Input Privacy vs. Output Privacy: a reflection from the perspective of official statistics

Fabio Ricciato
Unit A5 'Methodology; Innovation in Official Statistics'
Eurostat

Workshop at Garante Per la Protezione dei Dati Personali
21. September 2022

Caveat

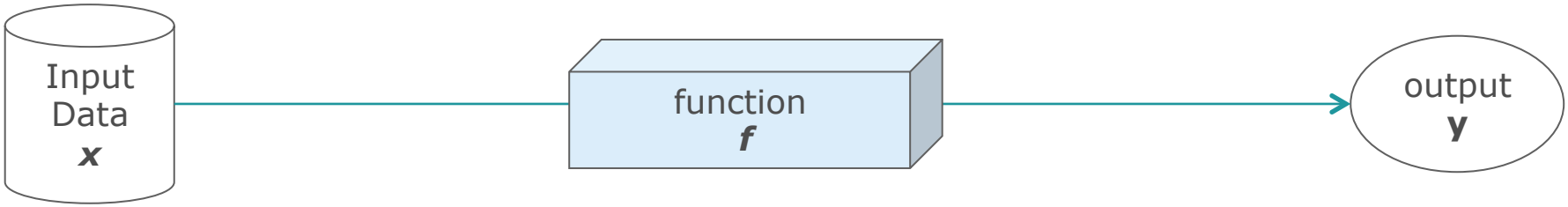
The information and views set out in this presentation are those of the author and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Goal

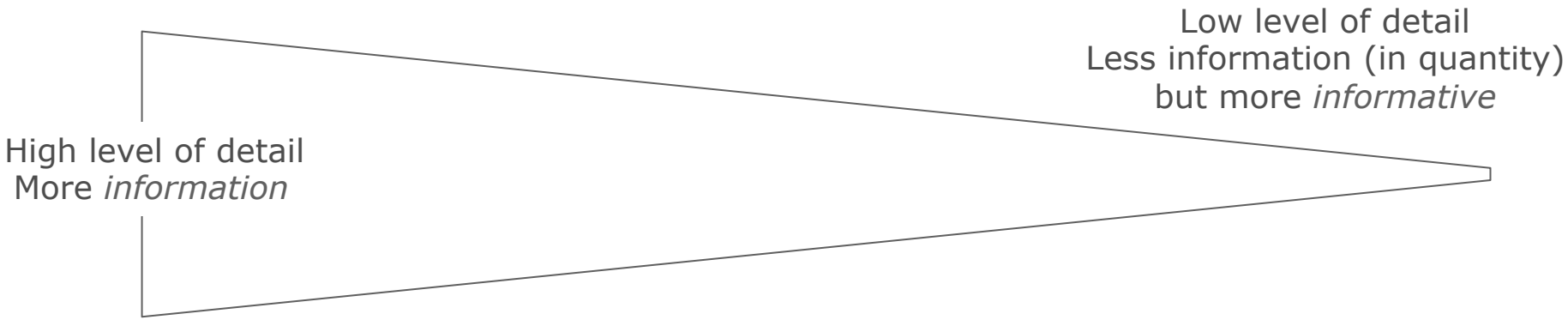
Propose an abstract bird's eye view of Input Privacy vs Output Privacy concepts and their relevance for official statistics …

… serving as basis for discussion (do not expect conclusive words)

