Removing Clutter - Killing Dimension
a new feature set for streamed data

Terry Lyons

Mathematical Institute, Oxford
Alan Turing Institute

Hao Ni, Thomas Cass, Wiexin Yang, Imanol Perez, Paul More, ….

Oxford
Mathematics

Rough paths - streamed data

Mathematics of Information

Period of rapid change

- in our understanding and modelling of functions
- in the role of optimisation
- In how we use information

So dramatic that we might ask what is left for mathematics?

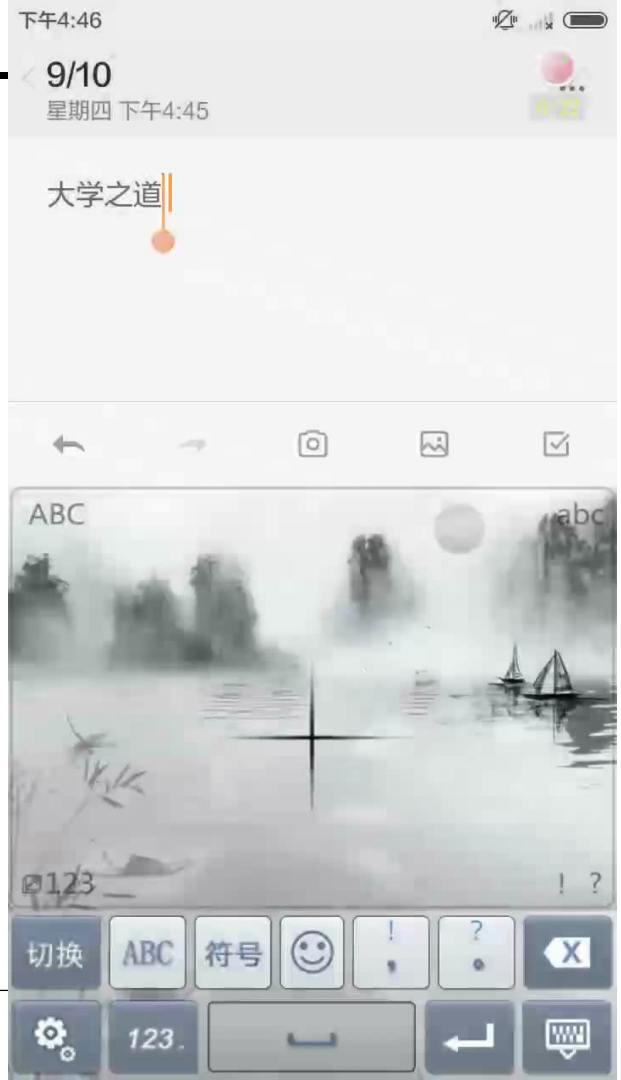
- Will describe three data science challenges
- Where a core piece of new mathematics adds value
    - a feature set that removes an infinite dimensional invariance

Understanding Multi-Modal Data for Social and Human Behaviour at the Isaac Newton Institute, Cambs. Tuesday 27th November

Rough paths - streamed data

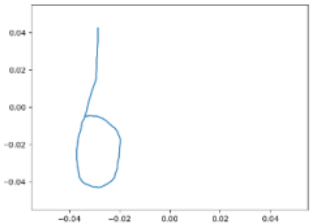
- a character drawn on the screen of an iphone
- an order book
- a piece of text
- progression through hospital record
- astronomical data

- a facial expression evolving - smiling -

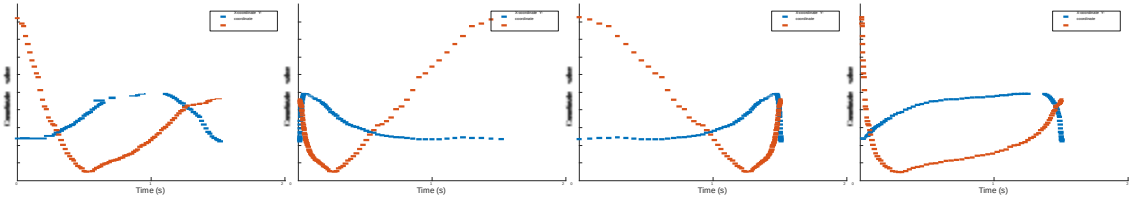


Rough paths: *unparameterised* streamed data

– The letter “b” is drawn from top to bottom by hand.



