MNC3 "Databases"

marco milanesio
UCA - MSI - INRIA
1-12-2017

summary

- Introduction
- Databases
  - UK Biobank
  - ADNI
  - MAPT
- On going
  - INSIGHT
- Next steps

introduction

- Overview of the DBs actually @INRIA
- **Heterogeneity**
  - Study
  - Cohort
  - Type
  - Size
- Content
  - Images
  - Genetics
  - Clinical data
  - Questionnaires

introduction

- Collections of objects
  - CSV tables
  - XML summaries
- Mostly unstructured
- No common entry point
  - web access to download / query (some APIs)
  - text-based retrieval
    - user-id [+date] [+exam-id] […] as file name
- Template: what it is / what we have

uk biobank

- Prospective cohort of > 500K subjects (40-69)
  - Blood, urine and saliva samples
  - Physical measurements
  - Questionnaire on health and lifestyle
  - Genetic data
  - Heart and body images
  - Brain images
    - T1, rfMRI, tfMRI, T2_Flair, dMRI, SWI

brain images

- T1-weighted
  - Contrast grey/white matter
  - Interaction of water to surrounding tissues
- rfMRI: resting-state functional MRI time series
  - Changes in blood oxygenation
- tfMRI: task functional MRI time series
  - Same as rfMRI but while performing tasks
- T2_Flair
  - Contrast dominated by signal decay from interactions between water molecules
  - Alterations to tissue compartments associated to pathologies
- dMRI: diffusion MRI
  - Ability of water molecules to move within their local environment
  - Integrity of tissues (local), structural connectivity (tractography)
- SWI: susceptibility-weighted imaging
  - Magnetized tissue constituents
  - Venous vasculature, microbleeds, microstructure

downloaded data

- Questionnaire
  - 21 Gb
  - 10581 attributes
  - 1978 unique
  - 3 time points
- Data dictionary
- Images
  - T1 / T2_Flair / fMRI
- Genetics

data dictionary

ValueType	count
Categorical single	920
Compound	9
Text	196
Integer	516
Time	47
Continuous	1408
Categorical multiple	111
Date	23

ValueType	Instances	Array	count
Continuous	1	1	1212
Categorical single	3	1	285
Categorical single	1	1	273
Categorical single	5	1	260
Integer	1	1	214

only showing top 5 rows

ValueType	Instances	Array	count
Categorical multiple	1	30	1
Categorical single	3	14	1
Integer	1	5	1
Categorical multiple	1	11	1
Categorical single	3	32	1

only showing top 5 rows

data dictionary

ValueType	Instances	Array	count
Categorical single	30	1	1
Categorical single	32	1	4
Categorical single	4	1	1
Categorical single	5	1	260

ValueType	Instances	Array	count
Categorical single	1	23	1
Categorical single	3	29	1
Categorical single	3	32	1
Categorical single	1	33	4
Categorical single	1	35	1
Categorical single	1	40	15

data dictionary

ValueType	Instances	Array	count
Categorical single	30	1	1
Categorical single	32	1	4
Categorical single	3	32	1
Categorical single	1	33	4
Categorical single	1	35	1
Categorical single	1	40	15

