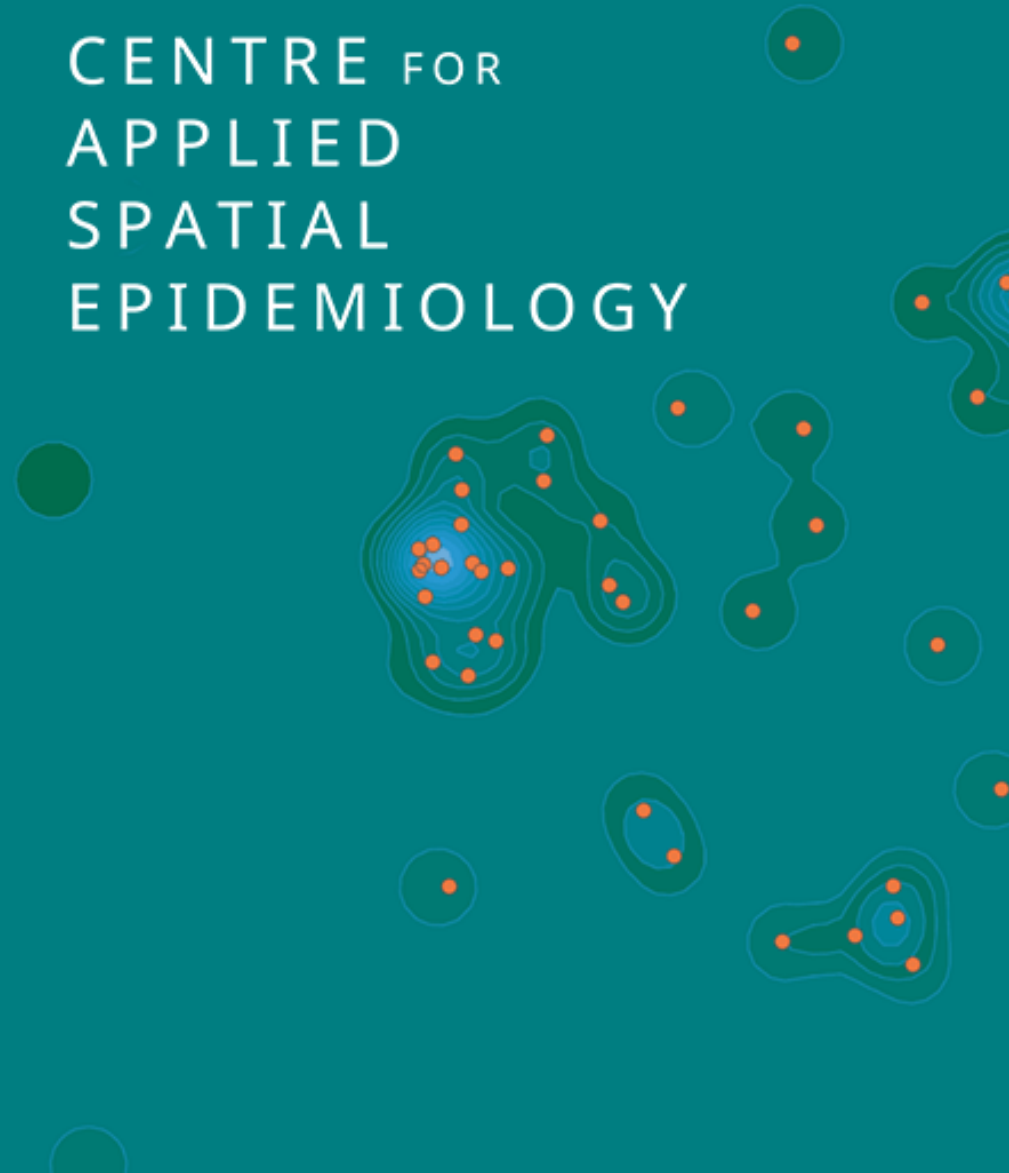
# Pakistan TB hackathon

**Sandra Alba**
**KIT Epidemiologist**

**TB MAC WHO Annual Meeting**
**Istanbul 2 October 2019**

KIT Royal Tropical Institute

CASE
CENTRE FOR APPLIED SPATIAL EPIDEMIOLOGY



# Tuberculosis hackathon

Partnership between

- Pakistan's National Tuberculosis Control Program
- KIT Centre for Spatial Epidemiology (CASE)

