Feature and Variable Selection

Context for Representation Learning in RL

Patrick M. Pilarski — RLAI & AICML, University of Alberta

5th Barbados Workshop on Reinforcement Learning, April 6th 2010

Feature Selection in Brief

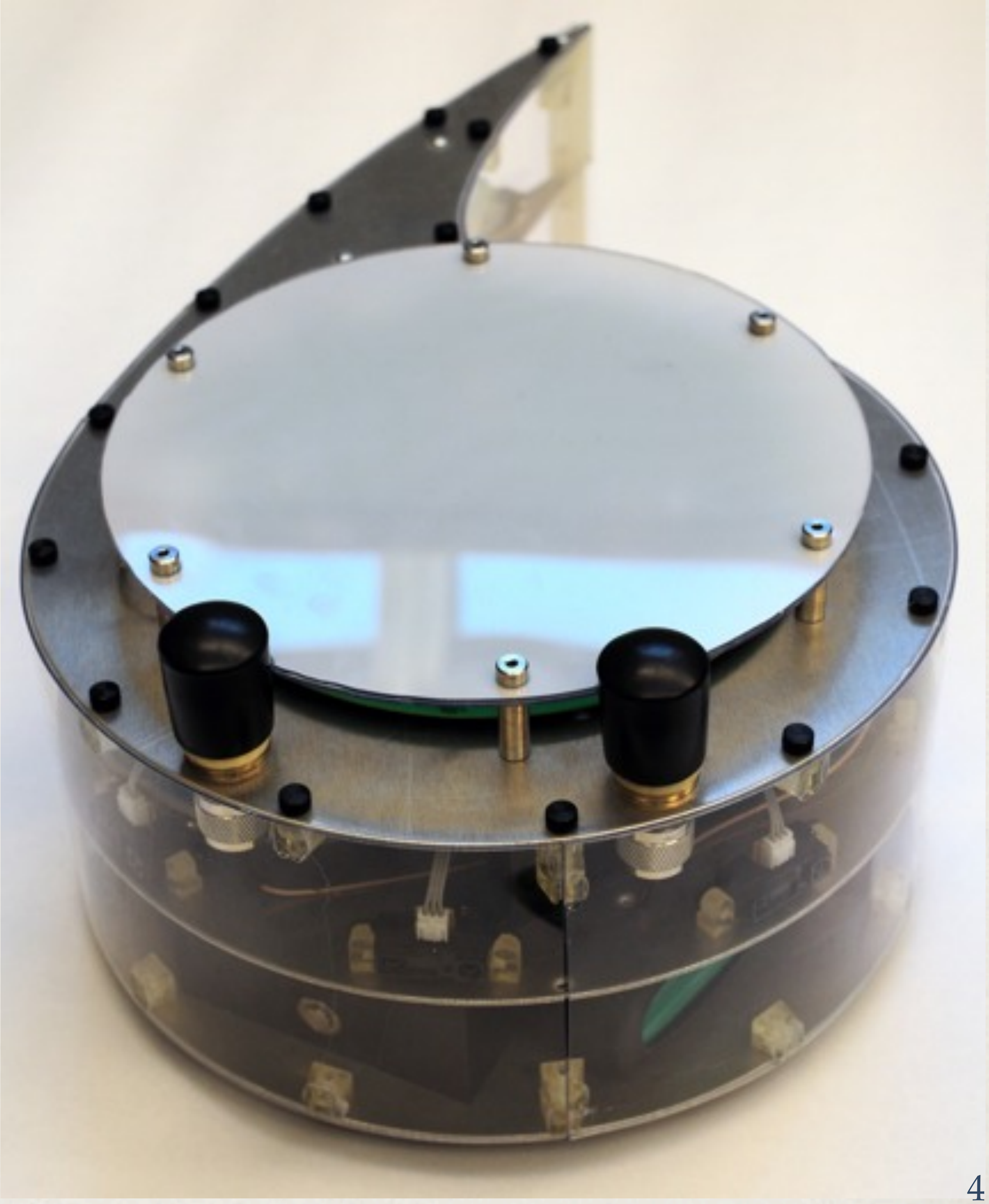
- House painting analogy...
- **The Basic Idea:**
  - Many possible inputs for use in learning, control, and knowledge discovery.
  - How to decide which inputs contain valuable information?
  - How to choose which inputs should be used for a given task (and how)?

Feature Selection Applications

- **Gene array data processing and drug discovery** *(Guyon et al. 2003)*
  - Thousands of genes, limited labeled patient samples... need to determine how gene expression corresponds to disease, or produces proteins that can be targeted with new drugs.
- **Text classification** from a bag-of-words with over 10k variables.
- **Biomedical image analysis** *(Pilarski et al., Proc. SPIE, 2009)*

Relationship to RL

- Many inputs, e.g. sensors, each with unknown value to learning and system operation.
- The features used to describe *state* can have a dramatic impact on performance (as will be discussed in the next talk.)
- Advantageous to have ways for a learner to automatically evaluate and compare features and combinations of features.



Outline

- **Part 1:** Common terminology, definitions, and examples from the feature selection literature.
    - *Variables* and *Features*
    - *Relevance*, *Redundancy*, and *Correlation*
    - *Subset* v.s. *Ranking* methods, *Filters* and *Wrappers*
    - *Feature Construction*
- **Part 2:** Feature selection on the RLAI Critterbot.
- **Part 3:** Conclusions and a summary of core concepts.

Part 1

Common Terminology and Definitions in Feature Selection

The Nature of "The Input"

- **Variable:** a raw input signal.
  - *x1, x2, x3, ...*
  - *e.g.* raw 8-bit sensor signals, line voltage or current measurements
- **Feature:** a processed version a variable (or combination of variables).
  - f(*x1*) , f(*x1*,*x2*,...)
  - *e.g.* norm(*x1*), avg(*x1*), log(*x1*+*x3*), max[FFT(*x1*) | *w*>*1e6*]

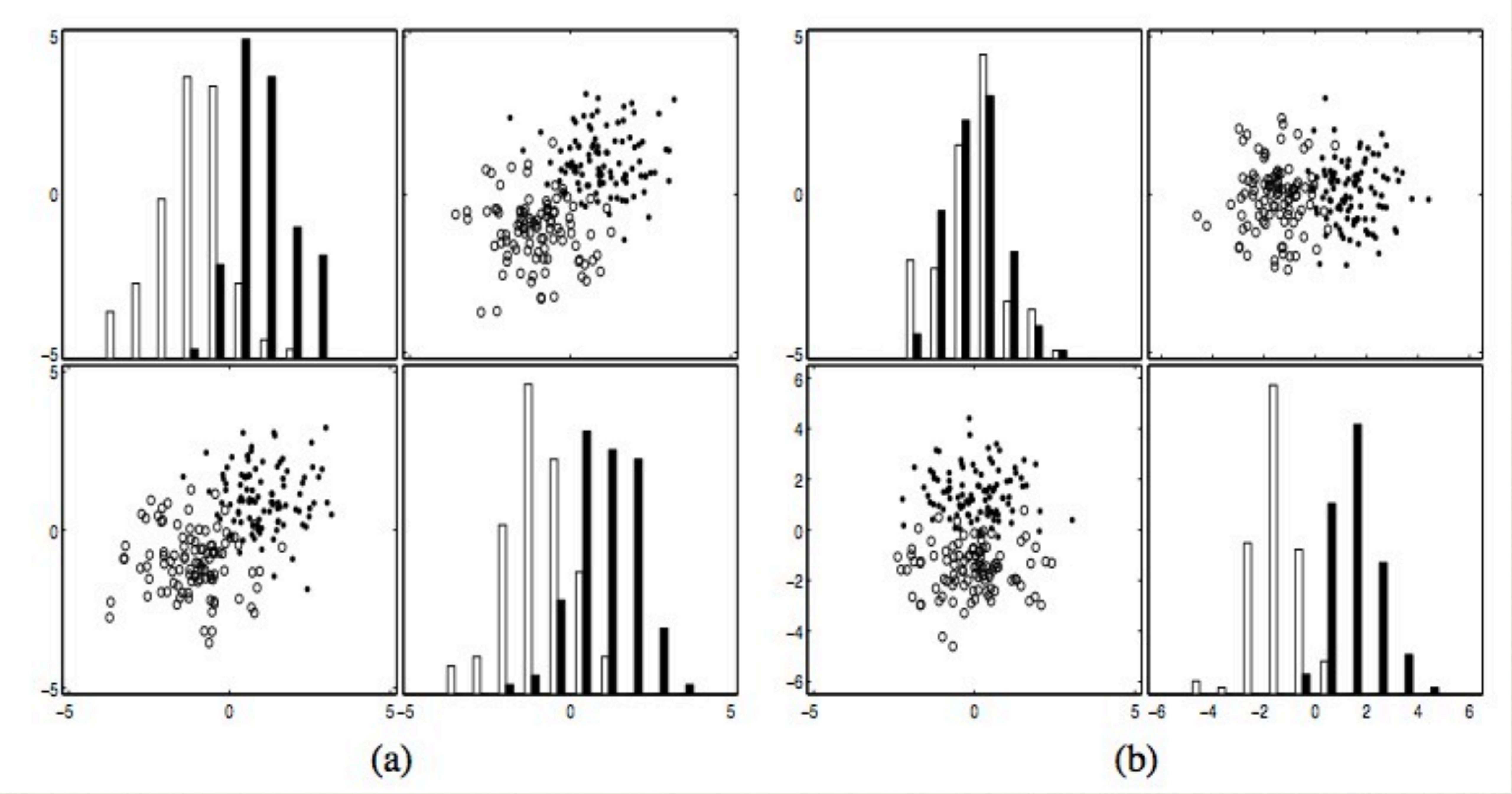
Guyon and Elisseeff, JMLR, Vol. 3, 2003.

Redundancy and Correlation

- **Can redundant variables be used to improve performance?**
  - *Yes:* noise reduction and improved class separation can be achieved by combining two presumably redundant variables.

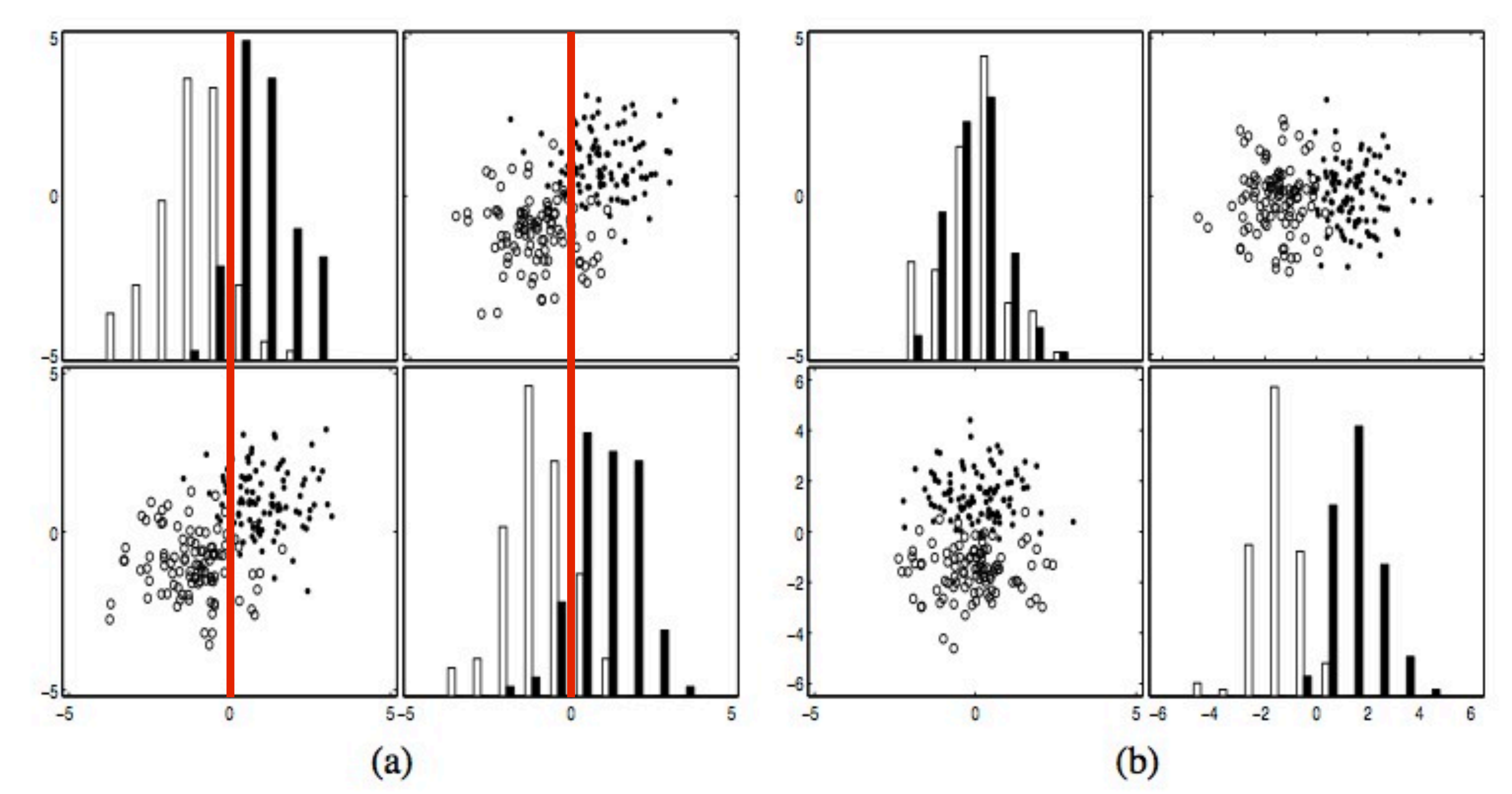
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



