CoBrain Project: Analysis of Big Data in *Clinical* Neuroscience

02.12.16

Mikhail Belyaev

Center for Data Intensive Science and Engineering,

Skolkovo Institute of Science and Technology

Skoltech

Skolkovo Institute of Science and Technology

Outline

- Overview: Big Data
- Big Data in Clinical Neuroscience
- Deep Learning for Neuroimaging
- CoBrain project

Big Data: three aspects (V3s)

- **Volume** - Volume describes the amount of data generated by organizations or individuals. Big Data is usually associated with this characteristic.

- **Velocity** - Velocity describes the frequency at which data is generated, captured and shared.

- **Variety** - Big data means much more than rows and columns. It means unstructured text, video, audio that can have important impacts on company decisions – if it's analyzed properly in time.

Big Data: examples

- eBay uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising.
- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes of data.
- 1 PB = 1 000 TB = 1 000 000 GB
- A modern PC has 8-16 GB of memory and 1-2 TB of storage

What about clinical data? An example: Enroll-HD

Enroll-HD - an ongoing, prospective, open-ended, globally standardized, longitudinal, observational study of Huntington Disease

Key facts:

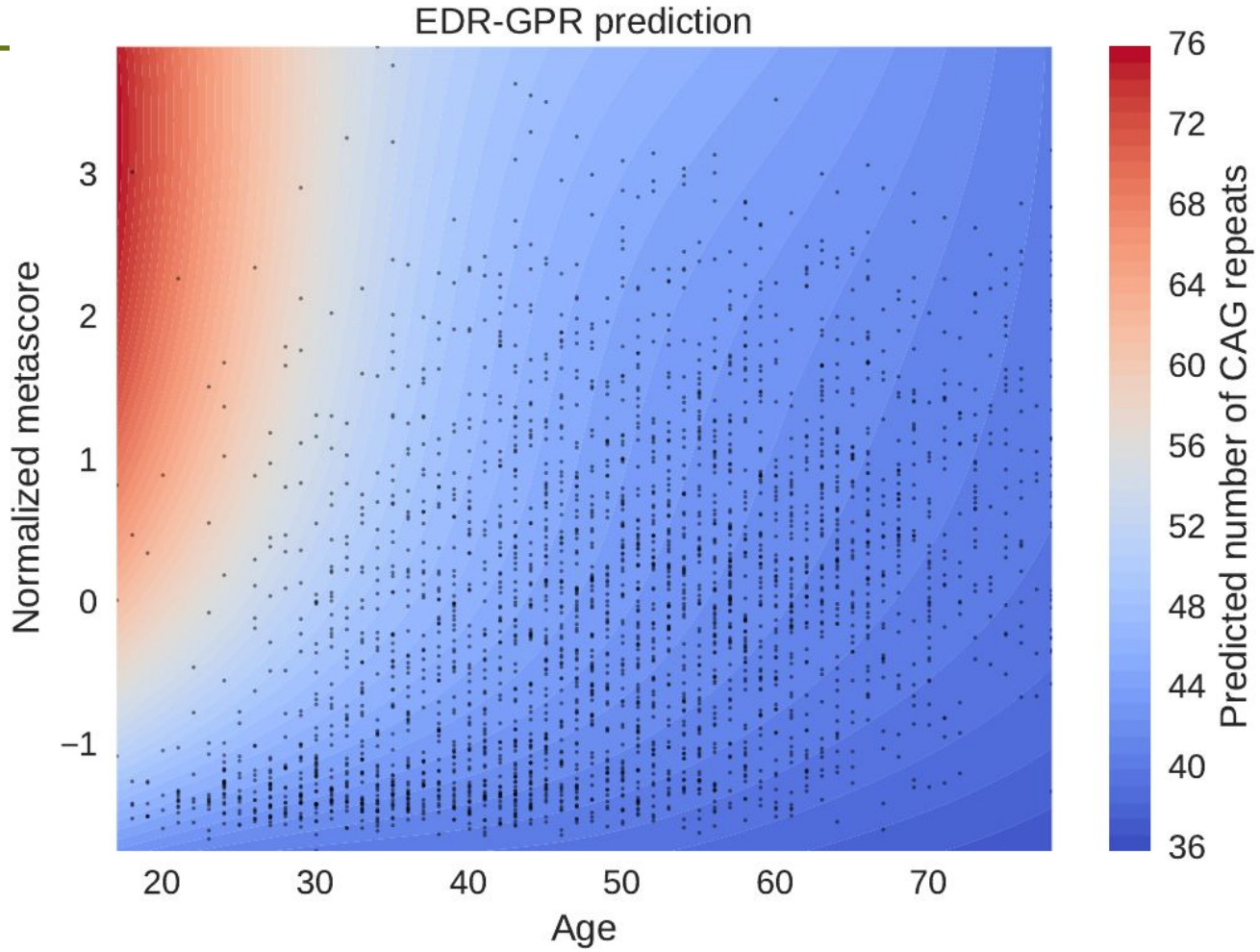
- **12142** participants
- **140** clinical sites worldwide
- **14** nations participating
- More than **200** features: various tests (motor, cognitive, etc), demographic data, medical history, etc

