Machine Learning and Cyber Security: It's the Data, not the Algorithm

Associate Professor Mike Johnstone

Security Research Institute, Edith Cowan University

m.johnstone@ecu.edu.au

"Prediction is very difficult, especially if it is about the future"
(Niels Bohr, 1885-1962)

Session Agenda

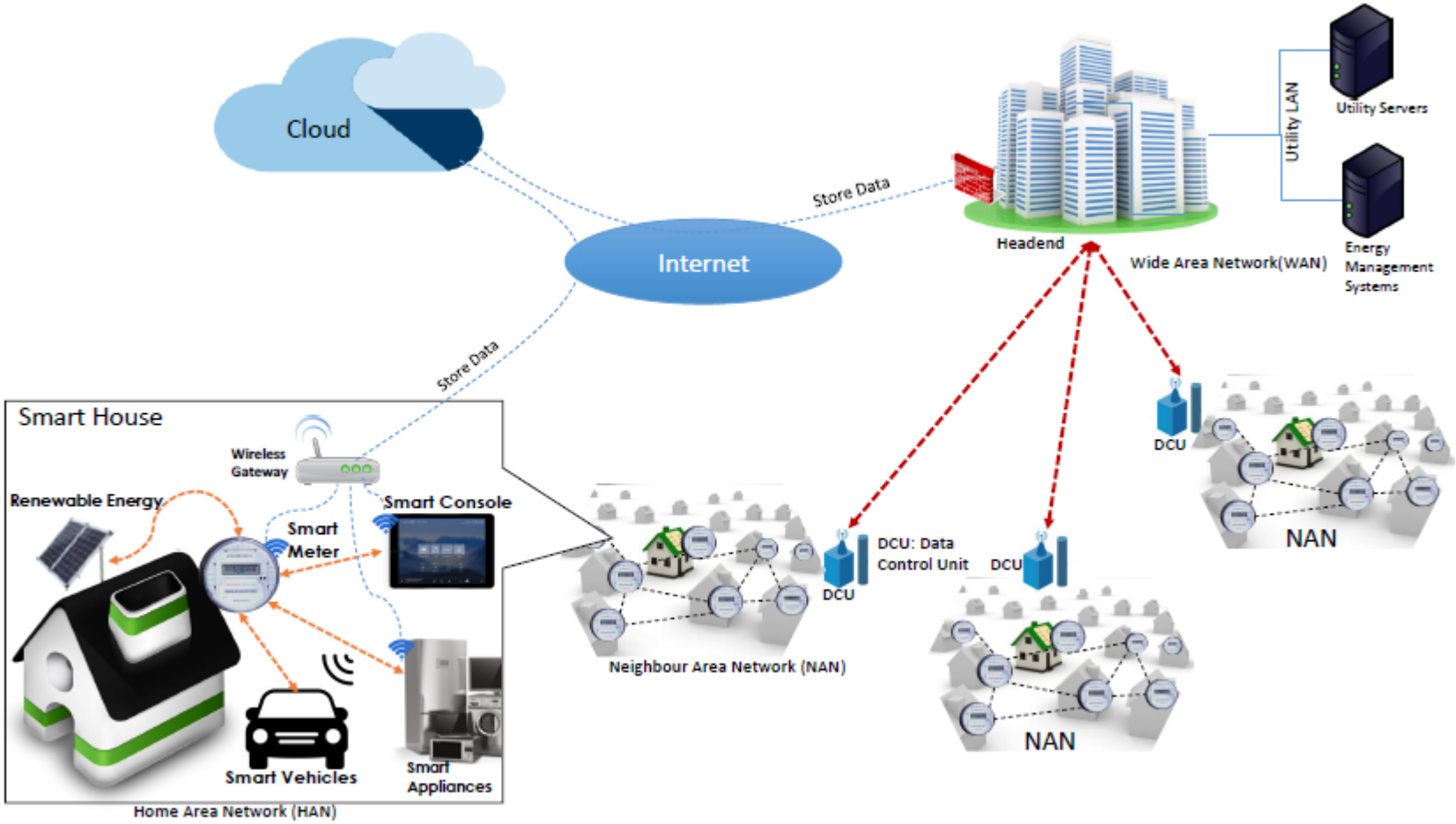
- Introduction
- Problems in machine learning
- Case studies

Growth of the “Internet of Things”

8.4 Billion Devices in 2017

20.4 Billion Devices by 2020

Everything is connected…



Why Machine Learning (ML)?

- We wish to solve a problem (or class of problems) that is not amenable to treatment via a fixed (deterministic) human-written program
- What is the problem? Can it be solved by:
  - Regression (need to predict a value)?
  - Classification (need to find out the category of a value)?

What People Think

- “AI is the biggest risk we face as a civilisation” (Elon Musk)
- Facebook AI chat bots develop their own language

Machine Learning to the Rescue

"By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it."

Eliezer Yudkowsky

Machine Learning to the Rescue (or not)

"Artificial intelligence has the same relation to intelligence as artificial flowers have to flowers.

From a distance they may appear much alike, but when closely examined they are quite different."

