Time-Series Prediction: A Challenge to the Neural Network Field

- NSF funding support via Guyon, interest
- Neural network people need to respond, but only in the right way
- Need to develop, teach and use the fundamental statistical principles which make brain-like “cognitive” prediction possible.
- How to win: lessons from past competitions, formal and informal

Better universal prediction is a core goal of science

Search on "COPN" at www.nsf.gov

Pr(A|B)=Pr(B|A)* Pr(A)/Pr(B)

Prediction

Memory

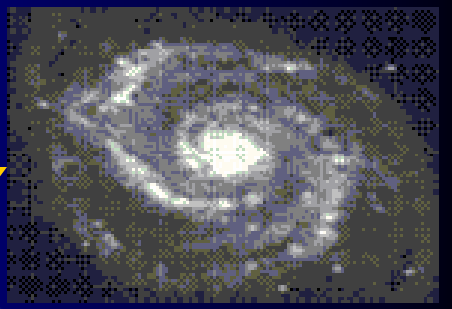
. . .

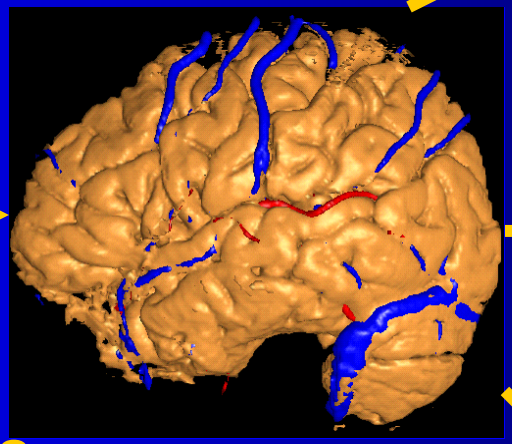
Clustering

Optimization

J(t)=Max<J(t+1)+U>

$$\frac{\partial^{+} z_n}{\partial z_i} = \frac{\partial z_n}{\partial z_i} + \sum_{j=i+1}^{n-1} \frac{\partial^{+} z_n}{\partial z_j} \frac{\partial z_j}{\partial z_i}$$





Example of the Wrong Way to do Competition and the Right way

- "Deep Blue" competition taught us almost nothing about intelligence because it was domain-dependent and used cheating
- Fogel (Proc IEEE 2004) did the first master-class chess player which how to play well
- In economics, we need to "cheat" – use prior information – but we do best if our way of learning from data is as powerful as possible without cheating; in other words, combine best empirical information with best prior information
- I applaud this competition for demanding a test of general-purpose methods

Advice to Neural Net Engineers

- Don't let this competition distract you from critical prediction tasks in engineering – clean, flexible car engines; power grids; batteries; manufacturing plants; chemical plants, etc. (www.werbos.com)
- Keep your eyes on the multivariate case –causal relations to enable control, brain-like complexity
- Fill in your weakness in general-purpose modular software (MatLab⇒C ⇒chip). Example: why do people use 10,000-crash broom-balancers instead of no-crash balancers?
- Create software which makes it quick and easy for you to compete here, and learn and disseminate
- Learn to improve your accuracy in the general case by "high level debugging" analysis, extracting general principles
- Learn & teach the underlying statistical principles – simple but crucial points, not well-known even to most statisticians

"Bayes" versus "Vapnik": today's debate

- Theorem: Pr(A|B) = Pr(B|A)*Pr(A)/Pr(B)
- Platonic Bayes:
  - Predict by using stochastic model Pr($\underline{\mathbf{x}}$(t)|past)
  - Find model with highest probability of being true: Pr(Model_W|database) = Pr(database|Model_W)* Pr(Model_W)/Pr(database)
  - Neural $\underline{\mathbf{x}}$(t+1)=$\underline{\mathbf{f}}$($\underline{\mathbf{x}}$(t),…,W)+$\underline{\mathbf{e}}$(t) is just another stochastic model, with full NL regression statistics
  - Many variations; e.g. "Box-Jenkins" ARMA methods
  - "anything else is Las Vegas numerology"
- Vapnik says NO. "New" philosophy: if you want $, not truth, pick Model_W which would have maximized $ in the past (database)

