Für Mensch & Umwelt

KI-Lab
Anwendungslabor für künstliche Intelligenz und Big Data

Umwelt Bundesamt

Fachtagung Bodenindikatoren (17.04.2024)

Wie kann KI dabei helfen, Datenlücken bei der Entwicklung bodenbezogener Indikatoren zu schließen?

Viola Rädle
Data Scientist
KI-Lab Leipzig, Umweltbundesamt
ki-anwendungslabor@uba.de

Das KI-Lab am Umweltbundesamt

Anwendungslabor für Künstliche Intelligenz und Big Data

Informieren • Involvieren • Begeistern

Ziele: Forschung, Entwicklung, Kompetenzaufbau, Beratung, Wissenstransfer

Welche Arten von Datenlücken gibt es?

Bsp: Umfrage zu Einkommen

Umfrage

Geschlecht:

☐ m ☐ w ☐ d

Haarfarbe: _____

Alter: _____

Schuhgröße: _____

Einkommen: _____

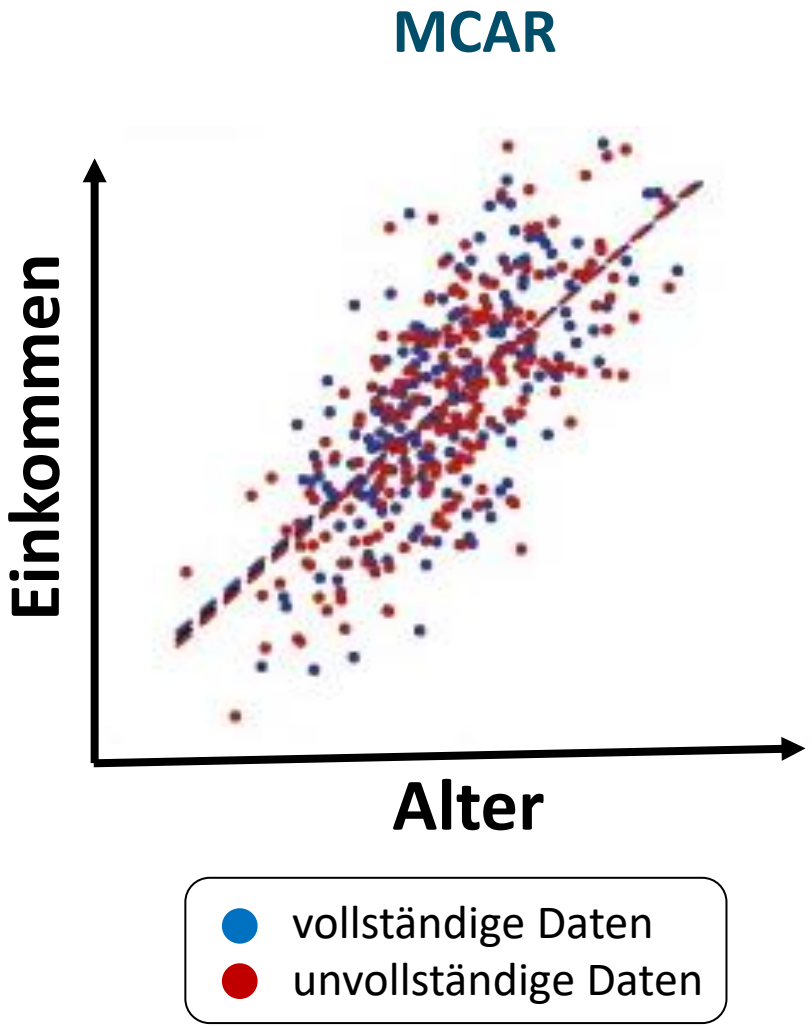
Quelle: https://de.vecteezy.com/vektorkunst/554935-lust-auf-kugelschreiber-vektor-icon

Welche Arten von Datenlücken gibt es?

Bsp: Umfrage zu Einkommen

Missing completely at random (MCAR):

- **kein Zusammenhang** mit erfassten Daten
- Bsp: technische Fehler, Übersehen eines Feldes



Quelle: https://hushuli.github.io/Metabox-Blog.github.io/posts/2018-11-08-missing-value/

Welche Arten von Datenlücken gibt es?

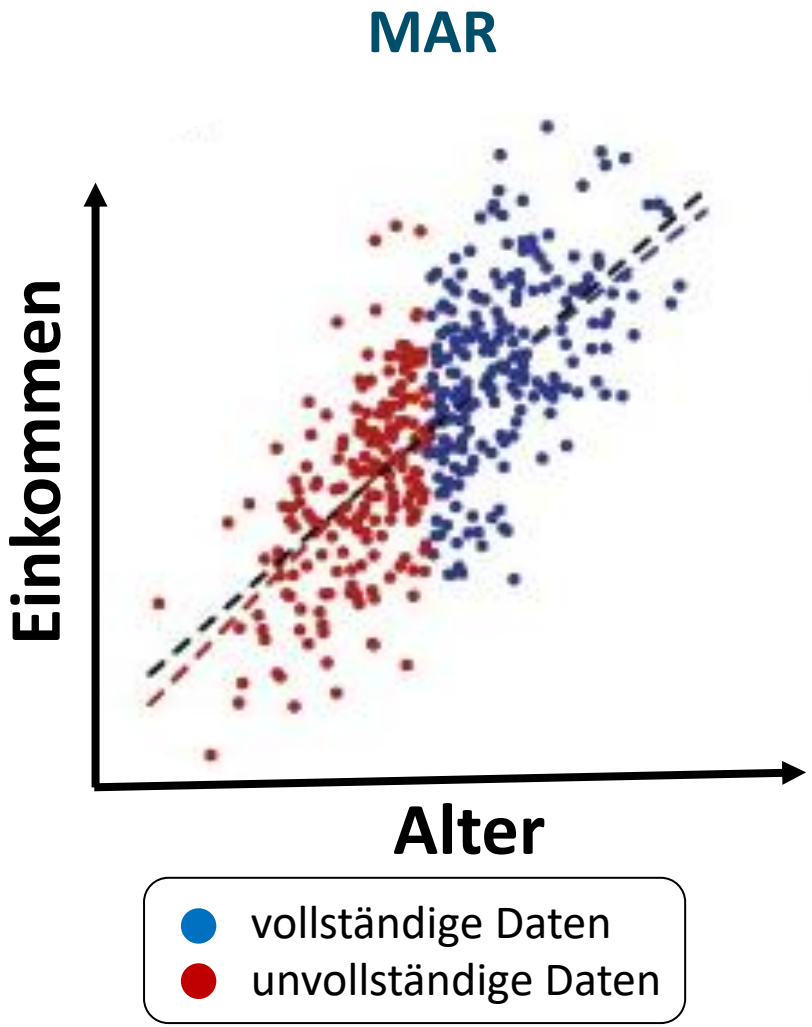
Bsp: Umfrage zu Einkommen

Missing completely at random (MCAR):

- **kein Zusammenhang** mit erfassten Daten
- Bsp: technische Fehler, Übersehen eines Feldes

Missing at random (MAR):

- Zusammenhang mit **anderer Variable**
- Bsp: Jüngere Menschen wollen keine Auskunft über ihr Einkommen geben



Quelle: https://hushuli.github.io/Metabox-Blog.github.io/posts/2018-11-08-missing-value/

Welche Arten von Datenlücken gibt es?

