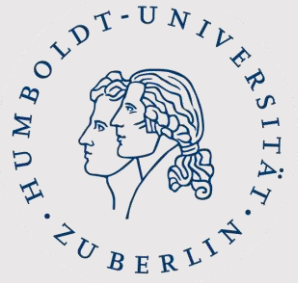
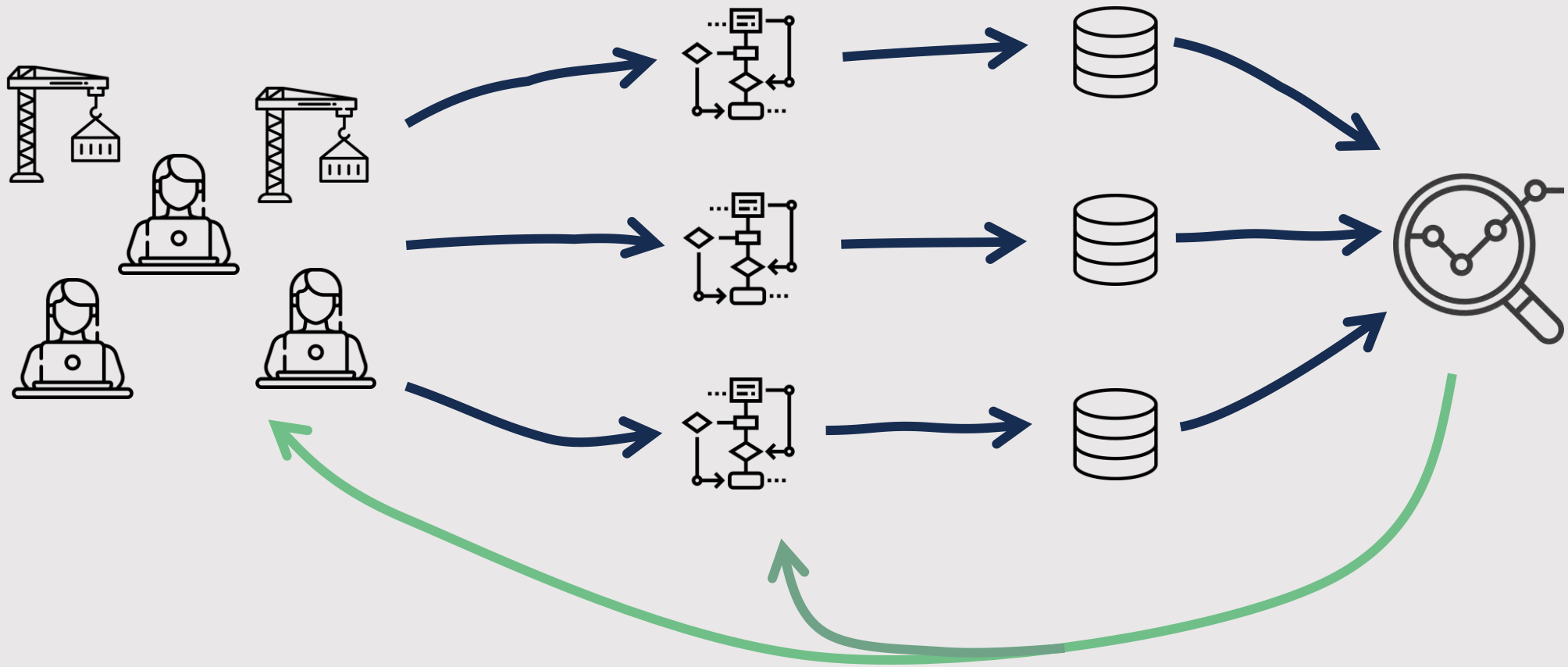
Addressing the Log Representativeness Problem using Species Discovery

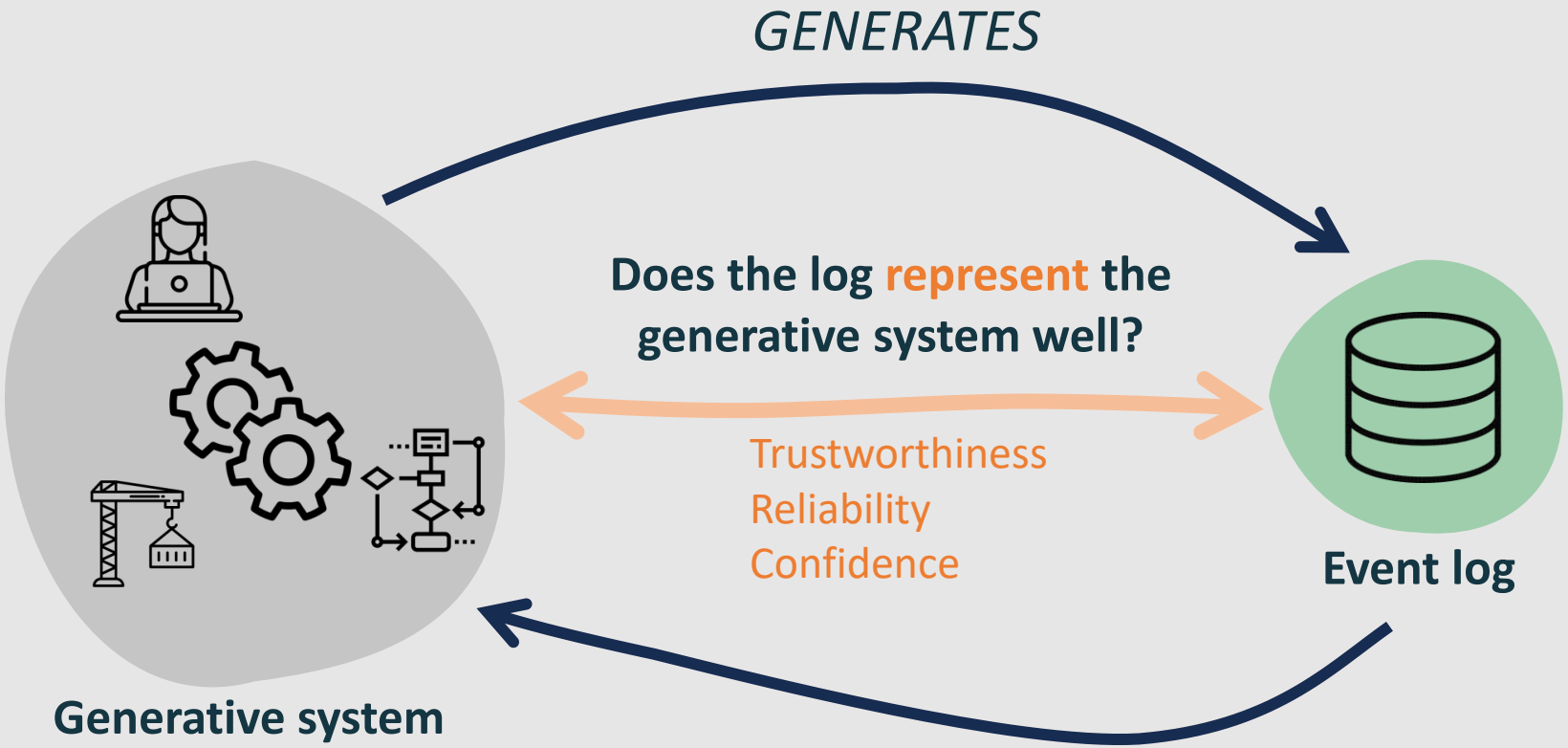
Martin Kabierski, Markus Richter, Matthias Weidlich