Some Guidelines from Platonic Bayesian Approach

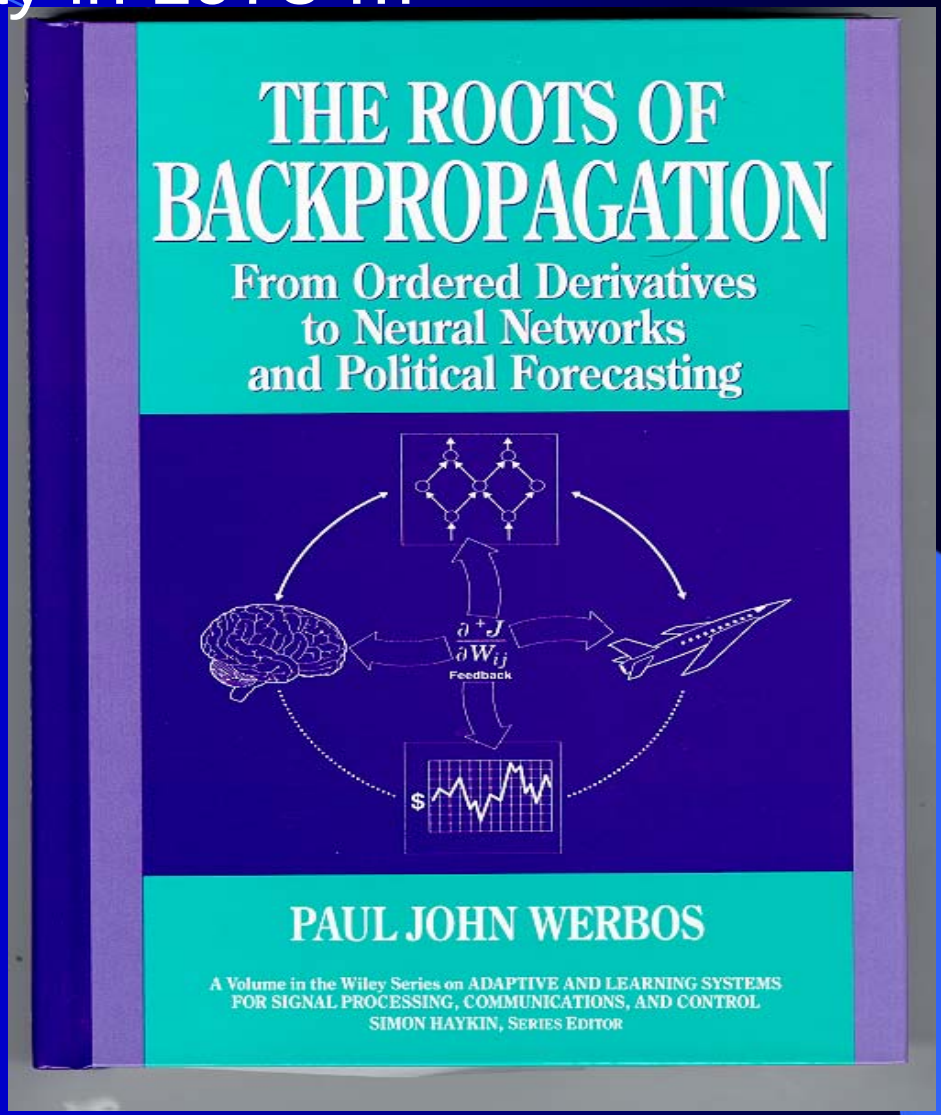
- Given two families of models or topologies, **g**(W_1) and **f**(W_2), if every model in **g** is close to a model in **f** but not vice-versa, then f is more powerful. "Almost-free lunch."
- Given enough data or given the right priors (favoring **g**-like points in **f**), **f** should always do much better than **g** or almost as well
- Examples:
  - ARMA beats AR: $x(t)+bx(t-1)=e(t)+ce(t-1)$, $c \neq 0$
  - (ARMA fits partially observed or noisy underlying AR.)
  - TLRN beats ARMA: $x(t)=e(t)+f(x(t-1),R(t-1))$
- BehavHeuristics airline seat forecasting example
- Most powerful if **f** is most universal approximator, fewer parameters. Neural vs. translog, SRN versus MLP.

