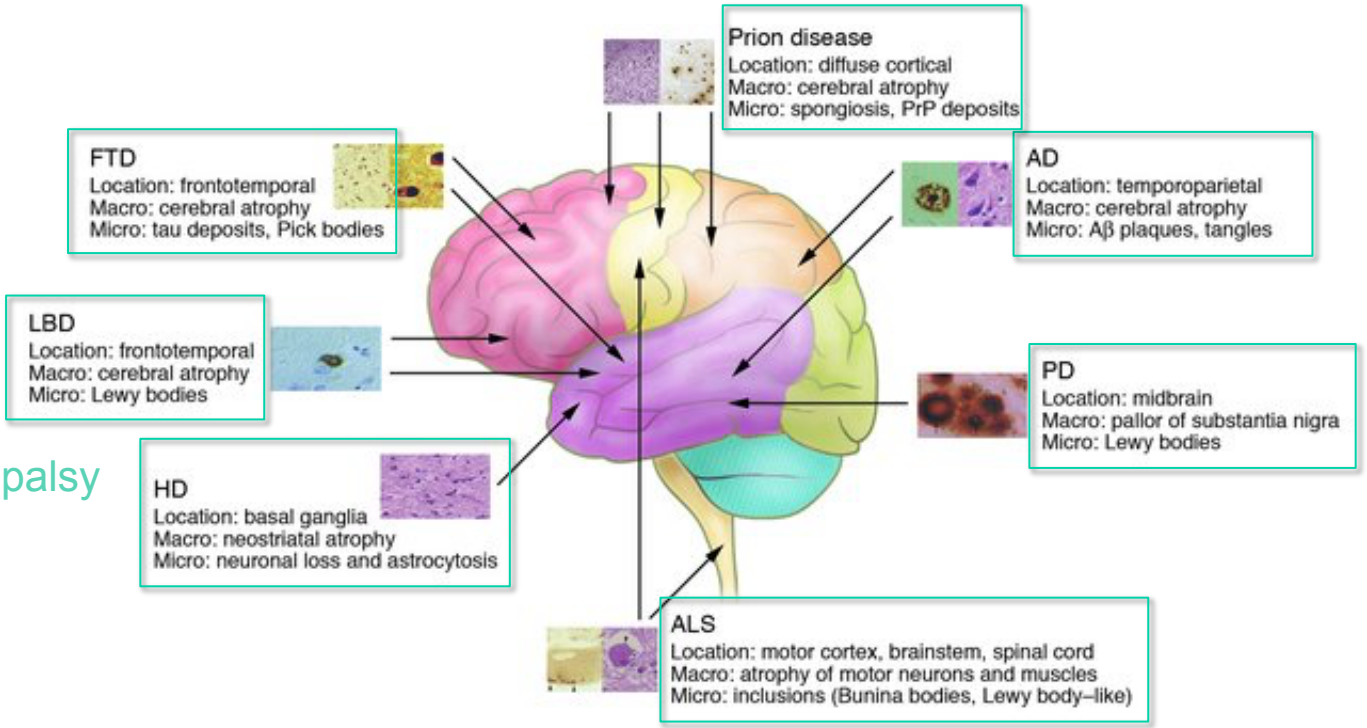
Novel drug target identification for the treatment of dementia using multi-relational association mining

Nguyen et al, 2015 Nature Sci Rep. **5**:11104

Critical paper review by
Jasleen Grewal

Dementia (…formerly, ‘senility’)

- Central Nervous System (CNS) diseases
- Degeneration and eventual death of neurons in brain, eye, spinal cord
- Long term (> 6 months), gradual decrease in one’s mental functioning



Progressive supranuclear palsy
Corticobasal dementia
Pick’s disease

Cures for dementia?

- Of the top 10 causes of death in USA, dementias are the only ones for which no therapies exist that can halt or even slow disease progression (UCSF IND)[1]
- >10 years since last treatment approved[2]
- Known, pharmacologically validated targets[3]
  - Cholinesterase inhibitors – symptom alleviation

1. Neurodegenerative Disease Incentive Initiative, 2015.
2. The Dementia Consortium, UK, 2014.
3. New Prospects and Strategies for Drug Target Discovery in Neurodegenerative Disorders. Hilbush et al. NeuroRx. 2005 Oct; 2(4): 627-637

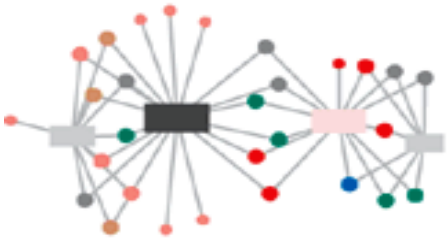
Drug development and dementia

- Studies to prove that new medicines are safe
  - Target brain degeneration[1]
  - Much longer, complex, costly, prone to failure[1]
- 98% of promising drugs prohibited by the blood-brain barrier (nasal mucosal grafting?)
- Diseasemodifying therapies needed, not present
- Bringing out a new drug = $6 billion cost, on average

1. Neurodegenerative Disease Incentive Initiative, 2015.

Finding drug targets

- Narrow our focus for drug discovery and research
- Proteins rarely function in isolation in a biological system
  - Interactomes
- Understand the interactions of drugs with their targets, in context of the networks[1]
- To understand determinants of disease expression[1]:
  - Molecular networks: protein interaction, metabolic, regulatory
  - Phenotypic networks: coexpression, genetic

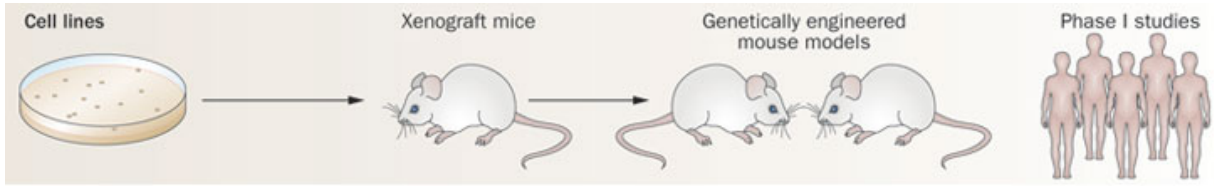


1. Drug-target network. Yildrim, M et al. Nat. Biotechnol. 2007 25: 1119-26

Methodology

What drugs already exist for dementia

- Lab (preclinical)
- Trial
- Marketed



What are their molecular targets

Build an annotated interactome of these targets

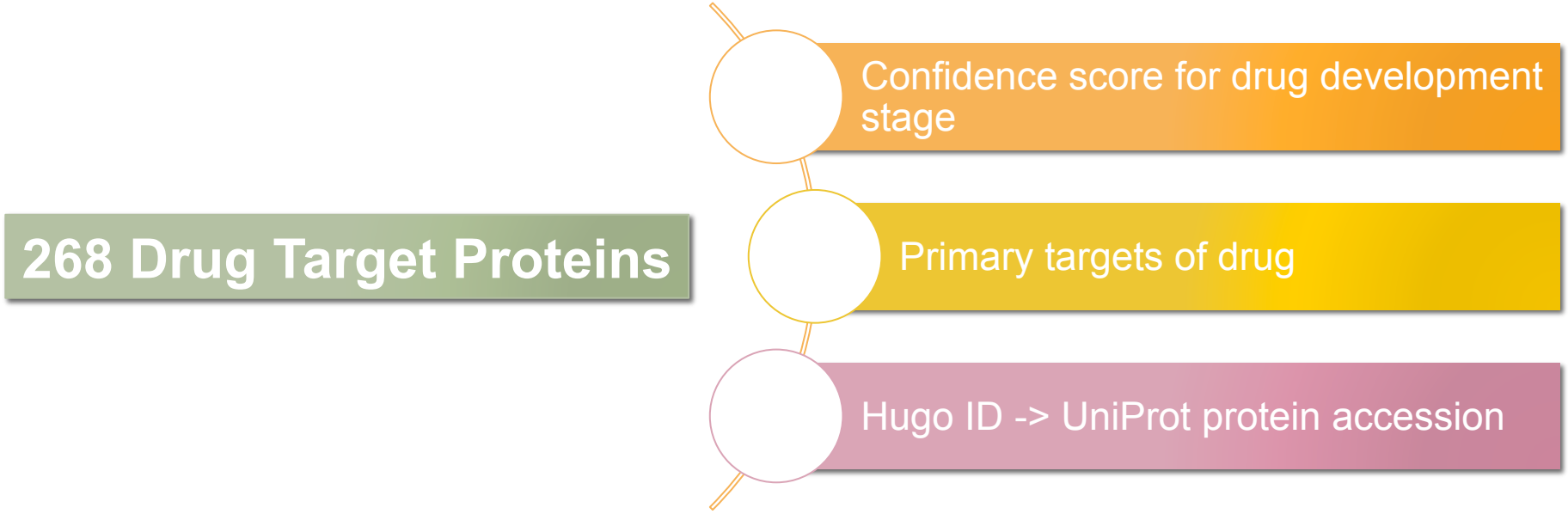
- Protein interactions
- Associated Pathways
- GO terms
- Protein domain information

