Differential privacy and noisy confidentiality concepts for European population statistics

2021 Joint UNECE/Eurostat Expert Meeting on SDC, 1 – 3 December 2021
Tabular data session

Fabian BACH
European Commission – Eurostat
Unit F2 – Population and migration

Outline

1. Intro: evolution of SDC (in population tables)
2. Noisy concepts: bottom-up and top-down
3. Risks: exploiting and massive averaging
4. Utility: (noise) tail wagging the (statistic) dog
5. Outro: the 2021 EU census picture

European Commission

Intro: evolution of SDC (in population tables)

20th century lore:

- must protect individuals
- therefore treat small counts…
- … and ensure consistency…
- … and ensure consistency…
- … and ensure consistency…

SEX \\ POB*	Total	Country	Outside
Total	42	35	7
Male	22	C	C
Female	20	C	C

* Place of birth (POB)

➔ looks easy, but is generally **neither simple nor efficient**

Intro: evolution of SDC (in population tables)

21st century state of the art:

- database reconstruction theorem (Dinur and Nissim, 2003)

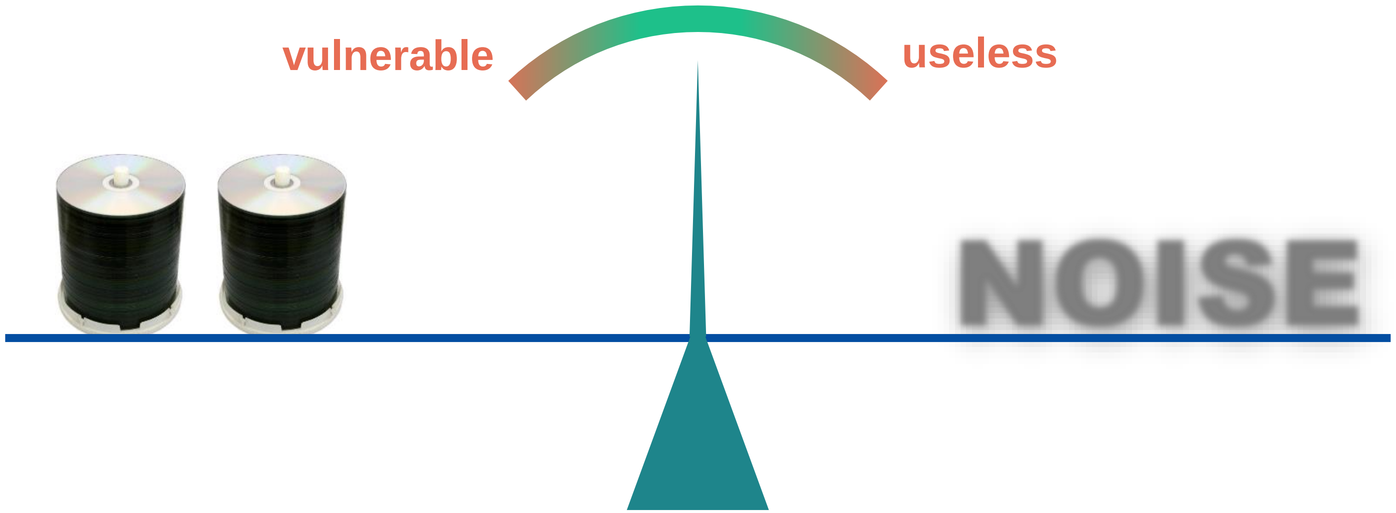
  Too many statistics, published too accurately, allow full & accurate reconstruction of all the input microdata…

  (example e.g. in U.S. Census Bureau, 2018a, 2018b)

Intro: evolution of SDC (in population tables)

21th century state of the art:

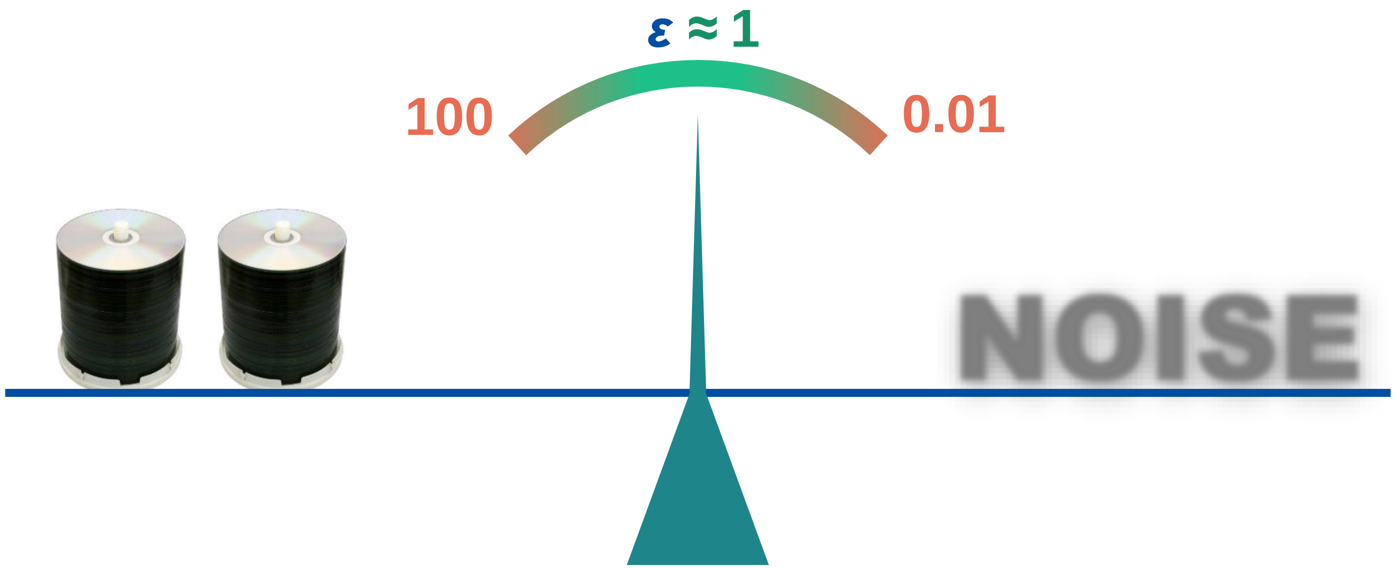
- database reconstruction theorem (Dinur and Nissim, 2003)



Noisy concepts: top-down

Differential privacy (DP) picture:

