Rough Paths

DataSig

A rough path between mathematics and data science

Terry Lyons

26 February 2021

with many others… but particularly Tom, Varun, Cris, Maud, Peter, Roly, Sam, Patricia

University of Oxford · The Alan Turing Institute · Imperial College London · UCL · UKRI Engineering and Physical Sciences Research Council

A Turing vision

DataSig

A rough path between mathematics and data science

> We channel our research around a number of ambitious challenges which represent areas in which AI and data science can have a game-changing impact for science, society, and the economy. These challenges will not be led by the Turing alone, but depend on significant collaboration and partnerships.

DataSig

A rough path between mathematics and data science

Modelling behavior of evolving systems

A collaboration

DataSig
A rough path between mathematics and data science

GCHQ

Thomas C

Jack D

National Cyber Security Centre
a part of GCHQ

David P

dstl

Varun C

DataSig

Terry Lyons

Cris Salvi

Patricia Andrew

Imanol Pérez Arribas

The Alan Turing Institute

Peter Foster

Maud Lemercier

Roly Perera

DataSig | an EPSRC/UKRI 5-year program grant

Mathematics

- rough path theory and signatures
- describing the interactions between complex systems from the top down
- extending the calculus of differential equations to complex contexts

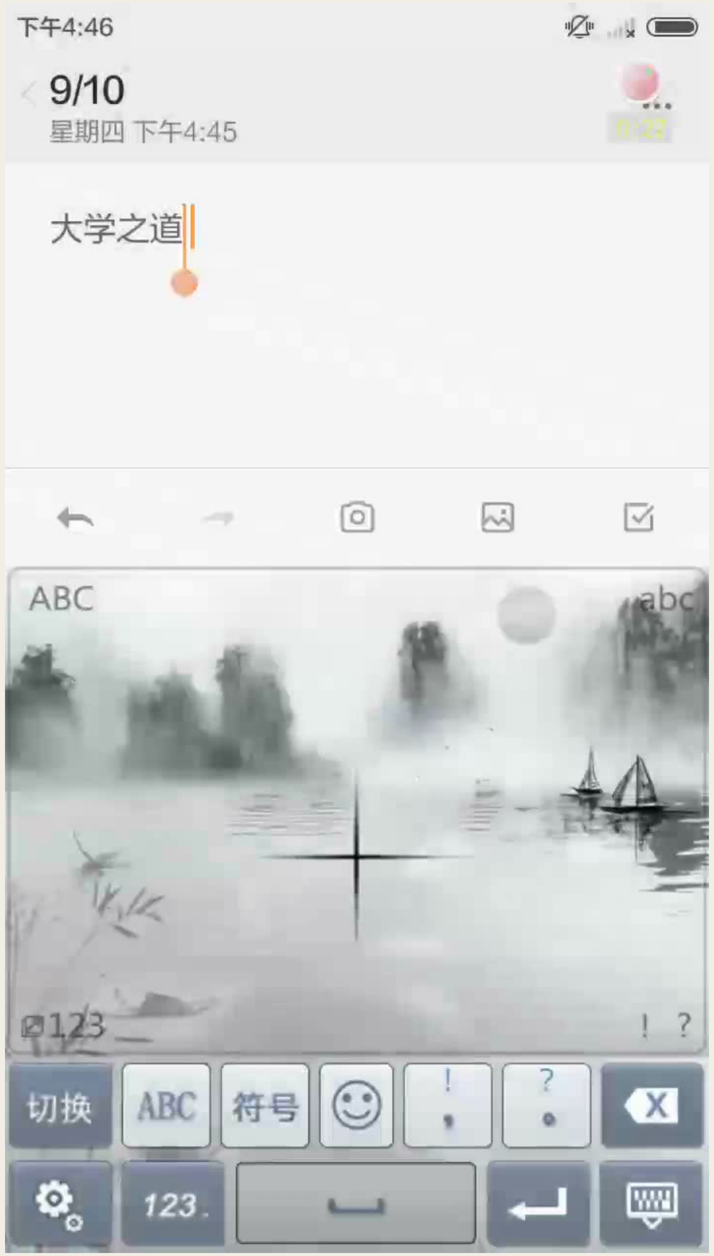
Data science

- the notion of an unparameterised path captured by the order of events
- clean and minimal universal feature sets
- the notion of a neural controlled differential equation
- The notion of a pde-kernel
- a principled mathematical framework that allows further innovation

Embedded contexts

- streamed data is everywhere; Chinese handwriting, hospital wards, …

DataSig
A rough path between mathematics and data science



Streamed data

- a character drawn on the screen of an iPhone
- an order book
- a piece of text
- progression through hospital record
- astronomical data
- video of a person moving
- an evolving stream of emotions
- ICU data to detect sepsis

Ensembles of streamed data

- the processes generated by malware
- the behaviour of crowds
- the evolution of cancer cell lines