Process Analysis in the Wild



The Representativeness Problem



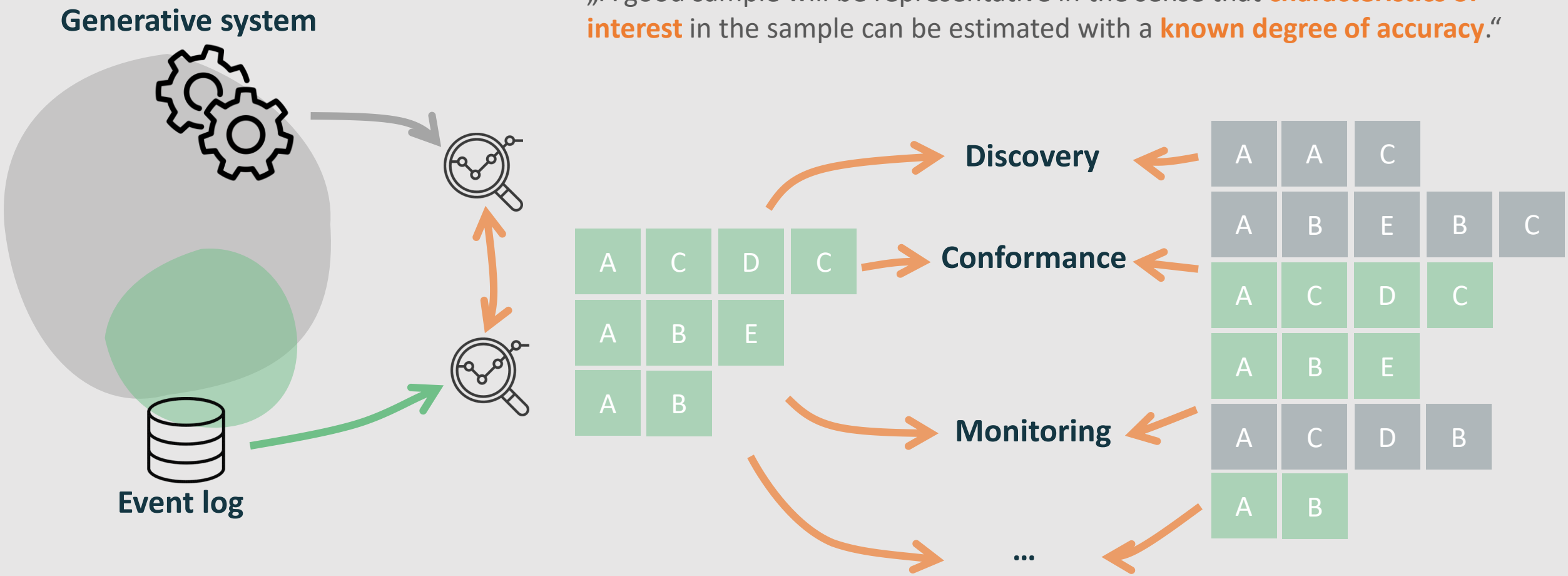
is a sample of generative system

- Incompleteness
- Noise
- Erroneous logging
- Process drifts
- …

The many Faces of Representativeness

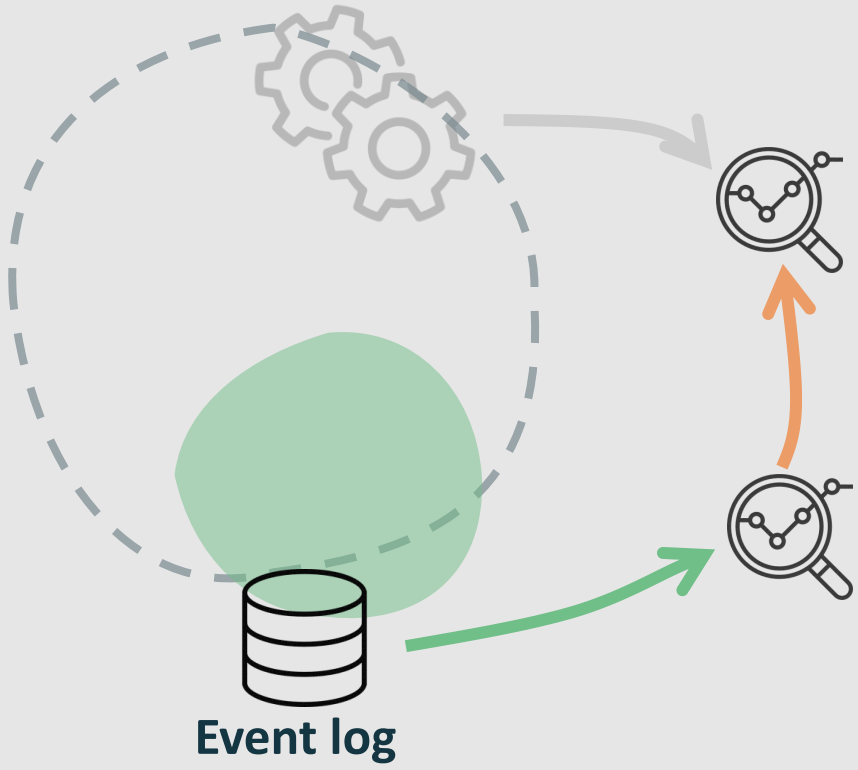
Representativeness [Sampling: Design and Analysis, Second Edition. Sharon L. Lohr. 2010]

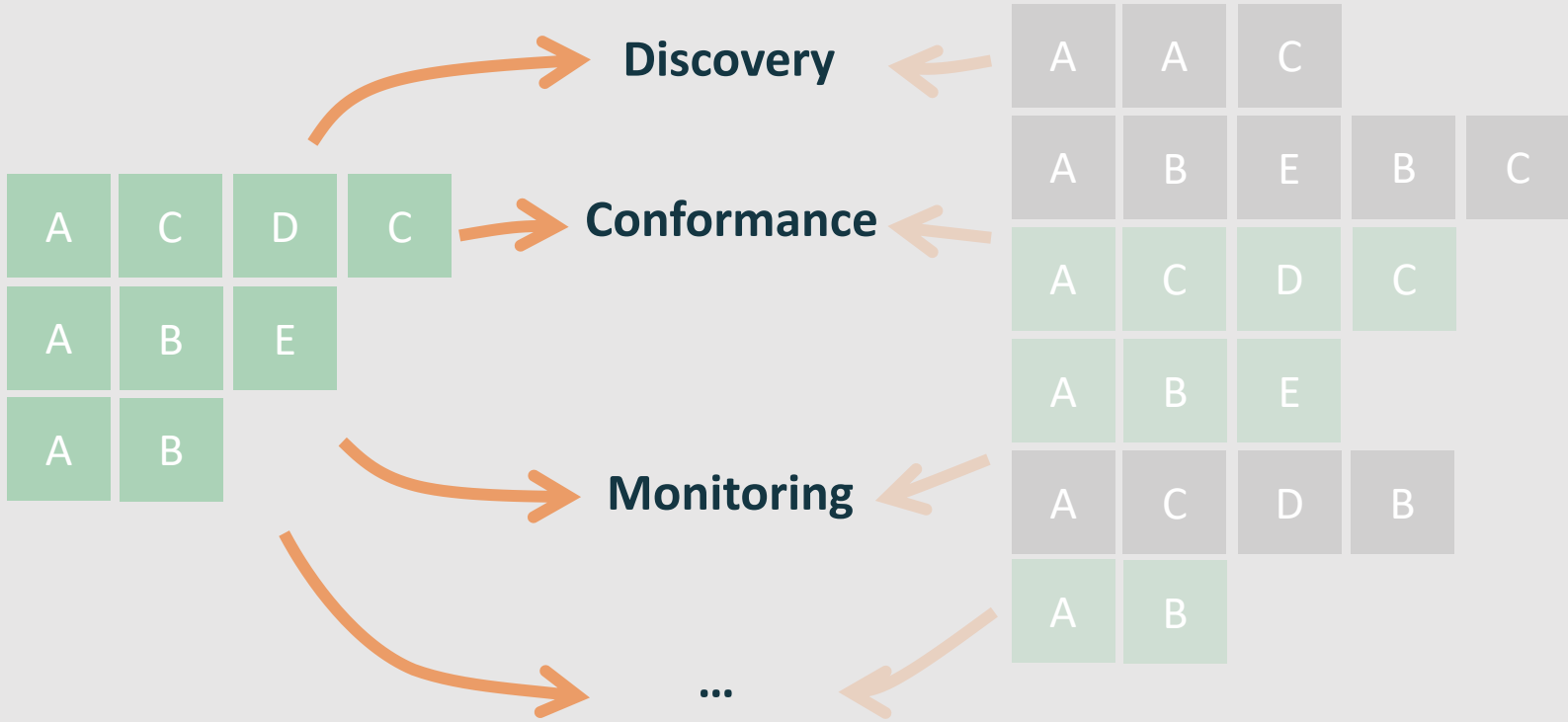
„A good sample will be representative in the sense that **characteristics of interest** in the sample can be estimated with a **known degree of accuracy**.“