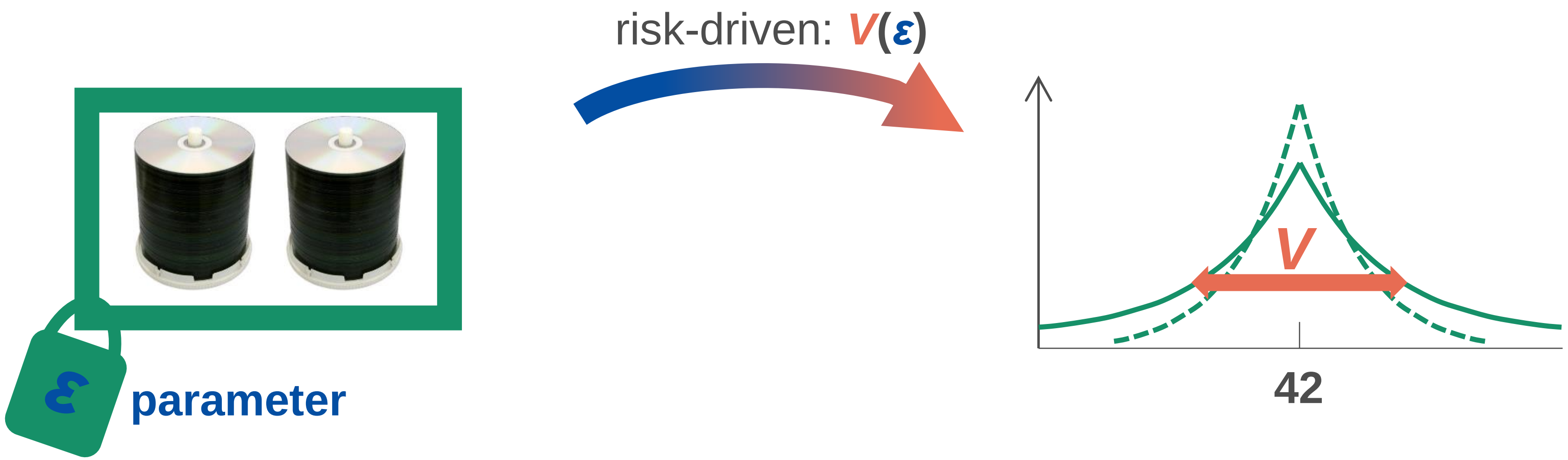
- introducing global privacy budget ε (Dwork et al., 2006)



Noisy concepts: top-down or *risk-driven*

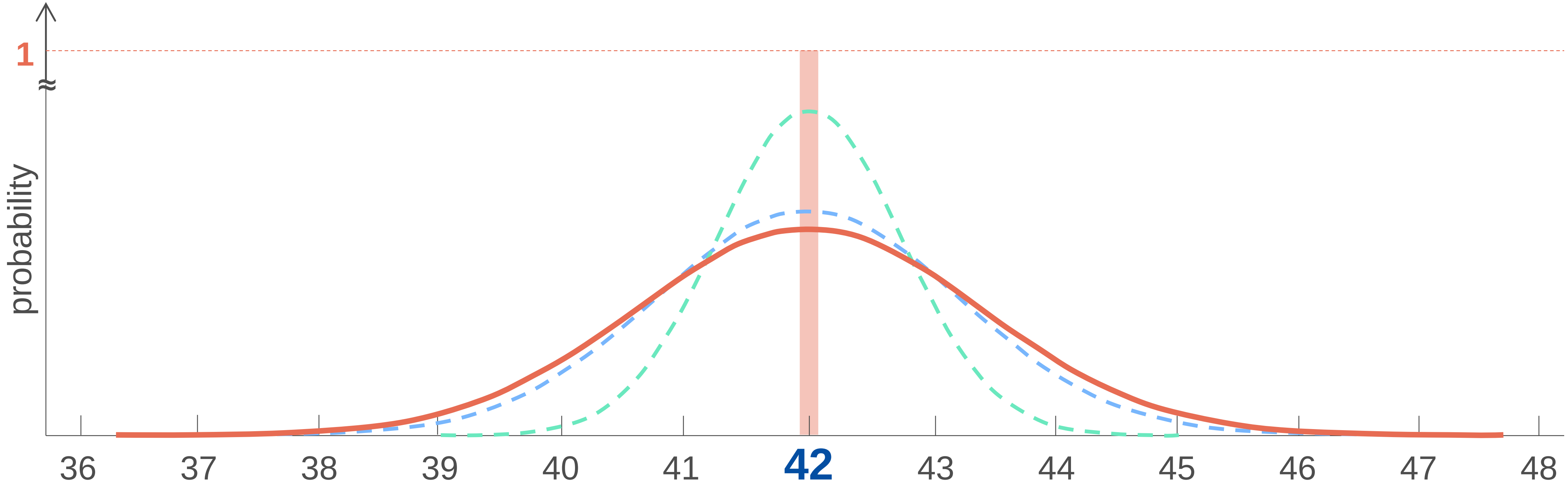
Differential privacy (DP) picture:

- introducing global privacy budget *ε* (Dwork et al., 2006)
- promise: strong global privacy guarantee … but local noise size?



Noisy concepts: bottom-up

… a closer look at **single statistic** level – e.g. total population in the area:



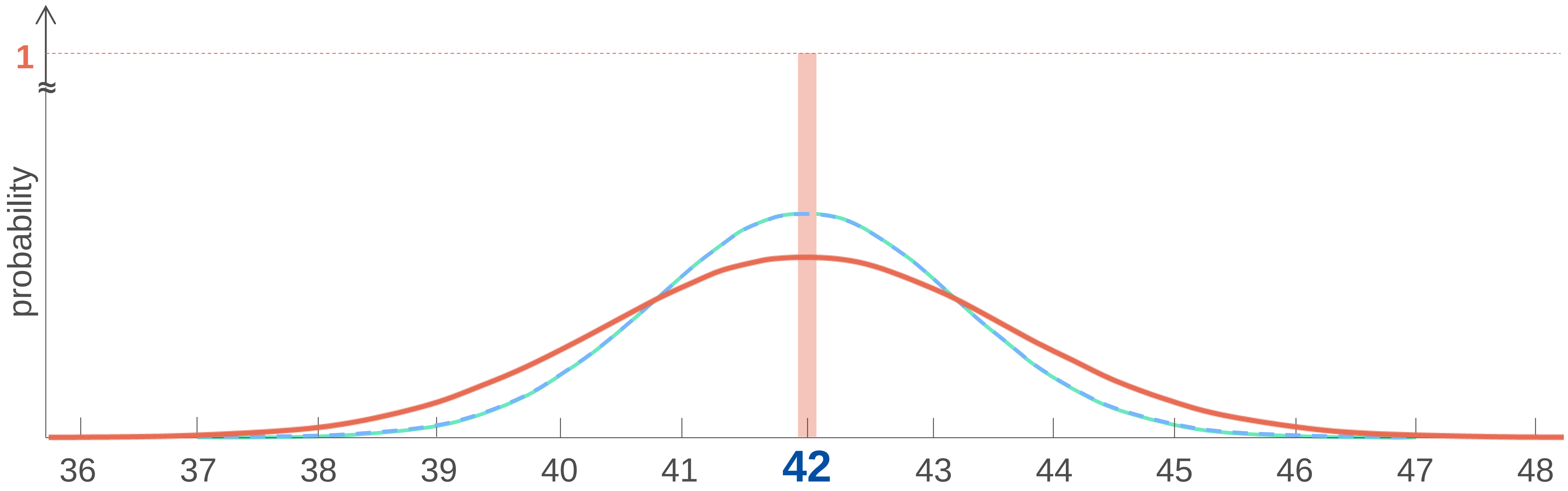
intrinsic uncertainty: ± 2

protective noise added: ± 1

total uncertainty: ± 2.2

Noisy concepts: bottom-up

… a closer look at **single statistic** level – e.g. total population in the area:



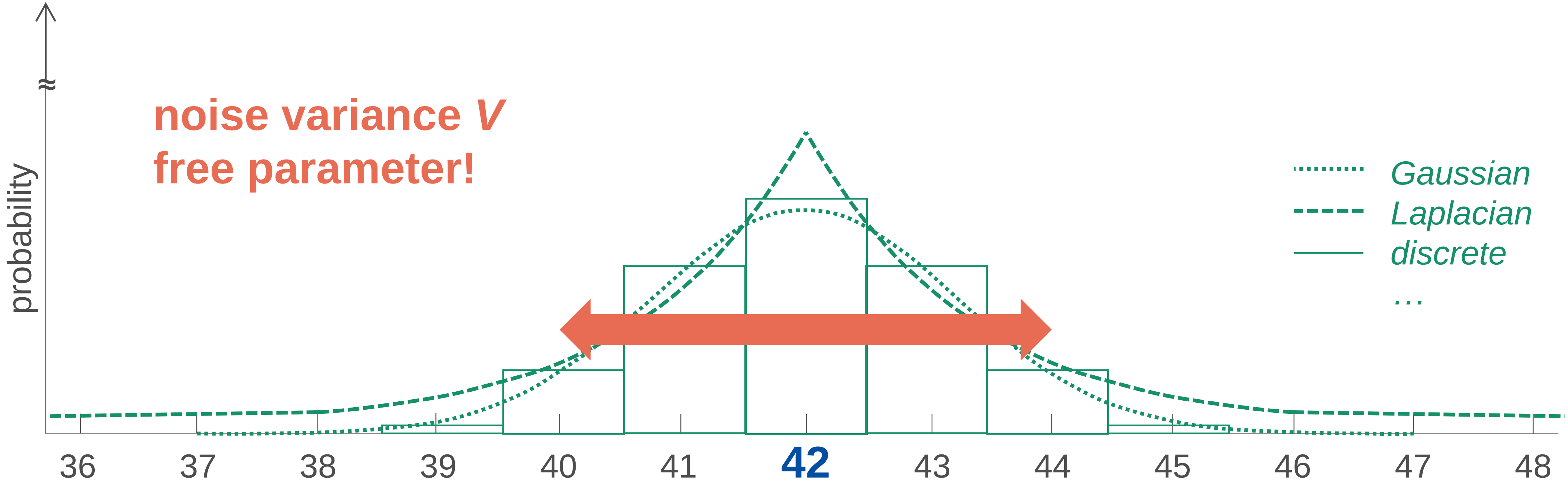
intrinsic uncertainty: ± 2

protective noise added: **± 2**

total uncertainty: ± 2.8

Noisy concepts: bottom-up or *utility-driven*

… a closer look at **single statistic** level – e.g. total population in the area:

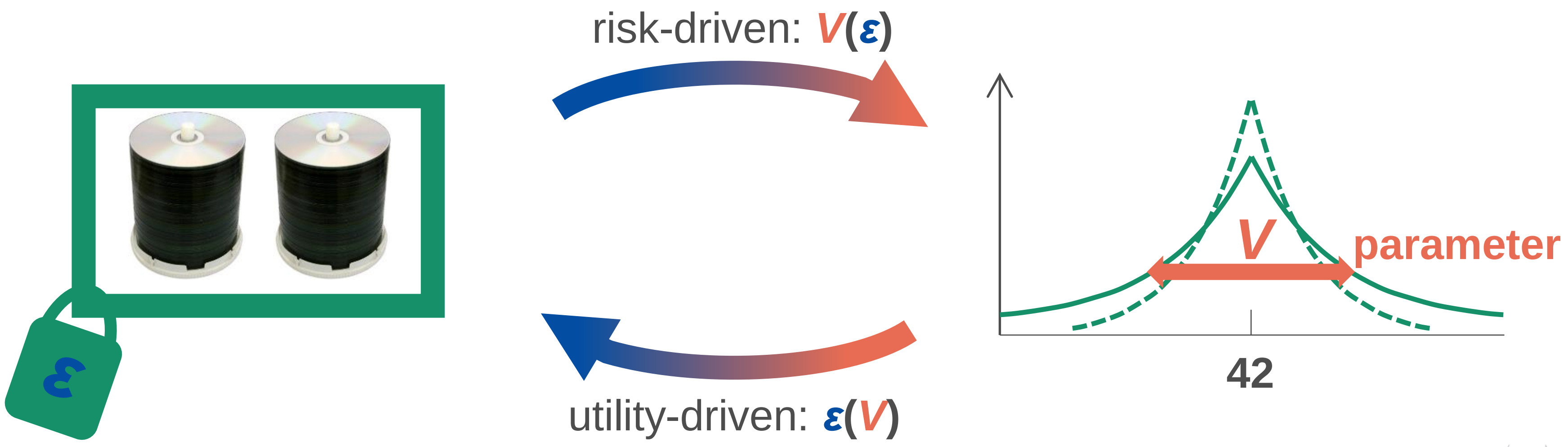


protective noise added: ± 2

Noisy concepts: bottom-up or *utility-driven*

Utility driven picture:

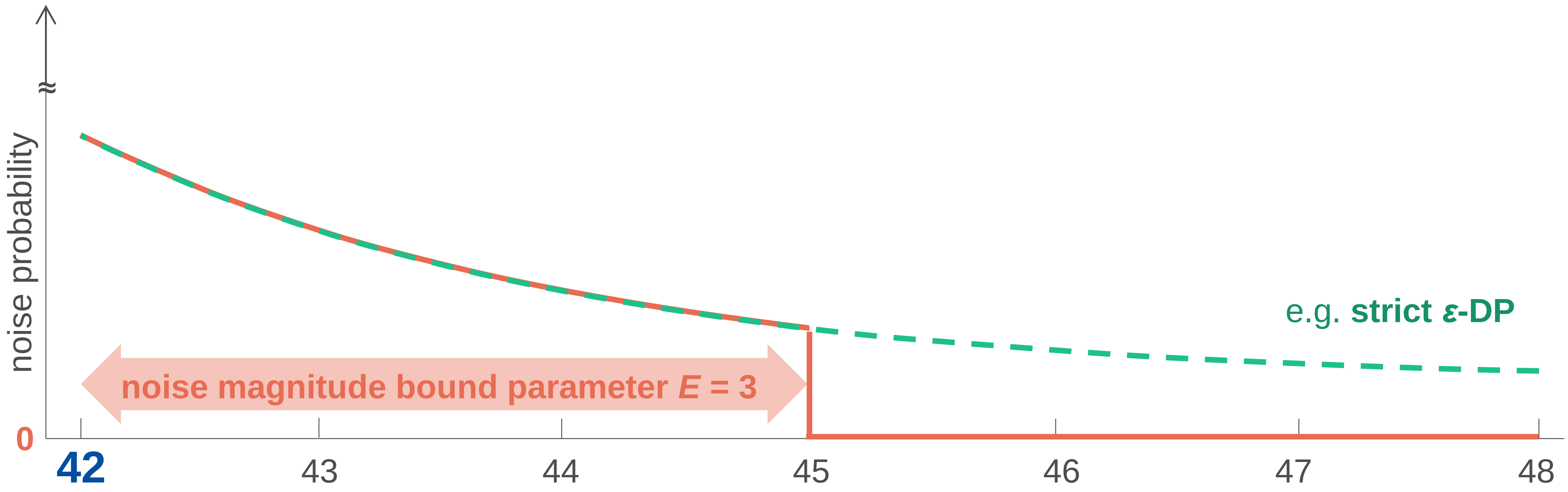
- parametrising local noise impact at single statistic level
- promise: strong utility guarantees … but global privacy level?