Induce association rules

- Predict novel drug targets

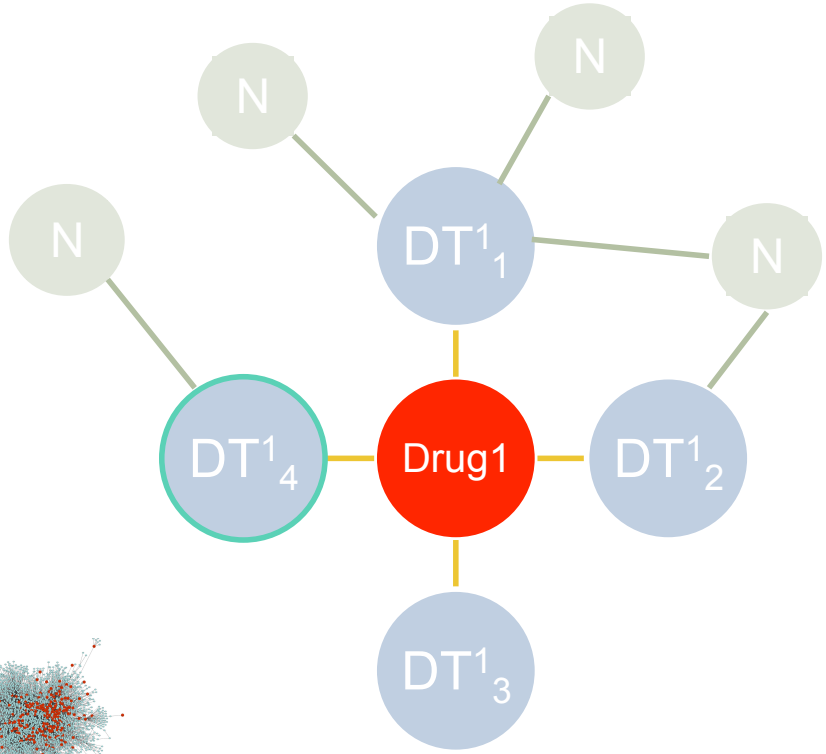
Finding pre-existing dementia targets

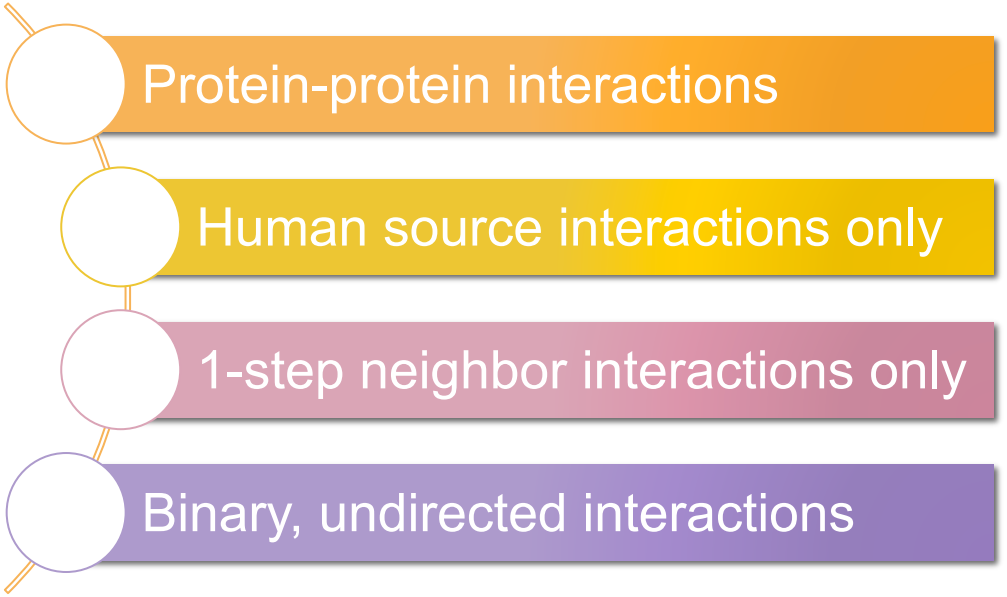
- DrugBank
- Clinical trial database (www.clinicaltrials.gov)
- Pharmaceutical company websites (how?)



Expanding on list of dementia targets

- Interologous Interaction Database (i2d)
- Source Interactions - curated from other data-sources
- 221 Drug Target proteins had atleast 1 neighbour



