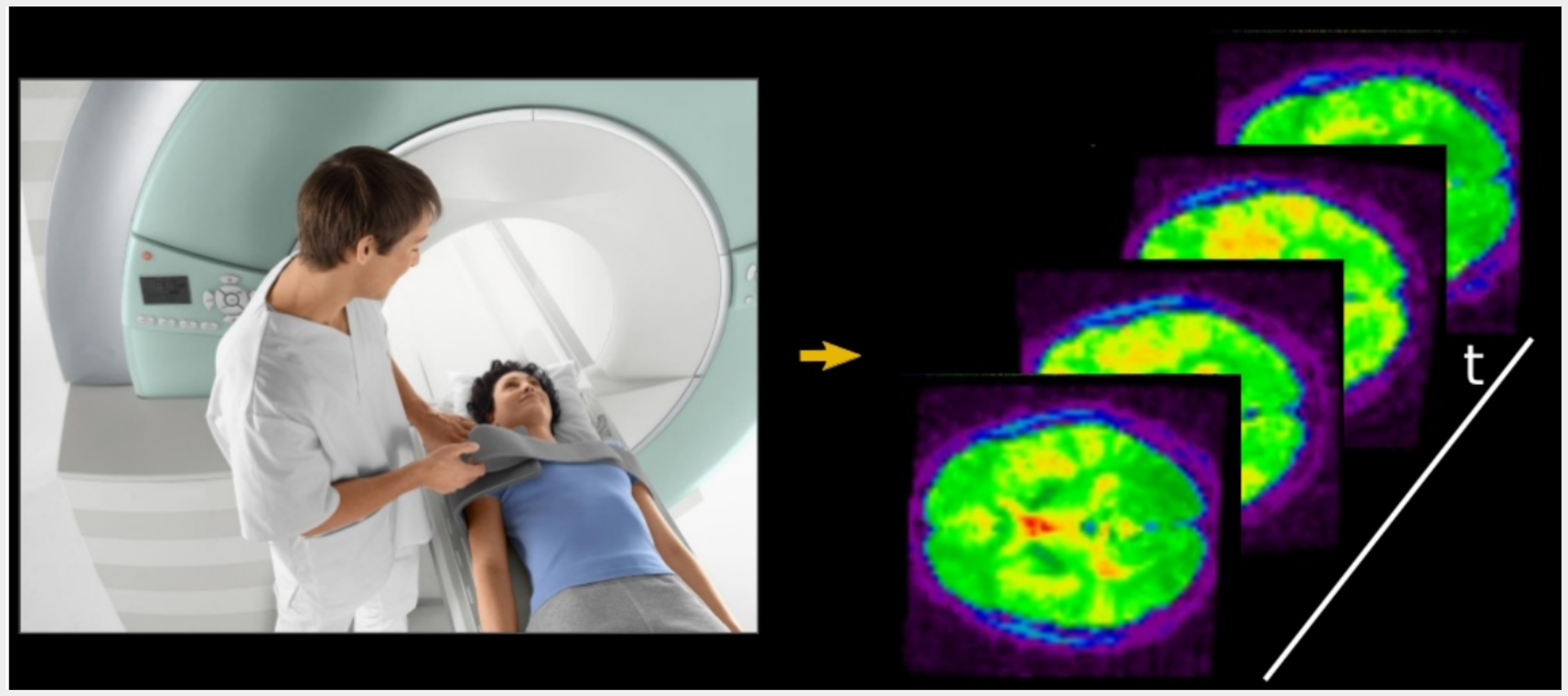
BIS BIG DATA PANEL

LOÏC ESTÈVE

LOIC.ESTEVE@INRIA.FR

PARIETAL TEAM @ INRIA



- try to learn a two-way mapping between brain activity and cognitive function

PARIETAL

Inria

INVENTORS FOR THE DIGITAL WORLD

Disclaimer: datasets up to ~TB but more typically 1-100GB

WHY PYTHON?

- interactive language, key for data exploration
- General purpose language
- Easy to read/write
- mature scientific Python stack (numpy, scipy, matplotlib, etc …)
- Performance through numpy, cython

SCIKIT-LEARN VISION

- an enabler: machine learning without having to learn the machinery
- High quality Pythonic software library: interfaces designed for users
- community-driven development: BSD licensed with diverse contributors

SCIKIT-LEARN OVERVIEW

- very rich feature set:
  - supervised learning: decision trees, linear models, SVM, …
  - unsupervised learning: clustering, dictionary learning, …
  - model selection: built-in cross-validation, parameter optimization
- performance matters
- used in production by data-driven companies (Spotify, Evernote, New York Times)

For more details see http://scikit-learn.org/

STRATEGIES FOR TACKLING BIG DATA

- feature reduction
    - randomized projections: embedding into a lower dimensional space that conserves distances
    - feature clustering: on images super-pixel strategy
- online learning: learn one sample at a time, e.g. IncrementalPCA

JOBLIB

- never recompute the same thing twice
- fast hashing of input numpy arrays
- helper functions for parallel computing

NOTES ON HARDWARE

- Parietal team workhorse: single server with 384 GB RAM, 48 cores, 70 TB storage (SSD cache on RAID controller)
- gets the work done faster than our 800 CPU cluster

Nobody ever got fired for buying an Hadoop cluster

SUMMARY

TODO: less than great

- scikit-learn: machine learning library focused on usability and performance
- specific strategies can be used to scale to big data
- joblib facilitate TODO
- single big memory server can outperform cluster