Biostat 202: Opportunities and Challenges of "Big Data": Case Study - classifying dementia based on volumetric MRI

Kirsten Rogers[1], Yann Cobigo[1], and John Kornak[2]

[1]UCSF Memory and Aging Center

[2]Department of Epidemiology and Biostatistics

09/01/2016

Other Important People

Kate Rankin (MAC)

Howie Rosen (MAC)

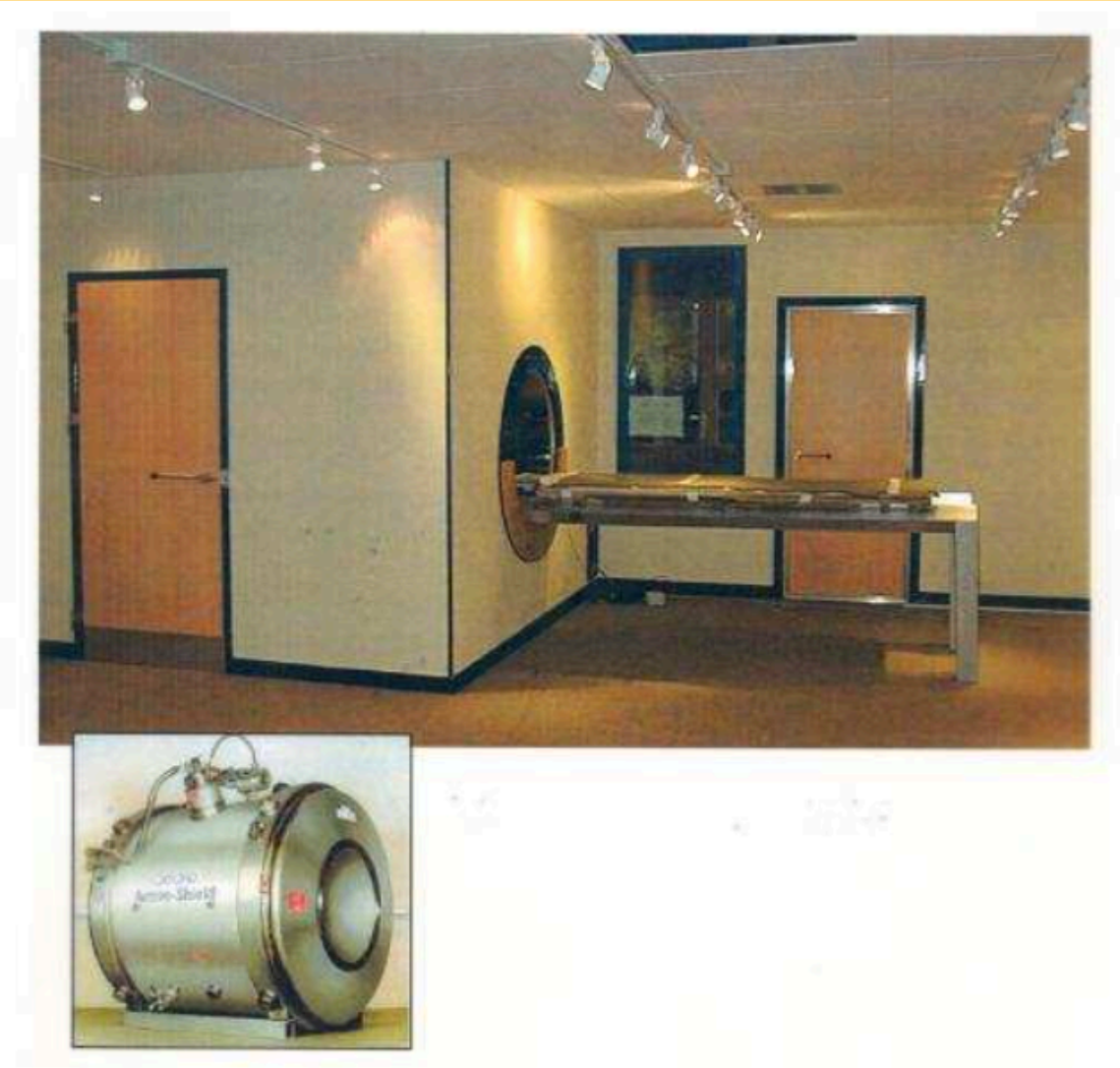
Chris Hess (Radiology)

And many others…

Outline

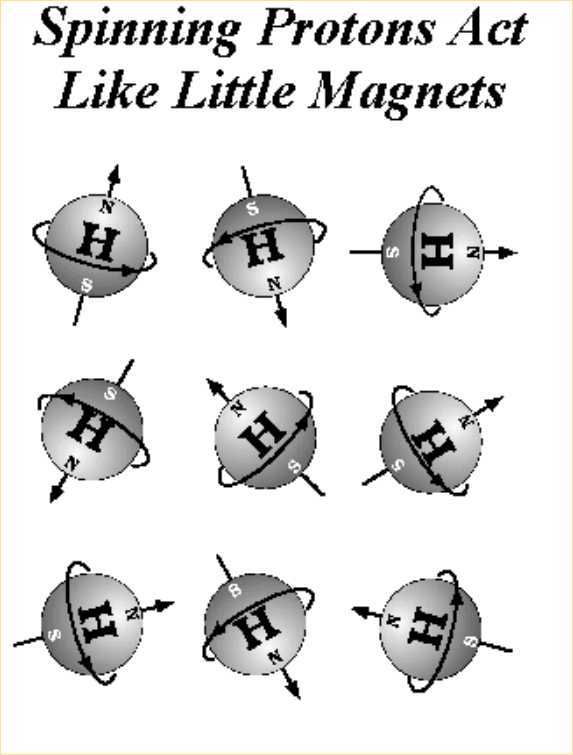
- Quick hand-waving introduction to Magnetic Resonance Imaging (MRI)
- Introduction to the Quest project (Kirsten)
- Pre-processing (Yann)
- Memory and Aging Clinic (MAC) dataset
- ROI approach – IBM Modeler example
- “Voxel-based” analysis (Yann)
- Future plans – validation, validation and more validation
- Summary

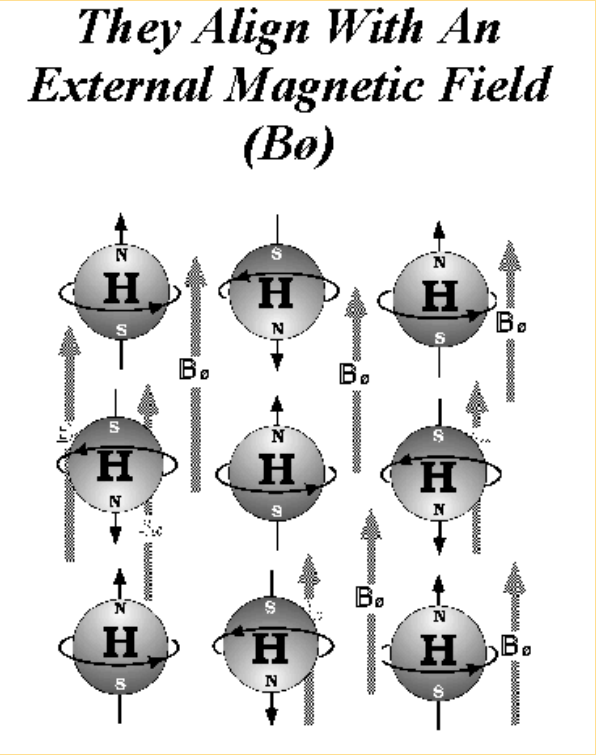
MRI scanner (old school) – University of Nottingham 3T

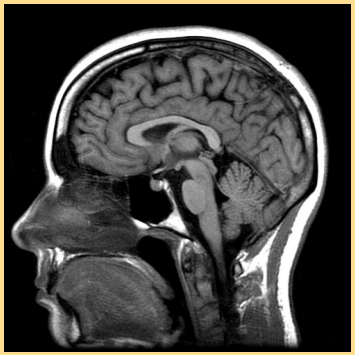


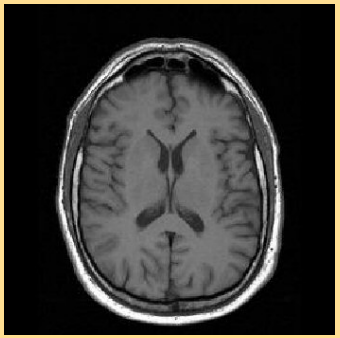
MRI Scanner – Mission Bay

Protons in a magnetic field and MRI









- Different tissues have different proportions of “free hydrogen” that align with the magnetic field and rotate like spinning tops (precess) together at a specific frequency – therefore different tissue types have differing signal intensities that can be detected and differntiated using a radio-frequency coil.
- Magnetic field gradients are applied in the x and y directions to perturb the signal to specify spatial location within a 2D slice – leads to an image – multiple slices 3D

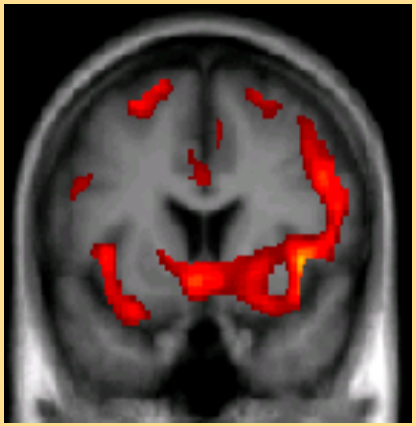
Why automate MRI diagnosis of dementia?

- Different dementias have different patterns of brain loss

AD

svPPA

bvFTD



- Current standard: neurologists and radiologists review images and generate a diagnosis from images, neuropsyche testing, and family interviews etc.
- Group patterns clear, but individual scans more difficult to interpret/classify
- Clinicians/radiologists often disagree on image-based diagnosis
- Imaging based classifier will provide useful tool to clinician – especially in smaller clinics with no expert neurologist – Kirsten will say more

The Quest project – Kirsten

An Integrated Diagnostic & Management Care Pathway for Dementia

September 1, 2016

Quest Diagnostics™

UCSF Medical Center

Memory & Aging Center

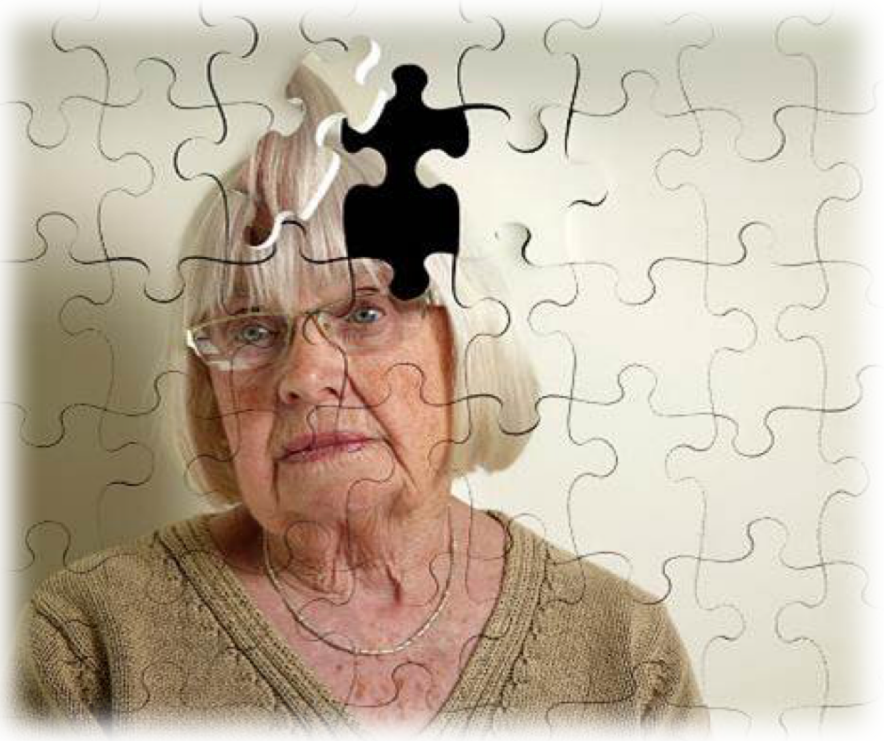
The Growing Dementia Epidemic

- ~15% of individuals >age 71 in the US have dementia, and the number of people > age 65 is expected to double to 70 million by 2030
- Most individuals with neurocognitive symptoms rely on primary care providers (PCPs) to perform a screening evaluation, and often to provide ongoing care
- Many PCPs report they don't have the appropriate resources or training to ensure timely and accurate diagnosis of their dementia patients

