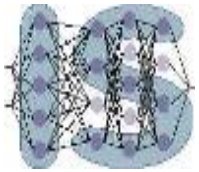
Federated, privacy-preserving learning of large-scale probabilistic graphical models in life sciences

Antal Péter

Computational Biomedicine (Combine) workgroup
Intelligent Systems research group
Department of Measurement and Information Systems,
Budapest University of Technology and Economics



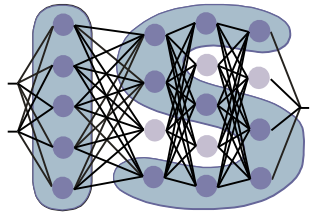
Méréstechnika és Információs Rendszerek Tanszék

MŰEGYETEM 1782

Overview

- Background
- Probabilistic graphical models, Bayesian networks
- Trends in biomedical data science
- Fedarated privacy-preserving learning in life sciences
- Fedarated privacy-preserving learning of Bayesian networks

Intelligent Systems Research Group



- Bachelor
  - Artificial Intelligence course (500<students/year)
- Master
  - Intelligent Systems MSc specialization
    - Probabilistic Inference and Decision Support Systems
    - Machine Learning, Complex probabilistic models for ML
    - Complex AI Systems
    - AI laboratory
  - Bioinformatics, Biostatistics & Health Informatics
- Ph.D.
  - Intelligent data analysis

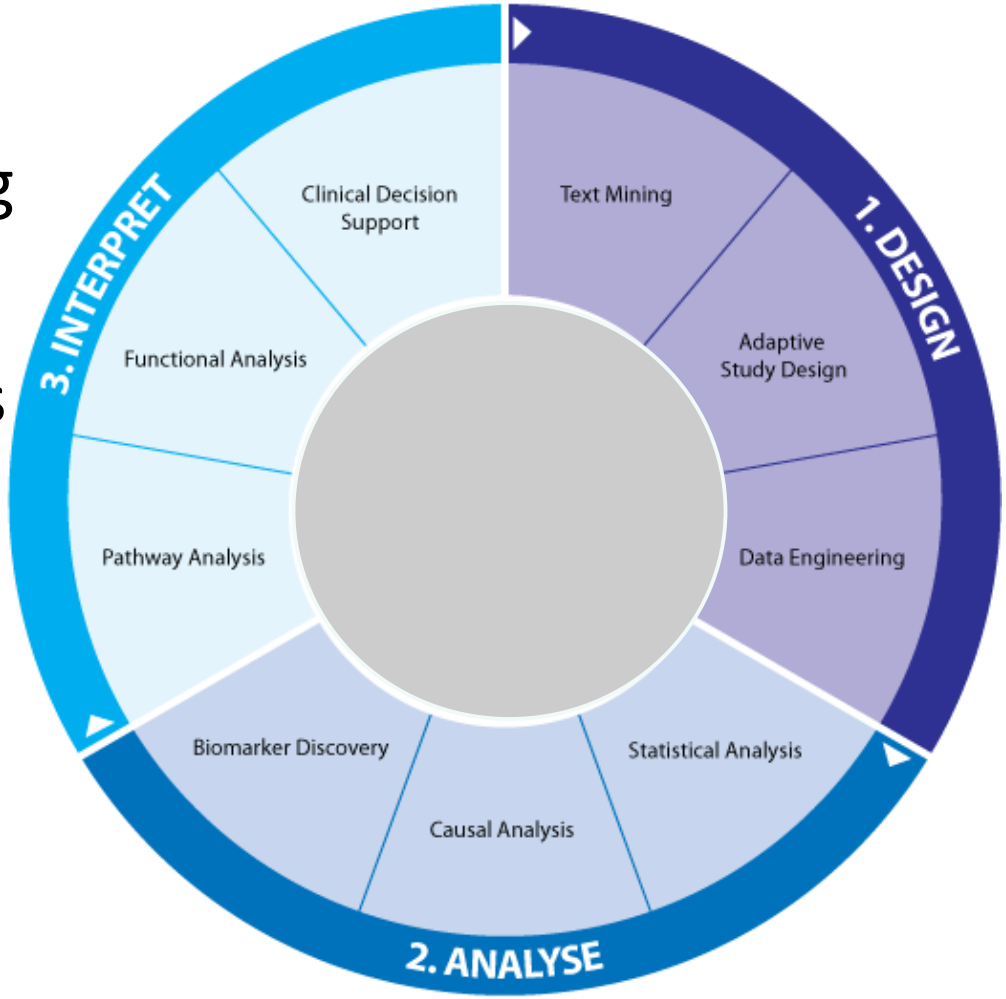
MŰEGYETEM 1782

What is AI:

Russell, S. J., & Norvig, P. (2002). Artificial intelligence: a modern approach

ComBineLab.hu: Bio&chemo-informatics

- Knowledge engineering
- Study design
- Genetic measurements
- Data engineering
- Data analysis
- Interpretation
- Decision support



ComBineLab.hu: profile

- A data-centered view
  - Genetic data
  - Clinical and patient-reported data
  - Drug/compound data
- A (statistical) methodological view
  - Probabilistic knowledge engineering
  - Systems-based/causal data analysis
  - Data and knowledge fusion
  - Biomarker analysis
  - Decision support systems

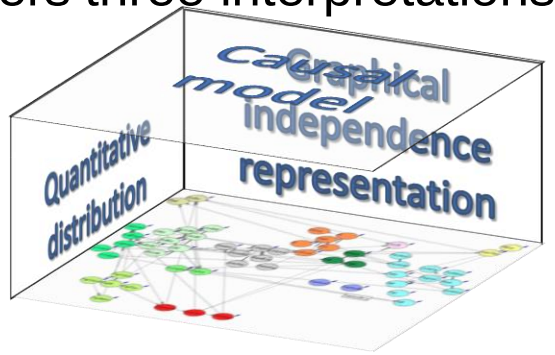
ComBineLab.hu: tools

- **BayesEye: Bayesian, systems-based data analysis**
  - Bayesian model averaging over Bayesian network structures.
- **BayesCube: Probabilistic decision support**
  - Semantically enriched Bayesian and decision network models.
- **BysCyc/QSF** (Bayesian Encyclopedia):
  - Large-scale probabilistic inference
- **QDF**: Kernel-based fusion methods for repositioning
  - Multi-aspect rankings and multi-aspect metrics in drug discovery
- **Variant Meta Caller**: precision NGS
  - Next-generation sequencing pipelines
- **VB-MK-LMF**: drug-target interaction prediction
  - Variational Bayesian Multiple Kernel Logistic Matrix Factorization

- ... see Tools @ http://bioinfo.mit.bme.hu/

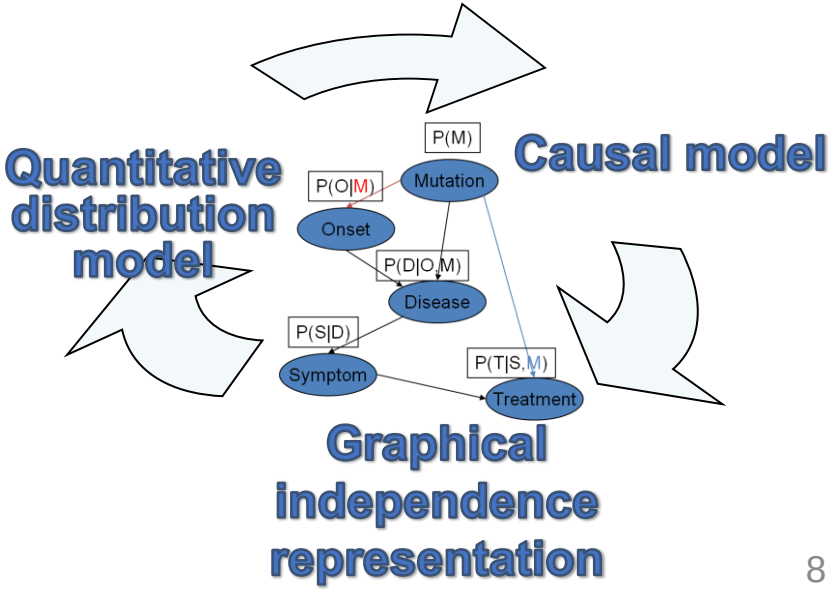
Probabilistic graphical models: Bayesian Networks

- A directed acyclic graph (DAG)
- Nodes are random variables
- Edges represent direct dependence (causal relationship)
- Local models: $P(X_i|Pa(X_i))$
- Offers three interpretations

