Empirical Modeling:
Current and Emerging Techniques

Ian Flood,

Research Foundation Professor,

Holland Professor,

Rinker School, College of Design, Construction and Planning,

University of Florida

(flood@ufl.edu)

Topics Covered:

- What is Empirical Modelling?
- Modelling Systems:
  - Interface Structure
  - Internal Structure
  - Development Schemes
- Empirical Modelling Methodology:
  1. Strategizing
  2. Collation and Evaluation of Data
  3. Model Development
  4. Model Evaluation and Final Selection
  5. Final Validation
  6. Implementation and Review
- Challenges and Emerging Solutions

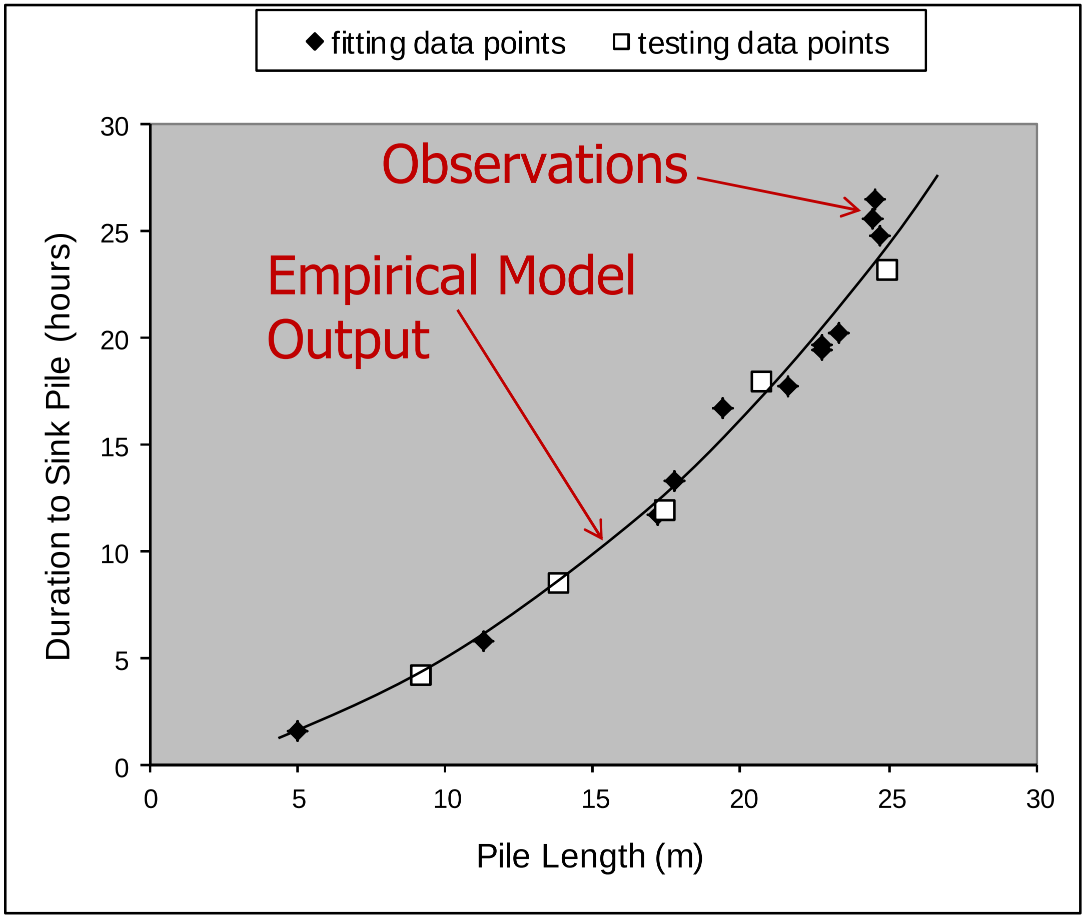
What is Empirical Modelling?

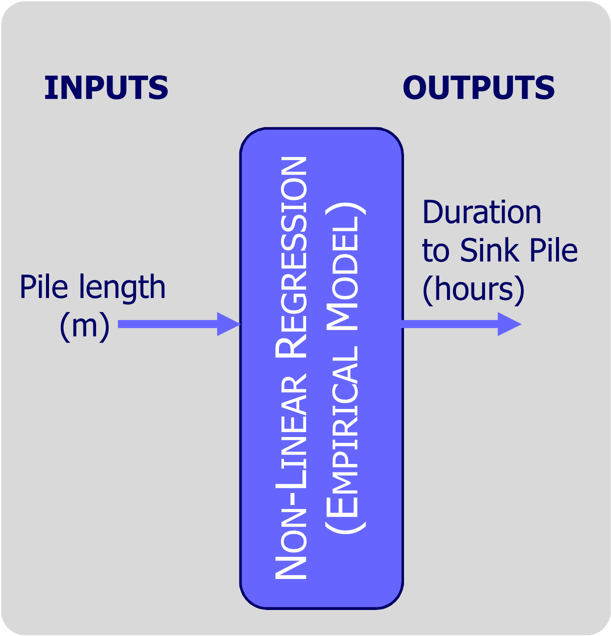
What is Empirical Modelling?

- Mathematical Models:
  - Abstractions of systems described using mathematical language:
    - Algebra
    - Statistics
    - Logic
    - Algorithms, etc…
- Usage (in all branches of science and technology):
  - Experimental tools used to extend our understanding of a system;
  - Predictive tools used in:
    - Decision-making
    - Automated systems control, etc…
- Important dichotomy in their development:
  - **Theoretical** - built from principles that govern the behaviour of the system
  - **Empirical** - developed by emulating/capturing characteristic behaviour observed in a system

What is Empirical Modelling?

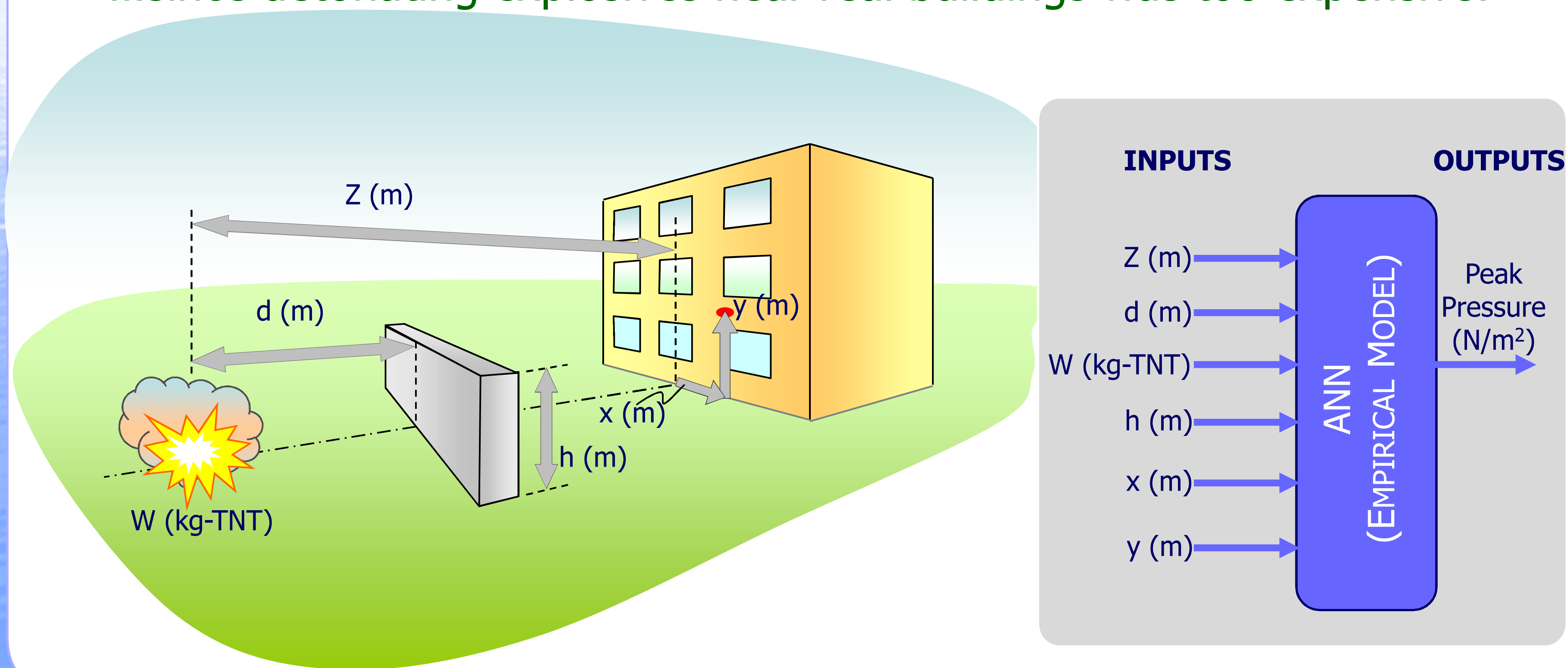
- A model developed from observations of the type of system under investigation
  - based on some measure of the quality of its output (replication, utility)
- simple example:





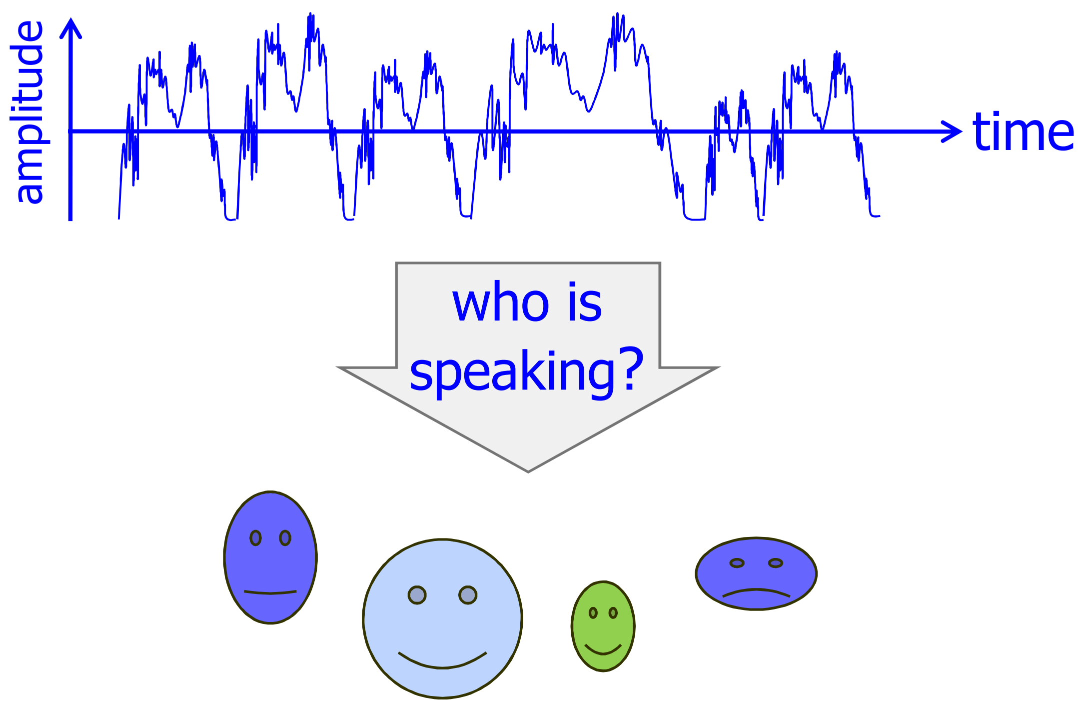
What is Empirical Modelling?

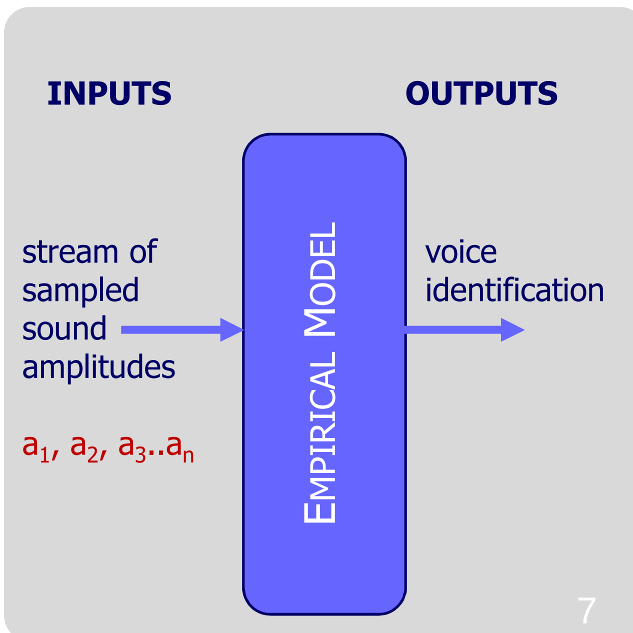
- …**or** one developed from observations of an **analog** of the system under investigation (a model of a model):
  - eg: neural net (ANN) for predicting bomb blast pressures on a structure … since the simulation model was too slow for use by engineers.[2]
  - used the simulation model to generate training patterns for the ANN …since detonating explosives near real buildings was too expensive!



What is Empirical Modelling?

