Hypothesis Generation by Interactive Visual Exploration of Heterogeneous Medical Data

Cagatay Turkay, Arvid Lundervold ,
Astri Johansen Lundervold, Helwig Hauser

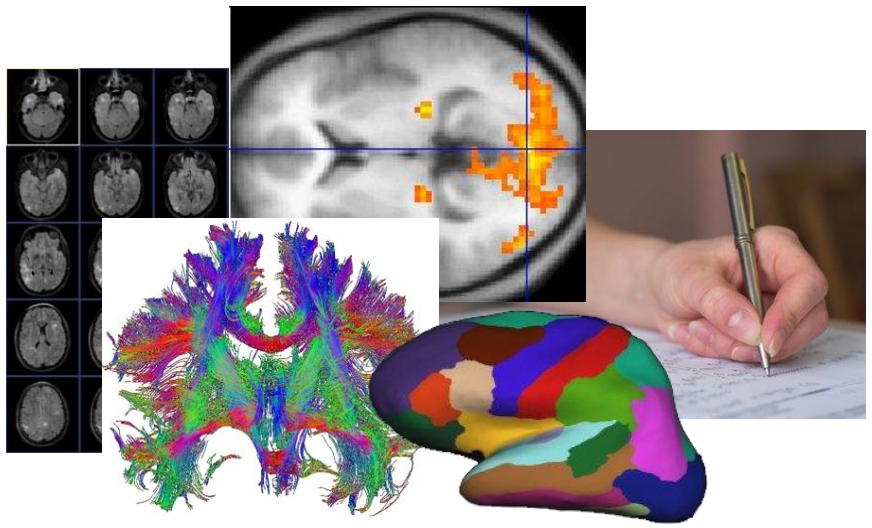
UNIVERSITY OF BERGEN

What you will hear today?

- **Interactive & visual** methods in data analysis
  - **Dual analysis** approach
- Deal with **complex** datasets
  - Many variables
  - Heterogeneous
  - Several modalities
- **Generating hypotheses** interactively
- Analyze medical data as a **multidisciplinary** group

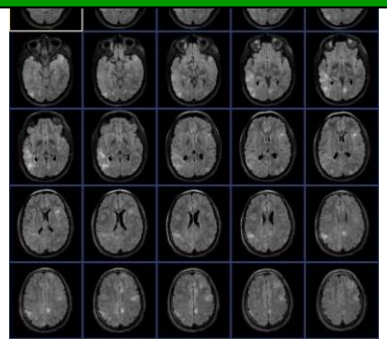
Problem Domain: Cognitive Aging Study Analysis

- Carried out by neuropsychology & biomedicine experts
- Analyze relations between **brain segments** vs. **cognitive decline**
- **Heterogeneous**: image statistics + test scores + patient data
  - Imaging modalities, MRI, DTI, fMRI
  - Neuropsychological examination: IQ, memory function, and attention/executive function
- **Longitudinal** study, 3 waves (**2005**, 2009, 2012)
  - ~100 participants



