Statistical models

Statistical modeling

Simplified mathematical formulation for statistical analysis

- stochastic
- somehow approximate reality
- explanatory

Different philosophies

- Generative models
  - describe how the data are generated
- Predictive models
  - models mainly for making predictions

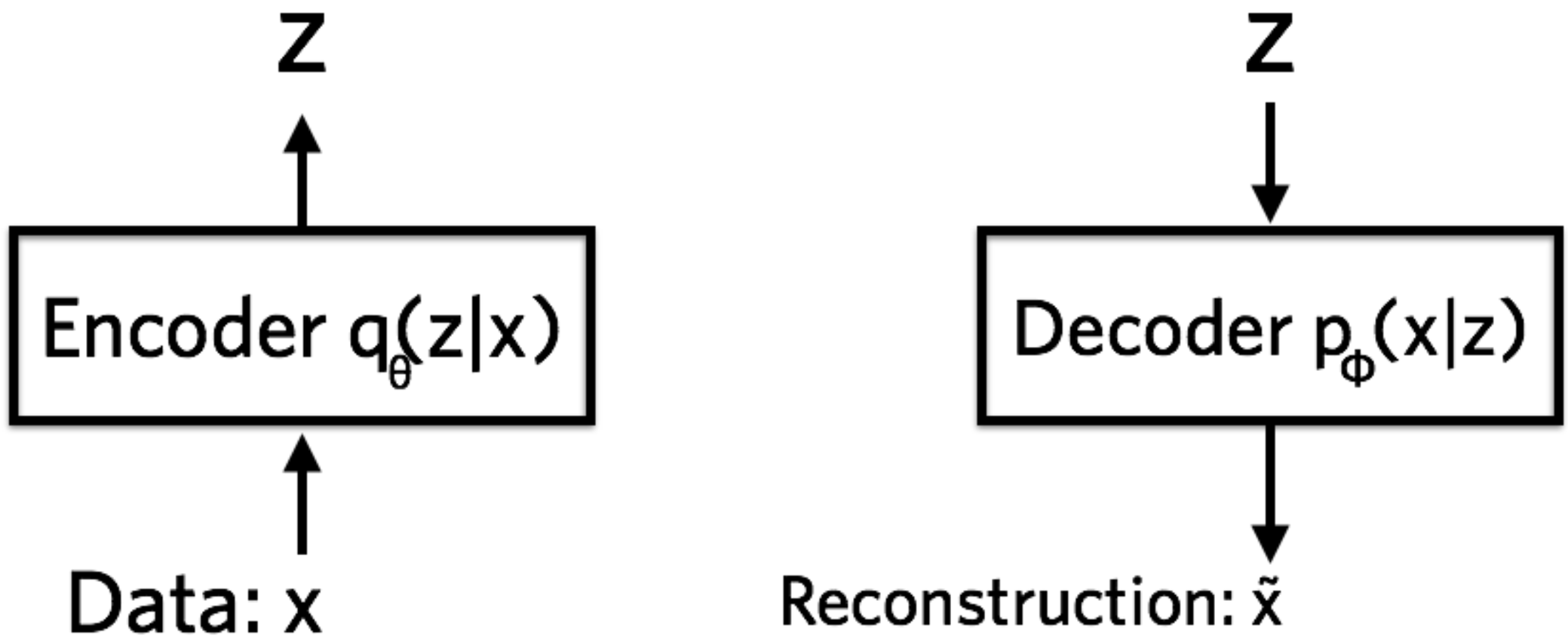
Generative models

Coin tossing

$X_i = \begin{cases} 1, \text{ with prob. } p & (\text{第 } i \text{ 次出現正面}) \\ 0, \quad " \quad 1-p & (\quad " \quad \text{反}) \end{cases}$