The many Faces of Representativeness

Representativeness [Sampling: Design and Analysis, Second Edition. Sharon L. Lohr. 2010]
„A good sample will be representative in the sense that **characteristics of interest** in the sample can be estimated with a **known degree of accuracy**.“



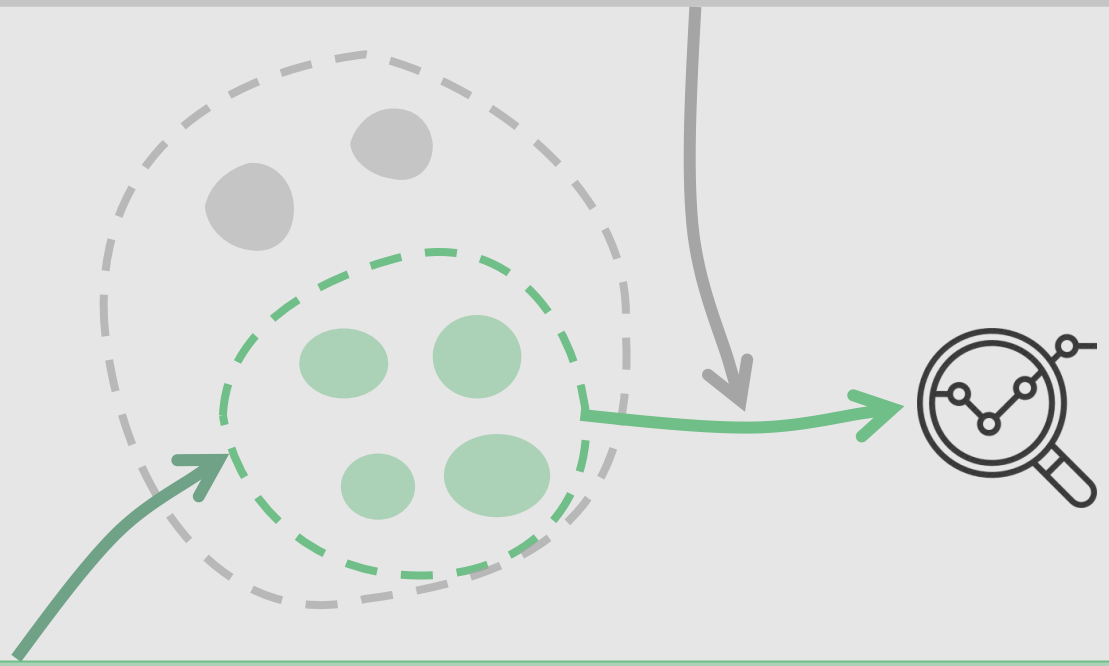


The many Faces of Representativeness

Functions on Event Logs

Estimating Function Convergence based on statistical tests

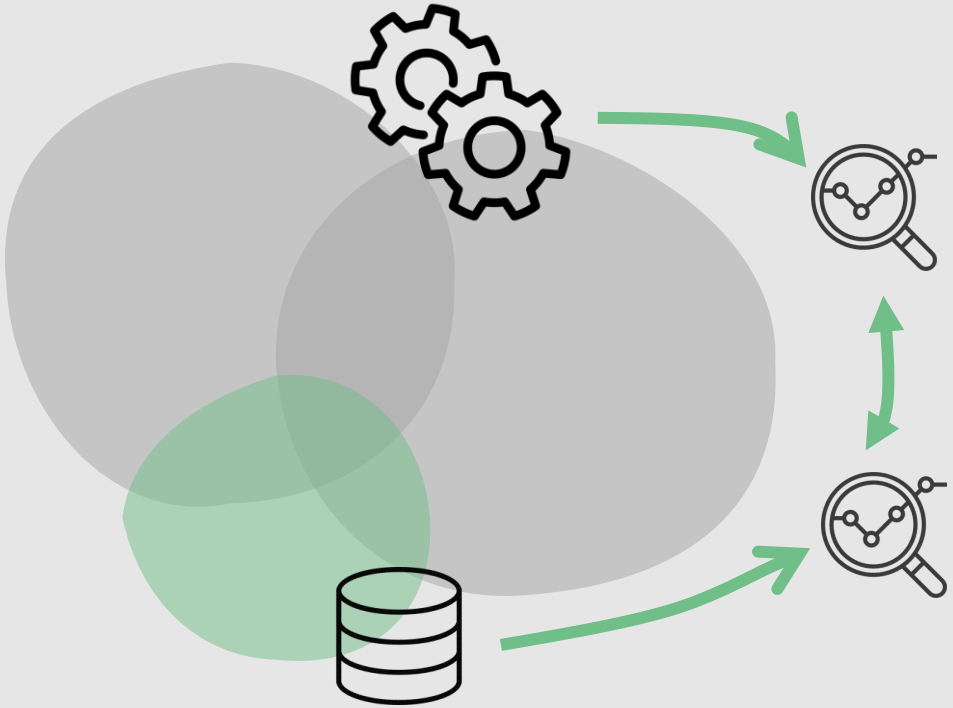
- Bauer, Senderovich, Gal, Grunske, Weidlich, How much Event Data is Enough? A statistical framework for process discovery.,CAiSE 2018
- Bauer, van der Aa, Weidlich, Estimating Process Conformance by Trace Sampling and Result Approximation., BPM 2019
- Bauer, van der Aa, Weidlich, Sampling and approximation techniques for efficient process conformance checking., Information Systems, 2022