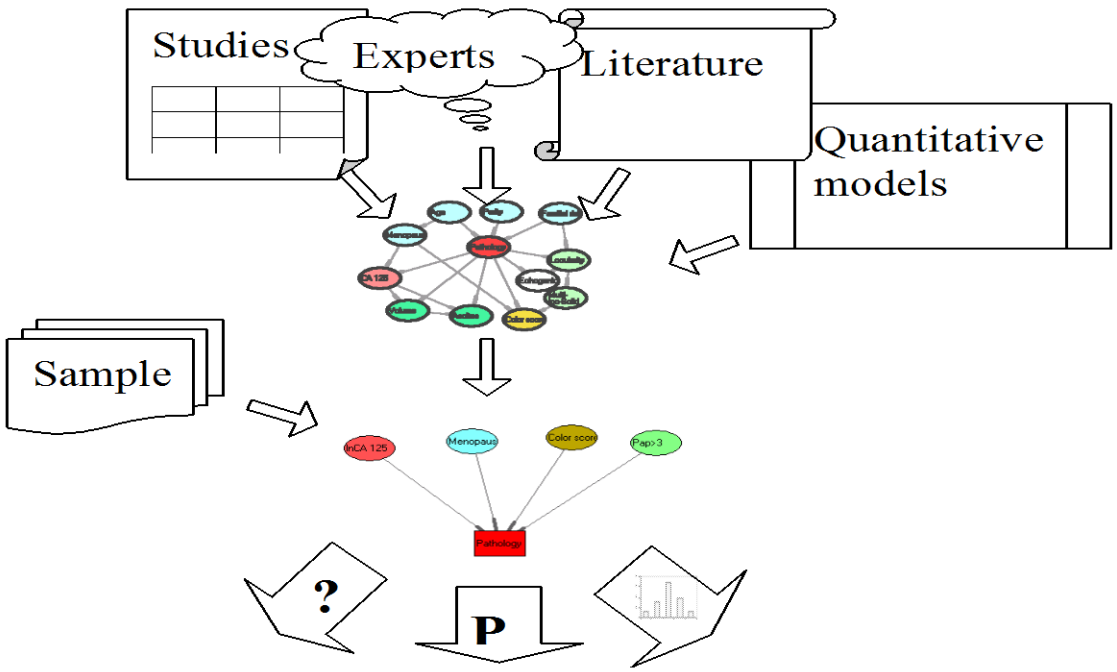


Thomas Bayes
(c. 1702 – 1761)

$$P(Model \mid Data) \propto P(Data \mid Model)P(Model)$$



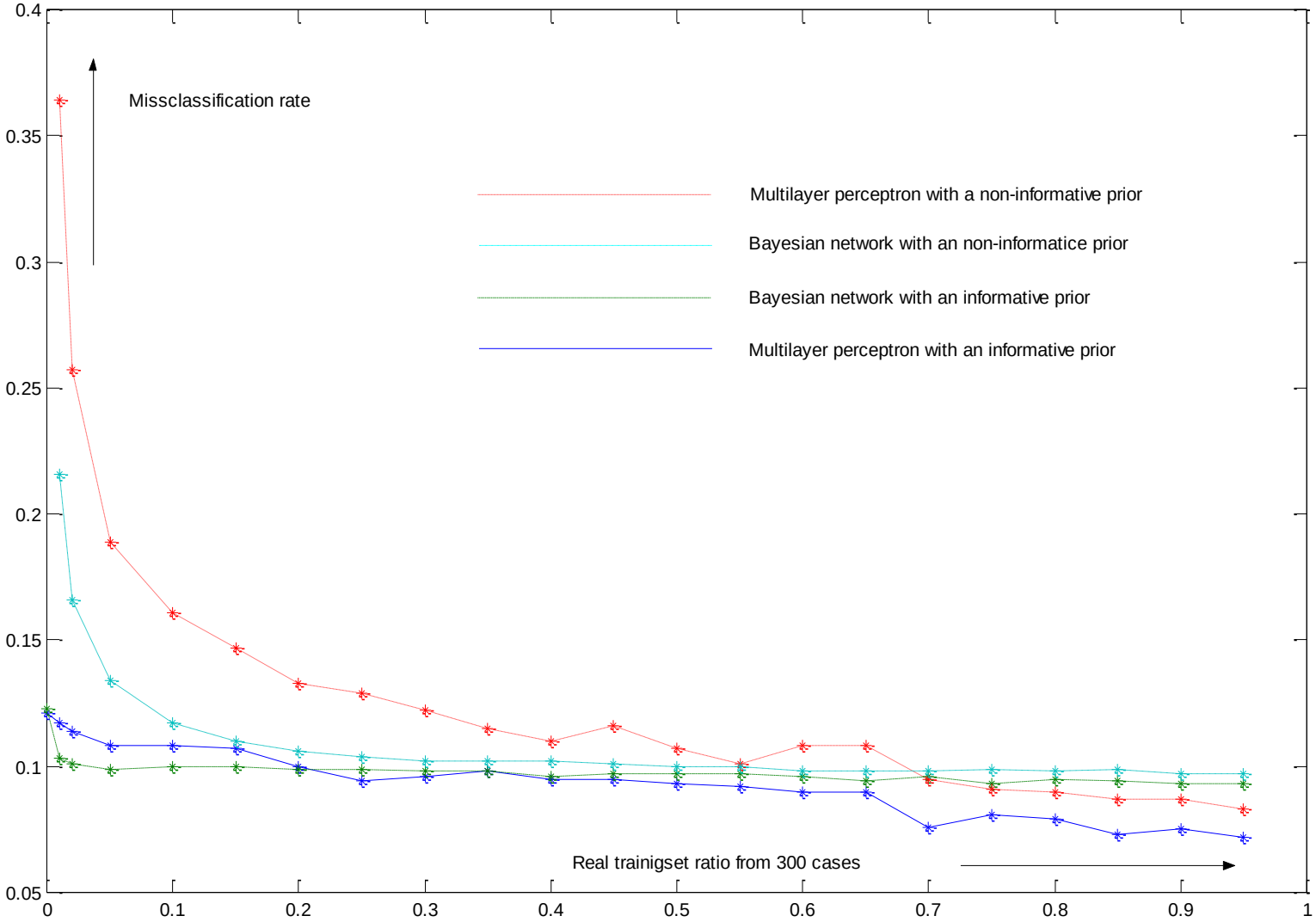
A.I. in health data analysis



- Non-invasive diagnostics of ovarian cancer, 1998

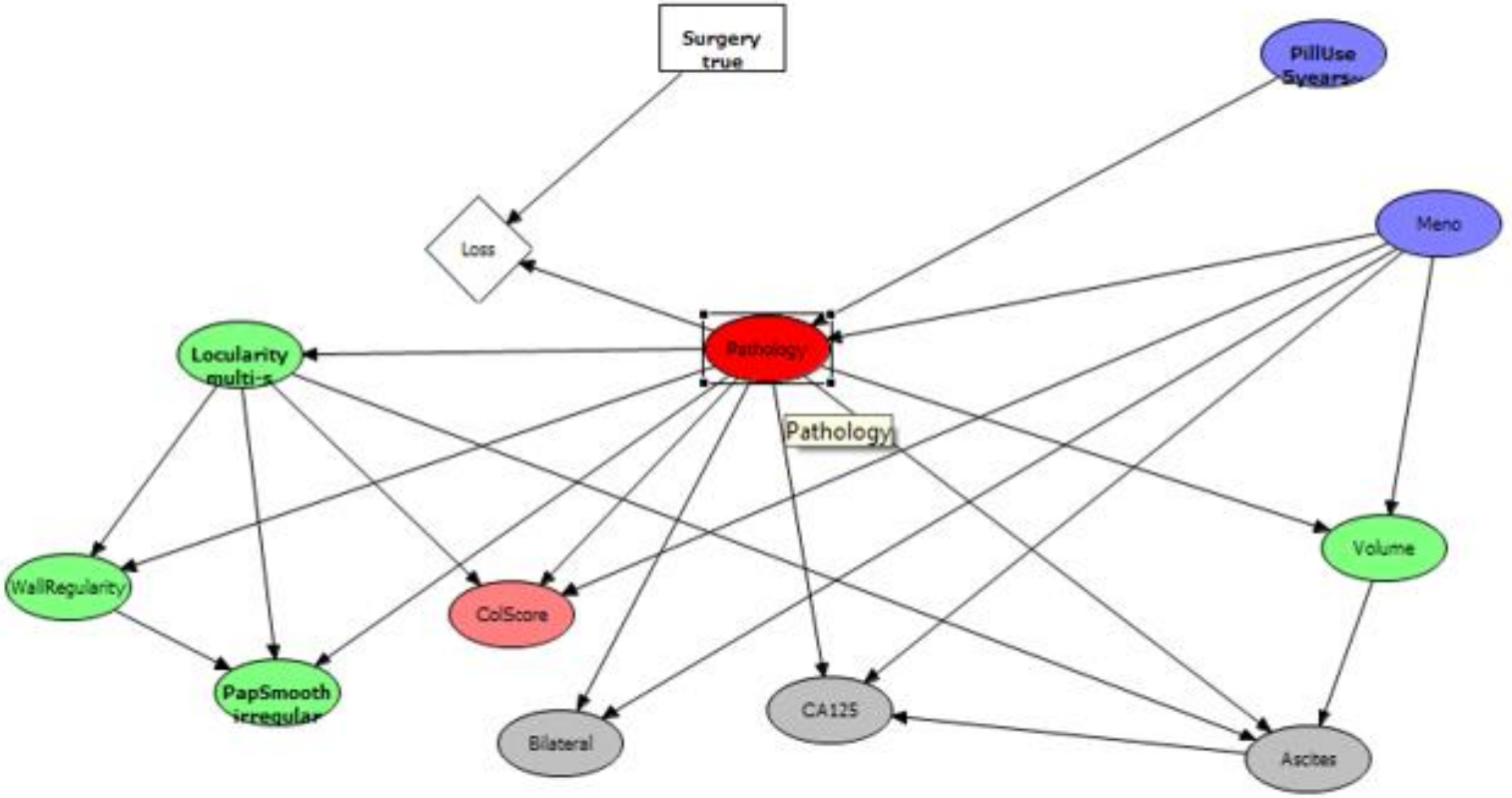
  International Ovarian Tumor Analysis (IOTA, http://www.iotagroup.org/)

„Informed" neural networks



P. Antal, G. Fannes, D. Timmerman, Y. Moreau, B. De Moor: Bayesian Applications of Belief Networks and Multilayer Perceptrons for Ovarian Tumor Classification with Rejection, *Artificial Intelligence in Medicine*, vol. 29, pp 39-60, 2003

Ovarian tumor diagnostics



Antal, P., Fannes, G., Timmerman, D., Moreau, Y. and De Moor, B., Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in medicine*, *30*(3), pp.257-281, 2004

Types of inference

- (Passive, observational) inference
  - P(Query|Observations, Observational data)
- Interventionist inference
  - P(Query|Observations, Interventions)
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

