Data-driven modelling of signal-transduction networks

Kevin A. Janes and Michael B. Yaffe‡*

Abstract | New technologies are permitting large-scale quantitative studies of signal-transduction networks. Such data are hard to understand completely by inspection and intuition. 'Data-driven models' help users to analyse large data sets by simplifying the measurements themselves. Data-driven modelling approaches such as clustering, principal components analysis and partial least squares can derive biological insights from large-scale experiments. These models are emerging as standard tools for systems-level research in signalling networks.

Matrix
A table of numbers. Alternatively, a matrix can be viewed as an arrangement of row or column vectors.

**Cell Decision Processes Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA, and Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.*
‡Cell Decision Processes Center, Center for Cancer Research and Departments of Biology and Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
Correspondence to M.B.Y.
e-mail: myaffe@mit.edu
doi:10.1038/nrm2041

The increasing availability of high-throughput and multiplex techniques for quantifying signalling and cellular responses[1–4] makes it immediately feasible to collect large data sets on protein abundance and activity[5–10]. The paradox for systems biology is that these large data sets by themselves often bring more confusion than understanding[11]. In this regard, reductionist experimental approaches seem to make more sense — simplify things down to the point that we can get our heads around them. However, the systems biologist has an important tool that can handle complexity: computation. To improve understanding without adding more confusion, however, computational models must (at least) be indicative of a mechanism and (at least) be based on experimental data[12].

Modelling approaches can be based on prior biological understanding of the molecular mechanisms involved (see the accompanying Review by Aldridge, Burke, Lauffenburger and Sorger in *Nature Cell Biology*). Alternatively, models can be constructed based solely on analysing the data itself, without having to make any assumptions about the underlying mechanisms. These so-called 'data-driven models' allow multivariate biological measurements to become tractable to our intuition and often reveal new surprising and unanticipated biological insights.

In this User's guide article, we introduce three data-driven modelling approaches that have brought understanding to signal-transduction networks and complex biology. First, we discuss clustering as a means for data organization. Then, we describe principal components analysis (PCA) as a method for data condensation. Last, we explain partial least squares (PLS) regression as a technique for data prediction. Because each approach uses the same underlying mathematics to analyse multivariate data sets, it is important to understand how biological measurements are portrayed analytically to data-driven modelling algorithms, and what these algorithms are trying to accomplish.

From data matrices to data-driven models

A systems biology approach to studying cell signalling entails measuring the levels, localization and activities of several proteins over a range of timescales and treatment conditions. Cellular signals are not static but dynamic[13,14], and time courses are perhaps the most fundamental type of signalling data set. Each measured kinase or substrate, for example, constitutes a signalling variable in the data set, and each time point constitutes an observation. For simplicity, we focus only on time points as observations, although different treatment conditions and perturbations that affect signalling variables could likewise be considered as observations. Together, the observations (as rows) and signalling variables (as columns) can be represented as a 'data matrix' (FIG. 1a).

We routinely use data matrices as the basis for constructing time-course plots (FIG. 1b). These plots help users to visualize changes in signals over time while retaining the quantitative information of the original data matrix. However, time-course plots are limiting for systems applications that track many signals together. The overlap and intersection of large numbers of time courses provides little insight into how signals operate as a network[15], because each measurement is plotted separately with respect to time. Individual signals vary dynamically, but can also co-vary with respect to one another.

An alternative view of the same data set is to consider each signalling variable at its own axis (FIG. 1c). Here, the position of a time point is determined by the

a

Variables →

Observations	AKT	JNK	MK2	...	AKT-sub
0 min	$AKT_{0\ min}$	$JNK_{0\ min}$	$MK2_{0\ min}$	...	$AKT\text{-}sub_{0\ min}$
5 min	$AKT_{5\ min}$	$JNK_{5\ min}$	$MK2_{5\ min}$	...	$AKT\text{-}sub_{5\ min}$
⋮	⋮	⋮	⋮		⋮
24 h	$AKT_{24\ h}$	$JNK_{24\ h}$	$MK2_{24\ h}$	...	$AKT\text{-}sub_{24\ h}$

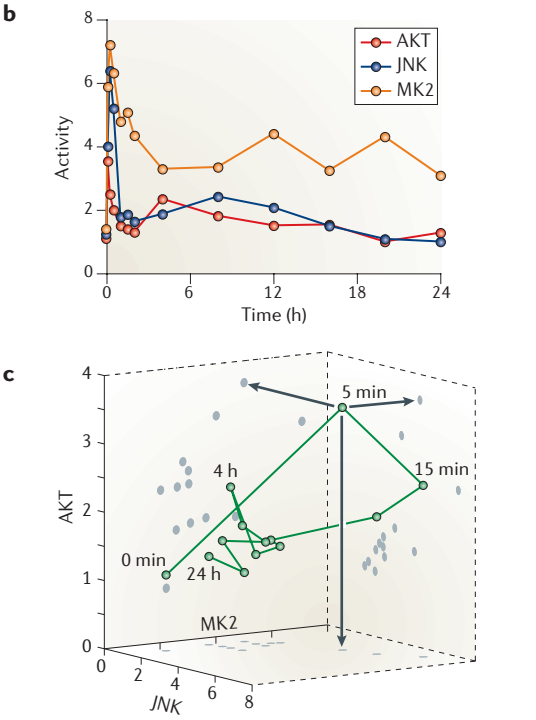


Figure 1 | **Alternative representations of a systems biology data set. a** | A data matrix of time points (considered here to be observations) and four intracellular signals (considered here to be signalling variables): *v-akt* murine thymoma viral oncogene homologue (AKT) activity, c-jun N-terminal kinase (JNK) activity, mitogen-activated protein kinase-activated protein kinase-2 (MK2) activity and AKT-substrate (AKT-sub) phosphorylation. **b** | Time-course plots of the data matrix. AKT, JNK and MK2 measurements are plotted against time. **c** | Data space defined by AKT, JNK and MK2 signalling. AKT, JNK and MK2 are plotted along their respective coordinates, and time is indicated next to each marker. AKT–JNK, AKT–MK2 and JNK–MK2 correlations are shown as projections onto the data space (see 5-min observation). For parts **b** and **c**, AKT-sub has been omitted for clarity. AKT, JNK and MK2 data are adapted from REF. 6.

individual signal-activation strengths (for example, of *v-akt* murine thymoma viral oncogene homologue (AKT), c-jun N-terminal kinase (JNK), and mitogen-activated protein kinase-activated protein kinase-2 (MK2) in FIG. 1c), which project the observation along the signalling axes. Each signalling axis functions as a dimension for a measured variable, exactly as *x*-, *y*- and *z*-axes are dimensions for measurements of position. A 'data space' consists of the complete set of signalling axes (and the projection of the observations along these axes; FIG. 1c).