Bsp: Umfrage zu Einkommen

Missing completely at random (MCAR):

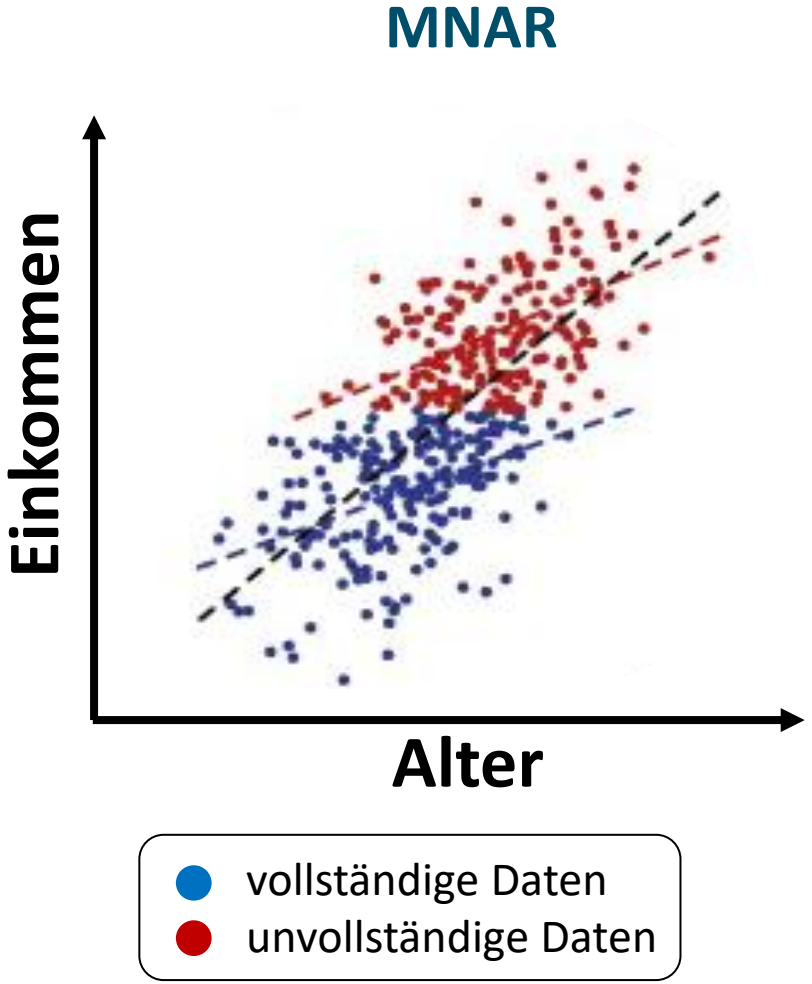
- **kein Zusammenhang** mit erfassten Daten
- Bsp: technische Fehler, Übersehen eines Feldes

Missing at random (MAR):

- Zusammenhang mit **anderer Variable**
- Bsp: Jüngere Menschen wollen keine Auskunft über ihr Einkommen geben

Missing not at random (MNAR):

- Zusammenhang mit **lückenhafter Variable**
- Bsp: Menschen mit höherem Einkommen wollen keine Auskunft geben



Quelle: https://hushuli.github.io/Metabox-Blog.github.io/posts/2018-11-08-missing-value/

Welche Arten von Datenlücken gibt es?

Beispiele aus der Umweltforschung

Missing completely at random (MCAR):

- technische Probleme / Stromausfall
- finanzieller Aufwand der Messung
- Messpersonal erkrankt

Missing at random (MAR):

- Monitoring der Wasserqualität in Industrienähe nicht möglich (Gelände abgesperrt)
- weniger Messstationen in schwer zugänglichen Gebieten als in Zivilisationsnähe
- Messfahrten aufgrund schlechter Witterungsbedingungen abgebrochen

Missing not at random (MNAR):

- Monitoring der Artenvielfalt: bedrohte Arten sind schwieriger zu finden
- Sensor kann Werte oberhalb des Messbereichs nicht mehr erfassen
- Radioaktivitätsmessungen: Probenahme in kontaminierten Gebieten zu gefährlich

data imputation

data dropping

lösche lückenhafte Messungen

Der einfachste Weg: Data dropping

Lösche alle Messungen mit fehlenden Werten

Datum	Temperatur in °C	Wassergehalt in %	Leitfähigkeit in dS/m	pH	Stickstoff in %
20.03.2023		22.6	1.2	4.9	0.09
01.04.2023	12.5	20.4	1.8	5.2	0.07
13.04.2023	14.2	25.3		6.3	0.10
02.05.2023	15.6	33.7	1.9	5.0	
24.05.2023	16.4	30.1	2.1	5.7	0.12
	19.0	28.3	2.3	6.5	0.14
02.09.2023	18.7	35.8	2.6	6.8	0.11
17.09.2023	17.2		2.0	6.2	0.06

Der einfachste Weg: Data dropping

Lösche alle Messungen mit fehlenden Werten

Datum	Temperatur in °C	Wassergehalt in %	Leitfähigkeit in dS/m	pH	Stickstoff in %
~~20.03.2023~~		~~22.6~~	~~1.2~~	~~4.9~~	~~0.09~~
01.04.2023	12.5	20.4	1.8	5.2	0.07
~~13.04.2023~~	~~14.2~~	~~25.3~~		~~6.3~~	~~0.10~~
~~02.05.2023~~	~~15.6~~	~~33.7~~	~~1.9~~	~~5.0~~	
24.05.2023	16.4	30.1	2.1	5.7	0.12
	~~19.0~~	~~28.3~~	~~2.3~~	~~6.5~~	~~0.14~~
02.09.2023	18.7	35.8	2.6	6.8	0.11
~~17.09.2023~~	~~17.2~~		~~2.0~~	~~6.2~~	~~0.06~~

data imputation

data dropping

lösche lückenhafte Messungen

single imputation

fülle die Lücken mit anderem Wert

univariat

- Hot-Deck
- Mittelwert/Median
- Regression

multivariat

- k-nearest-neighbors
- Random Forest
- Support Vector Machine
- multiple Regression
- Hauptkomponentenanalyse