David Parnas

Many choices-everyone has a favourite

- ANN: Image processing
- NBC: Spam detection
- HMM: Predictive text

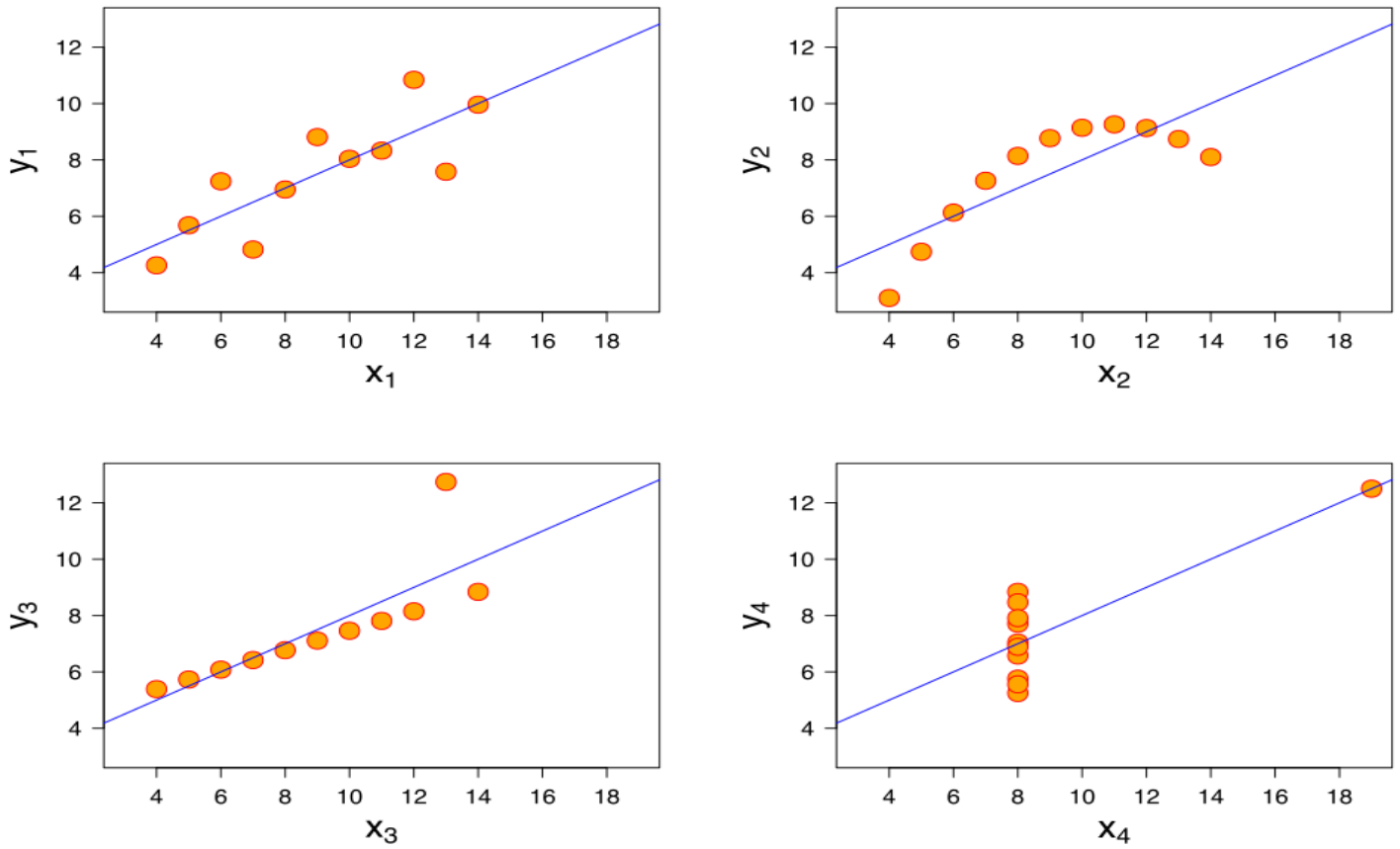
Anscombe's Quartet
(Anscombe, 1973)

- Four datasets with identical statistical properties:

Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $= 27.50$ (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