European Commission

Noisy concepts: bottom-up or *utility-driven*

- Noise distributions – part 2: how long is the **tail**?



e.g. **relaxed (ε,δ)-DP** or cell key method ([Marley and Leaver, 2011](#))

… but ~~**strict ε-DP**~~

Risks: exploiting table constraints

- Now would you bet all your money on a guess for the true count of the …
  - ❑ … total population?
  - ☑ … country-born males (= 17)
  - ❑ … total females?
  - ❑ … total foreign-born?

SEX \\ POB	Total	Country	Outside
Total	42	37 = 35+2	7
Male	23	15 = 17-2	4
Female	21	16 = 18-2	3

each count with noise variance $V = 1$
and noise bound $E = 2$

- How often does this happen?

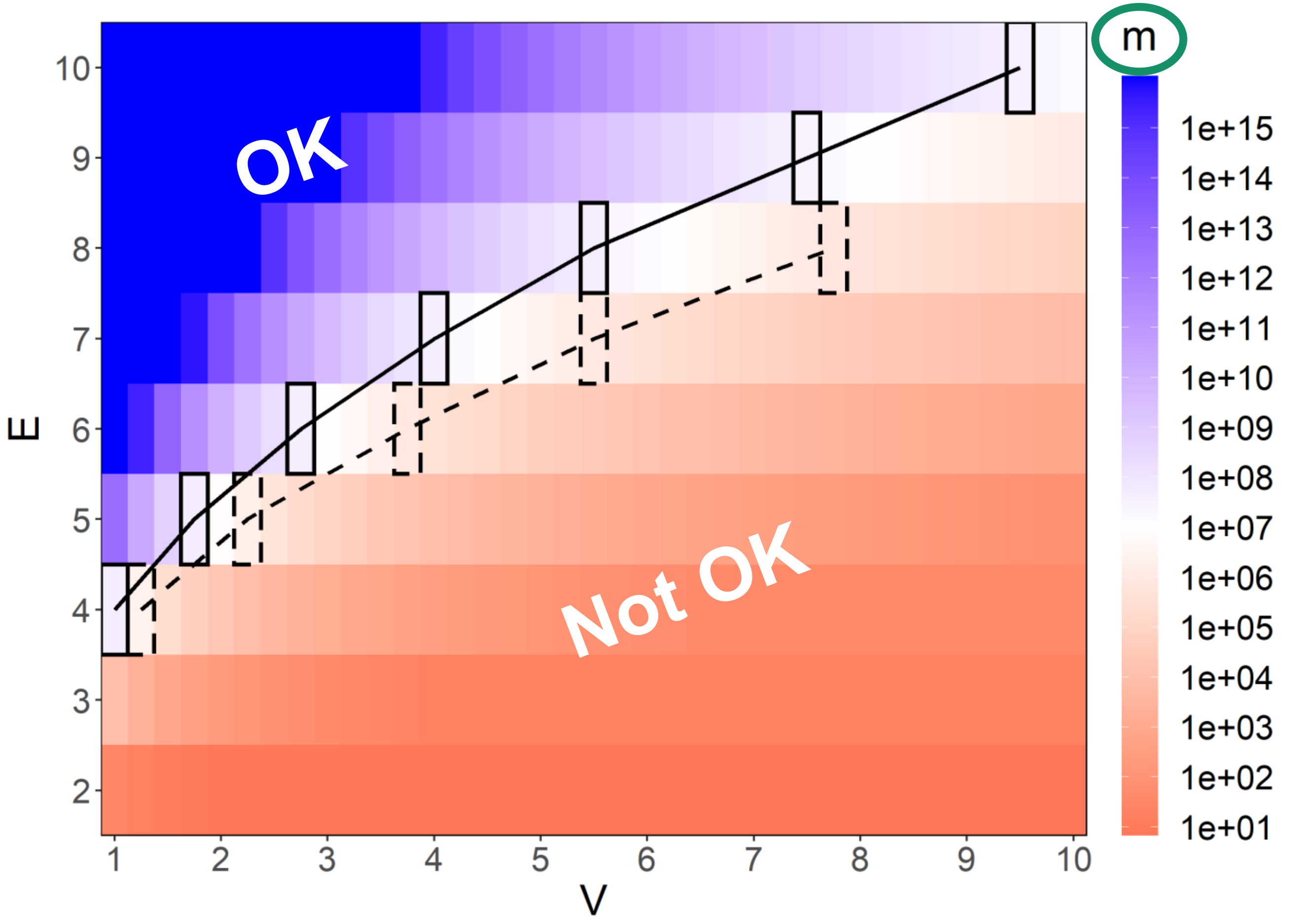
# of constraint ***n***-tuples in output x Pr(noise = $\pm E$)^{n}

fixed by output tables

fixed by noise parameters V and E

Risks: exploiting table constraints

➔ Knowing the full output, the risk can be quantified systematically – e.g. for the 2021 EU census output:



m: number of 3-tuples needed in output to get ca. one *E*-disclosive noise pattern

black boxes showing where *m* exceeds the number of available 3-tuples for Malta (dashed) and Germany (solid)

Risks: massive averaging

- How many independent observations t of “total population” are in this table?

  - ❑ $t = 1$
  - ❑ $t = 2$
  - ❑ $t = 3$
  - ✓ $t = 4$

SEX \\ POB	Total	Country	Outside
Total	**42**	**37** +	**7**
Male	**23** +	15 + +	4 +
Female	**21**	16 +	3

