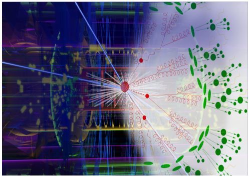
Data, Models, and Reality

Frank Krauss

Institute for Data Science
Institute for Particle Physics Phenomenology
Durham University

Victoria University, Wellington, 7.12.2023

disclaimer: my background



- many aspects of this talk outside my "core" competence
  (theoretical particle physicist by training)
- background in high-precision modelling for Large Hadron Collider
- relatively simple simulation task:
  first-principles theory, data-rich environment, high-quality data
- code: SHERPA (250,000 lines in public release, about 20,000 CPU years per year of simulation run by users)
- used to analyse data (by comparison with theory)
  (the work by our experimental colleagues)
- used to suggest new analyses/analysis strategies
  (routinely using ML techniques)

Outline

Data Science in Context

two (extreme) views of data science: goal-driven

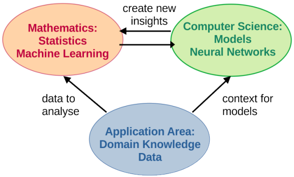
("how to extract knowledge from data")

- data science = statistical data analysis + computational methods
- self-contained field of study and research:
  "all data are created equal"

  (and are equally "good" - blind to provenance)

- relatively "blind" to domain knowledge
- challenge: select/create best suited method
  for data type and range:

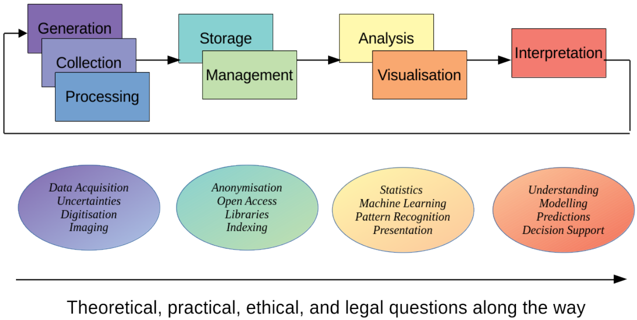
  e.g. training vs. validation



two (extreme) views of data science: method-driven

("how to treat data")

- data science = solutions in life-cycle of data
- mandatory collaboration of technology and domain knowledge



example applications of data science: bird's eye view

- purist (particle theory): near perfect, well-understood data
  mainly statistical interpretation and parameter fitting

  (e.g. discovery of new particle according to pre-defined statistical threshold . . .)

- opportunist (amazon): good data, not particularly well understood
  mainly pattern detection and optimisation of choices

  (important factor here: cost-benefit of storage, analysis, . . .)

- pragmatist (public health): messy data, often badly understood
  mainly understanding reality and models for decision support

  (challenges: provenance, quality, and context of data; complex, hard-to-model reality)

Monitoring Plant Health

coffee in Thailand

(work by my Durham colleague Anthony Brown)

- valuable traded good: coffee

  (total value about $36B)

- stable cash-crop for many LMICs
- threatened by coffee-leaf-rust (CLR): not curable, highly contagious
- infected plants must be quickly identified and destroyed
- plantations often on steep hills: impediment to inspection

  $\Longrightarrow$ UAVs (drones)

coffee in Thailand: bespoke solution

(work by my Durham colleague Anthony Brown)

green control | rust | green rust | arabica/geisha

- "different greens" – that will do
- quantitative identification through spectral analysis of reflected light
- database with 100's of labelled/identified reference spectra

(this data acquisition is the "manual" part of the project)

- PCA + ML $\longrightarrow$ 4 critical wavebands
- lovely, BUT …

coffee in Thailand: bespoke solution

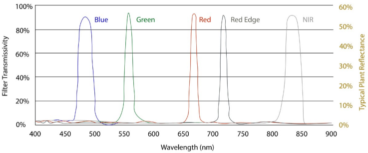
(work by my Durham colleague Anthony Brown)

green control

rust

green rust

arabica/geisha



- typical pass bands of commercial multi-spectral cameras

coffee in Thailand: bespoke solution

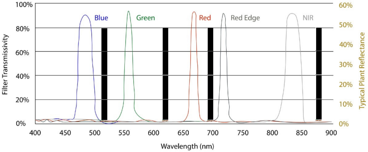
(work by my Durham colleague Anthony Brown)