- Biomedical applications
  - Prevention
  - Screening
  - Diagnosis
  - Therapy selection
  - Therapy modification
  - Evaluation of therapic efficiancy

Learning from multi-centric data

- Challenges
  - Data standardization
  - Representativity
  - Quality
  - Selection bias (confounding)

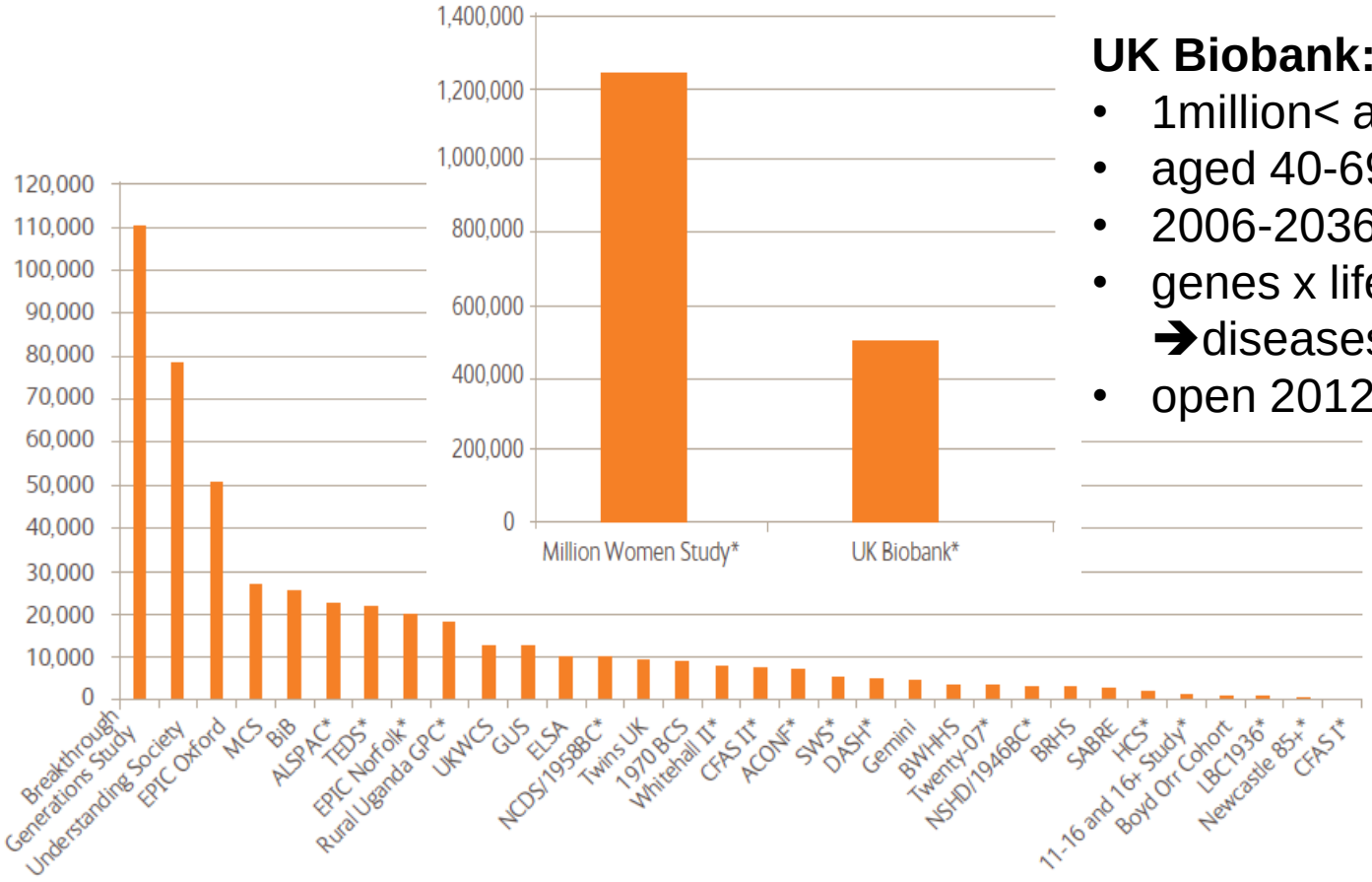
UK Biobank 2006-2010

UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. **It is following the health and well-being of 500,000 volunteer participants and provides health information**, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Scientists, please ensure you read the background materials before registering. **To our participants, we say thank you for supporting this important resource to improve health**. Without you, none of the research featured on this website would be possible.

Elliott, P., & Peakman, T. C. (2008). The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *International Journal of Epidemiology*, *37*(2), 234-244.

Collins, R. (2012). What makes UK Biobank special?. *The Lancet*, *379*(9822)

Large-scale cohorts in UK



UK Biobank:

- 1million< adults
- aged 40-69,
- 2006-2036<
- genes x lifestyle x environment ➔diseases
- open 2012-

UKBiobank: incidences

Table 1. Approximate numbers of incident cases of some exemplar conditions expected to accrue during the first 20 years of follow-up in UK Biobank.

Condition	2012	2017	2022	2027
Diabetes mellitus	10,000	25,000	40,000	68,000
MI and coronary death	7,000	17,000	28,000	47,000
Stroke	2,000	5,000	9,000	20,000
COPD	3,000	8,000	14,000	25,000
Breast cancer (female)	2,500	6,000	10,000	16,000
Colorectal cancer	1,500	3,500	7,000	14,000
Prostate cancer	1,500	3,500	7,000	14,000
Lung cancer	800	2,000	4,000	8,000
Hip fracture	800	2,500	6,000	17,000
Rheumatoid arthritis	800	2,000	3,000	5,000
Alzheimer's disease	800	3,000	9,000	30,000
Parkinson's disease	1000	3,000	6,000	14,000

PLOS Medicine | DOI:10.1371/journal.pmed.1001779 March 31, 2015

UKB: baseline assessment

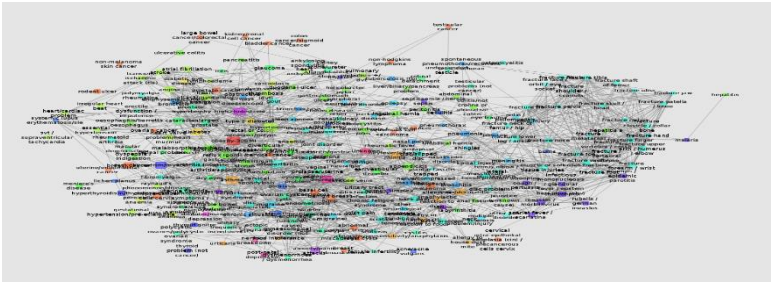
Table 2. Data collected at the baseline assessment.

Questionnaire and interview	
Sociodemographic	Social class; ethnicity; employment status; marital status; education; income; car ownership
Family history and early life exposures	Family history of major diseases; birth weight; breast feeding; maternal smoking; childhood body size; residence at birth
Psychosocial factors	Neurosis; depression (including bi-polar spectrum disorder); social support
Environmental factors	Current address; current (or last) occupation; domestic heating and cooking fuel; housing; means of travel; shift work; mobile phone use; sun exposure
Lifestyle	Smoking; alcohol consumption; physical activity; diet; sleep
Health status	Medical history; medications; disability; hearing; sight; sexual and reproductive history
Hearing threshold	Speech reception threshold*
Cognitive function	Pairs matching; reaction time; prospective memory*; fluid intelligence*; numeric memory†
Physical measures	
Blood pressure and heart rate	two automated measures, one minute apart
Grip strength	Left- and right-hand grip strength
Anthropometrics	Standing and sitting height; weight and bio-impedance; hip and waist circumference
Spirometry	Up to three measures
Bone density‡	Calcaneal ultrasound
Arterial stiffness¶	Pulse wave velocity
Eye examination§	Refractive index, intraocular pressure; acuity; retinal photograph; optical coherence tomography
Fitness test§	Cycle ergometry with electrocardiogram (ECG) heart rate monitoring