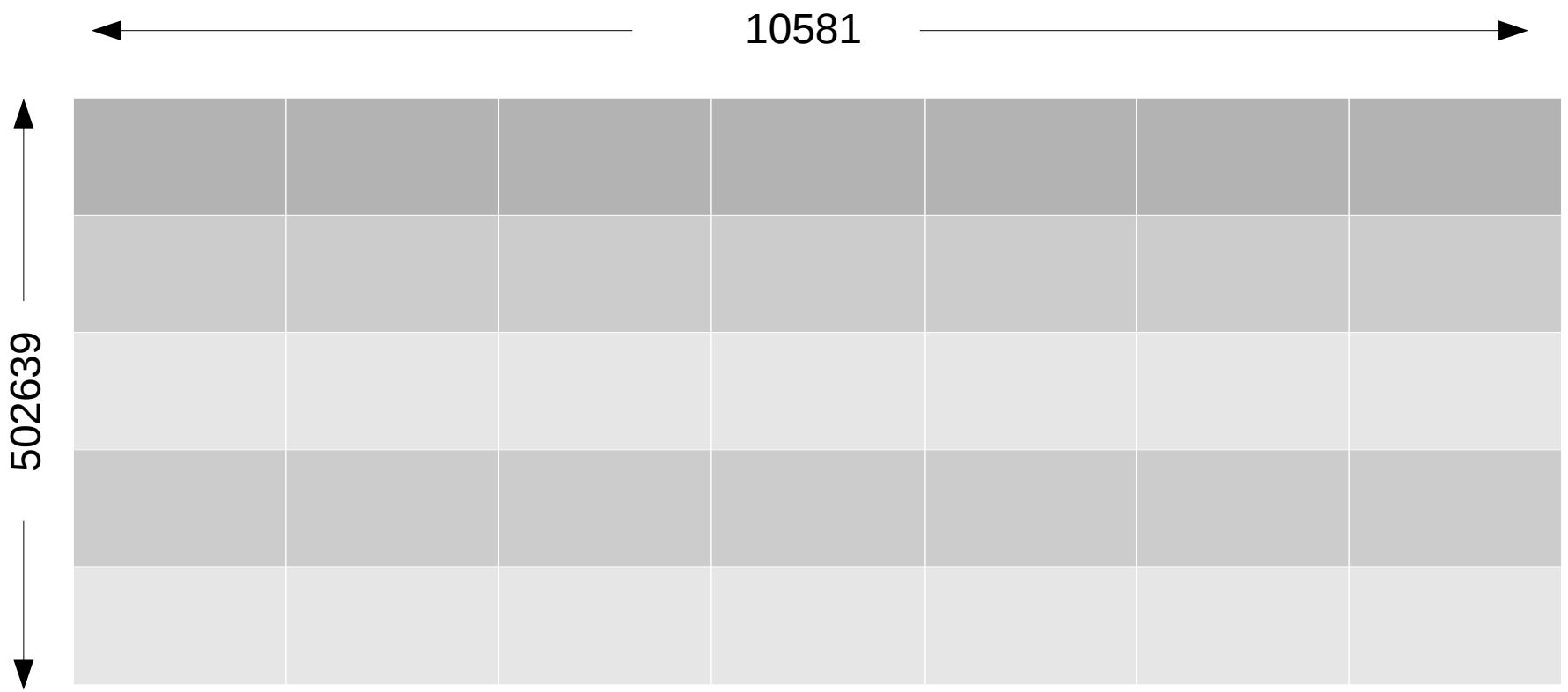
FieldID	Field	Instances
40013	Type of cancer: ICD9	30
40006	Type of cancer: ICD10	32
40012	Behaviour of cancer tumour	32
40011	Histology of cancer tumour	32
40019	Source of cancer report	32

data dictionary

FieldID	Field
22601	Job coding
22604	Work hours - lumped category
22606	Workplace very noisy
22607	Workplace very cold
22608	Workplace very hot
22609	Workplace very dusty
22610	Workplace full of chemical or other fumes
22611	Workplace had a lot of cigarette smoke from other people smoking
22612	Worked with materials containing asbestos
22613	Worked with paints, thinners or glues
22614	Worked with pesticides
22615	Workplace had a lot of diesel exhaust
22616	Breathing problems during period of job
22617	Job SOC coding
22620	Job involved shift work

