data basics

- observations, variables, and data matrices
- types of variables
- relationships between variables

Dr. Mine ,etinkaya-Rundel
Duke University

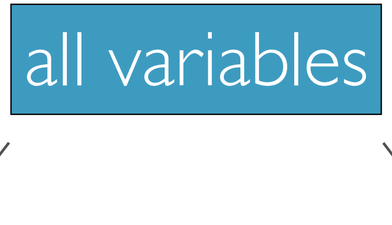
data matrix

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

observation (case)

variable

types of variables



numerical

(quantitative)

take on numerical values

sensible to add, subtract, take averages, etc. with these values

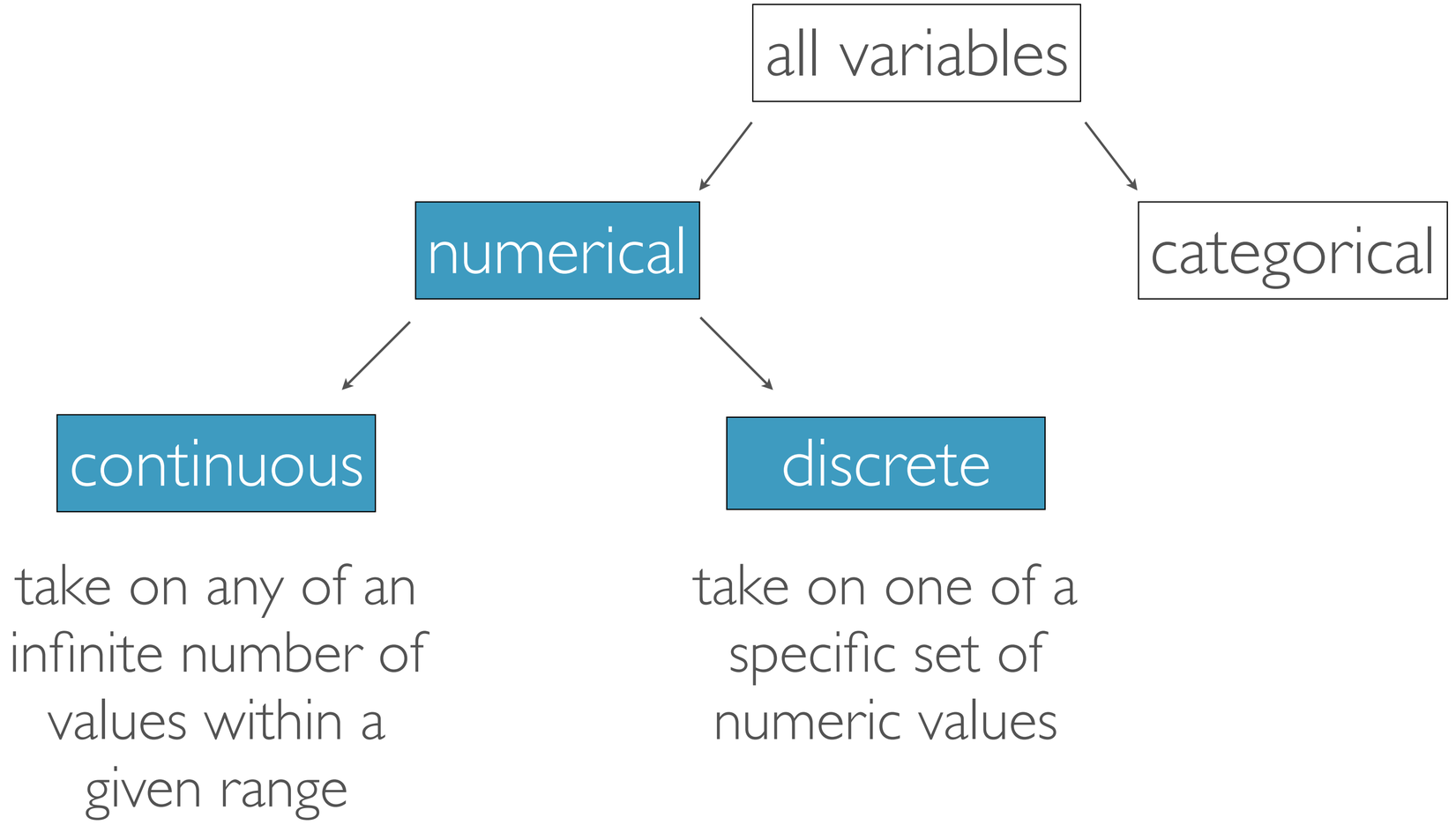
categorical

(qualitative)

