In development:
A package for deep learning of 'omics data

RORY STARK

PRINCIPAL BIOINFORMATICS SCIENTIST

CANCER RESEARCH UK CAMBRIDGE INSTITUTE
UNIVERSITY OF CAMBRIDGE

6 DECEMBER 2017

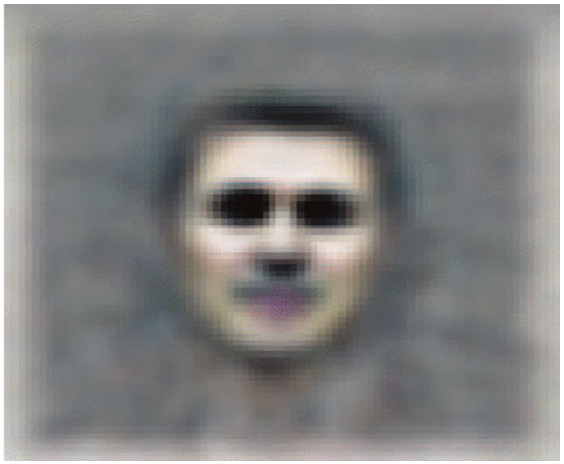
UNIVERSITY OF CAMBRIDGE

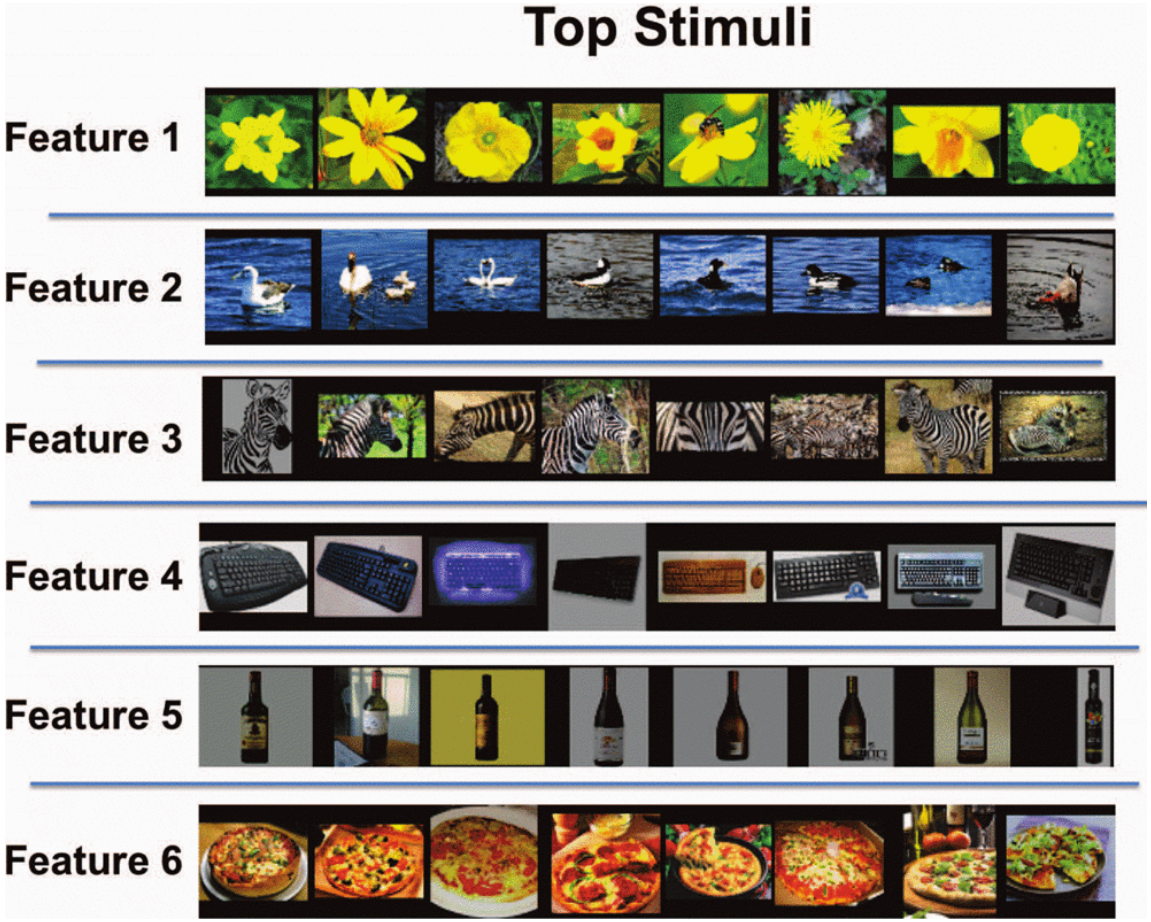
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Deep Learning

- **3rd wave of artificial neural networks**
- **Highly complex nonlinear predictors with 10^7 -10^9 parameters**
- **Beyond prediction/classification**
  - **Feature discovery**
    - Dimension reduction
    - Dimension expansion (re-mapping/sparse coding)
    - ANNs are *not* a black box!
  - **Content-addressable memory**
    - Robust to missing/incomplete/noisy/erroneous data
    - Easy to match a new pattern to a trained one
    - Can hypothesize “prototype” patterns

