Some comments on results reporting in neuroimaging

Overview

- Brief overview of the data shown here
- Example 0: statistics *and* effect estimates
- Example 1: better thesholding
- Example 2: peak voxel issues
- Example 3: information content and data “digestibility”
- Example 4: improving cross study comparisons
- Conclusions

Example 0:
Statistics *and* effect estimates

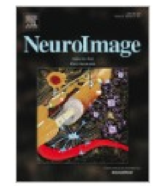
Are statistics the only results?

Some neuroimaging software packages only provide statistics from FMRI modeling. Wherever possible, AFNI provides both the effect estimate *and* the statistic. This raises the question:

ELSEVIER

NeuroImage

Volume 147, 15 February 2017, Pages 952-959



Is the statistic value all we should care about in neuroimaging?

Gang Chen, Paul A. Taylor, Robert W. Cox

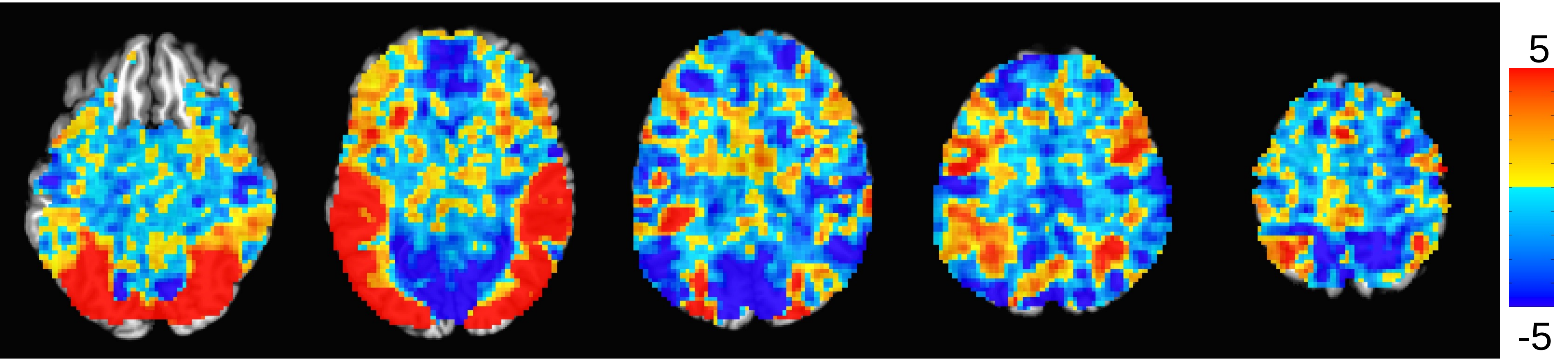
Abstract

Here we address an important issue that has been embedded within the neuroimaging community for a long time: the absence of effect estimates in results reporting in the literature. The statistic value itself, as a dimensionless measure, does not provide information on the biophysical interpretation of a study, and it certainly does not represent the whole picture of a study. Unfortunately, in contrast to standard practice in most scientific fields, effect (or amplitude) estimates are usually not provided in most results reporting in the current neuroimaging publications and presentations. Possible

Are statistics the only results?

Example: FT subj, vis stimulus

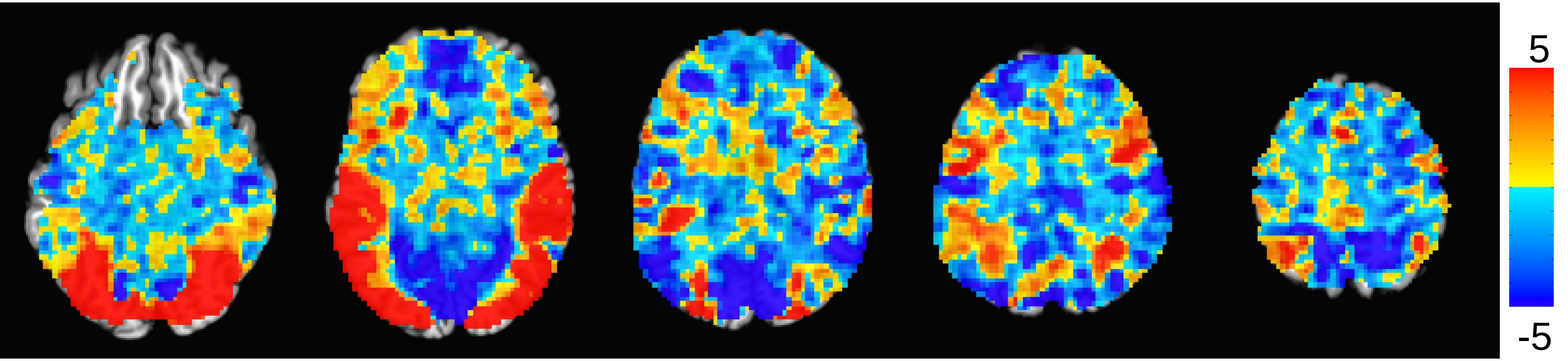
- Statistics image (t-stat, DF = 412):



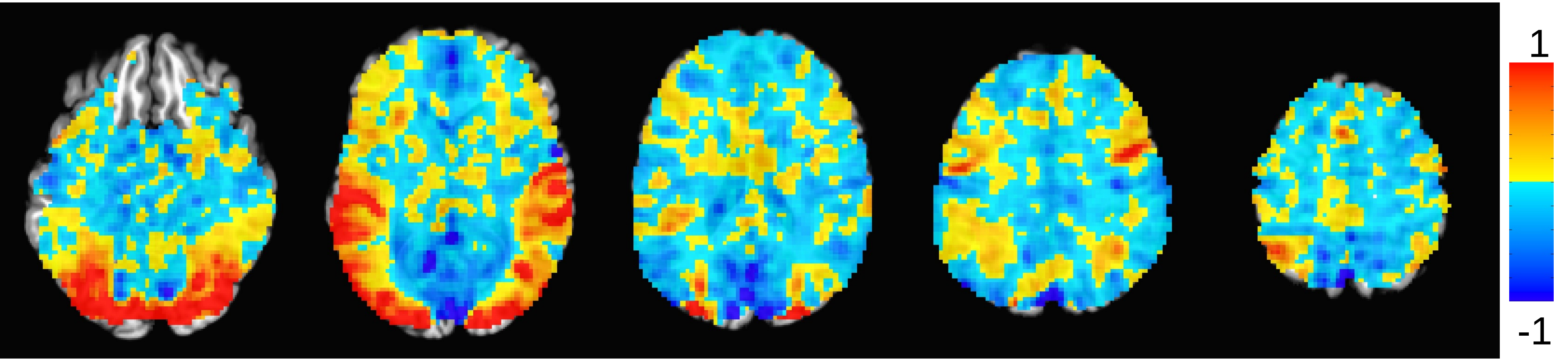
Are statistics the only results?

Example: FT subj, vis stimulus

- Statistics image (t-stat, DF = 412):



- Effect estimate image (BOLD % signal change):



Are statistics the only results?

- **TL;DR:** No, stats are *not* the only results.
- Effect estimates (or "point estimates", "betas", "coefs") are the physical response evidence
  → these have units like "BOLD % signal change"
- Statistics are the reliability or accuracy of estimate
  → no units; e.g., t is ratio of effect to uncertainty σ

Are statistics the only results?

- **TL;DR:** No, stats are *not* the only results.

- Effect estimates (or "point estimates", "betas", "coefs") are the physical response evidence
  → these have units like "BOLD % signal change"
- Statistics are the reliability or accuracy of estimate
  → no units; e.g., t is ratio of effect to uncertainty σ

- Consider difference between "statistical" and "practical" significance: need effect estimates to compare about latter

- Looking at effect estimates aids reproducibility comparisons
- Seeing effect estimates helps validate modeling/find problems

Are statistics the only results?

We *can* include both effects and stats in plots (and using each for what they are good at), so shouldn’t we do so?

- **effect estimate as overlay:** show size of effects
- **statistic as threshold:** highlight regions with higher reliability

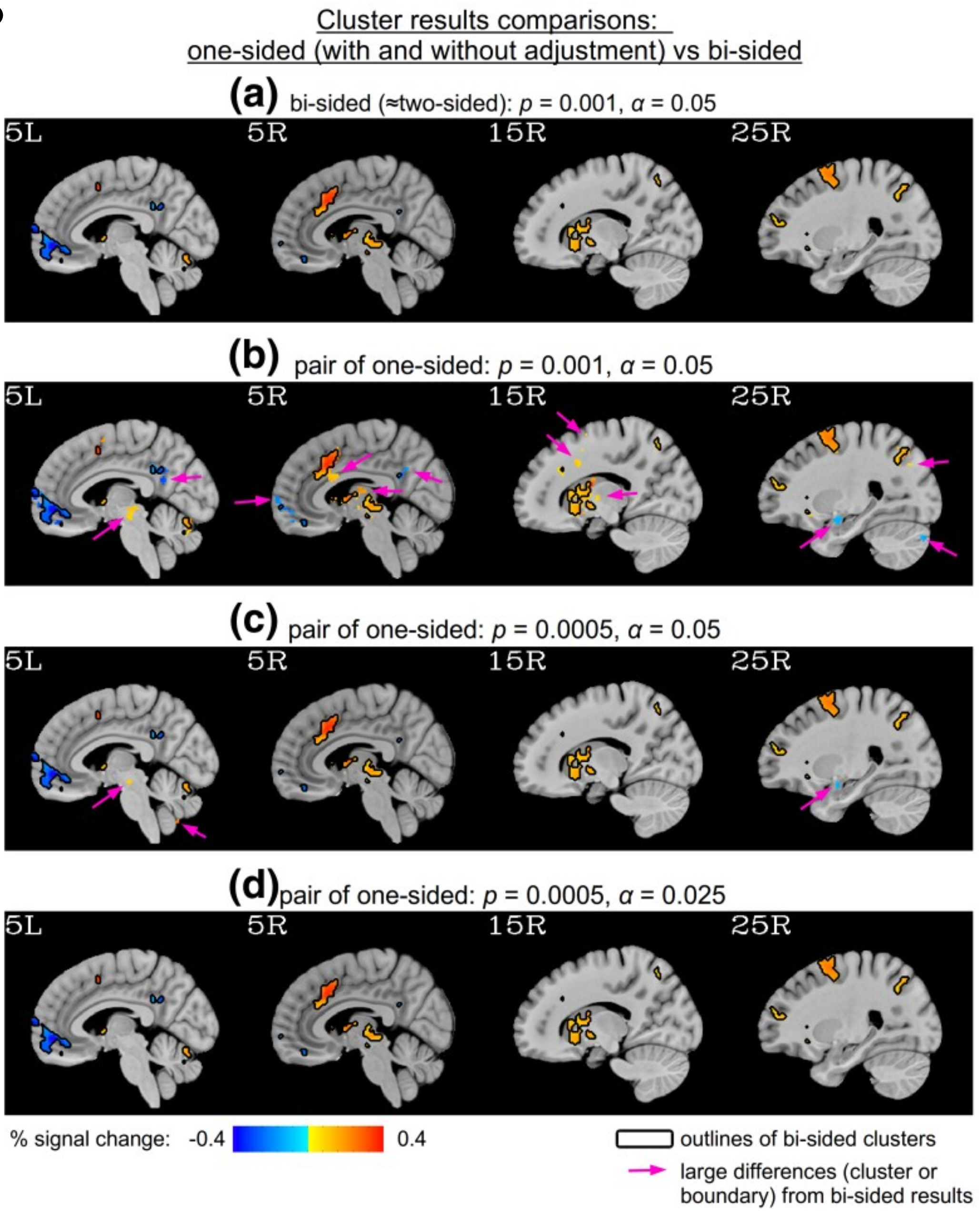
When reporting statistics, also include fundamental information:

- How many degrees of freedom?
- Were tests 1-sided or 2-sided?

Side(dness) note

Were tests 1-sided or 2-sided?

"One-sided t-tests are widely used in neuroimaging data analysis. While such a test may be applicable when investigating specific regions and prior information about directionality is present, we argue here that it is often mis-applied, with severe consequences for false positive rate (FPR) control."



Chen G, Cox RW, Glen DR, Rajendra JK, Reynolds RC, Taylor PA (2019). A tail of two sides: Artificially doubled false positive rates in neuroimaging due to the sidedness choice with t-tests. HBM 40:1037-1043.
https://pubmed.ncbi.nlm.nih.gov/30265768/

Data description for other examples

Example Data + Processing Summary

- **Data:** from NARPS project (Botvinik-Nezer et al., 2020)
  - 2 groups, each with 54 subj: 4 EPI runs (2 mm), 1 T1w anatomical
  - task-based FMRI: mixed gambling paradigm, with responses

nature > articles > article

Article | Published: 20 May 2020

Variability in the analysis of a single neuroimaging dataset by many teams

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M. W. J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, … Tom Schonberg + Show authors

Nature **582**, 84–88 (2020) | Cite this article

Example Data + Processing Summary

- **Group-level analysis here:** processing with AFNI (Cox, 1996) and afni_proc.py pipeline:
  - voxelwise analysis: nonlinear alignment to template, 4 mm blur, motion censoring, amplitude modulation in regr. model
  - also performed separate ROI-based analysis
  - details: *https://pubmed.ncbi.nlm.nih.gov/37116766/*
  - scripts: *https://github.com/afni/apaper_highlight_narps*

NeuroImage 274 (2023) 120138

Highlight results, don’t hide them: Enhance interpretation, reduce biases and improve reproducibility

Paul A. Taylor[a,*], Richard C. Reynolds[a], Vince Calhoun[b], Javier Gonzalez-Castillo[c], Daniel A. Handwerker[c], Peter A. Bandettini[c,d], Amanda F. Mejia[e], Gang Chen[a]

Example Data + Processing Summary

- **Group-level analysis here:** processing with AFNI (Cox, 1996) and afni_proc.py pipeline:
  - voxelwise analysis: nonlinear alignment to template, 4 mm blur, motion censoring, amplitude modulation in regr. model
  - also performed separate ROI-based analysis
  - details: *https://pubmed.ncbi.nlm.nih.gov/37116766/*
  - scripts: *https://github.com/afni/apaper_highlight_narps*

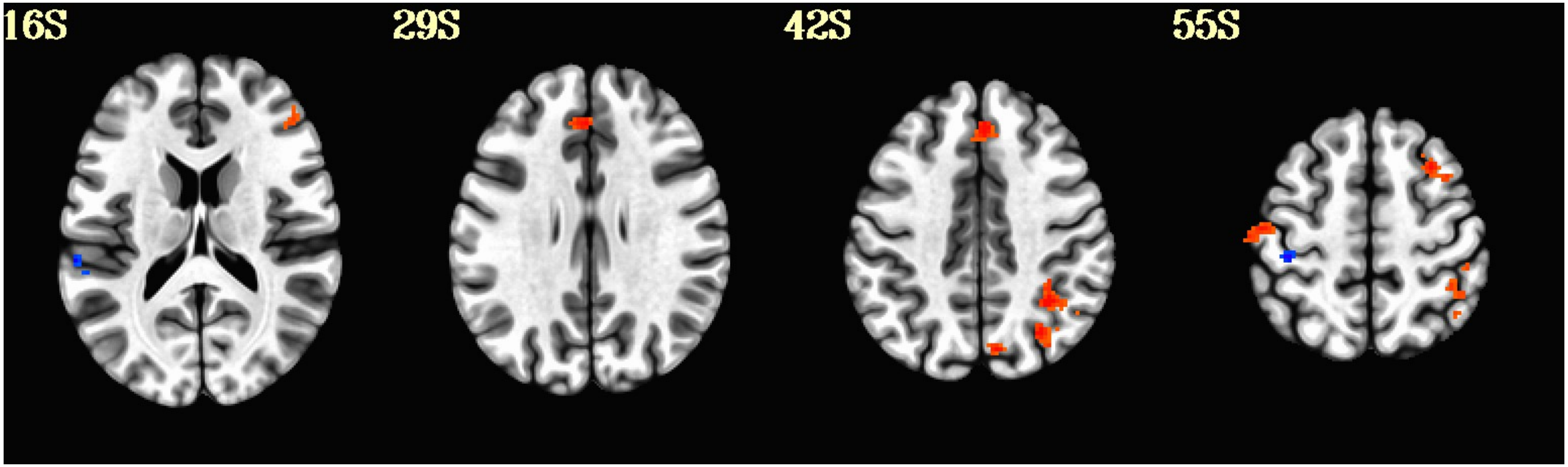
- **Cross-study comparisons:** from original submissions to NARPS
  - ~70 teams analyzed same data voxelwise, however each wanted
  - A) teams answered yes/no questions about certain hypotheses per group (specific ROIs, contrasts, directionality)
  - B) teams uploaded unthresholded stat maps (no effect maps ☹), available on NeuroVault (one link per team, see NARPS paper)

Example 1:
Better thesholding

What are “the results” of a study?