PLOS Medicine | DOI:10.1371/journal.pmed.1001779 March 31, 2015

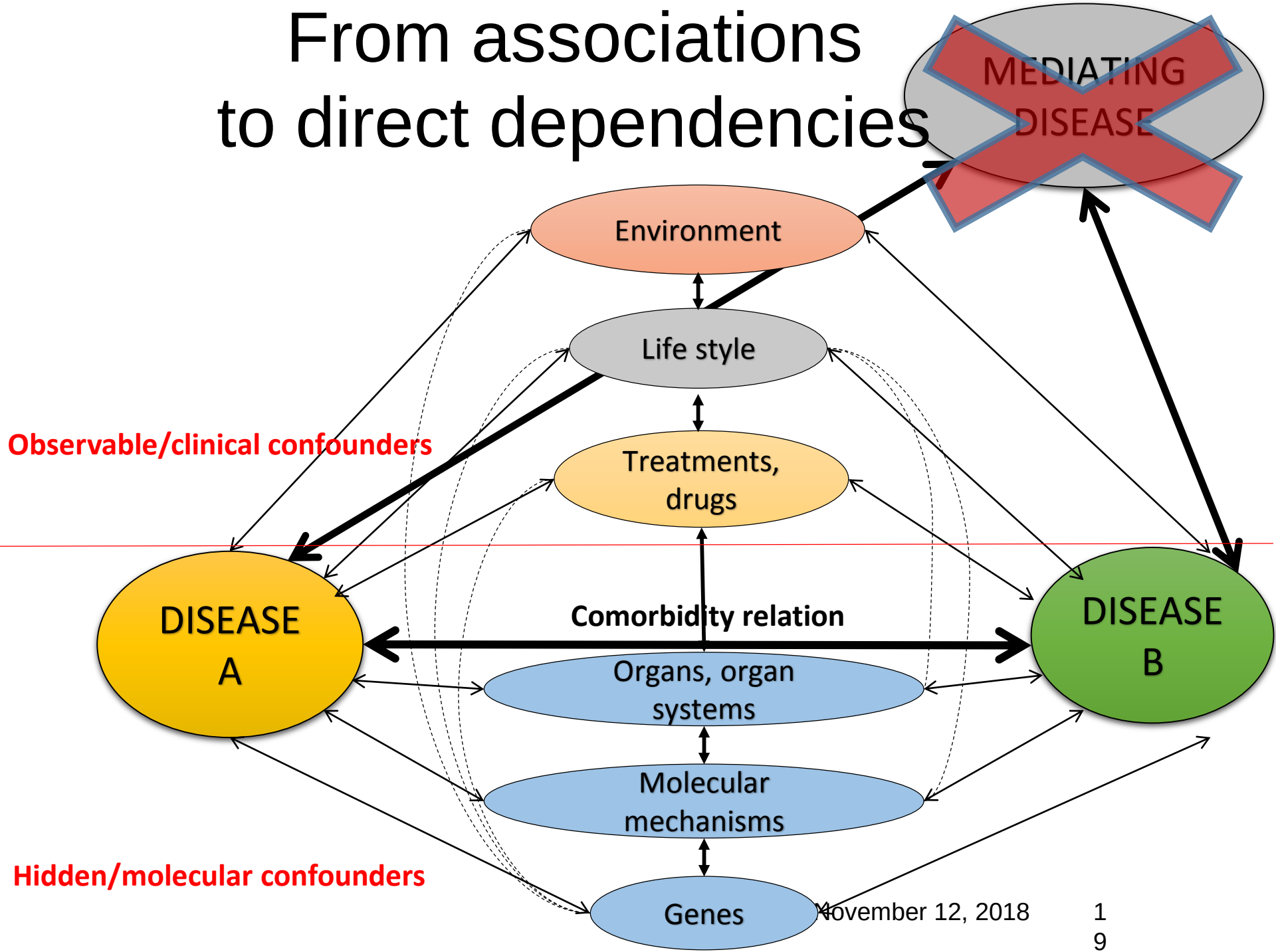
UKB-1602: Role of genetics, diet and comorbidities in depression

- UK Biobank – research project No.1602
- 2013-2017, 2017-2020
- PI: G. Juhasz, University of Manchester (UoM)
- Cooperation: SE-BME-UoM

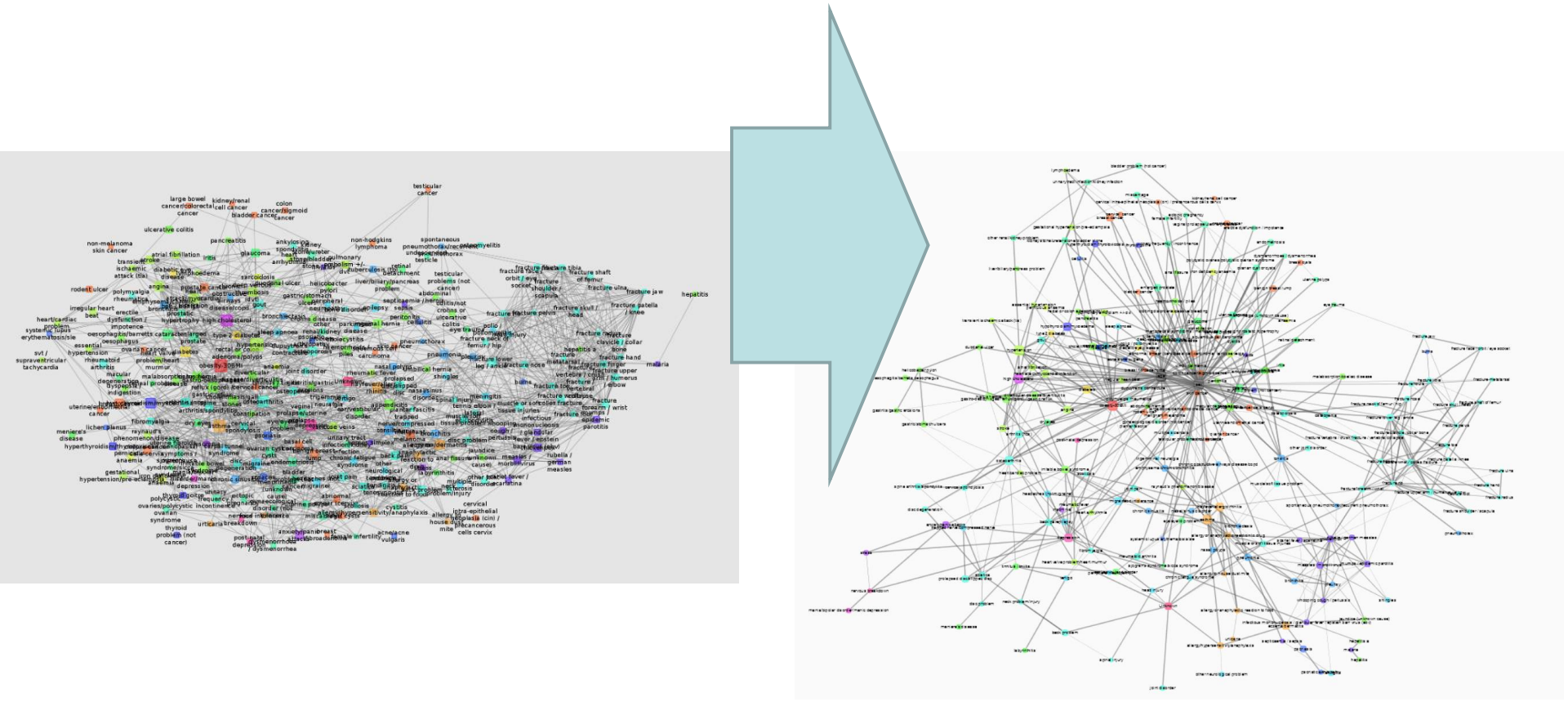


Marx, P., Antal, P., Bolgar, B., Bagdy, G., Deakin, B. and Juhasz, G., 2017. Comorbidities in the diseasome are more apparent than real: What Bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, *13*(6), p.e1005487.

From associations to direct dependencies

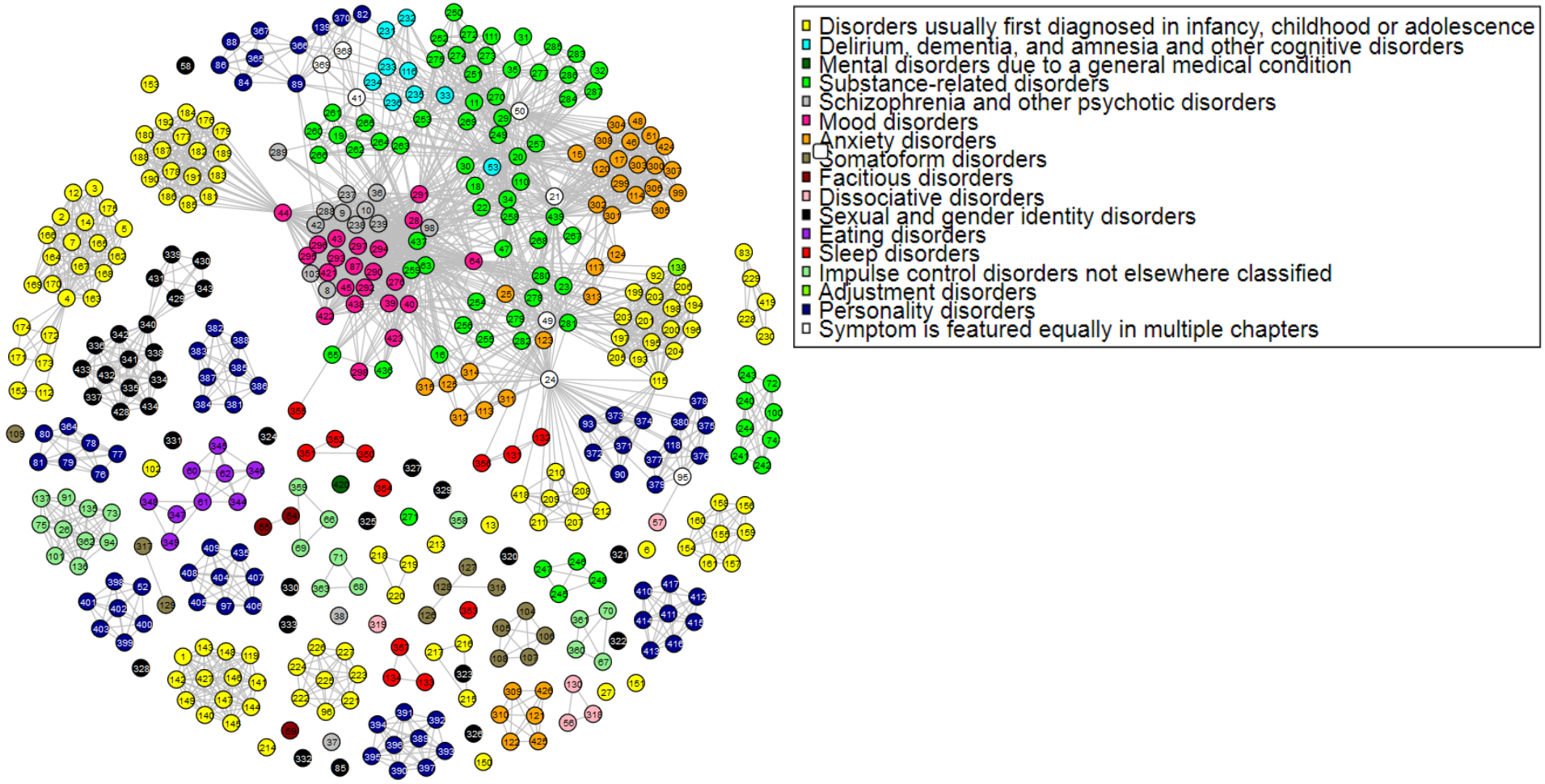


From associations to direct dependencies II. (off:80%)