Similar to a time-course plot, a data space is quantitatively equivalent to the starting data matrix (FIG. 1a). Importantly, however, covariation between signals is retained: flattening the data onto any surface of the data space gives the correlation between the two signals; the axes of these determine the surface (FIG. 1c). Representing data matrices as data spaces is also mathematically advantageous, because experimental observations are cast as vectors in a coordinate system that is defined by the measured signals. These vectors (as well as the coordinate system itself) can then be analysed by linear-algebraic techniques.

Applying linear algebra to data spaces. The use of linear algebra is beneficial for the types of data that are encountered in systems biology. One reason for this is because, after three measured signals, we run out of spatial dimensions to use for quantitatively viewing the data space. For instance, in FIG. 1c, where would an axis for the AKT-substrate (AKT-sub) measurements point relative to the AKT, JNK and MK2 axes? By contrast, there is no limit to the number of dimensions that can be accommodated by vectors or the mathematical transformations that involve them. A second reason is that by algebraically restructuring the data space, it becomes possible to identify a small number of 'optimal' dimensions in the experimental observations. These optimal dimensions are combinations of signalling variables that allow a simple, efficient approximation of the original data space and therefore constitute a data-driven model.

The goal of data-driven modelling is to mathematically specify how the optimal dimensions are defined and then extract them from the data space. Ideally, for data-driven models to help our intuition, the optimal dimensions must provide a new, quantitative perspective on the underlying biology that is fundamentally different from that of the original data set alone. The usefulness of any data-driven model is balanced by its predictive power and biological insight, both of which are determined by how an optimal dimension is defined.

Data organization through clustering

The purpose of clustering is to improve the arrangement of the data matrix by organizing rows or columns (or both) in a way that reveals potential biological meaning. For systems biology measurements, organization is achieved by grouping, for example, signalling proteins with similar behaviour. The key to successful clustering is defining what is meant by 'similar', which depends on how the data are framed. It is convenient to explain clustering through a vector-based approach, in which similarity is expressed as a 'distance' between either the vectors themselves or the parameters that are derived from the vectors. In general, how distance is defined is more important than the particular algorithm that is used to cluster the distances[16].

The two general types of clustering are divisive clustering and agglomerative clustering. Divisive clustering takes an entire set of data and divides it successively into groups, in which each group contains those vectors

Vector
A mathematical quantity that has both magnitude (or length) and direction. The entries of a vector specify the magnitude of its projection in different directions.

Linear algebra
A branch of mathematics that involves linear manipulations of vectors and matrices.

Transformation
A mathematical function that can be applied to vectors and matrices.

