CSE 6242 / CX 4242

Time Series

Nonlinear Forecasting; Visualization; Applications

Duen Horng (Polo) Chau
Georgia Tech

Some lectures are partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos, Le Song

Last Time

Similarity search

- Euclidean distance
- Time-warping

Linear Forecasting

- AR (Auto Regression) methodology
- RLS (Recursive Least Square)
  = fast, incremental least square

This Time

Linear Forecasting

- Co-evolving time sequences

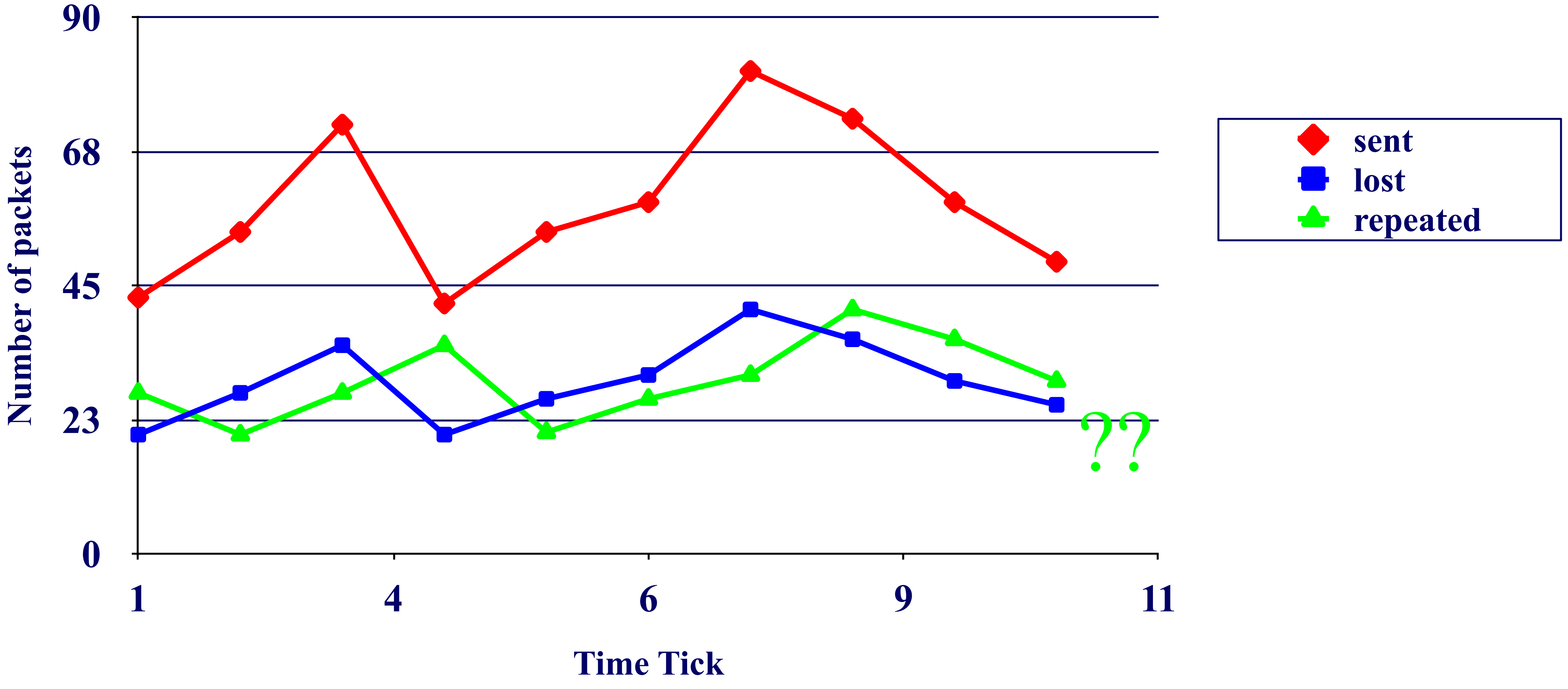
Non-linear forecasting

- Lag-plots + k-NN

Visualization and Applications

Co-Evolving Time Sequences

- Given: A set of **correlated** time sequences
- Forecast ‘**Repeated(t)**’



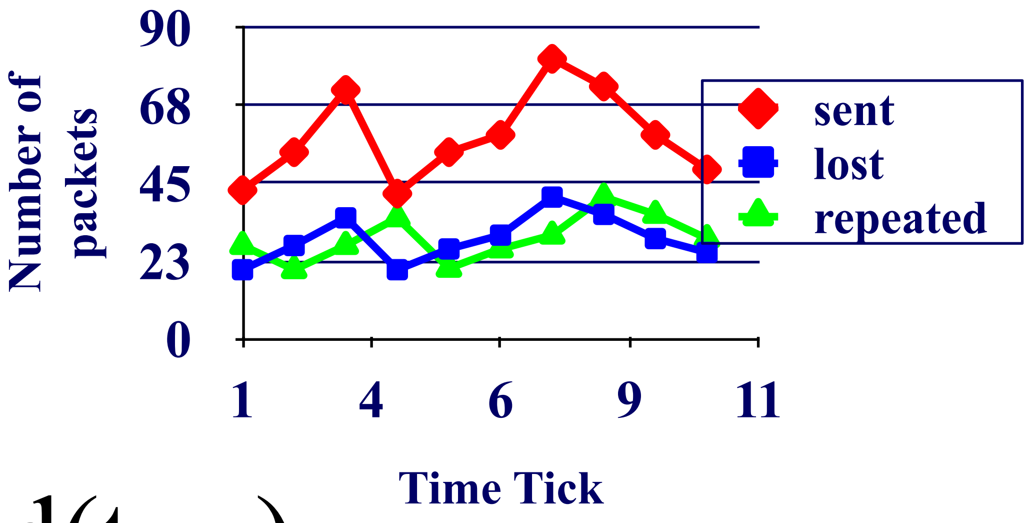
Solution:

Q: what should we do?

Solution:

Least Squares, with

- Dep. Variable: Repeated(t)
- Indep. Variables:
  - Sent(t-1) … Sent(t-w);
  - Lost(t-1) …Lost(t-w);
  - Repeated(t-1), Repeated(t-w)
- (named: ‘MUSCLES’ [Yi+00])



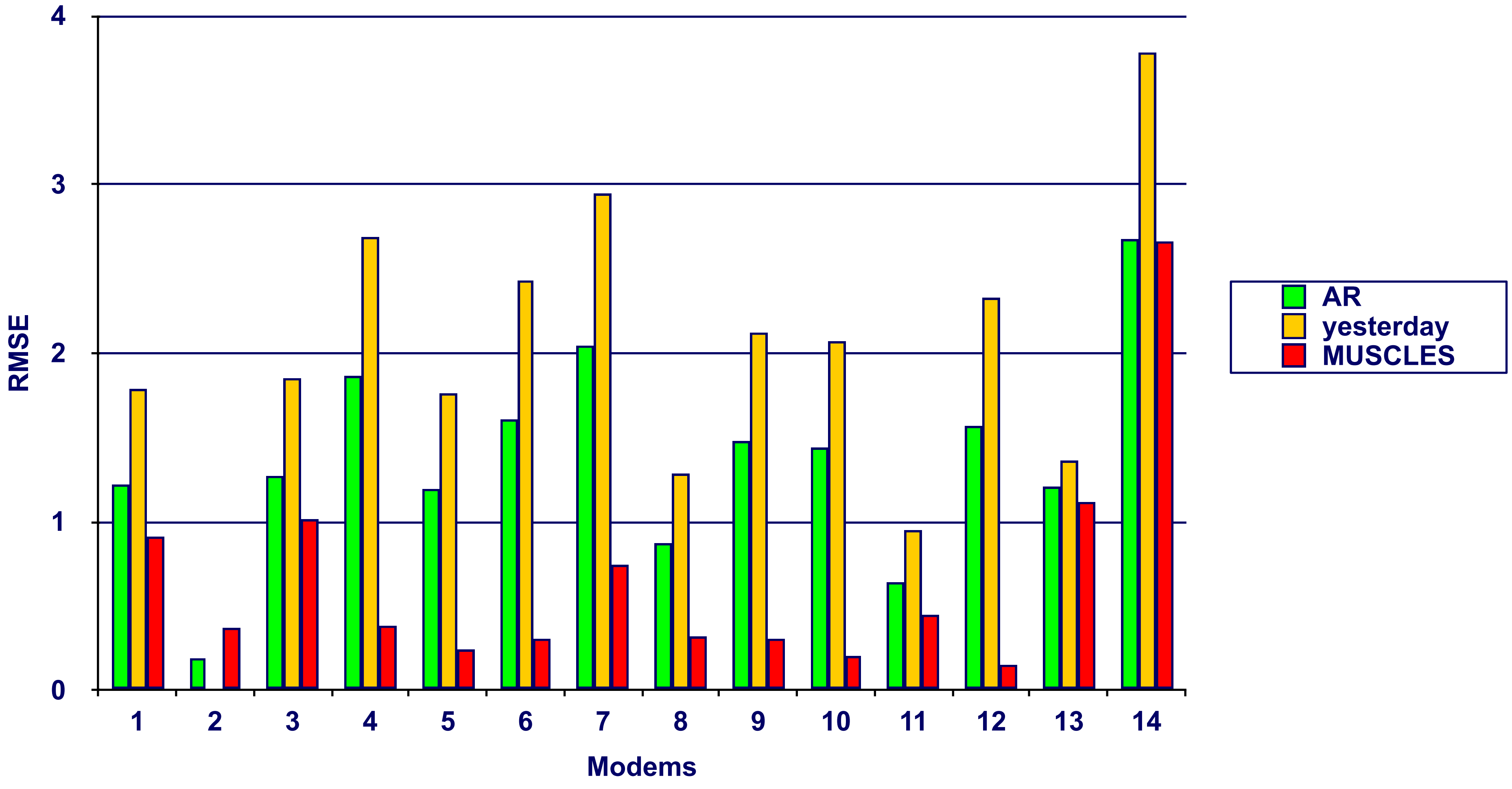
Forecasting - Outline

- Auto-regression
- Least Squares; recursive least squares
- Co-evolving time sequences
- Examples
- Conclusions

Examples - Experiments

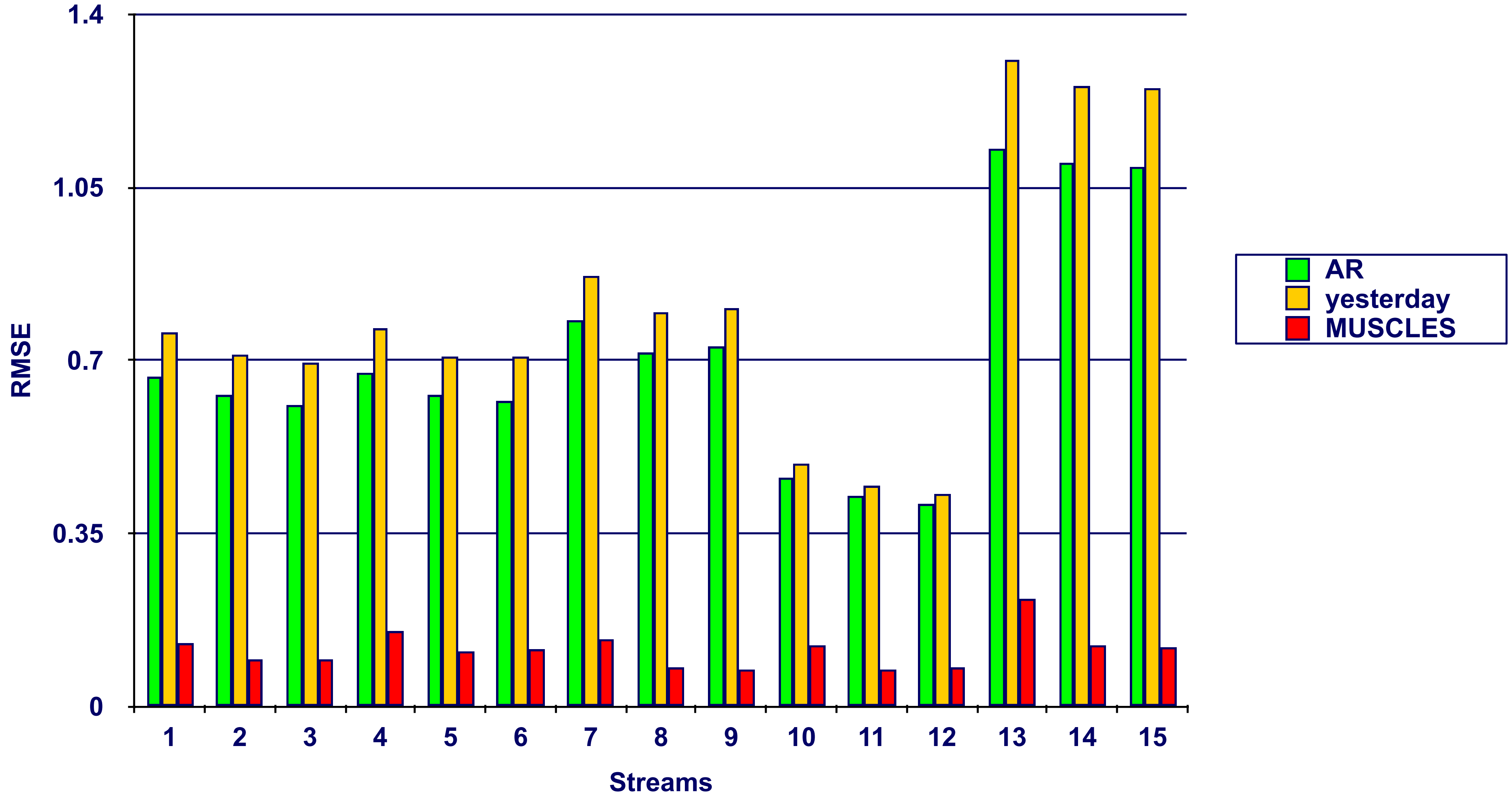
- Datasets
  - Modem pool traffic (14 modems, 1500 time-ticks; #packets per time unit)
  - AT&T WorldNet internet usage (several data streams; 980 time-ticks)
- Measures of success
  - Accuracy : Root Mean Square Error (RMSE)

Accuracy - “Modem”



MUSCLES outperforms AR & “yesterday”

Accuracy - “Internet”



MUSCLES consistently outperforms AR & “yesterday”

Linear forecasting - Outline

- Auto-regression
- Least Squares; recursive least squares
- Co-evolving time sequences
- Examples
- Conclusions

Conclusions - Practitioner’s guide

- AR(IMA) methodology: prevailing method for linear forecasting
- Brilliant method of Recursive Least Squares for fast, incremental estimation.