each count with noise variance $V = 1$

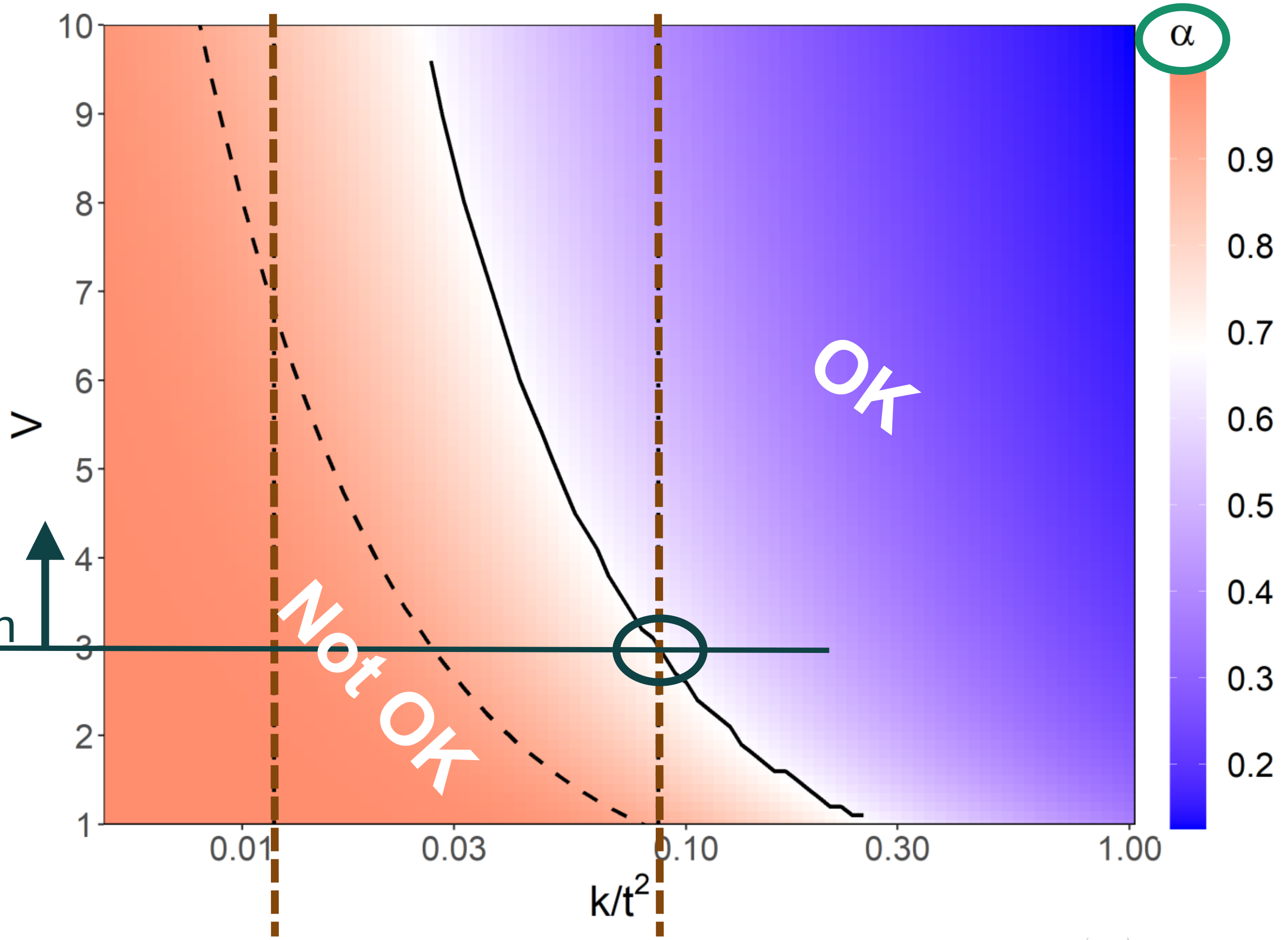
- average variance: $\boxed{\bar{V} = \frac{k}{t^2} V} = \frac{9}{4^2} 1 = 0.\bar{5}$

fixed by output tables

noise parameter

Risks: massive averaging

➔ Knowing the full output, also this risk can be quantified systematically – e.g. for the 2021 EU census output:



intersection of α = 68% contour with smallest k/t^2 value (with SPSN)

α: c.l. of obtaining correct rounded integer count after averaging

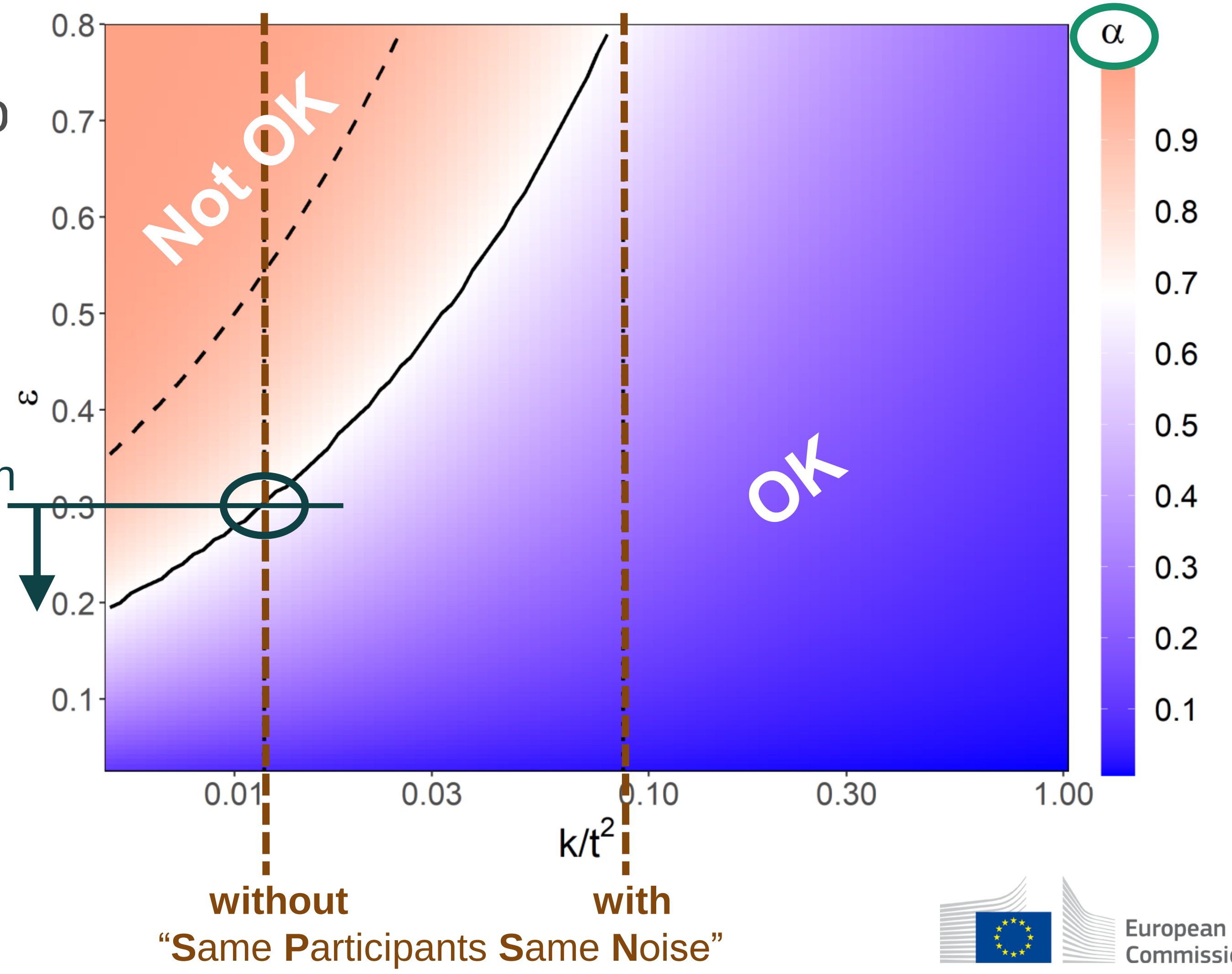
without **with**

"**S**ame **P**articipants **S**ame **N**oise"