$\Rightarrow X_i \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$

Variational autoencoder



Variational autoencoder

Failure rate model

State-space models for air pollution

depend on weather

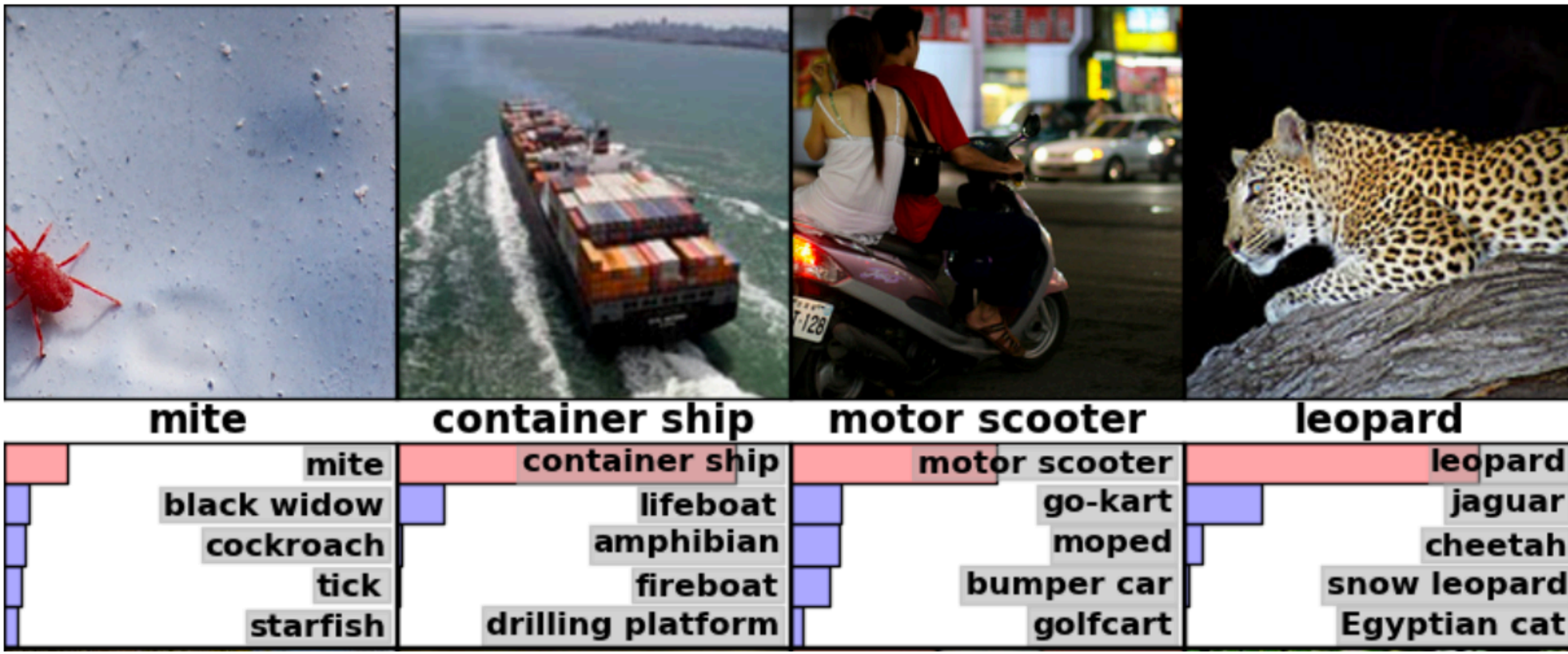
$$P_{t+1}(s) = \int D(s,s')\, P_t(s')\, ds' + N(s)$$

