Digitise, Optimise, Visualise: Data

Peter H. Gruber

July 1-5, 2019

▷ Intro
Intro
Partitions
Terminology
Random variables
What is data? Classical and alternative data
Main objective

Why data is useful

- Gold standard of every scientific endeavour.
- Entire industry around data: Google, Facebook, Bloomberg, ...
- Financial data industry is 28.5 billion dollars [Burton-Taylor 2017]
- New data sets $\rightarrow$ research and business opportunities

Why data is problematic

- Recorded, processed, transferred and converted by humans $\rightarrow$ inevitable errors
- Usually problematic: faulty, incomplete, censored, survey-based
- Most data problems are not IT problems
- Check sources, keep audit trail for any data usage

Intro
▷ Intro
Partitions
Terminology
Random variables
What is data? Classical and alternative data
Main objective

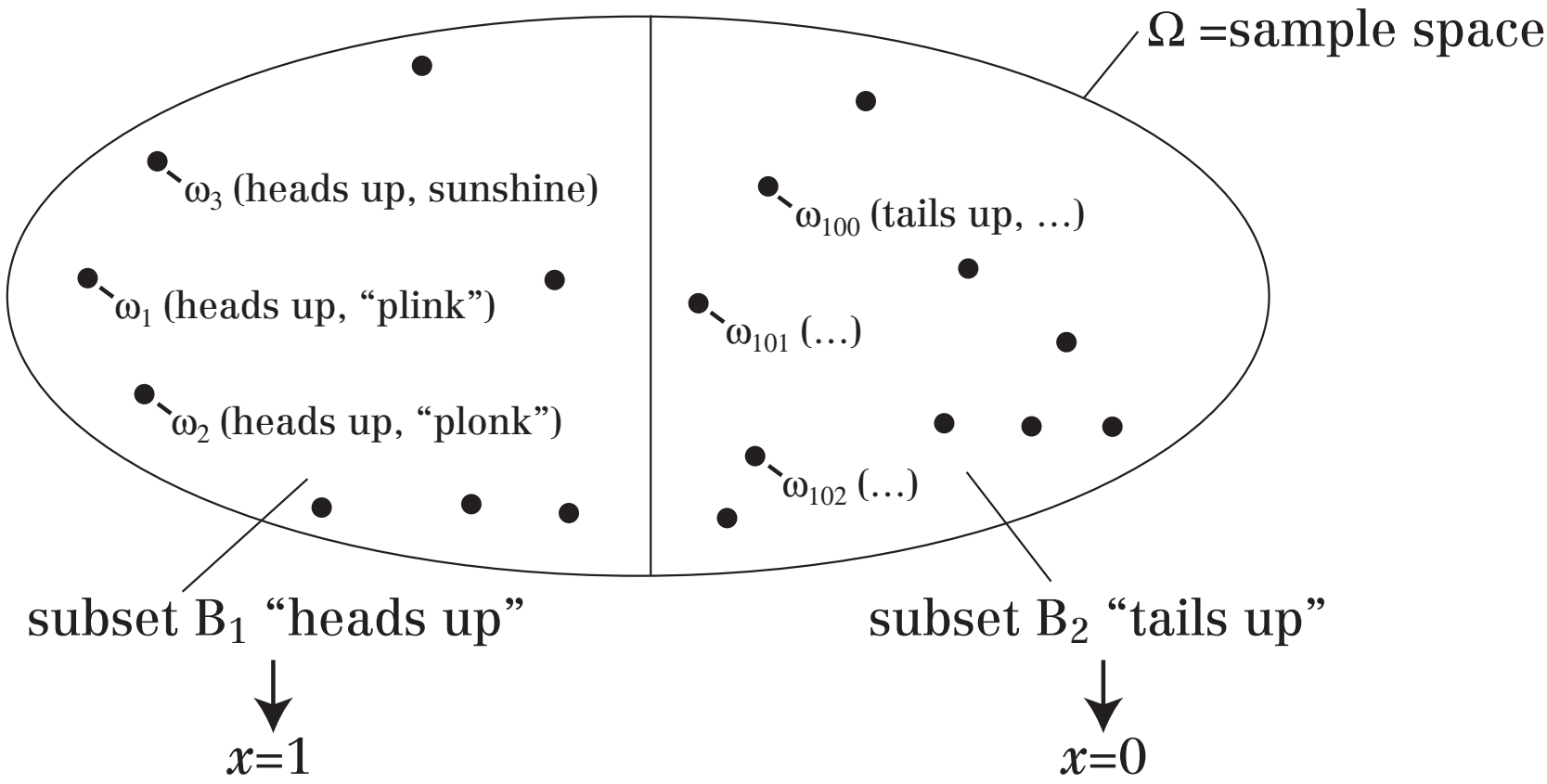
Data is power

- "That which is measured, improves" (K. Pearson or P. Drucker)
- Dickey Amendment (1996)
- Open and crowd data movements

Partitions

Intro
Intro
▷ Partitions
Terminology
Random variables
What is data? Classical and alternative data
Main objective

Spoiler. Simply put, a random variable is a function that assigns a real number to every possible state of nature.



Terminology

Intro
Intro
Partitions
▷ Terminology
Random variables
What is data? Classical and alternative data
Main objective

Sample space Ω. Set of all possible (relevant) events.

States of nature. Each possible outcome = state of nature, (ω_i).
Finite $(i \in \{1, 2, \ldots, N\})$ or infinite $(i \in \mathbb{N})$ number.

Partitions of Ω. Collection of subsets $\mathcal{P} = \{B_1, \ldots B_n\}$.
Two rules ("pizza slicing rules"):

1. Don't forget a part (or $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$)
2. Dont't count a part twice (or $B_i \cap B_j = \{\} \quad \forall i \neq j$).

NB: different $\mathcal{P}$ exist for every Ω. $\leftarrow$ choice is researcher's job

Sigma algebra $\mathcal{F} = \sigma(\mathcal{P})$ = set of subsets of Ω.
Rules ("pizza dish rules")

1. $\Omega \in \mathcal{F}$
2. $B \in \mathcal{F}$ implies $B^c \in \mathcal{F} \leftarrow$ thus $\{\} \in \mathcal{F}$