Beyond prediction: Feature Discovery





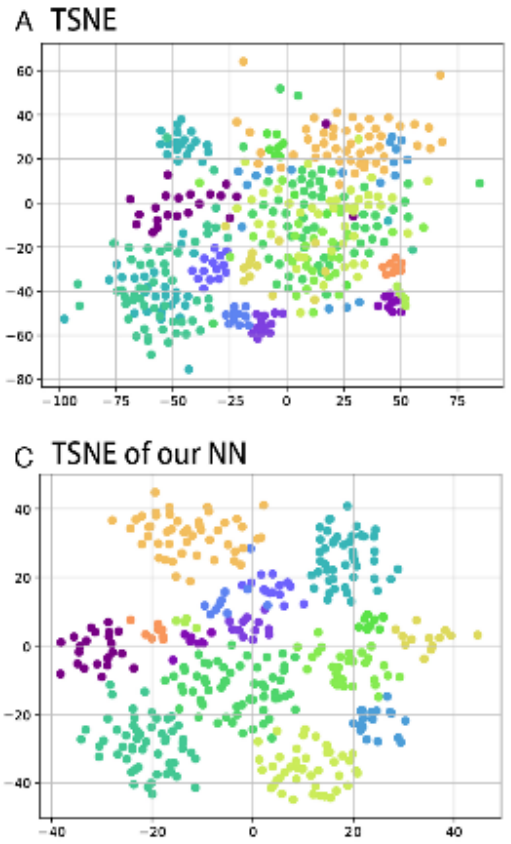
Follow the Data

– Key driver: data

- **Large numbers of samples**
  - – Best results with 10^5-10^7 samples; generally 10^4 minimum
  - – Optimally annotated (pre-classified)
  - – Greatest success:
    - » Images (spatial data): low resolution using CNNs
    - » Temporal (sequence data): LSTM
- ***Bulk 'omics data is problematic***
  - – Very high-dimensional samples
  - – Relatively few samples in bulk experiments (10^3-10^4)

– Single cell assays have higher sample counts

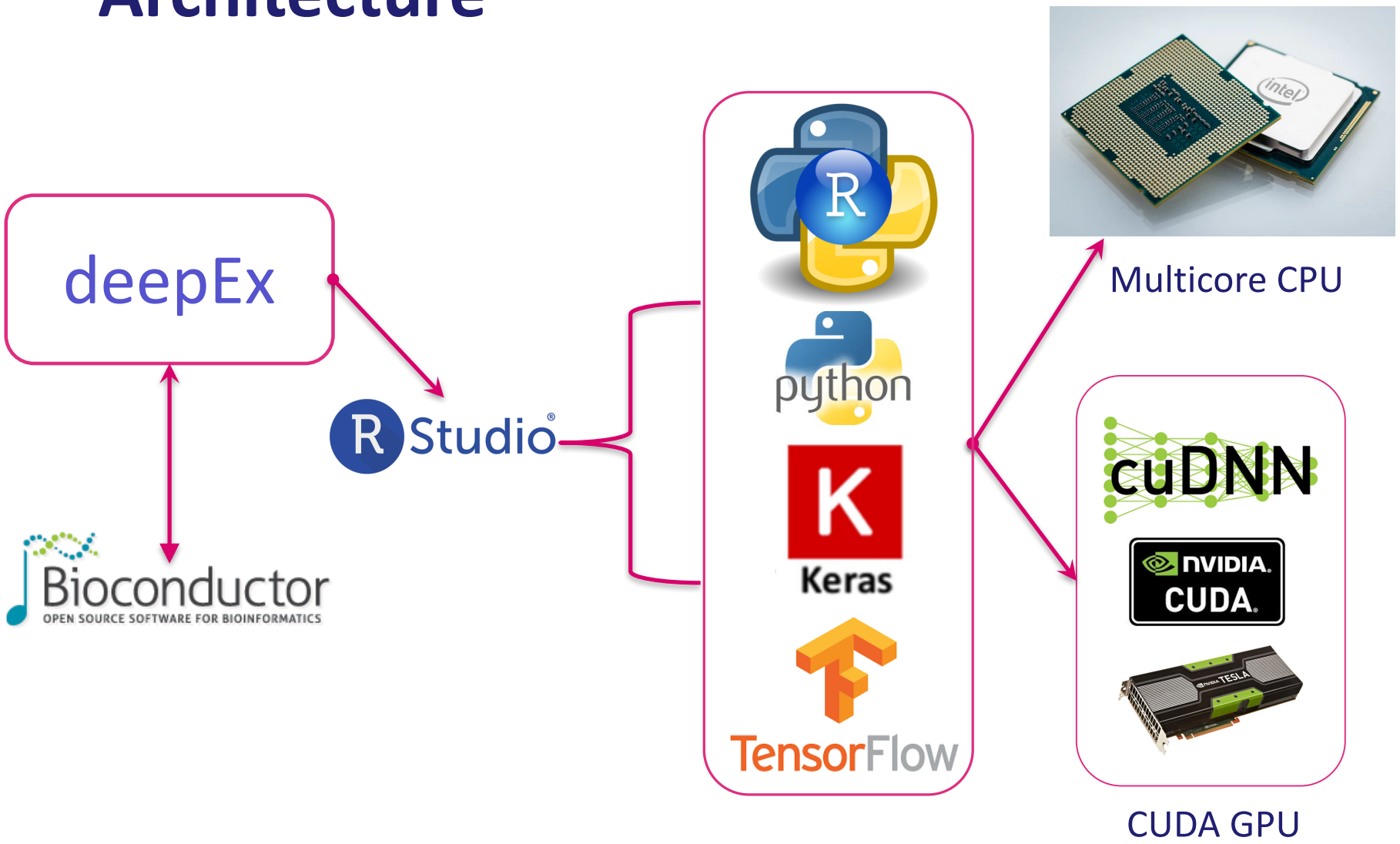
- *eg.* Lin *et al.* (2017) "*Using neural networks for reducing the dimensions of single-cell RNA-Seq data*"



