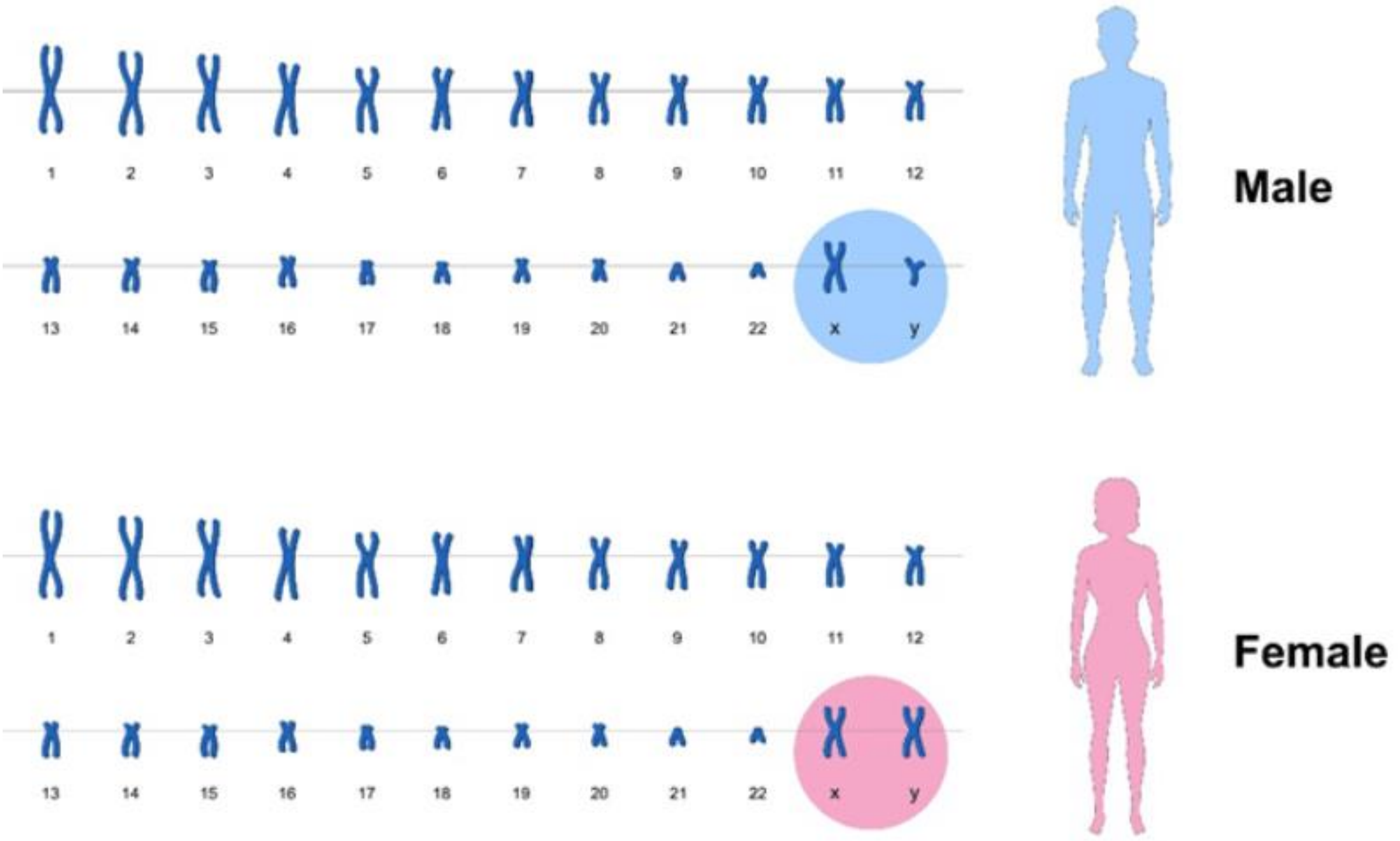
Medical/Bio Research Topics Ⅱ: Week 12 (19.11.2024)

Practical Implementation of AI Models for Classification (1): Dataset Exploration and Problem Formulation

분류 인공지능 모델 개발 실습 (1): 데이터 및 예측 문제

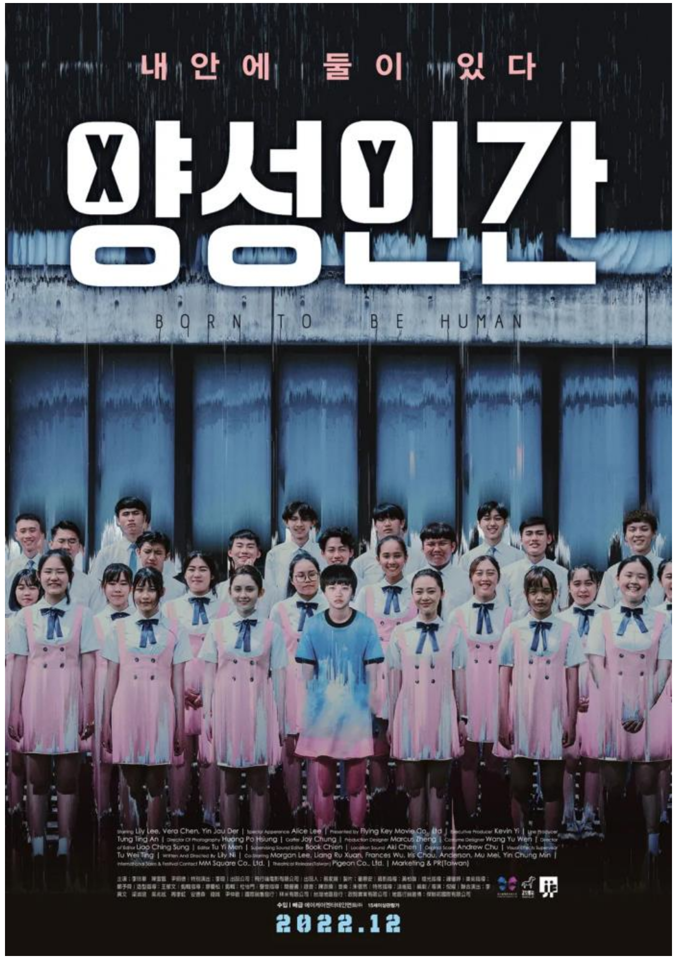
Sex and Gender

- Sex
  - Usually described by the terms "males" and "females"
  - Typically refers to the biological and physiological characteristics that define males and females
  - Determined by biological factors, primarily chromosomal (XX for females, XY for males) and anatomical differences



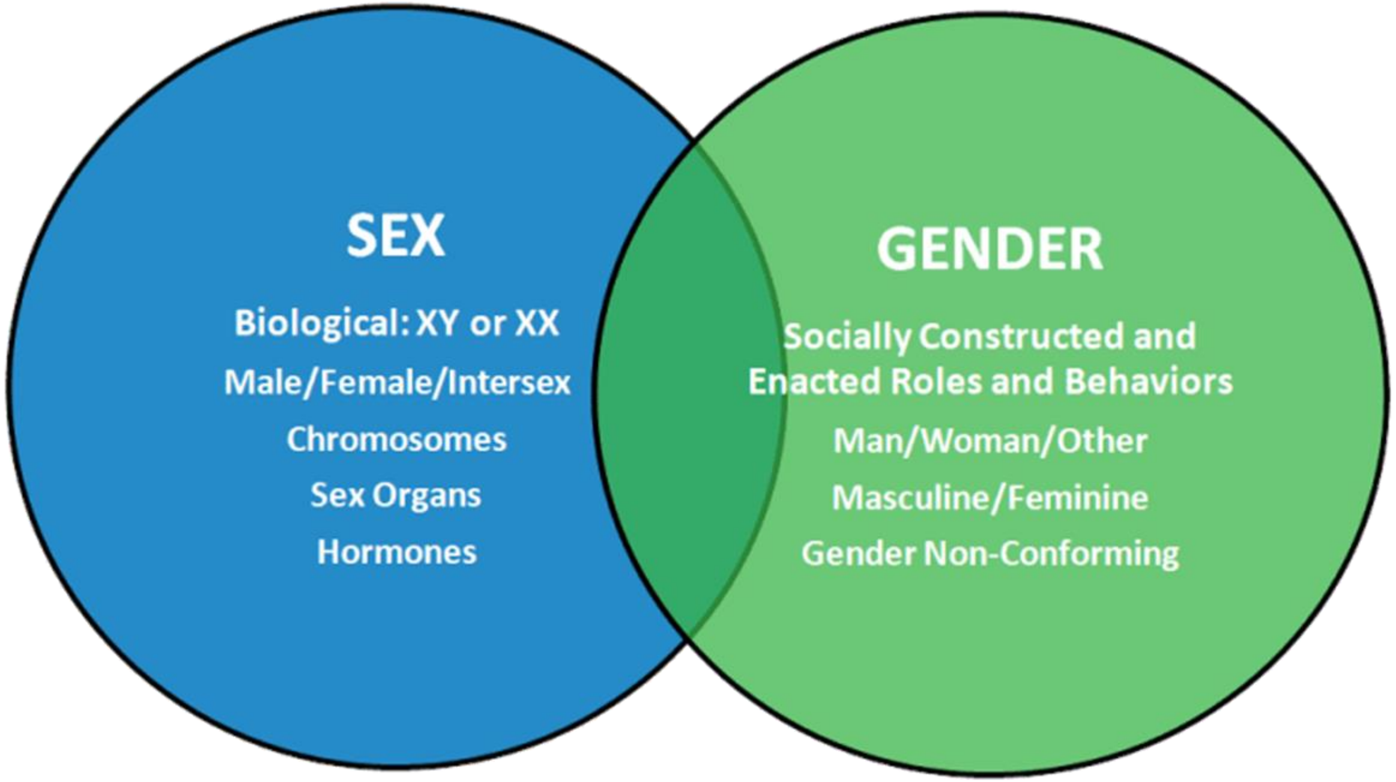
[https://www.shalom-education.com/courses/gcse-biology/lessons/genetic-variation-and-mutation/topic/sex-determination]

Sex determination through sex chromosomes



Intersex born with sex characteristics that do not fit typical binary notions

- Gender
  - Usually described by the terms "men" and "women"
  - Often refers to the socially constructed roles, behaviours, activities, and attributes that a society considers appropriate for individuals based on their sex
  - Related to how individuals perceive themselves and what they call themselves, which can be influenced by societal norms and personal experiences
- Both sex and gender exist on a spectrum, such that individuals may identify and express themselves in various ways that do not conform to traditional binary categories



[https://orwh.od.nih.gov/sex-gender]

Dimensions of sex and gender

Classification in Machine Learning

- Models the relationship between input features (predictors) and the target variable (class label)
- Purpose
  - Understanding the relationship between input features and categorical target variables (classes)
  - Predicting categorical target values (classes) for new sets of input features

- Supervised learning technique for predicting discrete output values (classes)
  - Traditional methods
    - Linear classifiers: Logistic Regression, Linear Discriminant Analysis (LDA)
    - Non-linear classifiers: k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Decision Trees
    - Probabilistic classifiers: Naive Bayes, Gaussian Discriminant Analysis
  - Ensemble methods
    - Bagging-based methods: random forests, extra trees
    - Boosting-based methods: AdaBoost (Adaptive Boosting), Gradient Boosting Machines (GBM), XGBoost (eXtreme Gradient Boosting), LightGBM, CatBoost (Categorical Boosting)
    - Stacking: combining predictions from multiple models

- Deep learning-based classification
  - Feedforward Neural Network (FNN) / Multilayer Perceptron (MLP)
  - Specialised architectures
    - Convolutional Neural Network (CNN) for spatial data classification
    - Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) for sequential data classification
    - Transformer-based models for complex sequential data classification
- Hybrid approaches
  - Combining traditional methods with ensemble techniques or neural networks
  - Automated machine learning (AutoML) systems incorporating various classification techniques

- Types of classification problems
  - Binary classification: classifying into two classes (e.g., spam vs. not spam)
  - Multi-class classification: classifying into more than two classes (e.g., different animal species)
  - Multi-label classification: assigning multiple labels to each instance (e.g., image tagging)

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

[https://en.wikipedia.org/wiki/Confusion_matrix]

Confusion matrix: a special kind of contingency table

- Classification performance
  - Accuracy
    - Proportion of correct predictions among the total number of cases examined: (TP + TN) / (TP + FP + TN + FN)
    - Range: 0 to 1 (higher is better)
    - Easy to interpret, but can be misleading for imbalanced datasets
  - Precision (True Positive Value (TPV))
    - Proportion of true positive predictions among all positive predictions: TP / (TP + FP)
    - Range: 0 to 1 (higher is better)
    - Focuses on the accuracy of positive predictions

- Recall (True Positive Rate (TPR) or sensitivity)
  - Proportion of true positive predictions among all actual positive cases: TP / (TP + FN)
  - Range: 0 to 1 (higher is better)
  - Focuses on the completeness of positive predictions
- F_1 score
  - Harmonic mean of precision and recall: 2 / ((1 / precision) + (1 / recall)) = 2TP / (2TP + FP + FN)
  - Range: 0 to 1 (higher is better)
  - Provides a single score that balances both precision and recall
  - Particularly useful for imbalanced datasets

- ROC (Receiver Operating Characteristic) curve
  - Graph showing the performance of a classification model at all classification thresholds
  - Plots TPR (TPR, sensitivity) against False Positive Rate (FPR = 1 - True Negative Rate (TNR, specificity))
  - Useful for visualizing the trade-off between sensitivity and specificity
- Area Under the ROC Curve (AUC-ROC)
  - Aggregate measure of performance across all possible classification thresholds
  - Range: 0.5 (random guessing) to 1 (perfect classification)
  - Provides a single score that summarizes the ROC curve

- Confusion matrix
  - Table showing the number of true positives, true negatives, false positives, and false negatives
  - Not a single metric, but a comprehensive view of the model's performance
  - Useful for deriving other metrics and understanding the types of errors made

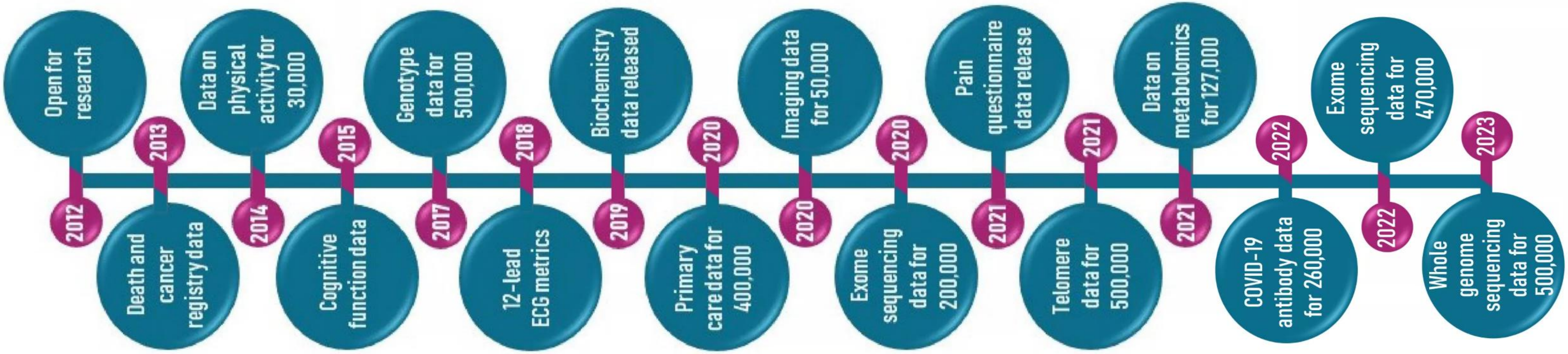
		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP), hit**	**False negative (FN), type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP), type I error, false alarm, overestimation**	**True negative (TN), correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2 PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

[https://en.wikipedia.org/wiki/Confusion_matrix]

Confusion matrix and its derived metrics

UK Biobank

- World’s largest health research database
  - Around 500,000 UK residents aged 40-69 years at recruitment
  - Initial recruitment from 2006 to 2010
  - Ongoing follow-up:
    - Regular updates of health outcomes for all participants
    - Ongoing imaging study, aiming to scan 100,000 participants
    - Periodic questionnaires on various health topics
- Aims to improve prevention, diagnosis, and treatment of various diseases