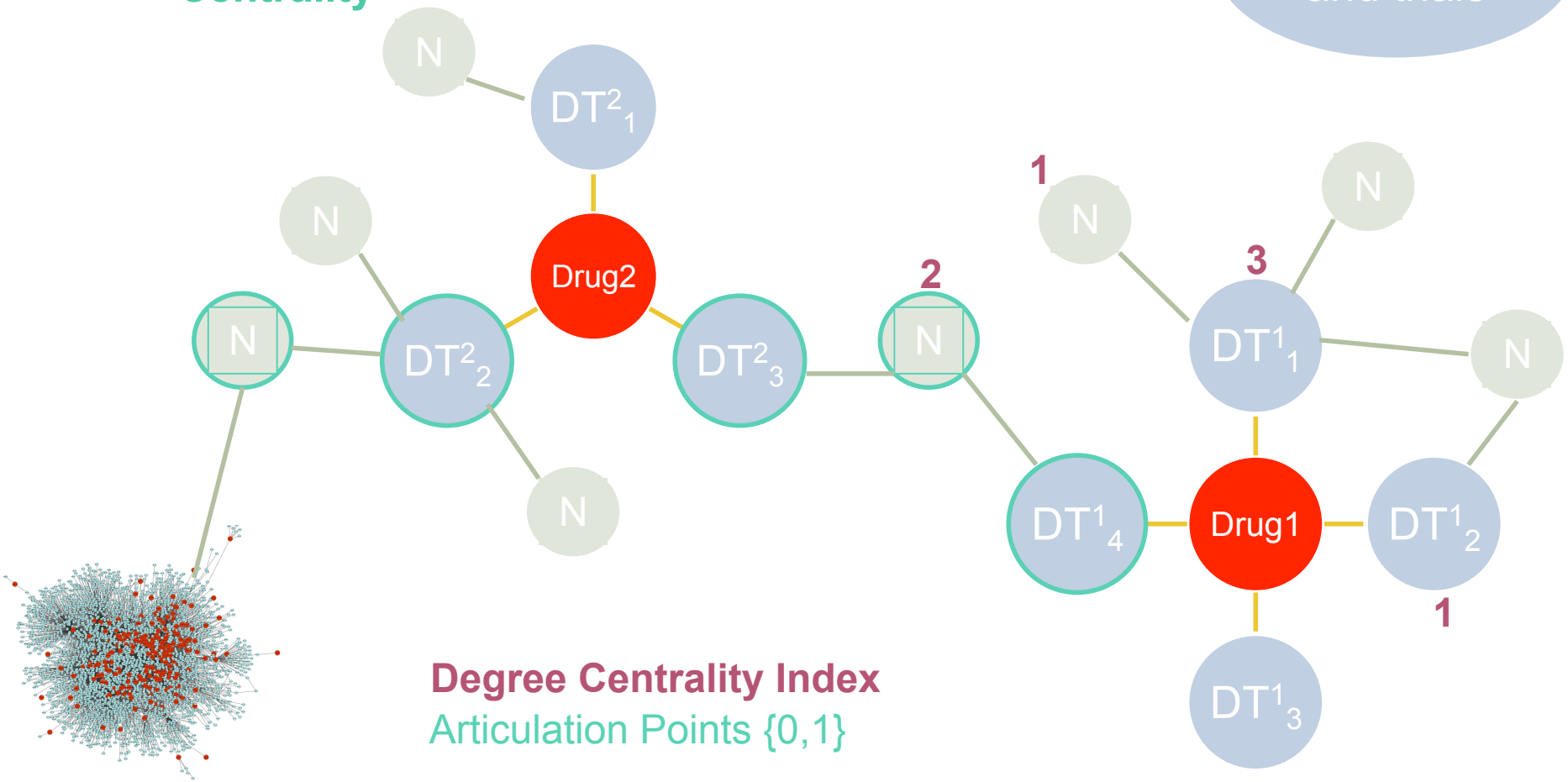


Build the network – (i)

Derived from i2d

DrugBank and trials

- Add topological characteristics
  - **Connectivity**
  - **Centrality**

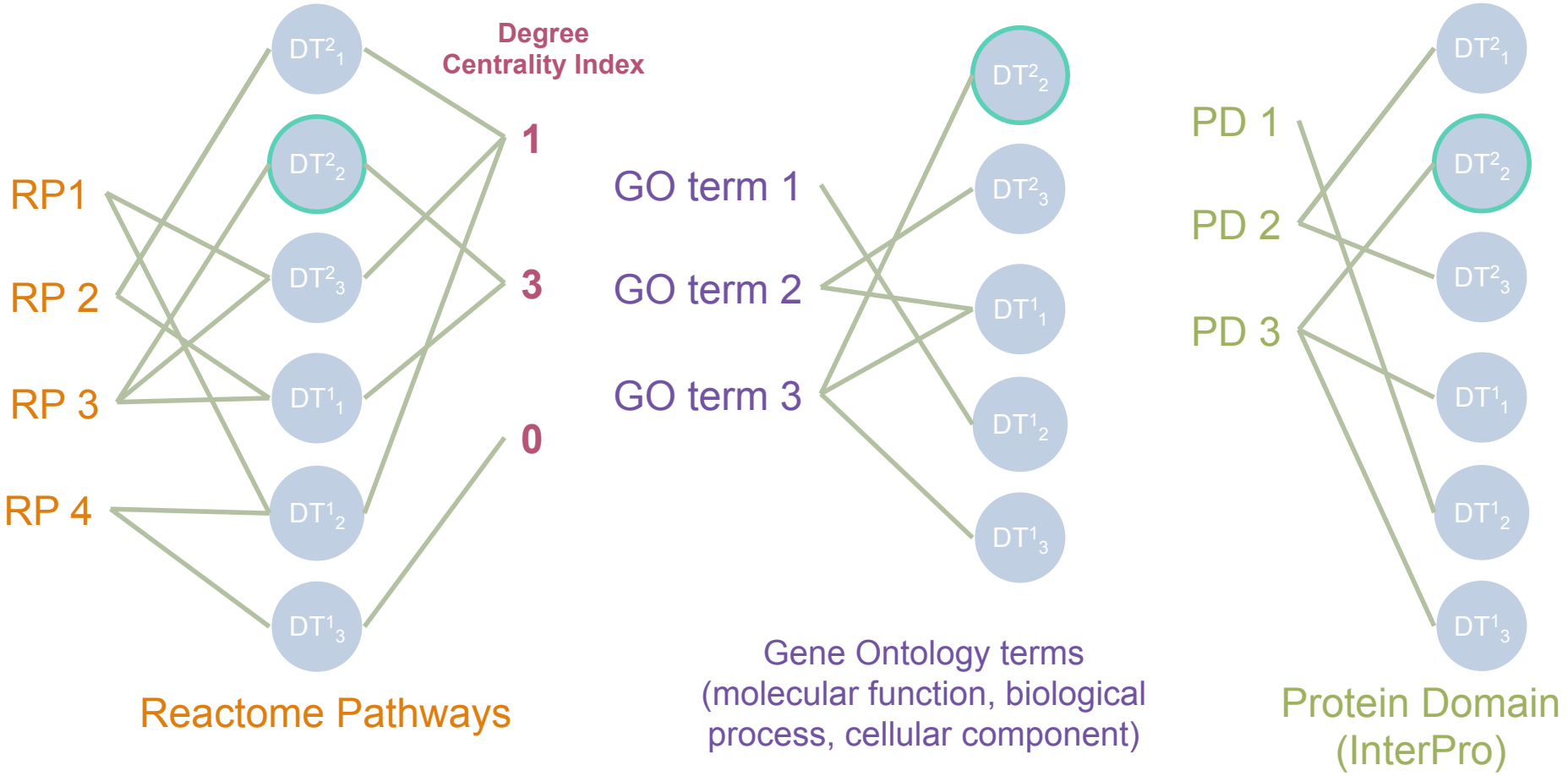


Degree Centrality Index

Articulation Points {0,1}

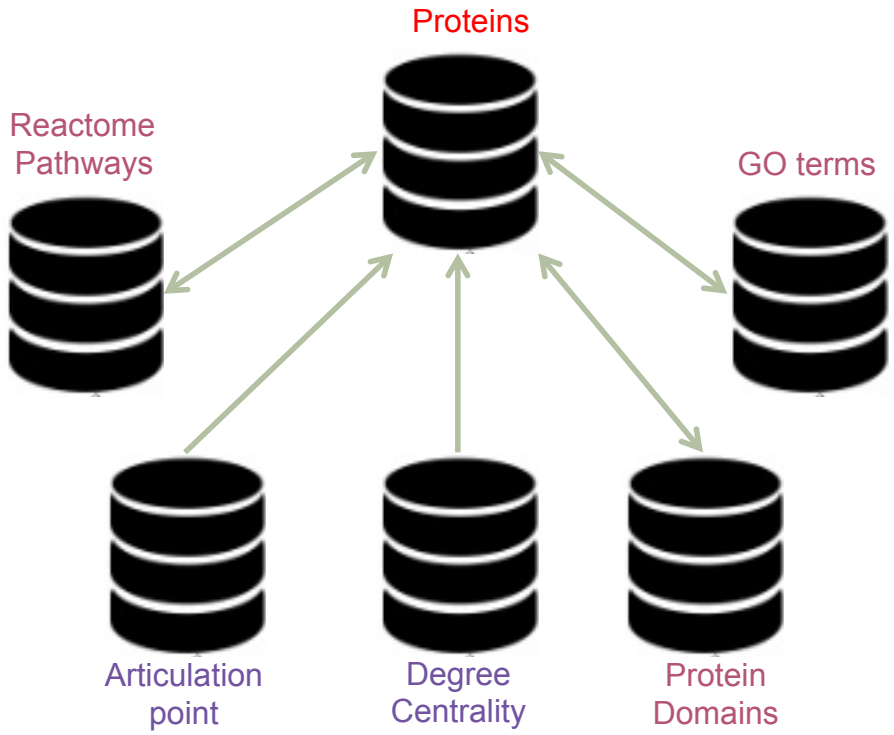
Build the network – (ii)

- Add functional data – multi relational database
- Many to many, one to many relationships (SQL)



Summary so far…..

