LV Datengestützte Analysemethoden

2. WAS SIND DATEN?

Sommersemester 2018
FH Joanneum Graz
Studiengang Journalismus und Public Relations

Lehrender: Stefan Kasberger

Stefan Kasberger
@stefankasberger

DATEN

Definition

"Daten sind ein Satz an Werten von qualitativen oder quantitativen Variablen."

Definition

"Daten sind ein Satz an Werten von qualitativen oder quantitativen Variablen."

Satz an Werten: ein Set an Objekten, an denen man interessiert ist.

Definition

"Daten sind ein Satz an Werten von qualitativen oder quantitativen Variablen."

Variablen: Ein Maß oder eine Charakteristik eines Wertes.

Definition

"Daten sind ein Satz an Werten von qualitativen oder quantitativen Variablen."

Qualitativ: Land, Geschlecht, Farbe

Definition

"Daten sind ein Satz an Werten von qualitativen oder quantitativen Variablen."

Quantitativ: Größe, Gwicht, Blutdruck

Definition

"Unter Daten versteht man im Allgemeinen Angaben, (Zahlen-)Werte oder formulierbare Befunde, die durch Messung, Beobachtung u. a. gewonnen wurden."

Differenzierung

Information: zusammengetragene Daten, mit Kontext.

Wissen: zusammengetragene Information, beeinflusst Denken und Handeln.

DATEN & WAHRHEIT
-
WIRKLICHKEIT & DATEN

Diskussion

1)Laptop zu und Tisch aufräumen
2)Gruppen (4-5 Studierende) finden.
Melden und Tisch suchen
3)ModeratorIn und PräsentatorIn
fixieren
4)„Sind naturwissenschaftliche Daten
objektiv“ diskutieren (3min)
5)„Sind sozialwissenschaftliche Daten
objektiv“ diskutieren (3min)
6) Präsentieren der Ergebnisse (je ~30s)