- Consider group-level results from a standard voxelwise analysis:
  - Threshold at voxelwise p = 0.001
  - Cluster-based threshold (multiple comparison adjustment) at FWE=5%

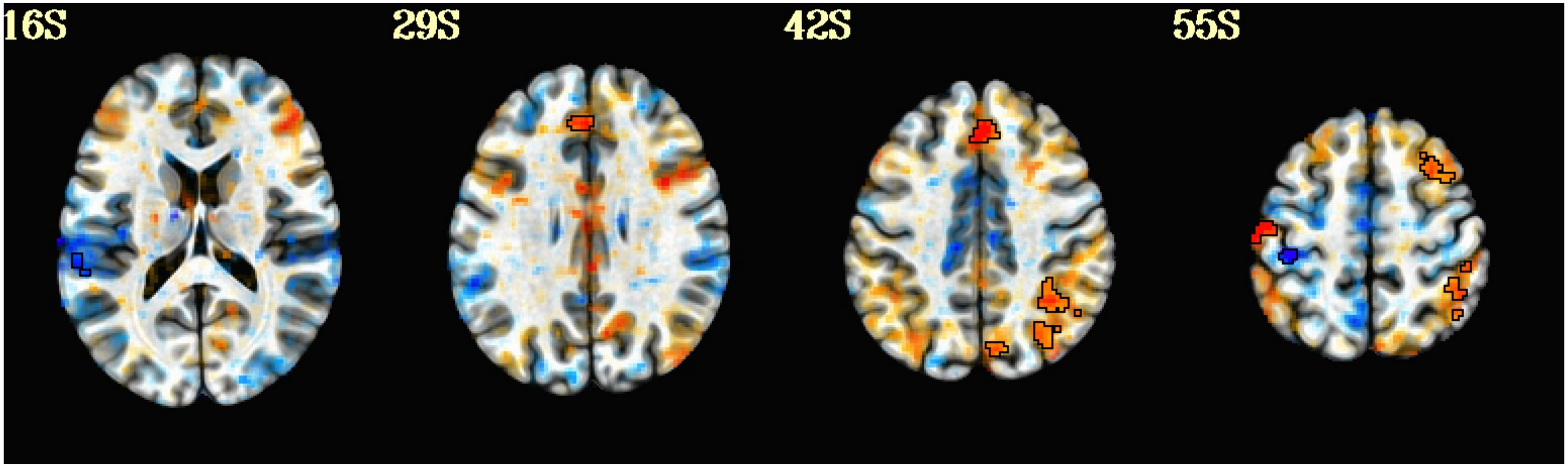
standard thresholding: info at suprathreshold, hide elsewhere



What are "the results" of a study?

- Consider group-level results from a standard voxelwise analysis:
  - Threshold at voxelwise *p* = 0.001
  - Cluster-based threshold (multiple comparison adjustment) at FWE=5%

*transparent thresholding: highlight suprathreshold, info everywhere**

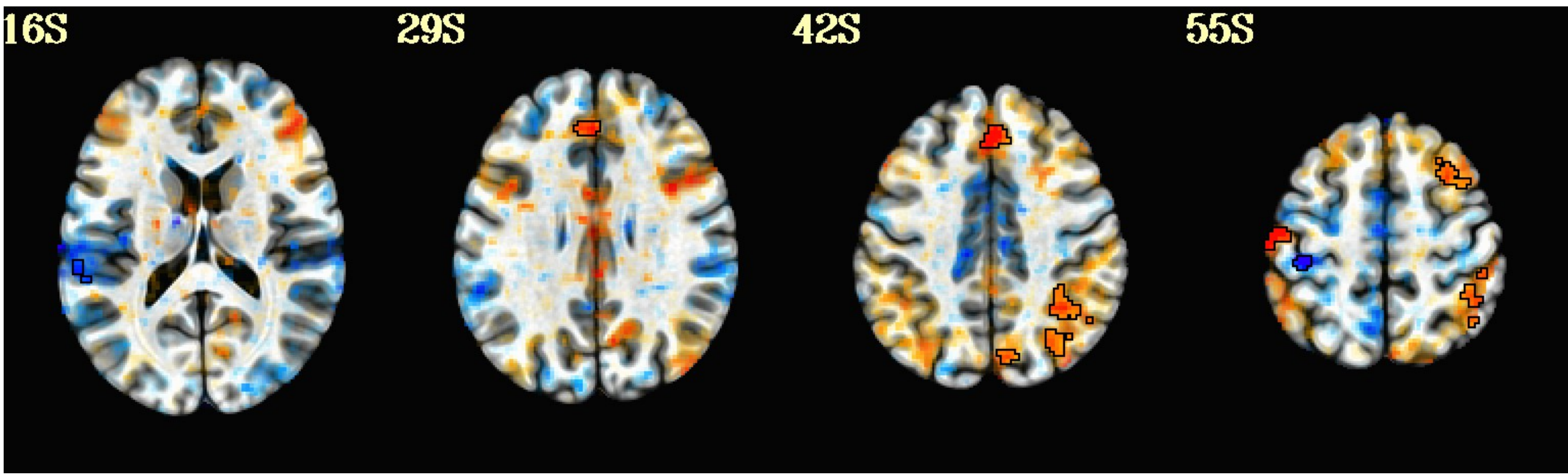


**overlay more transparent as statistic decreases (Allen et al., 2012)*

What are "the results" of a study?

- Consider group-level results from a standard voxelwise analysis:
  - Threshold at voxelwise *p* = 0.001
  - Cluster-based threshold (multiple comparison adjustment) at FWE=5%

transparent thresholding: highlight suprathreshold, info everywhere

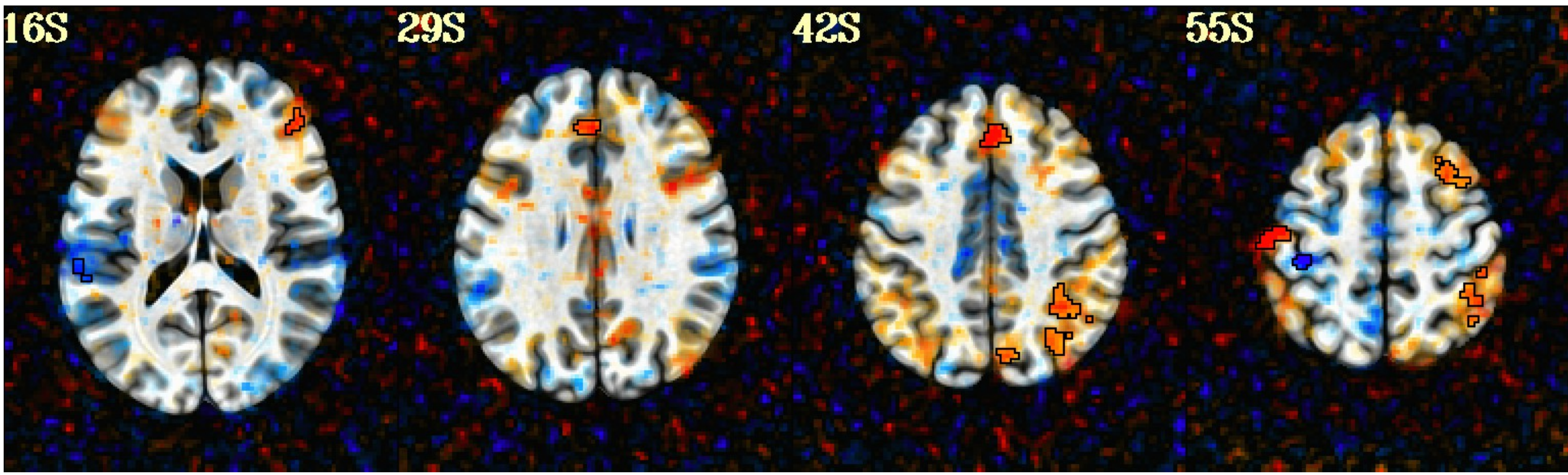


→ Once you see these more full results, does the standard image seem adequate?

What are "the results" of a study?

- Consider group-level results from a standard voxelwise analysis:
  - Threshold at voxelwise p = 0.001
  - Cluster-based threshold (multiple comparison adjustment) at FWE=5%

transparent thresholding: highlight suprathreshold, info ***everywhere***

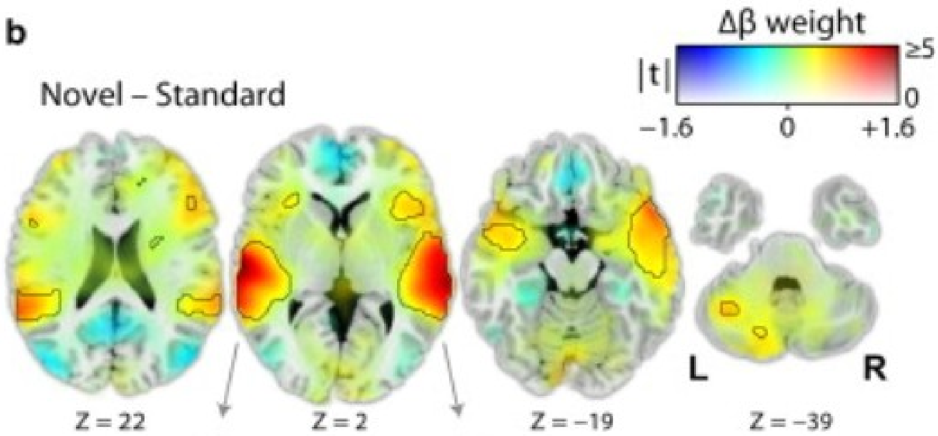


→ Once you see these more full results, does the standard image seem adequate?

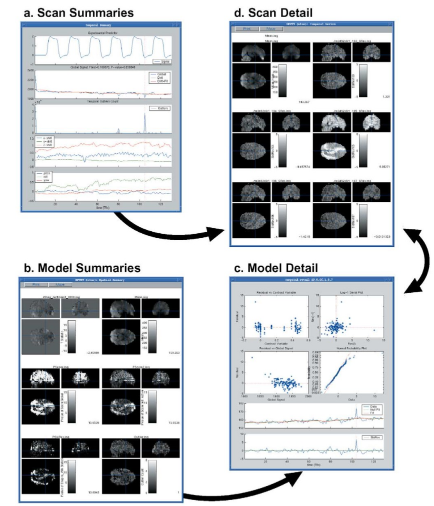
→ And why not see **everywhere**? Data were gathered in full FOV, and this helps quality control, avoiding artifacts, etc.

What are “the results” of a study?

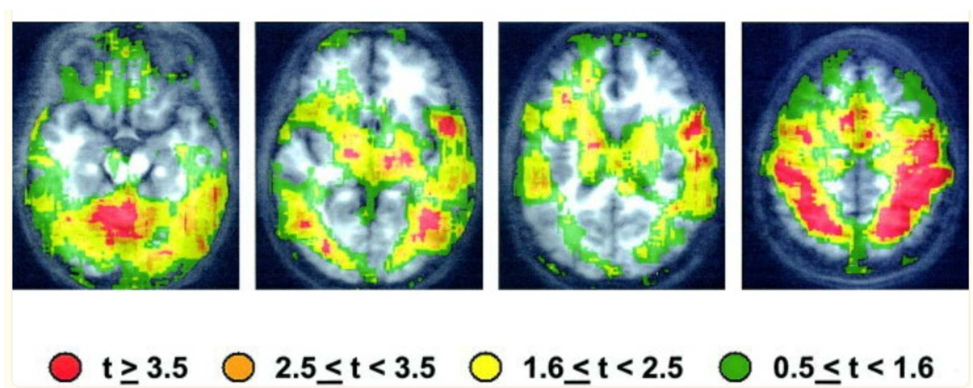
- NB: several previous neuroimaging works to "show more results", including:



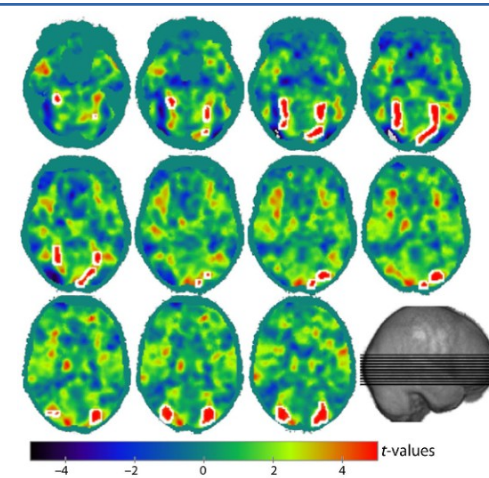
Transparent thresholds (Allen et al., 2012)



Model summaries (Luo & Nichols, 2003)



Color-range bands (Jernigan et al., 2003)

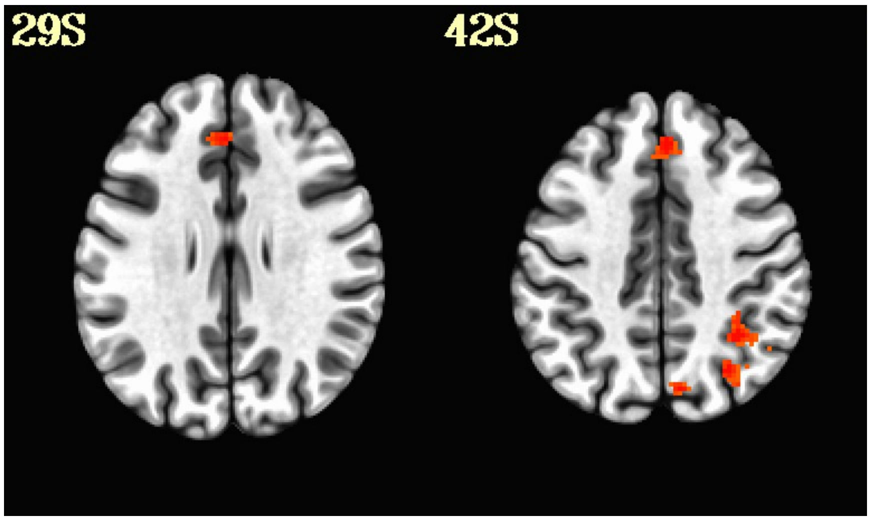


No threshold (Pernet & Madan, 2020)

What are "the results" of a study?