[https://www.ukbiobank.ac.uk/enable-your-research/about-our-data]

Extension of UK Biobank data

- Unique features:
  - Large sample size
  - Comprehensive data collection
  - Long-term follow-up
- Types of data collected:
  - Genetic information: whole genome sequencing, whole exome sequencing, and genotyping
  - Health-related records
  - Lifestyle questionnaires
  - Physical measurements
  - Imaging data (for a subset of participants)

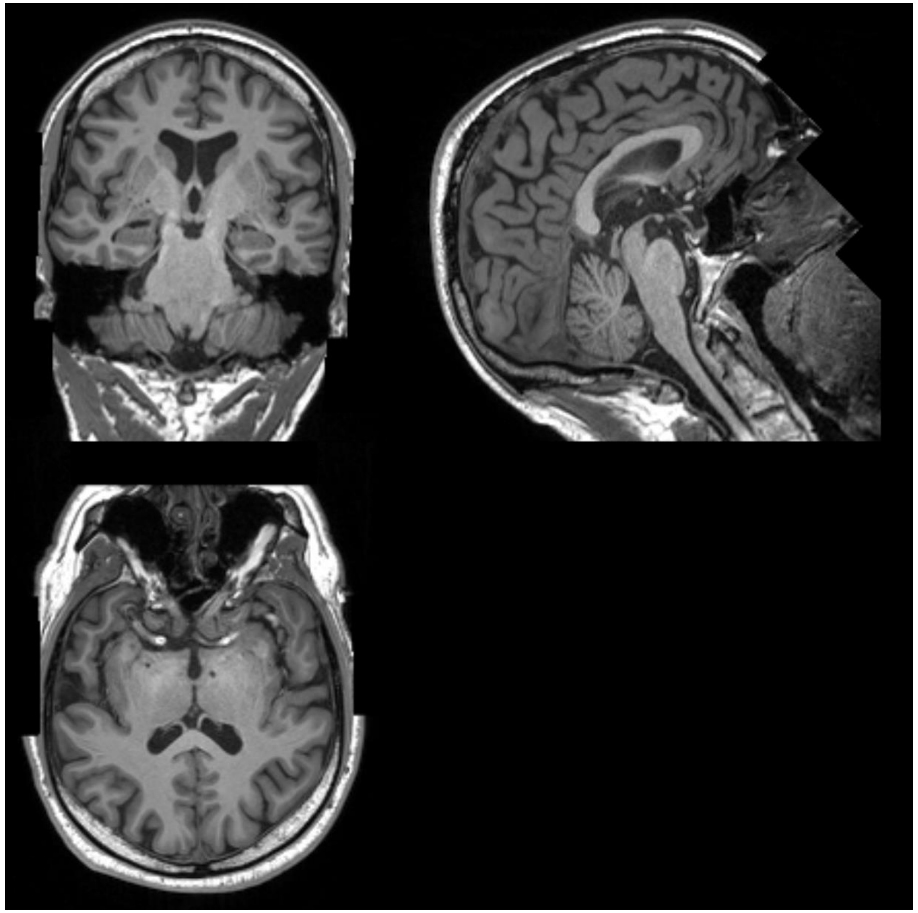
- Brain MRI data
  - Structural MRI (sMRI)
    - T1-weighted
    - FLAIR
  - Functional MRI (fMRI)
    - Task-based
    - Resting state
  - Diffusion-weighted MRI (dMRI)
  - Susceptibility-weighted MRI (swMRI)

Modality	Duration	Voxel, Matrix	Key Parameters	#Volumes/ #Timepoints
T1	4:54	$1 \times 1 \times 1$ mm $208 \times 256 \times 256$	3D MPRAGE, sagittal, R = 2, TI/TR = 880/ 2000 ms	1
T2 FLAIR	5:52	$1.05 \times 1.0 \times 1.0$ mm $192 \times 256 \times 256$	FLAIR, 3D SPACE, sagittal, R = 2, PF 7/ 8, fat sat, TI/ TR = 1800/5000 ms, elliptical	1
swMRI	2:34	$0.8 \times 0.8 \times 3.0$ mm $256 \times 288 \times 48$	3D GRE, axial, R = 2, PF 7/8 TE1/TE2/ TR = 9.4/20/27 ms	2
dMRI	7:08	$2.0 \times 2.0 \times 2.0$ mm $104 \times 104 \times 72$	MB = 3, R = 1, TE/ TR = 92/3600 ms, PF 6/8, fat sat, b = 0 s/mm^2 (5x + 3× phase-encoding reversed), b = 1 000 s/mm^2 (50×), b = 2000 s/mm^2 (50×)	105 + 6 (AP + PA)
rfMRI	6:10	$2.4 \times 2.4 \times 2.4$ mm $88 \times 88 \times 64$	TE/TR = 39/735 ms, MB = 8, R = 1, flip angle 52°, fat sat	490
tfMRI	4:13	$2.4 \times 2.4 \times 2.4$ mm $88 \times 88 \times 64$	Acquisition same as rfMRI. Task is faces/ shapes “emotion” task.	332

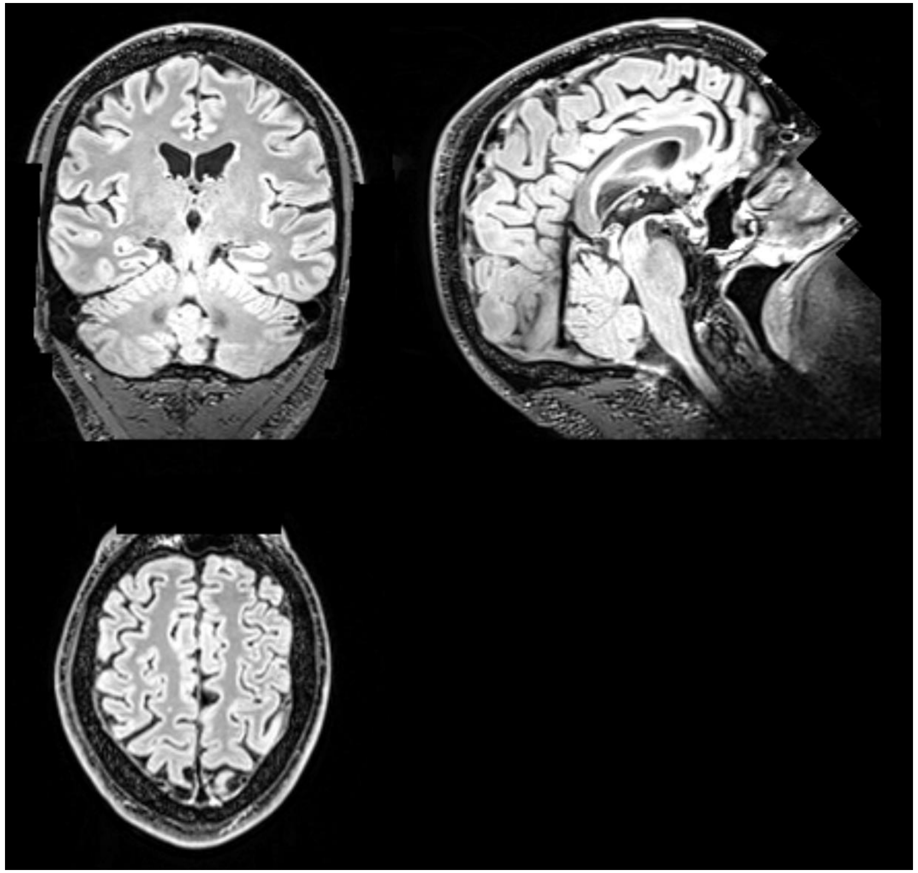
[Alfaro-Almagro et al., 2018]