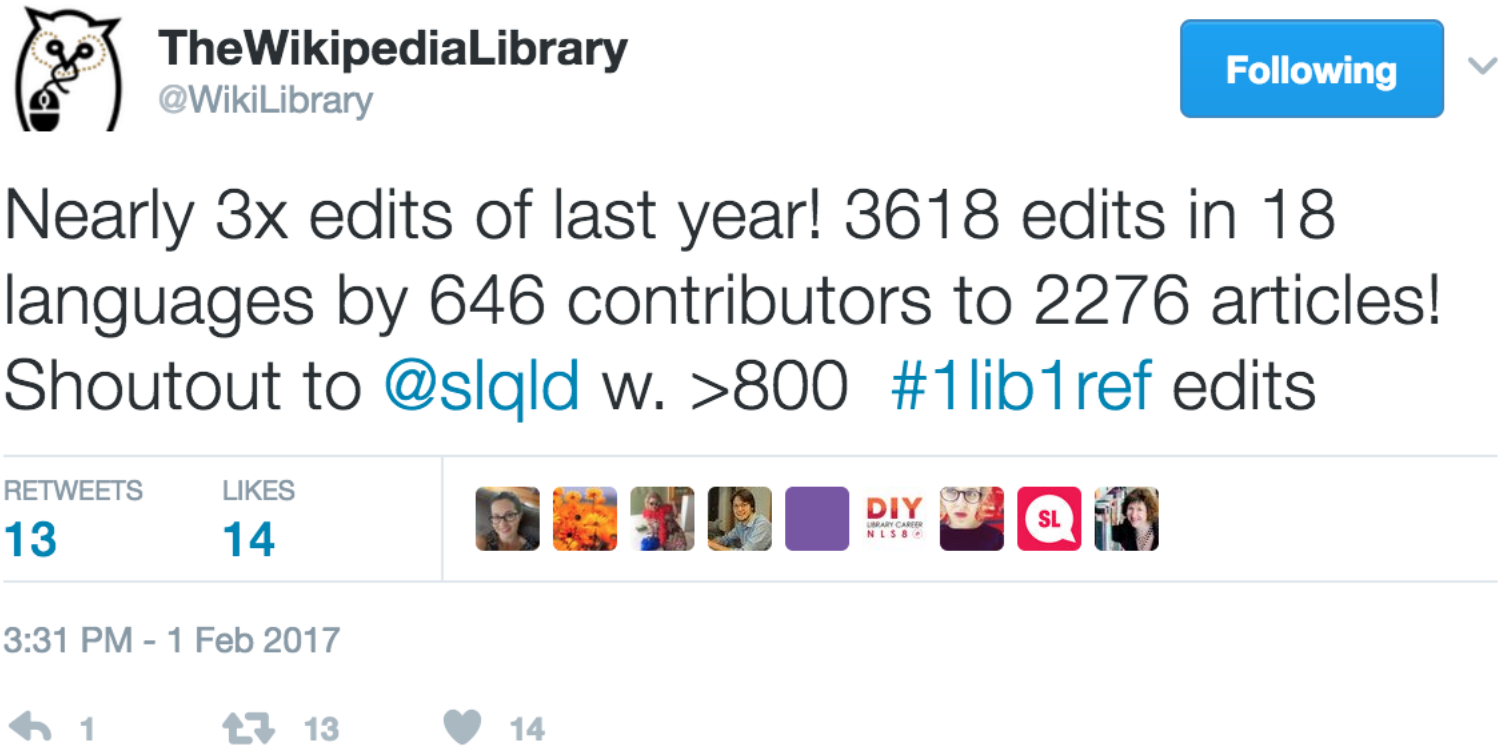
Zu beachten

Daten sind nicht die Wirklichkeit,
sondern nur ein kleiner Ausschnitt
daraus mit gewissen Verzerrungen dabei.