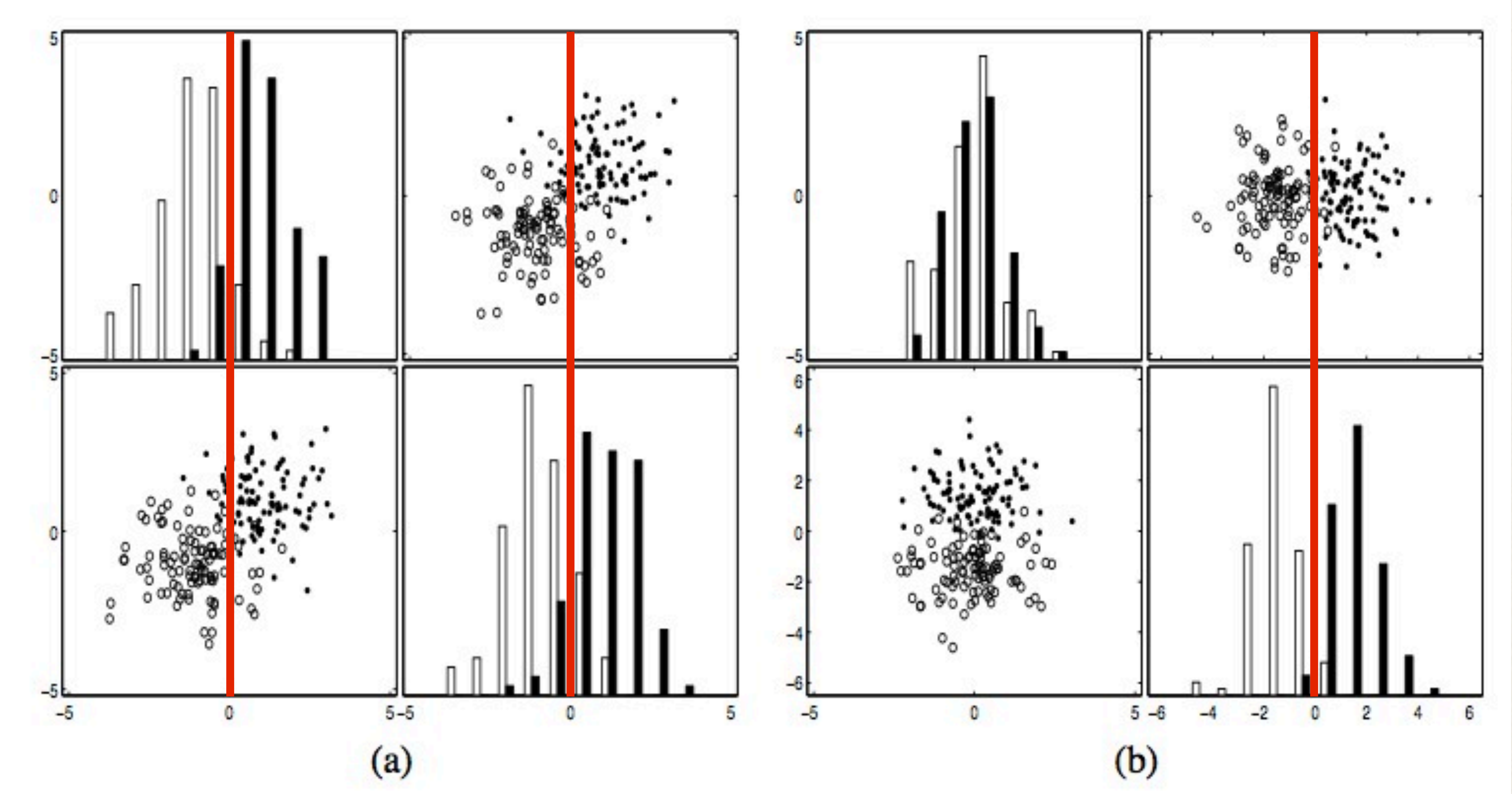
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



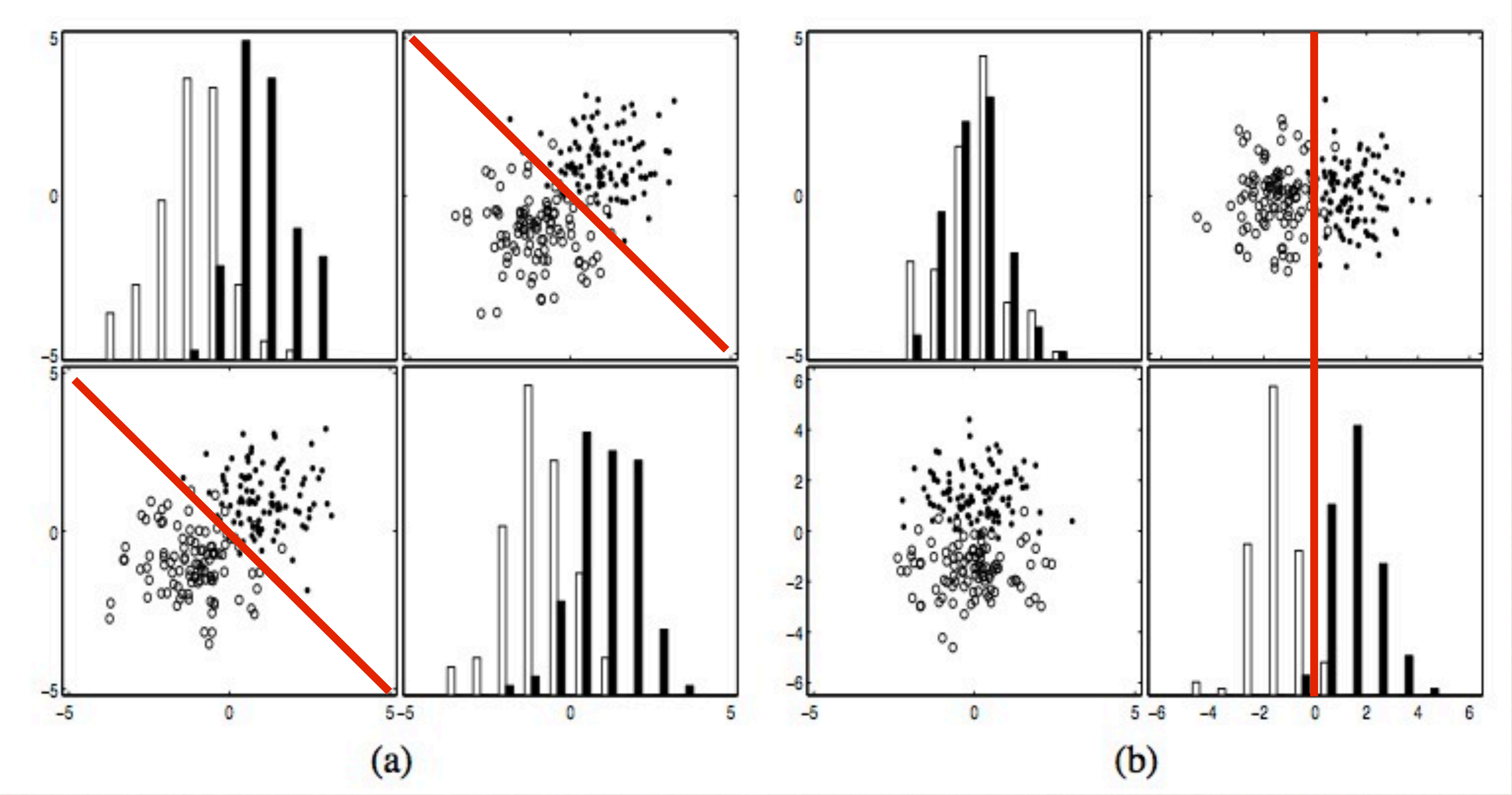
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



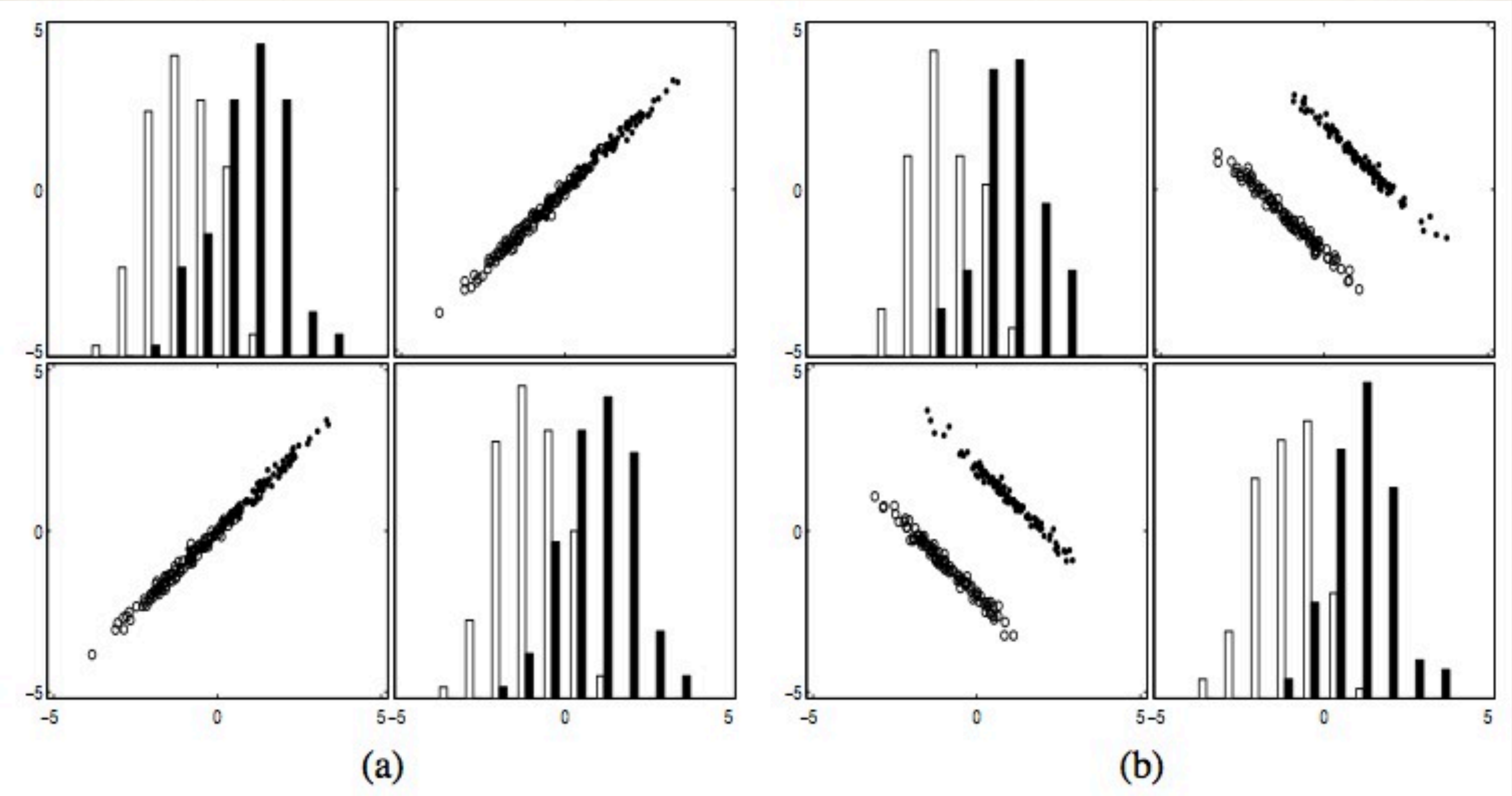
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation

- **What are the relationships between correlation and redundancy?**
  - *Perfectly correlated* variables are truly redundant in that no extra information is gained by having them.
  - However, *very high correlation* (or anti-correlation) does not mean that variables are not complementary.