Hot-Deck-Imputation

Fülle die Datenlücke mit einem zufälligen Wert aus der Spalte

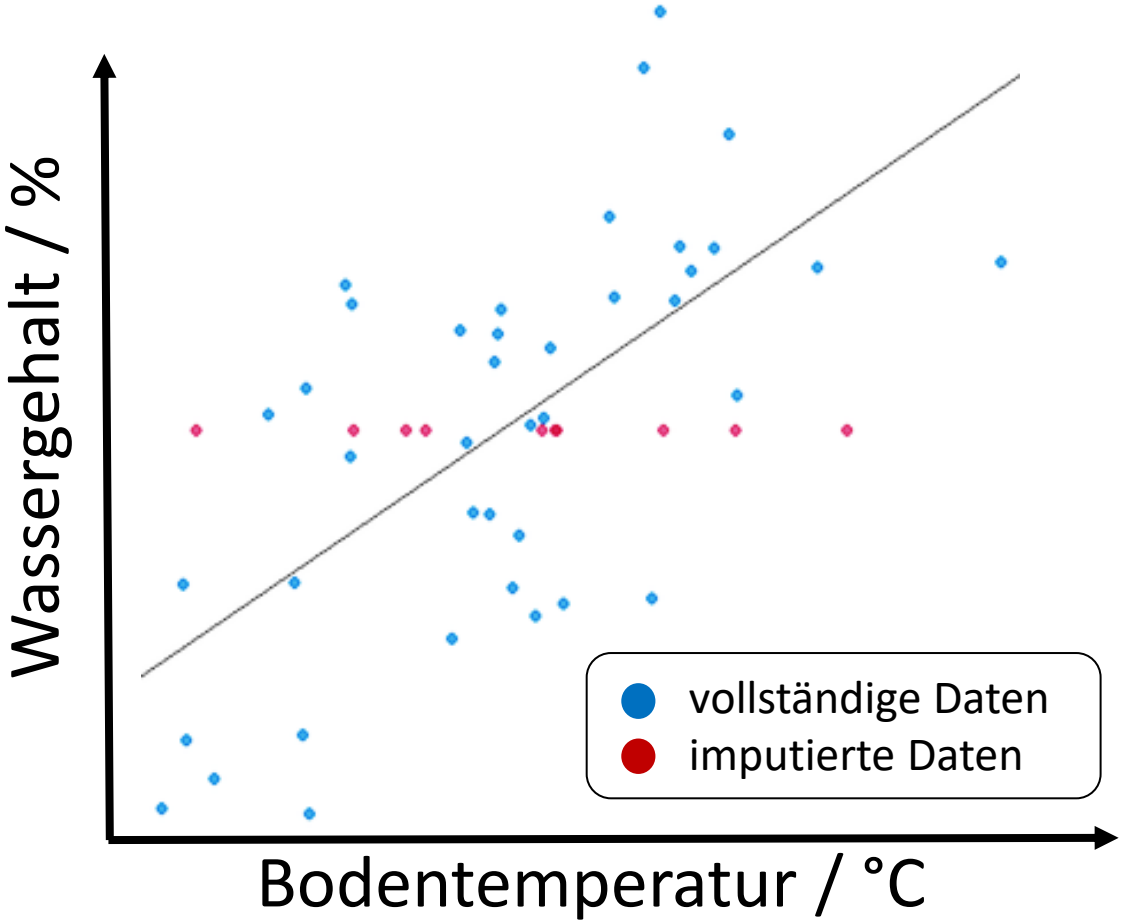
Datum	Temperatur in °C	Wassergehalt in %	Leitfähigkeit in dS/m	pH	Stickstoff in %
20.03.2023		22.6	1.2	4.9	0.09
01.04.2023	12.5	20.4	1.8	5.2	0.07
13.04.2023	**14.2**	25.3		6.3	0.10
02.05.2023	15.6	33.7	1.9	5.0	
24.05.2023	16.4	30.1	**2.1**	5.7	0.12
	19.0	**28.3**	2.3	6.5	0.14
02.09.2023	18.7	35.8	2.6	6.8	0.11
17.09.2023	17.2		2.0	6.2	**0.06**

Mean/Median Imputation

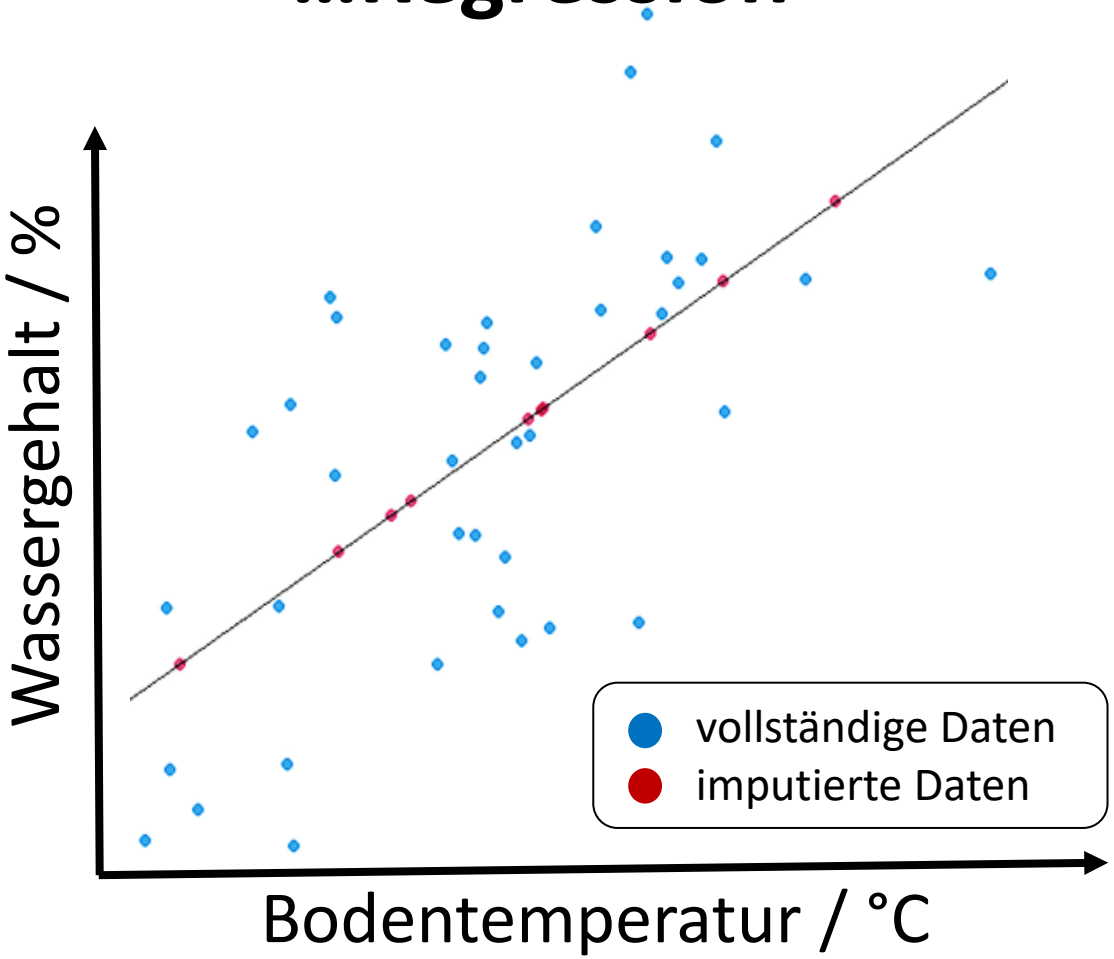
	Datum	Temperatur in °C	Wassergehalt in %	Leitfähigkeit in dS/m	pH	Stickstoff in %
	20.03.2023		22.6	1.2	4.9	0.09
	01.04.2023	12.5	20.4	1.8	5.2	0.07
	13.04.2023	14.2	25.3		6.3	0.10
	02.05.2023	15.6	33.7	1.9	5.0	
	24.05.2023	16.4	30.1	2.1	5.7	0.12
		19.0	28.3	2.3	6.5	0.14
	02.09.2023	18.7	35.8	2.6	6.8	0.11
	17.09.2023	17.2		2.0	6.2	0.06
Mean	**29.05.2023**	**16.2**	**28.0**	**2.0**		**0.1**
Median	**02.05.2023**	**16.4**	**28.3**	**2.0**		**0.1**