- Created a multi-relational database
- Next step: Multi-relational association mining (MRAM)
  - Learn what rules make a Drug Target!



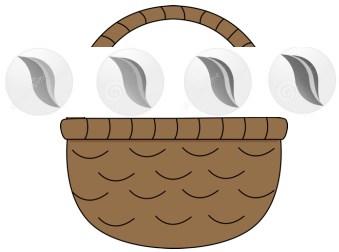
MRAM algorithm

- Predict dementia Drug Targets (rules to tell me what a potential DT for dementia might ‘look like’)

Feature vector
(for every marble)

- Centrality
- Articulation
- GO term
- Reactome Pathway
- Protein Domain

Known DTs
(Positive training examples)



Non-DTs
(Negative training examples)

MRAM algorithm

- Predict dementia Drug Targets (rules to tell me what a potential DT for dementia might 'look like')
- Like in any healthy relationship, a good association rule

  X => Y must have
  - **Support** : Y must occur above a certain threshold frequency in the given dataset (**Probability** of a result to occur)
  - **Confidence** : X => Y occurs at a certain frequency in the given dataset (**Importance** – log likelihood of X, given Y)
  - X and Y are disjoint
  - **Want High Probability (1) and Importance (>0.5)**

Results

- Learnt new rules (Probability = 1, Confidence > 0.5)
- MRAM: Apply rules on a network of proteins -> Get new putative Drug Targets
- Comparison with other methods
  - Decision Tree
  - Naïve Bayes
  - Neural Network

Results

- Learnt new rules
- Apply rules on a set of proteins -> Get new putative Drug Targets
- Comparison with other methods
- True cases = 221 DTs (1-step neighbours)
- False cases = 3 randomly selected sets of non-DTs, sizes 221, 500, 1000
- Network of 3112 proteins and 6541 interactions
  - 44433 GO terms
  - 11738 InterPro domains
  - 4240 Reactome pathways

Results

- Comparison with other methods
- True cases = 221 New Possible DTs (1-step neighbours)
- False cases = 3 randomly selected sets of non-DTs, sizes 221, 500, 1000
- Lift chart: Did we do better than a random guess approach
  - Randomly pick out possible Drug Target genes
  - Use the rules learnt by my model in order to pick out DT genes
  - Any improvement from random guess = ‘lift’!
  - Graph – poor resolution (can bash at end of presentation)

Measure Method	AUC	Likelihood Log Score	Likelihood Lift	Likelihood RMSE
$n_1 = 221$				
MRAM	**0.846**	**−0.259 ± 0.021**	**0.433 ± 0.020**	**0.211 ± 0.001**
Decision Tree	0.837	−0.405 ± 0.063	0.287 ± 0.063	0.213 ± 0.020
Bayesian Network	0.822	−0.540 ± 0.175	0.152 ± 0.175	0.284 ± 0.029
Neural Network	0.783	−0.416 ± 0.084	0.276 ± 0.084	0.224 ± 0.026

Table 2. Computational measures calculated for the 4 methods (n_1=221 is balanced set with equal number of true and false cases). 10 fold cross validation results.

Overfitting to their input data?

Measure Method	AUC	Likelihood Log Score	Likelihood Lift	Likelihood RMSE
n_3 = 1,000				
MRAM	**0.883**	**−0.211 ± 0.056**	**0.256 ± 0.054**	**0.063 ± 0.007**
Decision Tree	0.866	−0.265 ± 0.039	0.202 ± 0.040	0.166 ± 0.001
Bayesian Network	0.808	−0.263 ± 0.048	0.218 ± 0.047	0.130 ± 0.025
Neural Network	0.804	−0.293 ± 0.060	0.198 ± 0.057	0.174 ± 0.019

Table 2. Computational measures calculated for the 4 methods (n_1=221 is balanced set with equal number of true and false cases). 10 fold cross validation results.

MRAM Rules were learnt from the 268 Drug Targets
Positive set was derived from 221 of these 268 drug targets!

Results

- What are the new putative Drug Targets like?
- p38 MAPK has recently emerged as a disease treatment target for Alzheimer's
- Support for existing research – link between neurodegenerative dementia and metabolic disorders

Functional Enrichment Analysis	Pathway analysis
• GOrilla and REVIGO • GO *biological process terms* • Metabolic-related • Cell surface receptor signaling pathways (MAPK cascade) • Immune response • Apoptosis • Long-term memory	• DAVID • Similar results as GO • Alzheimer disease-amyloid secretase pathway • Type 2 Diabetes

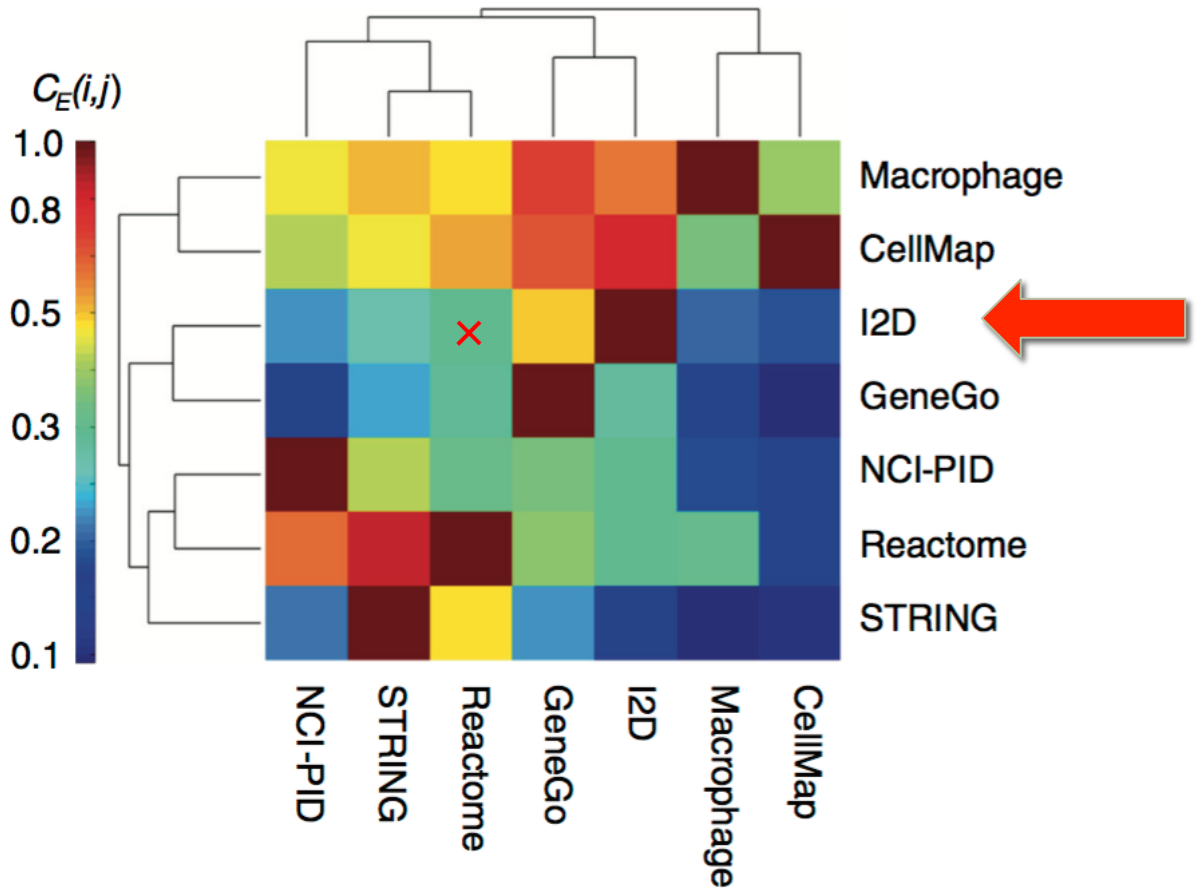
What did they do well?