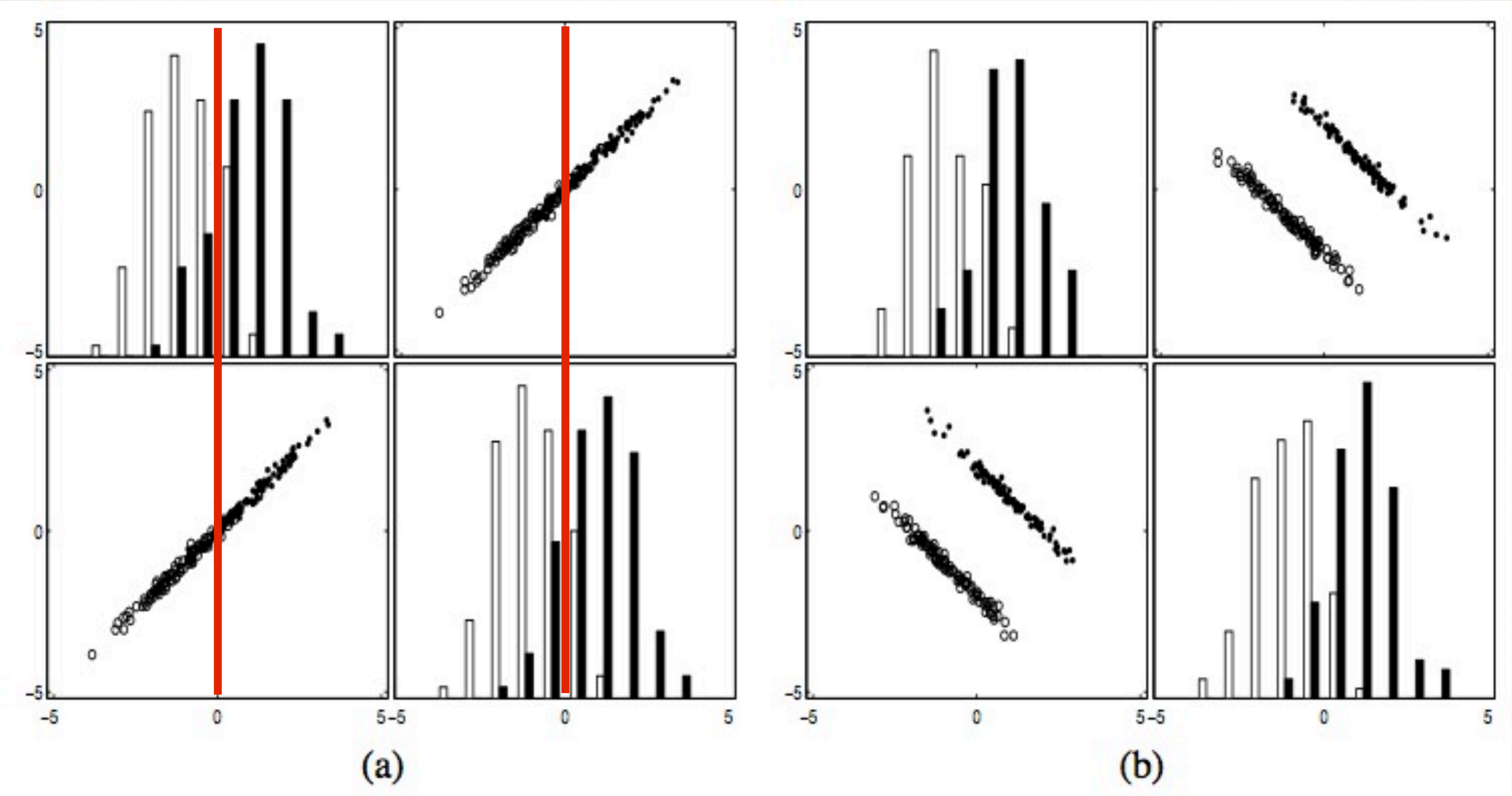
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



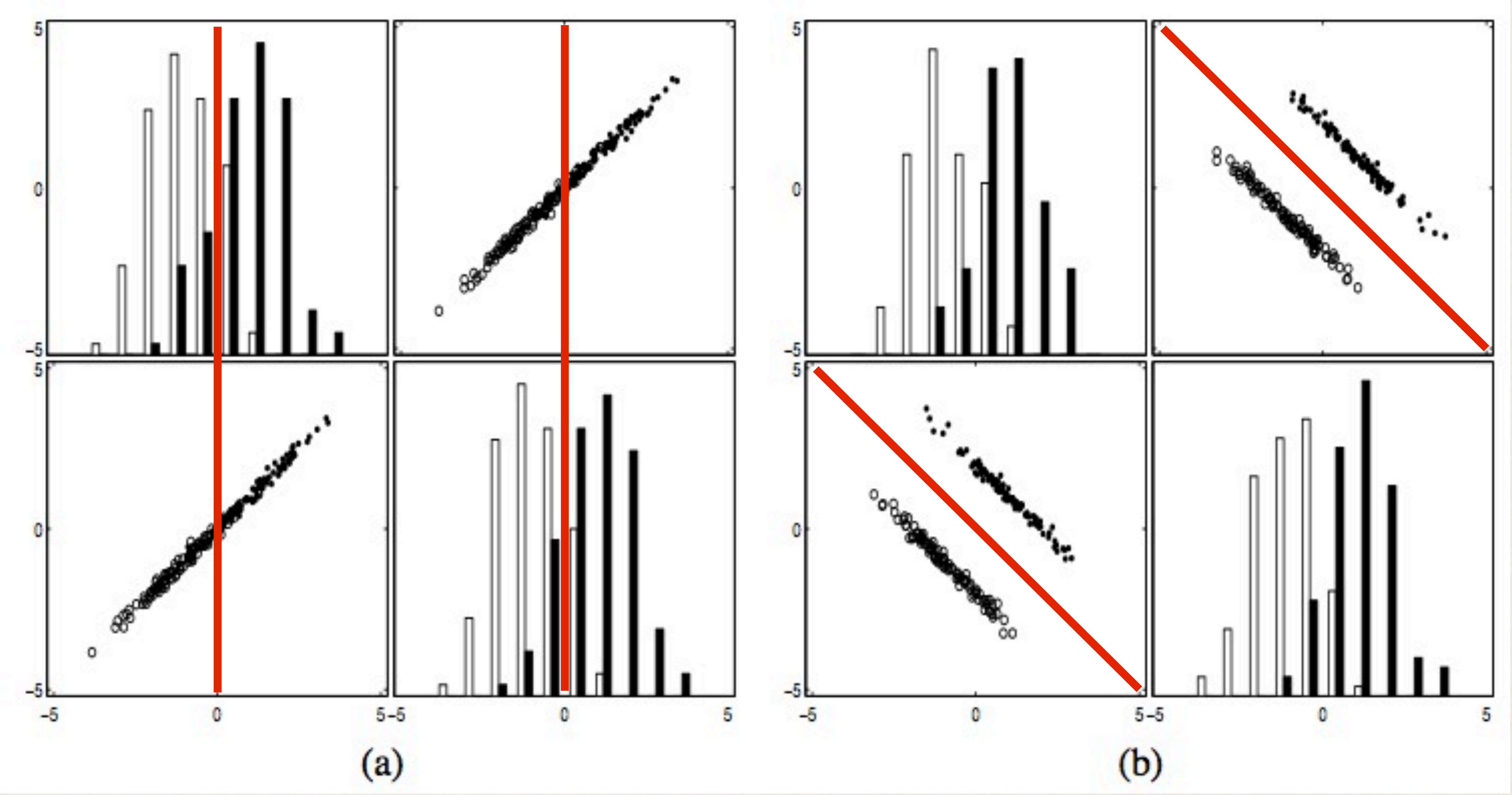
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



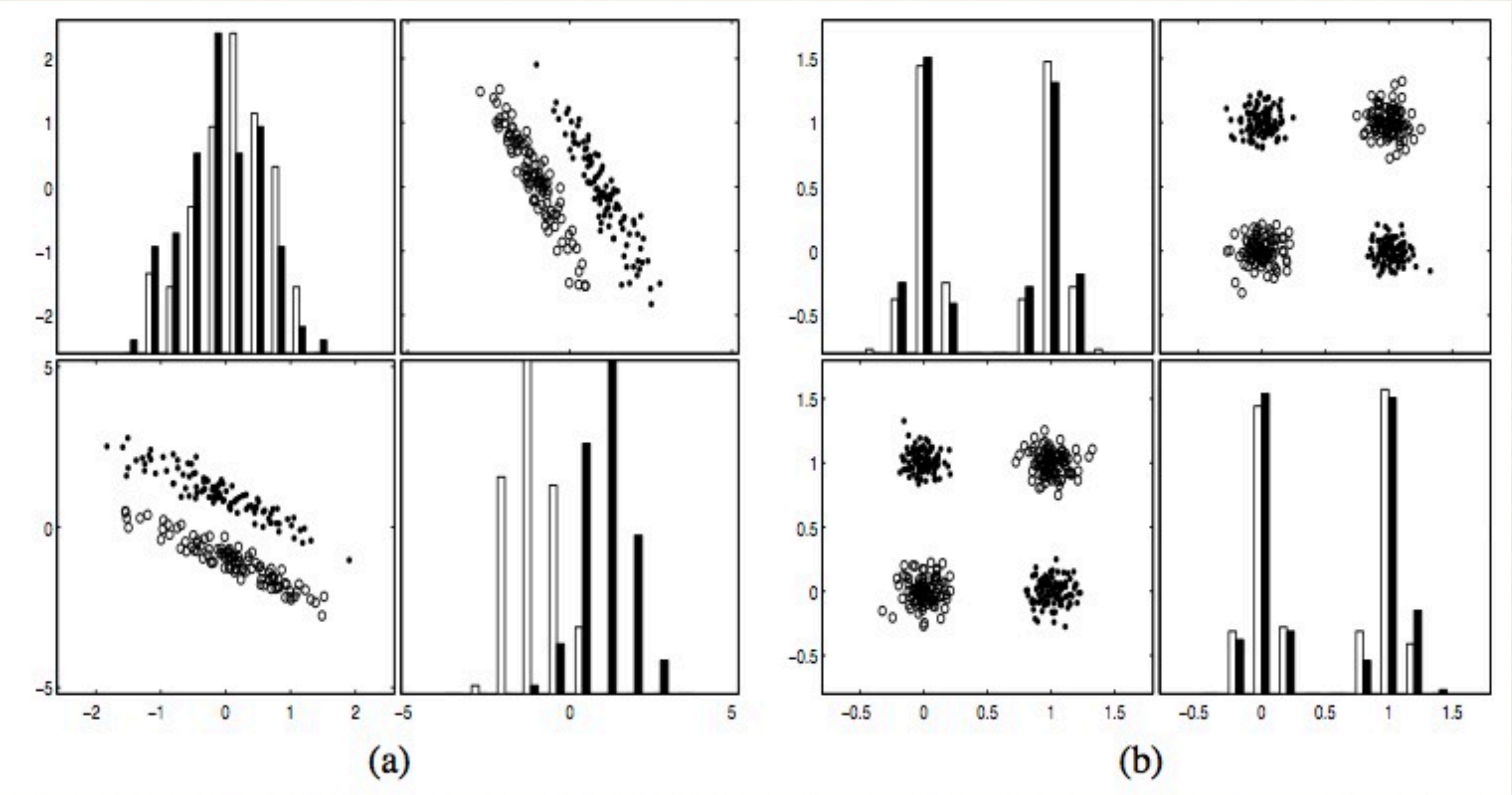
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation

- **Can variables that are individually useless be useful when combined together?**
  - *Yes:* a variable that is by itself useless can improve performance when combined with other useful variables.
  - *Yes:* variables that are individually useless can be useful when combined together.