green control

rust

green rust

arabica/geisha

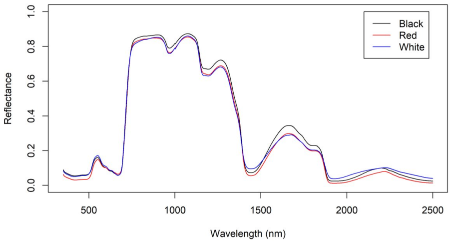


- typical pass bands of commercial multi-spectral cameras
- not covering critical regions
- need to build bespoke camera: mobile-phone cameras plus filters
- tests in the field as next step

mangrove surveying and identification in Suriname

(work by my Durham colleagues Anthony Brown and Isabella Bovolo)

- protecting and stabilising coastlines
- contributor to biodiversity
- but: threatened by climate change
  $\implies$ need to monitor
- three species with subtle differences in Suriname: black, red, white



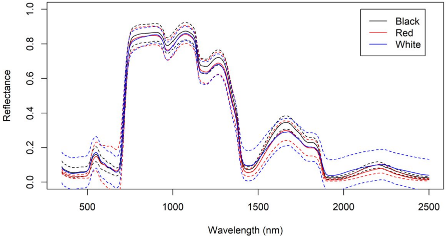
- multi-spectral, encore?

(tested tool-chain)

mangrove surveying and identification in Suriname

(work by my Durham colleagues Anthony Brown and Isabella Bovolo)

- protecting and stabilising coastlines
- contributor to biodiversity
- but: threatened by climate change
  $\Longrightarrow$ need to monitor
- three species with subtle differences in Suriname: black, red, white



- multi-spectral, encore?

(tested tool-chain)

- insufficient discriminatory power
  3-D point cloud $\leftrightarrow$ shapes?
- tests underway

Modelling Reality:

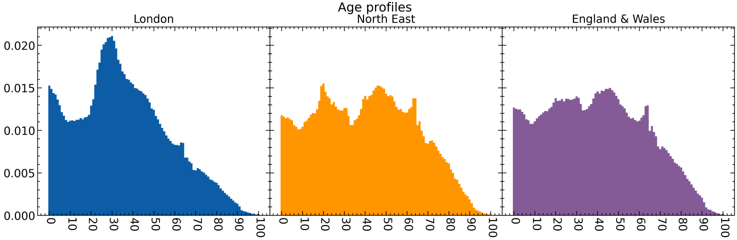
June & COVID-19 in England

motivation: why granularity matters

- impact of COVID=19 highly age-dependent

$\longrightarrow$ **need geographical granularity for regional planning**

(coincidence: Durham hosts & maintains England & Wales census data of past decades)

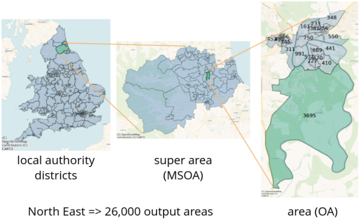


example data inputs: demographics

- last census (2011)

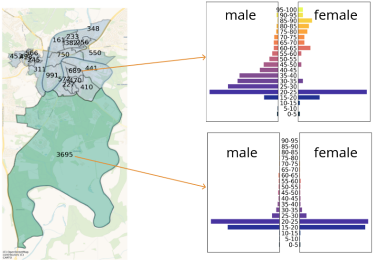
  (data freely available from Office for National Statistics)

- hierarchical data structure



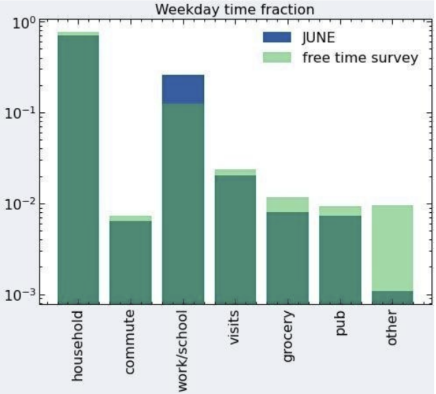
- OA's with $\sim$ 250 residents, with similar characteristics