Categorical single		1	40	15

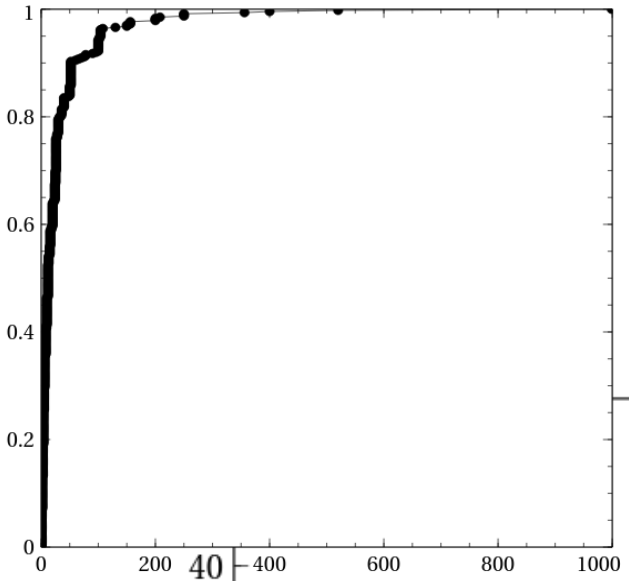
questionnaire



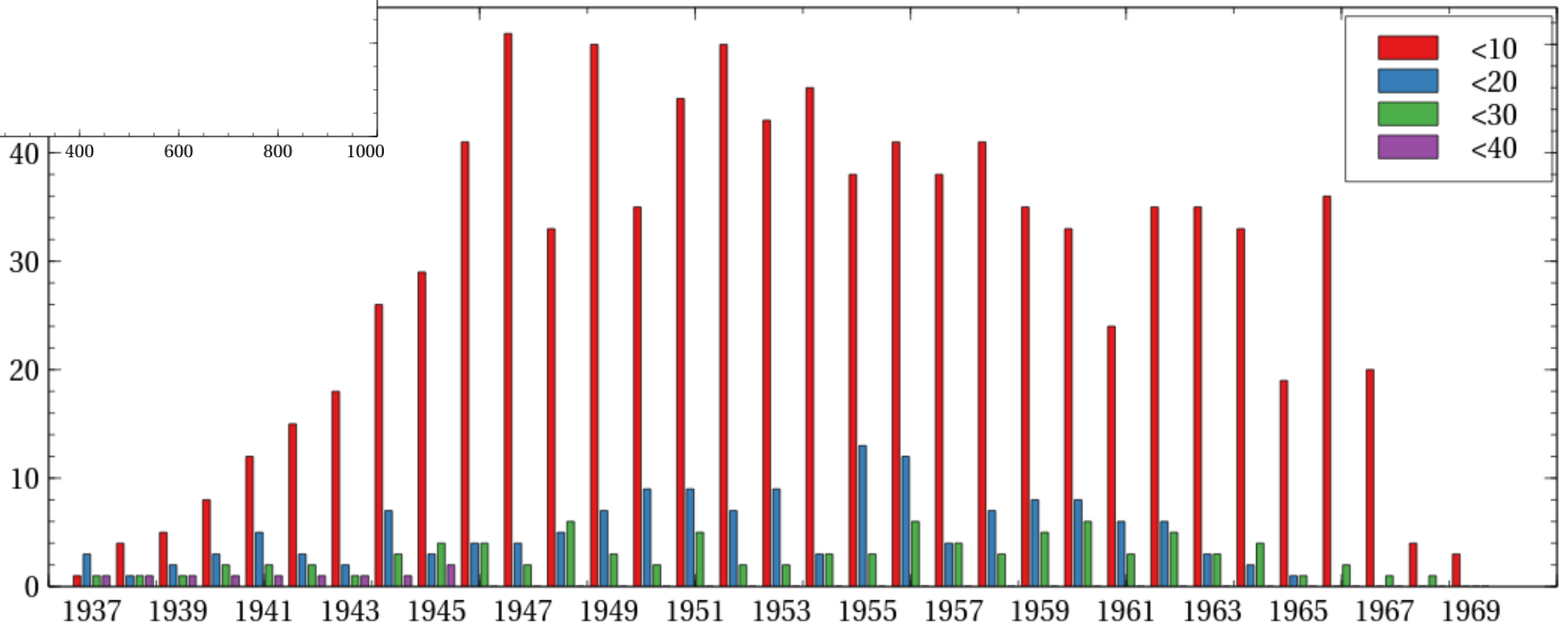
1 row = 1 subject

1 column = 1 attribute/value

attribute-based investigation



e.g., Longest period of unenthusiasm / disinterest (5375)



some numbers

- unique attributes
  - in header = 1977 (+ eid)
  - in dd = 3230
  - (in dd & not in header) = 1306
  - (in header & not in dd) = 53
    - genotyping intensities
    - 22014 ... 22352

downloaded images

- T1
  - 5544 subjects
  - 92 GB (compressed)
    - ~ 36 MB per subject
- T2_Flair
  - 5638 subjects
  - 176 GB (compressed)
- fMRI
  - 5767 subjects
  - 2.1 TB (uncompressed)

