UK-HDAN
Data Science for Health

The Alan Turing Institute

Turing/UK-HDAN Workshop on Health Data Analytics

Friday 3rd November 2017

Workshop Session Output

@Man_Inf @ESPRC #UKHDAN

EPSRC
Engineering and Physical Sciences
Research Council

Title: Identifying Sub-groups	**Group:** Brown Group (Session One)

Issues	Existing solutions/gaps
Headline: - Re-identifying disease using unsupervised methods. **Headline:** "Patients Like Me"	- Standard unsupervised clustering: distance based, model based etc. Lack of gold-standard validation.
Headline: - Unsupervised clustering, inter/intra clustering **Headline:** Uncertainty in clustering labelling	- Latent growth modelling for dynamic clustering Probabilistic inference - Hierarchical mixtures Fuzzy clustering Flexible clustering: hard vs. soft clustering
Headline: Response-based clustering - to action for the most positive outcome **Headline:** "Multi-objective" clustering	- Profile Regression -Integrative Clustering Methods
Headline: Stratifying disease **Headline:** Missing data within clustering	- Local modelling methods - Hierarchical: Global to Local - Multilevel Modelling
Headline: Drugs - Identifying tissue cascades to develop drug targets. - Identifying groups with worse/better side effects	-Individualised treatment effect (causal inference)
Headline: Identifying sub-populations in the context of clinical trials **Headline:** Interpretation of clusters & validation (Gap!)	

Contributors	Potential Contributors
Lydia Drumwright, Tingting Zhu, Andrey Kormitzlin, Shang-ming Zhou, Catalina Vallejos, Allan Tucker, Arianna Dagliati, Fotios Drenos, Hamza Javed, Joris Bucker, Jans Dattscher, Mihaela Van Der Schaar	

Title: Identifying Sub-groups

Group colour/number: Brown Group Session One

1: Describing, understanding & managing boundaries between clusters both within and across time.

2: Validation of methods for data driven approaches in sub-typing in the absence of a gold standard. Specifically those approaches that would be accepted by the medical community.

3: Methods/partnerships for interpreting subgroup profiles or identify globally accepted methods. Design across disciplines.

4: Partnership, cross training & common language development between HCW's & analysts. Training/Pilot Scheme?

5: Methods for managing the bias in the observational data.

6: Methods for multi-objective clustering.

Specific Use Case Examples:

A: Drug Development (See Headline #9 on main sheet).
B: Application to diseases with different time spans & progression over time (e.g Diabetes, IBD).

Contributors	**Potential Contributors**

Title: Linking & Integrating Heterogeneous Data	**Group:** Green Group (Session One)

Issues	Existing solutions/gaps
Headline: Linking Across scale, time and space, format/modality.	- RB2; Data shield distributed frameworks, implementation, temporal data is challenging.
Headline: Analyse the linked data (prior to or post linkage).	- Distributed Learning, hierarchical models.
Headline: Statistical linkage and statistical disclosure and associated uncertainty.	- Data perturbation, differential privacy.
Headline: Handle Conflicting Data	- New Logics
Headline: Real-time inference on continuous data	
Headline: Life-cycle of research data particularly categorical.	- FAIR data principles

Contributors	Potential Contributors
Ann Gledson, Goran Nenadic, Arianna Daguati, Emily Jefferson, Hamed Haddadi, Marcos Barreto, Jens Rittecher, Jan Wildenhain, Nophar Geifmen	

UK-HDAN
Data Science for Health

The Alan Turing Institute

Title: Untitled	**Group:** Orange Group (Session One)

Issues	Existing solutions/gaps
Headline: People	- Small Scale Efforts to engage patients but Gov/NHS pushing other way - Country does not work together - The infrastructure exists, but the formulation is not yet right
Headline: Policy/Law	Policy + Law not linked to norms Gap: interpretation by data controllers leadership Needs a long-term plan - 30 yr - but how to do this with a 5 yr Gov cycle and link long term research progress to Gov policy
Headline: Data Use Flexibility - Care is not the same as research but need links	- Making the NHS electronic and sharing between institutions - What is allowed and what is believed to be allowed? - Put the algorithm in the clinical space
Headline: Catastrophic Confounding - experimental design	Gap: aftercare linkage Policy for data linkage exists for point of care
Headline: Technical Solutions Synthetic Data Sets	- Multiple platforms exist but are not linked - Banks can do it! Medical records need translation to research
Headline:	