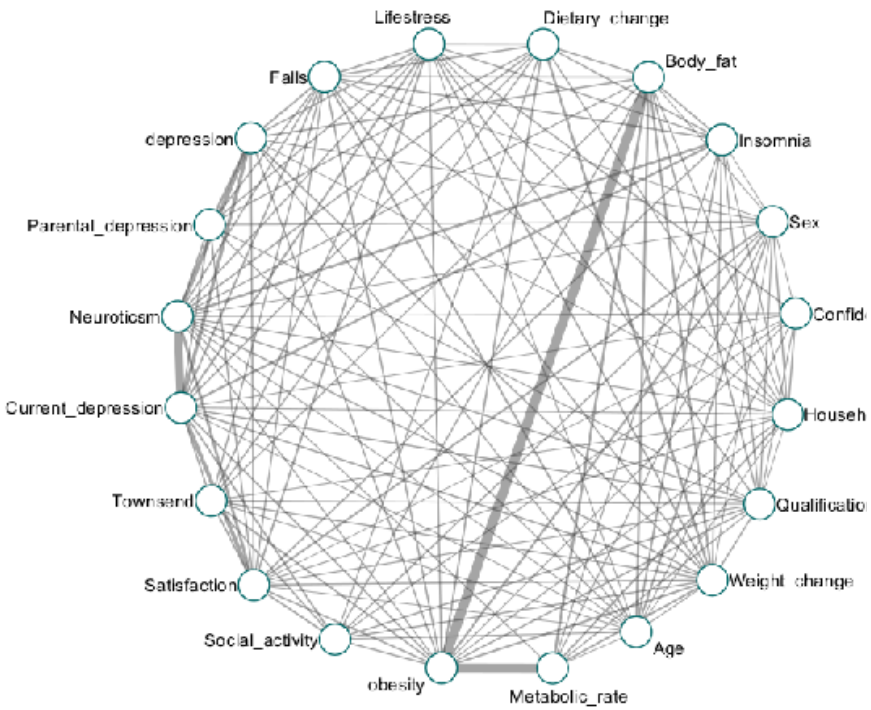
Marx, P., Antal, P., Bolgar, B., Bagdy, G., Deakin, B. and Juhasz, G., 2017. Comorbidities in the diseasome are more apparent than real. *PLoS computational biology*, *13*(6), p.e1005487.

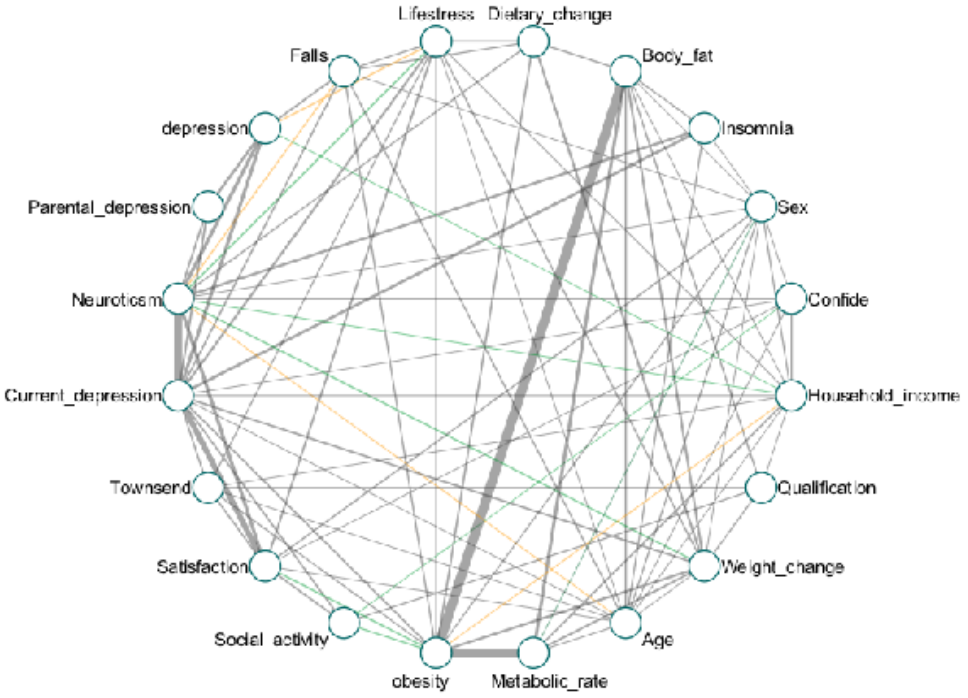
Markov networks, pairwise Markov Random Fields



Borsboom, D. and Cramer, A.O., 2013. Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, *9*, pp.91-121.

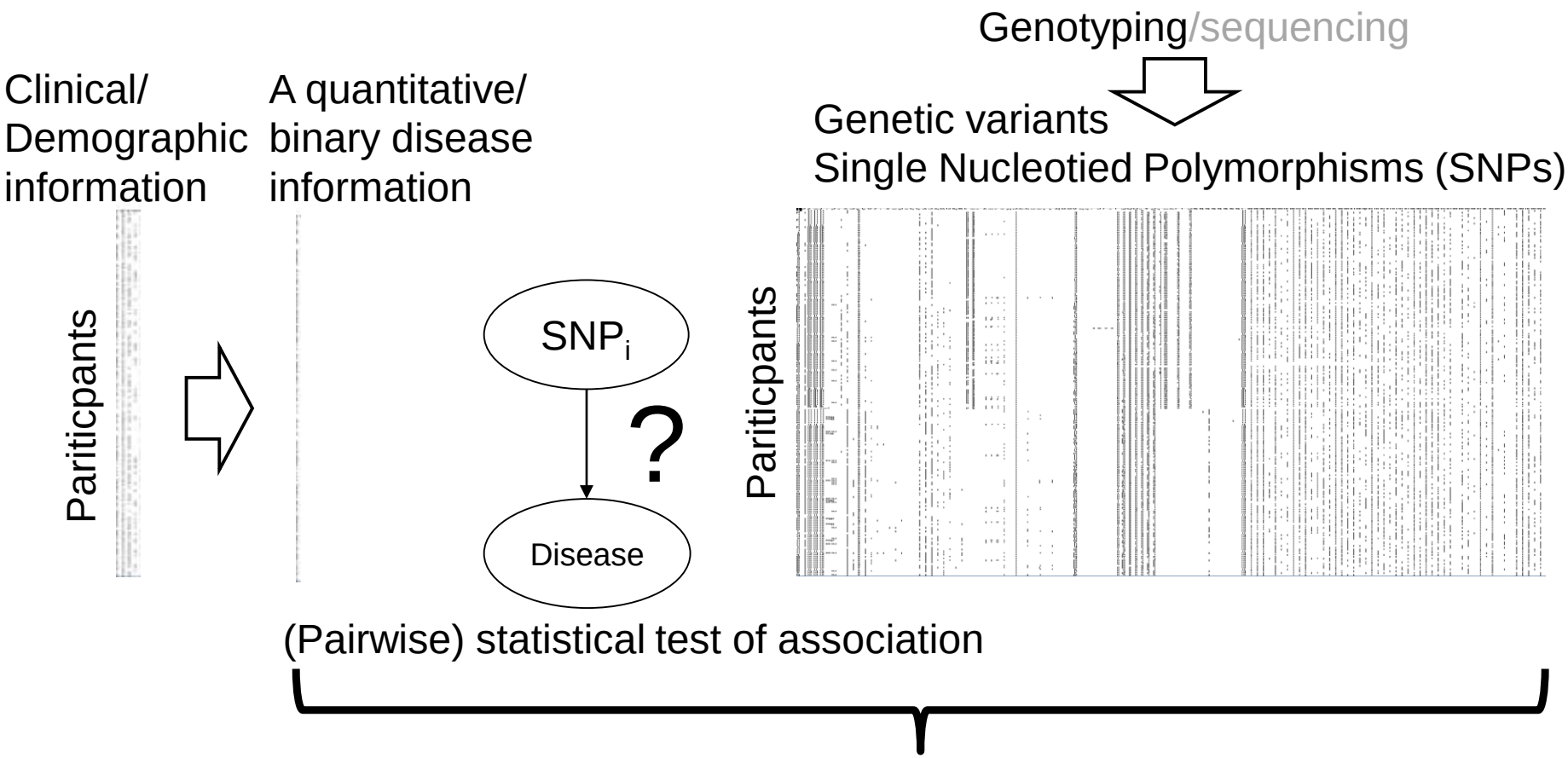
Bootstrapped PGMs vs. Bayesian PGMs





Hullam, G., Juhasz, G., Deakin, B. and Antal, P., 2017, August. Structural and parametric uncertainties in full Bayesian and graphical lasso based approaches: Beyond edge weights in psychological networks. In *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, 2017*

Genetic association studies (GAS)

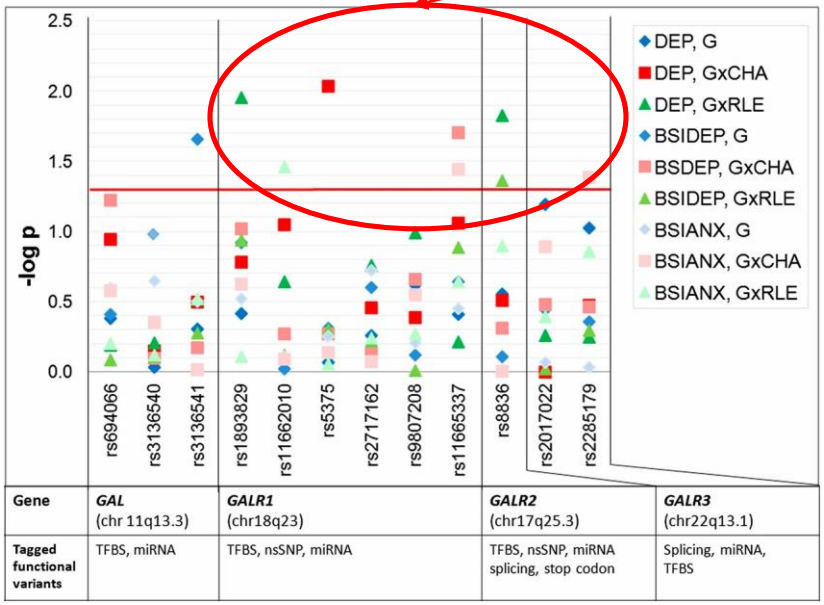


Relevant SNPs for a disease with complex genetic background.

