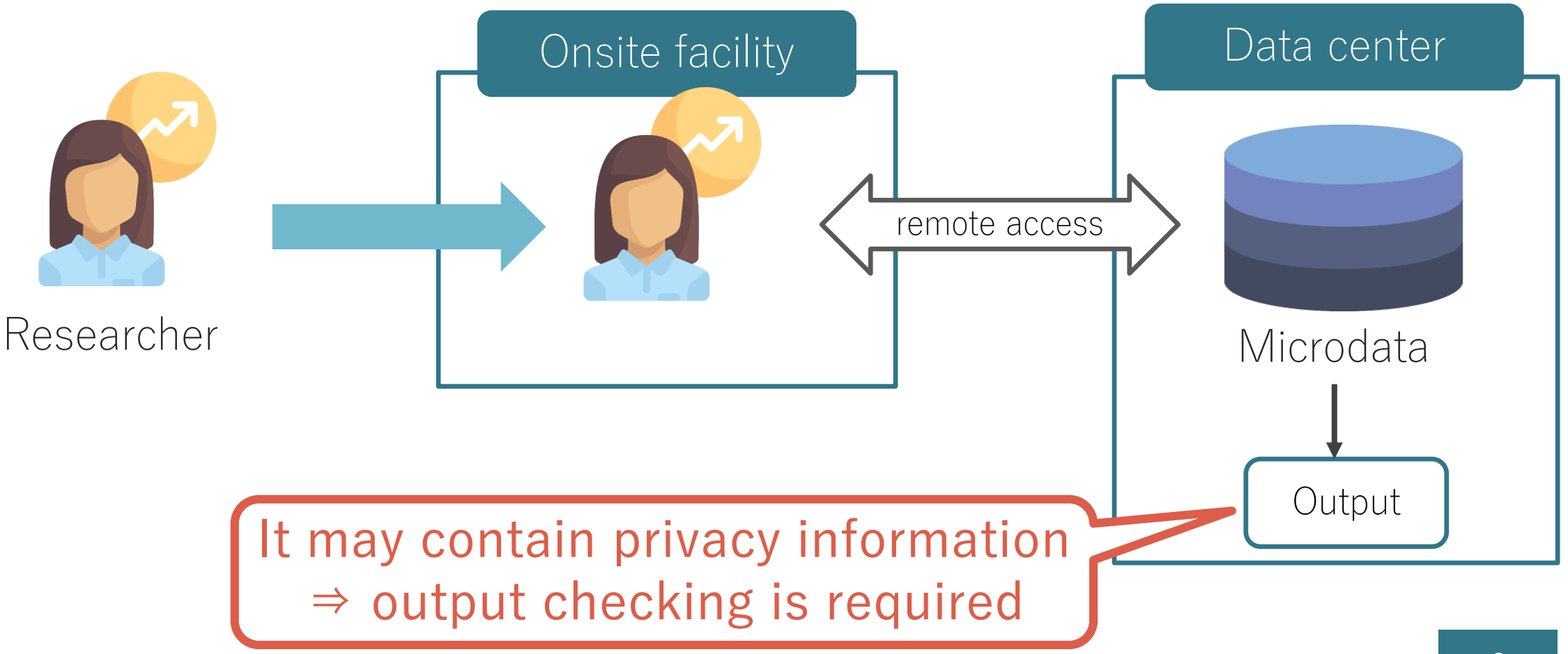
On-site Service and Safe Output Checking in Japan

Ryo Kikuchi
NTT corporation /
National statistics center

Kazuhiro Minami
Institute of statistical mathematics /
National statistics center