3. All unions of B_i are also elements of $\mathcal{F}$

Random variables

Intro
Intro
Partitions
Terminology
▷ Random variables
What is data? Classical and alternative data
Main objective

Measurability. Function $f : \Omega \to \mathbb{R}$ is measurable w.r.t. $\sigma(\mathcal{P})$ if the value of f is the same for all states of nature (ω) in a given B_i.

Note: $f(\cdot)$ does not have to take distinct values for every B_i.
Constant function $f(\omega) = 1$ is measurable w.r.t any σ-algebra.

Random variable. A measurable function from $(\Omega, \mathcal{F})$ to $\mathbb{R}$.

Interpretation: σ-algebra $\mathcal{F}$ determines how detailed our knowledge of the real world can be, given the result x of a random draw.

– Best: infer from x to a specific B_i.
– Sometimes: only infer to a set of B_i
– Never: more detailed information than element of partition $\mathcal{P}$.

Intro
Intro
Partitions
Terminology
Random variables
▷ What is data?
Classical and alternative data
Main objective

Data = collection of measurements of a property of an entity/individual.

Many sources of errors.

Classical and alternative data

Intro
Intro
Partitions
Terminology
Random variables
What is data?
Classical and alternative data
Main objective

Classical data

- Macro: GDP, consumption, employment, trade, ...
- Macro-Finance: inflation, interest rates, exchange rates, ...
- Micro: Socio-economic panel, education, health, social services
- Finance
  - Base: balance sheet, valuation, geography, people, ...
  - Aggregate: investments, fund flows, holdings, ...
  - Transact: price/volume/time of stocks/bonds/derivatives ...
  - Survey: analyst recommendations, prof. forecasters ...

Alternative data gains importance

- Physical world (satellites, electricity use, parking utilisation)
- Disclosure (firms, central banks) + language
- News
  - Traditional (papers, TV)
  - Alternative news and opinion (Twitter, Facebook)

Intro
Intro
Partitions
Terminology
Random variables
What is data? Classical and alternative data
▷ Main objective

Main objective: Replicability.

(Easier said than done.)