Resources: software and urls

- MUSCLES: Prof. Byoung-Kee Yi: http://www.postech.ac.kr/~bkyi/ or christos@cs.cmu.edu
- R
  http://cran.r-project.org/

Books

- George E.P. Box and Gwilym M. Jenkins and Gregory C. Reinsel, *Time Series Analysis: Forecasting and Control*, Prentice Hall, 1994 (the classic book on ARIMA, 3rd ed.)
- Brockwell, P. J. and R. A. Davis (1987). Time Series: Theory and Methods. New York, Springer Verlag.

Additional Reading

- [Papadimitriou+ vldb2003] Spiros Papadimitriou, Anthony Brockwell and Christos Faloutsos *Adaptive, Hands-Off Stream Mining* VLDB 2003, Berlin, Germany, Sept. 2003
- [Yi+00] Byoung-Kee Yi et al.: *Online Data Mining for Co-Evolving Time Sequences*, ICDE 2000. (Describes MUSCLES and Recursive Least Squares)

Outline

- Motivation
- ...
- Linear Forecasting
- Non-linear forecasting
- Conclusions

Chaos & non-linear forecasting

Reference:

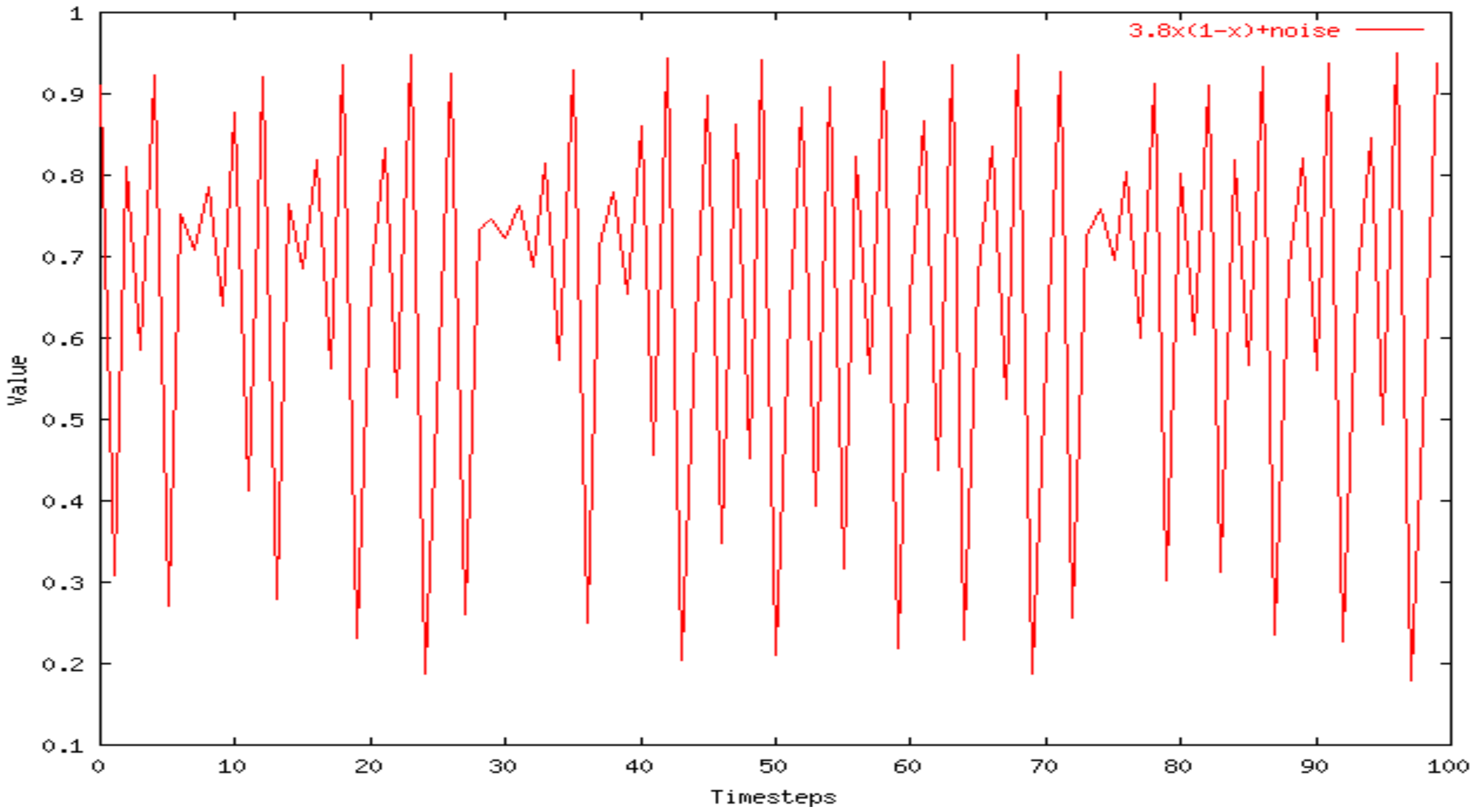
[Deepay Chakrabarti and Christos Faloutsos *F4: Large-Scale Automated Forecasting using Fractals* CIKM 2002, Washington DC, Nov. 2002.]

Detailed Outline

- Non-linear forecasting
  - Problem
  - Idea
  - How-to
  - Experiments
  - Conclusions

Recall: Problem #1

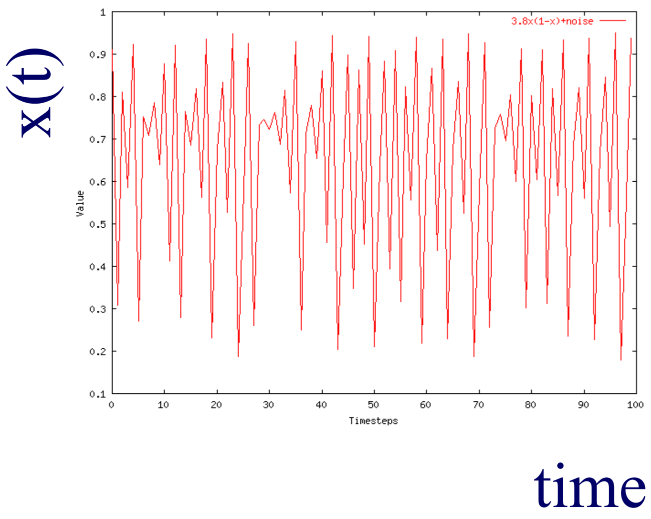
Value



Time

Given a time series $\{x_t\}$, predict its future course, that is, x_{t+1}, x_{t+2}, ...

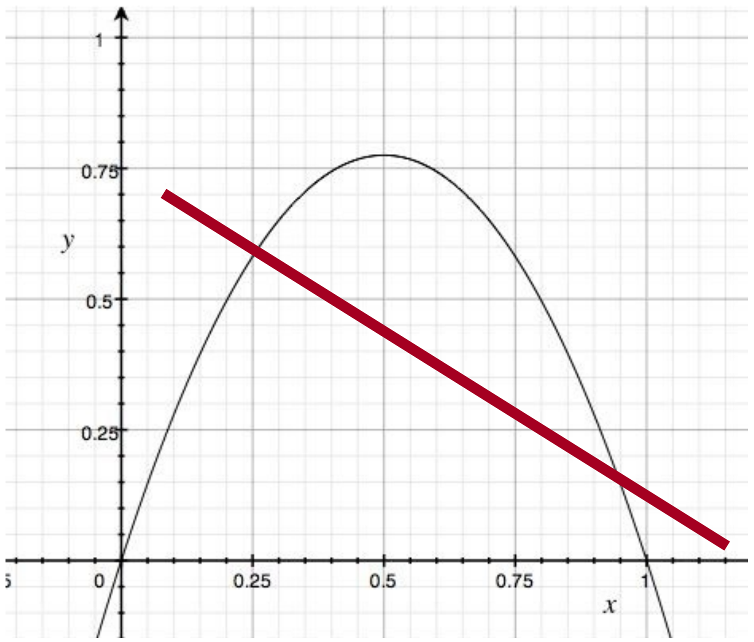
Datasets



Logistic Parabola:

$x_t = ax_{t-1}(1-x_{t-1})$ + noise

Models population of flies [R. May/1976]

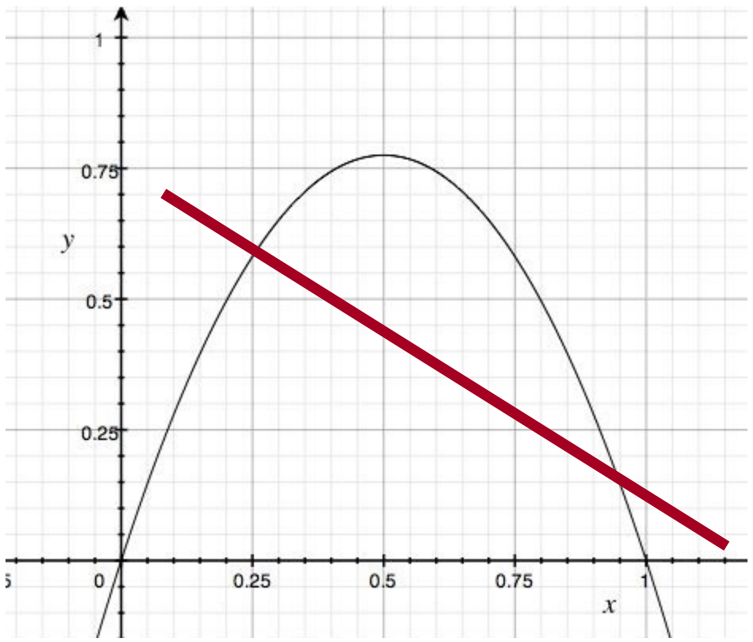


Lag-plot

ARIMA: fails

How to forecast?

- ARIMA - but: linearity assumption



Lag-plot

ARIMA: fails

How to forecast?

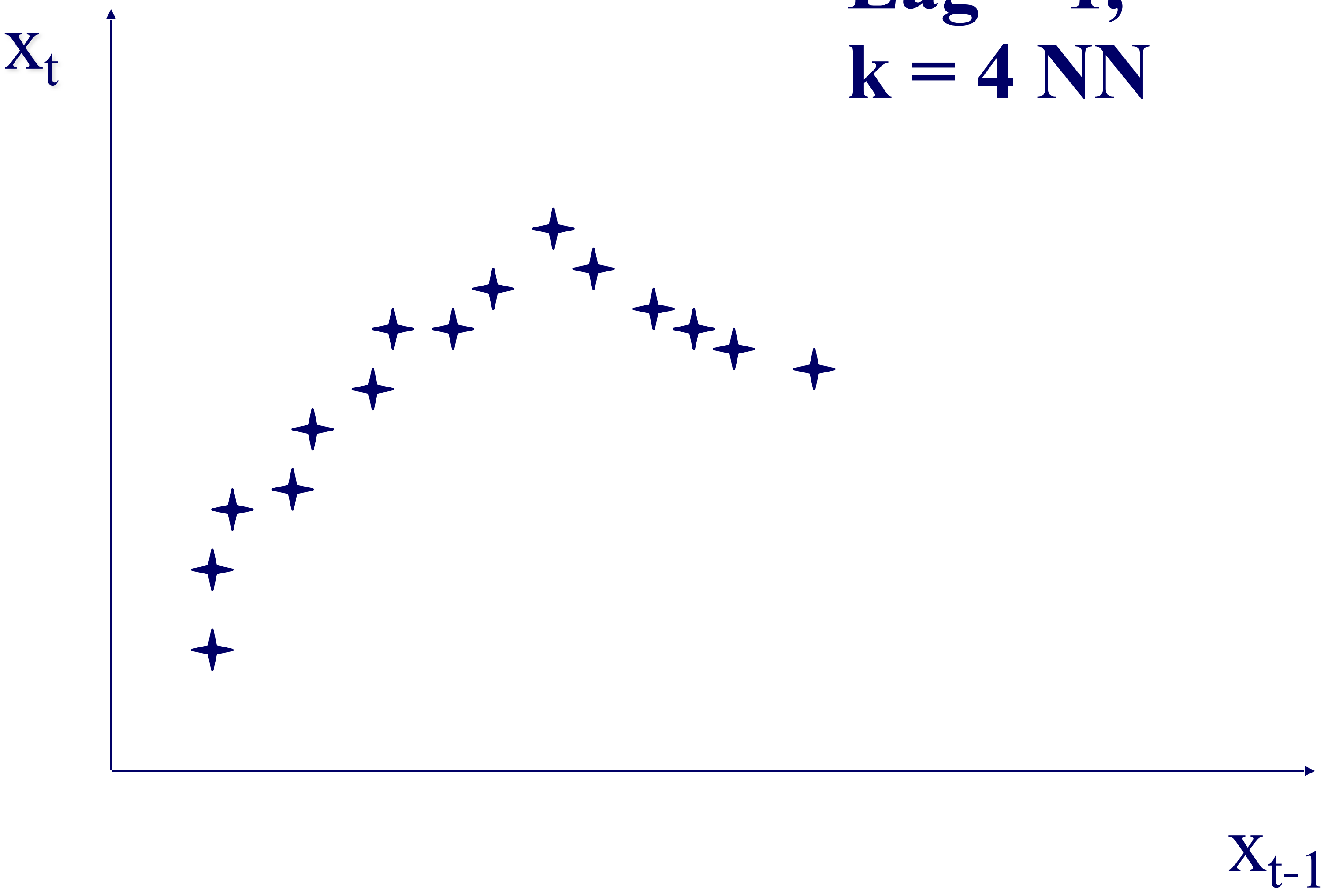
- ARIMA - but: linearity assumption

- ANSWER: ‘Delayed Coordinate Embedding’ = Lag Plots [Sauer92]