Key questions

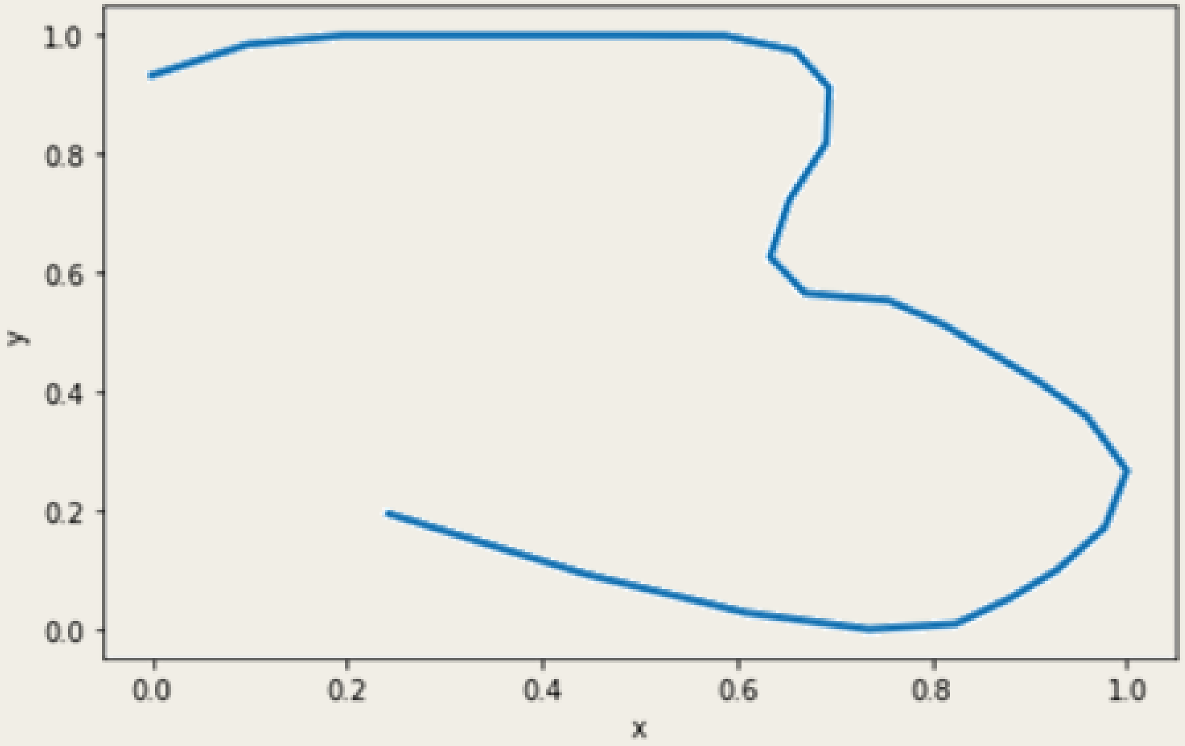
- understand what you have observed
- predict the distribution of what is happening next
- identify anomalies

