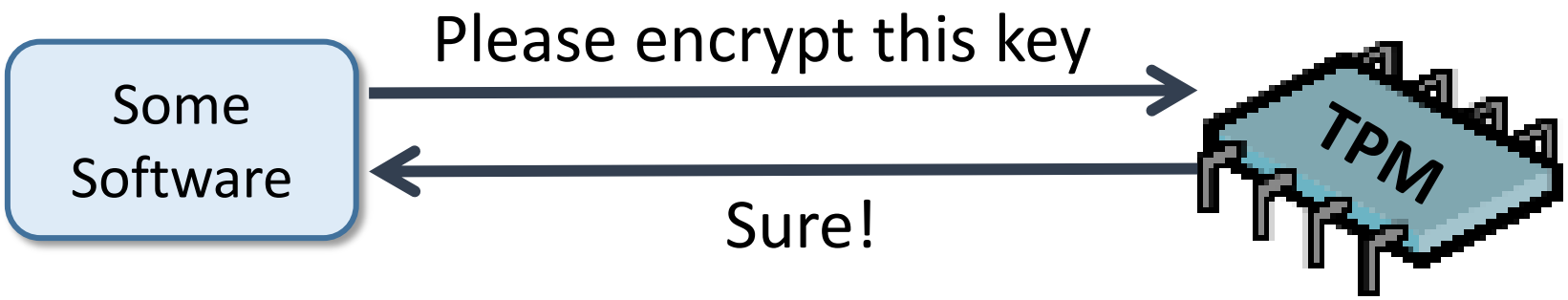
- With special hardware: Trusted Platform Module (TPM)

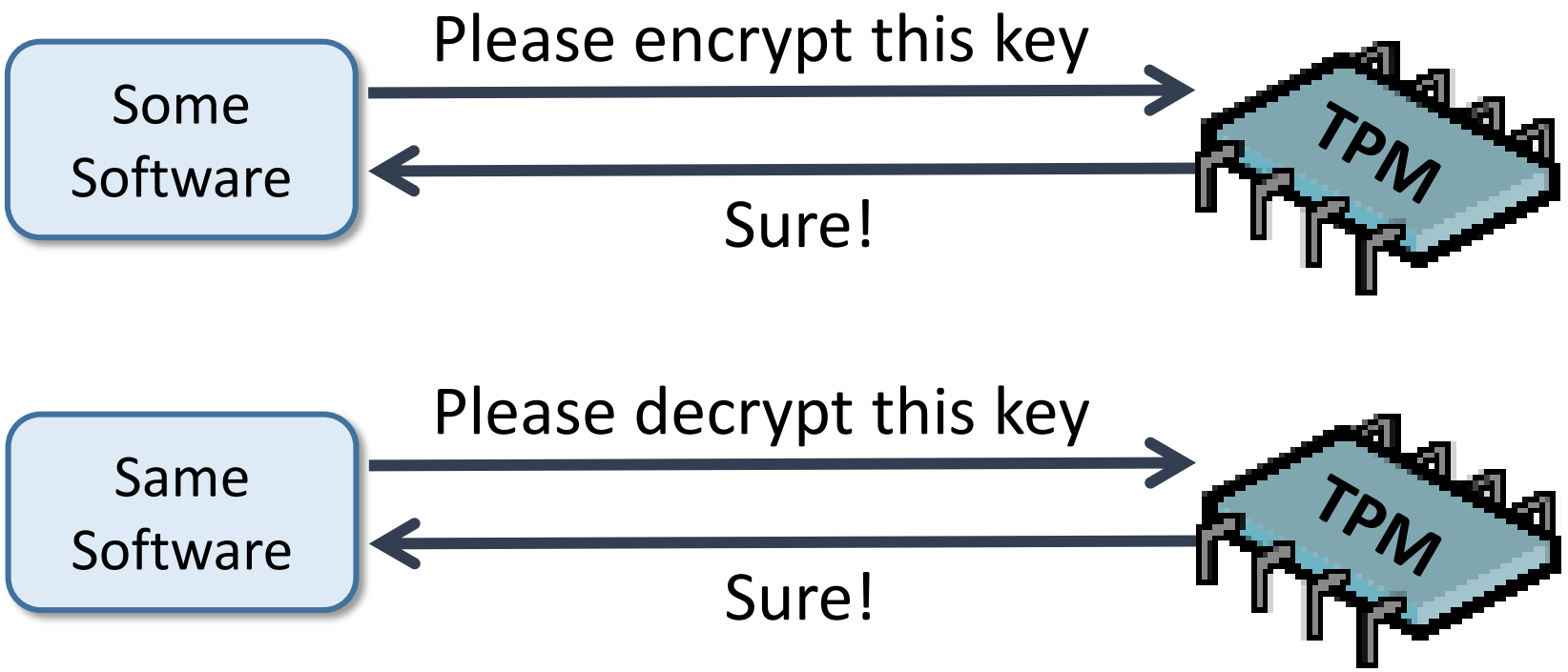
- TPM is a separate hardware chip
  - Can generate and store encryption keys
  - Can do encryption and decryption operations
  - Can ***identify*** the software on the machine!