genetics

- Genotype calls (based on genotyping array measurements) and related measurements
  - Calls (0.1TB)
  - Confidences (2.9TB)
  - Intensities (2.9TB)
  - CNV B-allele frequencies (1.5TB)
  - CNV log2ratios (2.3TB)
- All proprietary binary formats
- http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100315

downloaded genetics

- 152728 subjects
- 2.6 Tb
- Chromosomes 1-22
  - .bgen format
    - www.well.ox.ac.uk/~gav/bgen_format/bgen_format.html
  - single nucleotide polymorphisms (SNP)
  - 1.2 Tb
- Chromosomes 1-22, X, Y, MT
  - .cal .con .int
  - 1.4 Tb
  - information not found

ADNI

- Alzheimer's Disease Neuroimaging Initiative
- MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers

CLINICAL DATA	GENETICS DATA
◦ Demographics ◦ Clinical Assessments ◦ Cognitive Assessments	◦ Illumina SNP genotyping
MR IMAGE DATA	**PET IMAGE DATA**
◦ Raw, pre- and post- processed image files ◦ fMRI (ADNI GO/ADNI2) ◦ DTI (ADNI GO/ADNI2)	◦ Raw, pre- and post- processed image files ◦ PIB (ADNI1) ◦ FDG (ADNI1/GO/2) ◦ Florbetapir (ADNI GO/2)
IMAGE ANALYSIS RESULTS	**CHEMICAL BIOMARKER**
◦ Numeric results derived from image analyses ◦ MRI Analysis ◦ PET Analysis	◦ Laboratory Results ◦ Proteomic Analysis

ADNI

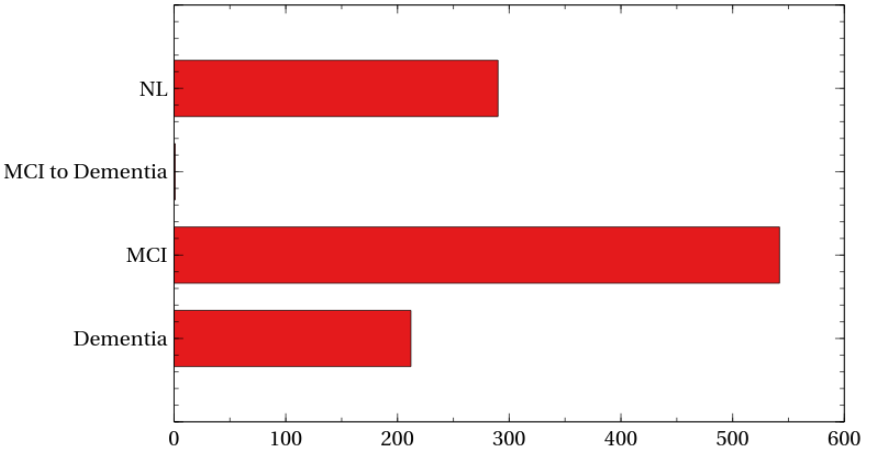
CN	**Normal Aging /Cognitively Normal**	ADNI 1/GO/2	CN participants are the control subjects in the ADNI study. They show no signs of depression, mild cognitive impairment or dementia.
SMC	**Significant Memory Concern**	ADNI 2	SMC participants score within normal range for cognition (or CDR = 0) but indicate that they have a concern, and exhibit slight forgetfulness. The informant does not equate this as progressive memory impairment nor considers this as consistent forgetfulness.
EMCI	**Early Mild Cognitive Impairment**	ADNI GO/2	MCI participants have reported a subjective memory concern either autonomously or via an informant or clinician. However, there are no significant levels of impairment in other cognitive domains, essentially preserved activities of daily living and there are no signs of dementia. Levels of MCI (early or late) are determined using the Wechsler Memory Scale Logical Memory II.
MCI	**Mild Cognitive Impairment**	ADNI 1	
LMCI	**Impairment**	ADNI GO/2	
AD	**Alzheimer's disease**	ADNI 1/GO/2	AD participants have been evaluated and meet the NINCDS/ADRDA criteria for probable AD.