DataSig
A rough path between mathematics and data science

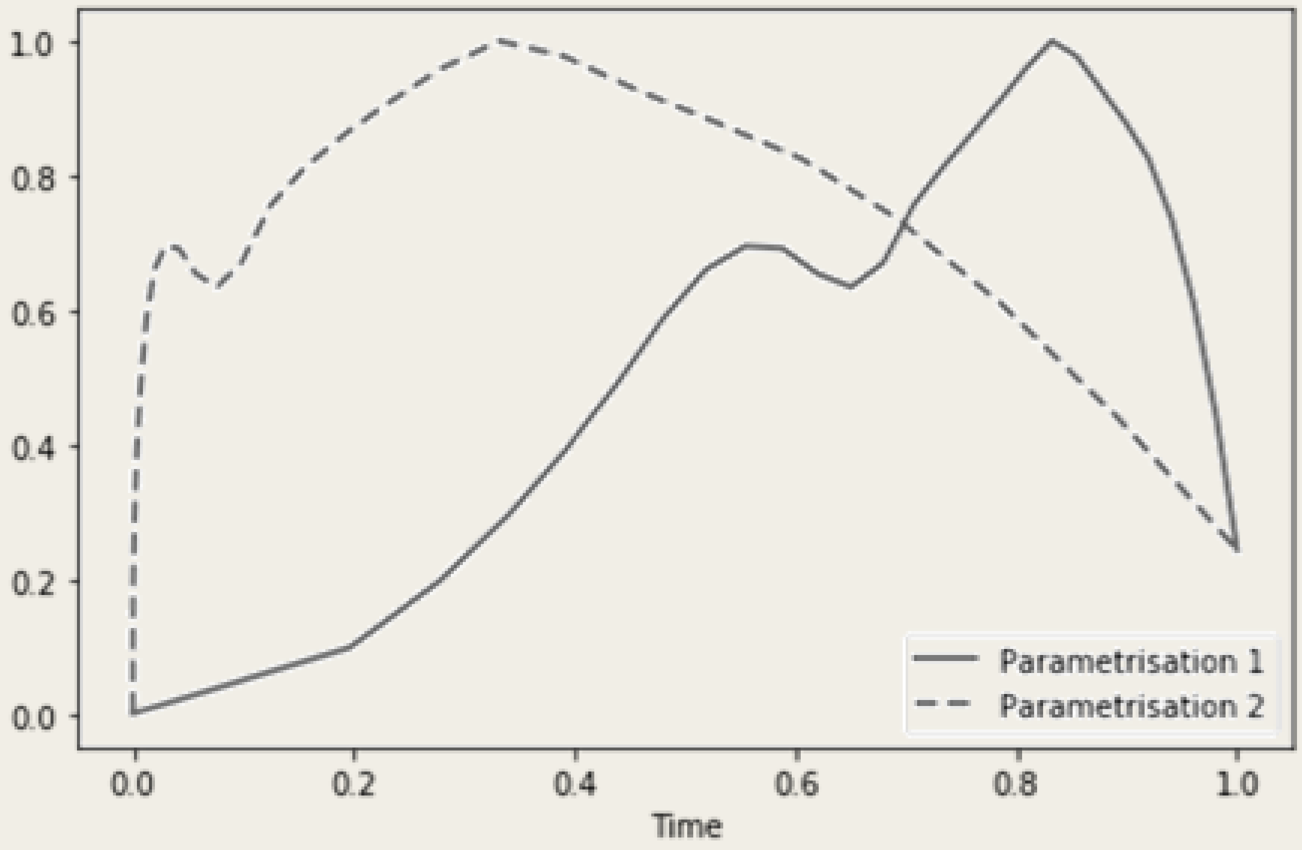
Some maths of evolving systems

Data science does not like symmetry

- Re-parameterisation is a huge symmetry group
- Multimodal streams modulo re-parameterisation form a group
- Representing this group in the tensor algebra provides a faithful feature set and removes the symmetry
- New tools signature and log signature, new maths describing the functions on streams



Different sampling procedures

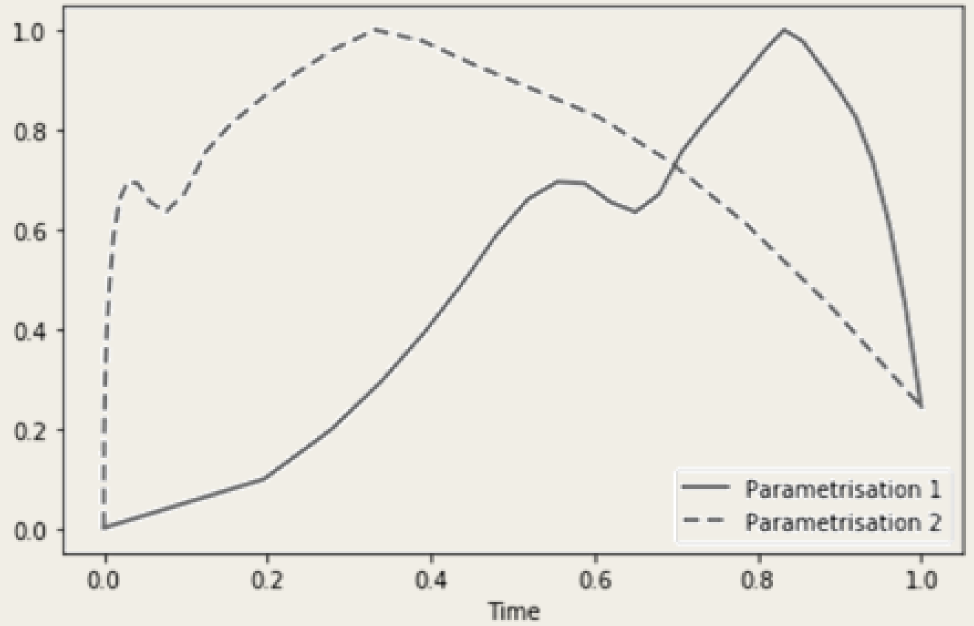


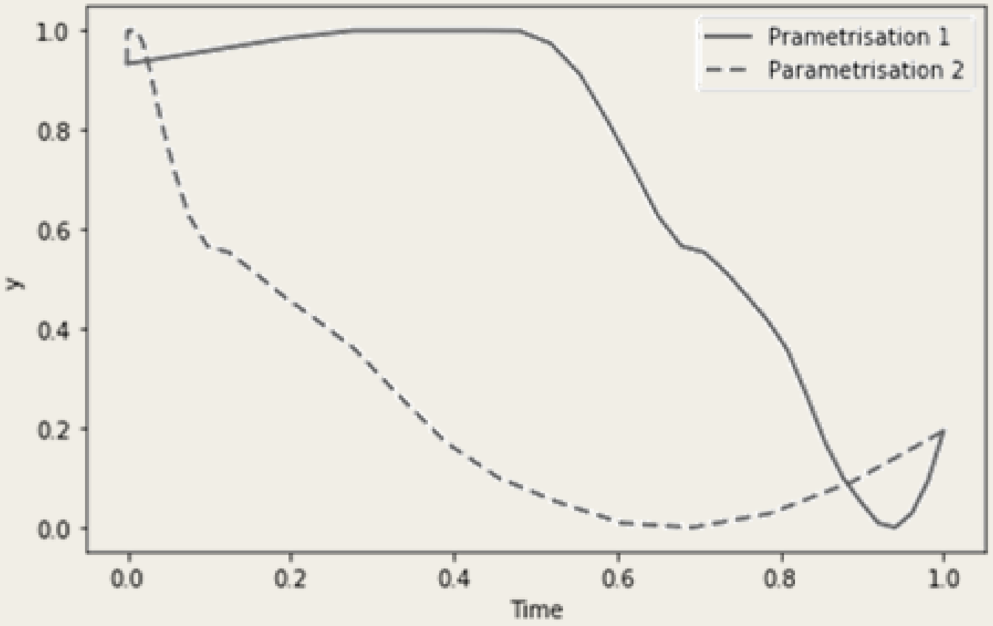
- The letter “3” is drawn from top to bottom
- The x coordinate of the evolving symbol sampled differently (at uneven speeds)

Different sampling procedures

The number “3” x, y coordinates – same picture drawn at two different speeds

- no consistent wavelets
- reparameterisations do not form a linear space!