Setting the scene: direct computation

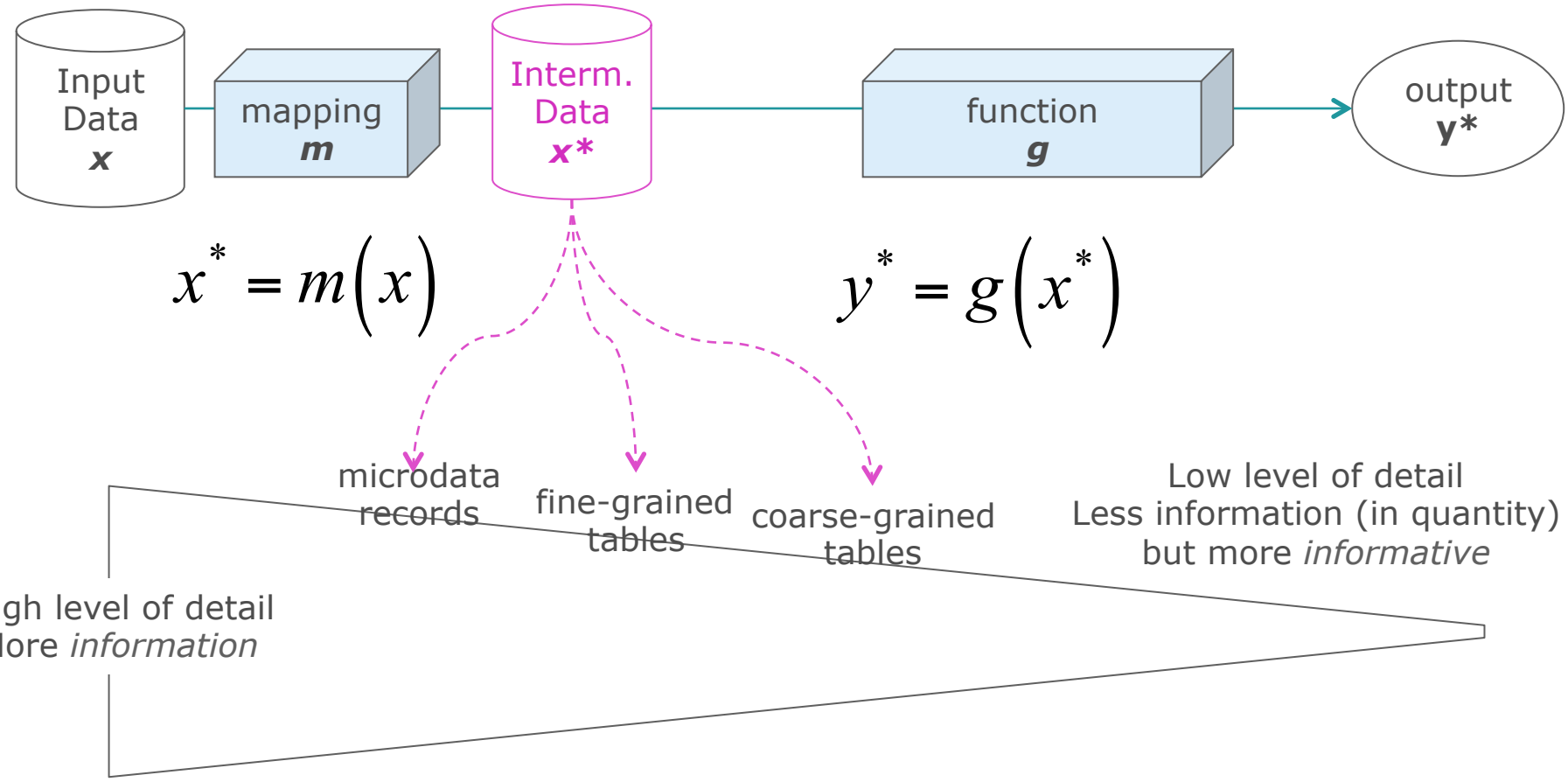


$$y = f\left(x\right)$$



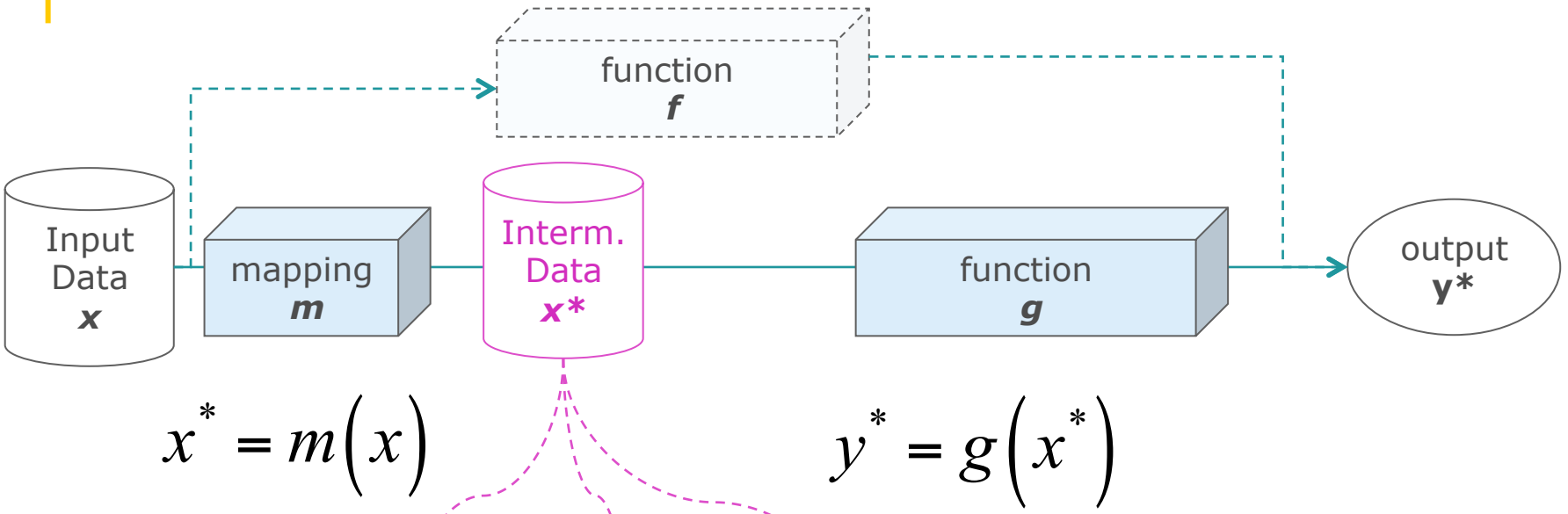
Computation as information reduction process

Setting the scene: mediated computation



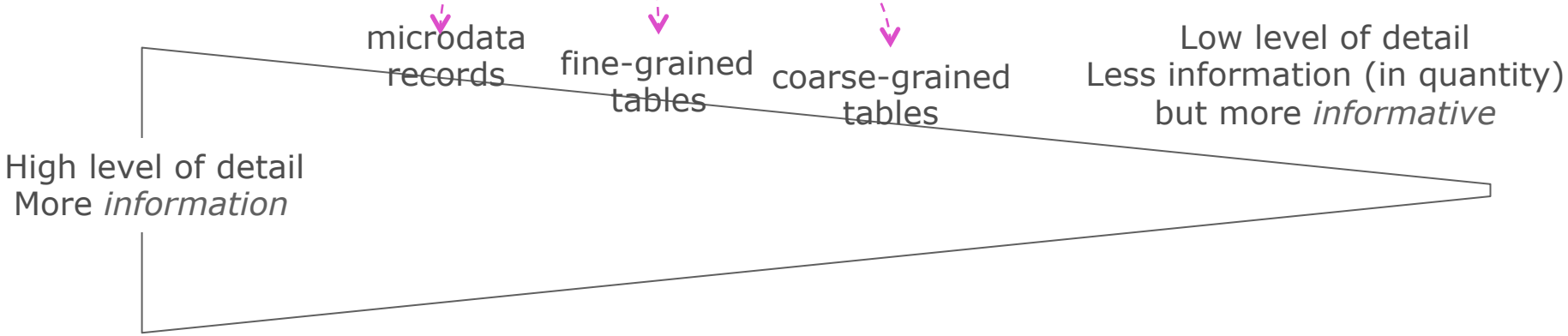
Computation as information reduction process