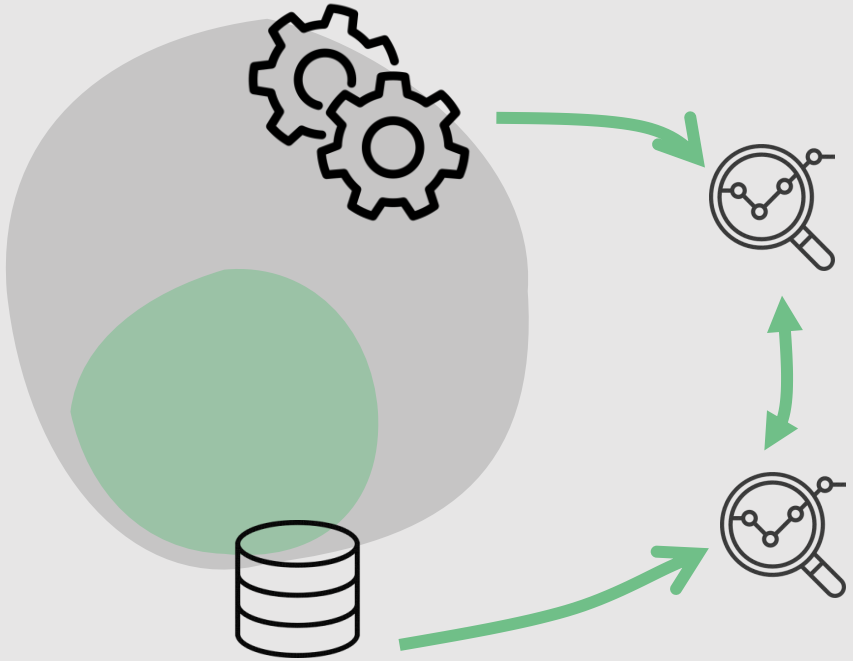


Completeness

Does the log contain all values present in the generative system?

Assumptions: no Drifts & no Errors in the Log





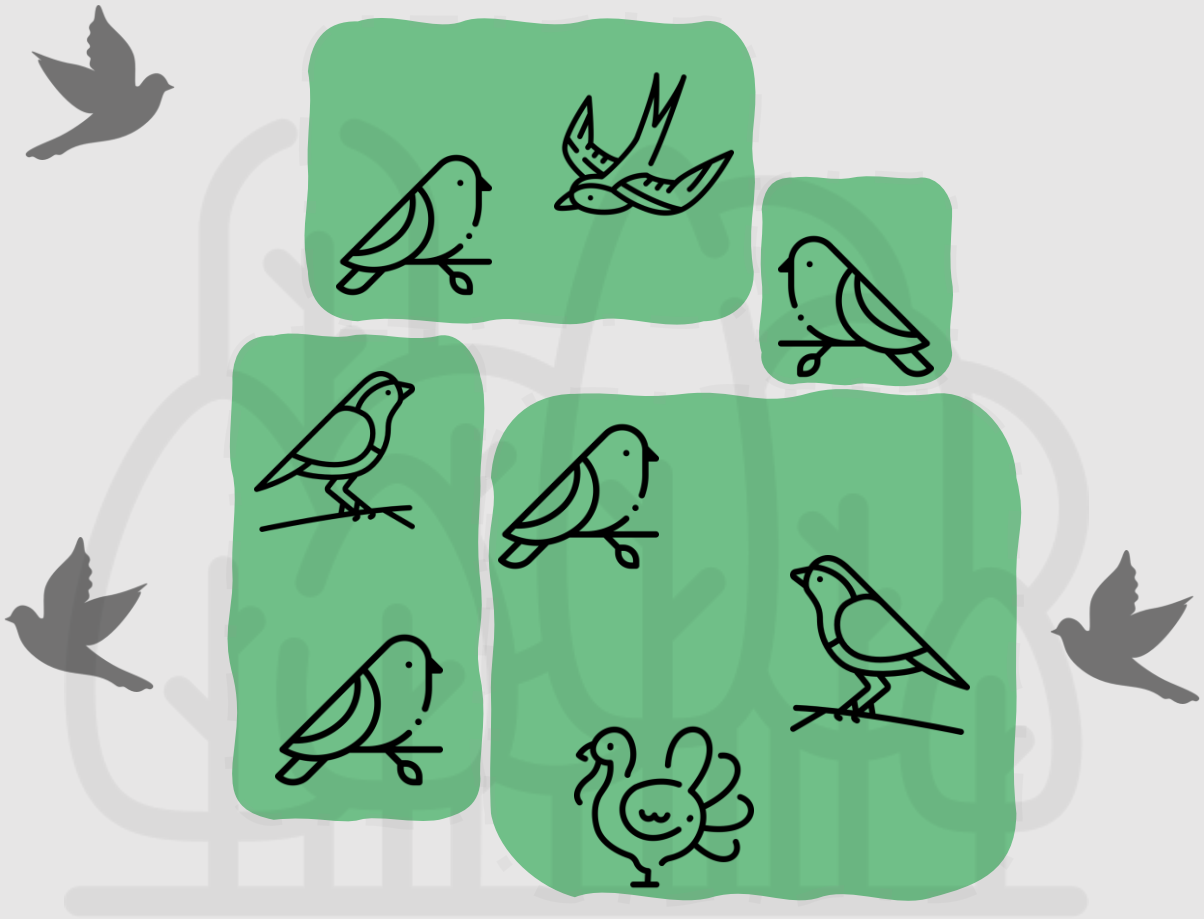
A different View: Biodiversity Analysis

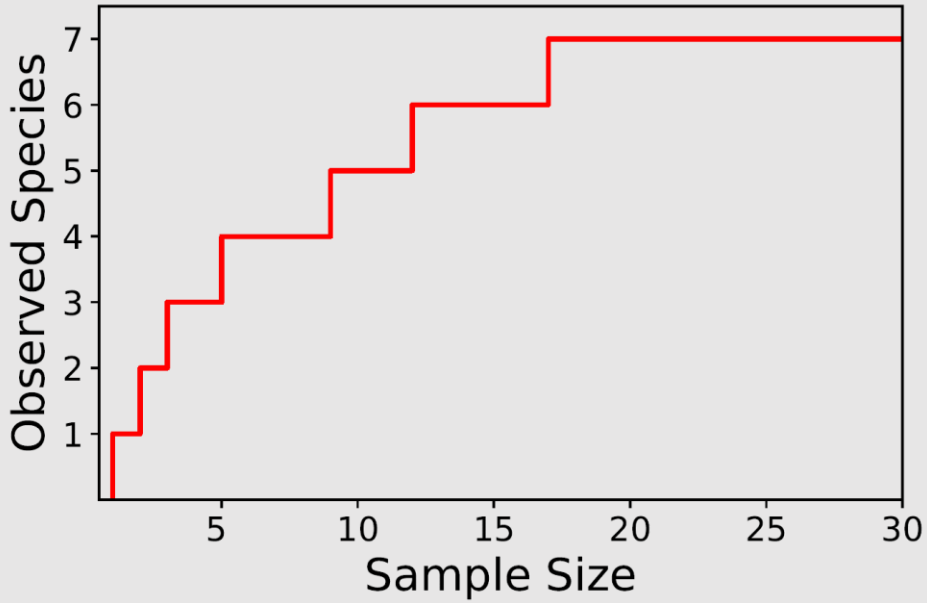
Bird Populations Are in Meltdown
[wired.com, 20.06.2023]

With an estimated 5.5 million species, insects are the most diverse group of animals on the planet. More than one million have been named by scientists — and many more have yet to be discovered.
[Florida Museum]