- Integrated multiple different types of data
  - Interpolate data relations without merging them into one unanimous ‘mega-relation’
  - Good proof of concept
- Learnt new **rules** to predict possible genetic targets, instead of clustering different genes/proteins together

Methodology critique

- Use of interaction databases as prior knowledge
  - i2d source interactions (String? GeneGO?)
  - High-throughput experimental platforms (ex. yeast-two hybrid) known to have high false positive and false negative rates, and platform specific biases[1]
  - Compendium sources? (WikiPathways, Pathway Commons)

1. Kirouac, Daniel C., Julio Saez-Rodriguez, Jennifer Swantek, John M. Burke, Douglas A. Lauffenburger, and Peter K. Sorger. 2012. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. BMC Systems Biology 6:29.



[1]**Figure 4. Edge consistency between interaction databases.** The fractional edge overlap scroes (proportion of interactions consistent between 2 databases) are represented as a hierarchically clustered heatmap.

1. Kirouac, Daniel C., Julio Saez-Rodriguez, Jennifer Swantek, John M. Burke, Douglas A. Lauffenburger, and Peter K. Sorger. 2012. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. BMC Systems Biology 6:29.

Results critique

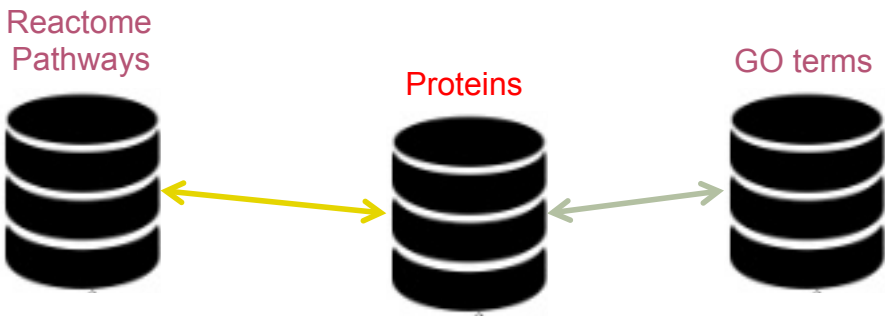
- False cases sets
  - Randomly selected sets of non-Drug Targets
  - What if these are n-step neighbours of known Drug Targets (n > 1)?
  - i.e. Excludes pathway interactions beyond 1-neighbour
- Accuracy of prediction
  - Rules were learnt from the 268 Drug Targets
    - Positive set was derived from 221 of these 268 drug targets!
    - Overfitting?
- GOrilla terms enrichment is based on the ‘top’ of a ranked gene list.

Questions?


WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS

Multi relational data mining[1]

- Traditional Data Mining (Knowledge Discovery in Databases)
  - Discover a high level knowledge from lower levels of relatively raw data
- MRDM: Handle multiple relations at the same time (Inductive Logic Programming)
- **Classification**: Separate data into distinct classes, supervised.
- **Clustering**: Separate data into distinct classes, unsupervised.
- **Association**: Discover relations between variables in the data, using real valued prediction variables. Rule learning.



1. Domingos, P. Prospects and challenges for multi-relational data mining. *ACM SIGKDD Explor. Newsl.* **5**, 1–4 (2003).

InterPro

- Uses predictive signatures (from numerous databases) in order to classify proteins, and predict the presence of domains and important sites.
- Combines protein function recognition methods of member databases of InterPro into one application.

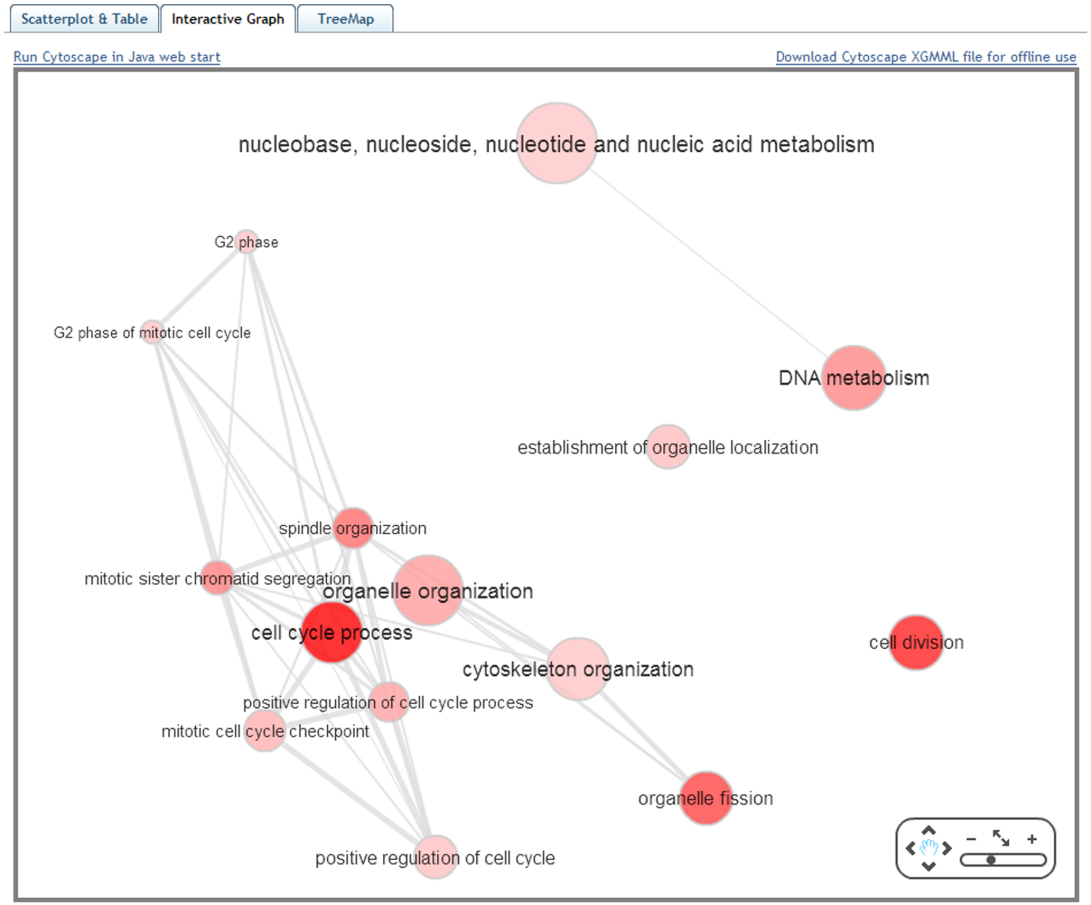
InterProScan: protein domains identifier

Table 1
Database members and their applications

Database	Application
ProDom (6)	BlastProDom (Blastall) (4)
PRINTS (7)	FingerPrintScan (8)
SMART (9)	Hmmpfam (http://hmmer.wustl.edu/)
TIGRFAMs (10)	Hmmpfam (http://hmmer.wustl.edu/)
Pfam (11)	Hmmpfam (http://hmmer.wustl.edu/)
PROSITE (12)	ScanRegExp + ProfileScan (13)
PIRSF (14)	Hmmpfam (http://hmmer.wustl.edu/)
SUPERFAMILY (15)	Hmmpfam (http://hmmer.wustl.edu/)
CATH (16)	Hmmpfam (http://hmmer.wustl.edu/)
PANTHER (17)	Hmmsearch (http://hmmer.wustl.edu/)
SignalPHMM	SignalPHMM (18)
Transmembrane	TMHMM2.0 (19)

REVIGO

- Takes in list of GO terms (with associated p-values)
- Finds representative subset of terms using a clustering algorithm



GOrilla

- Process/Function/Component/All ontologies
- Identifies the list of GO Terms enriched in a list of genes
- Statistical model to identify which GO terms are enriched significantly at the *TOP* of a ranked gene list