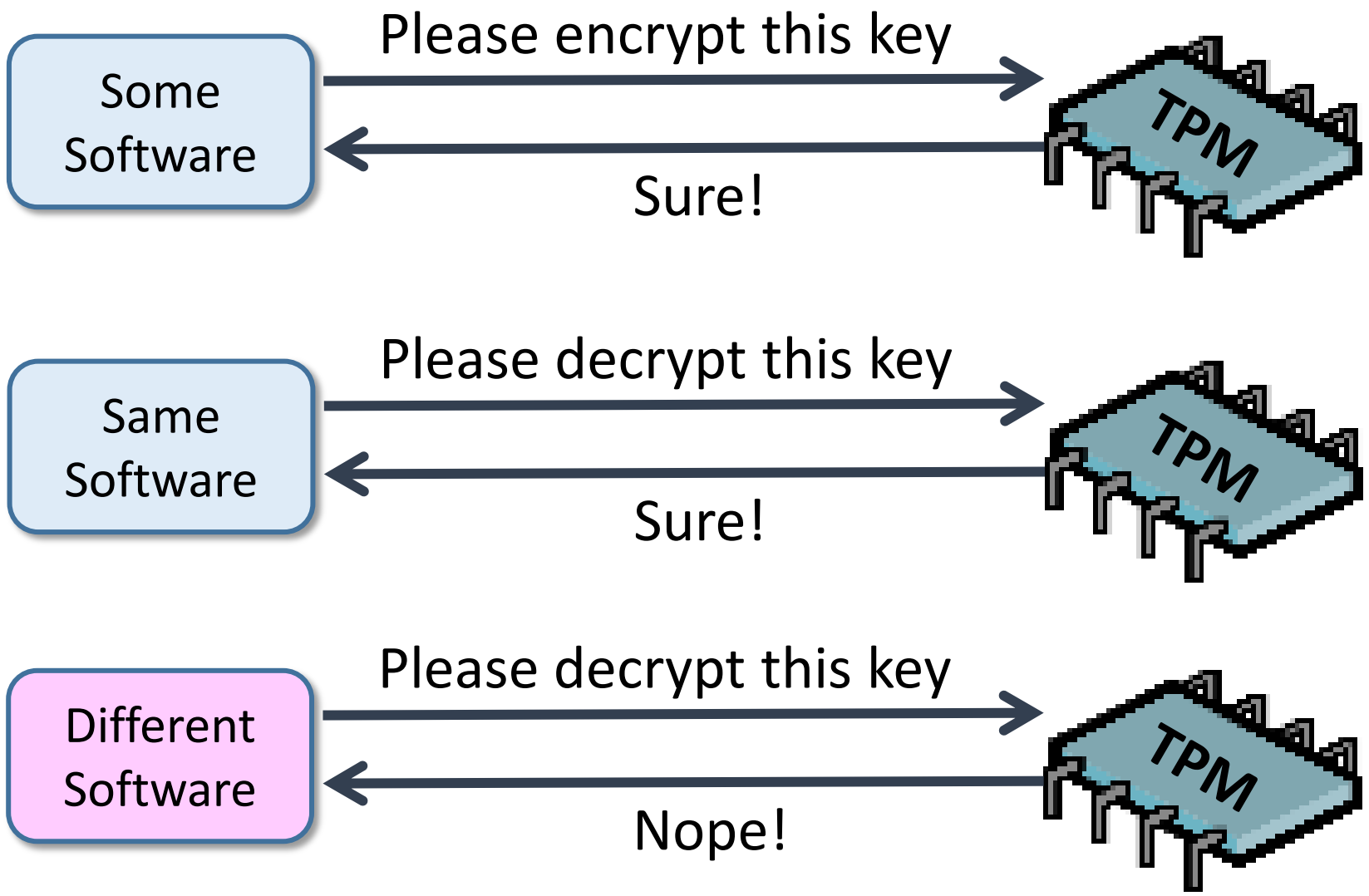
- Special crypto operation called “Sealing”
  - Only the same software that requested encryption can request decryption