With the support of the StopTB Partnership

CASE
CENTRE FOR APPLIED SPATIAL EPIDEMIOLOGY

TOGETHER 6
National TB Control Program

Stop TB Partnership

KIT Royal Tropical Institute

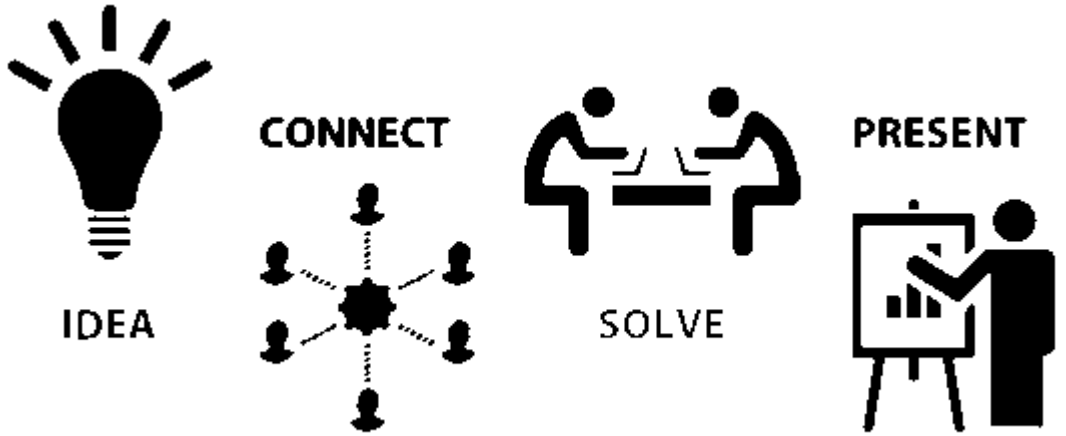
# What is a hackathon?

#millennials love #hackathons and #junfkood

- Collaborative problem solving event – with a competitive twist
- Originally: computer programmers and others involved in software development collaborate intensively on software projects.
- Goal: create a functioning product by the end of the event.

# Why hackathons are conducted

- Increasingly proposed as a model to solve complex problems
- Leveraging mixed skills of a group of people.
- Core idea: Collaborate to co-create something greater than the sum of the parts.






KIT Royal Tropical Institute

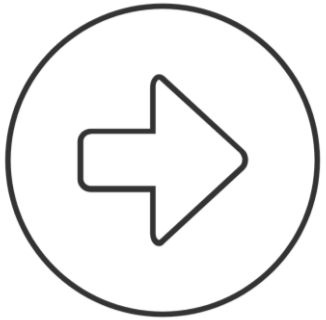
# Tuberculosis hackathon challenge

Sub-national TB estimates

- Valuable for NTPs to optimize resource allocation
- Pakistan NTP among many programs worldwide who value this data, but do not have reliable estimates.
- Find missed people with TB: Who are they and where are they?

TB prevalence surveys

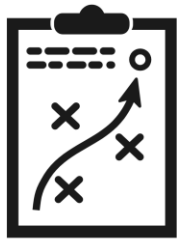
- TB prevalence surveys usually powered to provide national prevalence estimates.
- Yet also contain a wealth of sub-national information that could benefit decision makers and populations.



*Can we use TB prevalence survey data to predict sub-national TB burden in Pakistan?*

# Objectives

*Bring together research groups to collaborate on a joint modelling exercise to estimate sub-national TB burden in Pakistan in 2018*



**NTP perspective:**

Provide Pakistan NTP with data to tailor their TB control efforts to different sub-national contexts.

**TB modelling perspective:**

Learn from each other by comparing different methods and further refine our approaches.