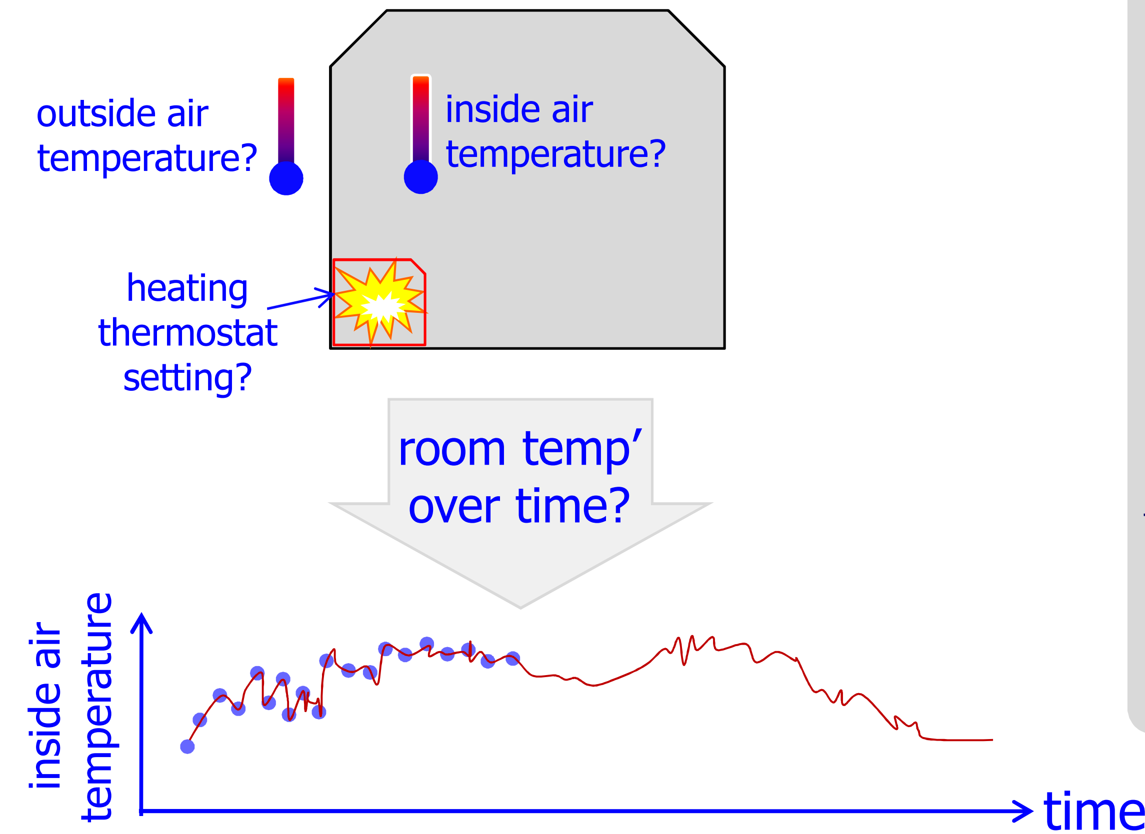
- …can receive **streams** of input (time-wise input)
- …and/or generate streams of outputs:
  - eg: for voice identification, a stream of inputs representing sound amplitude are integrated by the model to generate a single conclusion:

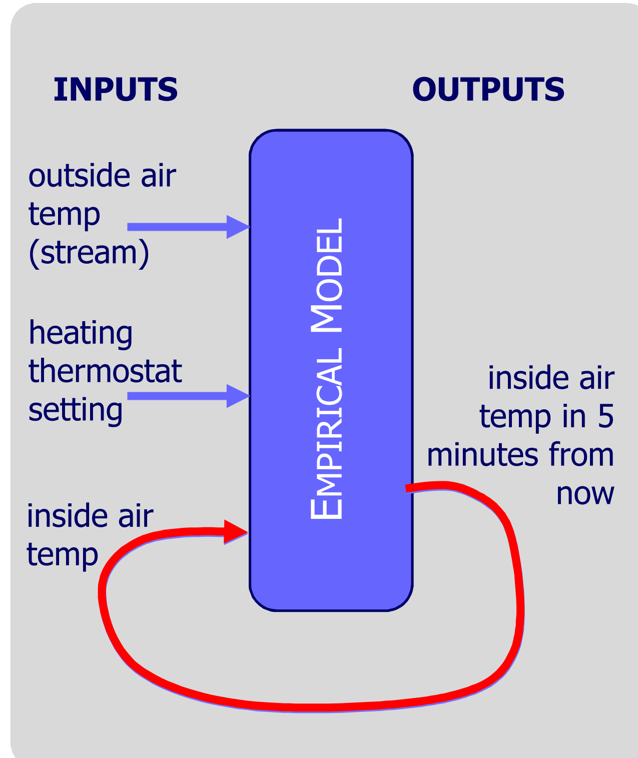




What is Empirical Modelling?

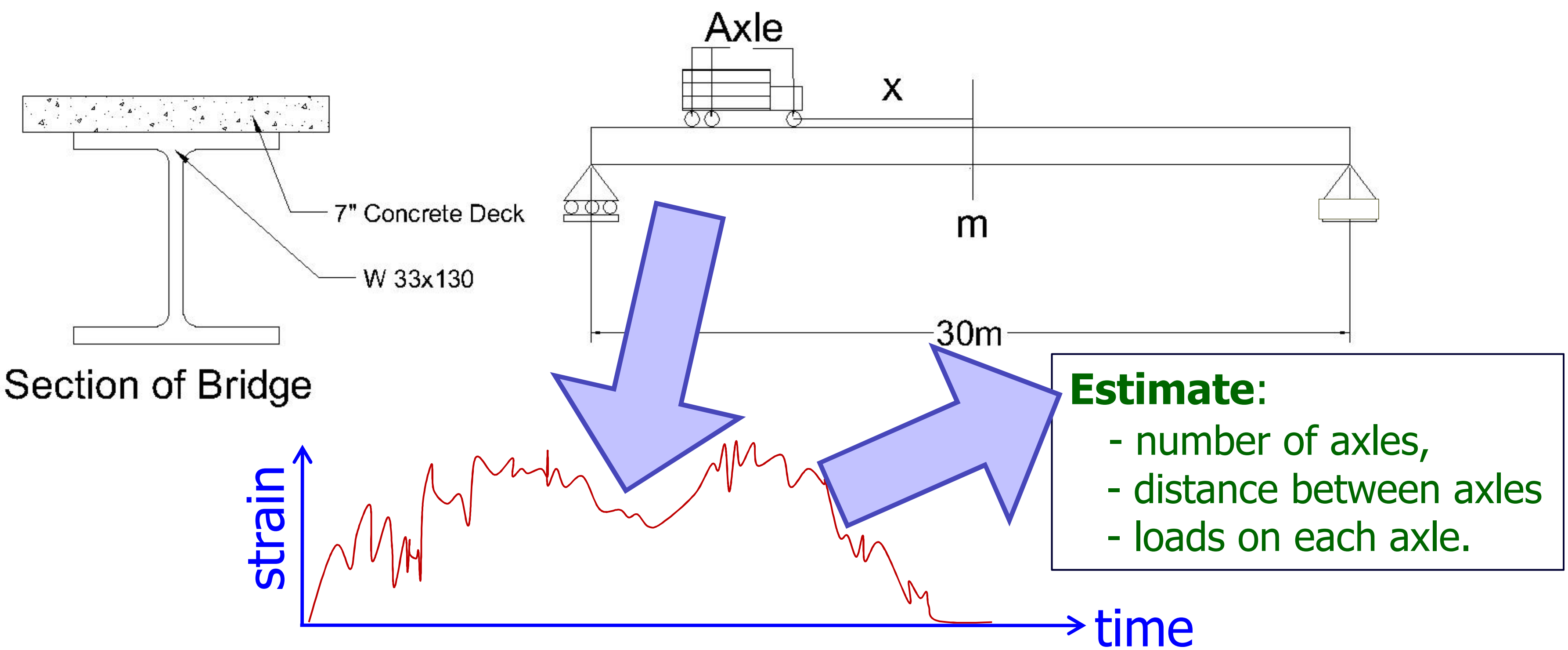
- ...can operate **recursively** (self-feedback), a special case of streaming input and streaming output:
  - eg: predicting room temperature over time:





What is Empirical Modelling?

- …can have a rich **internal structure**:
  - maybe developed directly by the modeler (handcrafted & modular)
  - …or developed automatically (such as by a genetic algorithm)
  - eg: determining truck attributes from the strain they induce on a bridge

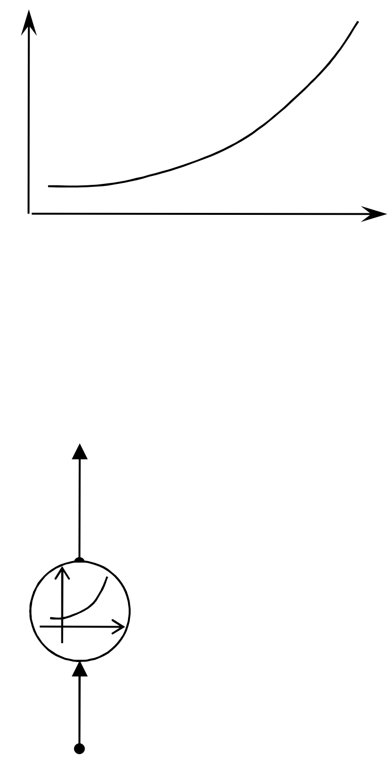


What is Empirical Modelling?

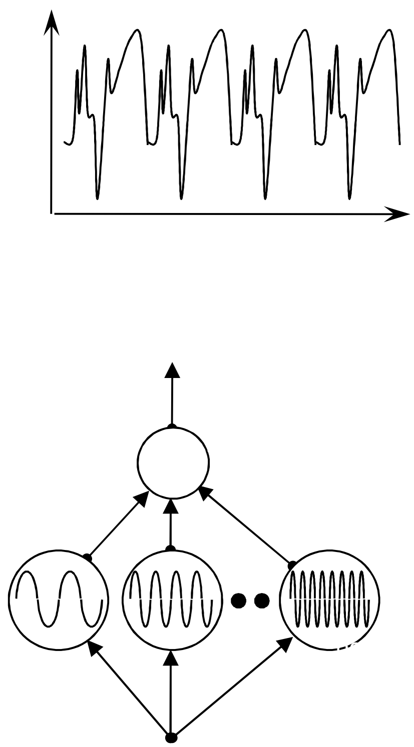
- Why use Empirical instead of Theoretical modelling?
  - Many problems have no theory or limited theory (are poorly understood)
  - Theoretically derived models can be computationally expensive where an empirically derived model can provide rapid solutions
- Traditional view of limitations of Empirical vs. Theoretical:
  - Empirical models are black box devices
    - provide limited understanding of the rationale behind their solutions
  - Empirical models are less accurate
  - Empirical models are limited in scope by the set of observations used in their development
    - can only interpolate (not extrapolate);
    - are not extensible to new configurations of the problem
  - Experience a geometric explosion in the size of the data set required for training with respect to the number of independent variables
- However, these are not fundamental limitations, but rather challenges for empirical modelling **(discuss later)**.

What is Empirical Modelling?

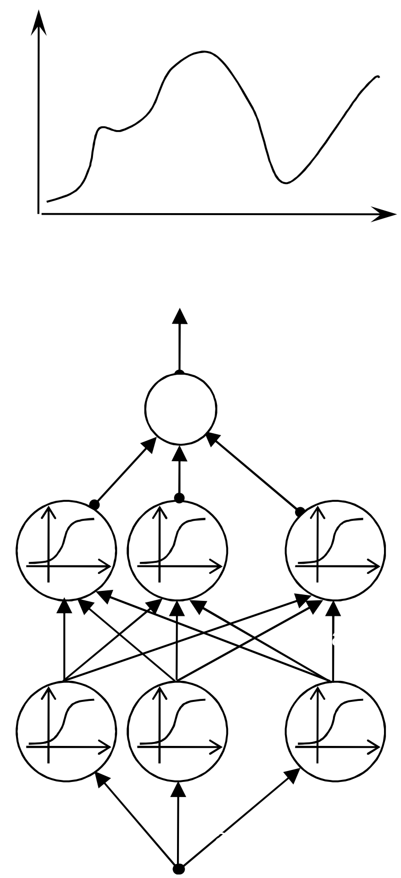
Currently, most applications use relatively simple direct mapping models:



Regression



Fourier Series



Artificial Neural Network

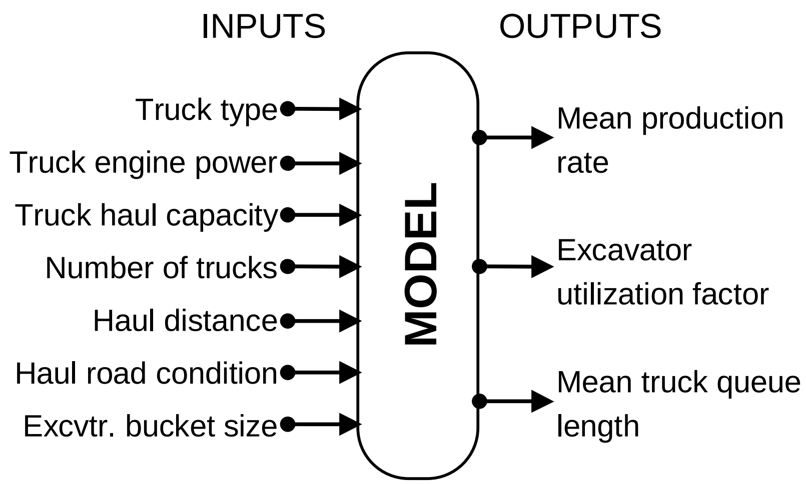
Modelling Systems

Modelling Systems: Introduction

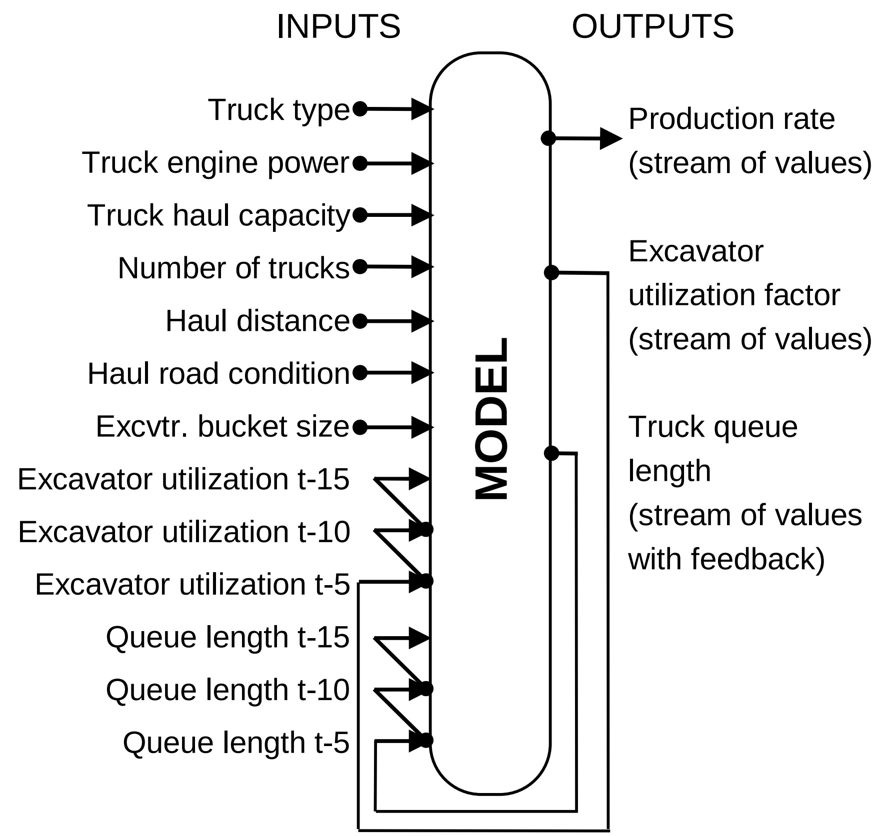
- Empirical methods can be used to develop models far more sophisticated than the above simple mapping devices…
- …greatly extending the scope and performance of applications.
- For any empirical modelling study, there are two broad issues that must always be considered:
  - **structure** of the model, which comprises two aspects:
    - interface (input and output)
    - internal structure
  - **development (training)** scheme used to develop the model, for which there are many types. Typically iterative in nature, in which case a common dichotomy:
    - supervised
    - unsupervised
    - …But also direct derivation

Modelling Systems: Interface Structure

Static vs. dynamic modelling interface structures. *Example: modelling the performance of an excavator-truck earthmoving system:*



Static Model



Dynamic Model

Modelling Systems: Interface Structure

- Important decision is determining the input variables to include:
  - obviously only include those that are significant in terms of affecting the output values
  - however, often these are not known at the outset of the study. Determination may be by:
    - expert judgement,
    - published work,
    - experimentation with combinations of input variables.
- Some input variables my be relevant but not significant…
- …while others may have overlap/redundancy between each other.
- Consider: '*truck type*', '*engine power*', '*and haul capacity*':
  - The first may implicitly define the 2nd and 3rd and imply additional important other information such as '*truck weight*'…
  - …however, '*truck type*' is an enumerative type with no progressive order of values:
    - this introduces a discontinuity in the solution function which can be problematic for many model types (e.g. neural nets)

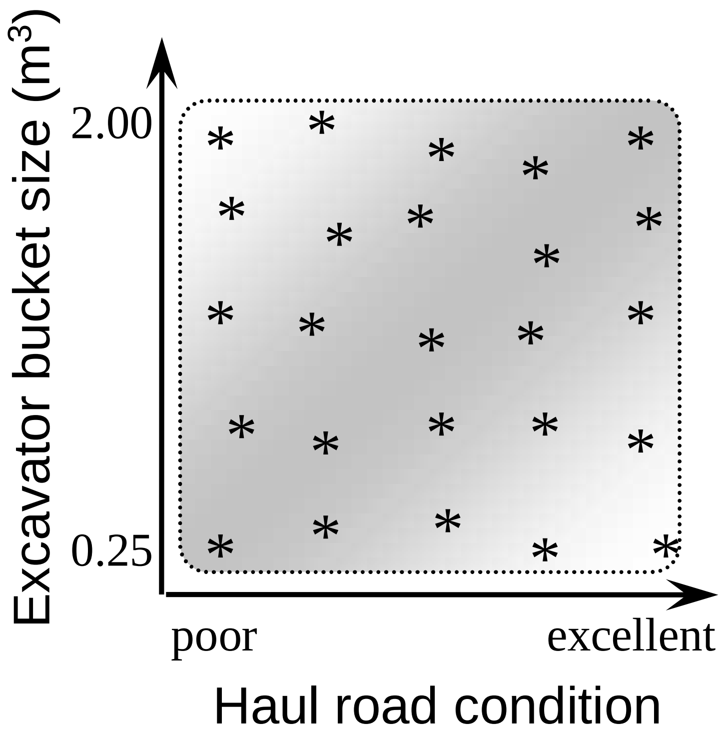
Modelling Systems: Interface Structure

- Obviously cannot include variables for which data are not available
- For many modelling approaches, the number of observations required tends to **increase exponentially** with the number if input variables…
- …but not where there is correlation between those input variables:

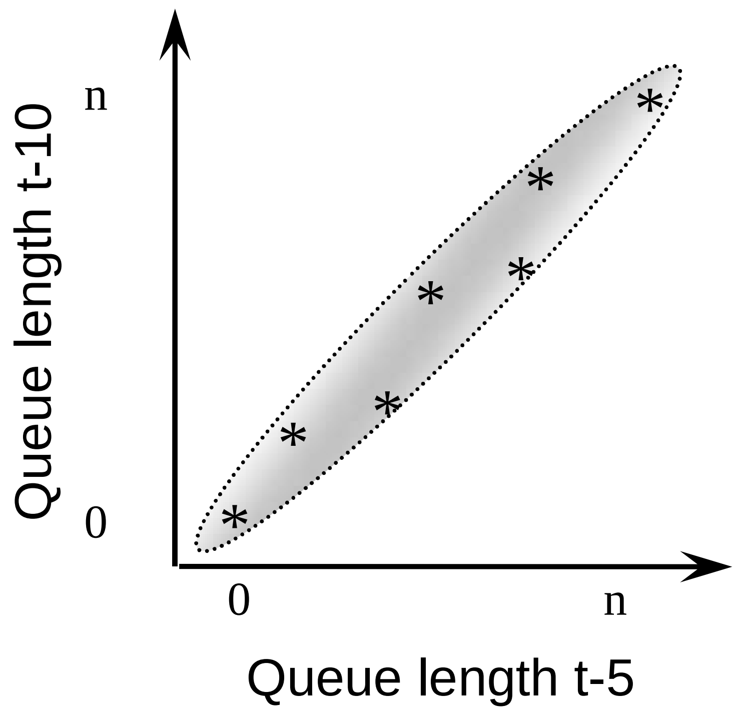
Modelling Systems: Interface Structure

Relationship between correlation and number of training patterns required

Key: = problem domain * = observation data points



Uncorrelated Input Variables



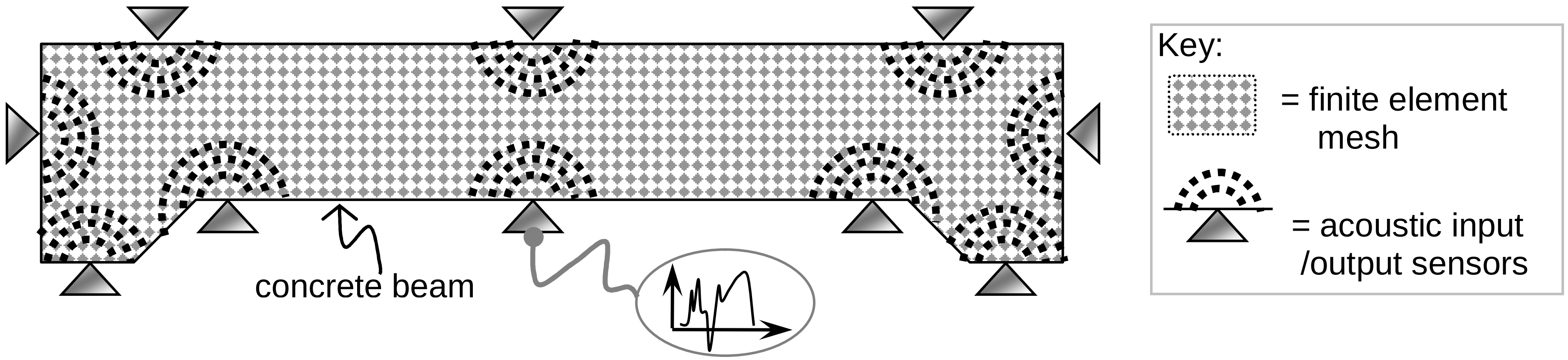
Correlated Input Variables

Modelling Systems: Internal Structure

- Internal structure of a model can be:
  - Defined implicitly by the type of model used (e.g. regression)
  - Derived automatically by the model development algorithm
  - Hand-crafted by the model developer (many neural net studies).
- Most studies concerned with determining a set of output values that correspond to a set of input values…
- …however, a potentially powerful yet under exploited application focuses on the resultant internal structure following model development:
  - could tell us something about the structure of the problem being studied, or
  - provide a set of rules or principles that can be used to solve related problems.
    - consider the problem of detecting the location of reinforcing steel in a concrete structure from its acoustic responses across multiple positions:

Modelling Systems: Internal Structure

Simulated evolution based development of internal structure of model until it replicates the behaviour of the observed system (using FEM elements)



Modelling Systems: Development Schemes

- Model development includes:
  - Determining an appropriate internal structure:
    - e.g. neural net layers, nodes in each layer, connectivity and activation function
  - Determining the values for the models attributes/coefficients:
    - e.g. neural net weights and base values
- Ideally this will all be determined automatically…
- …often the internal structure has to be hand crafted:
  - alternative structures may be tested using sensitivity analyses of the performance.
- Most model development algorithms operate iteratively:
  - Progress is measured and directed by an objective function, e.g:
    - to minimize errors when attempting to replicate a set of observed input to output mappings (supervised training)
    - to maximize utility such as the production rate in an excavation system (unsupervised training)

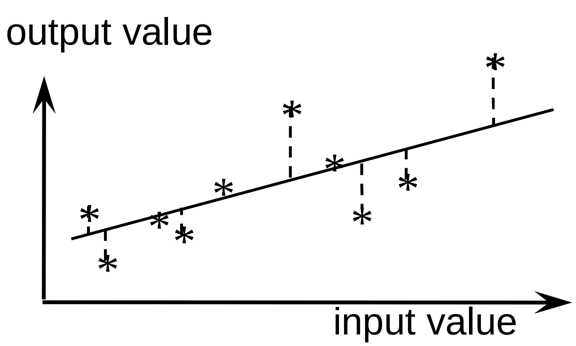
Modelling Systems: Development Schemes

- Performance usually requires the model to be evaluated for a different set of examples than that used for training:
  - This set must be fully **representative** of the types of problem to which the model will be applied;
  - Performance should be measured in a way that is relevant to the way the model be used:
    - e.g. a dynamic model will be used iteratively and may experience compounding errors, so the testing should be made for complete run sequences, not just the first iteration;
    - this is illustrated in the following:

Modelling Systems: Development Schemes

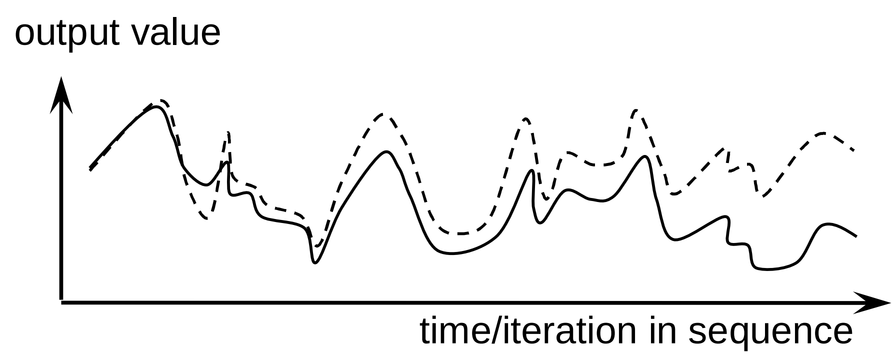
Comparison of Errors for Static versus Dynamic Models

Key: * = target outputs;
= model's output;
= errors.



Errors for Multiple Example Problems in a Static Model

Key: = target for one example problem;
= model's output for the problem;
= error.