Contributors	Potential Contributors
John Parry, Nigel Birch, Rachel Furner, Lydia Drumwright	

Title: Untitled

Group colour/number: Orange Group Session One

- De-identification: How to anonymise (remove data) whilst still retaining usefulness.

- What is the status of linkages between NHS datasets and what are the restrictions?

- The law makes assumptions about what people want which aren't necessarily correct. Disconnect between patient/ delivery of care and legal/policy.

- Conflict between use of data, data control officers and info commission

- Format of date

- Policy Issues

- People Issues (Data Owners)

- Data Use, research and care, what are the links?

Contributors	**Potential Contributors**

UK-HDAN
Data Science for Health

The Alan Turing Institute

Title: a ƻŘŜťŕ𝕪Ǝ ¢ŜYǇƻ𝕎l·ť 5l·ǘl·

Group: tŕyʅ Group (Session ¢øƻ)

Issues	Existing solutions/gaps
Headline: - Irregular Sampling - Purposive Sampling	- Sliding Windows - Data Imputation but MNAR and UNK links - PROMS & Experience Measures
Headline: - Range of Time Scales	- Gaussian Process Models - Recurrent Neural Networks - Hidden Marker Models
Headline: - Anonymisation by removing absolute time stamps. "Fuzzing"	- Privacy rather than anonymisation. Data behind firewall analysis
Headline: - Stratifying disease Headline: - Missing data within clustering	- Local modelling methods - Hierarchical: Global to Local - Multilevel Modelling
Headline: - Quality of Time Capture	-Individualised treatment effect (causal inference)
Headline: - Identifying sub-populations in the context of clinical trials Headline: - Interpretation of clusters & validation (Gap!)	

Contributors	Potential Contributors
Lydia Drumwright, Tingting Zhu, Andrey Kormitzlin, Shang-ming Zhou, Catalina Vallejos, Allan Tucker, Arianna Dagliati, Fotios Drenos, Hamza Javed, Joris Bucker, Jans Dattscher	

Title: Modelling Temporal Data	**Group colour/number:** Pink Group Session Two

How do we model/analyse longitudinal data.

Irregular sampling & purposeful sampling (consultation for a reason).

Range of time scales (Daily/seasonal/shorter)

Anonymisation by removing absolute time stamps (e.g for hour of the day, for month of the year).

Data Quality of date stamps - difference in linked data (e.g DoD)

System date does not equal event date and time stamps not right. Messy

Using the past to predict the future. Is this a reliable premise for (e.g training algorithms).

- Fast moving tech development
- Confounding context. Capture this richly.

Outlier patients (modelling without observing individuals).

- Similarity across patient pathways
- Modelling disease trajectories
- Trajectory clustering

Time Series:
Treat a time line as a sentence. Synatactic approach, borrow techniques from NLP community.

Using time to predict time. "Time to event" as an outcome.

Understanding human gaming of the systems.

Separating a path into "pathlets"

Understanding the drivers of timing of data recording.

It is easier to go from time-course data to action than build a model in between.