Imputation durch…

…den Mittelwert



…Regression

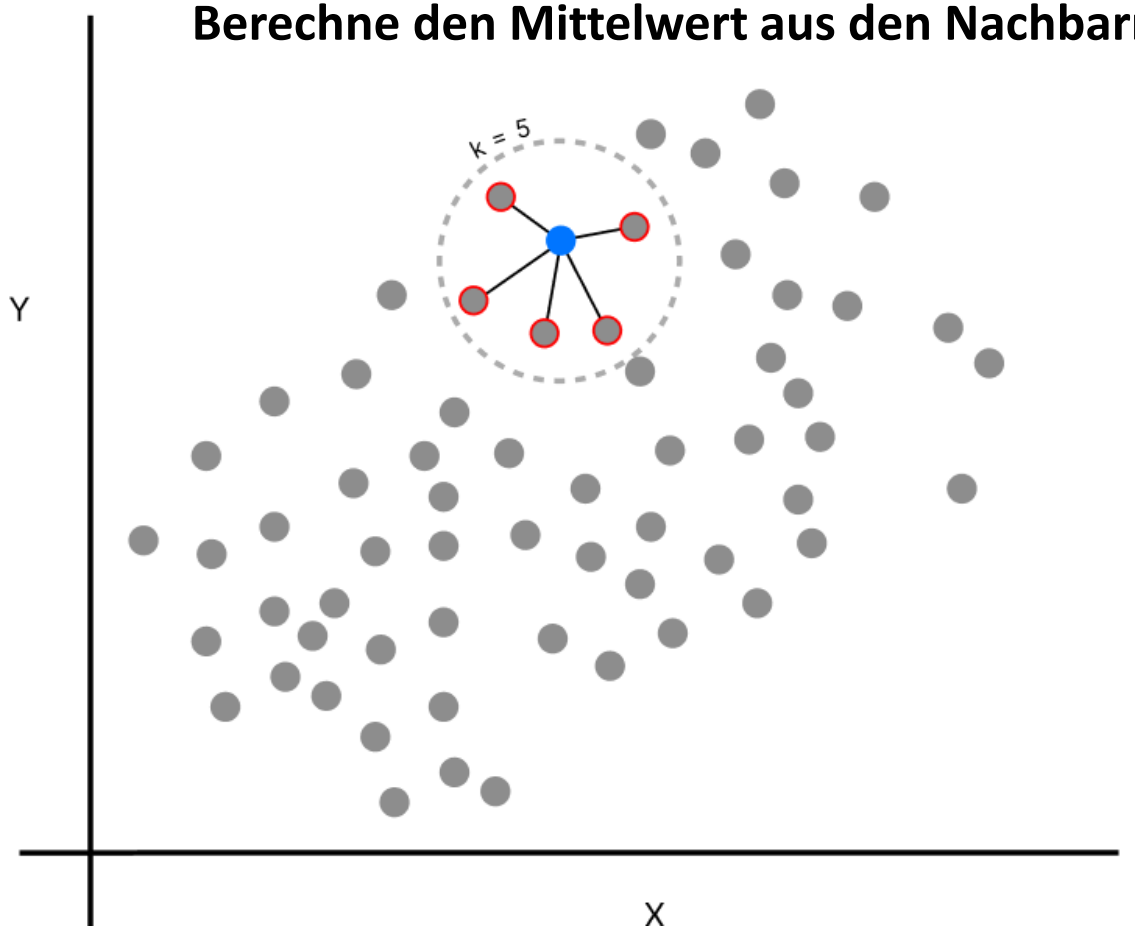


Quelle: https://www.missingdata.nl/missing-data/missing-data-methods/imputation-methods/