Errors for 1 Example Problem in a Dynamic Model

Empirical Modelling Methodology

- Development and implementation of an empirical model must follow a rigorous set of procedures to ensure validity:
  - Can recognize 6 steps common to all studies:

Empirical Modelling Methodology: Step 1: Strategizing

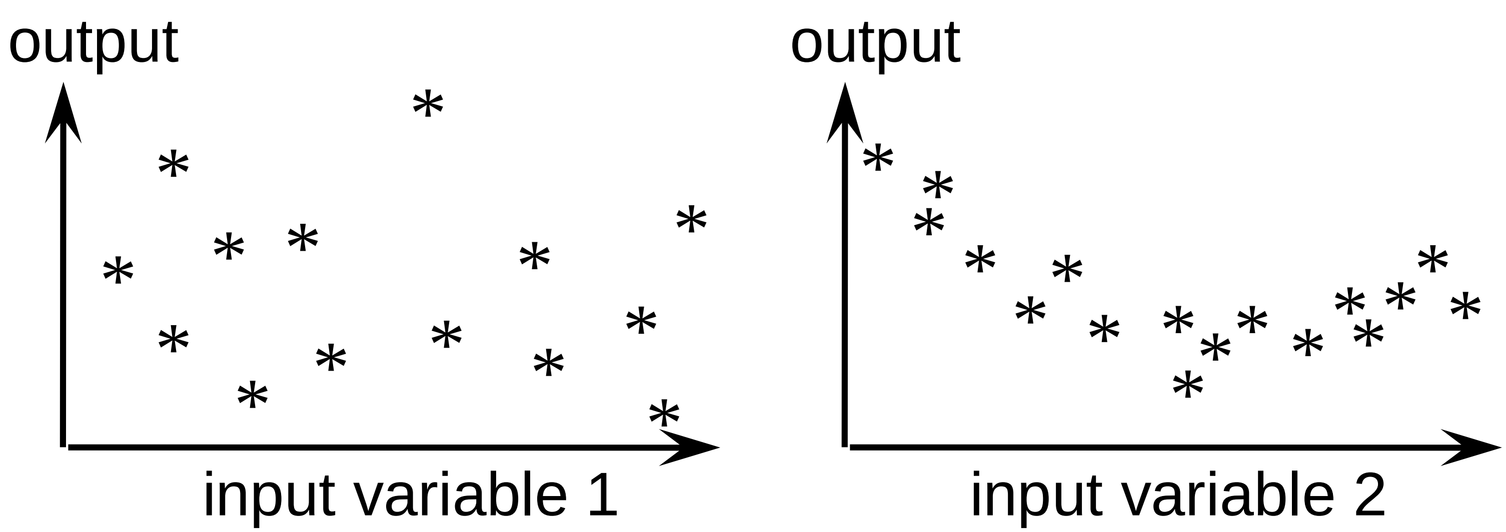
- The aims of strategizing are:
  - Identify the objectives of the study
  - Determine a likely appropriate set of input variables
  - Gain a feel for how the system being modelled responds to different variables, e.g.
    - Linear vs non-linear;
    - Stochastic vs. deterministic, etc…
- Questions to be answered at this stage:
  - What type and structure to adopt for the model?
  - What development algorithm to adopt?
  - What is the objective function?
  - What are the sources for information and what new studies will be required to acquire the necessary data for training, model selection, and validation.
- A pilot study may be required to help answer these questions and to determine feasibility.

Empirical Modelling Methodology: Step 1: Strategizing

- Gaining a graphical understanding of the problem can be extremely useful at this stage:
  - Plotting each output variable against each of the input variables:
    - Relevance of each input variable
    - Complexity of the response of the system – e.g. linear vs. non-linear
    - Existence of unexplained variance in the response of the system
  - Plotting each of the input variables against each other
    - Determine correlation between inputs
  - Both approaches illustrated in the following two figures:

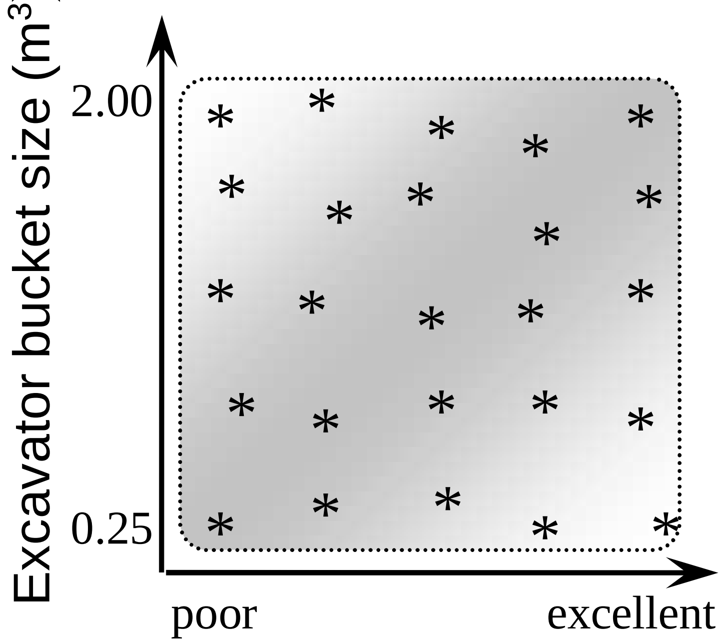
Empirical Modelling Methodology: Step 1: Strategizing

Plotting **Output** vs. **Input** for a Set of Existing Observations of the Response of a System

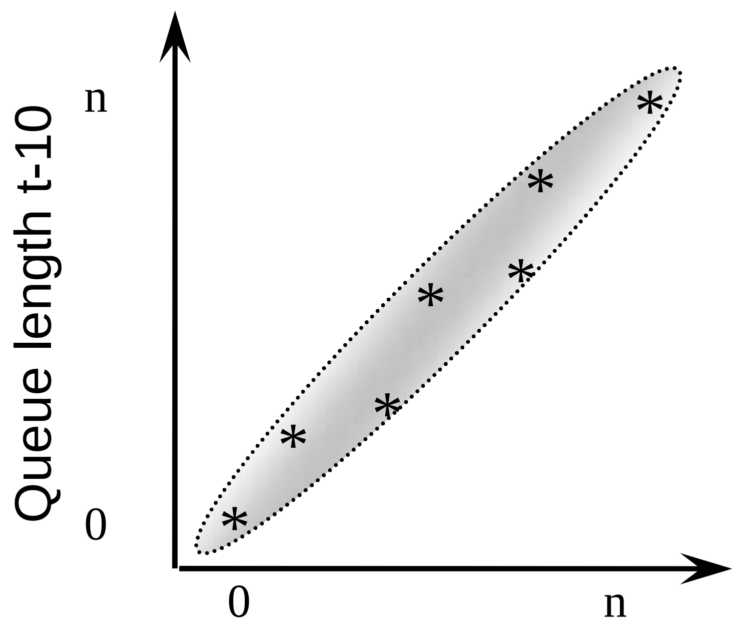


Empirical Modelling Methodology: Step 1: Strategizing

Plotting **Input** vs. **Input** for a Set of Existing Observations of the Response of a System



Haul road condition



Queue length t-5

Uncorrelated

Correlated

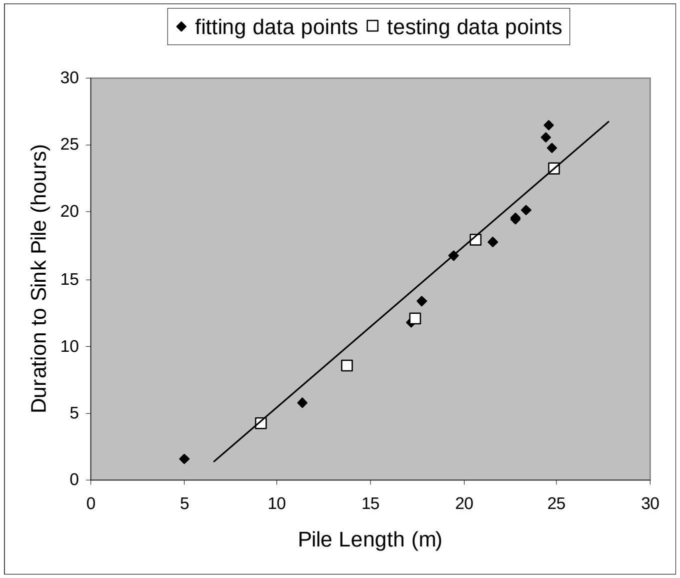
Empirical Modelling Methodology: Step 1: Strategizing

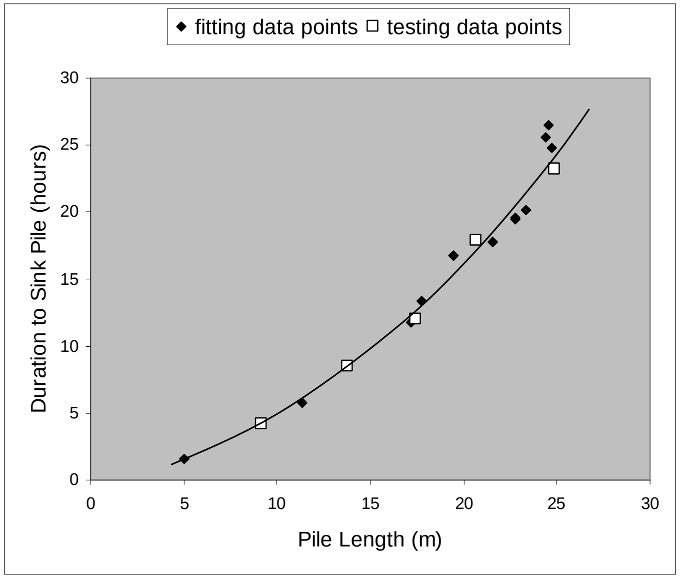
- Understanding a problem is critical to selecting an appropriate type of model:
  - Consider the following:

Empirical Modelling Methodology: Step 1: Strategizing

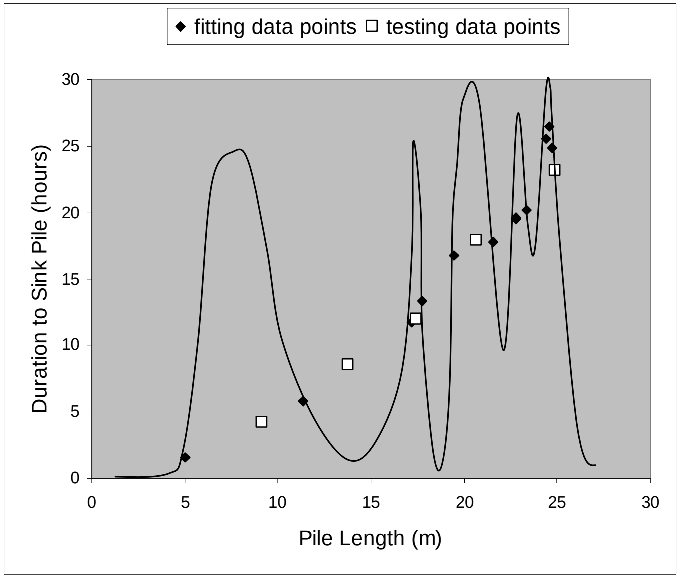
Fitting Functions of Different Complexity to a Set of Observations

Straight Line





Simple Curve



Convoluted Curve
(many degrees of freedom)

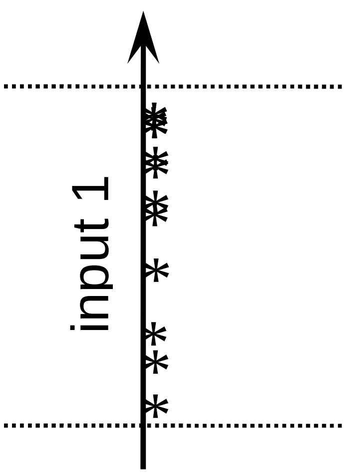
Empirical Modelling Methodology:
Step 2: Data Collation and Evaluation

- Most empirical modelling studies require 3 sets of data:
  - Training data set – used to develop the model
  - Testing data set – used to compare the performance of alternative models and variants of the model
  - Validation data set – used to make a final validation of the performance of the final model
- Each of these data sets must be assessed or designed to make sure that it is representative of the problem.
- An appropriate data set **size** is dependent on:
  - complexity of the problem…
  - …and may be determined through sensitivity analyses
- An appropriate data set **distribution** is dependent on:
  - form of the problem (some areas may require higher density of observations)…
  - …and may be assessed using graphical plots:

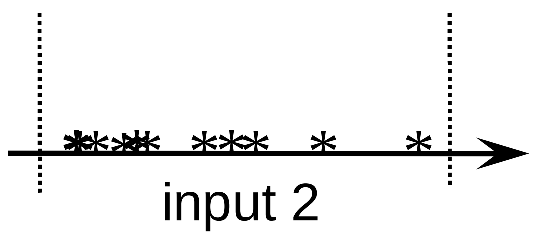
Empirical Modelling Methodology:
Step 2: Data Collation and Evaluation

Distribution of 12 Observations Across the Problem Domain

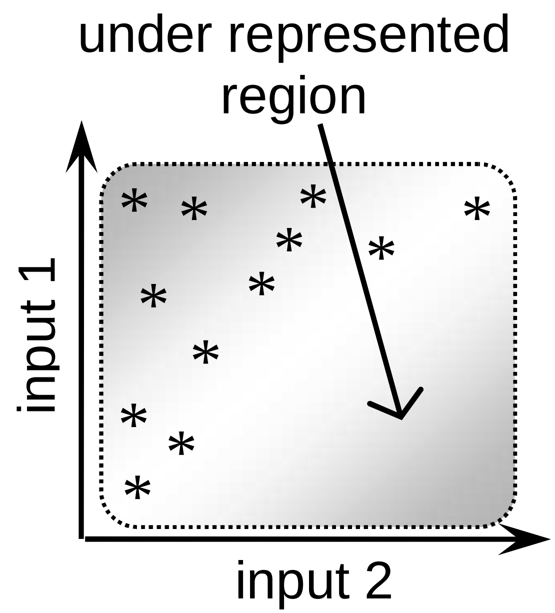
Key: = problem domain; * = observation data points