Brain MRI data in UK Biobank

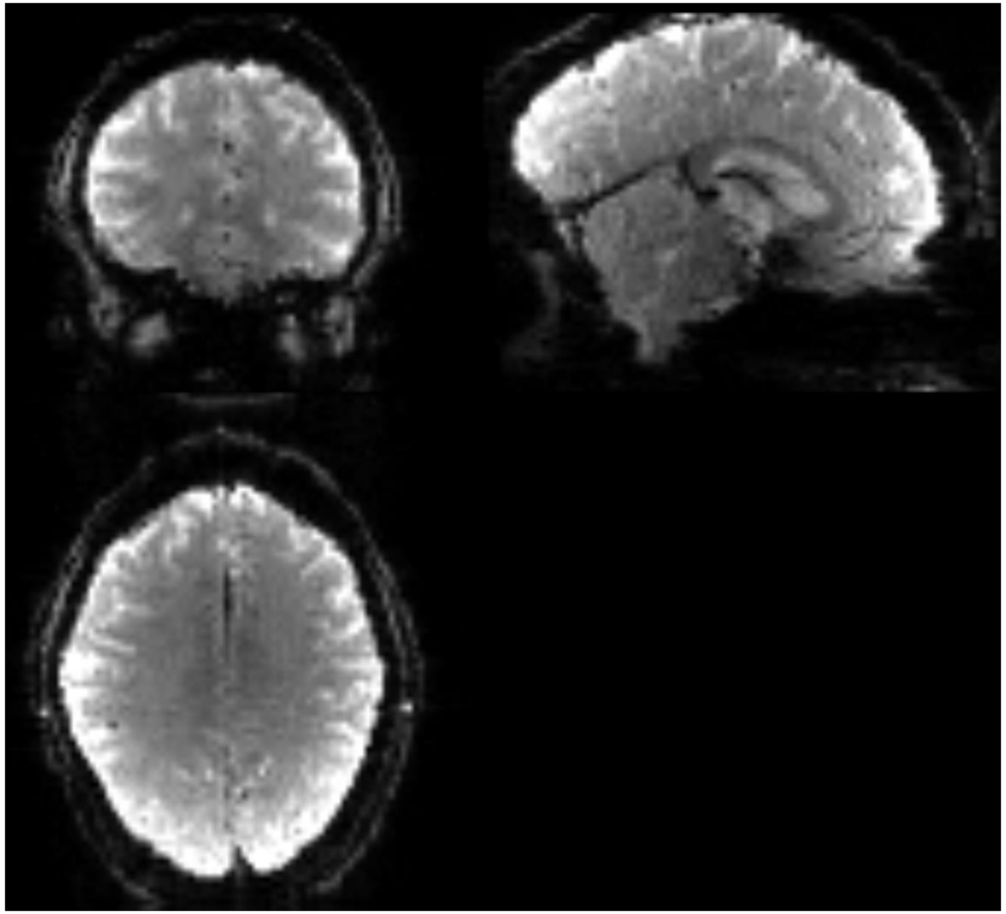
T1-weighted sMRI



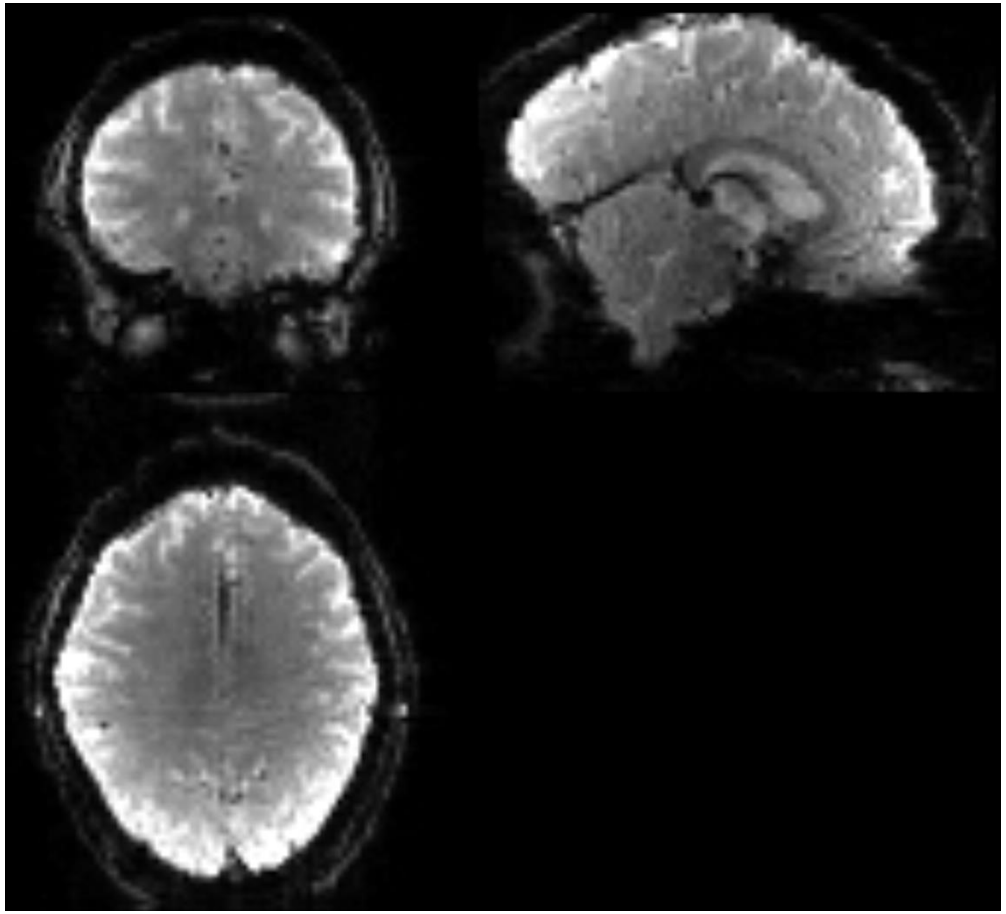
FLAIR sMRI



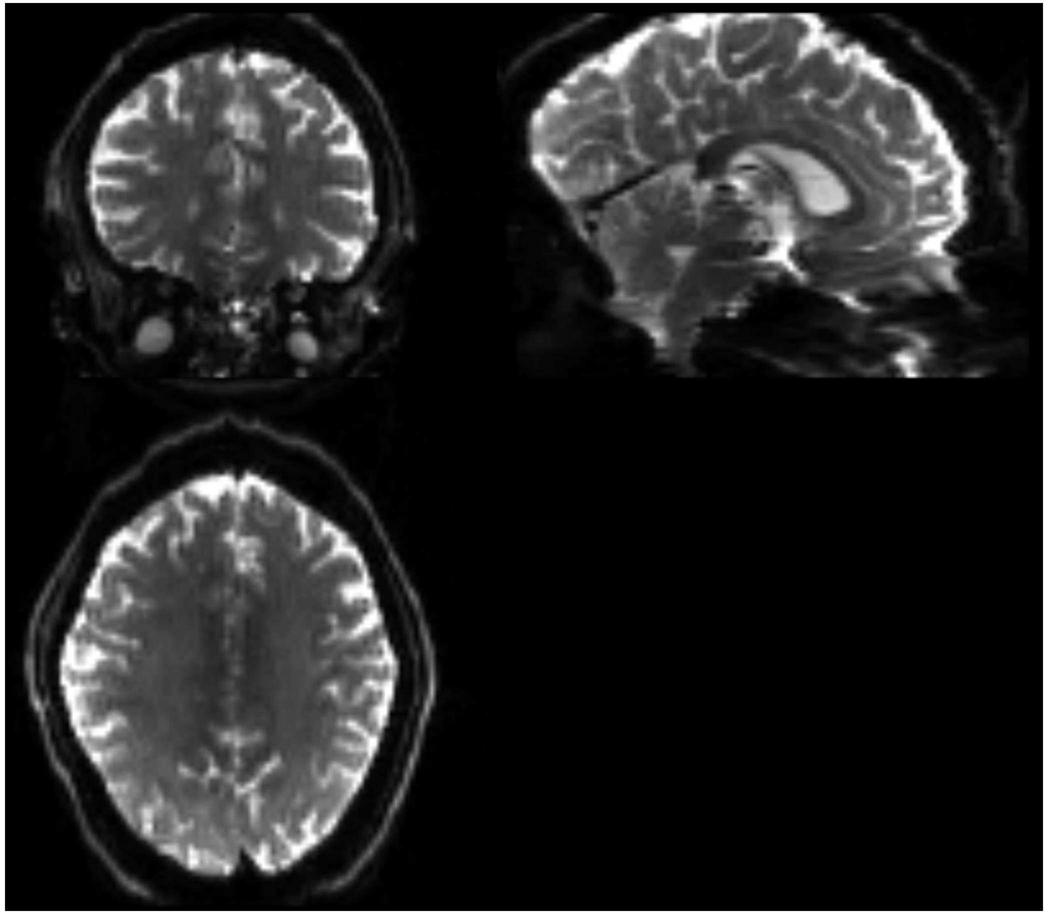
Task-based fMRI



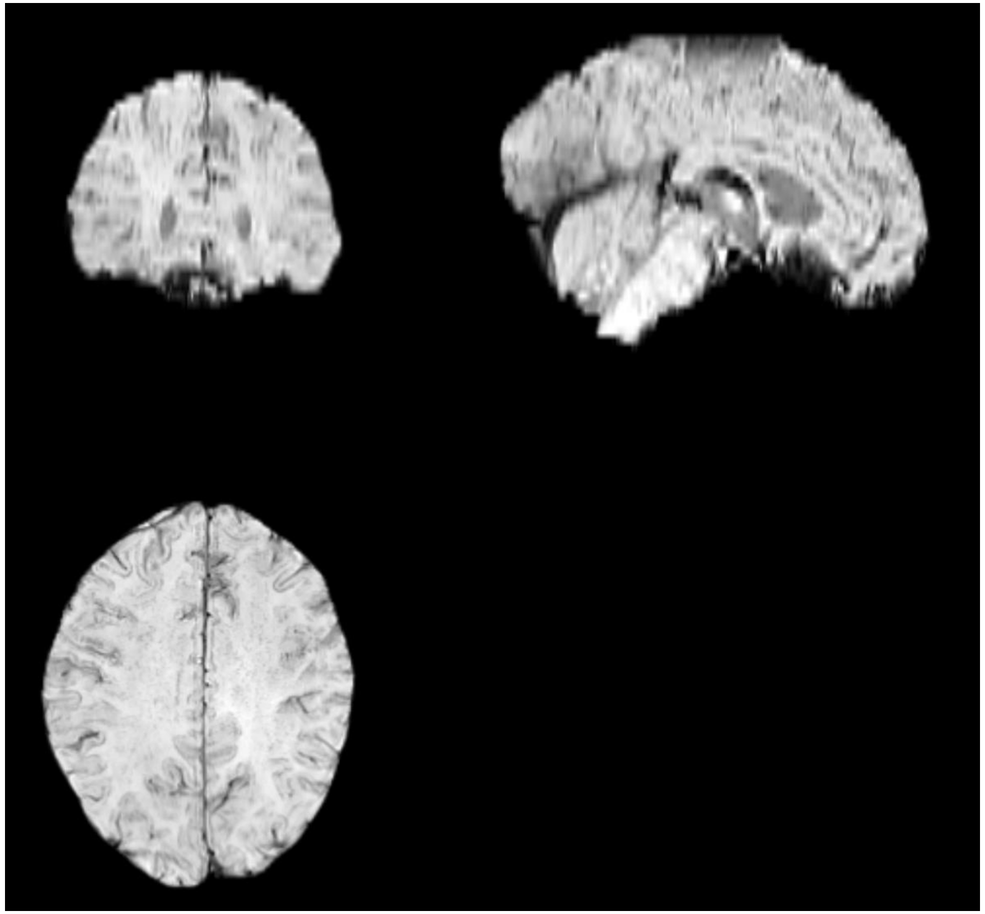
Resting state fMRI



dMRI



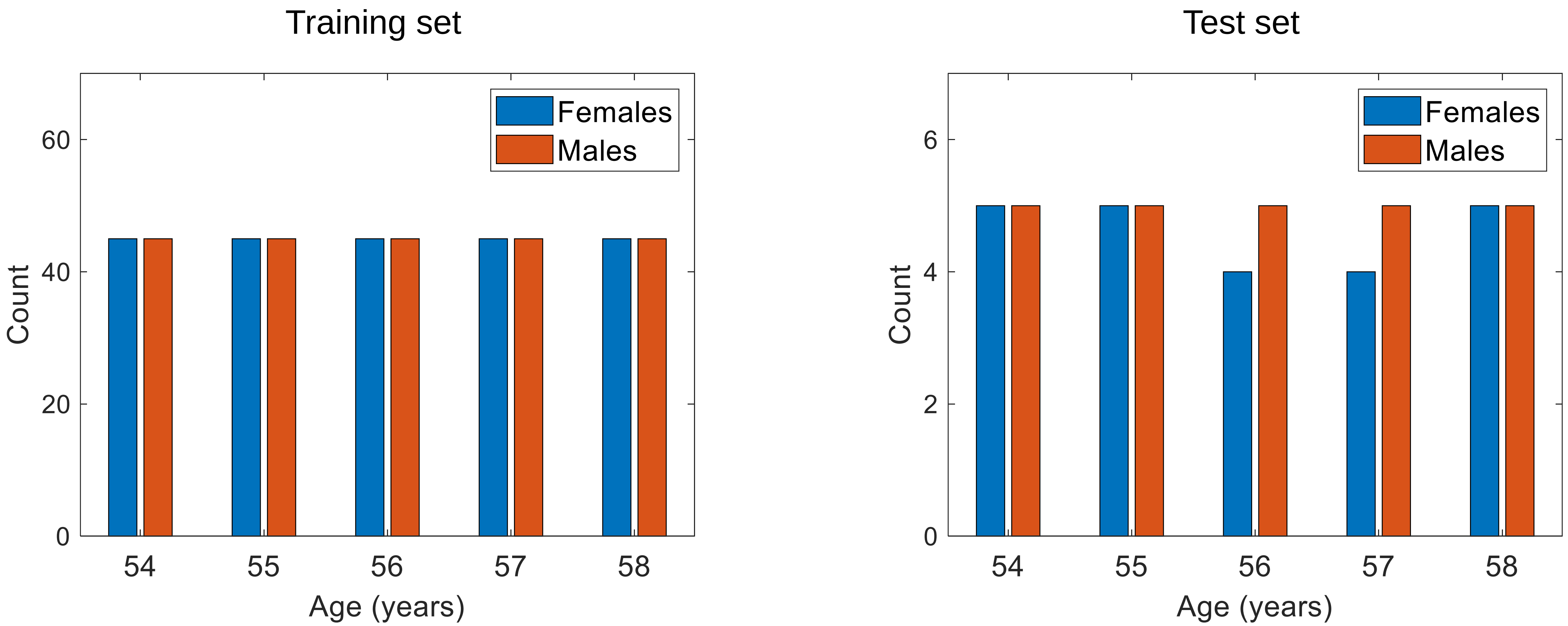
swMRI



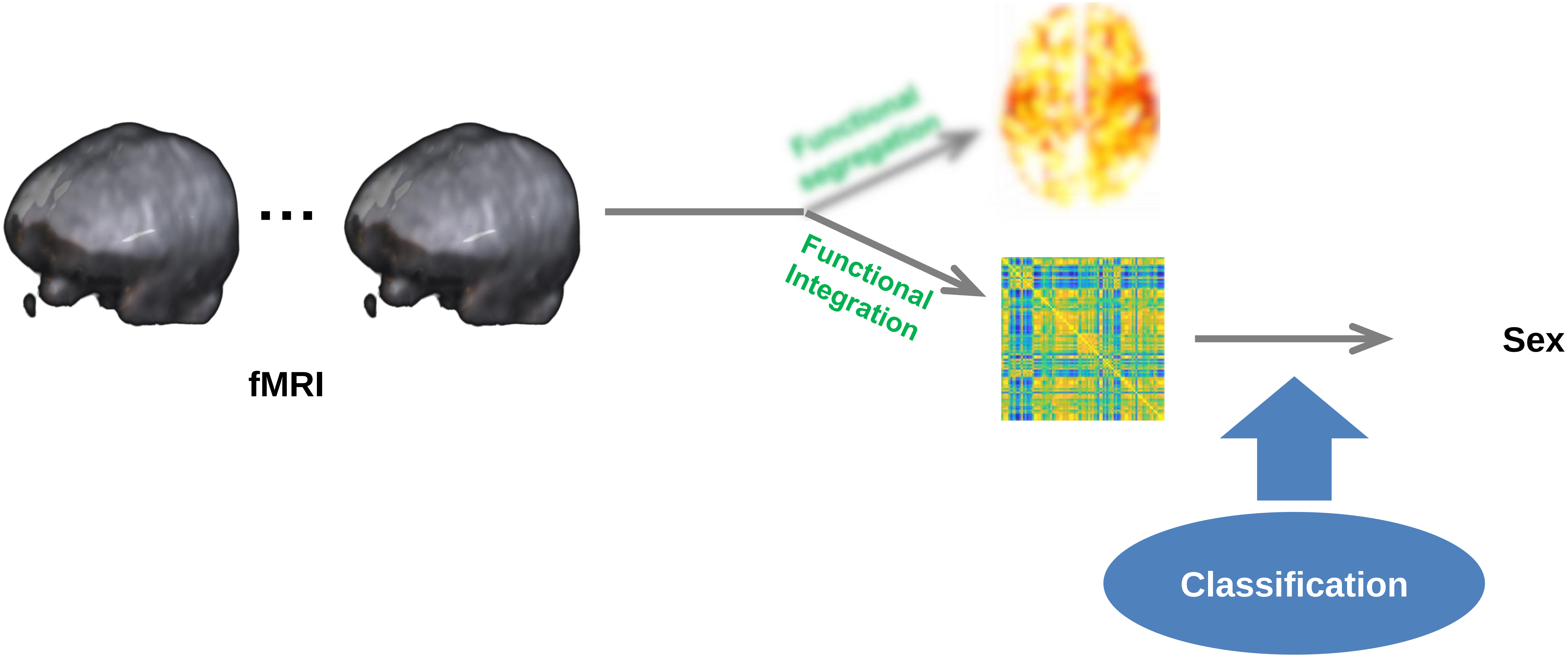
Dataset

- Part of UK Biobank dataset (n = 498)
  - Training set: n = 450
    - Functional brain networks from resting state fMRI data: train/{PC,GC}/001-450.csv
    - Network metrics for functional brain networks: train/{FuncBU_GE,FuncBU_LE,FuncBD_GE,FuncBD_LE}/001-450.csv
    - Structural brain networks from dMRI data: train/{Count,FA,MD,AD,RD}/001-450.csv
    - Network metrics for structural brain networks: train/{StruBU_GE,StruBU_LE}/001-450.csv

    - Age (in years): train/Subjects.csv: Age
    - Sex (0: female, 1: male): train/Subjects.csv: Sex
- Test set: n = 48
    - Functional brain networks from resting state fMRI data: test/{PC,GC}/001-048.csv
    - Network metrics for functional brain networks: test/{FuncBU_GE,FuncBU_LE,FuncBD_GE,FuncBD_LE}/001-048.csv
    - Structural brain networks from dMRI data: test/{Count,FA,MD,AD,RD}/001-048.csv
    - Network metrics for structural brain networks: test/{StruBU_GE,StruBU_LE}/001-048.csv
    - Age (in years): test/Subjects.csv: Age
    - Sex (0: female, 1: male): hidden



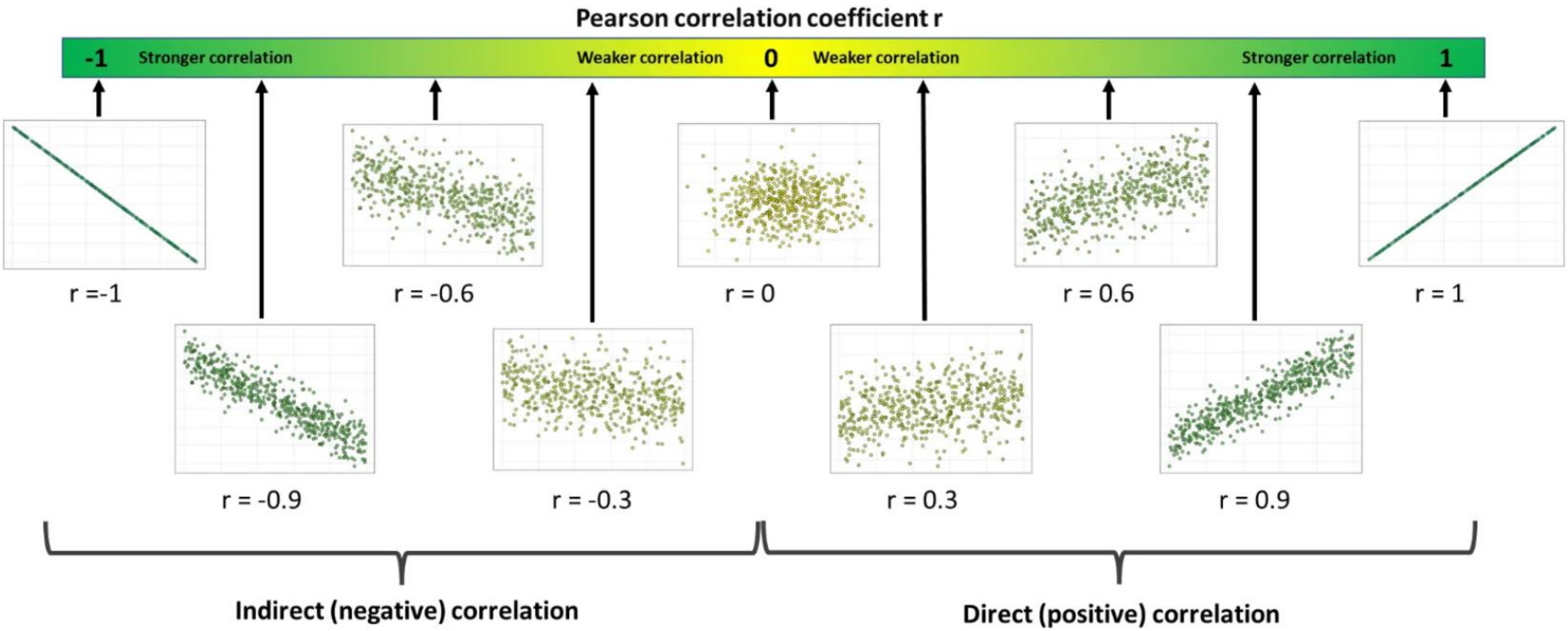
Distribution of age and sex for training and test sets



fMRI
Functional segregation
Functional Integration
Sex
Classification

- Preprocessing of resting state fMRI data
  - Correction for slice timing difference, head motion, and susceptibility artifact (B0 inhomogeneity-induced distortion)
  - Extraction of time series
  - Construction of functional brain networks
    - Nodes: pre-defined brain regions
    - Edges: connectivity (correlation or causality) between brain regions

- Pearson correlation
  - Measures the linear relationship (strength and direction) between two variables
  - Characteristics:
    - Range: -1 to 1
    - Symmetric
    - Does not imply causation
  - Computed by effectively standardizing the covariance between two variables (how the two variables change together) by dividing it by the product of the two standard deviations
  - Useful for identifying potential associations between variables, but not accounting for time lags or temporal order