Cognitive Aging Study Data

MR Imaging



Anatomical Segmentation

-- 45 brain **segments,** e.g., *cerebellum, white matter, …*

-- 7 features for each segment e.g., *number of voxels , volume, …*

+

Personal/Clinical Data

+

Neuropsychological Examination

2D data table

82 *X* 373

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
ID	BirthDate	Gender	Education	cvlt_tri_tc	cvlt_tri_tc	cvlt_sd_fr	cvlt_sd_fr	cvlt_ld_fr	cvlt_ld_fr	cvlt_rec_h	cvlt_rec_h	cvlt_rec_f	cvlt_rec_f	cvlt_tot_d	cvlt_tot_d	cwi_1_col	cwi_1_col	cwi_2_wo
501	1951	2	14	70	73	15	1.5	16	1.5	16	0.5	0	-1	4	1.5	30	10	22
507	1948	2	15	50	51	11	0	9	-1	15	0	0	-1	3.7	1	33	9	20
508	1947	1	15	52	64	11	1	10	0.5	15	0.5	1	-1	3.4	1	27	12	21
510	1956	1	11	56	62	12	1	12	1	13	-0.5	4	0.5	2	-0.5	37	7	28
512	1935	2	17	61	68	15	1.5	16	2	16	0.5	0	-1	4	1.5	34	10	26
514	1930	1	18	40	54	9	0.5	10	1	15	0.5	1	-1	3.4	1.5	37	9	20
517	1949	1	13	54	60	13	1.5	12	1	15	0.5	5	0.5	2.5	0	24	13	20
518	1942	1	19	45	56	9	0	11	1	14	0	1	-1	3	0.5	38	8	20
519	1944	2	14	61	65	13	1	14	1	16	0.5	1	-0.5	3.7	1	38	8	24
520	1946	1	18	33	41	6	-0.5	7	-0.5	11	-2	3	0	1.8	-1	29	12	24
523	1946	1	19	75	87	16	2.5	16	2.5	16	1	0	-1	4	2	29	12	20
524	1950	2	16	63	66	15	1.5	14	1	16	0.5	0	-1	4	1.5	26	12	21
526	1945	1	18	59	71	12	1	11	1	15	0.5	3	0	2.9	0.5	28	12	22
527	1945	2	12	63	66	15	2	14	1	14	-1	1	-0.5	3	0	27	12	18
529	1942	2	19	65	68	14	1.5	13	0.5	14	-1	2	0	2.7	-0.5	28	12	19
530	1949	1	12	50	56	10	0.5	10	0	15	0.5	0	-1	3.7	1.5	24	13	21
532	1949	1	15													34	8	25
533	1948	2	9	58	62	12	1	12	0.5	15	0	0	-1	3.7	1	26	13	21
537	1951	2	15	61	64	12	0.5	16	1.5	16	0.5	0	-1	4	1.5	27	11	20
538	1956	2	16	66	69	15	1.5	16	1.5	16	0.5	0	-1	4	1.5	28	11	19
539	1950	1	15	62	68	15	2	16	2	16	1	0	-1	4	1.5	37	7	26
542	1954	2	15	58	61	11	0	13	0.5	16	0.5	1	-0.5	3.7	1	35	8	21
543	1946	2	18	65	68	15	2	16	1.5	16	0.5	1	-0.5	3.7	1	35	9	23
544	1945	2	12	51	55	10	0	12	0.5	15	0	0	-1	1.7	1	30	11	22
545	1952	2	15	62	65	12	0.5	13	0.5	14	-0.5	0	-1	3.3	0.5	24	13	21
547	1940	2	17	66	70	16	2	16	1.5	15	0	0	-1	3.7	1	24	14	18
549	1951	2	12	75	80	16	2	16	1.5	16	0.5	0	-1	4	1.5	22	13	19
551	1938	2	10	58	62	15	2	15	1.5	15	0	1	-0.5	3.4	0.5	27	12	26
555	1947	1	16	49	60	11	1	14	2	16	1	0	-1	4	2	42	6	29
556	1932	1	15	32	45	6	-0.5	6	-0.5	11	-2	5	0.5	1.5	-1	35	9	23
558	1933	2	18	61	68	14	1.5	14	1	16	0.5	0	-1	4	1.5	32	11	23
559	1944	1	16	50	62	7	-0.5	9	0	14	0	4	0	2.3	-0.5	19	16	17
560	1951	2	20	56	58	10	-0.5	12	0	16	0.5	0	-1	4	1.5	33	9	21
564	1935	2	10	67	74	11	0.5	13	1	15	0	1	-0.5	3.4	0.5			
566	1945	2	10	53	57	13	1	12	0.5	16	0.5	0	-1	4	1.5	38	8	27
567	1949	2	14	69	72	14	1	15	1.5	16	0.5	0	-1	4	1.5	34	8	22
569	1945	2	10	57	61	12	1	13	0.5	16	0.5	8	2	2.5	-1	38	8	30
573	1958	2	16	70	73	15	1.5	15	1.5	16	0.5	0	-1	4	1.5	23	13	19
577	1958	2	16	60	63	16	2	16	1.5	16	0.5	0	-1	4	1.5	26	11	19
579	1957	2	14	65	68	15	1.5	16	1.5	16	0.5	0	-1	4	1.5	27	11	17
581	1935	2	11	56	63	15	1.5	15	1.5	15	0	0	-1	3.7	1	28	12	20
583	1943	1	13	53	65	12	1	13	1.5	16	1	0	-1	4	2	35	9	26
586	1947	2	19	74	79	16	2	16	1.5	16	0.5	0	-1	4	1.5	32	10	19

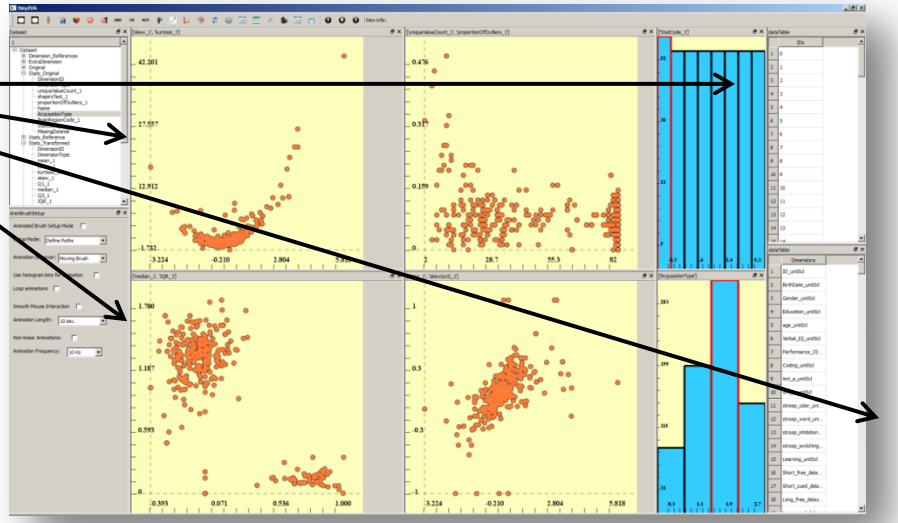
Problems in the analysis process

- **Slow** analysis pipeline
- Analysis **limited to a priori hypothesis,** i.e., already published research
- Relating **different types** of data (variables) is challenging
- Work on a **subset** of data at each **iteration** of the analysis, lose the **overall picture**
- Computational tools are often **black-boxes**

Interactive Visual Analysis Methods (In a Nutshell)