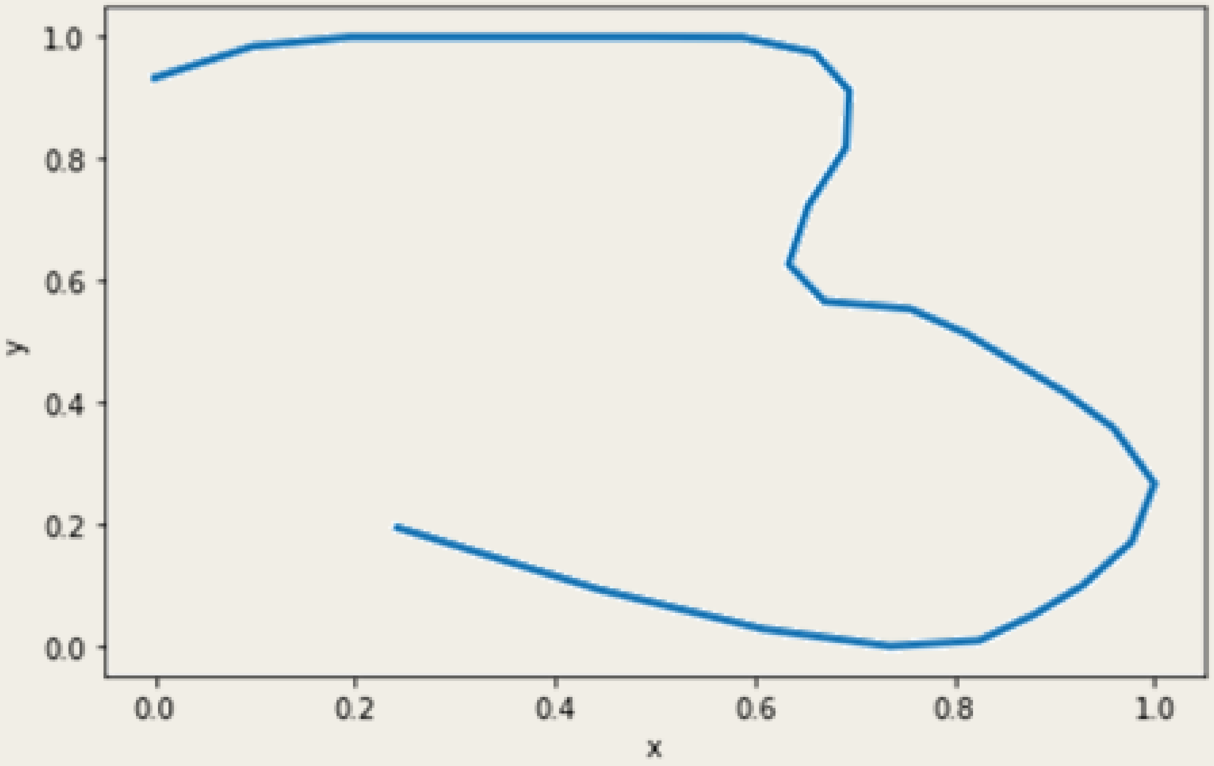


Different sampling procedures

DataSig

A rough path between mathematics and data science

- The letter “3” is drawn from top to bottom
- How does one describe the three or any path modulo the symmetry of parametrisation?



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

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

It filters out the infinite dimensional noise of resampling allowing prediction and classification with *much* smaller learning sets.

It gives fixed dimensional feature sets regardless of the sample points.*

* missing data/varying parameterisation not issues although inadequacy may be

Analysis, geometry, combinatorial Hopf/dendriform/sensor algebras

DataSig

A rough path between mathematics and data science

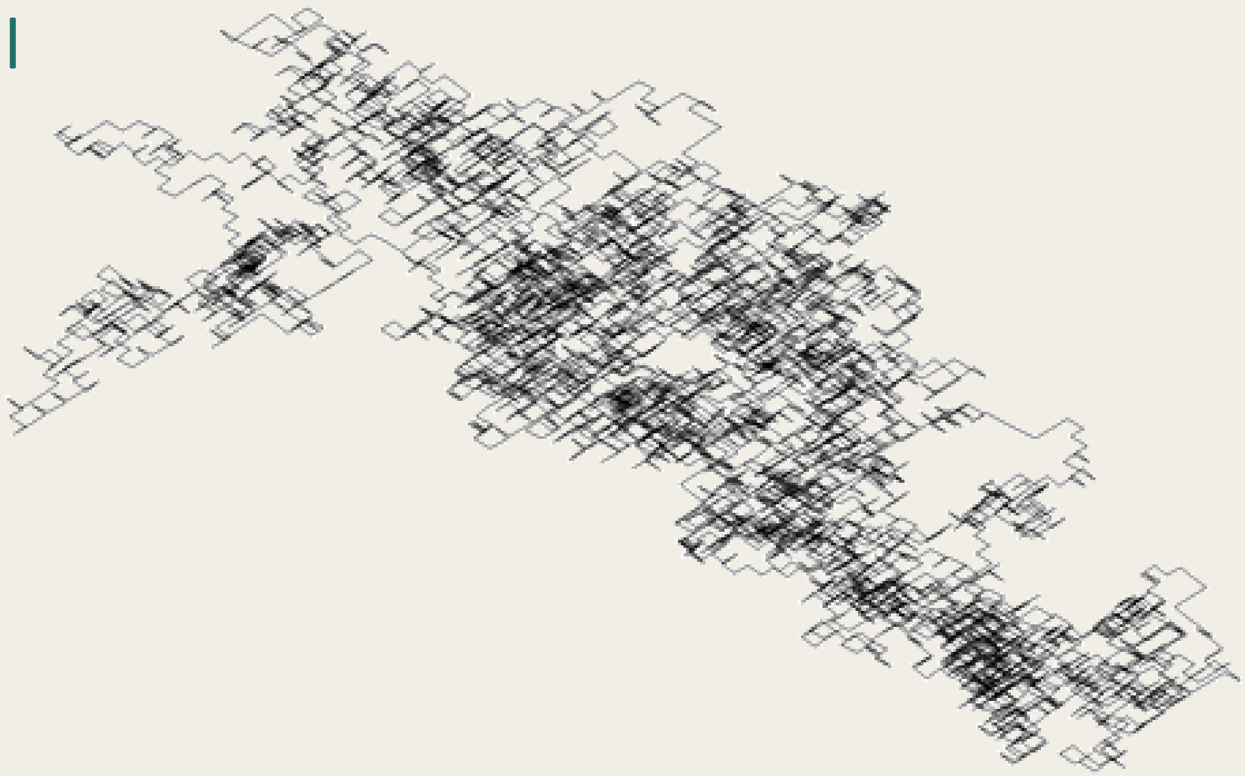
Signature leads to linear space of real valued functionals $\langle \boldsymbol{e} | \boldsymbol{S_I} \rangle$ on streams