– Raw x and y coordinates of the letter plotted against time at different speeds. (speed increased at start, at end, at start and end. (Paul Moore)



The signature – faithful and universal features describing an unparameterised stream

The signature of a stream γ over $I = [s, t]$ defined by

$\sum_{k=0}^{\infty} S_k$ where $S_0 = 1$ and

$$S_k(\gamma, I) := \iint_{s<u_1<\cdots<u_k<t} d\gamma_{u_1} d\gamma_{u_2} \dots d\gamma_{u_k}$$

These “Fourier-like” coefficients exactly describe the *un-parameterised* stream. (Hambly Lyons Annals Math 2010)

The signature – faithful and universal features describing an unparameterised stream

Fundamental tool in Rough path theory; foundational mathematics redefining how one talks about paths; extends calculus of Newton and Ito to broad new ranges of complex interacting systems. An under-pinning contribution to Fields medal winning mathematical work of Hairer.

Crucially for data applications, it removes parametrisation and can reduce the dimensionality of streamed data quite dramatically.

There is a world wide (PyPy) availability of a Python package “esig”.

The signature of a path describes an unparameterised stream γ

Signature is a *top down* description for unparameterised paths that describes a path segment through its effects of stylised nonlinear systems

$$dS = S \otimes d\gamma$$

removing an infinite dimensional invariance allowing prediction and classification with *much* smaller learning sets.

gives fixed dimensional feature sets regardless of the sample points (missing data/common parameterisation not issues).

The signature of a stream – a faithful and universal feature set for a stream

Suppose γ is a stream or path $I = [s,t] \to E$ with values in $E := \mathbb{R}\langle e_1, \dots, \rangle$ a vector space, whose basis we refer to as the alphabet, and I is an interval of information. Then any word $\boldsymbol{e} = (\boldsymbol{e_{i_1}}, \dots, \boldsymbol{e_{i_n}})$ defines a real valued feature of the stream γ over I:

$$\boldsymbol{\phi_e}(\boldsymbol{\gamma}, \boldsymbol{I}) := \iint_{\boldsymbol{s<u_1<\cdots<u_k<t}} \boldsymbol{\langle e_{i_1} | d\gamma_{u_1} \rangle \langle e_{i_2} | d\gamma_{u_2} \rangle \dots \langle e_{i_k} | d\gamma_{u_k} \rangle}$$

$$= \boldsymbol{\langle e | S_I \rangle}.$$

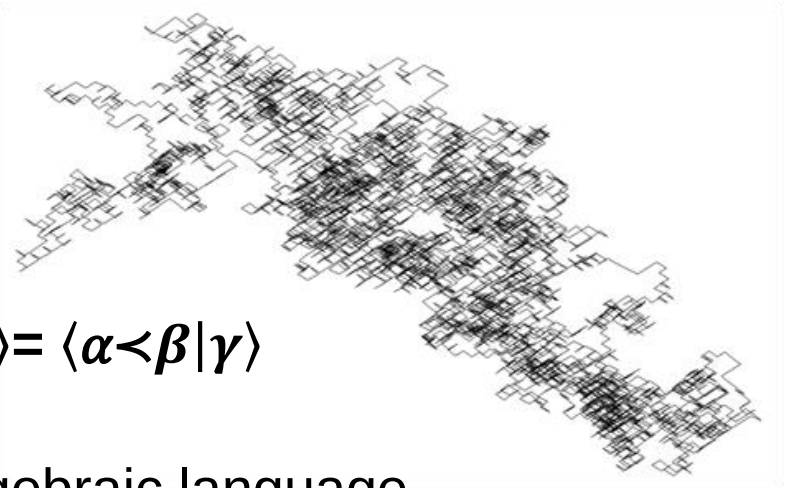
Words are independent and ?span? numerical functions on streams.

Analysis, Geometry, Combinatorial Hopf \ Dendriform \ Sensor Algebras

Signature leads to linear space of real valued functionals on streams

Pointwise multiplication and integration of these functionals

$\langle \alpha|\gamma\rangle\langle \beta|\gamma\rangle = \langle \alpha \sqcup\!\sqcup \beta|\gamma\rangle \qquad \int \langle \alpha|\gamma\rangle d\langle \beta|\gamma\rangle = \langle \alpha \prec \beta|\gamma\rangle$

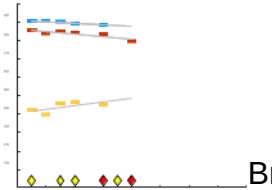


can usefully be described in purely algebraic language

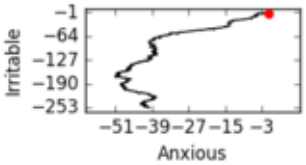
Leveraging the mathematics

Finger drawn characters

Moving matchstick man



Brain Weights



Evolving mood

Chinese handwriting

Action Recognition

Alzheimer’s Disease

Complex social data

Understanding *evolving* human action samples small and noisy – dimension high

Reducing video to pose is now at a very impressive!