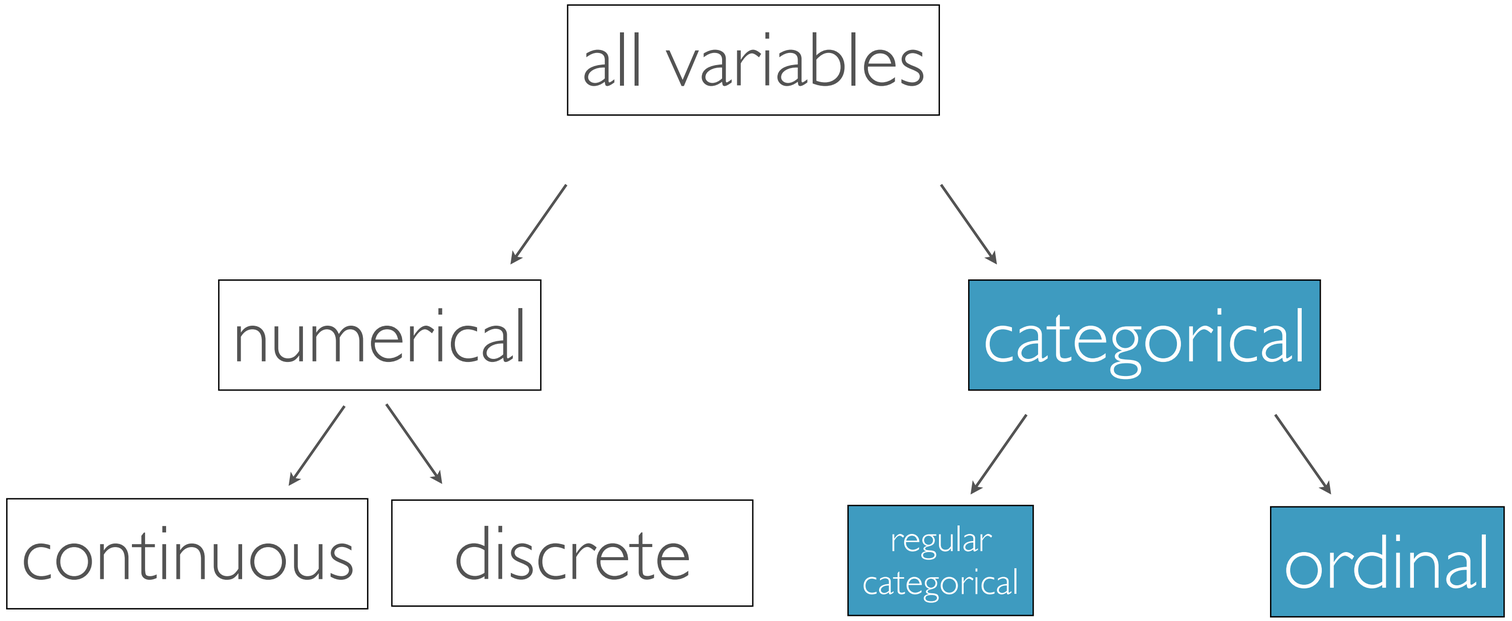
take on a limited number of distinct categories

categories can be identified with numbers, but not sensible to do arithmetic operations

numerical variables



categorical variables



levels have an inherent ordering

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

`country`: Name of the country

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

`cr_req`: Number of content removal requests made to Google

discrete numerical

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

`cr_comply`: Percentage of content removal requests Google complied with

continuous numerical

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

`ud_req`: Number of user data requests as part of a criminal investigation

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

`ud_comply`: Percentage of user data requests Google complied with **continuous numerical**