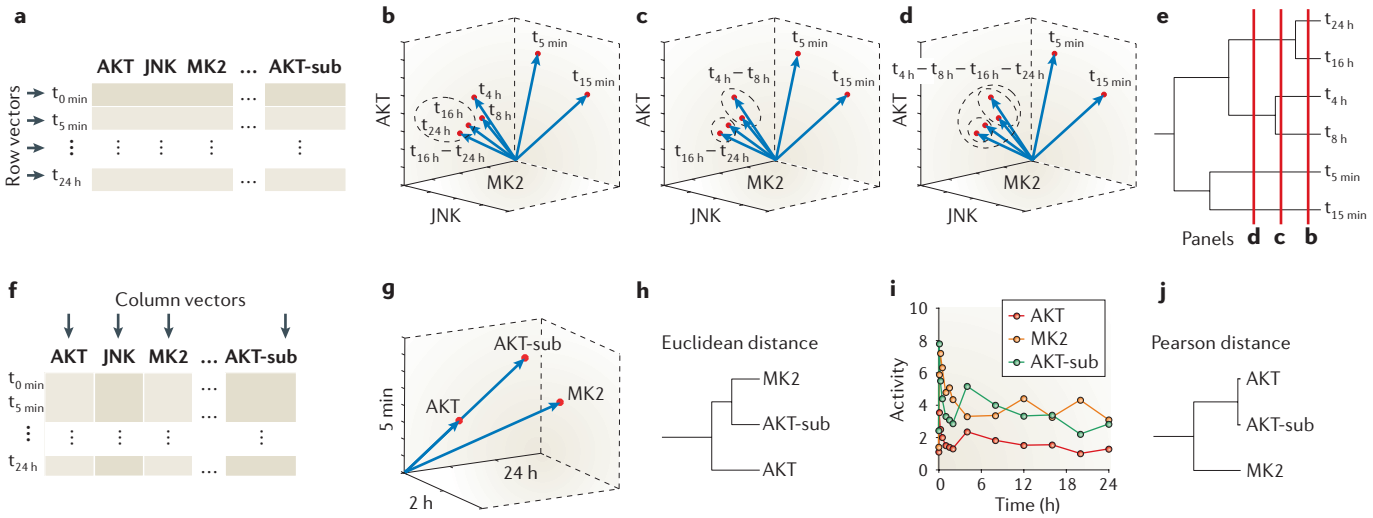


Figure 2 | **Clustering of row and column vectors by different distance metrics. a** | Row–vector representation of the time points (t) shown in FIG. 1a. **b** | The first hierarchical cluster, $t_{16\,h} - t_{24\,h}$ (dashed circle), identified by Euclidean distance. Each vector represents one time point and its position in the data space is defined by the magnitude of each signalling variable on the respective axes. **c**,**d** | Second and third hierarchical clusters, $t_{4\,h} - t_{8\,h}$ and $t_{4\,h} - t_{8\,h} - t_{16\,h} - t_{24\,h}$ (dashed circles), identified by Euclidean distance. **e** | Complete dendrogram for the clustering of the time points shown in part **b**. Vertical lines (red) indicate the clusters identified in parts **b**, **c** and **d**. **f** | Column–vector representation of the measured signalling proteins shown in FIG. 1a. **g** | The projection of AKT, MK2 and AKT-substrate (AKT-sub) phosphorylation onto three time-axis dimensions. **h** | Euclidean-distance-based dendrogram of the measured signalling proteins shown in part **g**. **i** | Comparison of AKT, MK2 and AKT-sub time-course plots. Note that AKT-sub is geometrically closer to MK2 (based on Euclidean distance), but its pattern of activation is closer to AKT (based on Pearson distance). **j** | Pearson-distance-based dendrogram of the measured signalling proteins shown in part **g**. Note that the clusters are different from those shown in part **h**. AKT, JNK and MK2 are defined in FIG. 1.

that are most closely related. By contrast, agglomerative clustering starts with each vector individually and builds clusters by grouping together those vectors that are similar. Divisive clustering is more computationally involved and less commonly used. Therefore, we focus here on agglomerative clustering and refer the reader elsewhere for more detailed methods[17,18]. To show the importance of distance measures, we compare clustering using row vectors and Euclidean distances to clustering using column vectors and Pearson distances. Ultimately, the ability to extract meaning from clustering depends on the user's prior biological understanding of the objects that are organized (see below).

Row vectors and Euclidean distance. The dynamic trajectory of a signalling time course, as shown in FIG. 1b, can alternatively be viewed along the signalling axes themselves (FIG. 1c). Individually, each observation (or time point) can be treated as a row vector (t; FIG. 2a) that points from a common origin to a position in the data space that is based on the variables that have been measured (FIG. 2b shows a subset of row vectors). The number of row vectors in a data matrix equals the number of observations (time points), and each row vector contains the projection of the signalling variables for that observation (FIG. 2a).

The simplest way to cluster observations is to group row vectors that lie closest together based on their geometric (or Euclidean) distance. For instance, the following equation calculates the Euclidean distance (edist) between row vectors $t_{5\,min}$ and $t_{15\,min}$ in the AKT–JNK–MK2 data space:

$$\text{edist}(t_{5\,min}, t_{15\,min}) = \sqrt{\left((AKT_{5\,min} - AKT_{15\,min})^2 + (JNK_{5\,min} - JNK_{15\,min})^2 + (MK2_{5\,min} - MK2_{15\,min})^2\right)} \quad (1)$$

For hierarchical clustering, the two vectors with the smallest Euclidean distance are combined into a new single group (for example, group $t_{16\,h} - t_{24\,h}$ in FIG. 2b). Then, the distance between the remaining row vectors and this new group is measured and compared with the distance between the remaining vectors themselves. The next two most similar vectors form a new group (group $t_{4\,h} - t_{8\,h}$ in FIG. 2c). As clusters accumulate, it becomes necessary to consider the distances between clusters, as well as vector–vector and vector–cluster distances. Eventually, the next most similar cluster will involve the combination of two prior clusters into a larger group (that is, group $t_{4\,h} - t_{8\,h} - t_{16\,h} - t_{24\,h}$ in FIG. 2d). The distance between clusters can be measured in different ways (see REF. 16 for details).

We can diagrammatically represent each step in the clustering process by generating a hierarchical 'tree' (or dendrogram, FIG. 2e) in which all of the vectors are eventually combined into a single large group. This tree compresses the high-dimensional similarity relationships between each of the individual row vectors (FIG. 2b) into a simple two-dimensional graph of clusters (FIG. 2e).

Row vector
A vector that is composed of one entire row of a matrix with dimensions that are specified by the matrix columns.

Euclidean distance
A mathematical quantity that calculates the measurable geometric distance between two vectors pointing from a common origin.

Column vector
A vector that is composed of one entire column of a matrix with dimensions that are specified by the matrix rows.

Pearson distance
A mathematical quantity that calculates the difference in direction between two vectors pointing from a common origin.

Drawing a vertical line down the tree divides the data into *k* discrete clusters based on the number of branches that intersect with the vertical line. Moving the line to the left on the tree decreases *k* (identifying larger clusters that group many observations), whereas moving the line to the right increases *k* (smaller clusters with fewer groups of observations).

Varying *k* therefore shows all of the possible ways the individual observations can be related to each other, given the information in the data set. Meaningful clusters emerge when they match existing knowledge about the observations, such as early versus late time points or distinct treatment conditions. If there is a very strong prior expectation about the number of clusters, this can be specified directly through other clustering techniques that are not hierarchical, such as *k*-means clustering (Supplementary information S1 (box)).

Column vectors and Pearson distance. The preceding example used row vectors to cluster together time points that had similar signalling activities. Often, however, we wish to understand how one signalling variable (for example, measured kinase activity or substrate phosphorylation) relates to another signalling variable. To do this, we need to re-evaluate the data matrix from the point of view of column vectors (FIG. 2f). Here, the observations (rather than the signalling variables) constitute the dimensions of the data space. Each signalling variable then forms a column vector that projects along the observation (time) axes based on the activity of that variable for the different observations (see FIG. 2g for the 5-min, 2-h and 24-h time-point axes).