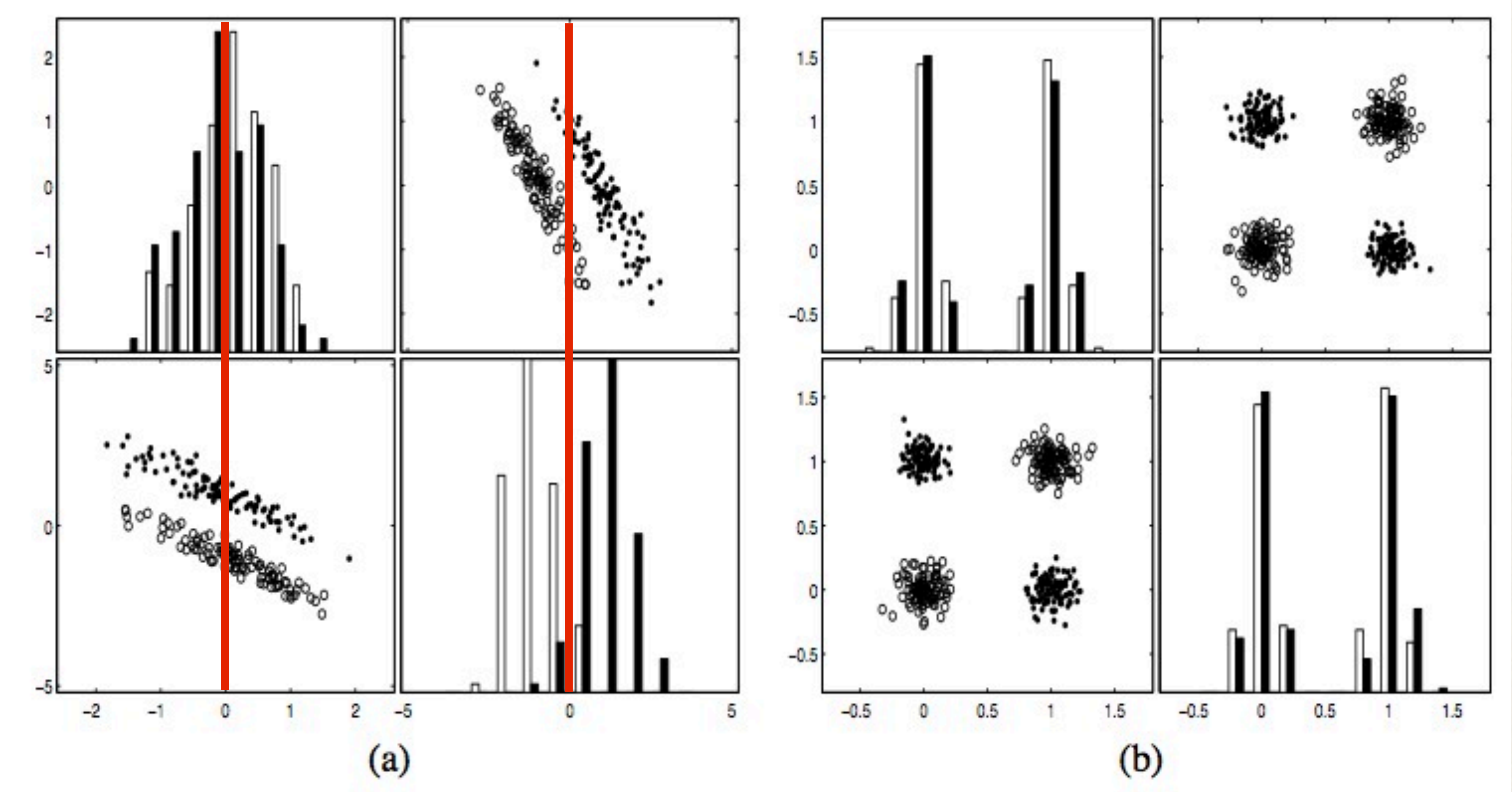
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



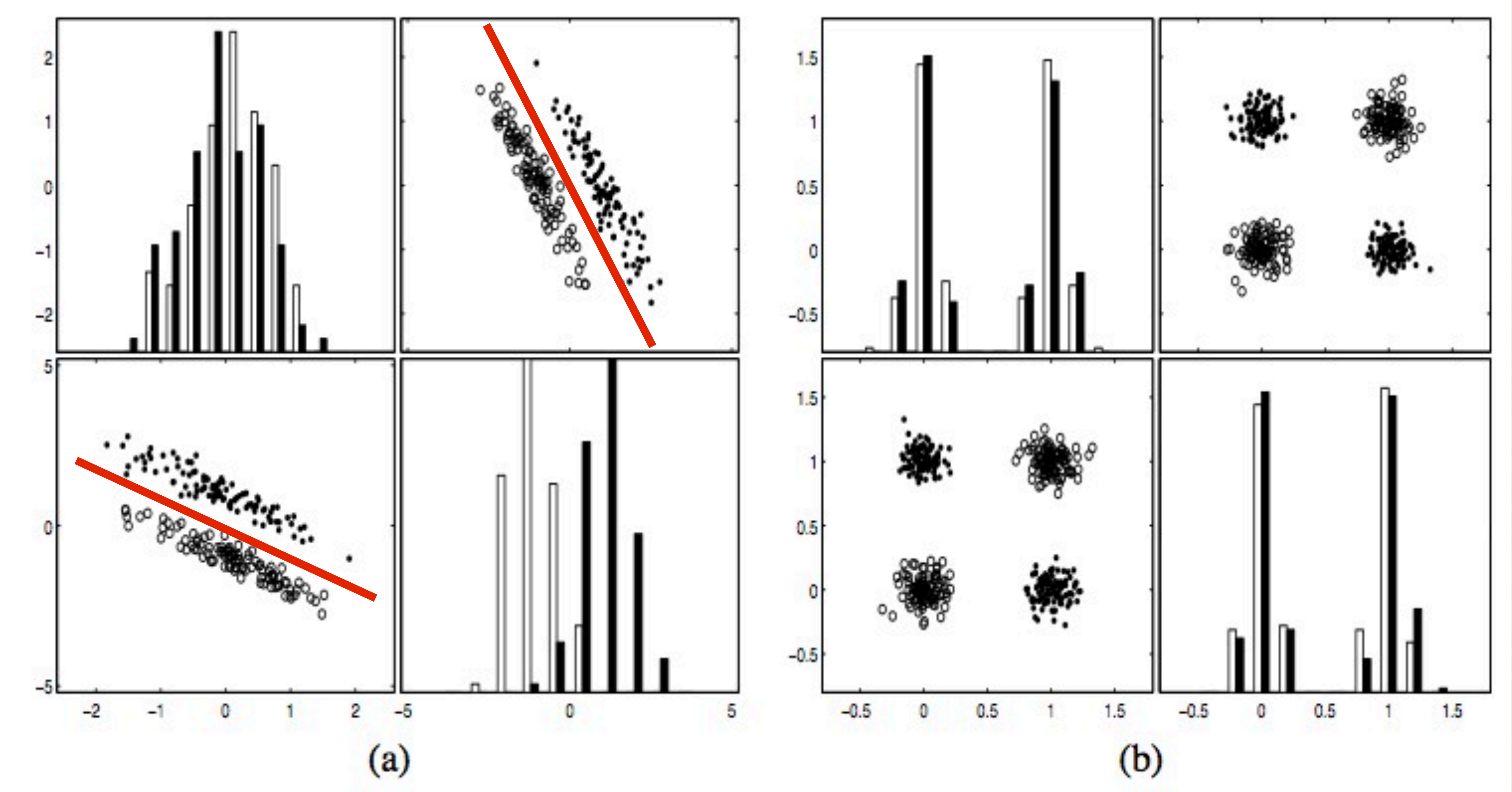
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



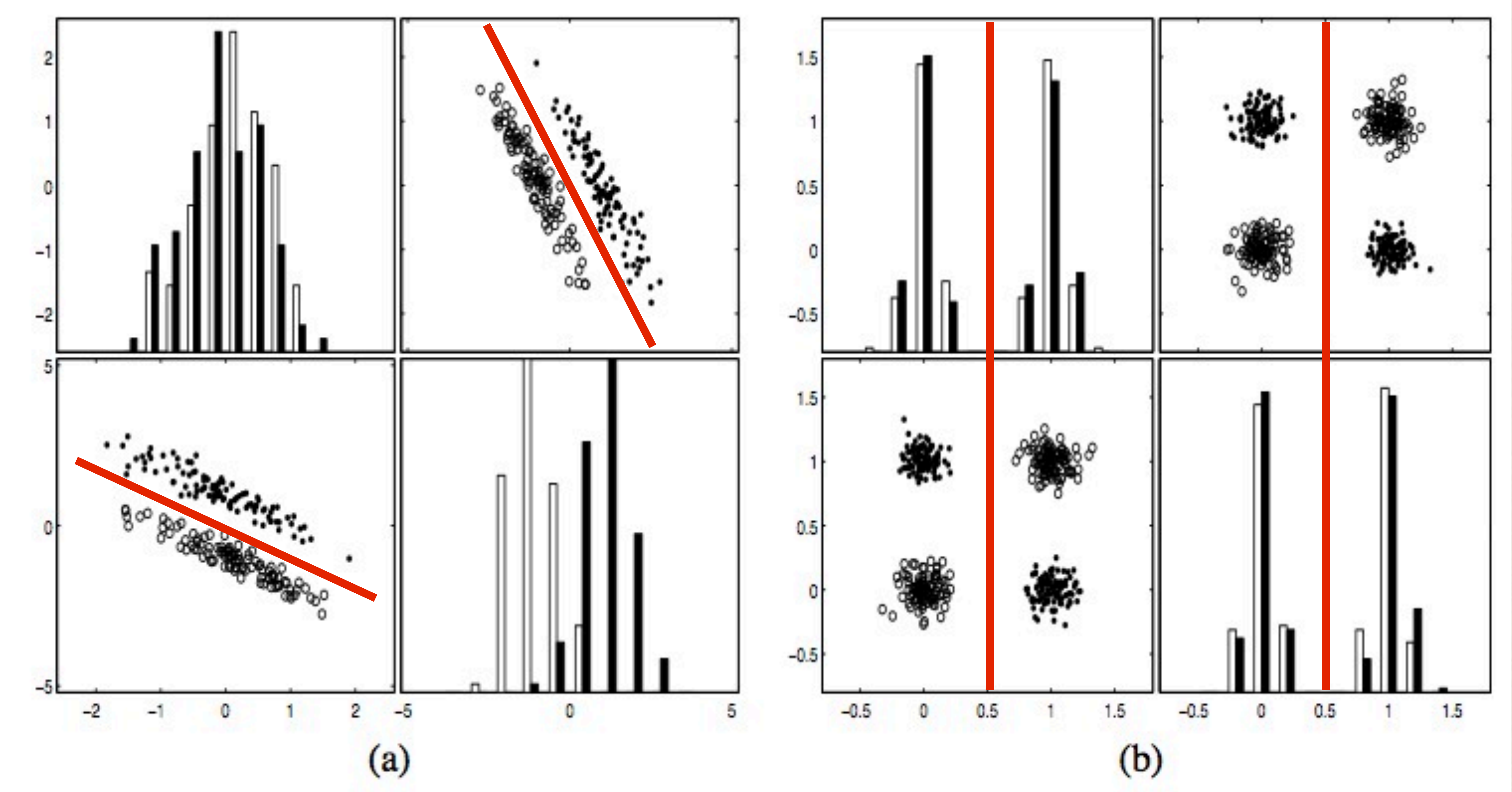
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