Imputation durch…

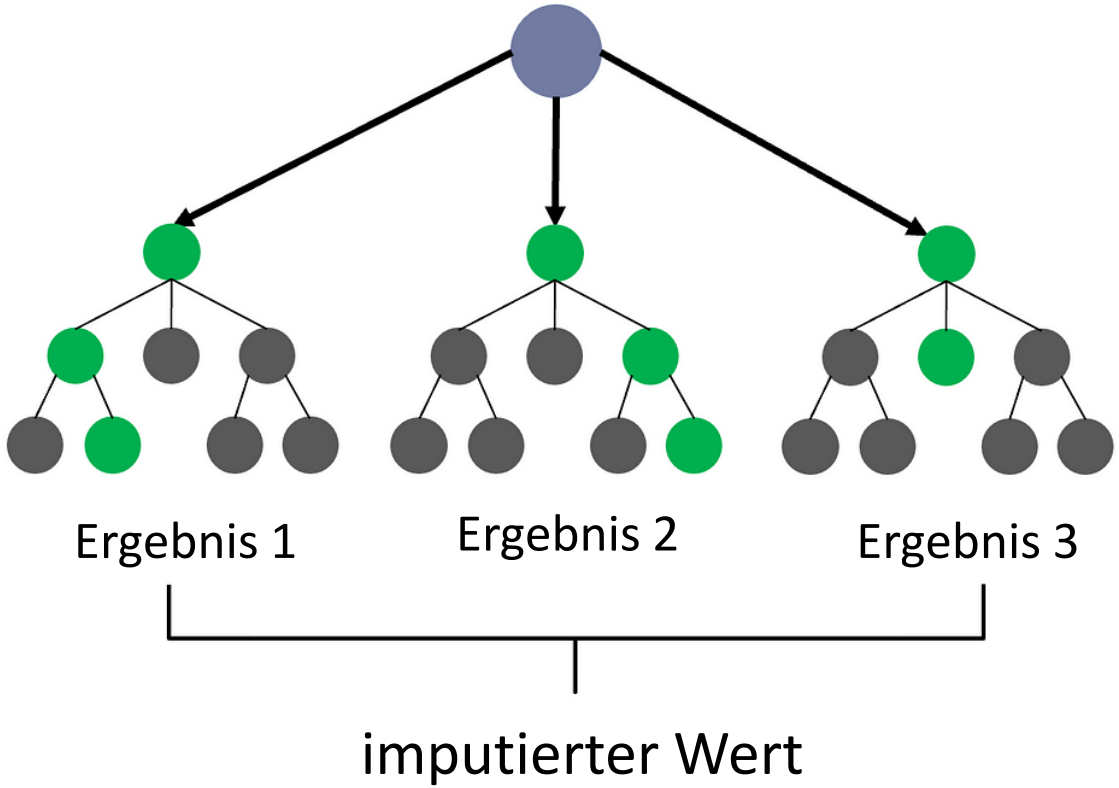
k-nearest-neighbours

Berechne den Mittelwert aus den Nachbarn



random forest

Mittle über mehrere Entscheidungsbäume



Quellen: https://pub.towardsai.net/understanding-k-nearest-neighbors-a-simple-approach-to-classification-and-regression-e4b30b37f151, https://medium.com/@roiyeho/random-forests-98892261dc49

data imputation

data dropping

lösche lückenhafte Messungen

single imputation

fülle die Lücken mit anderem Wert

univariat

- Hot-Deck
- Mittelwert/Median
- Regression

multivariat

- k-nearest-neighbors
- Random Forest
- Support Vector Machine
- multiple Regression
- Hauptkomponentenanalyse

multiple imputation

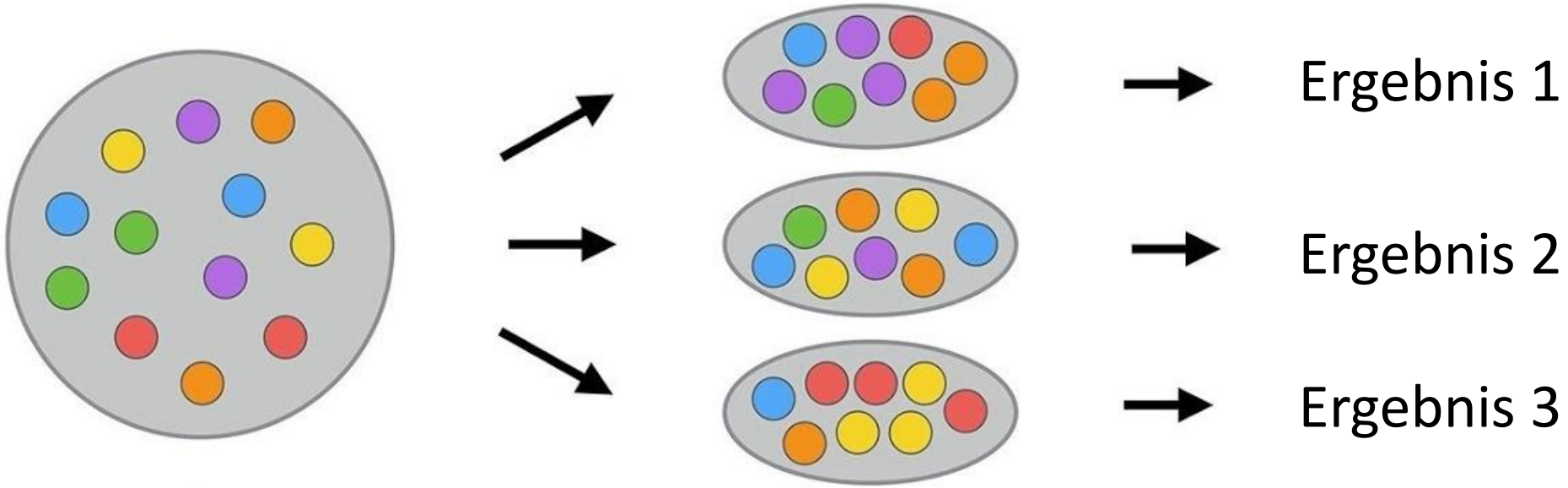
fülle Lücken & beachte Unsicherheiten

- Bootstrap
- Bayesianische Statistik
- Markov Chain Monte Carlo
- Expectation Maximization
- MICE (Multiple Imputation by Chained Equations)

Imputation durch…

Bootstrapping

Resampling zur Unsicherheitsabschätzung



Quelle: https://www.youtube.com/watch?app=desktop&v=d3mcuJycJfI

data imputation

data dropping
lösche lückenhafte Messungen

single imputation
fülle die Lücken mit anderem Wert

univariat
- Hot-Deck
- Mittelwert/Median
- Regression

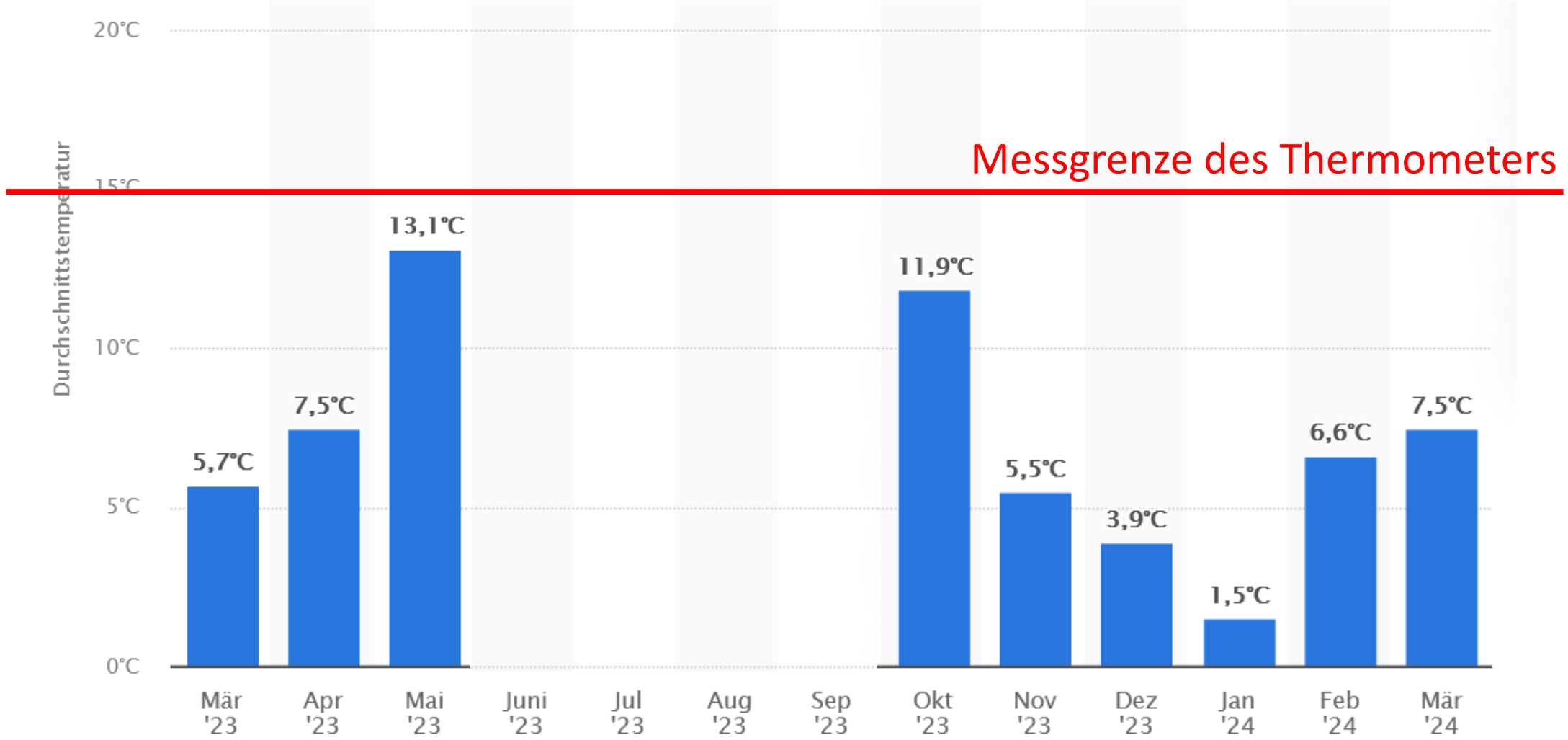
multivariat
- k-nearest-neighbors
- Random Forest
- Support Vector Machine
- multiple Regression
- Hauptkomponentenanalyse