But Platonic Bayes fails very badly in some ways, as I learned the hard way in 1973 …

Vector ARMA (f) had twice the prediction error of simple extrapolator (g), on 100-year political data and simulated dirty datasets

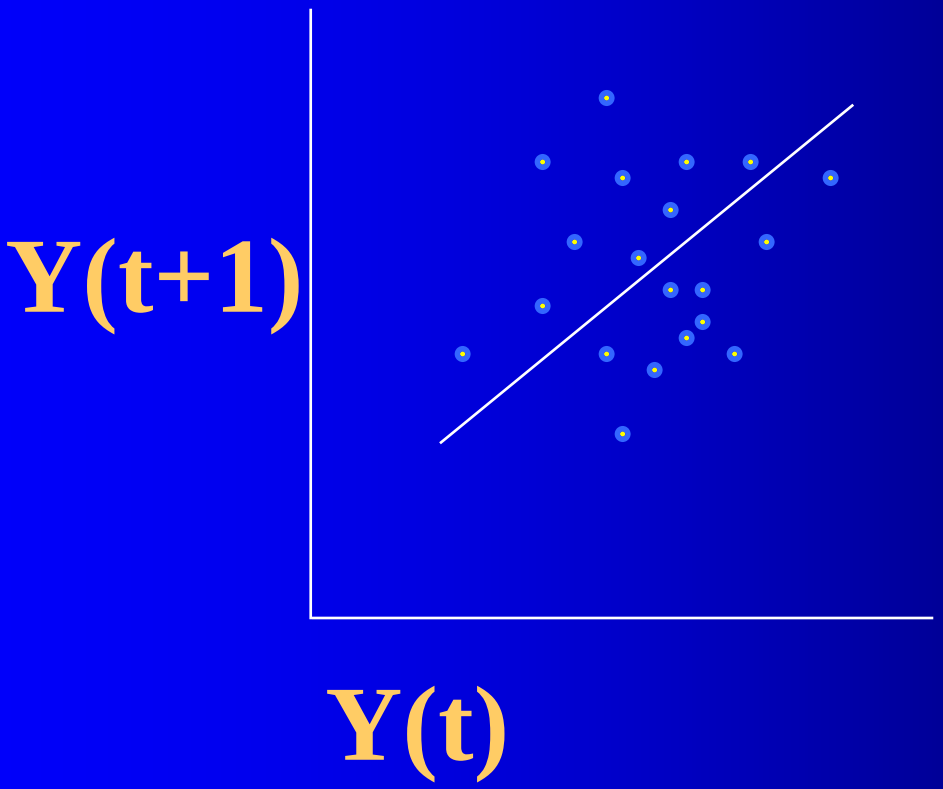
"Vapnik" style "pure robust method"

BRAINS absolutely require multiperiod robustness beyond what Platonic Bayes offers