Pointwise multiplication and integration of these functionals

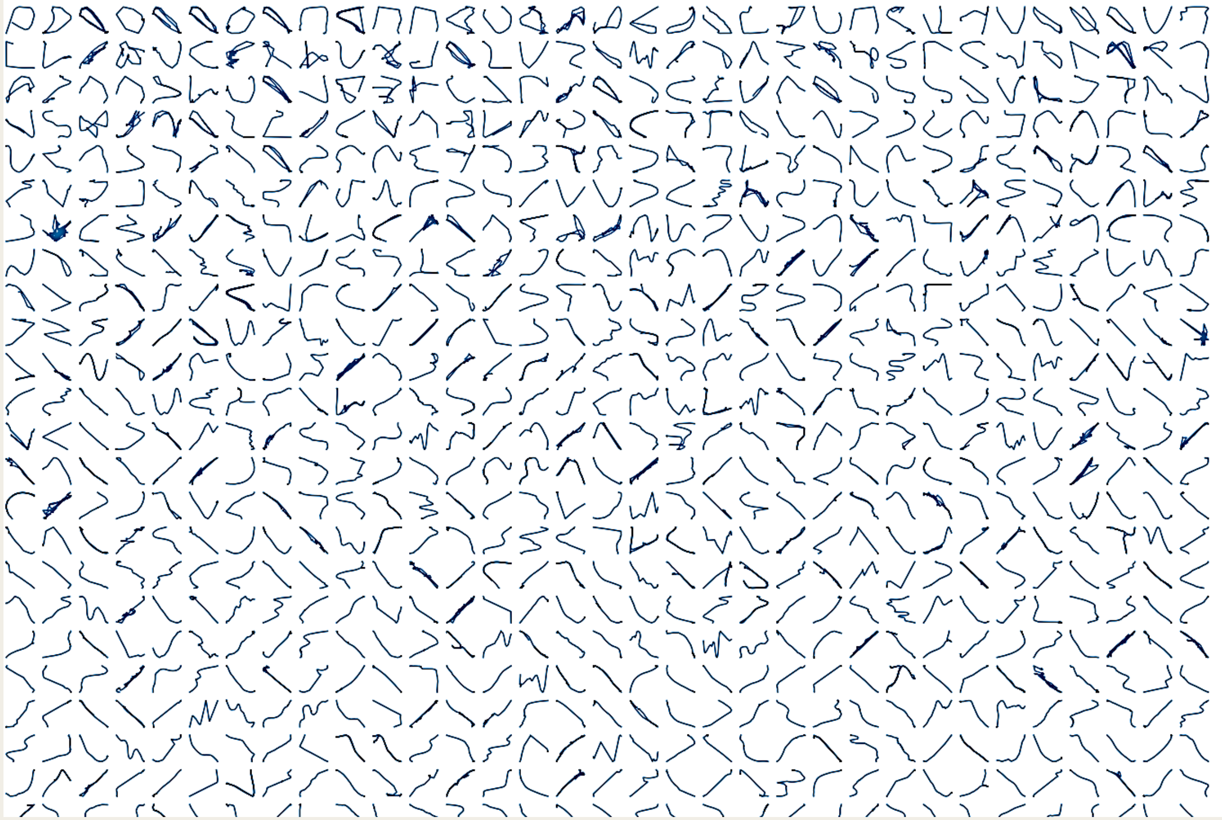
$$\langle \boldsymbol{\alpha} | \boldsymbol{\gamma} \rangle \langle \boldsymbol{\beta} | \boldsymbol{\gamma} \rangle = \langle \boldsymbol{\alpha} \Cup \boldsymbol{\beta} | \boldsymbol{\gamma} \rangle$$