Goals of package under development

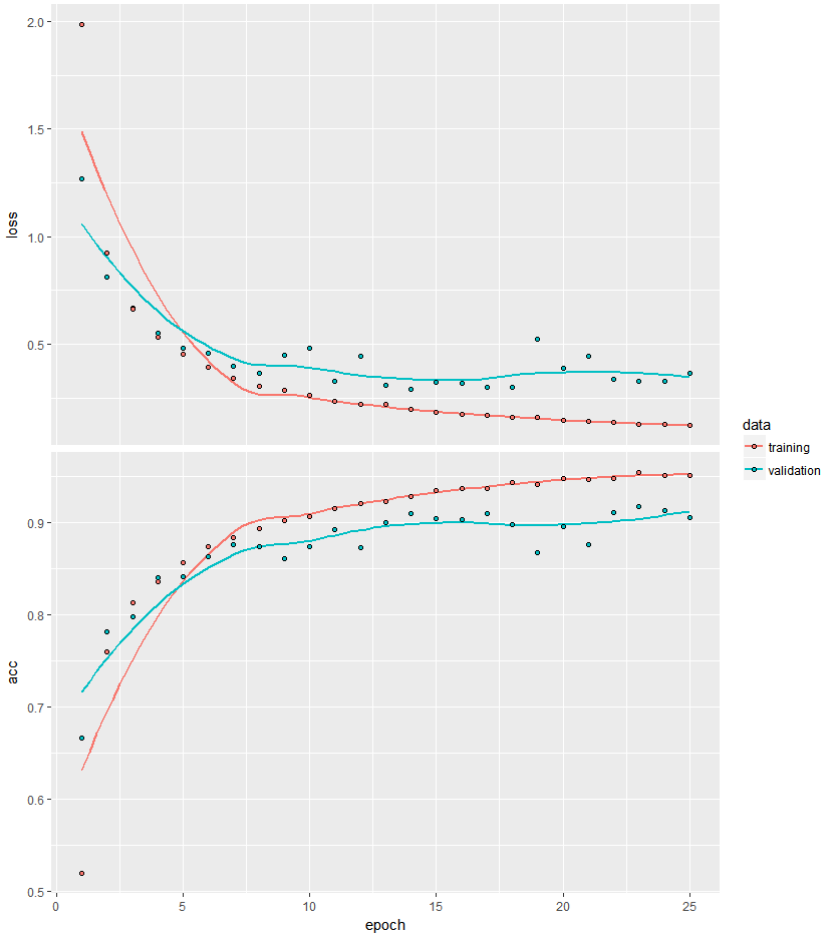
- Flexible, high-level R interface to deep learning tools
- Connect Bioconductor datatypes to deep networks
  - SummarizedExperiment
  - RangedSummarizedExperiment
  - SingleCellExperiment
  - MultiAssayExperiment
- Simplify complete analysis
  - Model building (AE/DAE/SDAE + RBM)
  - Classification (MLP, CNN)
  - Feature discovery, clustering (t-SNE)
  - Biological associations (GO/GSEA/Leading Genes)

Architecture



"DeepExperiment"

- `deepEx`
  - Experiment
  - classes (missing = AE)
  - hidden=c(100,sqrt,.5)
  - activation
  - loss
  - optimizer
  - initializer
- `fit`
  - deepEx
  - assay
  - train/test
  - dropoutRate
  - learningRate
  - batchSize
  - noise
  - epochs
- `represent`
  - Retrieve hidden unit activations
- `plot`
  - PCA/t-SNE/heatmap
  - Tensorboard
- `analyze`
  - Top samples per unit
  - Top genes per unit
  - GO/GSEA



Issues

- **Could this be a Bioconductor package?**
  - **DEPENDS on keras/tensorflow/reticulate packages on RStudio github (JJ Alaire)**
  - **Python/Keras/Tensorflow would optimally be pre-installed on build machines**