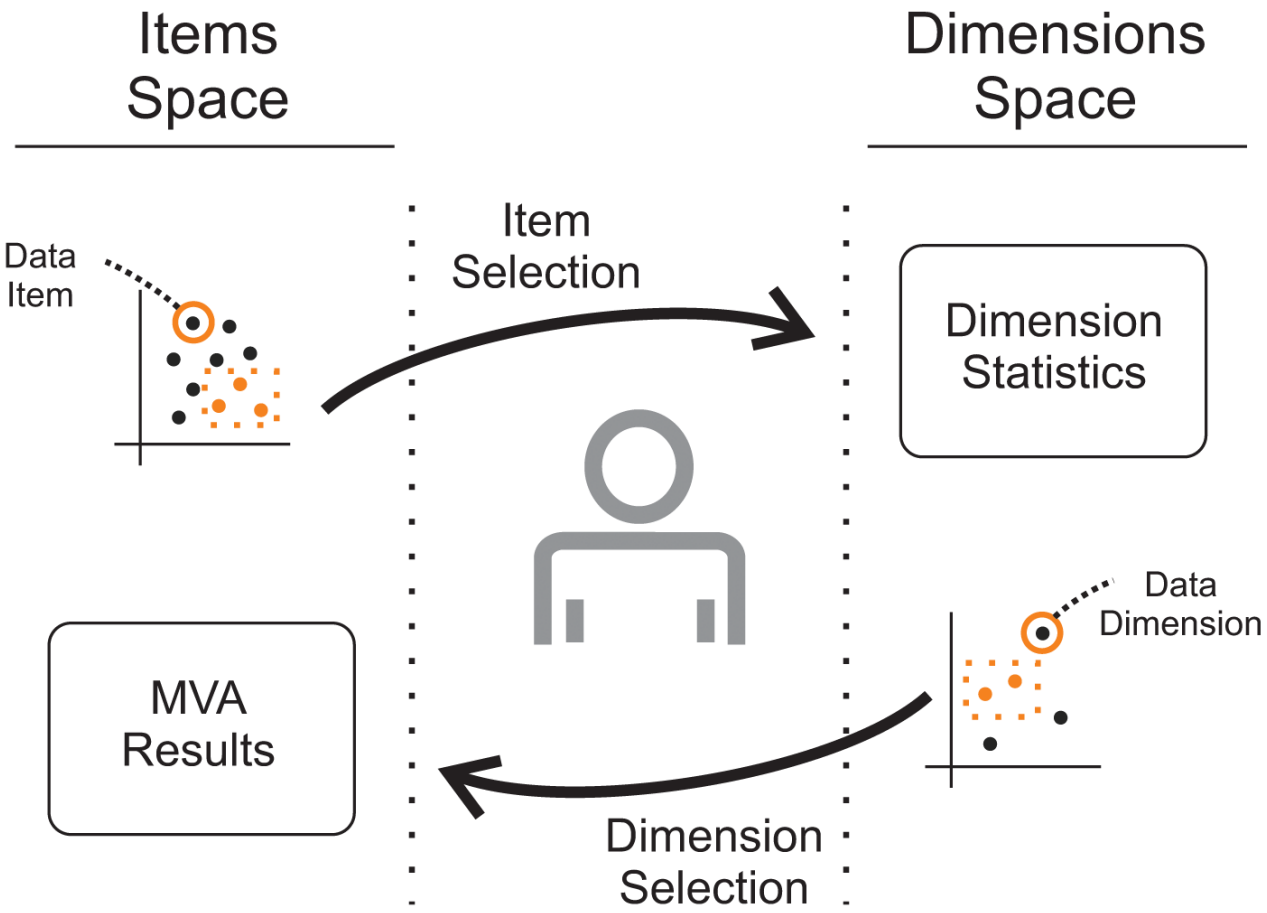
- **Multiple** visualizations of data
- Selections denoted as **focus + context**
- **Linked** selections within views
- **Integrated** use of computational tools
  - "R for Statistical Computing"
  - PCA, MDS, Clustering, Regression, etc…

Different views



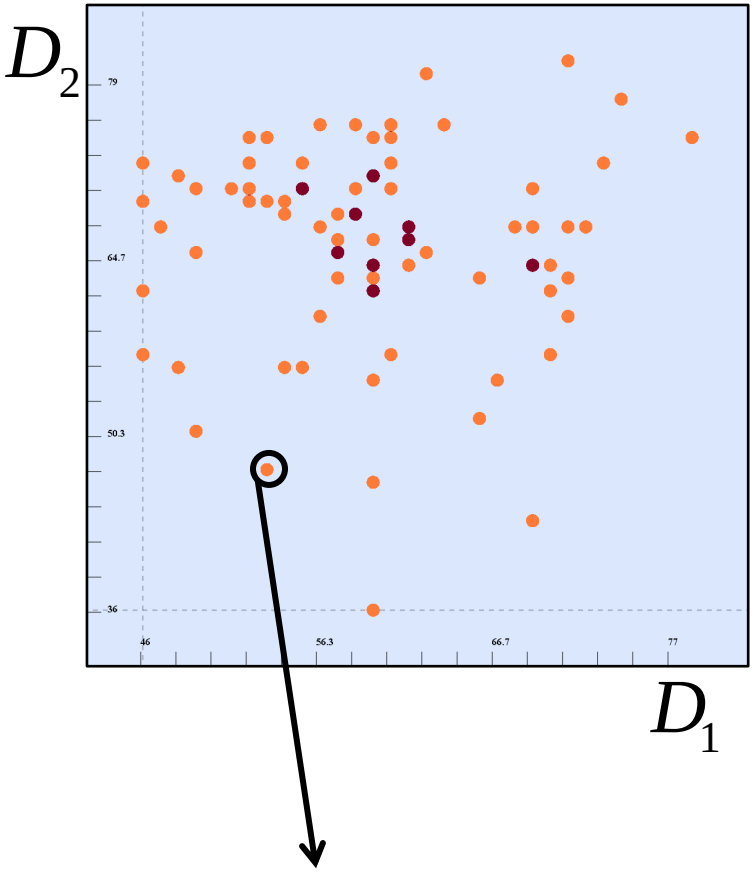
Dual Analysis Method

- Treat variables as **first-order analysis objects**
- Interactive visual analysis in **two linked spaces**