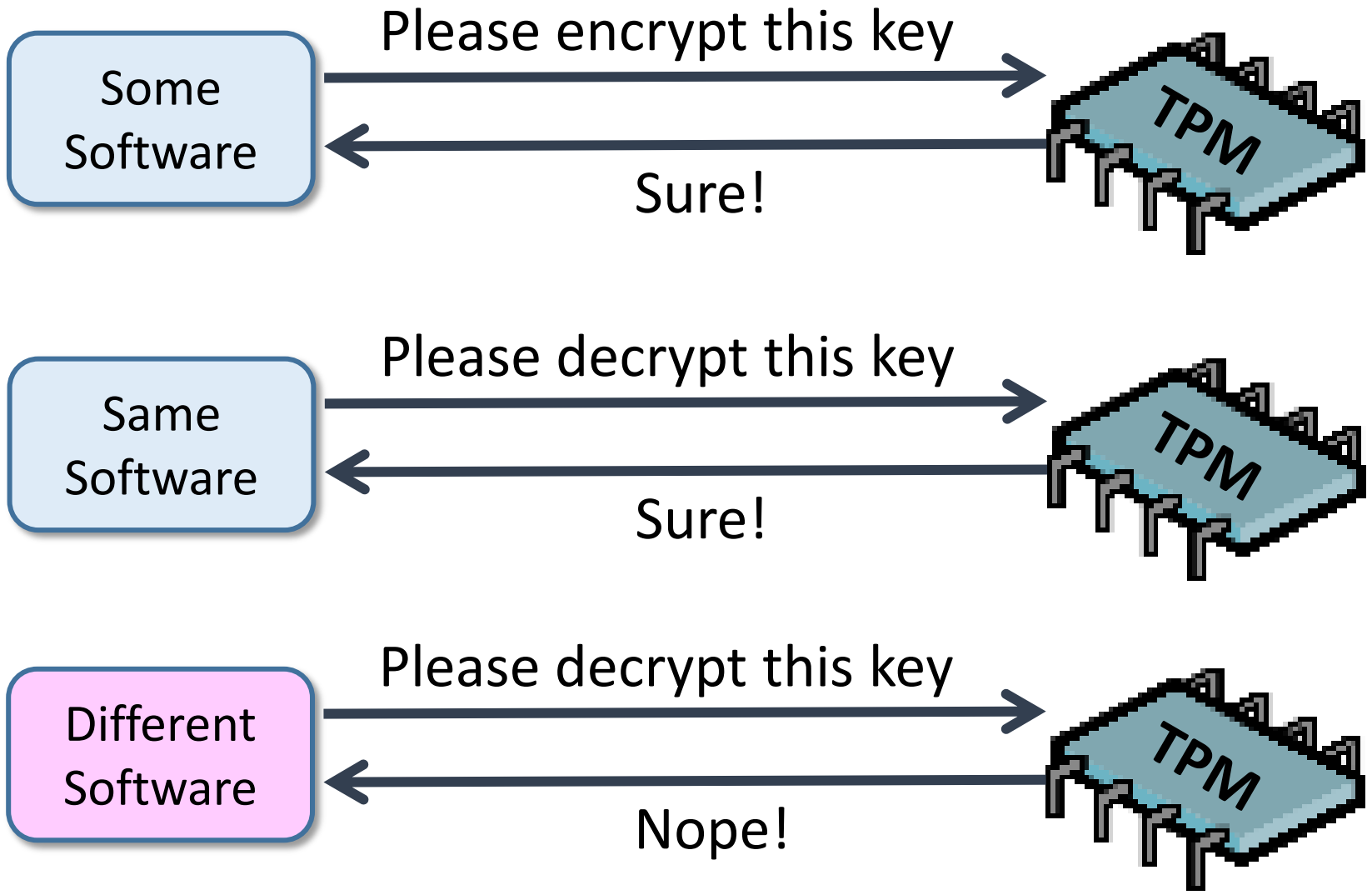
aircloak

Max Planck Institute for Software Systems


Some Software
Please encrypt this key
Sure!
TPM


Some Software
Please encrypt this key
Sure!
TPM
Same Software
Please decrypt this key
Sure!
TPM


Some Software
Please encrypt this key
Sure!
TPM
Same Software
Please decrypt this key
Sure!
TPM
Different Software
Please decrypt this key
Nope!
TPM



The software is the password!

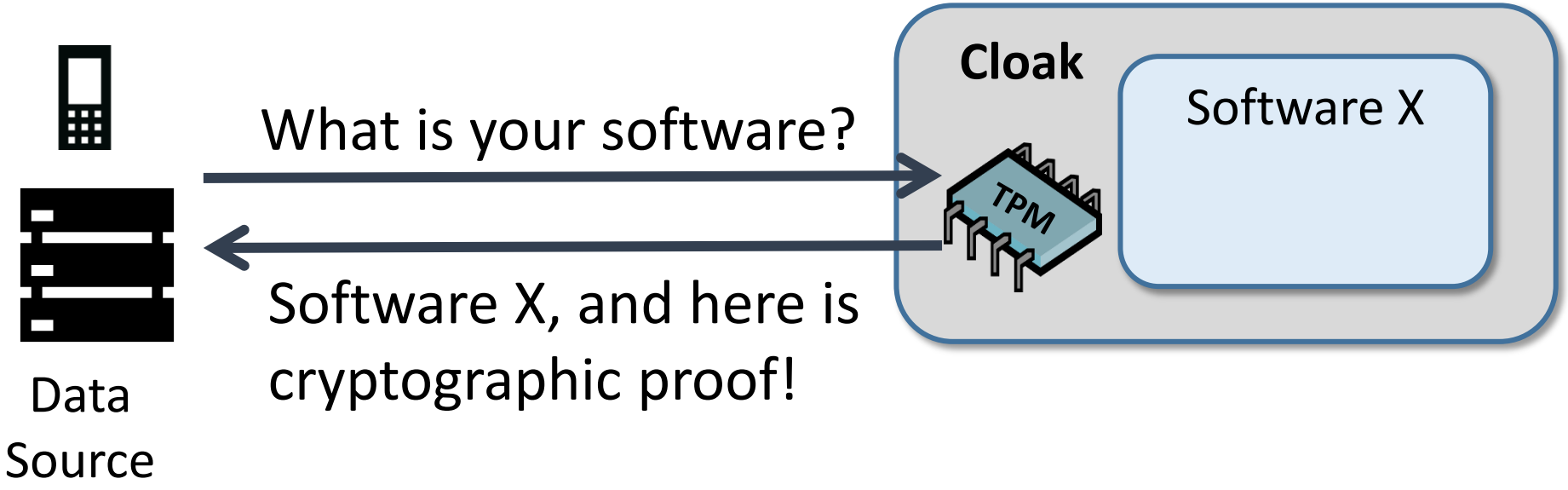
Why should we trust the software?