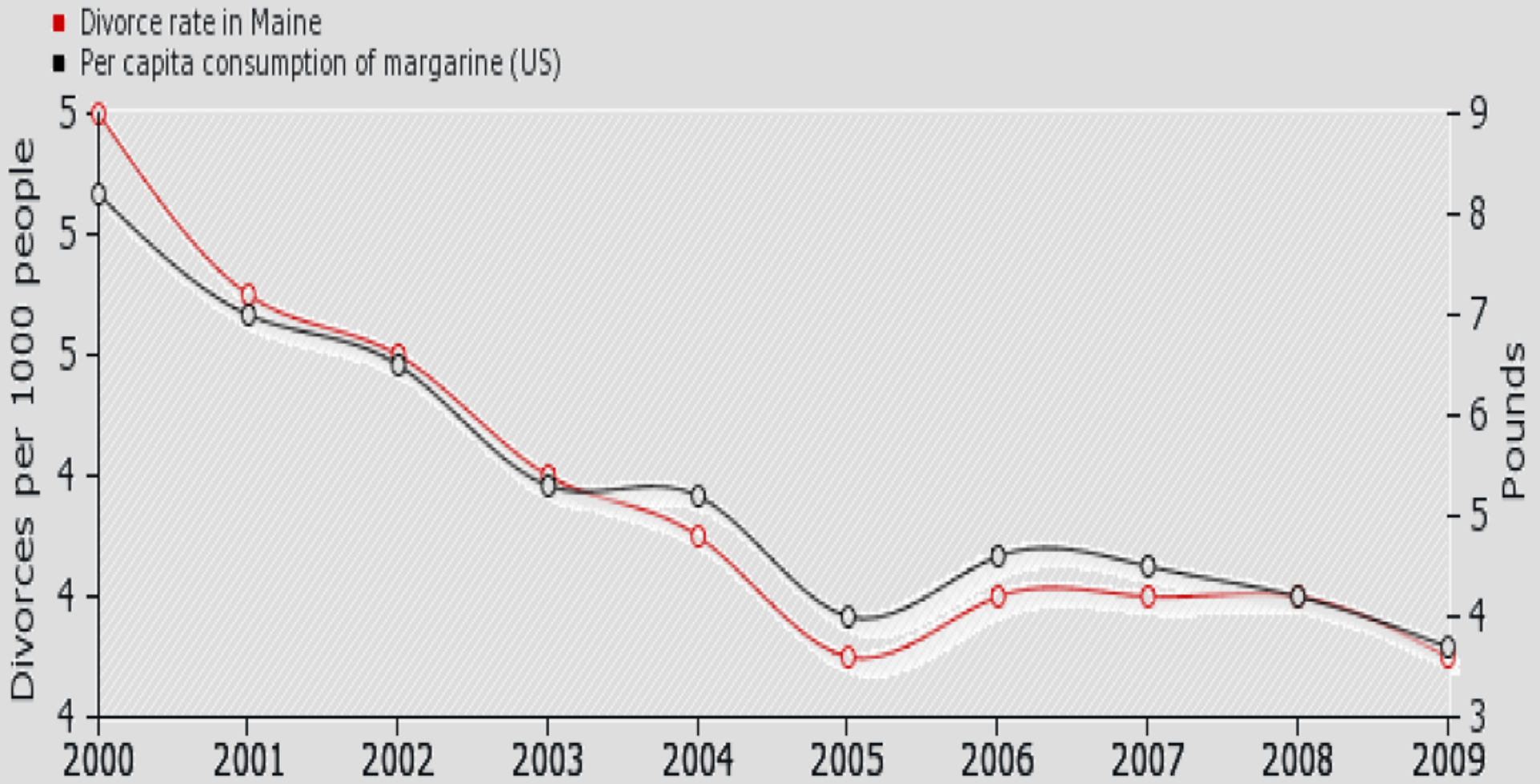
Anscombe's Quartet

- This is why visualisation is a good idea



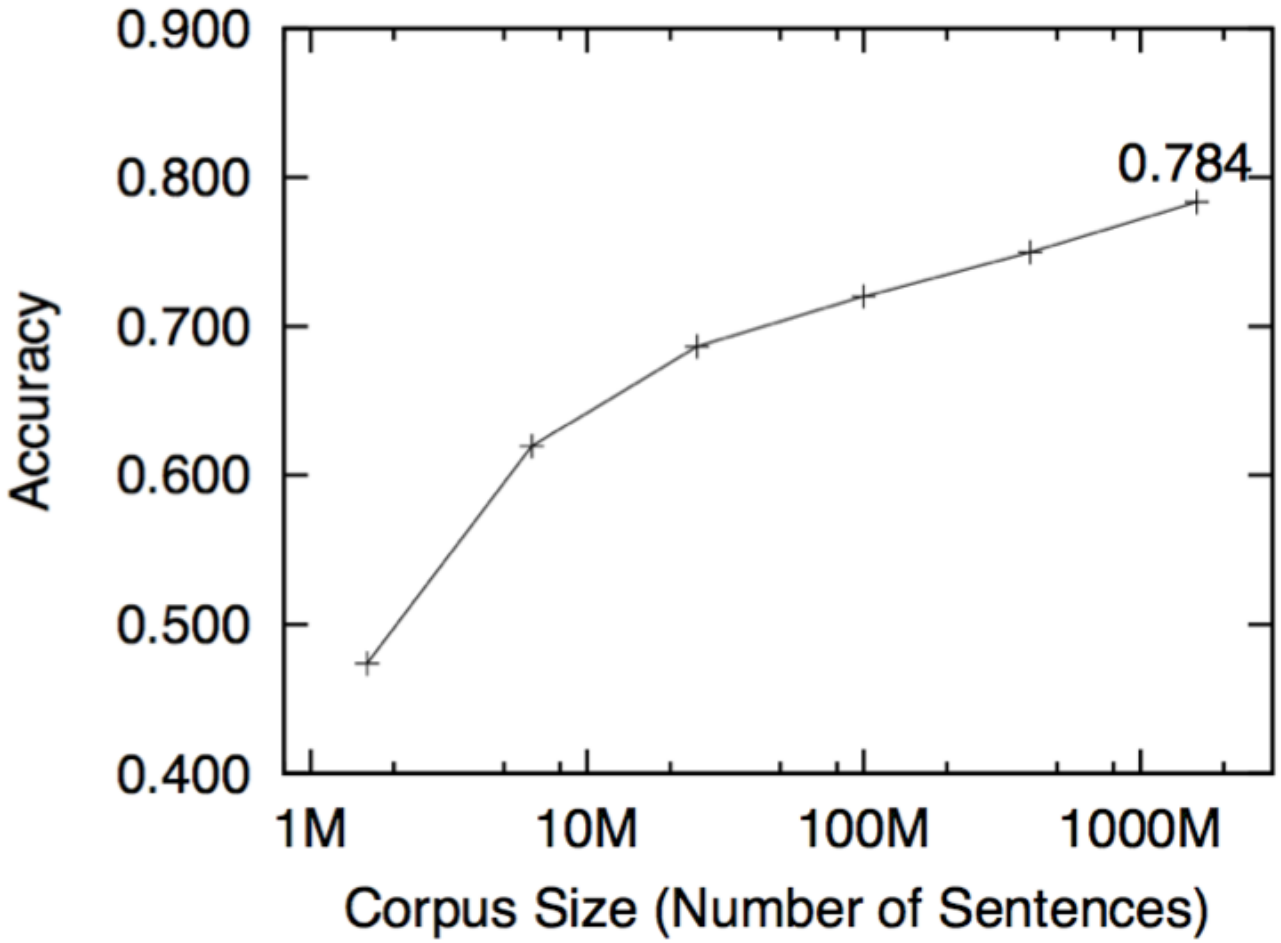
Correlation is not causation

(Vigen, n.d.)