$$y^* \approx y \Leftrightarrow g\left(x^*\right) \approx f\left(x\right)$$



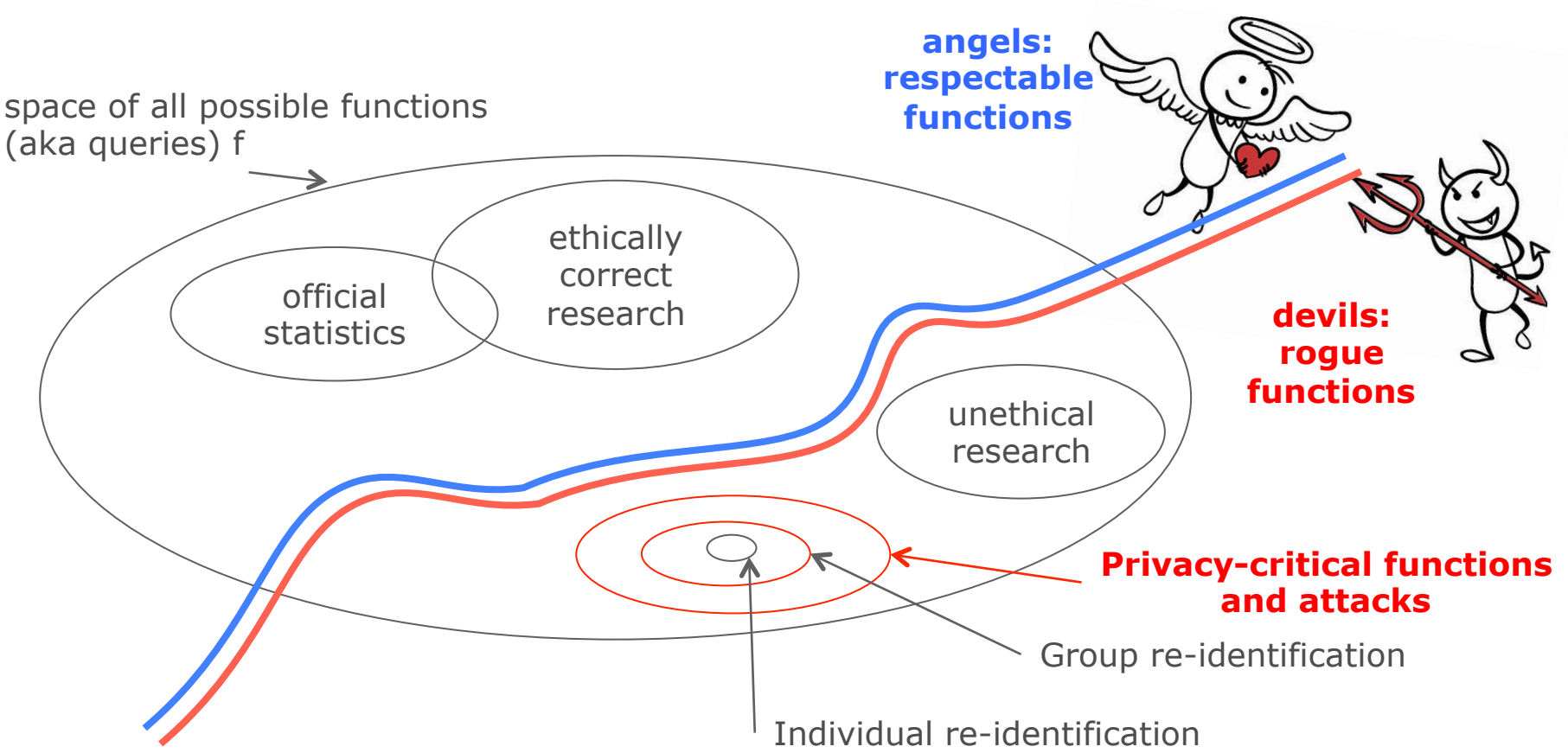
$$x^* = m\left(x\right)$$

$$y^* = g\left(x^*\right)$$

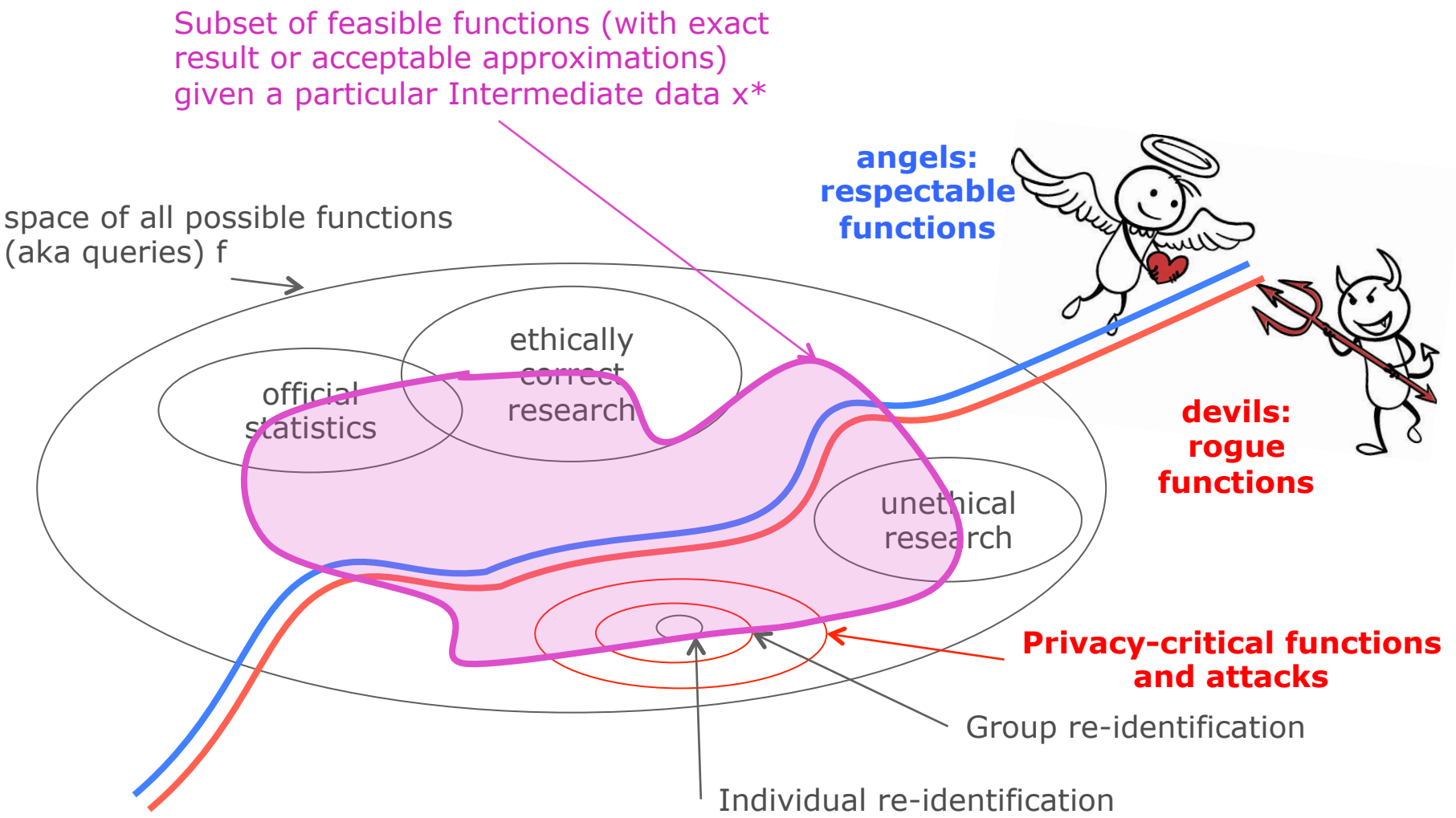


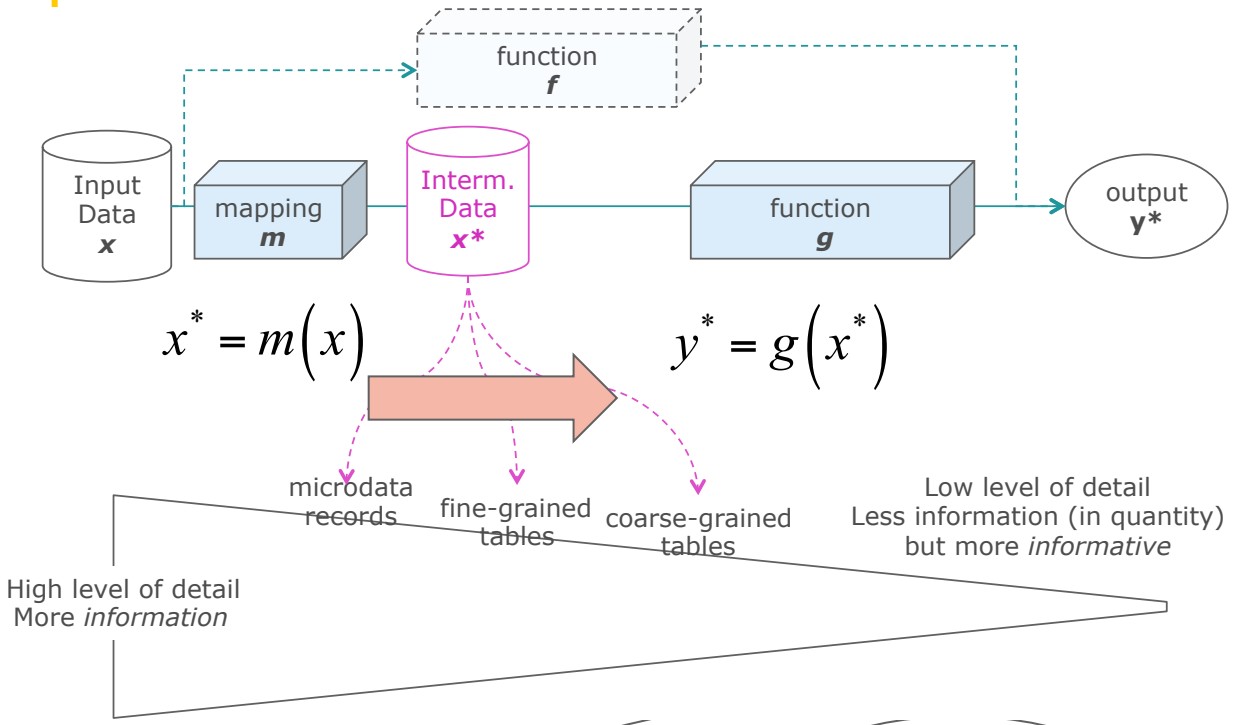
Computation as information reduction process