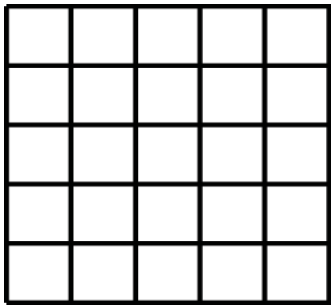
Dual Analysis Method

Items

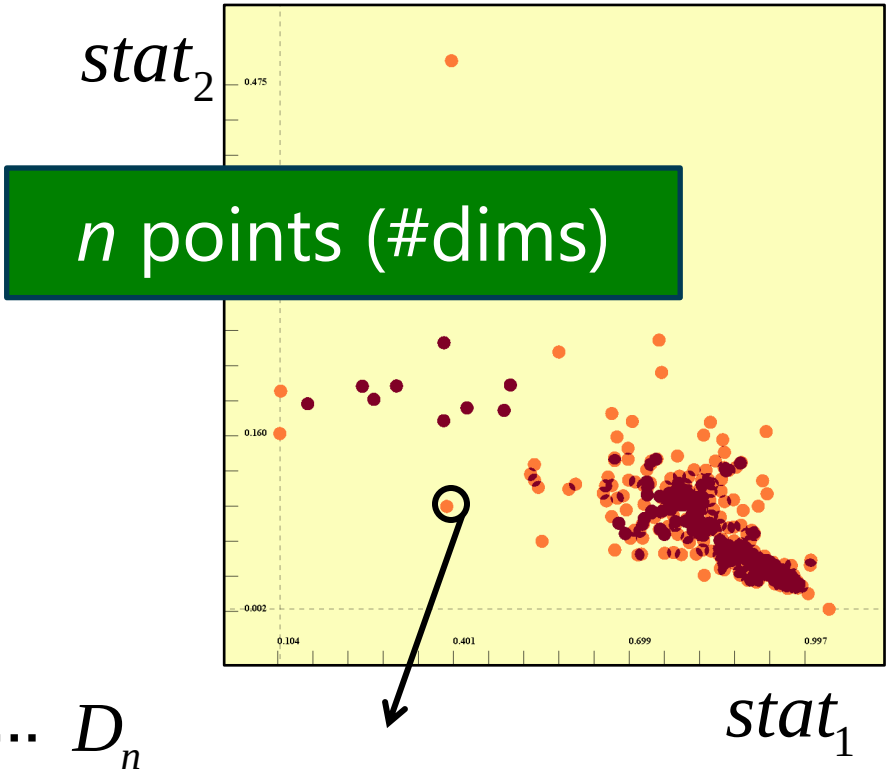


A single data item

D_1 D_2 ... D_n



Variables



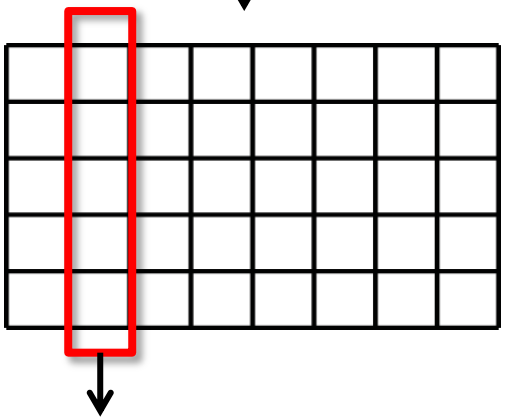
n points (#dims)