Late or Inaccurate Dementia Diagnosis is a Major Problem

No diagnosis or the wrong diagnosis leads to:

- Wrong tests ordered, wrong treatment given
- More medical costs as the correct diagnosis is sought
- Missed opportunity for curative treatments
- Burden on caregivers who need answers and resources
- Patients lose opportunity to guide their own life planning
- Doctors lose confidence and self-efficacy treating dementia



Quest Diagnostics™

Dementia is the most expensive healthcare problem

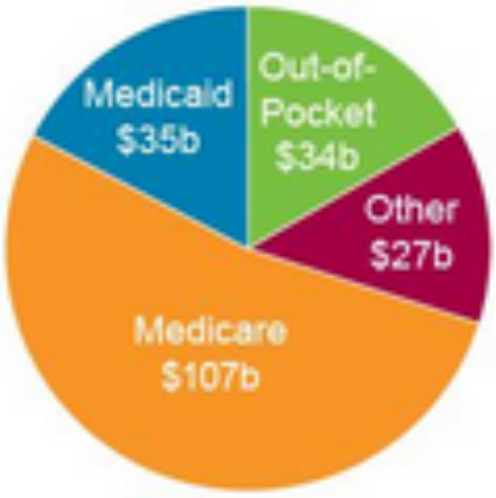
- The direct U.S. medical costs of dementia care are > $200 billion / year (NEJM article 2014)
- Greater than the costs of heart disease or cancer
- $1 out of every $5 Medicare dollars is spent on people with Alzheimer‘s and other dementias

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

Monetary Costs of Dementia in the United States

Michael D. Hurd, Ph.D., Paco Martorell, Ph.D., Adeline Delavande, Ph.D., Kathleen J. Mullen, Ph.D., and Kenneth M. Langa, M.D., Ph.D.



Quest Diagnostics™

What is Dementia?

Dementia is a clinical syndrome defined as cognitive or behavioral decline that leads to an inability to complete daily tasks independently.

Dementia has many causes some of which are reversible, such as metabolic disorders, while some are progressive such as Alzheimer's Disease (AD)

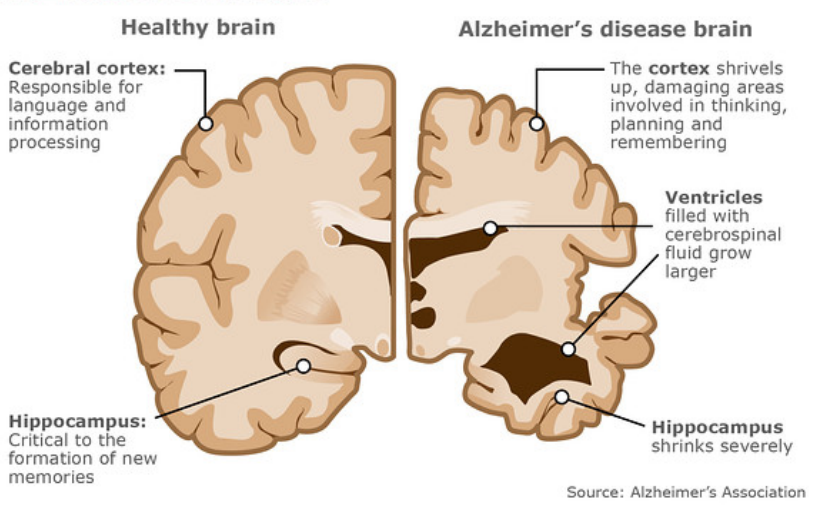
Symptoms of dementia may include:

- Memory loss
- Language impairment
- Loss of social function
- Parkinsonism
- Behavioral or personality changes
- Visuospatial deficits

Types of dementia

Alzheimer's Disease

- Alzheimer's is the most common form of dementia, accounting for between 60-80 percent of cases.
- The cause of Alzheimer's Disease is unknown, but is associated with the accumulation of amyloid plaques and tau neurofibrillary tangles
- The majority of patients with AD present with mild memory problems that gradually interfere with the patient's ability to manage daily tasks.
- AD is prevalent in older adults, but can also affect people in their 40s and 50s
- The progression usually occurs slowly over decades
  - May develop other symptoms such as difficulty recognizing familiar people, dysphagia, or gait disturbances.
- No cure yet, but there are medications that help manage the symptoms



Quest Diagnostics™

Frontotemporal Dementia

- Condition that slowly damages the front or temporal part of the brain
  - Behavioral variant frontotemporal dementia (bvFTD) is one subtype of that condition
- Characterized by alterations in social decorum and personal regulation
  - May affect how a person deals with everyday situations as well as cognitive skills
- One of the most common causes of dementia in people less than 65 years old
  - Cause is unknown, scientist know that bvFTD is the result of a large build-up of one of three kinds of proteins (Tau, TDP-43, Fused-in sarcoma (FUS))
- Most people with bvFTD start having symptoms in their 50s, although some people have shown signs earlier

Primary Progressive Aphasia

- Primary progressive aphasia (PPA) is an isolated language deficit with an insidious onset and gradual progressive impairment of language production, object naming, syntax, or word comprehension.
- Aphasia is the most prominent deficit at symptom onset and with disease progression, although other cognitive deficits may be affected later in the disease.
- Activities of daily living are unaffected except those related to language.
- Variants
  - Non-fluent Variant (nfvPPA) – people have increasing trouble getting words out
  - Logopenic Variant (lvPPA) – people have increasing trouble thinking of the words they want to say
  - Semantic Variant (svPPA) – people have increasing difficulty understanding the meaning of words, finding words, or naming people and objects.
- Cause is unknown

Integrated Solutions

The Dementia Care Pathway

- Step by step dementia diagnosis and care guidelines through interactive insights
- Tablet-based cognitive testing with instant interpretive reports
- Guidance for expanded lab ordering
- Imaging
  - Streamlining ordering of dementia-specific MRI
  - Automated quantification of structural brain abnormalities
  - Parameters for dementia specific MRI interpretation
- Diagnosis-specific educational modules to guide treatment and patient care

Summary

- The dementia pathway aims to:
  - Shorten the amount of time to correctly diagnosing a dementia patient
  - Better identify patients with treatable dementias
  - Enable better patient and physician education and treatments consistent with best practices
  - Lower costs by eliminating redundant or improper testing

Pre-processing – Yann

Brain MRI

3D cube

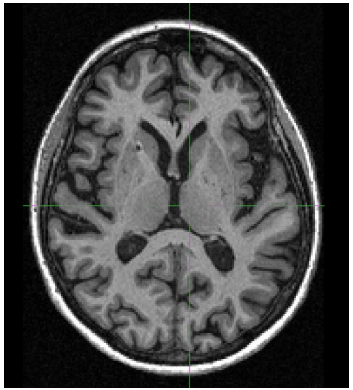
- Vector = 160 x 240 x 256 voxels (volume elements 1x1x1 mm3)
- Several tissues: gray matter, white matter, CSF, muscles, fat, bones, ...

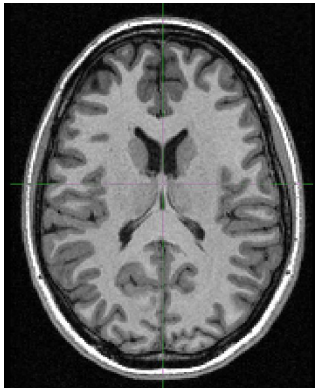
T1 structural image

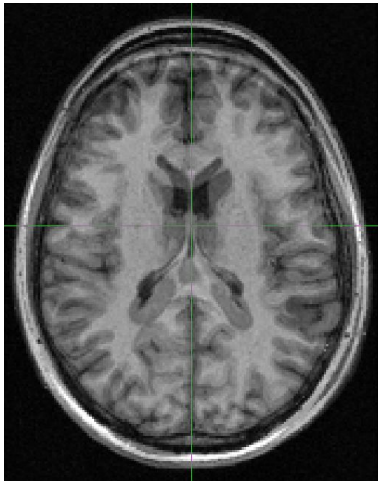
Brain MRI

3D cube

- Vector = 160 x 240 x 256 voxels.
- Several tissues: gray matter, white matter, CSF, muscles, fat, bones, ...
- Complex real world classification challenge.
    - Algorithm treating images in the native space
    - Normalization of the images







Brain Normalization

Brain Normalization

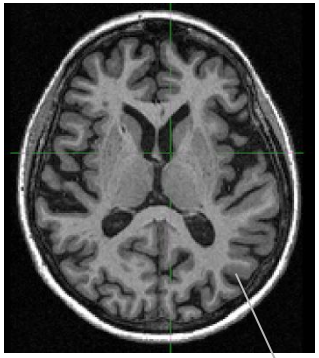
- Transformation of *e.g.* gray matter or white matter into a standard/ideal brain MNI (Montreal Neurological Institute)
    - Increase the likelihood of a voxel to have the same meaning across images
    - Conserve the physical properties of the native space head

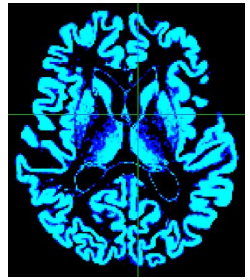
MNI gray matter tissue probability

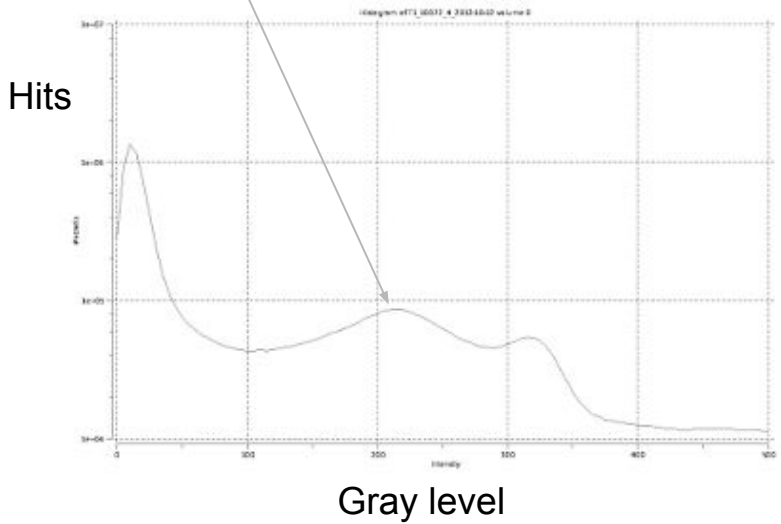
Brain Normalization

Segmentation

- Clusterization of grey scale image
    - Every voxel of a MRI structural image is projected in an histogram (2 dimensions)
    - Each bump is potentially a tissue: gaussian mixture
- Refinements: 3 dimensions information; other contrast (T2, FLAIR, ...)