It's an impressive and very complex effort.

https://www.enroll-hd.org

What about medicine? An example: Enroll-HD

The Enroll-HD dataset is large enough to find interesting facts, i.e. predict genotype based on set of phenotypic features.



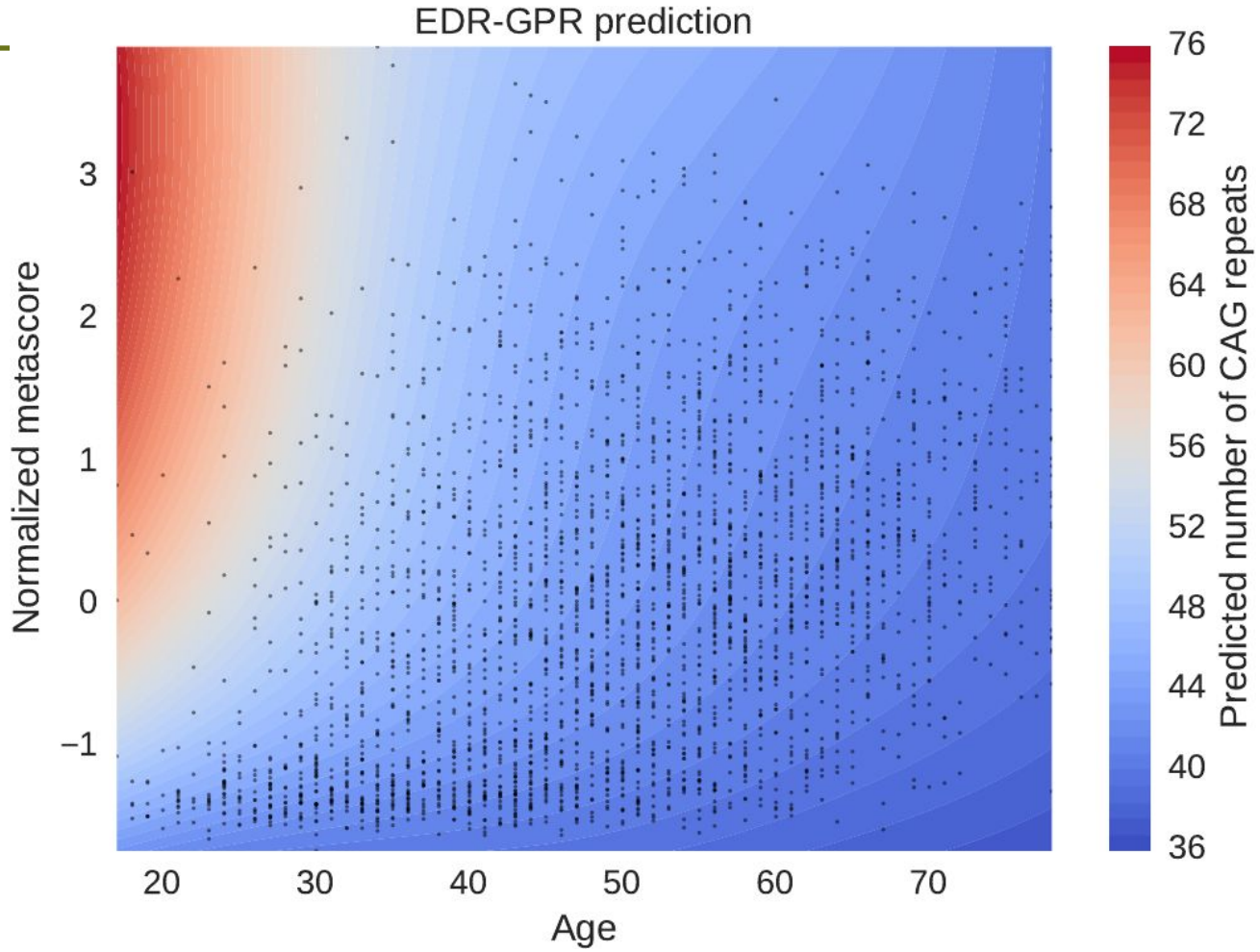
The Size of the CAG-Expansion Mutation Can Be Predicted in HD Based on Phenotypic Data Using a Machine Learning Approach 9th European Huntington's Disease Network Plenary Meeting - 2016 Y. Seliverstov, Enroll-HD investigators, S. Illarioshkin, B. Landwehrmeyer, M. Belyaev