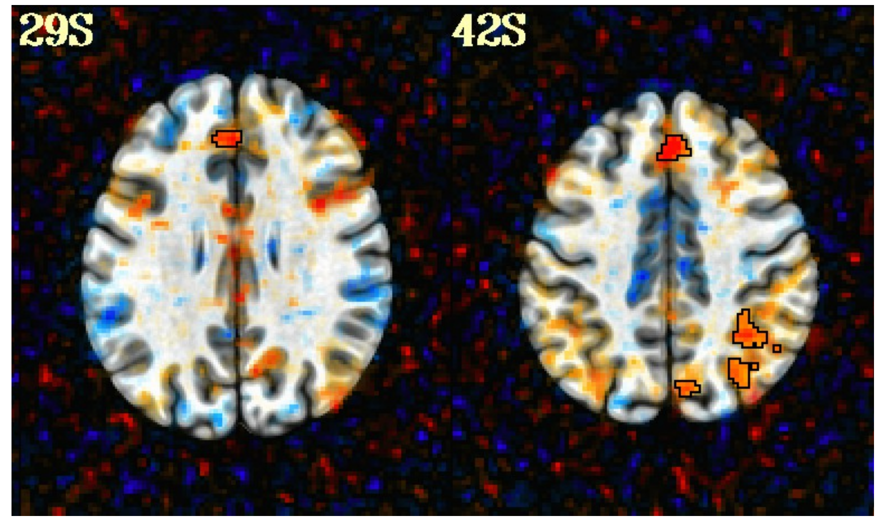
How you visualize results shapes interpretation and understanding, for both you and readers: *highlight, don't hide.*

all-or-nothing thresholding



- strong dependence on thr value
- biases due to thresholding
- waste information
- hinder interpretation
- unrealistic view of noise
- hide poor modeling
- harm reproducibility

transparent thresholding



- more consistent with biology
- more scientific (fuller results)
- assist quality control
- deeper comparisons
- less influence of arbitrariness
- provide more evidence
- improve meta analyses

Example 2:
Peak voxel issues

Peak voxel issues

How do cluster results typically get presented?
It is common to use: A) peak (statistic) voxel location:

Cluster	Peak Voxel (mm, RAI-Dicom)			
	x	y	z	ROI location (dist)
1	-43.0	71.0	37.0	R_Area_PGs
2	41.0	27.0	61.0	L_Primary_Sensory_Cortex (1 mm)
3	-3.0	-33.0	43.0	R_Area_8BM
4	-23.0	-19.0	49.0	R_Area_6m_anterior
5	51.0	13.0	57.0	L_Area_1 (1 mm)
6	-9.0	75.0	45.0	R_Parieto-Occ*_Sulcus_Area_2
7	17.0	105.0	-9.0	L_Primary_Vis*_Cortex (5 mm)
8	-49.0	35.0	57.0	R_Area_PFm_Complex
9	69.0	35.0	25.0	L_Area_PF_Complex (2 mm)
10	-55.0	37.0	-19.0	R_Area_TE2_anterior
11	-3.0	35.0	35.0	R_Area_dorsal_23_a+b

Peak voxel issues

How do cluster results typically get presented?
It is common to use: B) the cluster center of mass:

Cluster	Center of Mass Voxel (mm, RAI-Dicom)			
	x	y	z	ROI location (dist)
1	-39.4	59.1	44.2	R_Area_PFm_Complex
2	38.2	26.2	63.6	L_Area_1
3	0.7	-34.3	37.8	L_Area_8BM
4	-28.7	-15.9	51.7	R_Area_6m_anterior
5	51.8	16.5	53.9	L_Area_1
6	-8.2	74.1	44.3	R_Parieto-Occ*_Sulcus_Area_2
7	14.5	101.6	-8.4	L_Primary_Vis*_Cortex (1 mm)
8	-51.4	32.5	50.1	R_Area_IntraParietal_2 (1 mm)
9	58.3	33.7	21.4	L_PeriSylvian_Lang*_Area (1 mm)
10	-58.0	40.5	-13.5	R_Area_TE2_anterior
11	-1.2	34.3	35.2	R_Area_dorsal_23_a+b

Peak voxel issues

How do volumetric results typically get presented?

Note how the atlas attribution differs (gray lines) with either form of single-voxel summary:

Cluster	Peak Voxel (mm, RAI-Dicom)				Center of Mass Voxel (mm, RAI-Dicom)			
	x	y	z	ROI location (dist)	x	y	z	ROI location (dist)
1	-43.0	71.0	37.0	R_Area_PGs	-39.4	59.1	44.2	R_Area_PFm_Complex
2	41.0	27.0	61.0	L_Primary_Sensory_Cortex (1 mm)	38.2	26.2	63.6	L_Area_1
3	-3.0	-33.0	43.0	R_Area_8BM	0.7	-34.3	37.8	L_Area_8BM
4	-23.0	-19.0	49.0	R_Area_6m_anterior	-28.7	-15.9	51.7	R_Area_6m_anterior
5	51.0	13.0	57.0	L_Area_1 (1 mm)	51.8	16.5	53.9	L_Area_1
6	-9.0	75.0	45.0	R_Parieto-Occ*_Sulcus_Area_2	-8.2	74.1	44.3	R_Parieto-Occ*_Sulcus_Area_2
7	17.0	105.0	-9.0	L_Primary_Vis*_Cortex (5 mm)	14.5	101.6	-8.4	L_Primary_Vis*_Cortex (1 mm)
8	-49.0	35.0	57.0	R_Area_PFm_Complex	-51.4	32.5	50.1	R_Area_IntraParietal_2 (1 mm)
9	69.0	35.0	25.0	L_Area_PF_Complex (2 mm)	58.3	33.7	21.4	L_PeriSylvian_Lang*_Area (1 mm)
10	-55.0	37.0	-19.0	R_Area_TE2_anterior	-58.0	40.5	-13.5	R_Area_TE2_anterior
11	-3.0	35.0	35.0	R_Area_dorsal_23_a+b	-1.2	34.3	35.2	R_Area_dorsal_23_a+b

- Peak statistic locations will be sensitive and unstable to noise levels, to adding/subtracting subjects, as well as to minor acquisition and/or processing choices. Reproducibility…?(!!??)
- Also, what if clusters overlap multiple atlas regions?

→ ***Is there a better way to summarize clusters?***

Peak voxel issues

How ~~do~~*could* volumetric results typically get presented?

Perhaps a table of overlaps (say, >10%),

… and can even show sub-thr clusters of interest:

Cluster	Size	Overlap	ROI location
Primary clusters (p=0.001)			
1	551	19.5%	R_Area_PFm_Complex
		15.7%	R_Area_IntraParietal_1
		13.3%	R_Area_PGs
2	189	30.1%	L_Primary_Motor_Cortex
		28.8%	L_Primary_Sensory_Cortex
3	163	31.0%	R_Area_8BM
		24.3%	L_Area_dorsal_32
4	114	39.9%	R_Inferior_6-8_Transitional_Area
		19.5%	R_Area_8Av
		17.9%	R_Area_6_anterior
5	70	46.8%	L_Area_1
		34.2%	L_Primary_Sensory_Cortex
6	69	80.5%	R_Parieto-Occipital_Sulcus_Area_2
7	64	62.4%	L_Primary_Visual_Cortex
8	64	41.3%	R_Area_IntraParietal_2
9	62	32.3%	L_Area_PF_Complex
		25.0%	L_Area_PFcm
		16.5%	L_Area_PF_opercular
10	52	66.8%	R_Area_TE1_posterior
		18.8%	R_Area_TE2_anterior
11	40	40.8%	R_Area_dorsal_23_a+b
		32.6%	L_Area_dorsal_23_a+b

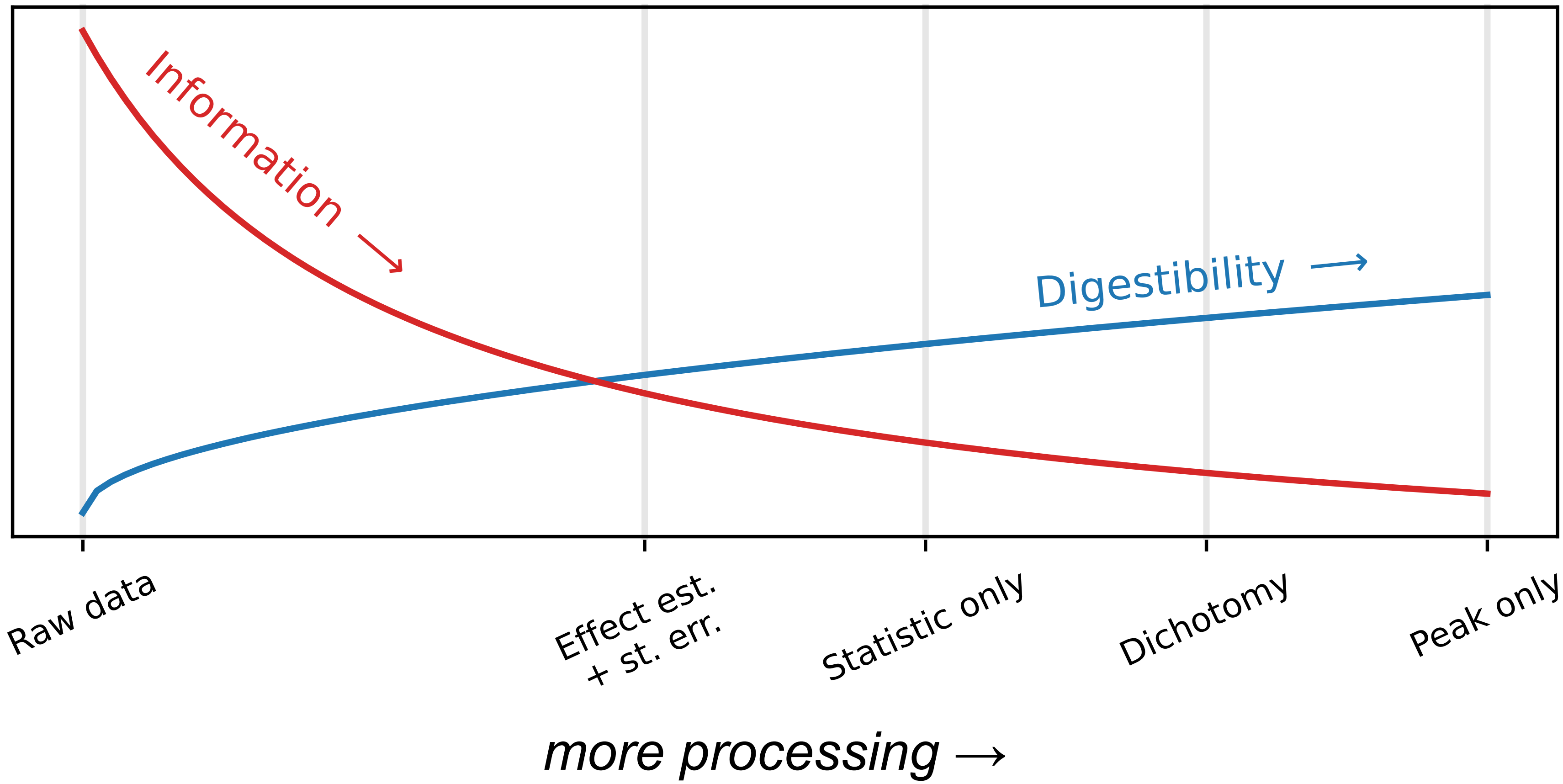
Additional clusters of interest (p=0.01)			
A-7	279	24.0%	R_Rostral_Area_6
		17.4%	R_Area_IFJa
		16.0%	R_Area_IFJp
		12.1%	R_Area_8C
A-27	77	31.4%	L_Area_8C
		26.6%	L_Area_IFJp
		17.5%	L_Area_IFJa
		14.8%	L_Rostral_Area_6

Example 3:
Information content
and data “digestibility”

- Much of this example's discussion from Chen et al. (2022):

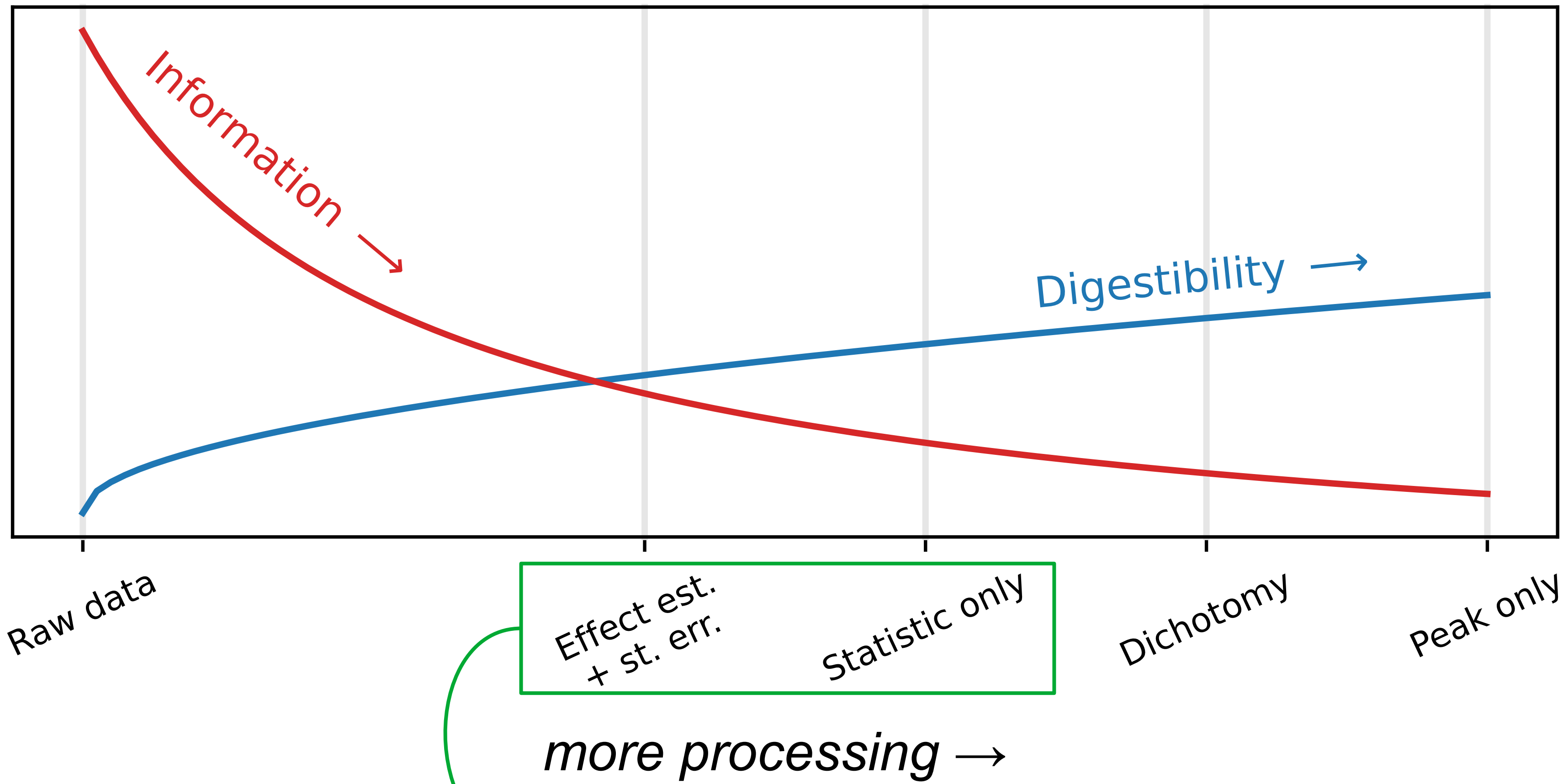
Neuroimaging analysis as information extraction

With data processing and analysis, there is a trade-off between:
information reduction and ease of interpretability.