Input 1



Input 2



Input 1 vs. Input 2

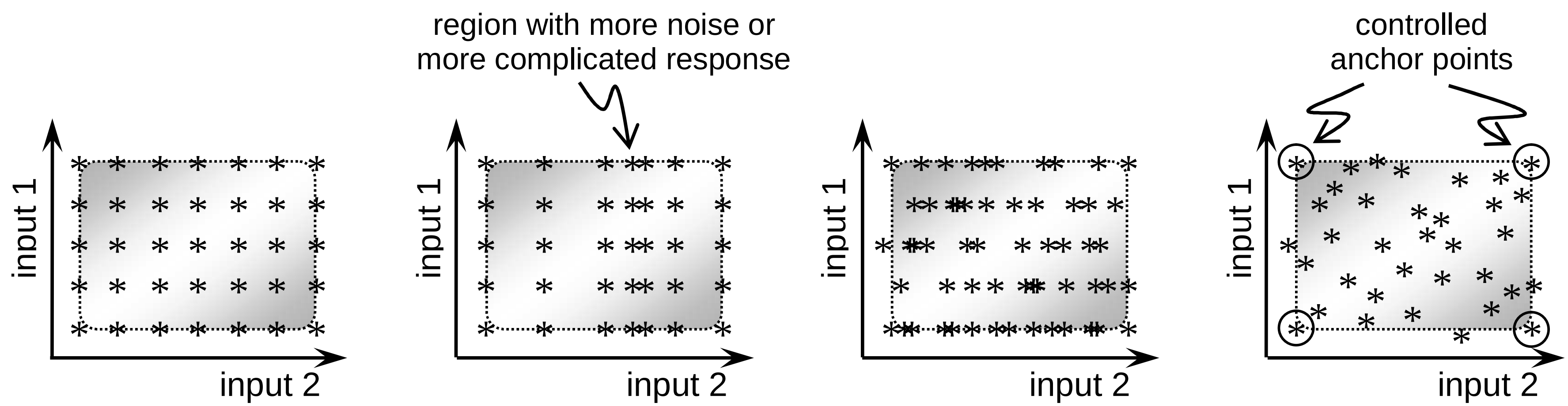
Empirical Modelling Methodology:
Step 2: Data Collation and Evaluation

- Where you can control the set of observations used for modelling:
  - Make sure all observations cover the entire problem domain
  - Many layout schemes are available, but make sure appropriate for the problem at hand
  - If use a regular grid, the testing and validation sets should normally still be randomly positioned

Empirical Modelling Methodology:
Step 2: Data Collation and Evaluation

Distribution of Observations Collected from Controllable Systems

Key: ⬚ = problem domain; * = observation data points



Regular grid. Variable density grid. Regular & random. Random.

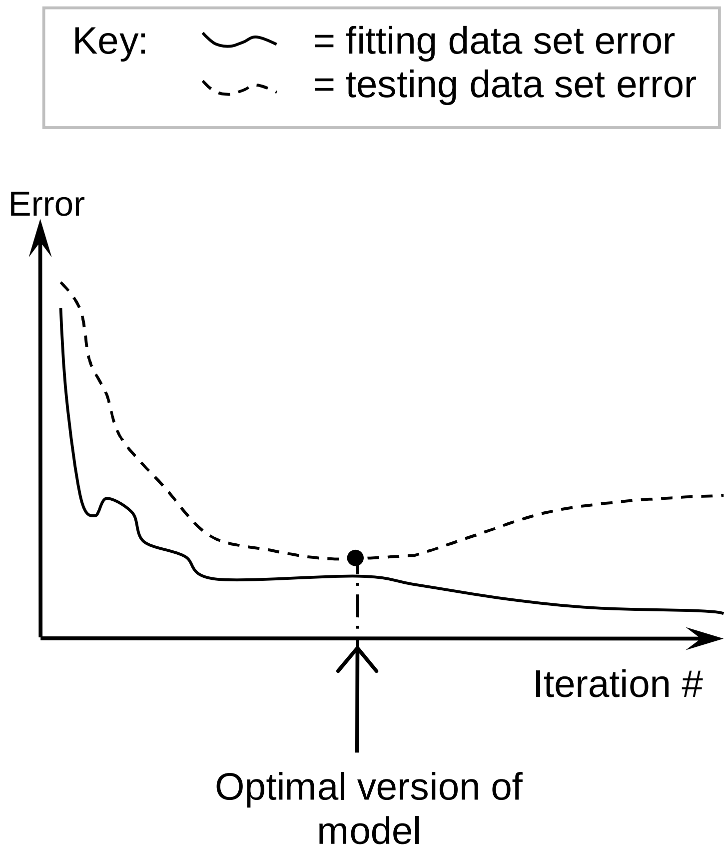
Empirical Modelling Methodology:
Step 3: Model Development

- Whereas step 1 (strategizing) identified a conceptual design for the model,…
- …step 3 develops the finalized design for the model.
- Progress in training can be monitored for both the training data set and the testing data set:
  – Training terminates where the testing data set performs optimally…
  – …going beyond this point can cause ‘overtraining’ (memorization);
  – consider the following:

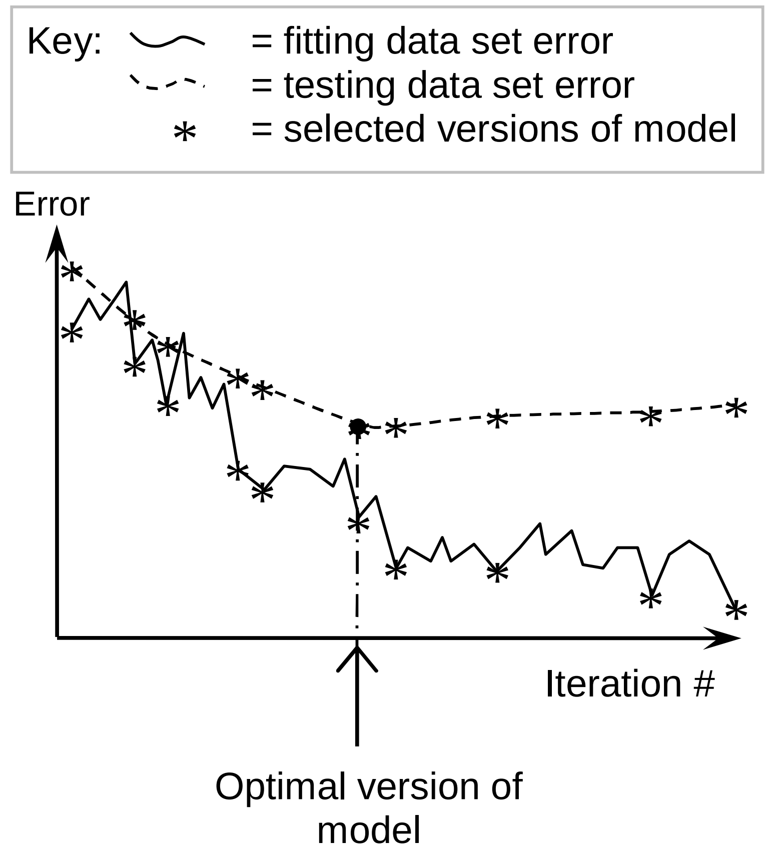
Empirical Modelling Methodology:

Step 3: Model Development

Progress in Model Development for Studies that use Search Algorithms



Error gradient descent applied to model coefficients.



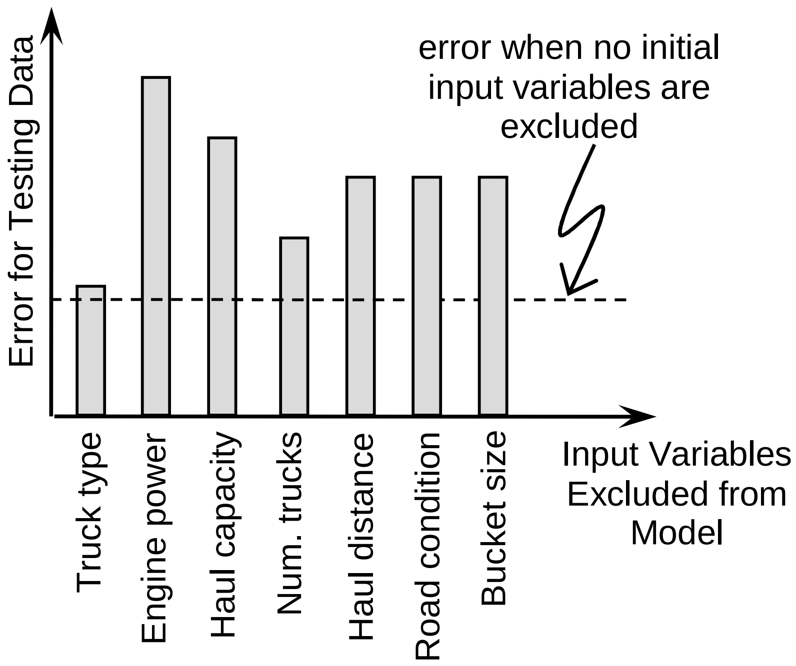
Simulated evolution applied to model structure.

Empirical Modelling Methodology:
Step 3: Model Development

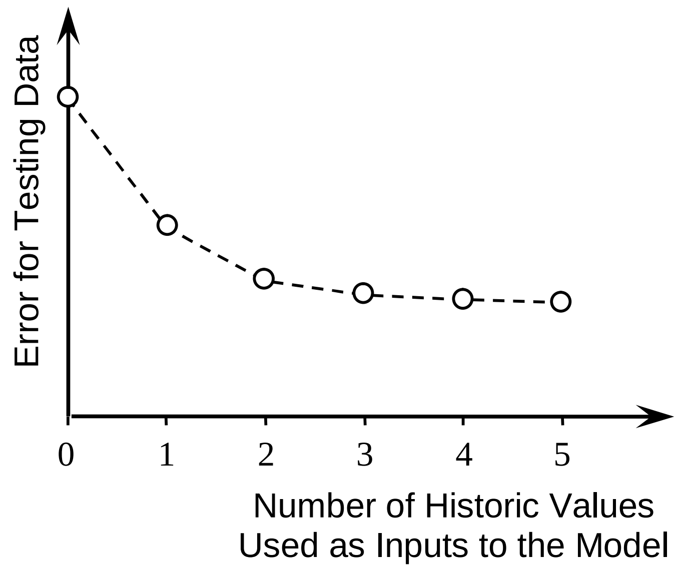
- Some model parameters are not adjusted by the model development/training algorithm, e.g.:
  - Number of layers in a neural net
  - Number of neurons in a layer of a neural net
  - Number of observations used for training
  - Set of input variables used, etc…
- These will need to be adjusted manually, and in a methodical way:

Empirical Modelling Methodology:
Step 3: Model Development

Searching for an Input Configuration for a Model (Excavation) that Minimizes the Testing Error



Alternative sets of input variables.



Alternative numbers of historic input values.

Empirical Modelling Methodology:

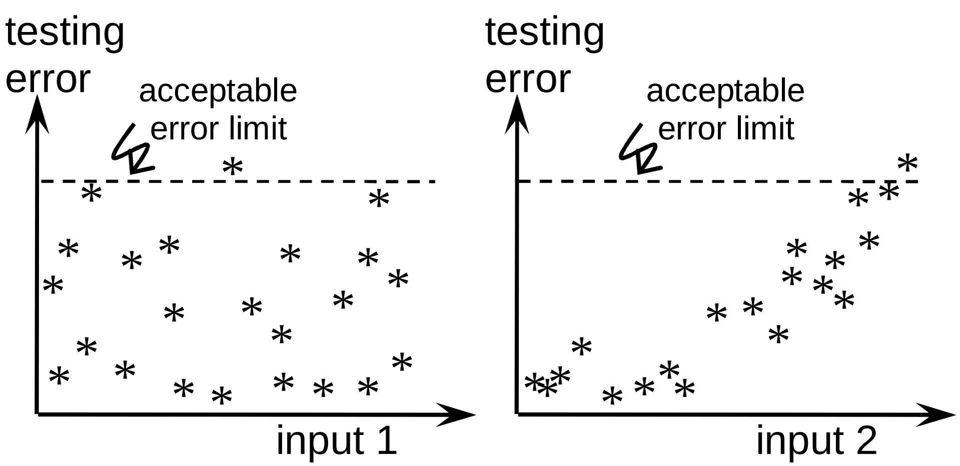
Step 4: Model Evaluation and Final Selection

- The study at this stage may have generated several candidate models
- These should be thoroughly evaluated using the **testing data set** to select the best
- Performance should not be based just on the objective function…
- …the performance across the problem domain should also be considered to look for consistency in performance:

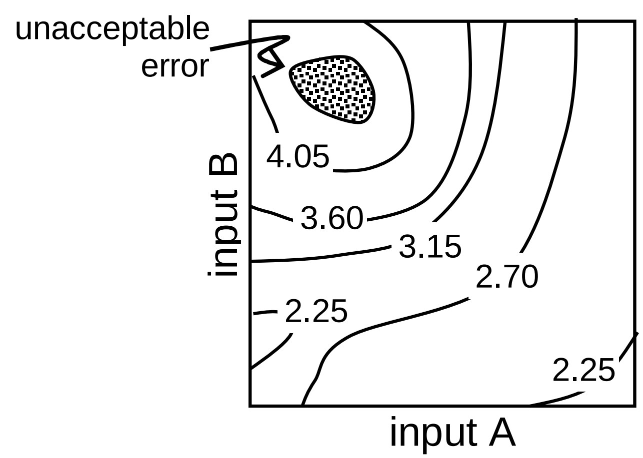
Empirical Modelling Methodology:

Step 4: Model Evaluation and Final Selection

Evaluating Error across the Problem Domain



Error plotted against input variable.



Error plotted as a contour map.

Empirical Modelling Methodology:

Step 5: Final Validation

- At this stage we have the final version of the model
- This needs to be validated:
  - to get an accurate assessment of its performance
  - to see whether further development may be required
- Should not use the testing data set for this as the model may have some bias towards it
- Requires a 3rd independent data set.