The matchstick men and women are data streams in 30-75 dimensions

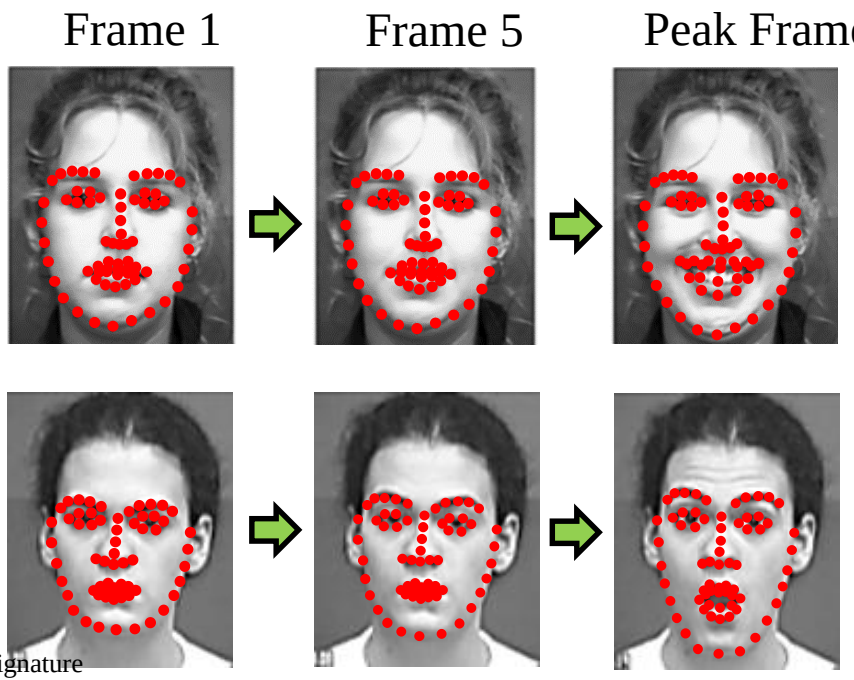
Can we recognise what is happening

http://mvig.sjtu.edu.cn/research/alphapose.html

Understanding *evolving* human action samples small and noisy – dimension high

Yang, W., Lyons, T., Ni, H., Schmid, C., Jin, L. and Chang, J., 2017.
Leveraging the Path Signature for Skeleton-based Human Action Recognition. *arXiv preprint arXiv:1707.03993*.

Facial Expression Recognition on CK+ database [1]



Train set:	248 clips (6 classes)
Test set:	79 clips
Feature:	6834 dimensional Path Signature Feature
Classifier:	6-layer fully-connected neural net
Accuracy:	96.21% (vs. 94.83% in [2])

[1] Cohn-Kanade AU-Coded Expression Database. www.pitt.edu/~emotion/ck-spread.htm
[2] D. Ghimire, et al. (2016) Facial expression recognition based on local region specific features and SVM.

Small noisy evolving data everywhere – component brain volumes (Paul Moore, John Gallacher)