Risch, N. and Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science*, *273*(5281), pp.1516-1517.

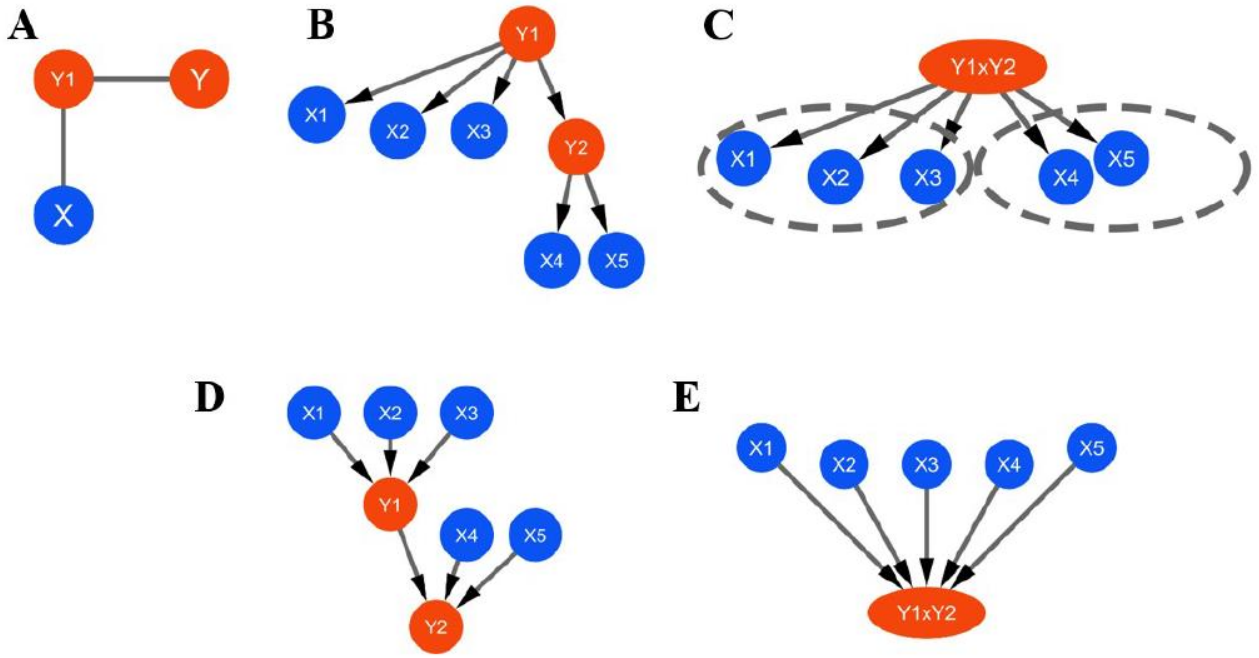
GAS challenge: interactions

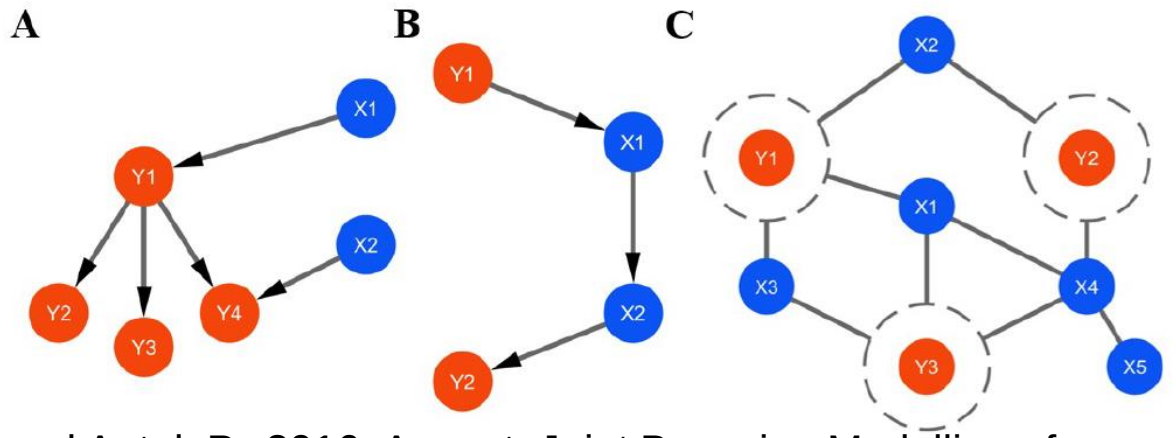
gene x environment (GxE) interactions



Juhasz, G., Hullam, G., Eszlari, N., Gonda, X., Antal, P., Anderson, I.M., Hökfelt, T.G., Deakin, J.W. and Bagdy, G., 2014. Brain galanin system genes interact with life stresses in depression-related phenotypes. *Proceedings of the National Academy of Sciences*, *111*(16), pp.E1666-E1673.

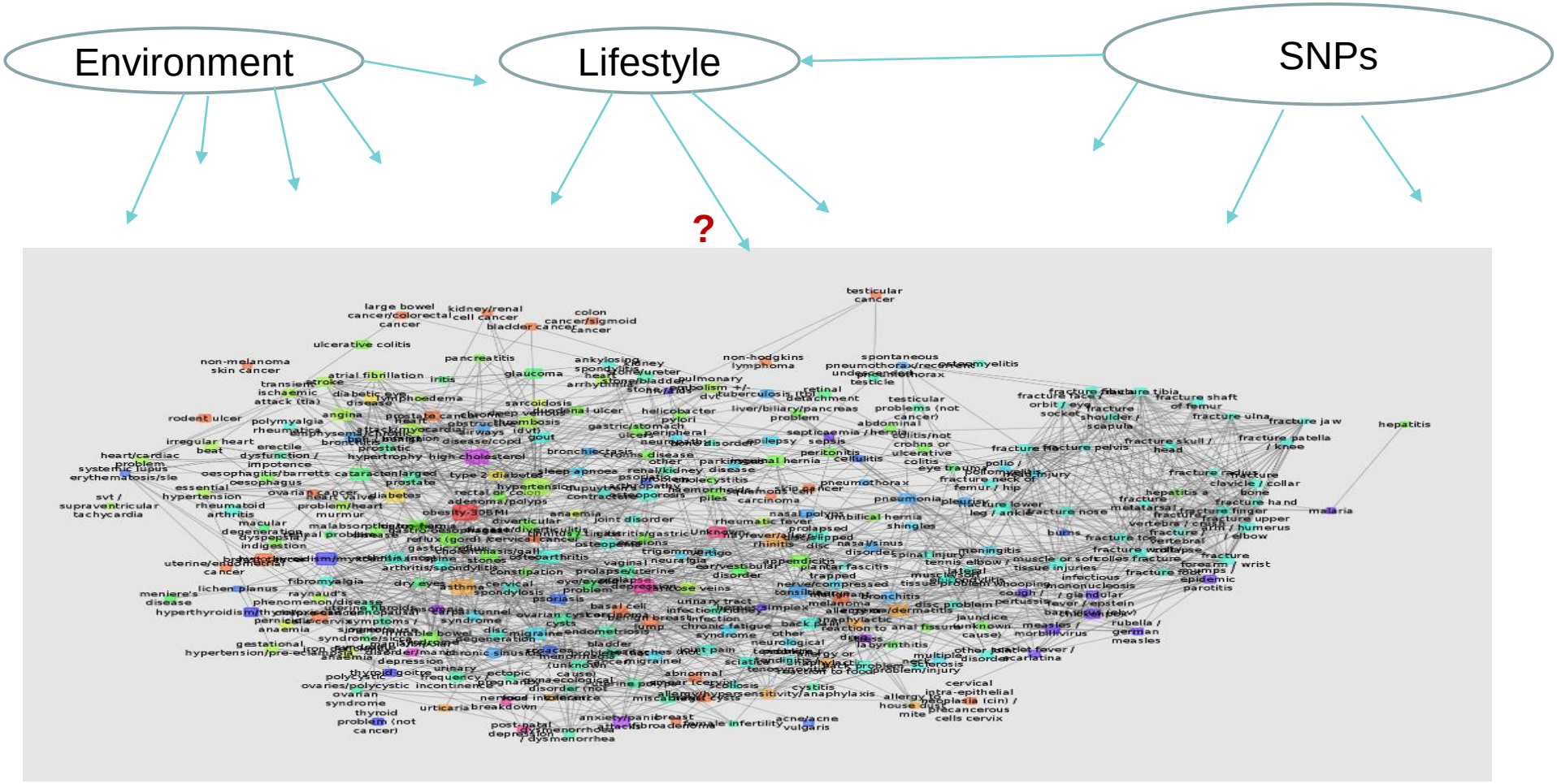
GAS challenge: multiple targets





Marx, P., Millinghoffer, A., Juhász, G. and Antal, P., 2016, August. Joint Bayesian Modelling of Internal Dependencies and Relevant Multimorbidities of a Heterogeneous Disease. In *Conference on Probabilistic Graphical Models*

GAS: GxExLS to diseasome?