A single variable

Visualizations in the dimensions space

- Dimensions are the main **visual entities !!**

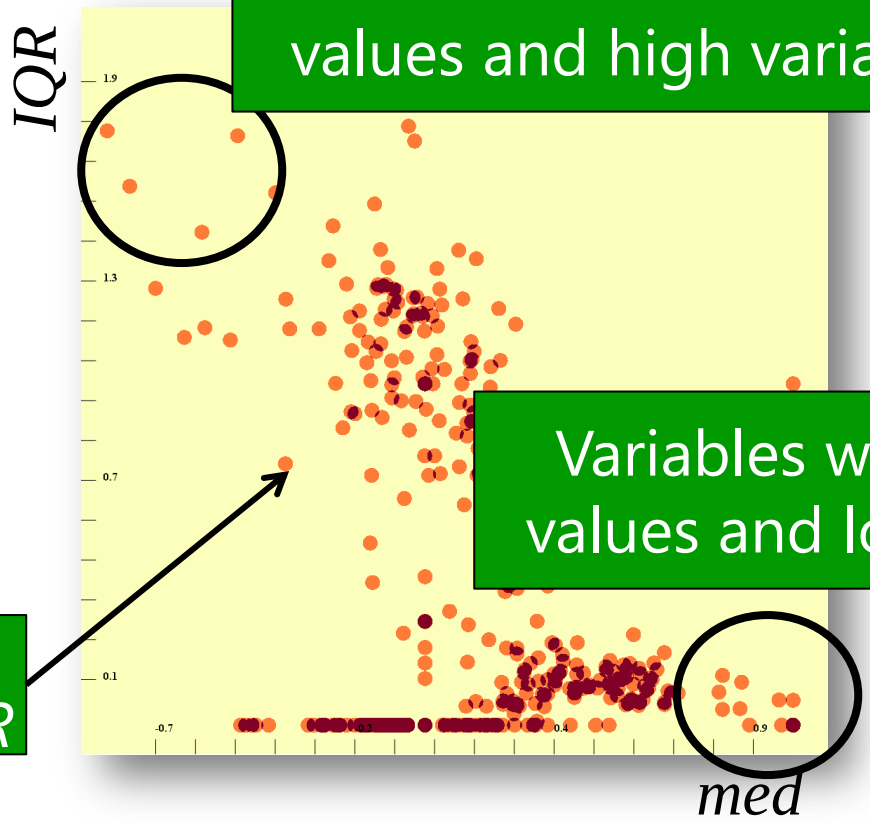
Normalize data first



For each column, compute *med* and *IQR*

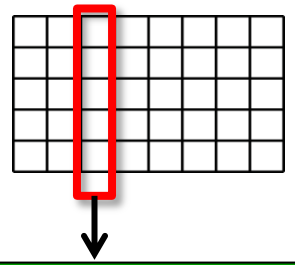
Variables with smaller values and high variance

Variables with higher values and low variance

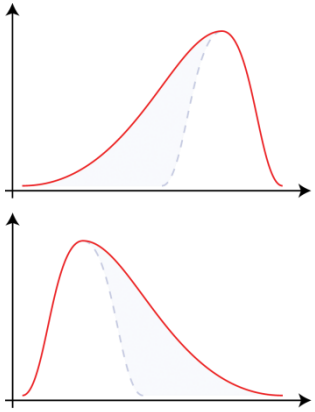


Rich statistics set = rich analysis

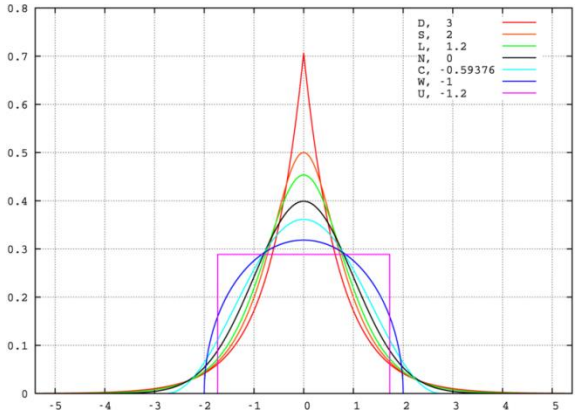
- Different statistics for different insights
  - **Descriptive** statistics, e.g., skewness, kurtosis
  - **Robust** statistics: e.g., median, IQR, etc.
  - Distribution **test scores**, e.g., normality
  - **Correlation** relations
  - …
- Include also the **meta-data**