multiple imputation
fülle Lücken & beachte Unsicherheiten

- Bootstrap
- Bayesianische Statistik
- Markov Chain Monte Carlo
- Expectation Maximization
- MICE (Multiple Imputation by Chained Equations)

Limitierungen von Imputation

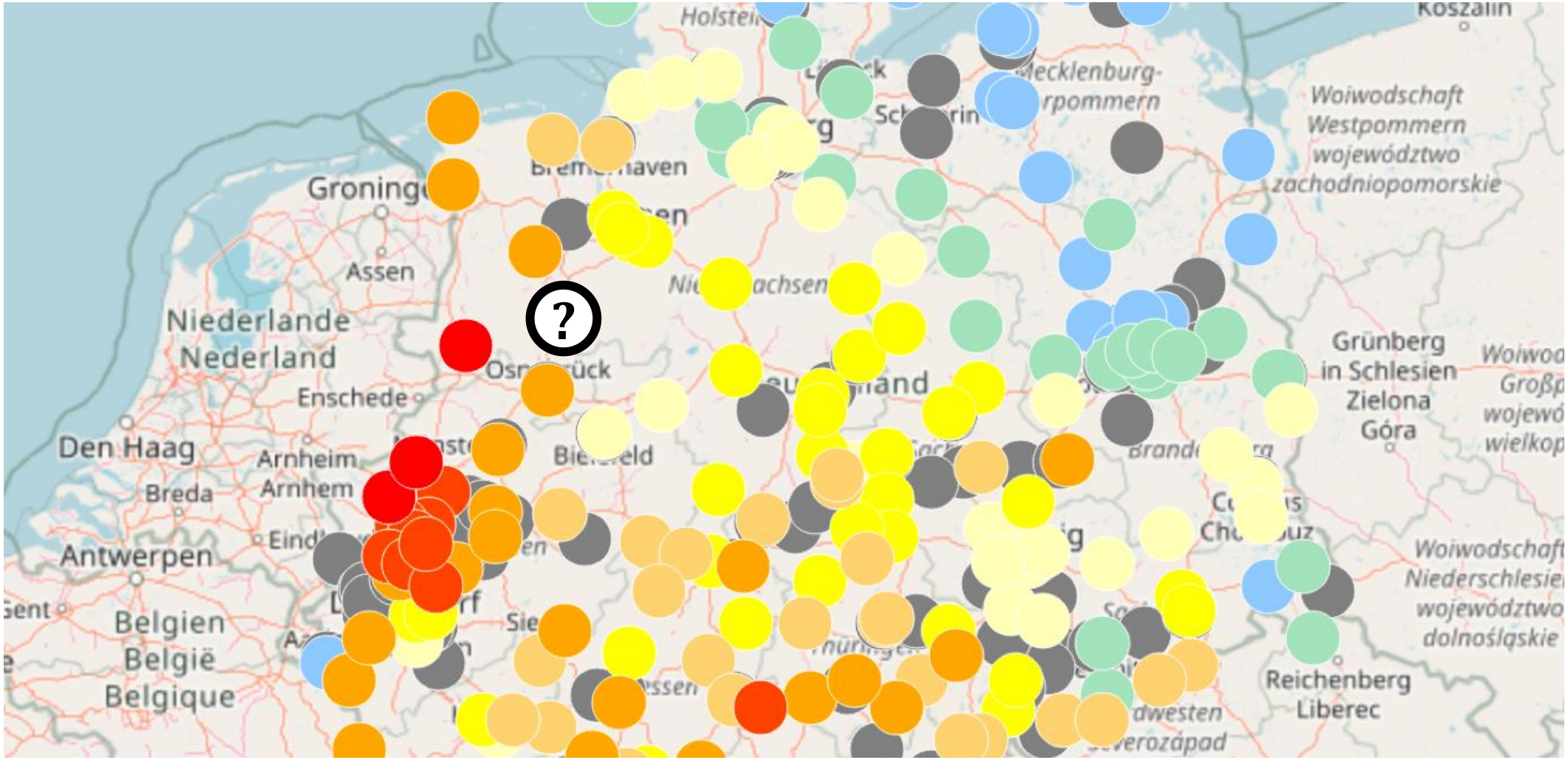
Entzauberung von Künstlicher Intelligenz



Quelle: https://de.statista.com/statistik/daten/studie/5564/umfrage/monatliche-durchschnittstemperatur-in-deutschland/