Skoltech
Skolkovo Institute of Science and Technology

What about medicine? An example: Enroll-HD

The Enroll-HD dataset is large enough to find interesting facts, i.e. predict genotype based on set of phenotypic features.

Total size of the dataset is **8MB**



The Size of the CAG-Expansion Mutation Can Be Predicted in HD Based on Phenotypic Data Using a Machine Learning Approach 9th European Huntington's Disease Network Plenary Meeting - 2016 Y. Seliverstov, Enroll-HD investigators, S. Illarioshkin, B. Landwehrmeyer, M. Belyaev

Skoltech
Skolkovo Institute of Science and Technology

Big Data in Medicine

- Typical Big Data problems are based on “automatic” generation of data -> it’s possible to generate millions of data entries
- Medical data usually require some manual activities by a doctor (i.e. to perform a test) -> the number of data entries is relatively low
- What if each data entry is large?

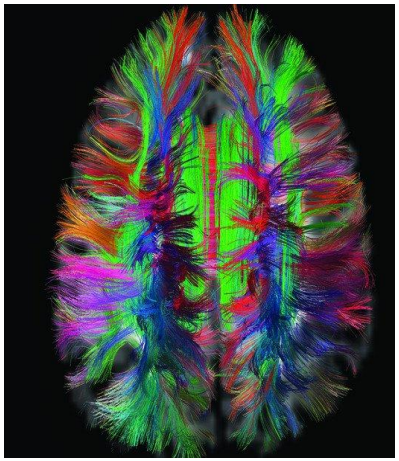
Big Data in Clinical Neuroscience

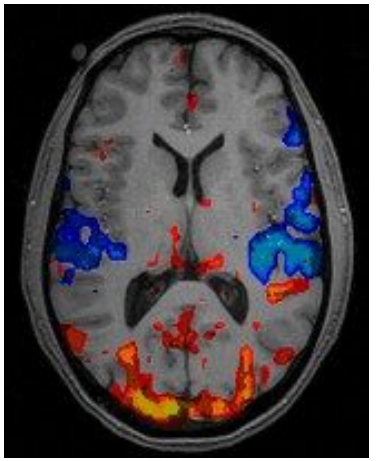
- Typical Big Data problems are based on “automatic” generation of data -> it’s possible to generate millions of data entries
- Medical data usually require some manual activities by a doctor (i.e. to perform a test) -> the number of data entries is relatively low
- What if each data entry is large?
  - Whole genome sequencing
  - Mass-spectrometry (proteomics, lipidomics, metabolomics)
  - **Neuroimaging data**

Neuroimaging Data

Data: various types of Magnetic Resonance Imaging data:

- structural MRI (like 3D image, each voxel contains a number),
- diffusion MRI (each voxel contains a tensor),
- functional MRI (time series of 3D images).