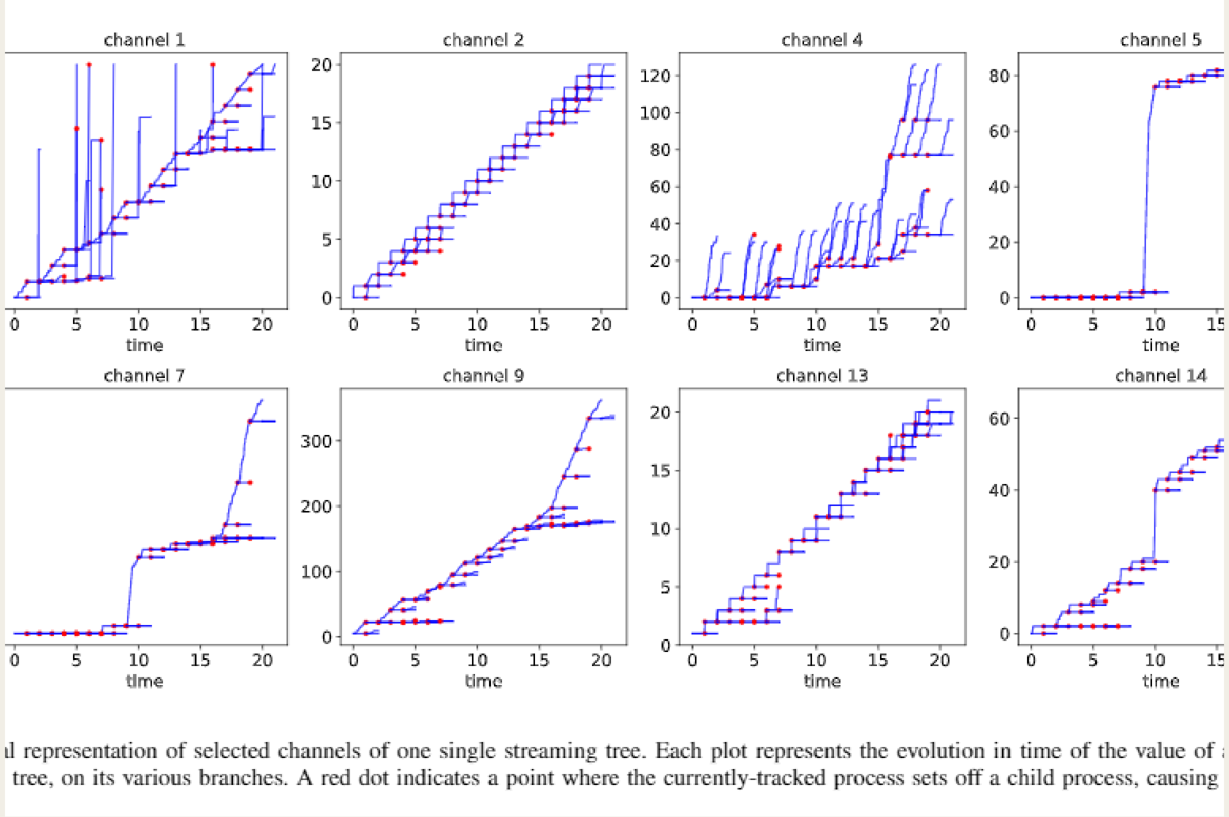
$$\int \langle \boldsymbol{\alpha} | \boldsymbol{\gamma} \rangle \, \boldsymbol{d} \langle \boldsymbol{\beta} | \boldsymbol{\gamma} \rangle = \langle \boldsymbol{\alpha} \prec \boldsymbol{\beta} | \boldsymbol{\gamma} \rangle$$

can usefully be described in purely algebraic language.