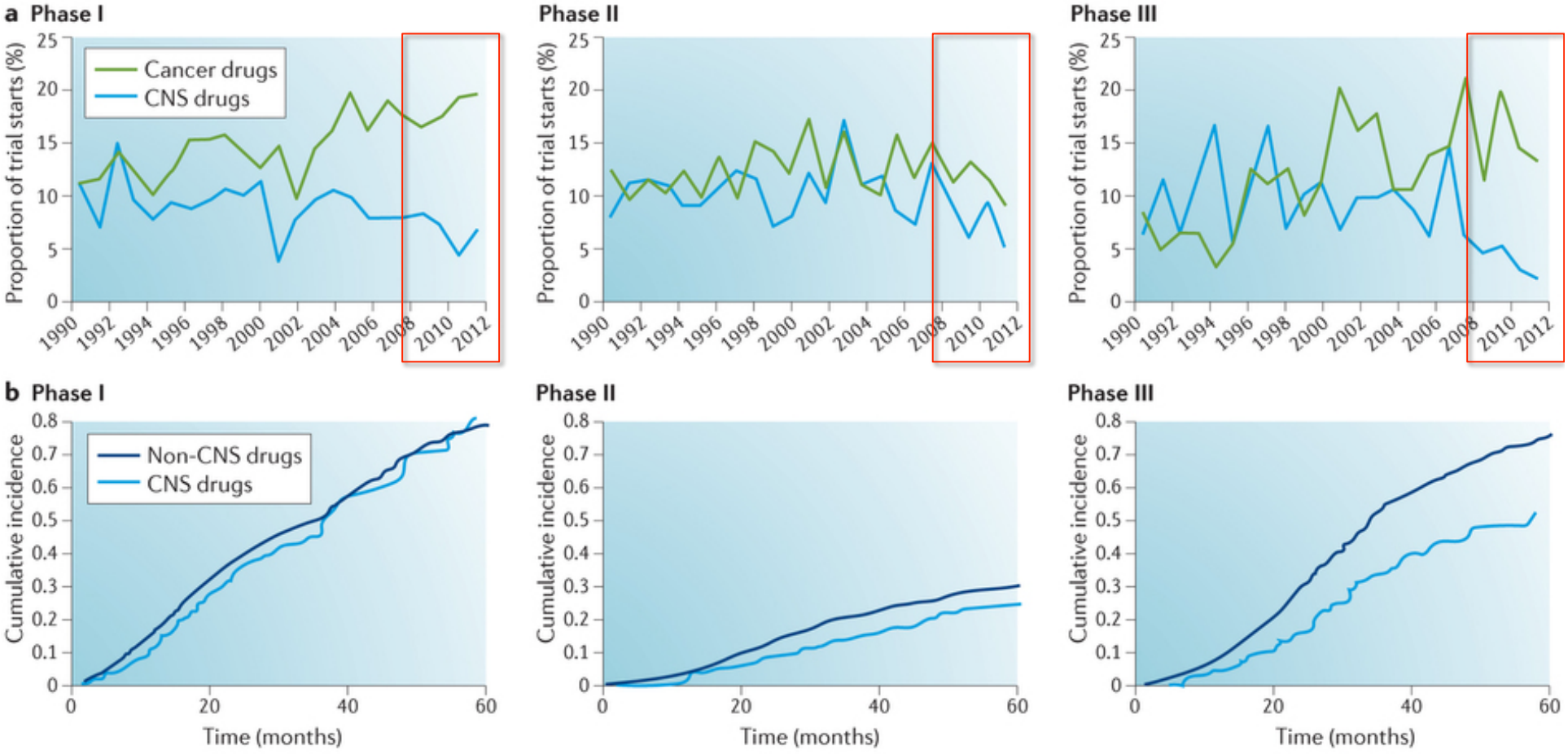
DAVID

- Database for Annotation, Visualization, and Integrated Discovery
- Annotation tools for Gene function
- Can identify which biological ‘themes’ are enriched (GO terms)
- Which pathways are enriched (KEGG and BioCarta)
- Uses a population background for enrichment analysis
  - Corresponding genome-wide genes with atleast one annotation in the analyzing categories

Finding drug targets

- Known, pharmacologically validated targets[1]
  - Classical neurotransmitter or neuropeptide systems
- Except for Parkinson's, no logical relationship exists between any transmitter system and the circuitry affected by a neurodegenerative process[1]
- No disease modifying therapies approved for Alzheimer's or Parkinson's.
- To understand determinants of disease expression:
  - Molecular networks: protein interaction, metabolic, regulatory
  - Phenotypic networks: coexpression, genetic

1. New Prospects and Strategies for Drug Target Discovery in Neurodegenerative Disorders. Hilbush et al. NeuroRx. 2005 Oct; 2(4): 627-637



Trends in drugs entering clinical trials for central nervous system disorders compared with other therapeutic areas: 1990–2012.

1. Kesselheim et al. Two decades of new drug development for central nervous system disorders. Published online 20 November 2015.

Alzheimer's Research UK

The Power to Defeat Dementia

How are we developing treatments for dementia?

1 Making a discovery

Understanding the biology of diseases that cause dementia.

- Is this process important in the disease?
- Could it be a target for new treatments?

2 Validating the discovery

Potential targets are studied more closely.

- What does the target do?
- Can we test it in the lab?

3 Searching for compounds that hit the target

Libraries of chemicals are screened to find those that could be developed into experimental drugs.

4 Developing hits to leads

Hit compounds are modified by chemists to improve their properties and activity.

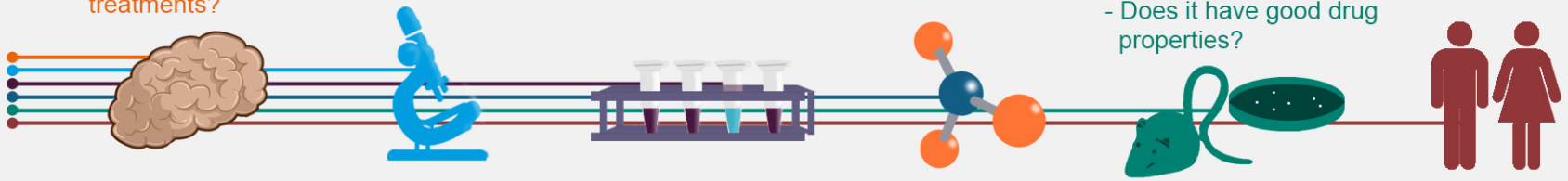
5 One step towards the clinic

Hit compounds go through many rounds of tweaking and re-testing.

- Does it work in a living system?
- Does it have good drug properties?

6 Clinical trials

Potential new treatments must pass through three phases of clinical trials in people to test if they are safe and effective.