- build virtual population in OA: age, gender, ethnicity, deprivation index
- example: Durham



example data inputs: daily activities

- time spent on activities known from ONS surveys (this changes under lock-down)
- translate into age-dependent probabilities for activities



example data inputs: social mixing matrices

- social mixing matrices from PolyMod and BBC Pandemics project

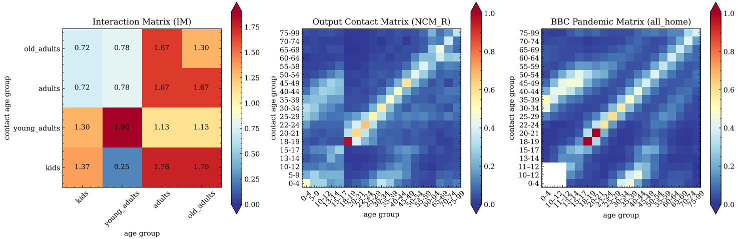
J.Mossong et al., PLoS Med 5(3) e74, https://doi.org/10.1371/journal.pmed.0050074;
P.Klepac et al., https://www.medrxiv.org/content/10.1101/2020.02.16.20023754v2

- denote number of contacts of person with age i with person of age j
- tricky: averages over full population (good for compartment models)
- broad agreement with input from surveys: important closure test

(in June contacts also depend on composition of environment)

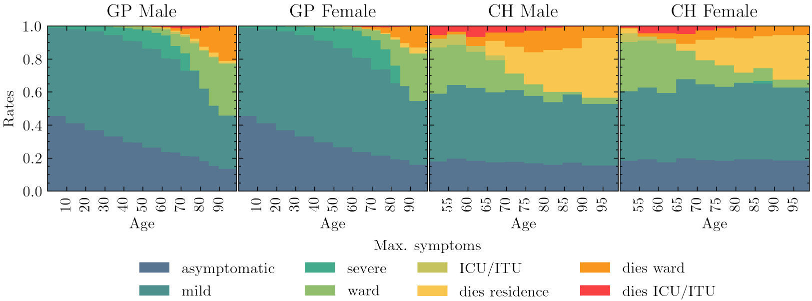
- example: household interactions vs. BBC pandemics project

(census has 4 categories of residents: kids, young adults, adults, old adults)

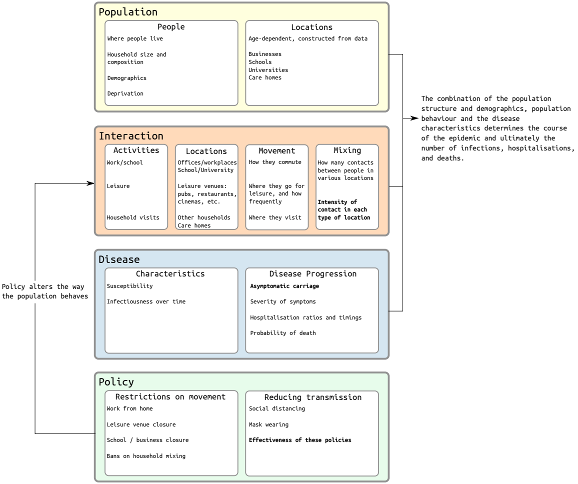


example data inputs: outcomes of infection

- tiring data-mining exercise with inconsistent and often contradictory data
- extra difficulty: include care homes (CH) vs. general population (GP)