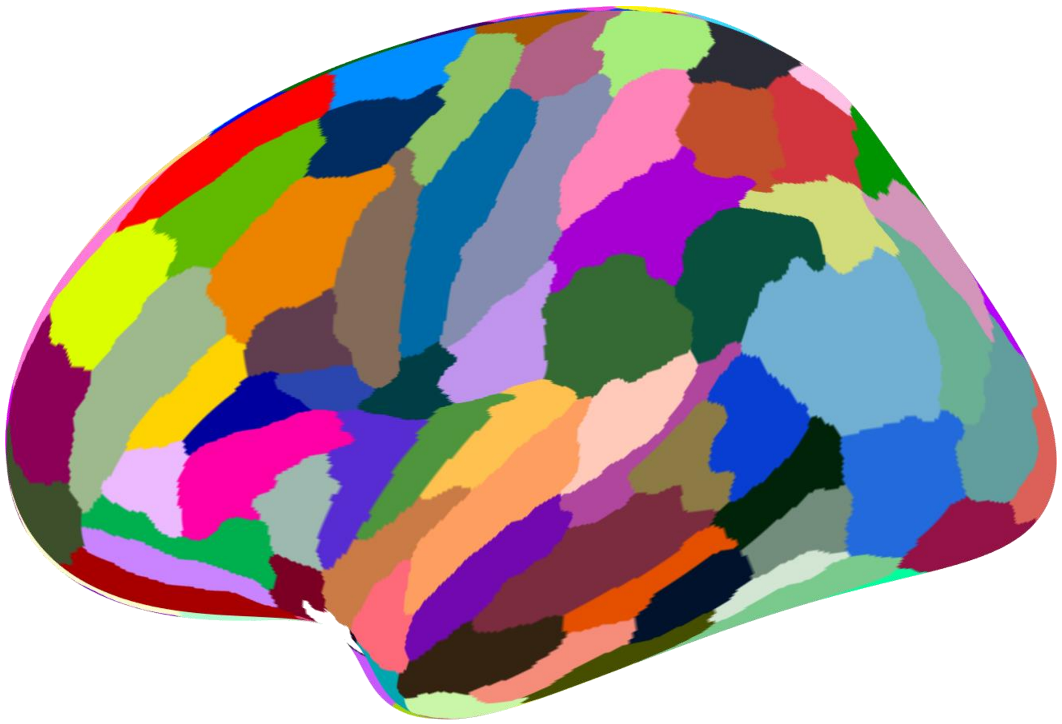


[https://medium.com/@anthony.demeusy/pearson-correlation-methodology-limitations-alternatives-part-1-methodology-42abe8f1ba90]

Range of a Pearson correlation coefficient

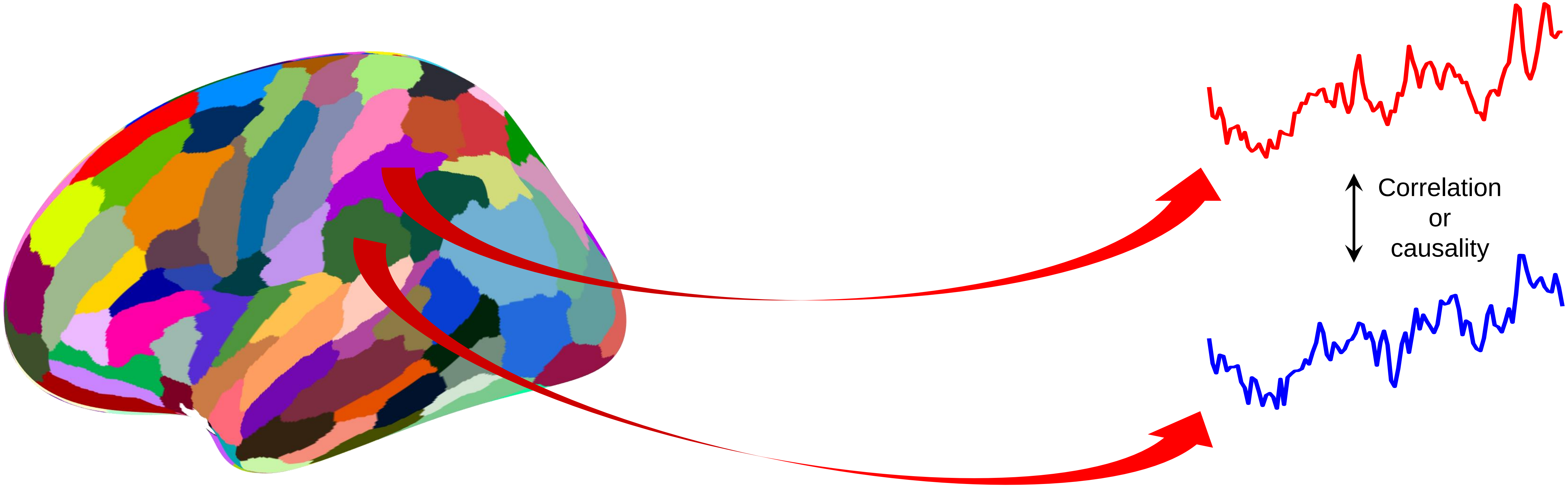
- Granger causality
  - Tests predictive causality (whether one time series is useful in forecasting another time series)
  - Characteristics:
    - Asymmetric
    - Considers temporal order and lags
  - Useful for understanding temporal relationships in time series, but not necessarity implying true causation

- Computed by comparing models with and without the potential causal variable based on autoregressive models that predict a variable's current value based on its own past values
  - Choose a maximum lag length
  - For each lag length up to the maximum:
    - Create two regression models: restricted model that predicts Y using only past values of Y (autoregressive) vs. unrestricted model that predicts Y using past values of both Y and X (augmented autoregressive)
    - Estimate both models and compute the *F*-statistic and its associated *p*-value
  - Compare the *p*-values to the chosen significance level
    - If the *p*-value is below the significance level (typically 0.05), conclude that X Granger-causes Y for that lag length
  - Repeat the process across different lag lengths

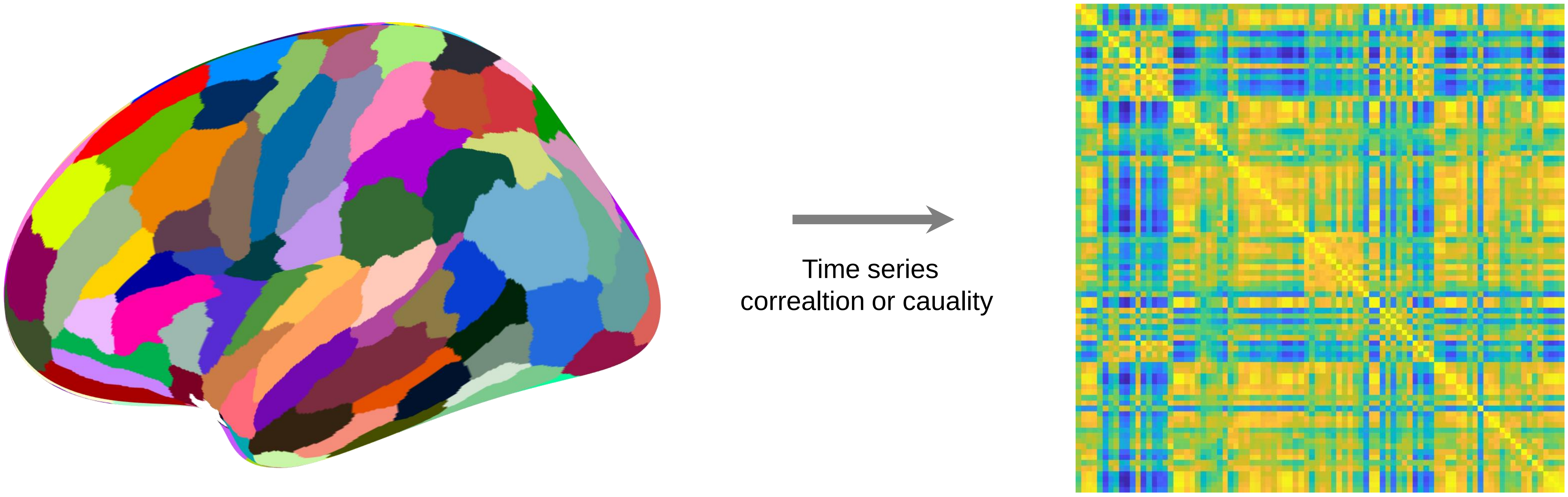


[Fan et al., 2016]

Definition of 246 nodes based on the Brainntome atlas

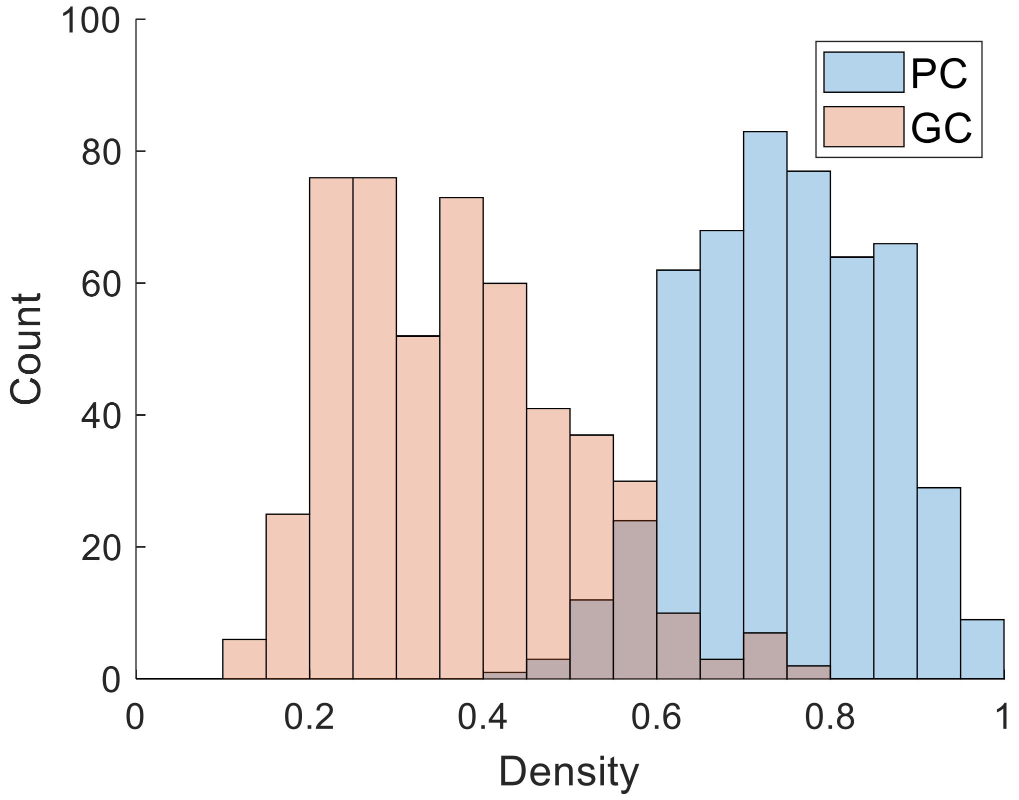


Estimation of edges between 246 nodes

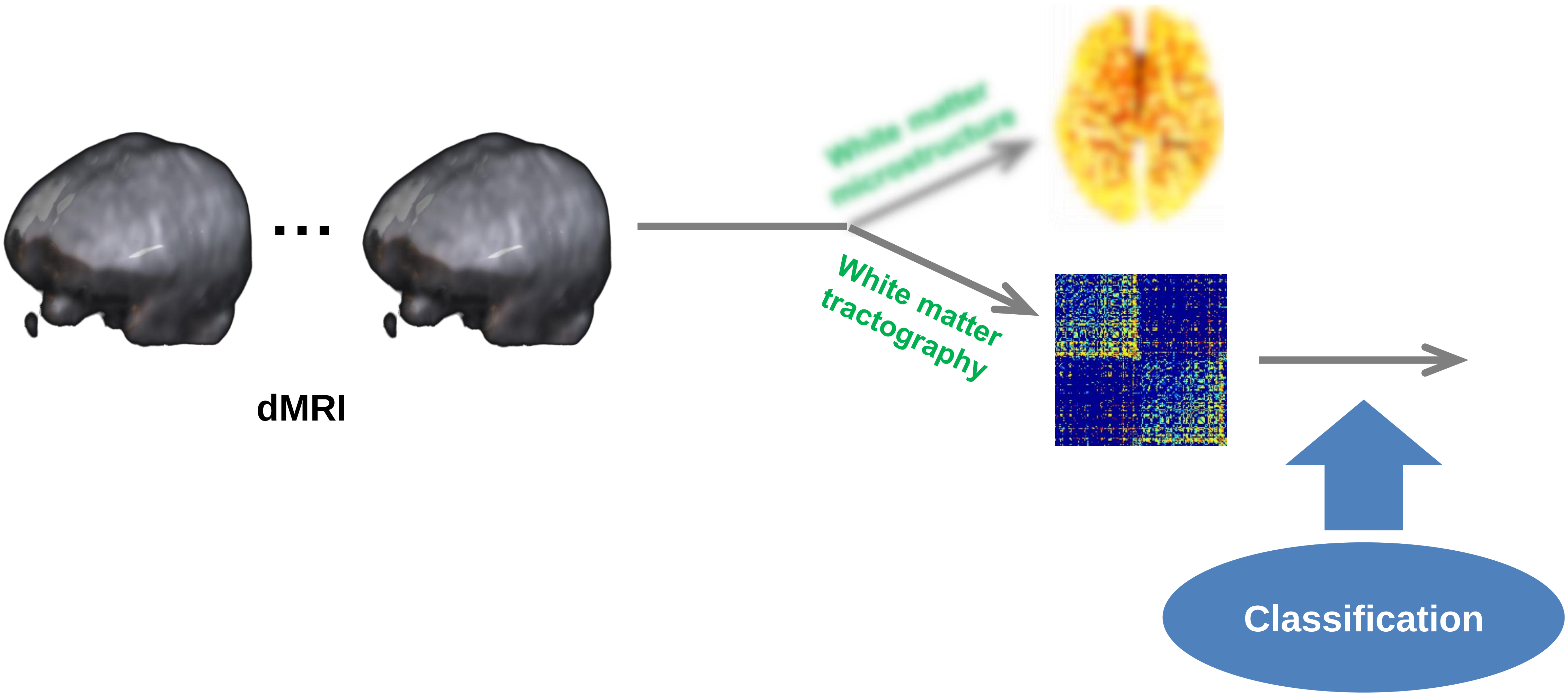


Functional brain network

- Functional brain networks from resting state fMRI data
  - Functional connectivity (Pearson correlation (PC)) network
    - Density = 0.748±0.108 (0.438 ~ 0.971)
  - Effective connectivity (Granger causality (GC)) network
    - Mean time lag = 2.064±0.165 (1.481 ~ 2.458)
    - Density = 0.369±0.130 (0.110 ~ 0.779)



Density of functional brain networks

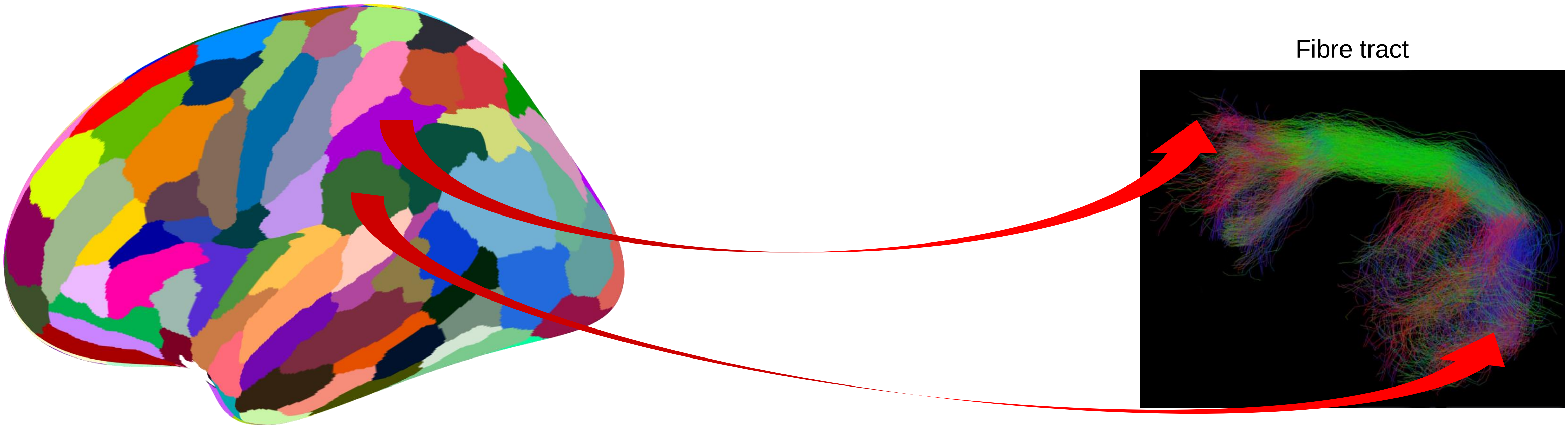


dMRI
White matter microstructure
White matter tractography
Sex
Classification

- Preprocessing of dMRI data
  - Correction for head motion, eddy current-induced distortion, and susceptibility artifact (B0 inhomogeneity-induced distortion)
  - Diffusion tensor modelling
  - White matter tractography
  - Construction of structural brain networks