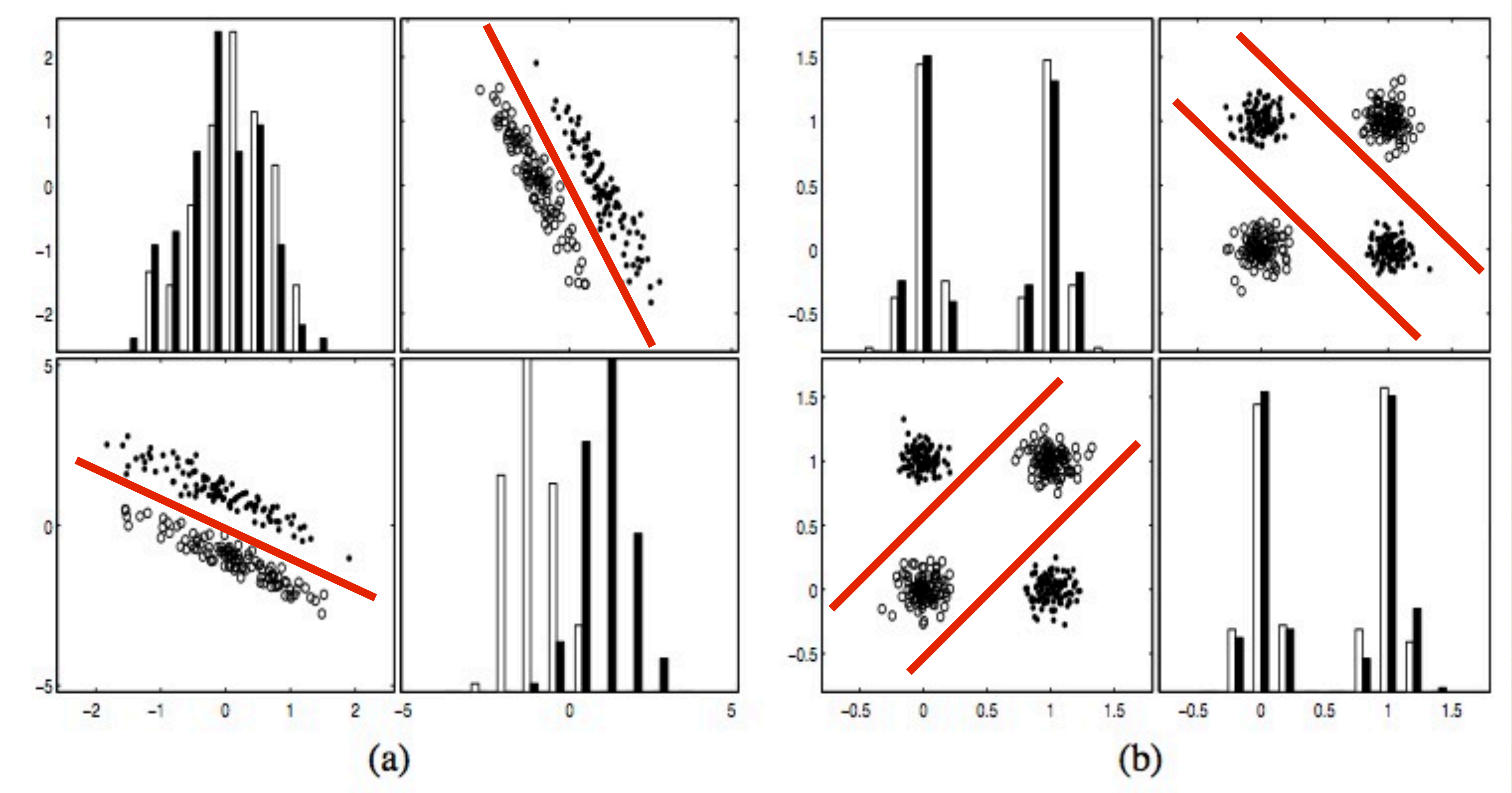
Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

Redundancy and Correlation



Figures from *Guyon and Elisseeff, JMLR, Vol. 3, 2003.*

What is “Relevance”?

- **Relevant to what?**
  - *Is it relevant to a target concept, sample or distribution*—can it help make distinctions? Do samples differ only in terms of a single feature and their label (or do when features are removed)?
  - *Is it relevant to specific algorithms—e.g.* ″usefulness″ to a constructor.
  - Relevance in terms of *saliency, entropy, density, smoothness, reliability*.
- This is a *problem of focus*—selection of relevant features to represent data, and selection of relevant examples to drive the learning process. (*n.b. using irrelevant attributes means more training examples are needed!*)

Blum and Langley, Artificial Intelligence 97, 1997;
Guyon and Elisseeff, JMLR, Vol. 3, 2003.

A Basis in Regression and Weights

- Calculate coefficient for variable '*i*' using an **estimate of correlation:**

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}}$$

- **Information theory approach:** use mutual information between the variables '*i*' and the target (very difficult for cases without nominal targets, however, since it is hard to estimate densities):

$$I(i) = \int_{x_i}\int_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} dx dy$$

Guyon and Elisseeff, JMLR, Vol. 3, 2003.

A Basis in Regression and Weights

- Calculate coefficient for variable *'i'* using an **estimate of correlation:**

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i}-\bar{x}_i)(y_k-\bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i}-\bar{x}_i)^2 \sum_{k=1}^{m}(y_k-\bar{y})^2}}$$

- **Information theory approach:** use mutual information between the variables *'i'* and the target (very difficult for cases without nominal targets, however, since it is hard to estimate densities):

$$I(i) = \sum_{x_i}\sum_{y} P(X=x_i, Y=y)\log\frac{P(X=x_i, Y=y)}{P(X=x_i)P(Y=y)}$$

Guyon and Elisseeff, JMLR, Vol. 3, 2003.

Ranking Methods

- **Summary:** Evaluate the merit of individual features *in isolation*. Rank order features based on *individual predictive power*.

- **Upside:** can be fast, simple, scalable, good empirical success; great for knowledge discovery setting, *e.g.* finding genes that indicate disease.

- **Downside:** variables in isolation can give poor class separation, may miss crucial relationships between individually useless variables. Promotes selection of redundant features.

- **Examples:** Relief-F, InfoGain, ...

Witten and Frank, Data Mining (Morgan Kaufmann, 2005).
Guyon and Elisseeff, JMLR, Vol. 3, 2003.

Subset Methods

- **Summary:** Evaluate the combinations of features to find subsets that *together* have good predictive power.
- **Upside:** can identify complex relationships; removes truly redundant variables; helps find a minimal set that still gives good prediction.
- **Downside:** many methods are computationally complex; unclear how to search the subset space—exhaustive search only possible for small # of variables; unclear how to best guide/halt the search process.
- **Examples:** Correlation-based Feature Selection (CFS), Consistency, WrapperANN, ...

Witten and Frank, Data Mining (Morgan Kaufmann, 2005).
Guyon and Elisseeff, JMLR, Vol. 3, 2003.

Wrappers and Filters

- **Wrapper:** a subset selection technique that has a specific "black box" machine learning component which is used to identify the performance of a given subset, usually with cross-validation.

- **Filter:** subsets are selected via a preprocessing routine independent of a predictor. Arguably faster than wrappers & may reduce overfitting.

- **Embedded:** incorporate variable selection as part of a classifier's training process, *e.g.* decision trees like CART. Search guided by estimating changes in objective function, or direct obj. optimization.

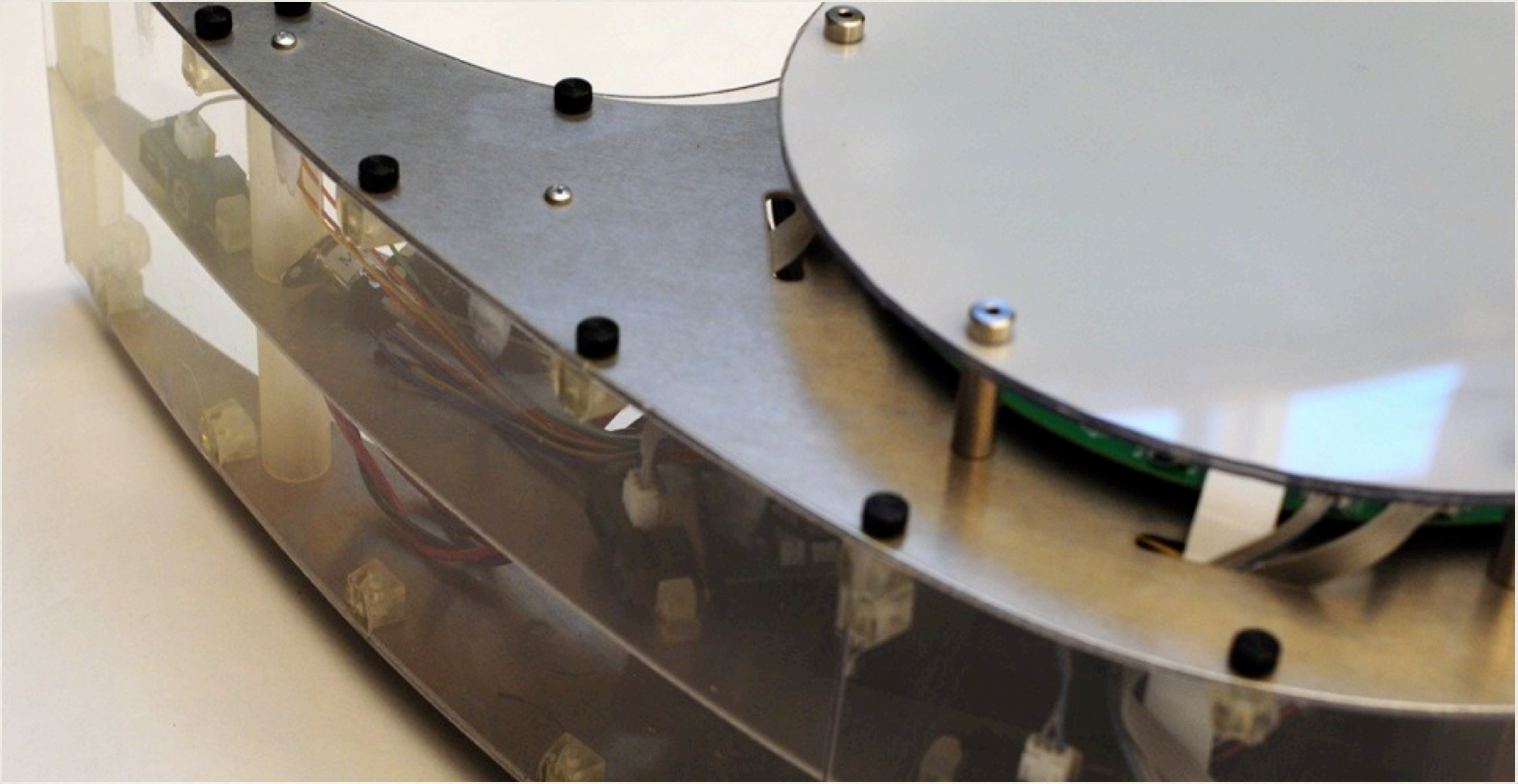
- Can build subsets with *forward selection* or *backward elimination*.

Guyon and Elisseeff, JMLR, Vol. 3, 2003.

Feature Construction

- Goals include achieving the best reconstruction of the data, or being the most efficient in making predictions.
- Both supervised and unsupervised methods for constructing features. (See 2003 JMLR special issue for a focus on feature construction.)
- Tied to ideas of compression and dimensionality reduction, and many algorithms are shared across these fields.
- **Examples:** clustering; basic linear transforms like PCA/SVD; more complex linear transforms like FFT; simple functions applied to subsets of variables; matrix factorization.

Guyon and Elisseeff, JMLR, Vol. 3, 2003.