Neuroimaging: sizes of files

HCP Data Sizes (per Subject)		
Session	**Format**	**.zip File Size**
Structural	Unprocessed	71 MB
	Preprocessed	1 GB
Resting State fMRI (each of 2 runs)	Unprocessed	2 GB
	Preprocessed	5.8 GB
	FIX (compact)	3.9 GB
	FIX_extended	4.2 GB
Task fMRI (avg per Task)	Unprocessed	490 MB
	Preprocessed	1.9 GB
	Analyzed	400 MB
All 7 Tasks	Unprocessed	3.4 GB
	Preprocessed	13.1 GB
	Analyzed	2.8 GB
Diffusion	Unprocessed	2.6 GB
	Preprocessed	1.2 GB

http://www.humanconnectome.org/documentation/S900/HCP_S900_Release_Reference_Manual.pdf

Neuroimaging: how many scans?

- **Public databases**:
  - Alzheimer's Disease Neuroimaging Initiative (approx. 1000 subjects)
  - Parkinson's Progression Markers Initiative (600 subjects)
  - Autism Brain Imaging Data Exchange (approx. 1100 subjects)
  - etc
- **ENIGMA - the leading international collaboration:**
  - More than 50 000 subjects, ~20 different pathologies
  - In the basic model, data isn’t shared between sites

Poldrack, Russell A., and Krzysztof J. Gorgolewski. "Making big data open: data sharing in neuroimaging." *Nature neuroscience* 17.11 (2014): 1510-1517.

Thompson, P. M., Andreassen, O. A., ... & Cohen, R. A. (2015). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage*.

Big Data in clinical Neuroscience

- **Volume**

  The typical size of public databases is several TB (up to 64TB for HCP). For large collaborations like Enigma total size of the dataset can be up to 1PB

- **Velocity**

  The only way to increase velocity is to aggregate data from many sources, like Enigma does.

- **Variety**

  At least, we can use all modalities of MRI data & basic info from clinical tests. At most, we can integrate all data available (i.e. genetics, metabolomics, etc)

Why do we need big data sets?

Two key factors drove recent breakthroughs in data science

- ➔ Continuous development of machine learning algorithms
- ➔ Unique data sets were collected and become available

An example: Deep Learning

Deep Learning: ImageNet

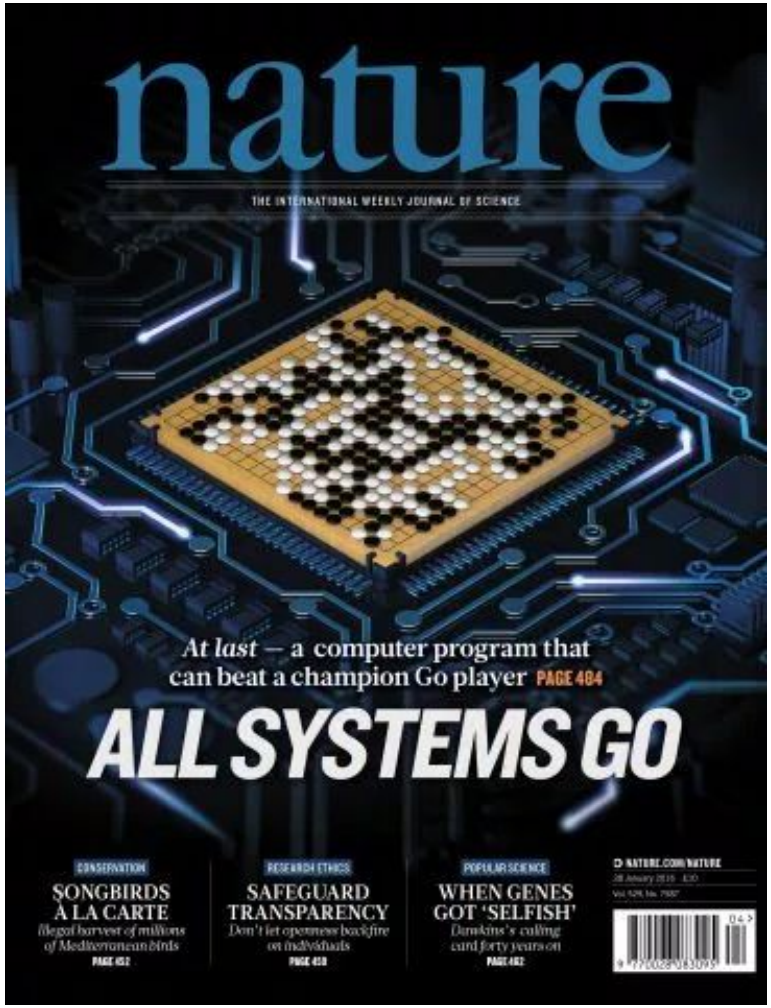
- ImageNet 1k dataset: 1000 of classes, 1.2 millions of images