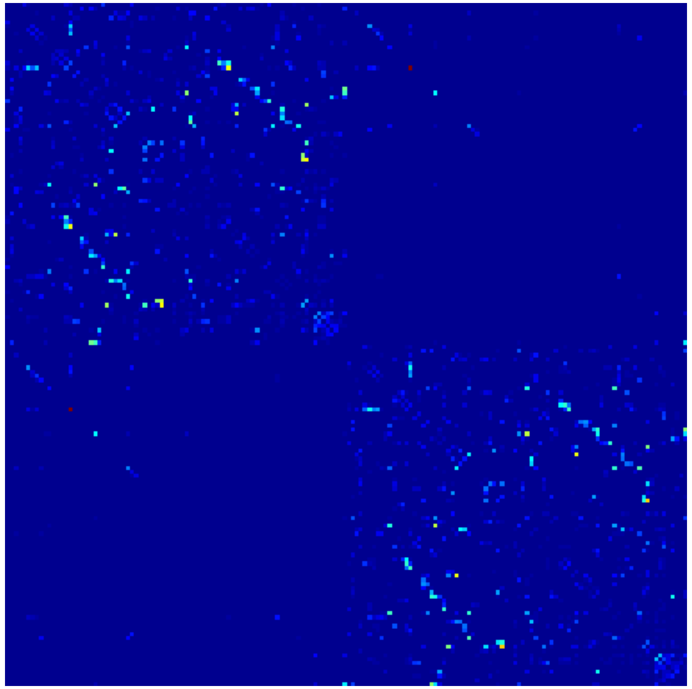
[Fan et al., 2016]

Definition of 246 nodes based on the Brainntome atlas



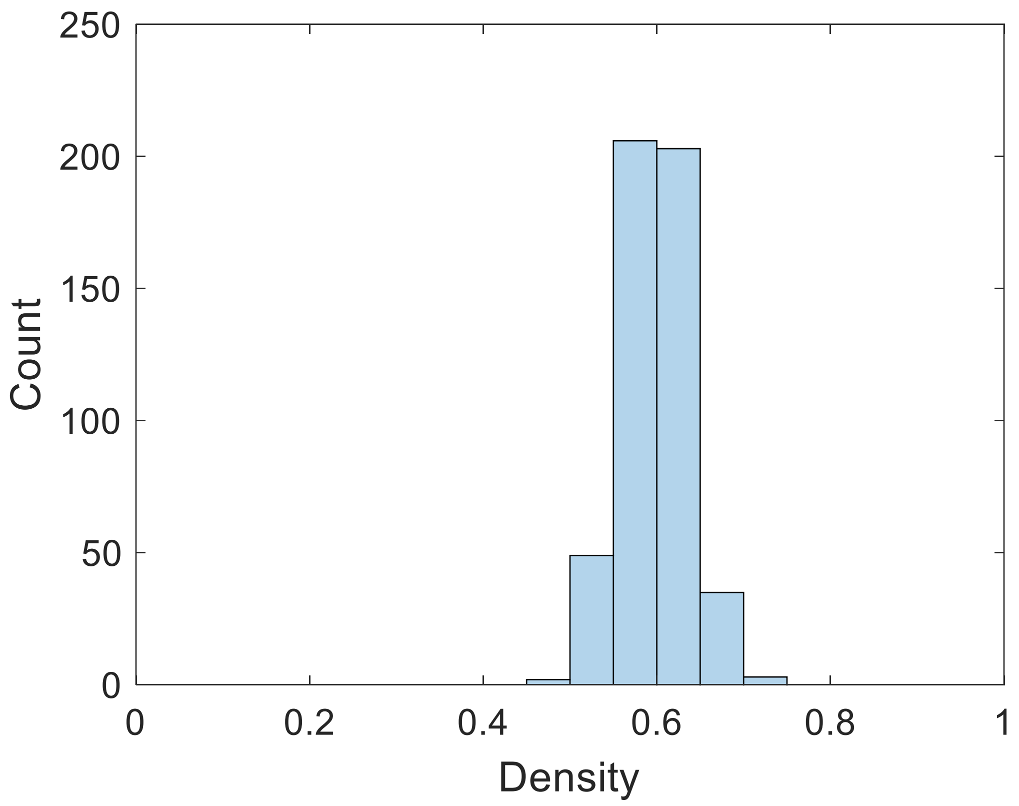
Estimation of edges between 246 nodes

White matter tractography

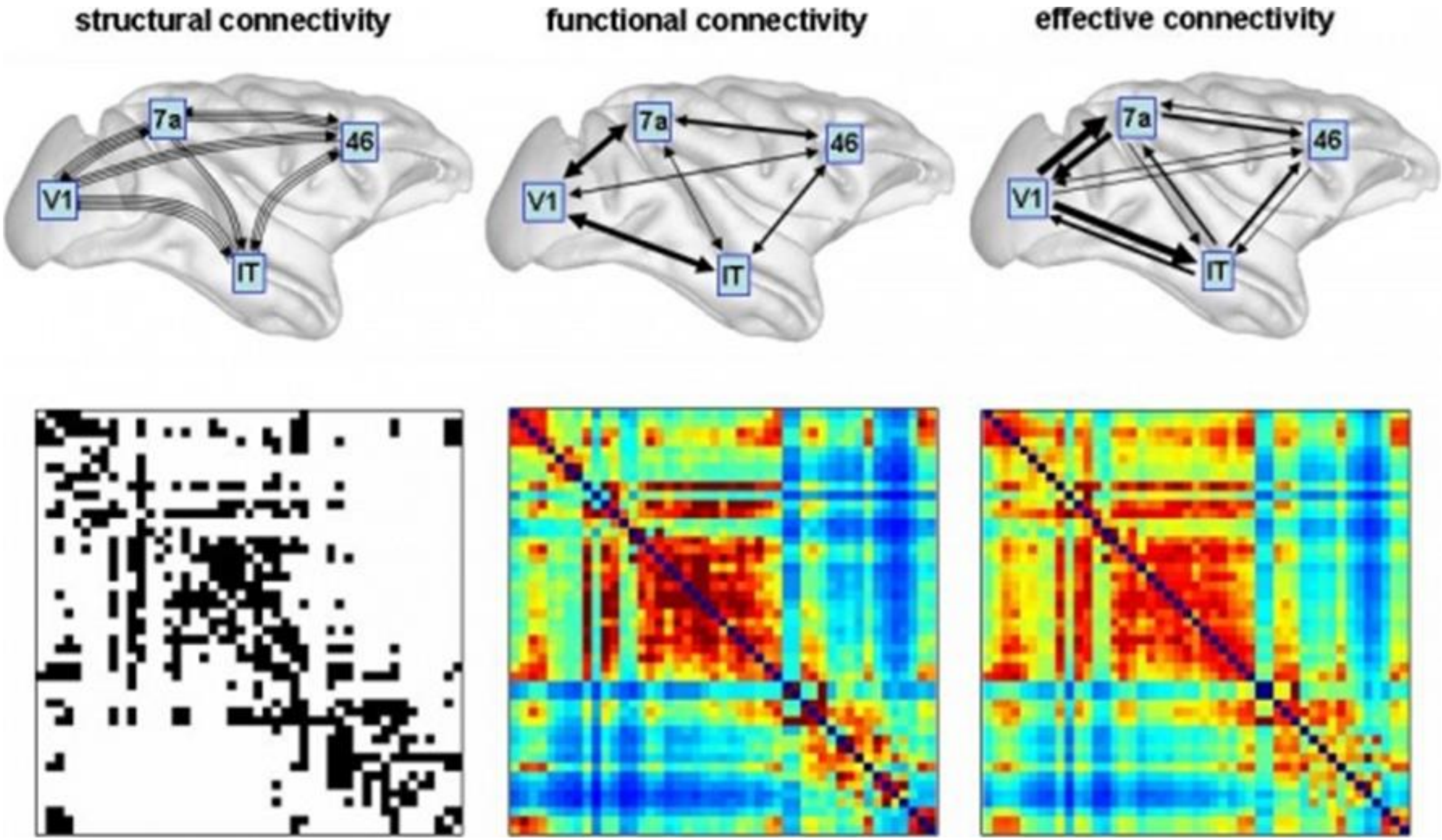


Structural brain network

- Structural brain networks from dMRI data
  - Structural connectivity (Tract Count (Count), Fractional Anisotropy (FA), Mean Diffusivity (MD), Axial Diffusivity (AD), or Radial Diffusivity (RD)) network
    - Density = 0.598 ± 0.038 ($0.491 \sim 0.728$)



Density of structural brain networks



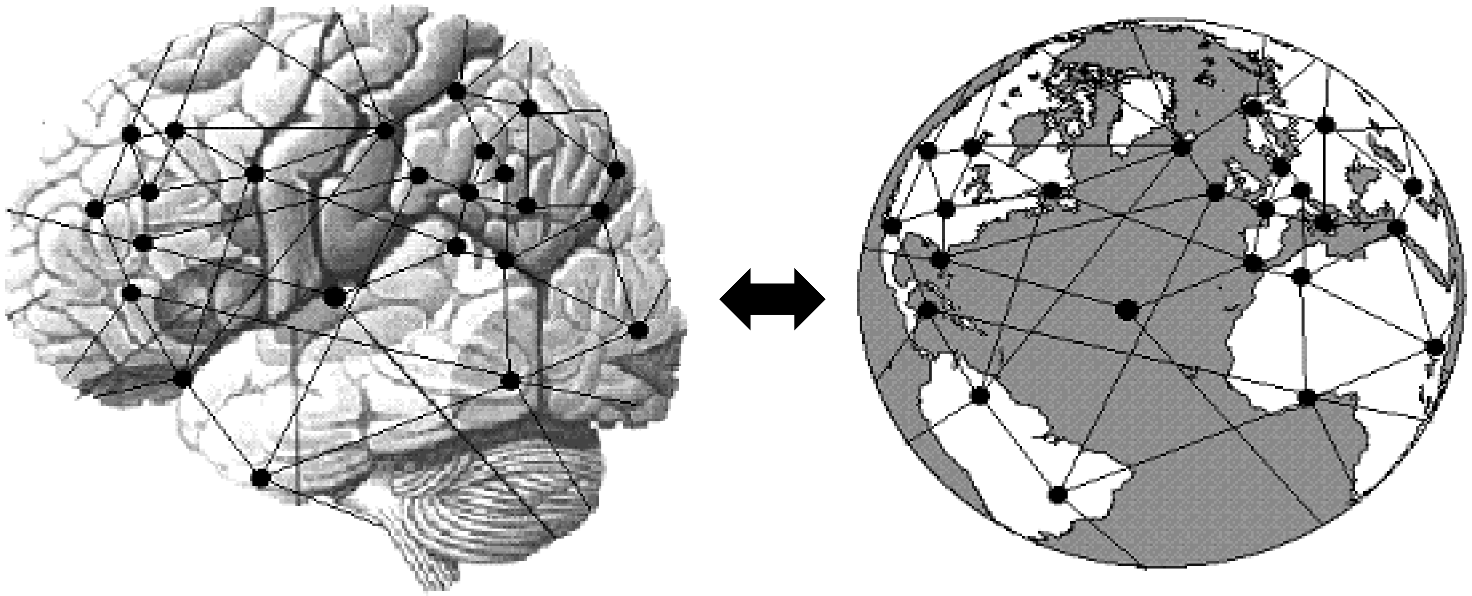
[Honey et al., 2007]

Three modes of brain connectivity

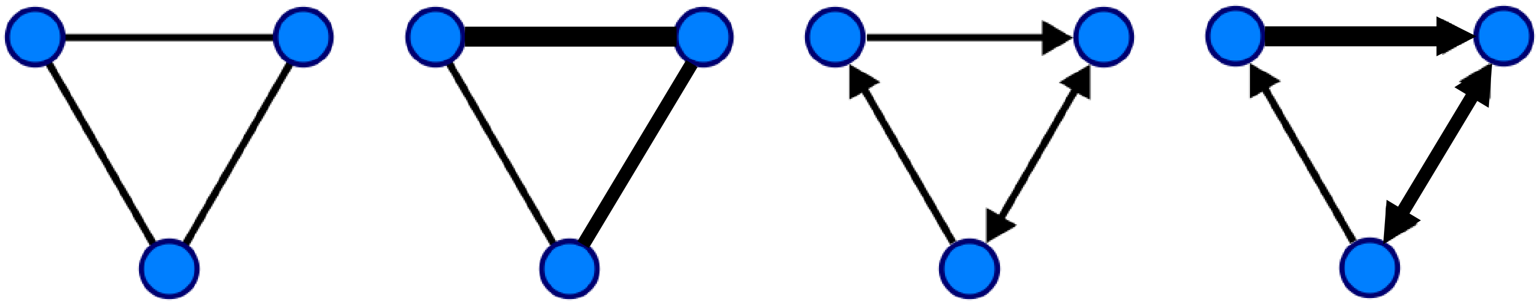
- Target variable
  - Sex (0: female, 1: male)
- Sex classification performance
  - Accuracy for the test set ($n = 48$)
    - Proportion of correct classifications
    - Ranges from 0 to 1 (higher is better)

Graph-theoretical Analysis

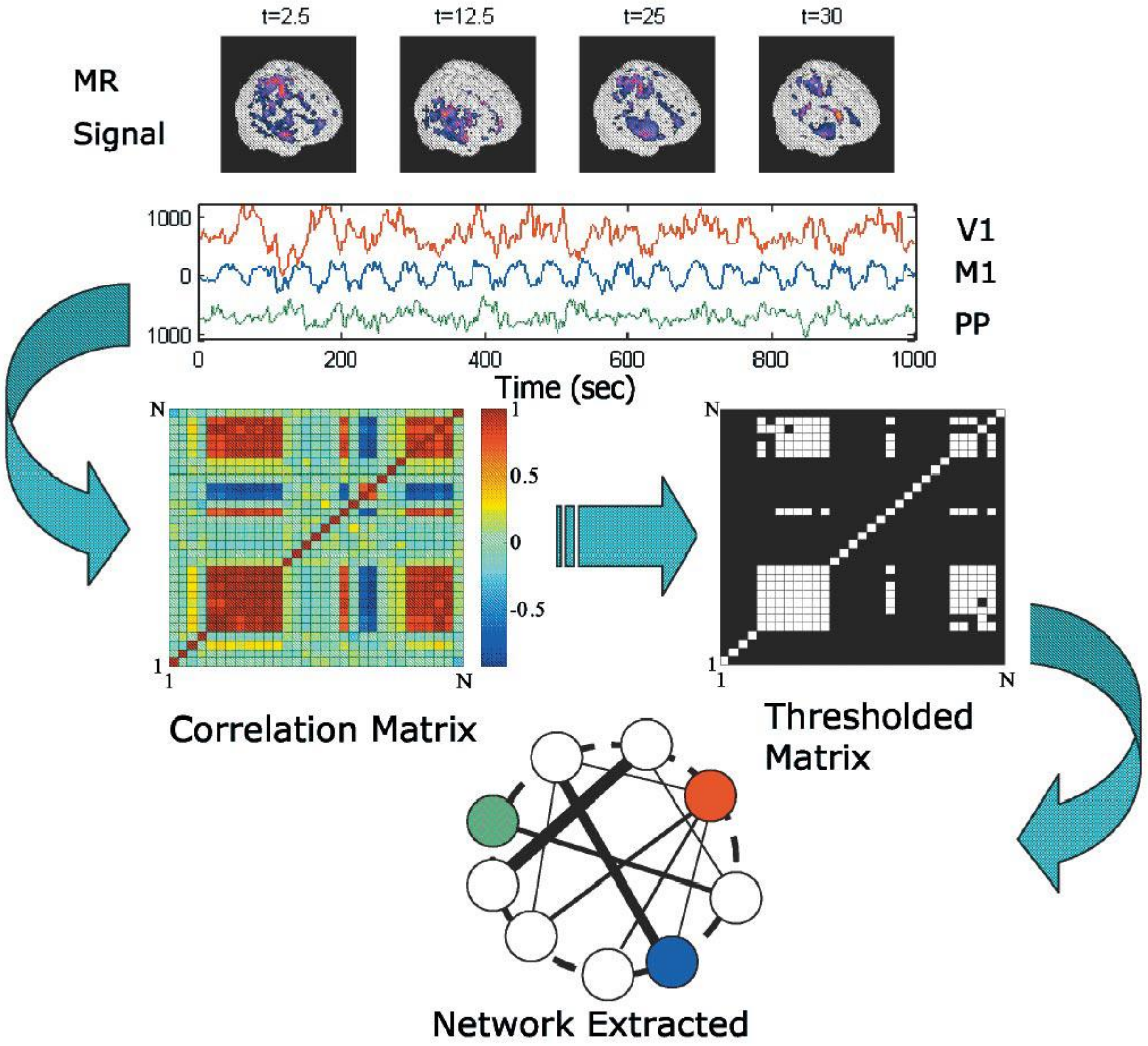
- Allows to conceptualise and analyse the brain as a complex network, similar to other complex networks