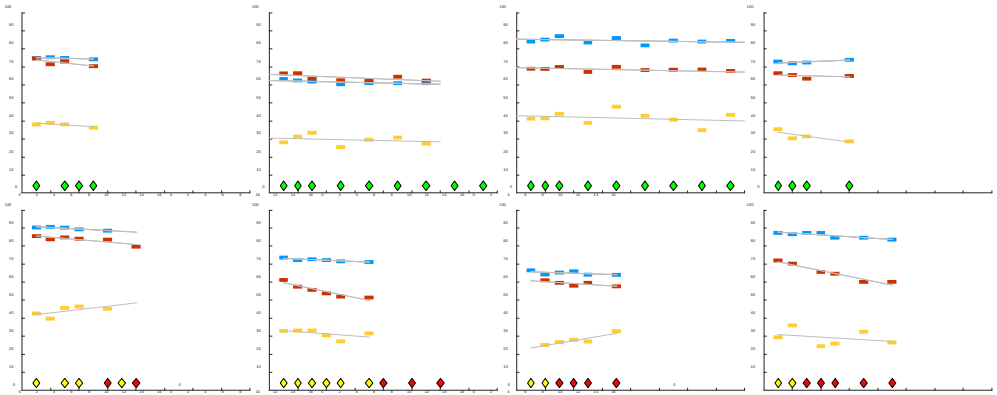


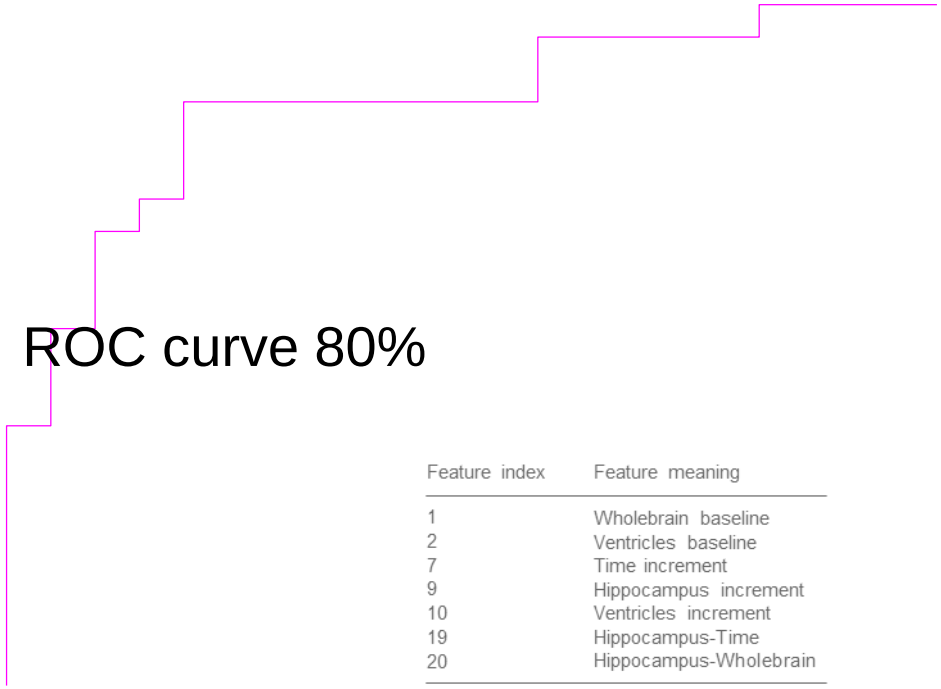
Figure : Sample plots by time (months) of scaled brain volumes for 8 patients, top row healthy and bottom row Alzheimer's disease. Data is taken from the Alzheimer's Disease Neuroimaging Initiative (ADNI) http://adni.loni.usc.edu/

- In each graph: Whole brain (blue), Hippocampus (red), Entorhinus (yellow).
- Diagnosis points are diamonds: Healthy (green), MCI (yellow), Alzheimer's disease (red).

Brain regions - use machine learning to preselect participants for early signs of dementia?

14/14

Experiments in foreseeing AD



ROC curve 80%

Feature index	Feature meaning
1	Wholebrain baseline
2	Ventricles baseline
7	Time increment
9	Hippocampus increment
10	Ventricles increment
19	Hippocampus-Time
20	Hippocampus-Wholebrain

The 1737 participants are split into those who get a diagnosis of Alzheimer's disease (AD) at some point, and those who remain healthy and with no memory problems. We follow Alzheimer's patients up the the last time point before diagnosis and create a matched set from healthy participants.

The experiment is to distinguish the two groups using changes in relative brain volumes. For each set a feature vector is derived from the initial values and a path signature formed from time and the three brain volume variables.
To select from these inputs we use Lasso. A test set of 20% is held out for error estimation and the remaining 80% is used to train a logistic regression model. We use binary logistic regression on the signature features.