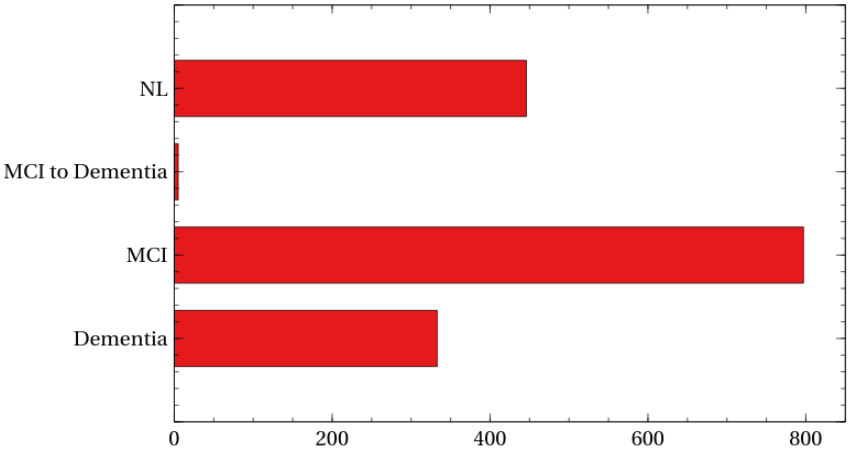
cohorts

- ADNI1
  - 800 (50-95) subjects (200 CN, 400 MCI, 200 AD)
- ADNI-GO
  - 200 (55-90) subjects (mildest symptomatic AD, early amnestic MCI (LMCI)
  - 500 subjects (LMCI and CN) from ADNI1
- ADNI2
  - 550 subjects (150 CN, 100 EMCI, 150 LMCI, 150 mild AD)
  - 500 subjects (CN and LMCI) from ADNI1
  - 200 subjects (EMCI) from ADNI-GO
- ADNI3 (starting june 2017)
  - 1070-2000 subjects in 3 cohorts (CN, MCI, mild AD)
    - 700-800 subjects from ADNI2
    - 370-1200 new subjects (55-90)

downloaded data

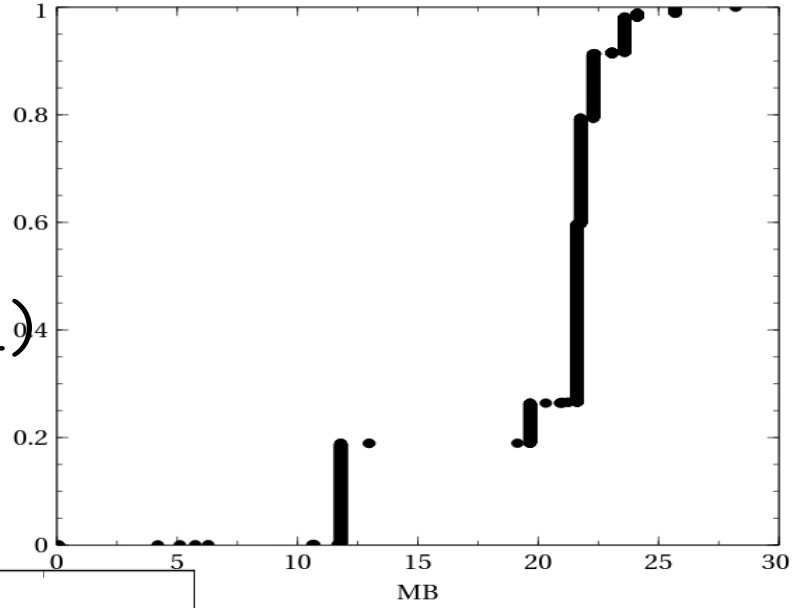
- Parts downloaded for different on-going projects
  - MRI
    - T1-weighted
  - PET
    - AV45 / FDG / MRI