Risks: massive averaging – *DP picture*

➔ Knowing the full output, also this risk can be quantified systematically – e.g. for the 2021 EU census output:

intersection of α = 68% contour with smallest k/t^2 value (without SPSN)

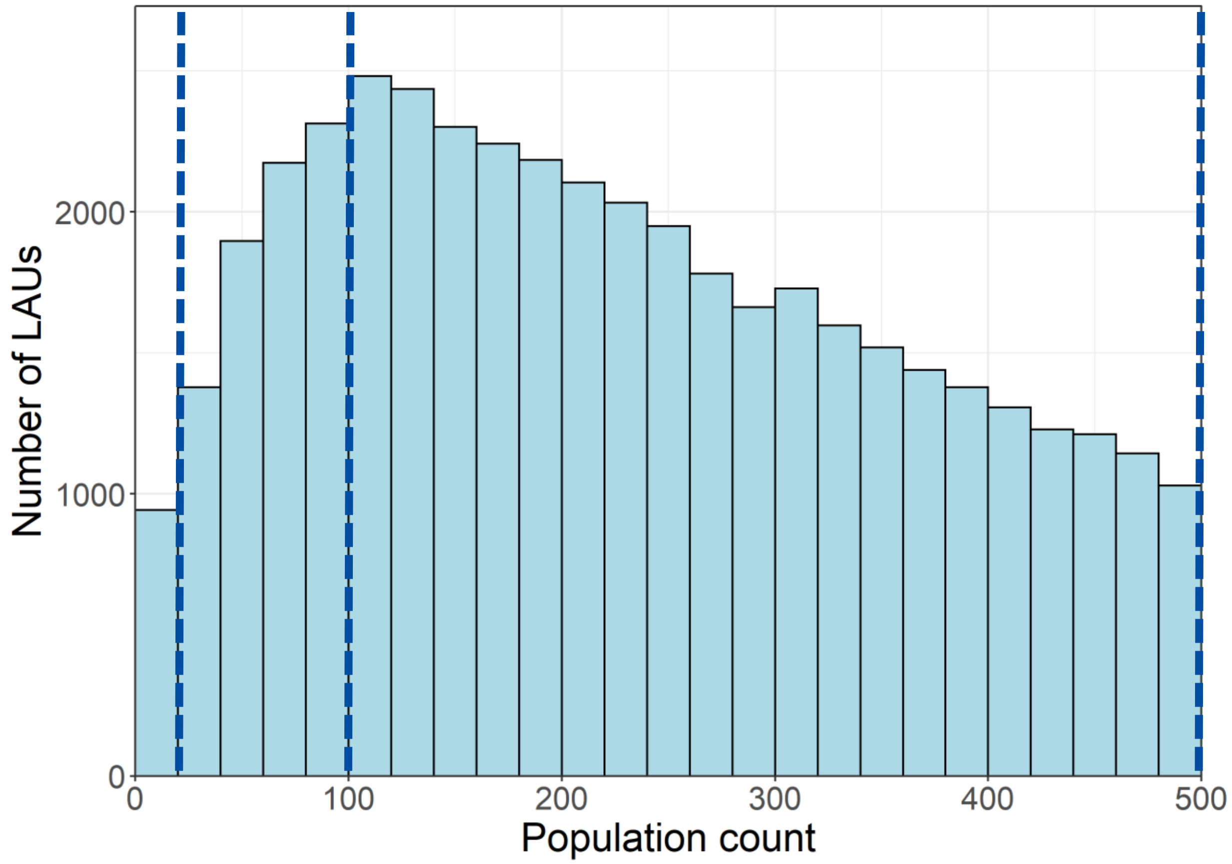


α: c.l. of obtaining correct rounded integer count after averaging

Utility: (noise) tail wagging the (statistic) dog

- 2021 EU census: ca. 110 000 **L**ocal **A**dministrative **U**nits (~ municipalities), of which
  - 43 395 with <500 people
  - 8 502 with <100 people
  - 866 with <20 people
- Could we accept here e.g. Pr(|noise|>100) = 0.1% or more?

❑ **Yes** ☑ **No**

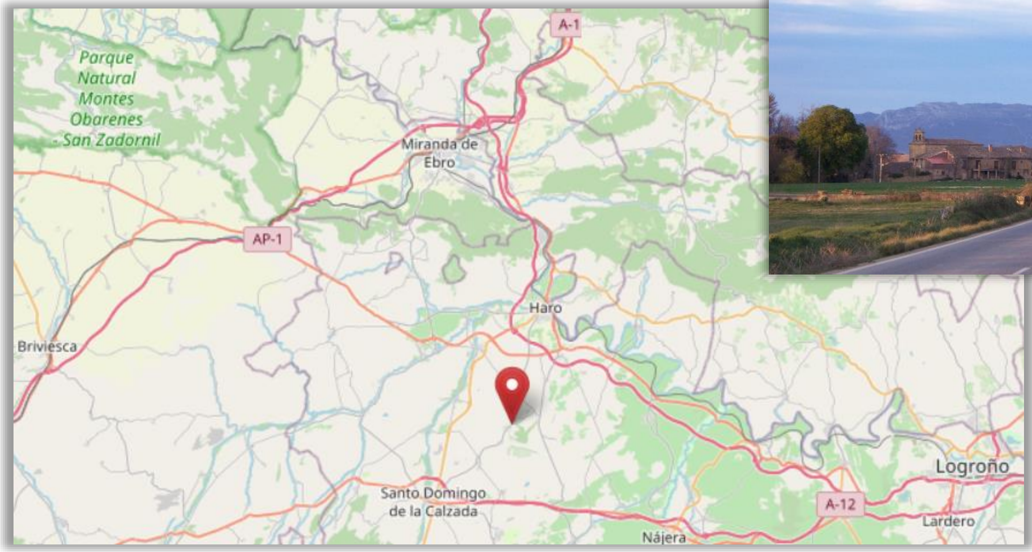


Utility: (noise) tail wagging the (statistic) dog

- mainly a problem of **unbounded noise**

  Recall: *Noise magnitude bound parameter* ***E****, "cutting off" the tail, is* ***forbidden*** *in strict ε-DP*