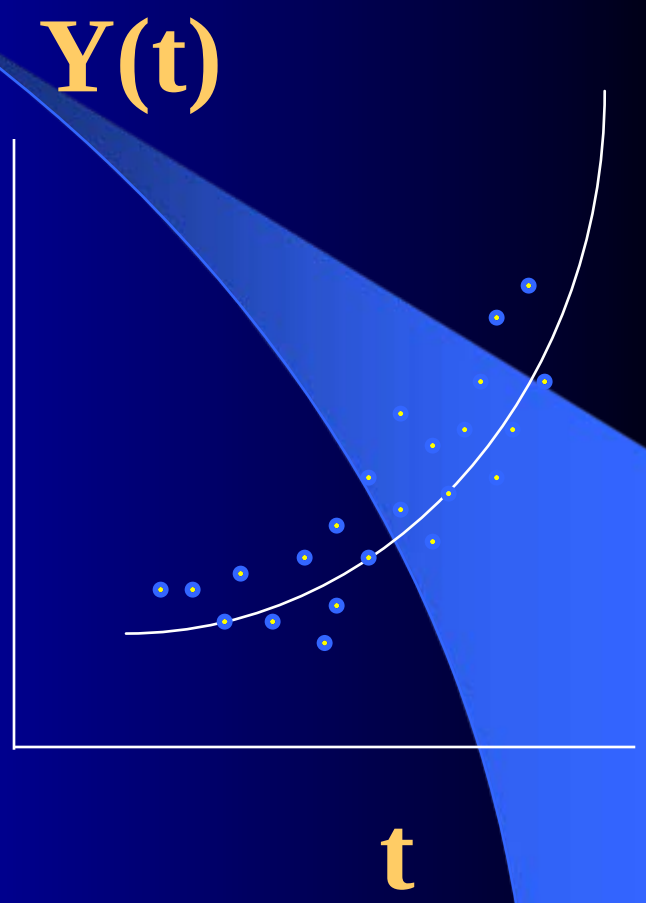


1974 Harvard PhD in subject of statistics, Mosteller on committee (Dempster help)

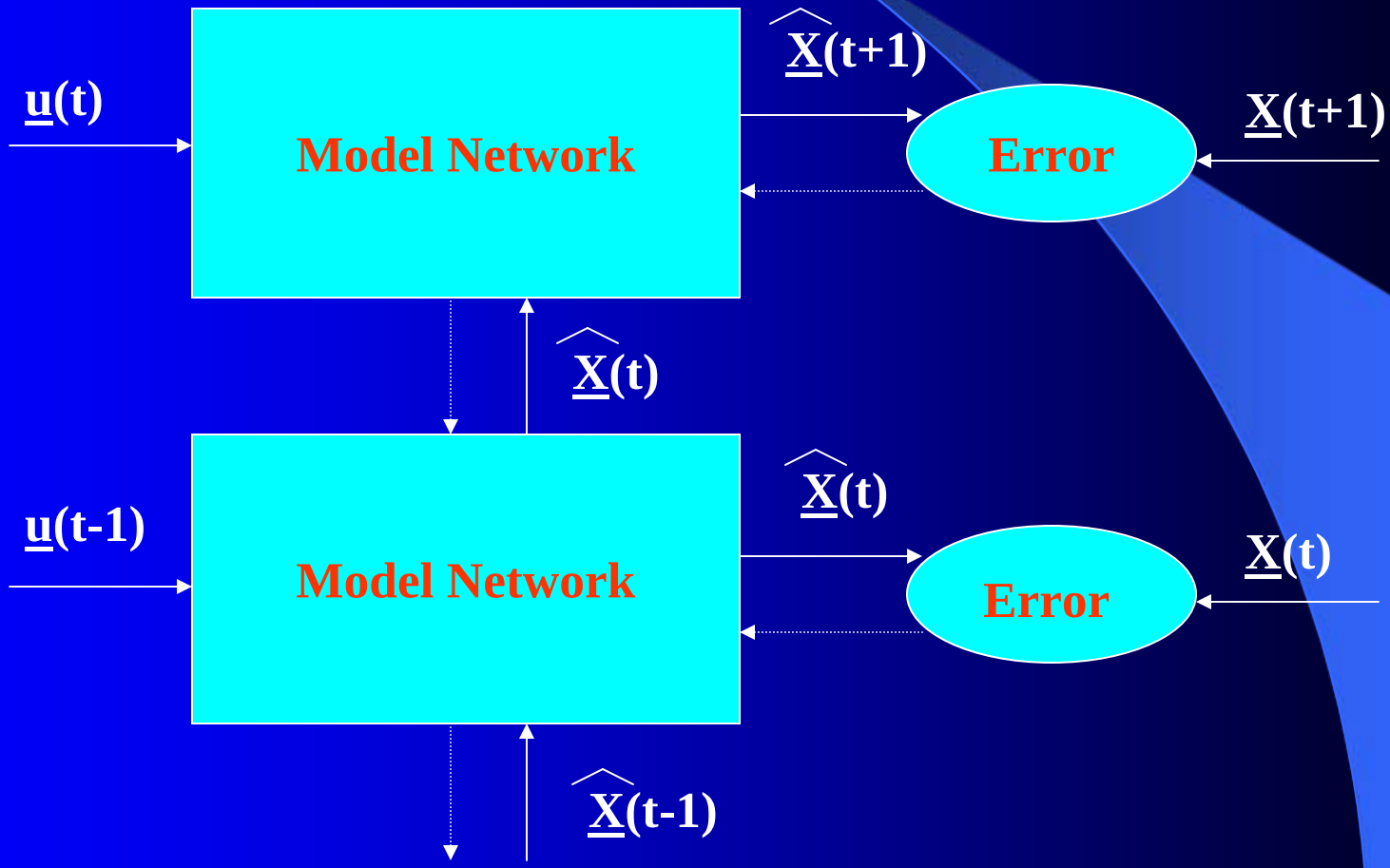
Conventional Least Squares

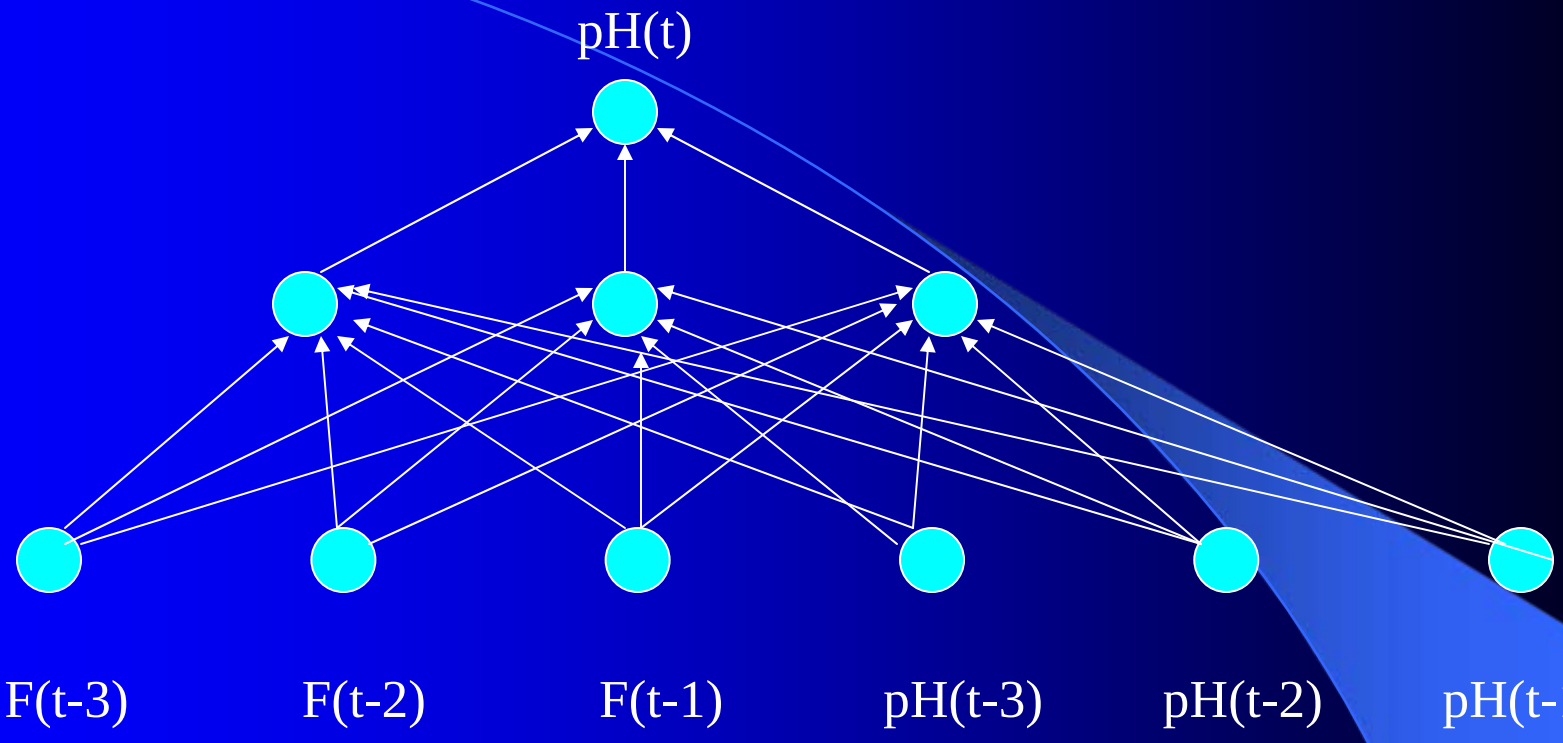


Pure Robust



PURE ROBUST METHOD

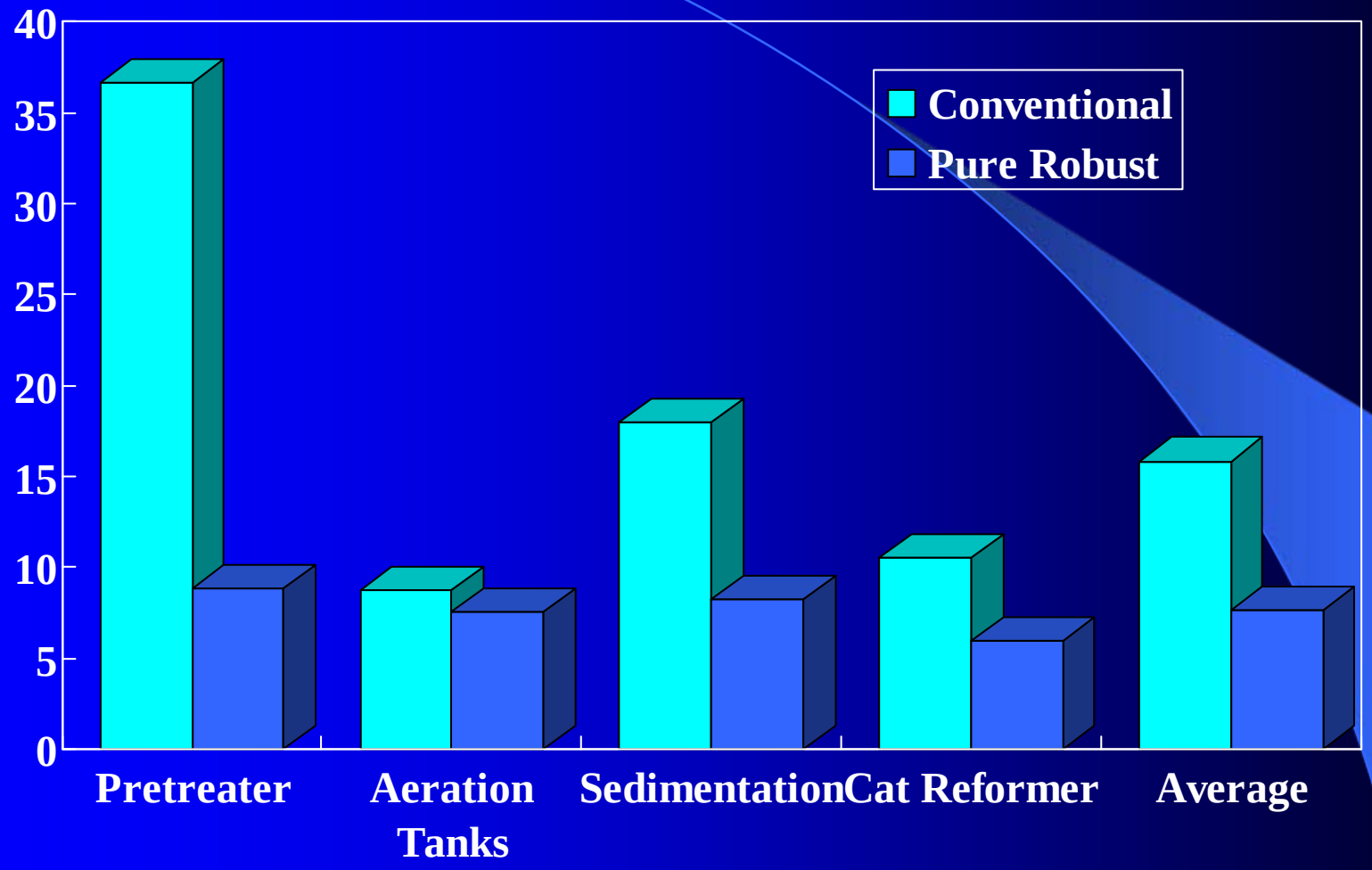




Example of TDNN used in HIC, Chapter 10

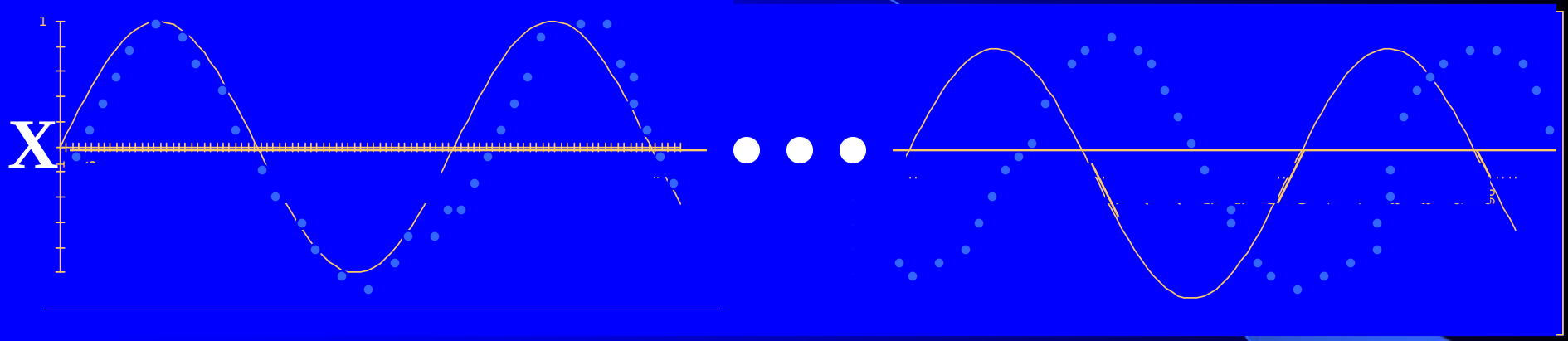
TDNNs learn NARX or FIR Models, not NARMAX or IIR

Prediction Errors (HIC p.319)



- Greatest advantage on real-world data (versus simulated)
- Full details in chapter 10 of HIC, posted at www.werbos.com.