Neuroimaging analysis as information extraction

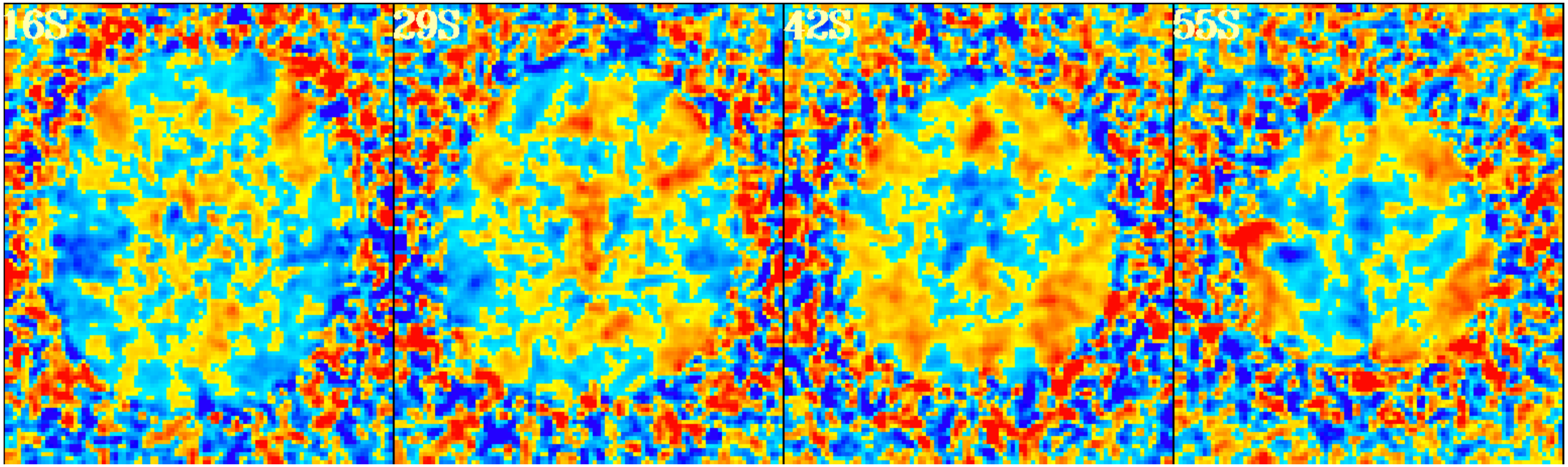
With data processing and analysis, there is a trade-off between:
information reduction and ease of interpretability.

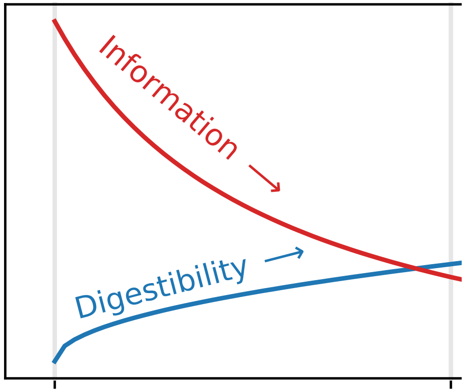


Sidenote: for this issue, see *Is the statistic value all we should care about in neuroimaging?* Chen et al. (2017)

How much should we reduce data in figures?

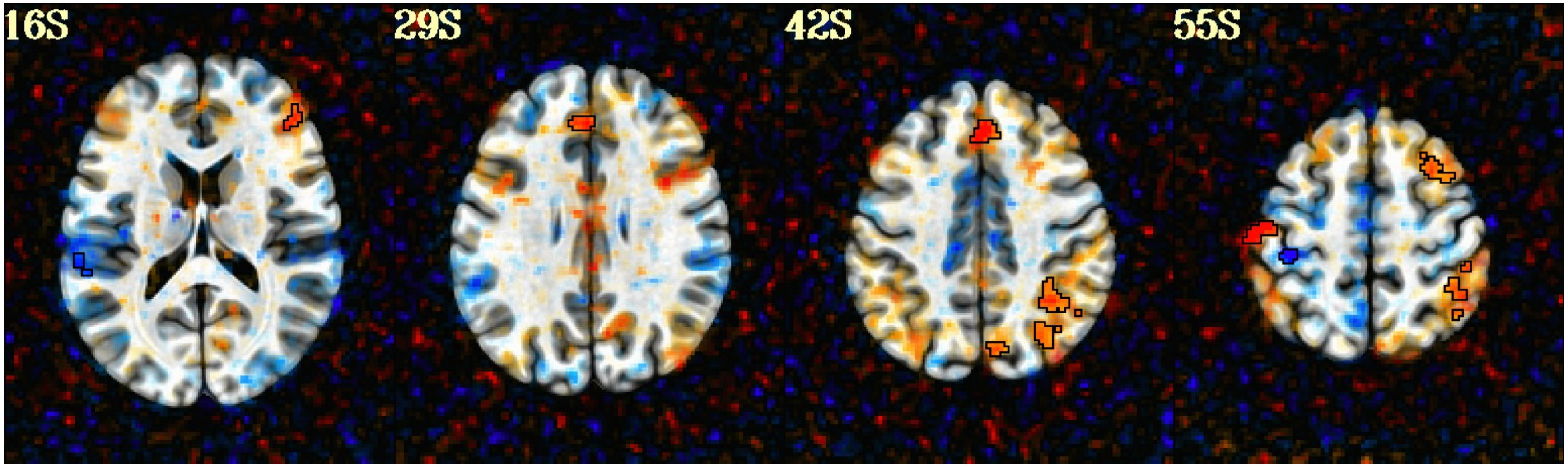
Group-level results:
unthresholded

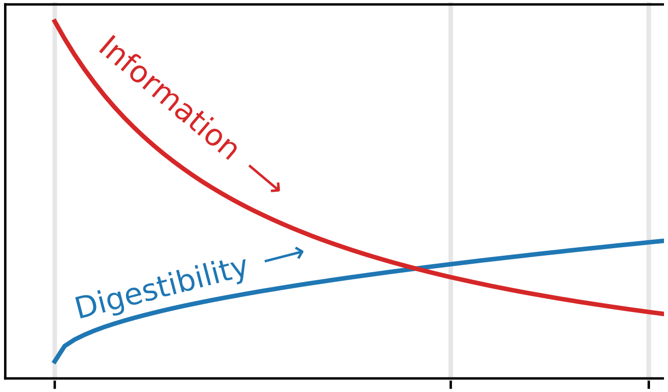




How much should we reduce data in figures?

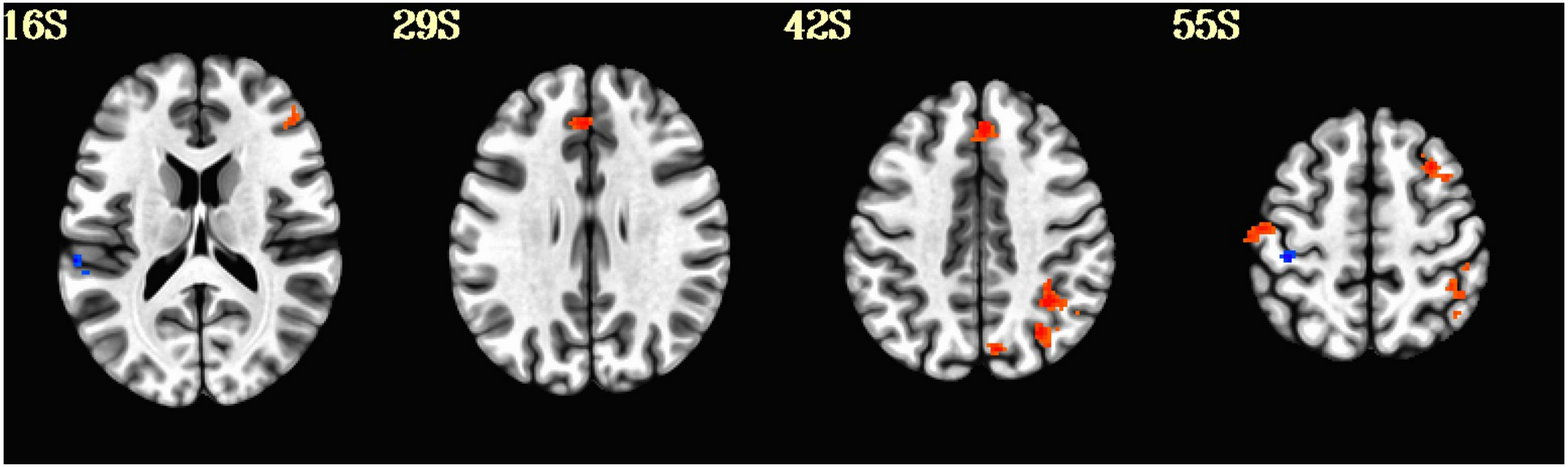
Group-level results:
transparent threshold (p = 0.001, FWE = 5%)

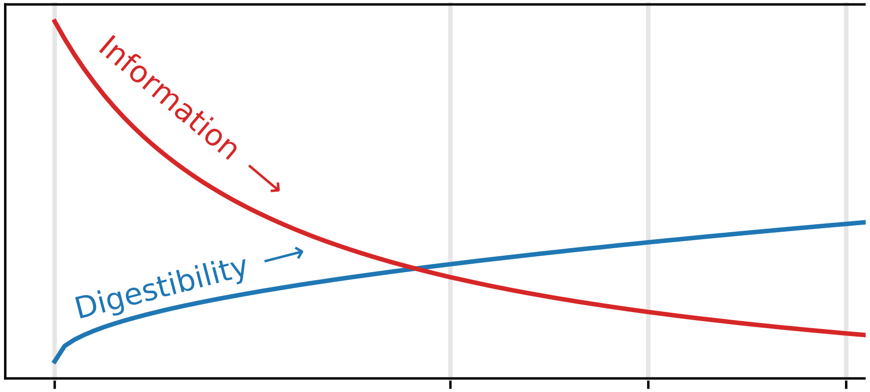




How much should we reduce data in figures?

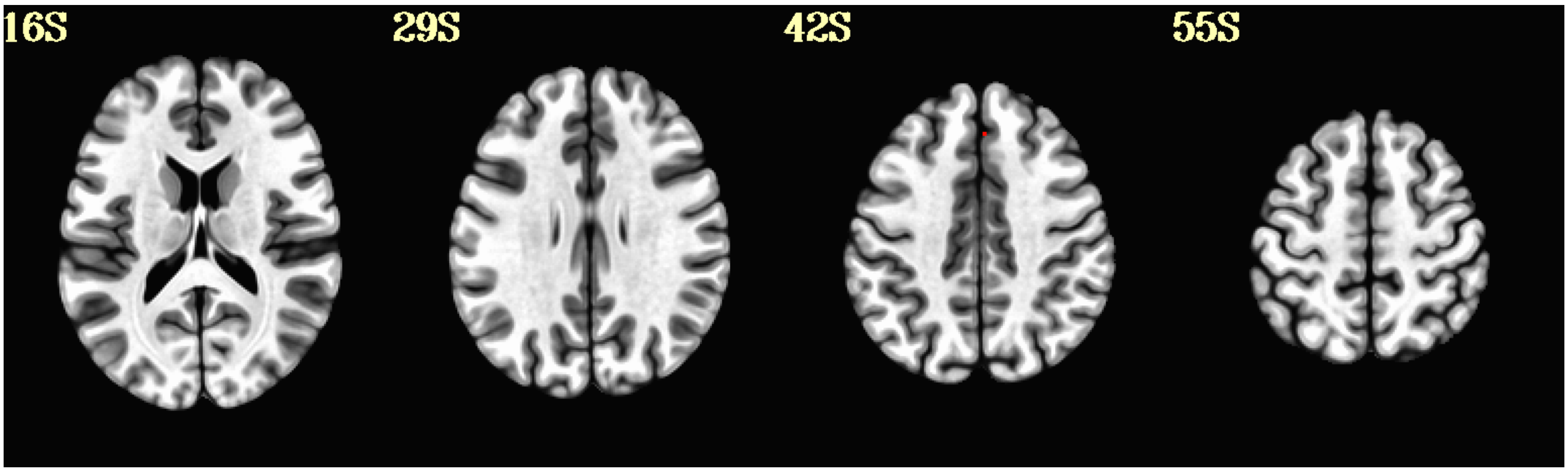
Group-level results:
all-or-nothing threshold (p = 0.001, FWE = 5%)