Brain Normalization

Normalization and modulation

- Linear transformation (global)
  - Rigid body: 3 rotations + 3 translations
  - Affine: Rigid body + 3 scaling + 3 Skews/Shears

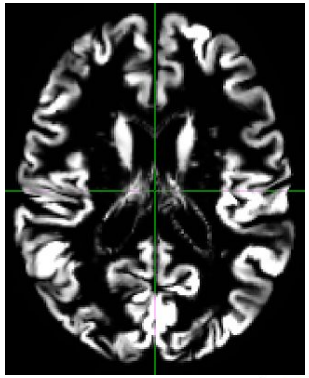
Affine transformation

Brain Normalization

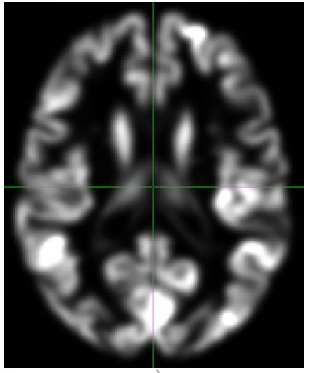
Smoothing

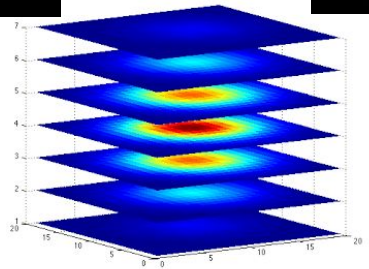
- Gaussian kernel
    - Blurring the image decrease subject features
- Increase the likelihood of a voxel to have the same meaning across images

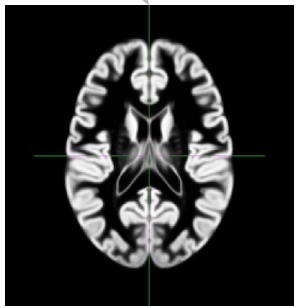
Unsmoothed



Smoothed







MNI

MAC Dataset and IBM modeler – John

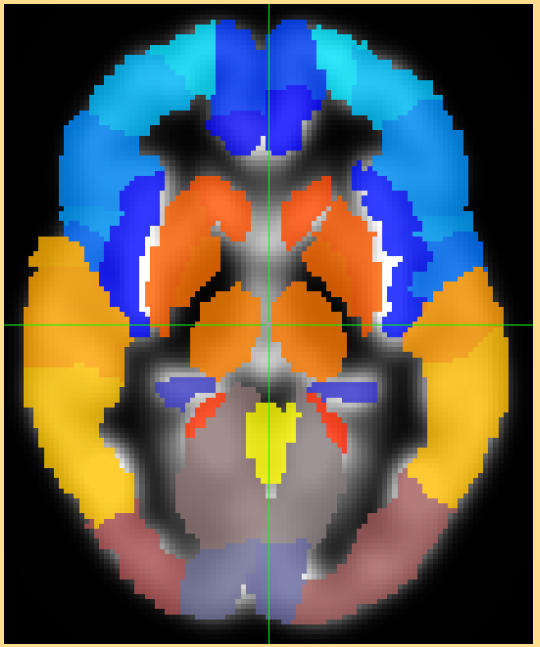
UCSF Memory and Aging Clinic (MAC) Dataset

- Structural (volumetric) MRI – multiple dementia studies – but in same scanner and same protocol
- Single time point
- Gold standard Dx – autopsy (if dead) UCSF clinical expert joint decision (otherwise)

Field	CN	SV	BV	AD	MCI	DLB	PSP	nfPPA
3.0 T	291	44	68	180	68	**0**	44	38

- Objective is to classify patients as having or not having dementia and if yes, what type of dementia?

Data reduction: ROIs

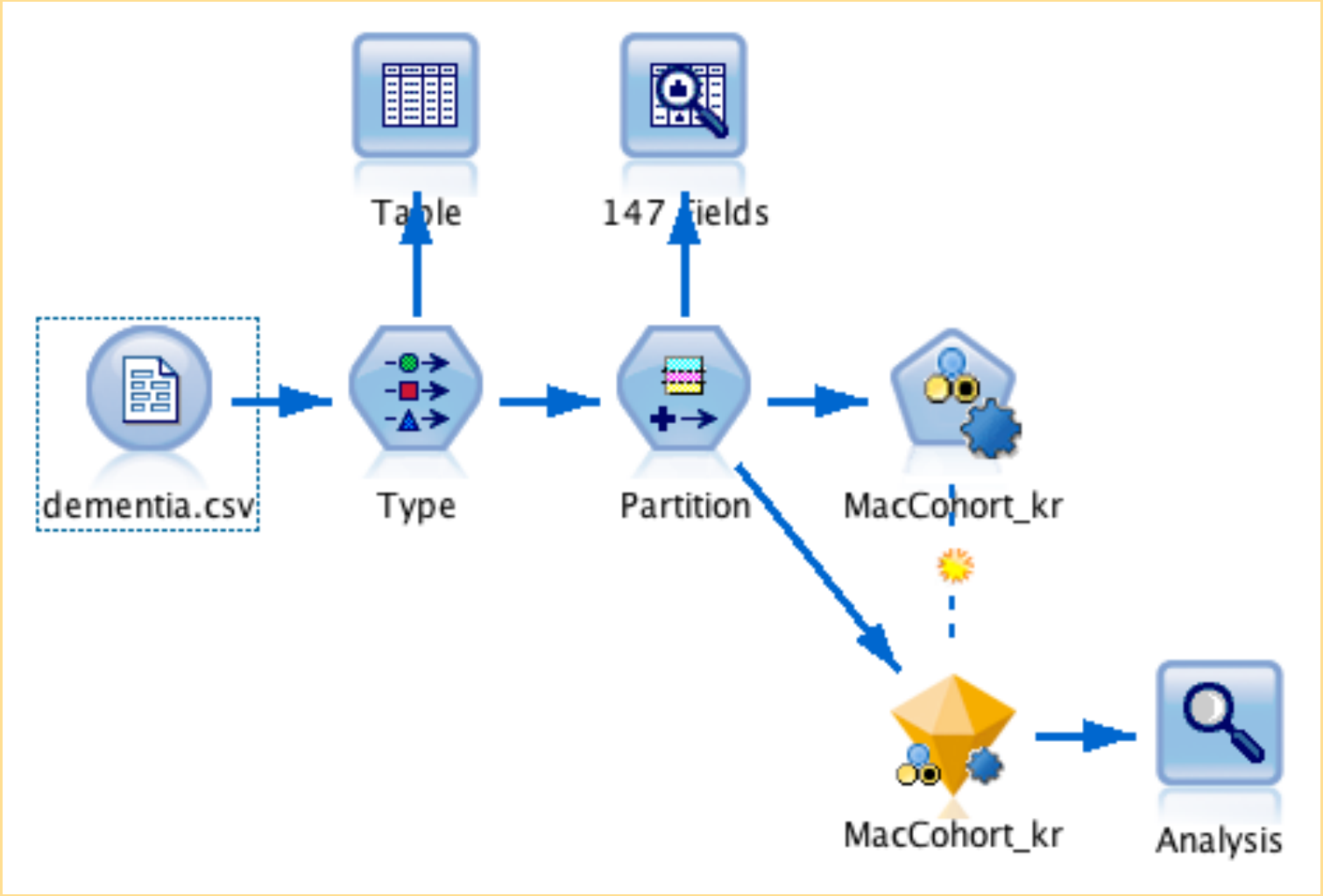


- Regions-of-interest (ROIs) as predictors: Insight Segmentation and Registration Toolkit (ITK) library extraction of ROIs in MNI space – each ROI is a predictor: 136 ROI predictors – big data?
- ROIs are defined in terms of anatomical regions and landmarks in the human brain that are common to all individuals – often have specific functional roles
- Yann will later discuss more complex voxel-based analysis

IBM Modeler analysis

- Classification analysis – predicting dementia existence and type based on voxel-based-morphometry gray matter maps
- Using region of interest (ROI) estimates as predictors
- Training-test-validation approach
- Use auto-classifier to look at many models

IBM Modeler stream



Var. File node

dementia.csv

Preview Refresh

/Users/kornak/Documents/Teaching/DataScienceCourse/LecturesLabsAndHW/My...

File Data Filter Types Annotations

File: /Users/kornak/Documents/Teaching/DataScienceCourse/LecturesLabsAndH ...

```
"","PIDN","DCDate","SourceID","ScanNo","vbmID","gm","wm","csf","tiv","Magne
"1",65,2011-06-23,"ABN00248-1",1,4951,595.483,521.213,284.383,1401.08,3,70,
"800",3576,2009-08-11,"NIC060-1",1,4562,559.454,549.746,292.001,1401.2,3,72
"915",4042,2010-02-08,"NIC096-1",1,4247,687.109,551.16,298.904,1537.17,3,73
```

☑ Read field names from file ☐ Specify number of fields 1

Skip header characters: 0 EOL comment characters:

Strip lead and trail spaces: ◉ None ○ Left ○ Right ○ Both

Invalid characters: ◉ Discard ○ Replace with

Encoding: Stream default Decimal symbol: Stream default

☐ Line delimiter is newline character Lines to scan for column and type: 50

Field delimiters

☐ Space ☑ Comma ☐ Tab

☑ Newline ☐ Other

☐ Non-printing characters

☐ Allow multiple blank delimiters

☑ Automatically recognize dates and times

☐ Treat square brackets as lists

Quotes

Single quotes: Discard

Double quotes: Discard

OK Cancel Apply Reset

Var. File node – Preview

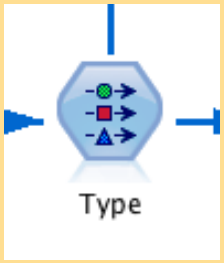
Preview from dementia.csv Node (163 fields, 10 records)

File Edit Generate

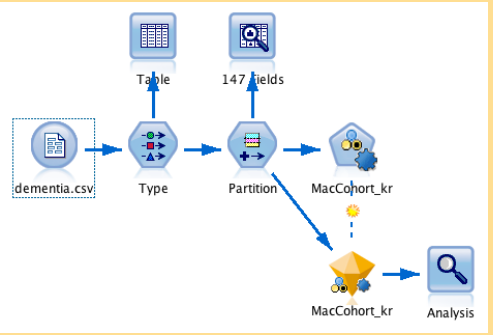
Table Annotations

	field1	PIDN	DCDate	SourceID	ScanNo	vbmID	gm	wm	csf	tiv	MagnetStrength	ageAtScan	ScannerID	ScanType	ROWNUM	InstrID	ScannerID_2	MacCohort_kr	PIBPos
1	1	65	2011-06-23	ABN00248-1	1	4951	595.483	521.213	284.383	1401.0...	3	70	NIC 3T MRI	MP-LAS-ABN	1	2106...	NIC 3T MRI	CONTROL	NA
2	800	3576	2009-08-11	NIC060-1	1	4562	559.454	549.746	292.001	1401.2...	3	72	NIC 3T MRI	T1-long	1493	1486...	NIC 3T MRI	AD	NA
3	915	4042	2010-02-08	NIC096-1	1	4247	687.109	551.160	298.904	1537.1...	3	73	NIC 3T MRI	T1-long	1696	1618...	NIC 3T MRI	AD	NA
4	1063	4992	2011-05-26	NIC225-1	1	4549	644.104	586.846	515.496	1746.4...	3	78	NIC 3T MRI	T1-long	1972	2078...	NIC 3T MRI	AD	NA
5	1154	5480	2009-02-05	GHB087-1	1	1781	550.418	422.150	285.142	1257.7...	3	80	NIC 3T MRI	T1-long	1153	1328...	NIC 3T MRI	CONTROL	NA
6	1201	5730	2012-02-14	NIC253-1	1	4020	445.386	451.453	356.003	1252.8...	3	78	NIC 3T MRI	T1-long	2105	2306...	NIC 3T MRI	AD	NA
7	1233	5896	2009-10-14	NIC072-1	1	3684	567.450	547.406	311.409	1426.2...	3	71	NIC 3T MRI	T1-long	2165	1552...	NIC 3T MRI	AD	NA
8	1237	5927	2010-07-08	ADRC0054-1	1	2697	504.363	391.272	336.735	1232.3...	3	82	NIC 3T MRI	T1-long	2172	1777...	NIC 3T MRI	AD	NA
9	1321	6438	2010-01-05	GHB212-1	1	2583	589.887	468.907	264.074	1322.8...	3	65	NIC 3T MRI	T1-long	1320	1578...	NIC 3T MRI	CONTROL	NA
10	1327	6565	2008-04-18	GHB005-1	1	835	558.285	435.448	215.101	1208.8...	3	66	NIC 3T MRI	T1-long	1326	1264...	NIC 3T MRI	CONTROL	NA