  ~ nearest-neighbor search, for past incidents

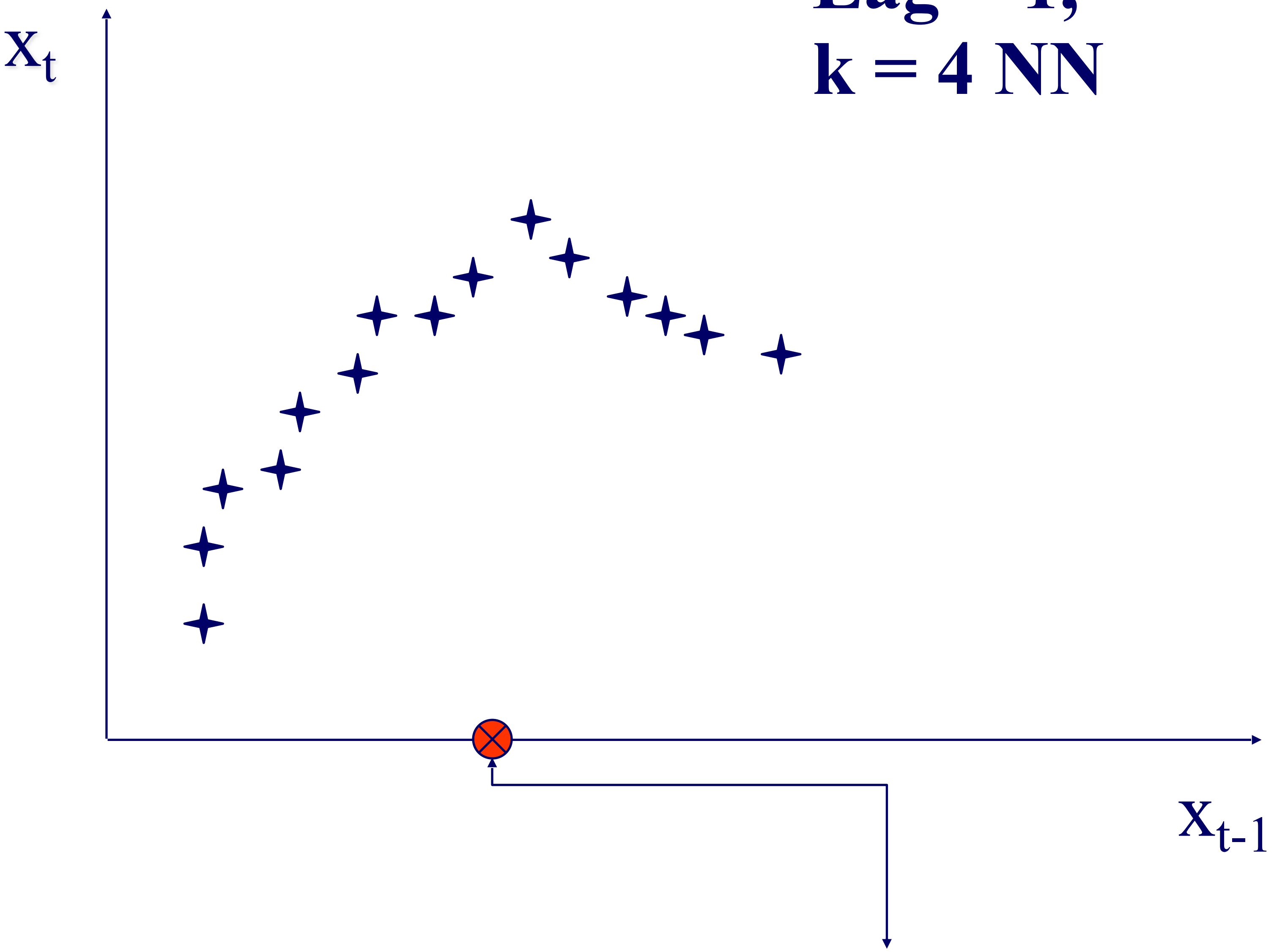
General Intuition (Lag Plot)

Lag = 1,
k = 4 NN



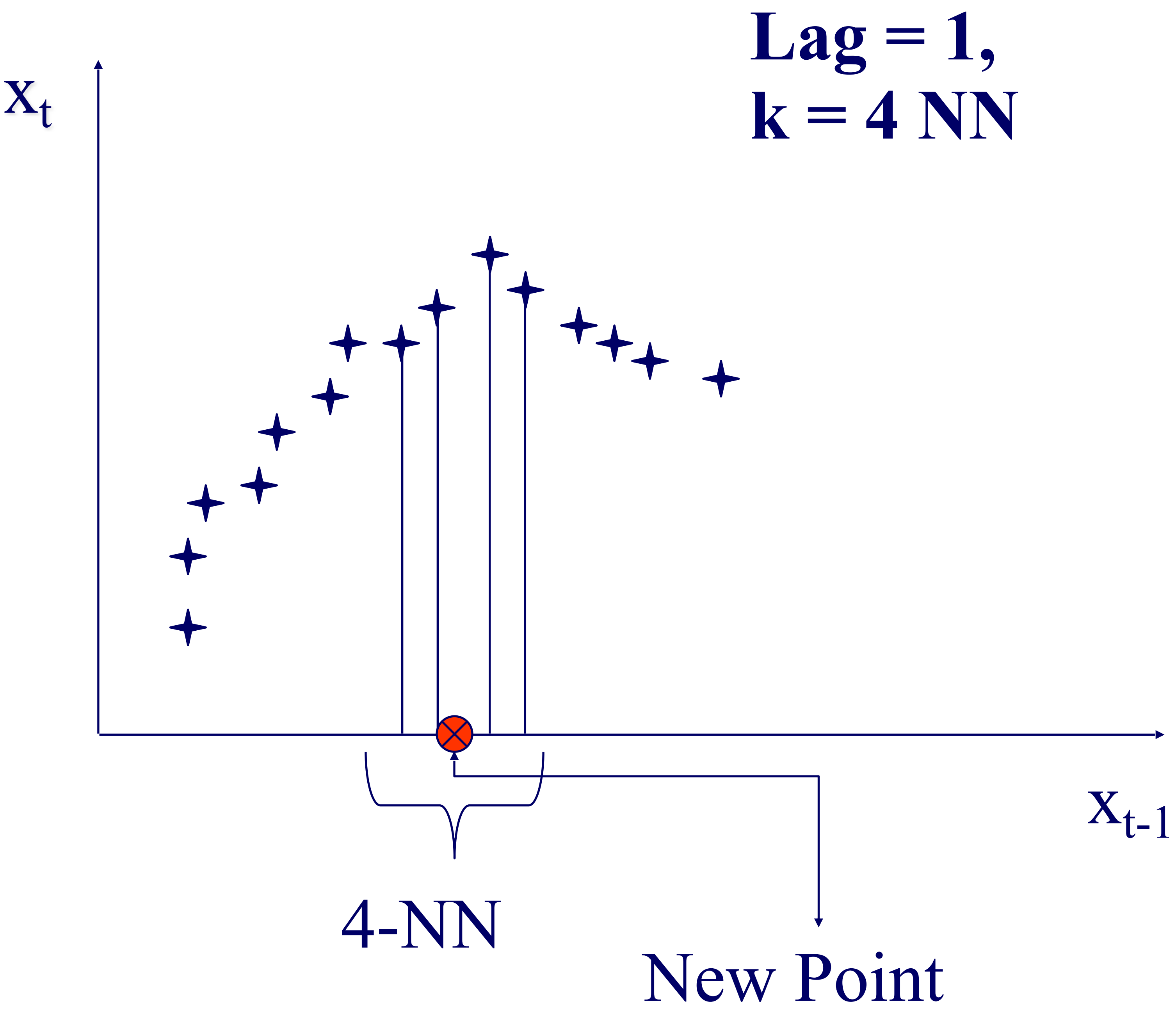
General Intuition (Lag Plot)

Lag = 1,
k = 4 NN



New Point

General Intuition (Lag Plot)



Lag = 1,
k = 4 NN

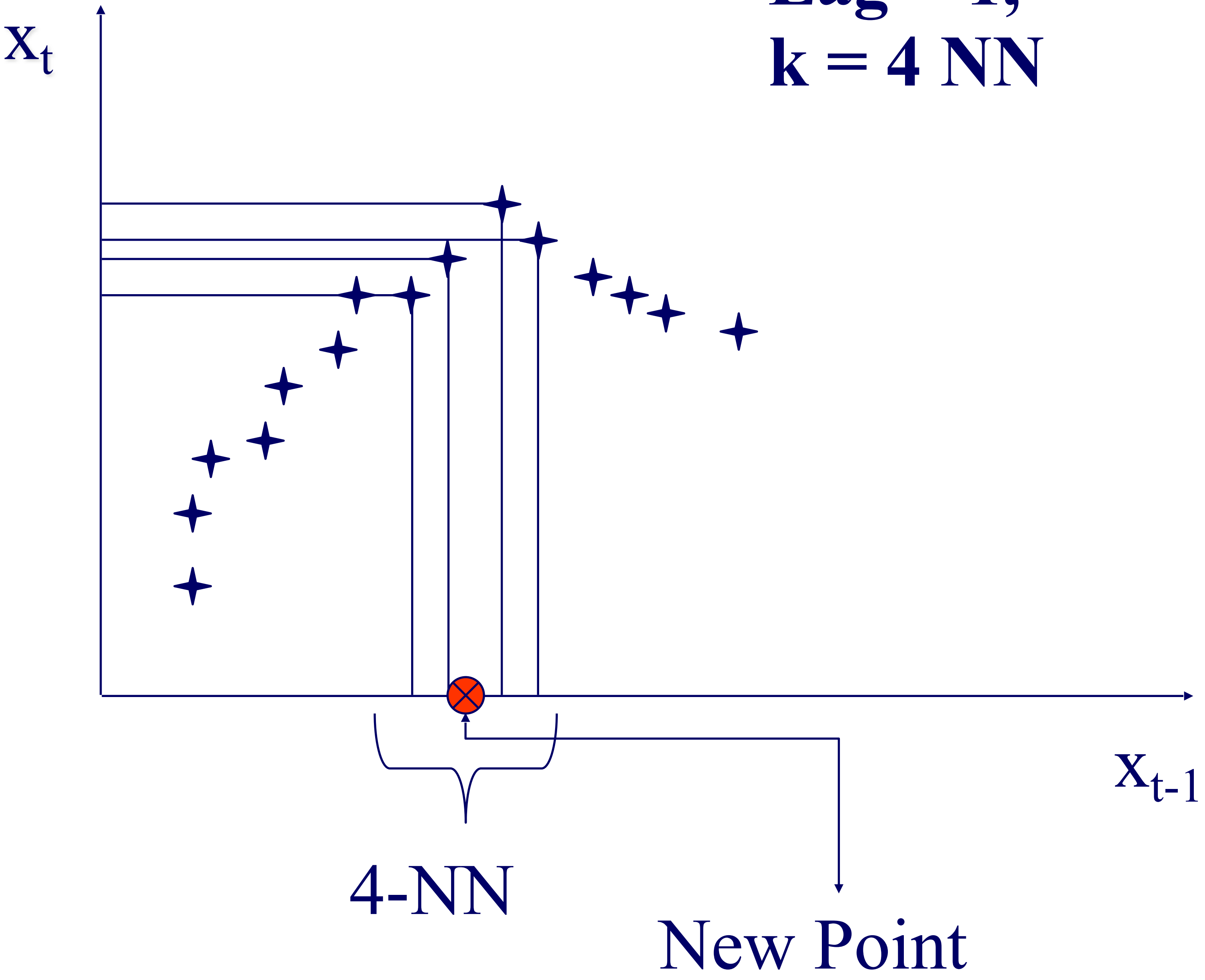
x_t

x_{t-1}

4-NN

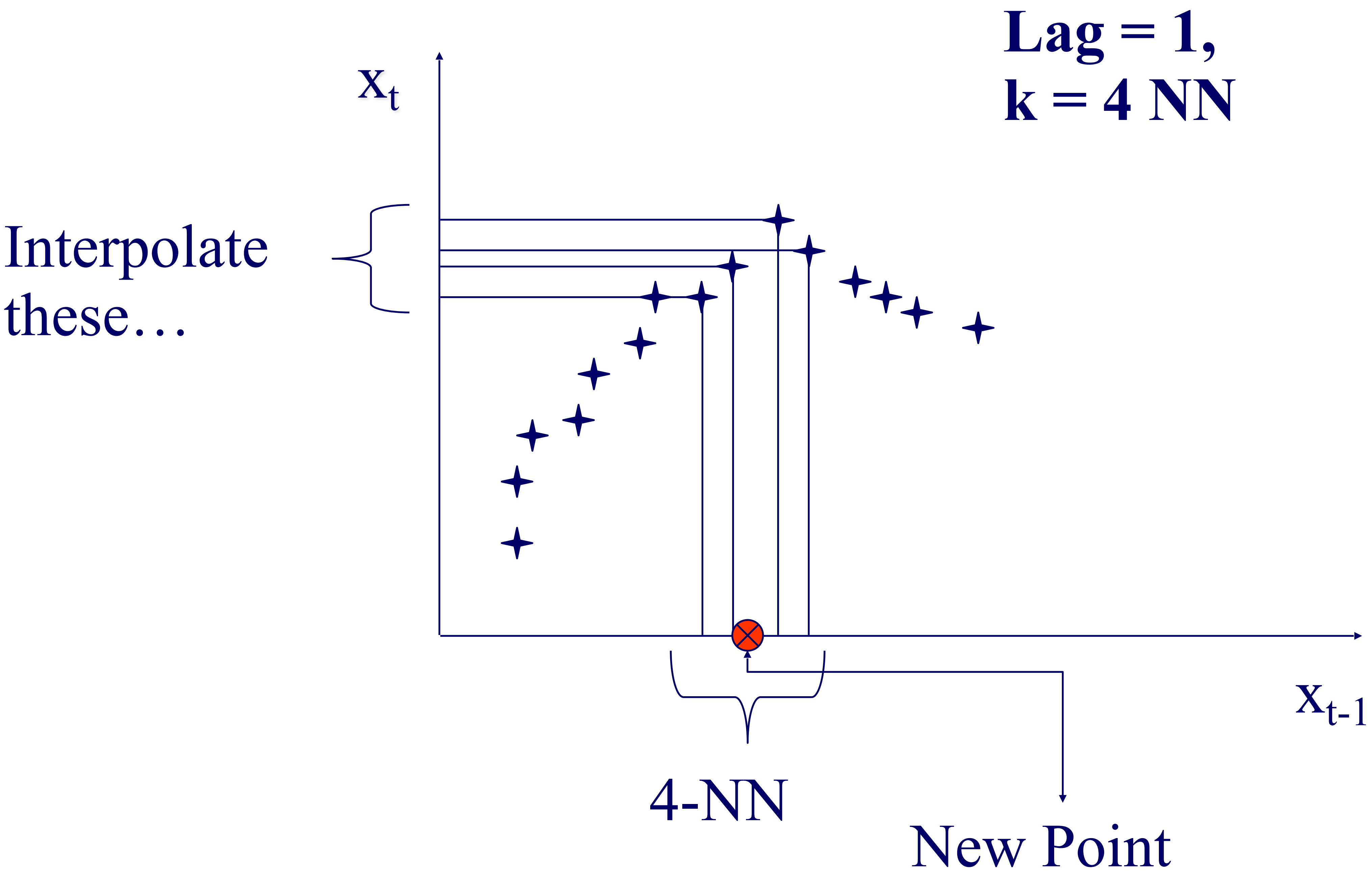
New Point

General Intuition (Lag Plot)