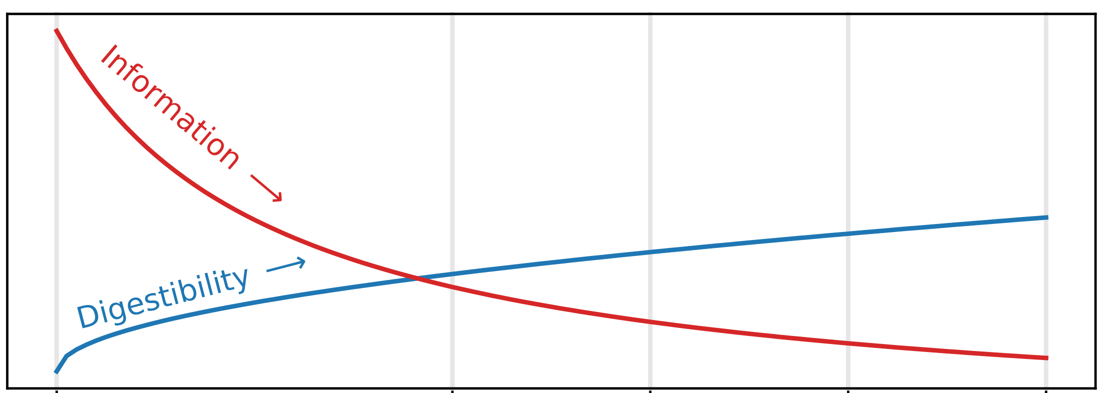




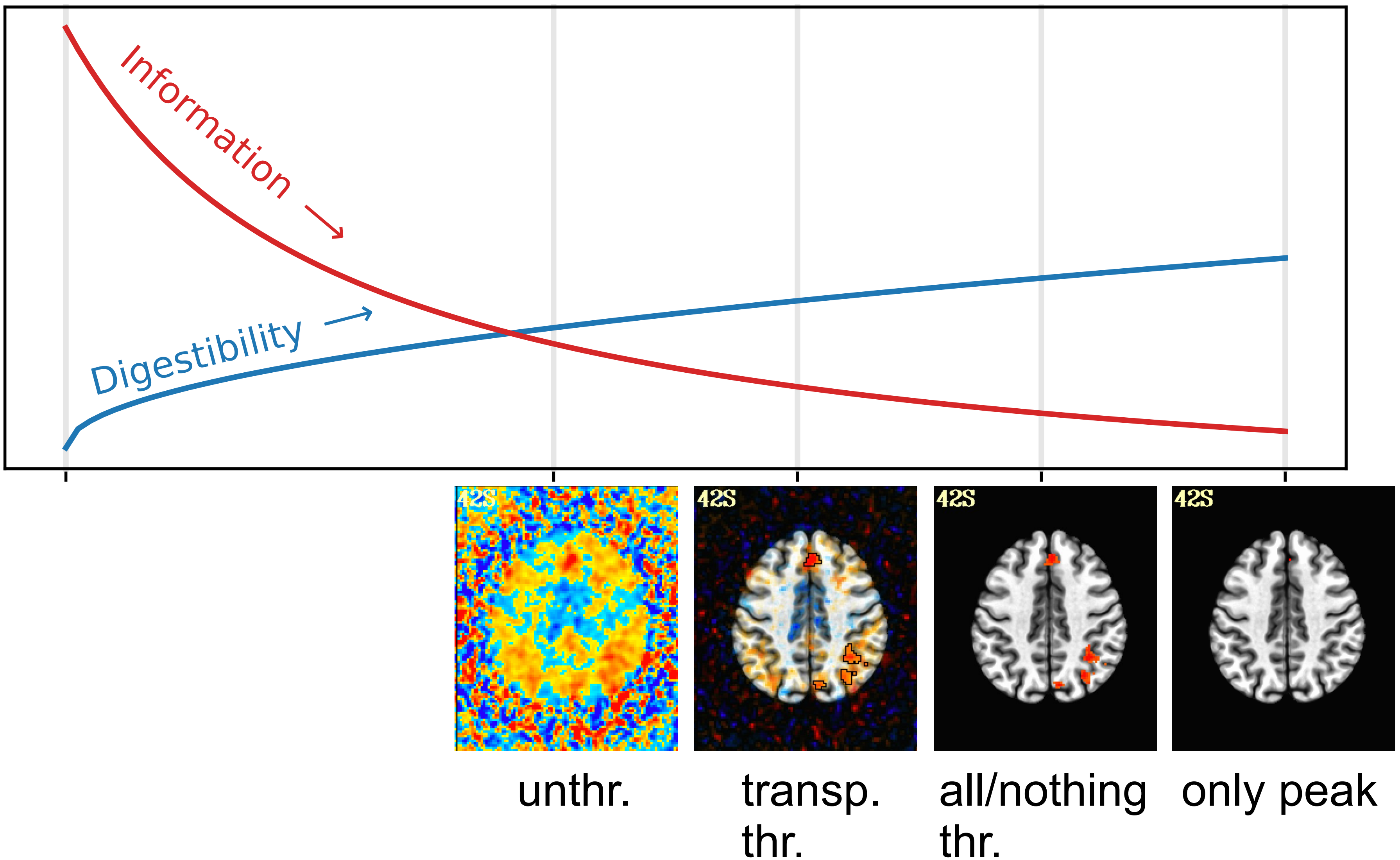
How much should we reduce data in figures?

Group-level results:
all-or-nothing threshold (p = 0.001, FWE = 5%) → peak voxels

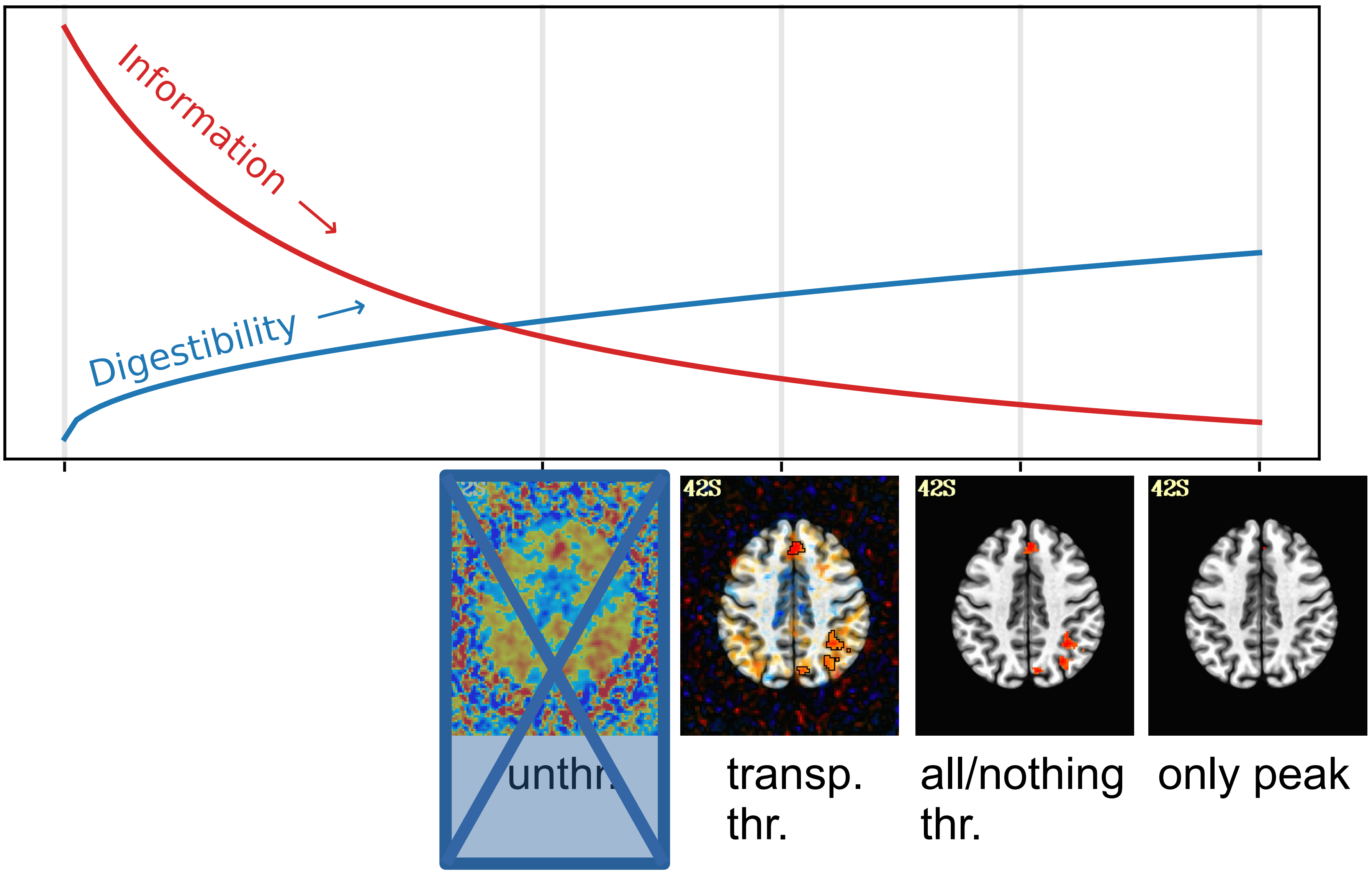




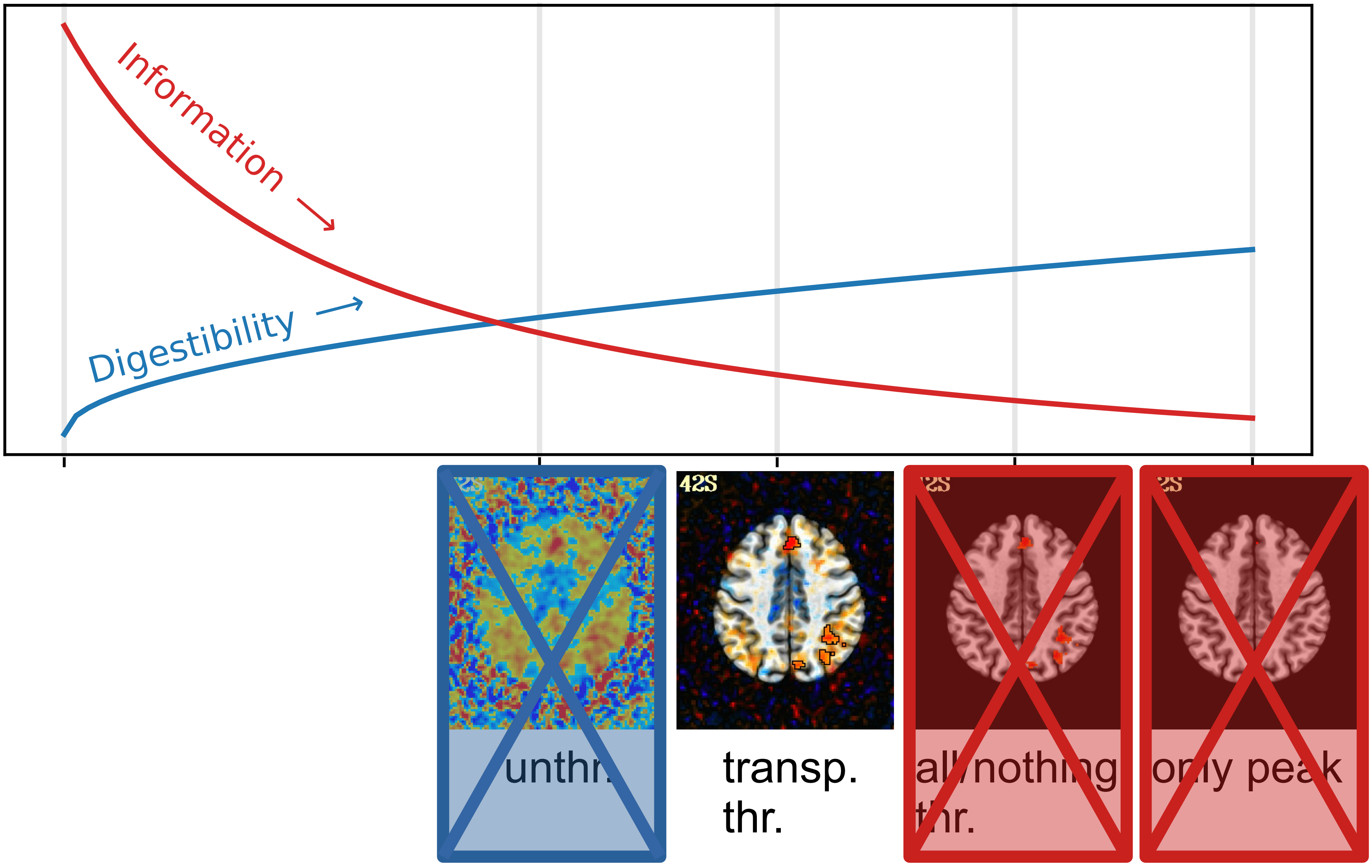
How much should we reduce data in figures?



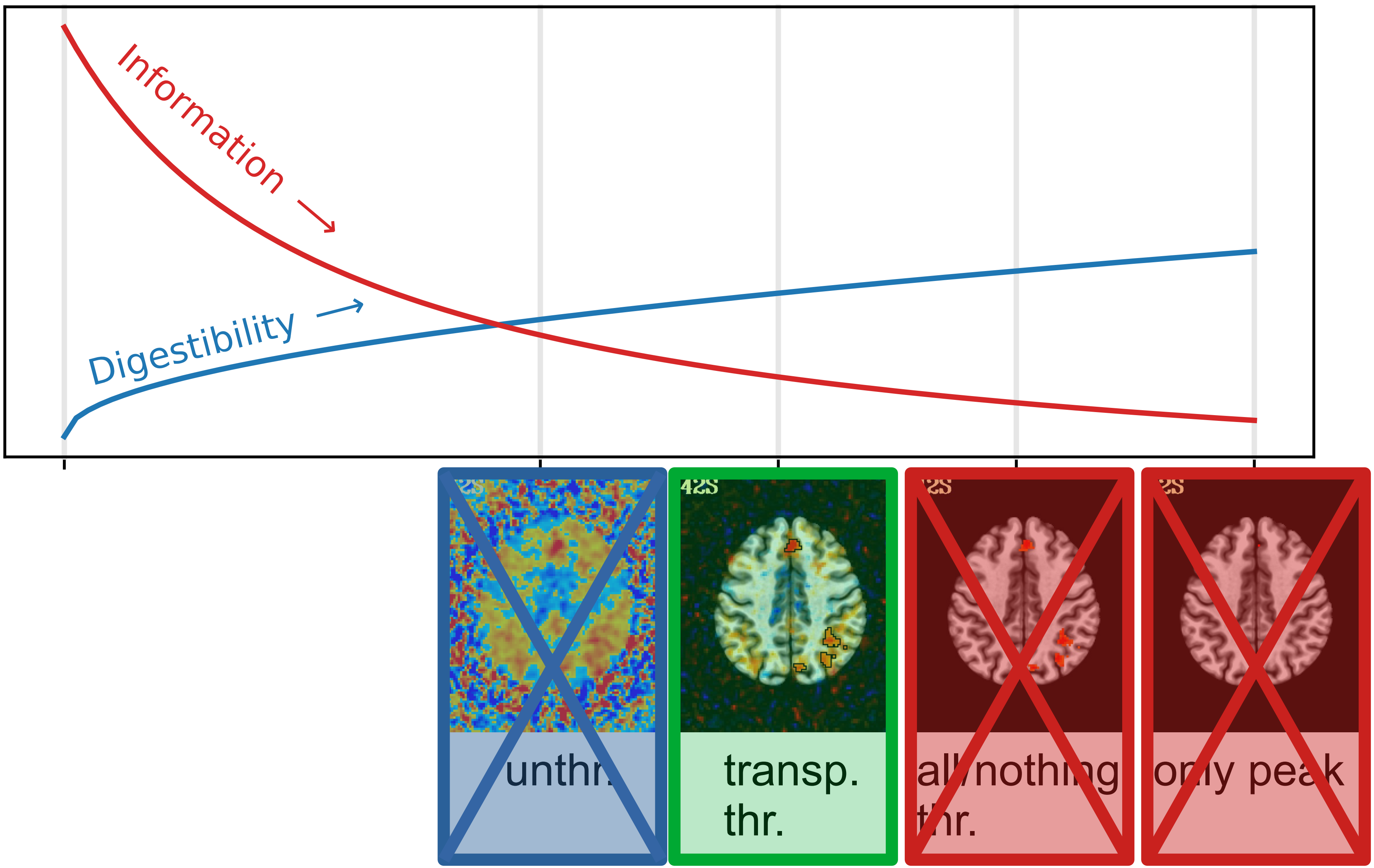
How much should we reduce data in figures?