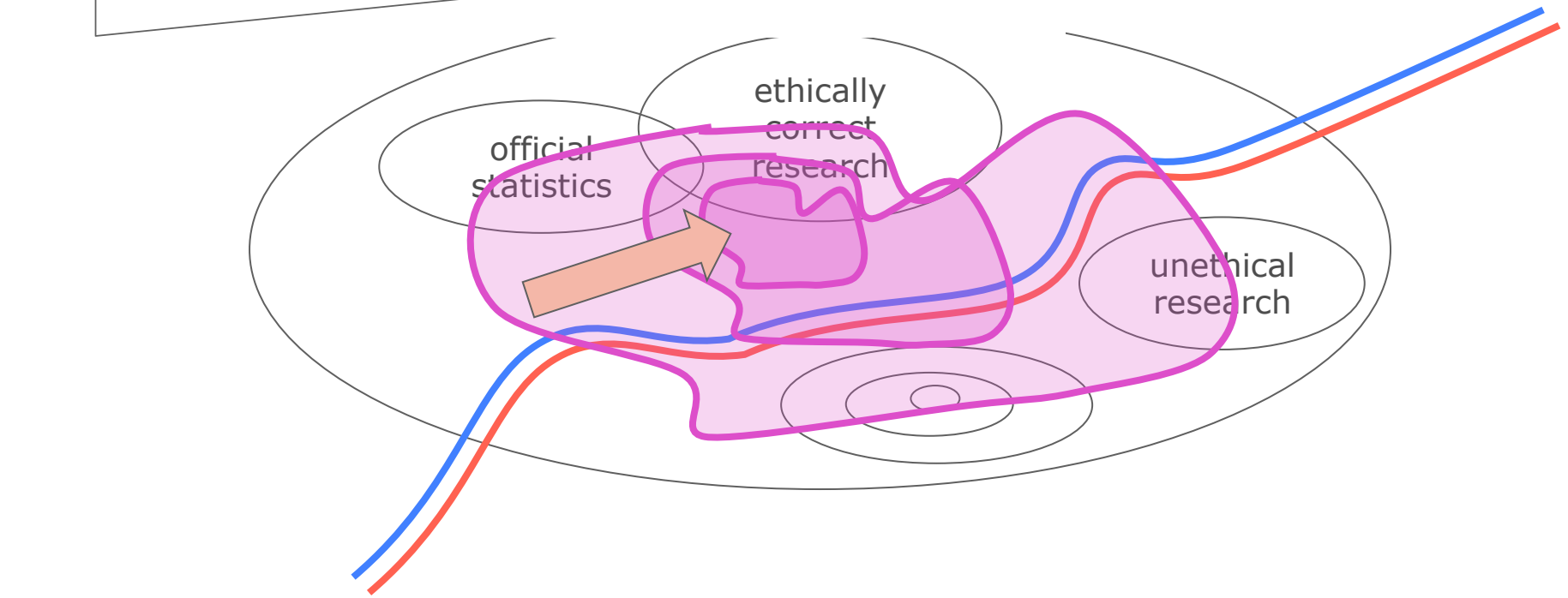
Setting the scene: angels & devils



Setting the scene: angels & devils




function
f
Input
Data
x
mapping
m
Interm.
Data
x*
function
g
output
y*
$x^* = m(x)$
$y^* = g(x^*)$
microdata
records
fine-grained
tables
coarse-grained
tables
Low level of detail
Less information (in quantity)
but more informative
High level of detail
More information

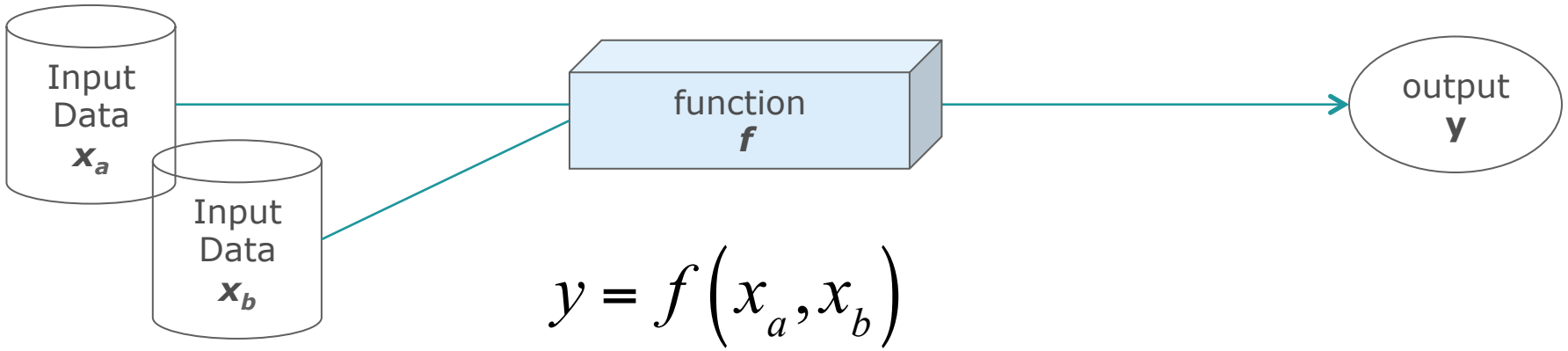

official
statistics
ethically
correct
research
unethical
research

Important questions

- Who is/are the holder(s) of the input data x?
- Who has access to the intermediate data x*?
- Who will be the recipient(s) of the output result y?
- Who defines the function *f* ? Can it be checked?

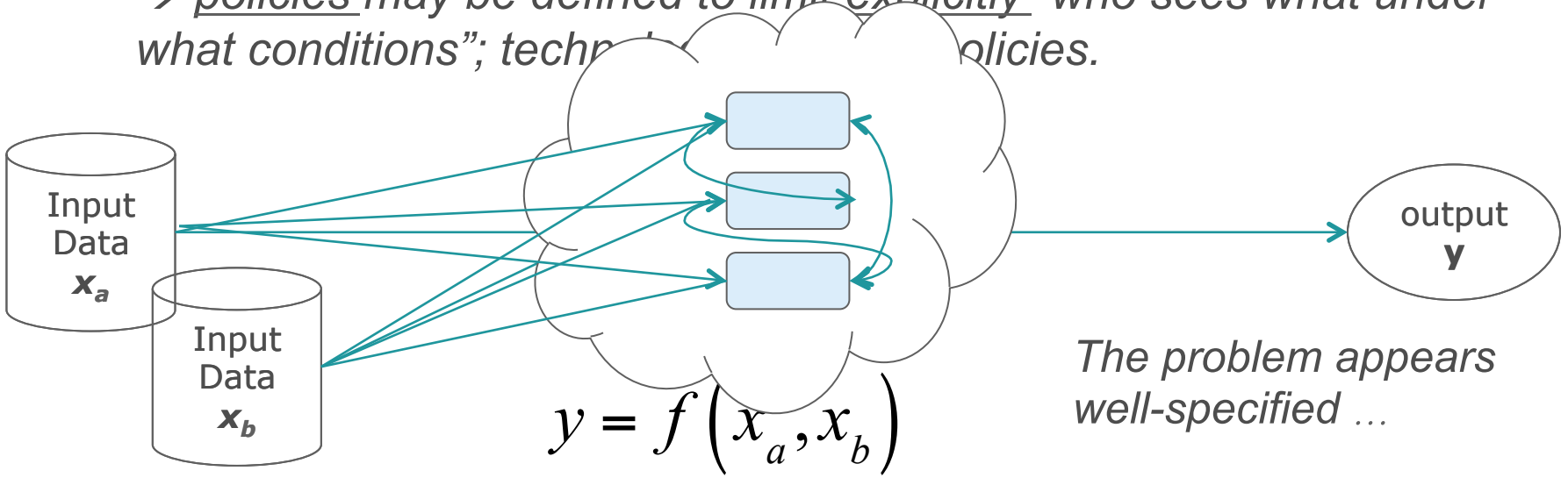
Input Privacy (IP)