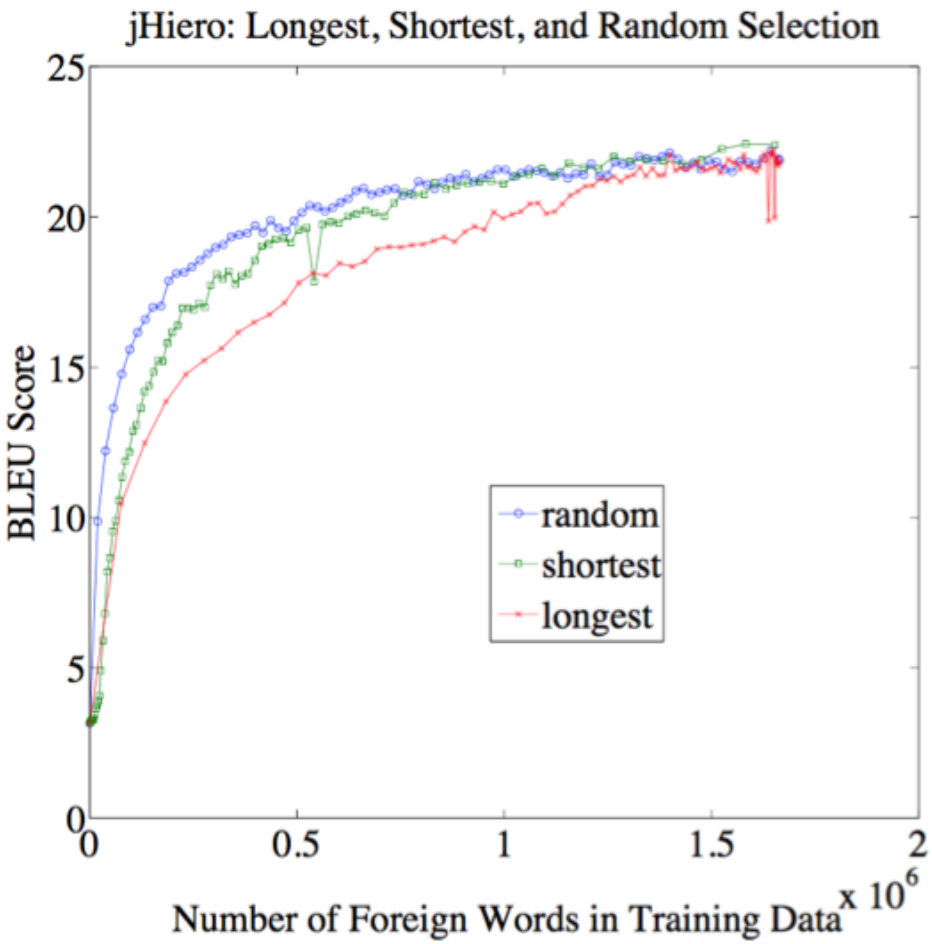
Discourse Analysis with Case Frames

(Sasano et al., 2009)



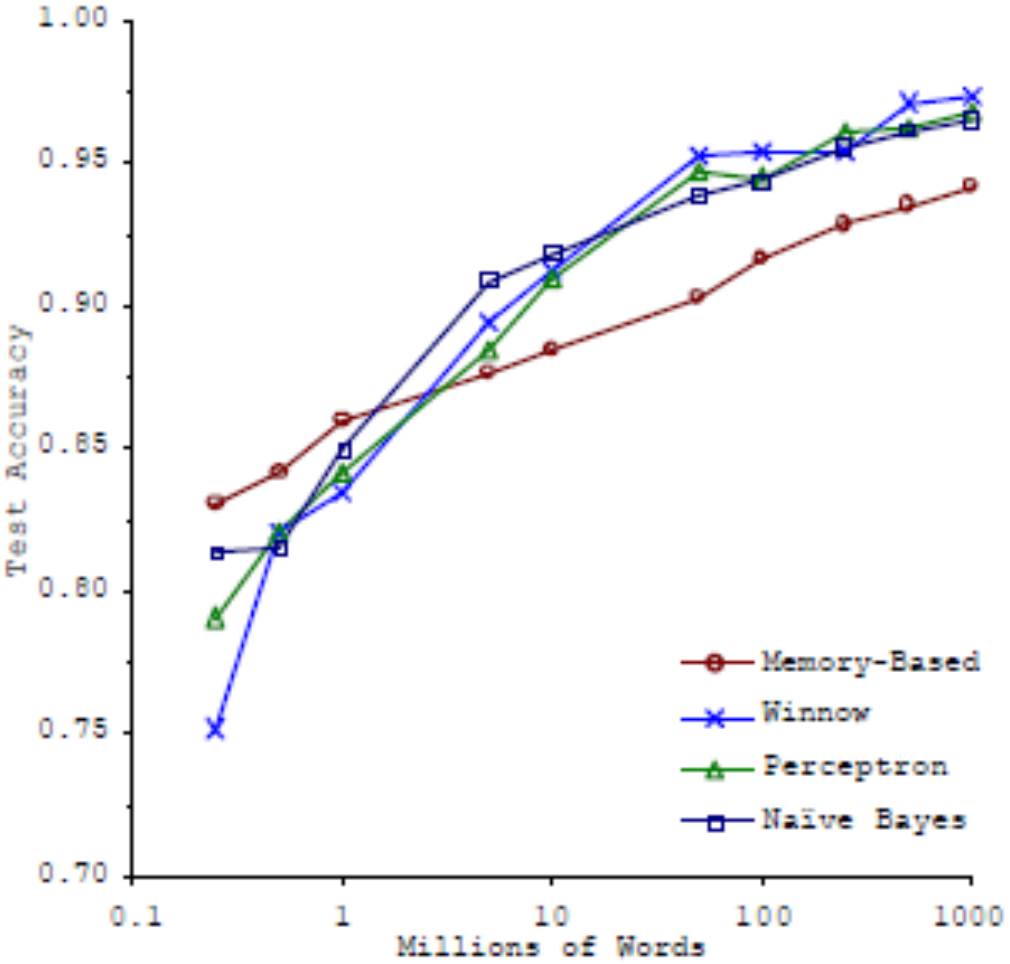
Diminishing Returns in Statistical Machine Translation

(Bloodgood and Callison-Burch, 2010)



Word Sense Disambiguation

(Banko and Brill, 2001)