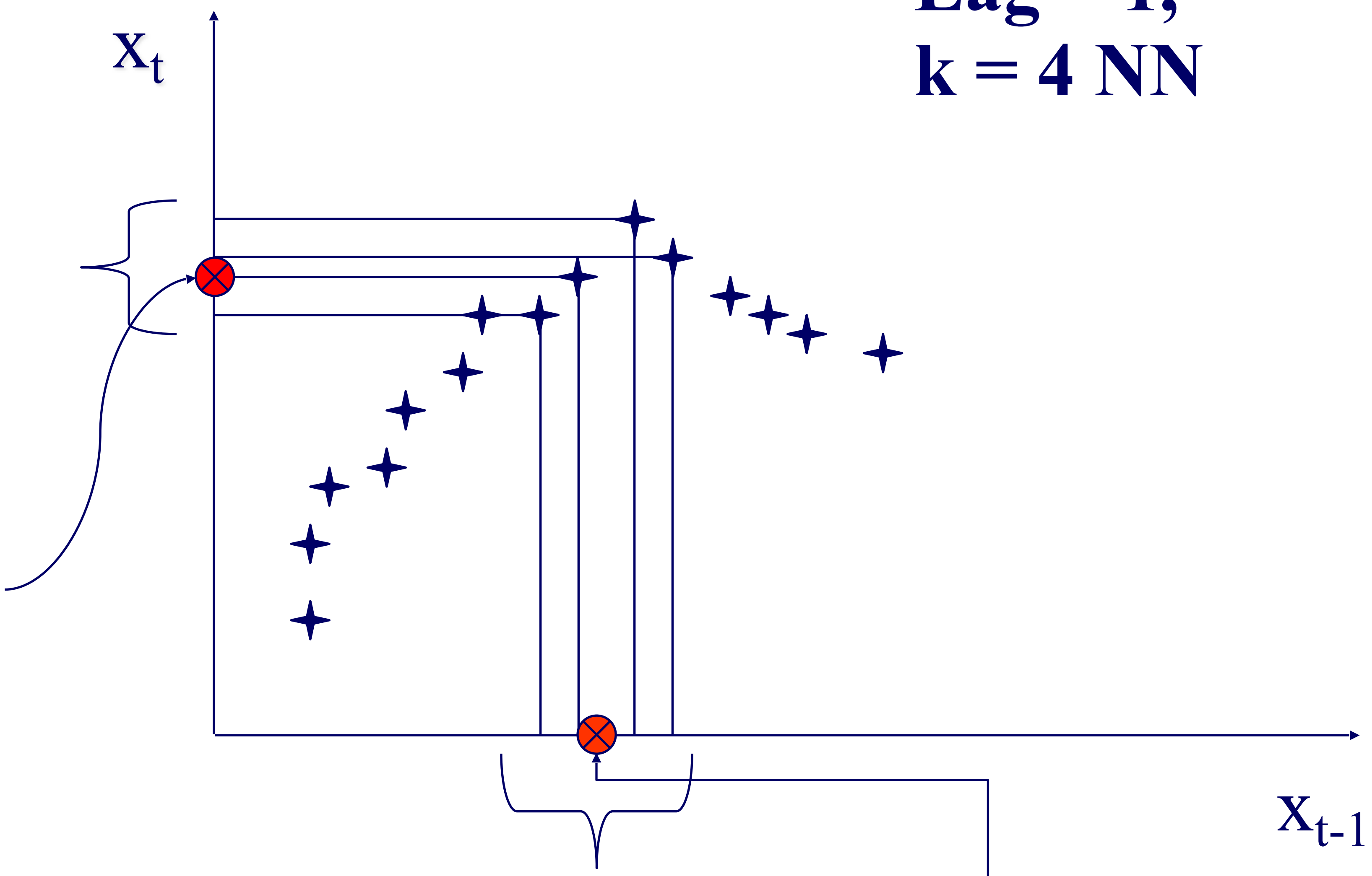
General Intuition (Lag Plot)

Lag = 1,
k = 4 NN



General Intuition (Lag Plot)

Lag = 1,
k = 4 NN



Interpolate these…

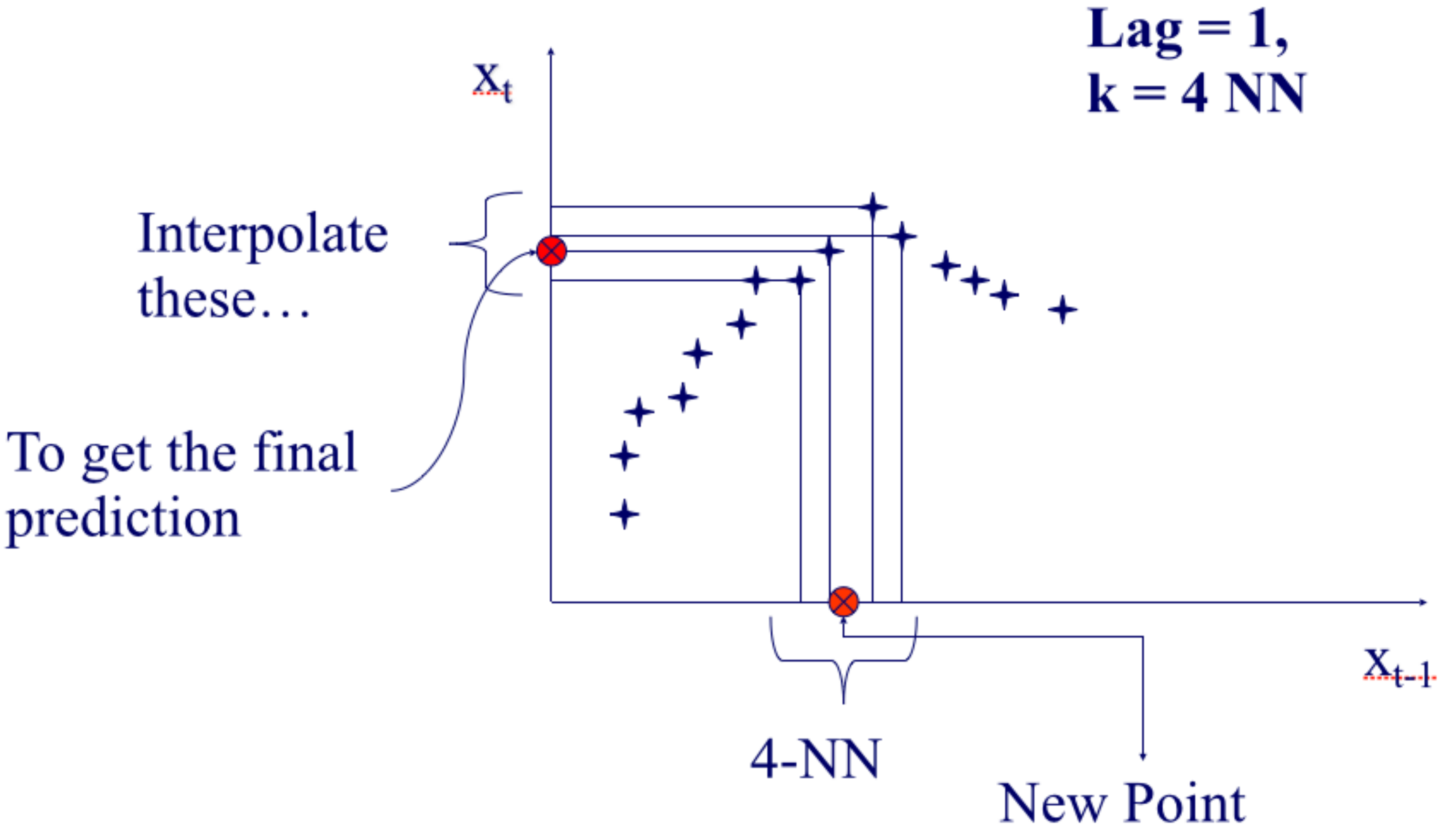
To get the final prediction

4-NN

New Point

Questions:

- Q1: How to choose lag *L*?
- Q2: How to choose *k* (the # of NN)?
- Q3: How to interpolate?
- Q4: why should this work at all?