Empirical Modelling Methodology:

Step 6: Implementation and Review

- Education of end-users:
  - Collection and organization of input data to ensure model validity
  - Interpretation of the output from the model
  - Usage of the model for problem solving
- Where possible, feedback from use to continue validation and improvement of the model.

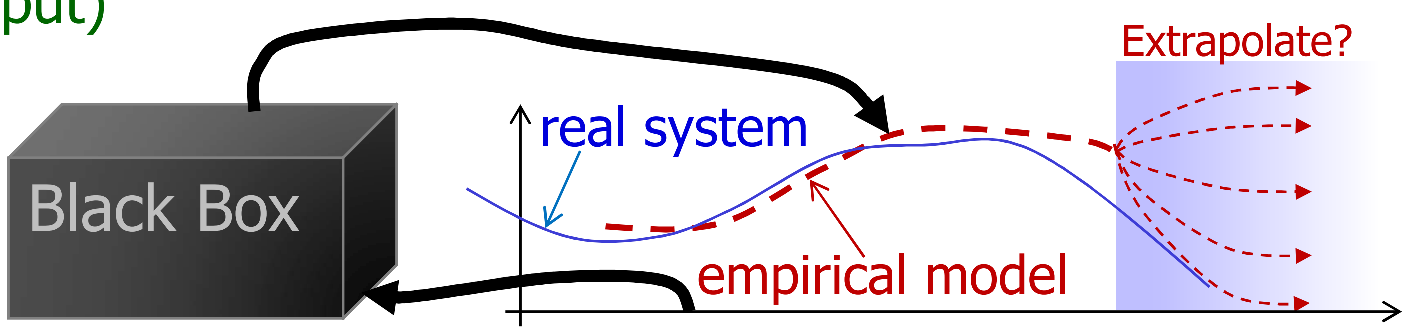
Challenges and Emerging Solutions

Disadvantages of empirical modeling:

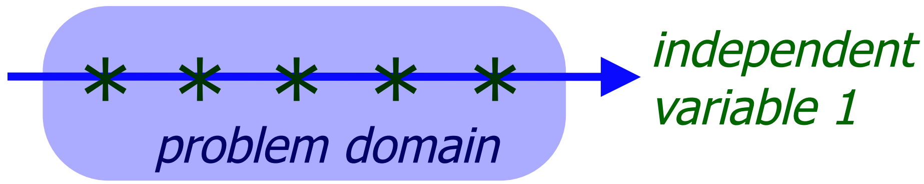
- many disadvantages are cited…
- …however, these are not absolute limitations but rather **challenges** that we are attempting to overcome:

Challenges:

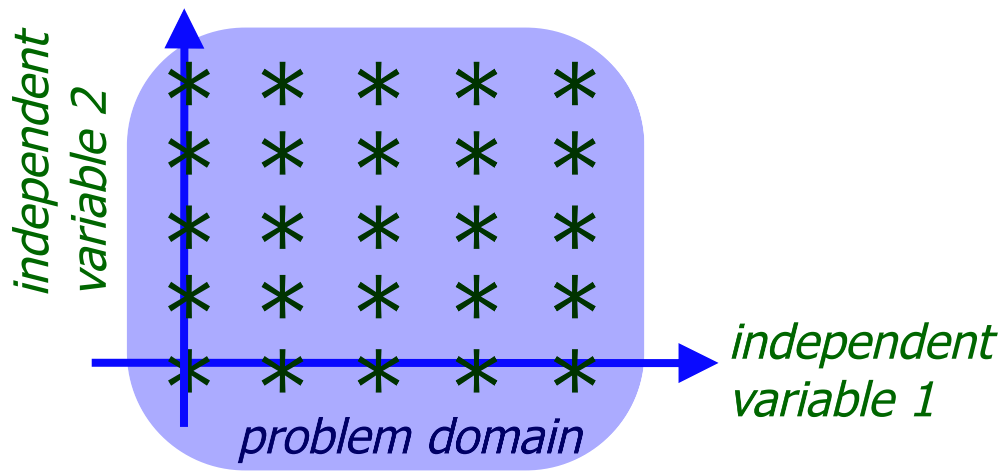
- can interpolate but less accurate than theoretical models (in truth is often more accurate than theoretical modelling)
- limited ability to extrapolate (beyond the set of solutions used in their development)
- are *black box* devices (providing no explanation of their output)



- the number of observations required increases geometrically with the number of independent variables:
  - say we need a density of **5** training examples across the range of an independent variable:



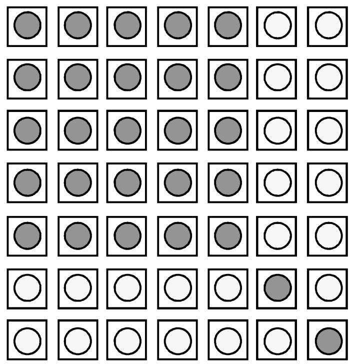
  - with two independent variables this increases to 5^2=**25** examples:



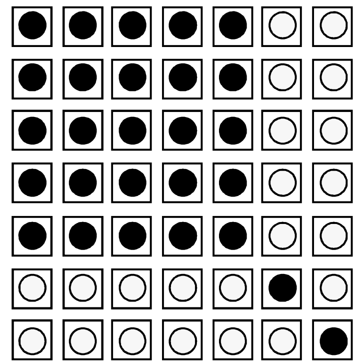
  - for ANN's a practical limit is typically **5 or 6** independent variables, say 5^6=**15,625** observations, otherwise training is too slow.

# independent variables:	1	2	3	4	5	6	7	8	9
# observations (5/variable):	5	25	125	625	3,125	15,625	78,125	390,625	1,953,125

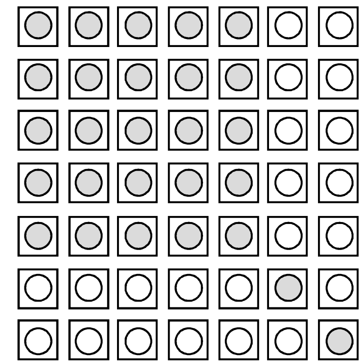
- a need to handle various types of variance, such as:
  - value/amplitude variance for spatially distributed inputs:



Base values (amplitudes)

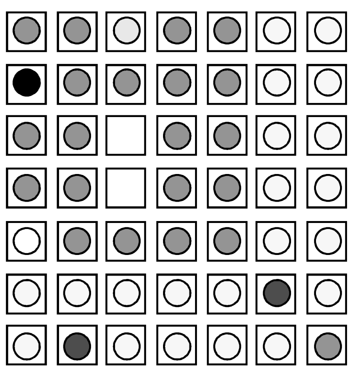


Values (amplitudes) increased



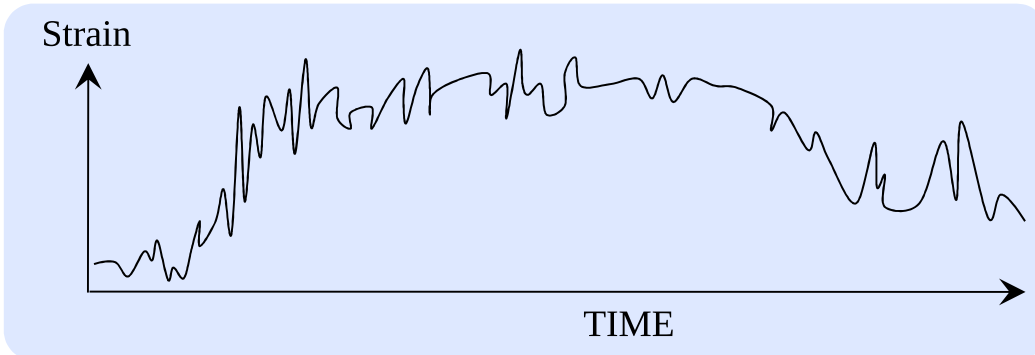
Values (amplitudes) decreased

  - stochastic variance and error for spatially distributed inputs:

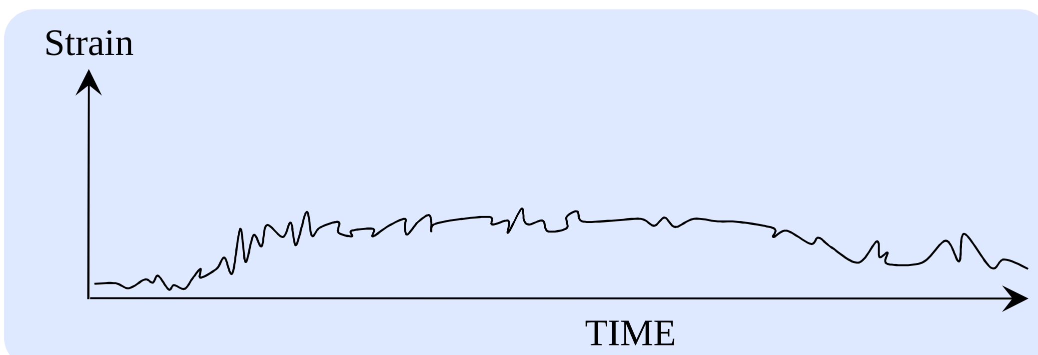


Noisy and missing values

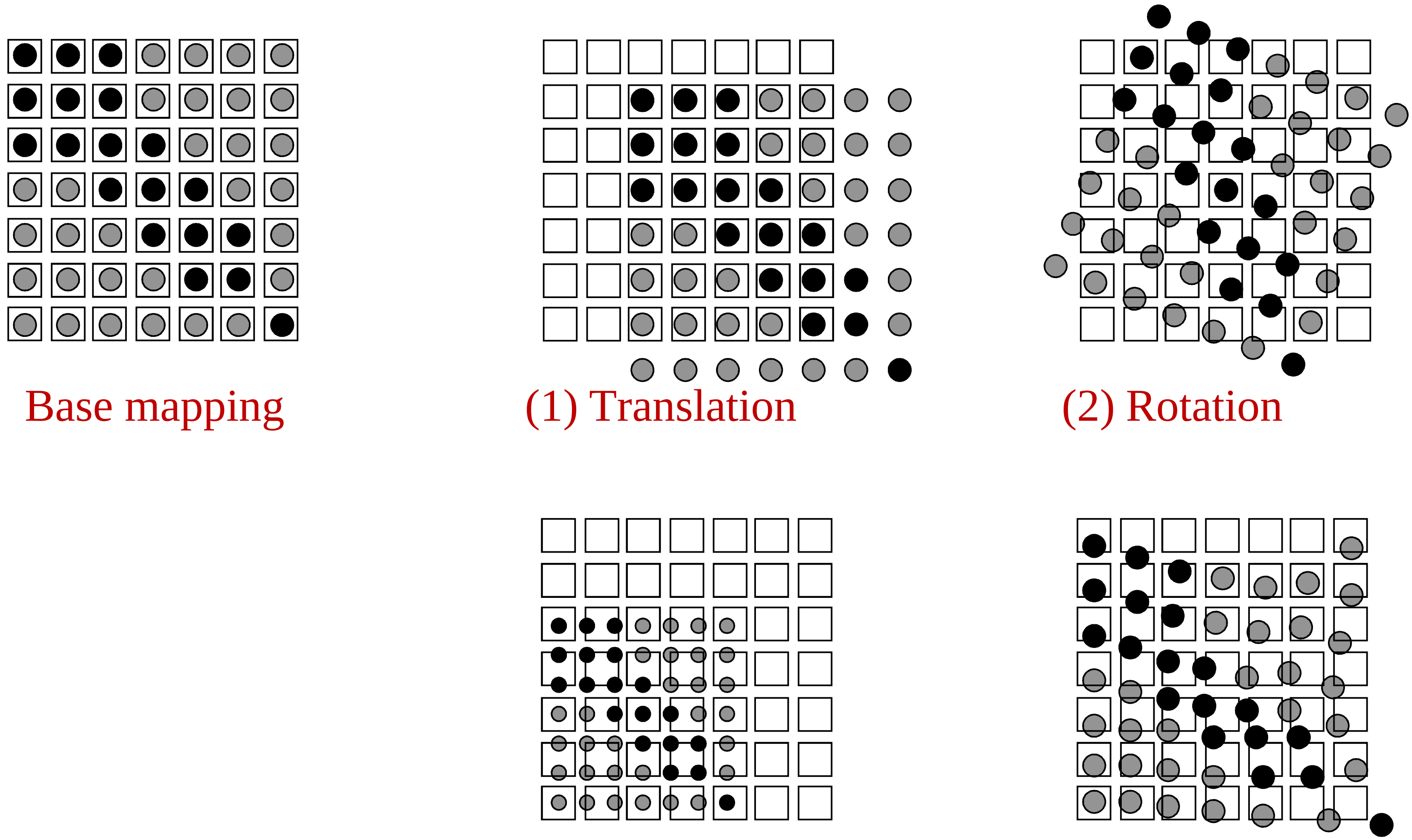
- value/amplitude variance for streams of input
- eg: strain in girder induced by truck crossing bridge



lower amplitude could be due to lighter loads
OR due to truck travelling in adjacent lane
- ambiguous



- a need for flexibility in the input format:
  - empirical models usually restricted to a fixed layout of the input values
  - …yet many problems require variation in the presentation of the inputs
  - variation may be for spatially distributed inputs:



Base mapping

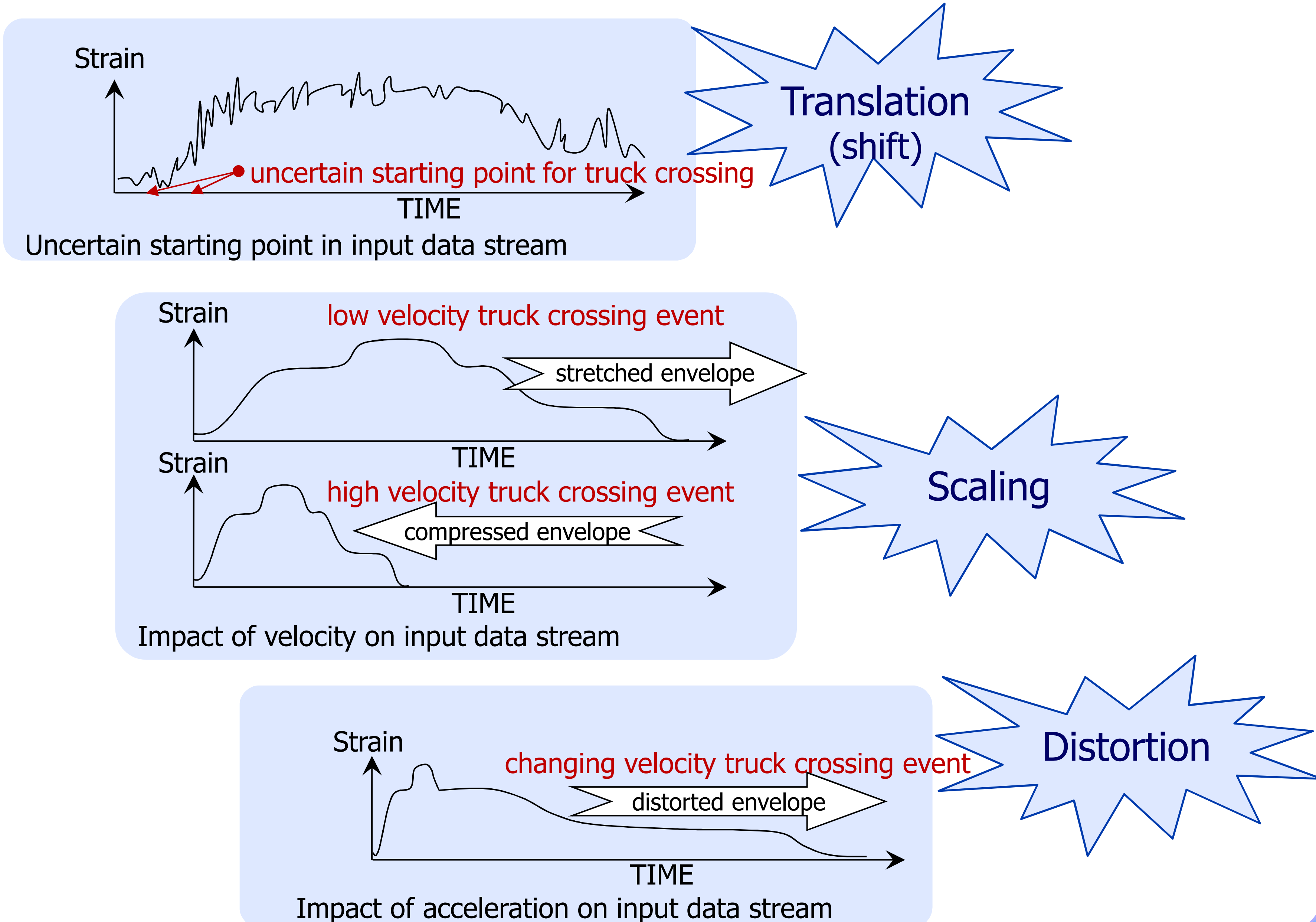
(1) Translation

(2) Rotation

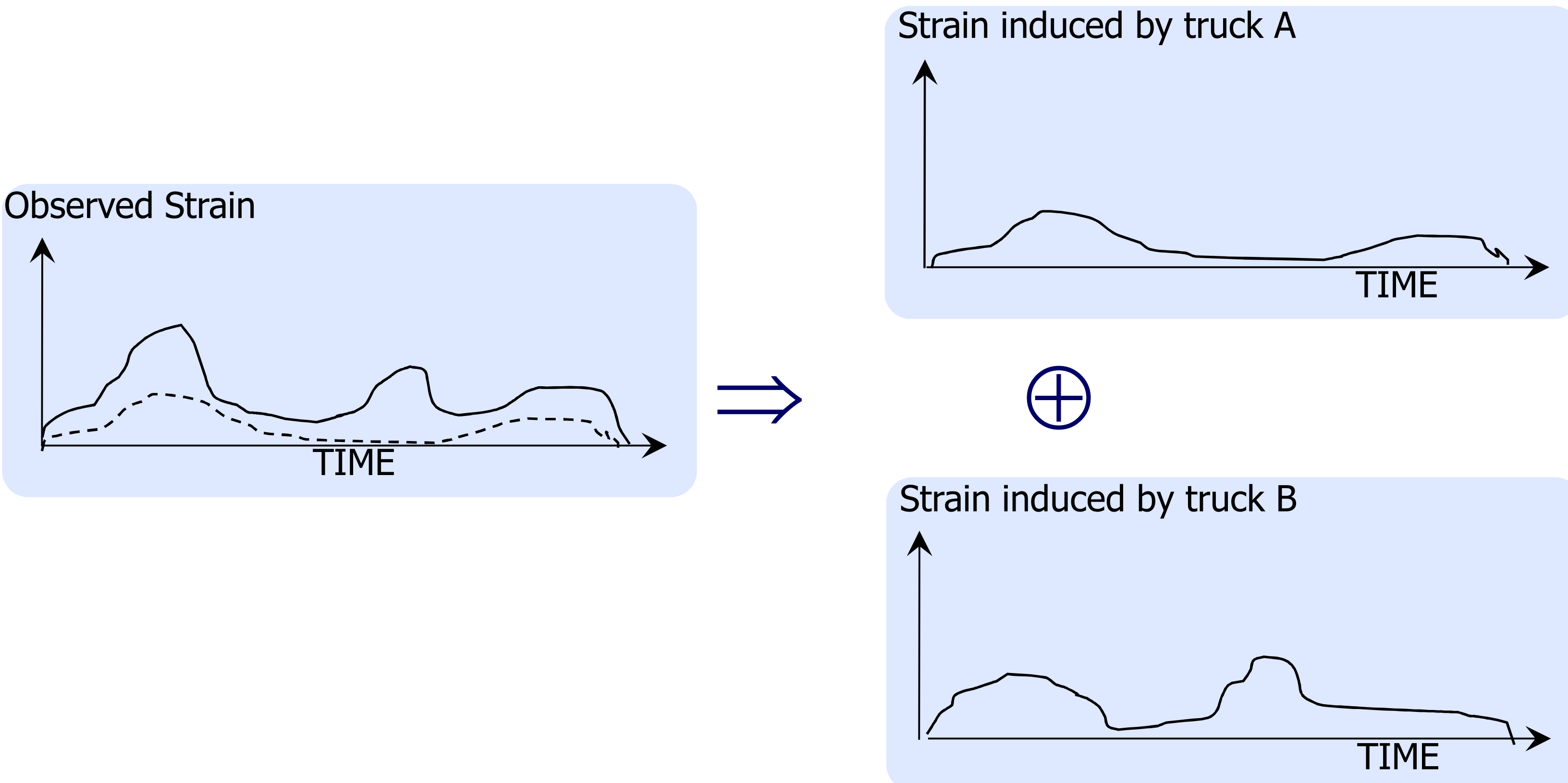
(3) Scaling

(4) Distortion

– ...or variation may be for streams of inputs
– eg: strain in girder induced by truck crossing bridge

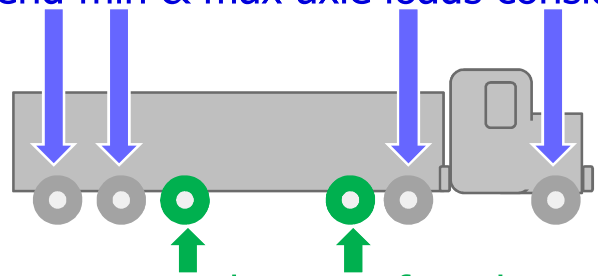


- uncoupling data sets:
  - many data sets/streams comprise two or more **overlapping** (or partially overlapping) data sets/streams
  - …we often need to **uncouple** them to handle them separately
  - eg: strain induced in girder by 2 trucks crossing bridge simultaneously



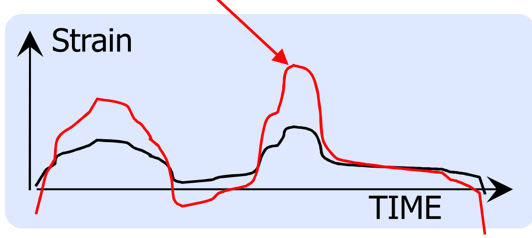
- extendibility of a model:
  - empirical models are developed to solve a class of problems
  - …often there is a need to **extend** the class of problems solved (increase the functionality of the model)
  - …eg: determining truck attributes from bridge strain data:

extend min & max axle loads considered (extend values of dependent variables)



extend range of truck types considered
(extend model internal structure, extend number of dependent variables)

extend range of values for strain readings considered (extend values of independent variables)



Others:
extend bridge lengths considered, extend number of lanes, etc…

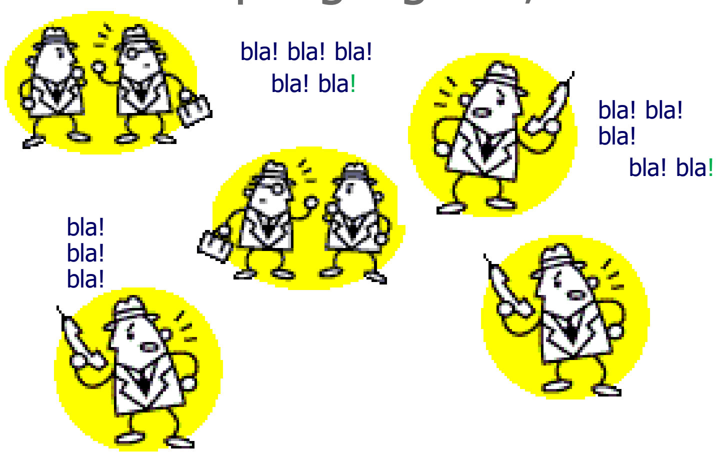
  - extension should be achievable without the model-user having to rebuild the existing model

APPROACH TO THESE CHALLENGES:

A rich future source of inspiration for empirical modelling is **the brain**:

- provides effective **empirically derived** solutions to many complex problems
- overcomes many of the challenges identified earlier:
  - eg: face recognition: spatial interpolation, translation, rotation, scaling, distortion, amplitude, noise:
  - eg: following a single conversation amongst a chattering crowd:

    uncoupling signals, etc…

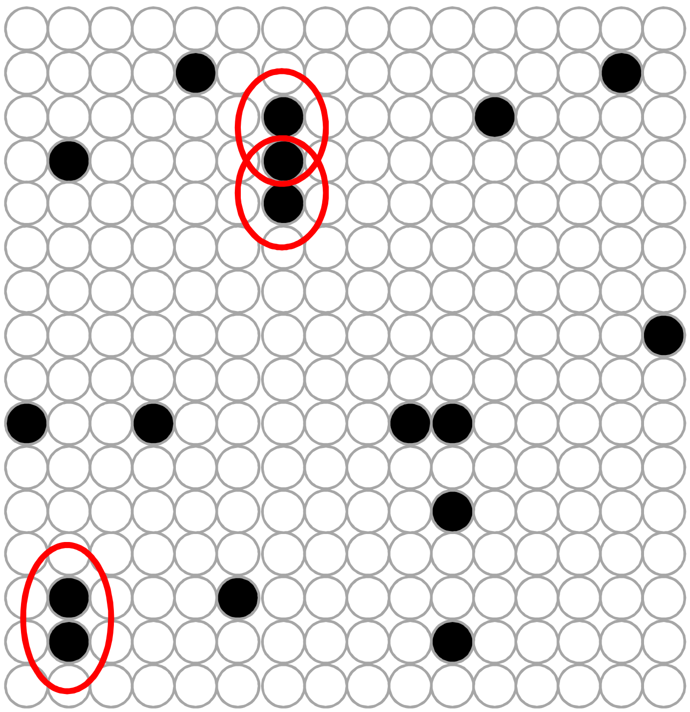


Which US politician do you recognize?

