Machine Learning for Scientific Discovery

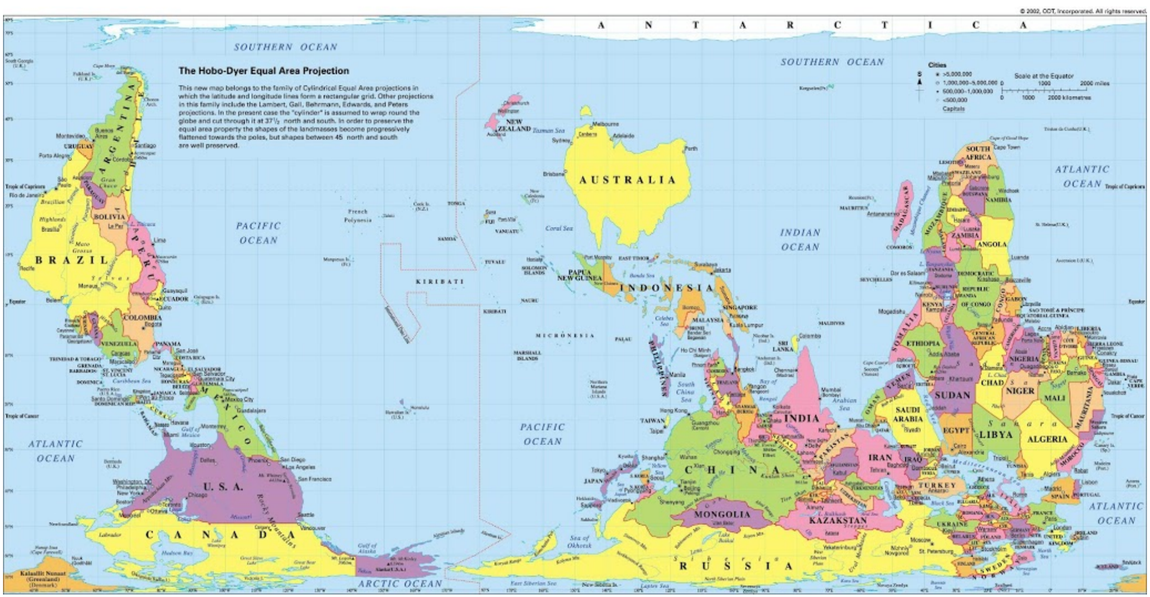
Cheng Soon Ong

Machine Learning Research Group
Data61 | CSIRO, Canberra

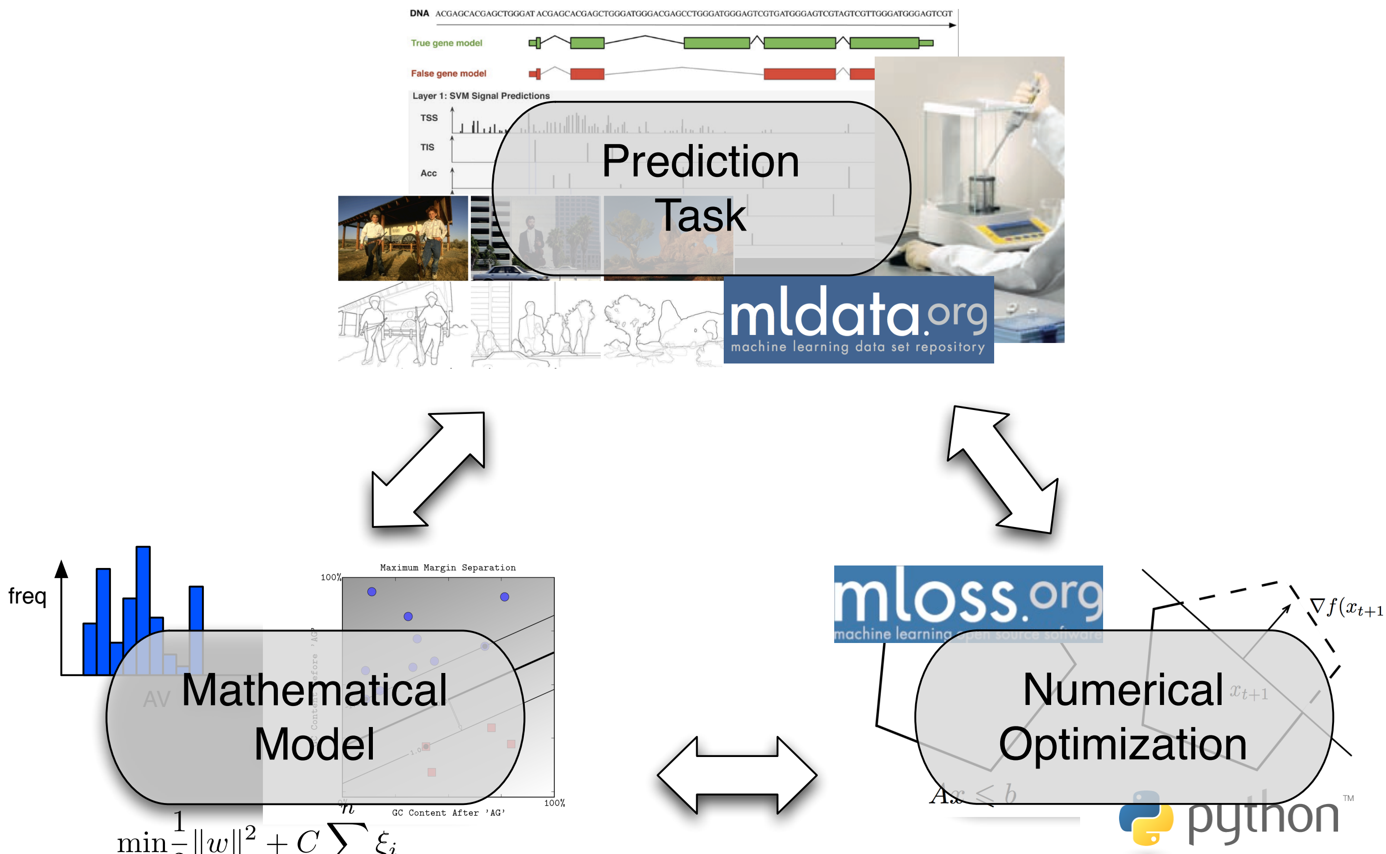
25 November 2016
Faculté Informatique et Communications, EPFL

Data61

- NICTA merger
- Part of CSIRO, focus on ICT
- Approx 1000 researchers, PhD students and university staff



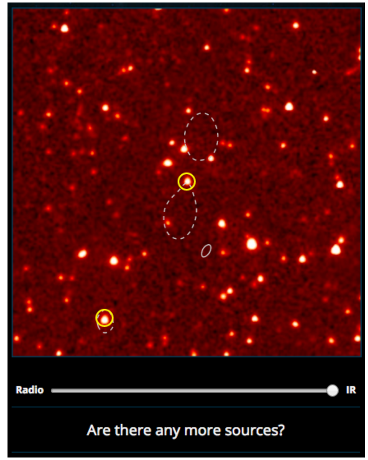
Applications - Optimization - Models

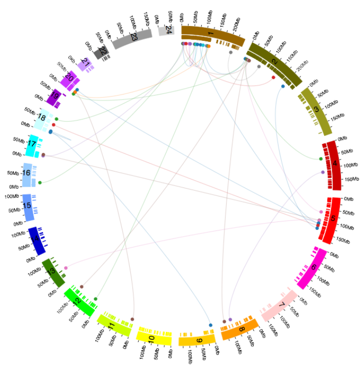


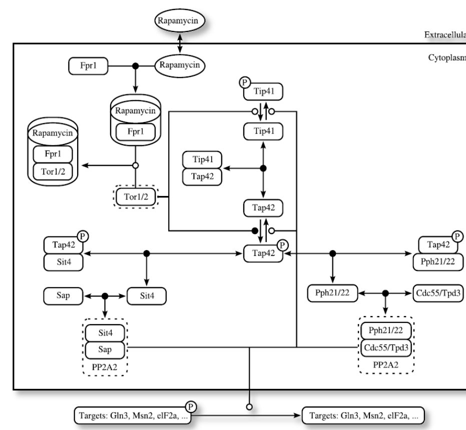
$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) \geqslant 1 - \xi_i$$

Machine Learning and Science







What is machine learning?

Machine learning is about prediction

Examples/features	$x_1, \ldots, x_n \sim \mathcal{X}$
Labels/annotations	$y_1, \ldots, y_n \sim \mathcal{Y}$
Predictor	$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$

Estimate best predictor = training

Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.