Bycroft, Clare, et al. "The UK Biobank resource with deep phenotyping and genomic data." *Nature* 562.7726 (2018): 203.

Canela-Xandri, O., Rawlik, K., & Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. *Nature Genetics*, 1.

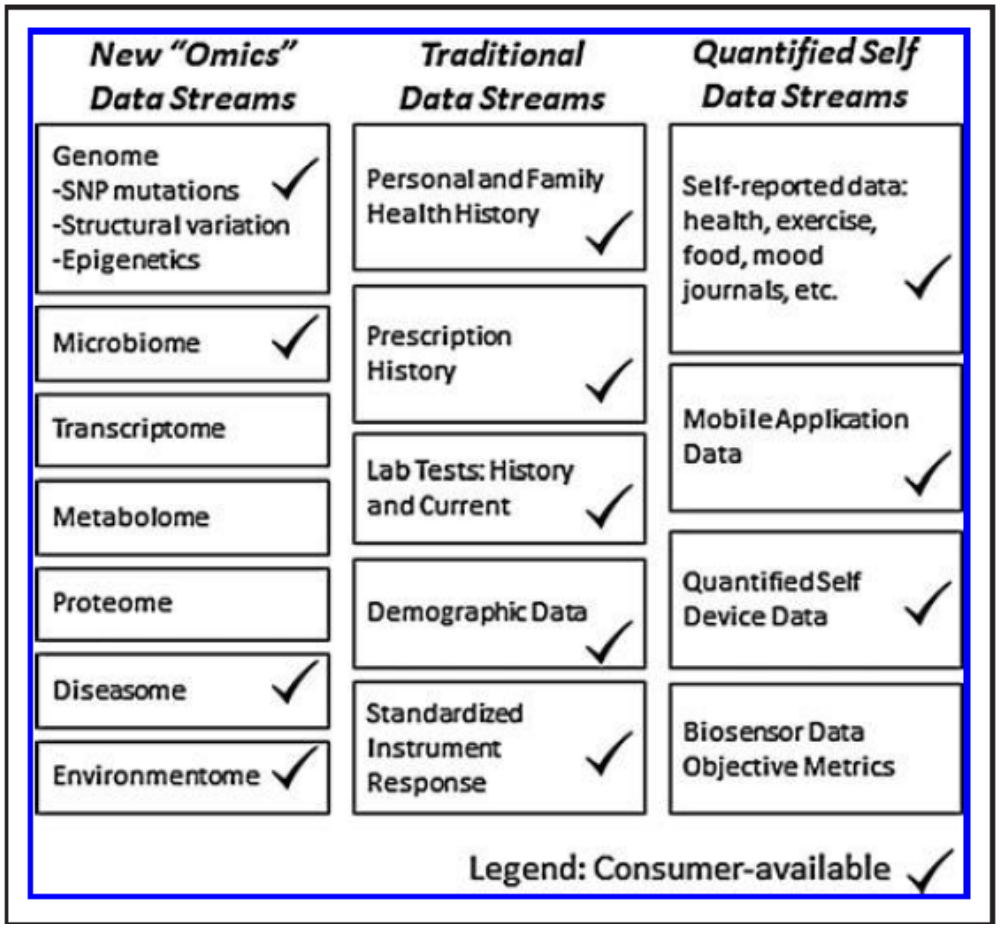
Further national biobanks: FinnGen

- https://www.finngen.fi/en
- 500k participants
- 2017-
- **Personalized medicine project**
- genome information (WGS) + digital health care data
- **The study is funded[!!!] by Business Finland and seven international pharmaceutical companies**: Abbvie, AstraZeneca, Biogen, Celgene, Genentech (a member of the Roche Group), Merck & Co., Inc., Kenilworth, NJ, USA and Pfizer.

Further health data

- FlatIron Health (acquired by Roche):
  - 7 major academic research centers
  - 280+ community oncology practices
  - top 15 therapeutic oncology companies
  - 2500 clinicians
  - 2.1 million active patient records
    - complete, electronic health records
    - +patient-reported data

Big health data streams

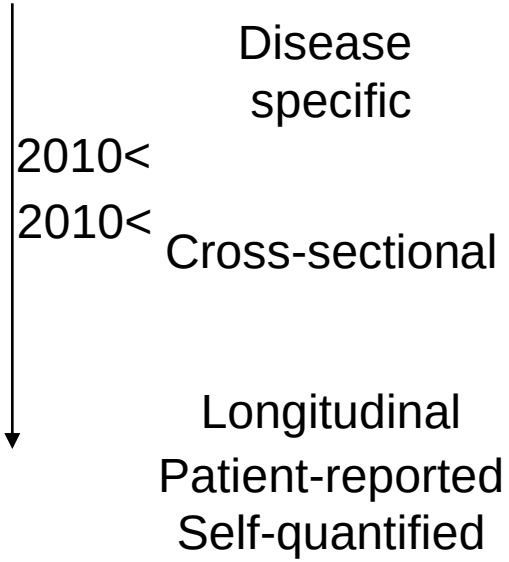


M.Swan: THE QUANTIFIED SELF: Fundamental Disruption in Big Data Science and Biological Discovery, Big data, Vol 1., No. 2., 2013

On the thresholds of data: health

- Local datasets: 1k➔10k participiants
- International datasets: 10k➔100k
- National biobanks: <1million
- International biobanks: x1million
- Regular health records: 100 million

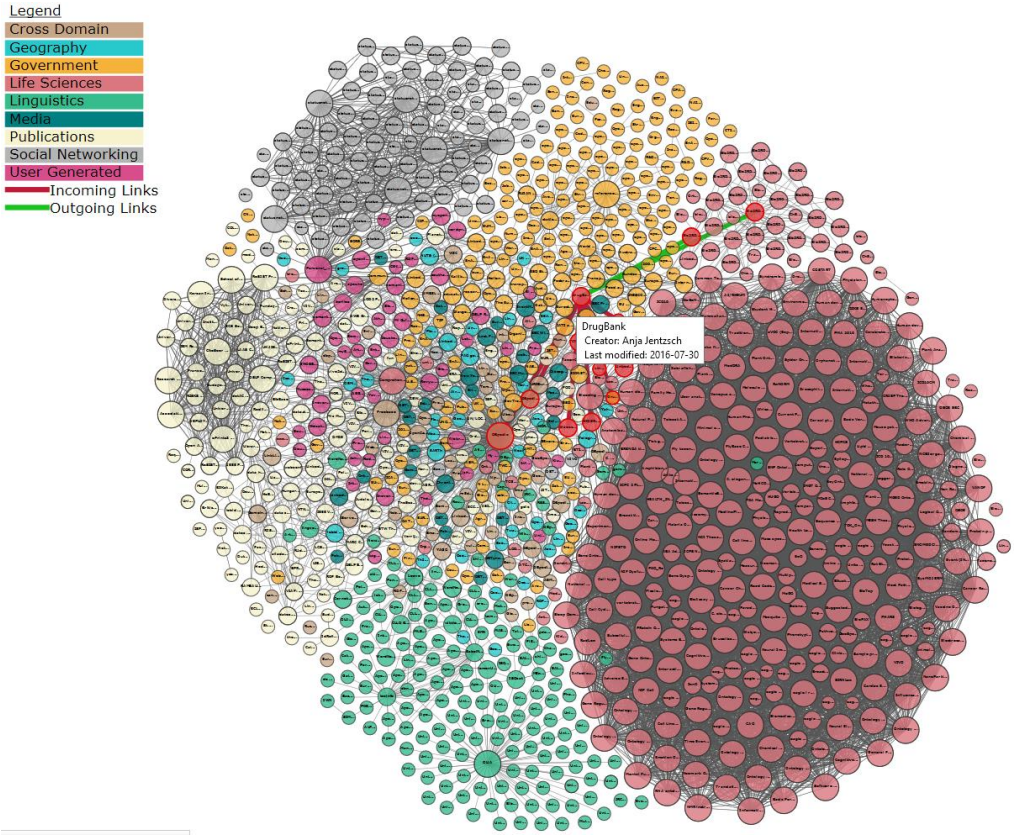
(➔Meta-analysis using summary statistics)

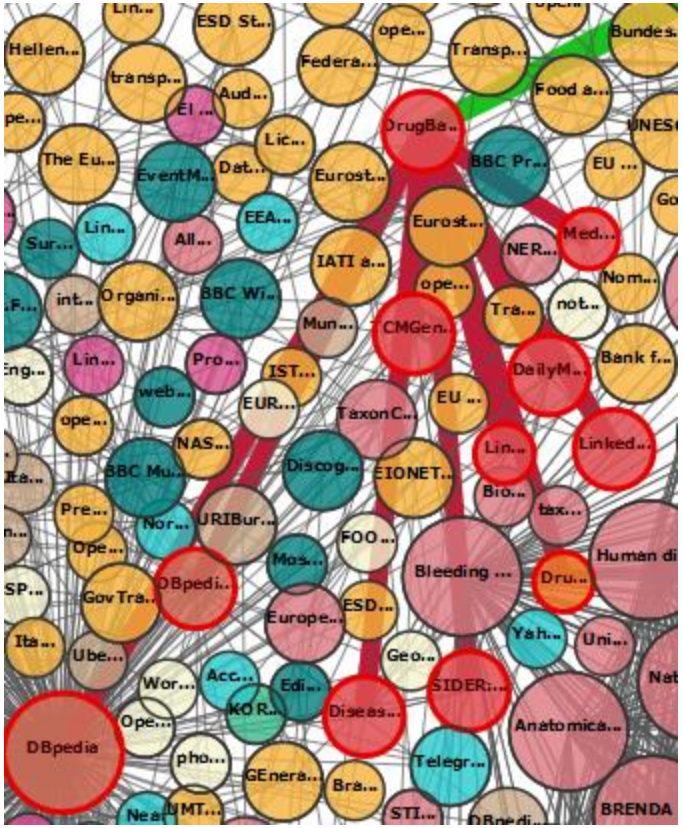


➔Federated learning: separation of data and model

1. Data is standardized (using ontologies)
2. Stays at the institutes/individuals
3. Model updates are communicated
4. Using privacy-preserving techniques