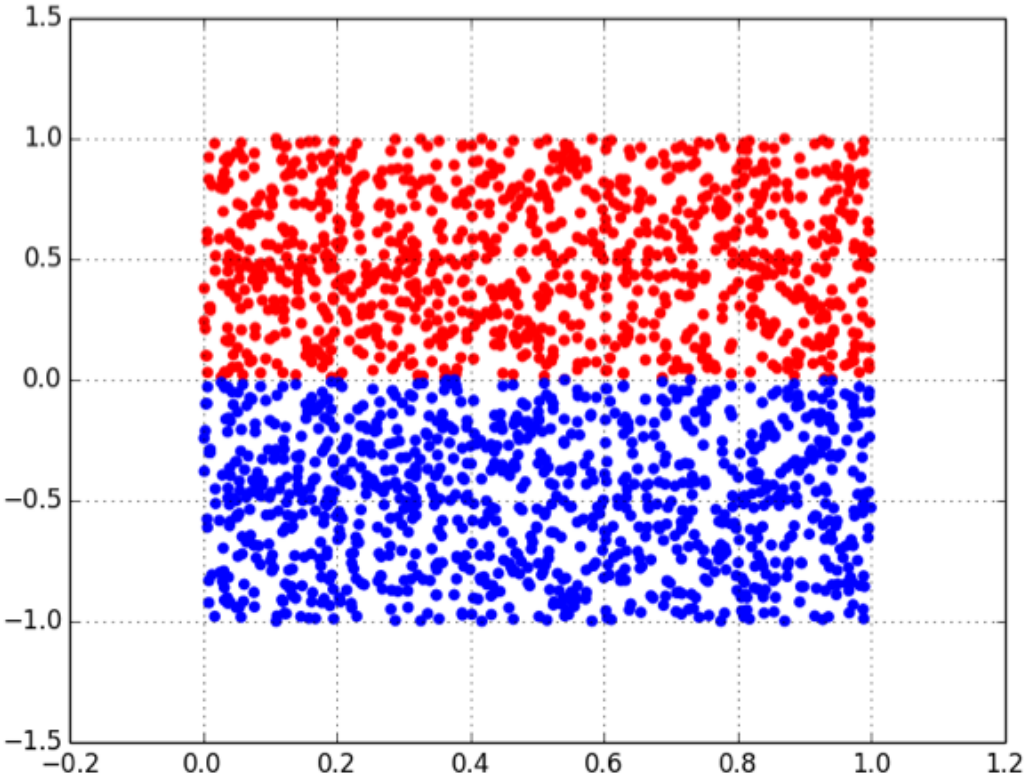


Bag of Visual Words

(Hentschel and Sack, 2014)

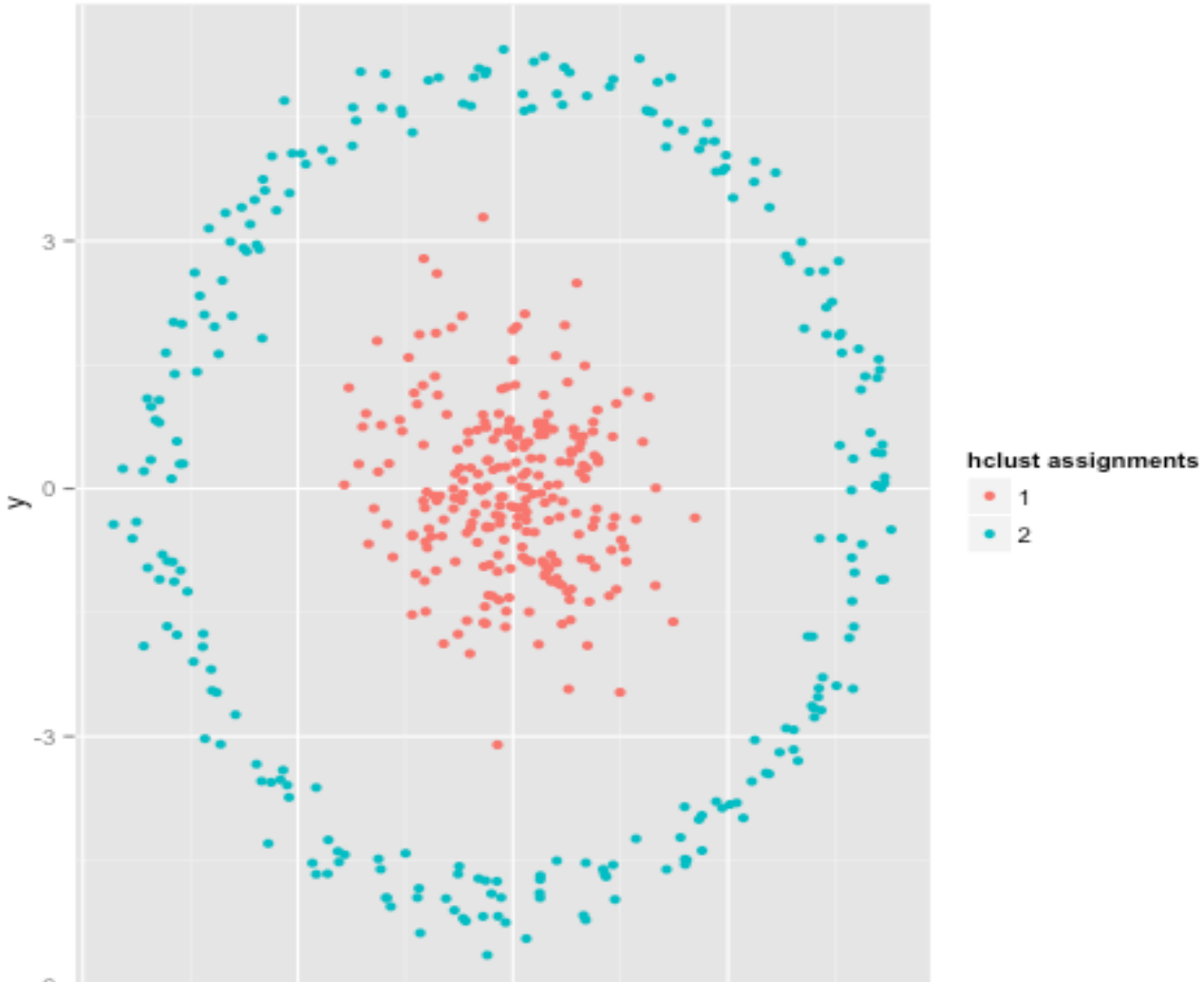
Classifier	Hyperparameters	mAP
Naïve Bayes	α (smoothing parameter)	0,480
k nearest neighbors	k (no. of nearest neighbors)	0,524
Logistic Regression	C (regularization)	0,548
linear SVM	C (regularization)	0,554
RBF kernel SVM	C (regularization), γ (kernel coefficient)	0,593
Random Forest	n (no. of decision trees)	0,612
AdaBoost	n (no. of decision trees), d (depth of each decision tree)	0,632
χ^2-kernel SVM	C (regularization)^5	0,674