- No mechanistic model of the phenomenon
- There is relatively large amounts of data (examples, x usually $\mathbb{R}^d$)
- The outcomes (labels, y usually binary) are well defined

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

Today: focus on the predictor

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

Label: Finding black holes

- Exist physical models, we directly use images
- There is relatively large amounts of data (examples)
- Object localisation with crowd labels

Feature: Finding genetic associations

- No mechanistic model of the phenomenon
- High dimensional low sample size
- Stability of feature selection

Predictor: Finding good experiments

- Partial mechanistic model of the phenomenon
- Estimate the expected information gain

Discuss challenges to applying machine learning

Not standard binary classifcation

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

Finding black holes

Goal: Automate radio cross-identification, a problem in astronomy

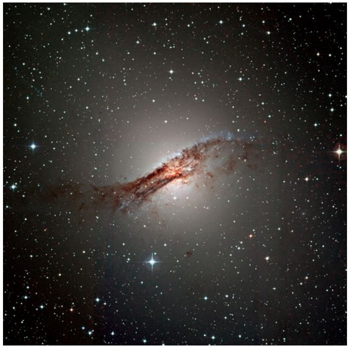
Too much data

- Collaboration with ANU, ANTF, CAASTRO
- Square kilometer array (South Africa and Australia)

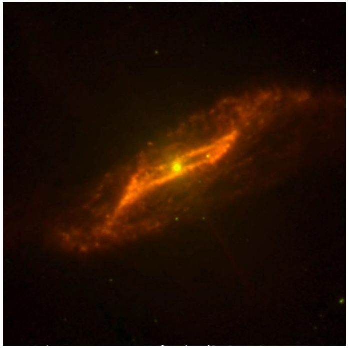
Labelled by non-experts

- Convert object localisation to binary classification
- Deal with label noise

Radio cross-identification

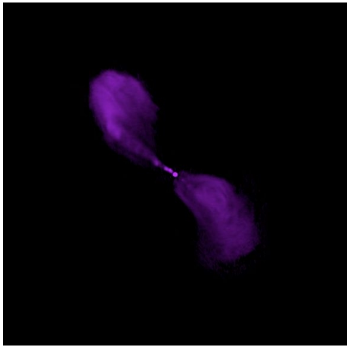


Optical



Infrared

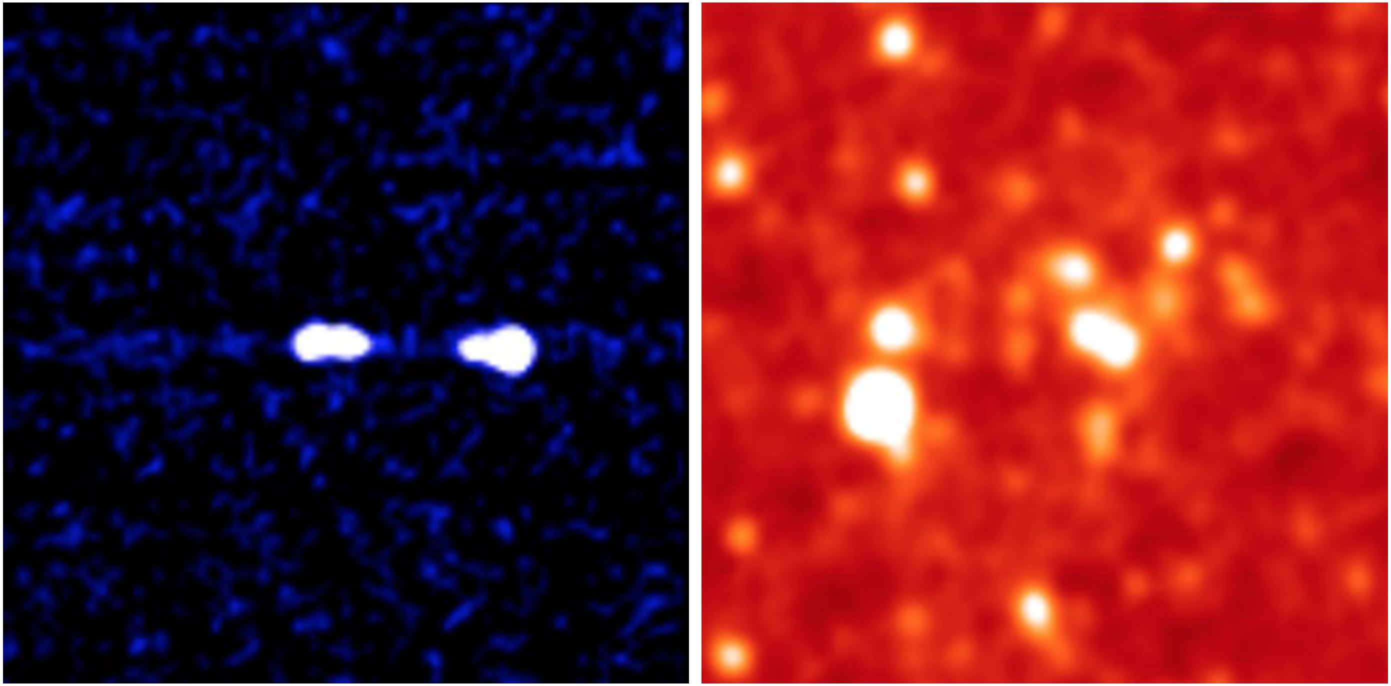
X-ray



Radio

Images of Centaurus A at different wavelengths.

The real data

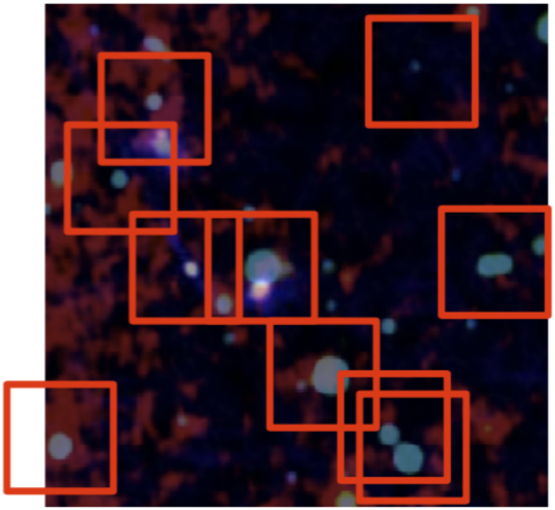


The same patch of sky in both radio (left) and infrared (right)

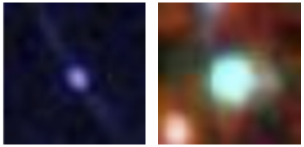
Localisation as binary classification

Galaxy catalogue as candidates

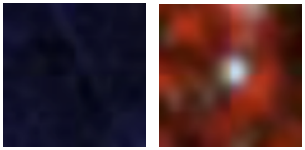
Could scan a patch across the sky



Classify pairs of images



positive



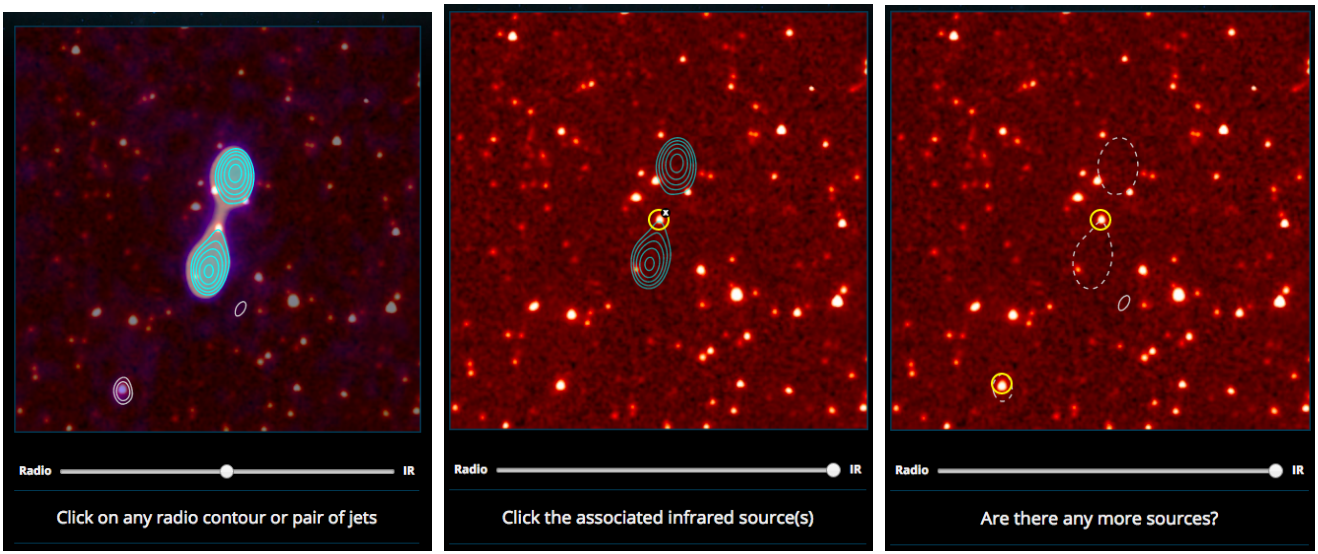
negative

Features: Neural network image features, fluxes, radial distance

https://github.com/chengsoonong/crowdastro

Crowdsourcing labels

Radio Galaxy Zoo:
citizen science project to cross identify radio galaxies



Radio Galaxy Zoo

About 100000 of 177000 image pairs labelled.