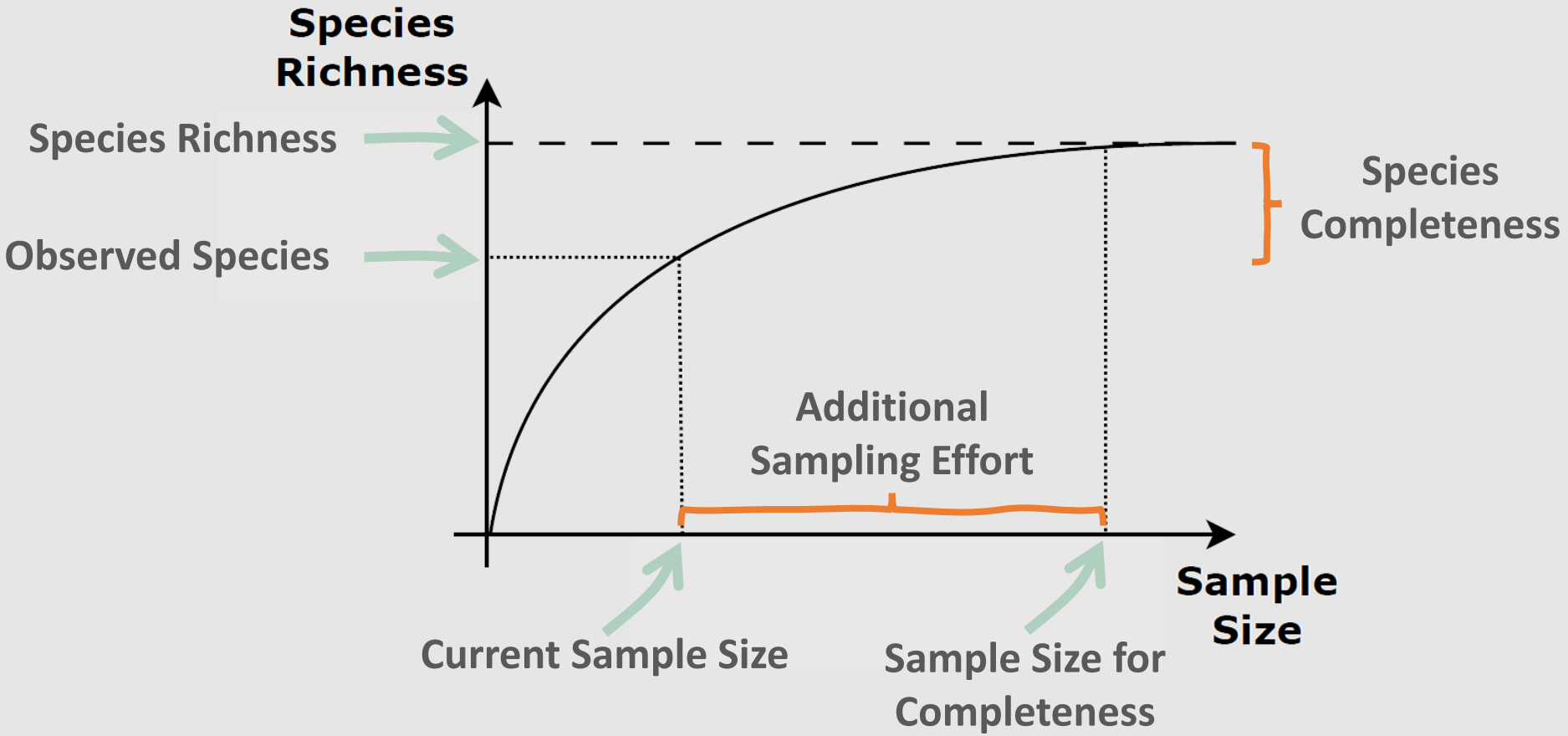
[…] we estimate the total tree species richness at global, continental, and biome levels. Our results indicate that there are ~73,000 tree species globally, among which ~9,000 tree species are yet to be discovered
[Gatti et. al, The number of tree species on Earth, 2022]

A different View: Biodiversity Analysis





Species Richness Curves



The Bernoulli Space Model

Sampling Process

- i.i.d. sampling
- species have unknown, fixed observation probability
- one observation may contain multiple species

Chao2-Estimator [A. Chao. 1984]