OK



Type node – initial



Type

Preview

Types | Format | Annotations

Read Values | Clear Values | Clear All Values

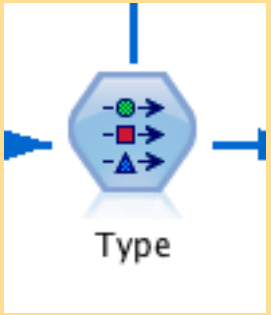
Field	Measurement	Values	Missing	Check	Role
field1	Continuous	<Read>		None	Input
PIDN	Continuous	<Read>		None	Input
DCDate	Continuous	<Read>		None	Input
SourceID	Categorical	<Read>		None	Input
ScanNo	Continuous	<Read>		None	Input
vbmID	Continuous	<Read>		None	Input
qm	Continuous	<Read>		None	Input
wm	Continuous	<Read>		None	Input
csf	Continuous	<Read>		None	Input
tiv	Continuous	<Read>		None	Input
MagnetStreng...	Continuous	<Read>		None	Input
ageAtScan	Continuous	<Read>		None	Input
ScannerID	Categorical	<Read>		None	Input
ScanType	Categorical	<Read>		None	Input
ROWNUM	Continuous	<Read>		None	Input
InstrID	Continuous	<Read>		None	Input
ScannerID_2	Categorical	<Read>		None	Input
MacCohort_kr	Categorical	<Read>		None	Input
PIBPos	Categorical	<Read>		None	Input
Sex	Categorical	<Read>		None	Input
Educ	Continuous	<Read>		None	Input
Race	Categorical	<Read>		None	Input
DateofDeath	Continuous	<Read>		None	Input
autopsy.	Categorical	<Read>		None	Input
PIBDone.	Categorical	<Read>		None	Input
X3rd_Ventricle	Continuous	<Read>		None	Input
X4th_Ventricle	Continuous	<Read>		None	Input
Right_Accum...	Continuous	<Read>		None	Input
Left_Accumbe...	Continuous	<Read>		None	Input
Right_Amygd...	Continuous	<Read>		None	Input
Left_Amygdala	Continuous	<Read>		None	Input
Brain_Stem	Continuous	<Read>		None	Input
Right_Caudate	Continuous	<Read>		None	Input
Left_Caudate	Continuous	<Read>		None	Input
Right_Cerebel...	Continuous	<Read>		None	Input
Left_Cerebell...	Continuous	<Read>		None	Input
Right_Cerebel...	Continuous	<Read>		None	Input
Left_Cerebell...	Continuous	<Read>		None	Input
Right_Cerebr...	Continuous	<Read>		None	Input

Left_PCu_prec...	Continuous	<Read>		None	Input
Right_PHG_pa...	Continuous	<Read>		None	Input
Left_PHG_par...	Continuous	<Read>		None	Input
Right_PIns_po...	Continuous	<Read>		None	Input
Left_PIns_post...	Continuous	<Read>		None	Input
Right_PO_pari...	Continuous	<Read>		None	Input
Left_PO_parie...	Continuous	<Read>		None	Input
Right_PoG_po...	Continuous	<Read>		None	Input
Left_PoG_post...	Continuous	<Read>		None	Input
Right_POrG_p...	Continuous	<Read>		None	Input
Left_POrG_po...	Continuous	<Read>		None	Input
Right_PP_plan...	Continuous	<Read>		None	Input
Left_PP_planu...	Continuous	<Read>		None	Input
Right_PrG_pre...	Continuous	<Read>		None	Input
Left_PrG_prec...	Continuous	<Read>		None	Input
Right_PT_plan...	Continuous	<Read>		None	Input
Left_PT_planu...	Continuous	<Read>		None	Input
Right_SCA_su...	Continuous	<Read>		None	Input
Left_SCA_sub...	Continuous	<Read>		None	Input
Right_SFG_su...	Continuous	<Read>		None	Input
Left_SFG_sup...	Continuous	<Read>		None	Input
Right_SMC_su...	Continuous	<Read>		None	Input
Left_SMC_sup...	Continuous	<Read>		None	Input
Right_SMG_su...	Continuous	<Read>		None	Input
Left_SMG_sup...	Continuous	<Read>		None	Input
Right_SOG_su...	Continuous	<Read>		None	Input
Left_SOG_sup...	Continuous	<Read>		None	Input
Right_SPL_su...	Continuous	<Read>		None	Input
Left_SPL_supe...	Continuous	<Read>		None	Input
Right_STG_su...	Continuous	<Read>		None	Input
Left_STG_sup...	Continuous	<Read>		None	Input
Right_TMP_te...	Continuous	<Read>		None	Input
Left_TMP_tem...	Continuous	<Read>		None	Input
Right_TrIFG_t...	Continuous	<Read>		None	Input
Left_TrIFG_tri...	Continuous	<Read>		None	Input
Right_TTG_tra...	Continuous	<Read>		None	Input
Left_TTG_tran...	Continuous	<Read>		None	Input
drop	Categorical	<Read>		None	Input
Dementia	Categorical	<Read>		None	Input

View current fields | View unused field settings

OK | Cancel | Apply | Reset

Type node – after



Type

Preview

Types | Format | Annotations

Read Values | Clear Values | Clear All Values

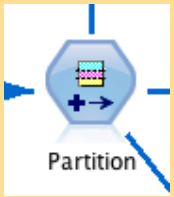
Field	Measurement	Values	Missing	Check	Role
field1	Continuous	[1,3095]		None	None
PIDN	Continuous	[65,19260]		None	Record ID
DCDate	Continuous	[2007-11...		None	None
SourceID	Typeless			None	None
ScanNo	Ordinal	1		None	None
vbmID	Continuous	[5,4963]		None	None
qm	Continuous	[362.558,...		None	Input
wm	Continuous	[354.125,...		None	Input
csf	Continuous	[189.1,51...		None	Input
tiv	Continuous	[1007.5,1...		None	Input
MagnetStre...	Nominal	3		None	None
ageAtScan	Continuous	[33,99]		None	Input
ScannerID	Flag	"NIC 3T M...		None	None
ScanType	Nominal	MP-LAS-...		None	None
ROWNUM	Continuous	[1,3095]		None	None
InstrID	Continuous	[126430,...		None	None
ScannerID_2	Flag	"NIC 3T M...		None	None
MacCohort_...	Nominal	AD,CONT...		None	Target
PIBPos	Nominal	"0","1",NA		None	None
Sex	Flag	male/fem...		None	Input
Educ	Continuous	[4,32]		None	Input
Race	Nominal	"Asian In...		None	Input
DateofDeath	Continuous	[2010-06...		None	None
autopsy.	Nominal	"","AUTOP...		None	None
PIBDone.	Flag	Yes/No		None	None
X3rd_Ventri...	Continuous	[58.7565...		None	Input
X4th_Ventri...	Continuous	[91.3478...		None	Input
Right_Accu...	Continuous	[31.5960...		None	Input
Left_Accum...	Continuous	[21.7114...		None	Input
Right_Amyg...	Continuous	[84.1284...		None	Input
Left_Amygd...	Continuous	[68.7341...		None	Input
Brain_Stem	Continuous	[506.597...		None	Input
Right_Caud...	Continuous	[126.647...		None	Input
Left_Caudate	Continuous	[117.645...		None	Input
Right_Cere...	Continuous	[5573.55...		None	Input
Left_Cerebe...	Continuous	[4978.18...		None	Input

Left_PP_pla...	Continuous	[145.689...		None	Input
Right_PrG_...	Continuous	[881.938...		None	Input
Left_PrG_pr...	Continuous	[1013.28...		None	Input
Right_PT_pl...	Continuous	[105.518...		None	Input
Left_PT_pla...	Continuous	[181.947...		None	Input
Right_SCA_...	Continuous	[64.3291...		None	Input
Left_SCA_s...	Continuous	[65.6406...		None	Input
Right_SFG_...	Continuous	[1022.29...		None	Input
Left_SFG_su...	Continuous	[813.819...		None	Input
Right_SMC_...	Continuous	[353.486...		None	Input
Left_SMC_s...	Continuous	[302.230...		None	Input
Right_SMG_...	Continuous	[565.565...		None	Input
Left_SMG_s...	Continuous	[854.651...		None	Input
Right_SOG_...	Continuous	[228.080...		None	Input
Left_SOG_s...	Continuous	[170.802...		None	Input
Right_SPL_s...	Continuous	[934.200...		None	Input
Left_SPL_su...	Continuous	[765.355...		None	Input
Right_STG_...	Continuous	[314.631...		None	Input
Left_STG_su...	Continuous	[542.556...		None	Input
Right_TMP_...	Continuous	[163.171...		None	Input
Left_TMP_te...	Continuous	[65.9406...		None	Input
Right_TrIFG...	Continuous	[179.009...		None	Input
Left_TrIFG_t...	Continuous	[97.0478...		None	Input
Right_TTG_...	Continuous	[74.6414...		None	Input
Left_TTG_tr...	Continuous	[161.437...		None	Input
drop	Flag	FALSE/FA...		None	None
Dementia	Flag	TRUE/FAL...		None	None

View current fields | View unused field settings

OK | Cancel | Apply | Reset

Partition node



Partition

Generate Preview

Settings Annotations

Partition field: Partition

Partitions: Train and test | Train, test and validation

		Label:	Value =
Training partition size:	40	Training	"1_Training"
Testing partition size:	30	Testing	"2_Testing"
Validation partition size:	30	Validation	"3_Validation"

Total size: 100%

Values: Use system-defined values ("1", "2" and "3")

Append labels to system-defined values

Use labels as values

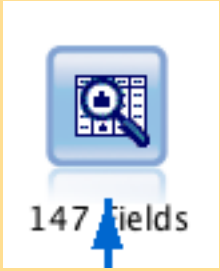
Repeatable partition assignment

Seed: 1234567 Generate

Use unique field to assign partitions:

OK Cancel Apply Reset

Audit node



147 Fields

Settings | Quality | Output | Annotations

Default | Use custom fields

Fields:

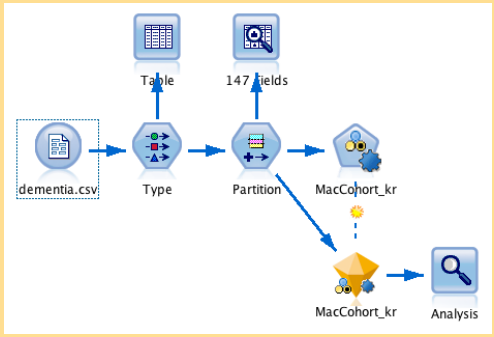
Overlay:

Display

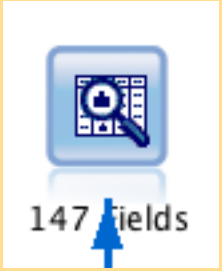
Graphs | Basic statistics | Advanced statistics

Calculate median and mode (may slow performance on large datasets)

OK | Run | Cancel | Apply | Reset



Audit node



Data Audit of [147 fields]

File Edit Generate

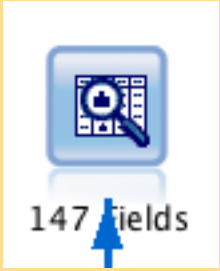
Audit Quality Annotations

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
PIDN		Continuous	65	19260	12372.7...	3514.176	-0.021	--	660
gm		Continuous	362.558	826.936	568.708	65.967	0.031	--	660
wm		Continuous	354.125	738.869	492.911	61.072	0.478	--	660
csf		Continuous	189.100	515.496	300.274	55.397	0.496	--	660
tiv		Continuous	1007.500	1843.110	1361.893	132.946	0.437	--	660
ageAtScan		Continuous	33	99	66.698	8.945	0.082	--	660
MacCohort_kr		Nominal	--	--	--	--	--	6	660
Sex		Flag	--	--	--	--	--	2	660

[1] Indicates a multimode result [2] Indicates a sampled result

OK

Audit node



Data Audit of [147 fields]

File Edit Generate

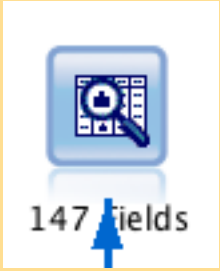
Audit Quality Annotations

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Left_STG_superior_temporal_gyrus		Continuous	542.556	1828.959	1282.501	210.340	-0.742	--	660
Right_TMP_temporal_pole		Continuous	163.171	2649.916	1797.301	381.095	-1.262	--	660
Left_TMP_temporal_pole		Continuous	65.941	2461.309	1655.974	386.556	-1.317	--	660
Right_TrIFG_triangular_part_of_the_inferior_frontal_g...		Continuous	179.010	872.055	602.290	103.471	-0.795	--	660
Left_TrIFG_triangular_part_of_the_inferior_frontal_gyr...		Continuous	97.048	813.283	529.719	102.303	-0.758	--	660
Right_TTG_transverse_temporal_gyrus		Continuous	74.641	525.473	302.991	58.504	-0.026	--	660
Left_TTG_transverse_temporal_gyrus		Continuous	161.437	533.214	342.375	66.575	0.047	--	660
Partition		Nominal	--	--	--	--	--	3	660