How much should we reduce data in figures?



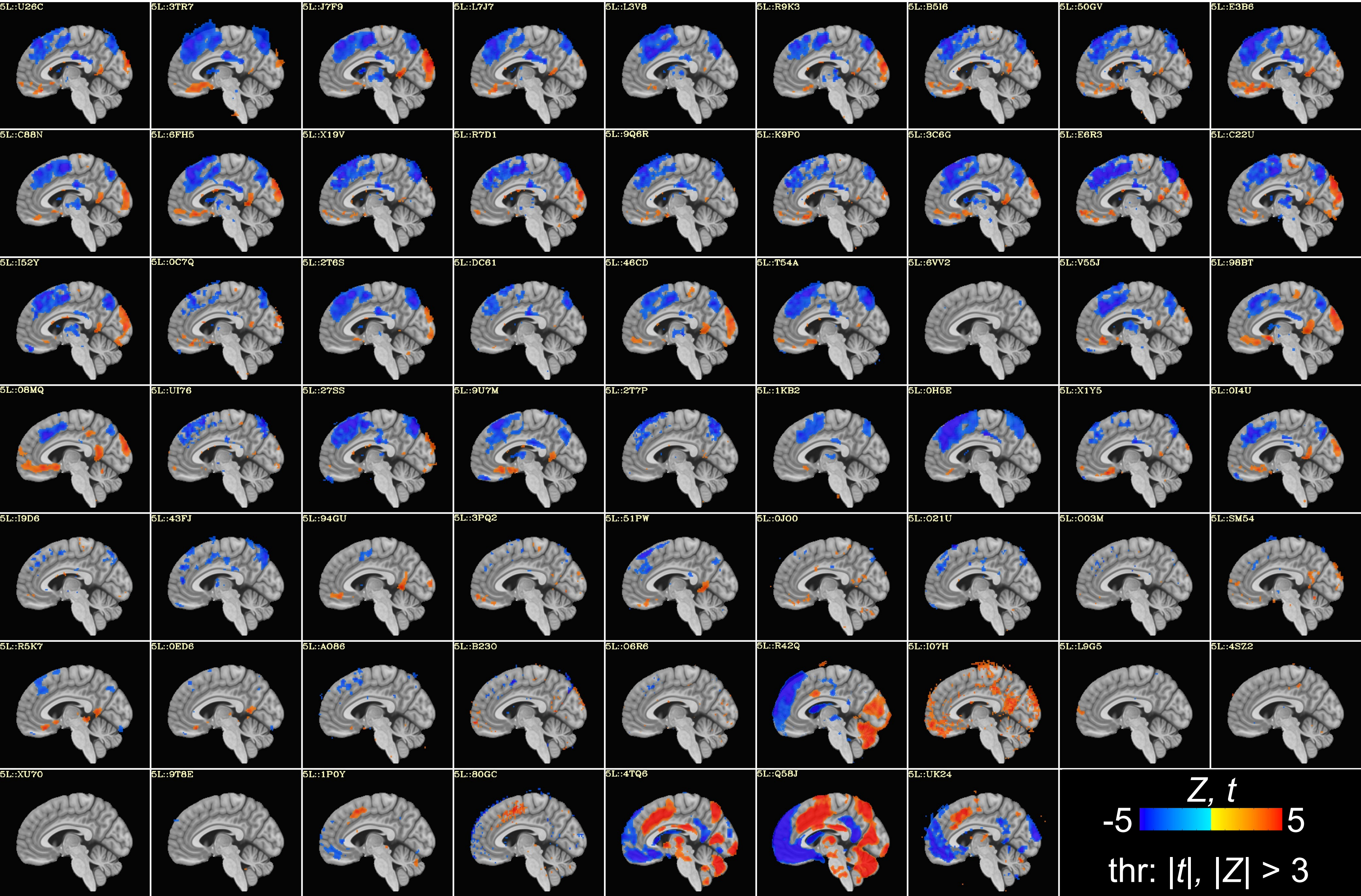
How much should we reduce data in figures?



Example 4:
improving cross study comparisons

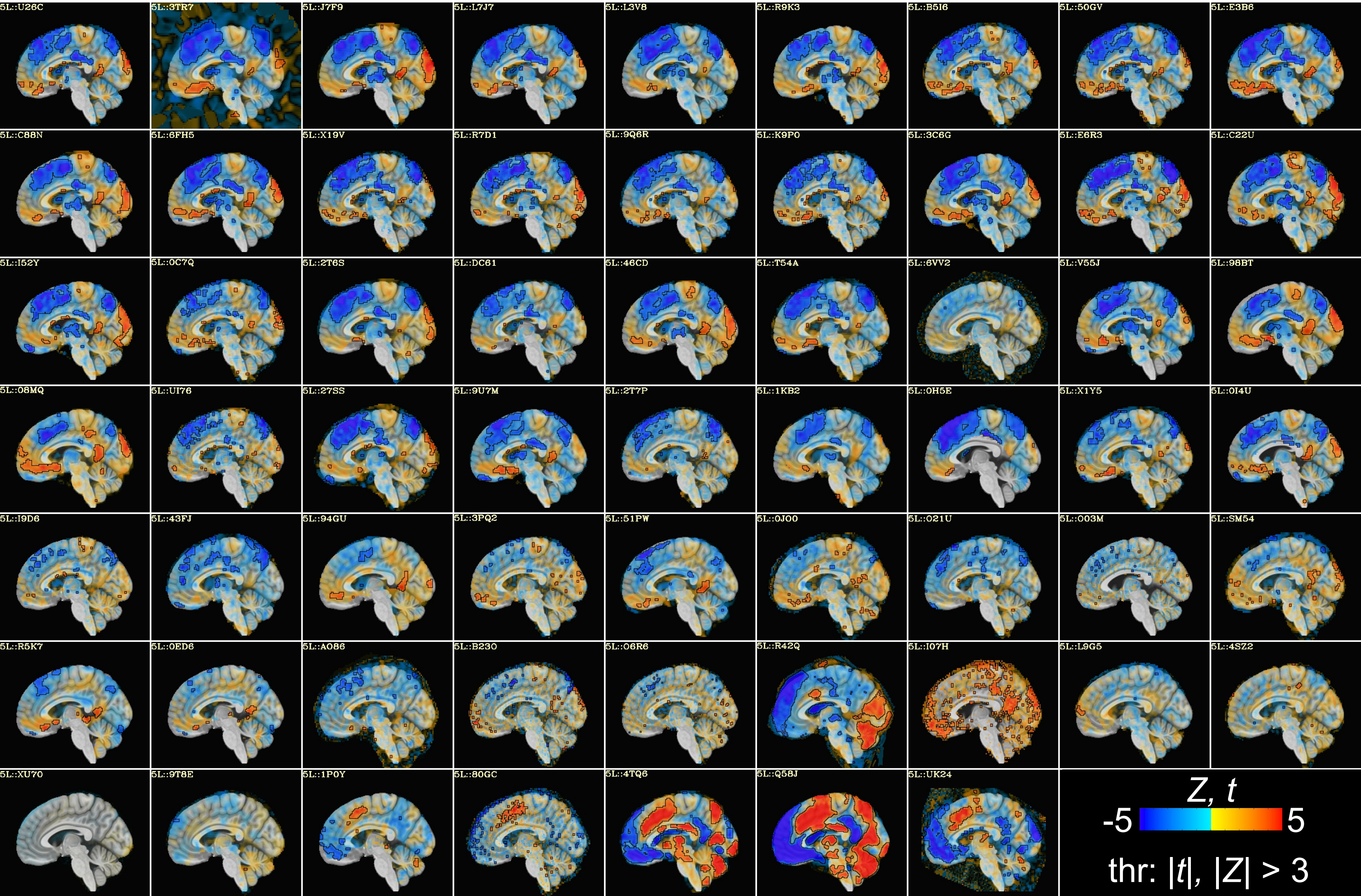
How does this affect cross study comparisons?

NARPS teams' results (Hyp #3): **all-or-nothing** thresholding



How does this affect cross study comparisons?

NARPS teams' results (Hyp #3): **transparent** thresholding

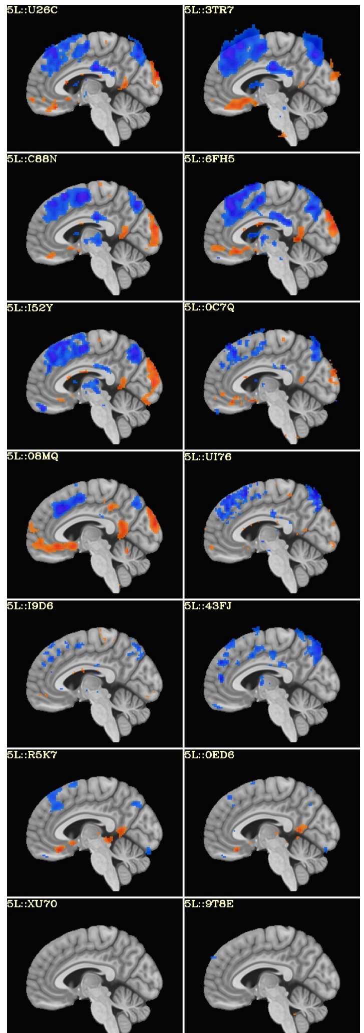


How does this affect cross study comparisons?

Researcher's choice of what to compare (decided in large part by thresholding filter) greatly affects perceived+measured outcomes (!)

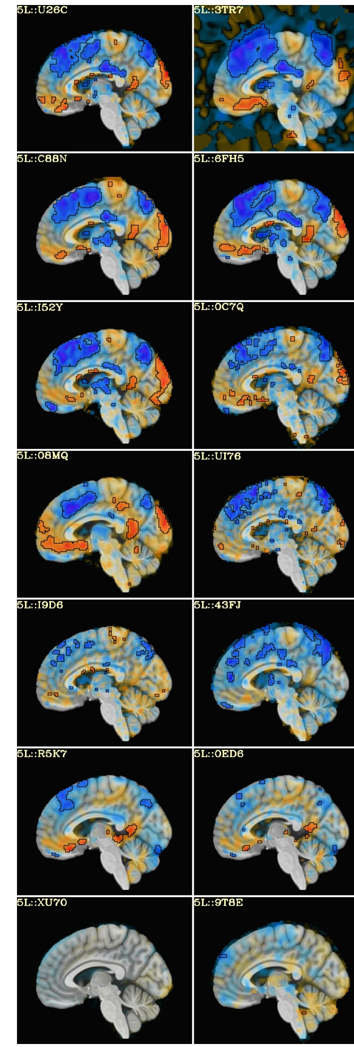
Opaque thresholding

Match blobs
→
Overlap in first rows, but then perceive *notable disagreement* and large variability, perhaps *“lack of reproducibility”*



Opaque thresholding

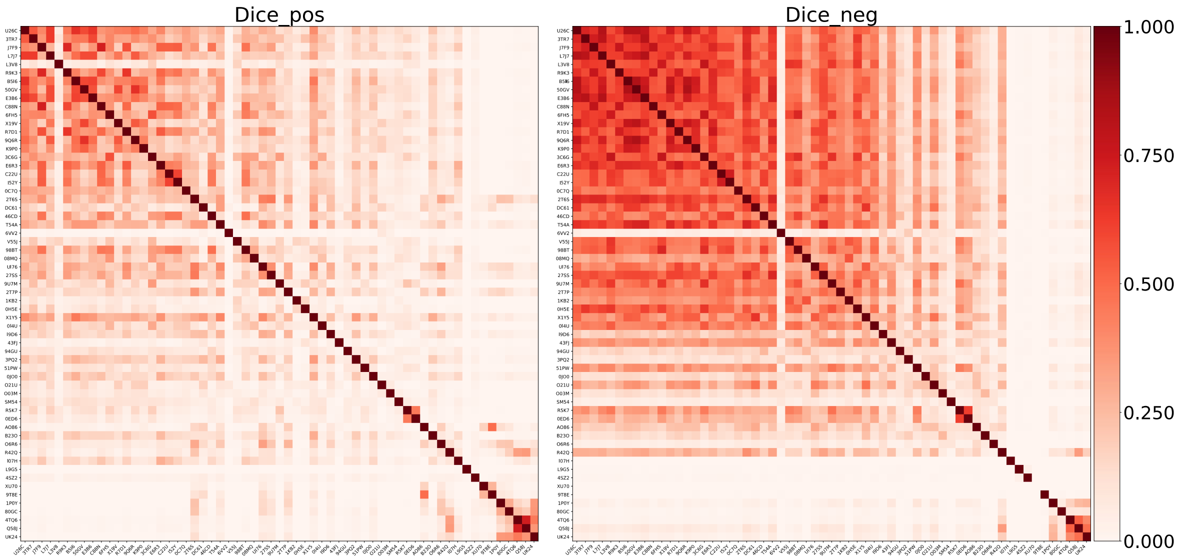
Match colors
→
Similar hot-cold patterns (except last 1.5 rows): *general agreement with varied strength* (befits allowed flexibility)



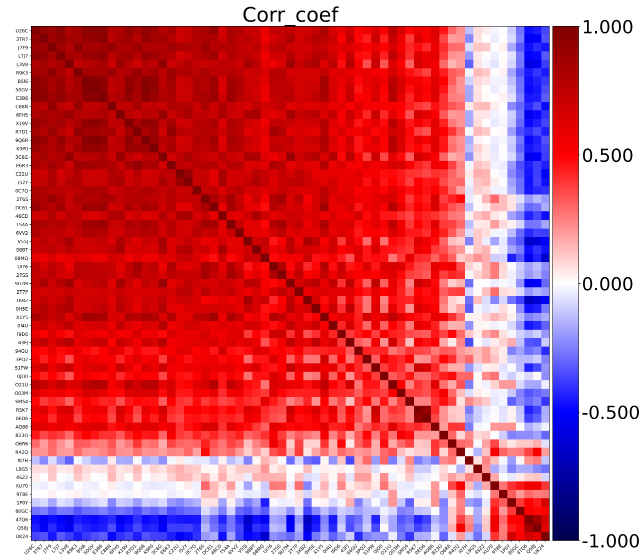
How does this affect cross study comparisons?