We could attempt to cluster column vectors by Euclidean distance as was done for row vectors (FIG. 2h). However, because the magnitudes of the column vectors that reflect the amplitude of the signalling responses often differ significantly, clustering that is based on Euclidean distance might give misleading results. For example, the magnitude of the phosphorylation response of AKT-sub might lie closer to the activity of MK2 by Euclidean distance in time-axis space, but its pattern of activation parallels the kinase activity of AKT (FIG. 2i). We might therefore wish to cluster column vectors that have similarly shaped trajectories, regardless of their Euclidean proximity in the data space (FIG. 2g). To match patterns of activation, we must consider the covariation between column vectors. The covariance (cov) between AKT and JNK column vectors, for example, is defined as:

$$\mathrm{cov}(\mathrm{AKT}, \mathrm{JNK}) = \frac{1}{\text{No. of time points}} \sum_{i=0\ \mathrm{min}}^{24\ \mathrm{h}} (\mathrm{AKT}_i - \overline{\mathrm{AKT}})(\mathrm{JNK}_i - \overline{\mathrm{JNK}}) \quad (2)$$

$\overline{\mathrm{AKT}}$ and $\overline{\mathrm{JNK}}$ are the AKT and JNK activities averaged across all observed time points. The covariance of the column vectors are then divided by the square root of the product of their individual variances to obtain the Pearson correlation coefficient, which can be used as the basis for clustering[16].

The values of the Pearson correlation coefficient range from –1 to +1, with +1 indicating strong positive correlation between the variables, –1 indicating that the variables are anti-correlated, and 0 reflecting no correlation. The Pearson distance, defined as one minus the Pearson correlation coefficient, can be used for clustering exactly as the Euclidean distance was used for organizing data matrices by hierarchical clustering (FIG. 2j) and *k*-means clustering (Supplementary information S1 (box))[16]. More refined methods of clustering are also available, such as model-based clustering[17], in which the clusters are approximated as a mixture of Gaussian distributions, and the row or column vectors are treated as random samplings from the clusters. In model-based clustering, the relevant distance measurement is called the Mahalanobis distance, which is the Euclidean distance scaled by the covariance between vectors[16]. So far, however, biological applications of model-based clustering have been limited to studies of gene expression[17].

Deriving biological insight. Clustering is most successful when the observations and variables in the data matrix contain a mix of known and unknown biological mechanisms. Recent phenotypic[19], pharmacological[20] and RNA interference (RNAi)-based[21] screens have used clustering to indicate the functions of unknown genes or small molecules by their association with recognized perturbations. Clustering time-course data with many variables can also be used to separate kinetic clusters that are possibly subject to the same upstream regulators[22]. These 'guilt by association' studies benefit from clustering by focusing the user's attention to the subset of data within the clusters, rather than the entire starting data set.

Principal components analysis

Clustering is a quantitative way to inspect the overall organization of the data matrix, but the technique does not simplify the large number of dimensions in the data space. PCA achieves dimensionality reduction by finding new axes, called principal components, that identify the linear combinations of signalling axes most tightly connected with one another. Principal components function as super-axes, and allow users to view the entire data space in just two or three dimensions that capture the most important information in each of the original signalling axes. Mathematically, this approach amounts to analysing the data matrix (after appropriate scaling and normalization; see Supplementary information S2 (box)) for overall covariance of each signalling variable with every other signalling variable (BOX 1). The principal components are defined by weighting signals with high covariance and de-emphasizing signals that show little covariation with other signals. In this way, PCA condenses measurements to highlight the global patterns in the data set as reflected by just two or three dimensions that capture the maximal covariation between all of the signals. Mathematically, the prioritization of signals based on covariation is accomplished by quantifying the importance of unique combinations of signals to the overall covariance of the data set by using eigenvalues (BOX 1).

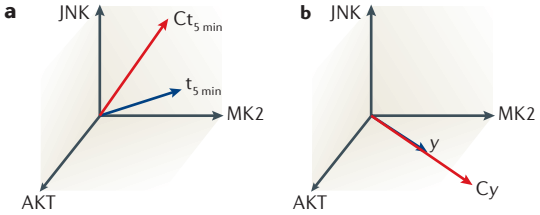
k-means clustering
A clustering technique in which observations are grouped into a fixed number of pre-specified clusters called centroids.

Eigenvalue
A mathematical quantity that provides the scaling factor for an eigenvector of a given transformation. For PCA, eigenvalues quantify the contribution of different portions of the data set to the overall measured variation.

Box 1 | Eigenvalues and eigenvectors

Overall covariance is itself a matrix (C) that describes the covariation (cov) between each measured signal and every other measured signal. For the example data set, C is defined as:

$$C = \begin{bmatrix} \text{cov(AKT, AKT)} & \text{cov(AKT, JNK)} & \text{cov(AKT, MK2)} \\ \text{cov(JNK, AKT)} & \text{cov(JNK, JNK)} & \text{cov(JNK, MK2)} \\ \text{cov(MK2, AKT)} & \text{cov(MK2, JNK)} & \text{cov(MK2, MK2)} \end{bmatrix}$$



C can be viewed as a linear transformation that can be applied to any vector in the data space. For instance, to transform the 5-min row vector $t_{5\ min}$, which contains $AKT_{5\ min}$, $JNK_{5\ min}$ and $MK2_{5\ min}$ (FIG. 2a) with C, we would multiply C by $t_{5\ min}$ to give us $Ct_{5\ min}$. In general, this transformation takes the starting vector and changes both its magnitude and its direction (see the figure, part **a**). For certain special vectors (*y*) in the data space, however, the transformation of C leaves the direction of the vector unchanged — only its magnitude is altered. This is equivalent to multiplying the vector *y* by a scalar λ: $Cy = \lambda y$ (see figure, part **b**). These special vectors (*y*) of C are called eigenvectors and the corresponding scalars (λ) are called eigenvalues. For the covariance matrix, the direction of an eigenvector identifies a fraction of each measured signal that covaries with all the others. The eigenvector in part **b**, for example, consists mostly of AKT and MK2 activity, with a small contribution from JNK. The corresponding eigenvalue quantifies the strength of the global covariation that is specified by the eigenvector. The eigenvector of C with the largest eigenvalue identifies the direction in the data space that can capture the most information (or variance) from the original observations. Likewise, the eigenvector with the second largest eigenvalue identifies the direction that can capture the next largest amount of information that was not captured by the first eigenvector, and so on. By using only the eigenvectors with the largest eigenvalues and omitting those with the smallest, the data matrix can be approximated by its most salient variations. Importantly, the ranked eigenvalue–eigenvector pairs identify the principal components of the data set, and these form the foundation of principal component analysis (see main text).