Lag = 1,
k = 4 NN

Interpolate these…

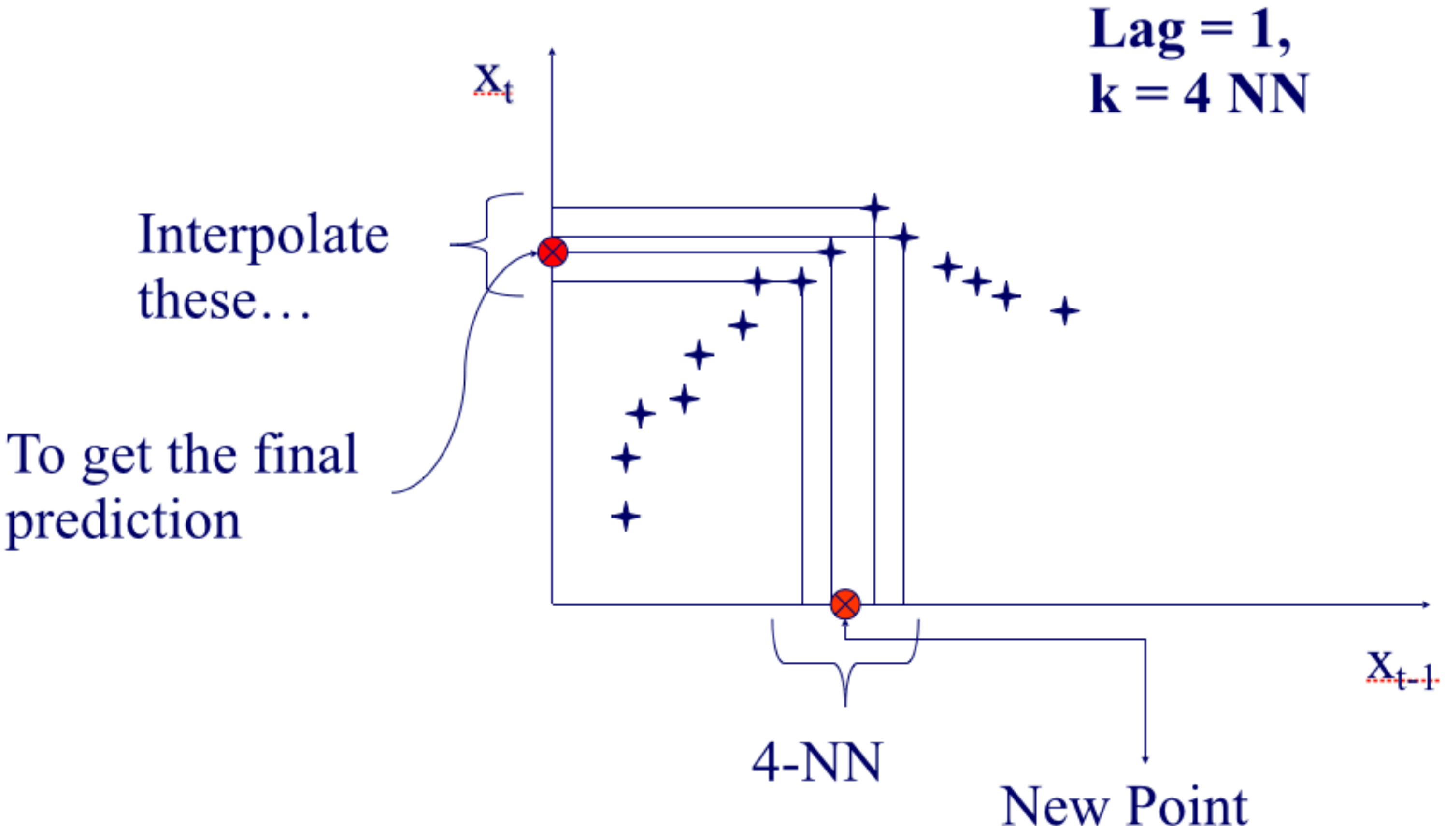
To get the final prediction

4-NN

New Point

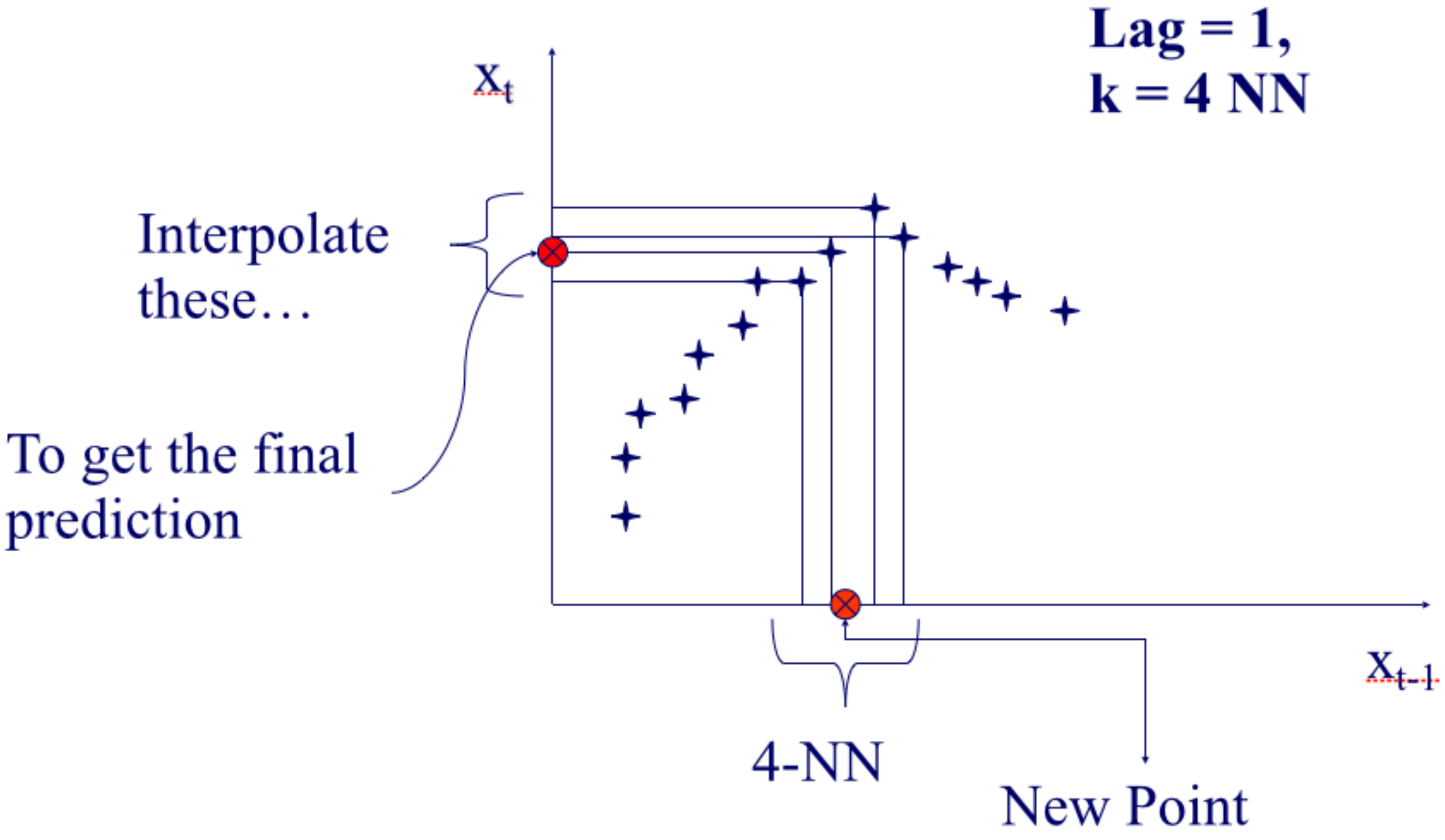
Q1: Choosing lag *L*

- Manually (16, in award winning system by [Sauer94])



Q2: Choosing number of neighbors *k*

- Manually (typically ~ 1-10)

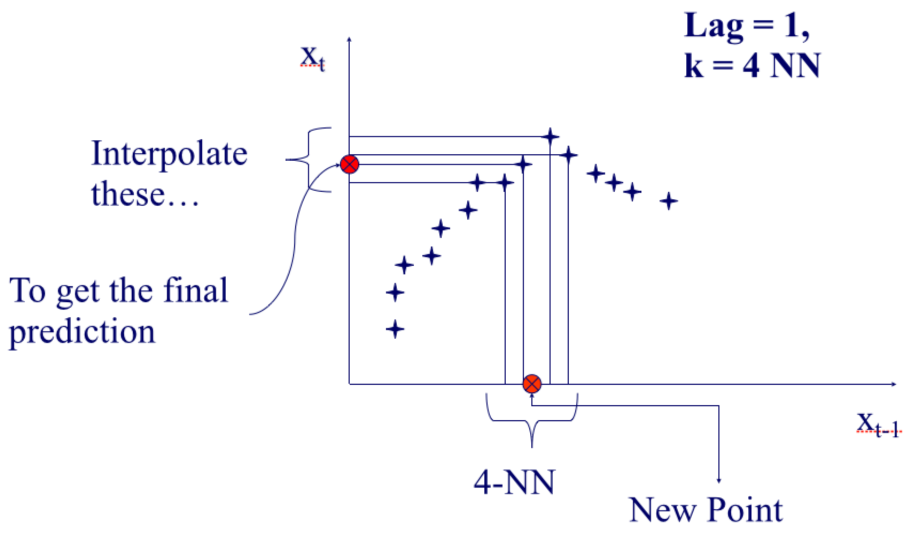


Q3: How to interpolate?

How do we interpolate between the *k* nearest neighbors?

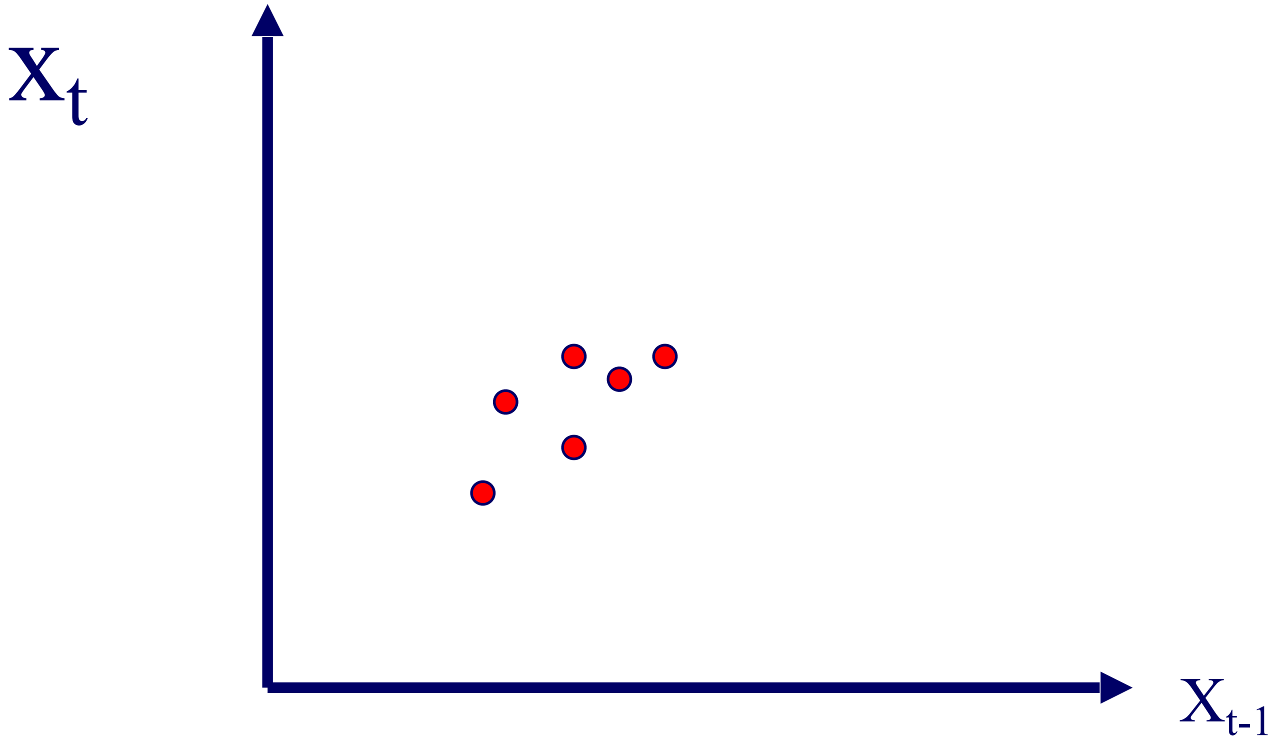
A3.1: Average

A3.2: Weighted average (weights drop with distance - how?)



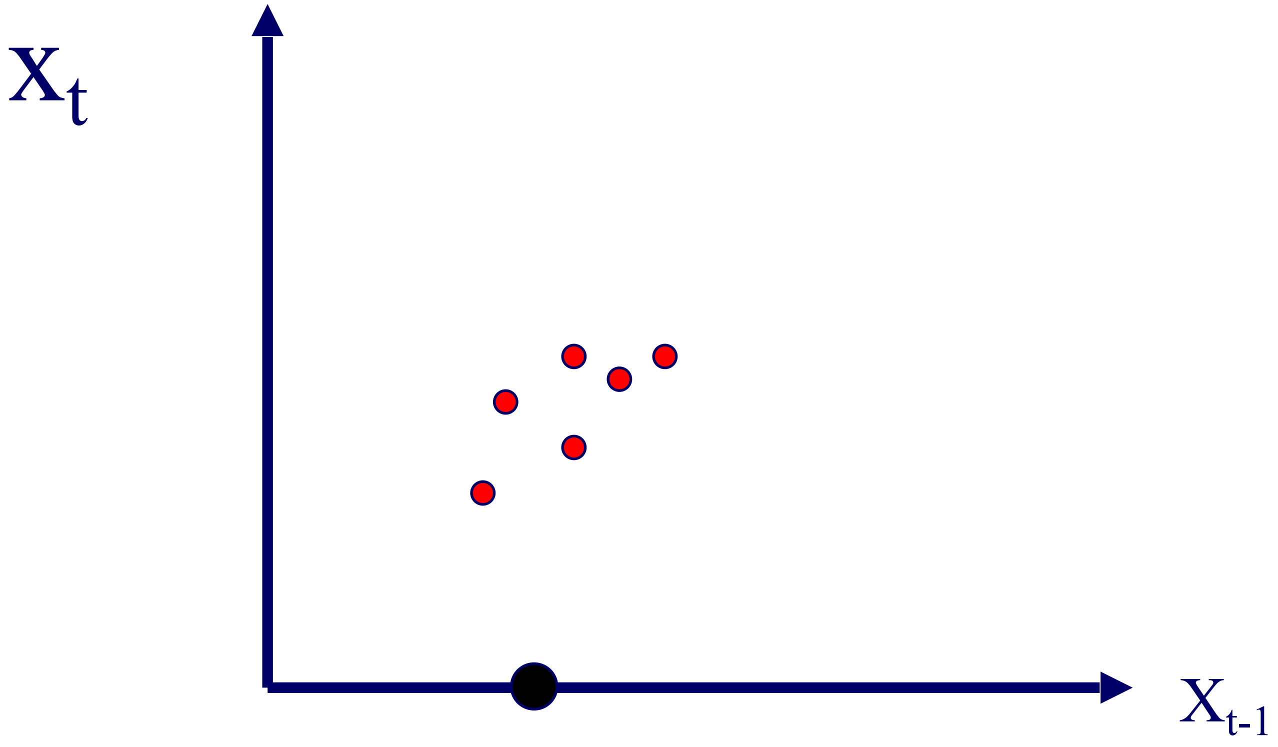
Q3: How to interpolate?

A3.3: Using SVD - seems to perform best ([Sauer94] - first place in the Santa Fe forecasting competition)



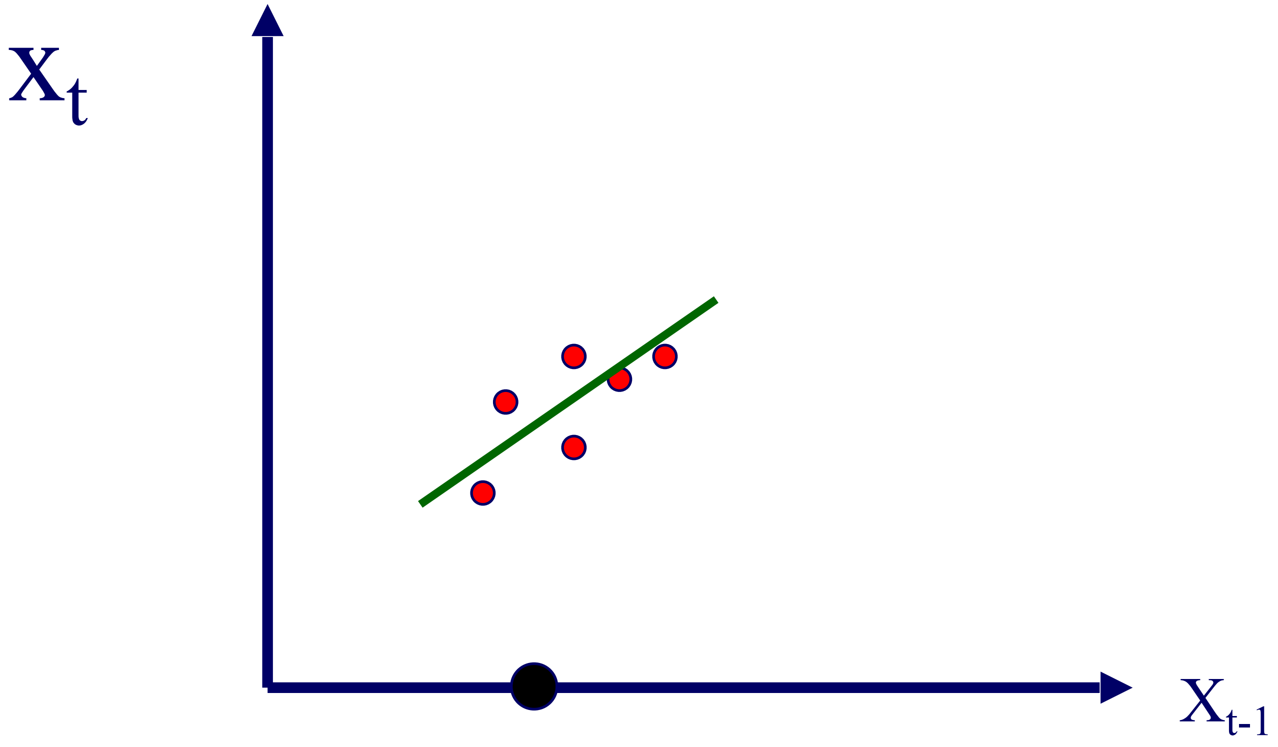
Q3: How to interpolate?