**Hopefully everyone:**

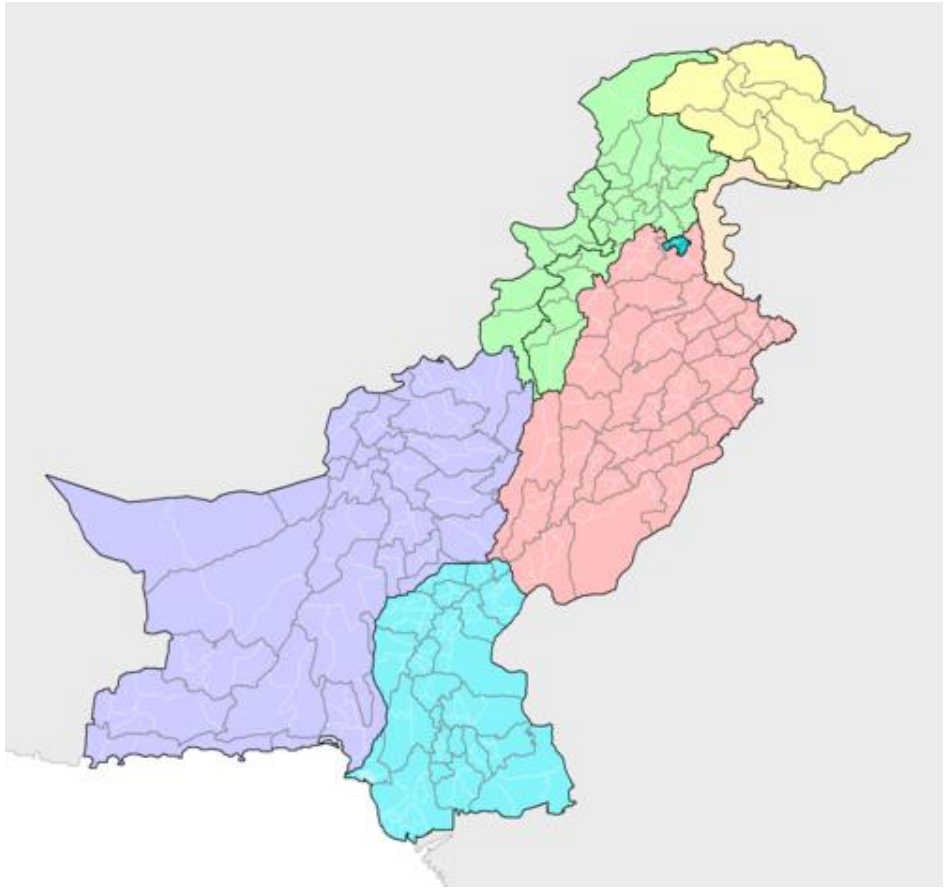
Have fun!

# In practice

- Pakistan NTP provided their 2010-2011 TB prevalence survey data and other TB program data

- Virtual hackathon: each group (or individual) was given three months and the same set of data to develop their own models to estimate sub-national TB burden.
  - District-level
  - Bact+ among adults (>15yrs)

- Participants were invited to use these data, and any other publicly available data for their modelling approach to succeed



# Data made available to modellers by Pakistan NTP

- **Prevalence survey full database (2010-2011)**
- All forms notifications (district level) 2009 – 2018, disaggregated quarterly
- New and relapse notifications (district level) 2009 – 2018, disaggregated quarterly
- Laboratory confirmation data (% notifications bac+) (district level) (2009-2018)
- Laboratory EQA data (district level) (2009-2018)
- DS-TB treatment outcomes data (district level; 2009-2018), disaggregated quarterly
- HIV testing and positivity rates among TB cases (district level; 2009-2018), disaggregated quarterly
- MDR/RR-TB notifications (district level; 2009-2018), disaggregated quarterly
- Master list of TB facilities: diagnostic and treatment disaggregated
- Number of presumptive TB cases (district level) any years available
- Number of presumptive TB cases tested (district level) any years available
- Number of slides tested (workload) (district level) any years available
- Contribution of private sector to notifications (district level) any years available
- HIV notifications  (2009-2018), at district/cluster level, disaggregated quarterly and annual estimates