$$S_{Chao2} \approx S_{obs} + Q_1^2/(2Q_2)$$

Species Completeness

„What fraction of all species have we observed?“

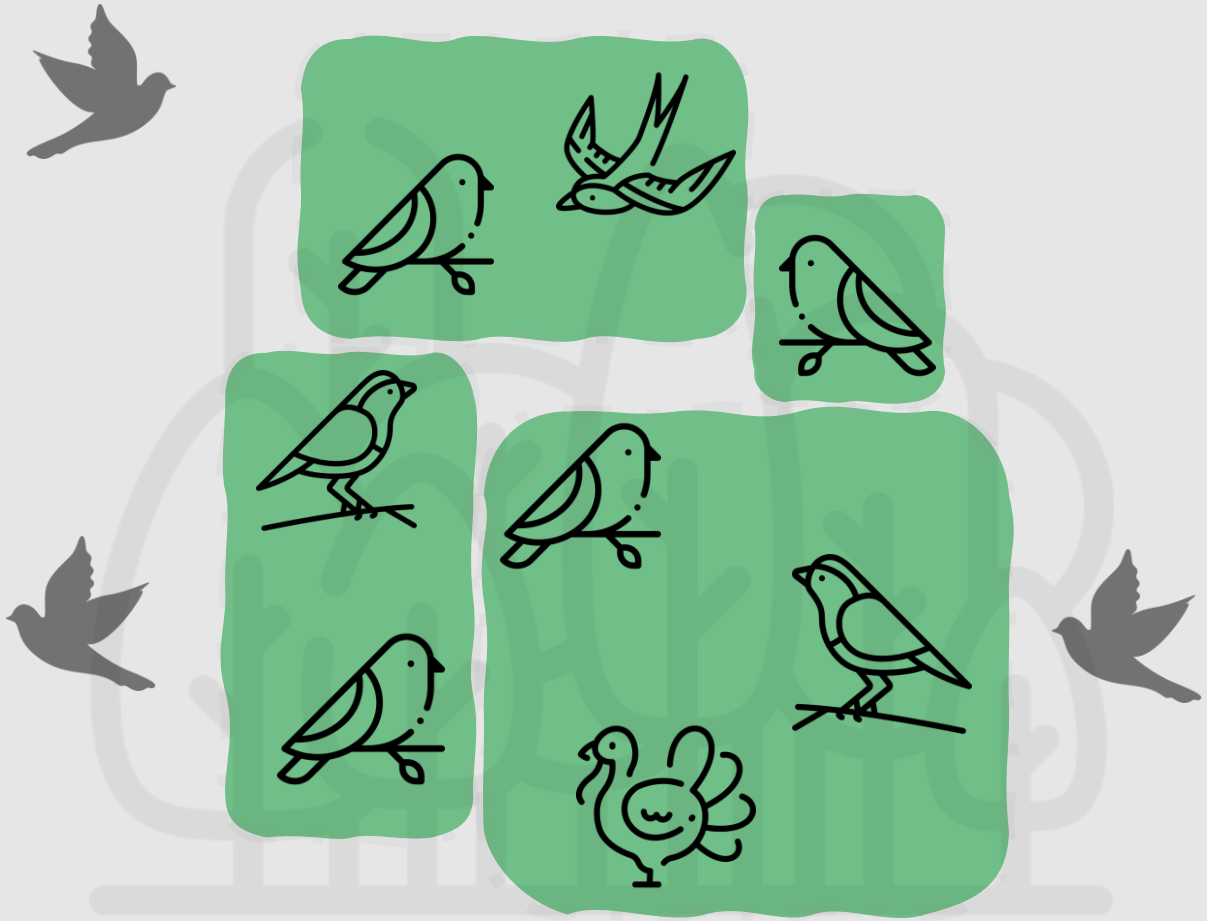
Species Coverage

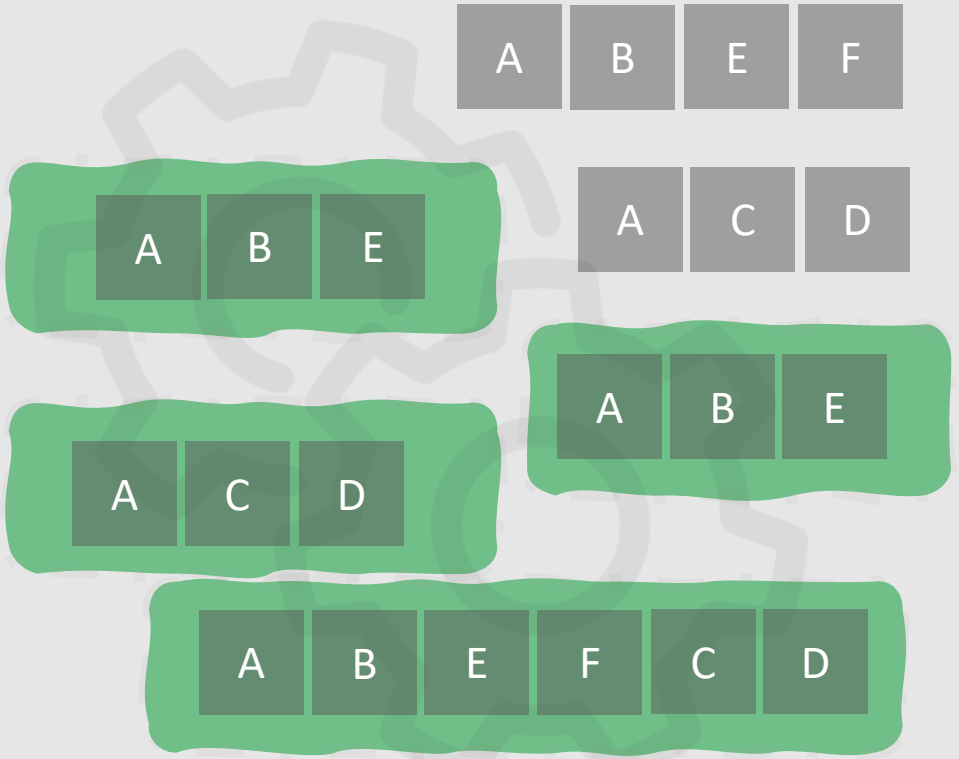
„How much of the probability space do the unobserved species cover?“

Sampling Extrapolation

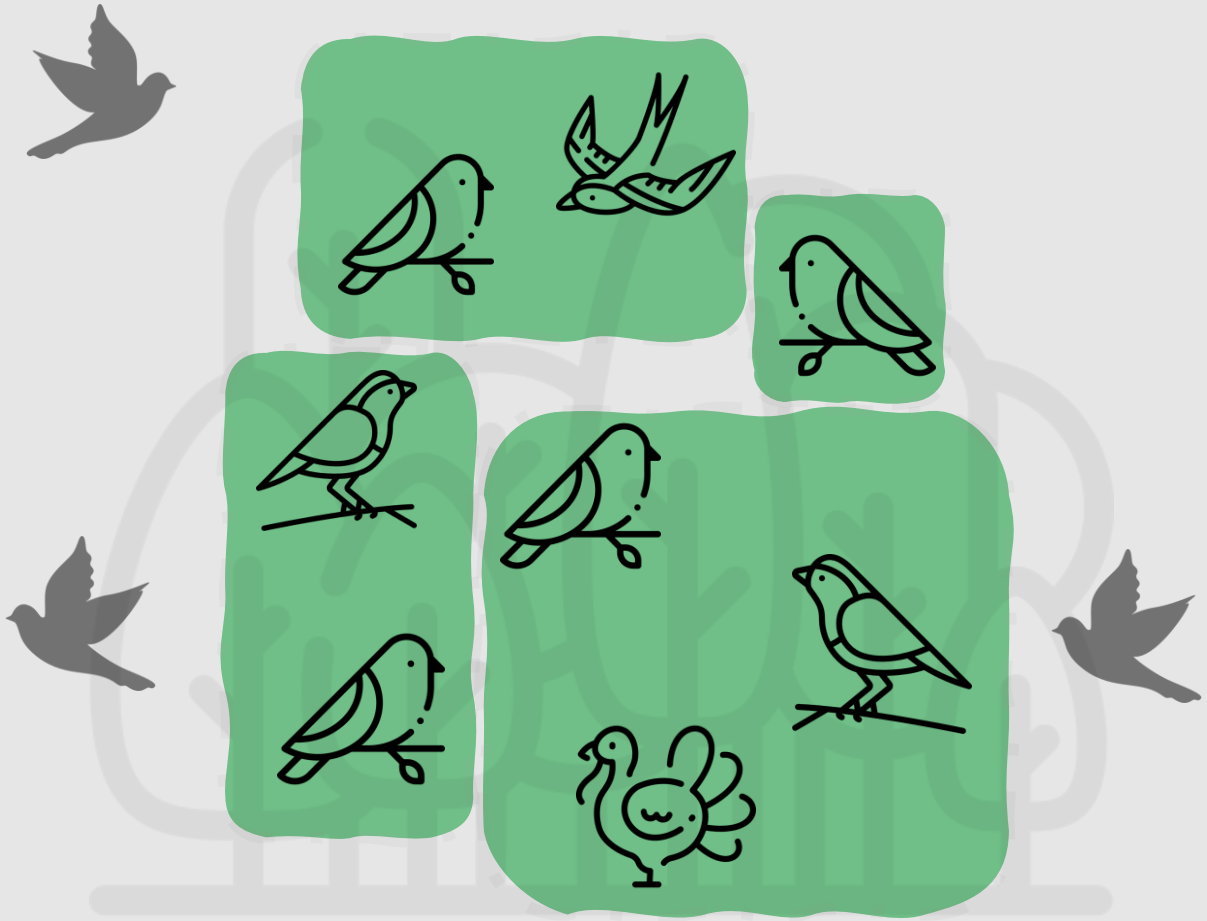
„How much longer do we need to observe birds until we observe new species?“