Why should we trust the software?

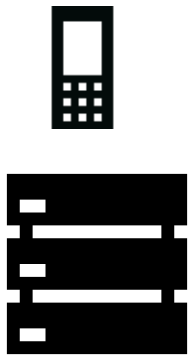
TPM: Remote Attestation

Why should we trust the software?

TPM: Remote Attestation



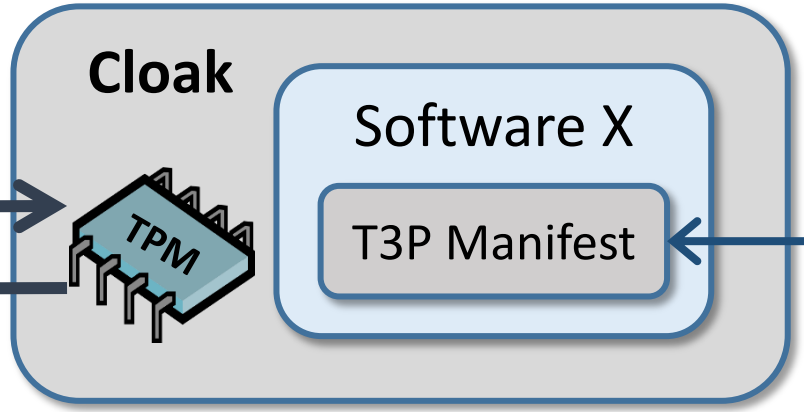
Cryptographically signed Manifest



Data Source

What is your software?

Software X, and here is cryptographic proof!



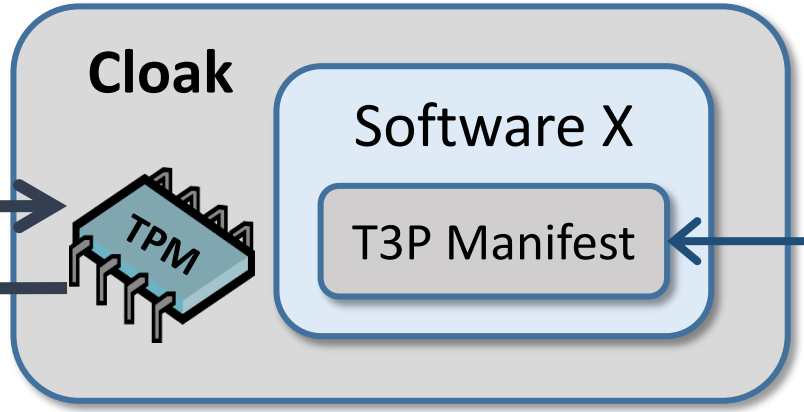
- I trust the T3P
- The T3P trusts Software X
  (and I trust the T3P signature)
- The TPM proves that it is Software X

Cryptographically signed Manifest

Data Source

What is your software?

Software X, and here is cryptographic proof!



Answer Anonymization

- High-level take-away, combination of:
  - filters
  - noise-adding
    - Gaussian, standard deviation around 3
    - Option of removing noise in special cases
  - active monitoring of answers
- Active monitoring:
  - Often detect and silently defeat attack
  - Otherwise, block analyst when too much suspicious behavior
    - “False positives” very rare

Aircloak Status

- Early stage:
  - 2 years old, 7 people, support from German government
- Pilot projects
  - Cisco Berlin innovation center
    - Lamp-post sensors (smart city), Indoor WiFi location service
  - Starting research projects
- Preliminary certification as “legally anonymous” in Germany
- Several patents
- Many conversations (transportation, health, smart city, finance)
- No revenue yet

aircloak

Max Planck Institute for Software Systems

Summary

- Fundamentally new approach to anonymized analytics
- High-quality analytics and strong anonymization
- Eliminates complexity of anonymization design, enables any use case
- sebastian@aircloak.com

Thank-you!