AKT, JNK and MK2 are defined in FIG. 1.

Principal components by scores and loadings. Eigenvalues allow the identification and ranking of the importance of combinations of signals to the overall covariance of the data set. However, eigenvalues can be difficult to interpret because they do not directly relate back to the original data matrix (FIG. 1a). Rather than explicitly calculating eigenvalues, therefore, most practical applications of PCA identify principal components by breaking down (or factorizing) the data matrix into a sum of vector products. Just as there are many ways to factorize numbers (for example, $16 = (2 \times 2) + (4 \times 3)$ and $16 = (2 \times 4) + (8 \times 1)$), there are many ways to factorize the data matrix. Therefore, the challenge for PCA is to find a factorization that recapitulates the eigenvalue profile of the data matrix (BOX 1).

How can we guarantee that the factorization converges to the principal components of the data set? First, we define two types of vector, called scores vectors and loadings vectors, that will be multiplied together to form a principal component for the factorization. A scores vector is similar to a column vector (FIG. 2a) and contains the projection of each observation (time point) along the principal component. A loadings vector is similar to a row vector (FIG. 2f) and contains the linear combination of signalling variables that defines the principal component. To select the loadings vector of the first product, we use a search algorithm that identifies the direction in the data space that captures the maximal variance from all signals in the starting data set (FIG. 3a). For the example data set, the first principal component points towards the JNK and MK2 axes because of the strong, concerted activation of these pathways. The search to maximize the variance that is captured is equivalent to selecting the eigenvector with the largest eigenvalue in the covariance matrix (BOX 1). Next, the first scores vector is identified by projecting the observed data set onto the first loadings vector to provide a one-dimensional approximation of the data matrix. Together, the first pair of scores–loadings vectors identifies the first principal component of the data matrix.

The derivation of subsequent principal components takes advantage of another constraint that is imposed on loadings vectors. Like *x*-, *y*- and *z*-axes in spatial dimensions, loadings vectors are required to be linearly independent (or orthogonal) from one another. Because information must be separated into distinct loadings vectors, this allows the second principal component to be calculated iteratively. The variance that is captured by the first principal component is subtracted from the data matrix, and the second loadings vector is optimized to capture as much of the residual variance as possible. For the example data set, the second principal component points strongly towards AKT and away from JNK and MK2, capturing the unique multiphase activation of AKT. The residual is projected on the second loadings vector to calculate the second scores vector that, together with the first principal component, gives a two-dimensional approximation of the data matrix. Additional principal components can continue to be iteratively defined up to the number of observations or signalling variables (whichever is smaller).

Computer implementations of PCA use numerical approaches to maximize the functions defining the loadings vectors and to identify the final number of principal components[23]. The end result is the factorization of the data matrix into a series of scores–loadings

Scores vector
The principal component vector that describes how strongly each observation projects along the principal component.

Loadings vector
The principal component vector that describes how strongly each measured signal contributes to the principal component.

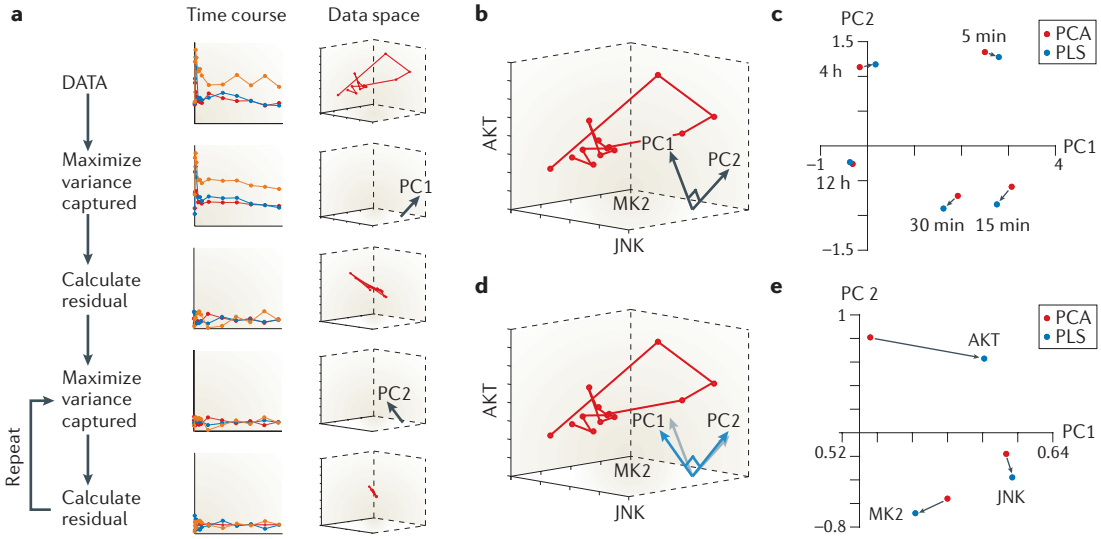


Figure 3 | **Principal components identified by PCA and PLS. a** | Numerical approach for calculating principal components (PC) for principal components analysis (PCA). The first principal component is selected to maximize the variance that is captured, and this contribution is subtracted from the starting data set to calculate the residual. Subsequent principal components maximize the variance that is captured in the residual until only small, uninformative residuals remain. Each step in the procedure is shown as a time-course plot and as vectors in the data space. **b** | Loadings vectors for the first two principal components (PC1 and PC2) identified by PCA for the AKT–JNK–MK2 data space. **c** | Scores vectors projected along the first two principal components identified by PCA (red circles) and partial least squares (PLS; blue circles) for the AKT–JNK–MK2 data space. **d** | Loadings vectors for PC1 and PC2 identified by PLS using the AKT–JNK–MK2 data space to predict AKT-substrate (AKT-sub) phosphorylation (blue) differ from those obtained by PCA (grey). **e** | Scores vectors projected along PC1 and PC2 identified by PCA (red circles) and PLS (blue circles). For parts **c** and **e**, note that the PCA scores–loadings (red circles) are different from the PLS scores–loadings (blue circles) because PLS uses AKT, JNK and MK2 to predict AKT-sub. AKT, JNK and MK2 are defined in FIG. 1.