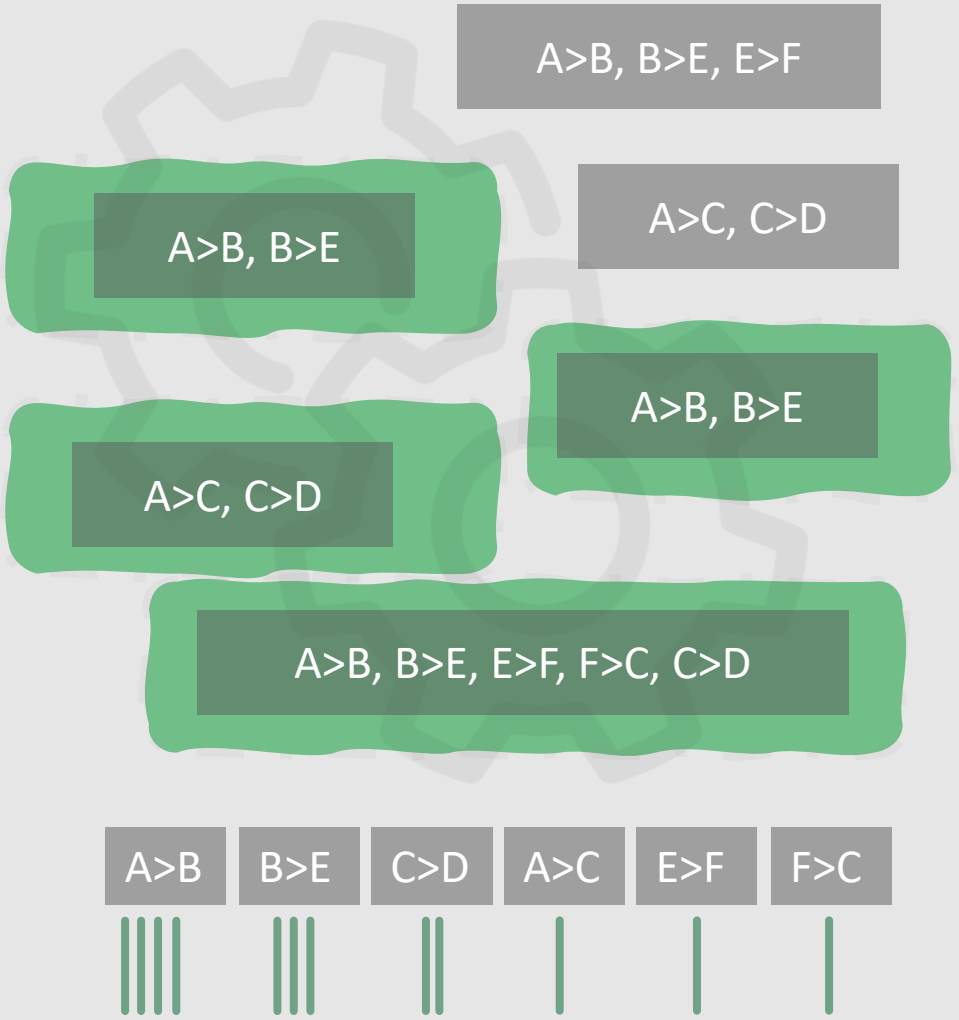
From Birds to Event Log Species





From Birds to Event Log Species





Estimating Richness on complete Logs

Python-based Implementation of metrics and log species

Calculated all proposed metrics on public event logs

BPI-2012, BPI-2018, BPI-2019, Sepsis Cases

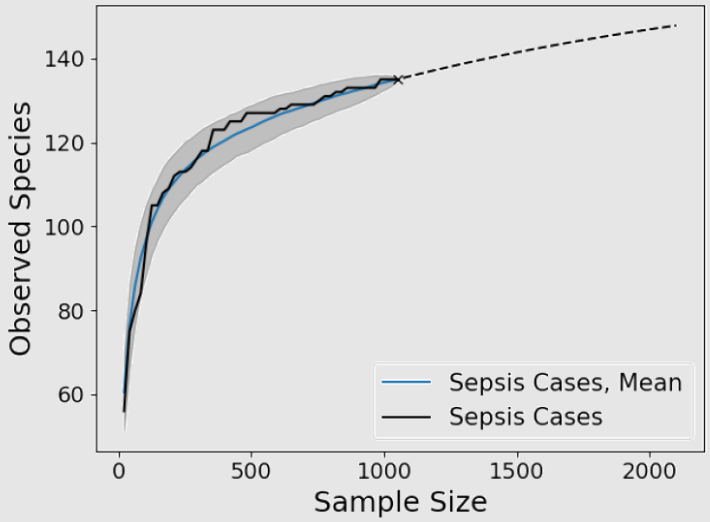
Considered log species

activities, df-relations, trace variants, activities + execution times

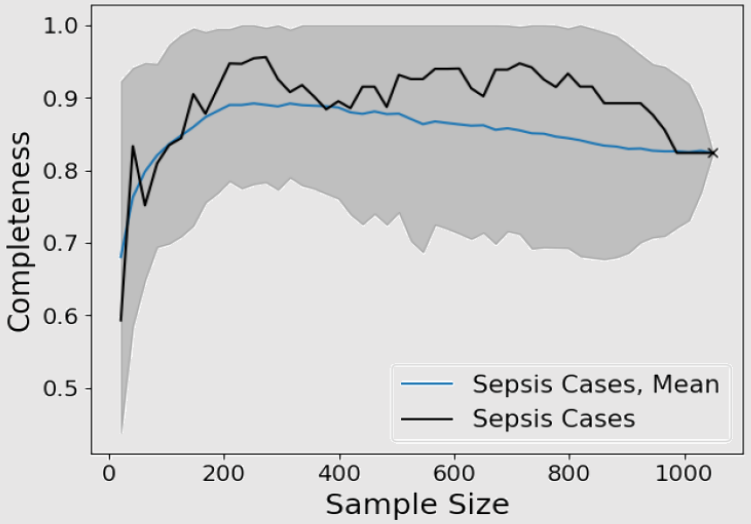
Table II: Species richness estimation, coverage, and completeness for four event logs, and seven species definitions.

Log	Species Def.	S_{obs}	S_{est}	Q_1	Q_2	Cov_{obs}	Com_{obs}	$l_{.99}$	$l_{.95}$	$l_{.90}$	$l_{.80}$
BPI-2012	ζ_{act}	24	24	0	0	1.0	1.0	-	-	-	-
	ζ_{df}	149	161	7	2	0.999	0.925	46435	9577	-	-
	ζ_{tv}	4336	30346	3727	267	0.715	0.143	406521	259527	196219	132912
	ζ_{t1}	958	2816	535	77	0.996	0.340	190458	117290	85779	54267
	ζ_{t5}	487	1164	268	53	0.998	0.418	134446	81196	58263	35329
	ζ_{t30}	210	288	74	53	0.999	0.729	45666	23401	13812	4223
	ζ_{te2}	112	112	2	5	0.999	1.0	-	-	-	-

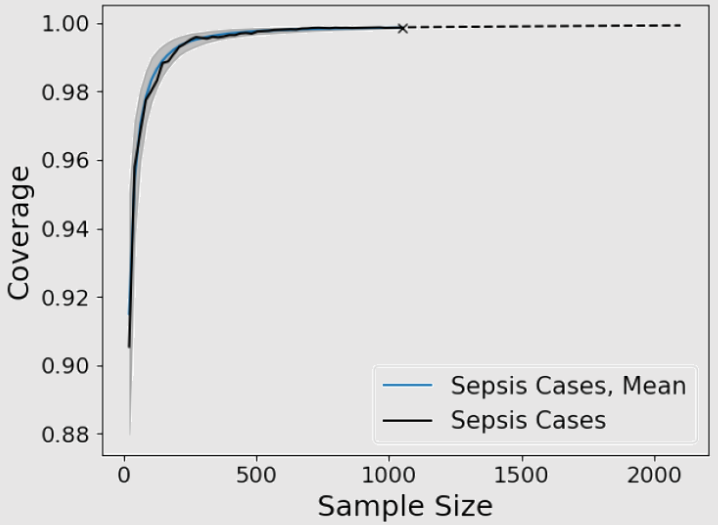
Estimating Richness on Log samples



(f) S_{obs} for ζ_{df}



(d) Com_{obs} for ζ_{df}

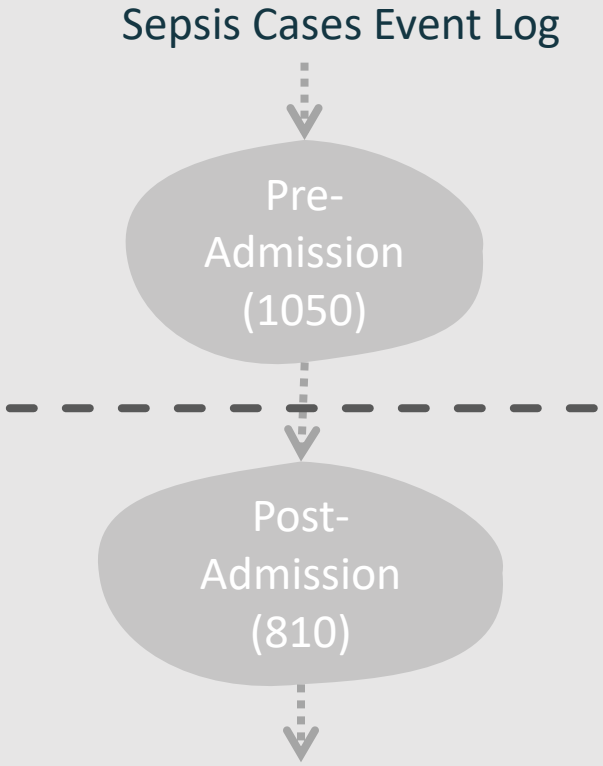


(b) Cov_{obs} for ζ_{df}

Estimating Richness on Subprocesses

Control-Flow based Splitting:

Does log completeness differ for different phases of the process?