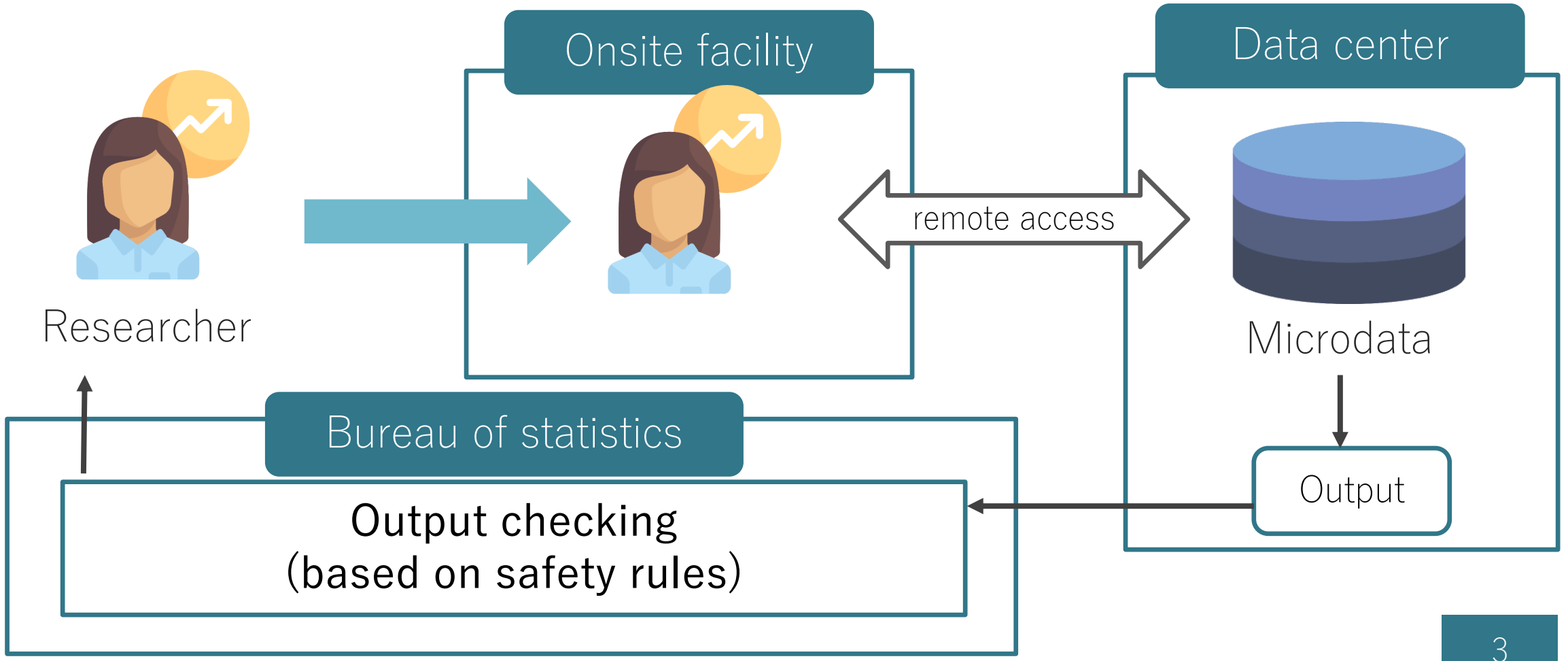
On-site service in Japan

- Researcher can access microdata at an on-site facility



On-site service in Japan

- Researcher can access microdata at an on-site facility



Design of output checking

Assumption and preconditions

- Assumption: researcher does not intend to cheat
  - Assuming malicious researcher is unreasonable
  - Reader of a paper including outputs can be an adversary

- Preconditions:
  - Applying the same set of rules to both intermediate and final outputs

    We do not relax the rules for intermediate outputs
  - Final responsibility is imposed to researcher

    Purpose of output checking is to catch unsafe outputs by an inexperienced researcher