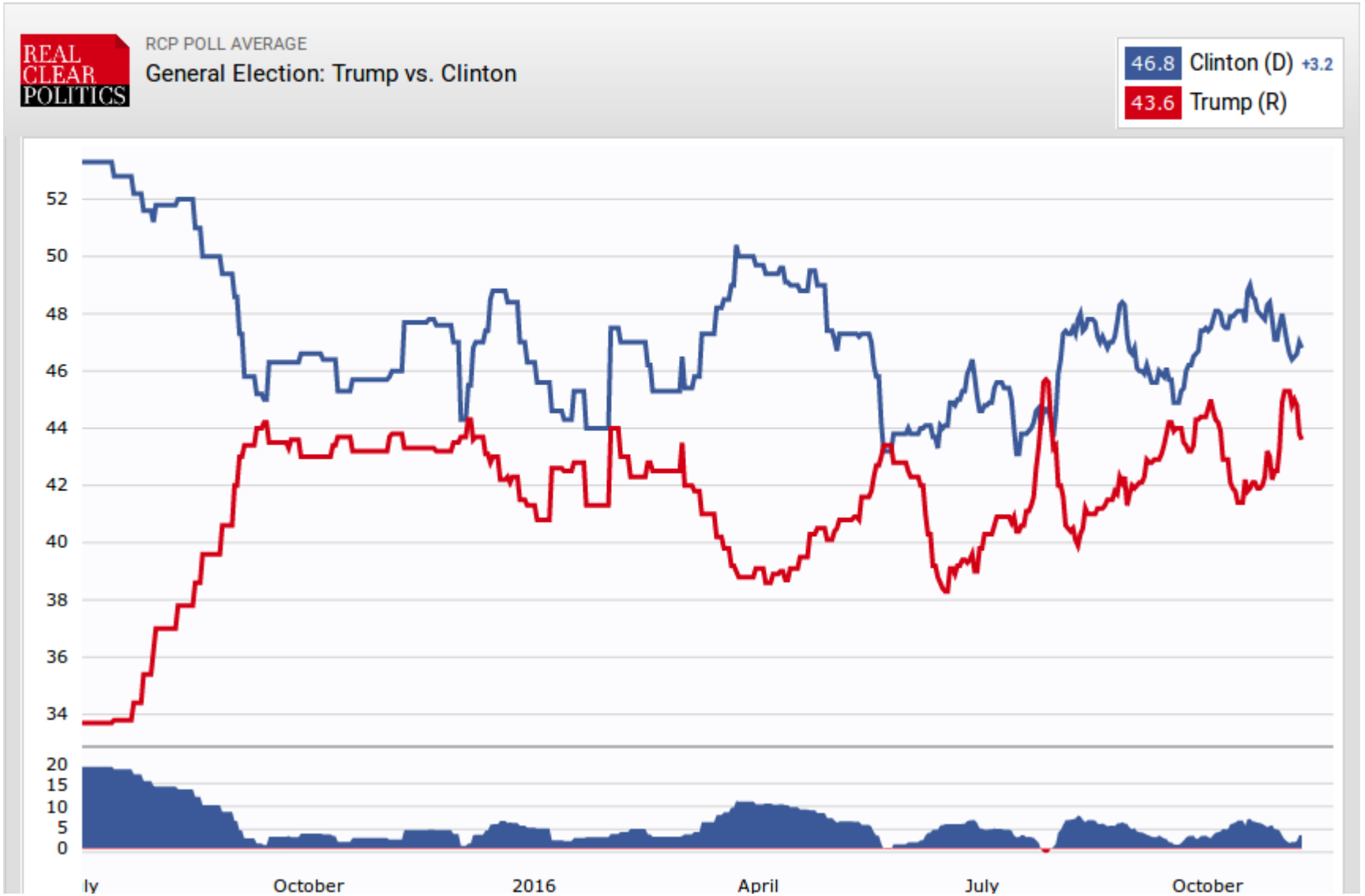
Zu beachten

- Kontext bei Erhebung
- Entscheidungen Forschungsdesign
- Messmethode und Messinstrumente
- Daten die neuen Götter!?
- Messmethode erblickt nur einen Mini-Teil des Universums