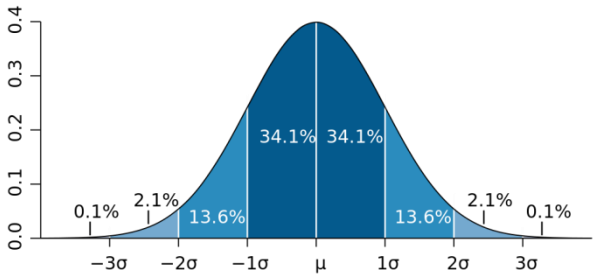
For each column, compute *k* statistics



Skewness



Kurtosis

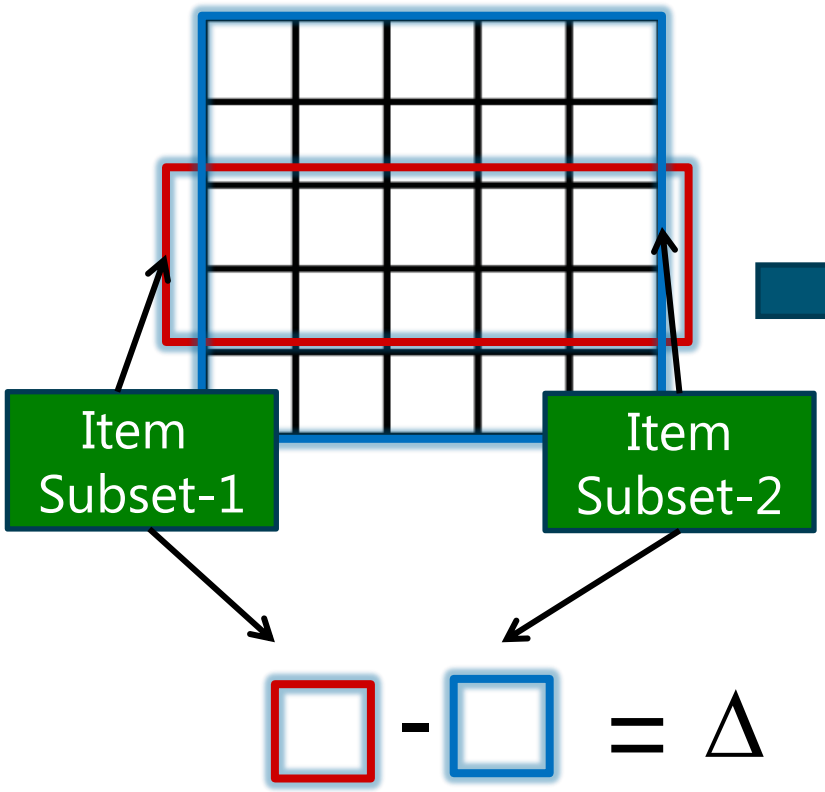


Normality

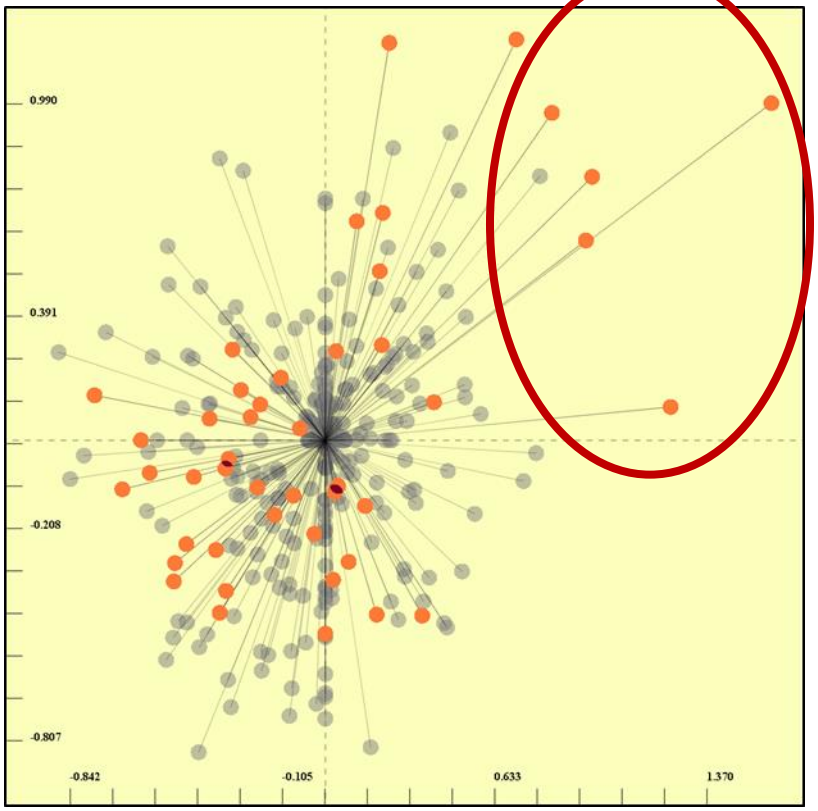
Deviation Plot

Compute "μ" & "α" values using two subsets of items

Higher values for the selection



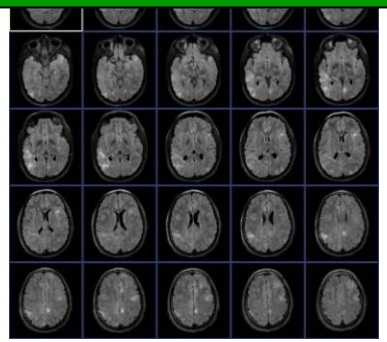
Change in "α" values

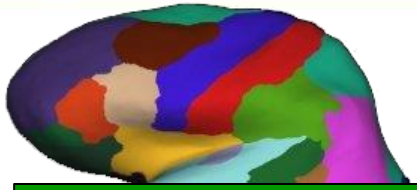


Change in "μ" values

Cognitive Aging Study Data

MR Imaging





Anatomical Segmentation

-- 45 brain **segments,** e.g., *cerebellum, white matter, …*

-- 7 features for each segment e.g., *number of voxels , volume, …*

+