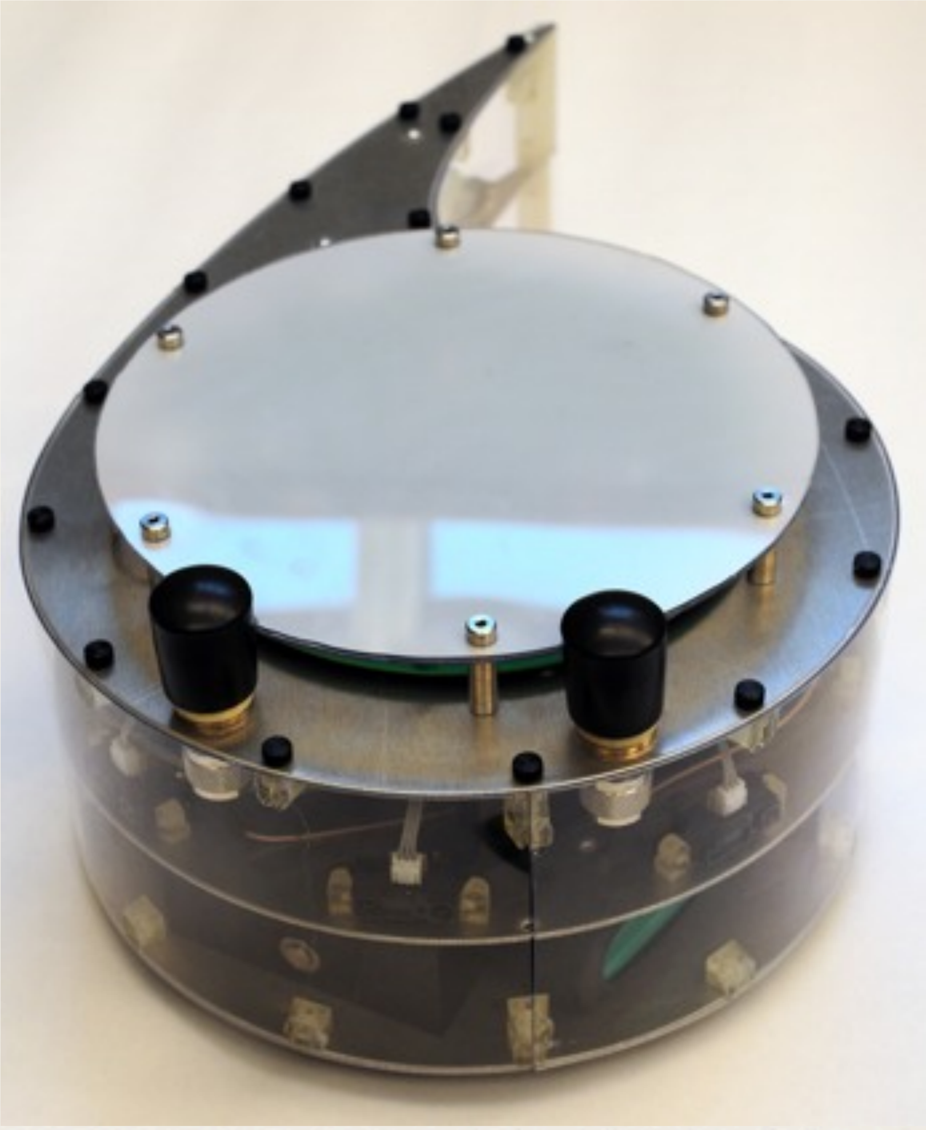
Part 2

Feature Selection on the RLAI Critterbot

Experimental Setup: Critterbot

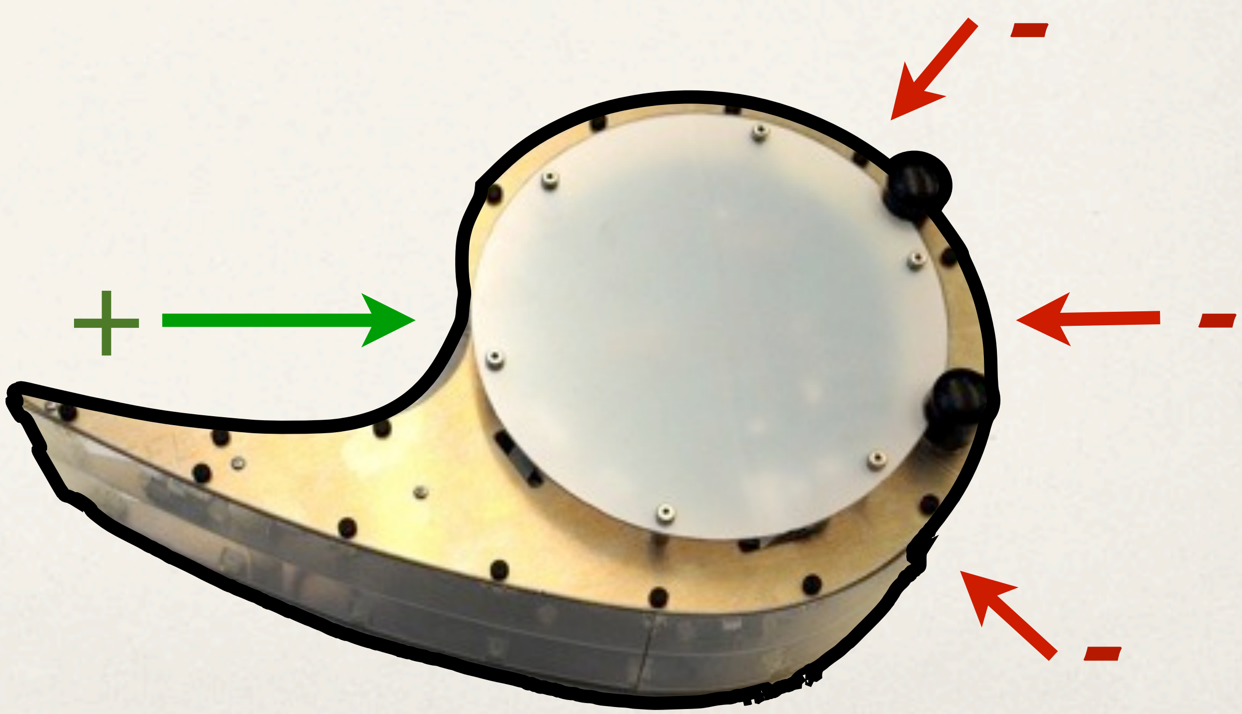
- Critterbot is a mobile robotic platform with 40+ sensors and three drive wheels. http://critterbot.rl-community.org

- Rich sensory data was gathered during a day-long run where Critterbot performed random *options* (macro-actions) / attempted to return to charging station when power became low.



Experimental Setup

- Used day-long Critterbot log file (see talk by Thomas Degris).
- **Paranoid Agent.** Data files were appended with a negative reward signal corresponding to the three forward distance sensors (IR0,IR1,IR6) and a positive reward signal tied to the magnitude of the tail distance sensor (IR8). Cumulative reward was made discrete.
- For this preliminary study, the log was divided into 35 slices; 100,000 steps per slice.
- Each slice was processed using the CFS Subset Feature Selection Algorithm.



Correlation-based Feature Selection (CFS)

- **Big Picture:** balances predictive value with redundancy, favouring high correlation within class and low intercorrelation (*Hall 2000*).

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

- s : subset.
  k : number of features.
  r_{cf} : average feature-class correlation.
  r_{ff} : average feature-feature intercorrelation.

Hall, ICML 17, pp. 359-366, 2000.

Results Using Subset Methods

#	Option Index	Light 0	Light 1	Light 2	Light 3	IR Light0	IR Light1	IR Light2	IR Light4	IR Light6	IR Light7	Accel X	Accel Y	Accel Z	Rot. Vel	Mag X	Mag Y	Mag Z	Therm. 1	Therm. 2	Therm. 3	Therm. 4	Therm. 5	Motor0 Speed	Motor1 Speed	Motor2 Speed	Motor0 Temp.	Motor1 Temp.	Motor2 Temp.	Bat. 1	Bus Voltage	Power Source
0	1																															
1	1	1	1	1	1								1			1		1														
2	1	1	1		1							1	1			1																
3	1	1	1	1	1							1	1			1		1														
4	1	1	1	1	1							1				1																
5	1	1	1		1							1	1			1																
6	1	1	1	1	1								1			1																
7	1	1	1	1	1							1	1			1											1				1	
8	1	1	1	1	1							1	1																			
9	1	1	1		1							1	1			1													1			
10	1	1	1	1	1								1			1																
11	1	1	1	1	1							1	1														1					
12	1	1	1	1	1				1				1			1																
13	1	1																														
14	1	1			1																											
15	1	1	1	1	1					1	1	1	1			1		1					1									
16	1	1	1		1		1					1	1	1		1		1														
17	1	1	1	1	1							1	1																			
18	1	1	1	1	1							1	1			1																
19	1	1	1	1	1							1				1												1				
20	1	1	1	1		1						1	1			1		1														
21	1	1	1	1	1								1			1		1														
22	1	1	1	1									1			1																
23	1	1		1								1				1		1														
24	1	1	1	1									1			1		1														
25	1	1	1	1	1							1	1			1																
26	1	1		1								1	1			1																
27	1	1	1	1									1			1																
28			1	1	1			1					1			1		1	1		1							1		1	1	1
29				1	1										1	1	1	1										1				
30					1								1		1	1	1	1										1				
31	1														1	1		1											1			
32	1															1		1														1
33															1		1											1				
34	1														1	1		1										1				
35								1							1	1				1	1	1		1	1	1						

** IR Distance sensors were excluded from the CFS input set.*

Results Using Subset Methods

#	Option Index	Light 0	Light 1	Light 2	Light 3	IR Light0	IR Light1	IR Light2	IR Light4	IR Light6	IR Light7	Accel X	Accel Y	Accel Z	Rot. Vel	Mag X	Mag Y	Mag Z	Therm. 1	Therm. 2	Therm. 3	Therm. 4	Therm. 5	Motor0 Speed	Motor1 Speed	Motor2 Speed	Motor0 Temp.	Motor1 Temp.	Motor2 Temp.	Bat. 1	Bus Voltage	Power Source
0	1																															
1	1	1	1	1	1								1			1		1														
2	1	1	1		1							1	1			1																
3	1	1	1	1	1							1	1			1		1														
4	1	1	1	1	1							1				1																
5	1	1	1		1							1	1			1																
6	1	1	1	1	1								1			1																
7	1	1	1	1	1							1	1			1											1				1	
8	1	1	1	1	1							1	1																			
9	1	1	1		1							1	1			1													1			
10	1	1	1	1	1								1			1																
11	1	1	1	1	1							1	1														1					
12	1	1	1	1	1				1				1			1																
13	1	1																														
14	1	1			1																											
15	1	1	1	1	1					1	1	1	1			1		1					1									
16	1	1	1		1		1					1	1	1		1		1														
17	1	1	1	1	1							1	1																			
18	1	1	1	1	1							1	1			1																
19	1	1	1	1	1							1				1												1				
20	1	1	1	1		1						1	1			1		1														
21	1	1	1	1	1								1			1		1														
22	1	1	1	1									1			1																
23	1	1		1								1				1		1														
24	1	1	1	1									1			1		1														
25	1	1	1	1	1							1	1			1																
26	1	1		1								1	1			1																
27	1	1	1	1									1			1																
28			1	1	1			1					1			1		1	1		1							1		1	1	1
29				1	1										1	1	1	1										1				
30					1								1		1	1	1	1										1				
31	1														1	1		1											1			
32	1															1		1														1
33															1		1											1				
34	1														1	1		1										1				
35								1							1	1				1	1	1		1	1	1						