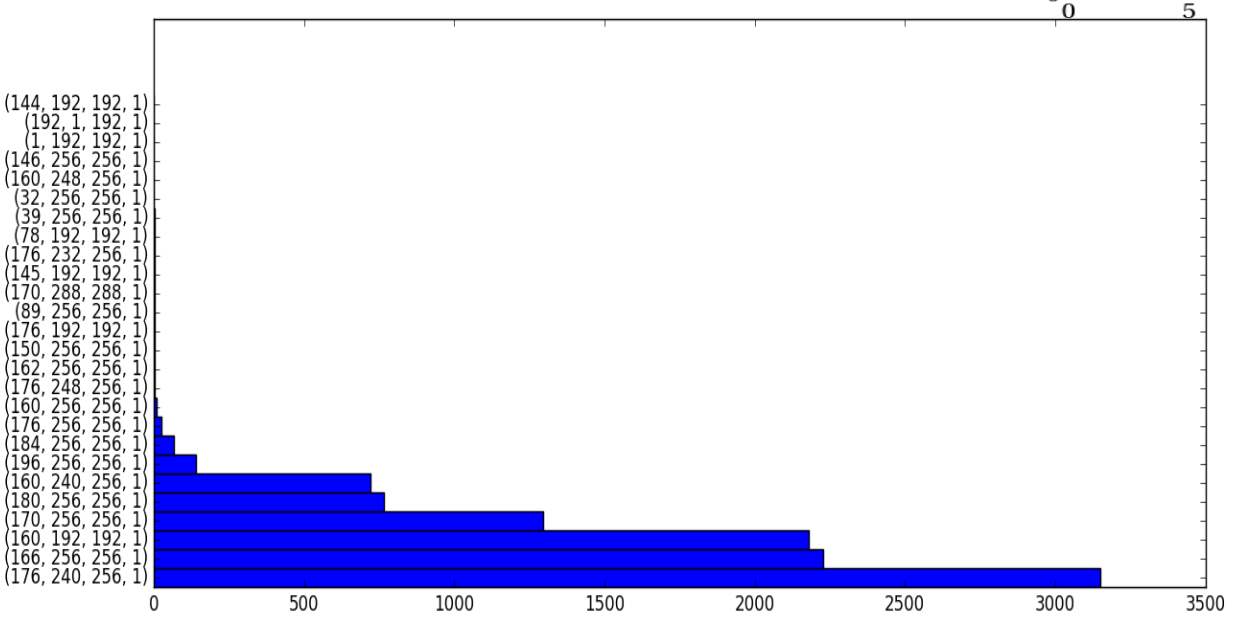




MRI - MPRAGE

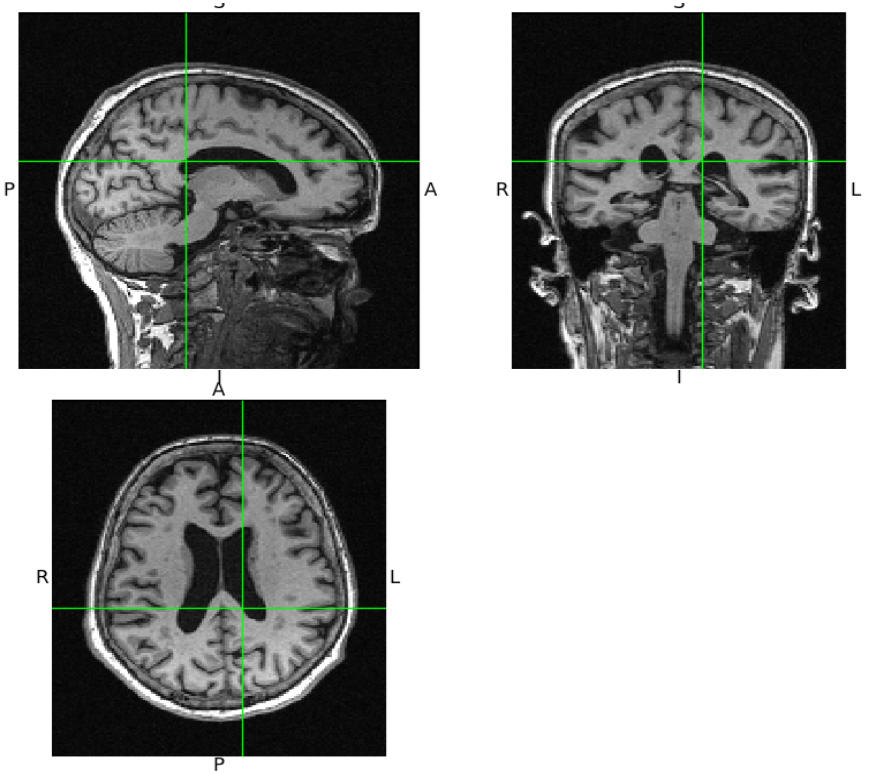
- 13975 xml files
  - 6505 summary files
- 6519 nii images (uncompressed)
  - ~ 130 Gb