Easy to separate



Not so easy to separate

(Robinson, 2017)

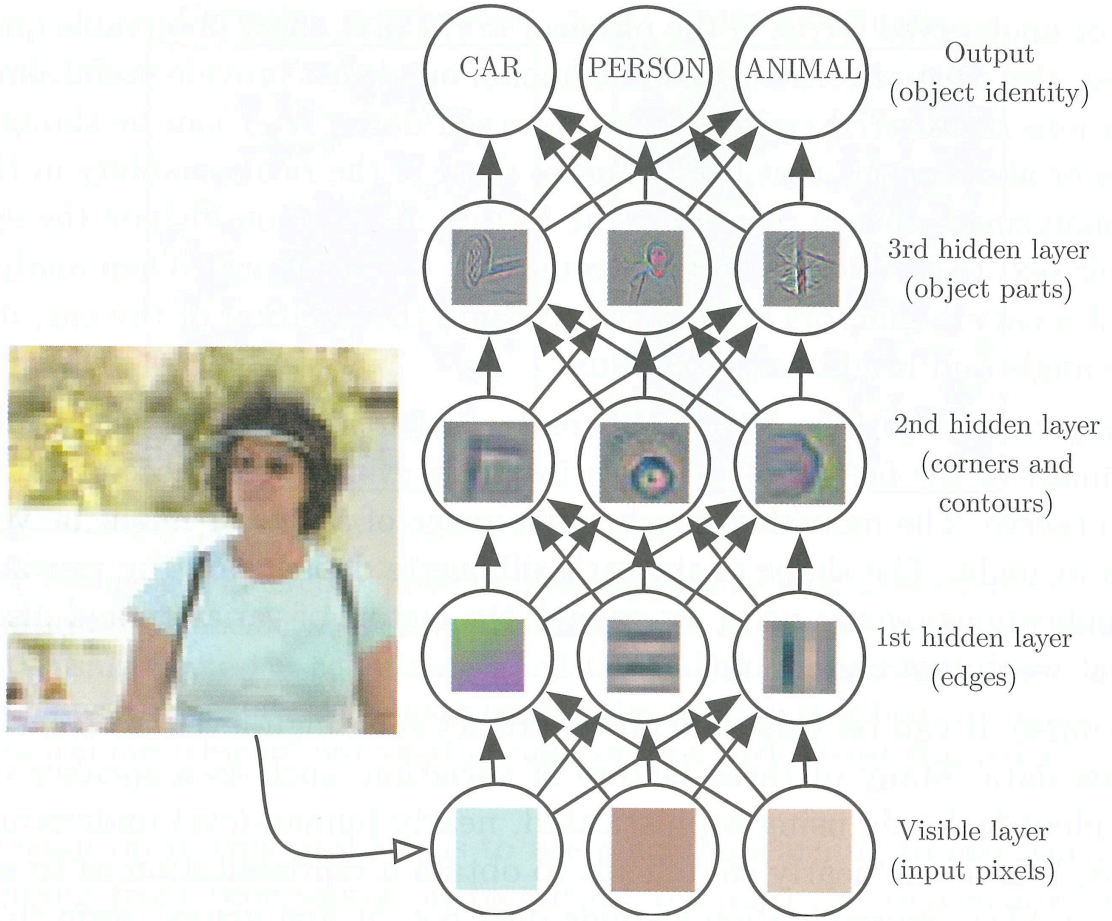


Case Study: Sounds

- Can machine learning algorithms assist or supplant human analysts in detecting specific motor vehicle sounds?
- Needs significant pre-processing
- What is the feature set?

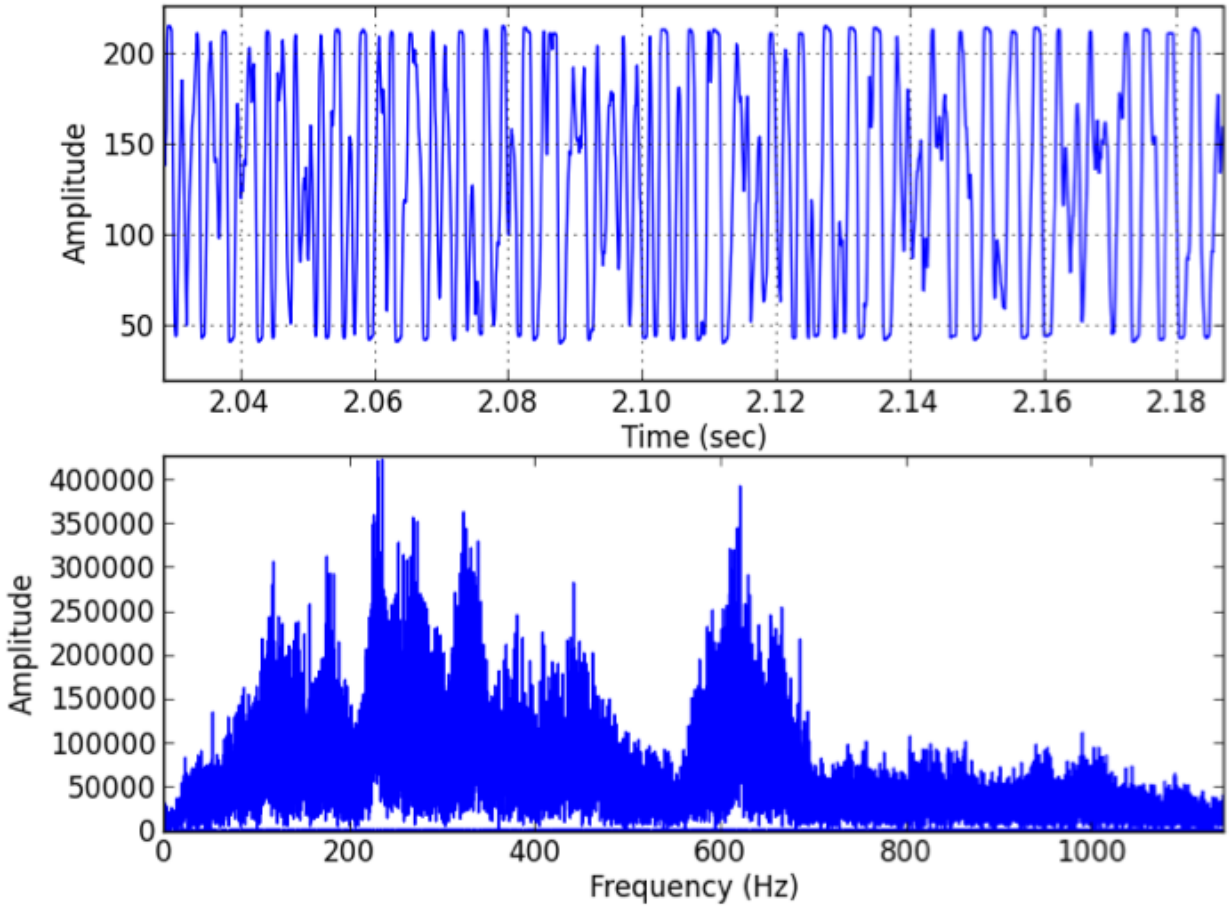
I think you'll find it's a bit more complex

