Birkbeck
UNIVERSITY OF LONDON

PDkit: A Scalable Computational Data Science Toolbox for High-Frequency Assessment of Parkinson's Disease

George Roussos
g.roussos@bbk.ac.uk

LONDON'S EVENING UNIVERSITY

Birkbeck
UNIVERSITY OF LONDON

Overview

- Motivation
- PDkit data processing pipeline
- PDkit user and contributor support
- Longitudinal processing with dataflow
- CUSSP study
- Future work

Motivation

1. Google Scholar: 1,000+ papers on Parkinsonian tremor using accelerometers and ML in 2018-19
   - Impossible to replicate and to compare results
   - Differences in data processing and algorithm implementation
   - In most cases, insufficient details provided to replicate algorithm used
2. Common pattern emerging:
   - Machine Learning processing pipeline
   - From raw data to severity assessment (often using MDS-UPDRS scores)

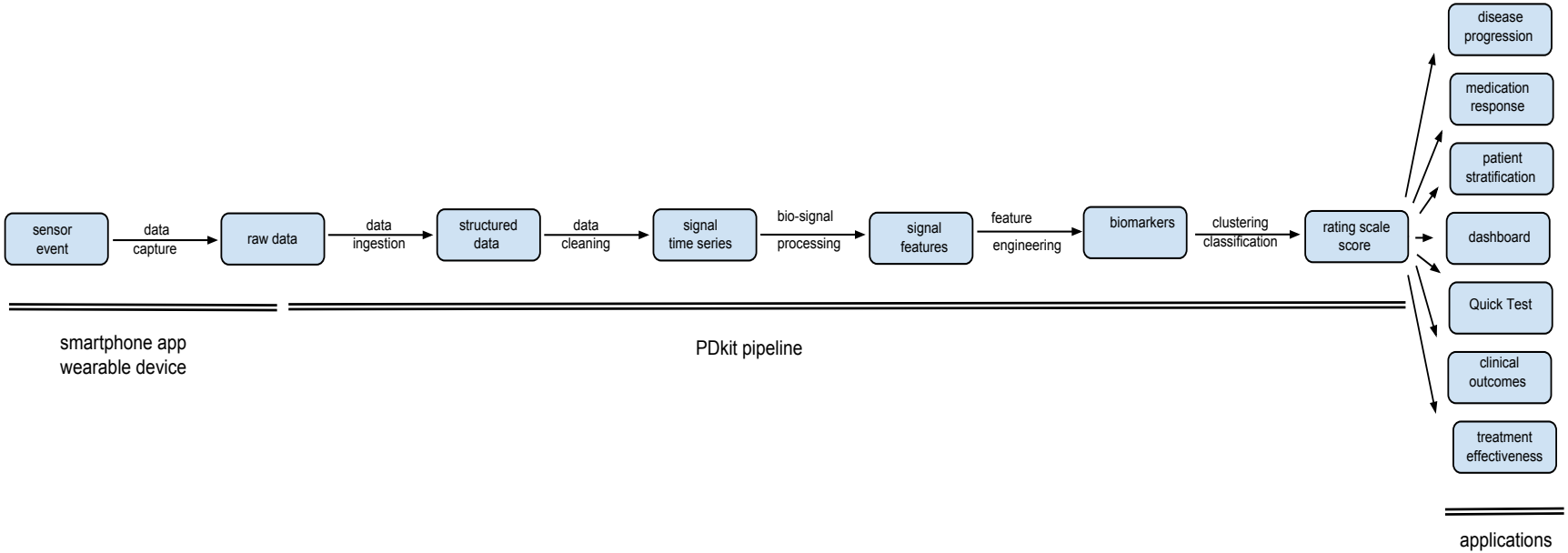
Birkbeck
UNIVERSITY OF LONDON

PDkit

- Specify digital analytics protocol precisely using software templates
- Unified extensible implementation of digital biomarkers
- Designed using a data processing pipeline approach
- API design follows common python machine learning approach scikit-learn

Source code `https://github.com/pdkit/pdkit`
Documentation `https://pdkit.readthedocs.io/`
Project info `https://pdkit.github.io/`
Docker images `https://hub.docker.com/r/pdkit/pdkit`

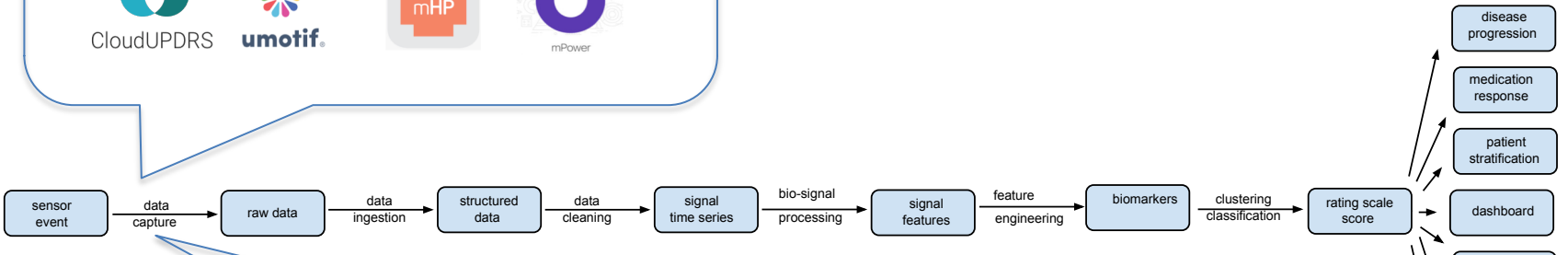
PDkit Processing Pipeline



PDkit Processing Pipeline

Smartphone apps
Modality: batch processing
Format: JSON, plain TXT

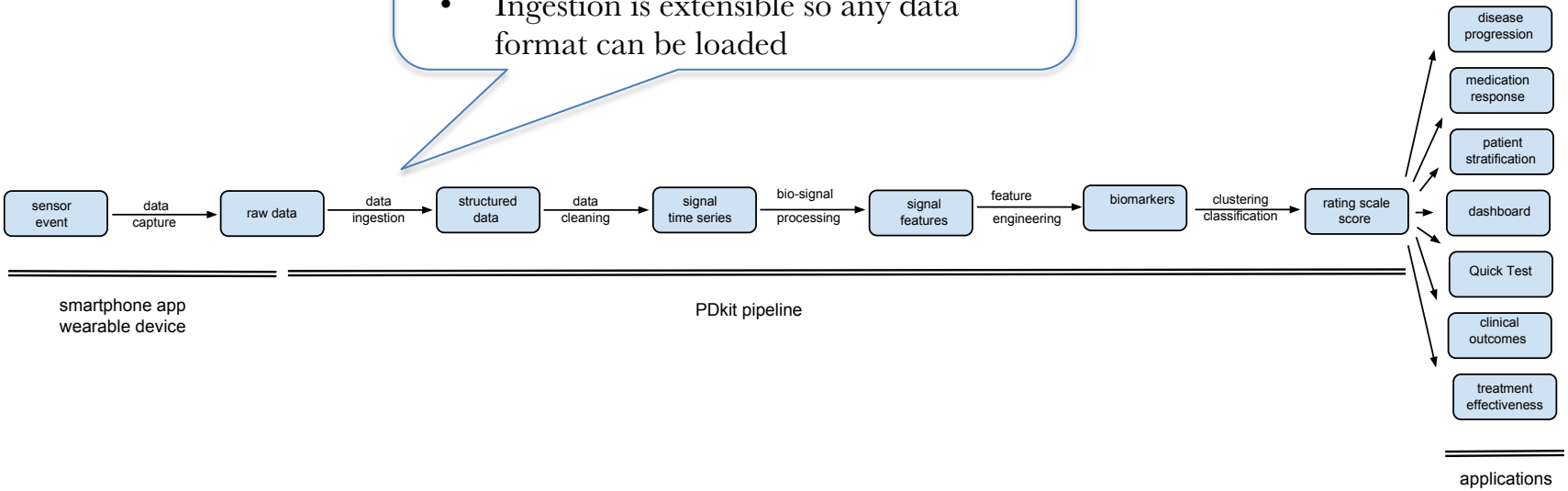
CloudUPDRS umotif. mHP mPower



Wearables
Modality: stream processing (edge or cloud)
Format: MQTT. Pub-Sub

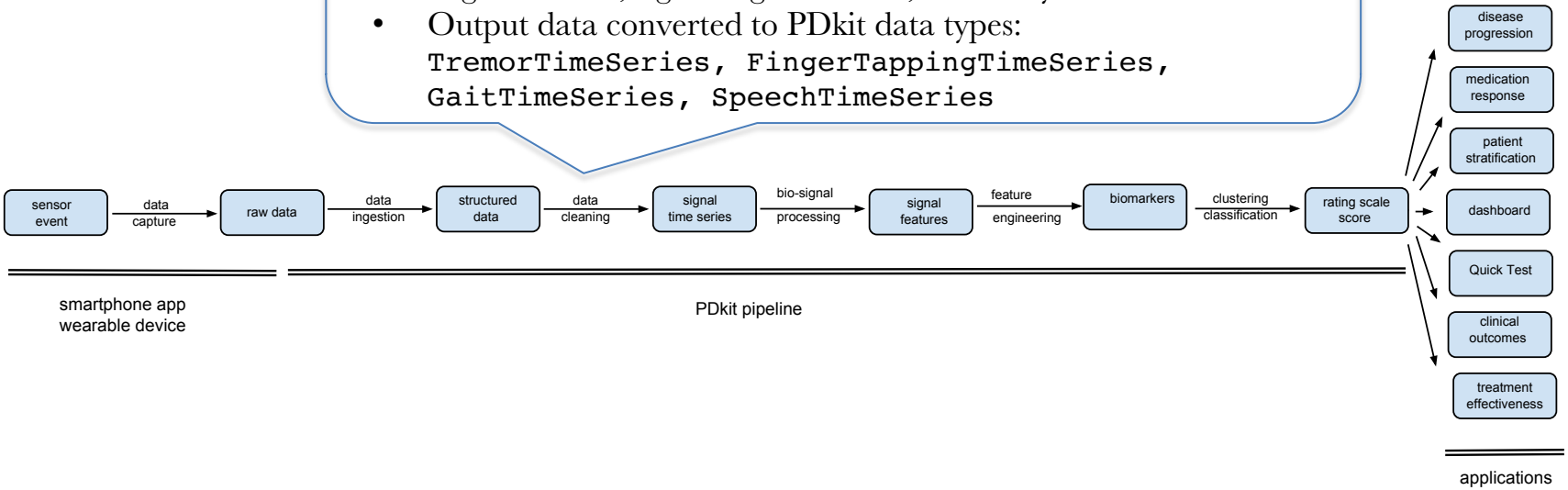
PDkit Processing Pipeline

- Data converted to standardized representation based on PANDAS dataframes
- Ingestion is extensible so any data format can be loaded



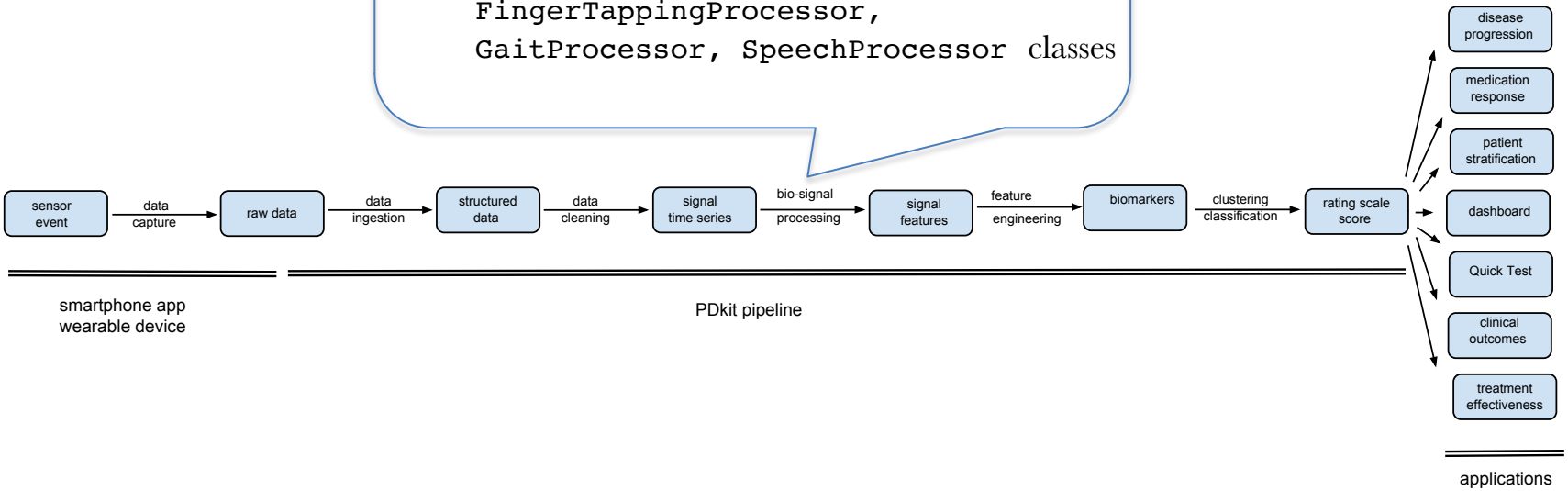
PDkit Processing Pipeline

- Missing and out of range values; data type normalization; indexing; standardised labelling; signal resampling
- Advanced features : gesture verification with DL, data augmentation, signal segmentation, extremity exclusions
- Output data converted to PDkit data types: `TremorTimeSeries, FingerTappingTimeSeries, GaitTimeSeries, SpeechTimeSeries`



PDkit Processing Pipeline

- Feature extraction
- 500+ tremor, bradikynesia, tapping, gait, turning and speech features
- Implemented in the `TremorProcessor`, `FingerTappingProcessor`, `GaitProcessor`, `SpeechProcessor` classes



PDkit Processing Pipeline

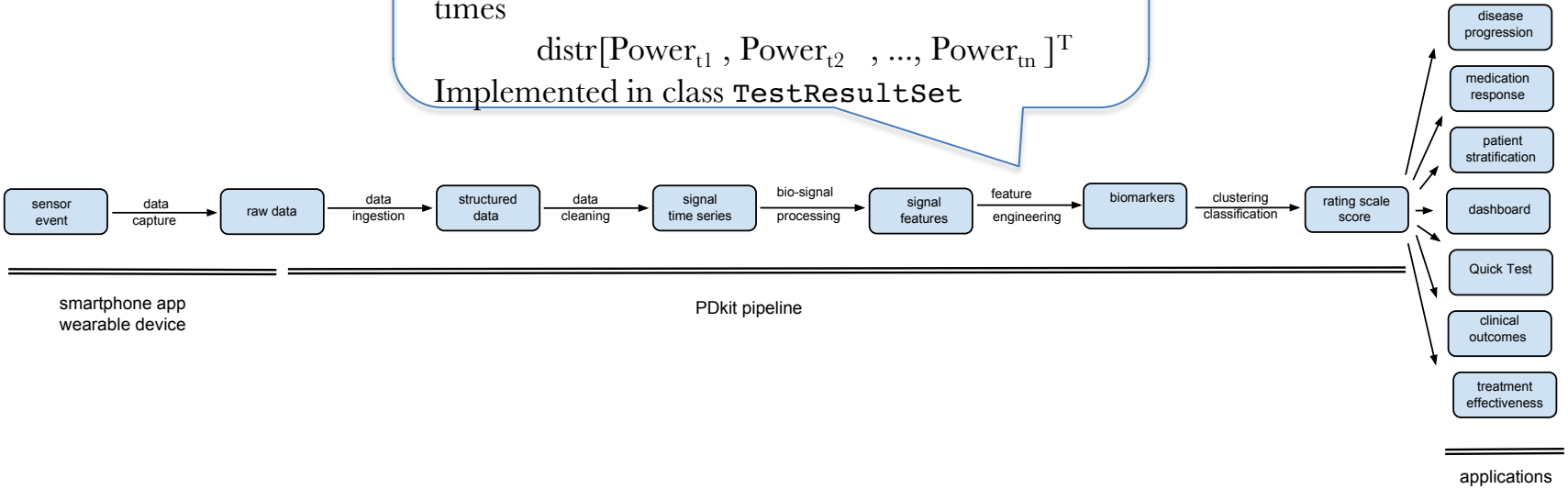
Classic unitary and vector instantaneous biomarkers

$$[\text{Power, Amplitude}]^T$$

Novel longitudinal biomarkers introducing temporal element e.g. same feature calculated at different times

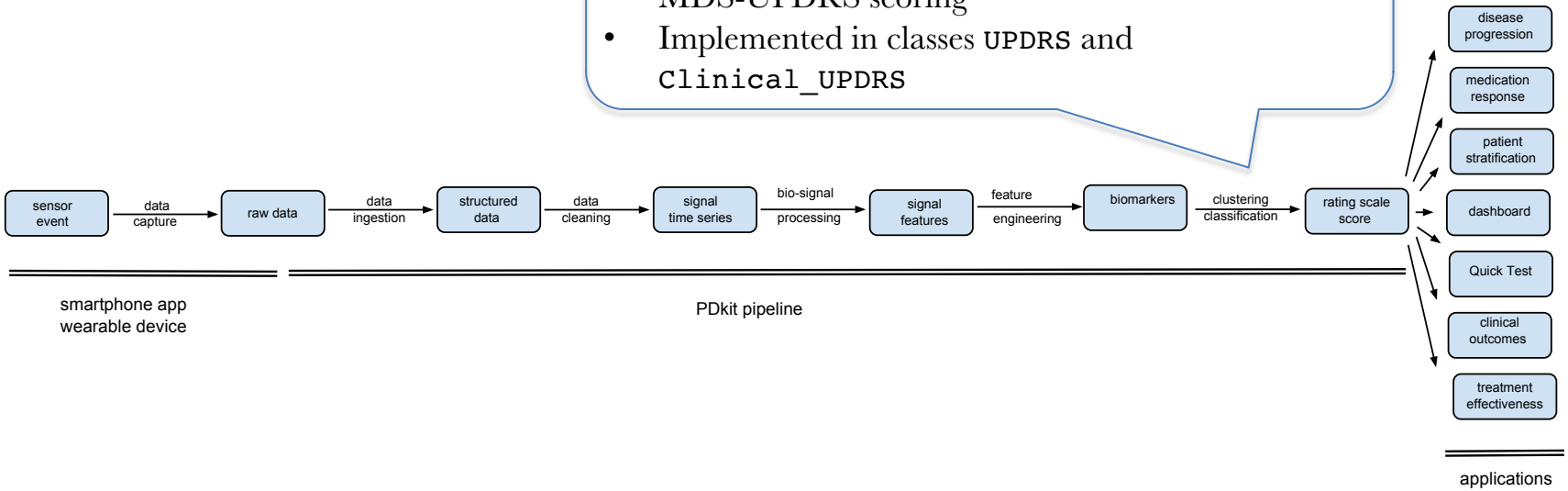
$$\text{distr}[\text{Power}_{t1}, \text{Power}_{t2}, ..., \text{Power}_{tn}]^T$$

Implemented in class `TestResultSet`



PDkit Processing Pipeline

- Supervised and unsupervised learning for scoring
- Clinical labels or classes created from the data only
- MDS-UPDRS scoring
- Implemented in classes `UPDRS` and `Clinical_UPDRS`



Birkbeck
UNIVERSITY OF LONDON

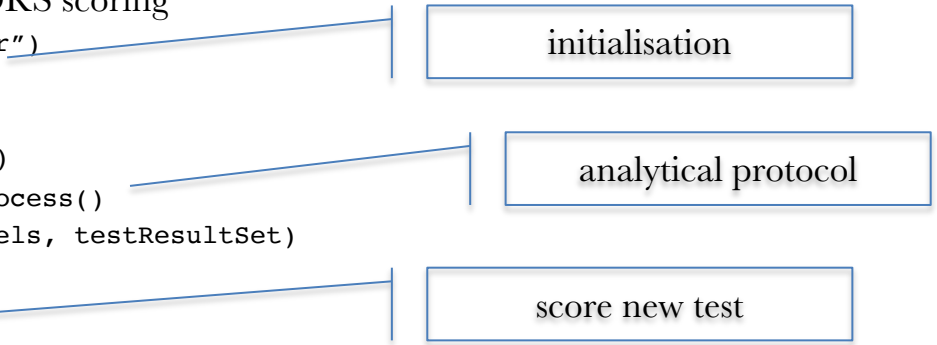
Examples

Train a model with clinical labels for MDS-UPDRS scoring

```
>> tp = pdkit.TremorProcessor("raw_data_folder")
>> ts = pdkit.TremorTimeSeries()

>> amplitude, freq = tp.amplitude(ts, 'welch')
>> welch_biomarker = pdkit.TestResultSet().process()
>> clinical_UPDRS =  pdkit.Clinical_UPDRS(labels, testResultSet)

>> clinical_UPDRS.predict(new_measurement)
```

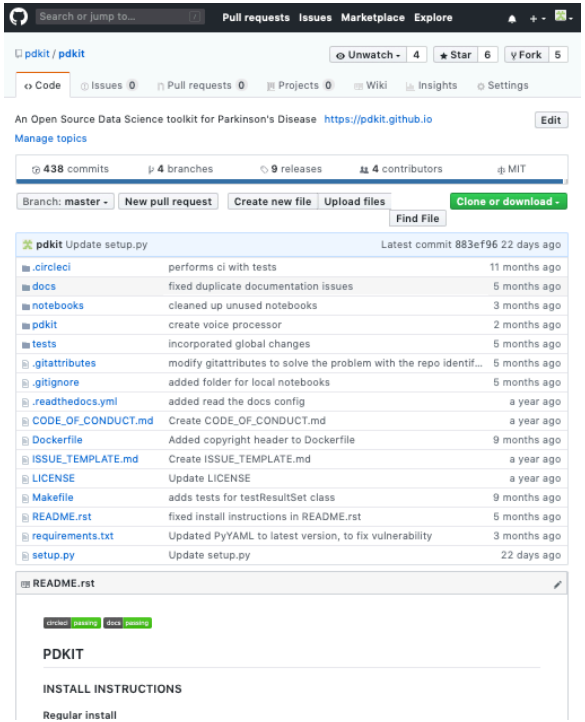


Calculate key gait features on a single data file:

```
>> ts = pdkit.GaitTimeSeries().load(filename)
>> gp = pdkit.GaitProcessor()
>> step_regularity, stride_symmetry = gp.walk_regularity_symmetry(ts)
```

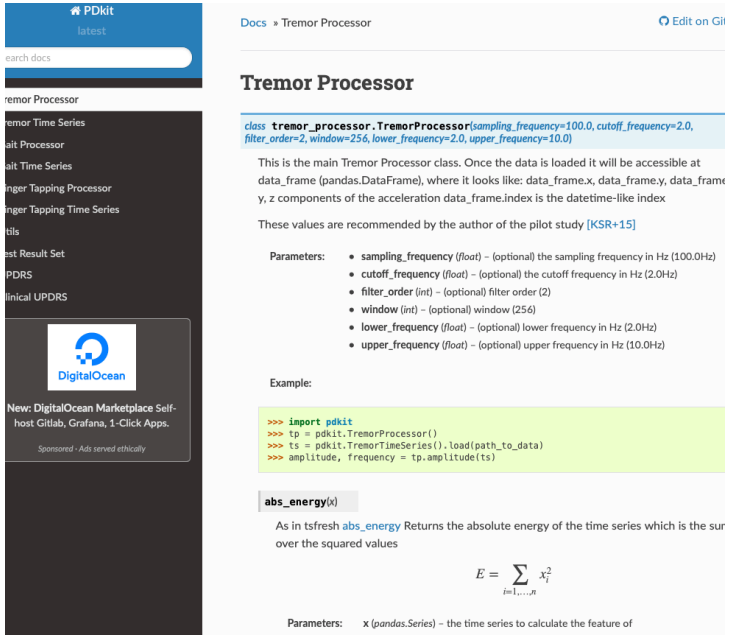
Availability

- Source code on github
  - Continuous Integration ensures working code
  - MIT License



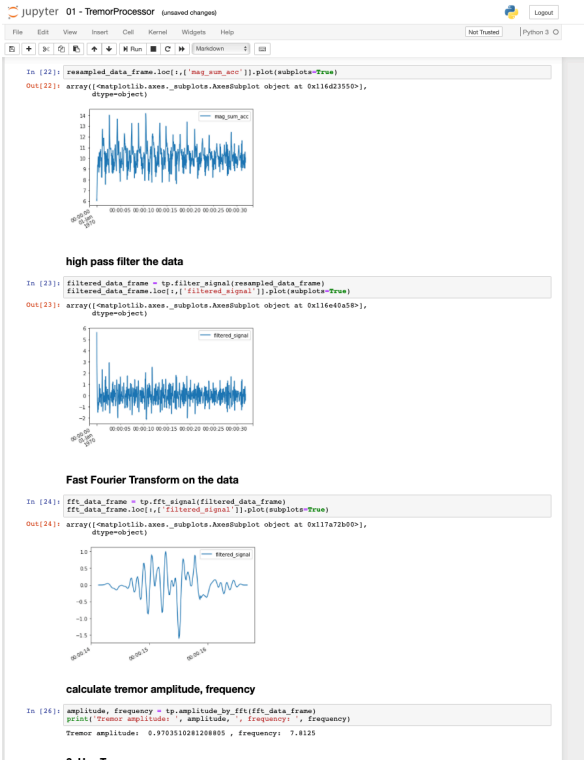
Availability

- Source code on github
  - Continuous Integration ensures working code
  - MIT License
- Documentation on Read the Docs
  - References to literature
  - Explanation of calculations
  - Auto-generated from the source code



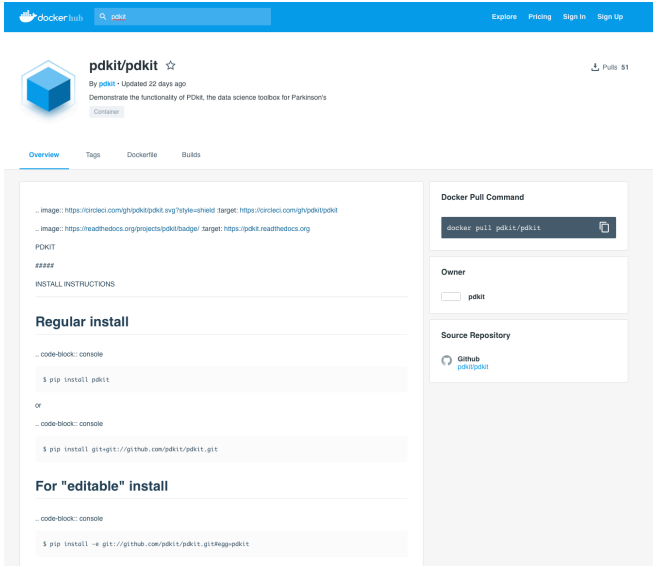
Availability

- Source code on github
  - Continuous Integration ensures working code
  - MIT License
- Documentation on Read the Docs
  - References to literature
  - Explanation of calculations
  - Auto-generated from the source code
- Jupyter Notebooks demonstrate functionality
  - Notebooks can run on Binder, Deepnote and Colaboratory



Availability

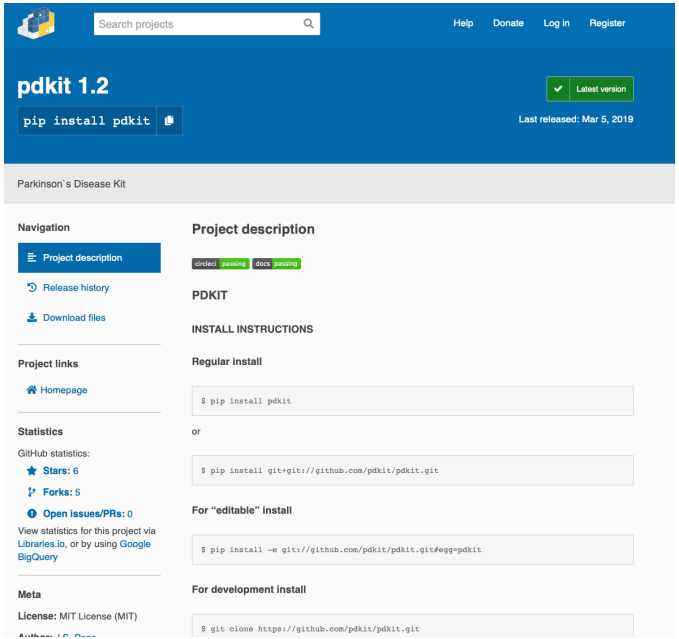
- Source code on github
  - Continuous Integration ensures working code
  - MIT License
- Documentation on Read the Docs
  - References to literature
  - Explanation of calculations
  - Auto-generated from the source code
- Jupyter Notebooks demonstrate functionality
  - Notebooks can run on Binder, Deepnote and Colaboratory
- Docker images for quick deployment
  - Publicly available at Docker Hub



Birkbeck
UNIVERSITY OF LONDON

Availability

- Source code on github
  - Continuous Integration ensures working code
  - MIT License
- Documentation on Read the Docs
  - References to literature
  - Explanation of calculations
  - Auto-generated from the source code
- Jupyter Notebooks demonstrate functionality
  - Notebooks can run on Binder, Deepnote and Colaboratory
- Docker images for quick deployment
  - Publicly available at Docker Hub
- Available on PyPI the python package registry



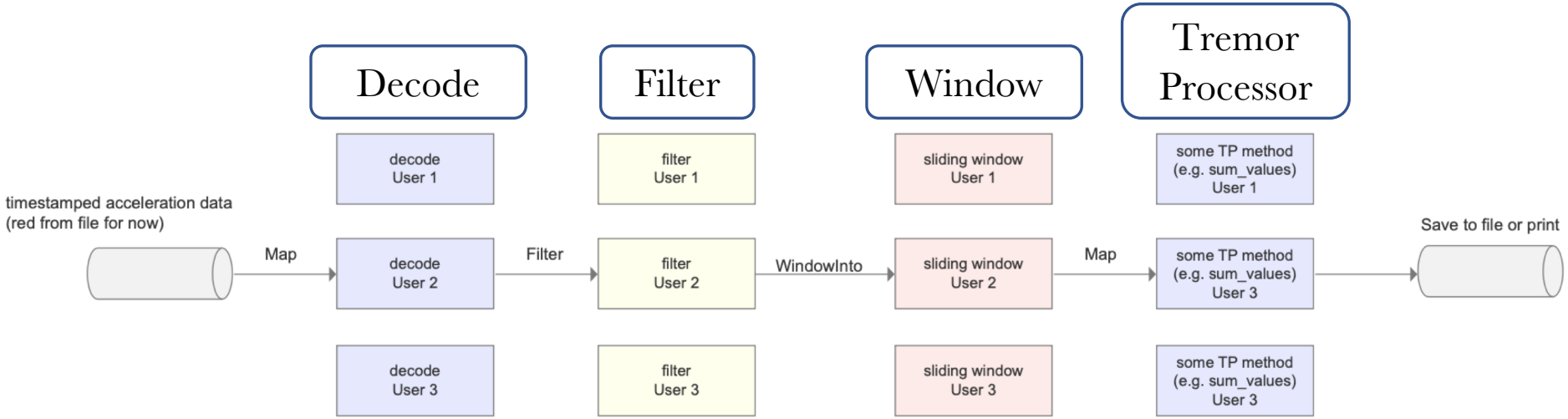
Birkbeck
UNIVERSITY OF LONDON

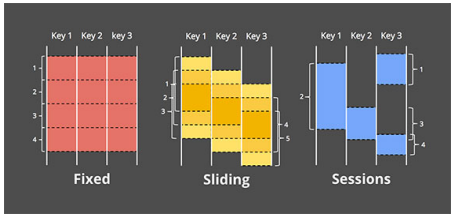
Dataflow programming

- Advanced unified programming model:
  - stream processing from wearables
  - batch processing from smartphones
- Can deal with infinite out of order datasets
- Scaling out on cloud architectures
- Platform selected Apache BEAM
- Describe calculations as a directed graph which is mapped automatically by BEAM to the underlying infrastructure
- Infrastructure mapped using runners
- Demonstrate streaming PDkit

Source code `https://github.com/pdkit/pdkit-beam`

Tremor processing with Beam-PDkit





Birkbeck
UNIVERSITY OF LONDON

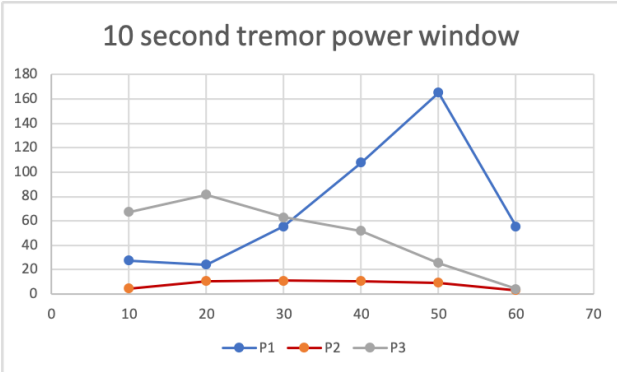
Example

```
with beam.Pipeline(options=options) as p:
        messages = (p
                        | 'Read from Stream' >> beam.io.gcp.pubsub.ReadFromPubSub(topic=known_args.input_topic)
                        .with_output_types(bytes)
                        )
        windowed_data = (messages
                         | 'ParseMagSumAcc' >> ParseMagSumAcc(30,10)
                         )
        grouped = (windowed_data
                    | 'GroupWindowsByUser' >> beam.GroupByKey()
                    | 'Calculate pdkit method' >> CalculatePDkitMethod()
                    | 'UserScoresDict' >> beam.ParDo(UserDict())
                    )
        welch = (grouped
                  | 'Parse it' >> beam.Map(lambda elem: (elem['user'], elem['result'], elem['start'], elem['end']))
                  )
```

Birkbeck
UNIVERSITY OF LONDON

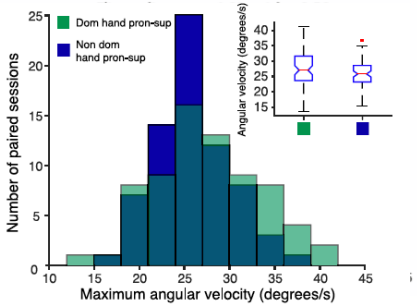
Case study

- Experiment:
  - 100,000 concurrent patients
  - 3d wrist sensor sampled at 50 Hz
  - 12 hours monitoring
  - Tremor power using Welch PSD calculated
- Show rapid tremor variations
- Total cost on GCE: $ 9.43

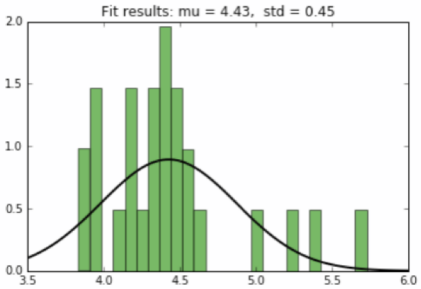


Longitudinal biomarkers

- Observation: Any feature can be turned to a longitudinal biomarker
- Consider the sample distribution rather than individual measurements
- Dataflow processing is the key ingredient towards a process sampling paradigm
- Early evidence suggests significantly higher consistency and precision



Pissadaki *et al* IBM Journal, 2018



July (43 samples)

Birkbeck
UNIVERSITY OF LONDON

CUSSP study

- CUSSP at the UCL Institute of Neurology and Homerton Hospital (UK)
  - Details https://clinicaltrials.gov/ct2/show/NCT02937324
  - Data collection completed in May
  - 74 patients
- 20 lines of PDkit source code specify processing protocol
- 2-3 hours of software development
- Can recreate results in 1 hour on standard laptop
- Clinical results to be published this summer

Current and future work

- Development of gait biomarkers (except symmetry) present distinct challenges
- Better objectives measures seem to be obtained by turning rather than straight line walking performance
- Fully automatic assessment of turning requires:
  - Signal segmentation with Bellman k-segmentation algorithm
  - Turn speed detection with Madgewick filter

Birkbeck
UNIVERSITY OF LONDON

Closing remarks

- PDKit is an effective tool to precisely specify analytical processing framework of a study
- Open and inclusive development model
- Challenge: Published code not always properly licensed
- Challenge: Increase user and contributor network

PDkit people

Joan Saez

Cosmin Stamate

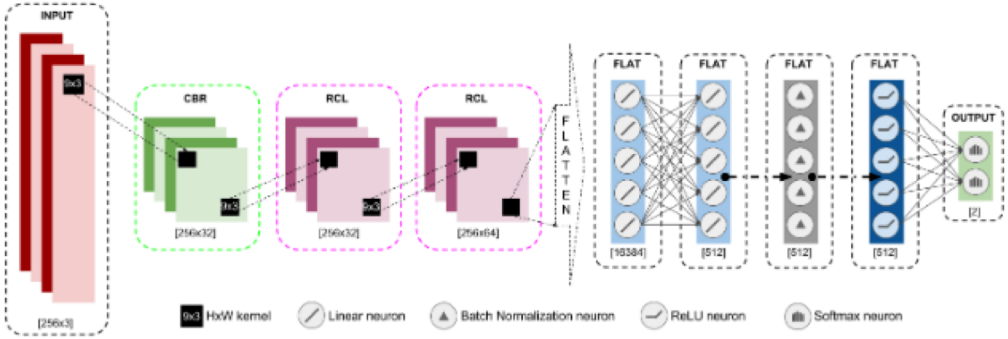
David Weston

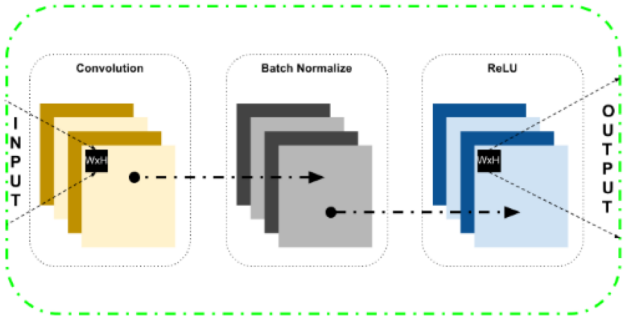
Ashwani Jha

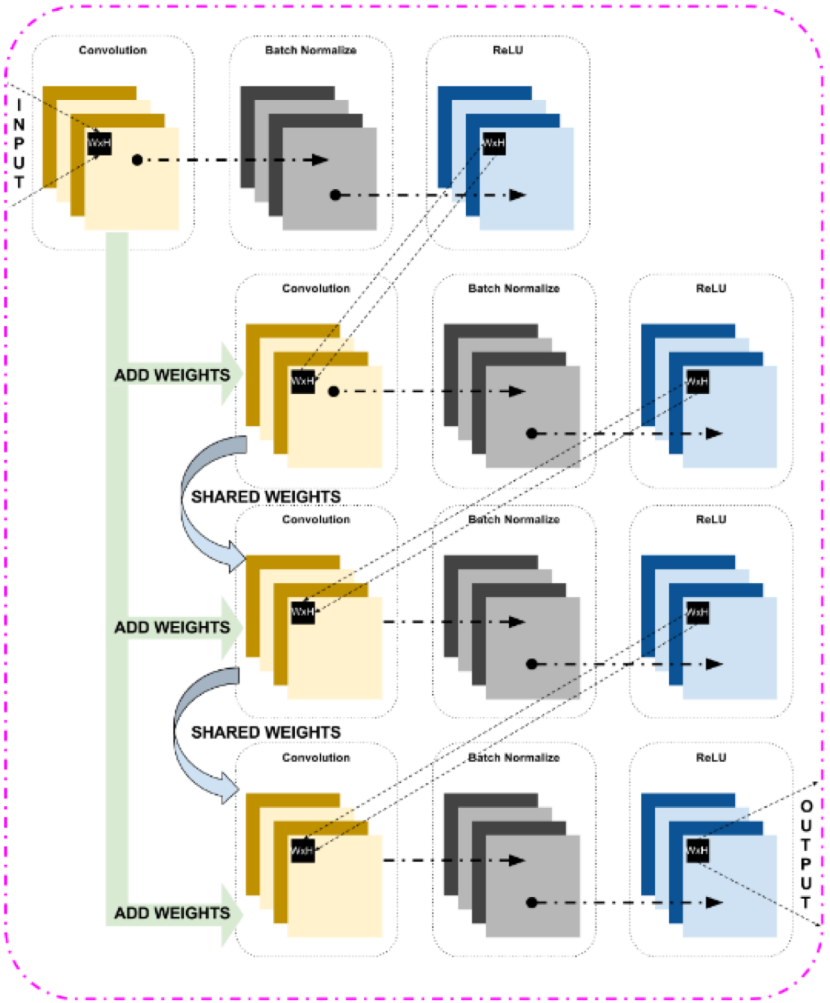
John Rothwell

Nikos Fragopanagos

Deep Learning Architecture



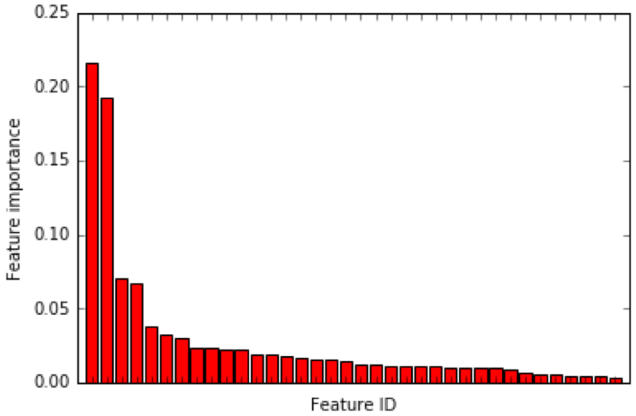




Birkbeck
UNIVERSITY OF LONDON

Quick Test

- UPDRS exhaustive search of all possible symptoms
- Each patient presents only a few
- Symptoms typically change slowly e.g. 6 months
- ~6 features are predictive of overall score
- Use ML to identify the specific tests that offer the highest inferential power
  - Observer five full tests
  - Apply standard ensemble of randomized decision tree method to rank tests according to predictive strength
  - Select top 3 tests for individualised quick test



Signal quality assesement