Following current safety standards

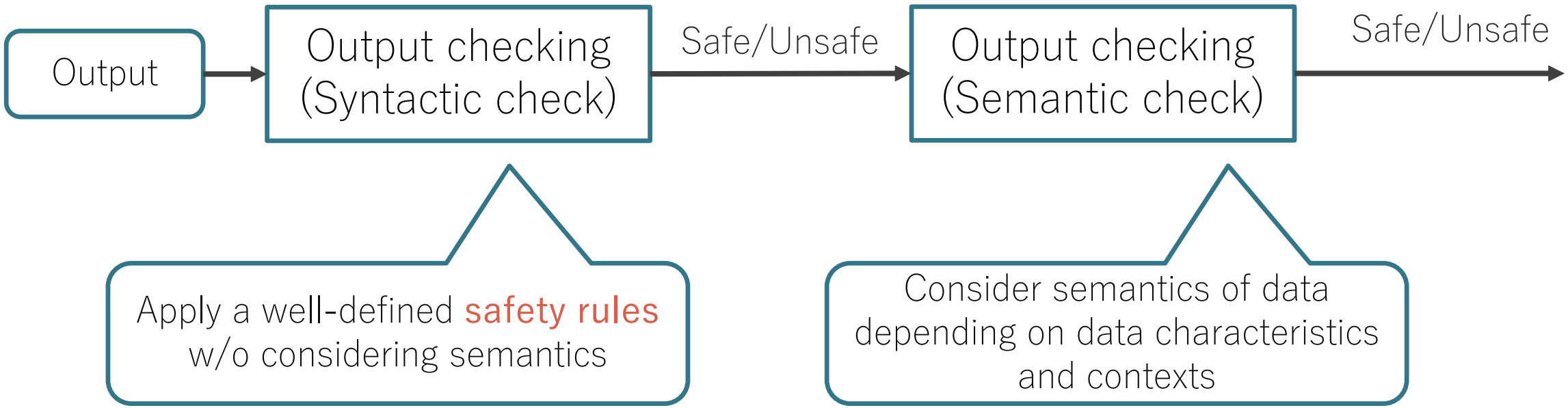
- Rule-of-thumb has played an important role

  All current safety standards are not necessarily **logically** deduced from scientific evidences

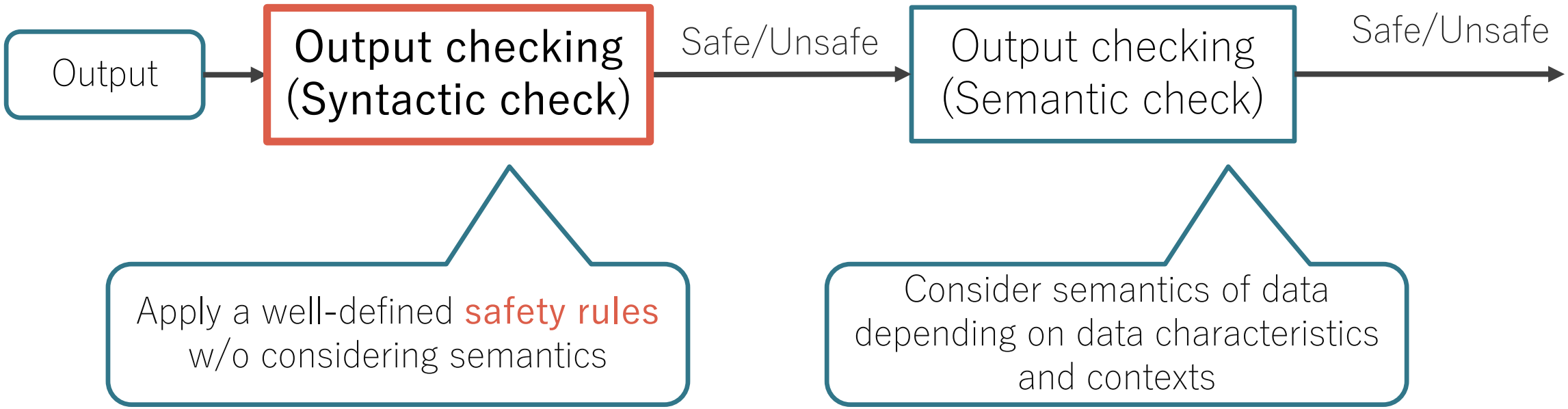
- Eurostat provides techniques relating output checking in order to share knowledge

  We follows Eurostat's guideline [BFG+]

Two steps of output checking



Two steps of output checking



Our safety rule for syntactic check

- Five principles (mainly derived from the Eurostat's guideline)
  1. Each individual value is confidential
  2. 10 units: all output must be an aggregate of at least 10 units
  3. 9 degrees of freedom: similar principle for statistics/models
  4. Group disclosure: a group of individuals should not belong to a certain group
  5. Dominance rule: the largest contributor should be smaller than 50%

- Our modifications
  - **confidentiality interval**, differencing attack, principle-based check for linear regression