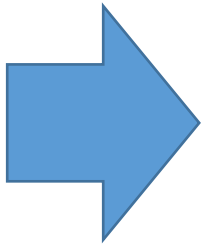
aircloak

Max Planck Institute for Software Systems

Max
Planck
Institute
for
Software Systems

New Trust Framework

- Trust the organization hosting the data
- Trust all the system and network administrators
- Trust the data analysts



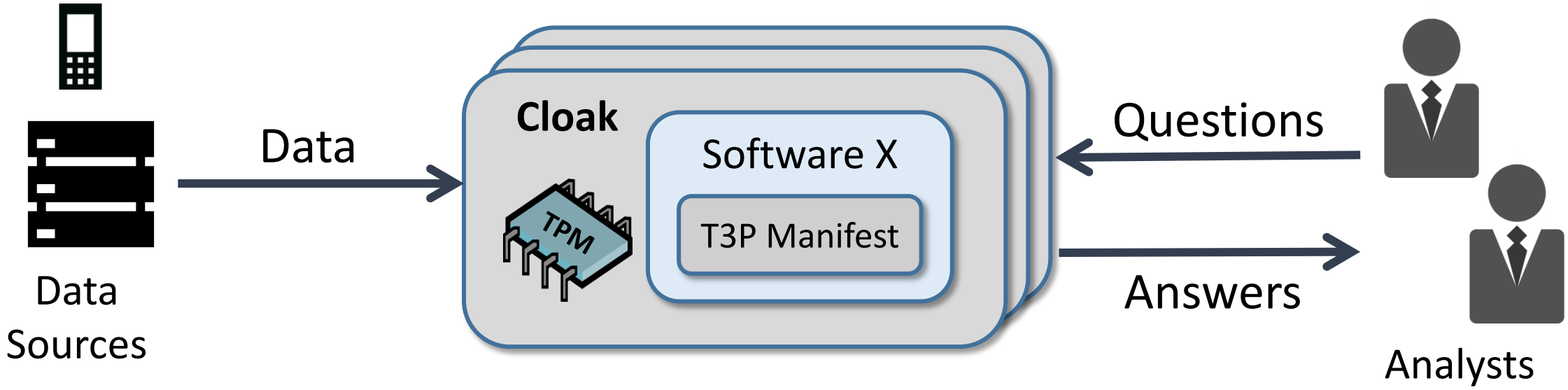
- Trust the “Trusted 3rd Party”

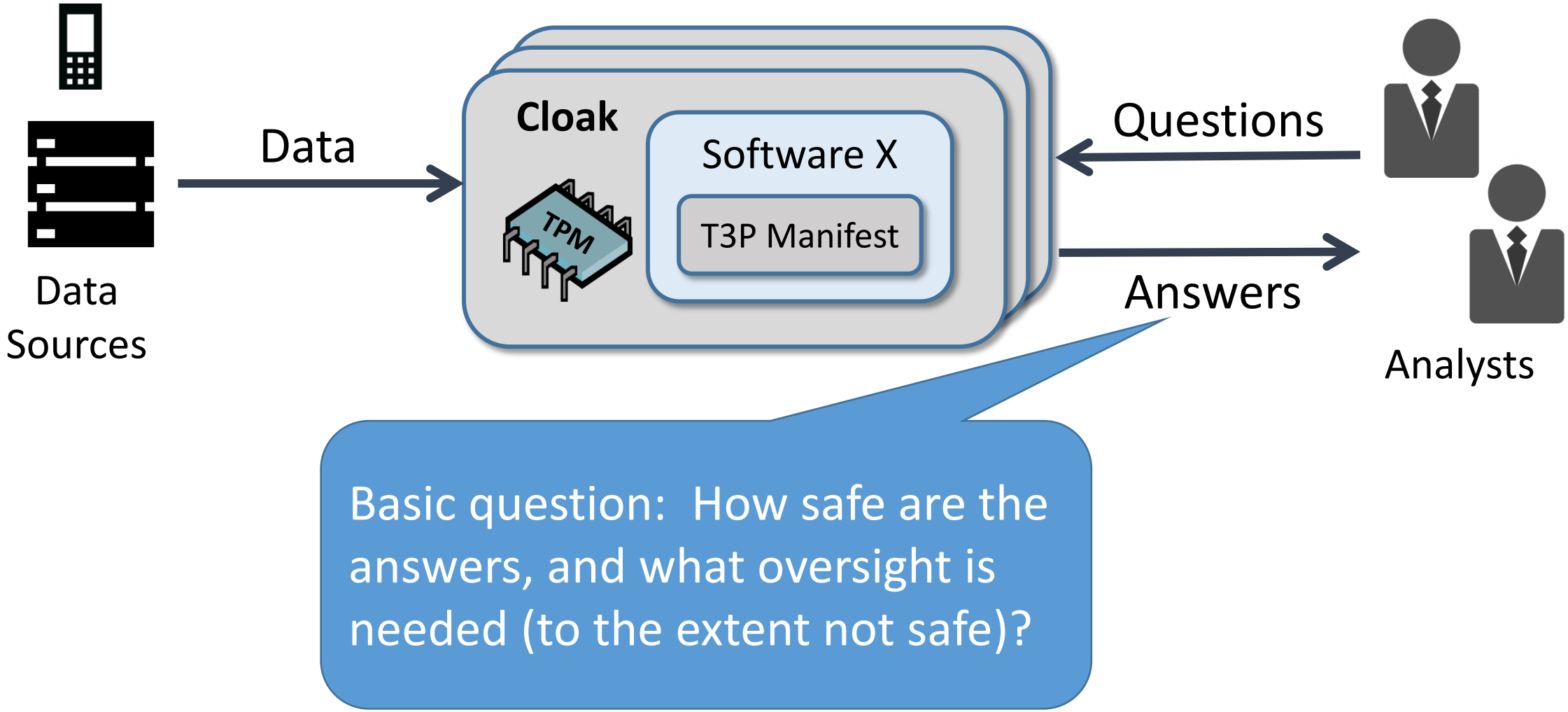
What can go wrong?