- Direct computation, some kind of multi-party settings otherwise it's no fun!
  - output party ≠ input party or N>1 input parties
- Different technical solutions to let the output party learn the exact result y without disclosing anything else to anybody else
  - Possibly involving additional "computing parties", e.g. secret sharing, multi-key schemes … software & hardware safeguards…
- *The function f is declared; the identities of the parties are known → policies may be defined to limit explicitly "who sees what under what conditions"; technology enforces policies.*



Input Privacy (IP)

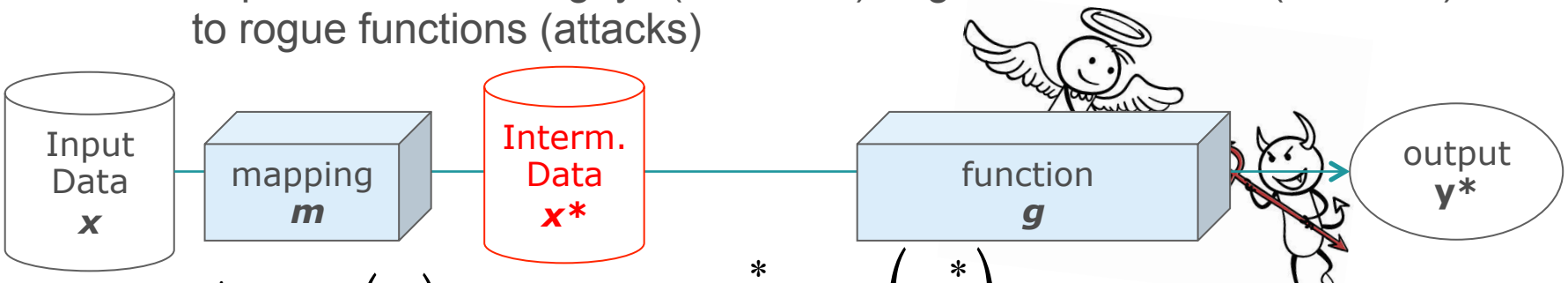
- Direct computation, some kind of multi-party settings otherwise it's no fun!
  - output party ≠ input party or N>1 input parties
- Different technical solutions to let the output party learn the exact result y without disclosing anything else to anybody else
  - Possibly involving additional "computing parties", e.g. secret sharing, multi-key schemes … software & hardware safeguards…
- *The function f is declared; the identities of the parties are known → policies may be defined to limit explicitly "who sees what under what conditions"; techn[illegible]olicies.*



The problem appears well-specified …

Output Privacy (OP)

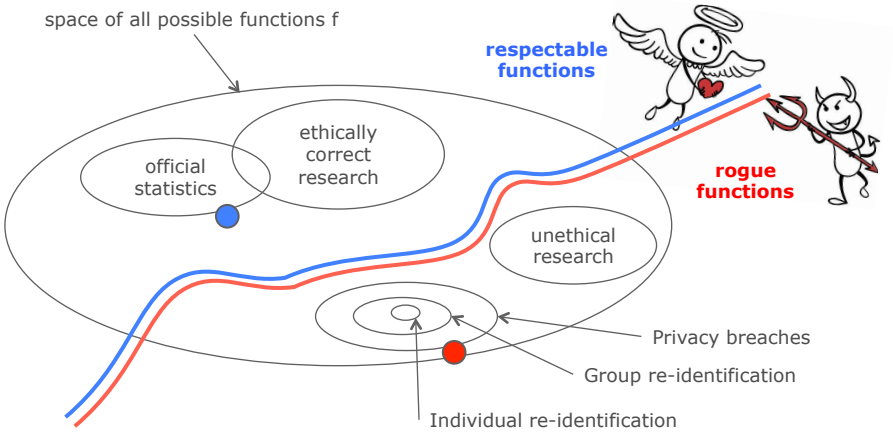
- Mediated computation, with Int. data" *x*=m(x)* released publicly or anyway handled to some non-trusted entity.
- Function *g* and identity of final recipients cannot be checked (and is typically unknown is x* is released publicly)
- Goal of OP methods: find a mapping *m()* such that "*angels will succeed and devils will fail*" $y^* \approx y \Leftrightarrow g(x^*) \approx f(x)$
  - x* enables good guys (analysts) to get "useful" result, i.e., "acceptable" approximations for "respectable functions"
  - x* prevents the bad guys (attackers) to get usable results (succeed) to rogue functions (attacks)



$$x^* = m(x) \qquad y^* = g(x^*)$$

Specifications?

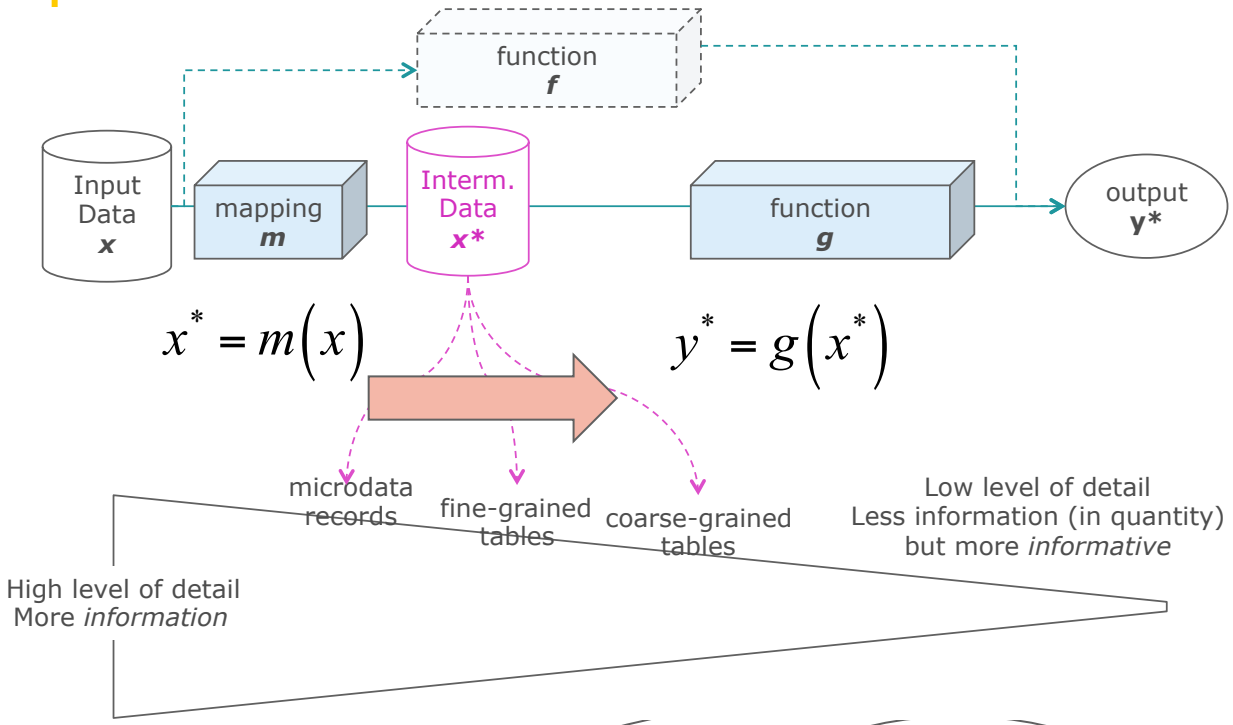
- In a **well-defined** setting one specifies narrowly both sides …
  - Specify one specific function to be supported (or a narrow class thereof) and what is “acceptable” approximation for it (= define the utility metric referred to a specific analysis)
  - Specify one specific type of attack (or narrow class thereof), and when it is successful (= risk metric referred to a specific attack)