- Graph or network
  - Set of nodes and edges
    - Binary or weighted
    - Undirected or directed

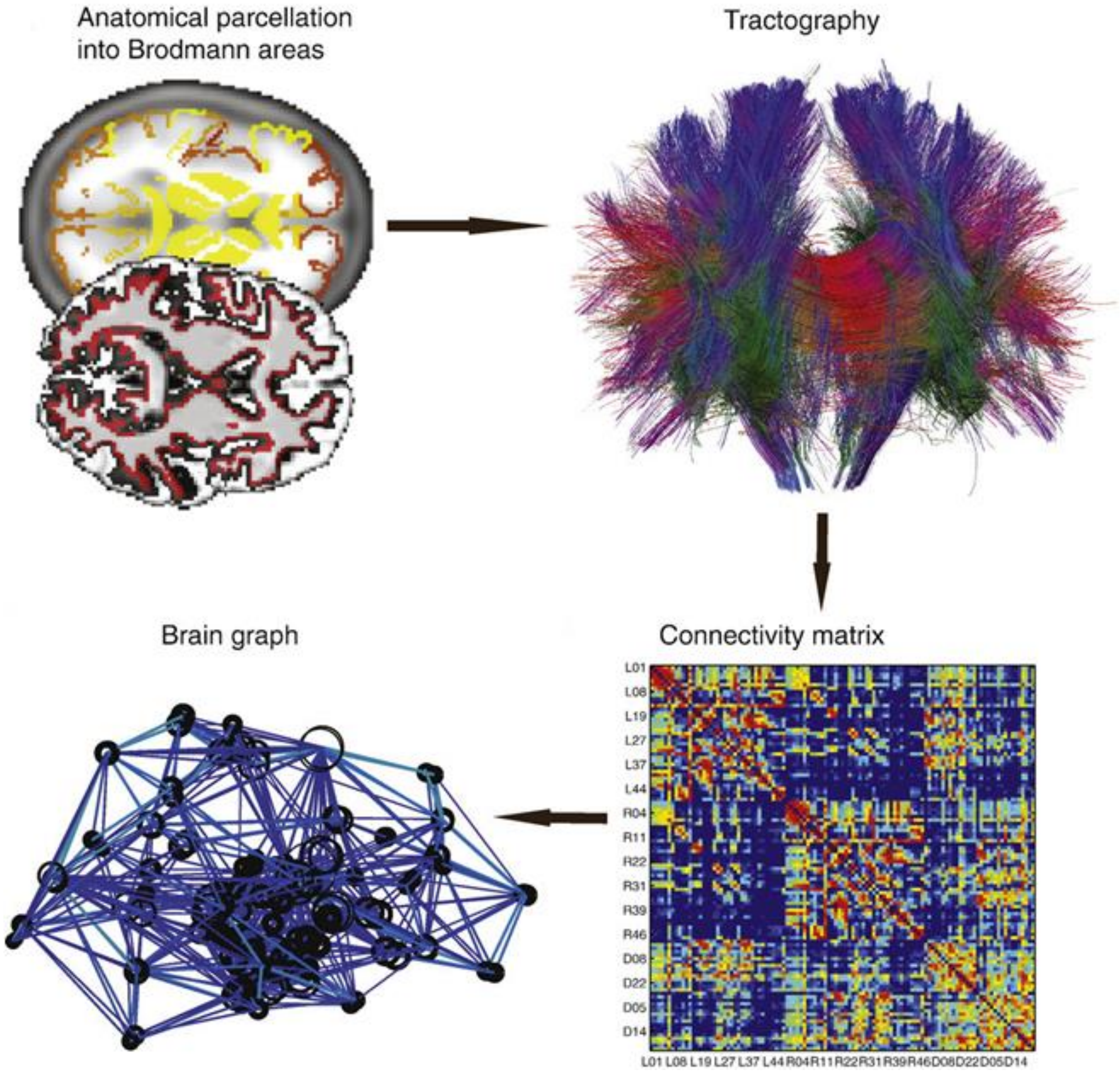


- Construction of brain networks
  - Define nodes
  - Estimate a continuous measure of association between nodes
  - Generate an association matrix by compiling all pairwise associations between nodes
  - If needed, apply a threshold to each element of the matrix to produce a binary matrix



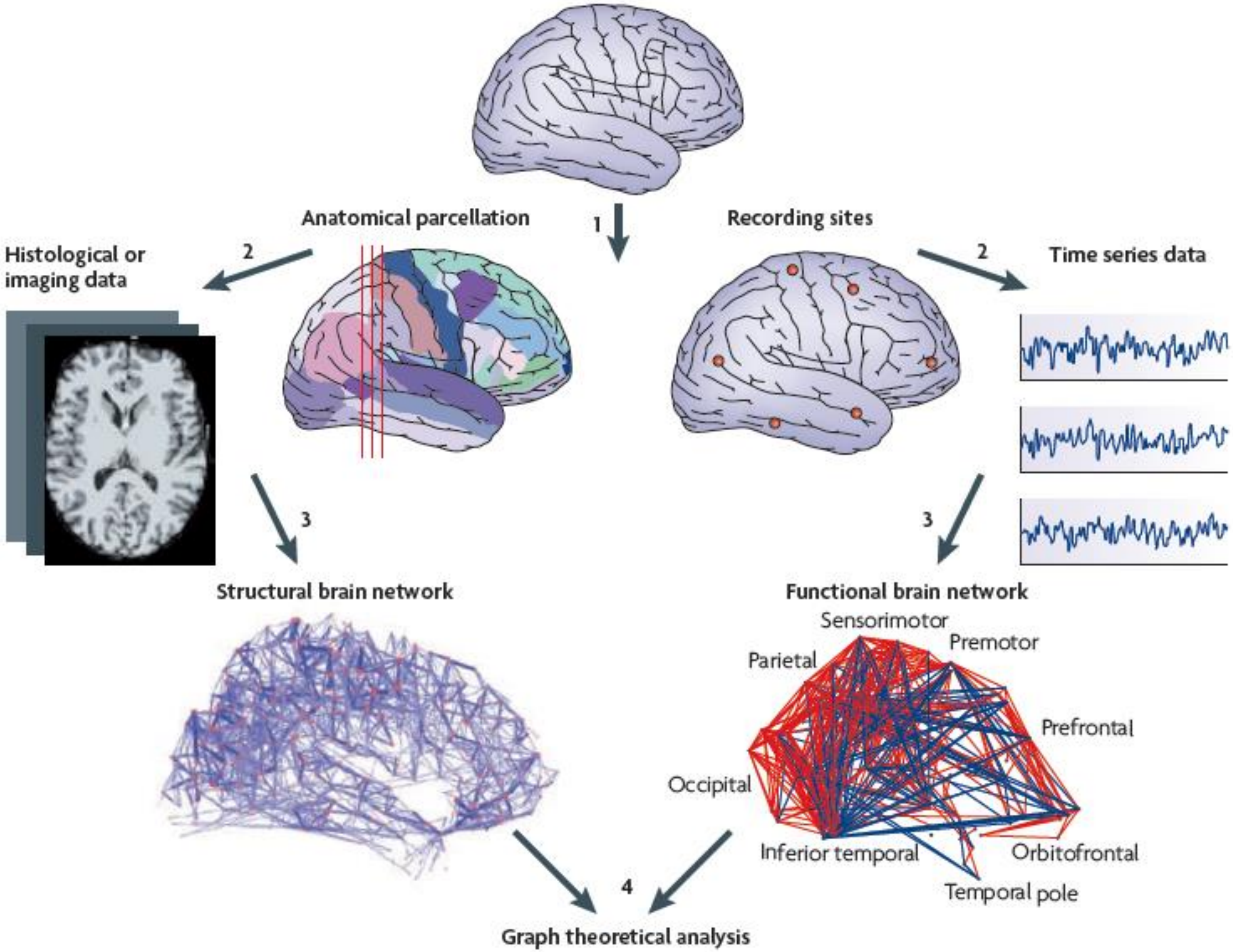
[Eguíluz et al., 2005]

Contruction of functional brain networks



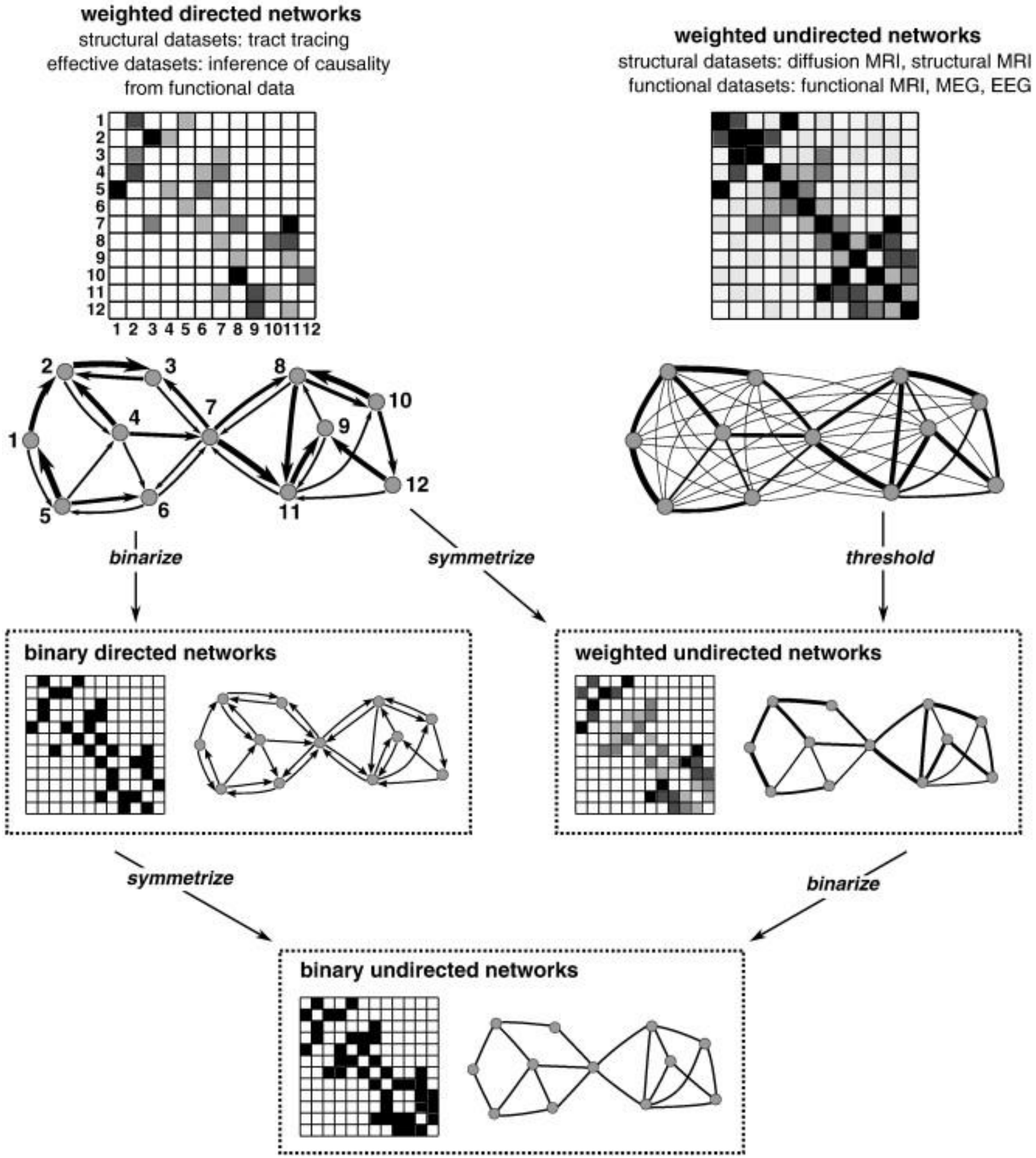
[Vaessen et al., 2010]

Construction of structural brain networks



[Bullmore & Sporns, 2009]

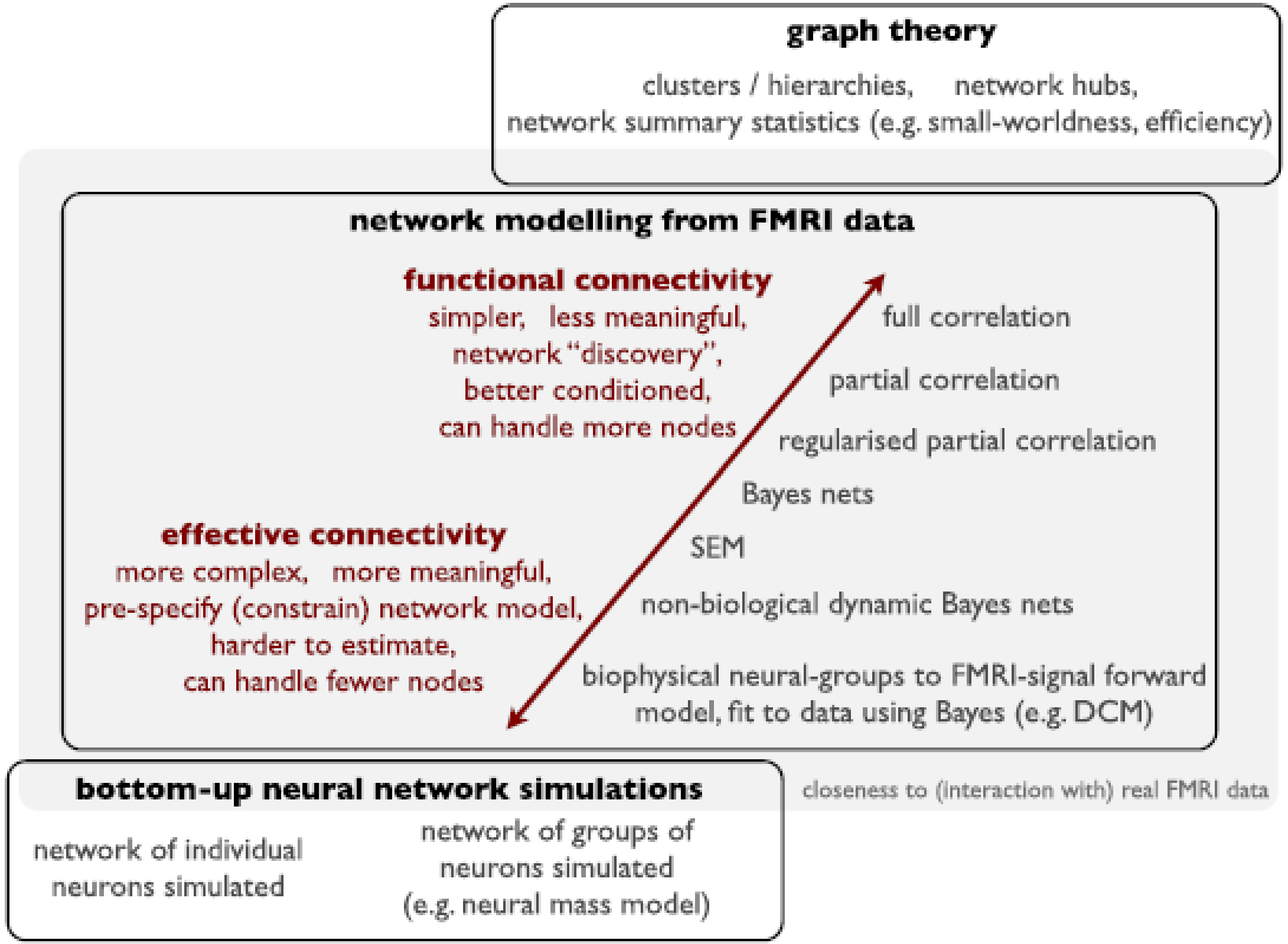
Parallel construction of functional and structural brain networks in an individual