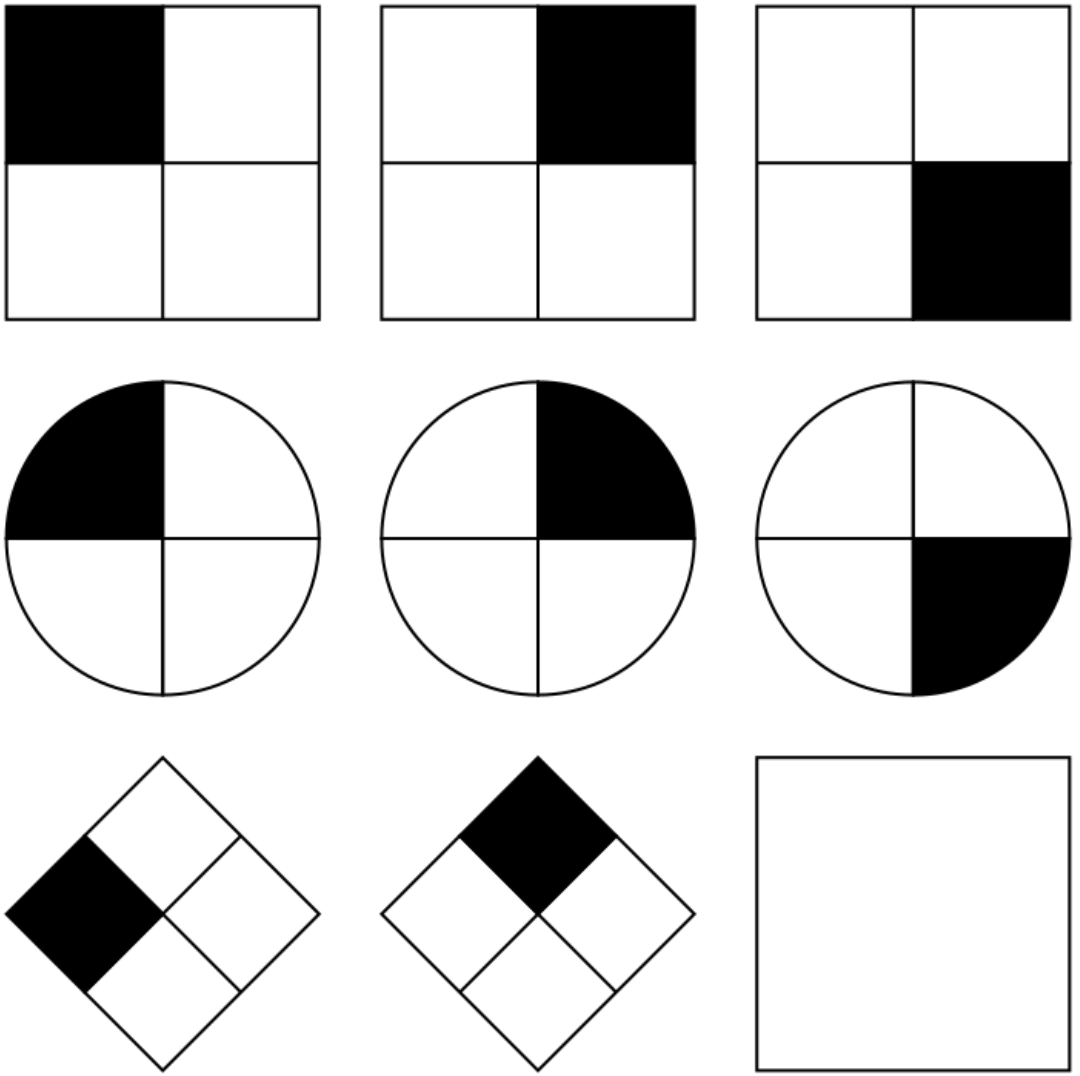
Beispiel: Text-Analyse @ Twitter



Beispiel: Wahlumfrage



Beispiel: IQ-Test



Beispiel: Fehler & Betrug

– Wikipedia

Bias

Kognitive Verzerrung (englisch cognitive bias oder cognitive illusions) ist ein kognitionspsychologischer Sammelbegriff für **systematische fehlerhafte Neigungen beim Wahrnehmen**, **Erinnern**, **Denken** und **Urteilen**. Sie bleiben meist unbewusst und basieren auf kognitiven Heuristiken.

– Wikipedia

Bias: Heuristik

Heuristik bezeichnet die Kunst, mit **begrenztem Wissen** (unvollständigen Informationen) und **wenig Zeit** dennoch zu **wahrscheinlichen Aussagen** oder **praktikablen Lösungen** zu kommen.

– Wikipedia

Bias

- Selection Bias: Trump & Twitter
- Observer Bias: Suggestion & Methoden
- Funding Bias
- Confirmation Bias
- Recall Bias: Urlaubs-Erinnerungen
- Racial Bias: Gorilla
- Gender Bias: Air Condition

https://de.wikipedia.org/wiki/Liste_von_kognitiven_Verzerrungen

WAS IST EINE DATENANALYSE?

Definition

„Eine Datenanalyse ist ein **Prozess**, welcher Daten inspiziert, bereinigt, transformiert, und modelliert, mit dem Ziel nützliche Informationen zu entdecken, Zusammenhänge zu finden, und bei Entscheidungen zu helfen.“

– Wikipedia

Definition

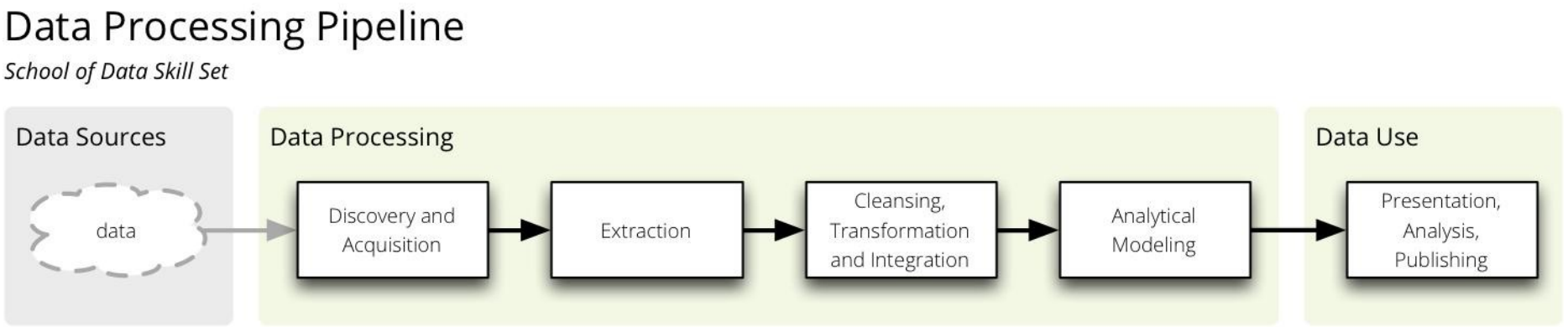
Die Datenanalyse hat verschiedene Facetten und Zugänge, umfasst diverse Techniken unter unterschiedlichen Namen, in verschiedenen Wirtschaftssektoren und Wissenschaften.