Contributors	**Potential Contributors**

Title: Effective Visualisation of Data	**Group:** Purple Group (Session One)

Issues	Existing solutions/gaps
Headline: Actionable Visualisations, communicating what people need to know and useful discoveries	- Education, training, software tools - Expensive, few UIS Experts in the UK
Headline: Availability of technical expertise and understanding to make visualisations useful (not necessarily pretty). Linking to semantics.	- Education, training - Funding, lack of standards, data quality
Headline: Availability and enthusiasm of workforce to interpret and value data. Multiple audiences - different understanding/ actions.	- Success stories curriculum. - Demonstrating value, medical safety and validation.
Headline: Interacting with high-dimensional data (geospatial, temporal, qualitative, quantitative, anatomical...)	- Cartographic Treemaps, research area. - Unsolved Technical Challenges.
Headline: Communicating Uncertainty & Trends.	- Existing Software Tools. - Generic Tools are Challenging.
Headline: Overlaying individual and population data for contextual interpretation. Real-time Visualisation	

Contributors	Potential Contributors
Mahmood Adil, Ann Blandford, Bob Laramee, Gary Leeming	

Title: 9ƚƚŜŎǘƛØŜ ±ƛăǳl·ŧƛăl·ǘƛƶʏ ƶƚ 5l·ǘl· **Group colour/number:** tdzNJLJŧŜ Group Session One

- Bridging Gaps between CS and Health. What's possible? What's Useful? Diagnostics, prognostic.

- Identify low-hanging fruit from data and from needs.

- Engagement vs. Comprehension.

Contributors	Potential Contributors

Title: Imperfect Data	**Group:** Red Group (Session One)

Issues	Existing solutions/gaps
Headline: Messy Data Missingness (MNAR), artifacts, units of measurement unknown	- Exploratory data analysis with domain experts - Rich models of observation process (including prior domain knowledge)
Headline: Missing Context e.g linking temporal events e.g environmental information for patient	Capture Meta-data
Headline: Inaccessible Data e.g free-text (not available) e.g constraints in collections	- With regards to free-text, issue is governance. - Pushing of NLP processing behind firewall
Headline: Lack of gold Standard/ground truth, difficulty in validating results.	Systems design of data collection.
Headline: Catastrophic Confounding, experimental Design	
Headline:	

Contributors	Potential Contributors
Magnus Rattray, Chris Williams, Sam Relton, Jian-Bo Yang, Hamza Javed, David Hogg, Kenan Direk, Liz Ford	

Title: Imperfect Data	**Group colour/number:** Red Group Session One

Expt. Design

5 Complete Confounding (Experimental Design) for inference of causal effects.

Str. of observations process (studies vs. routine observational date) Variable measured for a reason.

1 Missingness (not MAR), types of data (patient data vs. molecular).

1 Artifacts (incorporation in analysis).

4 Lack of gold standard (partially unlabelled).

2 Linking Temporal Events.

Accessibility of Data:
- Info in free-text (but this may not be available).
- Constraints of data collection & availability (was data collected? is it available to researchers?).

Incomplete Data.

Missing contextual information for observations (different state of person).

Data preparation process (80-90% of time), reproducibility.

Biases in recording outcome (and knowledge about context).

Probablistic Programming

Standard methods to map data - diagnosis
- Combining Data Sources
- Treat variables as noisy - use proxy variables/latent
- How to treat subjective variables (e.g; pain)
- Use of RL (reinforcement learning)

- Investigate variations of outcomes/ variables

- Symptom development over time semi-supervised learning

Latent variable for MNAR
Class for study adherence
Changes in recording patterns over time (and locations) e.g QOF.
Variation in GP's coding some interaction
SLAM obtained free-text for NLP Processing
How to create synthetic missing data, density models, GAN's

Contributors	**Potential Contributors**

UK-HDAN
Data Science for Health

The Alan Turing Institute

Title: Data & Knowledge Life Cycle	**Group:** Silver Group (Session One)

Issues	Existing solutions/gaps
Headline: Applied Intelligence - "Active" Data Analytics & DSS - Spectrum of analytics	- Integrate with Social Care - Data analytics life cycle - Not only descriptive, but also predictive and prescriptive
Headline: Meta-Data - Data Models - discrete date - Best Practice	
Headline: Knowledge Engineering - Context - Executable Guidelines/Pathway Models - Data/Knowledge Provenance	
Headline: How Knowledge Changes?	- Maintenance
Headline: Bringing data science and knowledge engineering together.	-Bridging data & Knowledge
Headline:	

Contributors	Potential Contributors
John Fox, Goran Nenadic, Emily Jefferson, Gary Leeming, Mahmood Adil	Jian-Bo Yang

Title: Data & Knowledge Life Cycle

Group colour/number: Silver Group Session One

Visualisation Issues:

1: Actionable Visualisations

- Questions people know they want to know & useful discovery

2: Availability of expertise to make visualisations useful but not necessarily pretty. 3a:

Extracting knowledge from data.