# Timeline

2019

**Apr** — Launch hackathon

**May** — Applications and data sharing agreements

**Sep** — Model development by participating teams

**Oct** — Model evaluation

**Nov** — Declare winner at Union conference

**Dec** — Dissemination in Islamabad

2020 — Publication(s)



KIT Royal Tropical Institute

# Participating teams

Team led by
**Pete Dodd**

**University of Sheffield**

Debebe Shaweno
Peter MacPherson

Team led by
**Stewart Chang**

**IDM**

Bradley Wagner, Karyn Sutton, Anna Bershteyn, William Trouleau, Amjad Khan, Jens Levy

Team led by
**Thys Potgieter**

**Epcon**

Zhi Zhen Qin, Yumna Moosa, Farihah Malik, Vincent Meurrens

Team led by
**Fulvia Mecatti**

**University of Milano-Bicocca**

Maeregu Arisido
Gaia Bertarelli

Team led by
**Jennifer Ross**

**IHME**

Nathaniel Henry, Emma Spurlock, Kate LeGrand, Bobby Reiner, Mingyou Yang, Brigette Blacker, Audrey Batzel

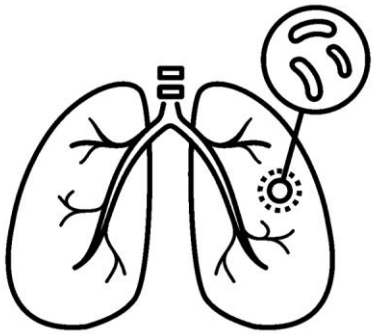
**Mystery team**

# Evaluation panel

## Co-chair KIT and NTP

- Ente Rood, KIT epidemiologist
- Abdulla, Manager Data National AIDS, TB & Malaria Control Programme

## Pakistan TB experts



- Javeriah Shamsi
- Rana Muhammad Safdar (National Coordinator (AIDS, TB and Malaria)

## Disease modelling experts

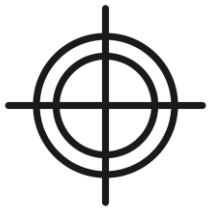
- Philippe Glaziou (WHO, Geneva)
- Federica Giardina (Erasmus MC, Rotterdam)

# Evaluation

- No gold standard for district-level TB prevalence…
- But quality is a multi-dimensional concept
- Qualitative properties of the models and expert opinion.


Source: [Quality Framework for OECD Statistical Activities]

# Evaluation grid

| Data quality dimension | Criteria | Points |
|---|---|---|
| **Validity**  | 1) Is the model scientifically sound (statistically or otherwise)? | 20 |
| | 2) Is the approach replicable? [code/pseudo-code] | 10 |
| | 3) Do model input data have known associations with TB prevalence? | 20 |
| | 4) Does the model have a sound methodology to derive credible/confidence intervals (CI) for the estimates (or other measures of sampling variability)? | 10 |
| **Accuracy**  | 5) What is the model's estimated predictive power based on 2010 data? [mean squared error (MSE) and relative square error (RSE) score from leave-one-out-cross-validation comparing actual and predicted 2010 cluster-level prevalence] | 20 |
| | 6) Do the estimates produce a distinction between high and low prevalence districts in 2018 in line with the local understanding of the epidemiology? | 10 |
| **Precision**  | 7) Is the model overfitting the prevalence survey data? | 10 |


SNEAK PEEK

# Sneak peek 1

Team led by
**Pete Dodd**

**University of Sheffield**

Debebe Shaweno
Peter MacPherson

**Data**

- **DHS 2017-2018 data**: mean household size, Wealth score (SES), indoor smoke, smoking, body mass index (BMI), weight-for-age Z-score (WAZ), prevalence of vaccination, prevalence of BCG vaccination, prevalence of chronic cough, self-reported tuberculosis (TB) prevalence, distance to nearest healthcare facility, and awareness of TB.
- **TB data:** Age-and sex-category specific bacteriologically confirmed TB notifications were available for years from 2009- 2012. The age-and sex-stratified per-capita notification rates were calculated by dividing the age-sex stratified bacteriologically confirmed case counts by the corresponding age-sex stratified population denominators

**Model:**

- Bayesian hierarchical regression models for prevalence (binomial logistic) fitted using Markov chain Monte Carlo approaches using Stan.
- Final retained model: Binomial distribution by cluster for age&sex included conditional autoregressive (CAR) priors to capture spatial effects
- To maintain consistency with published estimates, raw prevalence estimates central estimates and uncertainty were scaled to match..