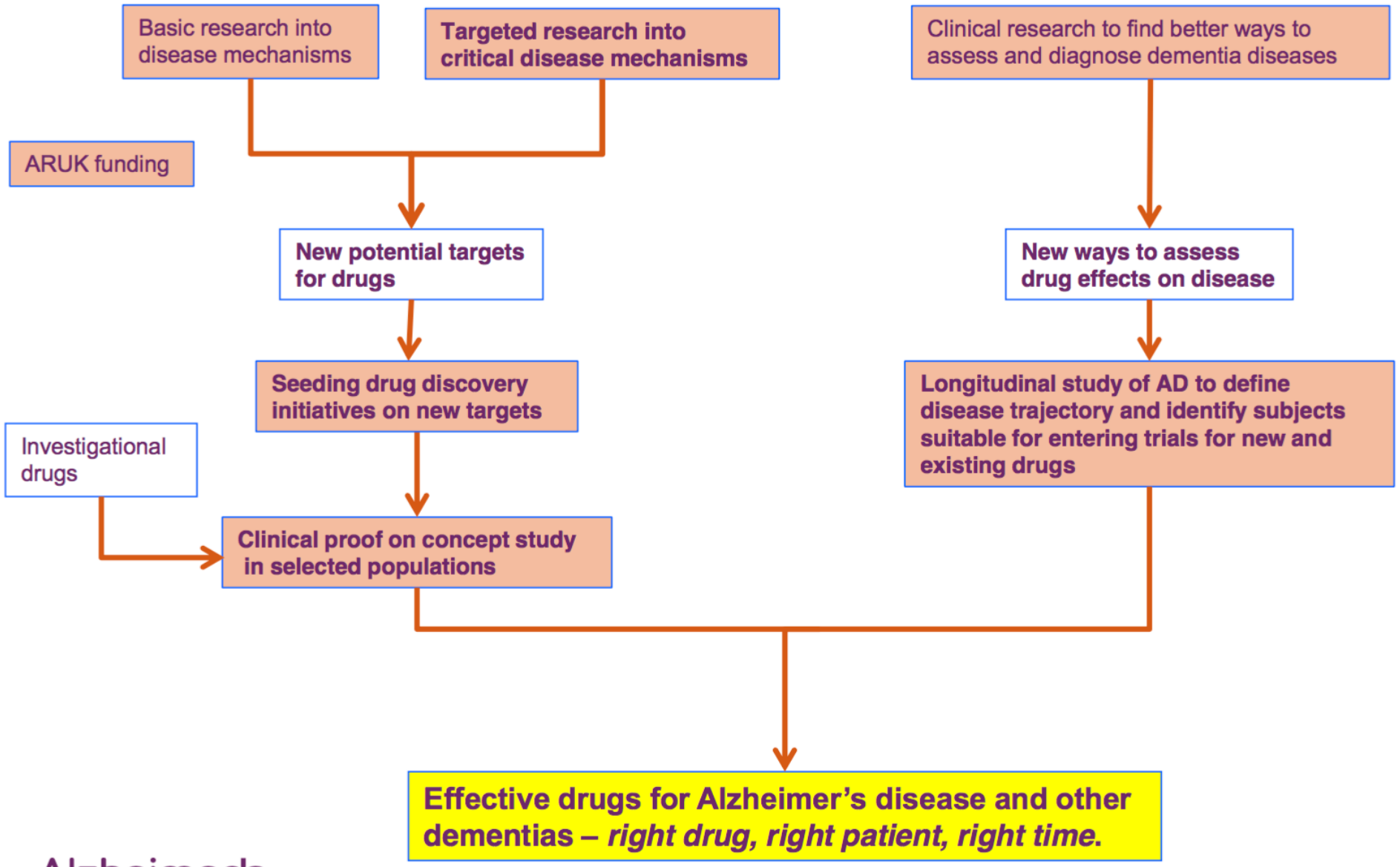


Where is Alzheimer's Research UK making a difference?

Our grant schemes fund all stages of the process with many focusing on the earliest discoveries.

Our Drug Discovery Alliance, Drug Discovery Fund, Dementia Consortium and Stem Cell Research Centre focus on turning these discoveries into promising leads.

Our Global Clinical Trials Fund has the potential to take the best leads into the clinic.



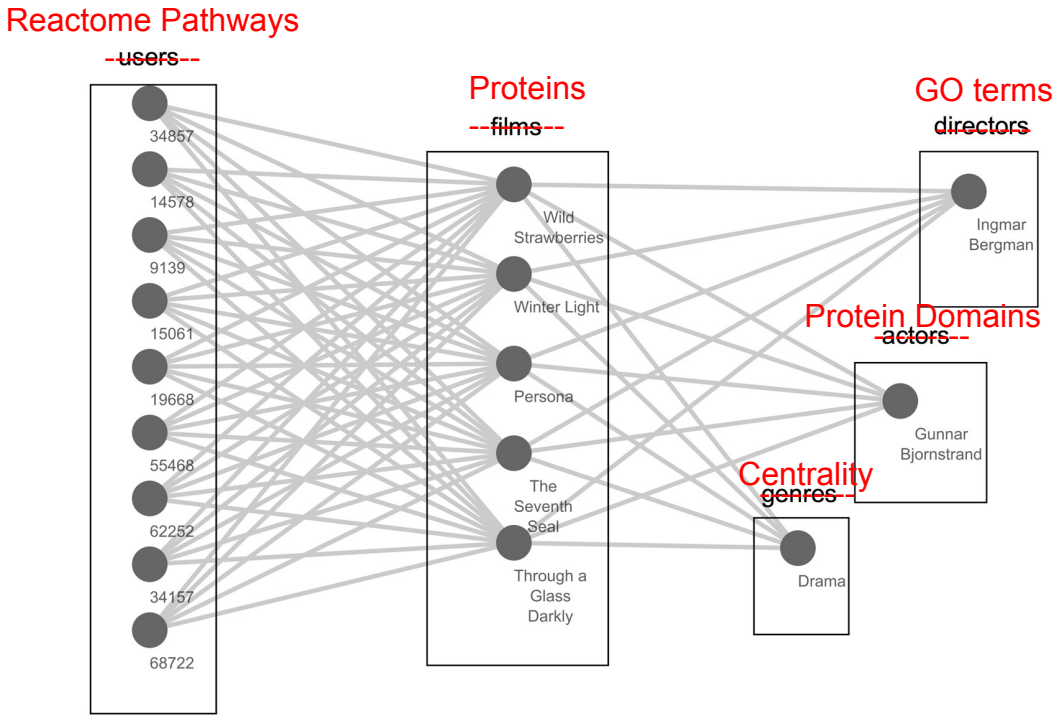
Alzheimer's Research UK
Defeating Dementia

Orphan Drug Act

- Offered incentives to pharmaceutical companies to invest in dieases where there would otherwise be no compleling economic argument to allocate resources
- 300 “orphan drugs” approved under its provisions
- Gives market exclusivity for 7 years (originally 5)

The novelty – multi relational association mining

- Created a multi-relational database
- Multi-relational data mining



Our novelty: Multi relational association mining

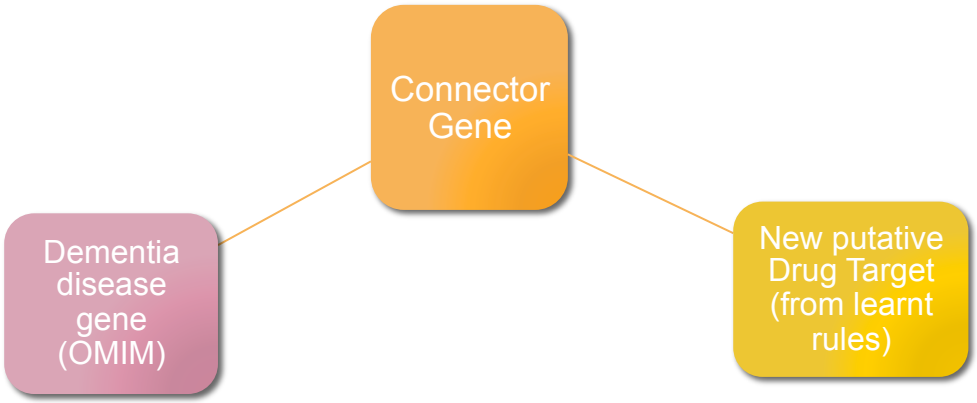
- Multi Relational Data Mining
  - Compact data representation
  - No pre-processing required (Integration with relational database)
  - Explore multiple data from multiple data sources
- Mining Association Rules
  - Given a database with a type of transactions,
  - Find all rules that correlate presence of one set of items with that of another set of items
  - *Classically, does not work for multiple, heterogeneous sets of transactions*
  - *Enter: MRAM* (part of SQL Server 2012 Analysis Services)
- Predict dementia Drug Targets (rules to tell me what a potential DT for dementia might 'look like')

Results

- Learnt rules that ‘make’ a potential Drug Target
- Use ‘connector genes’ from their previous paper, as input