[Rubinov and Sporns et al., 2010]

Weighted/binary directed/undirected networks

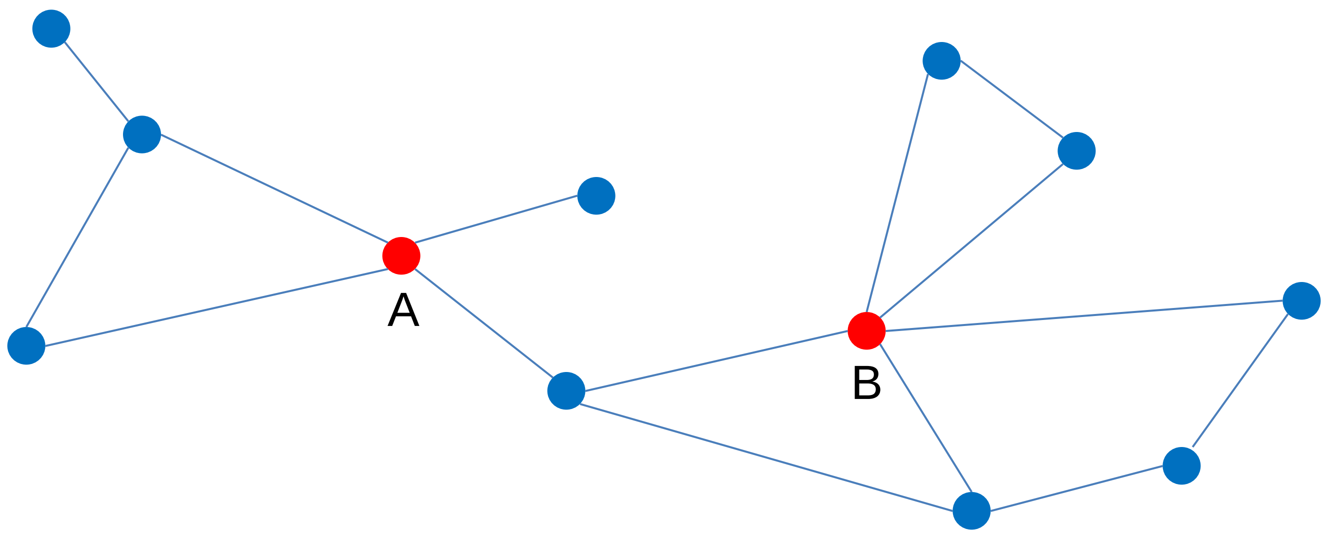
- Graph-theoretical analysis
  - Enables to study the brain's organisation using the same mathematical principles and metrics used to examine various complex networks across different domains
  - Provides a powerful tool for understanding the brain's connection topology in both health and disease states



[Smith, 2012]

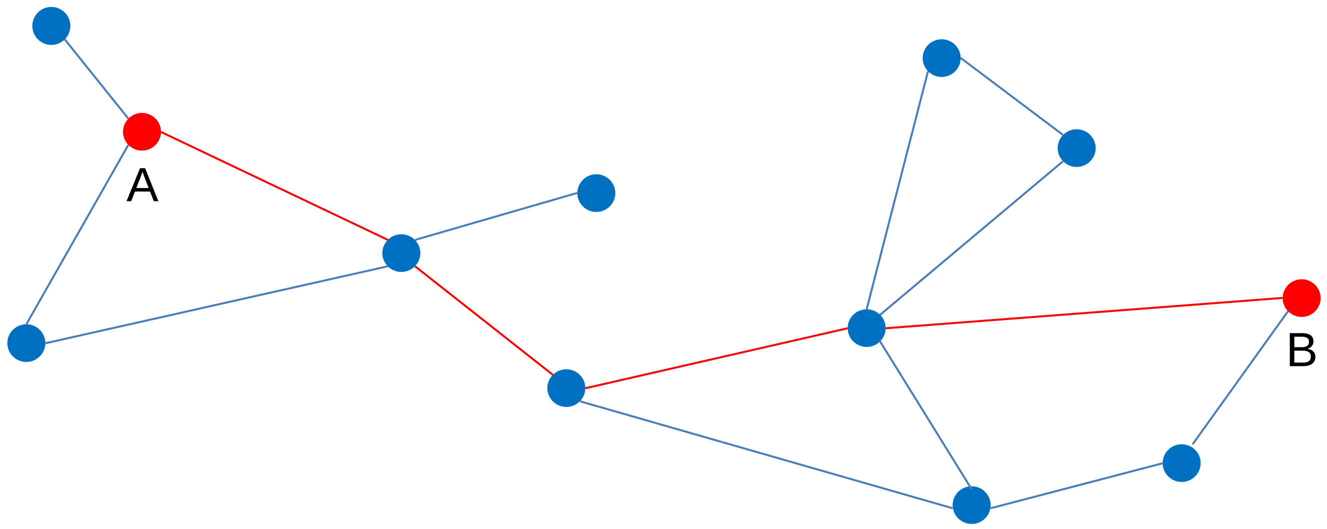
Bridging graph-theoretical analysis and network modelling of the brain

- Network metrics
  - Degree (of a node): number of edges incident to a node



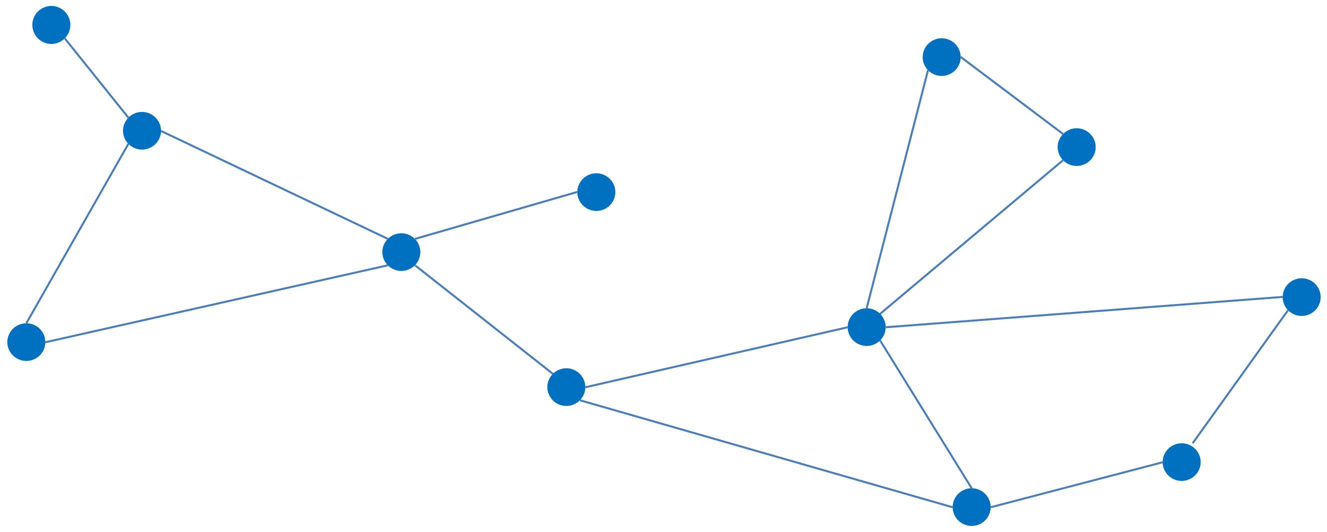
- Degree of node A = 4
- Degree of node B = 5

– Efficiency (in the communication between nodes): inverse of the shortest distance between a pair of nodes [Latora & Marchiori, 2001]



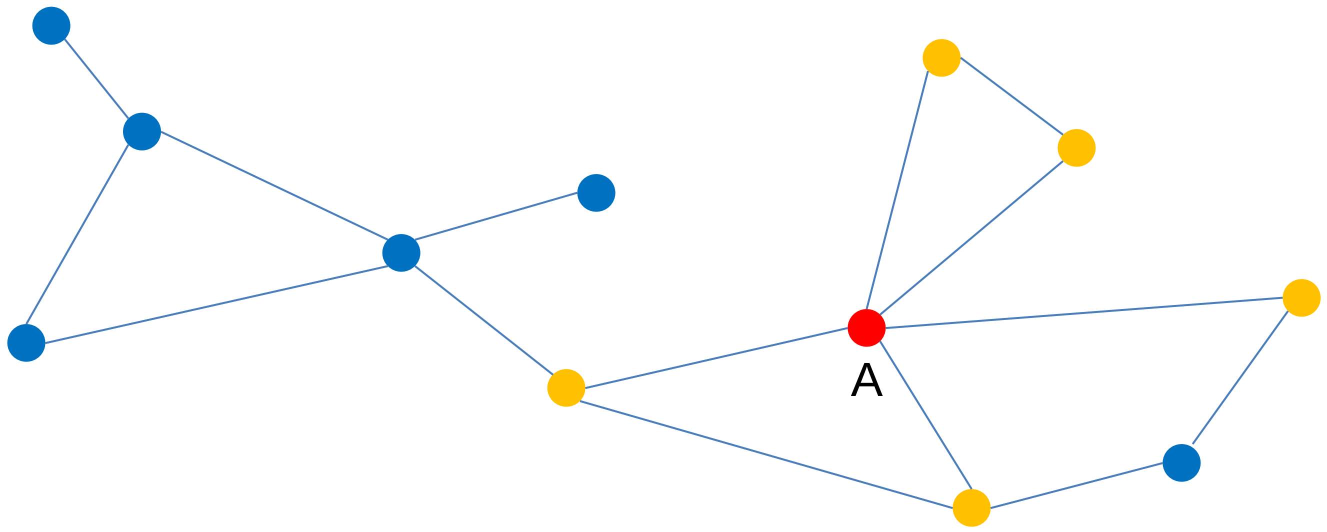
- Shortest distance between nodes A and B = 4
- Efficiency between nodes A and B = 1/4

– Global efficiency (of a network), E_{glob}: average efficiency across a whole network

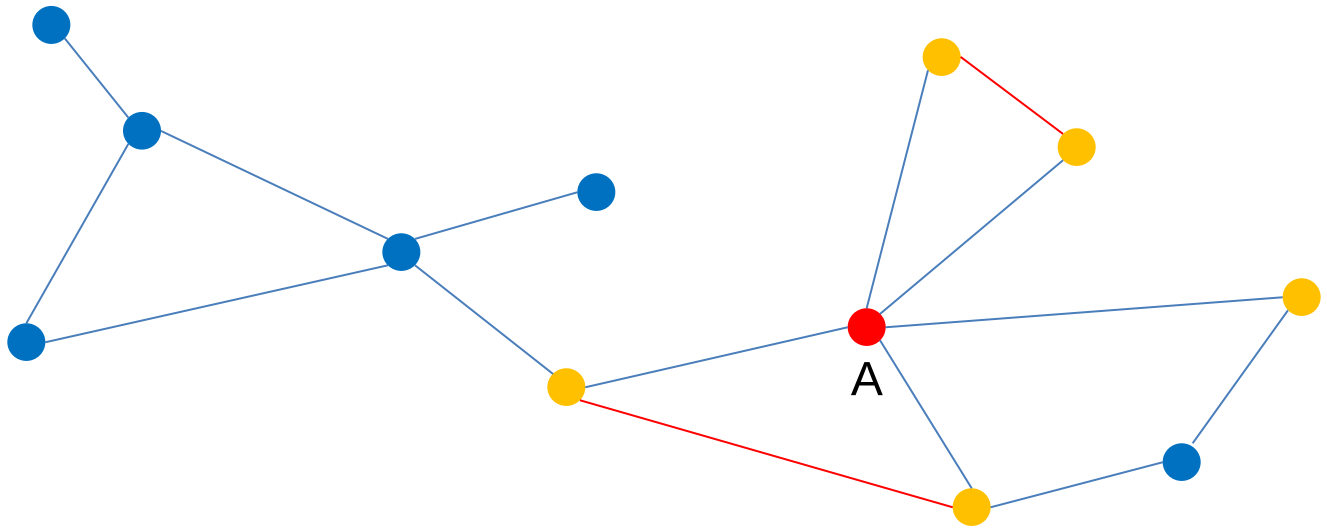


- Represents efficiency in information flow on a global scale

– Nearest neighbours (of a node): nodes adjacent to a node not including itself



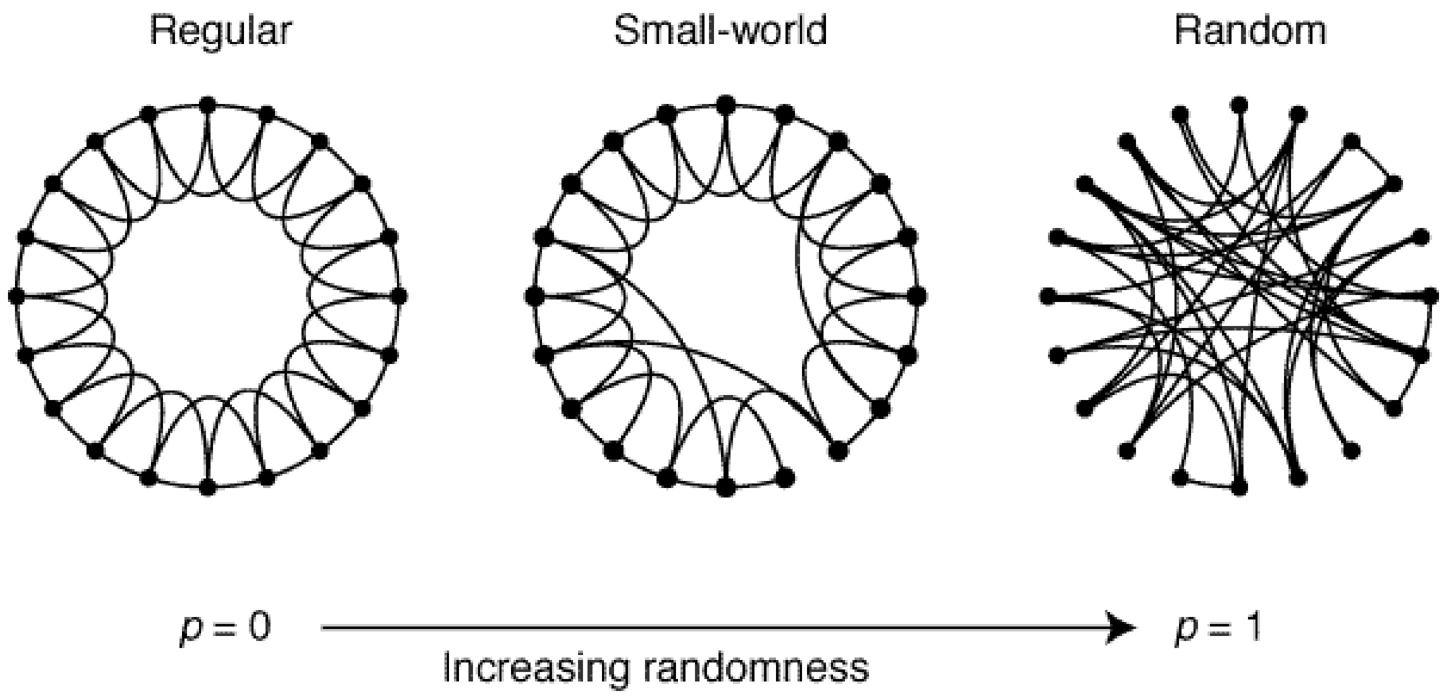
– Efficiency (of a subnetwork): average efficiency across the local subnetwork consisting of the nearest neighbours of a node