– Wikipedia

Workflow

1) Fragestellung und Anforderungen definieren
2) Daten sammeln
3) Daten bereinigen
4) Daten prozessieren
5) Daten analysieren
6) Ergebnisse interpretieren und dokumentieren

Data Pipeline



Data Pipeline: Rohdaten

- Orginalquelle der Daten
- oft schwierig zum Verwenden für die Analyse
- Rohdaten müssen zumeist 1-n Mal prozessiert werden für Analyse

Data Pipeline: prozessierte Daten

- Bereinigen von Fehlern
- Prozesse: Filtern, Gruppieren, Transformieren, Aggregieren, math. Operationen,
- Daten sind danach bereit für die Analyse

DATENQUELLEN

Datenquellen

1: nach bestehendem Datenset suchen

2: Selber sammeln

Orte für Fremd-Daten

- Internet: Websites, Repositories, API's
- Datenbanken
- Datenhändler (kaufen)
- Organisationen: Unternehmen, Universitäten, etc.

Eigene Daten

Formen der Datenerhebung:

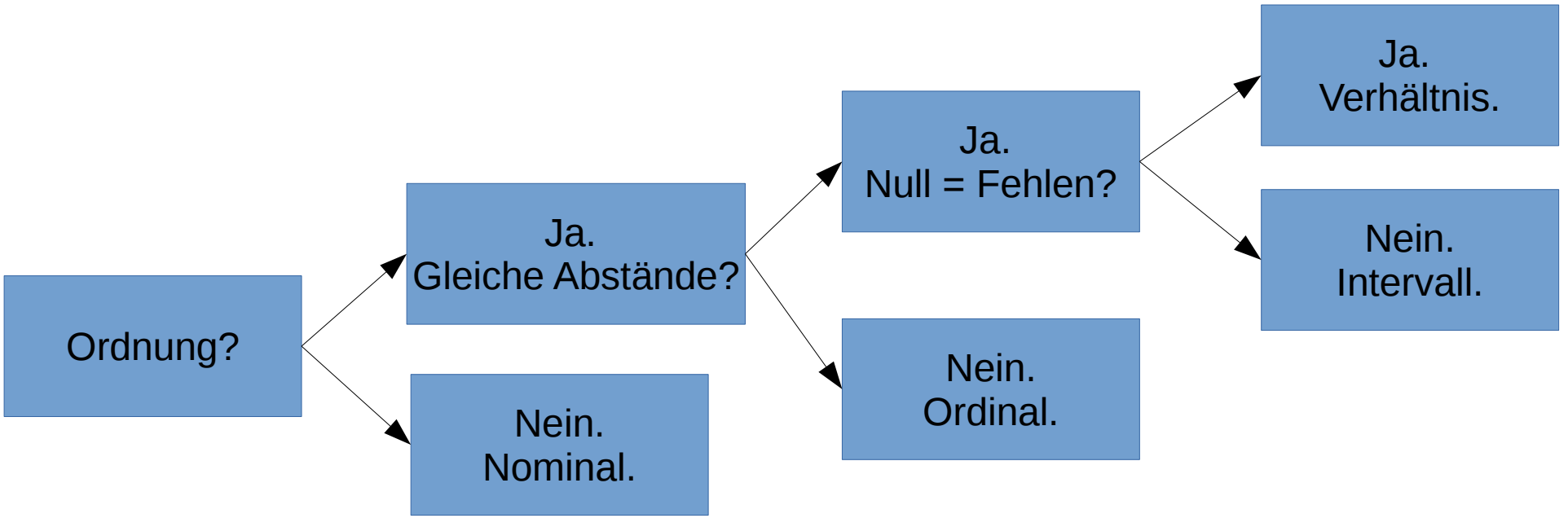
- Umfragen
- Interviews
- Beobachtung
- Recherche
- Messungen: Sensoren