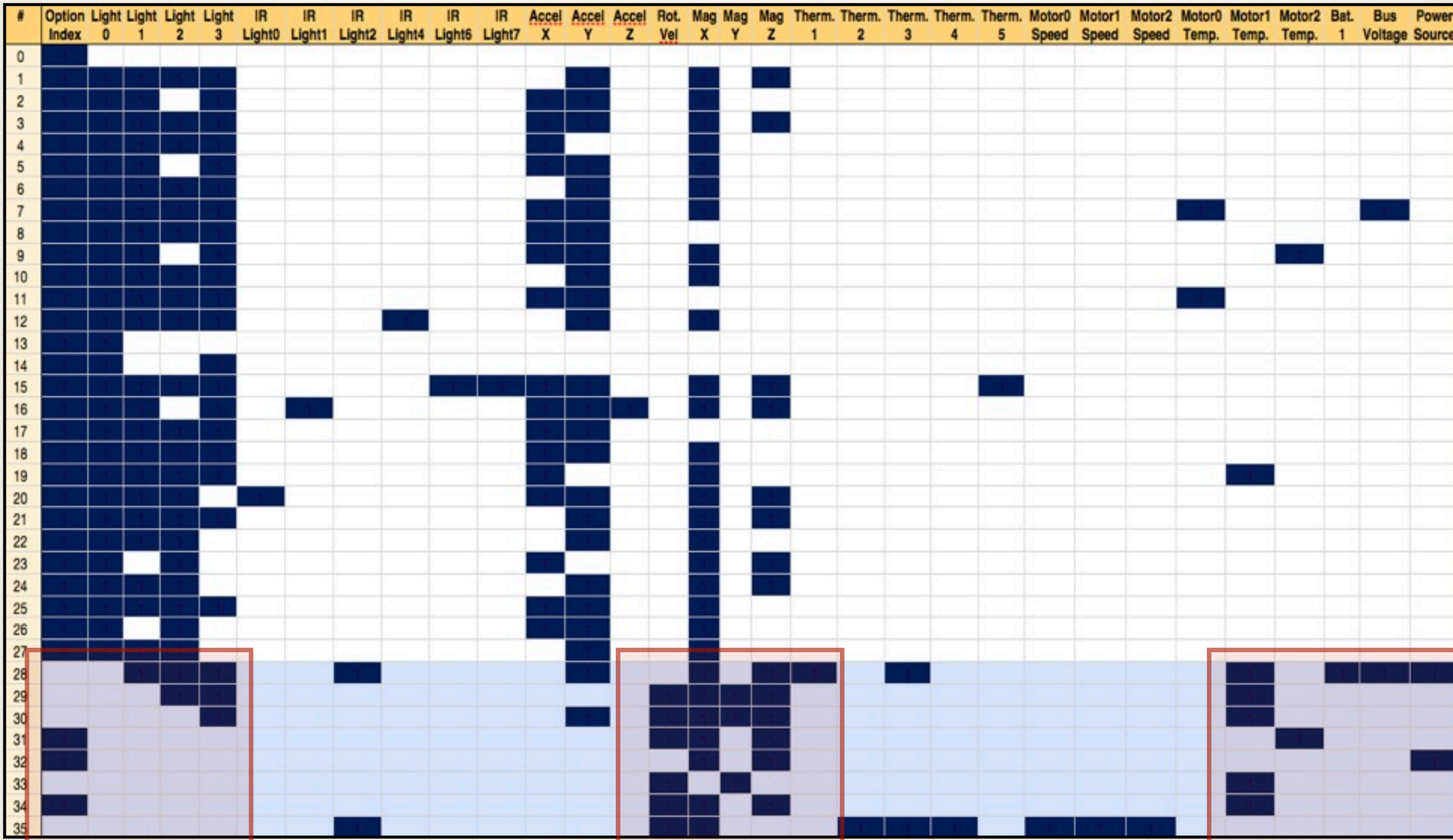
** IR Distance sensors were excluded from the CFS input set.*

Results Using Subset Methods

#	Option Index	Light 0	Light 1	Light 2	Light 3	IR Light0	IR Light1	IR Light2	IR Light4	IR Light6	IR Light7	Accel X	Accel Y	Accel Z	Rot. Vel	Mag X	Mag Y	Mag Z	Therm. 1	Therm. 2	Therm. 3	Therm. 4	Therm. 5	Motor0 Speed	Motor1 Speed	Motor2 Speed	Motor0 Temp.	Motor1 Temp.	Motor2 Temp.	Bat. 1	Bus Voltage	Power Source
0	1																															
1	1	1	1	1	1								1			1		1														
2	1	1	1		1							1	1			1																
3	1	1	1	1	1							1	1			1		1														
4	1	1	1	1	1							1				1																
5	1	1	1		1							1	1			1																
6	1	1	1	1	1								1			1																
7	1	1	1	1	1							1	1			1											1				1	
8	1	1	1	1	1							1	1																			
9	1	1	1		1							1	1			1													1			
10	1	1	1	1	1								1			1																
11	1	1	1	1	1							1	1														1					
12	1	1	1	1	1				1				1			1																
13	1	1																														
14	1	1			1																											
15	1	1	1	1	1					1	1	1	1			1		1					1									
16	1	1	1		1		1					1	1	1		1		1														
17	1	1	1	1	1							1	1																			
18	1	1	1	1	1							1	1			1																
19	1	1	1	1	1							1				1												1				
20	1	1	1	1		1						1	1			1		1														
21	1	1	1	1	1								1			1		1														
22	1	1	1	1									1			1																
23	1	1		1								1				1		1														
24	1	1	1	1									1			1		1														
25	1	1	1	1	1							1	1			1																
26	1	1		1								1	1			1																
27	1	1	1	1									1			1																
28			1	1	1			1					1			1		1	1		1							1		1	1	1
29				1	1										1	1	1	1										1				
30					1								1		1	1	1	1										1				
31	1														1	1		1											1			
32	1															1		1														1
33															1		1											1				
34	1														1	1		1										1				
35								1							1	1				1	1	1		1	1	1						

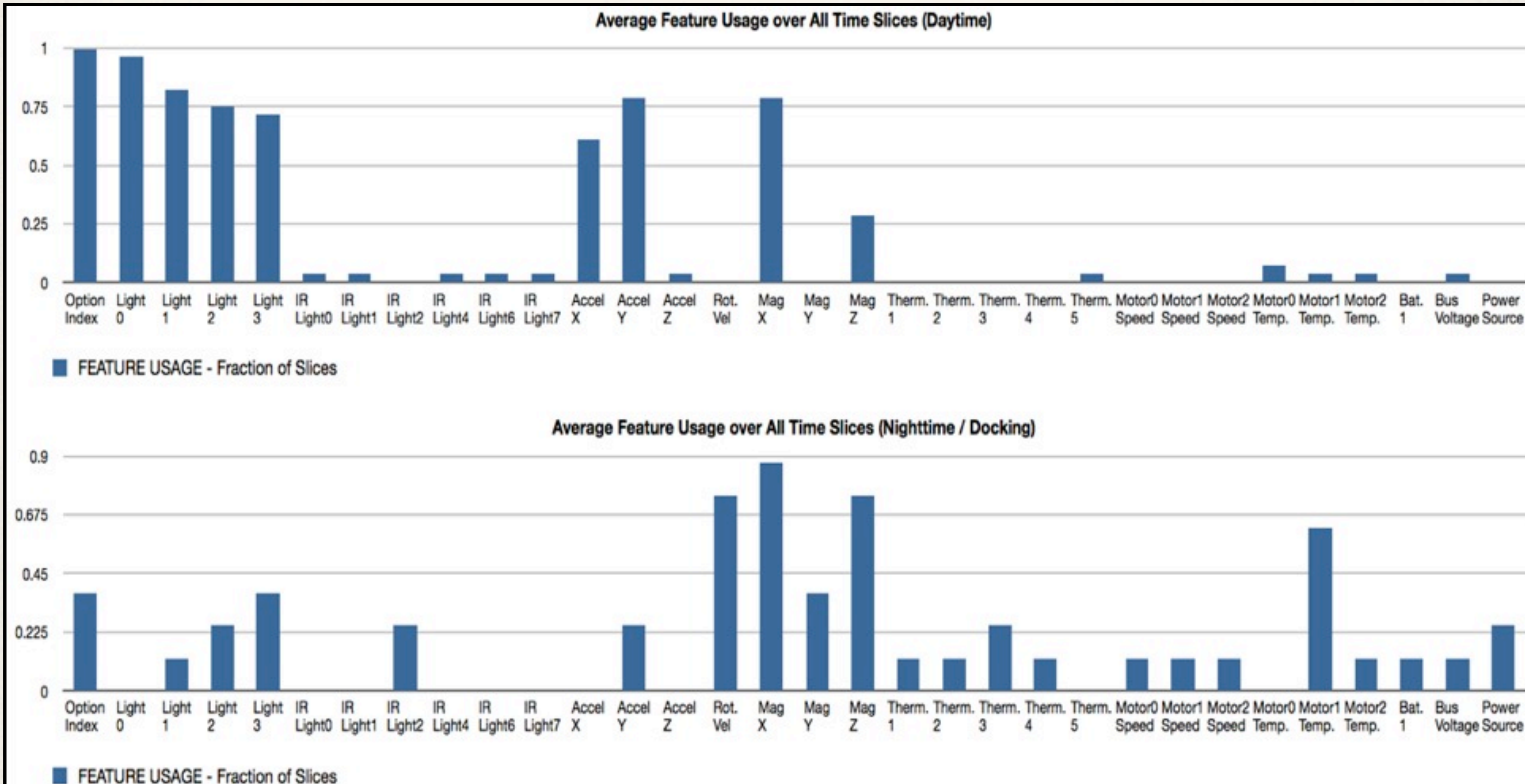
** IR Distance sensors were excluded from the CFS input set.*