[1] Indicates a multimode result [2] Indicates a sampled result

OK

Audit node



Data Audit of [147 fields]

File Edit Generate

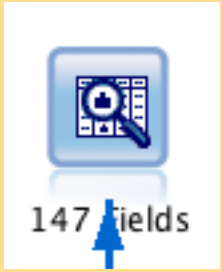
Audit Quality Annotations

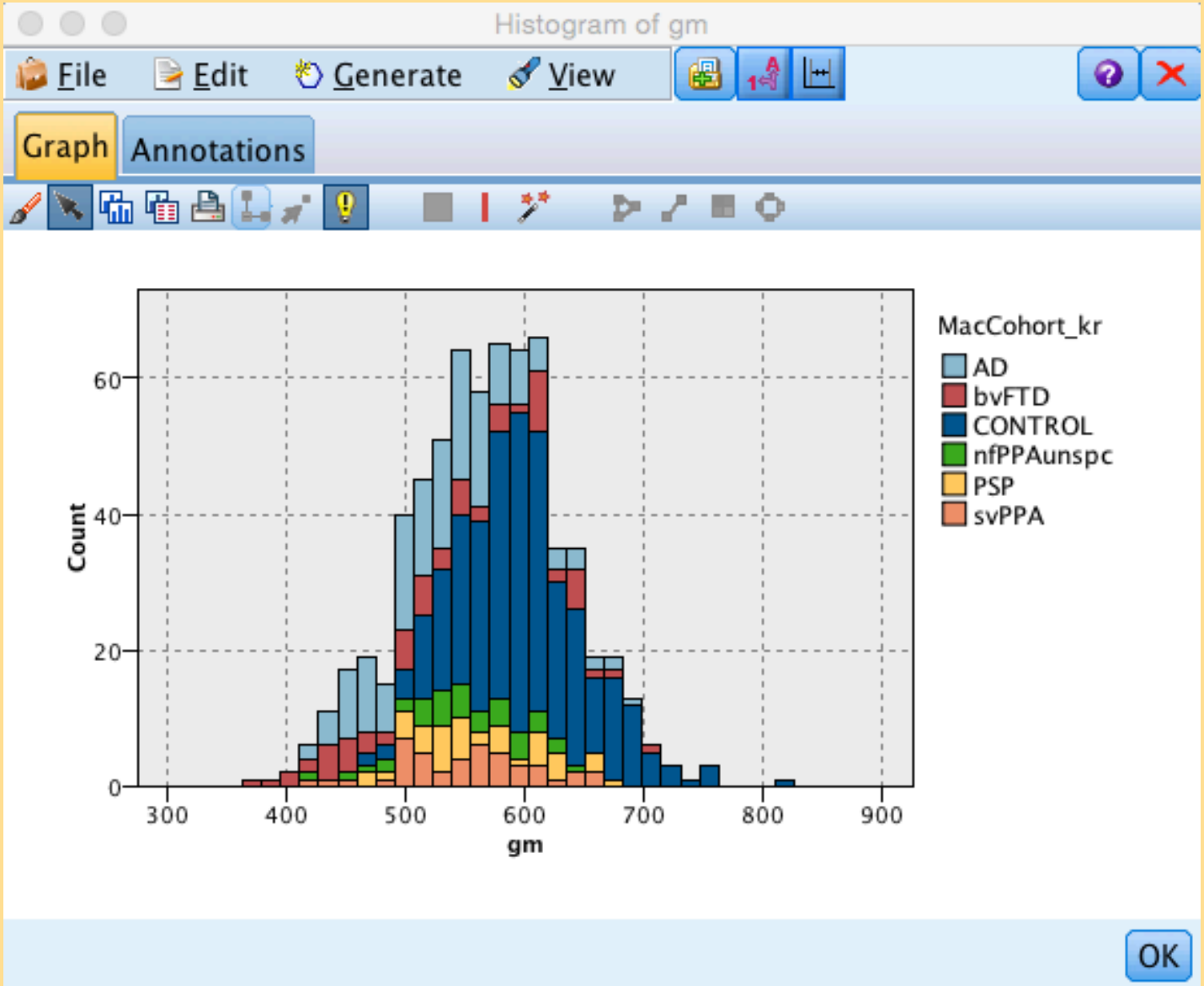
Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
PIDN		Continuous	65	19260	12372.7...	3514.176	-0.021	--	660
gm		Continuous	362.558	826.936	568.708	65.967	0.031	--	660
wm		Continuous	354.125	738.869	492.911	61.072	0.478	--	660
csf		Continuous	189.100	515.496	300.274	55.397	0.496	--	660
tiv		Continuous	1007.500	1843.110	1361.893	132.946	0.437	--	660
ageAtScan		Continuous	33	99	66.698	8.945	0.082	--	660
MacCohort_kr		Nominal	--	--	--	--	--	6	660
Sex		Flag	--	--	--	--	--	2	660

[1] Indicates a multimode result [2] Indicates a sampled result

OK

Audit node


147 fields


Histogram of gm
File
Edit
Generate
View
Graph
Annotations
Count
60
40
20
0
300
400
500
600
700
800
900
gm
MacCohort_kr
AD
bvFTD
CONTROL
nfPPAunspc
PSP
svPPA
OK

Audit node

Data Audit of [147 fields] #1

File Edit Generate

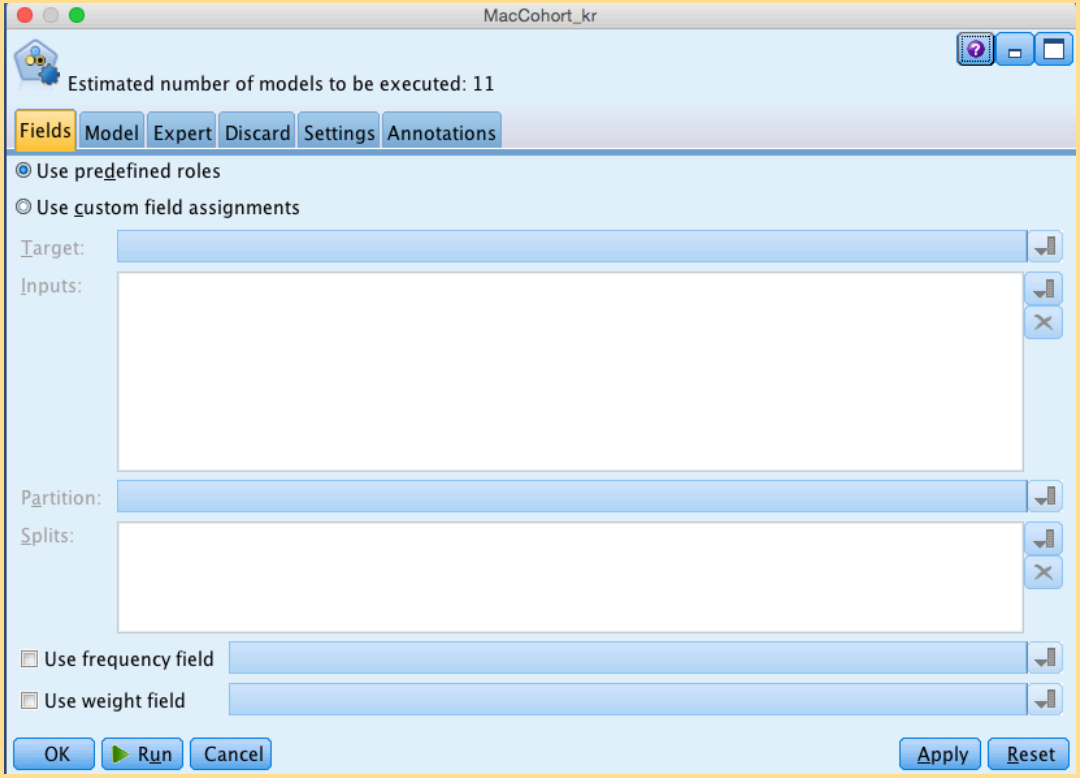
Audit Quality Annotations

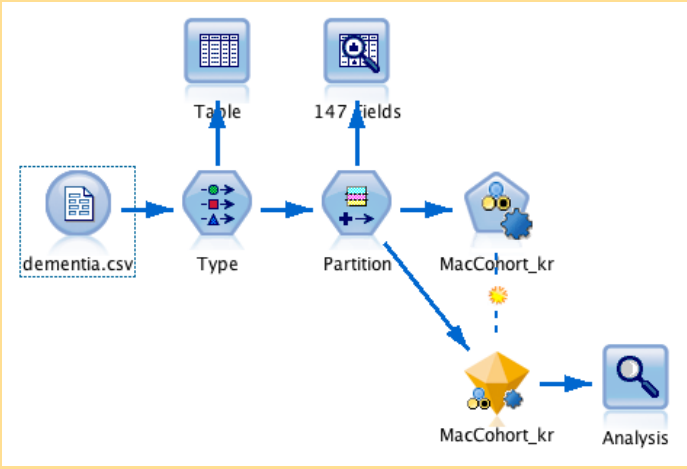
Complete fields (%): 99.32% Complete records (%): 96.82%

Field	Measurement	Outliers	Extremes	Action	Impute Missi...	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
PIDN	Continuous	1	0	None	Never	Fixed	100	660	0	0	0	0
gm	Continuous	2	0	None	Never	Fixed	100	660	0	0	0	0
wm	Continuous	5	0	None	Never	Fixed	100	660	0	0	0	0
csf	Continuous	4	0	None	Never	Fixed	100	660	0	0	0	0
tiv	Continuous	4	0	None	Never	Fixed	100	660	0	0	0	0
ageAtScan	Continuous	2	0	None	Never	Fixed	100	660	0	0	0	0
MacCohor...	Nominal	--	--	--	Never	Fixed	100	660	0	0	0	0
Sex	Flag	--	--	--	Never	Fixed	100	660	0	0	0	0
Educ	Continuous	10	1	None	Never	Fixed	96.818	639	21	0	0	0
Race	Nominal	--	--	--	Never	Fixed	100	660	0	0	0	0
X3rd_Ven...	Continuous	1	0	None	Never	Fixed	100	660	0	0	0	0
X4th_Vent...	Continuous	5	1	None	Never	Fixed	100	660	0	0	0	0
Right_Acc...	Continuous	7	0	None	Never	Fixed	100	660	0	0	0	0
Left_Accu...	Continuous	1	0	None	Never	Fixed	100	660	0	0	0	0
Right_Am...	Continuous	10	0	None	Never	Fixed	100	660	0	0	0	0
Left_Amy...	Continuous	4	0	None	Never	Fixed	100	660	0	0	0	0
Brain_Stem	Continuous	7	4	None	Never	Fixed	100	660	0	0	0	0
Right_Cau...	Continuous	5	0	None	Never	Fixed	100	660	0	0	0	0
Left_Caud...	Continuous	5	0	None	Never	Fixed	100	660	0	0	0	0
Right_Cer...	Continuous	7	1	None	Never	Fixed	100	660	0	0	0	0
Left_Cere...	Continuous	6	1	None	Never	Fixed	100	660	0	0	0	0
Right_Cer...	Continuous	4	1	None	Never	Fixed	100	660	0	0	0	0
Left_Cere...	Continuous	6	1	None	Never	Fixed	100	660	0	0	0	0
Right_Cer...	Continuous	1	0	None	Never	Fixed	100	660	0	0	0	0
Left_Cere...	Continuous	0	0	None	Never	Fixed	100	660	0	0	0	0
Cerebros...	Continuous	2	0	None	Never	Fixed	100	660	0	0	0	0
Right_Hip...	Continuous	5	0	None	Never	Fixed	100	660	0	0	0	0
Left_Hipp...	Continuous	5	0	None	Never	Fixed	100	660	0	0	0	0
Right_Inf_...	Continuous	8	0	None	Never	Fixed	100	660	0	0	0	0
Left_Inf_L...	Continuous	2	0	None	Never	Fixed	100	660	0	0	0	0
Right_Lat...	Continuous	15	1	None	Never	Fixed	100	660	0	0	0	0
Left_SCA_...	Continuous	12	0	None	Never	Fixed	100	660	0	0	0	0
Right_SFG...	Continuous	9	0	None	Never	Fixed	100	660	0	0	0	0
Left_SFG_...	Continuous	8	0	None	Never	Fixed	100	660	0	0	0	0
Right_SM...	Continuous	5	0	None	Never	Fixed	100	660	0	0	0	0
Left_SMC_...	Continuous	2	0	None	Never	Fixed	100	660	0	0	0	0
Right_SM...	Continuous	2	0	None	Never	Fixed	100	660	0	0	0	0
Left_SMG_...	Continuous	2	0	None	Never	Fixed	100	660	0	0	0	0
Right_SO...	Continuous	10	0	None	Never	Fixed	100	660	0	0	0	0
Left_SOG_...	Continuous	8	0	None	Never	Fixed	100	660	0	0	0	0
Right_SPL...	Continuous	11	0	None	Never	Fixed	100	660	0	0	0	0
Left_SPL_s...	Continuous	9	0	None	Never	Fixed	100	660	0	0	0	0
Right_STG...	Continuous	6	0	None	Never	Fixed	100	660	0	0	0	0
Left_STG_...	Continuous	5	0	None	Never	Fixed	100	660	0	0	0	0
Right_TM...	Continuous	14	0	None	Never	Fixed	100	660	0	0	0	0
Left_TMP_...	Continuous	15	0	None	Never	Fixed	100	660	0	0	0	0
Right_TrIF...	Continuous	7	0	None	Never	Fixed	100	660	0	0	0	0
Left_TrIFG...	Continuous	8	0	None	Never	Fixed	100	660	0	0	0	0
Right_TT...	Continuous	4	0	None	Never	Fixed	100	660	0	0	0	0
Left_TTG_...	Continuous	0	0	None	Never	Fixed	100	660	0	0	0	0
Partition	Nominal	--	--	--	Never	Fixed	100	660	0	0	0	0