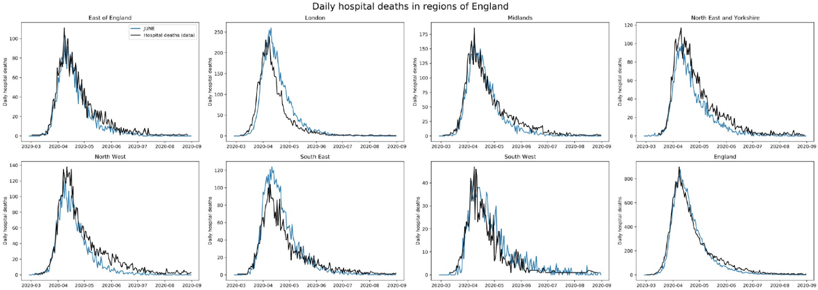


JUNE simulation content - summary



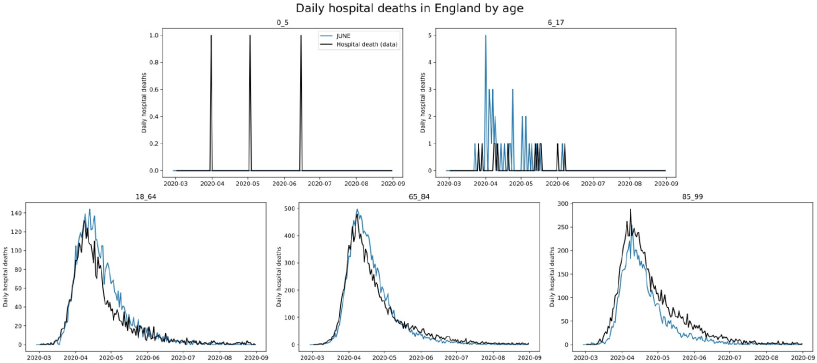
Results for 1^{st} wave: fatalities

- 1^{st} wave: deaths in hospitals - regional distribution



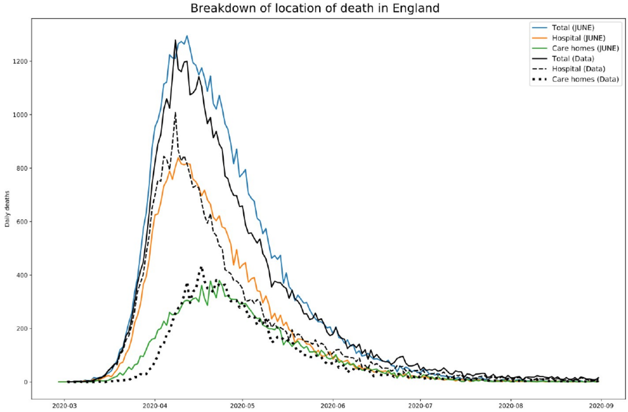
Results for 1$^{\rm st}$ wave: fatalities

- 1$^{\rm st}$ wave: deaths in hospitals - age distribution



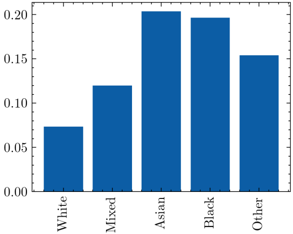
Results for 1$^{\text{st}}$ wave: fatalities

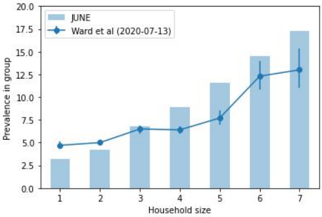
- 1$^{\text{st}}$ wave: all deaths - distribution of location

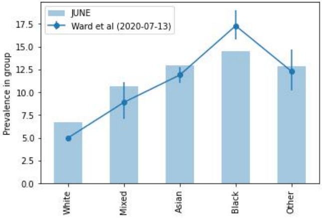


Results for 1$^{\text{st}}$ wave: social imbalances

- look at cumulative infection rates until July 2020 in dependence on
  - household size
  - ethnicity
- *nota bene*: all imbalances only due to differences encoded in census data







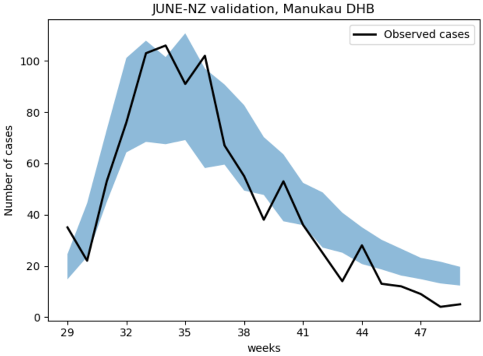
A spin-off: measles in New Zealand

(from a collaboration with ESR New Zealand)

- ESR team adapted JUNE to measles in New Zealand

(population & disease characteristics)

- validated model
- projected results from different vaccination regimes

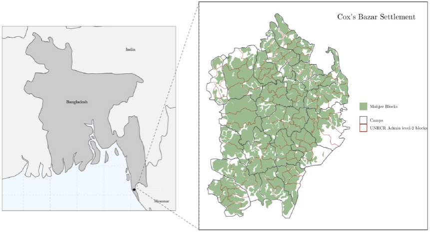


Modelling Reality:

Epidemics & JUNE in Cox's Bazar

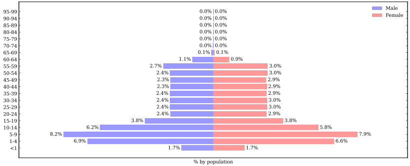
background: Cox's Bazar

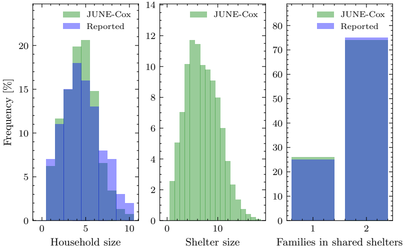
- largest settlement in the world
- in some areas, the settlement is denser than New York City
- high risk of COVID transmission



input data: demographics

- high-quality data thanks to WHO census
- need to adapt demography & distribute over households (=shelters)



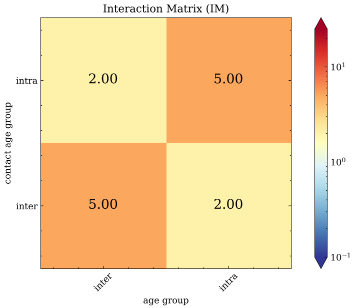


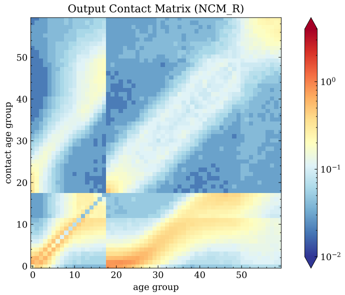
input data: social interactions

- no data available $\Longrightarrow$ mixed method approach: questionnaire in simplest categories + digital twin of population

(this is the first attempt ever in a refugee/IDP camp setting! important input for compartment models)

- different places for interactions: shelters (see below), distribution centres, communal kitchens, pump & latrines, mosques, etc.





input data: infer health impacts

- no data available $\implies$ infer from UK data
- need to account for difference in life expectancy (model!)

$$A_P = \begin{cases} A, & \text{if } A \leq A_{\text{cut-off}} \\ (A - A_{\text{cut-off}}) \left[\frac{LE(\mathrm{sex}) - A_{\text{cut-off}}}{LE_{\mathrm{uk}}(\mathrm{sex}) - A_{\text{cut-off}}} \right], & \text{if } A > A_{\text{cut-off}} \end{cases}$$

with A = age and LE = life expectancy.

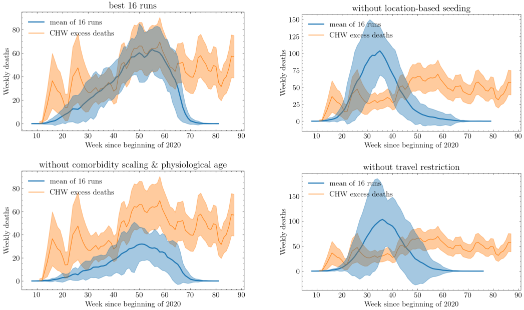
- need to account for co-morbidities in Cox' Bazar (CB):

$$P_{\text{CB}}(\mathrm{severe} \,|\, c,\, \mathrm{age},\, \mathrm{sex}) = \gamma\, \lambda_c\ P_{\text{UK}}(\mathrm{severe} \,|\, \mathrm{age},\, \mathrm{sex})$$

with γ overall scaling and λ_c risk multiplier

example results for wild-type (until March 2021)

- identify deaths in various ways: excess deaths (when camp was closed down) or "certified" by trained health visitors/workers
- wild-type was slowly replaced by Delta variant at about week 60



Summary & Outlook

summary

- data science
  - important addition in the scientific canon:
    permeating all fields of research: (nearly) everything is data
  - we live in the era of data: data science is here to stay
  - important to treat it with professional respect
- showed $2\frac{1}{2}$ applications of data science modelling:
  - monitoring of plant health and early warning of pathogens
  - large-scale modelling for public health

(direct ramifications as decision support for governments etc.)

some final thoughts on (data) science

- language: parametrizations vs. models of reality

  (black box vs. grey box or description vs. understanding)

- intellectual ownership: provenance, quality, meaning of data

  (added value of results without context/interpretation)

- uncertainties: how to have robust estimates

  (importance for decision support: necessity to estimate risk vs. reward)

- accuracy vs. precision

  (they are not the same! you can be precisely wrong . . .)

final-final thought