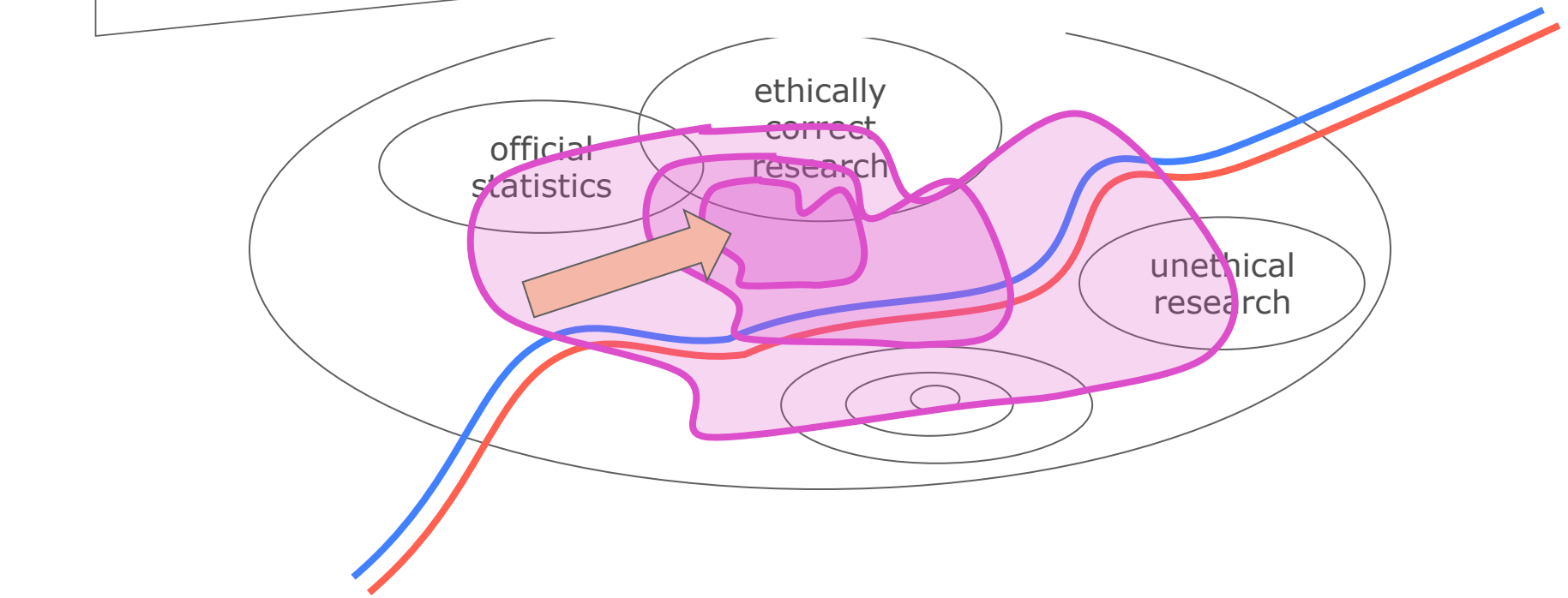


Under these conditions, it is perhaps possible in principle to find a mapping m() that does the job …

Specifications?

- In **semi-defined** settings, one specifies explicitly one side …
  - e.g. in traditional Statistical Disclosure Control schemes based on suppression, the “rogue” function to be counteracted is individual reidentification
  - The viability of SDC scheme depends on the level of detail of the intermediate data
    - Different solutions for m() cut the space of feasible function in different ways
    - In general, as the level of detail of x* decreases (more aggregation)
      → the subset of feasible functions shrinks
      → lower risks and lower utility


function
f
Input
Data
x
mapping
m
Interm.
Data
x*
function
g
output
y*
$x^* = m(x)$
$y^* = g(x^*)$
microdata
records
fine-grained
tables
coarse-grained
tables
Low level of detail
Less information (in quantity)
but more informative
High level of detail
More information


official
statistics
ethically
correct
research
unethical
research

Specifications?

- If the specifications are too loose (or non-existing), then the problem is under-specified, i.e. ill-posed

Consideration:

- When facing a difficult problem, let's question the problem: is it just complex, unfeasible (over-constrained) or ill-posed (underspecified)?
- The question matters when you need to decide where to invest your resources: keep trying **solving** it or better **reformulating** it?

Where to go?

- A good starting point is to spell clearly our goals (and constraints and costs)

- From the perspective of the statistical office: is the goal merely "***to release finer grained data***" or rather "***to enable more and better research***" based on the data?
  - Releasing "*finer-grained but noisier data*" does not necessarily implies enabling "*more and better research*" - some analysis functions will succeed, others may not …

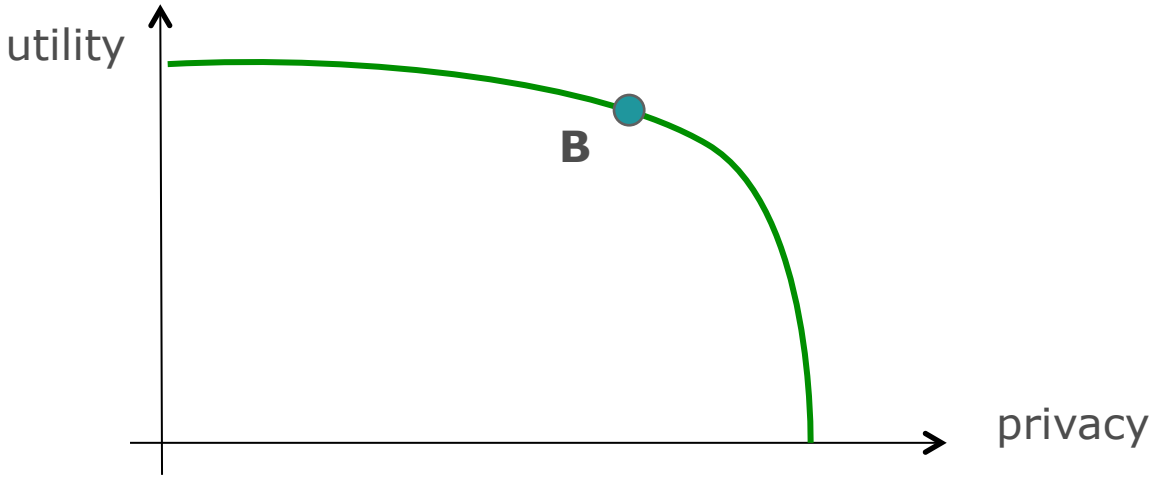
- Releasing data out is not the only way. An alternative is to bring computation in (e.g. safe data access environments)
  - Scale up safe data access mechanisms for researchers, invest in making them more widely available (remotisation), more robust and more cost-effective (automatisation of checks), possibly leverage input privacy technologies… (?)