OK

Auto-classifier node





Auto-classifier node

MacCohort_kr

Estimated number of models to be executed: 11

Fields | Model | Expert | Discard | Settings | Annotations

Model name: Auto Custom

Use partitioned data

Build model for each split

Rank models by: Overall accuracy

Rank models using: Training partition Test partition

Number of models to use: 8

Calculate predictor importance

Profit Criteria (valid only for flag targets)

Costs: Fixed 5.0 Variable

Revenue: Fixed 10.0 Variable

Weight: Fixed 1.0 Variable

Lift Criteria (valid only for flag targets)

Percentile to use for lift calculation: 30

OK | Run | Cancel | Apply | Reset

Auto-classifier node

MacCohort_kr

Estimated number of models to be executed: 7

Fields | Model | Expert | Discard | Settings | Annotations

Select models: All models

Use?	Model type	Model parameters	No of models
☑	C5	Default	1
☑	Logistic regression	Default	1
☑	Decision List	Default	1
☑	Bayesian Network	Default	1
☑	Discriminant	Default	1
☑	KNN Algorithm	Default	1
☐	LSVM	Default	1
☐	Random Trees	Default	1
☑	SVM	Default	1
☐	Tree-AS	Default	1
☐	CHAID	Default	1
☐	Quest	Default	1
☐	C&R Tree	Default	1
☑	Neural Net	Default	1

☐ Restrict maximum time spent building a single model to 15 minutes

Stopping rules... Misclassification costs...

OK | Run | Cancel | Apply | Reset

Auto-classifier node

MacCohort_kr

Estimated number of models to be executed: 7

Fields | Model | Expert | Discard | Settings | Annotations

Select models: All models

Use?	Model type	Model parameters	No of models
☑	C5	Default	1
☑	Logistic regression	Default	1
☑	Decision List	Default	1
☑	Bayesian Network	Default	1
☑	Discriminant	Default	1
☑	KNN Algorithm	Default	1
☐	LSVM	Default	1
☐	Random Trees	Default	1
☑	SVM	Default	1
☐	Tree-AS	Default	1
☐	CHAID	Default	1
☐	Quest	Default	1
☐	C&R Tree	Default	1
☑	Neural Net	Default	1

☐ Restrict maximum time spent building a single model to 15 minutes

Stopping rules... Misclassification costs...

OK | Run | Cancel | Apply | Reset

Auto-classifier node

MacCohort_kr

Estimated number of models to be executed: 7

Fields | Model | Expert | Discard | Settings | Annotations

Select models: All models

Use?	Model type	Model parameters	No of models
☑	C5	Default	1
☑	Logistic regression	Default	1
☑	Decision List	Default	1
☑	Bayesian Network	Default	
☑	Discriminant	Default	
☑	KNN Algorithm	Default	
☐	LSVM	Default	
☐	Random Trees	Default	
☑	SVM	Default	
☐	Tree-AS	Default	
☐	CHAID	Default	
☐	Quest	Default	
☐	C&R Tree	Default	
☑	Neural Net	Default	1

☐ Restrict maximum time spent building a single model to 15 minutes

Stopping rules... Misclassification costs...

OK | Run | Cancel | Apply | Reset

Algorithm settings - Logistic regression

Simple | Expert

Parameter	Options
Procedure	Multinomial
Include constant in equation	true
Multinomial Method	Enter
Base category	
Model type	Main Effects
Binomial Method	Enter

OK | Cancel | Help

Auto-classifier node

MacCohort_kr

Estimated number of models to be executed: 7

Fields | Model | Expert | Discard | Settings | Annotations

Select models: All models

Use?	Model type	Model parameters	No of models
☑	C5	Default	1
☑	Logistic regression	Default	1
☑	Decision List	Default	1
☑	Bayesian Network	Default	1
☑	Discriminant	Default	1
☑	KNN Algorithm	Default	1
☐	LSVM	Default	1
☐	Random Trees	Default	1
☑	SVM	Default	1
☐	Tree-AS	Default	1
☐	CHAID	Default	1
☐	Quest	Default	1
☐	C&R Tree	Default	1
☑	Neural Net	Default	1

☐ Restrict maximum time spent building a single model to 15 minutes

Stopping rules... Misclassification costs...

OK | Run | Cancel | Apply | Reset

Auto-classifier node

MacCohort_kr

Estimated number of models to be executed: 9

Fields | Model | Expert | Discard | Settings | Annotations

Select models: All models

Use?	Model type	Model parameters	No of models
☑	C5	Default	1
☑	Logistic regression	Default	1
☑	Decision List	Default	1
☑	Bayesian Network	Default	1
☑	Discriminant	Default	1
☑	KNN Algorithm	Specify...	3
☐	LSVM	Default	1
☐	Random Trees	Default	1
☑	SVM	Default	1
☐	Tree-AS	Default	1
☐	CHAID	Default	1
☐	Quest	Default	1
☐	C&R Tree	Default	1
☑	Neural Net	Default	1

☐ Restrict maximum time spent building a single model to 15

Stopping rules... | Misclassification costs...

OK | Run | Cancel | Apply | Reset

Algorithm settings - KNN Algorithm

Simple | Expert

Parameter	Options
Normalize range inputs	true
Use case labels	false
Case Label variable	
Identify focal record	false
Focal case variable	
Automatically select k	false
K	3; 6; 10
Minimum	3
Maximum	5
Distance Computation	Euclidean metric
Weight features by importance w...	true
Predictions for Range Target	Mean of nearest neighbor values
Perform feature selection	false
Stop when the change in the abs...	false
Number to select	10
Minimum change	0.01
Use field to assign cases	false
Number of folds	10
Set random seed	false
Seed	12345
Folds variable	

OK | Cancel | Help

Auto-classifier node

MacCohort_kr

Estimated number of models to be executed: 9

Fields | Model | Expert | Discard | Settings | Annotations

Do not keep models if:

- Overall accuracy is less than 80 %
- Number of fields is greater than 10

Valid only for flag targets

- Area under the curve is less than 0.8
- Lift is less than 2.0
- Profit is less than 200 %

Settings for lift and profit are as set in the Model tab

OK | Run | Cancel | Apply | Reset

Auto-classifier node

MacCohort_kr

Estimated number of models to be executed: 9

Fields | Model | Expert | Discard | Settings | Annotations

Ensemble Settings

Filter out fields generated by ensembled models

Set Target

Ensemble method: Confidence-weighted voting

If voting is tied, select value using:

- Random selection
- Highest confidence

OK | Run | Cancel | Apply | Reset

Golden nugget

MacCohort_kr

File Generate View Preview

Model Graph Summary Settings Annotations

Sort by: Use Ascending Descending Delete Unused Models View: Testing set

Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
☑		SVM 1	< 1	76.19	144
☑		Discriminant 1	< 1	67.196	142
☐		C5 1	< 1	66.138	22
☑		Neural Net 1	< 1	61.905	144
☑		Logistic regre...	< 1	43.386	144
☑		Bayesian Netw...	< 1	8.466	144

OK Cancel Apply Reset

Golden nugget

MacCohort_kr

File Generate View Preview

Model Graph Summary Settings Annotations

Sort by: Use Ascending Descending Delete Unused Models View: Training set

Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
☑		Discriminant 1	< 1	97.037	142
☑		Logistic regre...	< 1	96.296	144
☑		SVM 1	< 1	95.926	144
☐		C5 1	< 1	88.889	22
☑		Neural Net 1	< 1	71.111	144
☑		Bayesian Netw...	< 1	19.259	144

OK Cancel Apply Reset

Golden nugget

MacCohort_kr

File Generate View Preview

Model Graph Summary Settings Annotations

Sort by: Use Ascending Descending Delete Unused Models View: Testing set

Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
☑		SVM 1	< 1	76.19	144
☑		Discriminant 1	< 1	67.196	142
☑		C5 1	< 1	66.138	22
☑		Neural Net 1	< 1	61.905	144
☑		Logistic regre...	< 1	43.386	144
☑		Bayesian Netw...	< 1	8.466	144