(Goodfellow et al., 2016)



A Sound Sample-Ford Mustang V8

(Johnstone and Woodward, 2013)



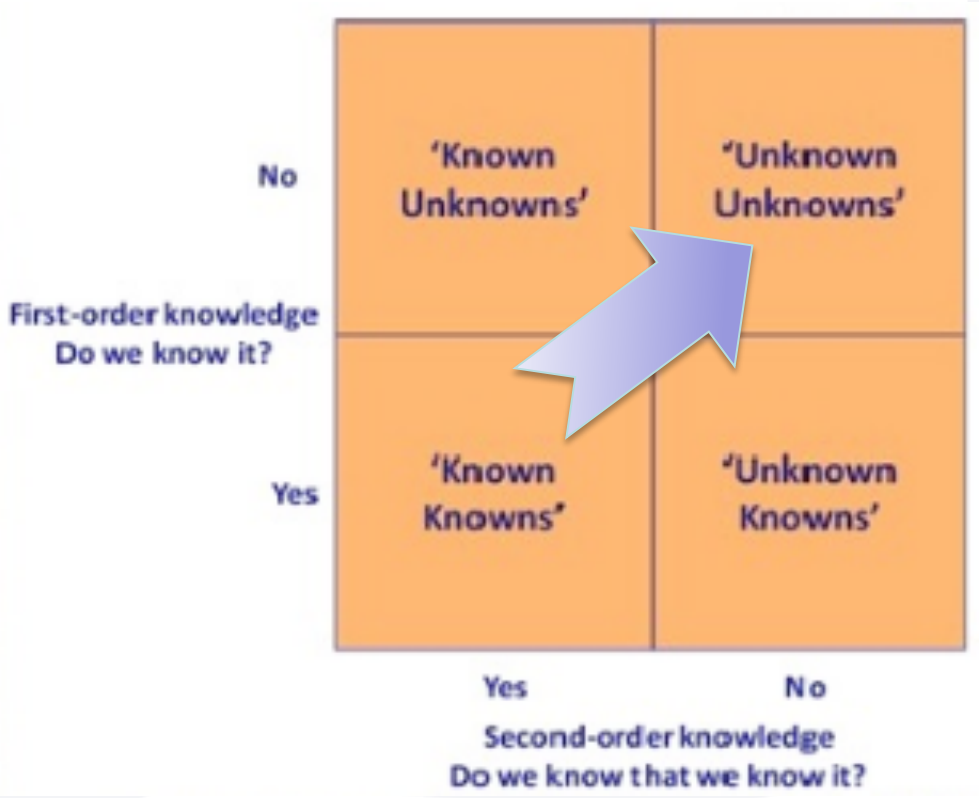
Case Study: BACnet

- A protocol designed for building automation systems
- Security as an addendum
- Problems in the protocol arise from second-order effects
- How to detect an “unknown-unknown”?
- What are the relevant features?