- Statistical theory (and how to do better) in second half of that chapter.

But Pure Robust ("Vapnik") Can Fail Badly Too: Phase Drift



$$R(t+1) = R(t) + w + e_p(t)$$

$$X(t) = \sin R(t) + e_m(t)$$

TINY

A unified method cut GNP errors in half on Latin American data, versus maximum likelihood and pure robust both (SMC 78, econometric).

“Vapnik” approach is not new even in the static case

- Utilitarian Bayes: google “Raiffa Bayesian”: pick model and weights W so as to minimize a loss function L.
- Example of the issue: to weight or not weight your regression (in actual DOE/EIA model and conflict model):

  Energy(state,year)=a*income(state,year)+e(year) (1)

  (energy(state,year)/income(state,year)=a+e(year) (2)

  If big states different, equation (1) is more consistent

  If big states few, (2) has more information, less random error

  Platonic approach: use F tests to see which is more true, but..

NonBayesian methods in econometrics for consistency under more general conditions

The Prior Term Pr($Model_W$) is crucial, in Bayesian or robustified statistics

- Not just specific domain knowledge, but key basic principles like Occam's Razor – that Pr($Model_W$) is greater for simpler models. See Emmanuel Kant: "apriori analytic." New jargon: "uninformative priors" and "metastatistics."
- Under old school "flat priors," human brain could not exist. Too many variables.
- 1977: to handle complexity (many input variables), ridge regression – empirical Bayes, estimated pr(W_i).
- For ANNs: penalty functions, robustified by allowing redundancy (Phatak); symmetry (see brain paper)…; and "syncretism," unification of memory and prediction. Symmetry+TLRN and proper loss function was how we got 6% per month above Dow in 1990's..