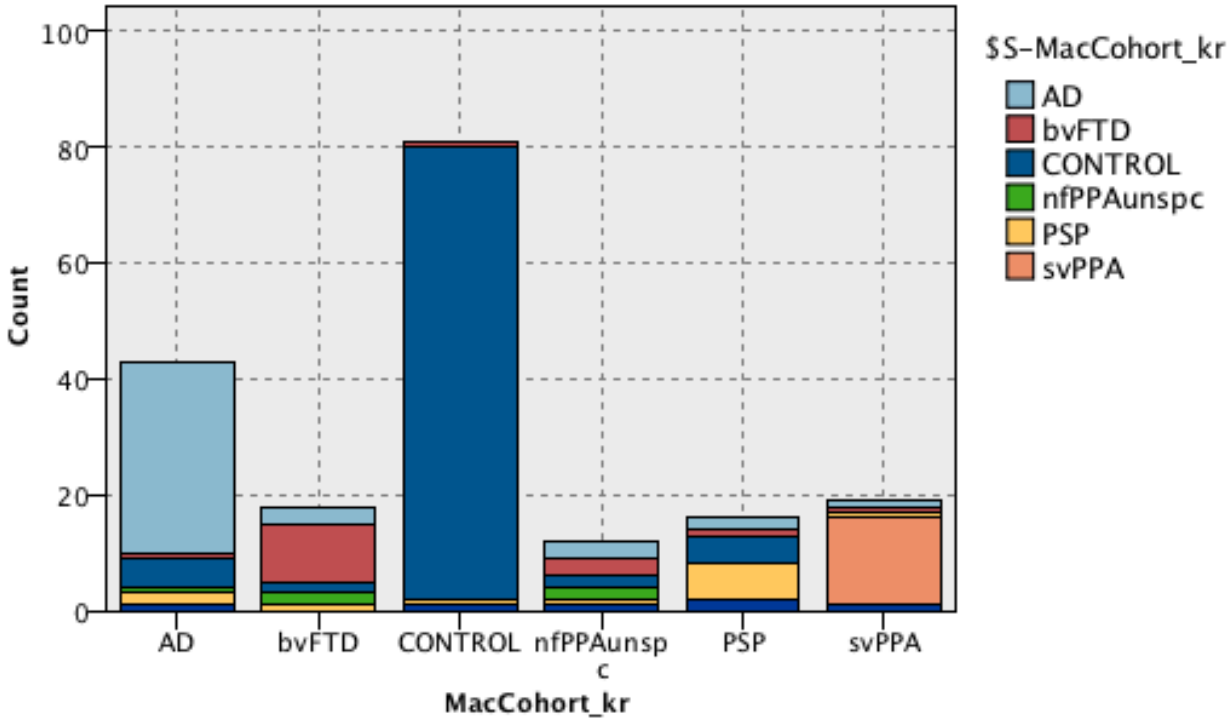
OK Cancel Apply Reset

Golden nugget

Graphboard of [MacCohort_kr $S-MacCohort_kr] #1

File Edit Generate View

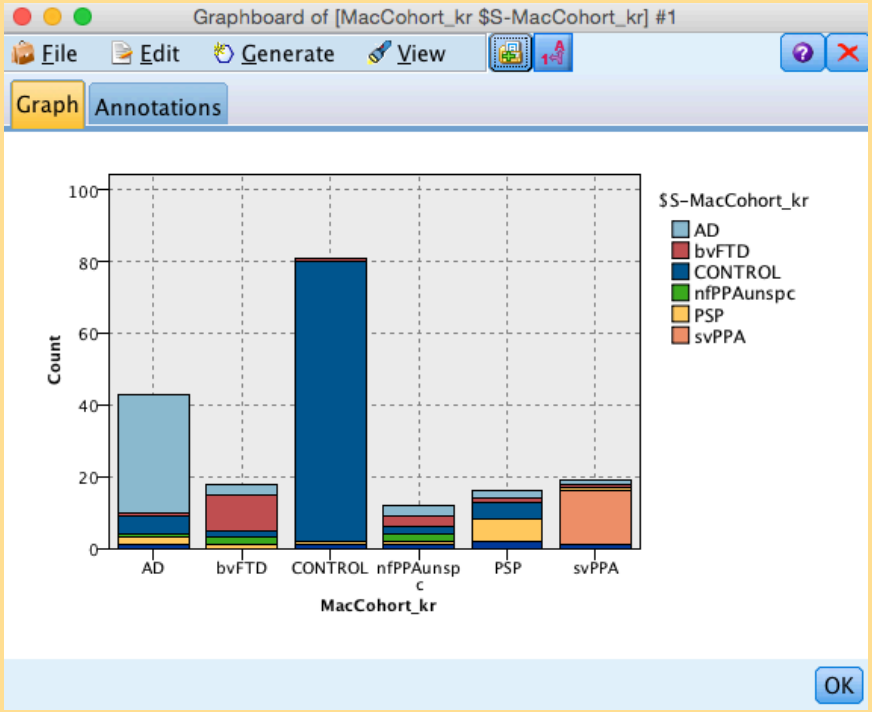
Graph Annotations



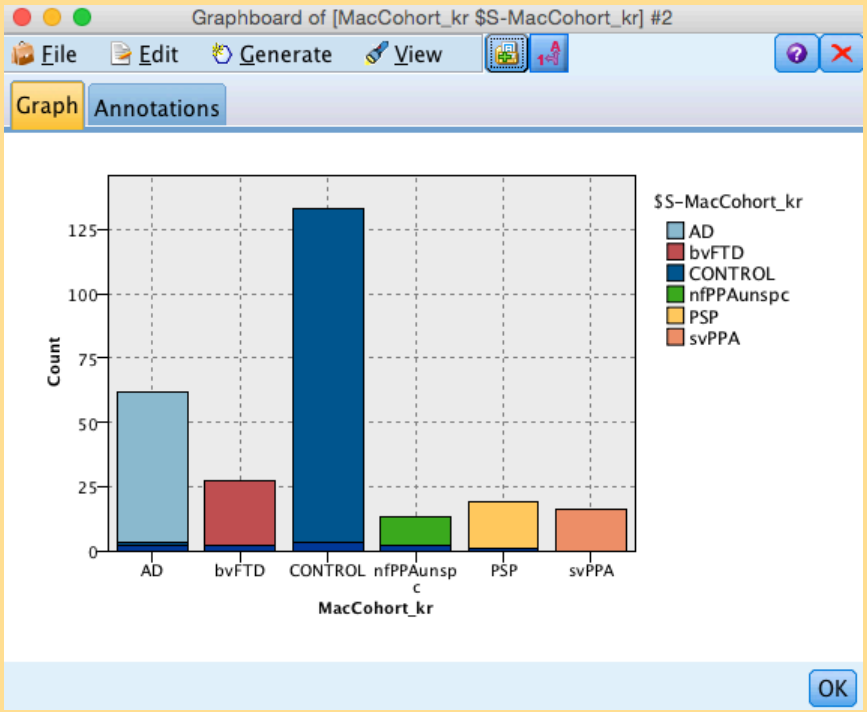
OK

Golden nugget

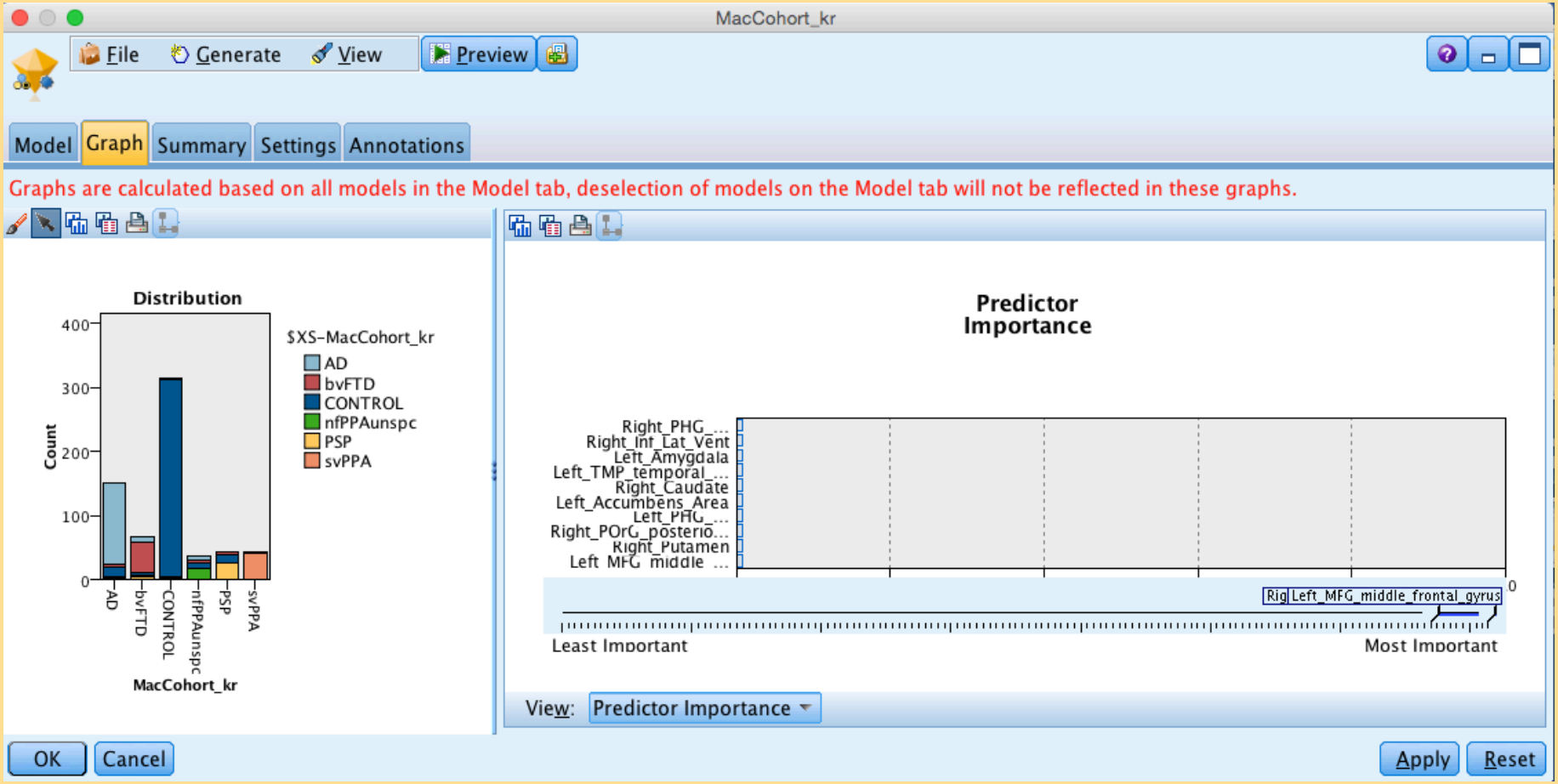
Test



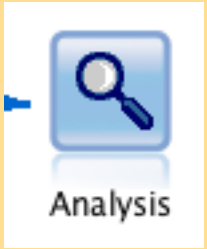
Training

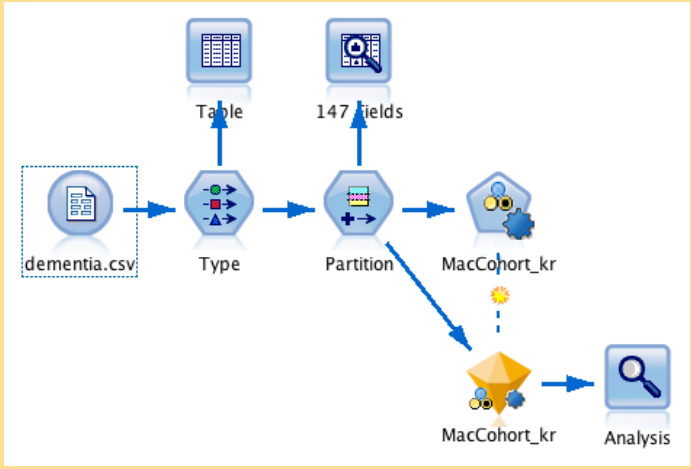


Golden nugget



Analysis node





Analysis

Analyze $XS-MacCohort_kr

Analysis | Output | Annotations

- [x] Coincidence matrices (for symbolic targets)
- [x] Performance evaluation
- [ ] Evaluation metric (AUC & Gini, binary classifiers only)
- [x] Confidence figures (if available)

Threshold for: 90 % correct

Improve accuracy: 2.0 fold

Find predicted/predictor fields using:

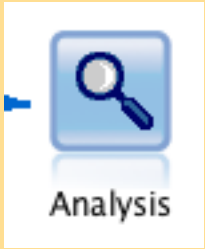
- (•) Model output field metadata
- ( ) Field name format (for example, '$<x>-<target field>')

- [x] Separate by partition
- [ ] User defined analysis — Define User Measure...