Results Using Subset Methods

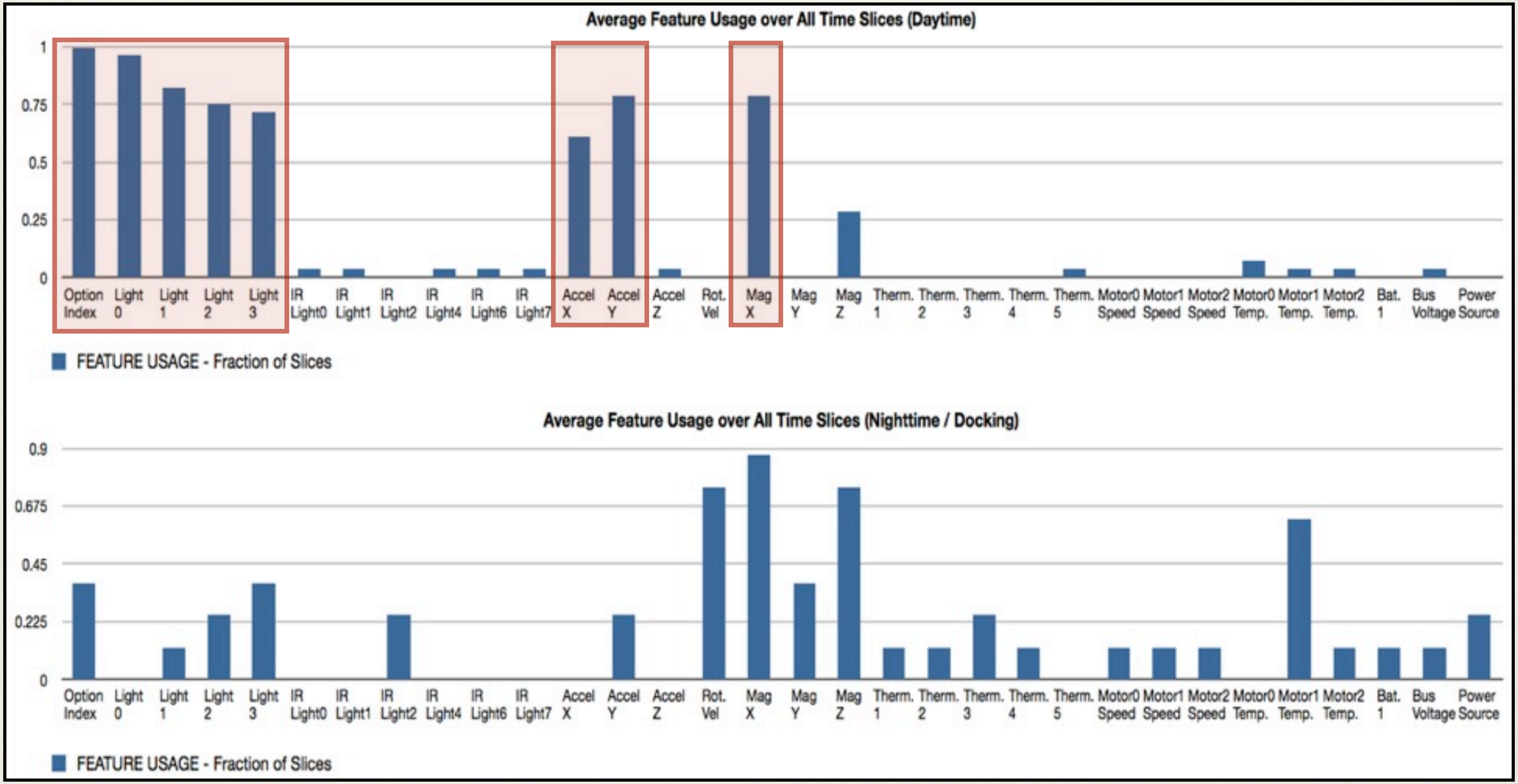


** IR Distance sensors were excluded from the CFS input set.*

Results Using Subset Methods

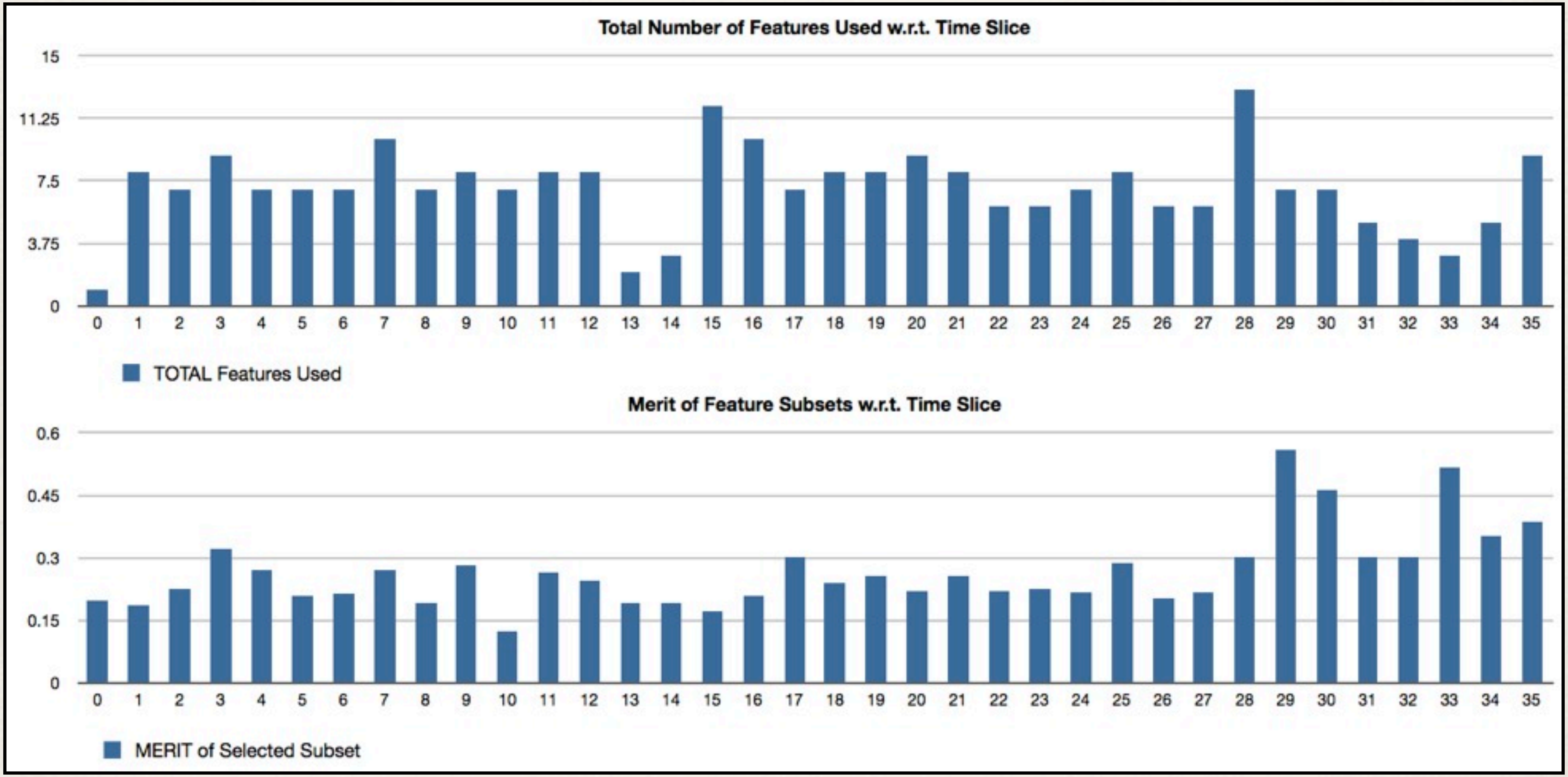


Results Using Subset Methods



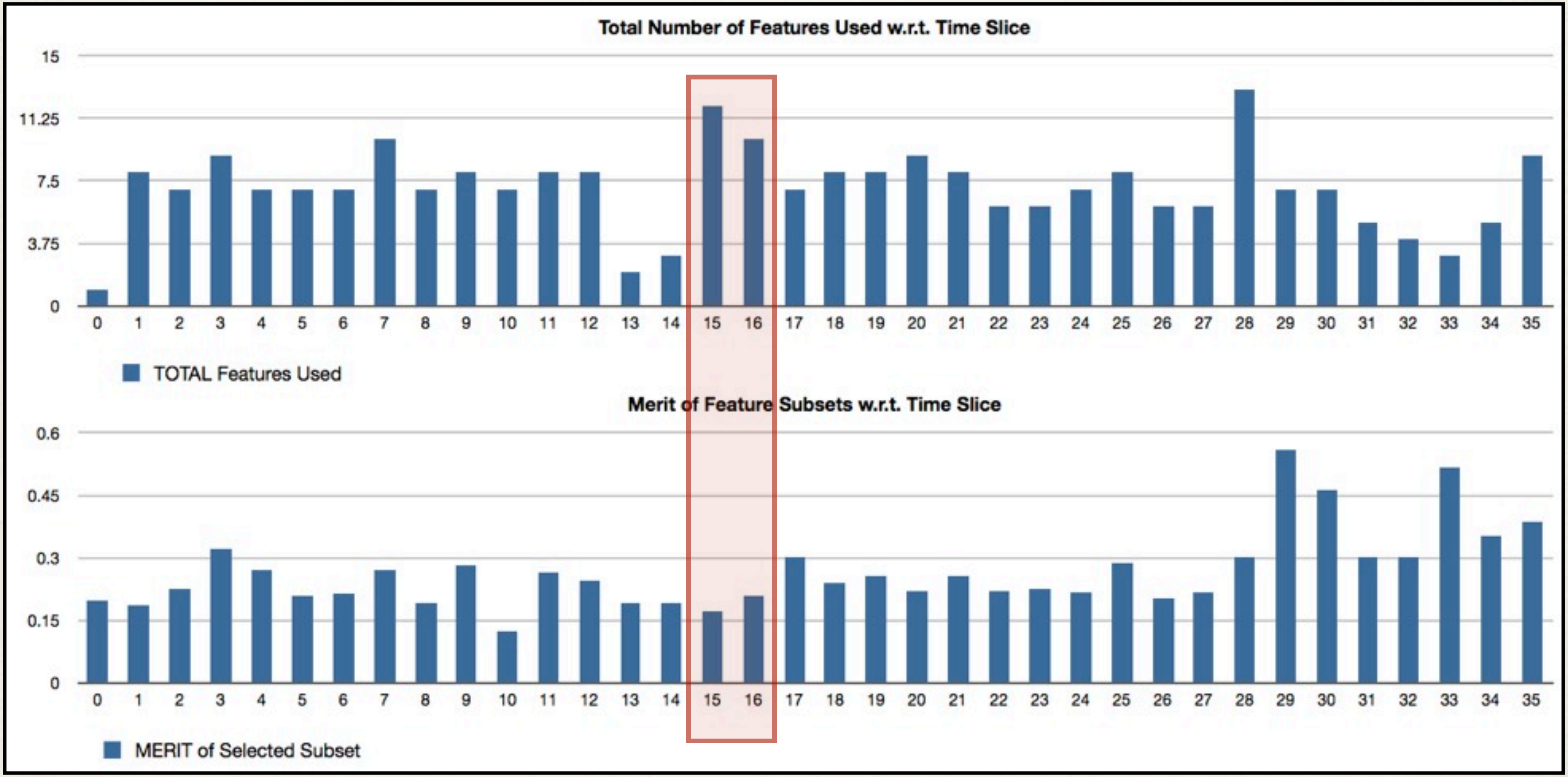
Results Using Subset Methods

Results Using Subset Methods

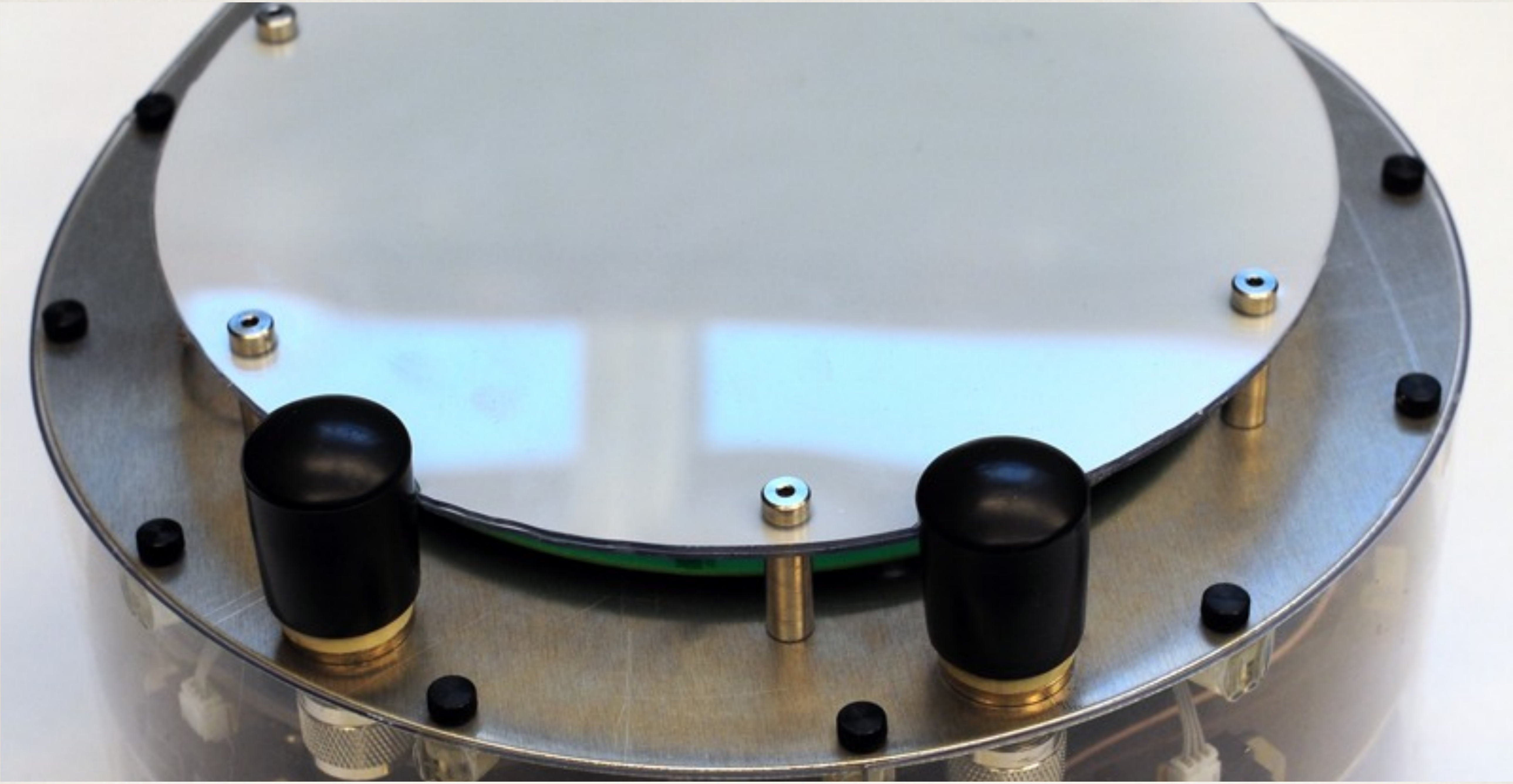


A large number of features does not guarantee good performance.

Results Using Subset Methods



A large number of features does not guarantee good performance.



Conclusions

Summary of Concepts and Key Messages to Leave With

Summary of Concepts

- *Variables* v.s. *features*
- The need to define *relevance* and *usefulness*.
- Interplay between *redundancy* and *correlation*.
- *Ranking methods* v.s. *subset methods*.
- *Wrappers*, *filters*, and *embedded methods*.
- *Feature construction*.

Key Messages to Leave With

- Feature selection is a key idea in AI and statistics.
- Feature selection is important for representation learning in RL, as it provides a way to evaluate and compare the worth of features.
- Feature selection alone is not a solution to RL representation learning
  — *e.g.* need for nominal targets, not incremental, requires stored data.
  — how to automatically construct new features from variables?
- Feature selection literature presents a foundation, context, and intuition to help develop automatic, incremental, life-long representation learning, but is only part of the picture.

One Final Thought

"The art of machine learning starts with the design of appropriate data representations."

(*Guyon and Elisseeff, JMLR, 2003*)

Thanks due to the **RLAI representation learning meeting group** and **Critterbot meeting group** for insight and suggestions, and specifically **Thomas Degris** for the extended Critterbot log data. Critterbot photos by **M. Sokolsky**.