pairs (principal components) that separate the measurements based on their contribution to the overall variance of the data set (FIG. 3a). For important technical considerations and variants of PCA, see Supplementary information S2 (box).

Deriving biological insight. Principal components define a reduced dimensionality that is optimal for capturing covariance in the data. For covariation to lead to biological understanding, however, it is essential to link optimized dimensions back to measured quantities. The scores and loadings vectors of PCA directly connect the data space to the principal component space. Each loadings vector contains the quantitative contribution of the signalling variables to the principal component. Together, the loadings vectors orientate the signalling axes of the principal component space (FIG. 3b). Reciprocally, each scores vector indicates how strongly the observations (time points) plot out along the corresponding principal component (FIG. 3c). Complex projections of observations in the data space (FIG. 1c) can therefore be visualized more simply by looking at how the same observations project along the principal components of the data set.

For viewing high-dimensional data spaces compactly, PCA has proved to be valuable for a number of biological applications. In neuroscience, PCA has been used successfully to discriminate sensory decisions[24] and odours[25] based on cellular firing patterns. PCA is also a common visualization technique in microarray analysis[26] as well as in fields such as chemical biology[27–29] and plant metabolism[30]. PCA has been shown to qualitatively discriminate apoptotic cell fates based on measured signalling profiles[15]. In systems biology, PCA has also been coupled with sensitivity analysis to help reduce the complexity of mechanistic signalling models[31]. As experiments that quantify networks of protein levels and interactions[5–10] become more common, PCA should become popular as a quick, intuitive technique for inspecting complex signalling data sets.

Unsupervised analysis
A type of computational learning approach in which the expected output is not specified. Hierarchical clustering and PCA are unsupervised analyses.

Partial least squares

PLS is similar to PCA in that it decomposes the data into a set of optimal dimensions based on scores and loadings vectors. What is the difference between PLS and PCA? In PCA, the entire data set is factored into principal components in an unsupervised and essentially automatic manner. Users obtain optimal dimensions that best capture how all of the signals co-vary with respect to one another, but it is not possible using PCA to generate or test a hypothesis in which some parts of the data set are causally related to other parts (for example, signalling responses observed at late times that can be predicted from events at earlier times).

By contrast, PLS identifies optimal, principal-components-based dimensions from a proposed relationship. These relationships are posed by splitting

the signalling variables of the data set into independent variables and dependent variables. The independent variables together form an independent 'block' (X) and the dependent variables form a dependent block (Y) for the proposed relationship: Y = F(X). For example, if several kinase activities and substrate-phosphorylation events involved in apoptosis were measured along with cell-death markers, we might designate the kinase-activity and substrate-phosphorylation events as the block of independent variables and the molecular markers of apoptosis as the block of dependent variables[32]. PLS then identifies a linear solution that relates the independent block to the dependent block.

Most systems biology data sets lack the large number of experimental observations that are needed for calculating a unique solution that estimates the contribution of each signalling variable in the independent block to those in the dependent block. It is in this context that principal components and PLS are most useful. Rather than performing the regression in the original data space, PLS reduces the dimensions to a principal-component space and regresses the independent and dependent principal components. This dimensionality reduction is important because it requires fewer unknown coefficients, and these are constrained better by the observations. These differences from ordinary regression techniques allow PLS to calculate a solution that is biased towards the independent variables that are connected most strongly with the dependent variables.

Covariance-based principal components. The identification of principal components in PLS regression occurs by simultaneous factorization of the independent and dependent blocks into their own scores and loadings vectors, as described above. An important addition for PLS is that the linear relationship between independent and dependent blocks is enforced by having both blocks use the same scores vector (with its length multiplied by a fixed number) to perform the factorization. Therefore, the prediction of phosphorylation of an AKT-sub from AKT, JNK and MK2 activity, for example, will define a principal component space in which the time-point observations of all four signalling variables point in the same direction.

The scores–loadings factorization is calculated iteratively by numerical algorithms, as described above for PCA[23]. The critical modification that distinguishes PLS from PCA is that the loadings vectors are optimized to capture the covariance between the independent and dependent blocks rather than simply the variance. Consequently, the principal components that are defined by PLS will be less efficient at capturing the data in X compared with PCA, but will be more accurate in predicting the data in Y. With the data in FIG. 1b, for example, AKT activity would be important for predicting AKT-sub, even though the magnitude of AKT activation is low compared with the rest of the data set. JNK and MK2 are filtered out of a PLS model that involves AKT-subs because their activation would not co-vary as strongly with AKT-sub. This reprioritization of the principal components based on the dependent variable(s) occurs by rotating both the loadings vectors and the scores vectors (FIG. 3d,e). For further technical considerations and variants of PLS, see Supplementary information S3 (box).

Deriving biological insight. PLS principal components extract the molecular-level evidence in the data that quantitatively support the hypothesis posed by the model. Both the efficacy of PLS for predicting data and the significance of the prediction are determined by the strength of the underlying hypothesis. Molecular-level studies that link sets of independent and dependent variables based on prior biological knowledge (for example, using kinase-activity data to predict substrate phosphorylation) are most likely to reveal direct mechanisms and new insights.