point for another discussion…

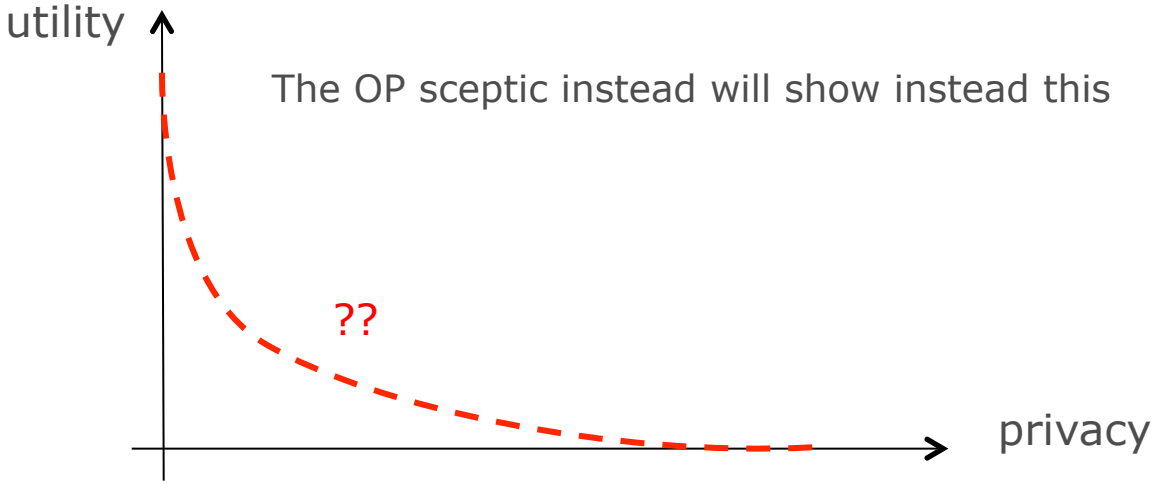
Thank you!