3: Ability/enthusiasm at workforce to understand/interpret data and value it.

4: Ontologies, high dimensional data - geospatial temporal, qualitative, quantitative. Communicating uncertainty trends.

"Active" Data Analysis

- Suitable representation of data models.
- Scale-up knowledge
- "Technology is not an issue"
- Managing Data Provenance
- Research is part of NHS landscape
- Health and Social Care Intelligence

"Applied" Data Analysis

- Meta-date is important? Interoperable?
- Two streams: Care & Research
- "Executable guidlines", modelling practice/pathways

Contributors	**Potential Contributors**

Title:	Predictive Modelling & Actionability	**Group:**	Yellow Group (Session One)

Issues	**Existing solutions/gaps**
Headline: Missing Data - Informative Missingness -Informative Censoring - Missing Context / Clinical Knowledge	- Knowledge Based Systems - MLI Stats Methods (Patterns and prior knowledge) - Causal Interference
Headline: - Prediction with observational Data - Optimal Treatment Prediction - Treatment Effect on Prediction	- Causal Inference (Propensity Scoring) - Mendelian Randomise - Machine Learning methods for individualised treatment effects
Headline: Imbalanced Data - Specially in the context of longitudinal data	- Prior Knowledge - Boosting Methods - Re-Weighting Methods - Synthetic Data - Transfer Data
Headline: Pre symptomatic prediction - Early Prediction	-Transfer Learning - Knowledge Engineering - Disease/Risk Trajectory - Wearables - State-space models
Headline: Dealing with Drifts or changes in practice	-Scoring Methods -Change Point Analysis -Unsupervised Learning -State Space Models
Headline: Action upon Predictive Models & Feedback	- Clinical Decision Support Systems - Causal Inference - Online Learning & Re-Calibration

Contributors	**Potential Contributors**
Catalina Vallejas, Mihaela Van Der Schaar, Tingting Zhu, Fotios Drenos, Lisa Koeppel, Joris Bucker, Robert Goudie, Shang-Ming Zhou, Allan Tucker, Andrey Kormilitzin, Maxine Mackintosh	

Title: Predictive Modelling & Acountability

Group colour/number: Yellow Group Session One

0: Bridging the gap between medical knowledge and modelling. 1:	John
Dealing with gradual shifts - Changing Features in the context of changing points - State-Space representations (latent models)	Rob
2: Interpretability vs predictive ability - Interaction between MLI stats approaches - Increases interpretability in ML settings - Clinical relevance vs prediction	Shang-Ming
3: Features selection in high-dimensional spaces	Michaela
4: Dealing with outliers & rare events on/off line	Cata
5: Rare Diseases & Unknown Features	Fotios
6: Co morbidities - how to incorporate them in predictive models & poli pharmacy	Mihaela, Catalina Shang-Ming & Tingting
7: Multiple pathways of care that interact (treatments, interventions)	Tingting

Contributors

Potential Contributors

Title: Predictive Modelling	**Group:** Yellow Group (Session Two)

Issues	Existing solutions/gaps
Headline: Trust Issues. Performance vs explainability trade-off	- Actionability
Headline: - Predicting the effects of interventions. "What if?"	- Causal Inference Methods - Control Engineering - Complexity? - Smart Cities?
Headline: - Predict Outcome (decision) of consultation	
Headline: - Holistic biology & behaviour. Predict health state based on corporate history.	
Headline: - Online vs. batch learning	- Trust/certification
Headline:	

Contributors	Potential Contributors
Niels Peek, Jian-Bo Yang, Jan Wildenhain, Chris Williams	