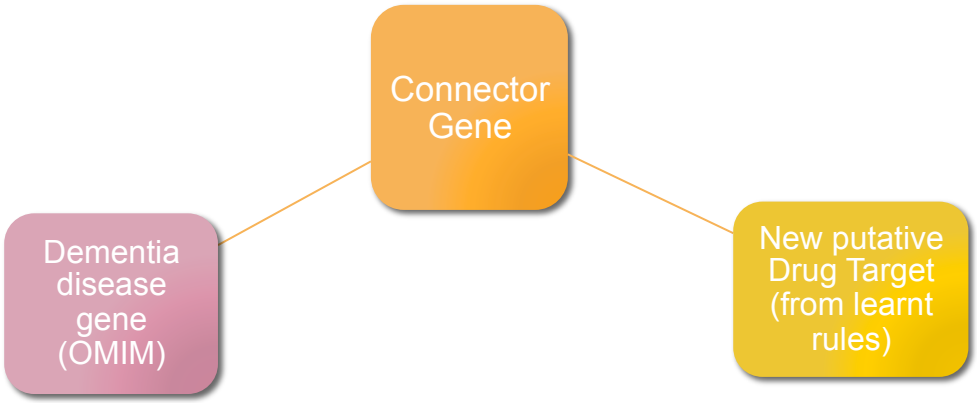
Results

- Learnt rules that maketh a potential Drug Target
- Use ‘connector genes’ from their previous paper, as input
- Which of these connector genes link a new putative drug target to a dementia disease gene (OMIM)?



Results

- Learnt rules that maketh a potential Drug Target
- Use ‘connector genes’ from their previous paper, as input
- Which of these connector genes link a new putative drug target to a dementia disease gene (OMIM)?



Results

- Are there particular data features (from the 5 different ones in our relational database), that the rules seem to use more of?
  - Leave-one-out experiment for each feature
  - Calculate likelihood lift subsequently

Experiment	$n_1 = 221$	$n_2 = 500$	$n_3 = 1{,}000$
Exp1: All data features excluding the topological data features	0.391	0.412	0.211
Exp3: All data features excluding the GO data feature	0.398	0.443	0.220
Exp4: All data features excluding the Reactome data feature	0.408	0.436	0.215
Exp5: All data features excluding the InterPro data feature	0.411	0.449	0.237
Exp6: All of investigated data feature	**0.433**	**0.469**	**0.256**

Table 3. Performance of MRAM in term of likelihood lift with different subsets of data features and the three sets of negative examples with different sizes.

Graphlet (Dominating Biological Networks)

- “Biologically central” genes are expected to exhibit some topological centrality to the rest of the proteins in the human PPI network.
- Graphlet = *induced* subgraph of a large network

Drugbank & Clinicaltrials.gov

- Drugbank
  - 4800 drug entries (FDA approved, experimental)
- Clinicaltrials.gov
  - NIH funded
  - Registry and results database
  - Publicly and privately supported clinical studies
  - Human participants
  - Intervention or retrospective studies also included
  - Query uses MeSH terms
- MeSH (Medical Subject Headings)
  - Tree with 16 main branches (organisms, diseases, drugs etc)
  - Index articles using a controlled vocabulary

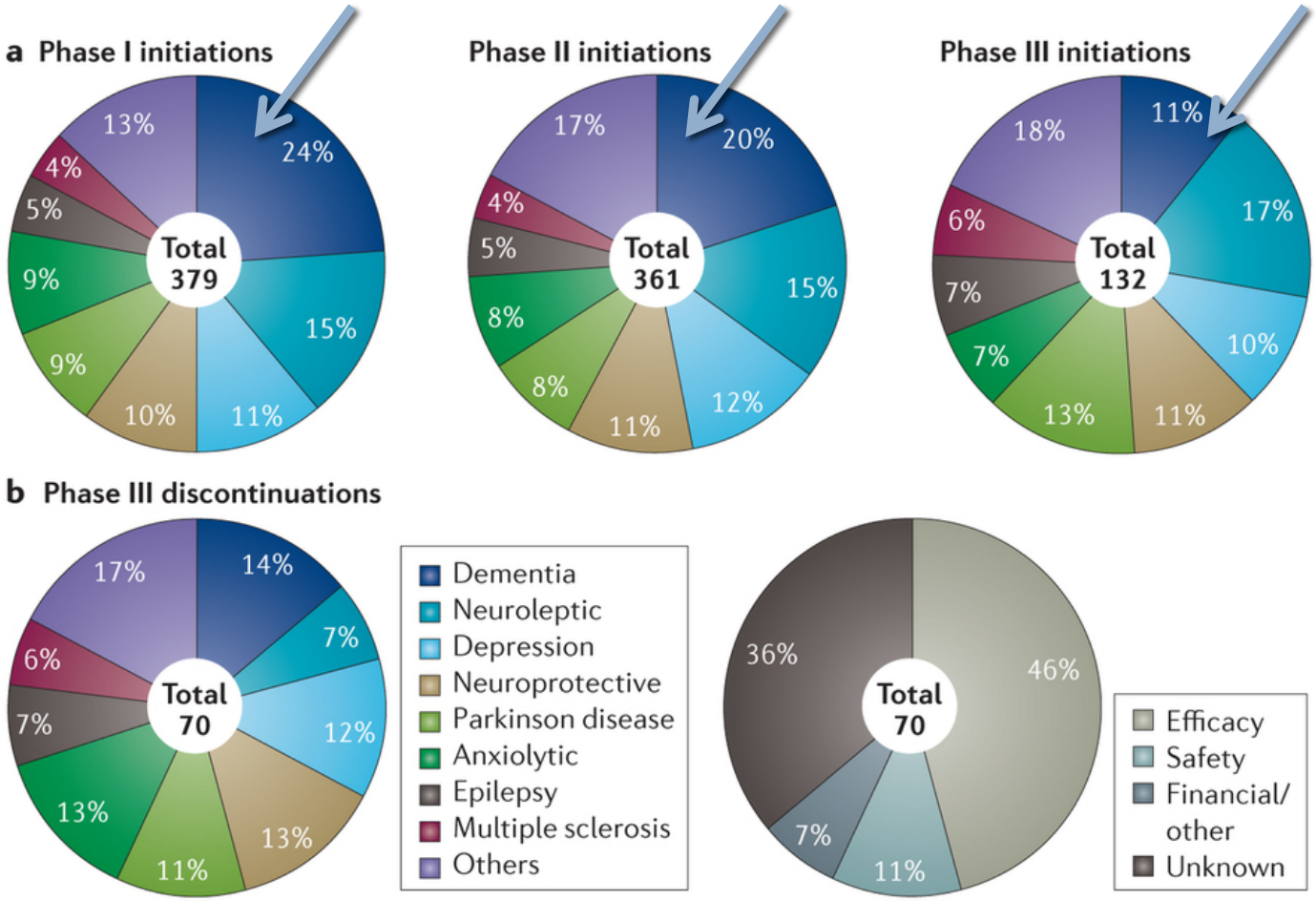
Reactome

- Curated, peer-reviewed knowledgbase of biological pathways
- Includes metabolic pathways, protein trafficking and signaling pathways
- Reactions : experimentally confirmed, manually inferred, and electronically inferred
- >2700 proteins, 2800 reactions ,and 860 human pathways.
- Has GeneOntology Ids for metabolic process of each reaction

Interologous Interaction Database

- Maps experimental interactions determined in model organisms, into human interactions
- Contains predicted interactions among human proteins

1. Kirouac, Daniel C., Julio Saez-Rodriguez, Jennifer Swantek, John M. Burke, Douglas A. Lauffenburger, and Peter K. Sorger. 2012. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. BMC Systems Biology 6:29.



Drugs in clinical trials for central nervous system disorders: 1990–2012.
46% (32) drugs were discontinued owing to inadequate efficacy (generally described as no improvement vs placebo)

1. Kesselheim et al. Two decades of new drug development for central nervous system disorders. Published online 20 November 2015.