- 5 volunteers per pair for compact sources
- 20 volunteers per pair for complex sources

How to find black holes

Prior catalogues

- Heuristic rules + expert human effort
  Norris et. al. 2006
- Annotation based on physical models
  Fan et. al. 2015
- Use set where both agree as gold standard

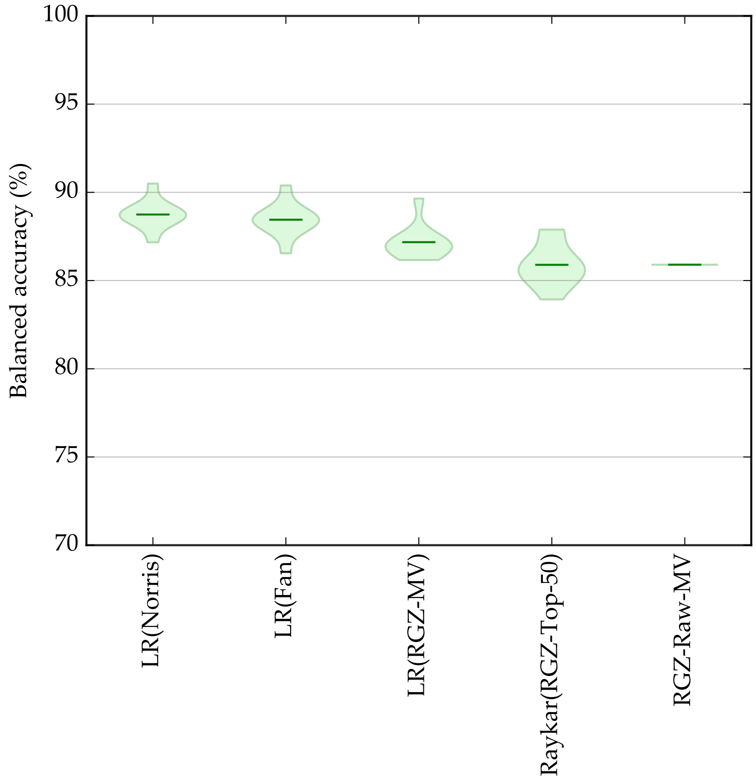
Many labels to one binary label

- Logistic regression from `sklearn`
- Majority vote
- EM style algorithm to estimate ground truth
  Raykar et. al. 2010, Yan et. al. 2010

Latent variable model

- Noisy labels = ground truth + biased coin flip

Results



Conclusion: Features meaningful, but pipeline can be improved.

Side note about label noise

Latent variable

Assume that there is a hidden ground truth label, and model it.

Alger, Banfield, Ong, (in preparation)

Learning with label noise

During training, pretend that labels are noiseless, and assume that the learning algorithm takes care of it.

Menon, van Rooyen, Ong, Williamson, ICML 2015

Model evaluation

How do we measure performance without ground truth?

What are good features?

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

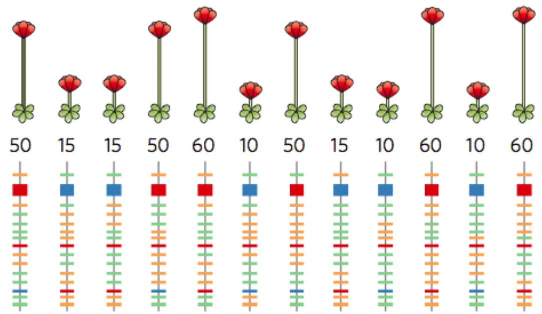
Genome wide association study

Case-control studies
A cohort of sick individuals (cases) and healthy individuals (controls) are genotyped and their corresponding binary phenotype are recorded.

We use the framework of hypothesis testing

Hypothesis testing Given a case control study, test whether a particular SNP is associated with the phenotype.

Good biomarker? If difference is statistically significant
$\Longrightarrow$
SNP is associated with the phenotype.



Epistatic Interactions

Genome Wide Interaction Search (GWIS)

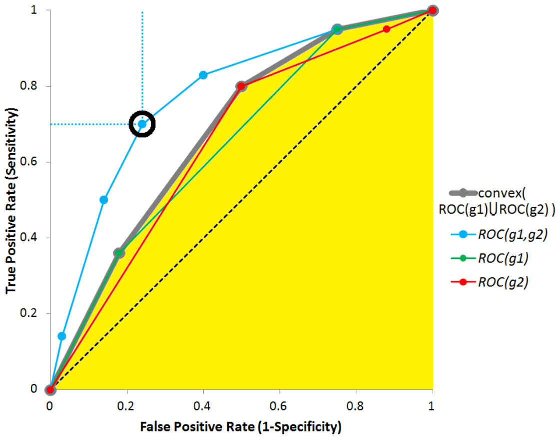
Consider the association of all pairs of genotypes to phenotypes

Large search space

- 5000 individuals, 500,000 SNPs (WTCCC)
- Need to tabulate 125 billion contingency tables

Classification based analysis

- Focus on SNPs in case control studies
- New statistical tests
- Consider specificity and sensitivity
- Gain over univariate ROC
- CPU ($\approx$ days) and GPU ($\approx$ hours)
- Store the top 1 million pairs



Web service

`http://gwis1.research.nicta.com.au/`

Goudey,...,Ong,...,Kowalczyk, BMC Genomics, 2013

p-values

Interpreting p-values

Is 10^{-10} probability of association very significant?

Quote

`... but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.`

Fisher, The Design of Experiments, 1947, p. 14

Stability of scoring

We consider p-values as a score of association.

- How stable is this score if we repeat the experiment?
- How do we combine scores?

Challenges

- Scores available for only the top-k examples
- Scores from different sources not calibrated

How to represent ranks?

Multiple ways to represent ranks

- Ordered list of n objects selected from Ω
- List of values $[1, \ldots, n]$ (the ranks of the object)
- Normalised ranks $\in (0, 1)$
- Permutation mapping $R : \Omega \to (0, 1)$

Measuring Overlap

Motivation

Given a set of replicated experiments, how do we measure overlap?

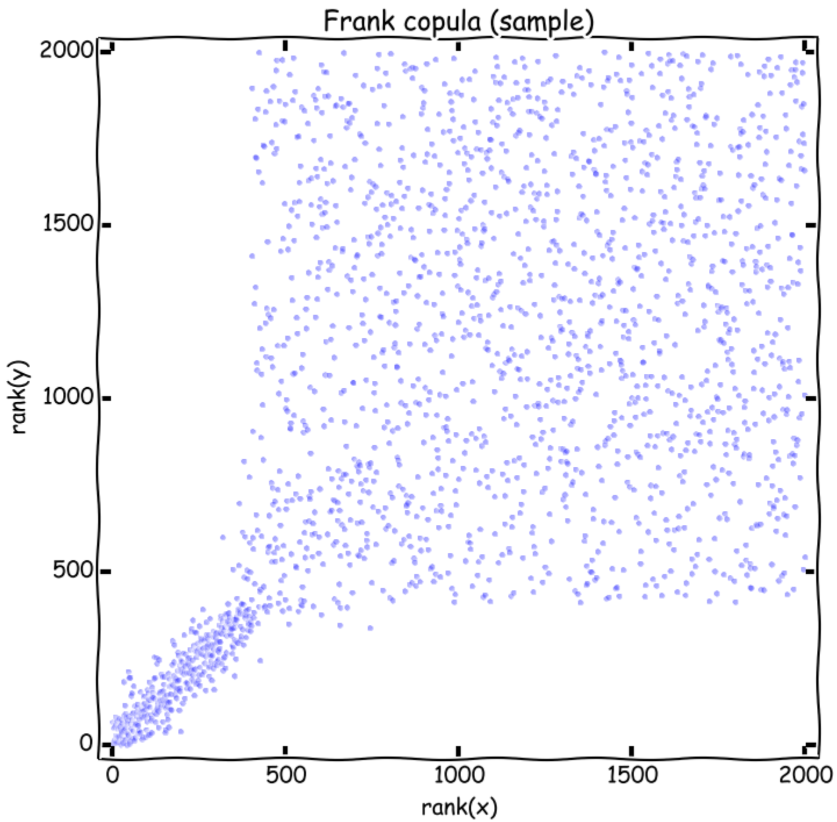
Examples

- Perform repeated splits of the data
- Experiments on different cohorts
- Multiple sources of information