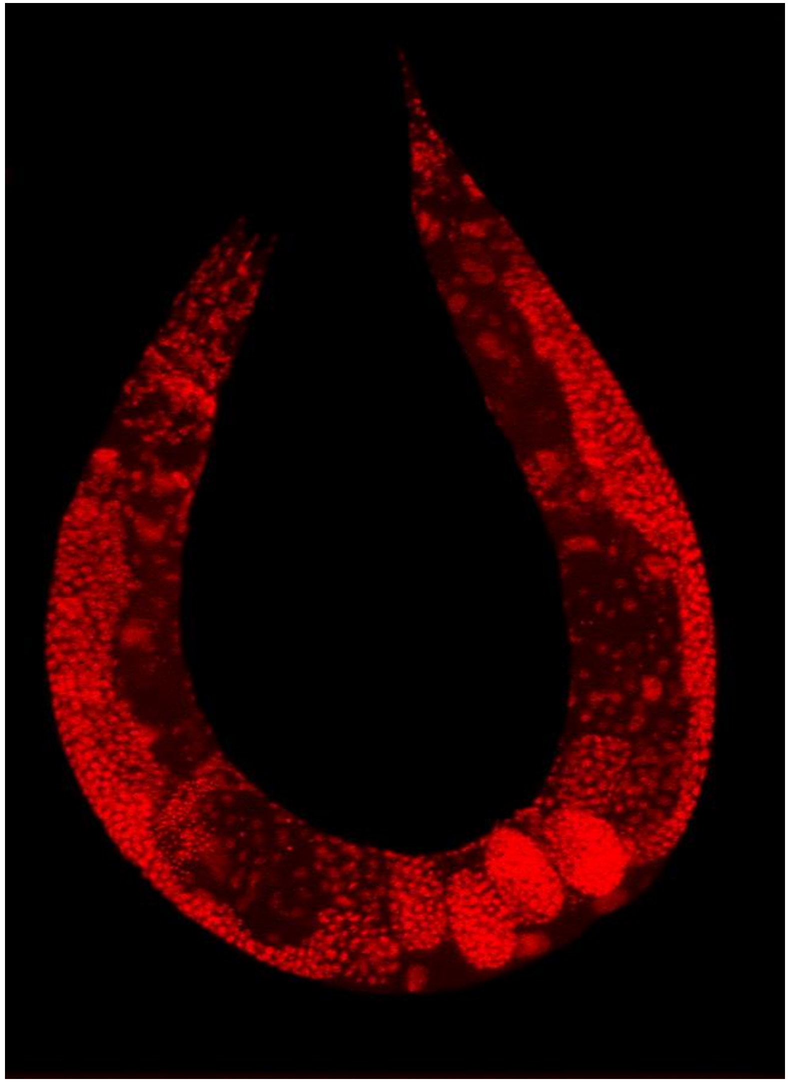
– Local efficiency (of a network), E_{loc}: average efficiency of local subnetworks

- Represents efficiency in information flow on a local scale

- Small-world network
  - Intermediate between regular and random networks
  - Has high global efficiency and local efficiency



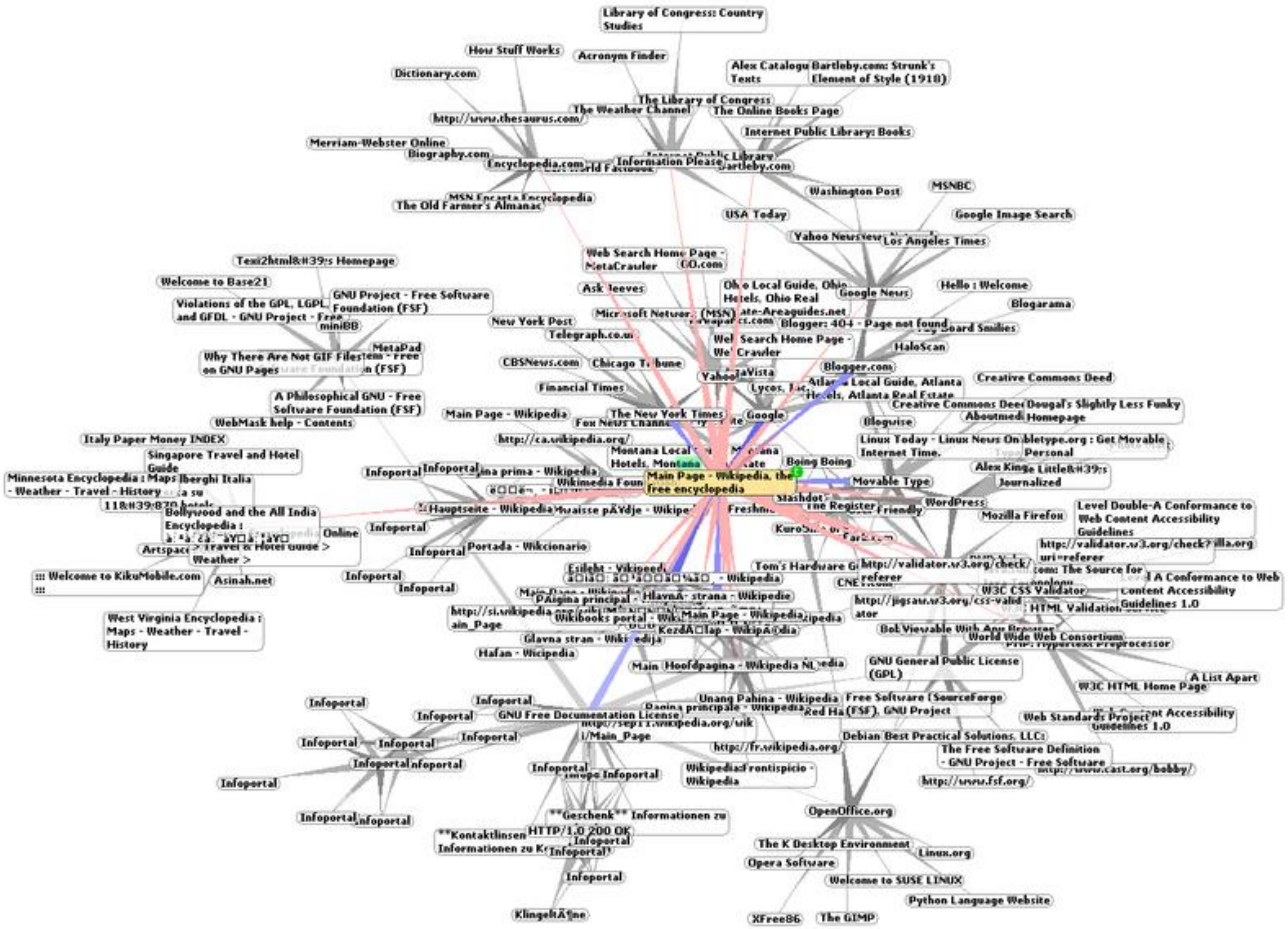
[Watts & Strogatz, 1998]



$E_{glob} = 0.46$
$E_{loc} = 0.47$

[Latora & Marchiori, 2001]

Small-worldness of neural networks: nervous system of C. elegans

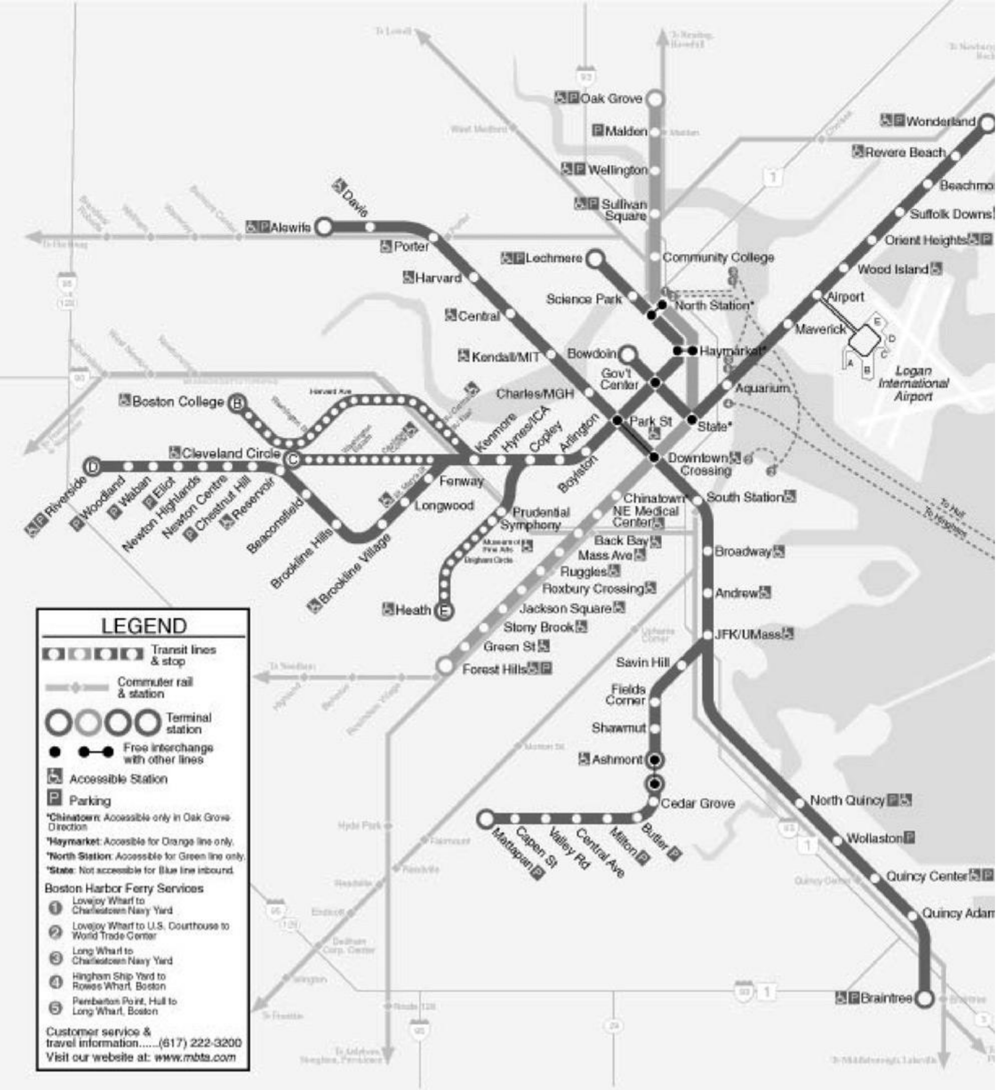


E_{glob} = 0.28

E_{loc} = 0.36

[Latora & Marchiori, 2001]

Small-worldness of communication networks: World Wide Web



Boston underground transportation system (weighted):

$E_{glob} = 0.63$

$E_{loc} = 0.03$

Boston underground transportation system + Boston bus transportation system (weighted):

$E_{glob} = 0.72$

$E_{loc} = 0.46$

[Latora & Marchiori, 2001]

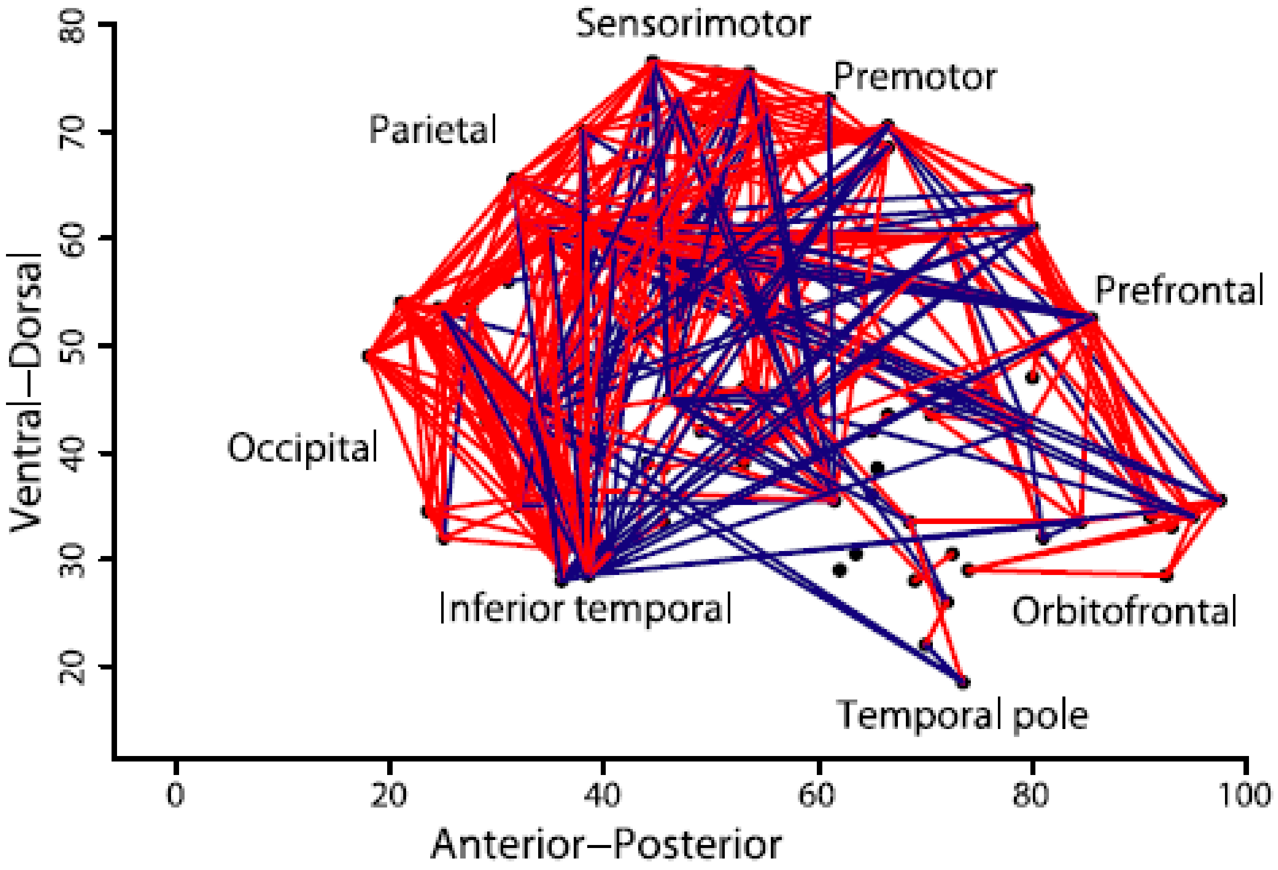
Small-worldness of transport networks: Boston transportation system

– Brain as a small-world network

Data	L	C	λ	γ	E_{glob}	E_{loc}	Cost
Macaque visual cortex	1.73	0.53	1.04	1.47	—	—	—
Macaque whole cortex	2.38	0.46	1.17	3.06	0.52	0.70	0.18
Cat cortex	1.81	0.55	1.06	1.77	0.69	0.83	0.38

Data	Connectivity Metric	*N*	*k*	*d*	*C*	*L*	λ	γ	σ
Macaque cortex (Stephan and others 2000)	Tract-tracing (binary)	39	6.1	0.15	0.38	2.17	1.01	2.46	2.44
Human MEG (Stam 2004)	Synchronization likelihood	126	15	0.12	~0.5	~5.0	~1.8	~4.2	~2.3
Human EEG (Micheloyannis and others 2006)	Synchronization likelihood	28	5	0.18	~0.4	~4.1	~1.0	~2.0	~2.0
Human fMRI (Eguiluz and others 2005)	Correlation	31,503	13.4	$4.3 \cdot 10^{-4}$	0.14	11.4	2.92	325	111
Human fMRI (Salvador, Suckling, Coleman, and others 2005)	Partial correlation	90	5.7	0.06	0.25	2.82	1.09	2.08	1.91
Human fMRI (Achard and others 2006)	Wavelet correlation	90	4.5	0.05	0.53	2.49	1.09	2.37	2.18

[Bassett & Bullmore, 2006]



[Achard et al., 2006]

Small-worldness of the human brain

- Network metrics for functional brain networks
  - Edge-wise efficiencies of a binary undirected network (FuncBU_GE)
  - Node-wise subnetwork efficiencies of a binary undirected network (FuncBU_LE)
  - Edge-wise efficiencies of a binary directed network (FuncBD_GE)
  - Node-wise subnetwork efficiencies of a binary directed network (FuncBD_LE)
- Network metrics for structural brain networks
  - Edge-wise efficiencies of a binary undirected network (StruBU_GE)
  - Node-wise subnetwork efficiencies of a binary undirected network (StruBU_LE)