Our data

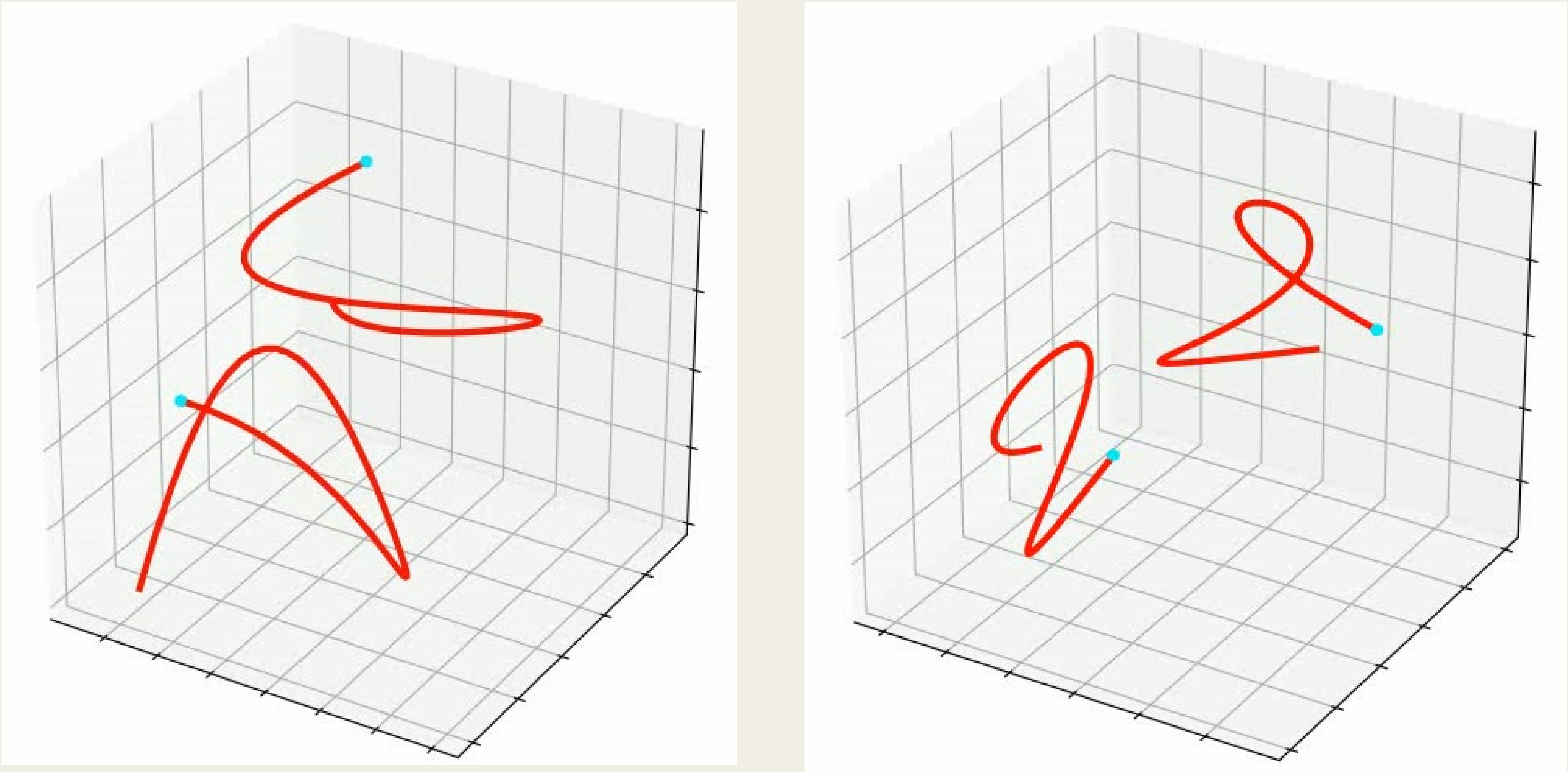



channel 1
channel 2
channel 4
channel 5
channel 7
channel 9
channel 13
channel 14
time
l representation of selected channels of one single streaming tree. Each plot represents the evolution in time of the value of
tree, on its various branches. A red dot indicates a point where the currently-tracked process sets off a child process, causing

Recovering the curves from the signature

DataSig

A rough path between mathematics and data science



Vectorisation of unsampled streams

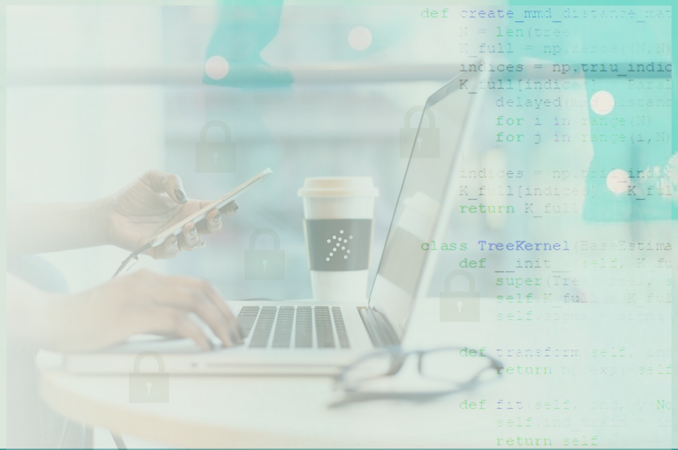
DataSig
A rough path between mathematics and data science

- signatures connect streams to the data science pipeline, but allows graduated approach
- expected signatures describe ensembles of paths
- pdes provide kernels on paths