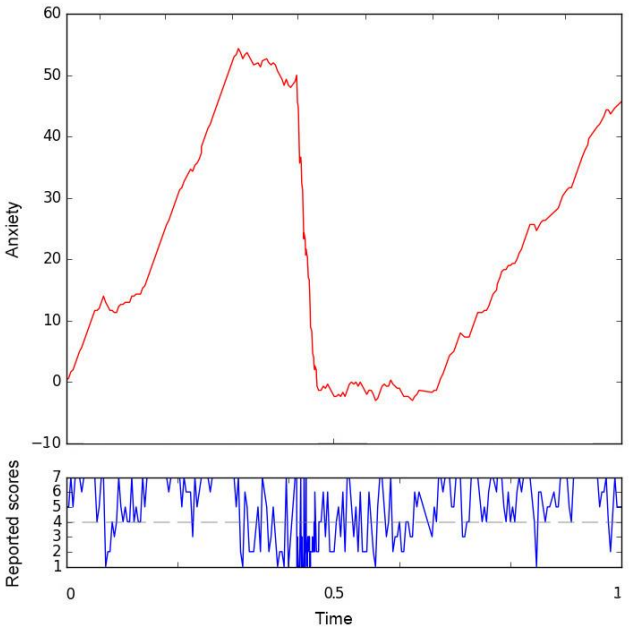
Triaging BP BP & N on the basis of mood zoom

In a clinical trial, mood zoom data was requested daily from three groups (total 130 people for a year) with distinct diagnoses: bipolar disorder, borderline personality disorder or healthy control. The data was noisy, missing.

This mood data was grouped into episodes of 20 sequential responses; a random forest classifier using second order signature features was trained on these episodes; it obtained very good separation of the three groups using one-cross validation.

Second order information was important. Signatures were critical to controlling the dimension given the sample size.

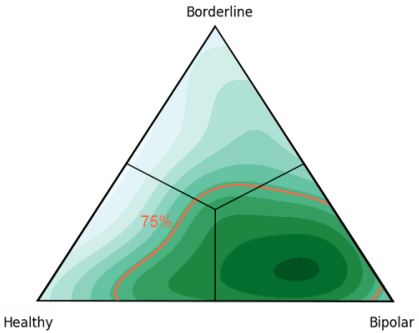
Imanol Perez Arribas, Kate Saunders, Guy Goodwin, and Terry Lyons. "A signature-based machine learning model for bipolar disorder and borderline personality disorder." Journal of Translational Psychiatry

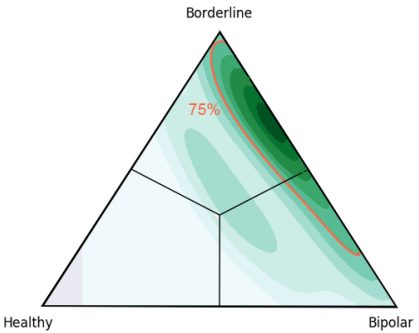


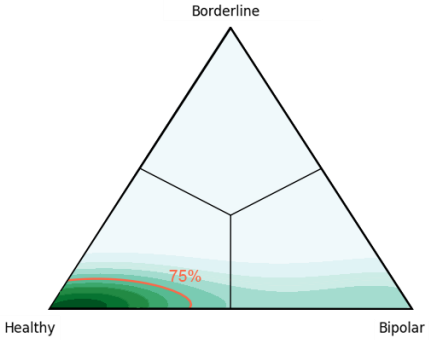
Evolution of anxiety scores for a participant diagnosed with bipolar disorder.

Triaging BP BP & N on the basis of mood zoom

Mood zoom data was requested daily from three groups (total 130 people for a year) with distinct diagnoses: bipolar disorder, borderline personality disorder or healthy control. The data was noisy, missing. Using higher order information (anger before depression …) captured by low dimensional signatures was able to classify on spectrum. The project is one of three being showcased and made reproducible at the ATI. (to appear: The Journal of Translational Psychiatry).







A current project jointly supported by Turing and Oxford Psychiatry:

Learn how to use speech, facial mood (from a phone app), with the aim of less intrusive and more reliable mood zooms that can be used to provide longer term information and feedback in clinical contexts (think Ha1cb).

Challenge the maths to do even better – blend signatures with natural language processing, harmonic analysis to further reduce dimension. Plan: create, validate, a reliable clinical tool for mental health triage

Low dimensional features that capture order information seem a key.

SCUT gPen, Sogou IME

- https://itunes.apple.com/cn/app/scut-gpen-shou-xie-shu-ru-fa/id957809824?l=en&mt=8
- Sogou IME, https://itunes.apple.com/cn/app/id917670924
- https://play.google.com/store/apps/details?id=net.hcilab.scutgPen.IME&hl=en 1-5 million downloads
- Sogou IME: https://play.google.com/store/apps/details?id=com.sohu.inputmethod.sogou 10-50 million downloads