Similarity matrices, derived from the preceding team results (whole brain)

all-or-nothing threshold:
Dice coefficients (pos & neg clusters)



transparent threshold:
Pearson correlation

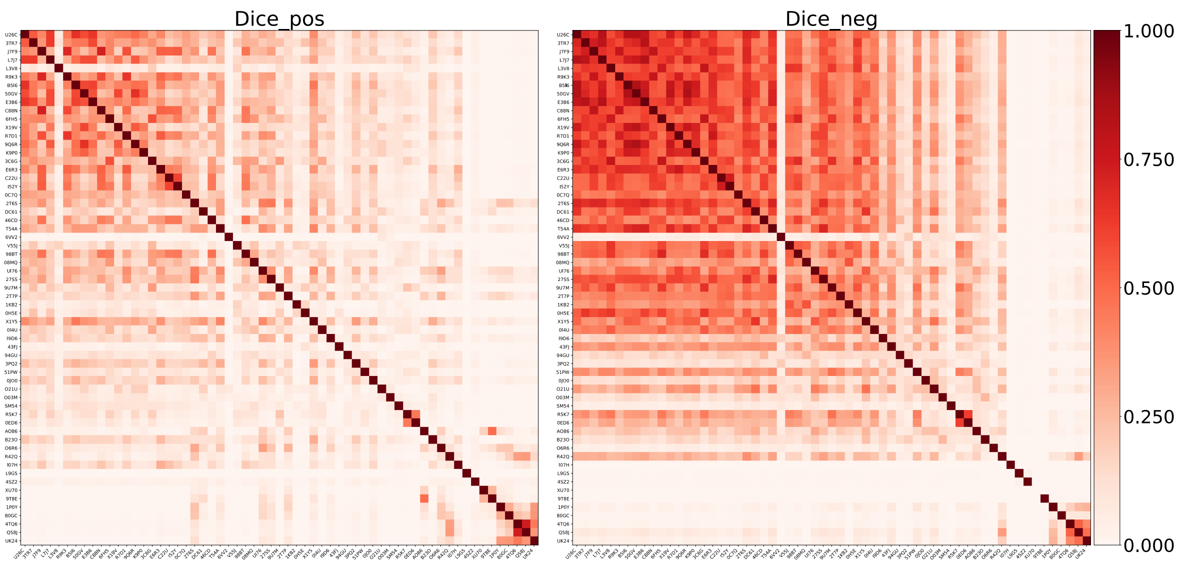


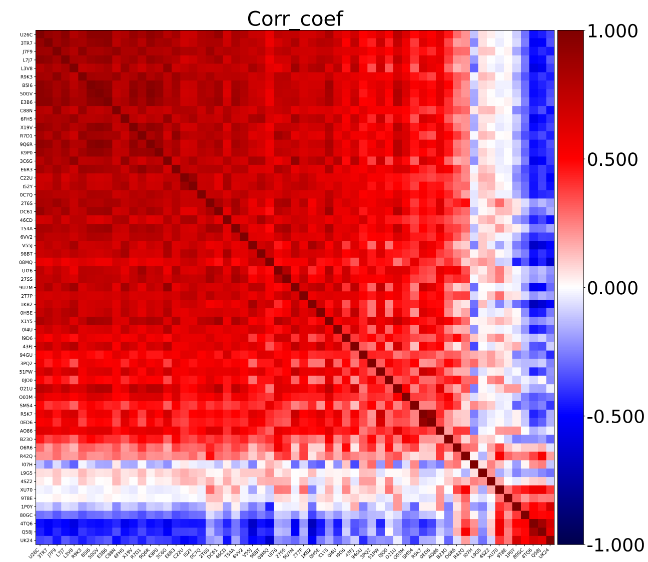
How does this affect cross study comparisons?

Similarity matrices, derived from the preceding team results (whole brain)

all-or-nothing threshold:
Dice coefficients (pos & neg clusters)

transparent threshold:
Pearson correlation





Might interpret as poor agreement or high variability of results → *crisis!*

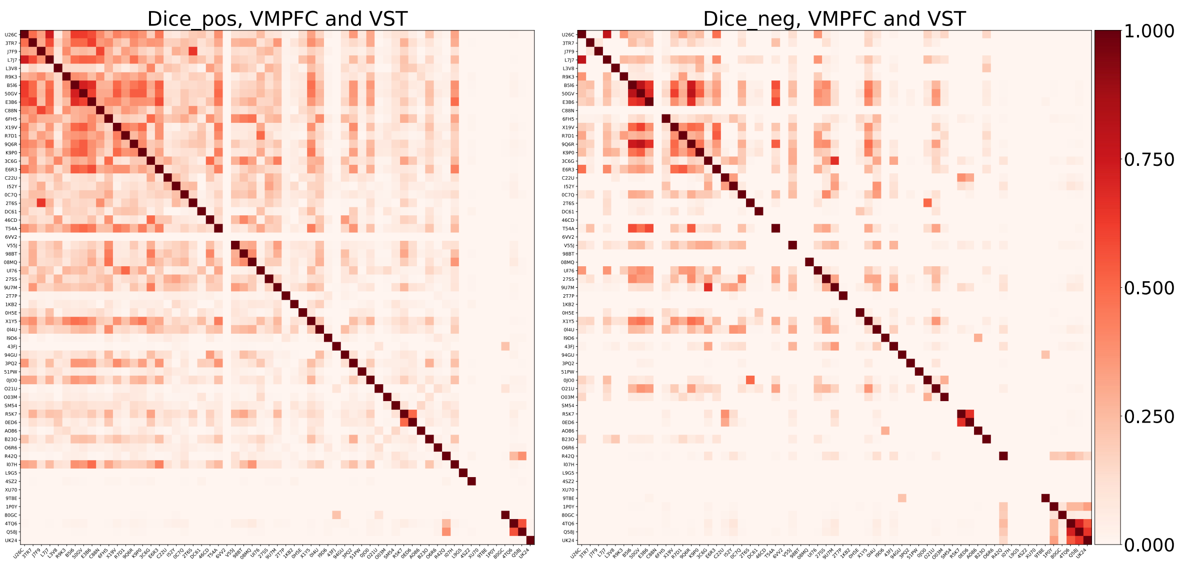
General agreement → *mainly consistent results (with varied strengths)*

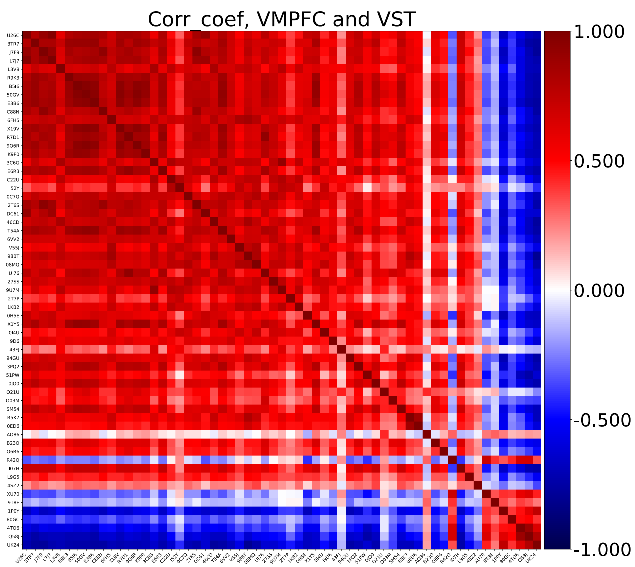
How does this affect cross study comparisons?

Similarity matrices, derived from the preceding team results (region specific)
→ NARPS design had specific ROIs per Hyp (here, VMPFC and VST)

all-or-nothing threshold:
Dice coefficients (pos & neg clusters)

transparent threshold:
Pearson correlation





Might interpret as poor agreement or high variability of results → *crisis!*

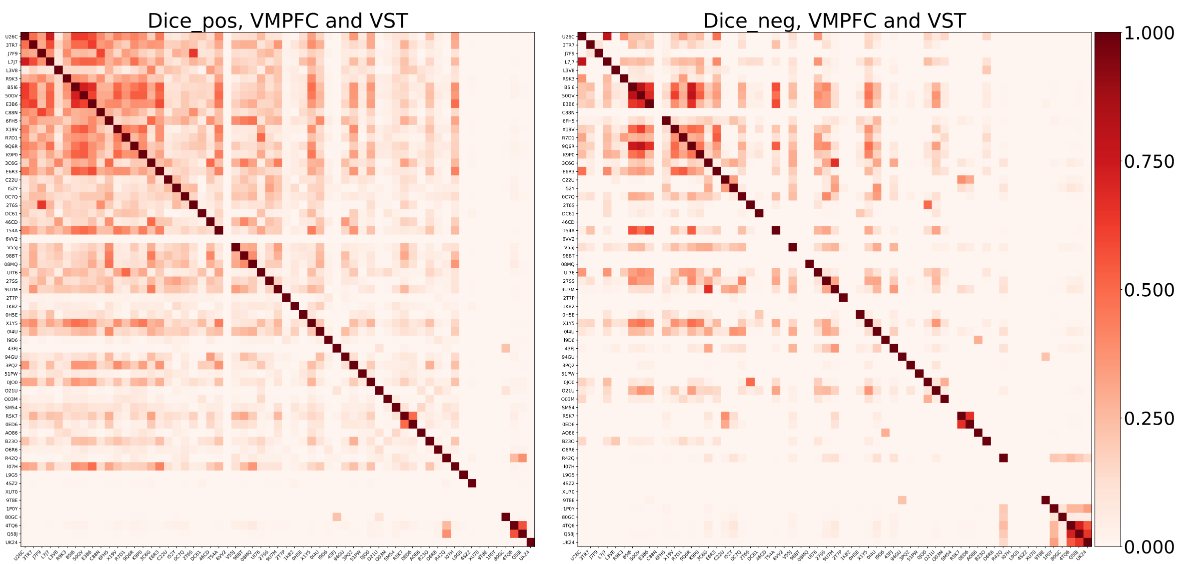
General agreement → *mainly consistent results (with varied strengths)*

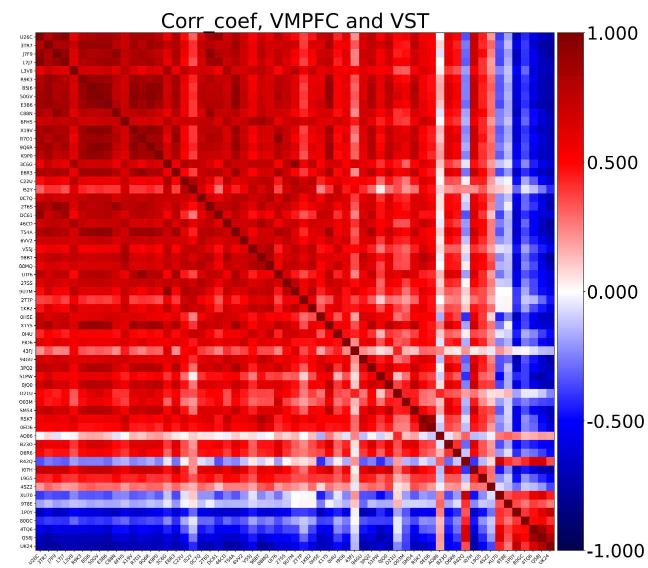
How does this affect cross study comparisons?

Similarity matrices, derived from the preceding team results (region specific)
→ NARPS design had specific ROIs per Hyp (here, VMPFC and VST)

all-or-nothing threshold:
Dice coefficients (pos & neg clusters)

transparent threshold:
Pearson correlation





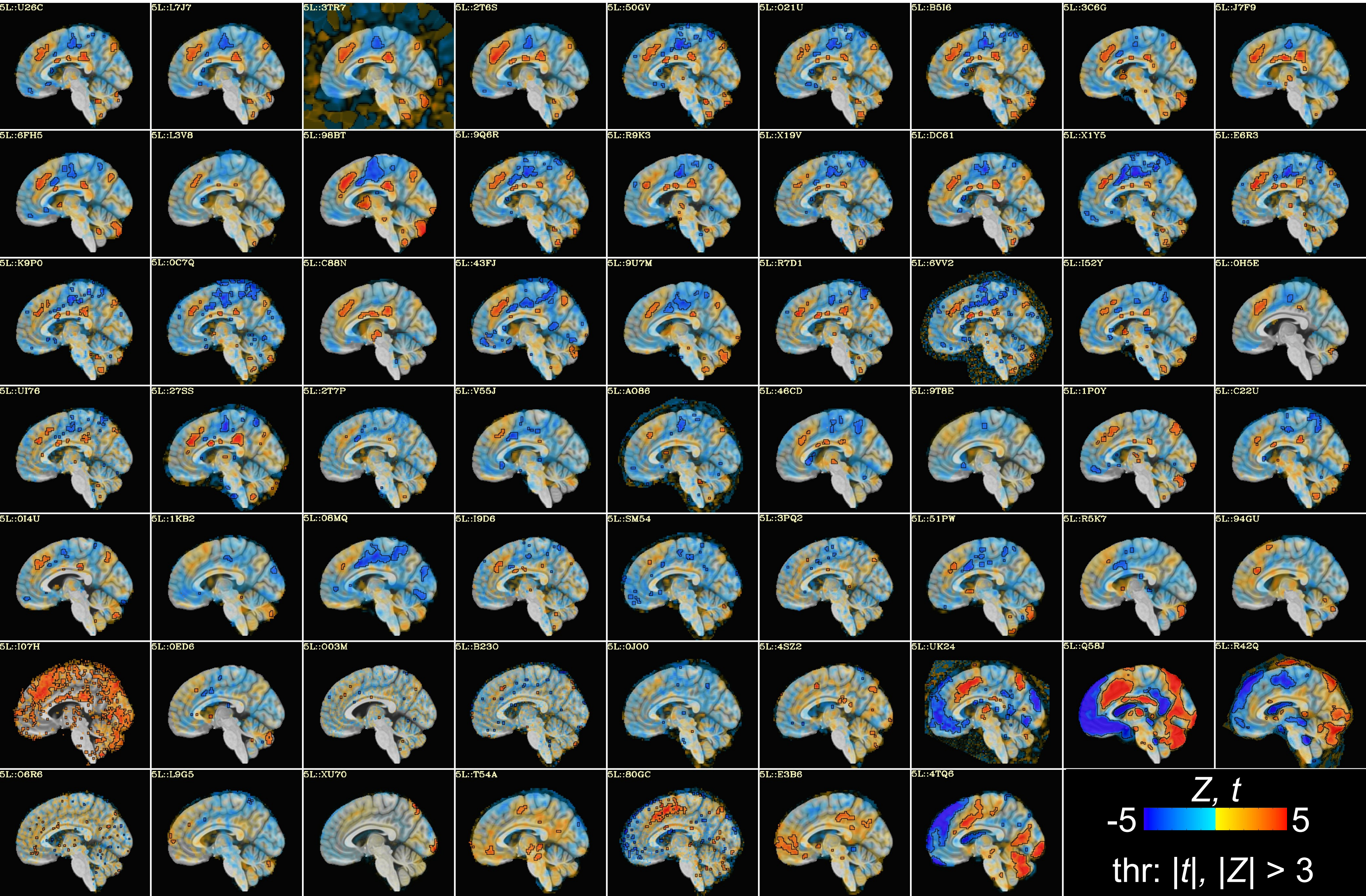
Might interpret as poor agreement or high variability of results → *crisis!*

General agreement → *mainly consistent results (with varied strengths)*

Thresholding before meta analysis removes valuable information!