A3.3: Using SVD - seems to perform best ([Sauer94] - first place in the Santa Fe forecasting competition)



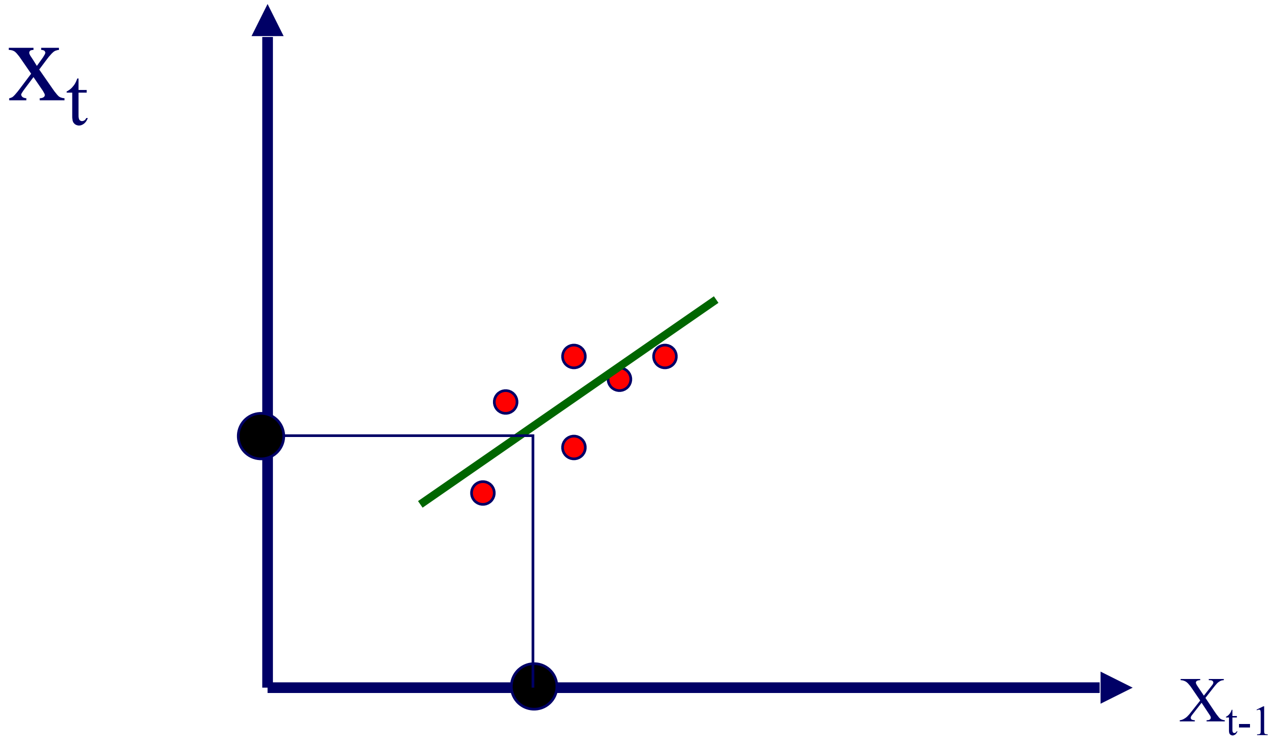
Q3: How to interpolate?

A3.3: Using SVD - seems to perform best ([Sauer94] - first place in the Santa Fe forecasting competition)



Q3: How to interpolate?

A3.3: Using SVD - seems to perform best ([Sauer94] - first place in the Santa Fe forecasting competition)



Q4: Any theory behind it?

A4: YES!

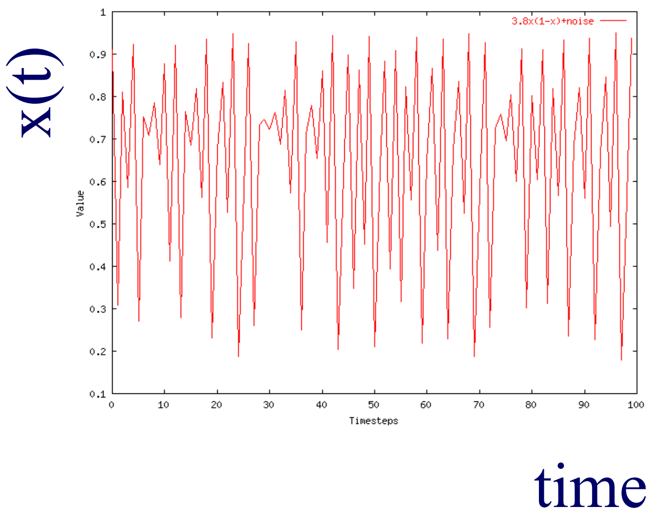
Theoretical foundation

- Based on the ‘Takens theorem’ [Takens81]
- which says that long enough delay vectors can do prediction, even if there are unobserved variables in the dynamical system (= diff. equations)

Detailed Outline

- Non-linear forecasting
  - Problem
  - Idea
  - How-to
  - Experiments
  - Conclusions

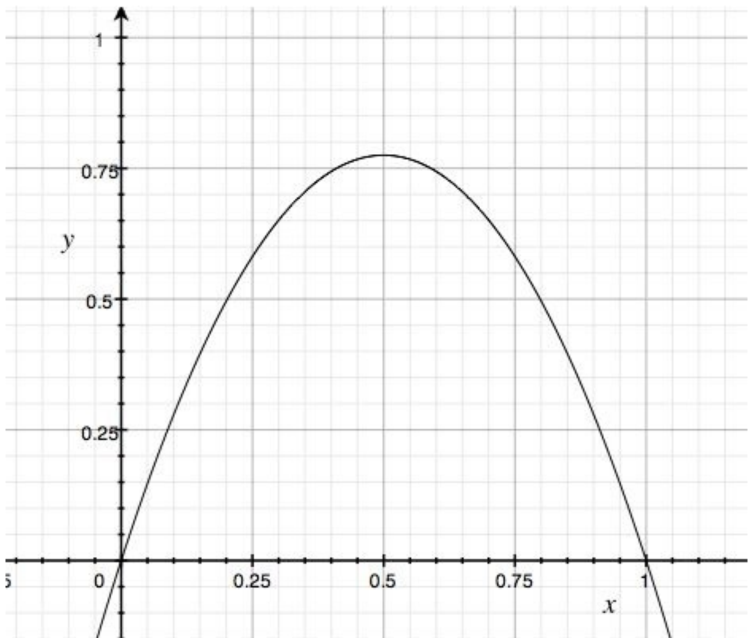
Datasets



Logistic Parabola:

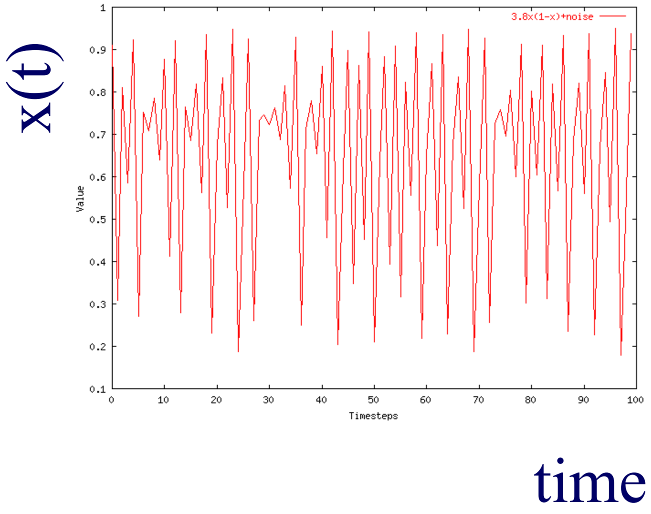
$x_t = ax_{t-1}(1-x_{t-1}) + \text{noise}$

Models population of flies [R. May/1976]



Lag-plot

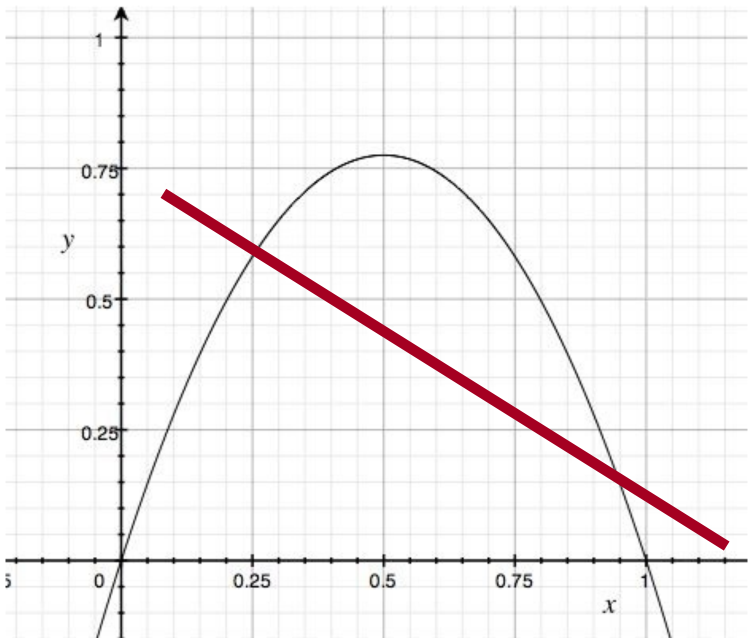
Datasets



Logistic Parabola:

$x_t = ax_{t-1}(1-x_{t-1})$ + noise

Models population of flies [R. May/1976]