- Malware exploit
  - Greatly mitigated by hardened (SELinux) cloak
  - Normal perimeter protection and software upgrade process
- Physical access to cloaks
  - Secure data center
- Malicious software developers
  - All software is signed by two developers
  - Checked by the T3P
  - Open to public inspection!

aircloak

Max Planck Institute for Software Systems


Data
Data Sources
Cloak
TPM
Software X
T3P Manifest
Questions
Answers
Analysts


Data
Data Sources
Cloak
TPM
Software X
T3P Manifest
Questions
Answers
Analysts
Basic question: How safe are the answers, and what oversight is needed (to the extent not safe)?

Current anonymization and analytics

- Designed to be quite general
- Designed to be very safe
  - Going for “legally anonymous” by German law

- Current analytics may not be adequate for many medical applications

aircloak

Max Planck Institute for Software Systems

Current anonymization and analytics

- Aggregate analytics only
  - No chance of revealing individual user information, even if desired
- Cloaks hold raw or pseudonymized data
  - Structured or unstructured
- Analyst queries run over raw data
  - Historic or real-time
- Queries can be arbitrary code, but limited to one user at a time
- Answers must be in the form of user counts (“how many users…..?”)
- Answers have a little noise added
  - Gaussian, zero-mean, standard-deviation around three

aircloak

Max Planck Institute for Software Systems

An example

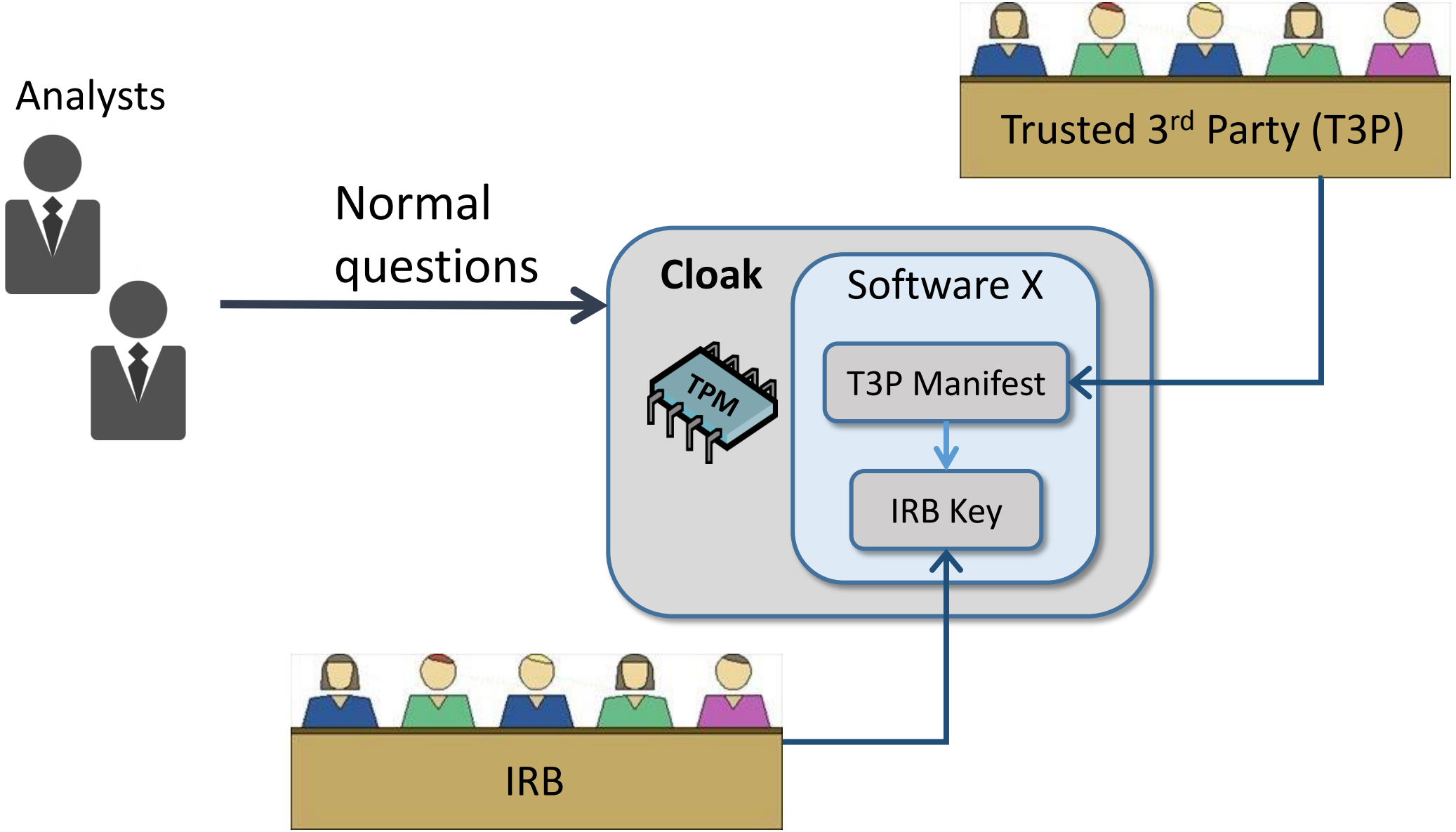
- Say we want to learn what factors lead to a certain condition X
- Generate a set of queries:
  - How many users with condition X have/don't have factor A?
    - Literally:
      - If user has condition X, and factor A, then count in bucket "has factor A"
      - If user has condition X, and not factor A, then count in bucket "does not have factor A"
  - How many users with condition X have/don't have factor B?
  - …….. Etc. (including for combinations of conditions, of course)
- Look for correlations in the answers
  - Counts should be >30 or so to be significant