Break down analysis by fields:

OK | Run | Cancel | Apply | Reset

Analysis node



Analysis of [MacCohort_kr]

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field MacCohort_kr

Comparing $XS-MacCohort_kr with MacCohort_kr

'Partition'	1_Training		2_Testing		3_Validation	
Correct	266	98.52%	145	76.72%	154	76.62%
Wrong	4	1.48%	44	23.28%	47	23.38%
Total	270		189		201	

Coincidence Matrix for $XS-MacCohort_kr (rows show actuals)

'Partition' = 1_Training	AD	CONTROL	PSP	bvFTD	nfPPAunspc	svPPA
AD	61	1	0	0	0	0
CONTROL	0	133	0	0	0	0
PSP	0	0	19	0	0	0
bvFTD	0	1	0	25	0	1
nfPPAunspc	1	0	0	0	12	0
svPPA	0	0	0	0	0	16

'Partition' = 2_Testing	AD	CONTROL	PSP	bvFTD	nfPPAunspc	svPPA
AD	31	8	1	1	1	1
CONTROL	0	79	2	0	0	0
PSP	0	7	5	3	1	0
bvFTD	4	2	1	10	0	1
nfPPAunspc	5	4	0	0	3	0
svPPA	2	0	0	0	0	17

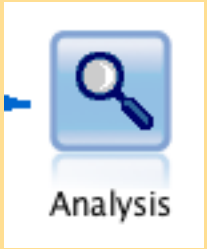
'Partition' = 3_Validation	AD	CONTROL	PSP	bvFTD	nfPPAunspc	svPPA
AD	35	6	2	3	0	0
CONTROL	1	95	2	2	1	0
PSP	0	5	2	2	0	0
bvFTD	4	2	2	13	2	0
nfPPAunspc	1	5	2	4	1	0
svPPA	1	0	0	0	0	8

Performance Evaluation

Confidence Values Report for $XSC-MacCohort_kr

OK

Analysis node



Analysis of [MacCohort_kr]

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field MacCohort_kr

Comparing $XS-MacCohort_kr with MacCohort_kr

'Partition'	1_Training		2_Testing		3_Validation	
Correct	266	98.52%	145	76.72%	154	76.62%
Wrong	4	1.48%	44	23.28%	47	23.38%
Total	270		189		201	

Coincidence Matrix for $XS-MacCohort_kr (rows show actuals)

Performance Evaluation

'Partition' = 1_Training	
AD	1.455
CONTROL	0.693
PSP	2.654
bvFTD	2.303
nfPPAunspc	3.033
svPPA	2.765

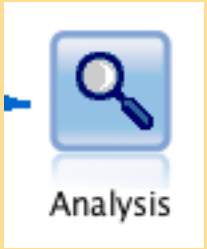
'Partition' = 2_Testing	
AD	1.177
CONTROL	0.612
PSP	1.881
bvFTD	2.015
nfPPAunspc	2.246
svPPA	2.186

'Partition' = 3_Validation	
AD	1.292
CONTROL	0.515
PSP	1.497
bvFTD	1.555
nfPPAunspc	1.352
svPPA	3.106

Confidence Values Report for $XSC-MacCohort_kr

OK

Analysis node



Analysis of [MacCohort_kr]

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field MacCohort_kr

Comparing $XS-MacCohort_kr with MacCohort_kr

'Partition'	1_Training		2_Testing		3_Validation	
Correct	266	98.52%	145	76.72%	154	76.62%
Wrong	4	1.48%	44	23.28%	47	23.38%
Total	270		189		201	

Coincidence Matrix for $XS-MacCohort_kr (rows show actuals)

Performance Evaluation

Confidence Values Report for $XSC-MacCohort_kr

'Partition' = 1_Training	
Range	0.356 - 0.978
Mean Correct	0.856
Mean Incorrect	0.604
Always Correct Above	0.963 (23.7% of cases)
Always Incorrect Below	0.356 (0% of cases)
98.52% Accuracy Above	0.0
2.0 Fold Correct Above	0.5 (99.62% of cases)

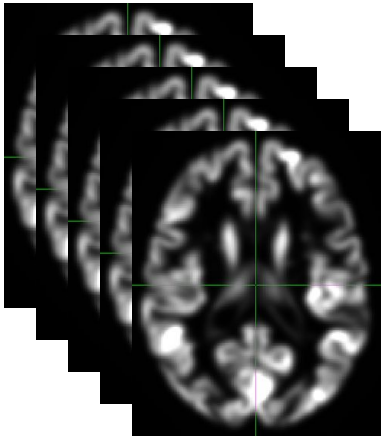
'Partition' = 2_Testing	
Range	0.282 - 0.978
Mean Correct	0.719
Mean Incorrect	0.561
Always Correct Above	0.975 (1.59% of cases)
Always Incorrect Below	0.282 (0% of cases)
90.11% Accuracy Above	0.722
2.0 Fold Correct Above	0.639 (88.57% of cases)

'Partition' = 3_Validation	
Range	0.2 - 1.0
Mean Correct	0.697
Mean Incorrect	0.452
Always Correct Above	0.902 (18.41% of cases)
Always Incorrect Below	0.249 (1% of cases)
90.14% Accuracy Above	0.486
2.0 Fold Correct Above	0.472 (88.67% of cases)

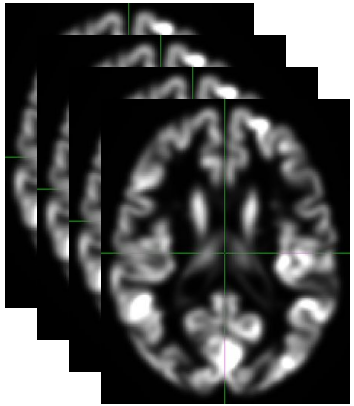
OK

Voxel-based analysis – Yann

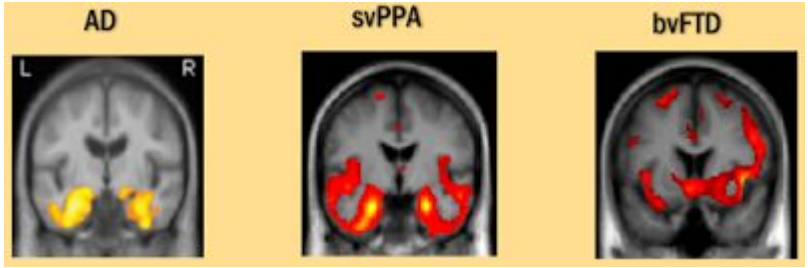
Classification



Group 1



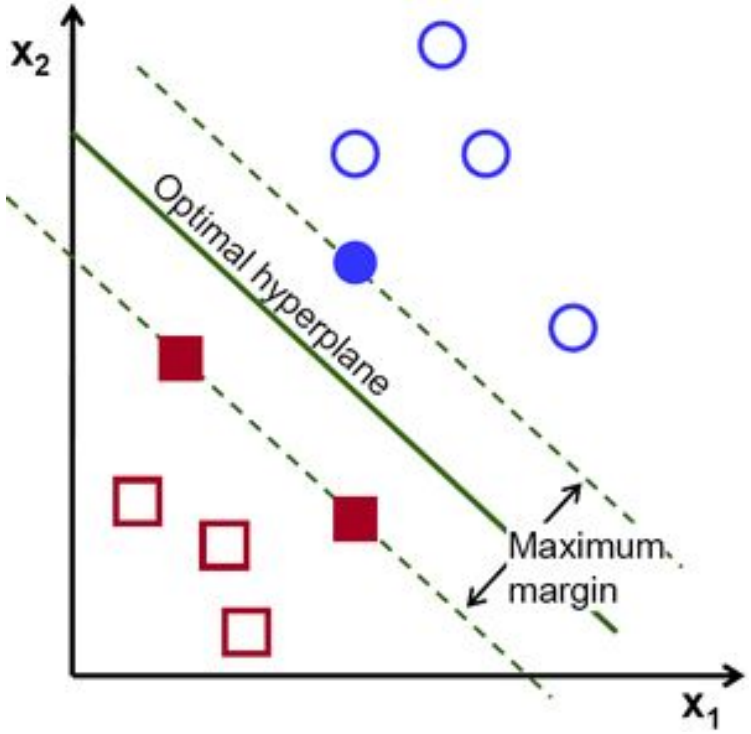
Group 2



Support Vector Machine

Algorithm

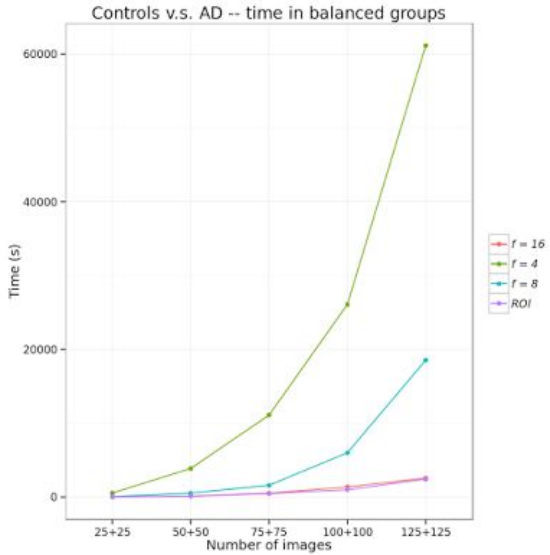
- Supervised parametric
  - Training and Testing
  - Cross-validation
- Maximization of margins
- Support vectors
- Feature space

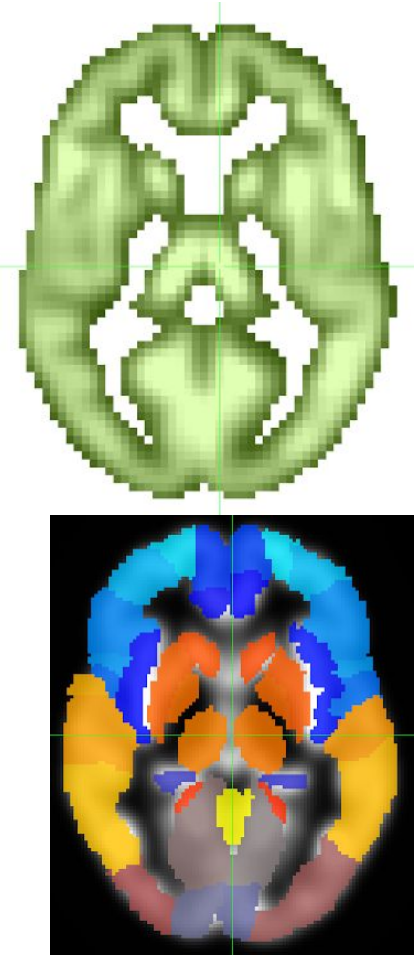


Voxel-based analysis

Vectors

- Voxel-based
    - Size of the vector managed by down-sampling and interpolation
- Region of interest
- Big data
    - The large size of images and the big number of images need low level programming

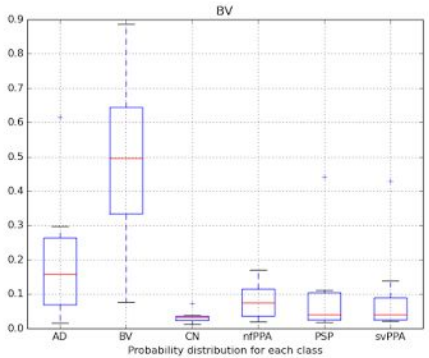


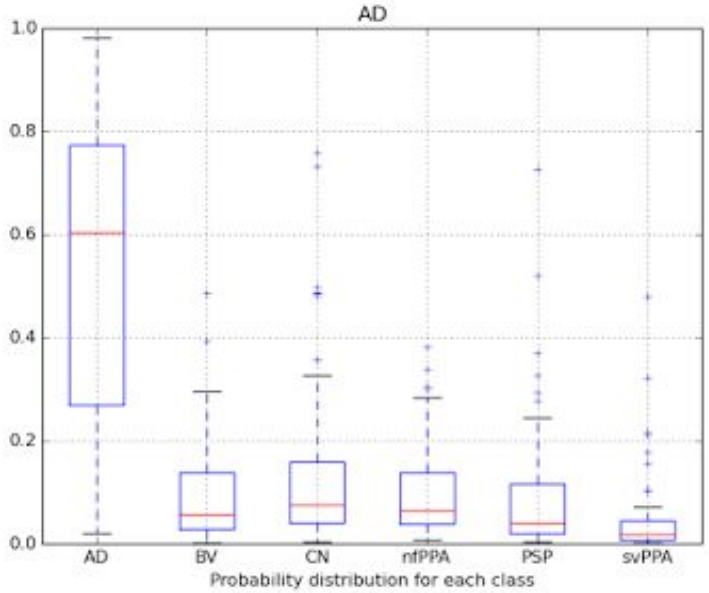


Support Vector Machine

Multi-class classification

- binary to multi-class classification
  - Probability score (J.C. Platt)
  - Multi-class (T. Hastie and R. Tibshirani)





Wrap up
– John

What next?

- Compare results across algorithms
- Validate on 1.5T scans – generalizability to multiple scanner field strengths
- Validate relative to different Dx proportions
- Validate on external data from another center – eventually multiple centers
- Add other clinical information: blood tests, neuropsyche testing etc.

New research directions?

- Use additional spatial information
- Consider sub-classes (unsupervised and unknown number)
- Incorporate longitudinal scans – clinically improve diagnostic accuracy through follow-up

Summary

- Automated volumetric MRI diagnosis is a lofty but important goal
- SVM on ROIs or down-sampled images can provide a feasible way to develop image-based classifiers on large datasets
- Multi-diagnosis classifiers with class probabilities can be developed – tool for clinicians
- **Thanks to Yann for all the hard work!**
- **And thanks for listening!**