diffusion

新污染 or 沈降

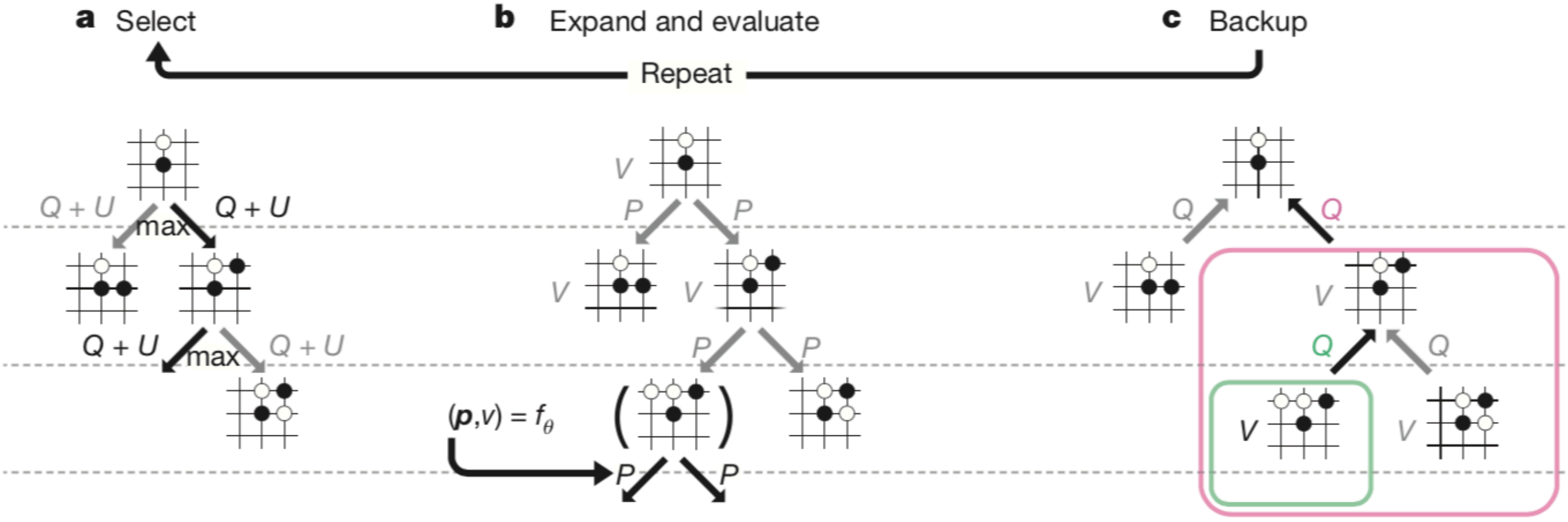
Predictive models

Image recognition



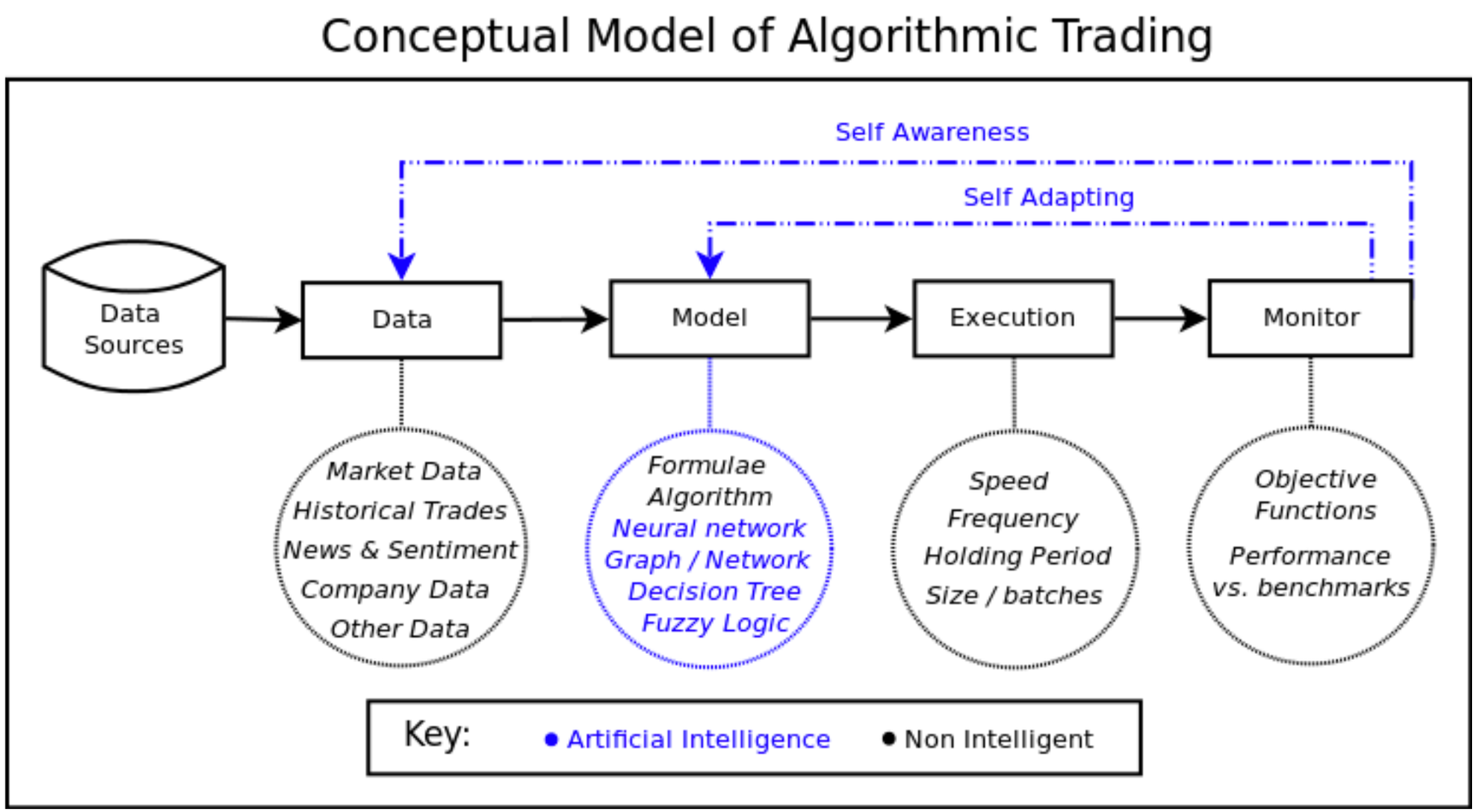
https://www.tensorflow.org/tutorials/images/image_recognition

AlphaGo



https://hackernoon.com/the-3-tricks-that-made-alphago-zero-work-f3d47b6686ef

Algorithmic trading



http://www.turingfinance.com/wp-content/uploads/2013/11/Algorithmic-Trading-Systems-Conceptual.png

Conclusion

- Most statistical models are represented by probability models with certain unknown parameters
  - Probability models are usually specified by human
  - Unknown parameters are estimated (trained) from data

- Does the model good enough?
    - goodness of fit
- Which model is better for a specific data science problem?
    - no free lunch theorem
    - model selection (cross–validation, information criteria, LASSO, etc.)
    - model averaging (ensemble learning)