Räumliche Datenlücken

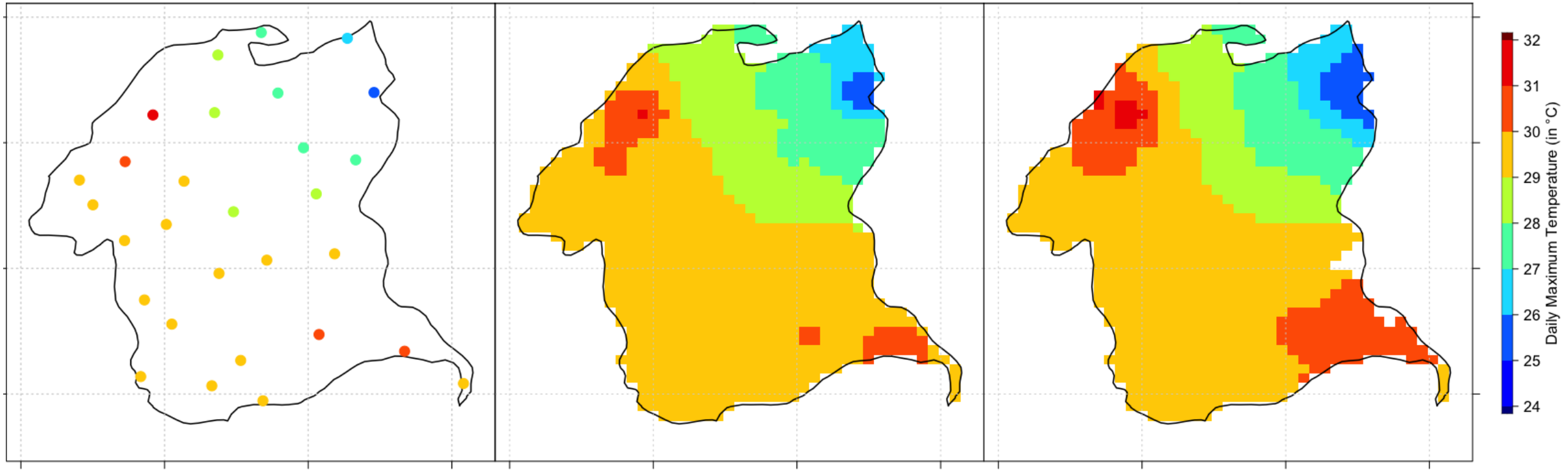
Bsp: Messstationen für Luftqualität



Quelle: https://www.umweltbundesamt.de/daten/luft/luftdaten

Räumliche Interpolation

global vs. lokal (Methode: *inverse distance weight*)

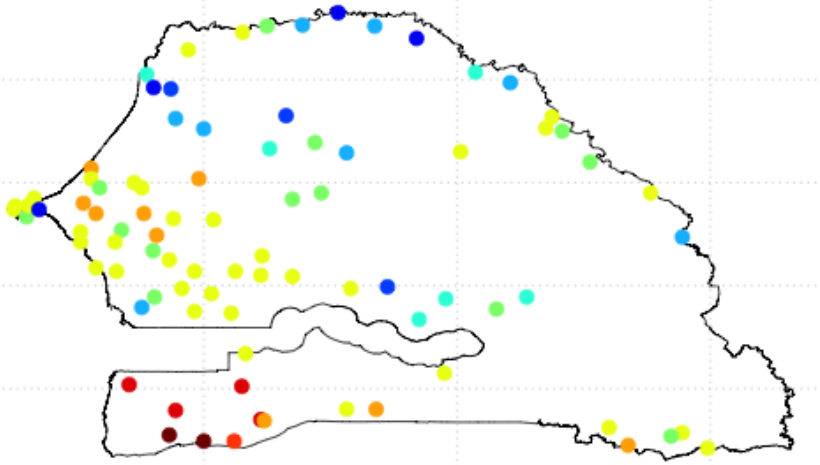


Messstationen | globale Interpolation | lokale Interpolation

Quelle: https://iri.columbia.edu/~rijaf/CDTUserGuide/html/interpolation_methods.html

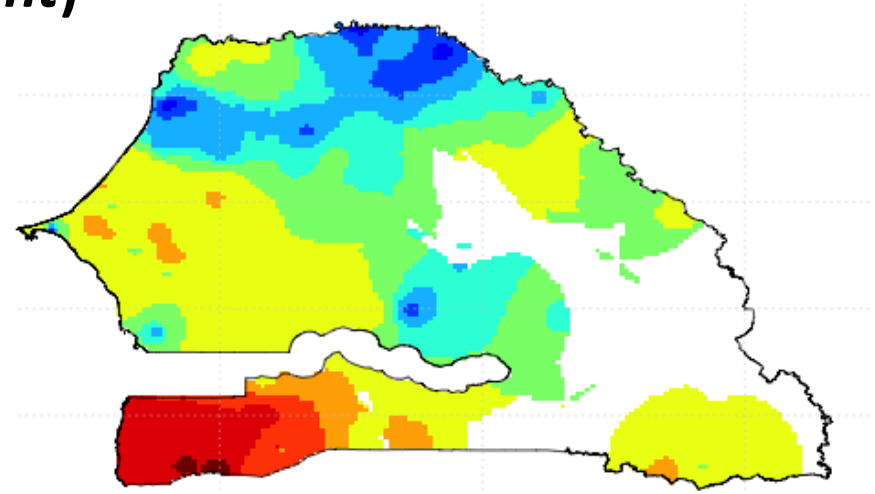
Räumliche Interpolation

Lokale Interpolation (Methode: *inverse distance weight*)

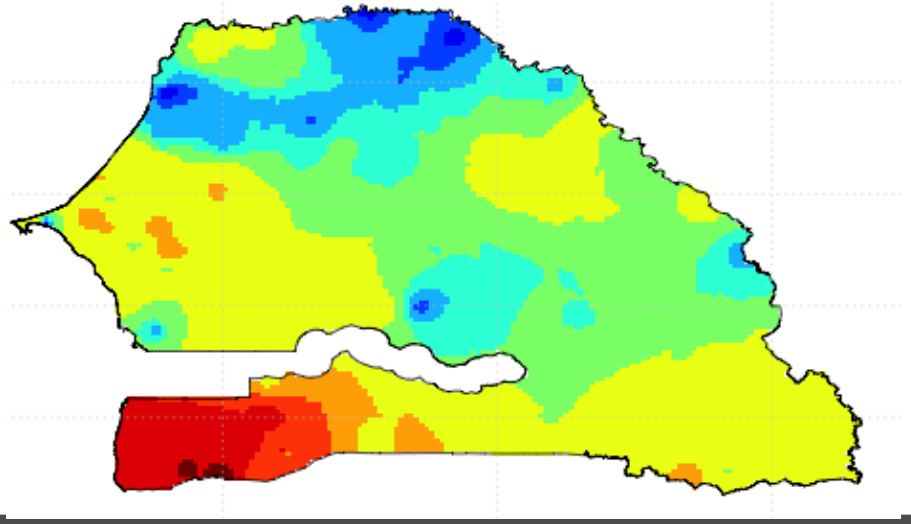


Messstationen

fester Radius r_{max}



variabler Radius r_{max}

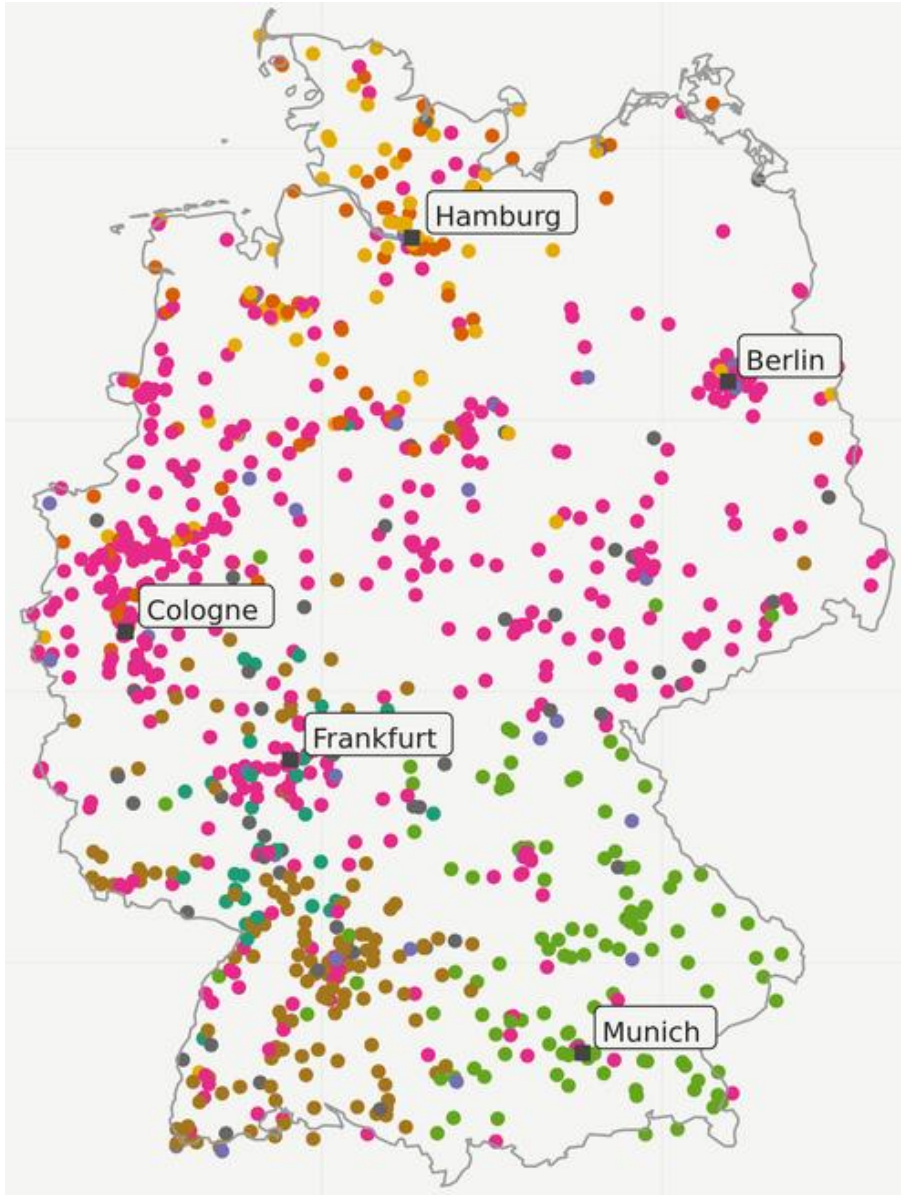


Quelle: https://iri.columbia.edu/~rijaf/CDTUserGuide/html/interpolation_methods.html

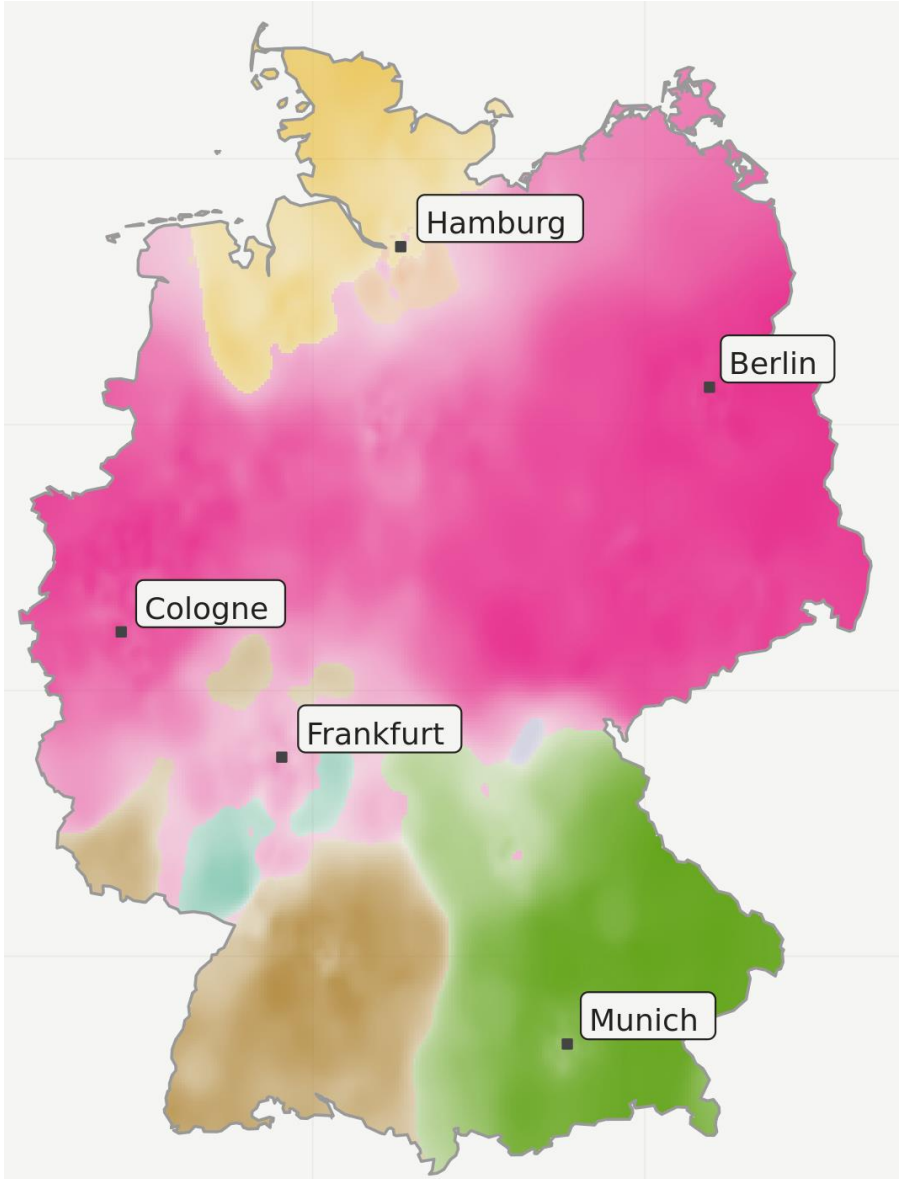
Danke fürs Zuhören ☺

Jetzt: Eure Fragen!

Wie kann KI dabei helfen, Datenlücken bei der Entwicklung bodenbezogener Indikatoren zu schließen?



Dialect
- babbeln
- klönen
- plaudern
- quatschen
- ratschen
- schnacken
- schwätzen



Quelle: https://timogrossenbacher.ch/categorical-spatial-interpolation-with-r/