- E.g. 2020 test setup of [U.S. Census Bureau (2019)]() with moderate tabular $\varepsilon = 0.1$



source: OpenStreetMap

source: Wikipedia

	2011 census	strict ε-DP
Total	**30**	**-17**
Male	20	-1
Female	15	-9

Cidamón, La Rioja, Spain
ES230_26048

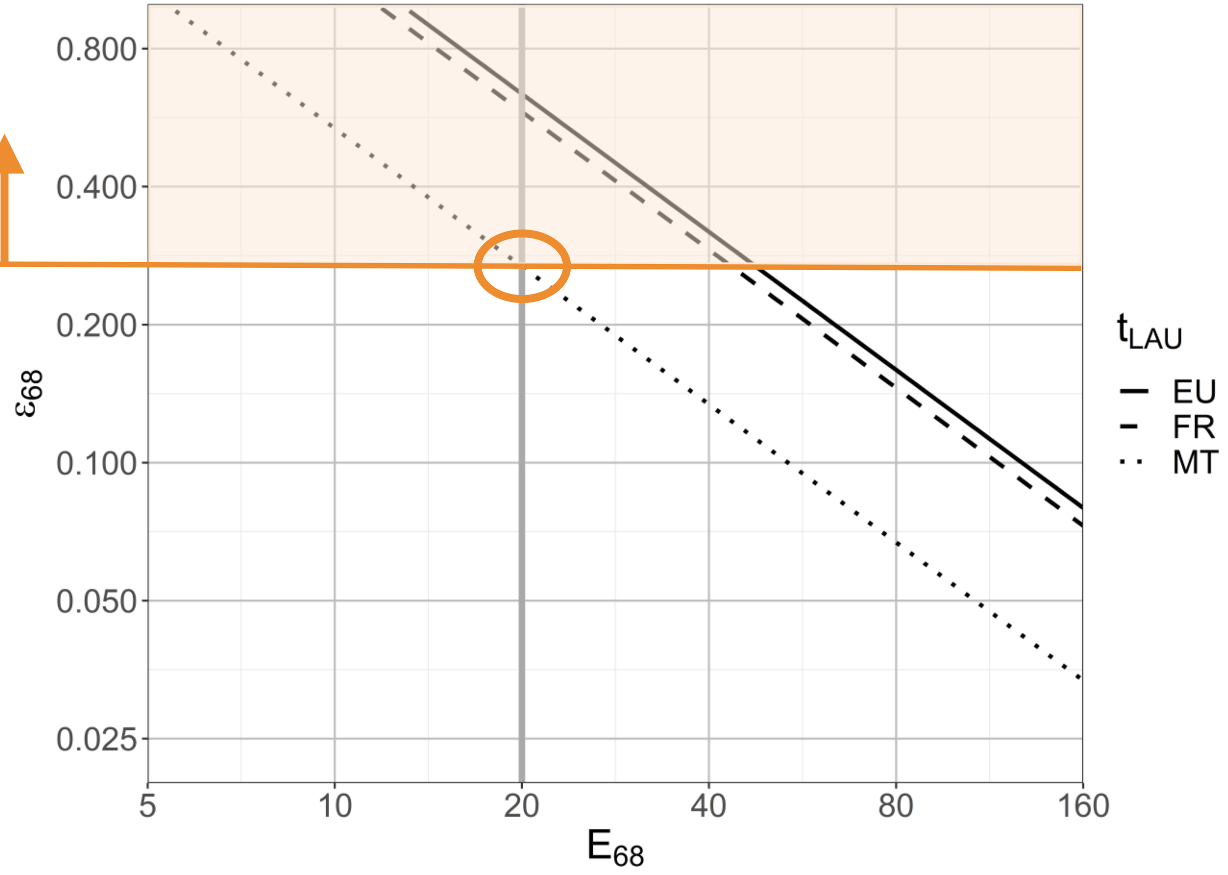
European Commission

Utility: (noise) tail wagging the (statistic) dog

- mainly a problem of **unbounded noise**

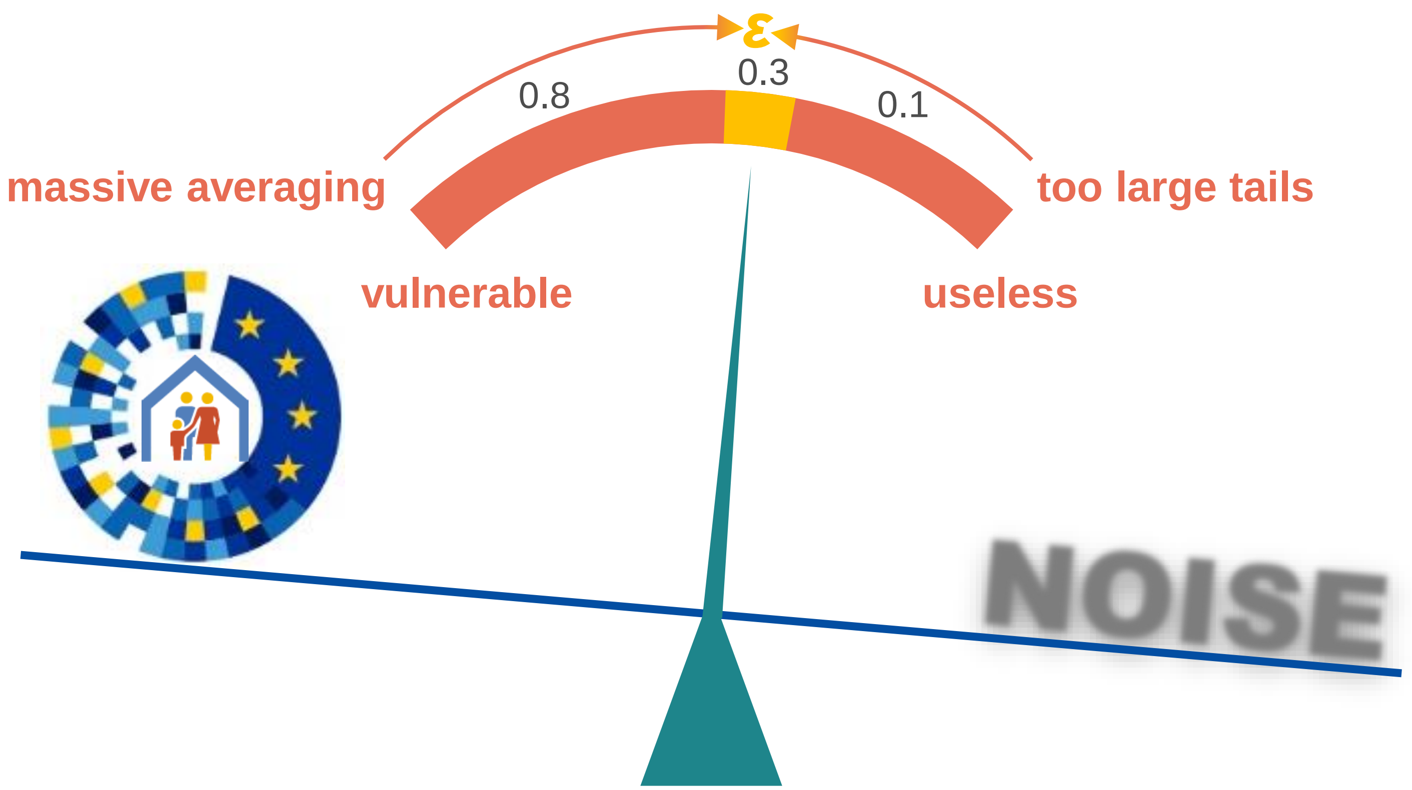
intersection of α = 68% contour for Malta with *E* = 20

α = 68% contours of getting ca. one LAU count noise > *E* for Malta (dotted), France (dashed), whole EU (solid)