Table 1 | **Comparison of data-driven modelling approaches**

Data-driven model	Model subtype	Optimal dimensions	Strengths	Weaknesses
Clustering	Hierarchical	Dendrogram 'branches'	Simple and unbiased; entire dendrogram can be scanned for assembly of clusters	Clusters must be assembled pairwise; some clusters might lack biological relevance; dendrogram does not simplify the data set
Clustering	*k*-means	Centroids	Clusters are assembled in groups; allows user to specify an expected number of biological classes; centroids provide a simplified representation of the data set	Requires user to specify initial number of centroids and their starting positions; some centroids might lack biological relevance
Principal components analysis (PCA)		Principal components	Simple and unbiased; scores and loadings vectors provide simplified representations of the data set	Cannot pose a hypothetical relationship within the data set; some principal components might lack biological relevance
Partial least squares (PLS)	Classification	Principal components	Allows user to specify an expected set of biological classes without the need for additional data	Class predictions are inherently qualitative; principal components might lack biological relevance when classes are too distantly related to the independent variables
PLS	Prediction	Principal components	Allows user to pose a biological hypothesis; predictions are quantitative	Often requires an additional data set of dependent measurements; assumes a linear relationship between independent and dependent variables

Early proof-of-principle studies showed that different samples could be classified by PLS based on transcriptional[33] and proteomic[34] measurements. However, there was no attempt to link the transcript-loadings and protein-loadings vectors to biologically meaningful differences between sample groups. This might be less important for systems applications that involve biomarkers[35], and PLS has recently been used to discriminate tumour outcomes based on serum profiles[36] and histological scores[37]. The opportunity to identify mechanistic linkages in these studies was limited because of the enormous biological distance between markers and patient response.

In contrast to most classifier-based models, quantitative PLS modelling can identify biological mechanisms when it is applied to complex but well defined signalling networks. We recently constructed a highly predictive PLS model that linked ~1,500 apoptosis measurements to ~8,500 measurements of the apoptotic signalling network, enabling users to predict cell-death responses to molecular perturbations and to identify the roles of important signalling intermediates[32]. As our knowledge of many signal-transduction pathways increases in complexity, quantitative PLS models will become increasingly important (see Supplementary information S4 (box) for other recent examples).

Conclusions

In this tutorial, we have described three different mathematical approaches for deriving data-driven models. All of these techniques analyse complex high-dimensional data spaces to reveal important biological information, but each method has specific advantages and disadvantages, depending on the type of data that was gathered and the biological questions that are being posed (TABLE 1).

One shortcoming of all the data-modelling techniques is an inability to incorporate prior information about the biological system. Network-components analysis (NCA)[38] is a data-modelling approach that was recently developed to include network topology. NCA biases the data-matrix decomposition towards the recognized or estimated connectivity 'strengths' between measured signals. The matrix of connectivity strengths must fulfil certain criteria that might not hold for certain systems, and NCA (unlike PCA) requires more observations than measured variables. Still, NCA adds an alternative approach for analysing signalling networks with detailed topologies.

Last, and most importantly, the modelling approaches described here will not 'fix' a badly designed experiment or a poorly posed hypothesis. Assays with nonlinear readouts are deceptively quantitative. Likewise, treatment conditions must be chosen so that they are informative. Data models are most useful when experimental conditions activate the measured network strongly and differently for each treatment. Furthermore, for PLS modelling, the specified hypothesis must have a strong biological foundation. Using large-scale data sets, it is possible to find covariation with many dependent variables (including those that make no sense). Starting with a hypothesis that is believably mechanistic will increase the probability of extracting new mechanisms from the resulting data model.

The data-driven modelling approaches that are described in this tutorial are common techniques in chemistry, physics and engineering[39,40]. So far, only clustering is widely used in biology, but we expect this to change. Just as sequencing and microarray studies demanded data organization through clustering, large-scale studies of signal transduction[41] will come to require techniques such as PCA and PLS[42]. Without them, the multivariate complexity of contemporary experiments will need to be trimmed by hand down to the level of our intuition. Long lists of disconnected observations might be easier to inspect, but they are unlikely to yield the type of understanding[43] that is promised by systems biology. The capability of data-driven models to analyse large-scale data sets simply, quantitatively and comprehensively ensures that these approaches will soon be standard tools for understanding signal-transduction networks.