DATENTYPEN

Skalierung

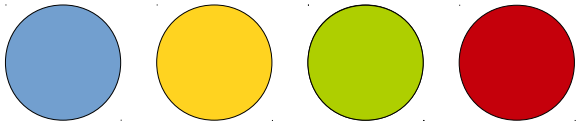
Skalenniveau	Operationen	Eigenschaften	Beispiele
Nominalskala	=, ≠	Häufigkeit	Farbe, Adresse, Beruf, Blutgruppe, Name,
Ordinalskala	=, ≠	Häufigkeit, Reihenfolge	Noten,Hotelklassen, Kleidung, Beaufortskala
Intervallskala	=, ≠, <, >, -	Häufigkeit, Reihenfolge, Abstand	Datum, Temperatur °C, IQ-Skala, Längengrad
Verhältnisskala	=, ≠,<, >, +, −, ÷	Häufigkeit, Reihenfolge, Abstand, Nullpunkt	Alter, Gewicht, Länge, Preis, Pprozent, Masse

Hilfe

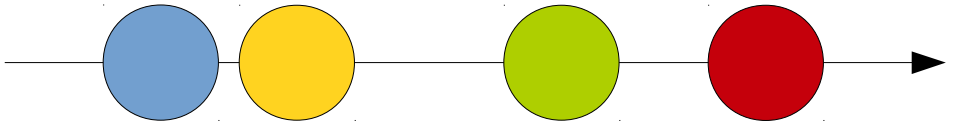

Ordnung?
Ja.
Gleiche Abstände?
Nein.
Nominal.
Ja.
Null = Fehlen?
Nein.
Ordinal.
Ja.
Verhältnis.
Nein.
Intervall.

Skalierung

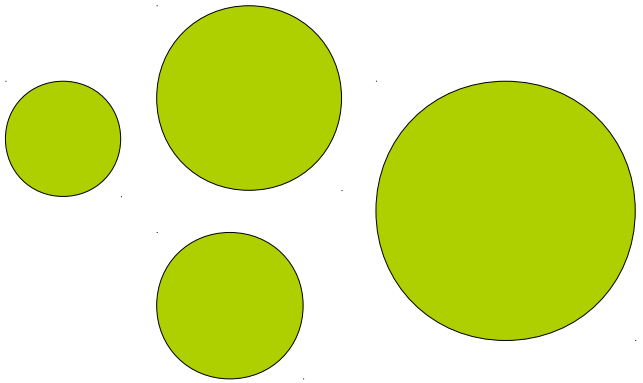
Nominal



Ordinal



Intervall oder Verhältnis



Intervall VS Verhältnis

Null Grad Celsius bedeutet nicht die Abwesenheit von Energie (kein echter Nullpunkt), und 80 Grad Celsius ist nicht zweimal so heiß wie 40 Grad Celsius (Intervall).

Intervall VS Ordinal

Rennen: Die Differenz zwischen erstem und zweiten Platz ist nicht notwendigerweise die selbe, wie zwischen zweitem und dritten Platz.

Skalierung

Skalenniveau	Operationen	Eigenschaften	Beispiele
Nominalskala	=, ≠	Häufigkeit	Farbe, Adresse, Beruf, Blutgruppe, Name,
Ordinalskala	=, ≠	Häufigkeit, Reihenfolge	Noten,Hotelklassen, Kleidung, Beaufortskala
Intervallskala	=, ≠, <, >, -	Häufigkeit, Reihenfolge, Abstand	Datum, Temperatur °C, IQ-Skala, Längengrad
Verhältnisskala	=, ≠,<, >, +, −, ÷	Häufigkeit, Reihenfolge, Abstand, Nullpunkt	Alter, Gewicht, Länge, Preis, Pprozent, Masse

Datum	Skalenniveau
"08/05/2018"	
173	
"32° 5′ 0″ N"	
True	
„Large“	

Skalierung

Skalenniveau	Operationen	Eigenschaften	Beispiele
Nominalskala	=, ≠	Häufigkeit	Farbe, Adresse, Beruf, Blutgruppe, Name,
Ordinalskala	=, ≠	Häufigkeit, Reihenfolge	Noten,Hotelklassen, Kleidung, Beaufortskala
Intervallskala	=, ≠, <, >, -	Häufigkeit, Reihenfolge, Abstand	Datum, Temperatur °C, IQ-Skala, Längengrad
Verhältnisskala	=, ≠,<, >, +, −, ÷	Häufigkeit, Reihenfolge, Abstand, Nullpunkt	Alter, Gewicht, Länge, Preis, Pprozent, Masse

Datum	Skalenniveau
"08/05/2018"	Intervallskala
173	Verhältnisskala
"32° 5′ 0″ N"	Intervallskala
True	Nominalskala
„Large“	Ordinalskala