Personal/Clinical Data

+

Neuropsychological Examination

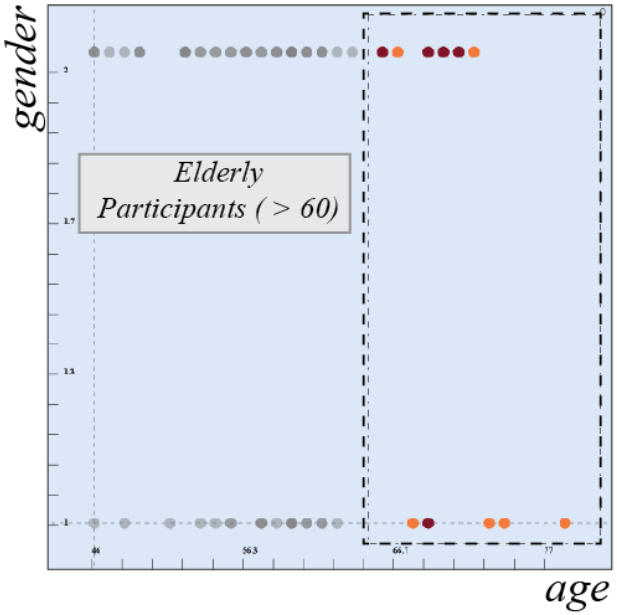
2D data table

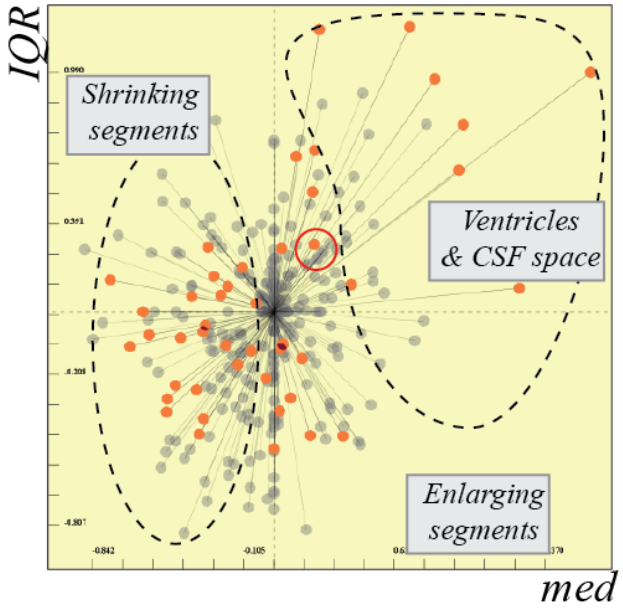
82 *X* 373

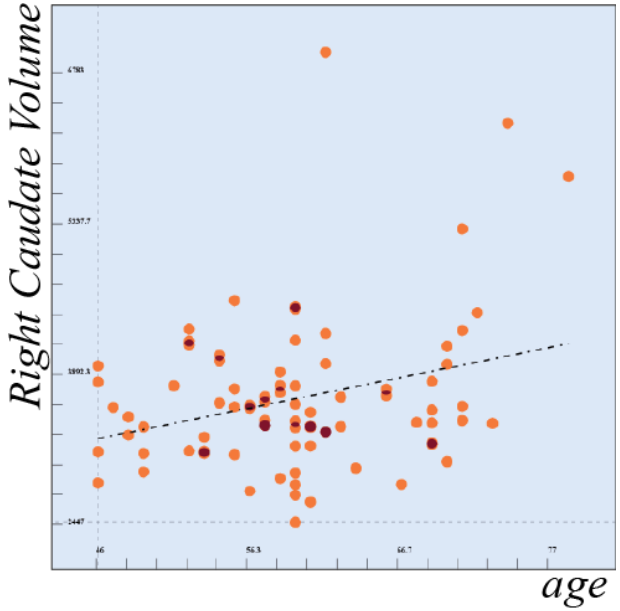
Analysis Process

- Generate new hypotheses **exploratively**
  - **Data-driven** process
  - Consider **a priori** expert **knowledge**
- Use **meta-data** on dimensions to steer analysis
  - Dependent / independent variables
- **5 hypotheses** in short sessions
  - Inter-relations in Test Results
  - Findings Based on Sex
  - **Findings Based on Age**
  - IQ & Memory Function vs. Brain Segment Volumes
  - **Relations within Brain Segments**