1. Janes, K. A. *et al.* A high-throughput quantitative multiplex kinase assay for monitoring information flow in signaling networks: application to sepsis-apoptosis. *Mol. Cell Proteomics* **2**, 463–473 (2003).
2. Kingsmore, S. F. Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nature Rev. Drug Discov.* **5**, 310–320 (2006).
3. Ong, S. E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nature Chem. Biol.* **1**, 252–262 (2005).
4. Irish, J. M., Kotecha, N. & Nolan, G. P. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. *Nature Rev. Cancer* **6**, 146–155 (2006).
5. Gaudet, S. *et al.* A compendium of signals and responses triggered by prodeath and prosurvival cytokines. *Mol. Cell Proteomics* **4**, 1569–1590 (2005).
References 3–5 are excellent reviews on emerging technologies for large-scale studies of signal-transduction networks.
6. Janes, K. A. *et al.* The response of human epithelial cells to TNF involves an inducible autocrine cascade. *Cell* **124**, 1225–1239 (2006).
This study applied data-driven modelling to a large-scale proteomic compendium and showed that tumour necrosis factor induces a regulated, interdependent cascade of autocrine cytokines.
7. Jones, R. B., Gordus, A., Krall, J. A. & MacBeath, G. A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* **439**, 168–174 (2006).
8. Blagoev, B., Ong, S. E., Kratchmarova, I. & Mann, M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature Biotechnol.* **22**, 1139–1145 (2004).
9. Irish, J. M. *et al.* Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217–228 (2004).
10. Natarajan, M., Lin, K. M., Hsueh, R. C., Sternweis, P. C. & Ranganathan, R. A global analysis of cross-talk in a mammalian cellular signalling network. *Nature Cell Biol.* **8**, 571–580 (2006).
The first data-driven analysis of the one- and two-ligand screens for macrophage signalling that was organized by the Alliance for Cell Signaling. The results show how crosstalk is widespread but not uniformly distributed across all ligands and signalling molecules.
11. Bray, D. Reasoning for results. *Nature* **412**, 863 (2001).
12. Janes, K. A. & Lauffenburger, D. A. A biological approach to computational models of proteomic networks. *Curr. Opin. Chem. Biol.* **10**, 73–80 (2006).
13. Pawson, T. Specificity in signal transduction: from phosphotyrosine–SH2 domain interactions to complex cellular systems. *Cell* **116**, 191–203 (2004).
14. Hunter, T. Signaling — 2000 and beyond. *Cell* **100**, 113–127 (2000).
15. Janes, K. A. *et al.* Cue-signal-response analysis of TNF-induced apoptosis by partial least squares regression of dynamic multivariate data. *J. Comput. Biol.* **11**, 544–561 (2004).
16. D'Haeseleer, P. How does gene expression clustering work? *Nature Biotechnol.* **23**, 1499–1501 (2005).
17. Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. & Ruzzo, W. L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987 (2001).
18. Yeung, K. Y., Haynor, D. R. & Ruzzo, W. L. Validating clustering for gene expression data. *Bioinformatics* **17**, 309–318 (2001).
19. Schuldiner, M. *et al.* Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507–519 (2005).
20. Perlman, Z. E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).

21. Bjorklund, M. *et al.* Identification of pathways regulating cell size and cell-cycle progression by RNAi. *Nature* **439**, 1009–1013 (2006).
22. Gilchrist, M. *et al.* Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature* **441**, 173–178 (2006).
23. Geladi, P. & Kowalski, B. R. Partial least-squares regression — a tutorial. *Anal. Chim. Acta* **185**, 1–17 (1986).
The classic review on partial least squares. The tutorial is presented in the context of spectroscopy, but the analytical approaches can be applied equally well to biological systems.
24. Briggman, K. L., Abarbanel, H. D. & Kristan, W. B. Jr. Optical imaging of neuronal populations during decision-making. *Science* **307**, 896–901 (2005).
25. Hallem, E. A. & Carlson, J. R. Coding of odors by a receptor repertoire. *Cell* **125**, 143–160 (2006).
26. Butte, A. The use and analysis of microarray data. *Nature Rev. Drug Discov.* **1**, 951–960 (2002).
27. Tanaka, M. *et al.* An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol.* **3**, e128 (2005).
28. Knight, Z. A. *et al.* A pharmacological map of the PI3-K family defines a role for p110α in insulin signaling. *Cell* **125**, 733–747 (2006).
29. Haggarty, S. J., Koeller, K. M., Wong, J. C., Butcher, R. A. & Schreiber, S. L. Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. *Chem. Biol.* **10**, 383–396 (2003).
30. Hirai, M. Y. *et al.* Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **101**, 10205–10210 (2004).
31. Liu, G., Swihart, M. T. & Neelamegham, S. Sensitivity, principal component and flux analysis applied to signal transduction: the case of epidermal growth factor mediated signaling. *Bioinformatics* **21**, 1194–1202 (2005).
32. Janes, K. A. *et al.* A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* **310**, 1646–1653 (2005).
33. Nguyen, D. V. & Rocke, D. M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, 39–50 (2002).
34. Jessen, F., Lametsch, R., Bendixen, E., Kjaersgard, I. V. & Jorgensen, B. M. Extracting information from two-dimensional electrophoresis gels by partial least squares regression. *Proteomics* **2**, 32–35 (2002).
These three papers are the first applications of PLS for classification (references 33 and 34) and prediction (reference 32) using biological networks.
35. Hood, L., Heath, J. R., Phelps, M. E. & Lin, B. Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**, 640–643 (2004).
36. Goncalves, A. *et al.* Postoperative serum proteomic profiles may predict metastatic relapse in high-risk primary breast cancer patients receiving adjuvant chemotherapy. *Oncogene* **25**, 981–989 (2006).
37. Linke, S. P., Bremer, T. M., Herold, C. D., Sauter, G. & Diamond, C. A multimarker model to predict outcome in tamoxifen-treated breast cancer patients. *Clin. Cancer Res.* **12**, 1175–1183 (2006).
38. Liao, J. C. *et al.* Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA* **100**, 15522–15527 (2003).
This paper is the first introduction of NCA and its proof-of-principle application to biological networks.
39. Martens, H. & Martens, M. *Multivariate Analysis of Quality: An Introduction* (John Wiley & Sons, Chichester, 2001).
40. Grossman, R. L., Kamath, C., Kegelmeyer, P., Kumar, V. & Namburu, R. *Data Mining for Scientific and Engineering Applications* (Kluwer Academic, Dordrecht, 2001).
41. Gilman, A. G. *et al.* Overview of the Alliance for Cellular Signaling. *Nature* **420**, 703–706 (2002).
42. Pradervand, S., Maurya, M. R. & Subramaniam, S. Identification of signaling components required for the prediction of cytokine release in RAW 264.7 macrophages. *Genome Biol.* **7**, R11 (2006).
43. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
44. MacQueen, J. B. in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* 281–297 (University of California Press, Berkeley, 1967).
45. Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum, New York, 1981).

Acknowledgements
The work cited in this review was supported by grants from the National Institutes of Health to M.B.Y. and an American Cancer Society postdoctoral fellowship to K.A.J.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Michael B. Yaffe's homepage:
http://web.mit.edu/biology/www/facultyareas/facresearch/yaffe.shtml

SUPPLEMENTARY INFORMATION
See online article: S1 (box) | S2 (box) | S3 (box) | S4 (box)
Access to this links box is available online.