Backup slides

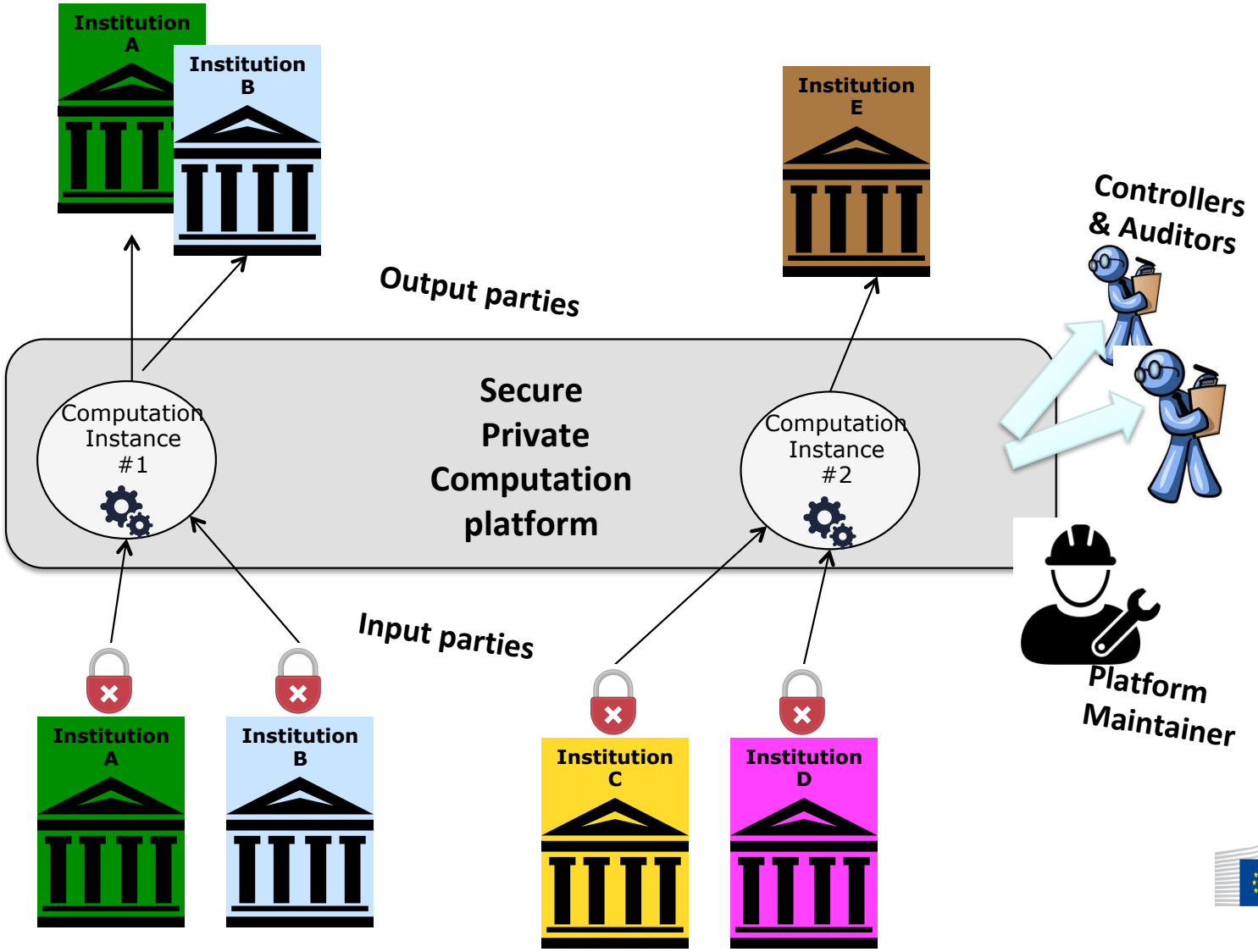
The OP enthusiast will show this



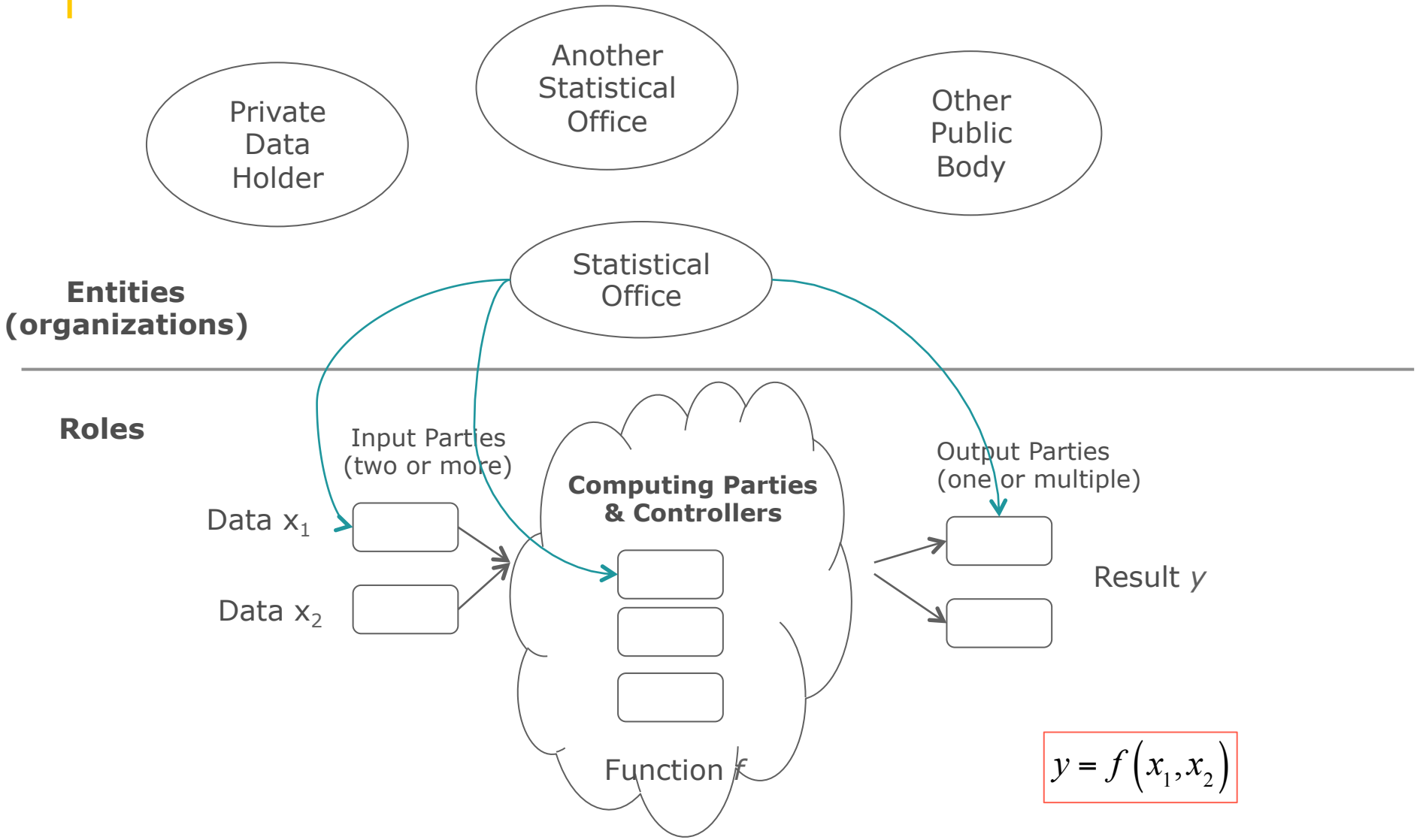
And at best present a concrete example, in a particular scenario of utility and privacy metrics for specific functions and attacks, tailored to specific use-cases, where the curve looks like like that ... but mind easy generalisations!



Secure Private Computing-as-a-service



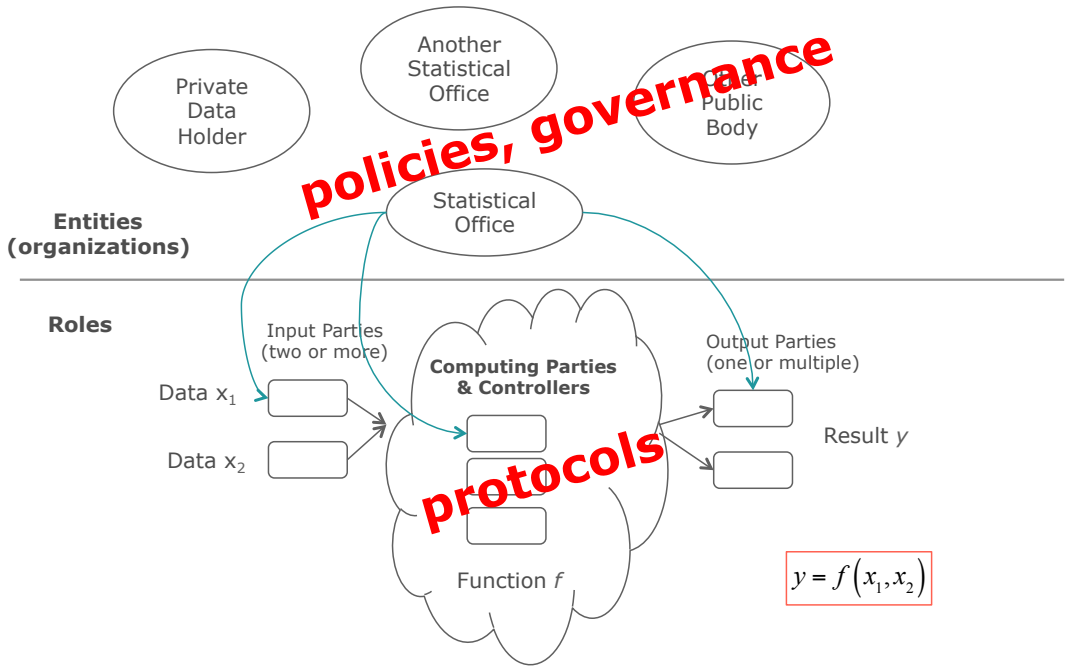
European Commission


Private Data Holder
Another Statistical Office
Other Public Body
Statistical Office
Entities (organizations)
Roles
Input Parties (two or more)
Data x_1
Data x_2
Computing Parties & Controllers
Function f
Output Parties (one or multiple)
Result y
$y = f(x_1, x_2)$

Trust model

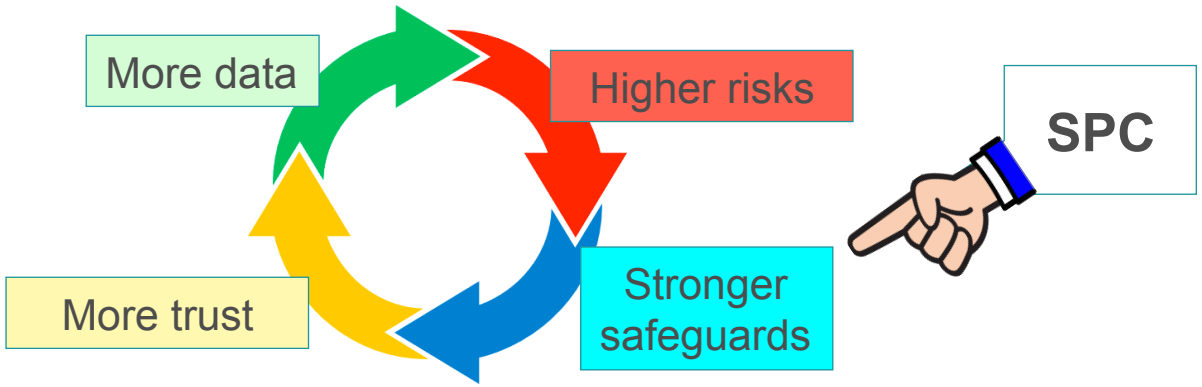
- The essential role of MPC is to enforce technologically the governance/policies (for data & code) defined among entities
- avoiding single-point-of-trust
  → trust the involved entities *collectively, not individually*

- The strength of MPC solution depends *jointly* on
- (i) the robustness of the policies/governance scheme;
- (ii) reliability of involved entities;
- (iii) strength of technology implementation



Why do we care?

- Increasing **appetite** for producing information (e.g., statistics, analyses) from the **combination of data held by different organizations** (private companies, public institutions) possibly in different Member States
  - Statistical authority/ies acting as output party, input party or both
- Increasing **pressure** to strengthen safeguards, "*technical and organisational measures*" for protecting the data
  - legal requirements by Data Protection Authorities
  - necessary condition to build public trust and public acceptance



The End.