MRI - MPRAGE

- 1301 subjects
- Range of time points:
  - '018_S_0450' → T0,T6,T12,T18,T24,T36
  - '128_S_0258' → T0




```
<metadata version="1.0">
    <subject id="128_S_0258"/>
    <study uid="5571"/>
    <series uid="S20932"/>
    <image uid="I27377"/>
</metadata>
```

xml summary files

- For each image
  - subjectIdentifier
  - researchGroup
  - subjectSex
  - visit
    - assessment name
    - assessment score
  - study
    - subjectAge
    - weigthKg
  - series
    - modality
    - imagingProtocol
    - protocolTerm
      - AcquisitionType
      - Pulse Sequence
      - Coil
      - Matrix X, Y, Z
      - Pixel spacing
      - ...

replication

- Lots of replication
  - And renames (!!!)
- On a total of 10622 .nii files:
  - 4110 conflicts (39%)
    - 3 times (14)
    - 2 times (4096)
- Inconsistent storage patterns
  - $ ~/ADNI
  - $ ~/ADNI/ADNI
  - $ ~/ADNI_201601

MPRAGE	4066
MP-RAGE	2121
MPRAGE_GRAPPA2	1006
Sag_IR-SPGR	50
Sag_IR-FSPGR	16
ASO-MPRAGE	14
MPRAGE_3dtfe	14
MPRAGE_Repeat	14
MP-RAGE_REPEAT	13
MPRAGE_SENSE	12
MPRAGE_SENS	11
MP_RAGE	9
MPRAGE_AUTOSHIM_ON	9
SAG_MP-RAGE	9
IR-SPGR	7
MPRAGE_ASO	6
MPRAGE_3dtf	5
MPRAGE_NO_ANGLE	5
SAG_3D_MPRAGE	4
mprage	3
ASO-MPRAGE_2	1
Localizer	1
MPRAGE_2ND	1
MPRAGEadni	1
MPRAGEASO	1
MP-RAGE-Repeat	1
MP-RAGE__SERIES_2_	1

PET

- ADNI 2:
  - Florbetapir (AV-45) PET and FDG PET imaging were performed on all newly enrolled participants on two separate days (minimum 12hr time lapse).
  - Scans were performed within two weeks before or two weeks after the in-clinic assessments at Baseline and at 24 months after Baseline.
  - ADNI 2 subjects had up to 3 florbetapir scans and up to 2 FDG scans, each acquired at 2 year time intervals.

downloaded PET

- 431 subjects, .nii, uncompressed
  - AV45
    - 1-4 images per subject
    - 20 GB
  - FDG
    - 1-2 images per subject
    - 42 GB
  - MRI
    - 1-5 images per subject
    - 0.9 TB
  - xml summaries (as before)

MAPT

MAPT

- Multidomain Alzheimer Preventive Trial
- Efficacy of
  - isolated supplementation with omega-3 fatty acid
  - isolated multidomain intervention (nutritional counseling, physical exercise, cognitive stimulation)
- 1680 subjects aged 70+
- M6, M12, M24, M36 checkpoints
- Impact of interventions on
  - celebral metabolism (FDG PET)
  - atrophy rate (MRI)
  - brain amyloïd deposit (AV45 PET)
- Protocol: https://www.clinicaltrials.gov/ct2/show/NCT00672685?term=NCT00672685&rank=1

MAPT

downloaded data

- 506 subjects
  - 379 (T0 T36)
  - 127 (T0)
  - T1-weighted
  - clinical data (csv)
  - ~ 8 Gb

next steps

- Insight
  - project P50: High dimensional statistical modeling of the relationship between heterogeneous information: behavior, actigraphy, cognition, and brain imaging
- no estimation (so far)

conclusions

- High heterogeneity within/between datasets
- Global merge hardly possible
  - > 7.3 Tb (+ replication) (+ temp results)
- Possible front-end to ease the query process
  - e.g. all individuals between 60 and 65 with MRI
- Current versions are incomplete
  - Do we need all datasets to be complete?
  - Store and manage only relevant parts?