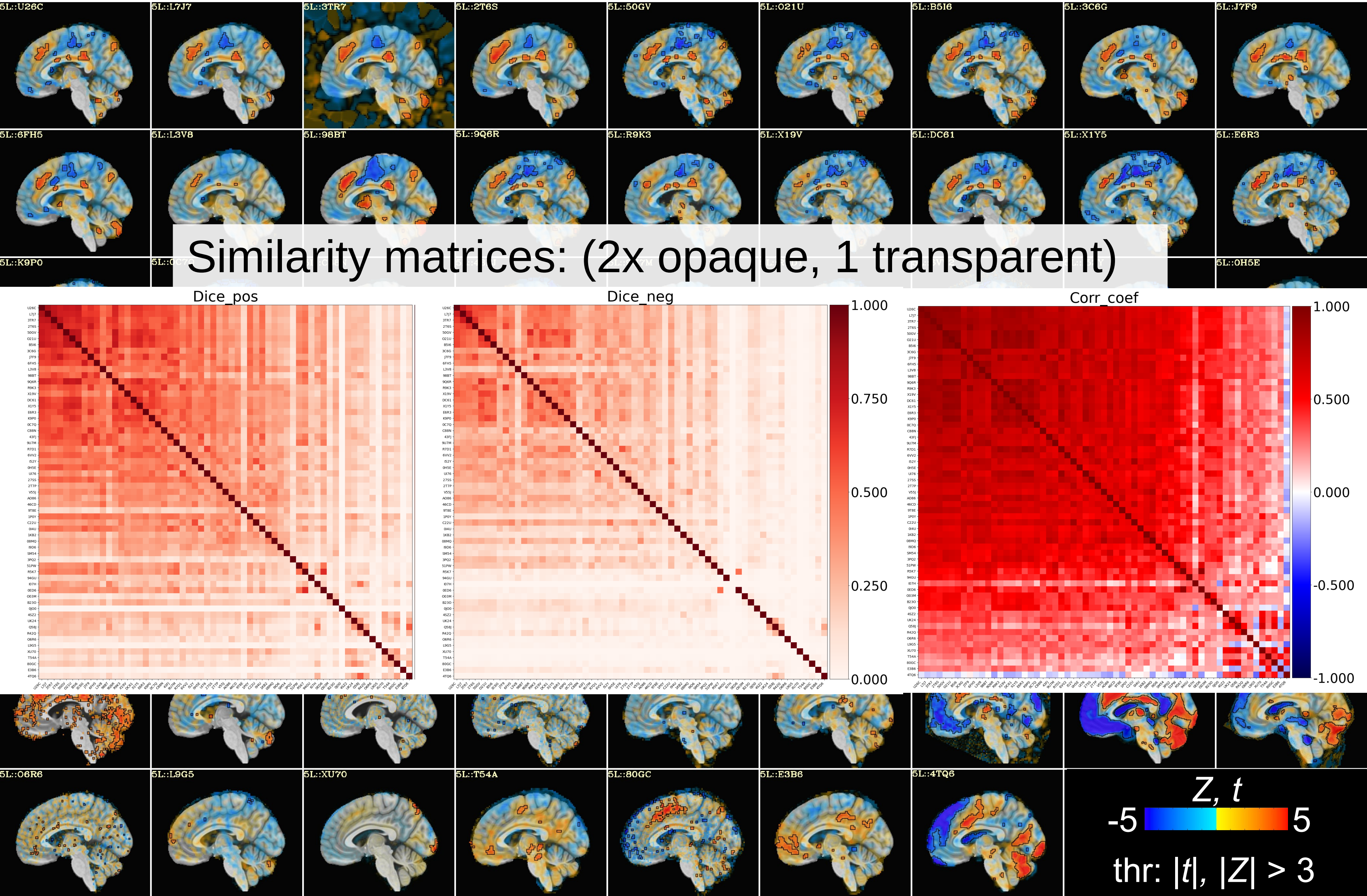
(and similar story across other NARPS Hyps.)

NARPS teams' results (Hyp #2&4): **transparent** thresholding



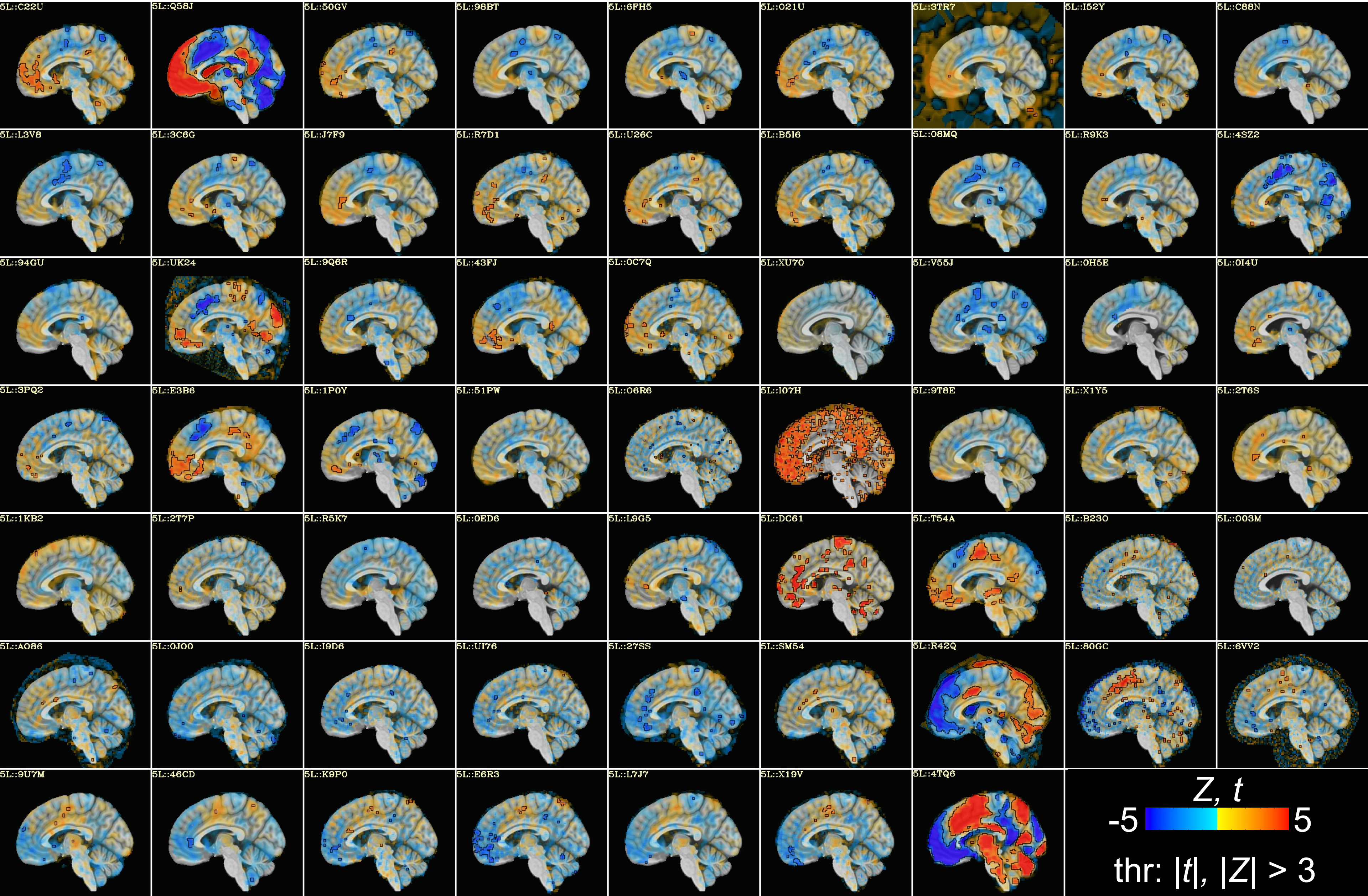
(and similar story across other NARPS Hyps.)

NARPS teams' results (Hyp #2&4): **transparent** thresholding



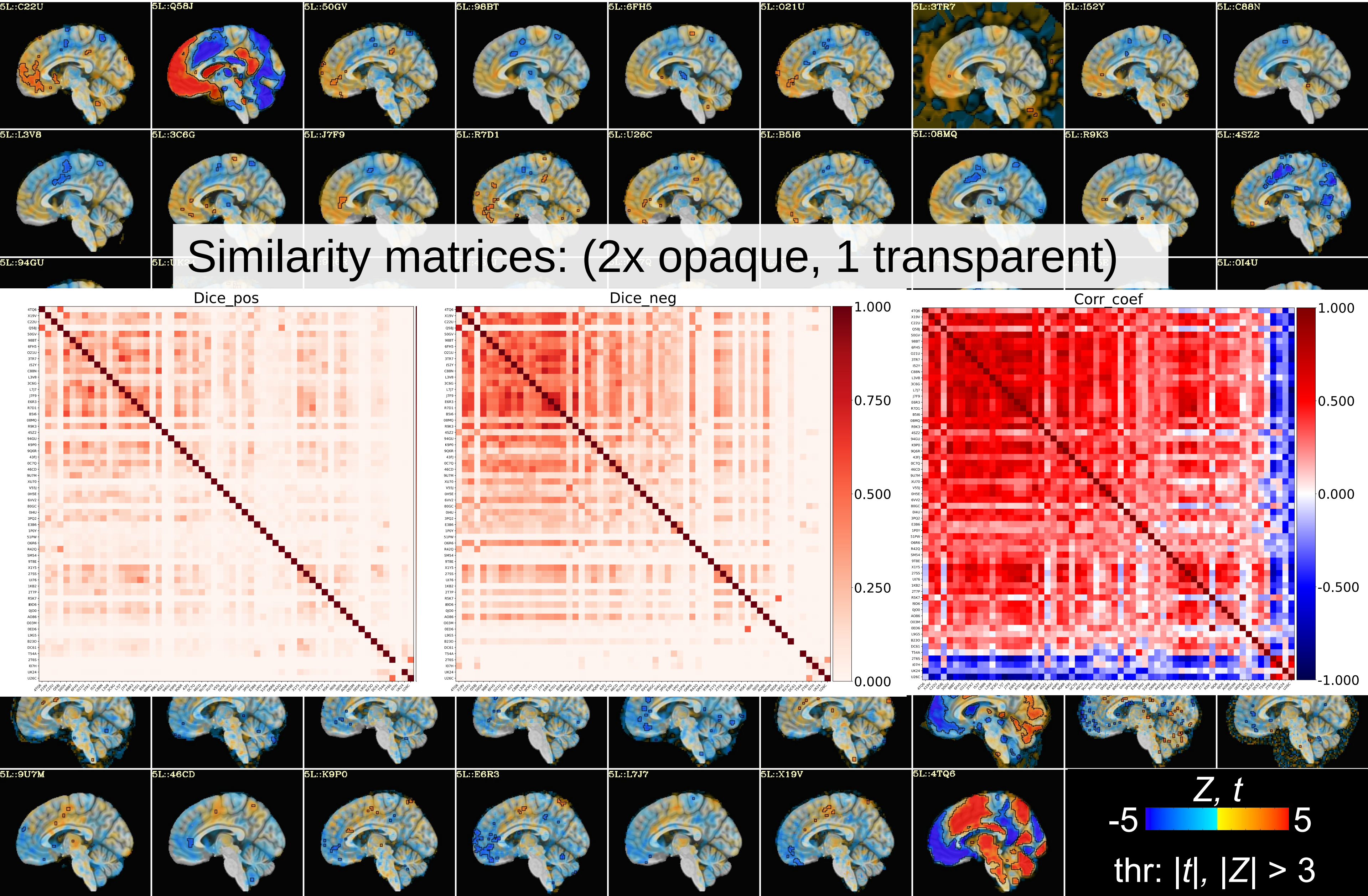
(and similar story across other NARPS Hyps.)

NARPS teams' results (Hyp #6): **transparent** thresholding



(and similar story across other NARPS Hyps.)

NARPS teams' results (Hyp #6): **transparent** thresholding



Conclusions

"Highlighting" is simple, improves information/interpretation within a study, and aids accuracy of comparisons across studies.

A study provides *evidence*, so show more full results.

- A study is part of a conversation, not 'the answer'.

Highlight key findings, and don't hide everything else.

- This philosophy applies beyond voxelwise studies.

All-or-nothing thresholding reduces information content, is sensitive to arbitrary values, hides artifacts, and more ☹

Cross study comparisons should be based on unthresholded results (and certainly not just peak voxels)

- Thresholding introduces strong biases into meta analyses
- This will improve the ability to assess reproducibility.

Bibliography

- Cox RW (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162-173.
- Jernigan TL, Gamst AC, Fennema-Notestine C, Ostergaard AL (2003). More "mapping" in brain mapping: statistical comparison of effects. Hum. Brain. Mapp. 19 (2), 90–95.
- Luo WL, Nichols TE (2003). Diagnosis and exploration of massively univariate neuroimaging models. Neuroimage 19 (3), 1014–1032.
- Allen EA, Erhardt EB, Calhoun VD (2012). Data visualization in the neurosciences: overcoming the curse of dimensionality. Neuron 74, 603–608 .
- Chen G, Taylor PA, Cox RW (2017). Is the statistic value all we should care about in neuroimaging? Neuroimage. 147:952-959. doi:10.1016/j.neuroimage.2016.09.066 https://pubmed.ncbi.nlm.nih.gov/27729277/
- Pernet CR, Madan CR (2020). Data visualization for inference in tomographic brain imaging. Eur. J. Neurosci. 51 (3), 695–705.
- Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582 (7810), 84–88.
- Lindquist M (2020). Neuroimaging results altered by varying analysis pipelines. Nature 582 (7810), 36–37.
- Chen G, Taylor PA, Stoddard J, Cox RW, Bandettini PA, Pessoa L (2022). Sources of information waste in neuroimaging: mishandling structures, thinking dichotomously, and over-reducing data. Aperture Neuro. 2: DOI: 10.52294/2e179dbf-5e37-4338-a639-9ceb92b055ea
- Taylor PA, Reynolds RC, Calhoun V, Gonzalez-Castillo J, Handwerker DA, Bandettini PA, Mejia AF, Chen G (2023). Highlight Results, Don’t Hide Them: Enhance interpretation, reduce biases and improve reproducibility. Neuroimage 274:120138. doi: 10.1016/j.neuroimage.2023.120138 https://pubmed.ncbi.nlm.nih.gov/37116766/