Our collaboration

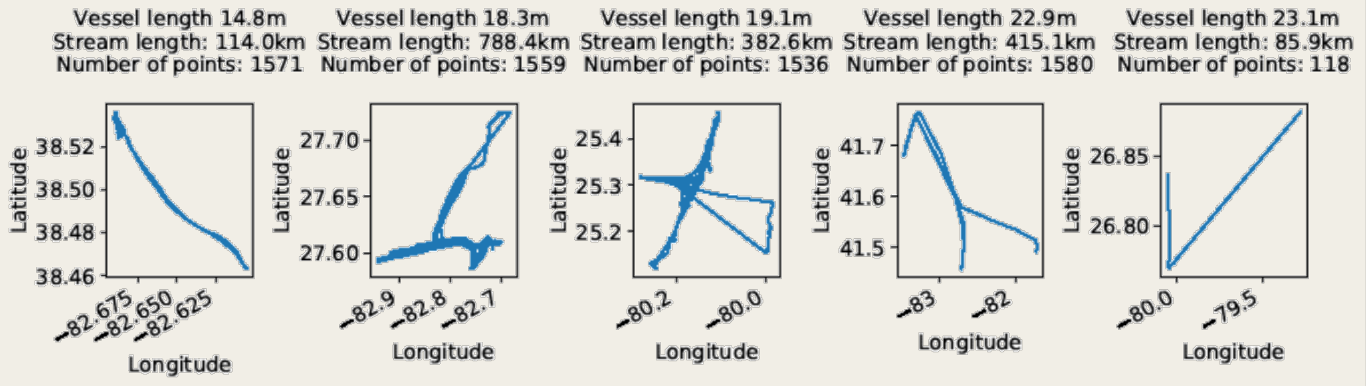
DataSig

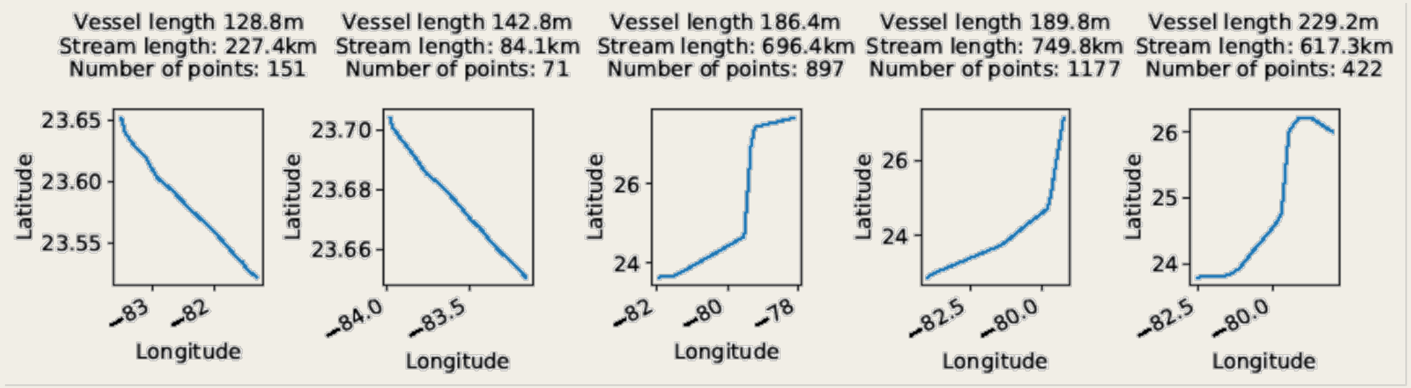
A rough path between
mathematics and data science



Signatures or Paths? |Long / Short vessel

- Demonstrated that novel math based method for anomaly detection is widely applicable and can detect anomalous streams where other methods fail.
- Applied this to the real-world shipping trajectories example.





Process tree example : Expected signatures of clouds of paths

Developed a way to apply expected signature techniques to host-based event logs, by viewing processes as trees evolving over time.

Applied this to malware classification: early results are that signature features improve classification of malware; further investigations are ongoing on richer data sets.

Process tree example : Expected signatures of clouds of paths

DataSig

A rough path between mathematics and data science

Developed SK-tree structure to apply standardised expected signature techniques to host-based event logs, by viewing processes as trees evolving over time analysed as expected signatures through a PDE kernel.

2102.07904.pdf (arxiv.org)

We demonstrate the SK-Tree to detect malicious events on a portion of the publicly available DARPA OpTC dataset, achieving an AUROC score of 98%

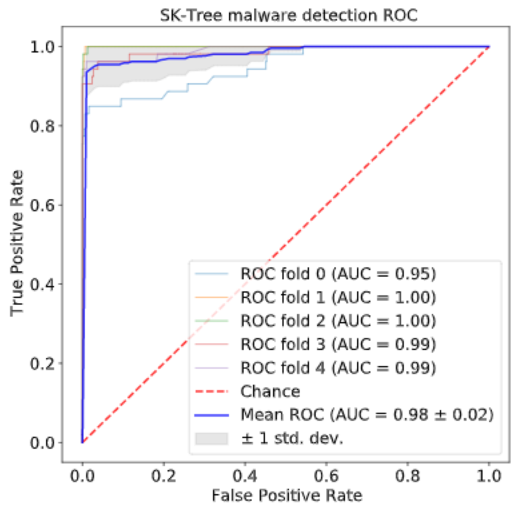
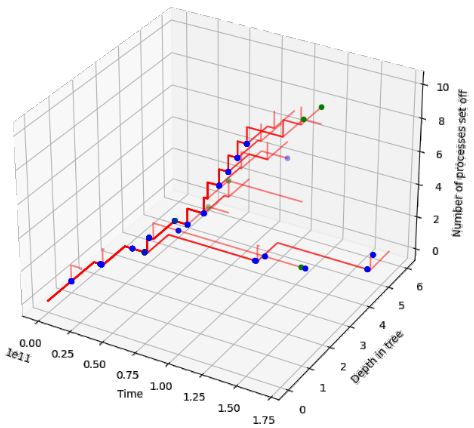
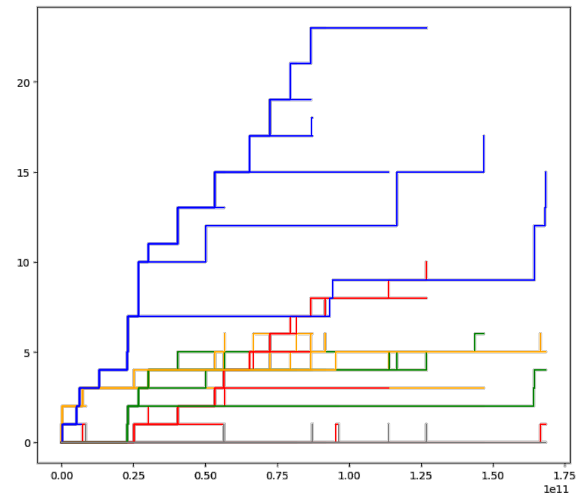


Fig. 2. ROC evaluation of the SK-Tree binary classifier on the OpTC data





Landmark based action recognition

A rough path between mathematics and data science

To communicate our methodology, we construct notebooks with introductory examples of what we can do.

People moving can easily be anonymized to landmarks. It is a static process. The moving stick people still contain information.

Peter Foster has put together a simple notebook you can run that demonstrates viable approaches to recognizing these actions that can be trained on small datasets.

Thanks for listening and over to Peter!

https://www.datasig.ac.uk/people