Confidentiality interval

- Every primarily suppressed cell must have enough uncertainty about its value

	T1	T2	T3	Sum
L1	2*	1*	60	63
L2	50*	60*	12	122
L3	60	11	60	131
Sum	112	72	132	312

- $x_{L1,T1} + x_{L1,T2} + 60 = 63$
- $x_{L2,T1} + x_{L2,T2} + 12 = 132$
- $x_{L1,T1} + x_{L2,T1} + 60 = 112$
- $x_{L1,T2} + x_{L2,T2} + 11 = 72$

Linear programming

$$0 \leq x_{L1,T1}, x_{L1,T2} \leq 3$$

Automation

Dominance rule

- The largest contributor should be smaller than 50%

Output

	research
M2	220

ID	Job	Region	Income
A	research	M2	**100**
B	research	M2	40
C	research	M2	60
D	research	M2	20

Dominance rule

- The largest contributor should be smaller than 50%

Output

	research
M2	220

ID	Job	Region	Income
A	research	M2	**130**
B	research	M2	10
C	research	M2	60
D	research	M2	20

Motivation of automation: Dominance rule

Output

	research
M2	220

ID	Job	Region	Income
A	research	M2	100
B	research	M2	40
C	research	M2	60
D	research	M2	20

Bureau of statistics

?

Output checking

	research
M2	220

Motivation of automation: Dominance rule

Output

	research
M2	220

ID	Job	Region	Income
A	research	M2	100
B	research	M2	40
C	research	M2	60
D	research	M2	20

Bureau of statistics

Output checking

	research
M2	220

1. Re-computation

☹ It is difficult to follow analysis scenario

Motivation of automation: Dominance rule

Output

	research
M2	220

	research
M2	100

Bureau of statistics

Output checking

	research
M2	220

	research
M2	100

ID	Job	Region	Income
A	research	M2	100
B	research	M2	40
C	research	M2	60

2. Prepare additional information

☹ A lot of burden for researchers

⇒ **automation**

Motivation of automation: confidentiality interval

- suppress cells less than 10 units
- secondary suppression satisfying confidentiality interval

	T1	T2	T3	T4	Sum
L1	7*	10*	60	13*	90
L2	11*	60	12*	60	143
L3	60	11*	60	12*	143
L4	14*	60	13*		
Sum	92	141	145		

$$x_{L1,T1} + x_{L1,T2} + 60 + x_{L1,T4} = 90$$
$$\vdots$$

Linear programming

$$7 \leq x_{L1,T1} \leq 7$$

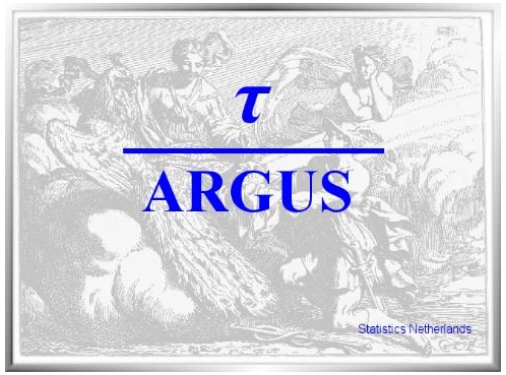
Counter-intuitive and difficult to check by hand
⇒ **automation**

Automation by using τ-Argus

- τ-Argus can suppress tabular data satisfying safety rules

Microdata /table →

metafile (format description) →



→

	T1	T2	T3	Sum
L1	*	*	60	140
L2	*	*	12	122
L3	60	11	60	131
Sum	140	121	132	395

- There are two ways to use τ-Argus
  1. Researcher uses τ-Argus

     ☹ s/he typically uses analytic tools, such as SAS, R, STATA, etc.
  2. Researcher uses analytic tools and an output checker uses τ-Argus

     ☹ a lot of burden for researchers

     have to prepare the largest contributor and metafile

Automation by tailor-maid functions in R

- Provide a tailor-made function for constructing a suppressed table in R
- Output a suppressed table satisfying safety rules
- Export additional information for an output checker to verify the safety of suppressed tables