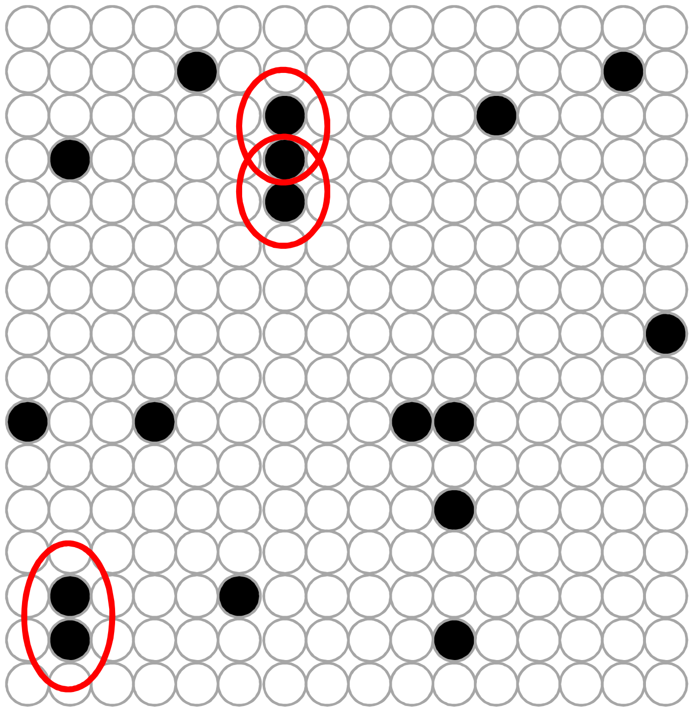
- arguably **the brain** is the ultimate black box

  ...but as we start to analyze its organization and operation we are discovering:
  - parts of the brain, at least, model the world as a set of **meaningful** features within a rich **hierarchical** structure
  - **lowest level in the visual system hierarchy** comprises detectors tuned to local features in an image such as orientation, spatial frequency, direction of movement, speed...
  - **second level in the visual system** integrates lowest level output with more specialized detectors tuned to features such as contours
  - **ultimately within the hierarchy** there are detectors tuned to very high level tasks such as recognition of a face (a US politician)
- similarly other brain systems, such as the auditory system, are based on a hierarchy of tuned feature detectors
  - ...(although there are many other sub-systems in the brain for which we currently have little or no understanding)
- so, empirical models do not have to be black boxes
  - they can develop **richly structured models** of the world
  - ...where the internal structure is an **insightful analog** of the internal structure of the problem represented

- what about exponential explosion in number of observations required?
  - richly structured models can resolve (or help resolve) this challenge
  - consider the simple problem of identifying vertical lines of two adjacent dots on a receptive matrix:



matrix size	total # of possible observations	# of 2 dot features that are vertical and adjacent
2 x 2	$4^2 = 16$	2
3 x 3	$9^2 = 512$	6
4 x 4	$16^2 = 65{,}536$	12
5 x 5	$25^2 = 3.36\text{x}10^7$	20
…		
16 x 16	$256^2 = 1.16\text{x}10^{77}$	240



matrix size	total # of possible observations	# of 2 dot features that are vertical and adjacent
2 x 2	$4^2 = 16$	2
3 x 3	$9^2 = 512$	6
4 x 4	$16^2 = 65{,}536$	12
5 x 5	$25^2 = 3.36\text{x}10^7$	20
…		
16 x 16	$256^2 = 1.16\text{x}10^{77}$	240

- a direct mapping model (mapping directly from input to output) would require a # of example observations proportional to column 2

  …the model would also be proportional in complexity to this

  …currently, most empirical models are implemented as direct maps
- a structured model (in this case with local feature detectors) would be proportional in complexity to column 3
- this is a simple example, but the argument extends to more complex patterns (if use a hierarchy of feature detectors)

- what about extendibility?
  - structured models are highly conducive to extension due to their inherent modularity
  - …extending the size of the receptive field (previous example) would just require an extension in the number of feature detectors
  - example is the coarse-grain modelling approach for the simulation of blast wave propagation around complex geometries:[4]
    - the spatial matrix through which the wave propagates is composed of empirically derived sub-models

      …allows model to be configured from a course mesh (1 m vs. 2 cm)

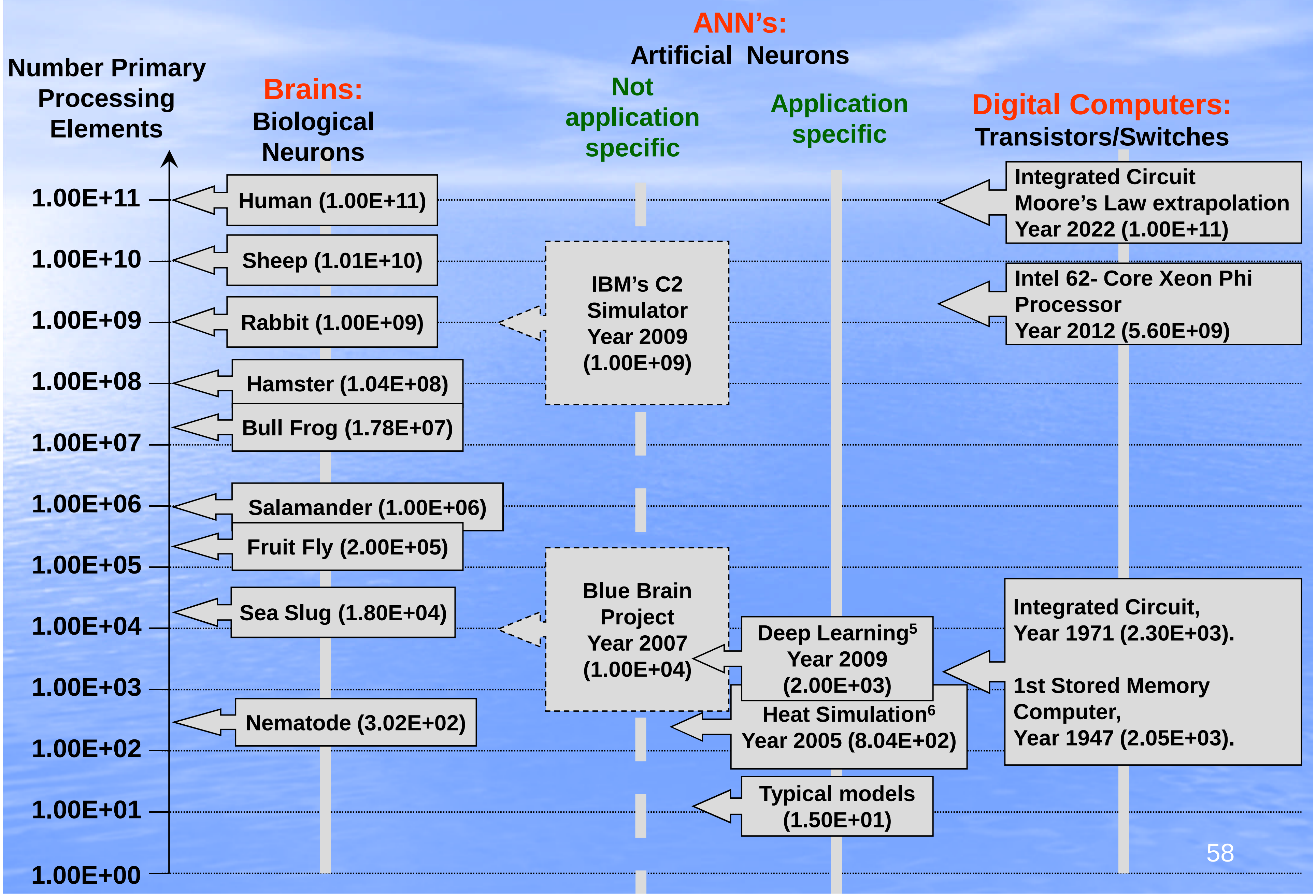
      …yet retains accuracy of conventional simulation

- indeed, the brain has provided modelling inspiration for 60/70 years in the fields of:
  - artificial intelligence (emulate intelligence at a high level)
  - …and in particular ANN’s (intelligence is an emergent property)

- …however, progress has been frustratingly slow
  - our knowledge of how the brain interprets, represents, and processes different types of information is still **rudimentary**
  - **practical applications** have similarly been limited in terms of the complexity of the problems solved

- compare the progress of ANNs with other devices:
  - **digital computing** has developed exponentially
  - can now build **massive ANNs** comparable in size to small mammalian brains (although operationally simplified)
  - …but not been able to exploit this in **practical applications**
  - biological model indicates a far **greater potential**

Compare progress: ANNs versus General Purpose Digital Computer



- then there is the question how to develop richly structured models:
  - need to learn their own **internal structure** and **representations**

    …these are not an explicit part of the observation data
  - for the brain:
    - parts of a model that are common to a broad range of problems may be developed through **evolution**
    - more novel aspects of a problem developed through direct experience (training)

    …how to apply either of these processes effectively within a computing environment is not clear

    …especially true for very large models (comprising say millions of neurons)
    - simulated evolution and other training methods are slow to converge for large models
- **Deep Learning** (Hinton et al.) is one of several attempts at developing models with rich internal structures
  - however, applications have been fairly limited (character recognition for example).[2]

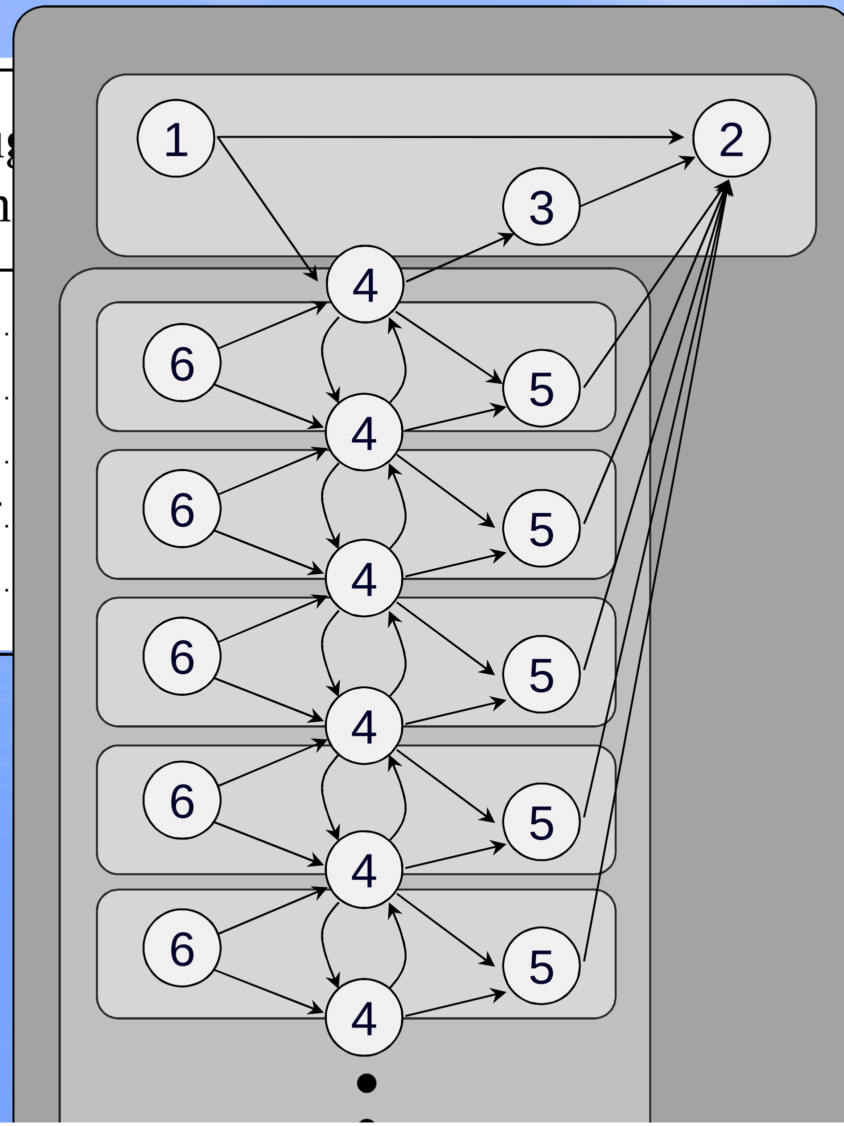
- an alternative approach for developing massive very complex model structures is **artificial embryogenesis** (growth algorithms)
  - simulated evolution would be applied to a **genotype**
  - the genotype is NOT the end model but rather a **code** used to direct the growth of the model
  - possibly well suited to structures that have a lot of **repetition**

    …only one version of the repeated element would have to be learned

- Consider the following simple growth table:

Neuron type:	1st daug neuron	er pe
1 ⇒	2	
2 ⇒		
3 ⇒	5	
4 ⇒	4	
5 ⇒	5	
6 ⇒		

...continue this process:



- could enhance this approach with multi-stage objective functions

SUMMARY AND CONCLUSIONS:

Empirical modeling:

- a very powerful means of modelling
  …but its potential has been largely untapped
- current models tend to be direct mapping devices:
  - no significant internal structure
  - provide no analog of the internal workings of the system under consideration
  - consequently restricted by issues such as:
    - **black box** devices
    - **number of observations** required for development = geometric function of number of independent variables
    - limited ability to handle **variance** in the presentation of a problem
    - limited ability to **extrapolate** and extend to new versions of a problem
- approach to overcoming these challenges:
  - inspiration from biology:
    - structure, operation, evolution, development, and learning in the brain

References:

1. Ian Flood and Raja R. A. Issa (2010). ”Empirical Modeling Methodologies for Construction.” *J. Constr. Eng. Manage.* 136, Special Issue: Research Methodologies in Construction Engineering and Management, ASCE, NY. pp. 36–48.
2. Ian Flood, Bryan T Bewick, Robert J Dinan and Hani A Salim (2009). “Modeling Blast Wave Propagation Using Artificial Neural Network Methods”, Advanced Engineering Informatics 23, Elsevier. pp 418–423
3. Nicolas Gagarin, Ian Flood and Pedro Albrecht (1994). "Computing Truck Attributes with Artificial Neural Networks", *Journal of Computing in Civil Engineering*, ASCE, Vol. 8, No. 2. pp 179-200.
4. Ian Flood, Bryan T Bewick and Emmart Rauch (2012). “Rapid Simulation of Blast Wave Propagation in Built Environments Using Coarse-Grain Simulation”, *International Journal of Protective Structures* , Vol. 3, No. 4. pp 431-448.
5. Ruslan Salakhutdinov and Geoffrey Hinton (2009). “Deep Boltzmann Machines.” Proc. 12th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, Florida, USA. Vol. 5 of JMLR: W&CP 5. pp448-455.
6. Ian Flood, Caesar Abi Shdid, Raja R. A. Issa and Nabil Kartam, (2007). “Rapid Multi-Dimensional Simulation of Transient Heat-Flow in Buildings Using Neural Network-Based Coarse-Grain Modeling”. *Journal of Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, AI-EDAM, Cambridge University Press, pp. 18.