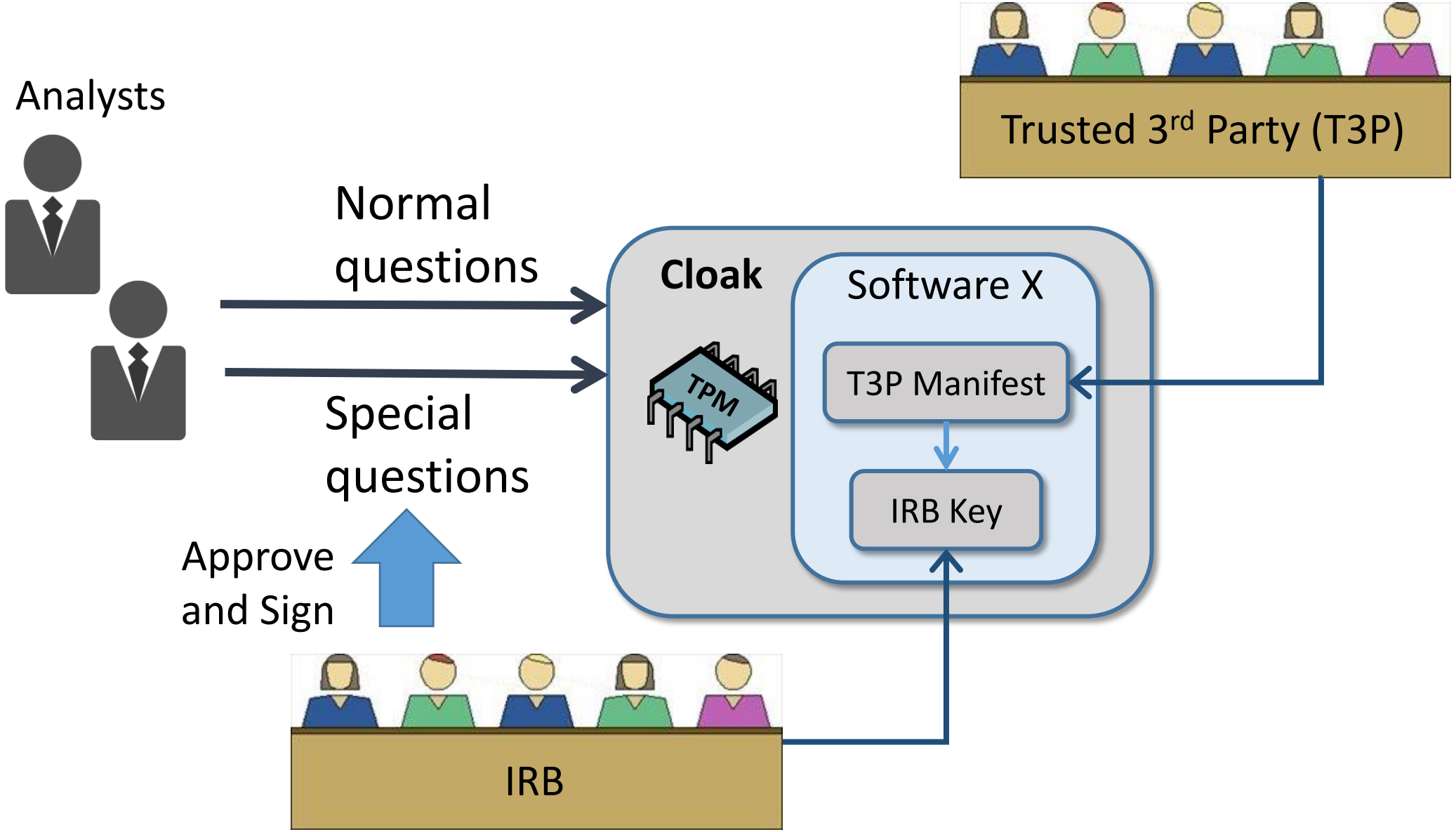
aircloak

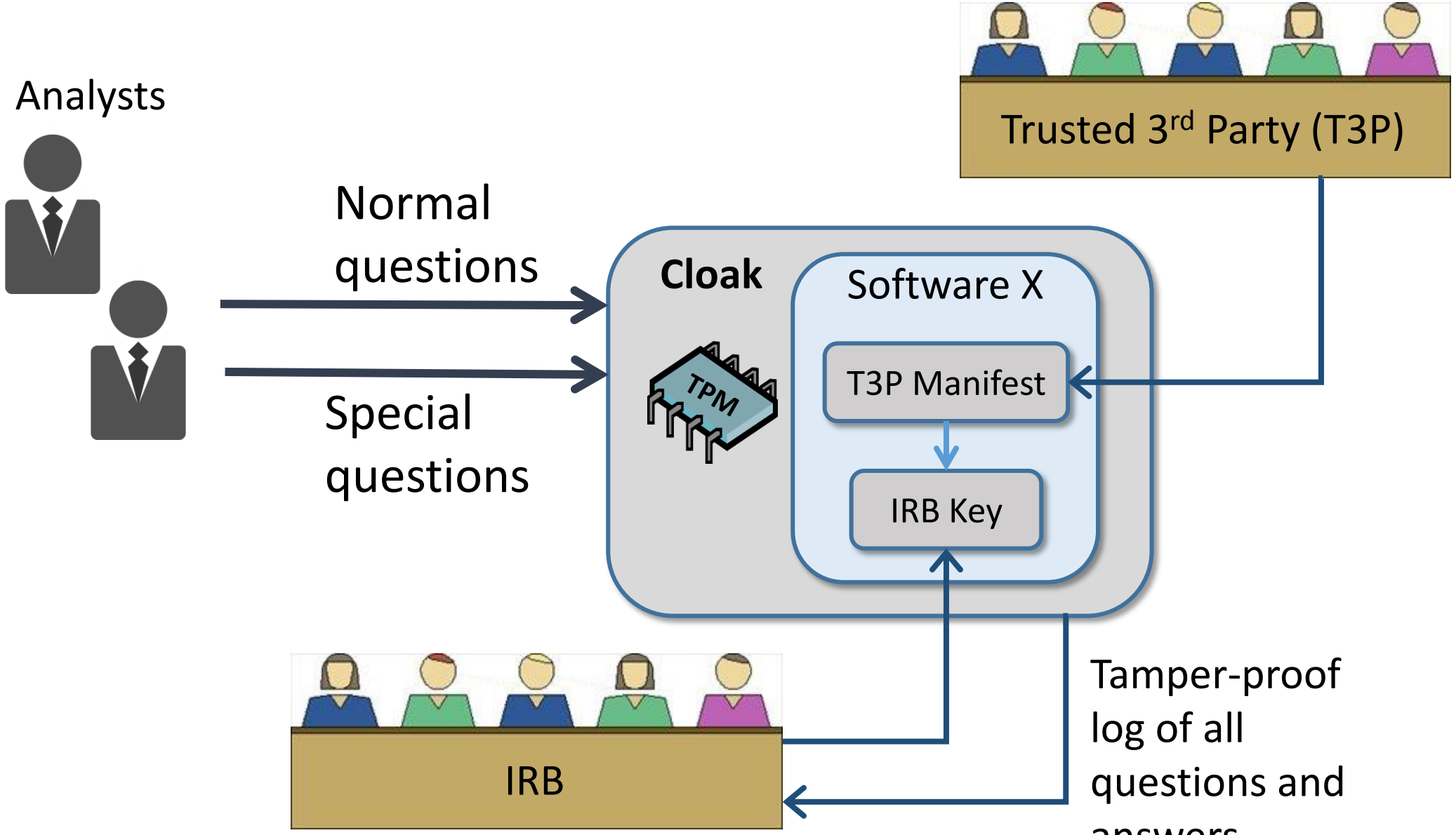
Max Planck Institute for Software Systems

Other options

- What if current approach too inefficient?
  - For instance, need some machine learning (ML) algorithm.
  - Need to understand privacy properties of ML answers
- What if SD=3 is still too much noise?
  - May need some query/answer oversight process to reduce noise in certain situations
- What if want to sometimes identify users (for instance, at risk)?
  - Again, query/answer oversight process needed
  - Need to modify system to base query input on users in database


Analysts
Normal questions
Trusted 3rd Party (T3P)
Cloak
TPM
Software X
T3P Manifest
IRB Key
IRB


Analysts
Trusted 3rd Party (T3P)
Normal questions
Special questions
Cloak
TPM
Software X
T3P Manifest
IRB Key
Approve and Sign
IRB


Analysts
Trusted 3rd Party (T3P)
Normal questions
Special questions
Cloak
TPM
Software X
T3P Manifest
IRB Key
IRB
Tamper-proof log of all questions and answers

Query/Answer Oversight

- Assume an IRB (institutional review board) has query/answer oversight
- Normal analytics does not require IRB approval: only when “interesting” scenarios arise
- T3P authorizes the IRB to the cloak
  - Signs IRB’s public key
- IRB authorizes (signs) queries that can, for instance:
  - Produce noise-less answers
  - Output user IDs

aircloak

Max Planck Institute for Software Systems

To summarize…

- Cloak system dramatically lowers trust requirements
- Current analytics/anonymization (probably soon will be) legally anonymous in Germany
- Current analytics/anonymization may not be adequate for some medical analytics scenarios
- Adequate analytics for medical may not be legally anonymous
- If not, additional oversight needed to satisfy medical community
- Cloak trust framework should dramatically lower cost of health analytics

aircloak

Max Planck Institute for Software Systems