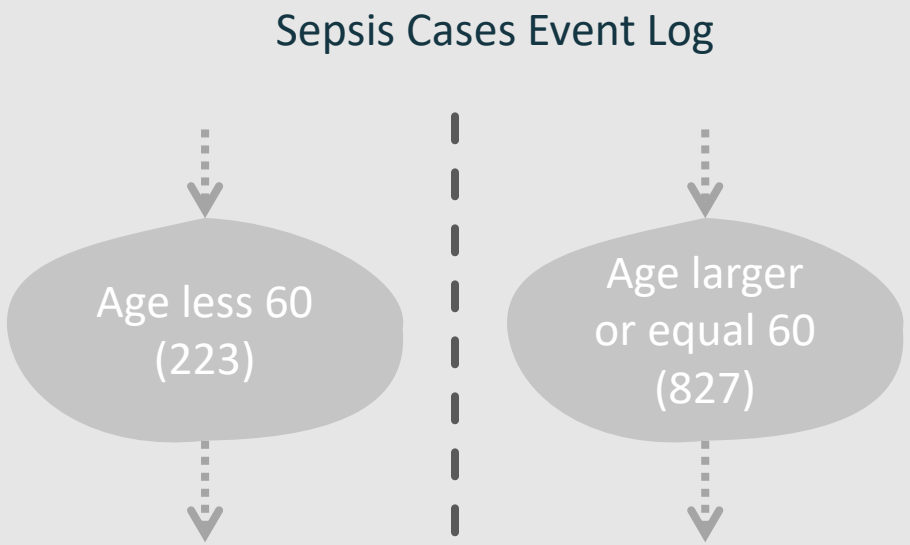


Spec.	Log	S_{obs}	S_{est}	Q_1	Q_2	Cov_{obs}	Com_{obs}	$l_{.99}$	$l_{.90}$
ζ_{act}	Pre	11	11	1	0	1.0	1.0	-	-
	Post	15	15	2	0	1.0	1.0	-	-
ζ_{df}	Pre	87	118	8	1	0.999	0.737	13811	4148
	Post	88	113	19	7	0.997	0.778	3426	897
ζ_{tv}	Pre	298	735	180	37	0.828	0.405	10427	4550
	Post	467	2955	393	31	0.514	0.158	22735	10926
ζ_{t1}	Pre	1041	1673	513	208	0.939	0.622	4699	1720
	Post	2392	13238	2058	195	0.646	0.181	18811	8980
ζ_{t30}	Pre	127	191	41	13	0.995	0.665	5821	2010
	Post	1120	2566	639	141	0.889	0.436	7392	3170
ζ_{te2}	Pre	132	144	14	8	0.998	0.917	1963	-
	Post	127	159	29	13	0.994	0.799	2717	638

Estimating Richness on Subprocesses

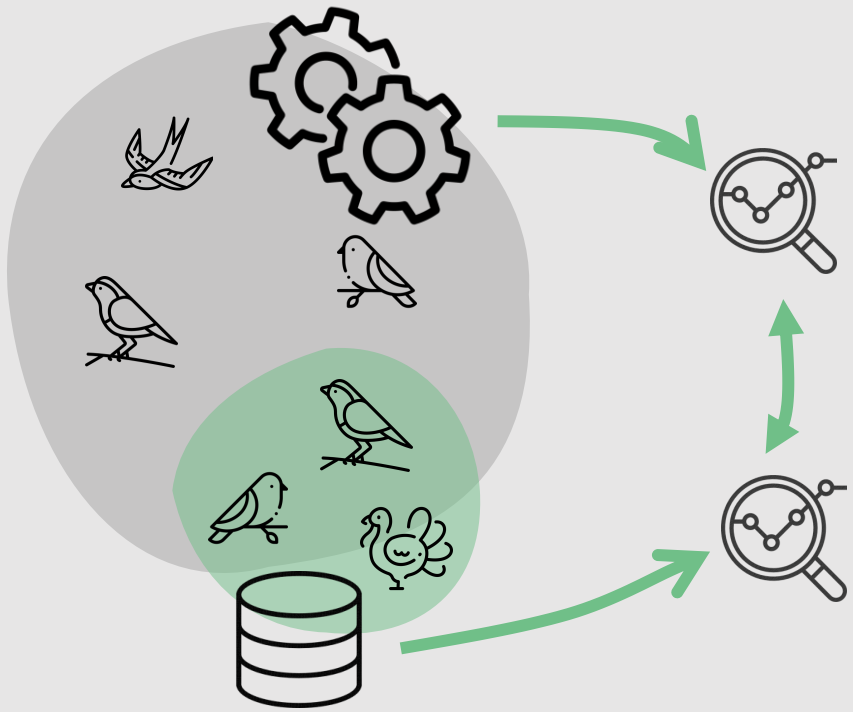
Attribute-based Splitting:

Does log completeness differ for different patient groups?



Spec.	Log	S_{obs}	S_{est}	Q_1	Q_2	Cov_{obs}	Com_{obs}	$l_{.99}$	$l_{.90}$
ζ_{act}	< 60	15	16	2	1	0.998	0.938	547	35
	≥ 60	16	16	0	0	1.0	1.0	-	-
ζ_{df}	< 60	101	113	15	9	0.994	0.894	444	17
	≥ 60	133	158	16	5	0.998	0.842	3675	631
ζ_{tv}	< 60	171	1946	158	7	0.287	0.088	11309	5539
	≥ 60	707	8185	659	29	0.202	0.086	42374	20763
ζ_{t1}	< 60	907	3574	706	93	0.704	0.254	3636	1694
	≥ 60	2858	10127	2138	314	0.793	0.282	12019	5543
ζ_{t30}	< 60	420	906	254	66	0.893	0.464	1703	718
	≥ 60	1081	2316	599	145	0.942	0.467	6786	2857
ζ_{te2}	< 60	149	171	25	14	0.989	0.871	509	51
	≥ 60	200	225	26	13	0.997	0.889	2018	114

Conclusion and Future Work



Conclusion

- representativeness of event logs
- log completeness estimation using species richness estimation
- well-known event logs are incomplete in many dimensions
- **enables assessment of log quality and analysis confidence**

Future Work

- dropping error- & drift assumptions
- automatically detect incomplete sub processes
- evaluation against ground truth dataset

Thank you for your Attention!

martin.kabierski@hu-berlin.de

Credits

Icons taken from:
Freepik
Smashicons from www.flaticon.com
Vitaly Gorbachev from www.flaticon.com