Challenges

- Scores available for only the top-k examples
- Scores from different sources not calibrated

Signal and Noise



Set based overlap

Running example (6 objects)

$$A = [a, b, c, d, e, f]$$

$$B = [a, b, e, f, c, d]$$

Jaccard Index

$$\text{overlap} = \frac{|A \cap B|}{|A \cup B|}$$

Measuring stability

- Easy to compute
- Works for top-k lists
  Consider the top-3 lists from above:

$$\text{Jaccard index} = \frac{|\{a, b\}|}{|\{a, b, c, e\}|} = \frac{1}{2}$$

- Ignores the order given by scores

Spearman's ρ

- Similar to Pearson's correlation for the measure of dependence
- Spearman's ρ is a correlation measure between ranked lists

$$\rho(A, B) := \frac{\sum_i (r_A^{(i)} - \bar{r}_A)(r_B^{(i)} - \bar{r}_B)}{\sqrt{\sum_i (r_A^{(i)} - \bar{r}_A)^2 \sum_i (r_B^{(i)} - \bar{r}_B)^2}},$$

- Running example:

$$\rho([a, b, c, d, e, f], [a, b, e, f, c, d]) = 0.543$$

(Jaccard index $= 1$)

- Need the same elements in A and B

$$\rho([a, b, c], [a, b, e]) \text{ ?}$$

Spearman's ρ on top k lists

Simple idea

Define Spearman's ρ for top k lists

Key observation

Any elements in list A that do not appear in list B must have a rank higher than the number of elements in B

Running example (top-3)

$$A = [a, b, c, d, e, f] \quad \text{and} \quad B = [a, b, e, f, c, d]$$
$$A_3 = [a, b, c] \quad \text{and} \quad B_3 = [a, b, e]$$
$$A_3^{\overset{B_3}{\rightarrow}} = [a, b, c, e] \quad \text{and} \quad B_3^{\overset{A_3}{\rightarrow}} = [a, b, e, c]$$

Spearman's $\rho = \rho(A_3^{\overset{B_3}{\rightarrow}}, B_3^{\overset{A_3}{\rightarrow}}) = 0.8$

Spearman's ρ on top k lists

Extend the list

We expand lists A and B to complete rankings over the same set of elements, denoting them as $A^{\overset{B}{\rightarrow}}$ and $B^{\overset{A}{\rightarrow}}$ respectively.

The missing values in the extension are given the average rank.

Running example (top-4)

$$A_4 = [a, b, c, d] \qquad \text{and} \qquad B_4 = [a, b, e, f]$$

$$A_4^{\overset{B_4}{\rightarrow}} = [1, 2, 3, 4, 5.5, 5.5] \qquad \text{and} \qquad B_4^{\overset{A_4}{\rightarrow}} = [1, 2, 5.5, 5.5, 3, 4]$$

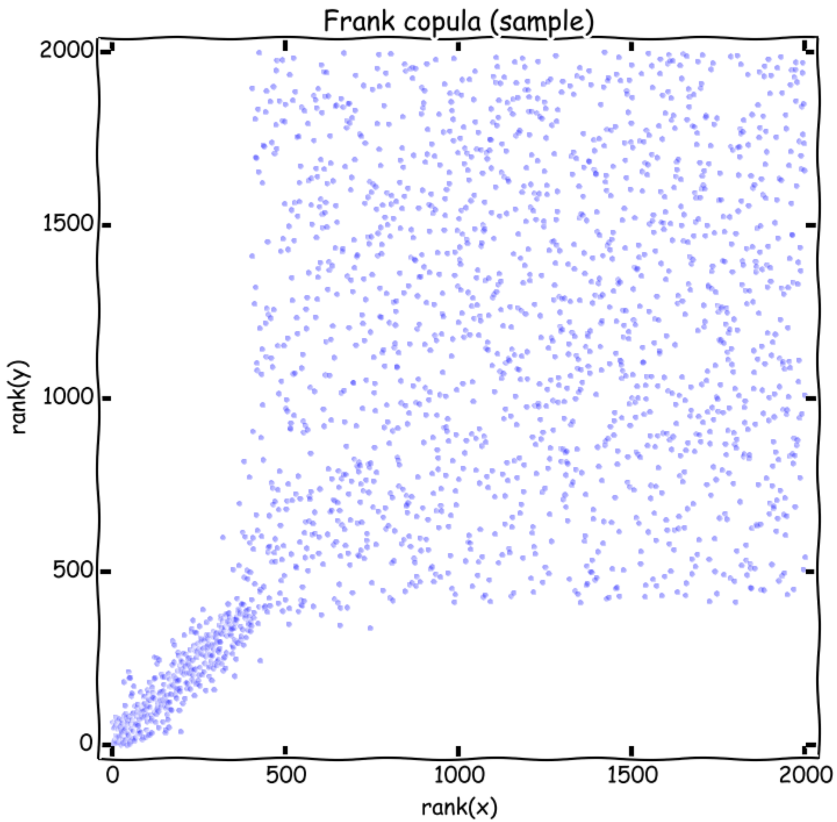
Makes no assumption about the order of the unranked objects

Other possible imputation approaches

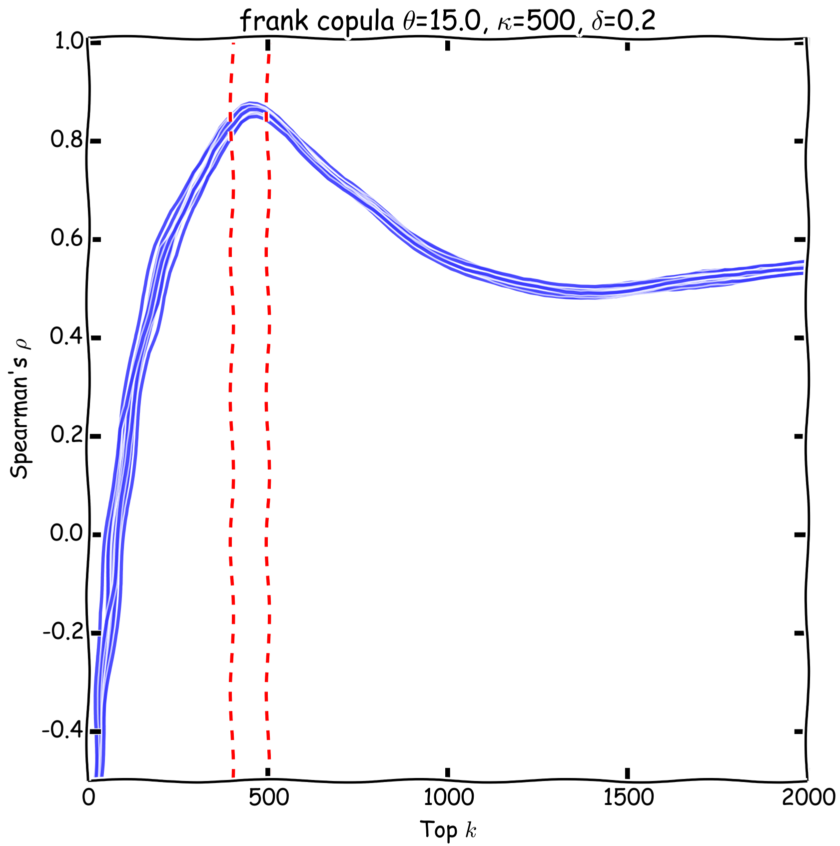
- Optimistic
- Worst case

Bedő, Rawlinson, Goudey, Ong, PLoS ONE, 2014

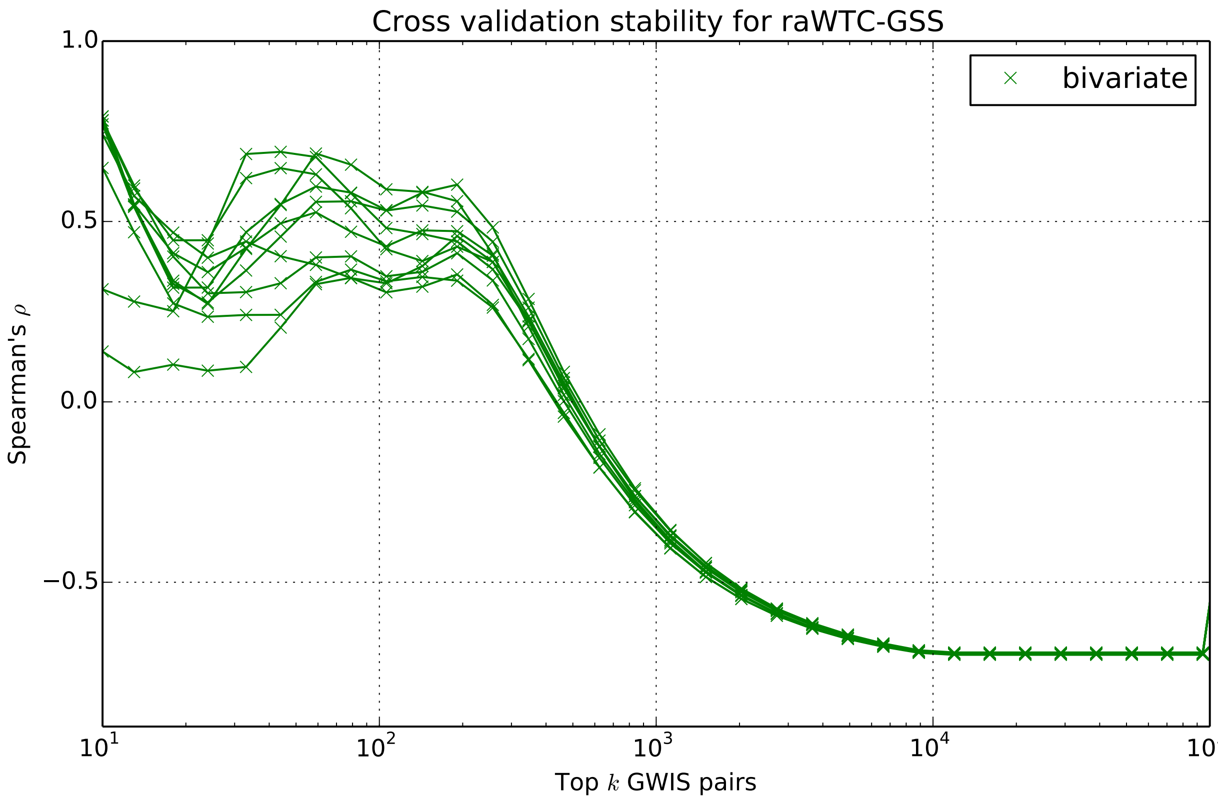
Signal and Noise



Spearman's ρ



Simulate two cohorts by splitting



Bedő, Rawlinson, Goudey, Ong, PLoS ONE, 2014

Measuring Overlap

Motivation

Given a set of replicated experiments, how do we measure overlap?

Challenges

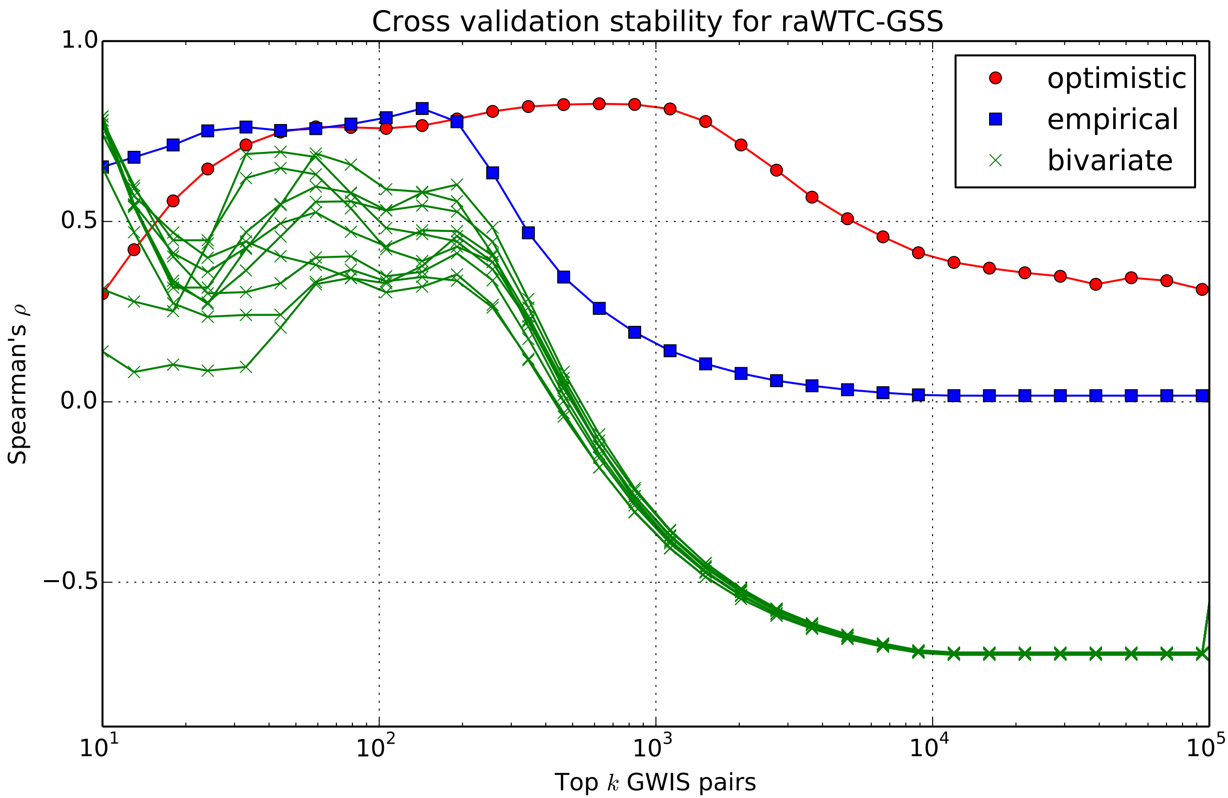
- Scores available for only the top-k examples
- Scores from different sources not calibrated

Model

- Ranked list Instead of just using set intersection, we can use the scores from GWIS to order the results
- top k Traditional methods (Spearman's ρ) requires ranks for the whole list. We have incomplete information, but we know our ranks are the top ones.
- Multivariate Textbook Spearman's ρ is for computing correlation between two ranks. We want to compute the correlation between multiple ranked lists.

  Bedő, Ong, JMLR (to appear)

Multiple replicates



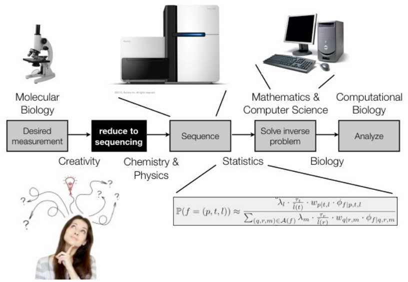
*-Seq

- dsRNA-Seq
- FRAG-Seq
- SHAPE-Seq
- PARTE-Seq
- PARS-Seq
- DMS-Seq
- ⋮

- Nucleo-Seq
- DNAse-Seq
- Sono-Seq
- ChIA-PET-Seq
- FAIRE-Seq
- NOMe-Seq
- ATAC-Seq
- ⋮

- GRO-Seq
- Quartz-Seq
- CAGE-Seq
- Nascent-Seq
- Cel-Seq
- 3P-Seq
- ⋮

https://liorpachter.wordpress.com/seq/

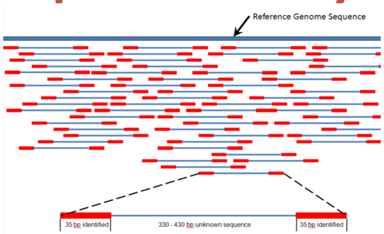


Integrating different sources of data

Association Study

	A	B	C	D	E	F	G	H	I	J	K	L
ID3023												
ID4454												
ID7675												
ID2283												

Sequence Analysis



Variation

- SNP
- Structural
- Methylation
- Expression
- …

Rank aggregation

Modeling using Spearman's correlation

Stability of feature selection

How to measure overlap?

$$\rho(R_1, \ldots, R_d)$$

Rank aggregation

How to combine different sources of information?

Macintyre, Yepes, Ong, Verspoor, PeerJ, 2014

Optimal aggregator: geometric mean

How to combine different sources of information?

We maximise multivariate correlation

$$R^* = \arg\max_R \rho(R, R_1, R_2, \ldots, R_d).$$

Theorem The aggregator that maximises multivariate Spearman's correlation is the product of the normalised ranks.

Use the geometric mean

NOT pairwise correlation

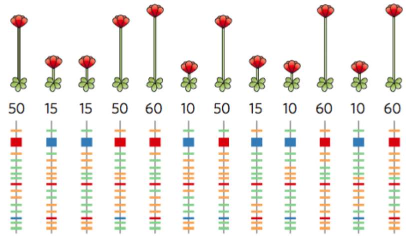
Instead of decomposing the association into a combination of pairwise similarities $\rho(R, R_1), \rho(R, R_2), \ldots, \rho(R, R_d)$.

Learning weighting of experts

We can also do supervised learning to rank

Bedő, Ong, JMLR (to appear)

What are good biomarkers?



Genome Wide Association Studies

- Which mutations are associated with tall poppies?
- Identify biomarkers with hypothesis tests

Finding stable biomarkers

- Split cohort into two (cross validation)
- Investigate rank correlation between scores

Integrating information via ranks

- Multivariate Spearman correlation using copulas
- Geometric mean is the optimal aggregator

What to measure?

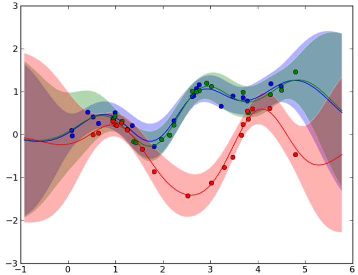
$$f_{\mathbf{w}}(x) : \mathcal{X} \rightarrow \mathcal{Y}$$

Active Learning / Expt. Design

Use predictor to identify good candidates

- Annotate top-k items
- Confidence interval improves performance
- Explore - exploit tradeoff

Krause, Ong, NIPS 2011



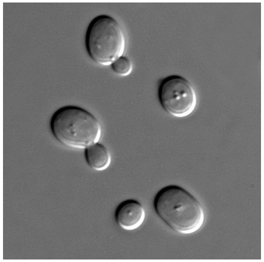
Finding black holes and redshifts

- Machine learning to classify images
- Show 10 candidates to expert daily

Collaboration with ANU, ANTF, CAASTRO

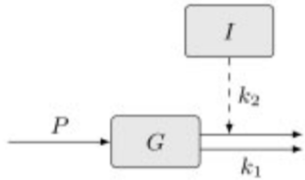
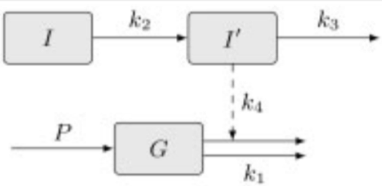
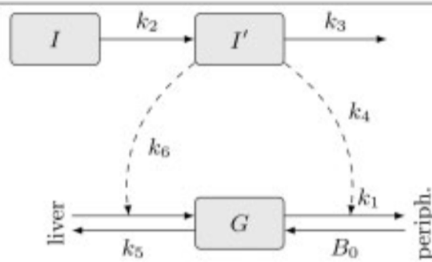
Glucose metabolism in Yeast

- Multiple possible models
- Design biological experiments that maximise information gain



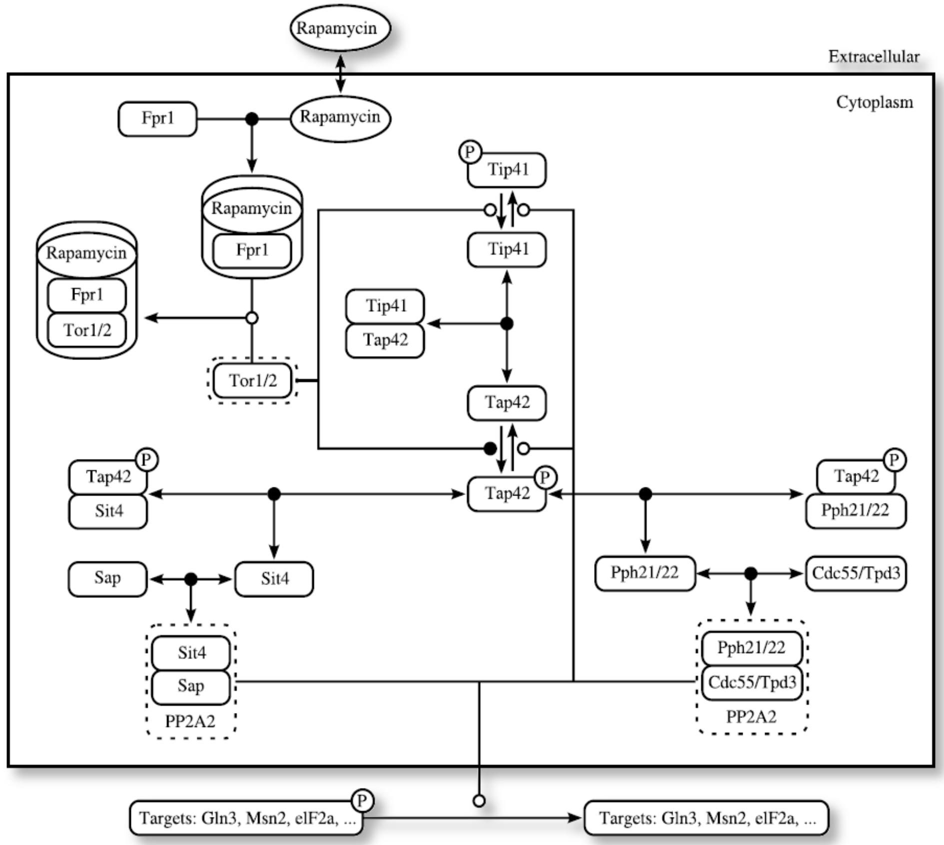
Collaboration of ETHZ with SystemsX Switzerland

What is a model?

No.	Model	ODE and parameters
I		$\dot{G} = \theta_1 G + \theta_2$ $\theta_1 = -k_1 \quad = (-1.09 \pm 0.11) \cdot 10^{-1}$ $\theta_2 = P \quad = 8.5 \pm 1.10$
IV		$\dot{G} = \theta_1 G + \theta_2 I + \theta_3$ $\theta_1 = -k_1 \quad = (-1.44 \pm 0.35) \cdot 10^{-1}$ $\theta_2 = -k_2 \quad = (9.15 \pm 4.0) \cdot 10^{-2}$ $\theta_3 = P \quad = 11.3 \pm 3.1$
V		$\dot{G} = \theta_1 G + \theta_2 X + \theta_3$ $\dot{X} = \theta_4 X + I$ $\theta_1 = -k_1 \quad = (6.50 \pm 0.73) \cdot 10^{-1}$ $\theta_2 = -k_2 k_4 \quad = (-9.10 \pm 1.73) \cdot 10^{-3}$ $\theta_3 = P \quad = 5.97 \pm 0.70$ $\theta_4 = -k_3 \quad = (-1.01 \pm 0.16) \cdot 10^{-1}$
VI		$\dot{G} = (\theta_1 - X)G + \theta_4$ $\dot{X} = \theta_2 X + \theta_3 I$ $X = I'/k_2$ $\theta_1 = -(k_1 + k_5) = (-4.90 \pm 0.97) \cdot 10^{-2}$ $\theta_2 = -k_3 \quad = (-9.10 \pm 1.20) \cdot 10^{-2}$ $\theta_3 = k_2(k_{4+5}) \quad = (8.96 \pm 1.88) \cdot 10^{-5}$ $\theta_4 = B_0 \quad = 4.42 \pm 0.74$

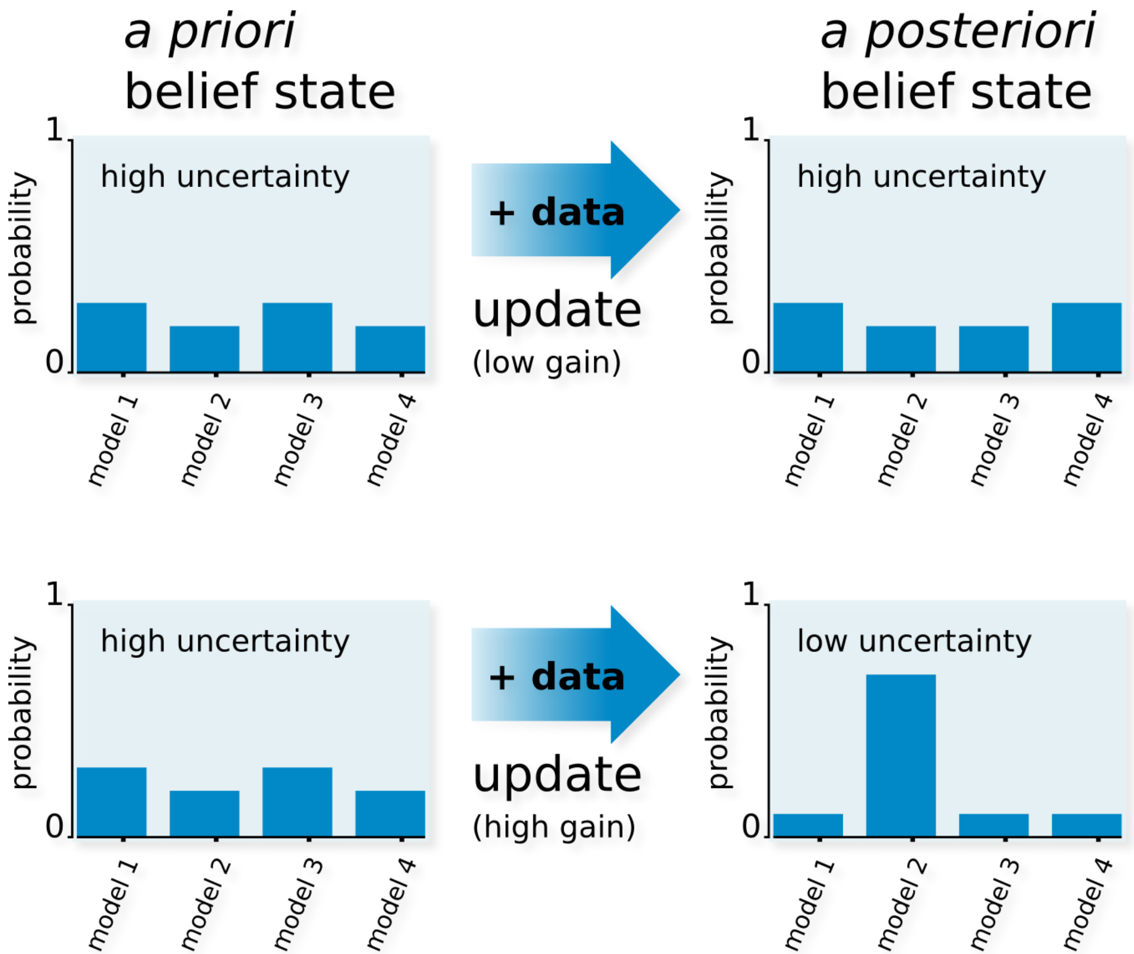
Bergman insulin dependent glucose metabolism model.

TOR pathway



Rapamycin
Extracellular
Cytoplasm
Fpr1
Rapamycin
P
Tip41
Rapamycin
Fpr1
Rapamycin
Fpr1
Tor1/2
Tip41
Tip41
Tap42
Tor1/2
Tap42
Tap42
P
Sit4
Tap42
P
Tap42
P
Pph21/22
Sap
Sit4
Pph21/22
Cdc55/Tpd3
Sit4
Sap
PP2A2
Pph21/22
Cdc55/Tpd3
PP2A2
Targets: Gln3, Msn2, elF2a, ...
P
Targets: Gln3, Msn2, elF2a, ...

Finding good models



Optimised experimental design

Measurements

Experiments produce readouts $y(t_i)$,
grouped into datasets Y_π for an experiment π.

Bayes rule

For a particular model f, (taking care of parameters)

$$p(f|Y_\pi) = \frac{p(Y_\pi|f)p(f)}{p(Y_\pi)}$$

Information gain

We want to take measurements that change model probabilities

$$D_{KL}[p(f|Y_\pi)||p(f)] = \sum_{f \in \mathcal{F}} p(f|Y_\pi) \log_2 p(f|Y_\pi)/p(f)$$

Marginalise over possible outcomes

Maximise expected information gain (tough computational problem)

$$\operatorname*{argmax}_{\pi} \mathbb{E}_{Y_\pi} D_{KL}[p(f|Y_\pi)||p(f)]$$

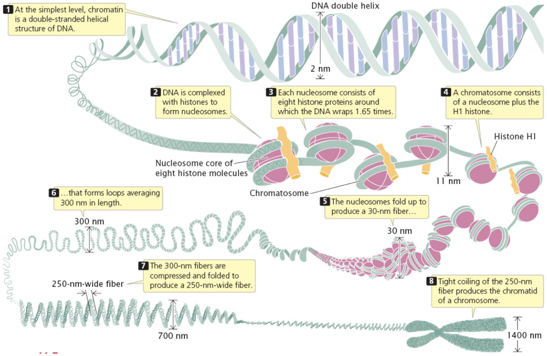
Experiments, experiments, ...

What is a biomarker?

How to measure?

Use adaptive experimental design to identify important time series.

Busetto et. al. Near-optimal experimental design for model selection in systems biology , 2013



What to measure?

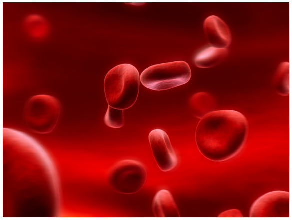
Combine various sources of information for robust decision making.

Macintyre et. al. Associating disease-related genetic variants in intergenic regions to the genes they impact, 2014

Where to measure?

Use expert domain knowledge to construct dynamical models.

Brodersen et. al. Generative embedding for model-based classification of fMRI data, 201 1



A more philosophical section...

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

Label: Finding black holes

- Exist physical models, we directly use images
- There is relatively large amounts of data (examples)
- Object localisation with crowd labels

Feature: Finding genetic associations

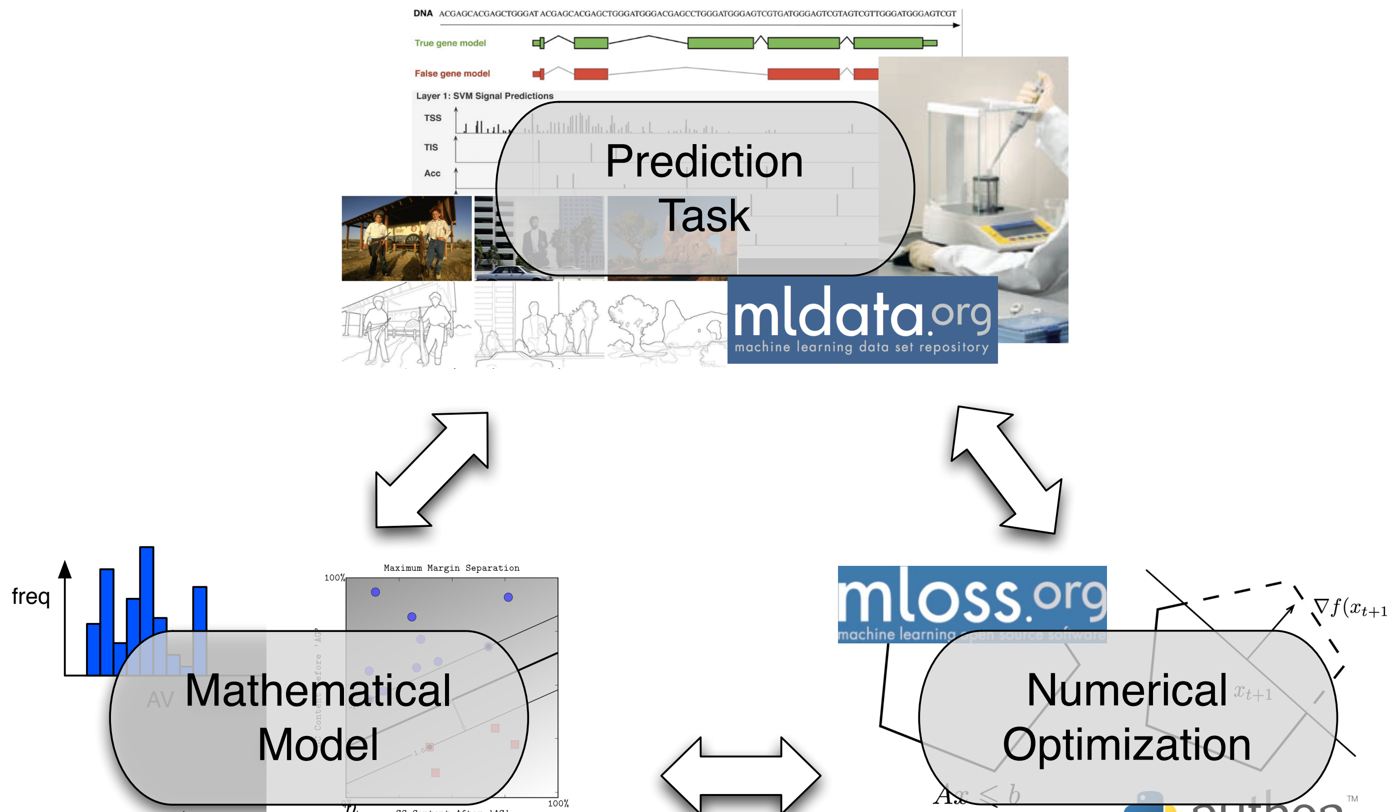
- No mechanistic model of the phenomenon
- High dimensional low sample size
- Stability of feature selection

Predictor: Finding good experiments

- Partial mechanistic model of the phenomenon
- Estimate the expected information gain

Discuss challenges to applying machine learning

Applications - Optimization - Models



$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t. } y_i(\langle w, x_i\rangle + b) \geqslant 1 - \xi_i$$

Scoring candidates - ABCDE

Active Learning

- Choose a particular example to label using heuristics
- Annotator assumed to provide ground truth

Bandits

- Select a choice from a set of actions
- Simple algorithms with theoretical guarantees
- Manage uncertainty with repeated sampling

Choice theory

- Aggregate set of ranks into one ordering
- Economics and social science, impossiblity theorems

Designing Experiments

- Choose a set of trials to measure
- Optimisation algorithms with theoretical analysis
- Information theory, real random variables

ML Open Source Software

Wider adoption of methods

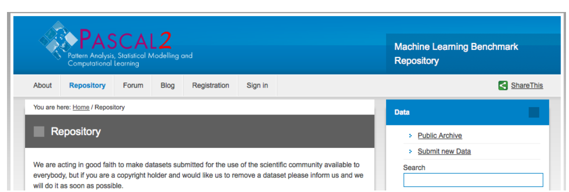
- Domain experts can use machine learning core
- Available for teaching

Scientific reproducibility

- Fair comparison of methods
- Access to scientific tools

Community growth

- “Given enough eyeballs, all bugs are shallow”
- Combination of advances



mloss.org mldata.org

Plug and Pray

Machine Learning Open Source Software

Do We Need Hundreds of Classifiers
to Solve Real World Classification Problems?

jmlr.org/papers/v15/delgado14a.html

Spoiler: No

Usability and Reproducibility

- (too much) focus on new algorithms
- Documentation, modularity issues
- Literate programming
  yihui.name/knitr jupyter.org
- Scientific computing workflows
  galaxyproject.org

Dream: App Bazaar for data science

Bumpy road to data science

Two classes of objects

Data

images, counts, raw sensor data, output of simulation, results

Analysis

visualisation, user interface, predictors, observational statistics

Multi-sided platform

- Decentralised architecture, not walled garden
- Enable direct interaction between data owner and analytics system
- Network effect: each new entrant benefits from whole network

Not just tech people

Domain experts, data managers, project management

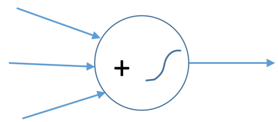
Wish list

We need an open federated framework for scientific discovery

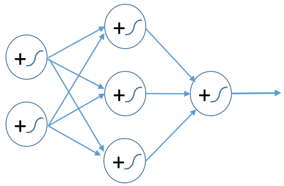
- Provenance, trust and reliability
- Management of legal rights
- Uncertainty propagation
- Confidentiality and privacy
- Complex workflows
- Late binding ontologies
- Cross organisation, jurisdiction, technical boundaries
- Decouple technique from problem
- No proprietary control
- *-as-a-service

One more challenge

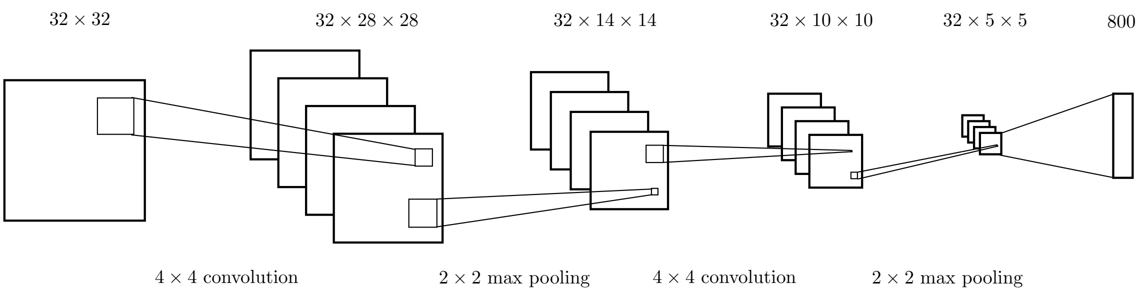
McCulloch and Pitts, 1943



Multilayer perceptron



Deep neural networks



One more challenge

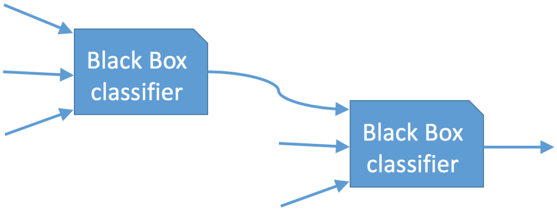
McCulloch and Pitts, 1943

Multilayer perceptron

Deep neural networks

Today's ML systems

How to analyse two systems?



Conclusion

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

Three extensions

- Not standard binary classification $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$
- What are good features? $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$
- What to measure? $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$

Plug and pray

- Software, software, software
- Build the road and rail for data science
- Understand combinations of machine learning components

Thank You

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

Three extensions

- Not standard binary classification $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$
- What are good features? $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$
- What to measure? $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$

Plug and pray

- Software, software, software
- Build the road and rail for data science
- Understand combinations of machine learning components

Please make your research open

www.ong-home.my

Copulas

Intuition

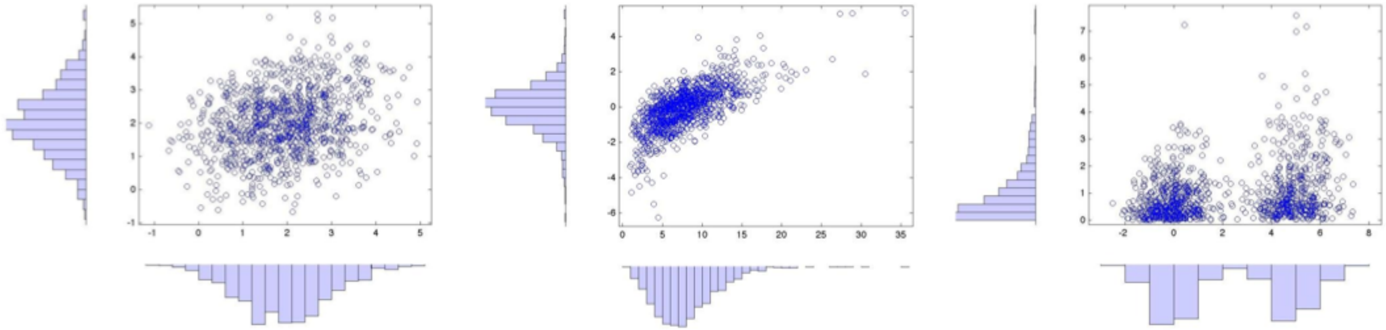
For continuous random variables, copulas model the dependence component after discounting for univariate marginal effects

Probabilistic definition

Let $U_1, \ldots, U_d$ be real random variables $\sim U([0,1])$.
A copula function $C : [0,1]^d \longrightarrow [0,1]$ is a joint distribution

$$C_\theta(u_1, \ldots, u_d) = P(U_1 \leqslant u_1, \ldots, U_d \leqslant u_d)$$

The same Gaussian copula function



Copulas and Spearman's ρ

Spearman's ρ can be expressed in terms of the copula

$$\rho(A,B) = 12 \int_{[0,1]^2} C(u,v)dudv - 3$$

Empirical copula

$$C_n(u,v) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \mathbf{1}\left(R(x) \leqslant u, S(x) \leqslant v\right)$$

Why do the math?

- Unclear how to extend formula for Spearman's correlation.
- Multivariate distributions $\Rightarrow$ multivariate copula.

Multivariate Spearman's ρ

A multivariate extension of Spearman's ρ

For a d dimensional set of random variables $\mathbf{u}$, the multivariate Spearman's ρ is given by

$$\rho(R_1, \ldots, R_d) = Q(C, \pi) = h(d)\left(2^d \int_{[0,1]^d} \pi(\mathbf{u})\, \mathrm{d}C(\mathbf{u}) - 1\right),$$

where

$$h(d) = \frac{d+1}{2^d - (d+1)}.$$

Empirical multivariate Spearman's corelation

$$\rho_n(R_1, \ldots, R_d) = h(d)\left[\frac{2^d}{n}\sum_x \prod_{j=1}^{d} R_j(x) - 1\right].$$

No negative correlation

As the number of dimensions increases, the lower bound of Spearman's ρ tends to zero