Key functionality: suppressFT

	T1	T2	T3	Sum
L1	30	50	60	140
L2	50	60	12	122
L3	60	11	60	131
Sum	140	121	132	395

Original table

- Unit frequency
- Row/Column sum dominance
- Confidentiality interval

Security threshold parameters

Function
suppressFT

- Primary suppressions
- Secondary suppressions

Files for output checking

	T1	T2	T3	Sum
L1	*	*	60	140
L2	*	*	12	122
L3	60	11	60	131
Sum	140	121	132	395

Secondary suppressed table
and other auxiliary tables

☺ confidential interval is automatically computed

☺ an output checker can verify the safety rules

Conclusion

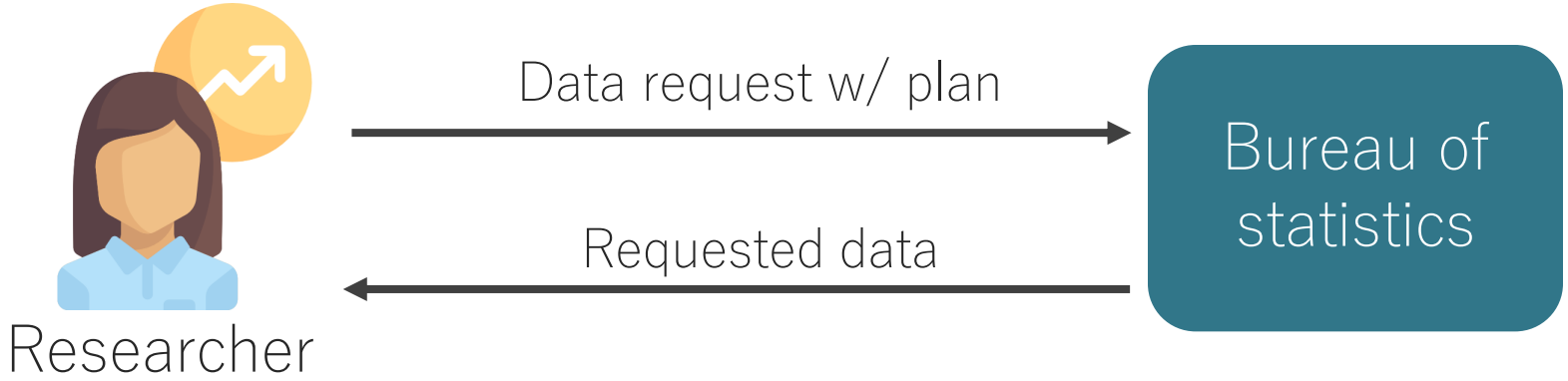
- On-site service in Japan at a trial stage
- Safety rule for output checking
  - based on Eurostat's except modifications including confidentiality interval
- Automation
  - To check dominance rules and confidentiality intervals is difficult to perform manually, which is a time-consuming and counter-intuitive task
  - Currently, τ-Argus has several issues when we adopt it into the output checking process of our onsite-use program
  - We developed a set of R functions that automatically produce safe tabular data and auxiliary files for output checking

Background: Survey data for secondary use

- Decision making based on data analysis

Current institution of secondary use in Japan

- Researchers can use microdata



- But;
  - Strict permission
    - Precise research objectives with a detailed plan
    - Long time-consuming review
  - Limited number of attributes
    Difficult to perform exploratory analysis requiring various attributes

Cons and possible solutions

Output (magnitude table)

	T1	T2	T3	Sum
L1	30	50	60	140
L2	50	60	12	122
L3	60	11	60	131
Sum	140	121	132	395

Aux. input for check

	T1	T2	T3	Sum
L1	20	20	10	140
L2	20	10	3	122
L3	21	3	25	131
Sum	140	121	132	395

...

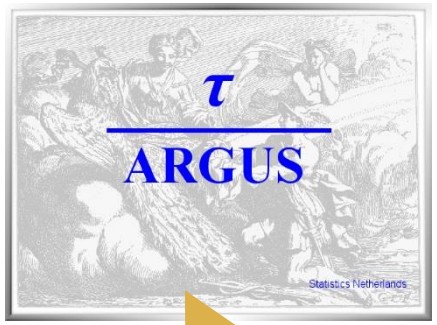
Researcher prepares all the information by itself

Specific format

⇒ some automatic conversion is possible

.tab file

metadata



difficult to use for non-experts

⇒ using batch mode

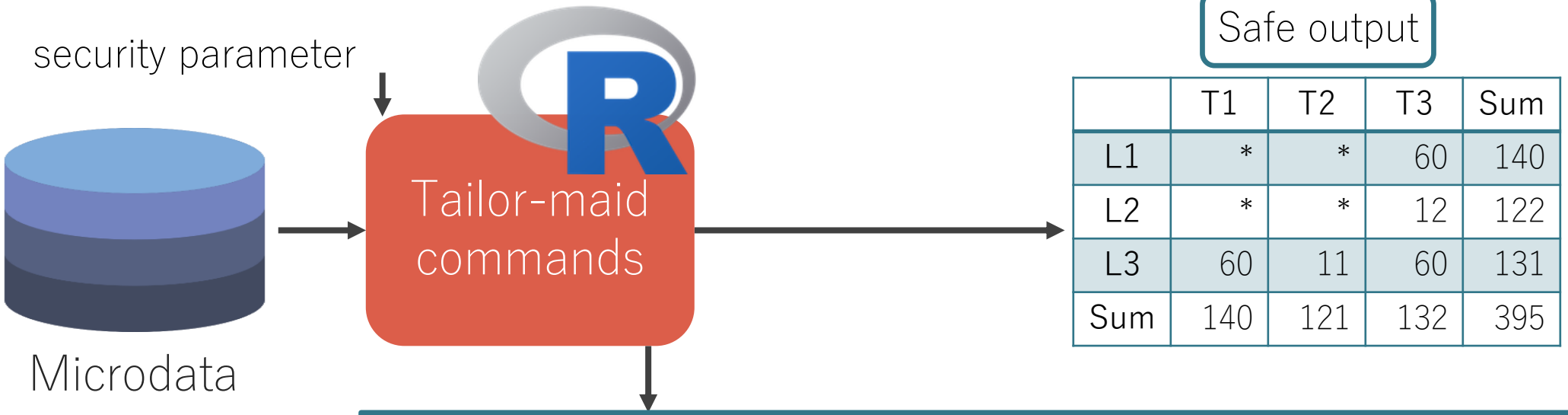
Safe output

	T1	T2	T3	Sum
L1	*	*	60	140
L2	*	*	12	122
L3	60	11	60	131
Sum	140	121	132	395

does not provide group disclosure

Tailor-maid commands in R

- substitutive commands for tabulation



Safe output

	T1	T2	T3	Sum
L1	*	*	60	140
L2	*	*	12	122
L3	60	11	60	131
Sum	140	121	132	395

Original output

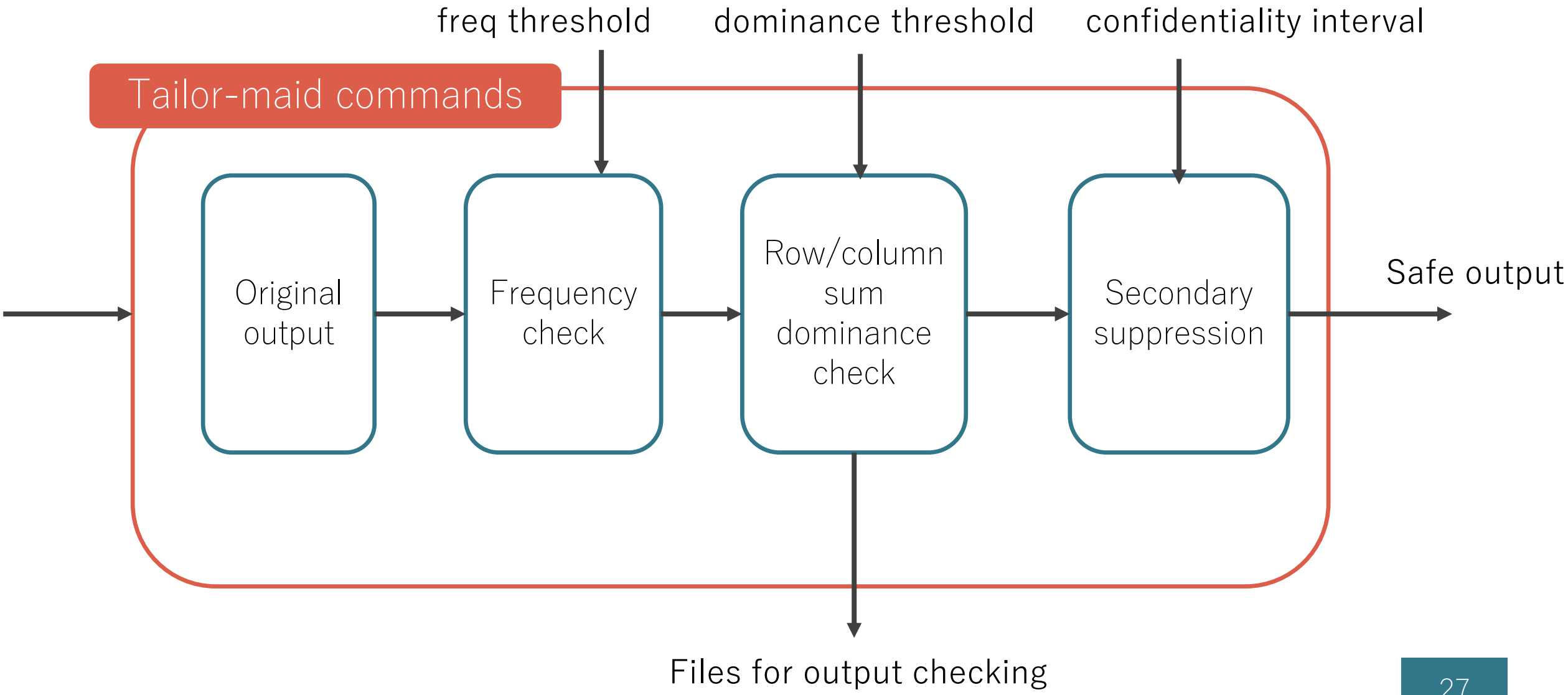
	T1	T2	T3	Sum
L1	30	50	60	140
L2	50	60	12	122
L3	60	11	60	131
Sum	140	121	132	395

Aux. input for check

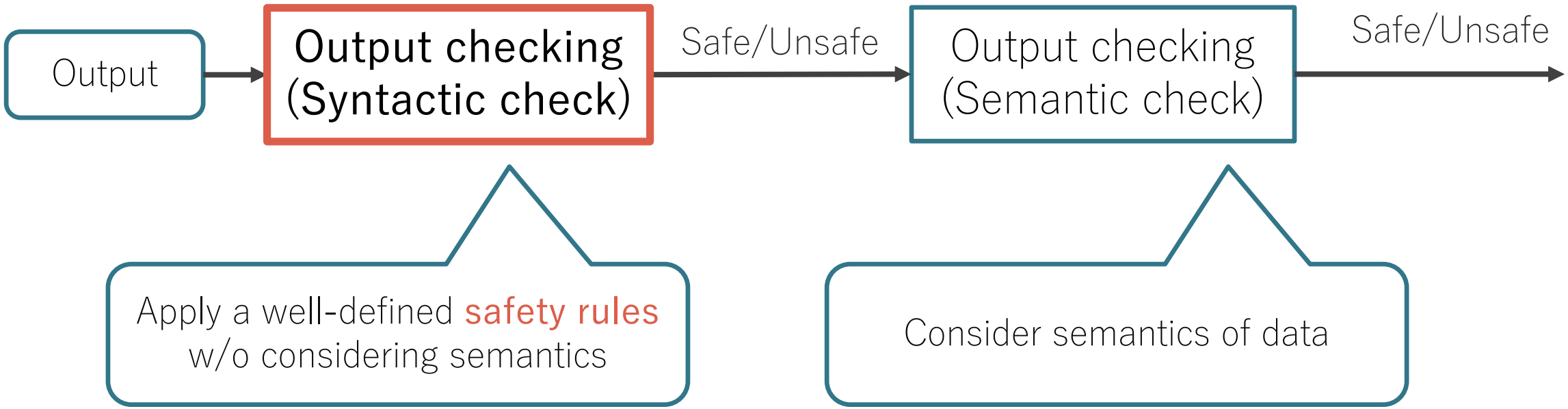
	T1	T2	T3	Sum
L1	20	20	10	140
L2	20	10	3	122
L3	21	3	25	131
Sum	140	121	132	395

…

More details of tailor-maid commands



Two steps of output checking



☺ easy for non-experts
☹ conservative and tends to be over-safety

☺ adequate check depending on data and situation
☺ difficult for non-experts