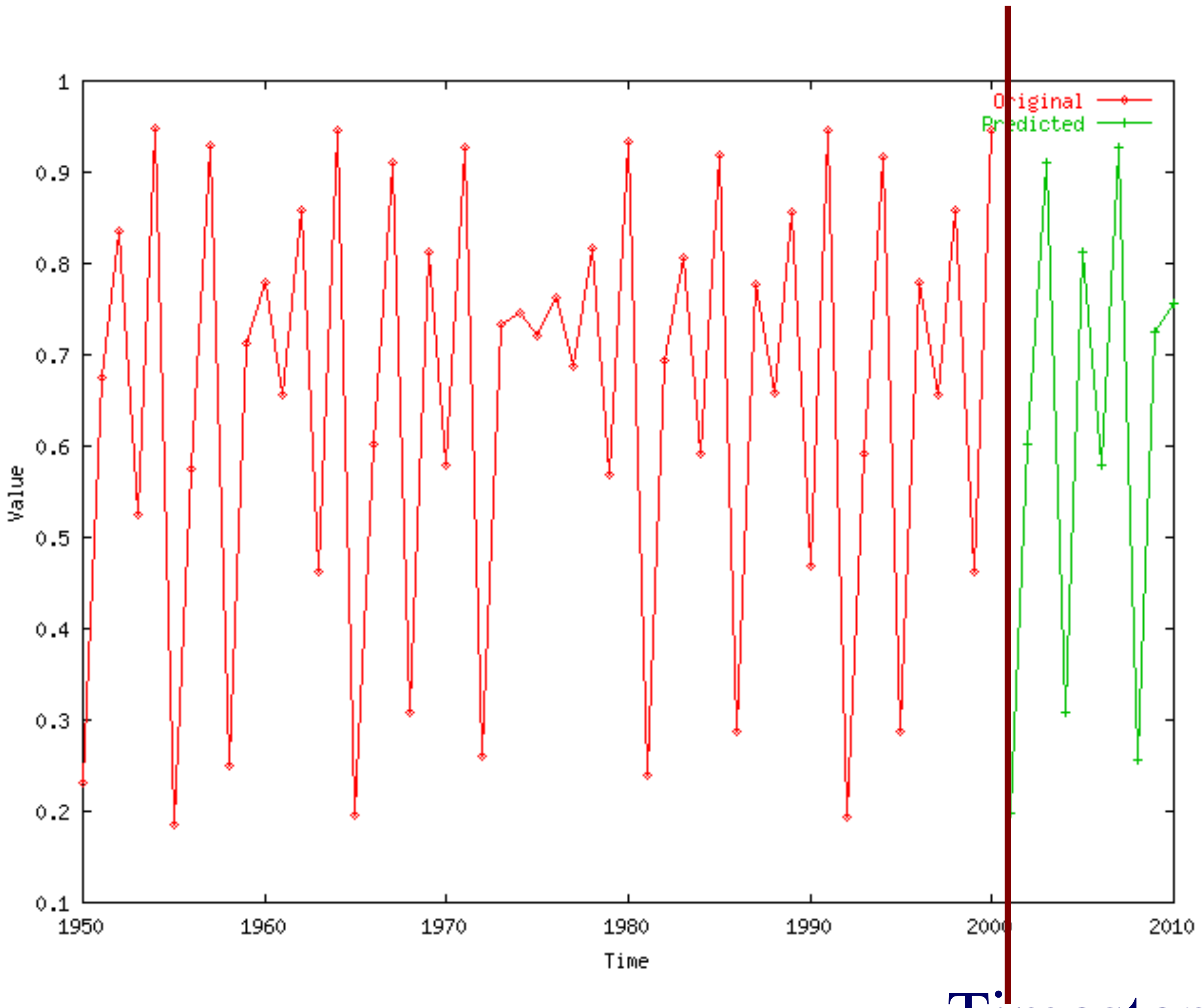
Lag-plot

ARIMA: fails

Logistic Parabola

Our Prediction from here

Value

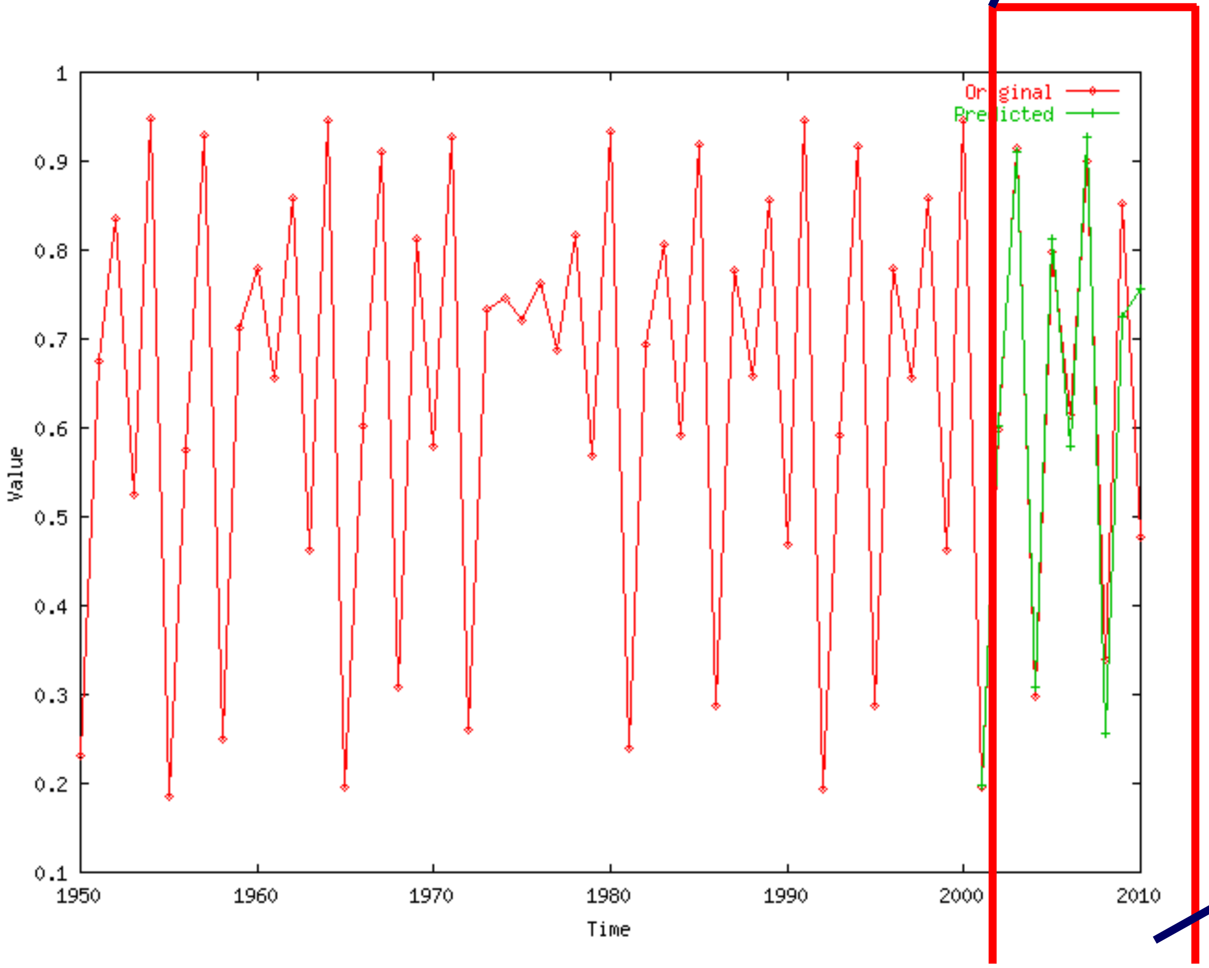


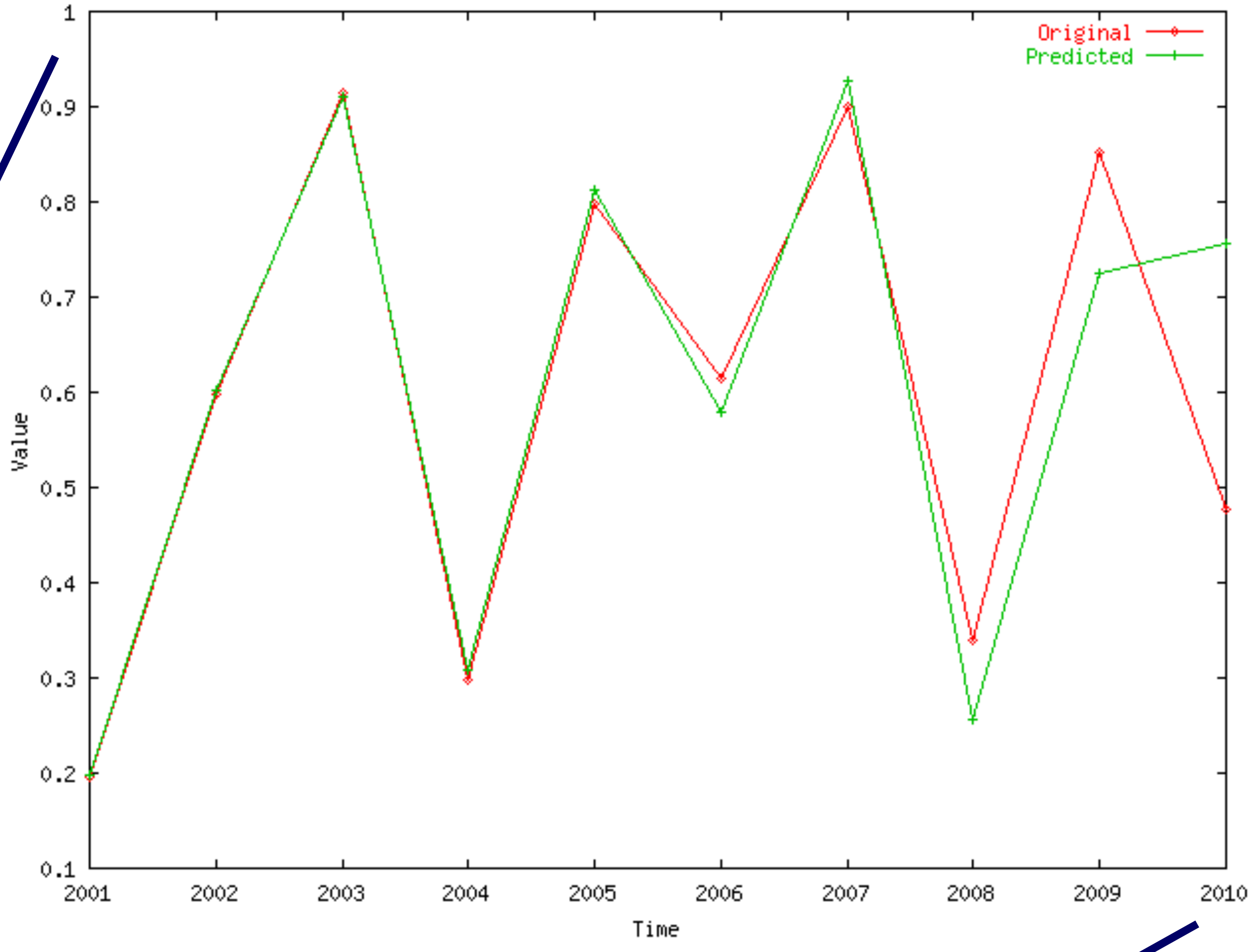
Timesteps

Value

Logistic Parabola

Comparison of prediction to correct values





Timesteps

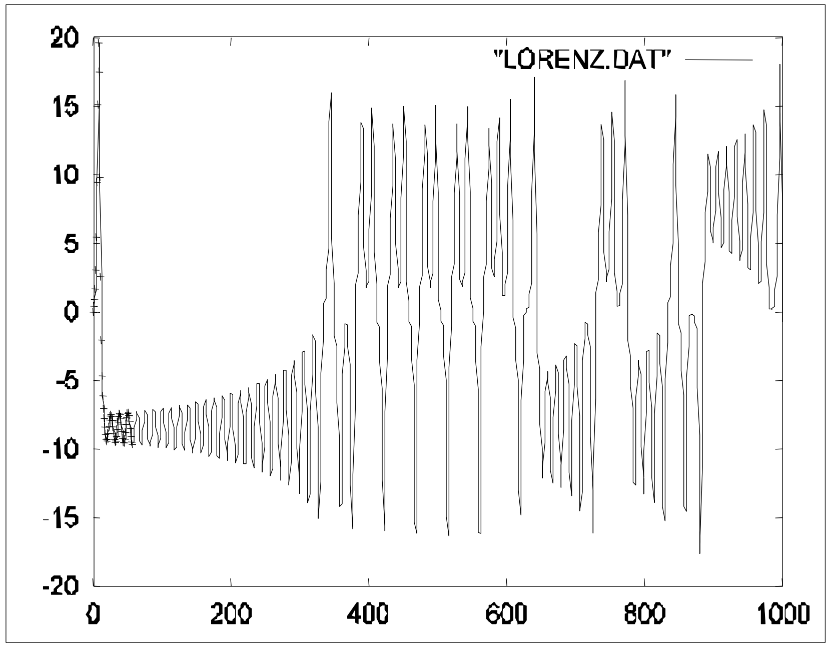
Datasets

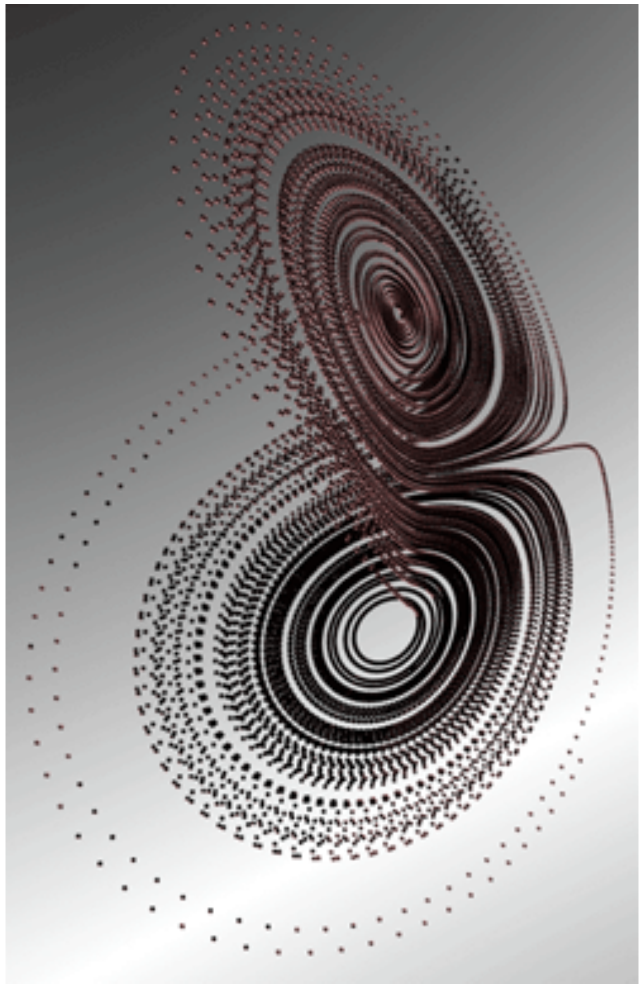
LORENZ: Models convection currents in the air

dx / dt = a (y - x)