Diskussion

1. Laptop zu und Tisch aufräumen
2. Gruppen (4-6 P) finden. Melden und Tisch suchen
3. ModeratorIn und PräsentatorIn fixieren
4. Frage 1 und 2 (je 3min) diskutieren
5. Frage Meta diskutieren (4min)
6. Präsentieren der Ergebnisse (1min)

Diskussion

Skalenniveau	Operationen	Eigenschaften	Beispiele
Nominalskala	=, ≠	Häufigkeit	Farbe, Adresse, Beruf, Blutgruppe, Name,
Ordinalskala	=, ≠	Häufigkeit, Reihenfolge	Noten,Hotelklassen, Kleidung, Beaufortskala
Intervallskala	=, ≠, <, >, -	Häufigkeit, Reihenfolge, Abstand	Datum, Temperatur °C, IQ-Skala, Längengrad
Verhältnisskala	=, ≠,<, >, +, −, ÷	Häufigkeit, Reihenfolge, Abstand, Nullpunkt	Alter, Gewicht, Länge, Preis, Pprozent, Masse

A (je 3min): Beispiele für...
Frage 1: Nominalskalierte Daten
Frage 2: Verhältnisskalierte Daten

B (je 3min): Beispiele für...
Frage 1: Ordinalskalierte Daten
Frage 2: Intervallskalierte Daten

Sum Up (3min):
Frage 3: Wie ist es gelaufen und welche Probleme gab es?

Computer

- Boolsch (boolean): true/false
- Numerisch (numeric): float, integer,
- Zeichenketten (string)
- zusammengesetzte Strukturen: Listen (Arrays), Dictionaries

Datum	**Skalenniveau**	**Computer**
"08/05/2018"	Intervallskala	
173	Verhältnisskala	
"32° 5′ 0″ N"	Intervallskala	
True	Nominalskala	
„Large“	Ordinalskala	

Computer

- Boolsch (boolean): true/false
- Numerisch (numeric): float, integer,
- Zeichenketten (string)
- zusammengesetzte Strukturen: Listen (Arrays), Dictionaries

Datum	**Skalenniveau**	**Computer**
"08/05/2018"	Intervallskala	String
173	Verhältnisskala	Integer
"32° 5′ 0″ N"	Intervallskala	String
True	Nominalskala	Boolean
„Large“	Ordinalskala	String

Dimension

- räumlich (spatial)
- zeitlich (temporal)
- thematisch (domain)

Datum	Skalenniveau	Computer	Dimension
"08/05/2018"	Intervallskala	String	
173	Verhältnisskala	Integer	
"32° 5′ 0″ N"	Intervallskala	String	
True	Nominalskala	Boolean	
„Large“	Ordinalskala	String	

Dimension

- räumlich (spatial)
- zeitlich (temporal)
- thematisch (domain)

Datum	Skalenniveau	Computer	Dimension
"08/05/2018"	Intervallskala	String	temporal
173	Verhältnisskala	Integer	domain
"32° 5′ 0″ N"	Intervallskala	String	spatial
True	Nominalskala	Boolean	domain
„Large“	Ordinalskala	String	domain

DATASET

Datasets

„Ein Datensatz ist eine Gruppe von inhaltlich zusammenhängenden (zu einem Objekt gehörenden) Datenfeldern.“

	A	B	C	D	E	F	G	H
1	wkurz	sprengel	ptname	ptlang	listenplatz	gesamt	unguel	gueltig
2	GRGRAZ03	101	ÖVP	Österreichische Volkspartei	1	277	1	276
3	GRGRAZ03	101	SPÖ	Sozialdemokratische Partei Österreichs	2	277	1	276
4	GRGRAZ03	101	FPÖ	Freiheitliche Partei Österreichs	3	277	1	276
5	GRGRAZ03	101	GRÜNE	Die Grünen - Die Grüne Alternative	4	277	1	276
6	GRGRAZ03	101	KPÖ	Kommunistische Partei Österreichs	5	277	1	276
7	GRGRAZ03	101	GVP	Grazer Verkehrspartei	6	277	1	276
8	GRGRAZ03	101	RWA	Reif für die Wirtschaft und Arbeit	7	277	1	276
9	GRGRAZ03	101	LIF	Liberales Forum	8	277	1	276

Datasets: Struktur

a) strukturierte Daten
b) semi-strukturierte Daten
c) unstrukturierte Daten

Datasets: Struktur

Topologische Struktur:

- Tabellen: Datenbank, CSV
- Texte
- Netzwerke / Graphen
  - Bäume
- Hierarchische Daten

DATEIFORMATE

Definition

Ein Dateiformat ist ein Standard, in welchem Informationen zum Speichern auf einem Computer encodiert werden. Es regelt wie die Zeichenketten (Bytes) aneinandergereiht werden und so die Information in sich tragen.