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

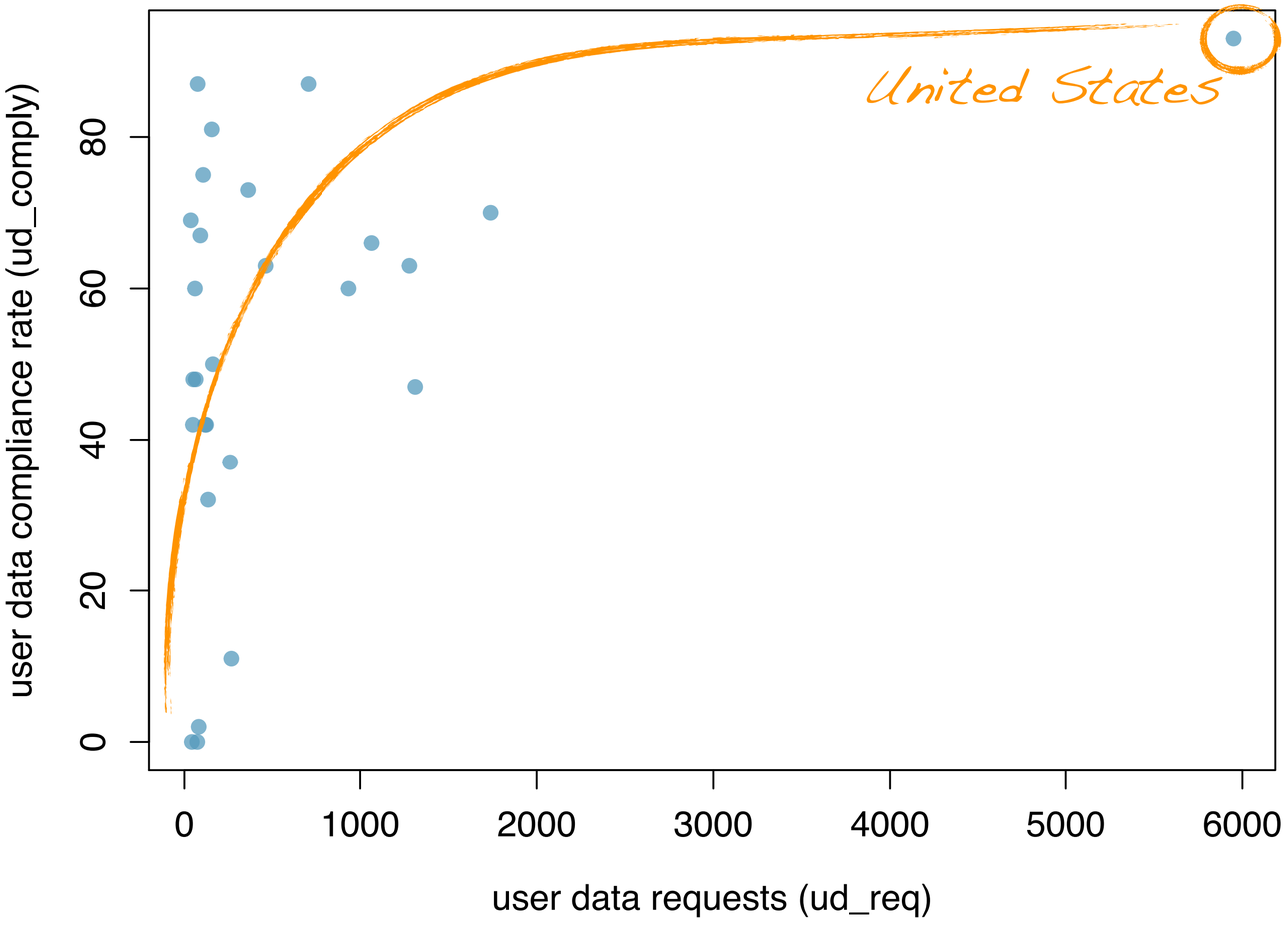
categorical

`hemisphere`: Hemisphere that the country is located in (southern, northern)

country	cr_req	cr_comply	ud_req	ud_comply	…	hemisphere	hdi
Argentina	21	100	134	32	…	southern	very high
Australia	10	40	361	73	…	southern	very high
Belgium	<10	100	90	67	…	northern	very high
Brazil	224	67	703	82	…	southern	high
…	…	…	…	…	…	…	…
United States	92	63	5950	93	…	northern	very high

`hdi`: Human Development Index
(very high, high, medium, low)

relationships between variables



- Two variables that show some connection with one another are called associated (dependent)
- Association can be further described as positive or negative
- If two variables are not associated, they are said to be independent