- Random guessing error is approx. 99%, human error is 5%
- Current DL error is ~3%

Deep Learning is an emerging technology

The game of Go has long been viewed as the most challenging of classic games for artificial intelligence owing to its enormous search space and the difficulty of evaluating board positions and moves.

AlphaGo, a computer system, achieved a 99.8% winning rate against other Go programs, and defeated the human Go champion

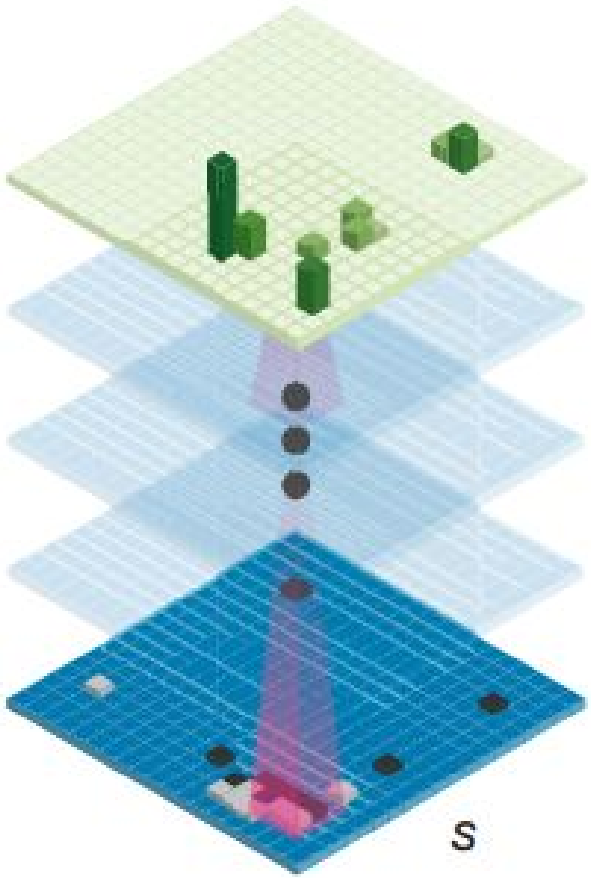


Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." Nature 529.7587 (2016): 484-489.

AlphaGO - is it a game changer?

They considered the board position as a 19×19 image

1. Predict the next move:
   a. Training data base: **30 millions positions** from human games
   b. Play with the previous version of the algorithm to **generate new positions**
2. Build DL networks to predict
   a. the next turn
   b. game outcome
3. Use Monte Carlo tree search to find the best possible turn



Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." Nature 529.7587 (2016): 484-489.

Deep Learning for Neuroimaging

Key research directions:

- ➔ Early prediction of neurodegenerative disorders *(Alzheimer's disease, Parkinson disease, Huntington disease, etc.)*
- ➔ Predict disease progression and treatment outcome *(neurodegenerative disorders, stroke, etc.)*
- ➔ Segmentation of brain metastases & outcome prediction

DTI & fMRI data isn’t used due to smaller datasets available and more complex nature of the data