Proprietär VS Open

Die Dateiformate können entweder frei, also von allen einlesbar (z. B. XML, JSON), oder proprietär und somit nur mit der passenden Software nutzbar sein (z. B. PDF).

Proprietär VS Open

Ein offenes Format ist eine publizierte Spezifikation zum Speichern digitaler Daten, welche ohne rechtliche oder technische Einschränkungen genutzt werden kann.

Warum offen?

- Interoperabilität: einfaches Austauschen
- für Mensch und Maschine verständlich
- offener Standard auf dem aufbauend allen es möglich ist Anwendungen dafür zu entwickeln
- Vendor Lock-In nicht möglich.

Wichtig für uns

- CSV
- Markdown
- JSON
- HTML/XML

→ alle offen

Übung

1. Sucht euch 5 verschiedene Dateitypen auf eurem Computer oder aus dem Inet.
2. Checkt in einem Editor, ob:
   a) offen oder proprietär, und
   b)welche Struktur die Daten haben.

Zeit: 5min

Metadaten

Definition

Metadaten oder Metainformationen sind strukturierte Daten, die Informationen über Merkmale anderer Daten enthalten.

Warum?

- Interoperabilität
- Maschinenlesbarkeit
- Informations-Suche
- Weiterverwendung

Wie?

Sich Frage stellen: Was braucht eine andere Person um die Daten verstehen zu können?

- eindeutige Variablennamen
- selbsterklärende Variablennamen

- insbesondere Daten zu:
  - Beobachtung: Datum, Ort, Methode, Sample Size, Frequenz, Sensor, Messgenauigkeit, etc.
  - Sozial: Person, Institution, Kontakt, Projektname
  - Datei: Typ, Datum, Datenset Version, Metadaten Version, Lizenz,

Beispiel Buch

- AutorInnen
- Publikations-Datum
- ISBN
- Kurz-Beschreibung
- Auflage
- Verlag
- UrheberInnenrechts-Lizenz

Typische Metadaten

- Titel
- Beschreibung
- Identifier
- Aktualisierungs-Zyklus
- Erstellungs-Datum
- AutorIn
- Messmittel mit Genauigkeit
- Version
- Lizenz
- Publisher
- Zeitlicher Raum
- Geographischer Raum
- Schlagwörter
- Metadaten Standard
- Encoding
- Sprache
- URI

RECHTLICHES

Datenschutz

- neue EU DSGVO ab 25. Mai
- Zweckmäßige Verwendung
- Personenbezogene Daten aufpassen (Name, IP, Email,…).

→ TODO: mit Rechtsabteilung sprechen

UrheberInnenrecht

- Regelt die Verwertungs- und Werknutzungsrechte
- Wird meist über Lizenzierungen des Werkes abgewickelt (Du darfst das Foto im Format Din A2 drucken und verkaufen, dafür zahlst du pro Stück 12.50€).
- Rechte können auch direkt geklärt werden.

→ TODO: mit Rechtsabteilung sprechen

UrheberInnenrecht

1. Fremde Daten nutzen:

- schauen ob UrheberInnenrechts-Lizenz eine Nutzung erlaubt:
  - nein: nicht verwenden bzw. wegen Nutzung direkt anfragen.
  - ja (= offene Lizenz): Verwendung mit erfüllen der Lizenz-Bedingungen (z. B. Werks-InhaberIn nennen).

UrheberInnenrecht

2. Eigene Daten erstellen:

- ihr seid UrheberIn: alle Rechte bei euch
- mit Rechtsabteilung sprechen
- Lizenz wählen
- eventuell Zustimmung einholen
- Lizenzieren

Kontakt

www.offenewahlen.at
@stefankasberger
stefan.kasberger@okfn.at
www.ofkn.at

UrheberInnenrecht:

Markenrecht: