Towards Computational Systems Biology with a Statistical Analysis Pipeline for Microarray Data

Neil Lawrence and Magnus Rattray
School of Computer Science
University of Manchester

October 31, 2007

Outline

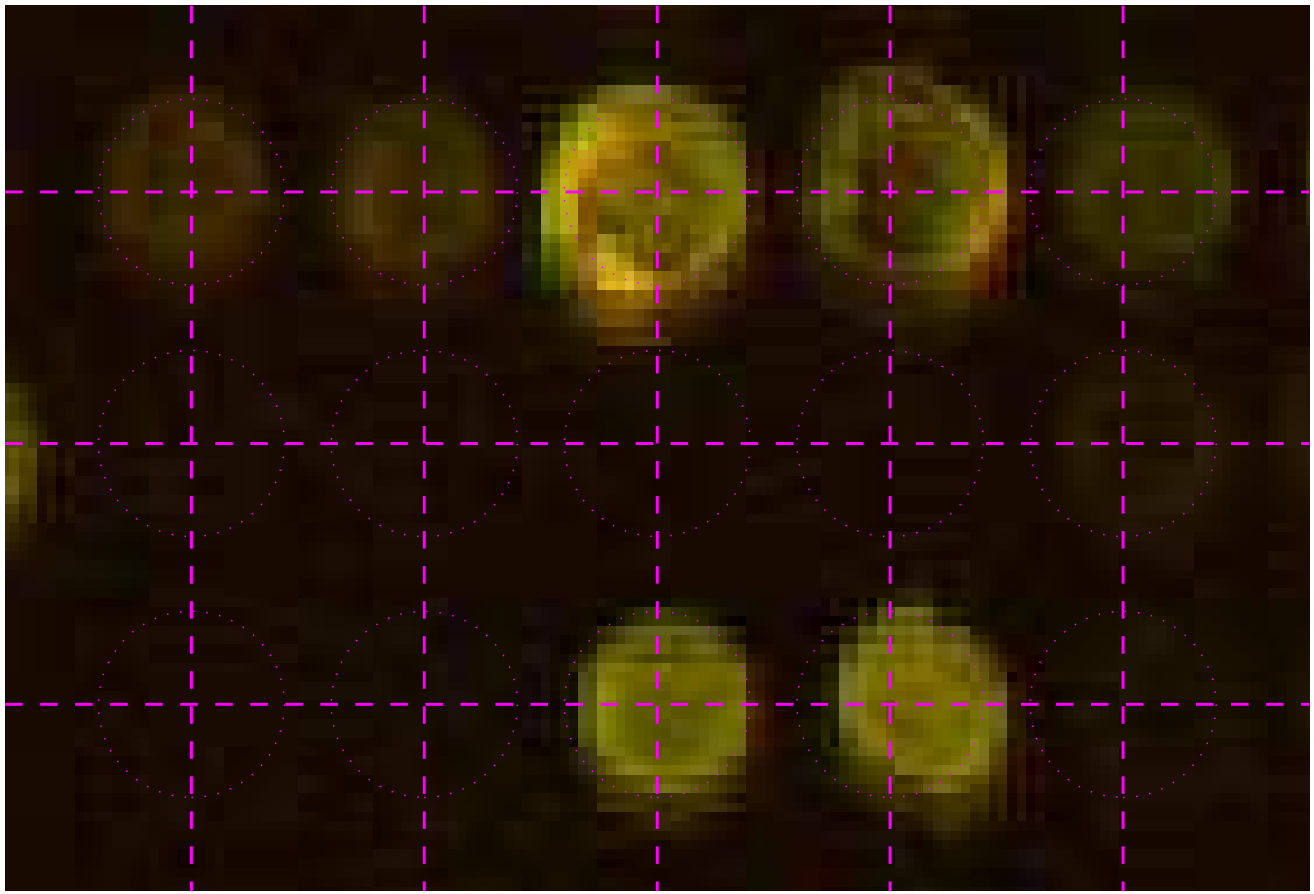
Online Resources

All source code and slides are available online

- This talk available from my home page (see talks link on side).
- PUMA Project main page (with links to software)
  - `http://bioinf.man.ac.uk/resources/puma/`.
- Additional project homepages
  - `http://www.cs.man.ac.uk/~neill/projects/pipeline/`.
  - `http://www.cs.man.ac.uk/~neill/projects/tigra/`

PUMA Project Outline

- Noisy Data $\rightarrow$ Useful inference
- My first contact with microarray: 2001 cDNA arrays with Niranjan, Pen Rashbass.



PUMA Project Outline

- Early work from Niranjan was about classification (e.g. tumour types) and feature selection (which genes are important in differentiating)
- Aim to build on this work with more complex modelling.
  - It was apparent that we needed to handle the noise in this data!
- Today our work focuses on determining the influence of *latent chemical species* on the data.
- We've developed techniques for both *genome wide* inference and *transcription factor specific* inference.
- All these techniques are embedded in a probabilistically rigorous handling of the data.

PUMA Project Outline

- Early work from Niranjan was about classification (e.g. tumour types) and feature selection (which genes are important in differentiating)
- Aim to build on this work with more complex modelling.
  - It was apparent that we needed to handle the noise in this data!
- Today our work focuses on determining the influence of *latent chemical species* on the data.
- We've developed techniques for both *genome wide* inference and *transcription factor specific* inference.
- All these techniques are embedded in a probabilistically rigorous handling of the data.

PUMA Project Outline

- Early work from Niranjan was about classification (e.g. tumour types) and feature selection (which genes are important in differentiating)
- Aim to build on this work with more complex modelling.
  - It was apparent that we needed to handle the noise in this data!
- Today our work focuses on determining the influence of *latent chemical species* on the data.
- We've developed techniques for both *genome wide* inference and *transcription factor specific* inference.
- All these techniques are embedded in a probabilistically rigorous handling of the data.

PUMA Project Outline

- Early work from Niranjan was about classification (e.g. tumour types) and feature selection (which genes are important in differentiating)
- Aim to build on this work with more complex modelling.
  - It was apparent that we needed to handle the noise in this data!
- Today our work focuses on determining the influence of *latent chemical species* on the data.
- We've developed techniques for both *genome wide* inference and *transcription factor specific* inference.
- All these techniques are embedded in a probabilistically rigorous handling of the data.

PUMA Project Outline

- Early work from Niranjan was about classification (e.g. tumour types) and feature selection (which genes are important in differentiating)
- Aim to build on this work with more complex modelling.
  - It was apparent that we needed to handle the noise in this data!
- Today our work focuses on determining the influence of *latent chemical species* on the data.
- We've developed techniques for both *genome wide* inference and *transcription factor specific* inference.
- All these techniques are embedded in a probabilistically rigorous handling of the data.

Future Perspectives

- Our work so far has focussed on *data driven* models.
- Examples include
  - Differential expression analysis
  - Hierarchical Clustering
  - Principal Component Analysis
- An alternative perspective is mechanistic models.
  - For example: differential equation models.
- There is always a balance between a *realistic* model and a *tractable* model.

Future Perspectives

- Our work so far has focussed on *data driven* models.
- Examples include
  - Differential expression analysis
  - Hierarchical Clustering
  - Principal Component Analysis
- An alternative perspective is mechanistic models.
  - For example: differential equation models.
- There is always a balance between a *realistic* model and a *tractable* model.

Future Perspectives

- Our work so far has focussed on *data driven* models.
- Examples include
  - Differential expression analysis
  - Hierarchical Clustering
  - Principal Component Analysis
- An alternative perspective is mechanistic models.
  - For example: differential equation models.
- There is always a balance between a *realistic* model and a *tractable* model.

Future Perspectives

- Our work so far has focussed on *data driven* models.
- Examples include
  - Differential expression analysis
  - Hierarchical Clustering
  - Principal Component Analysis
- An alternative perspective is mechanistic models.
  - For example: differential equation models.
- There is always a balance between a *realistic* model and a *tractable* model.

Future Perspectives

- Our work so far has focussed on *data driven* models.
- Examples include
  - Differential expression analysis
  - Hierarchical Clustering
  - Principal Component Analysis
- An alternative perspective is mechanistic models.
  - For example: differential equation models.
- There is always a balance between a *realistic* model and a *tractable* model.

Future Perspectives

- Our work so far has focussed on *data driven* models.
- Examples include
  - Differential expression analysis
  - Hierarchical Clustering
  - Principal Component Analysis
- An alternative perspective is mechanistic models.
  - For example: differential equation models.
- There is always a balance between a *realistic* model and a *tractable* model.

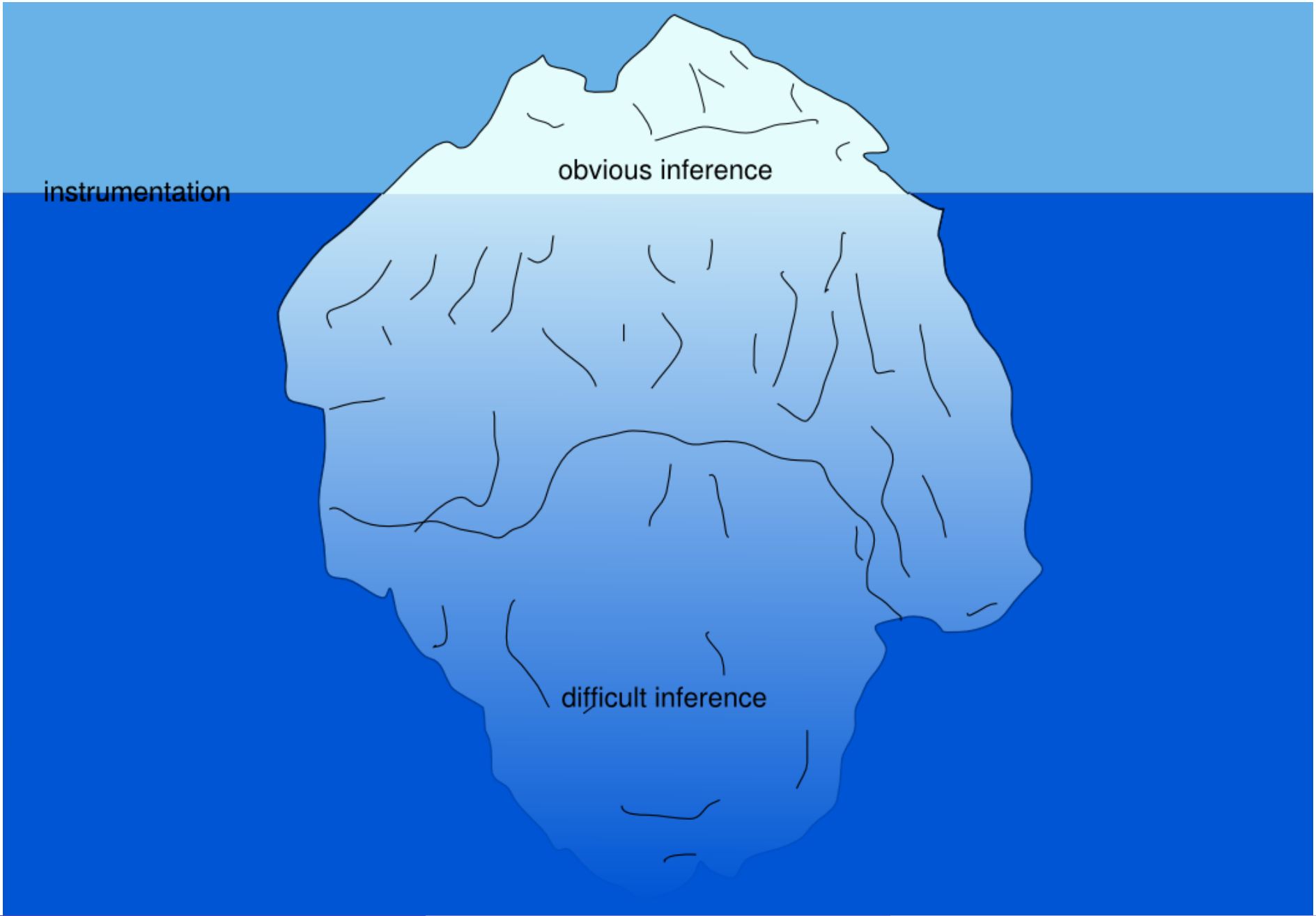
Future Perspectives

- Our work so far has focussed on *data driven* models.
- Examples include
  - Differential expression analysis
  - Hierarchical Clustering
  - Principal Component Analysis
- An alternative perspective is mechanistic models.
  - For example: differential equation models.
- There is always a balance between a *realistic* model and a *tractable* model.

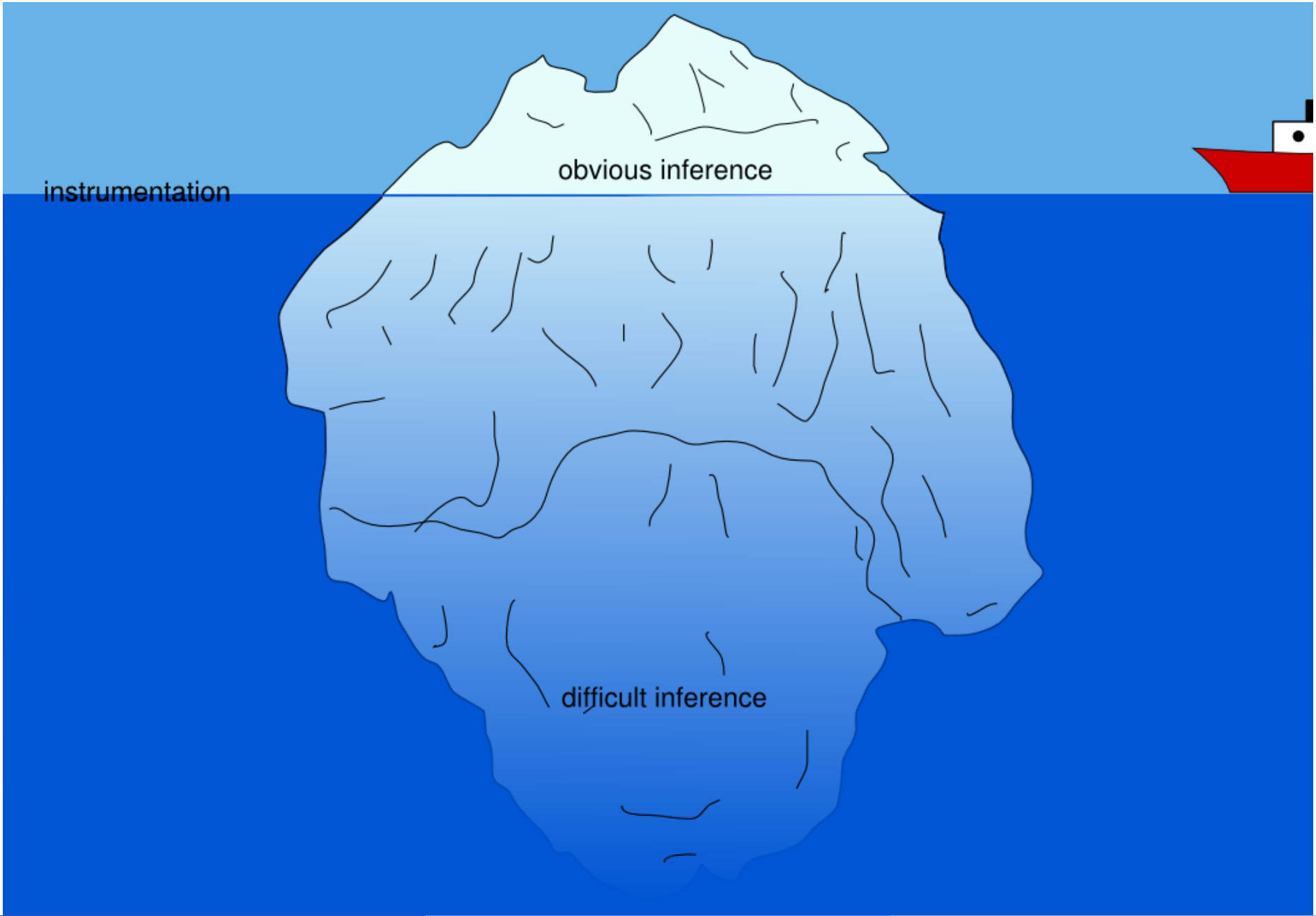
Future Perspectives

- Our work so far has focussed on *data driven* models.
- Examples include
  - Differential expression analysis
  - Hierarchical Clustering
  - Principal Component Analysis
- An alternative perspective is mechanistic models.
  - For example: differential equation models.
- There is always a balance between a *realistic* model and a *tractable* model.

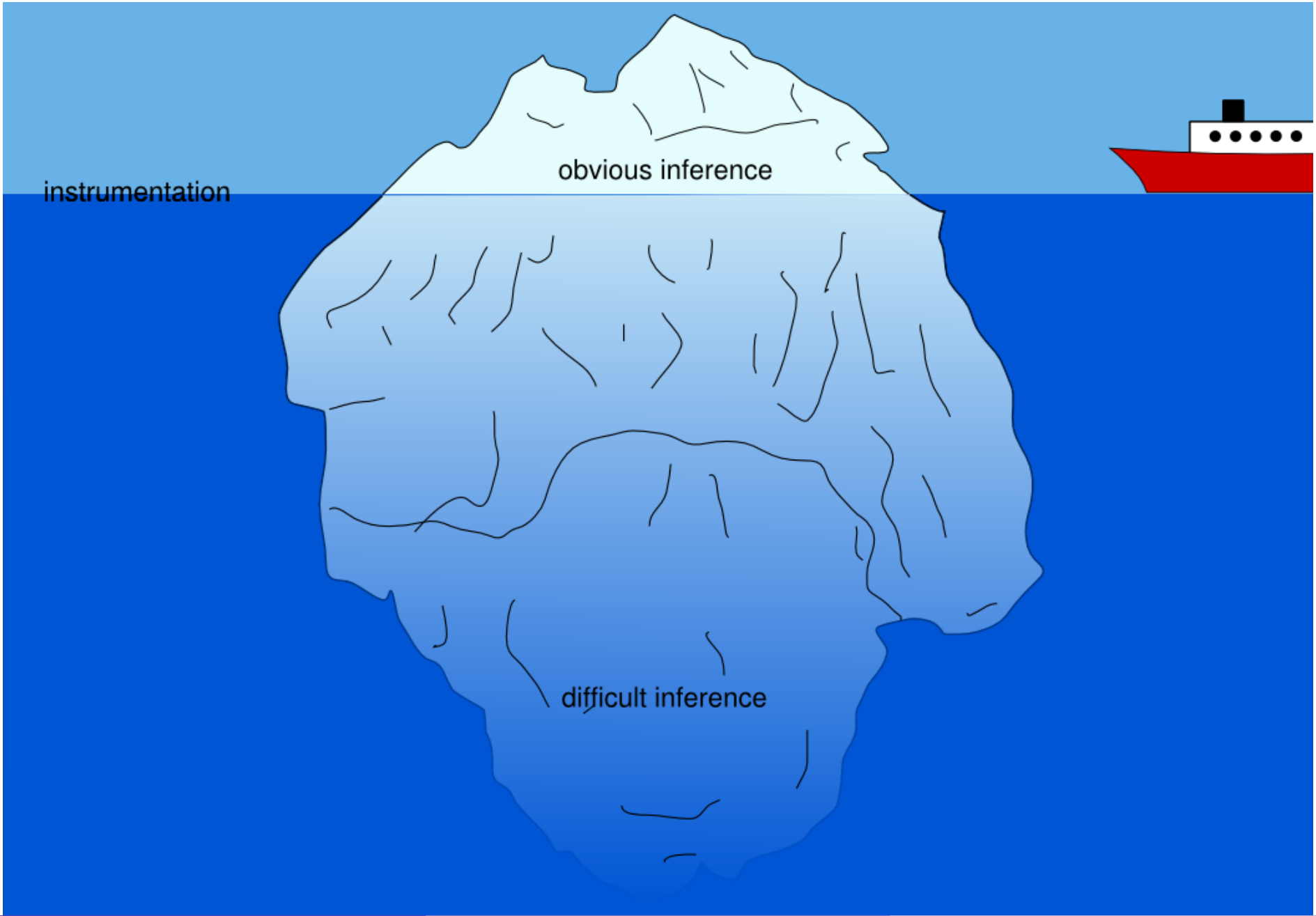
The Iceberg of Information



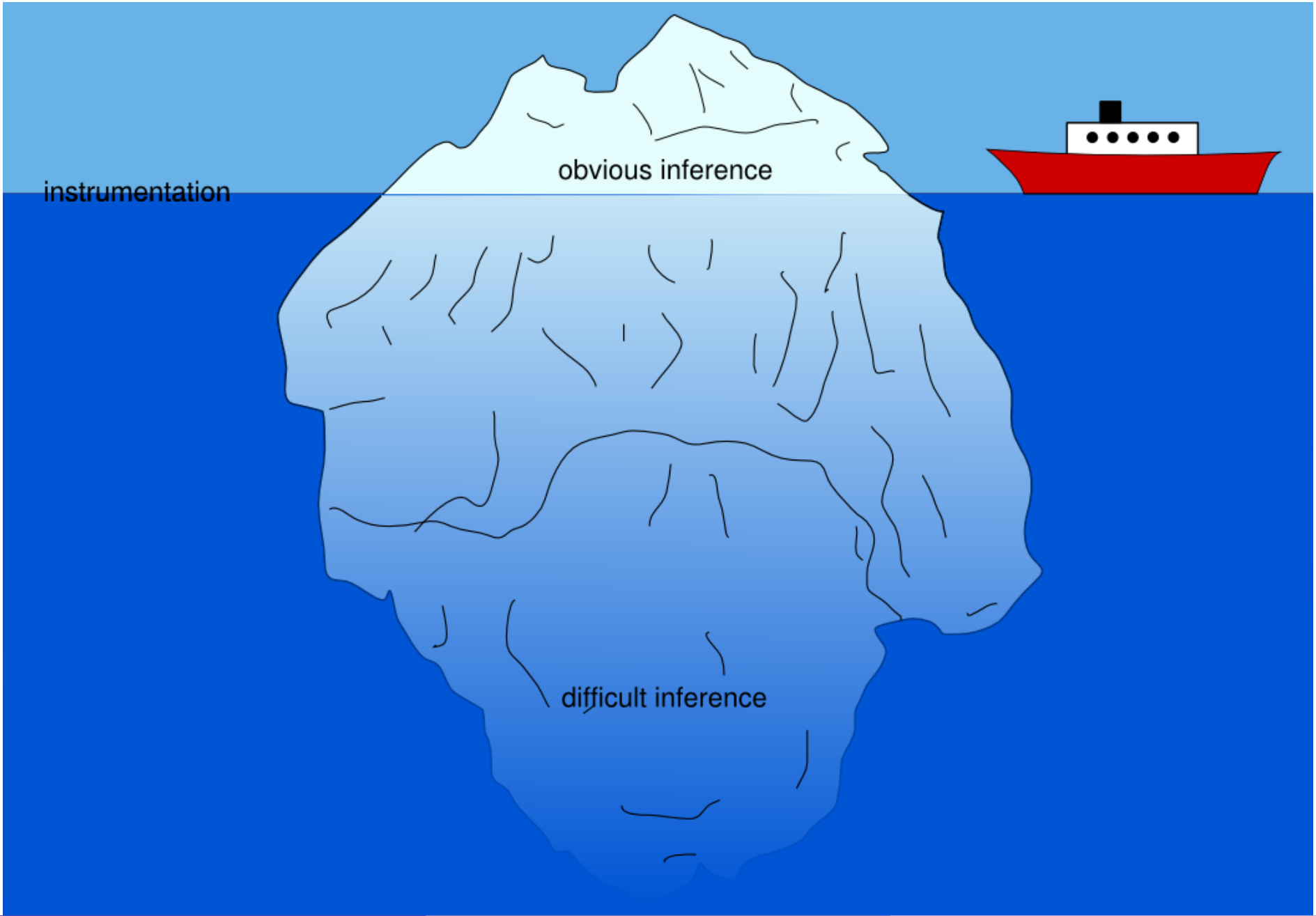
The Iceberg of Information



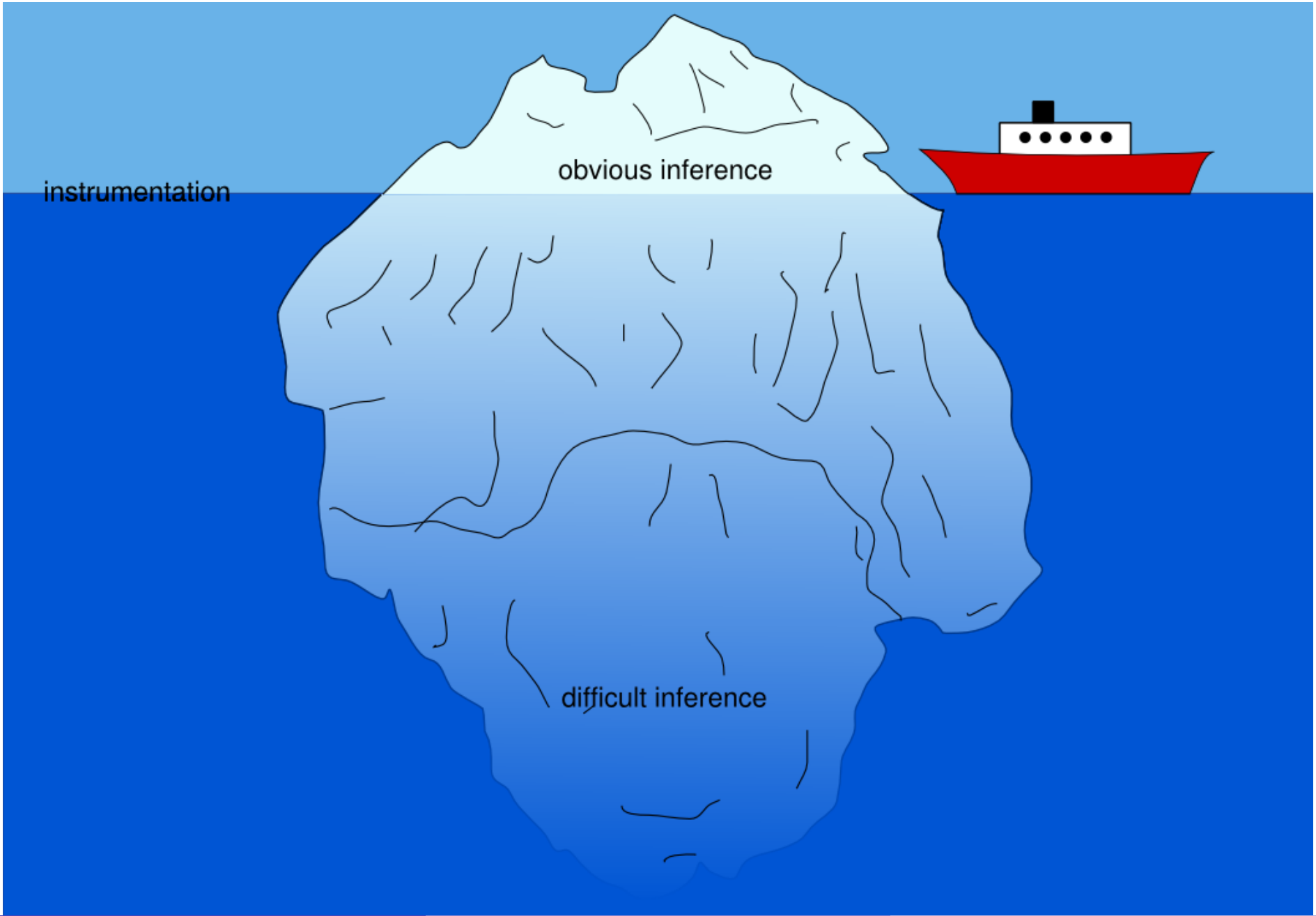
The Iceberg of Information



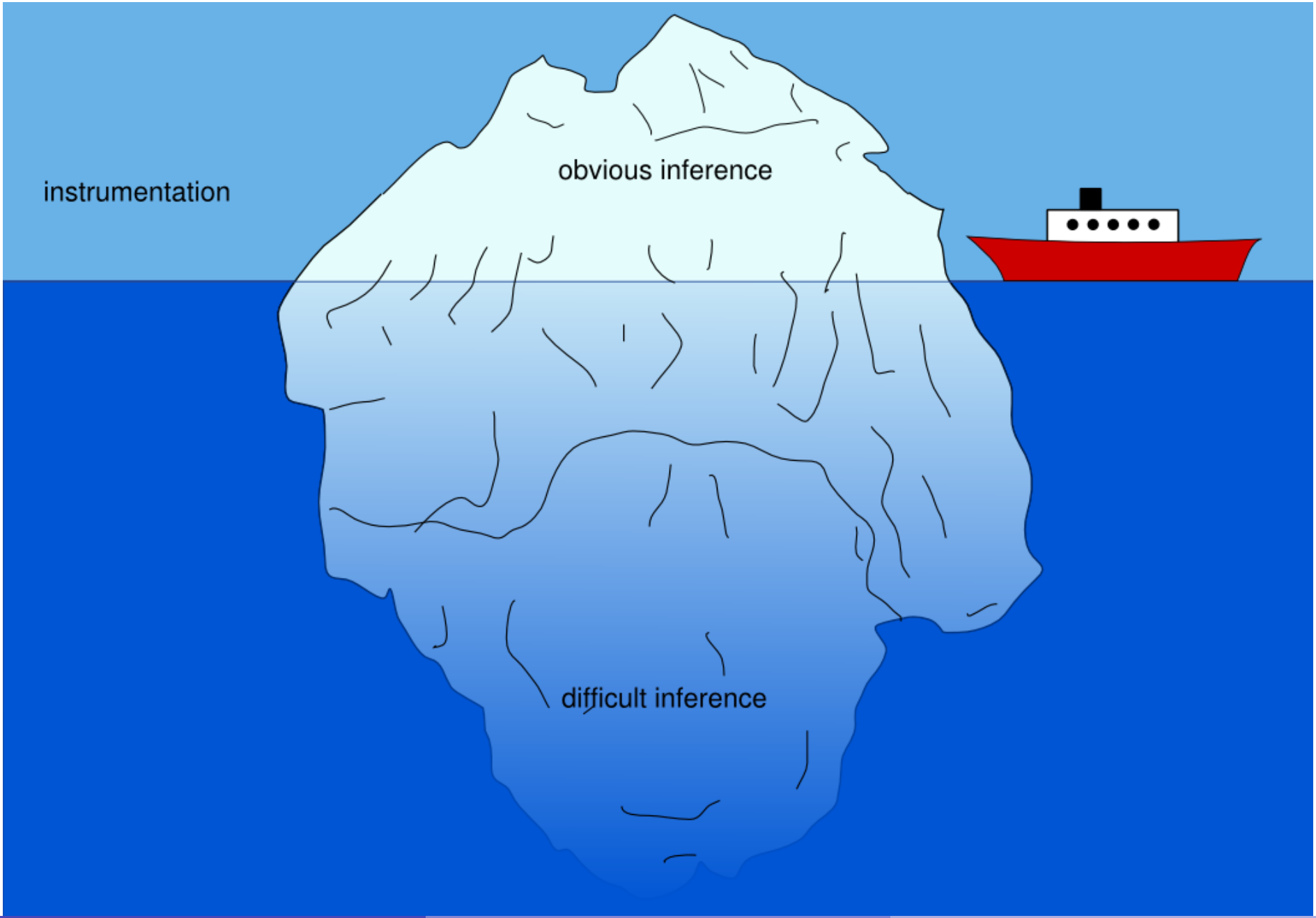
The Iceberg of Information



The Iceberg of Information



The Iceberg of Information



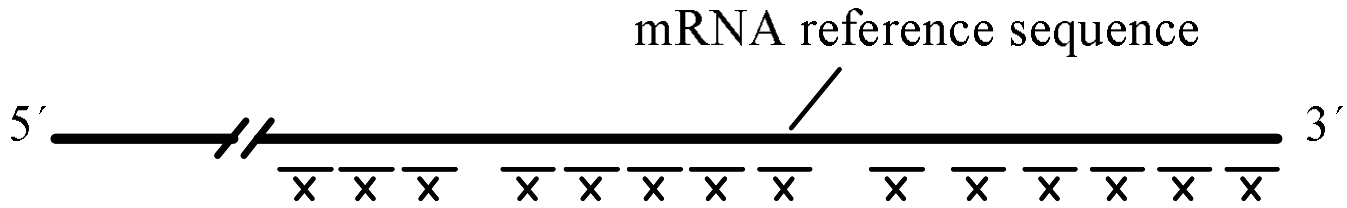
Affymetrix Arrays

- Working with Matthew Holley, our (Marta Milo, Niranjan and I) focus shifted to Affymetrix arrays.

Figure: Affymetrix arrays for human and mouse (image from Wikimedia Commons under GFDL).

- There are multiple probe pairs on an Affymetrix array; could we exploit this to estimate the noise?

Affymetrix Arrays



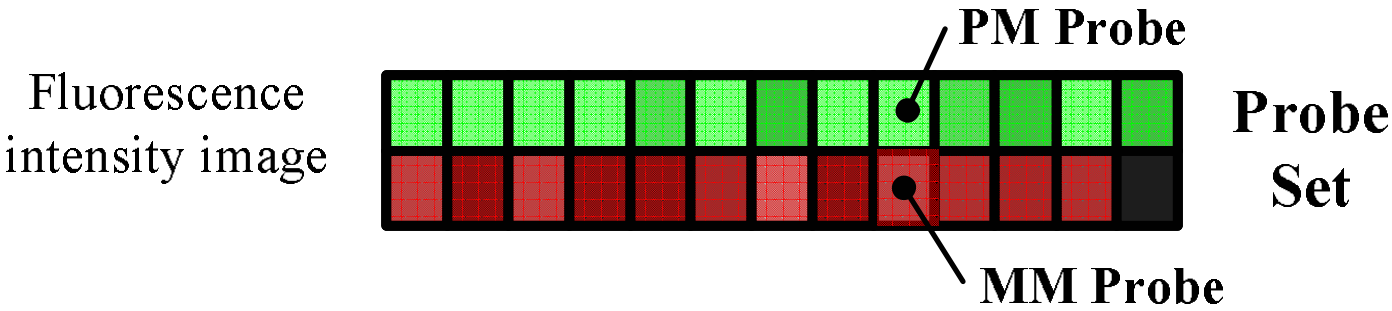


Figure: Affymetrix array schematic

Affymetrix Arrays

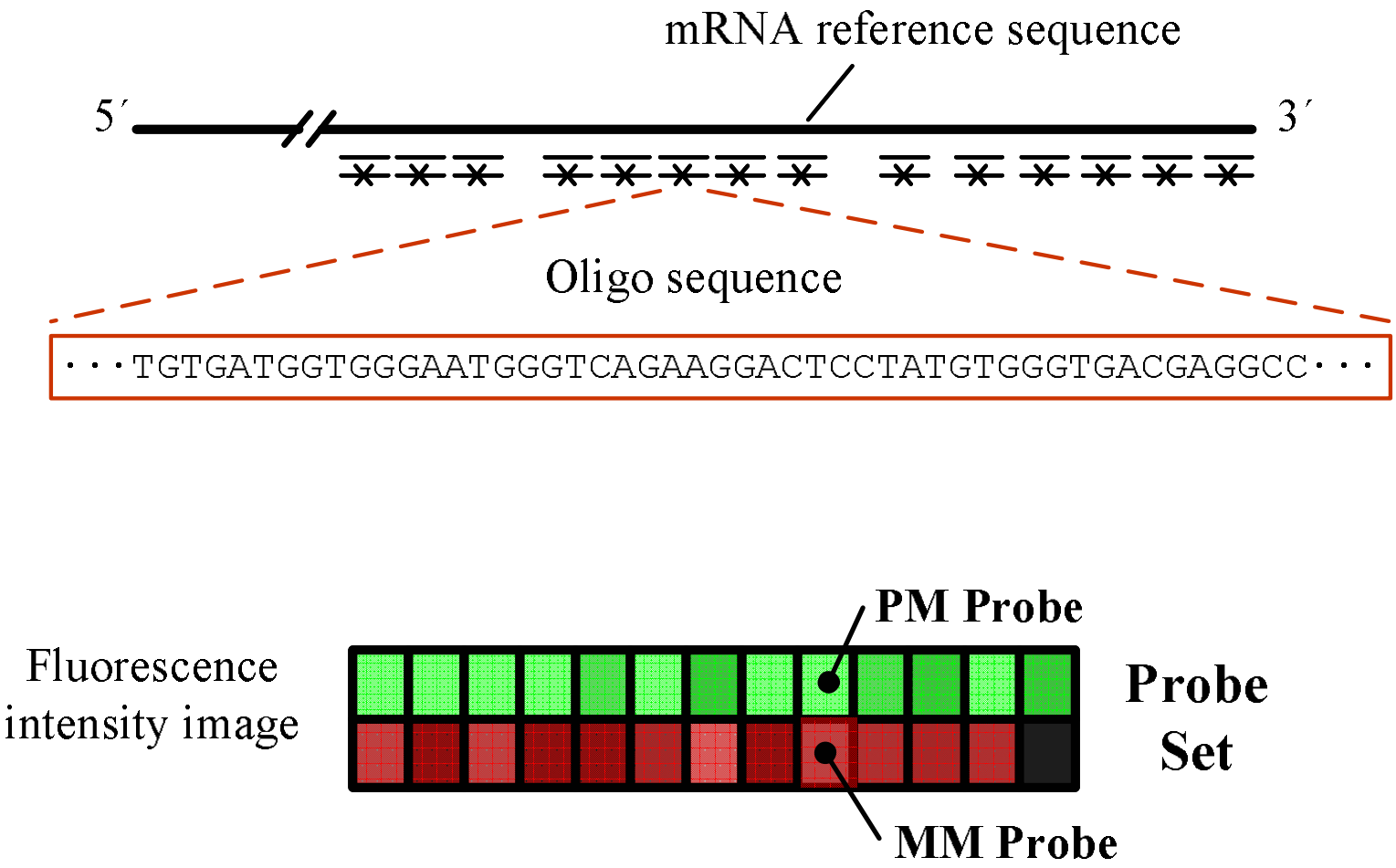


Figure: Affymetrix array schematic

Affymetrix Arrays

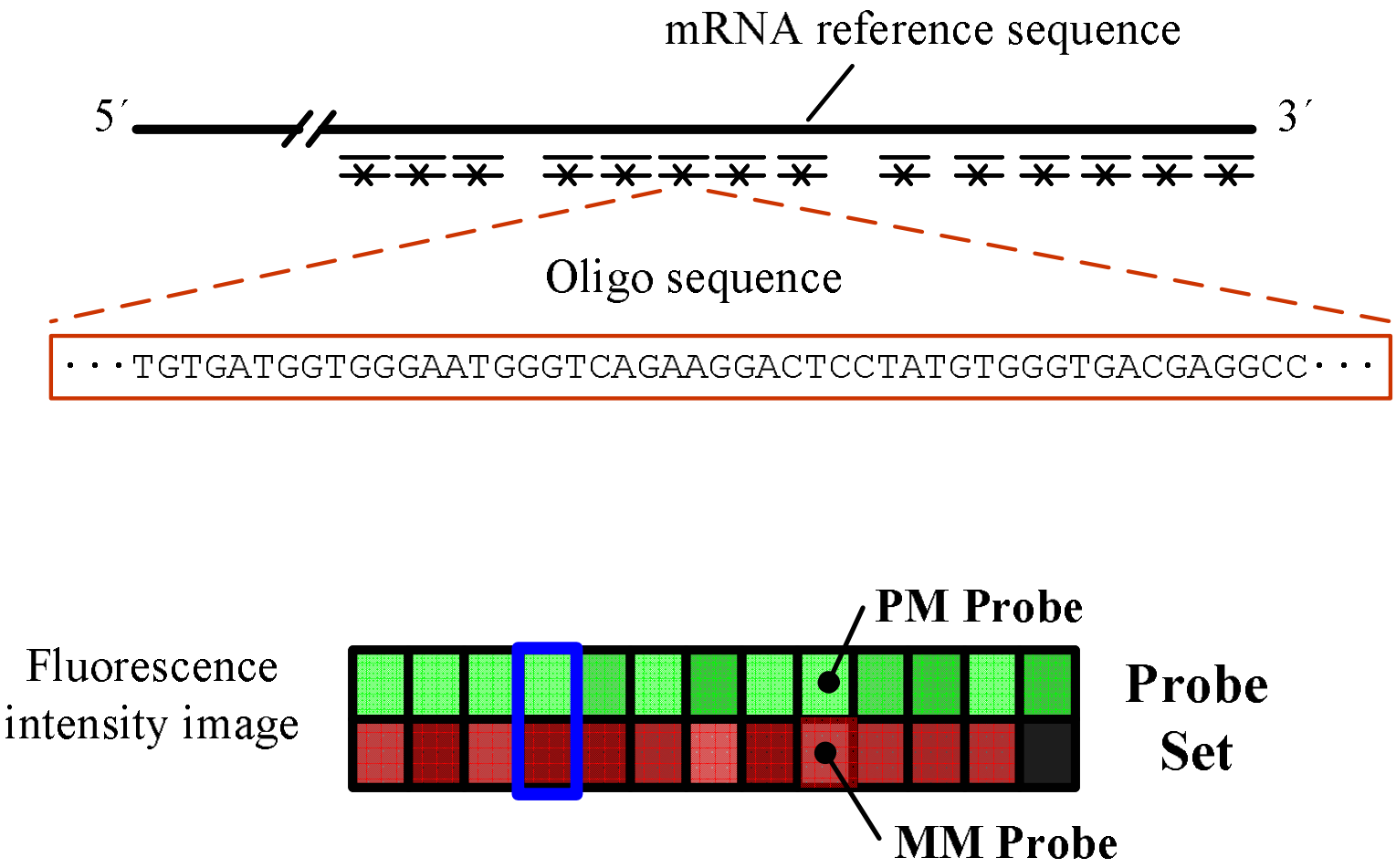


Figure: Affymetrix array schematic

Affymetrix Arrays

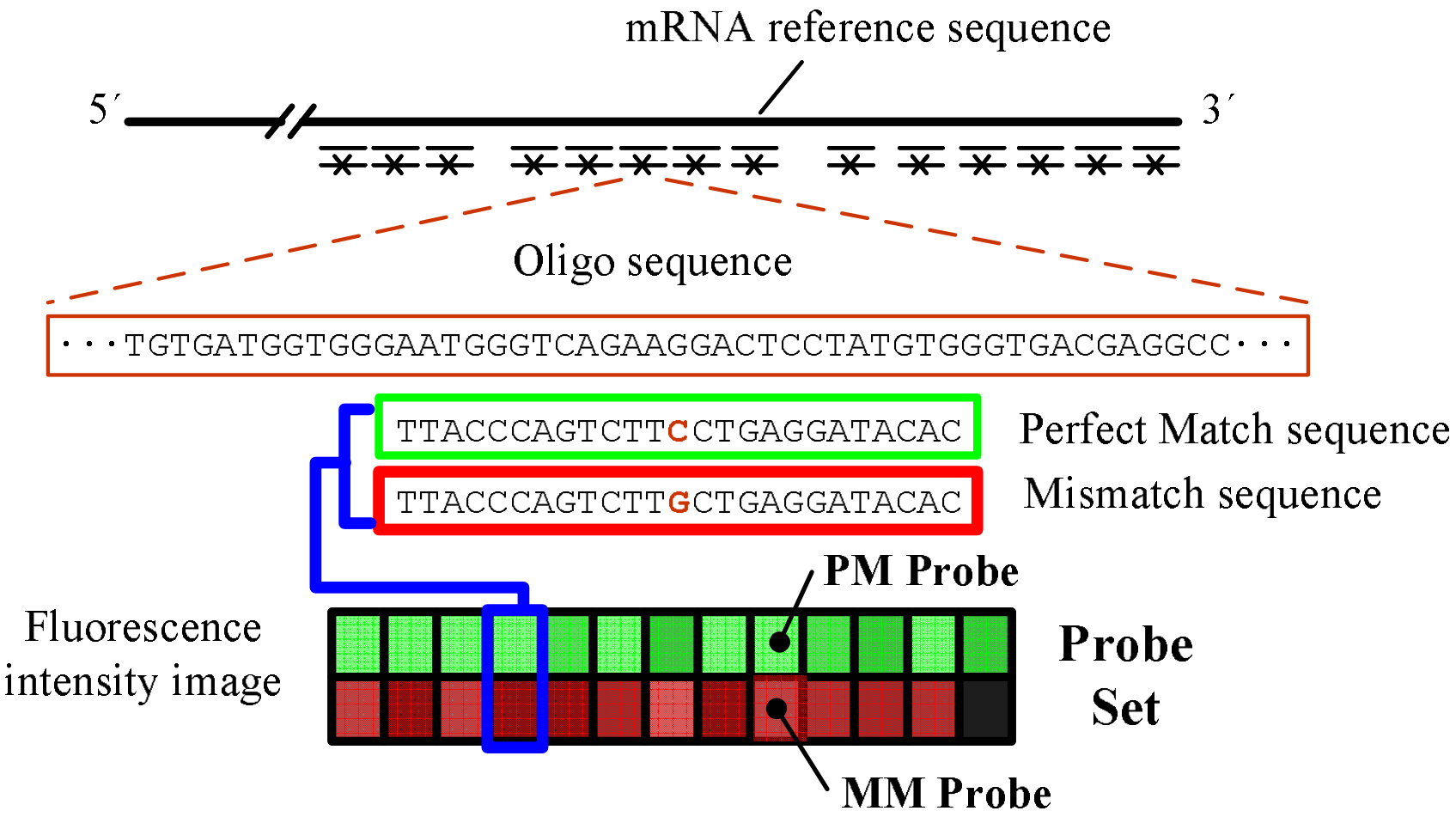


Figure: Affymetrix array schematic

gMOS Family of Methods

- gMOS — Gamma Model of Signal [Milo et al., 2003, Liu et al., 2005]
- Most methods return a single expression level estimate.
- The gMOS family of methods additionally provide *confidence intervals*.
- This confidence intervals can the be propagated through higher level analysis.

gMOS Family of Methods II

- Gamma Model of Signal

$$s_j \sim \text{Ga}(s_j|\alpha, b)$$

$$m_j \sim \text{Ga}(m_j|a, b)$$

$$y_j = m_j + s_j$$

$$y_j \sim \text{Ga}(y_j|a+\alpha, b)$$

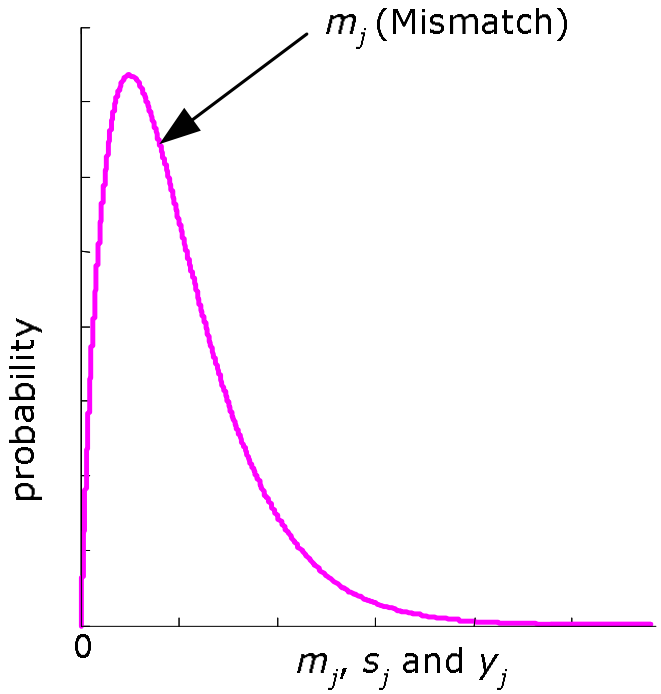


Figure: PDF of m_j, s_j and the implied distribution for y_j.

gMOS Family of Methods II

- Gamma Model of Signal

$$s_j \sim \text{Ga}(s_j|\alpha, b)$$

$$m_j \sim \text{Ga}(m_j|a, b)$$

$$y_j = m_j + s_j$$

$$y_j \sim \text{Ga}(y_j|a + \alpha, b)$$

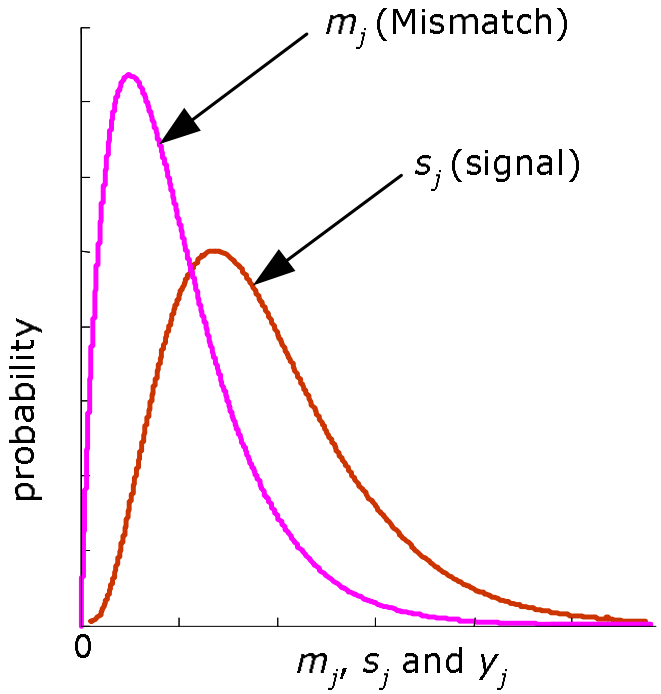


Figure: PDF of m_j, s_j and the implied distribution for y_j.

gMOS Family of Methods II

- Gamma Model of Signal

$$s_j \sim \text{Ga}\left(s_j|\alpha, b\right)$$

$$m_j \sim \text{Ga}\left(m_j|a, b\right)$$

$$y_j = m_j + s_j$$

$$y_j \sim \text{Ga}\left(y_j|a+\alpha, b\right)$$

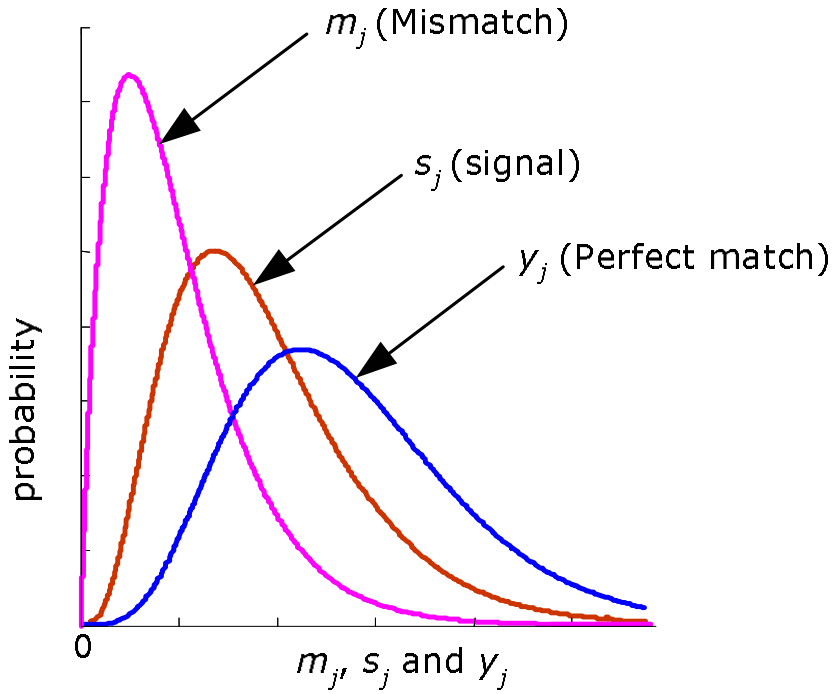


Figure: PDF of m_j, s_j and the implied distribution for y_j.

Modelling Probe Pair Affinity

- mgMOS
- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale.
  - Assume a probability distribution for the shared scale and 'marginalise'.

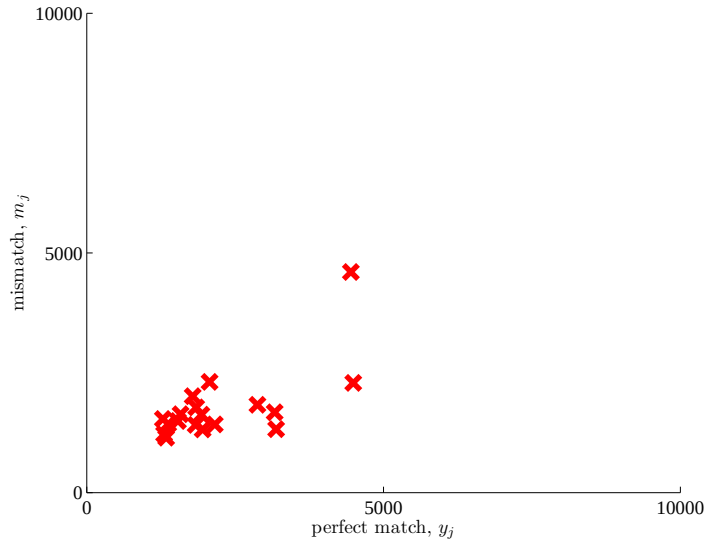


Figure: Correlation of PM (y_j) and MM (m_j).

Modelling Probe Pair Affinity

- mgMOS
- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale.
  - Assume a probability distribution for the shared scale and 'marginalise'.

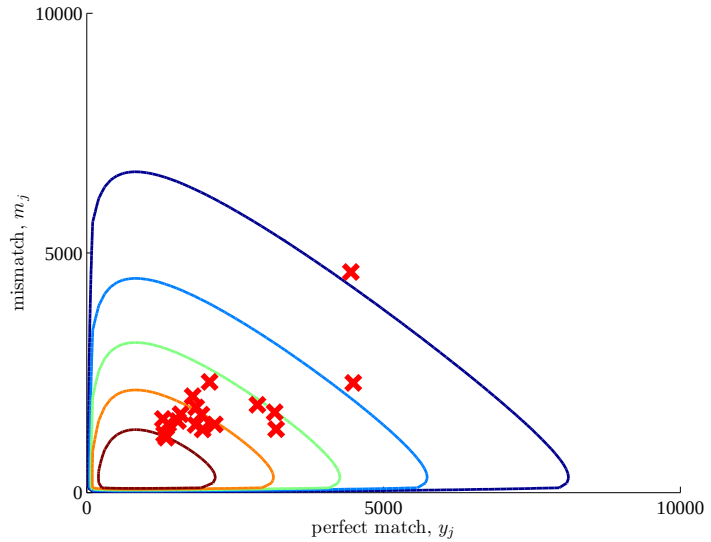


Figure: Correlation of PM (y_j) and MM (m_j).

Modelling Probe Pair Affinity

- mgMOS
- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale.
  - Assume a probability distribution for the shared scale and 'marginalise'.

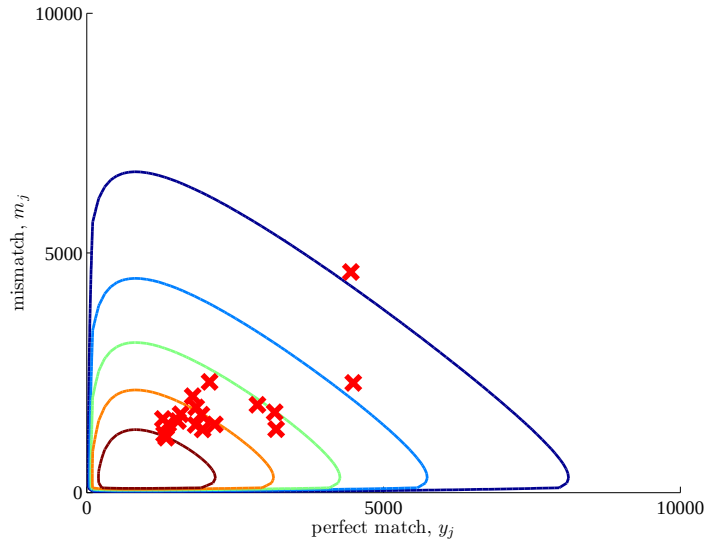


Figure: Correlation of PM (y_j) and MM (m_j).

Modelling Probe Pair Affinity

- mgMOS
- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale.
  - Assume a probability distribution for the shared scale and 'marginalise'.

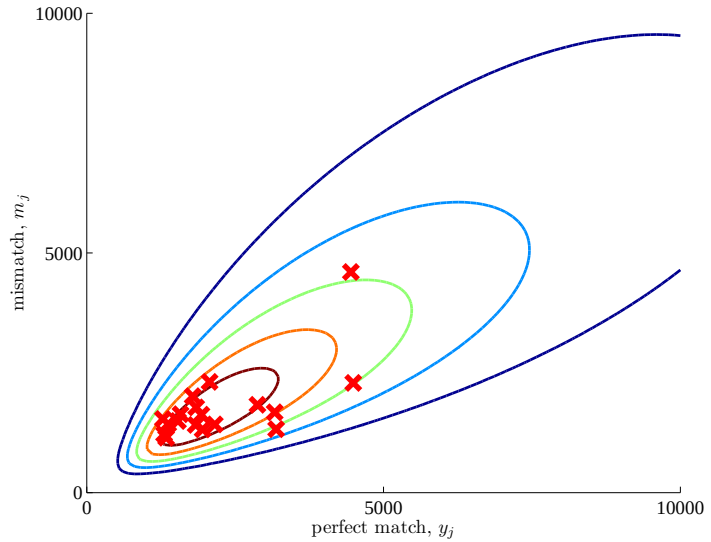


Figure: Correlation of PM (y_j) and MM (m_j).

Modelling Probe Pair Affinity

- mgMOS
- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale.
  - Assume a probability distribution for the shared scale and 'marginalise'.

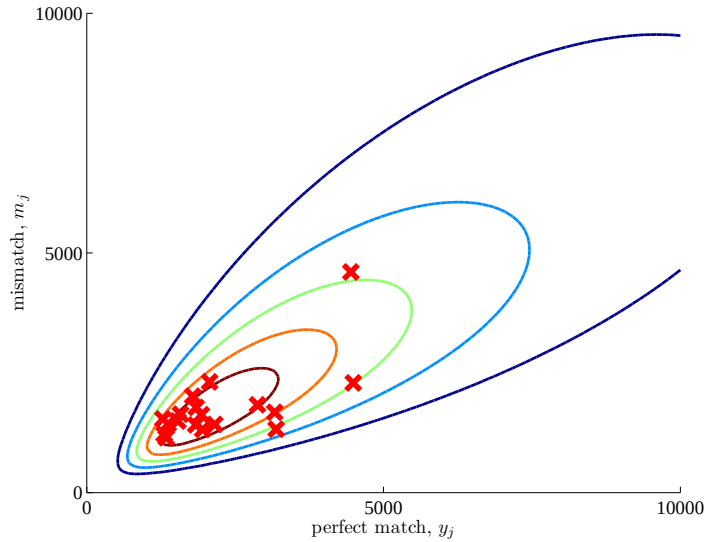


Figure: Correlation of PM (y_j) and MM (m_j).

Specific Binding to Mismatch

- Mismatch Effected by Signal
- Affymetrix Latin Square Spike-In data set.
- The perfect match responds to increasing mRNA.
- But so does the mismatch.

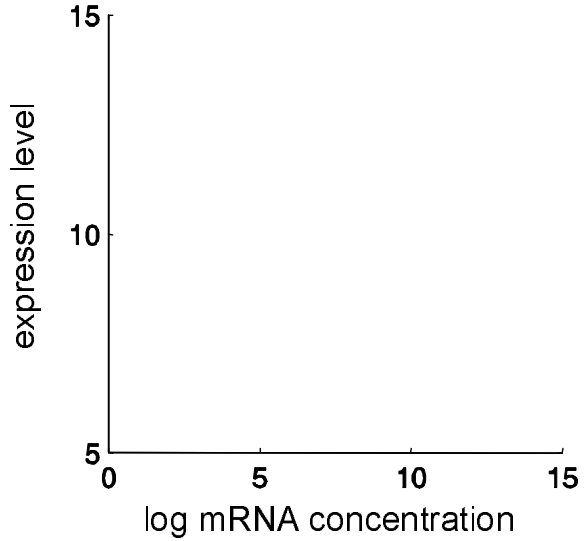


Figure: The perfect match goes up with the mRNA concentration as expected. But so does the mismatch.

Specific Binding to Mismatch

- Mismatch Effected by Signal
- Affymetrix Latin Square Spike-In data set.
- The perfect match responds to increasing mRNA.
- But so does the mismatch.

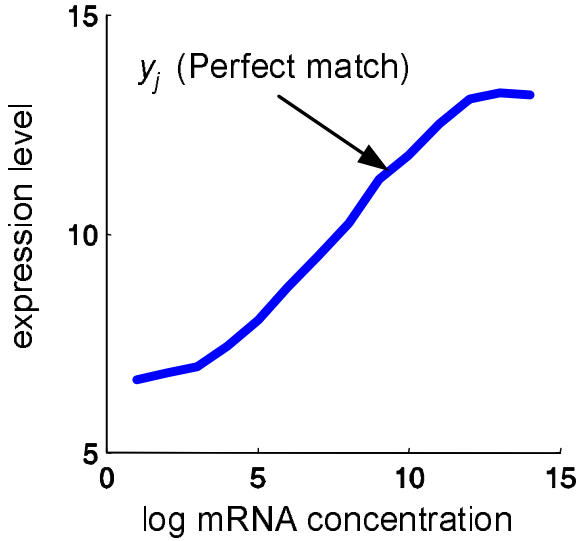


Figure: The perfect match goes up with the mRNA concentration as expected. But so does the mismatch.

Specific Binding to Mismatch

- Mismatch Effected by Signal
- Affymetrix Latin Square Spike-In data set.
- The perfect match responds to increasing mRNA.
- But so does the mismatch.

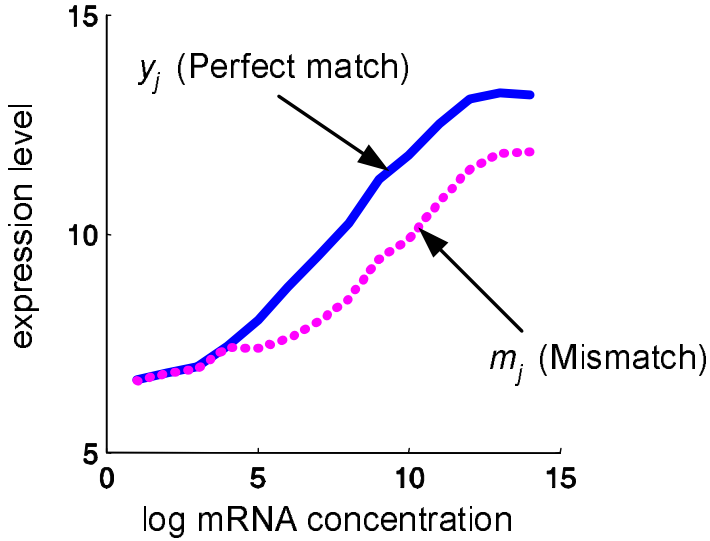


Figure: The perfect match goes up with the mRNA concentration as expected. But so does the mismatch.

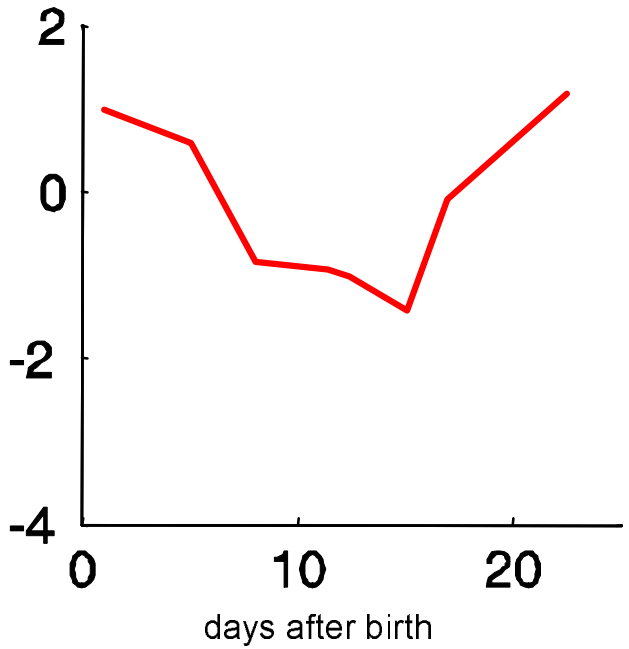
Specific Binding and Multiple Arrays

- multi-mgMOS
- Specific Binding to MM probe:
- An additional parameter is used to account for binding to MM probe.
- Multiple arrays:
  - Some of the parameters in the model are specific to the chip, not the sample.
  - Share these parameters across the arrays.

Mouse Data Set

- `http://www.ncbi.nlm.nih.gov/projects/geo`
- Mouse back skin mRNA expression profile for Dab2 [Lin et al., 2004].

RMSE	Root Mean Square Error	
	qr-PCR	x-probe set
MAS 5.0	0.656	0.360
GCRMA	0.694	0.370
multi-mgMOS	**0.601**	**0.233**

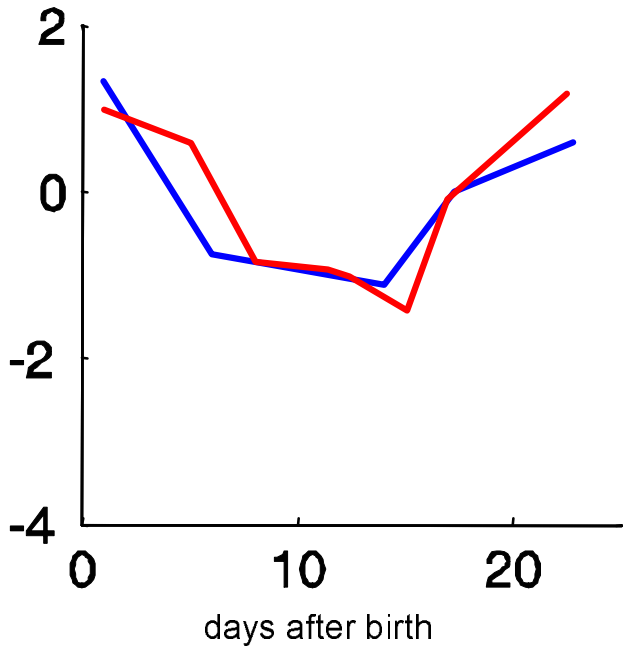


Prediction of Dab2 Expression level from qr-PCR, MAS 5.0, GCRMA and multi-mgMOS.

Mouse Data Set

- `http://www.ncbi.nlm.nih.gov/projects/geo`
- Mouse back skin mRNA expression profile for Dab2 [Lin et al., 2004].

RMSE	Root Mean Square Error	
	qr-PCR	x-probe set
MAS 5.0	0.656	0.360
GCRMA	0.694	0.370
multi-mgMOS	**0.601**	**0.233**

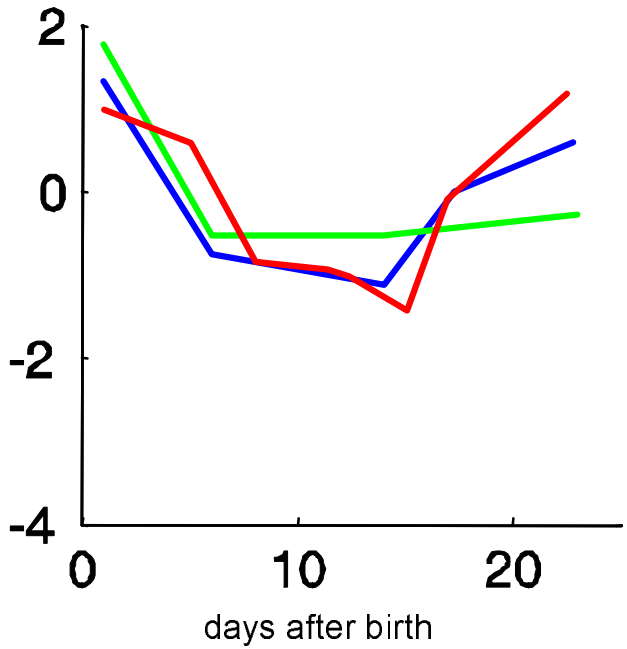


Prediction of Dab2 Expression level from qr-PCR, MAS 5.0, GCRMA and multi-mgMOS.

Mouse Data Set

- `http://www.ncbi.nlm.nih.gov/projects/geo`
- Mouse back skin mRNA expression profile for Dab2 [Lin et al., 2004].

RMSE	Root Mean Square Error	
	qr-PCR	x-probe set
MAS 5.0	0.656	0.360
GCRMA	0.694	0.370
multi-mgMOS	0.601	0.233

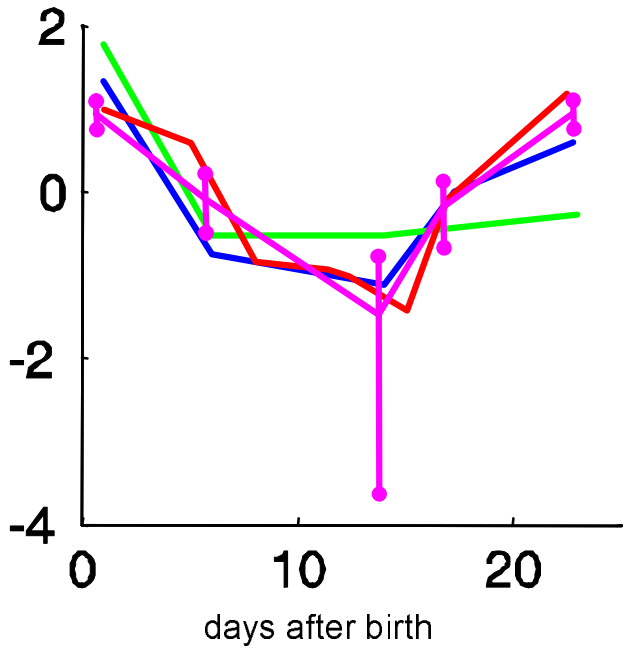


Prediction of Dab2 Expression level from qr-PCR, MAS 5.0, GCRMA and multi-mgMOS.

Mouse Data Set

- `http://www.ncbi.nlm.nih.gov/projects/geo`
- Mouse back skin mRNA expression profile for Dab2 [Lin et al., 2004].

RMSE	Root Mean Square Error	
	qr-PCR	x-probe set
MAS 5.0	0.656	0.360
GCRMA	0.694	0.370
multi-mgMOS	**0.601**	**0.233**



Prediction of Dab2 Expression level from qr-PCR, MAS 5.0, GCRMA and multi-mgMOS.

Differential Gene Expression

- Probability of Positive Log Ratio[Liu et al., 2006]
  - Differential gene expression is normally assessed with log ratios of gene expression.

$$r_{ij} = \log \frac{s_i}{s_j}$$

  - This measure is very sensitive to noise at low expresion levels.
  - Use variance of expression to obtain Probability of Positive Log Ratio (PPLR).

Differential Gene Expression

- Probability of Positive Log Ratio[Liu et al., 2006]
  - Differential gene expression is normally assessed with log ratios of gene expression.
$$r_{ij} = \log \frac{s_i}{s_j}$$
  - This measure is very sensitive to noise at low expresion levels.
  - Use variance of expression to obtain Probability of Positive Log Ratio (PPLR).

Differential Gene Expression

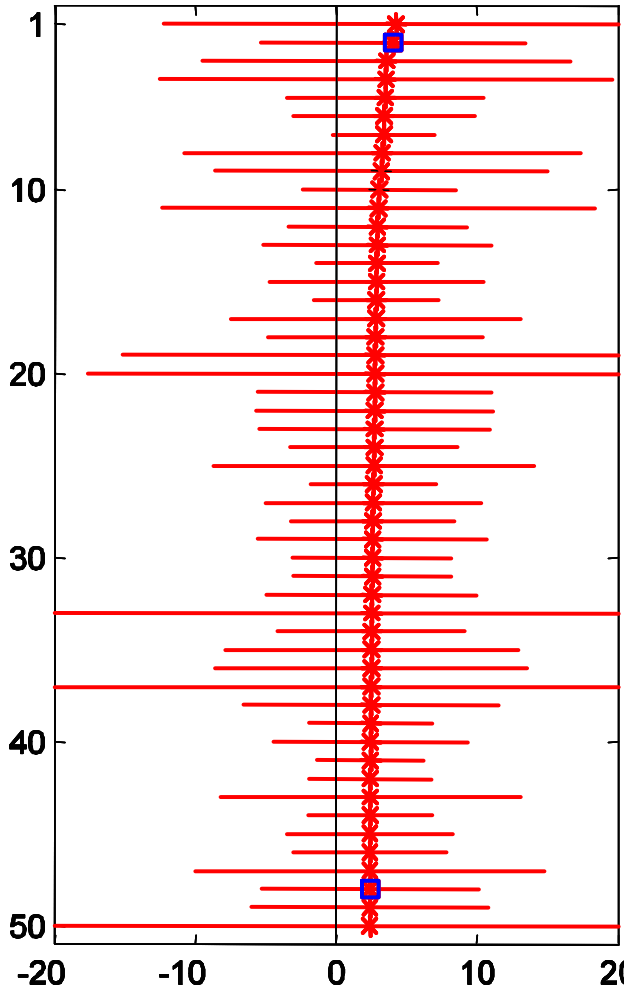
- Probability of Positive Log Ratio[Liu et al., 2006]
  - Differential gene expression is normally assessed with log ratios of gene expression.

$$r_{ij} = \log \frac{s_i}{s_j}$$

  - This measure is very sensitive to noise at low expresion levels.
  - Use variance of expression to obtain Probability of Positive Log Ratio (PPLR).

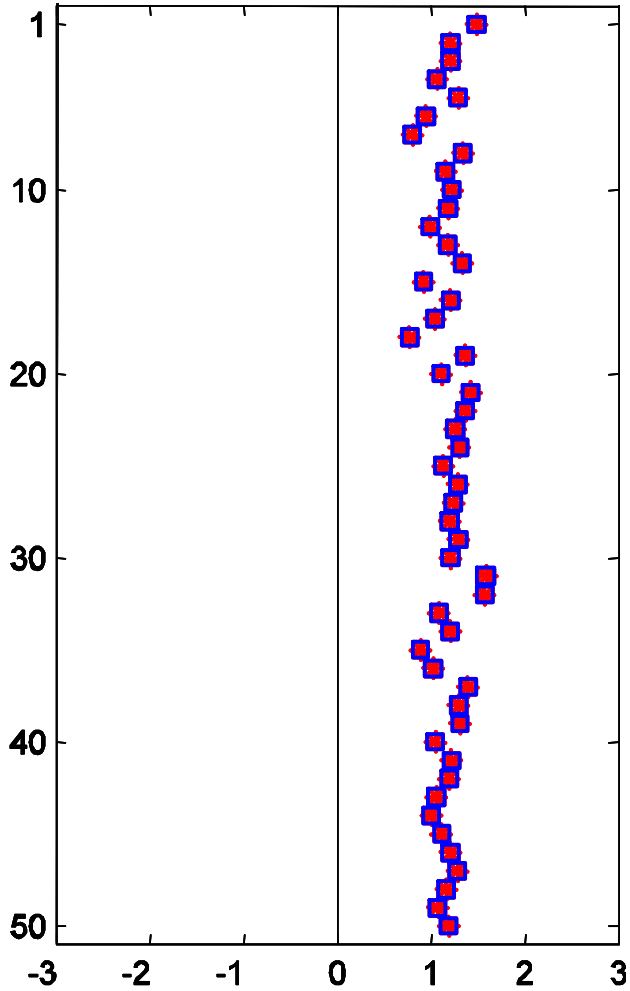
PPLR Results

- Golden spike-in dataset [Choe et al., 2005]
- Ranking (y-axis) against log ratio (x-axis) for.
  - **Ranking by Expected Log Ratio**.
  - Ranking by PPLR.
- Red stars indicate expected log ratio.
- Red lines indicate error bars.
- Blue squares indicates genes that were spiked-in.



PPLR Results

- Golden spike-in dataset [Choe et al., 2005]
- Ranking (y-axis) against log ratio (x-axis) for.
  - Ranking by Expected Log Ratio.
  - **Ranking by PPLR**.
- Red stars indicate expected log ratio.
- Red lines indicate error bars.
- Blue squares indicates genes that were spiked-in.



Cleaning up Profiles

- Converting Noisy Profiles to Clean
  - If we can 'clean up' the profiles we can use in other methods.
  - Construct a probabilistic model for the data and corruption process.
  - Work with posterior distribution over cleaned up profile.
  - We designed a *heteroschedastic probabilistic* PCA for doing this [Sanguinetti et al., 2005].

Heteroschedastic PPCA Results

- Mouse Cochlear Dataset
  - Data from a conditionally imortal cell lin extracted from mouse cochlear epithelial cells.
  - Twelve samples from 14 days of differentiation after extration at E13.5 [Rivolta et al., 2002].
  - Experimental setup:
    - Perform HPPCA/PCA on the data.
    - Extract 50 genes most associated with 2nd principal component
    - Cluster original profiles and reconstructed profiles.

Heteroschedastic PPCA Results

- Mouse Cochlear Dataset
  - Data from a conditionally imortal cell lin extracted from mouse cochlear epithelial cells.
  - Twelve samples from 14 days of differentiation after extration at E13.5 [Rivolta et al., 2002].
  - Experimental setup:
    - Perform HPPCA/PCA on the data.
    - Extract 50 genes most associated with 2nd principal component
    - Cluster original profiles and reconstructed profiles.

Heteroschedastic PPCA Results

- Mouse Cochlear Dataset
  - Data from a conditionally imortal cell lin extracted from mouse cochlear epithelial cells.
  - Twelve samples from 14 days of differentiation after extration at E13.5 [Rivolta et al., 2002].
  - Experimental setup:
    - Perform HPPCA/PCA on the data.
    - Extract 50 genes most associated with 2nd principal component
    - Cluster original profiles and reconstructed profiles.

Heteroschedastic PPCA Results

- Mouse Cochlear Dataset
  - Data from a conditionally imortal cell lin extracted from mouse cochlear epithelial cells.
  - Twelve samples from 14 days of differentiation after extration at E13.5 [Rivolta et al., 2002].
  - Experimental setup:
    - ★ Perform HPPCA/PCA on the data.
    - ★ Extract 50 genes most associated with 2nd principal component
    - ★ Cluster original profiles and reconstructed profiles.

Heteroschedastic PPCA Results

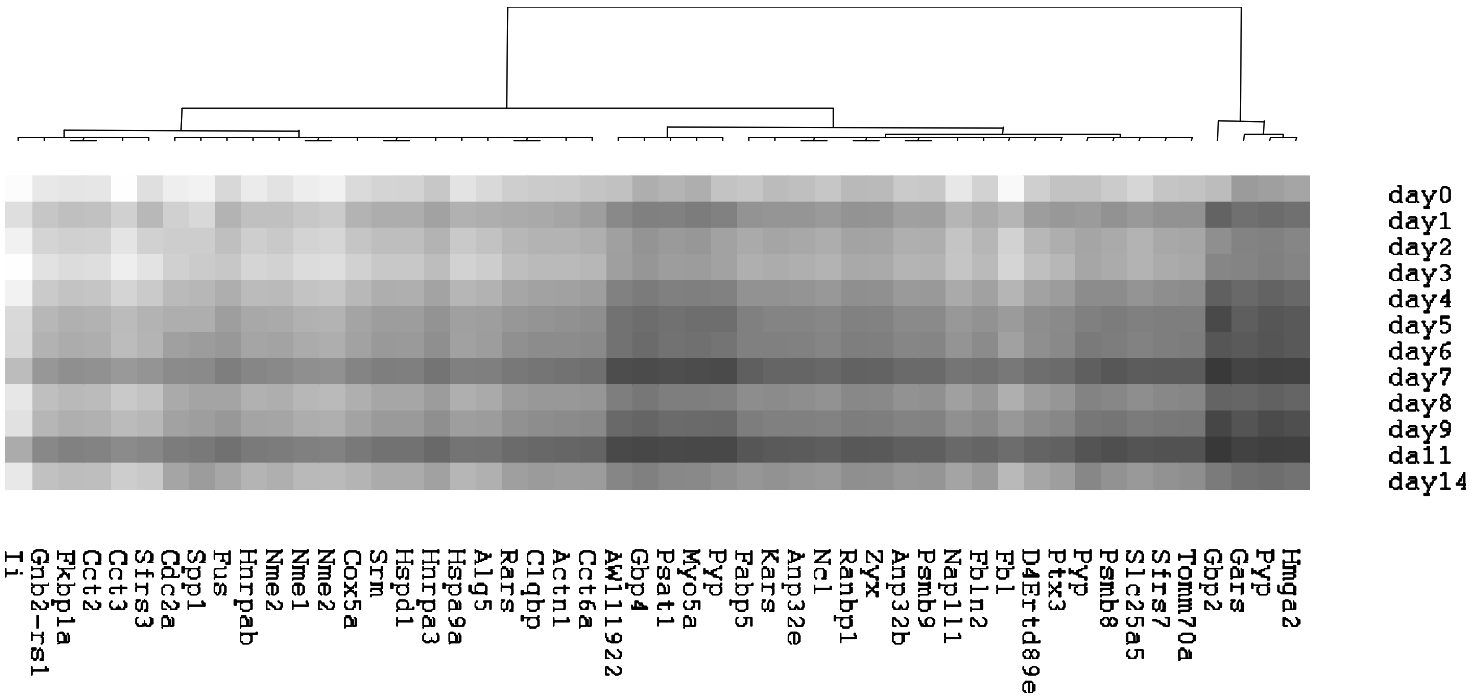


Figure: Hierarchical Clustering on Corrected Profiles.

Heteroschedastic PPCA Results

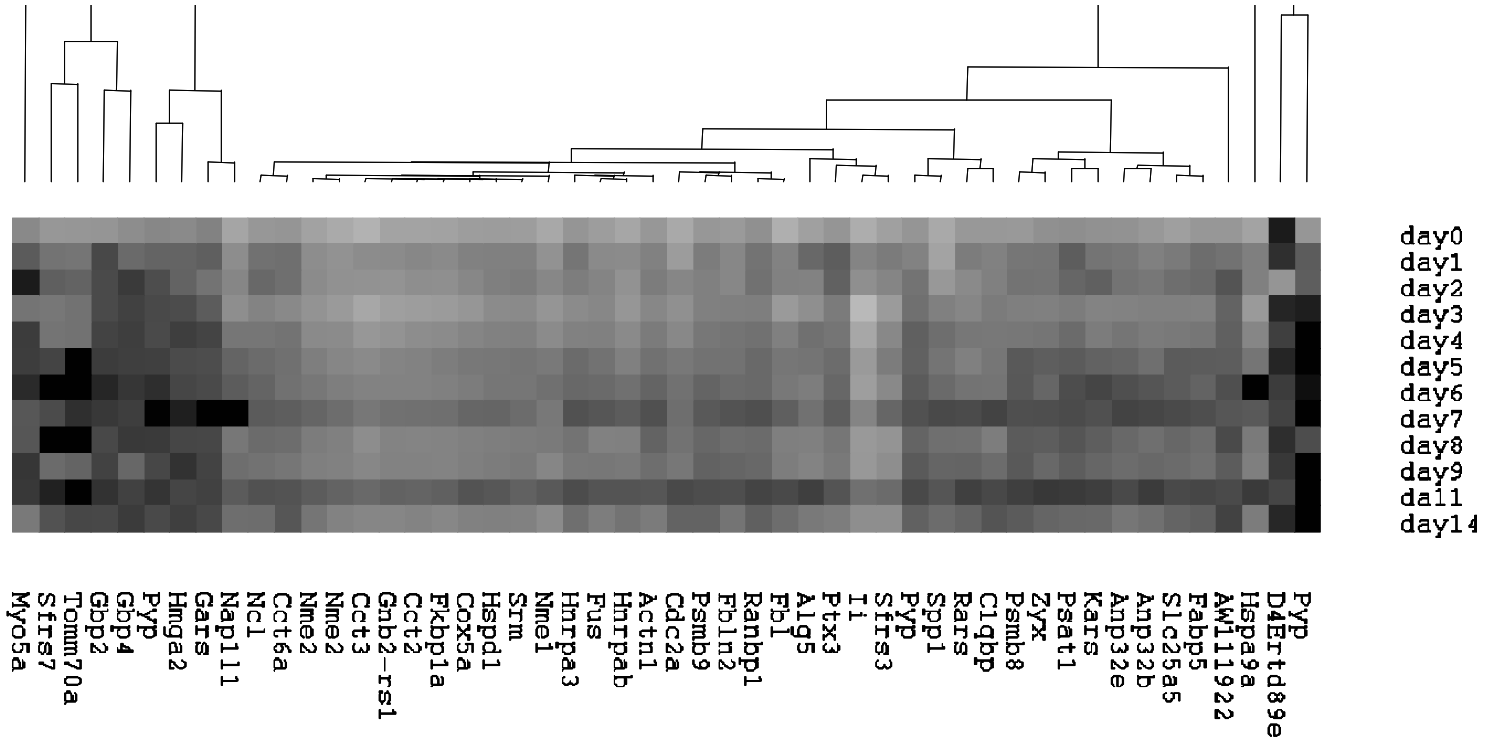


Figure: Hierarchical Clustering on Uncorrected Profiles.

Heteroschedastic PPCA Results

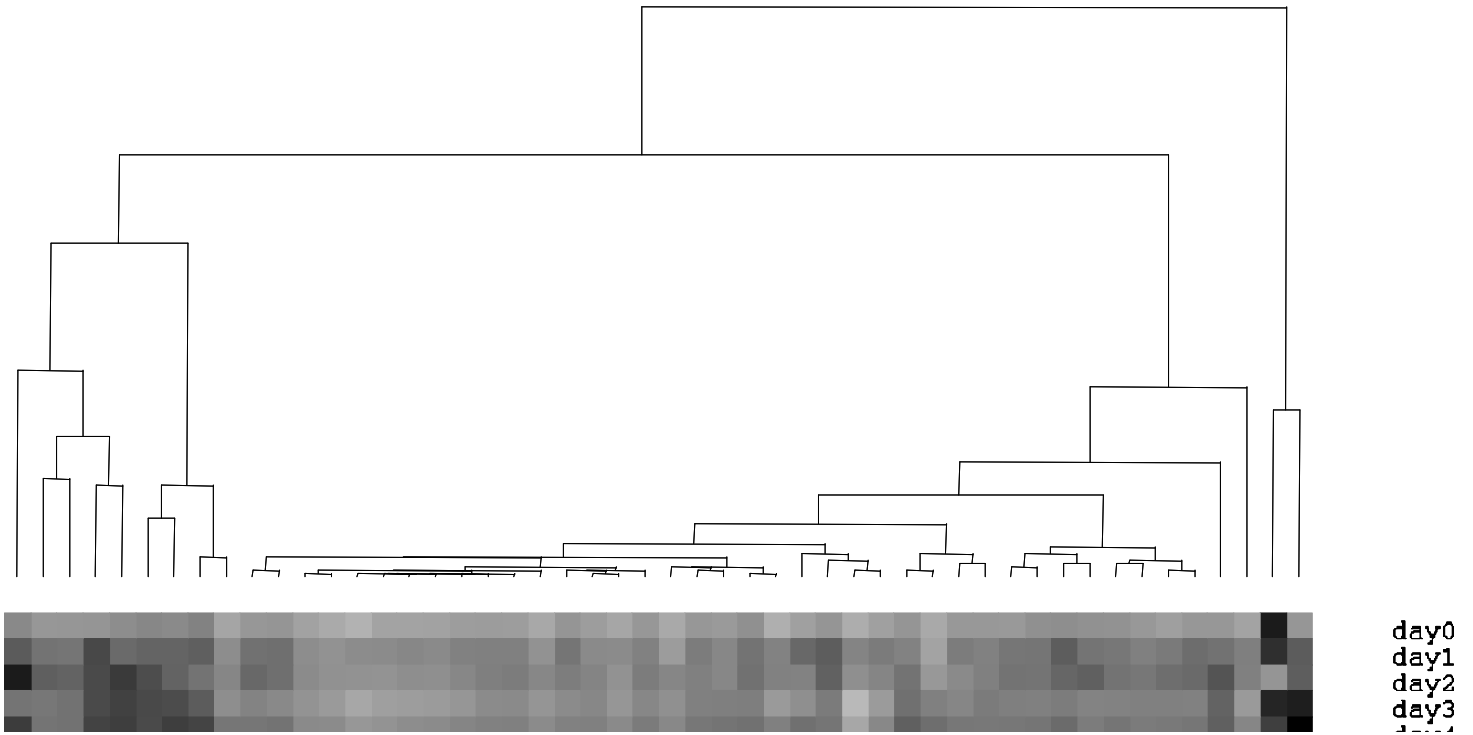


Figure: Hierarchical Clustering on Uncorrected Profiles.

Heteroschedastic PPCA Results

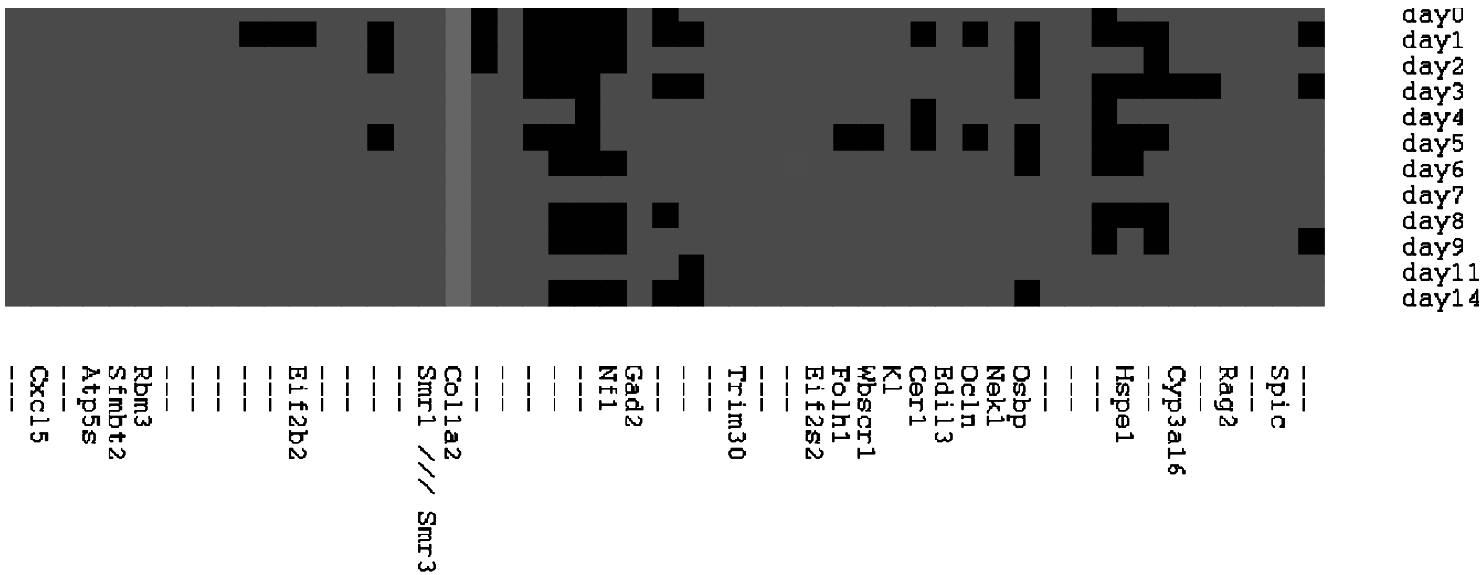


Figure: Hierarchical Clustering on genes selected by normal PCA.

Heteroschedastic PPCA Results

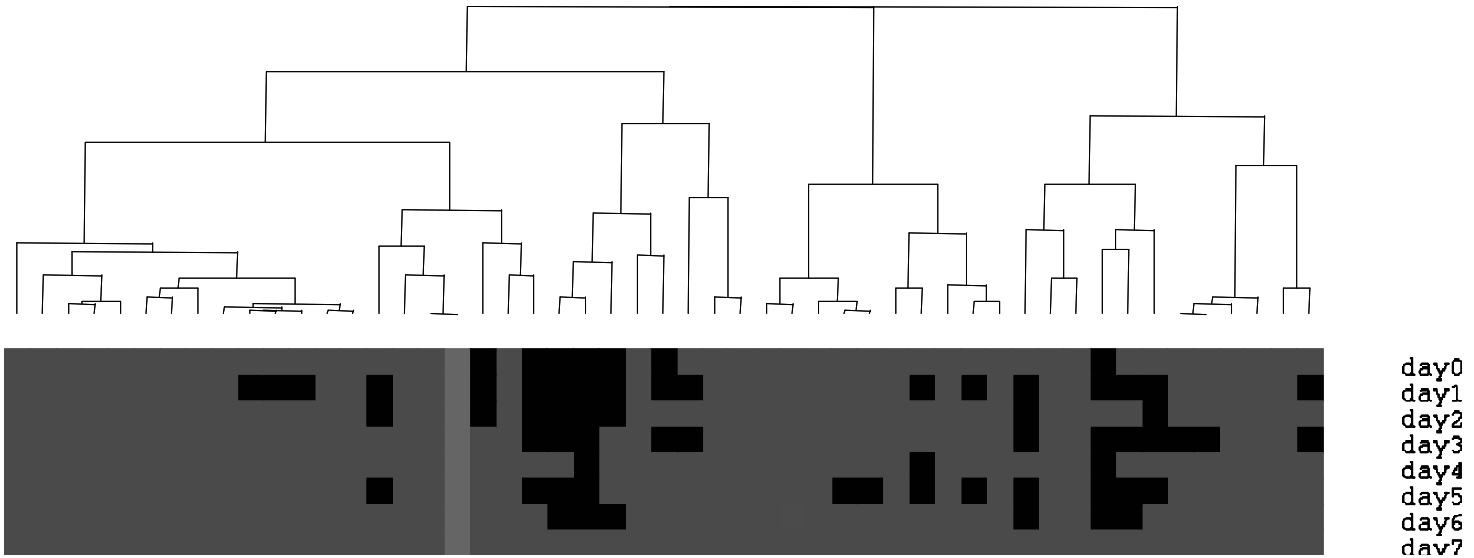
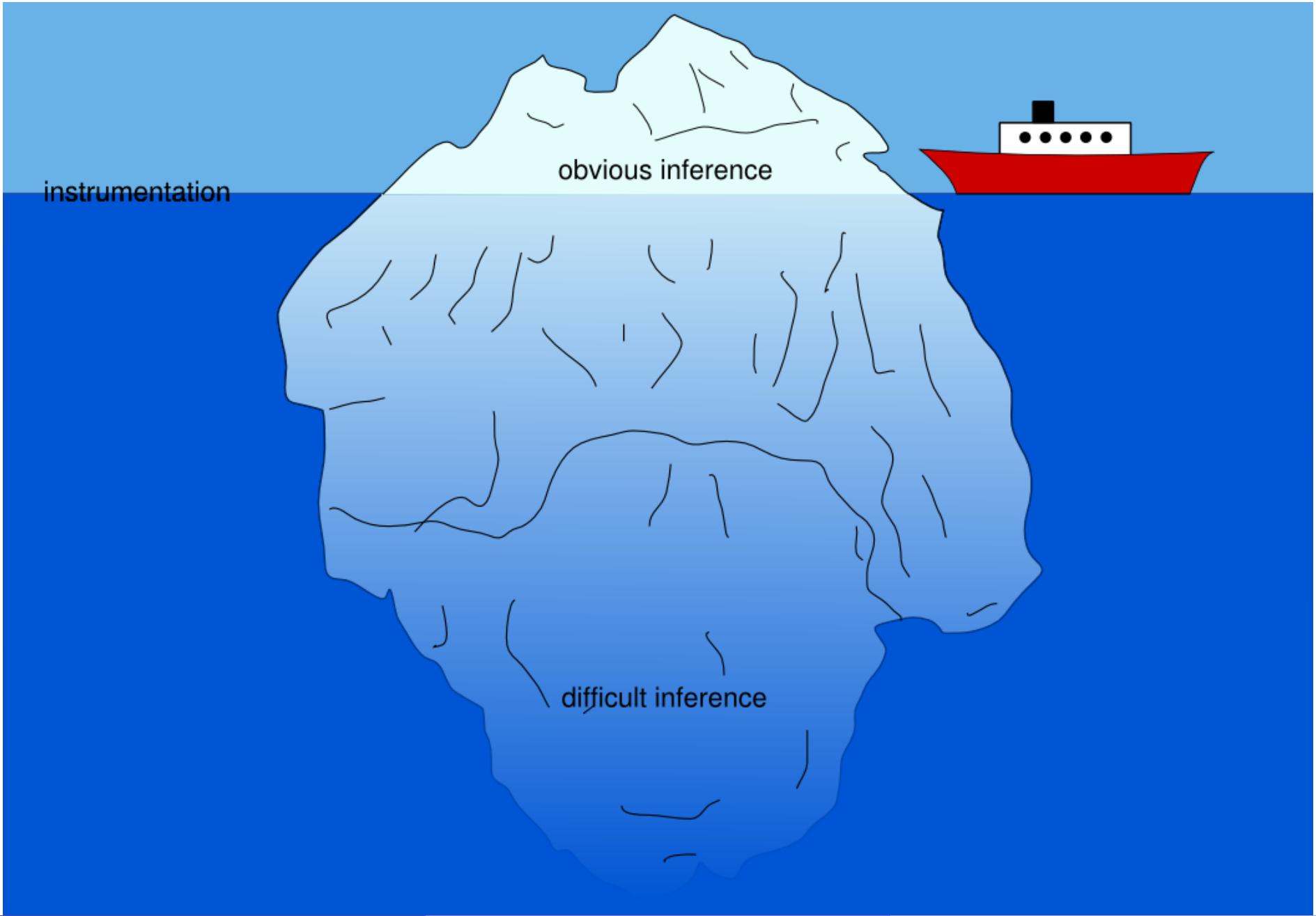
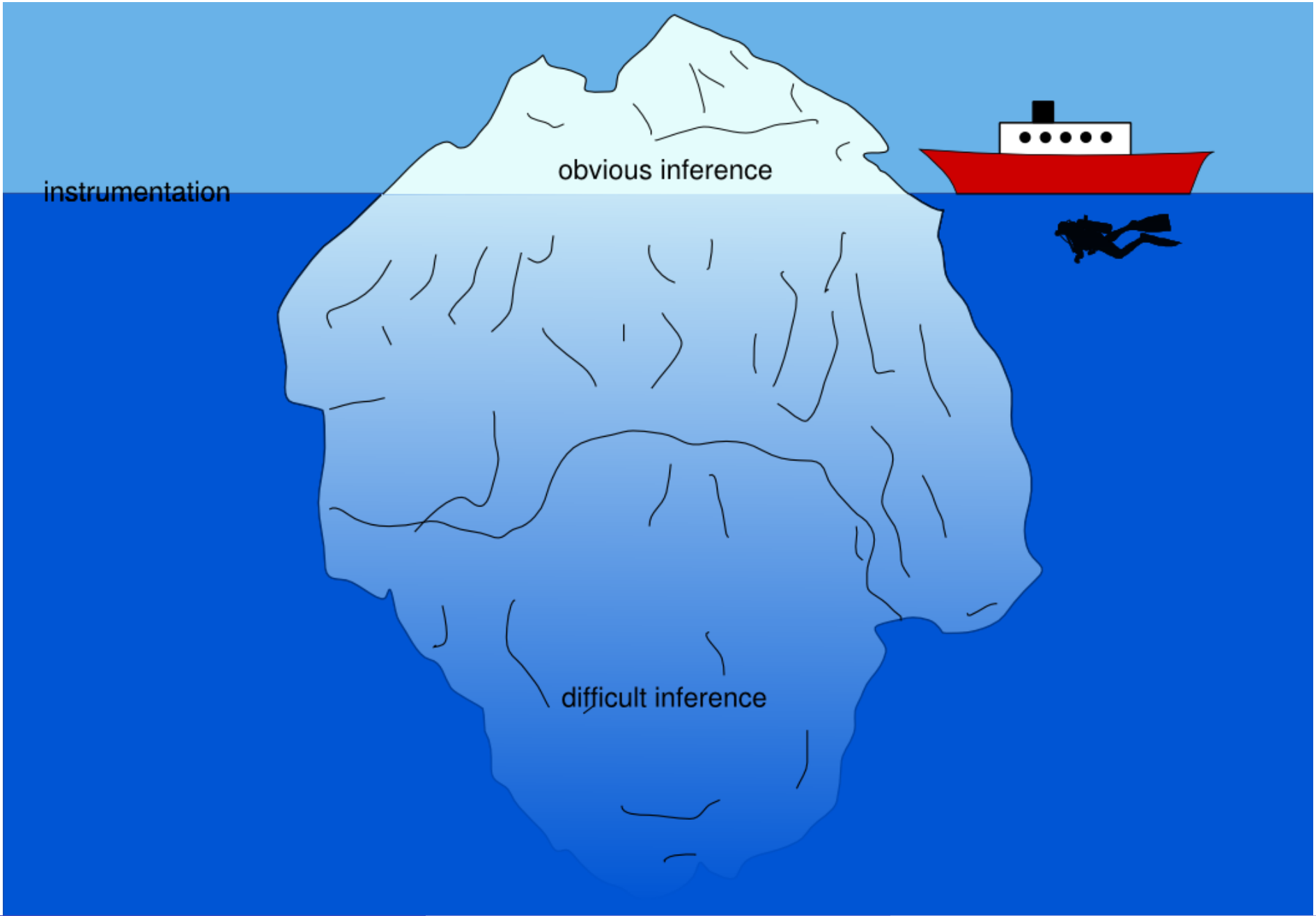


Figure: Hierarchical Clustering on genes selected by normal PCA.

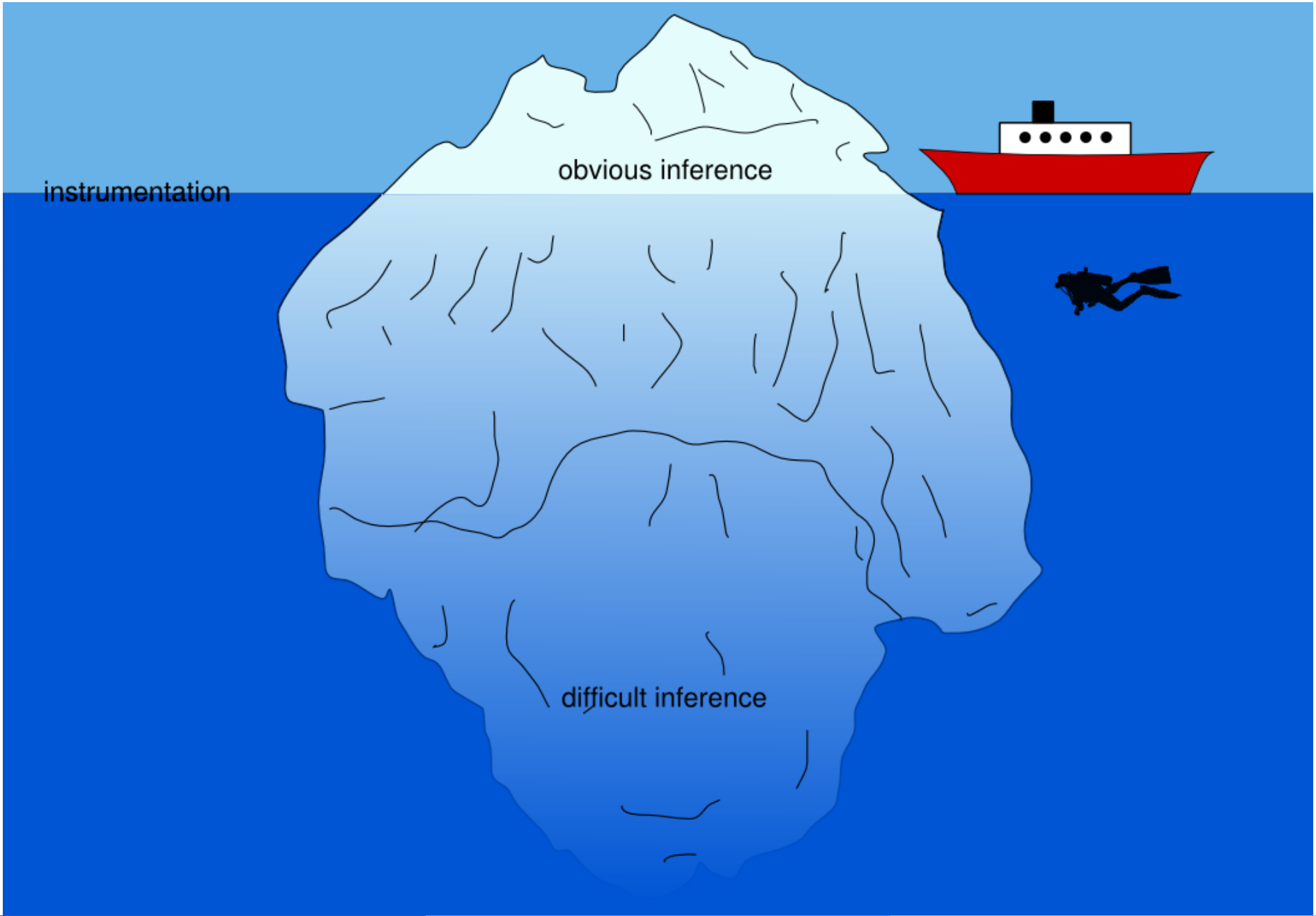
The Iceberg of Information



The Iceberg of Information



The Iceberg of Information



Transcription Factor Activities

- Transcription factors control the expression of genes.
- Knowledge of their 'activity' is key to understanding the mechanism behind biological processes.
- Transcription factors are proteins — activity is a combination of their concentration and effect.
- The mRNA concentration of a given transcription factor may be known but:
  - Transcription factors are often lowly expressed — mRNA concentrations difficult to measure.
  - Transcription factors are often post-transcriptionally regulated.
- We can see these TFs as *latent chemical species*.
  - Latent chemical species are a common issue in biological problems.

ChIP Microarrays

- Chromatine Immunoprecipitation (ChIP) Microarrays
  - ChIP Microarrays tell us which TFs bind to which genes under certain conditions.
  - In effect this gives a structure for the regulatory network.
  - We use *binary* output from ChIP arrays.
  - Combine this information with gene expression data to obtain transcription factor activities (TFA).
  - Approach also works for other sources of connectivity information (motifs).

Transcription Factor Activities

- Evaluating Activities of Transcription Factors
  - Several approaches based on regression [Liao et al., 2003, Gao et al., 2004, Boulesteix and Strimmer, 2005, Alter and Golub, 2004]
  - Assume a gene's expresion is given by a linear relationship

$$\mathbf{y}_i = \mathbf{B}\mathbf{x}_i + \boldsymbol{\epsilon}_i.$$

$\mathbf{y}_i$ is the expression profile of the ith gene with (contains T experiments),

$\mathbf{x}_i$ is binary: indicates which transctiption factors bind to the ith gene (there are q transcription factors total)

$\mathbf{B} \in \Re^{T \times q}$ is the matrix of TFAs.

ϵ_i is a noise term.

  - Intuition: $\mathbf{x}_i$ 'selects' which columns of $\mathbf{B}$ are switched on to recreate $\mathbf{y}_i$.
  - Problem: the matrix $\mathbf{B}$ is *not* gene specific. It gives average TFA across genes.

Gene Specific TFAs

- Associate TFAs to Genes [Sanguinetti et al., 2006]
  - Intoduce gene specific TFAs,

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{x}_i + \boldsymbol{\epsilon}_i.$$

  - Parameter Explosion — use Bayesian techniques to deal with the number of parameters in the model.
  - Time Course — encourage transcription factor activity to vary smoothly over time.
  - Temporal continuity parameter, γ, is between 0 and 1
    - When $\gamma = 0$ the experiments are unrelated to each other (in terms of time).
    - For $\gamma = 1$ the experiments are biological replicates.

Gene Specific TFAs

- Associate TFAs to Genes [Sanguinetti et al., 2006]
  - Intoduce gene specific TFAs,

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{x}_i + \boldsymbol{\epsilon}_i.$$

  - Parameter Explosion — use Bayesian techniques to deal with the number of parameters in the model.
  - Time Course — encourage transcription factor activity to vary smoothly over time.
  - Temporal continuity parameter, γ, is between 0 and 1
    - When $\gamma = 0$ the experiments are unrelated to each other (in terms of time).
    - For $\gamma = 1$ the experiments are biological replicates.

Gene Specific TFAs

- Associate TFAs to Genes [Sanguinetti et al., 2006]
  - Intoduce gene specific TFAs,

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{x}_i + \boldsymbol{\epsilon}_i.$$

  - Parameter Explosion — use Bayesian techniques to deal with the number of parameters in the model.
  - Time Course — encourage transcription factor activity to vary smoothly over time.
  - Temporal continuity parameter, γ, is between 0 and 1
    - When $\gamma = 0$ the experiments are unrelated to each other (in terms of time).
    - For $\gamma = 1$ the experiments are biological replicates.

Gene Specific TFAs

- Associate TFAs to Genes [Sanguinetti et al., 2006]
  - Intoduce gene specific TFAs,

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{x}_i + \boldsymbol{\epsilon}_i.$$

  - Parameter Explosion — use Bayesian techniques to deal with the number of parameters in the model.
  - Time Course — encourage transcription factor activity to vary smoothly over time.
  - Temporal continuity parameter, γ, is between 0 and 1
    - When $\gamma = 0$ the experiments are unrelated to each other (in terms of time).
    - For $\gamma = 1$ the experiments are biological replicates.

Gene Specific TFAs

- Associate TFAs to Genes [Sanguinetti et al., 2006]
  - Intoduce gene specific TFAs,

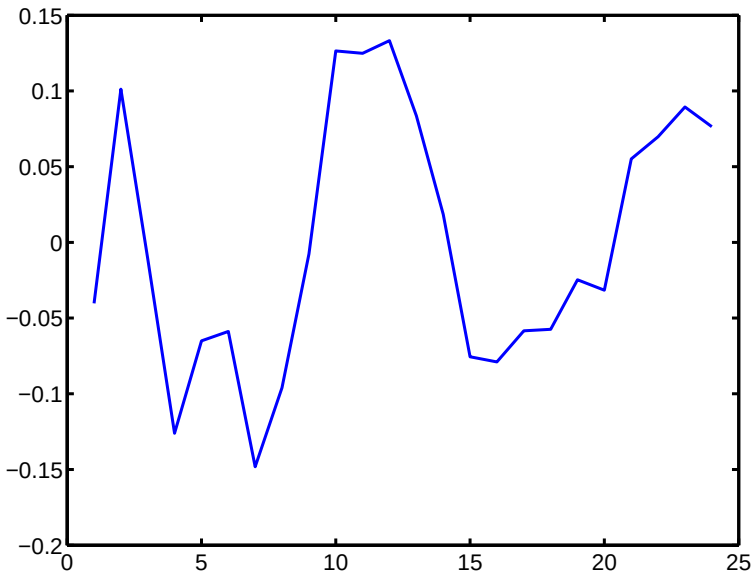
$$\mathbf{y}_i = \mathbf{B}_i \mathbf{x}_i + \boldsymbol{\epsilon}_i.$$

  - Parameter Explosion — use Bayesian techniques to deal with the number of parameters in the model.
  - Time Course — encourage transcription factor activity to vary smoothly over time.
  - Temporal continuity parameter, γ, is between 0 and 1
    - ⋆ When $\gamma = 0$ the experiments are unrelated to each other (in terms of time).
    - ⋆ For $\gamma = 1$ the experiments are biological replicates.

Results on TFAs

- Yeast Cell Cycle Data with ChIP-on-chip data
- Yeast cell cycle cdc15 data set [Spellman et al., 1998].
- ChIP on chip from 113 TFs [Lee et al., 2002].
- 24 experimental points in time series data.
- Compare with non-specific TFAs obtained by Regression.

Results on TFAs II



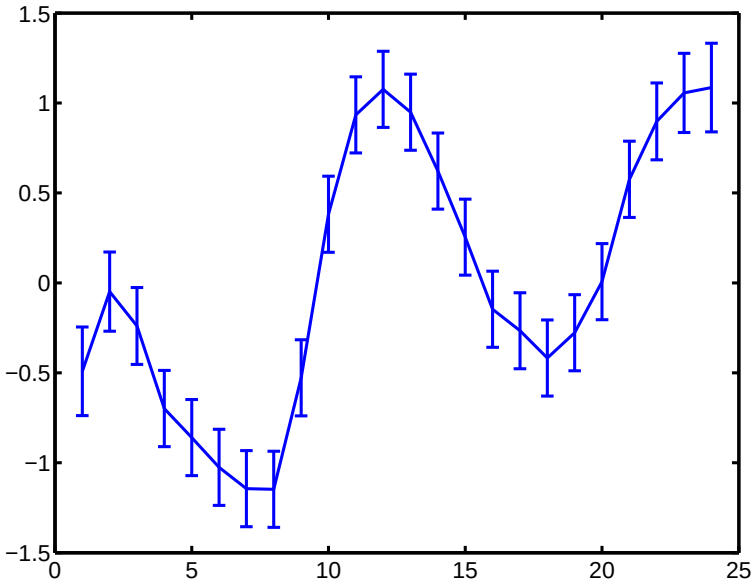
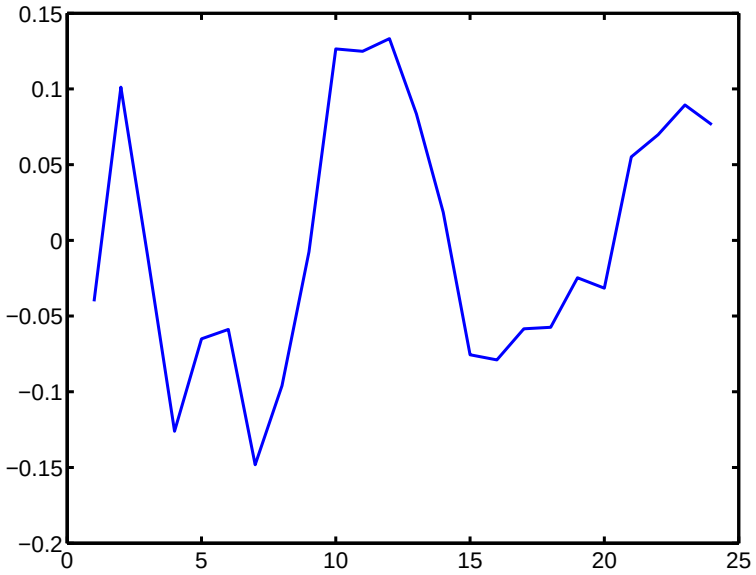


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA for averge of $\mathbf{B}_i$ across genes.

Results on TFAs II



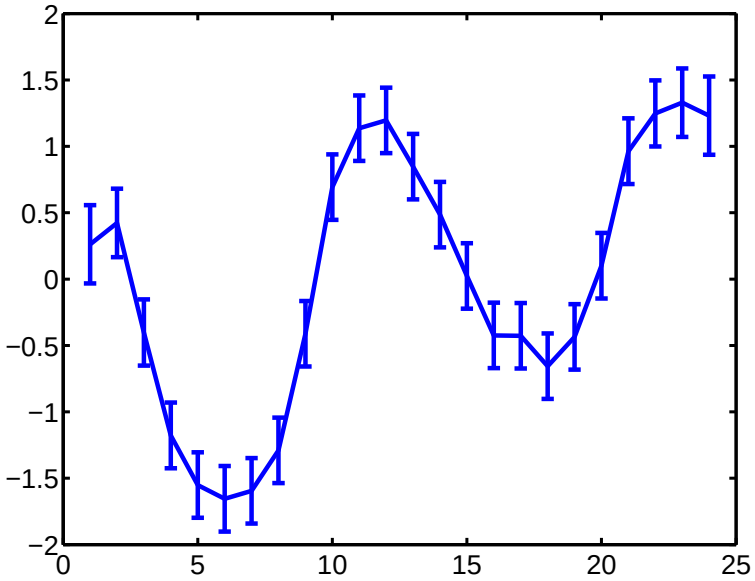
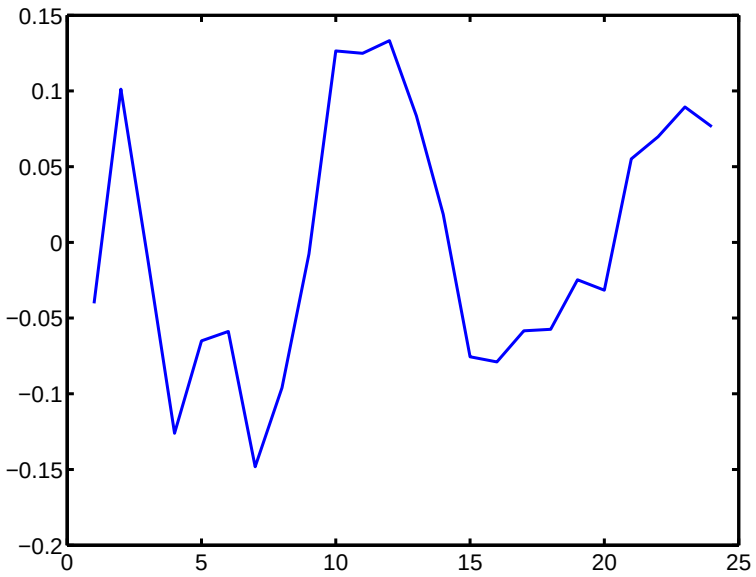


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA SCW11.

Results on TFAs II



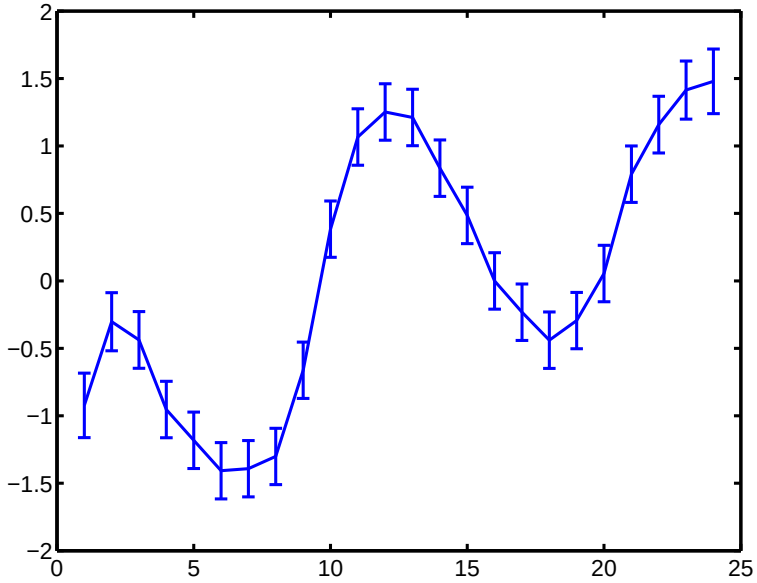
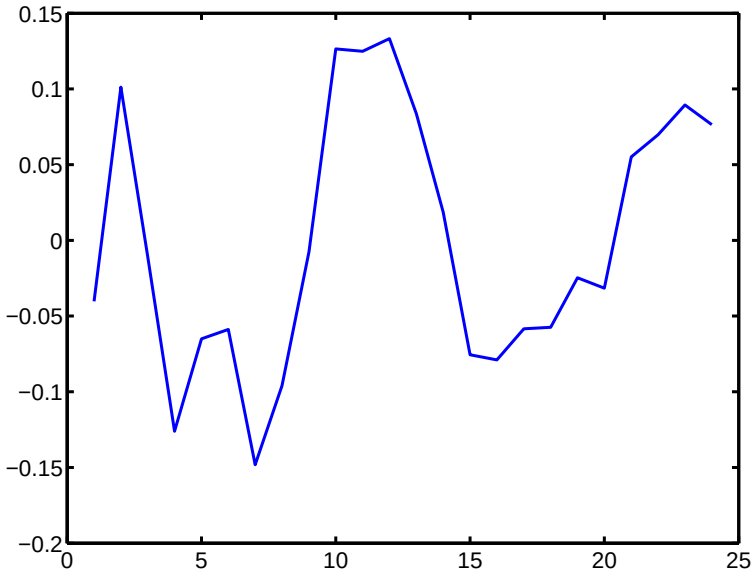


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA CTS1.

Results on TFAs II



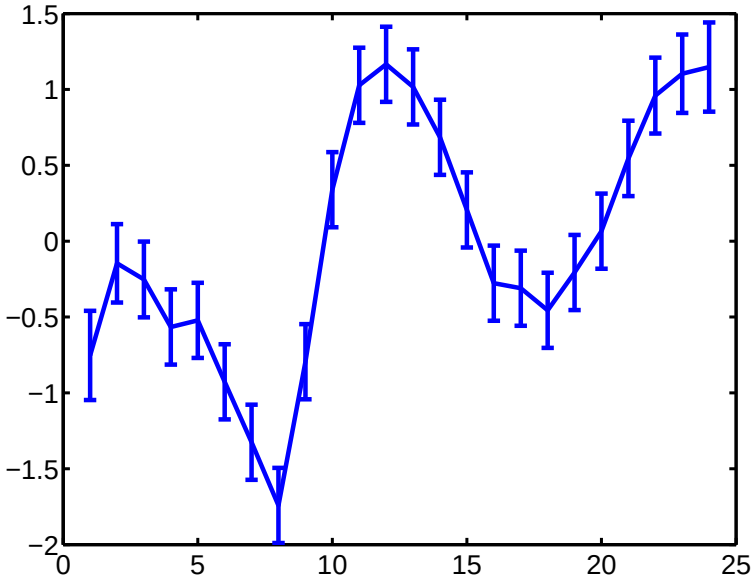
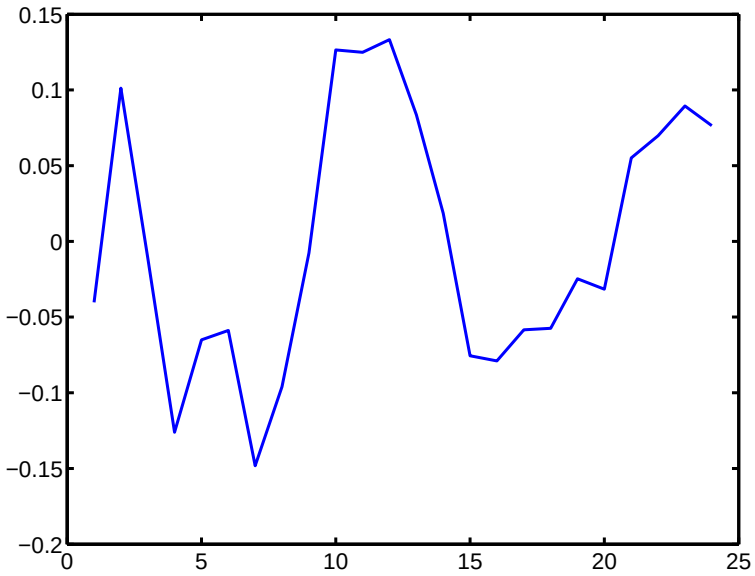


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA YER124C.

Results on TFAs II



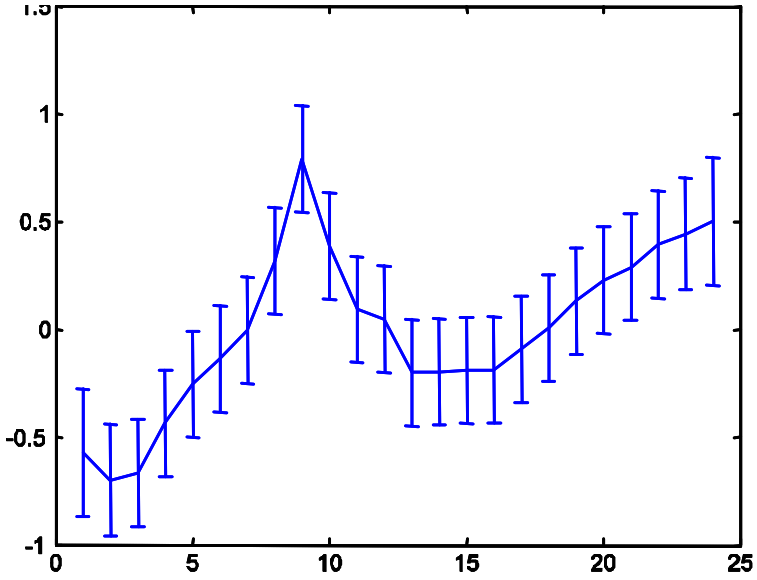


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA YKL51C.

Separation of Concentration and Effect

- Splitting the Activity into Component Parts
  - TFA is a combination of:
    - ★ TF 'concentration'.
    - ★ TF 'effect'.
  - Follow up model splits the TFA into its component parts.
    - ★ 'Concentration' is specific to the transcription factor (it's a time course).
    - ★ 'Effect' is specific to the gene (it's a single value — either positive (activation) or negative (repression)).
  - Bayesian treatment of $\mathbf{c}$ and $\mathbf{B}$ through a variational approach.

TF Concentration Results

Concentration of ACE2

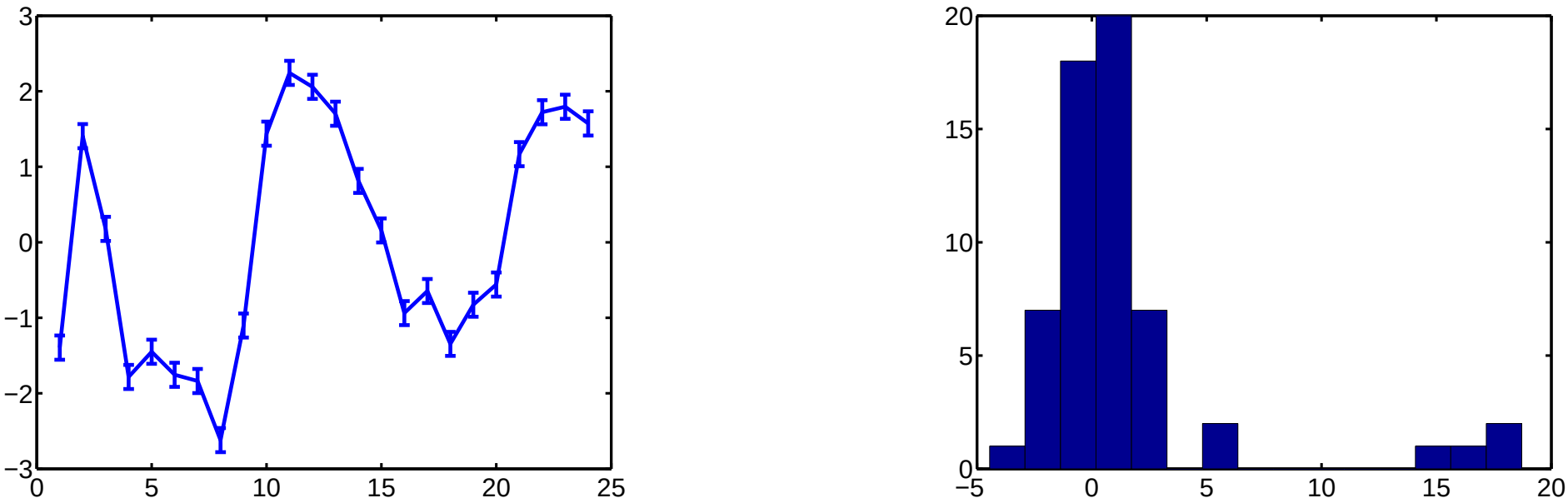
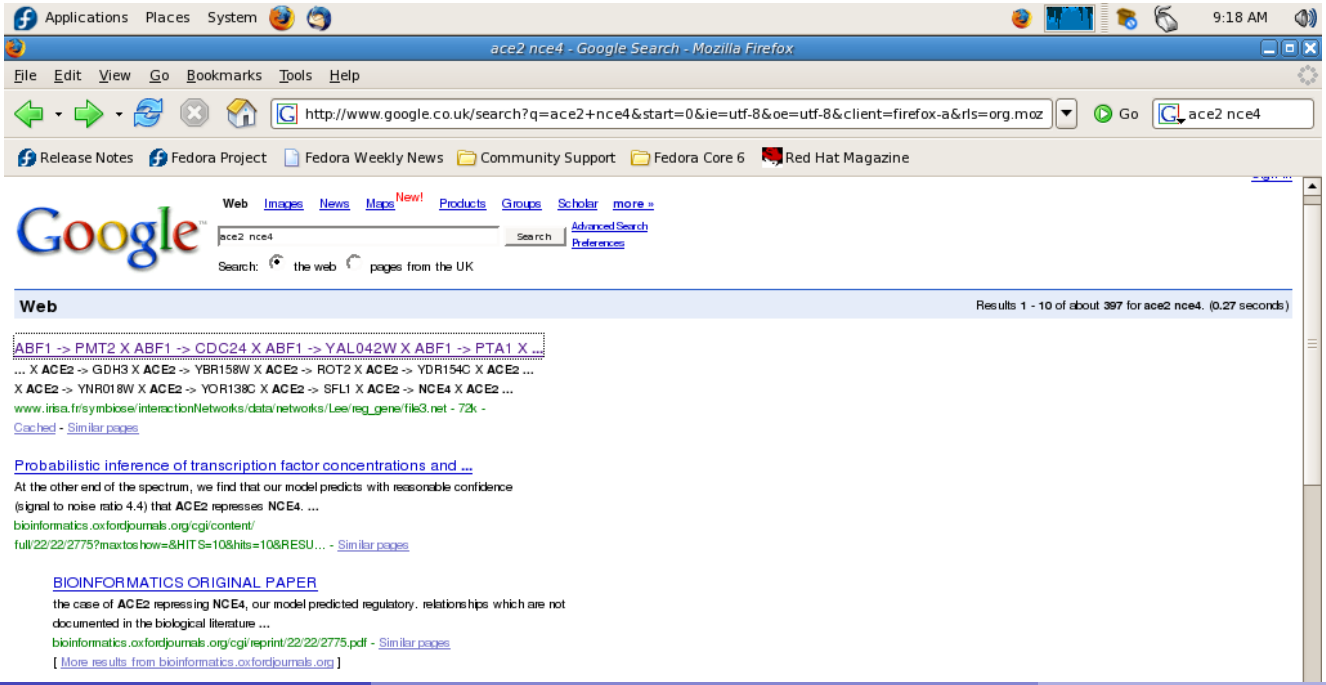


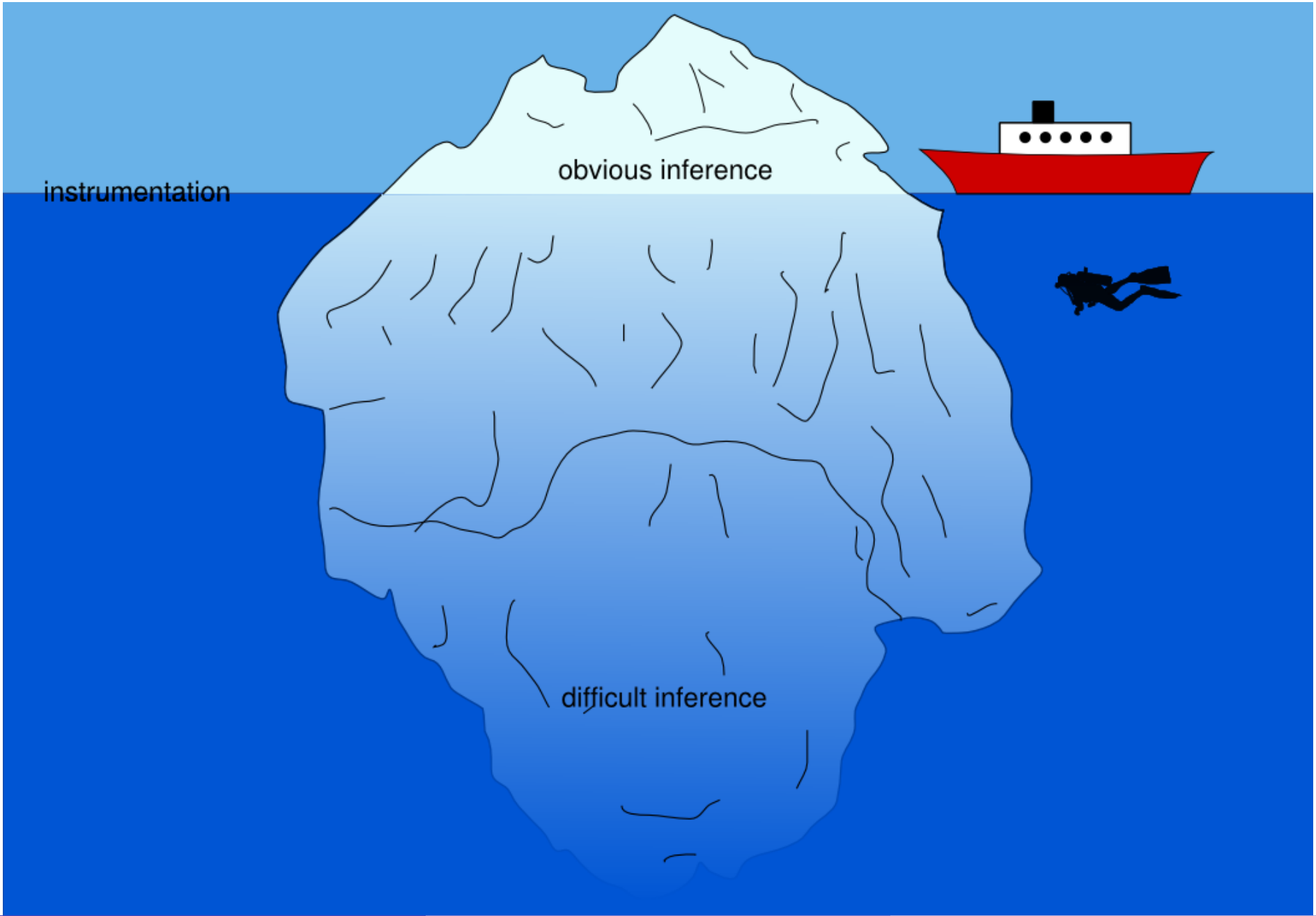
Figure: *Left:* concentration of ACE2 and *right*: effect of ACE2 on its target genes as a histogram.

TF Concentration Results II

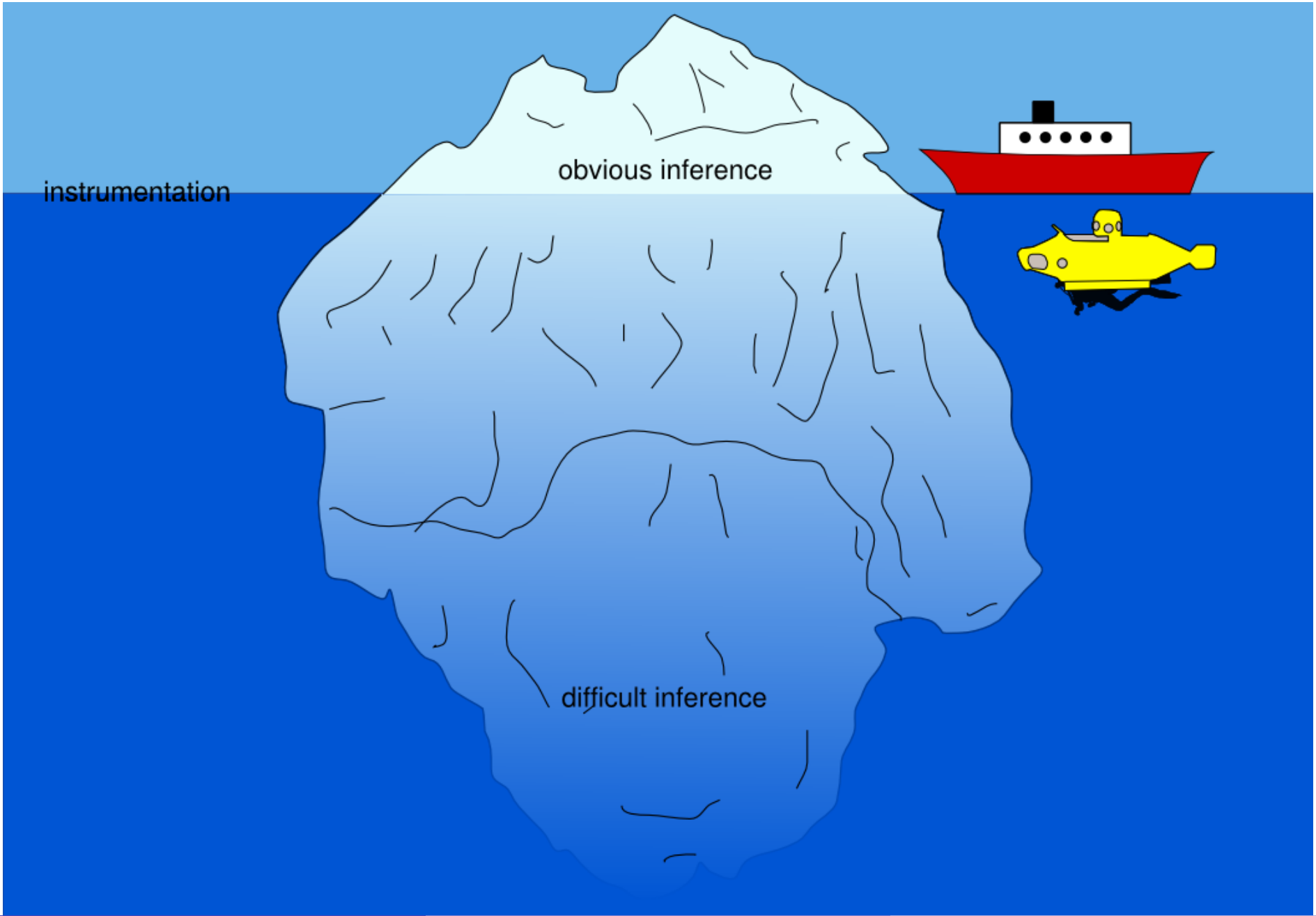
- Nice ACE2 Stories in Results
  - ACE2 four most significant targets: CTS1, DSE1, DSE2, SCW11.
    - ★ Evidence to back this up comes from CO data base.
    - ★ CTS1 relationship is known.
    - ★ DSE1 and DSE2 are involved in cell wall degradation causing daughter to seperate from parent.
    - ★ SCW11's function is unclear but protein is localised at cell wall.
  - Negative regulation of NCE4
    - ★ Wasn't documented — but now Google search leads to us!!



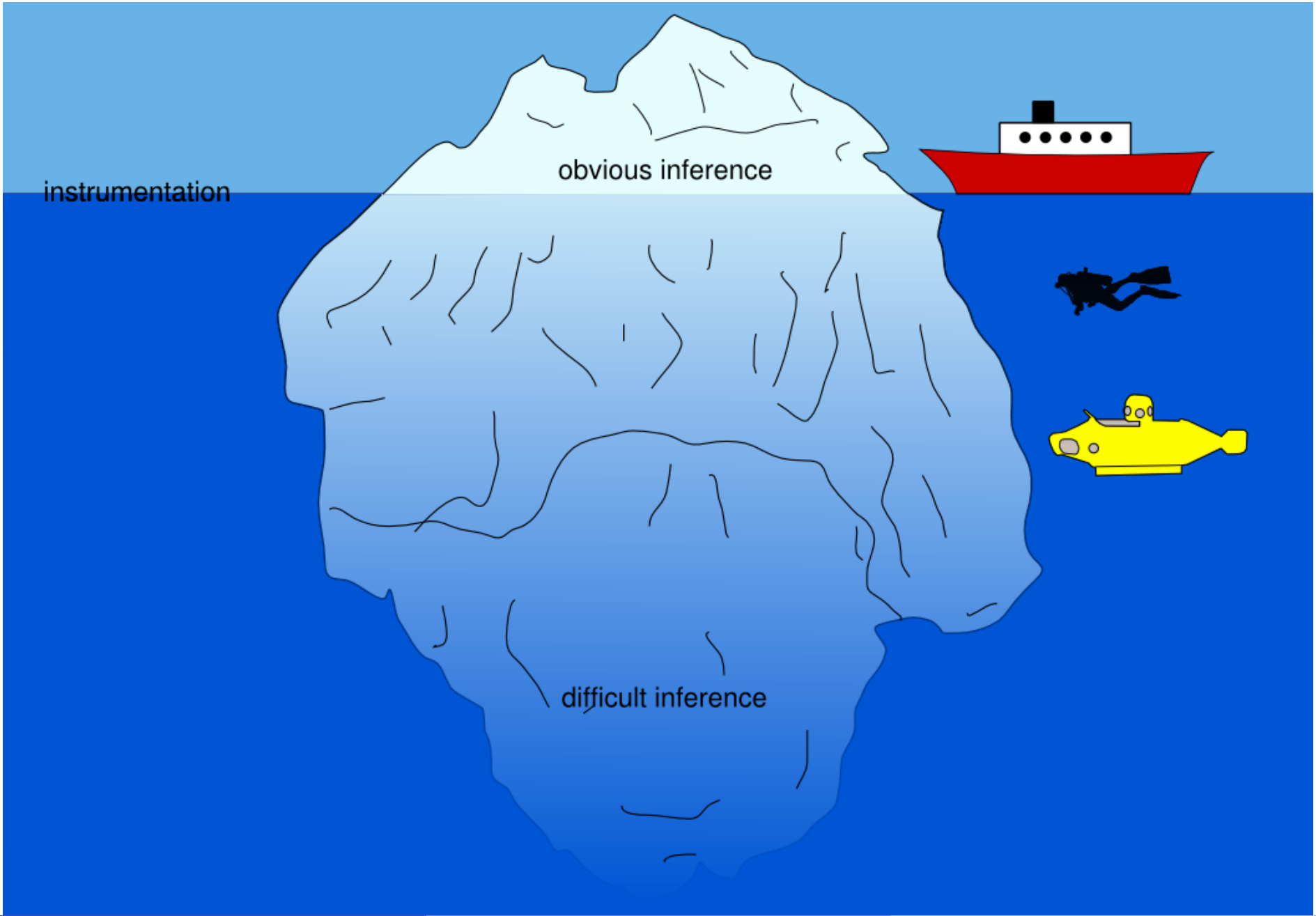
The Iceberg of Information



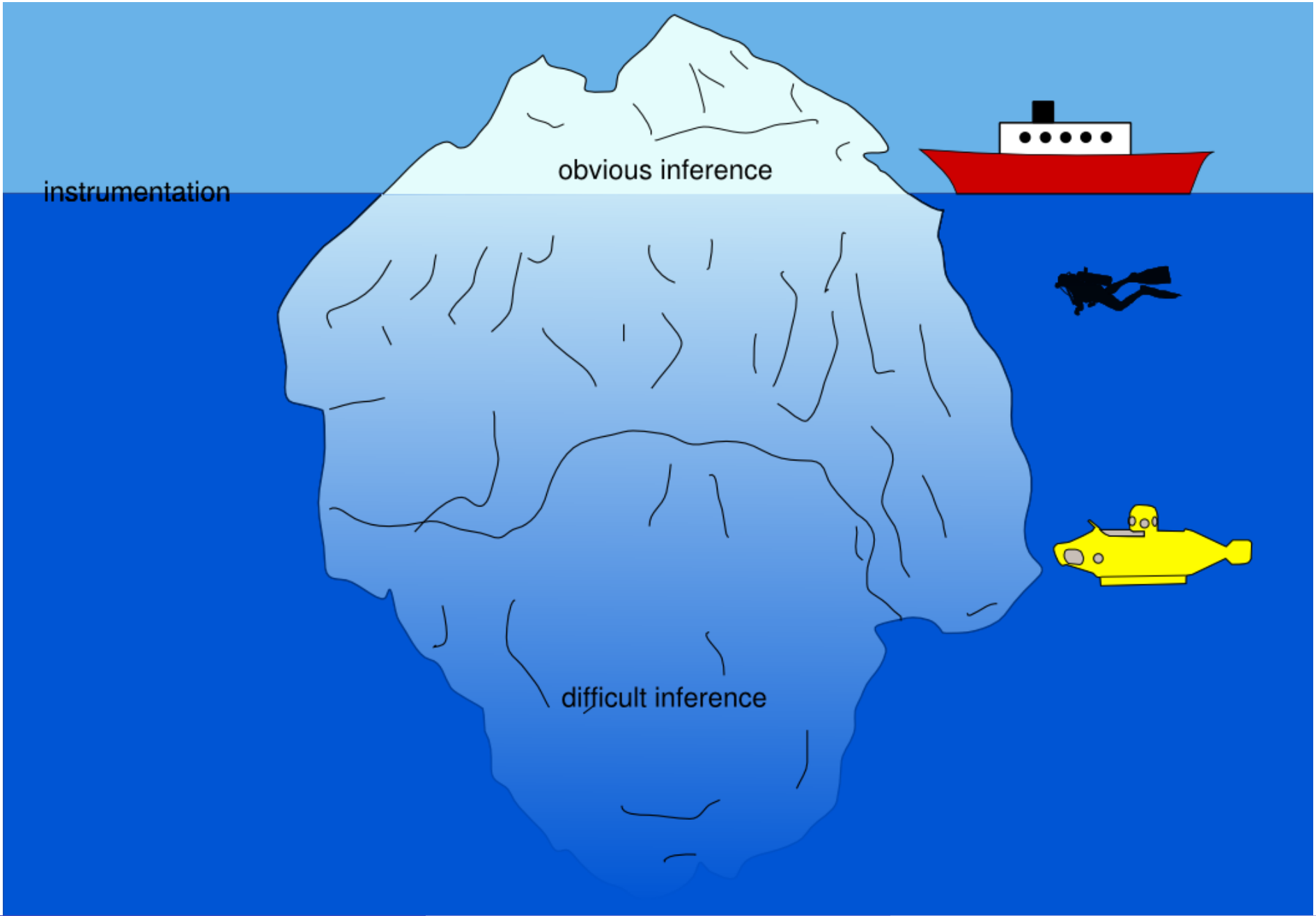
The Iceberg of Information



The Iceberg of Information



The Iceberg of Information



More Complex Model

- Complex Models on Small Networks
  - Simple linear models allow genome wide analysis of TFAs.
  - We now consider a more complex model on a much smaller network.
  - Differential Equation model
    - ⋆ Simple linear model differential equation model recently used by Barenco et al. [2006].
    - ⋆ Our inference methodology differs from theirs.

Differential Equation Model

$$\frac{dy_i(t)}{dt} = B_i + S_i f(t) - D_i y_i(t)$$

where:

$y_i(t)$	—	expression of the ith gene at time t.
$f(t)$	—	concentration of the transcription factor at time t.
D_i	—	gene's decay rate.
B_i	—	basal transcription rate.
S_i	—	sensitivity to the transcription factor.

- p53 is an tumour repressor.
  - Many targets of p53 are not shared with other TFs.
  - Consider more complex model in the simple p53 network.

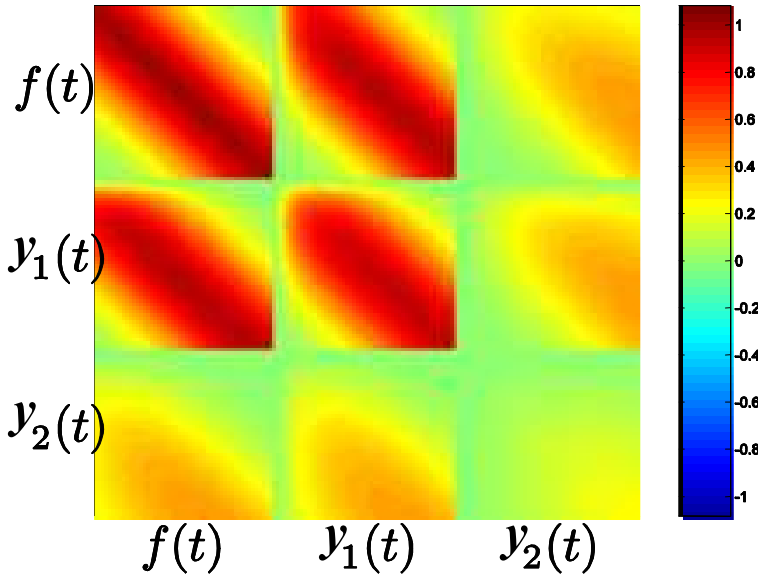
Covariance for Transcription Model

- RBF Kernel function for $f(t)$

$$y_i(t) = \frac{B_i}{D_i} + S_i \exp(-D_i t) \int_0^t f(u) \exp(D_i u)\, du.$$

- Joint distribution for $x_1(t)$, $x_2(t)$ and $f(t)$.
- Here:

D_1	S_1	D_2	S_2
5	5	0.5	0.5



Joint Sampling of $y(t)$ and $f(t)$ from Covariance

`gpsimTest`

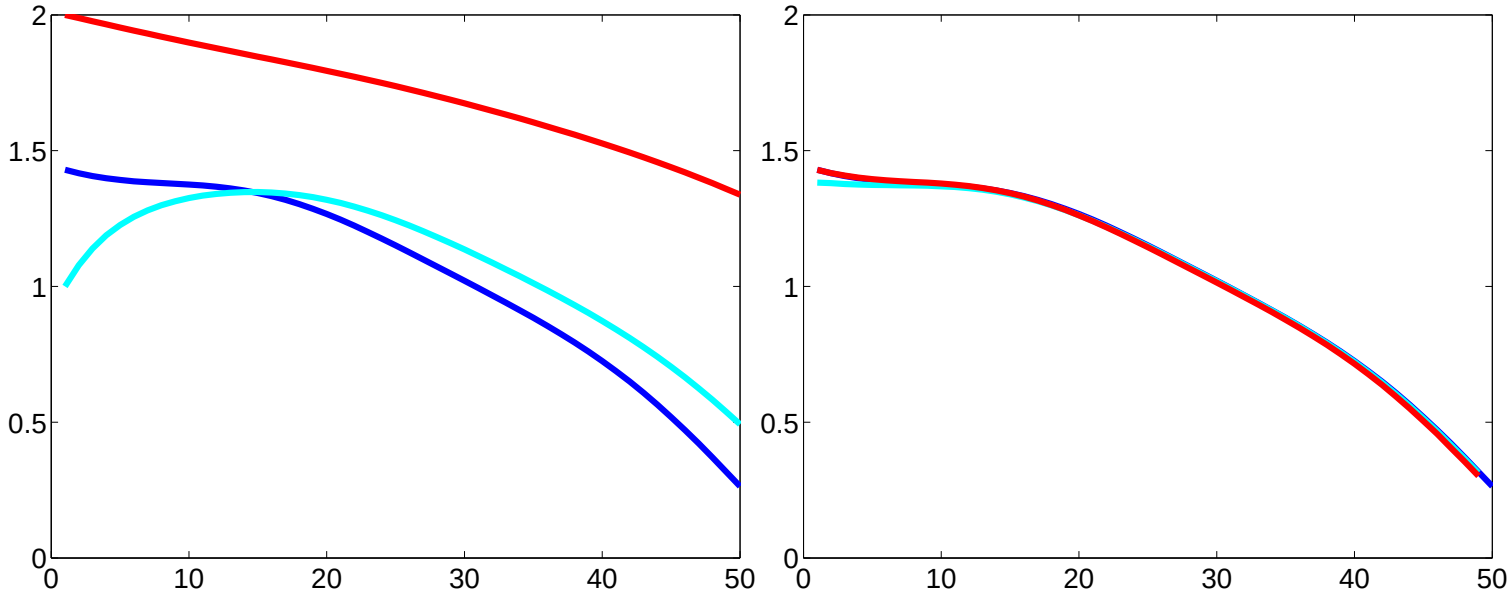


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Joint Sampling of $y(t)$ and $f(t)$ from Covariance

`gpsimTest`

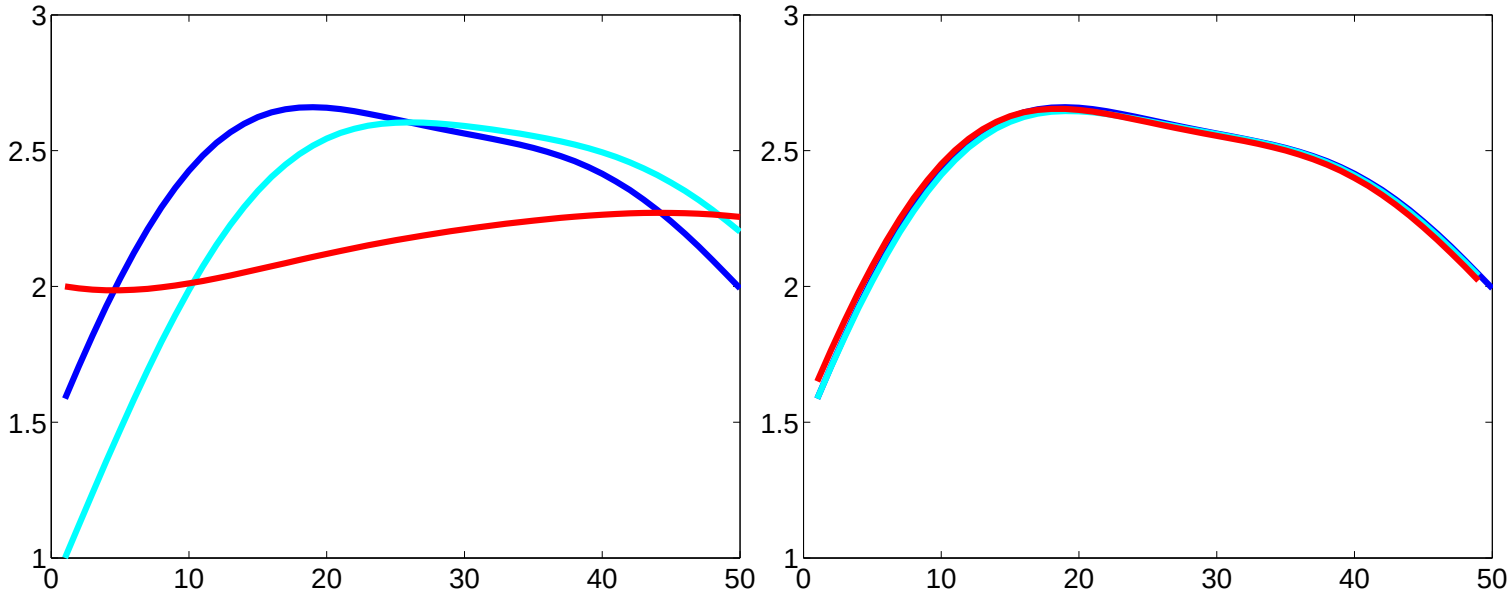


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Joint Sampling of $y(t)$ and $f(t)$ from Covariance

`gpsimTest`

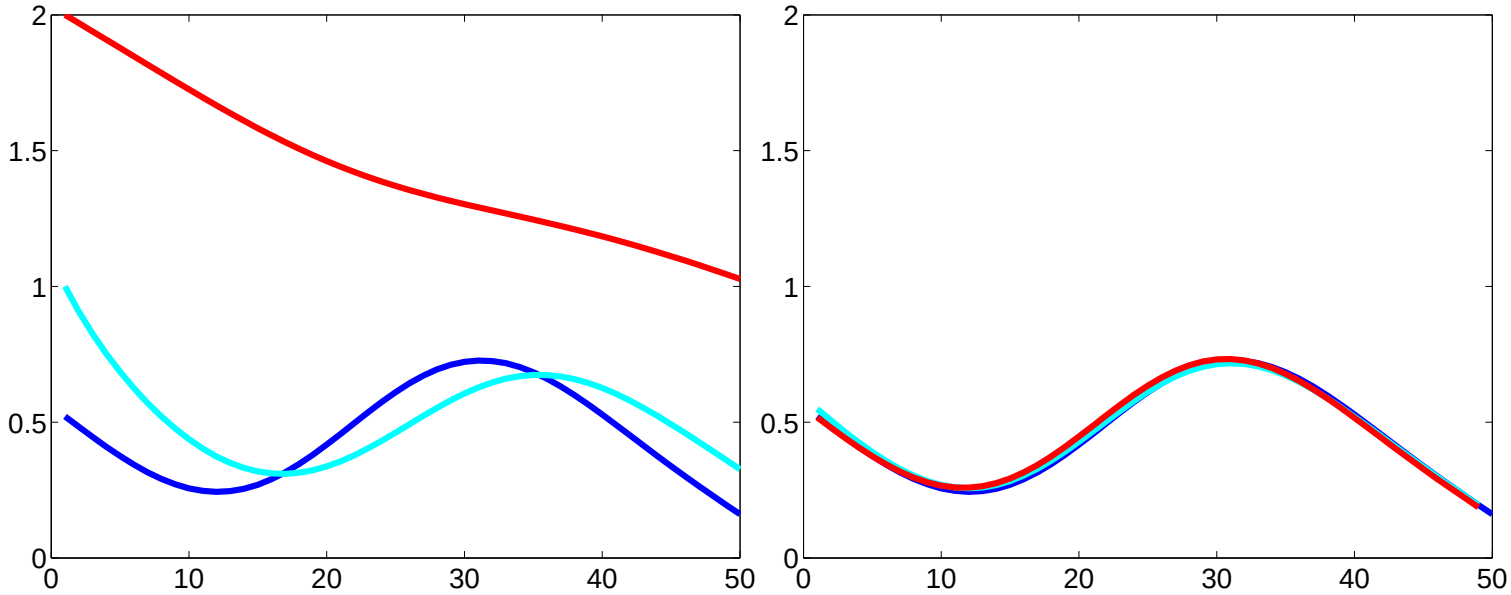


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Results — Transcription Rates

- Estimation of Equation Parameters `demBarenco1`

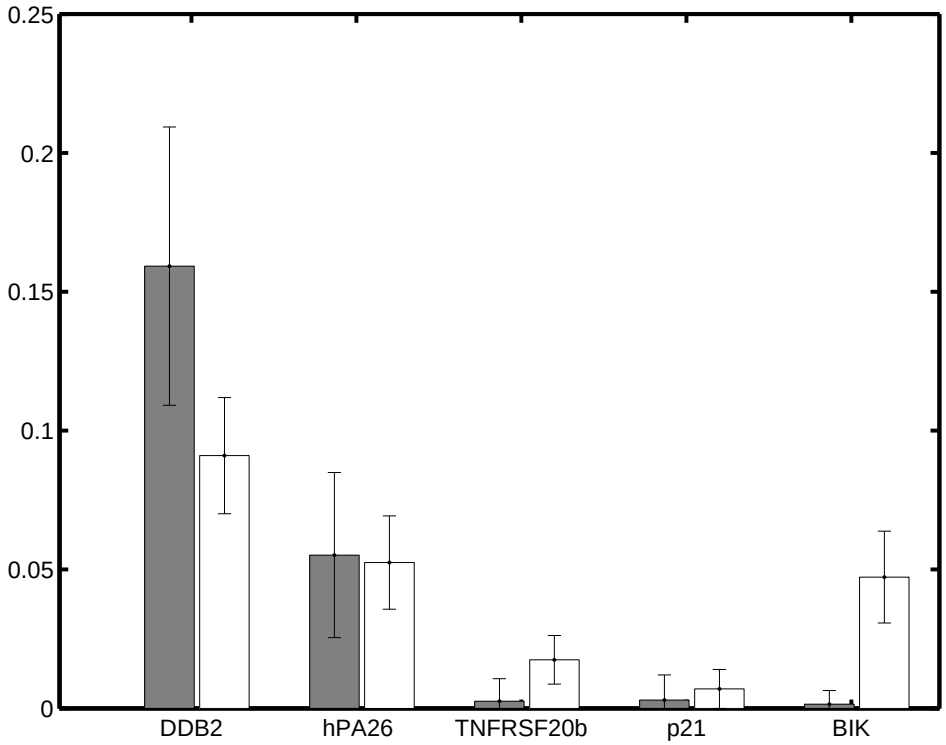


Figure: Basal transcription rates. Our results (black) compared with Barenco et al. [2006] (white).

Results — Transcription Rates

- Estimation of Equation Parameters `demBarenco1`

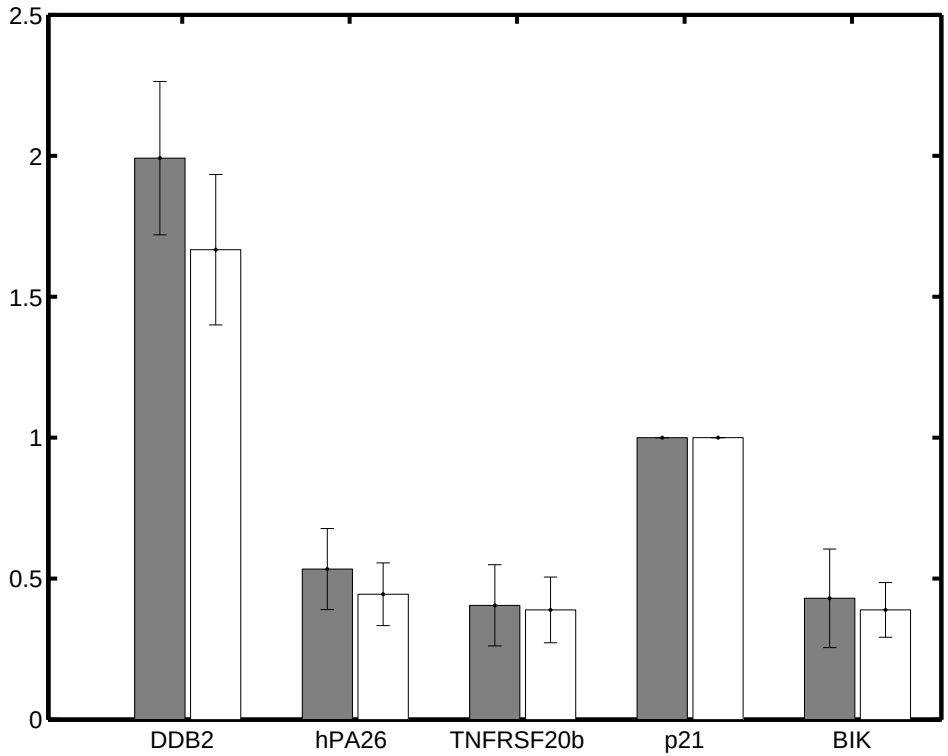


Figure: Sensitivities. Our results (black) compared with Barenco et al. [2006] (white).

Results — Transcription Rates

- Estimation of Equation Parameters `demBarenco1`

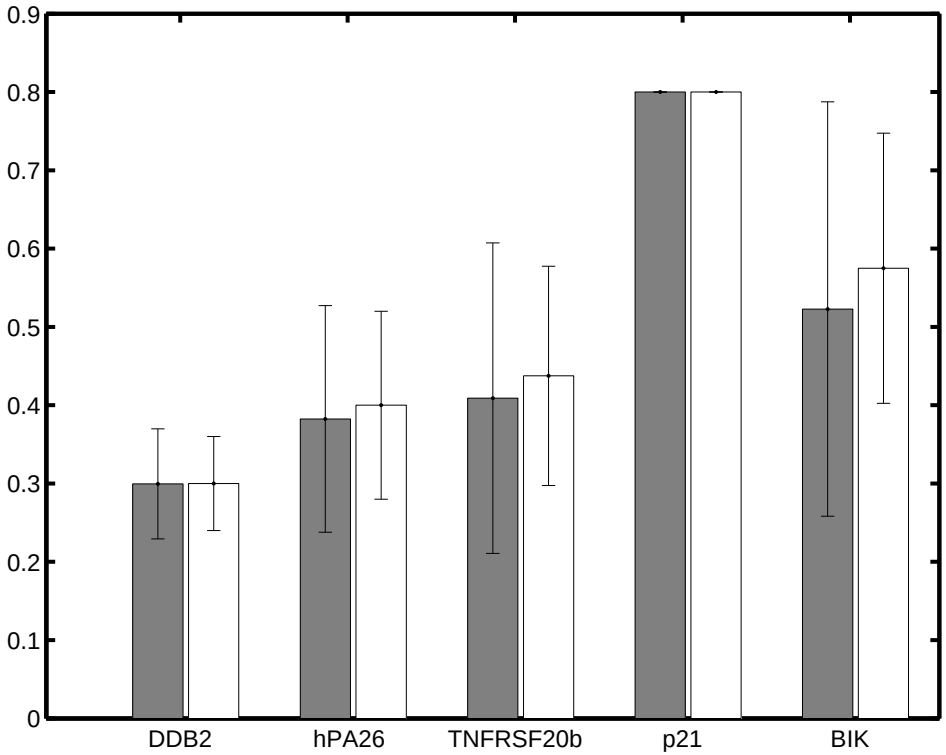
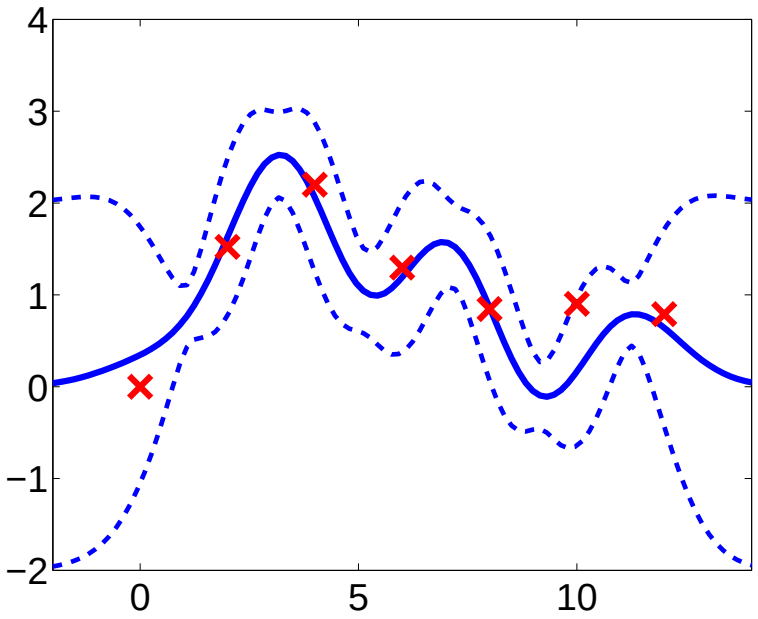


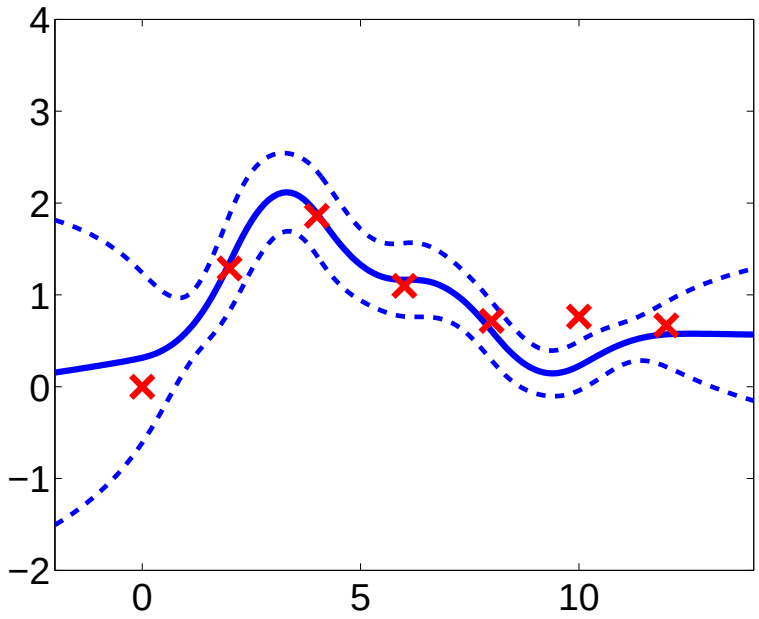
Figure: Decays. Our results (black) compared with Barenco et al. [2006] (white).

Results — Protein Concentration

- Prediction with error bars of protein concentration: $p\left(\mathbf{f}|\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\right)$



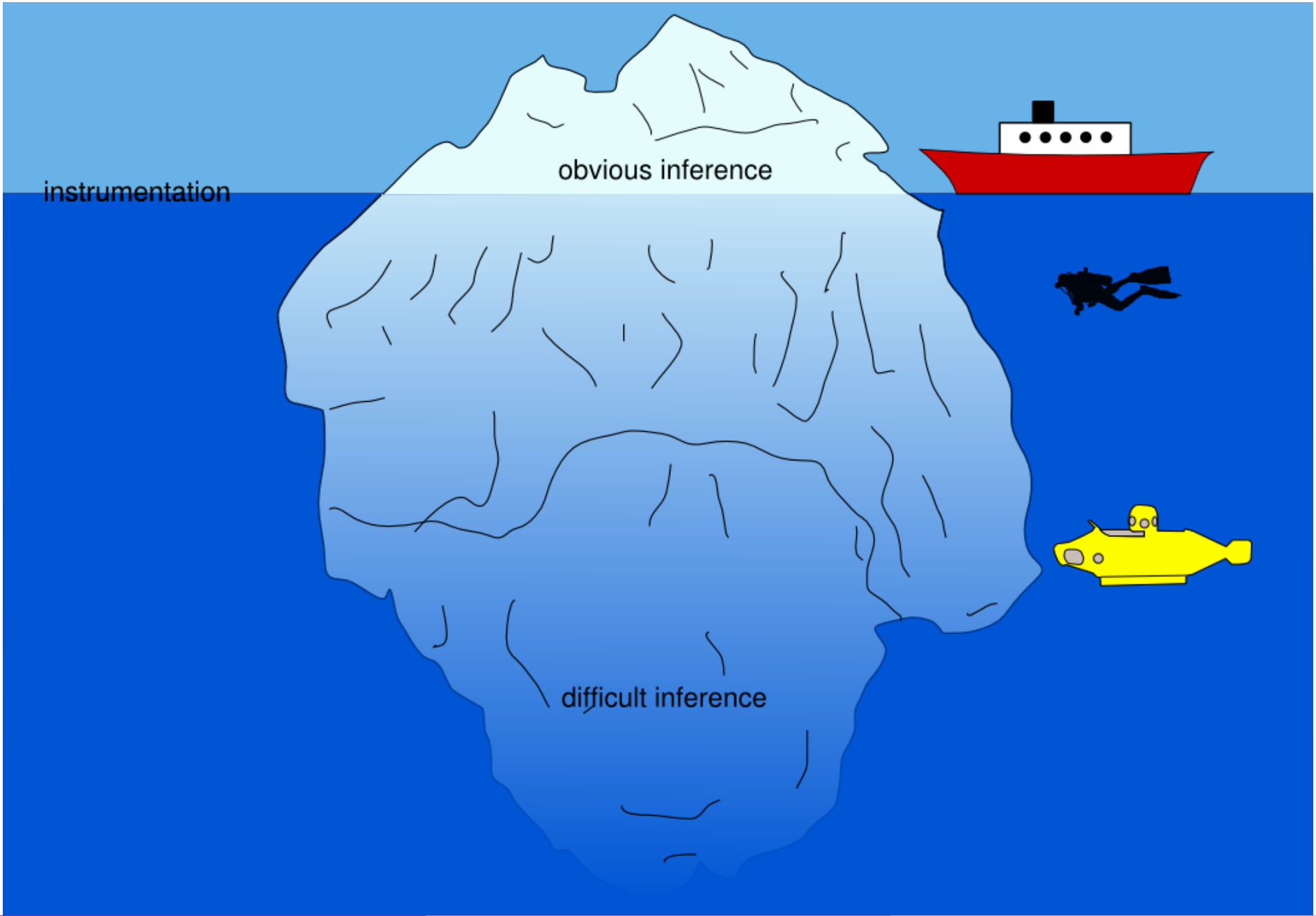
(a)



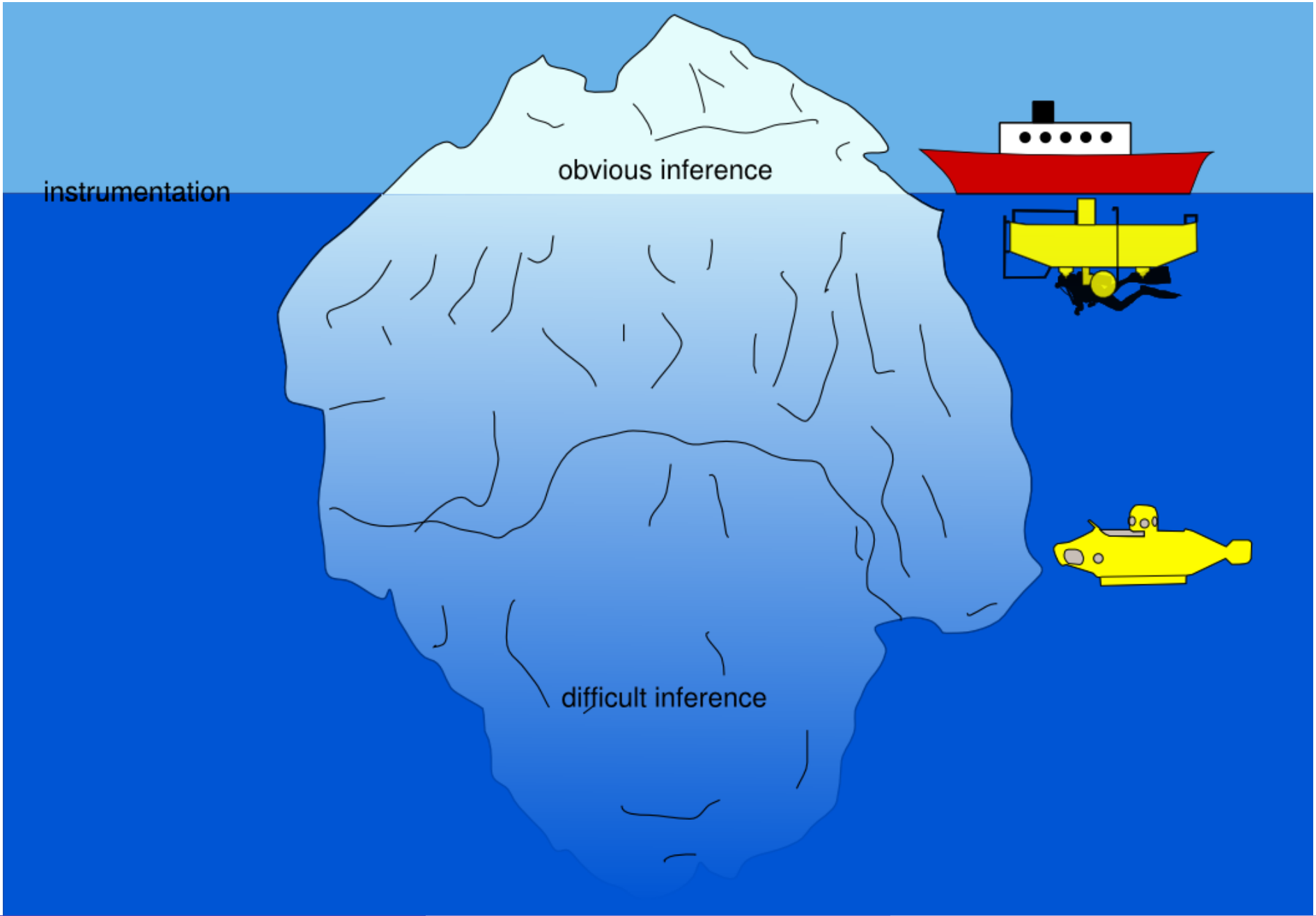
(b)

Figure: (a) RBF covariance function (b) MLP covariance function. Also included are results from Barenco et al. [2006] as crosses.

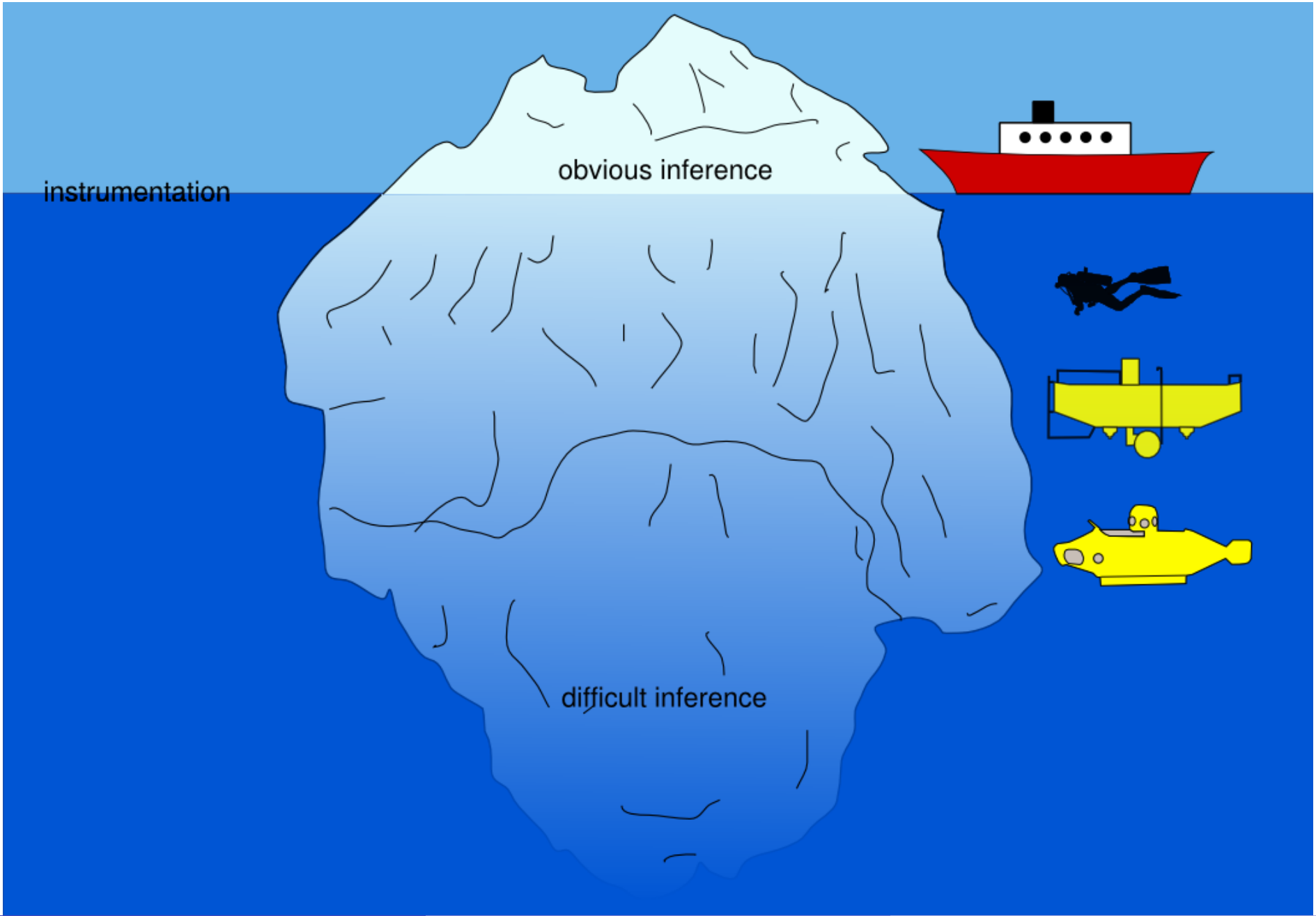
The Iceberg of Information



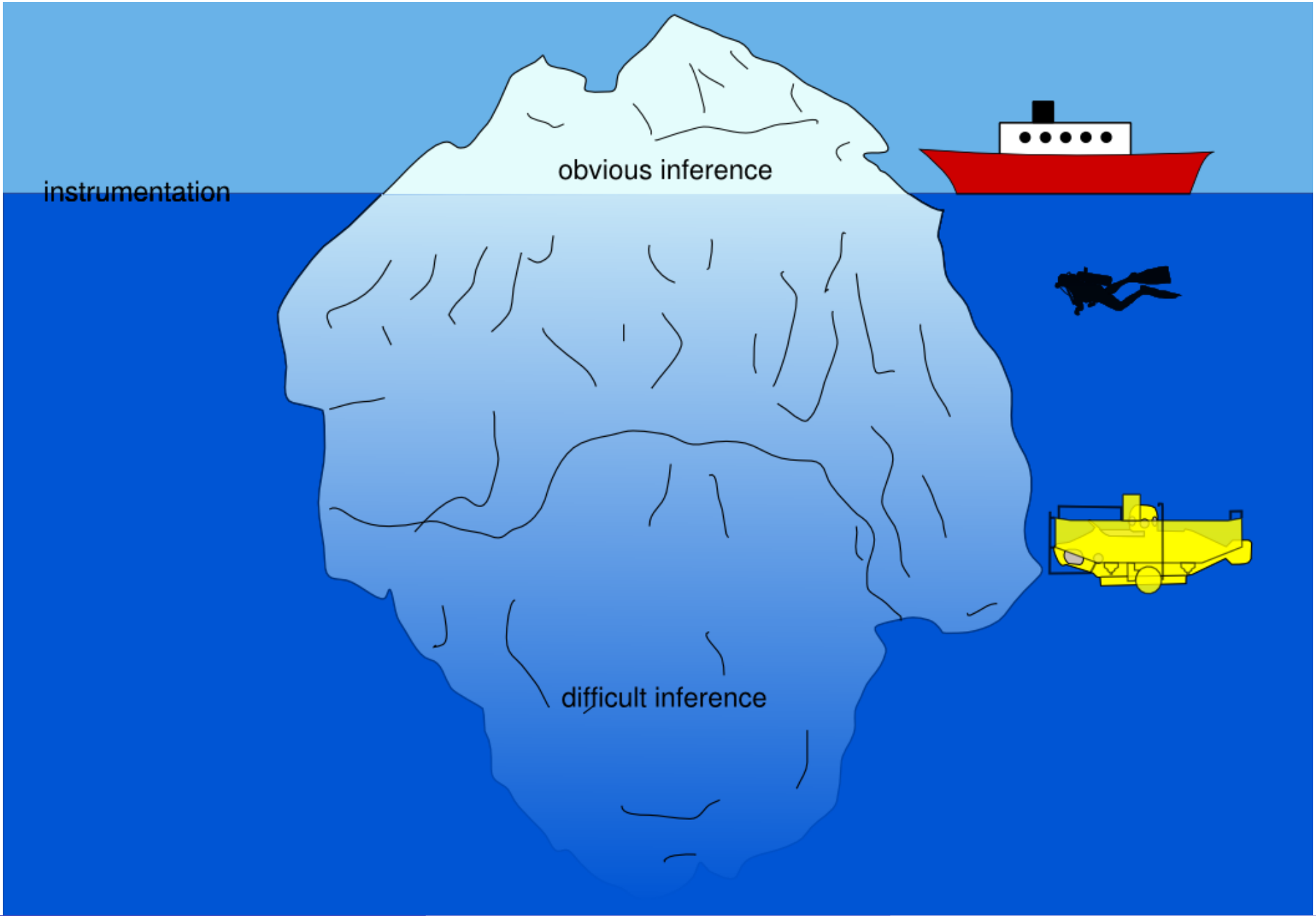
The Iceberg of Information



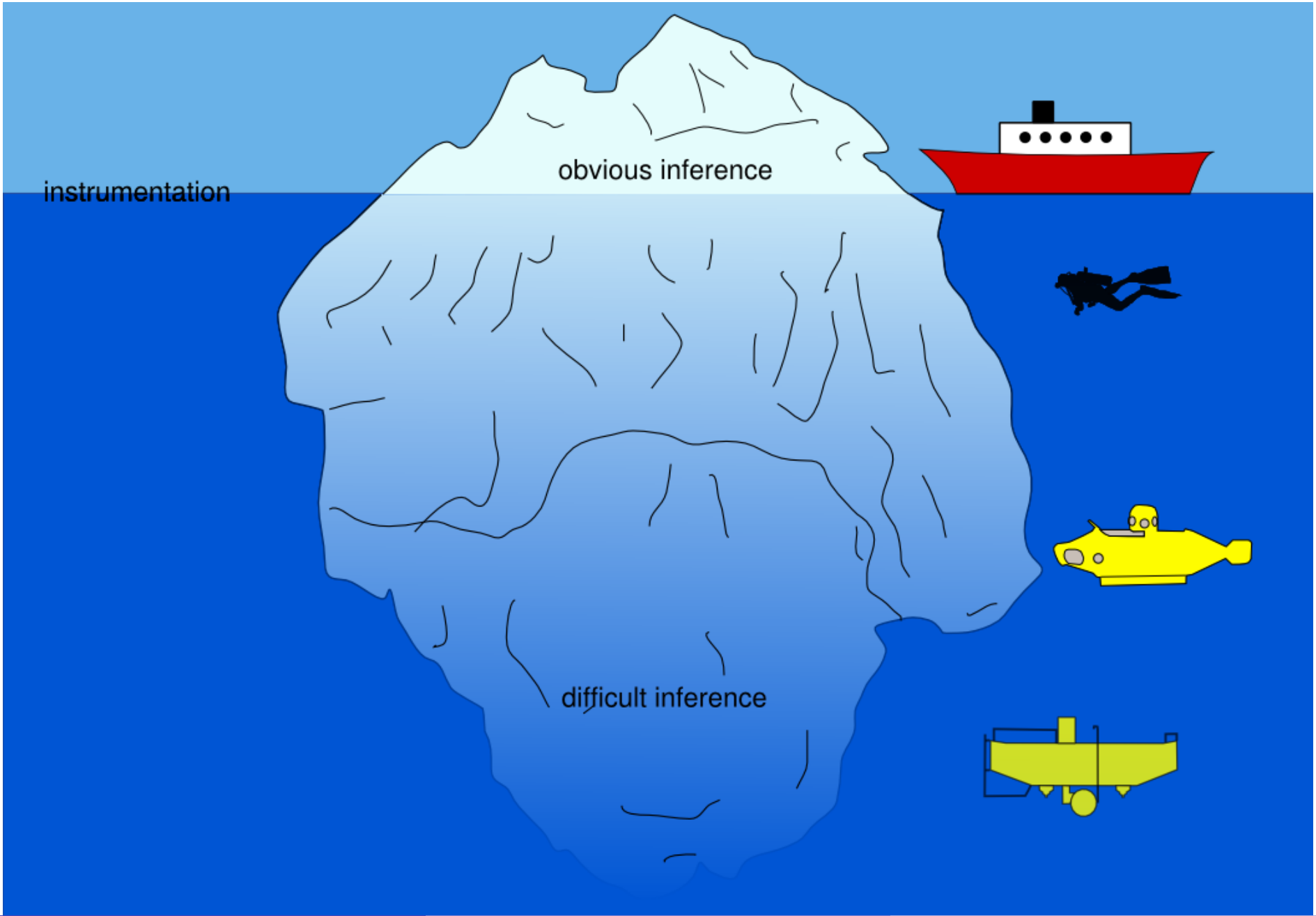
The Iceberg of Information



The Iceberg of Information



The Iceberg of Information



Summary

- PUMA: Propagation of Uncertainty in Microarray Analysis
  - Level of Noise in the Array can be Assesed (gMOS methods).
  - Probabilistic Models can:
    - ⋆ Improve selection of over-expressed genes (PPLR).
    - ⋆ Clean up gene expression profiles (NPPCA).
  - Simple (log-linear) probabilistic models can be used with network connectivity data to
    - ⋆ To infer *genome wide* transcription factor activities (chipdyno).
    - ⋆ To infer *genome wide* transcription factor protein concentrations (chipvar).
- Differential equation models
  - Deal with latent species using Gaussian processes.
  - Structural inference with Thermodynamic Intergration.

Acknowledgements

- Inspiration:
  - Martino Barenco, Mark Girolami, Mike Hubank, Dirk Husmeier, Andrew Millar, Nick Monk, Magnus Rattray
- Perspiration:
  - Investigators
    - ★ Neil Lawrence and Magnus Rattray
  - gMOS family of Methods and PPLR
    - ★ Xuejun Liu (ex PhD student) and Marta Milo (Wellcome Fellow)
  - Uncertainty Propagation through PCA
    - ★ Marta Milo (Wellcome Fellow), Richard Pearson (PhD student) and Guido Sanguinetti (ex post-doc)
  - Inference of Transcription Factor Activities
    - ★ Pei Gao (current post-doc), Michalis Titsias (new post-doc), Guido Sanguinetti (ex post-doc)
- Funding
  - BBSRC Grant No BBS/B/0076X (with Magnus)
  - EPSRC Grant No EP/F005687/1 (with Magnus, Johannes Yaeger and Nick Monk)

References

O. Alter and G. H. Golub. Integrative analysis of genome-scale data using pseudoinverse projection predicts novel correlation between dna replication and rna transcription. *Proceedings of the National Academy of Sciences USA*, 101(47): 16577–16582, 2004.

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006.

A.-L. Boulesteix and K. Strimmer. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor. Biol. Med. Model.*, 2(23):1471–16582, 2005.

S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(R16), 2005.

F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5(31):1471–2105, 2004.

T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298 (5594):799–804, 2002.

J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences USA*, 100(26): 15522–15527, 2003.

K. K. Lin, D. Chudova, G. W. Hatfield, P. Smyth, , and B. Andersen. Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. 101(45):15955–15960, 2004.

X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, 2005.

X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, 2006.

M. Milo, A. Fazeli, M. Niranjan, and N. D. Lawrence. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Transations*, 31(6):1510–1512, 2003.

M. N. Rivolta, A. Halsall, C. Johnson, M. Tones, and M. C. Holley. Genetic profiling of functionally related groups of genes during conditional differentiation of a mammalian cochlear hair cell line. *Genome Research*, 12(7):1091–1099, 2002.

G. Sanguinetti, M. Milo, M. Rattray, and N. D. Lawrence. Accounting for probe-level noise in principal component analysis of microarray data. *Bionformatics*, 21(19):3748–3754, 2005.

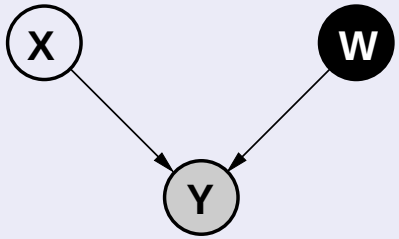
G. Sanguinetti, M. Rattray, and N. D. Lawrence. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*, 22(14):1753–1759, 2006.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization.

Probabilistic PCA

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- Latent variable approach:
  - Define Gaussian prior over *latent space*, $\mathbf{X}$.
  - Integrate out *latent variables*.



$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

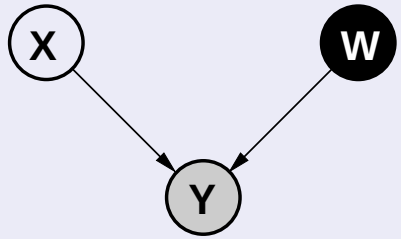
$$p(\mathbf{X}) = \prod_{i=1}^{n} N\left(\mathbf{x}_{i,:}|\mathbf{0}, \mathbf{I}\right)$$

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I}\right)$$

Probabilistic PCA

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- Latent variable approach:
  - Define Gaussian prior over *latent space*, $\mathbf{X}$.
  - Integrate out *latent variables*.



$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

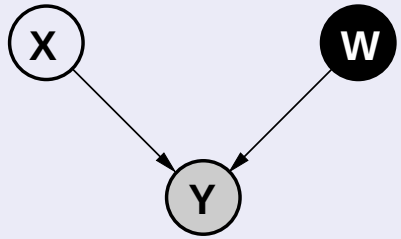
$$p(\mathbf{X}) = \prod_{i=1}^{n} N\left(\mathbf{x}_{i,:}|\mathbf{0}, \mathbf{I}\right)$$

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I}\right)$$

Probabilistic PCA

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- Latent variable approach:
  - Define Gaussian prior over *latent space*, $\mathbf{X}$.
  - Integrate out *latent variables*.



$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

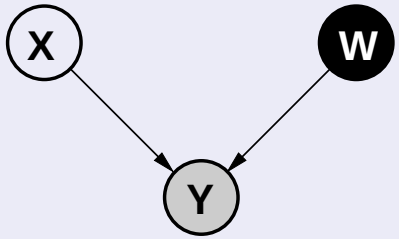
$$p(\mathbf{X}) = \prod_{i=1}^{n} N\left(\mathbf{x}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mu, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I}\right)$$

Probabilistic PCA

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- Latent variable approach:
  - Define Gaussian prior over *latent space*, $\mathbf{X}$.
  - Integrate out *latent variables*.



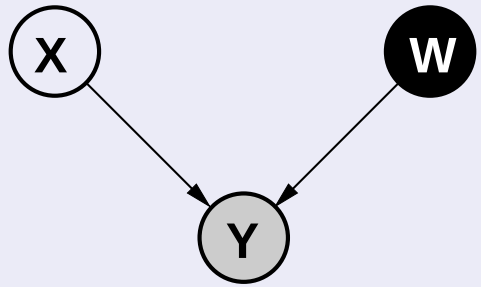
$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

$$p(\mathbf{X}) = \prod_{i=1}^{n} N\left(\mathbf{x}_{i,:}|\mathbf{0}, \mathbf{I}\right)$$

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I}\right)$$

Probabilistic PCA II

Probabilistic PCA Max. Likelihood Soln [Tipping and Bishop, 1999]



$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I}\right)$$

Probabilistic PCA II

Probabilistic PCA Max. Likelihood Soln [Tipping and Bishop, 1999]

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{W}\right) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\tilde{\mathbf{Y}}^{\mathsf{T}}\tilde{\mathbf{Y}}\right) + \text{const.}$$

Where $\tilde{\mathbf{Y}}$ is the matrix $\mathbf{Y}$ with $\boldsymbol{\mu}$removed. If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\tilde{\mathbf{Y}}^{\mathsf{T}}\tilde{\mathbf{Y}}$ and the corresponding eigenvalues are Λ_q,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{V}^{\mathsf{T}}, \quad \mathbf{L} = \left(\Lambda_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

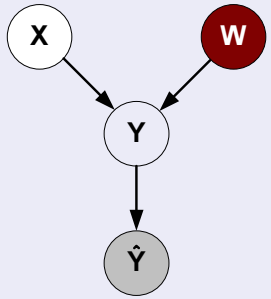
where $\mathbf{V}$ is an arbitrary rotation matrix.

$$\boldsymbol{\mu} = n^{-1}\sum_{i=1}^{n}\mathbf{y}_{i,:}$$

Heteroschedastic Probabilistic PCA

Heteroschedastic PPCA

- Define *linear-Gaussian relationship* between latent variables and $\mathbf{Y}$.
- Define a *further* Gaussian relationship to corrupted profiles $\hat{\mathbf{Y}}$.
  - $\mathbf{D}_i$ is a diagonal matrix of estimated variances.
- Integrate out *latent variables*.



$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}+\boldsymbol{\mu},\sigma^{2}\mathbf{I}\right)$$

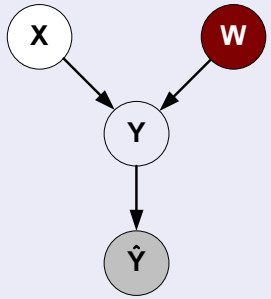
$$p\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:}\right)=N\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:},\mathbf{D}_{i}\right)$$

$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\mathsf{T}}+\sigma^{2}\mathbf{I}+\mathbf{D}_{i}\right)$$

Heteroschedastic Probabilistic PCA

Heteroschedastic PPCA

- Define *linear-Gaussian relationship* between latent variables and $\mathbf{Y}$.
- Define a *further* Gaussian relationship to corrupted profiles $\hat{\mathbf{Y}}$.
  - $\mathbf{D}_i$ is a diagonal matrix of estimated variances.
- Integrate out *latent variables*.



$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}+\boldsymbol{\mu},\sigma^{2}\mathbf{I}\right)$$

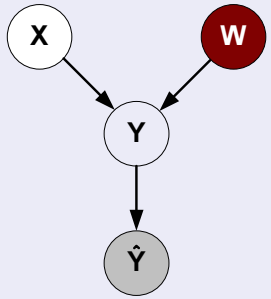
$$p\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:}\right)=N\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:},\mathbf{D}_{i}\right)$$

$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\mathsf{T}}+\sigma^{2}\mathbf{I}+\mathbf{D}_{i}\right)$$

Heteroschedastic Probabilistic PCA

Heteroschedastic PPCA

- Define *linear-Gaussian relationship* between latent variables and $\mathbf{Y}$.
- Define a *further* Gaussian relationship to corrupted profiles $\hat{\mathbf{Y}}$.
  - $\mathbf{D}_i$ is a diagonal matrix of estimated variances.
- Integrate out *latent variables*.



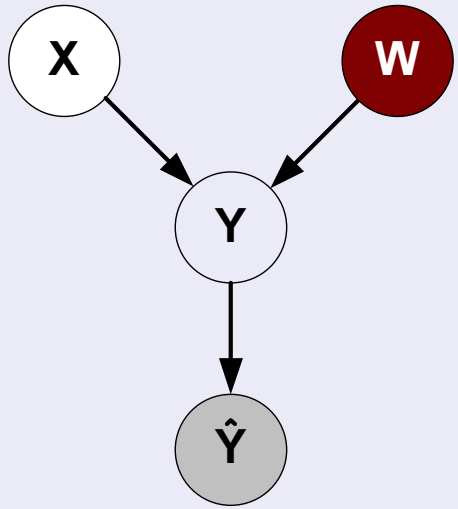
$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}+\boldsymbol{\mu},\sigma^{2}\mathbf{I}\right)$$

$$p\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:}\right)=N\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:},\mathbf{D}_{i}\right)$$

$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\mathsf{T}}+\sigma^{2}\mathbf{I}+\mathbf{D}_{i}\right)$$

Heteroschedastic PPCA II

Heteroschedastic PPCA Max. Likelihood Soln [Sanguinetti et al., 2005]



$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mu, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I} + \mathbf{D}_i\right)$$

Heteroschedastic PPCA II

Heteroschedastic PPCA Max. Likelihood Soln [Sanguinetti et al., 2005]

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I} + \mathbf{D}_i\right)$$

- Can no longer solve via eigenvalue problem.
- We use an EM algorithm.
  - A major problem is the strong correlation between $\mathbf{W}$ and $\boldsymbol{\mu}$.
  - We use some tricks to speed up convergence.
- Software available in R and MATLAB.