2010<: Linked open data (LOD) for cross-domain integration





Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Federated, privacy-preserving learning in life sciences

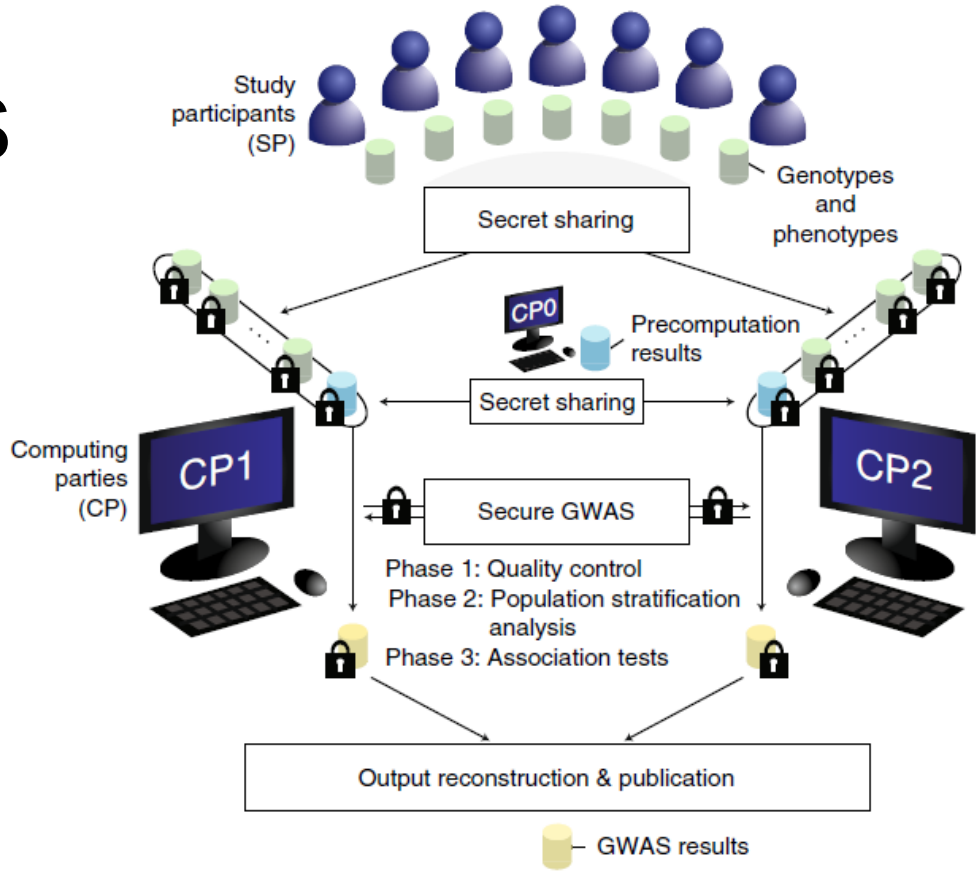
Mobile/IoT world

- Status recognition, action prediction
  - Predictive typing, recommendations
- Model users/data owners: 10^6 – 10^9
- Data size: ~1kB
- Data is uniform
- Background knowledge: common sense
- Online learning

Biomedicine

- Feature subset selection, prediction
  - Genetic association study, Pharma predictions
- Model users/data owners: 10 – 10^3
- Data size: ~1TB
- Data is heterogeneous, multimodal
- Voluminous prior knowledge!
- Online learning

FPPL in genetics



Cho, H., Wu, D.J. and Berger, B., 2018. Secure genome-wide association analysis using multiparty computation. *Nature biotechnology*, *36*(6), p.547.

Tkachenko, O., Weinert, C., Schneider, T. and Hamacher, K., 2018, May. Large-Scale Privacy-Preserving Statistical Computations for Distributed Genome-Wide Association Studies. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (pp. 221-235). ACM.

Raisaro, J.L., Choi, G., Pradervand, S., Colsenet, R., Jacquemont, N., Rosat, N., Mooser, V. and Hubaux, J.P., 2018. Protecting Privacy and Security of Genomic Data in i2b2 With Homomorphic Encryption and Differential Privacy. *IEEE/ACM transactions on computational biology and bioinformatics*.

Simmons, S., Sahinalp, C. and Berger, B., 2016. Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell systems*, *3*(1), pp.54-61.

Simmons, S. and Berger, B., 2016. Realizing privacy preserving genome-wide association studies. *Bioinformatics*, *32*(9), pp.1293-1300.

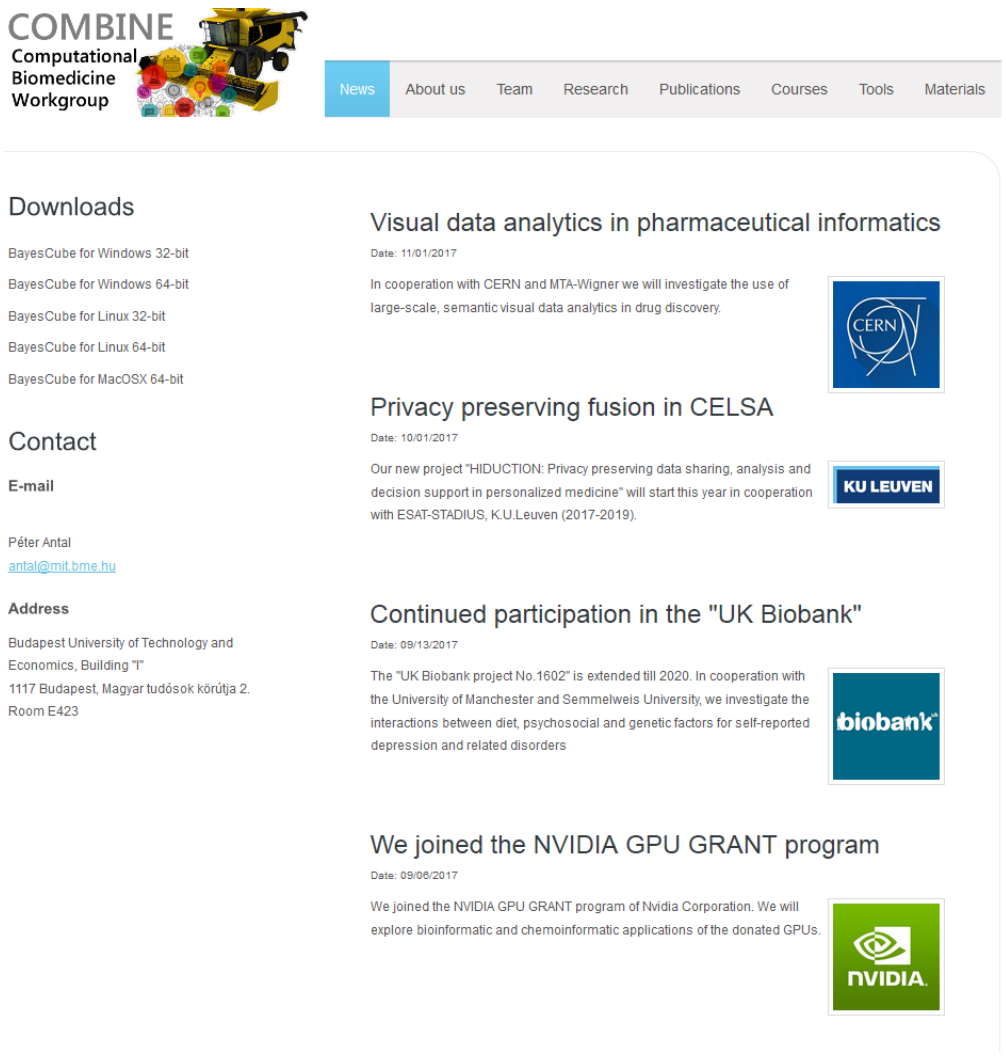
HIDUCTION: privacy in genetics

- CELSA: HIDUCTION: privacy preserving data sharing, analysis and decision support in personalized medicine
- 2017-2019
- Participants:
  - PI: Yves Moreau (KULeuven),
  - Levente Buttyán (BME),
  - Gabriella Juhász (SE),
  - Péter Antal (BME)

Summary

- Next stage of data analysis in life science
  - federated, privacy-preserving learning
- Different characteristics (w.r.t. ITC world)
  - Partners, data size, prior,..
- Scalable solutions exist
  - in genome-wide data analyis,
  - in drug-target interaction prediction.
- Probabilistic graphical models allow transparent and understandable combination of data and knowledge.

Computational Biomedicine (ComBine) lab



Team
Bence Bolgár
Bence Bruncsics
András Gézsi
Gábor Hullám
András Millinghoffer
Péter Sárközy
Péter Antal

http://bioinfo.mit.bme.hu/

Thank you for you attention!