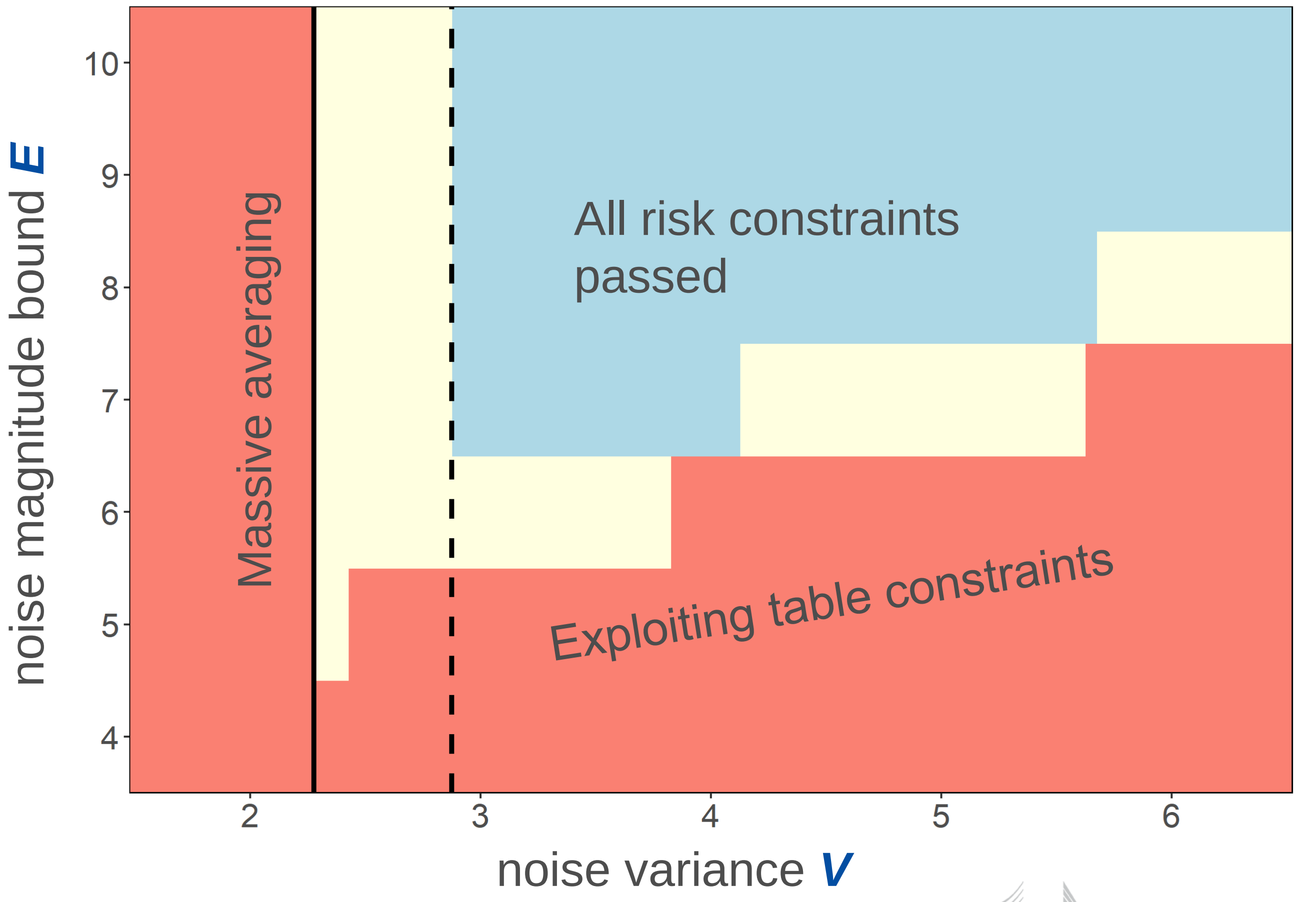
Outro: the 2021 EU census picture

- risk + utility constraints on tabular *ε* for whole 2021 EU census output



Outro: the 2021 EU census picture

- whole 2021 EU census output
- risk constraints on bottom-up parameter space ***V* – *E***
- utility controlled directly by ***V*** and ***E*** (utility-driven)
- e.g. cell key method recommended for 2021 EU census ([ESSnet, 2017, 2019](#))



Thank you

Key references (1)

Ashgar and Kaafar (2019)	H. J. Ashgar, D. Kaafar, *Averaging Attacks on Bounded Noise-based Disclosure Control Algorithms* (Proceedings on Privacy Enhancing Technologies; 2020 (2))
Dinur and Nissim (2003)	I. Dinur, K. Nissim, *Revealing Information while Preserving Privacy* (PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems)
Dwork et al. (2006)	C. Dwork, F. McSherry, K. Nissim, A. Smith, *Calibrating Noise to Sensitivity in Private Data Analysis* (Journal of Privacy and Confidentiality 7 (3):17-51; 2017)
ESSnet (2017)	Antal, L. et al., *Harmonised protection of Census data* (Centre of Excellence on Statistical Disclosure Control, Eurostat CROS portal, 2017)
ESSnet (2019)	De Wolf, P.-P. et al., *Perturbative confidentiality methods* (Centre of Excellence on Statistical Disclosure Control, Eurostat CROS portal, 2019 and github.com/sdcTools)
Marley and Leaver (2011)	J. K. Marley, V. L. Leaver, *A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis* (ISI Proc. 58th World Statistical Congress, 2011, Dublin (Session IPS060))
Petti and Flaxman (2019)	S. Petti, A. Flaxman, A. (2019), *Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff* (Gates Open Research 2020, 3:1722)
Rinott et al. (2018)	Y. Rinott, C. M. O'Keefe, N. Shlomo, C. J. Skinner, *Confidentiality and differential privacy in the dissemination of frequency tables* (Statistical Science, 33(3):358–385; 2018)

European Commission

Key references (2)

Ruggles et al. (2018)	S. Ruggles et al., *Differential Privacy and Census Data: Implications for Social and Economic Research* (AEA Papers and Proceedings, vol. 109, May 2019)
Santos-Lozada et al. (2020)	A. R. Santos-Lozada, J. T. Howard, A. M. Verdery, *How differential privacy will affect our understanding of health disparities in the United States* (PNAS June 16, 2020 117 (24))
Thompson et al. (2013)	G. Thompson, S. Broadfoot, D. Elazar, *Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics* (UNECE Work Session SDC, 2013)
U.S. Census Bureau (2018a)	S. L. Garfinkel, J. M. Abowd, C. Martindale, *Understanding Database Reconstruction Attacks on Public Data* (ACMQueue, Vol. 16, No. 5 (Sep/Oct 2018): 28-53)
U.S. Census Bureau (2018b)	J. M. Abowd, *Staring-Down the Database Reconstruction Theorem* (Joint Statistical Meetings, Vancouver, BC, Canada, July 30, 2018)
U.S. Census Bureau (2019)	L. Garfinkel, *Deploying Differential Privacy for the 2020 Census of Population and Housing* (Joint Statistical Meetings, US Census Bureau, Washington, DC, 2019)

European Commission