Second-order effects

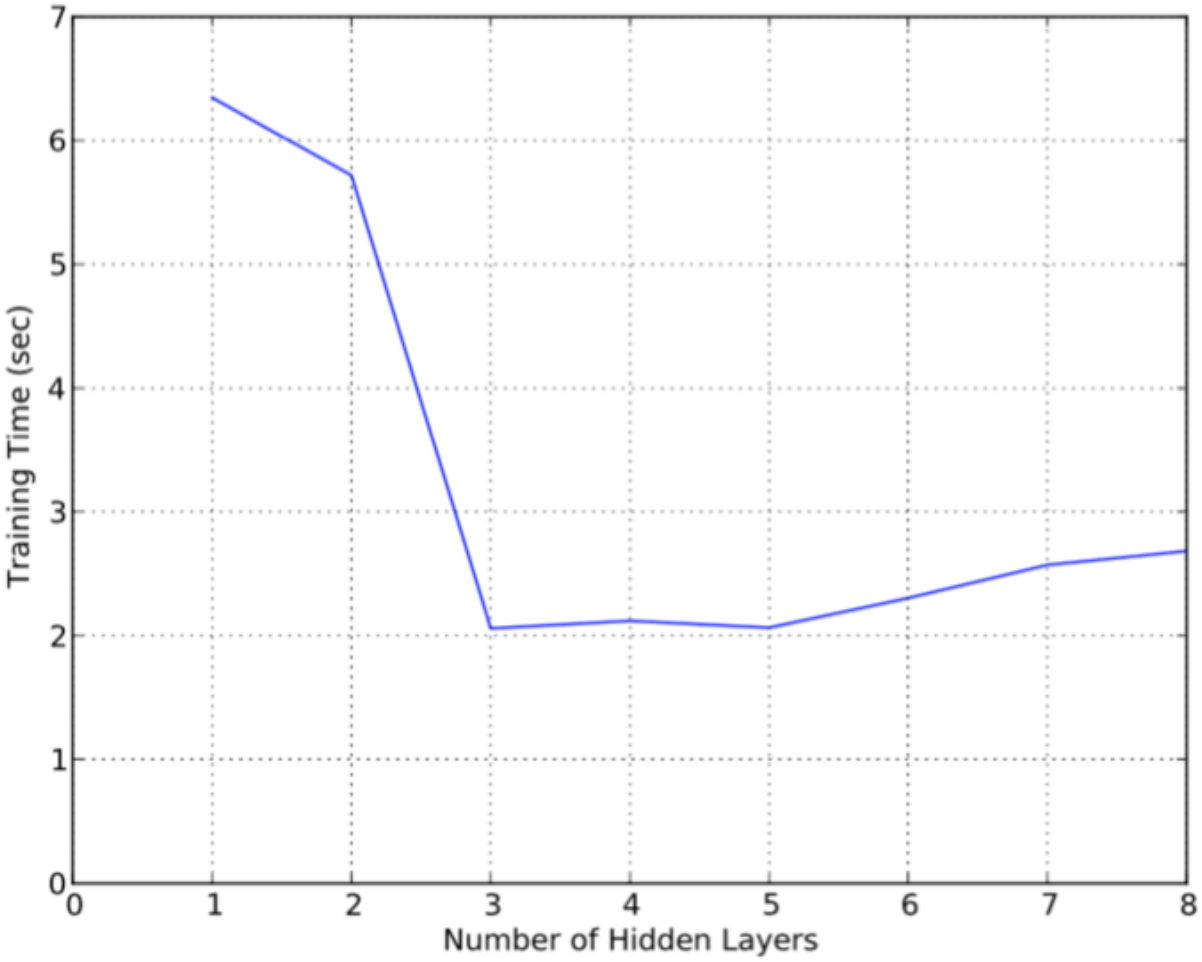
- Second-order effects an artifact of complex systems
- Post-9/11 travel patterns in America an example
- BMS and data centres an example

A quasi-Rumsfeldian Approach
(de Spiegeleire, 2009)

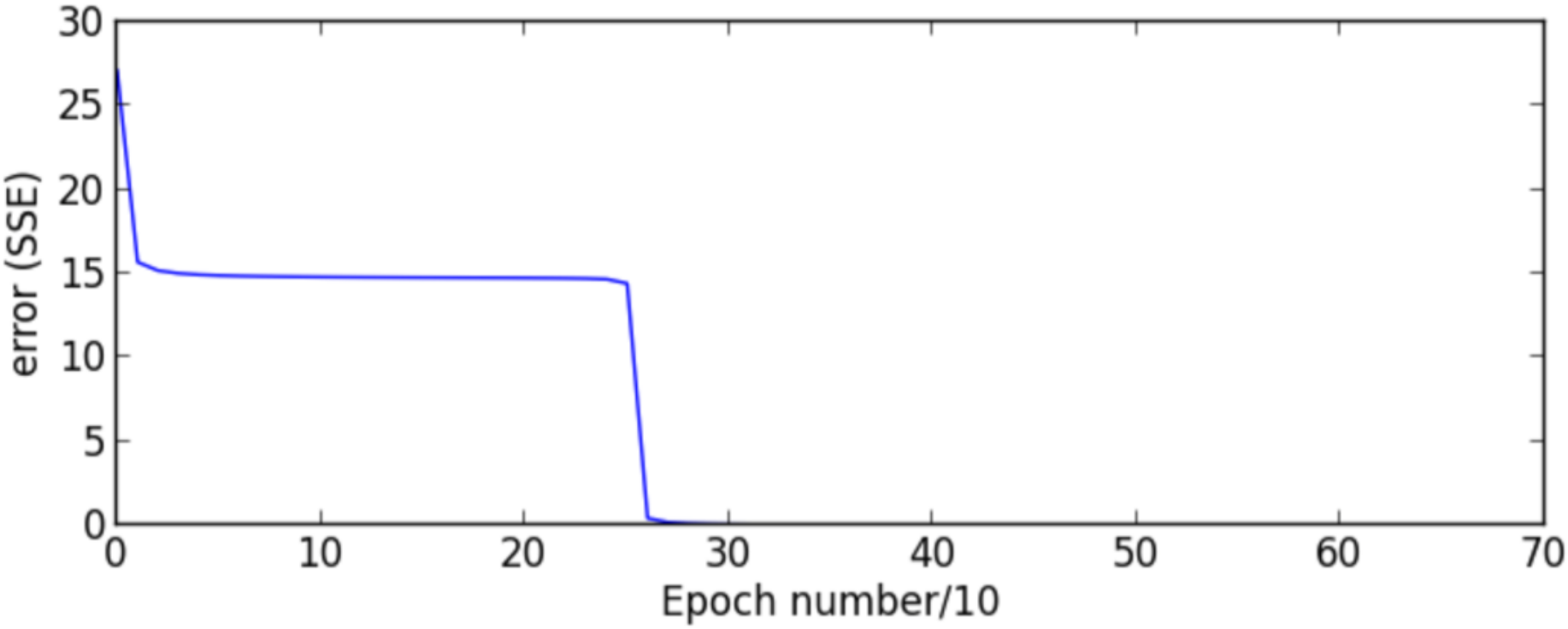


Case Study: BACnet

(Johnstone, Peacock and den Hartog, 2015)



Case Study: BACnet



Questions