KIT Royal Tropical Institute

# Sneak peek 2

Team led by
**Stewart Chang**

**IDM**

Bradley Wagner, Karyn Sutton, Anna Bershteyn, William Trouleau, Amjad Khan, Jens Levy

**Data**

- **TB data:** TB case notifications for 2009-2018 at district level,
- **Additional population data:** Shapefiles at tehsil (admin 3) and union council (admin 4) levels, Population estimates at a 100m level for 2010, Settlement density, Settlement type. Multidimensional poverty, Household cooking fuel type, wealth quintile, and urban/rural designation, Individual (female) underweight status and health visit frequency
- **DHS covariate extraction.** Household-level indicators including solid cooking fuel usage, wealth quintile status, and urban/rural classification were measured with GPS coordinate availability in DHS 2006 and 2017. In these cases, we downloaded the DHS microdata and derived a summary value for each DHS survey cluster, and generated a smoothed (kriged) surface for the entire country

**Model:**

- Binomial-logistic model with random effects
- Bayesian approach to model fitting

# Sneak peek 3

Team led by
**Jennifer Ross**

**IHME**

Nathaniel Henry, Kate LeGrand, Bobby Reiner, Mingyou Yang, Emma Spurlock, Brigette Blacker, Audrey Batzel

**Data**

- Population density
- Access to cities (Time required to travel to nearest settlement via surface travel.)
- Density of TB facilities
- Poverty
- Urban extents
- Locations of protests
- Locations of violent acts
- Aridity

**Model:**

- Bayesian spatially explicit binomial-logistic model (spatially-correlated residual variation) fitted using Template Model Builder
- Calibrated national-level prevalence estimates to Global Burden of Disease Study

# Sneak peek 4

**Mystery team**

**Data**

- Sex ratio, age structure, urban/rural from 2017 census

**Model**

- Bayesian Multi-level hierarchical model
- Binomial logistic
- Not spatially explicit
- Tehsil level predictions (Punjab only)

# Sneak peek 5

Team led by
**Fulvia Mecatti**

**University of Milano-Bicocca**

Maeregu Arisido
Gaia Bertarelli

**Data**

- Routine TB notification data (bac+, bac-, EP, All forms)
- Lab data (slides used, total errors committed)
- Census population data (households, male/female pop, urban/rural)
- HIV: total cases notified

**Model:**

- SAE-LM: integrating model-based Small Area Estimation (SAE) and Latent Markov (LM) modeling using Hierarchical Bayesian approach:
  1) True values of district prevalence considered a latent response variable, partially and indirectly measured (via sample data and longitudinal covariates) at successive time points, i.e. an underlying latent process;
  2) Distribution of latent process modelled both in space and time by means of a Markov chain

# Sneak peek 6

Team led by
**Thys Potgieter**

**Epcon**

Zhi Zhen Qin, Yumna Moosa, Farihah Malik,
Vincent Meurrens

**Data**

- Routine disease surveillance data
- Socio-economic data
- Spatial-Temporal Environmental Data (Quarterly consecutive precipitation (proxy for consecutive indoor days), other precipitation metrics, topographic height and population density recorded at each 25 km pixel covering Pakistan)

**Method:**

- Self-organizing maps (SOM) to extract patterns or features from multiple sources of spatial-environmental data that have been shown to influence incidence of tuberculosis.
- Machine learning (ML) engine to perform Bayesian learning
- Bayesian reasoning that allows inference and predictive what-if queries on newly observed variables based on prior learning.


# Contact

**KIT – Royal Tropical Institute**

Mauritskade 64

1092 AD Amsterdam

Sandra Alba

s.alba@kit.nl

KIT Royal Tropical Institute