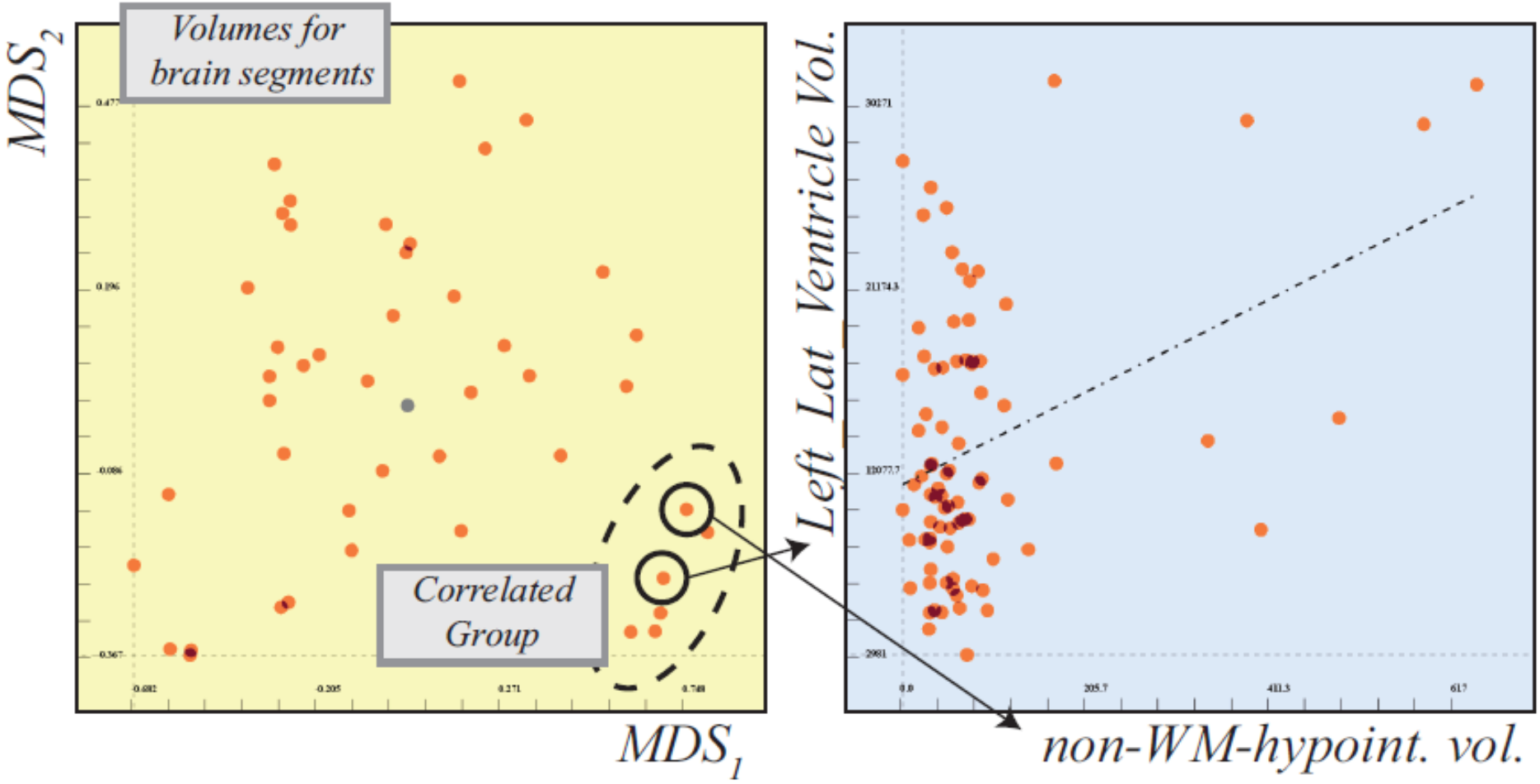
Findings Based on Age







Relations within Brain Segments



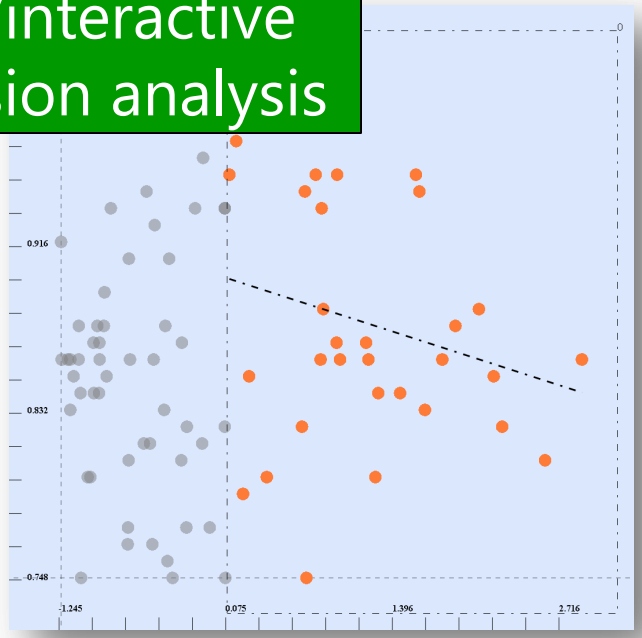
Observations & Limitations

- No need for limitations on a priori knowledge
- **Whole data** available along the analysis
- Change in working routine !
  - **Hypothesis driven analysis** to **hypothesis generation**
- Quickly **check for known** hypotheses – **data quality**?
- **Learning curve**? Understanding of statistics
- **Overfitting** to data / non-optimal solutions

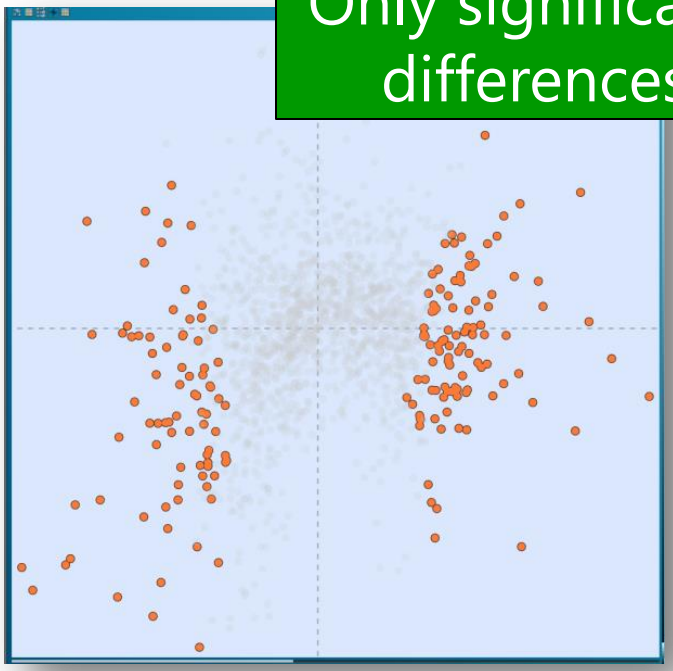
Lessons Learned (for the future)

- Need to incorporate **robust** methods / tools
- Enable more **accurate** readings
- Reduce **false positives**
- Improve **usability** & **visual guidance**

Local/interactive regression analysis



Only significant differences



Conclusions

- **Applicable/generalizable** methods to data from other scientific fields
- **Interactive** use of computational tools, **more reliable, easier to interpret**
- Quick hypotheses generation, **prototyping** ideas
  - Then use **robust (slow)** methods if necessary
- Sweet spot between "**hypothesis-driven**" & "**data-driven**" science

Acknowledgments

- Peter Filzmoser, TU Wien
- Julius Parulek, VisGroup @ UIB
- VisGroup @ UIB