Deep Learning for Neuroimaging

Deep learning for structural MRI: the core idea is that sMRI is just a 3D image

Typical 2D image classification problem

- ➔ approx. 1000 images per class
- ➔ approx. image size is 100 x 100

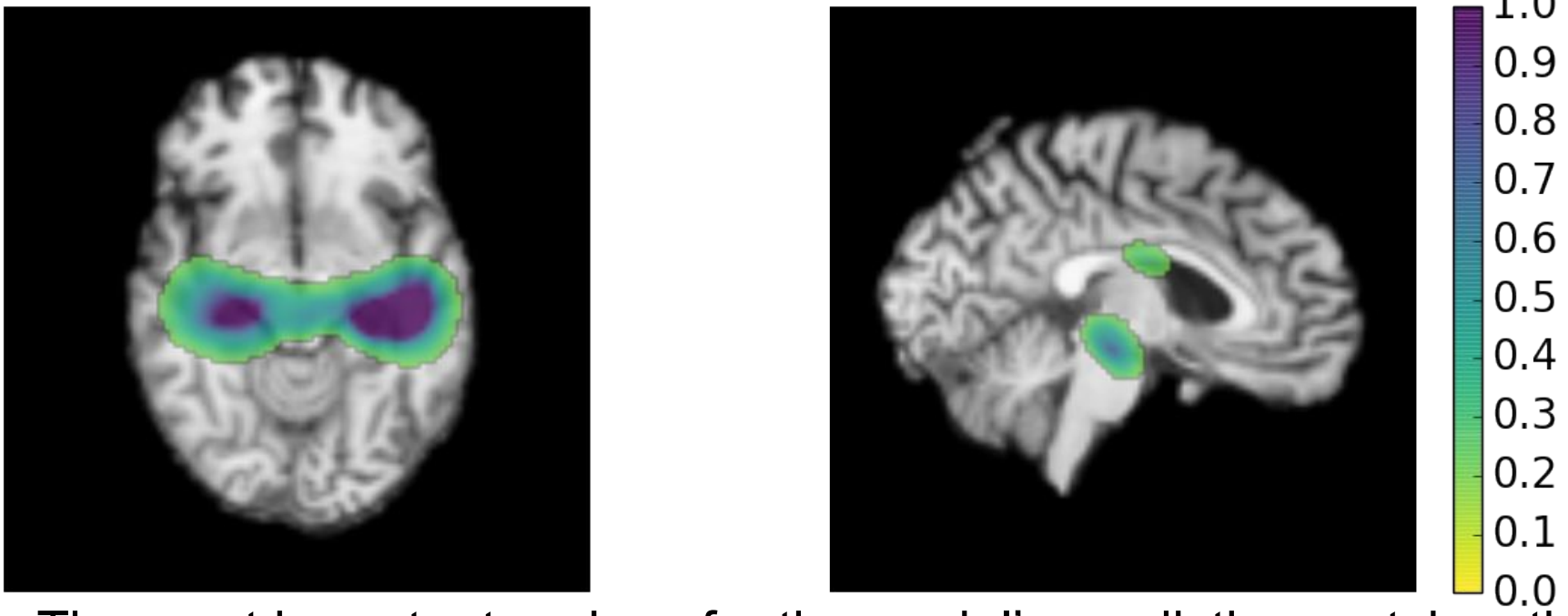
Typical 3D image classification problem

- ➔ approx. few hundreds of images per class
- ➔ approx. image size is 100 x 100 x100

Is it possible to adopt Deep Learning techniques to Neuroimaging data sets in such conditions?

Deep Learning for Neuroimaging: examples

A predictive model for the Alzheimer disease vs Normal Control classification problem and achieved 0.88 ± 0.08 area under ROC



The most important regions for the model's prediction matches the regions that are affected by Alzheimer's disease according to research on the molecular level.

Residual and Plain Convolutional Neural Networks for 3D Brain MRI Classification / S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova. Submitted to IEEE International Symposium on Biomedical Imaging 2017.

Skoltech
Skolkovo Institute of Science and Technology

Summary

To unlock potential of Deep Learning and Big Data technologies, we need a joint effort of Medical, Computer Science, Bioinformatics & IT specialists to

- A large & reliable set of clinical neuroscience related data
- Collaborate to build & evaluate modern data analysis algorithms for neuroscience
- Finally, push boundaries of neuroscience by joint research projects.

CoBrain project is recently supported initiative which addresses this challenge

CoBrain project

CoBrain is a platform for collecting and analysis of clinical neuroscience data.

- **Partners**: leading russian centers, including
  - Scientific Center of Neurology
  - Scientific Center for Child's Health
  - Scientific Research Neurosurgery Institute
  - And more
- **Algorithms**: specially developed algorithms for analysis of multimodal data (genomic data, lipidomics, metabolomics, neuroimaging**)**
- **Hardware**: 2 PB of storage & heavy computational power

Thank you!

Mikhail Belyaev

m.belyaev@skoltech.ru