dy / dt = x (b - z) - y

dz / dt = xy - c z

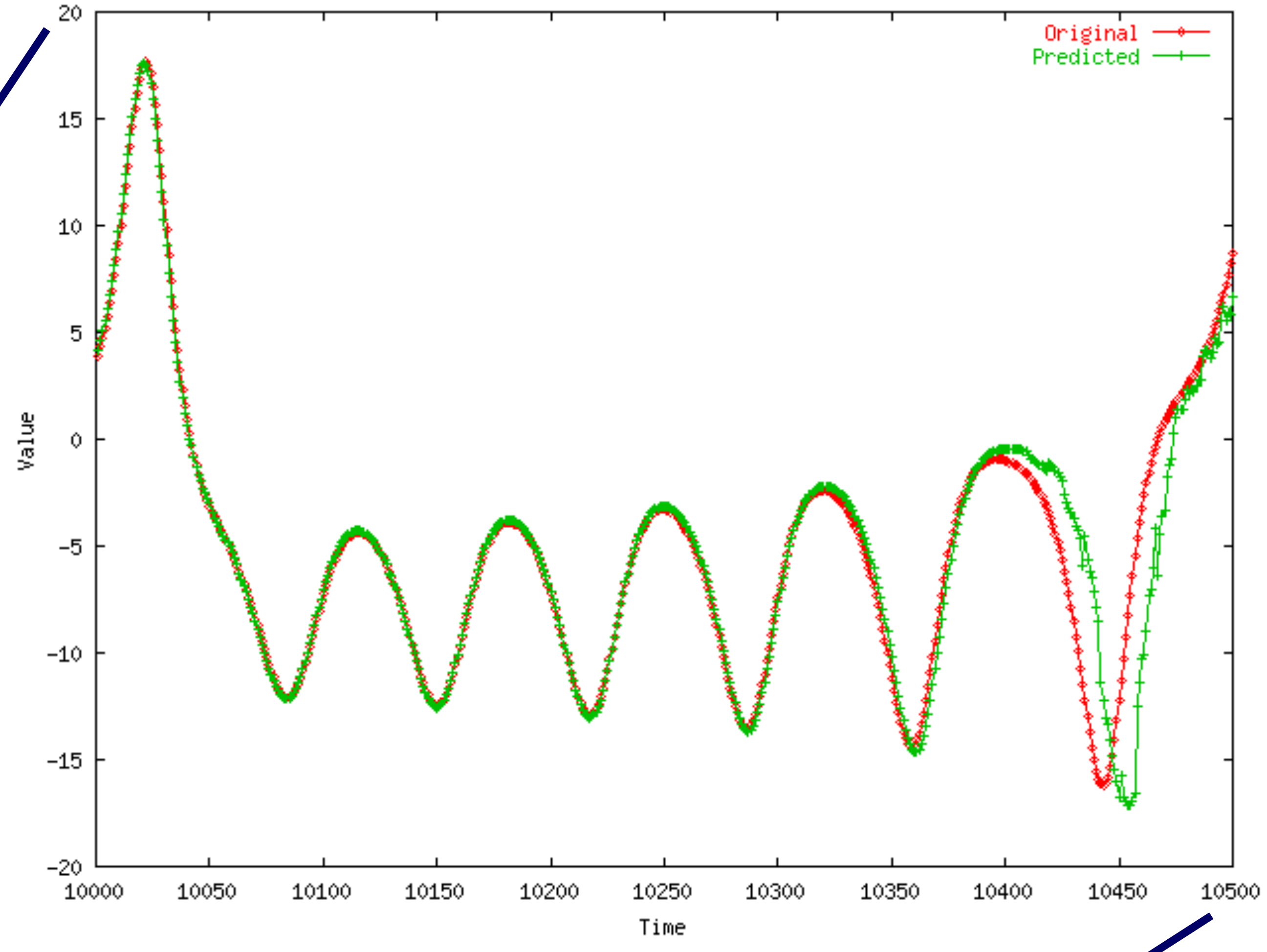




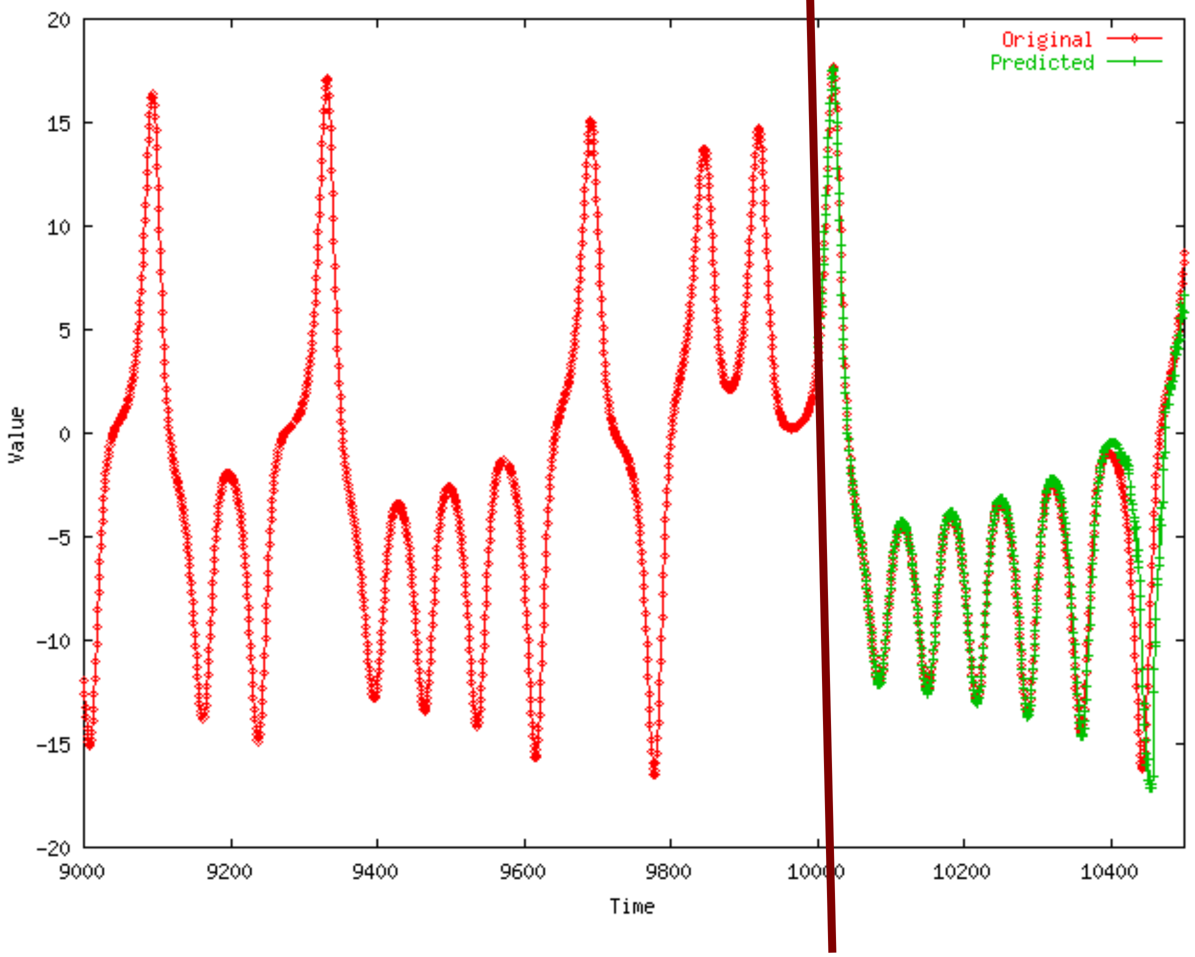
LORENZ

Comparison of prediction to correct values

Value

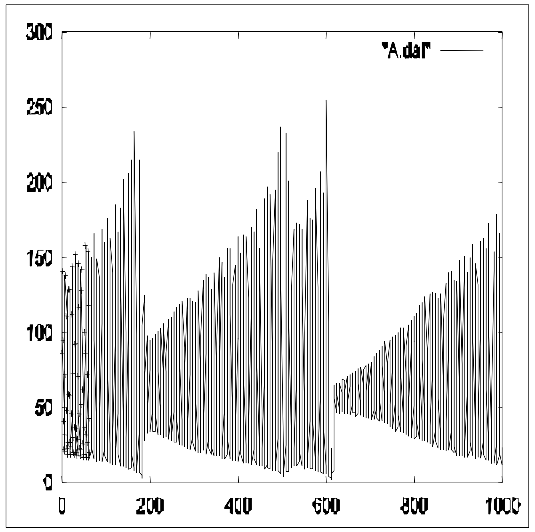


Timesteps



Datasets

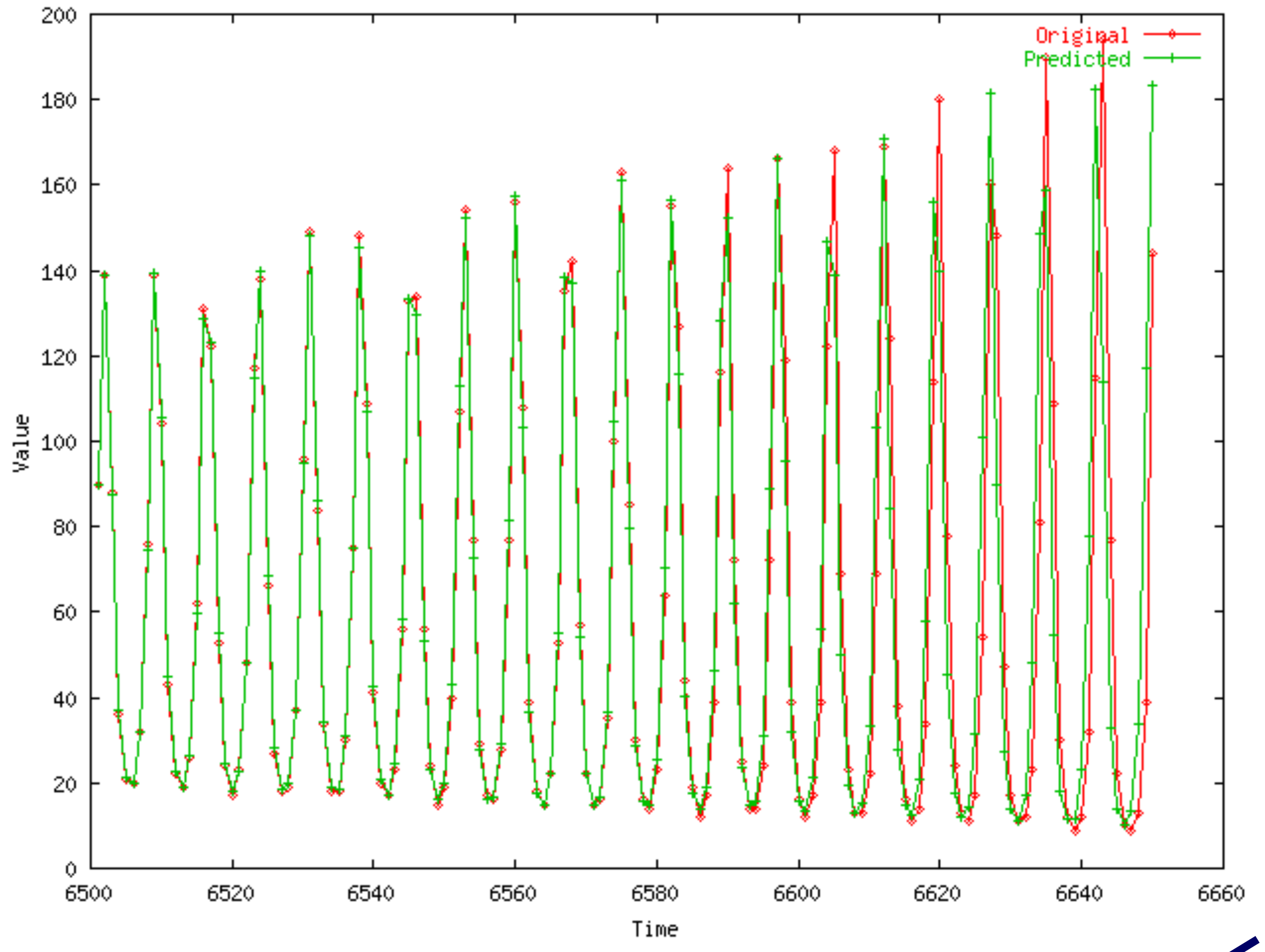
- LASER: fluctuations in a Laser over time (used in Santa Fe competition)



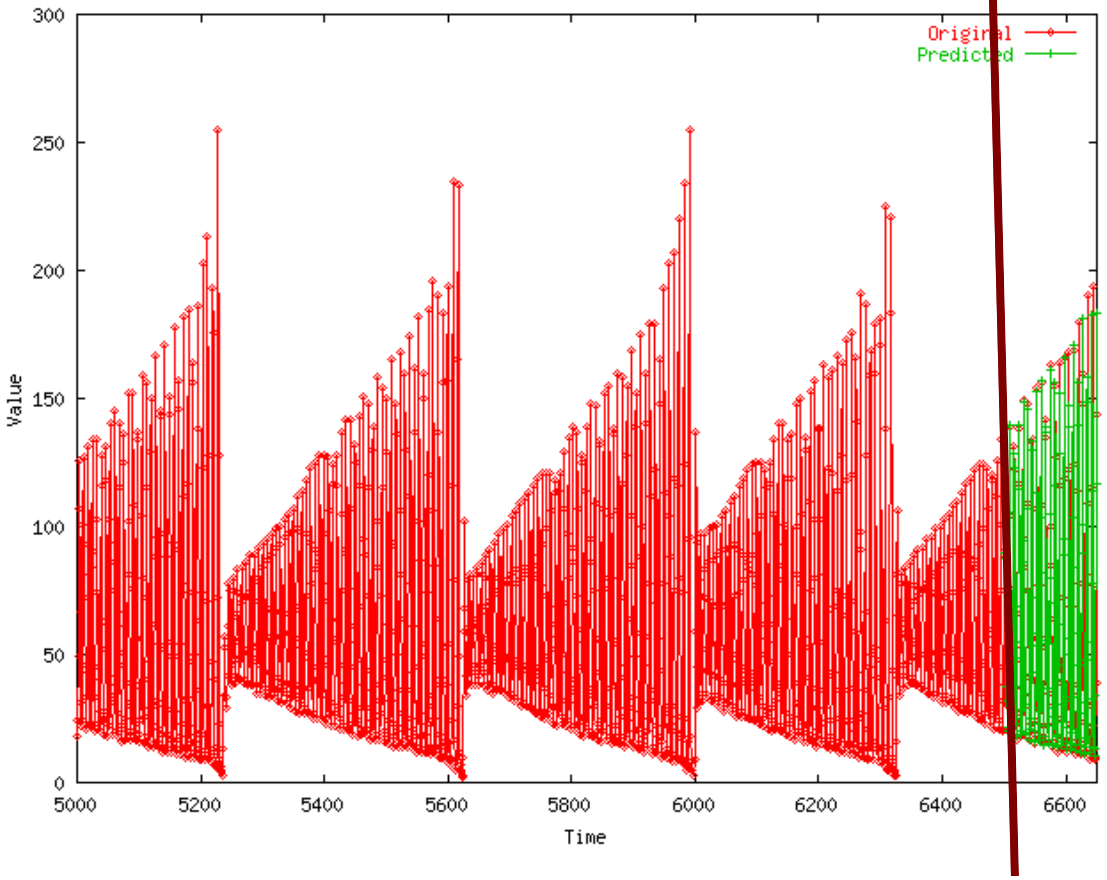
Laser

Comparison of prediction to correct values

Value



Timesteps



Conclusions

- Lag plots for non-linear forecasting (Takens' theorem)
- suitable for 'chaotic' signals

References

- Deepay Chakrabarti and Christos Faloutsos *F4: Large-Scale Automated Forecasting using Fractals* CIKM 2002, Washington DC, Nov. 2002.
- Sauer, T. (1994). *Time series prediction using delay coordinate embedding*. (in book by Weigend and Gershenfeld, below) Addison-Wesley.
- Takens, F. (1981). *Detecting strange attractors in fluid turbulence*. Dynamical Systems and Turbulence. Berlin: Springer-Verlag.

References

- Weigend, A. S. and N. A. Gerschenfeld (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison Wesley. (Excellent collection of papers on chaotic/non-linear forecasting, describing the algorithms behind the winners of the Santa Fe competition.)

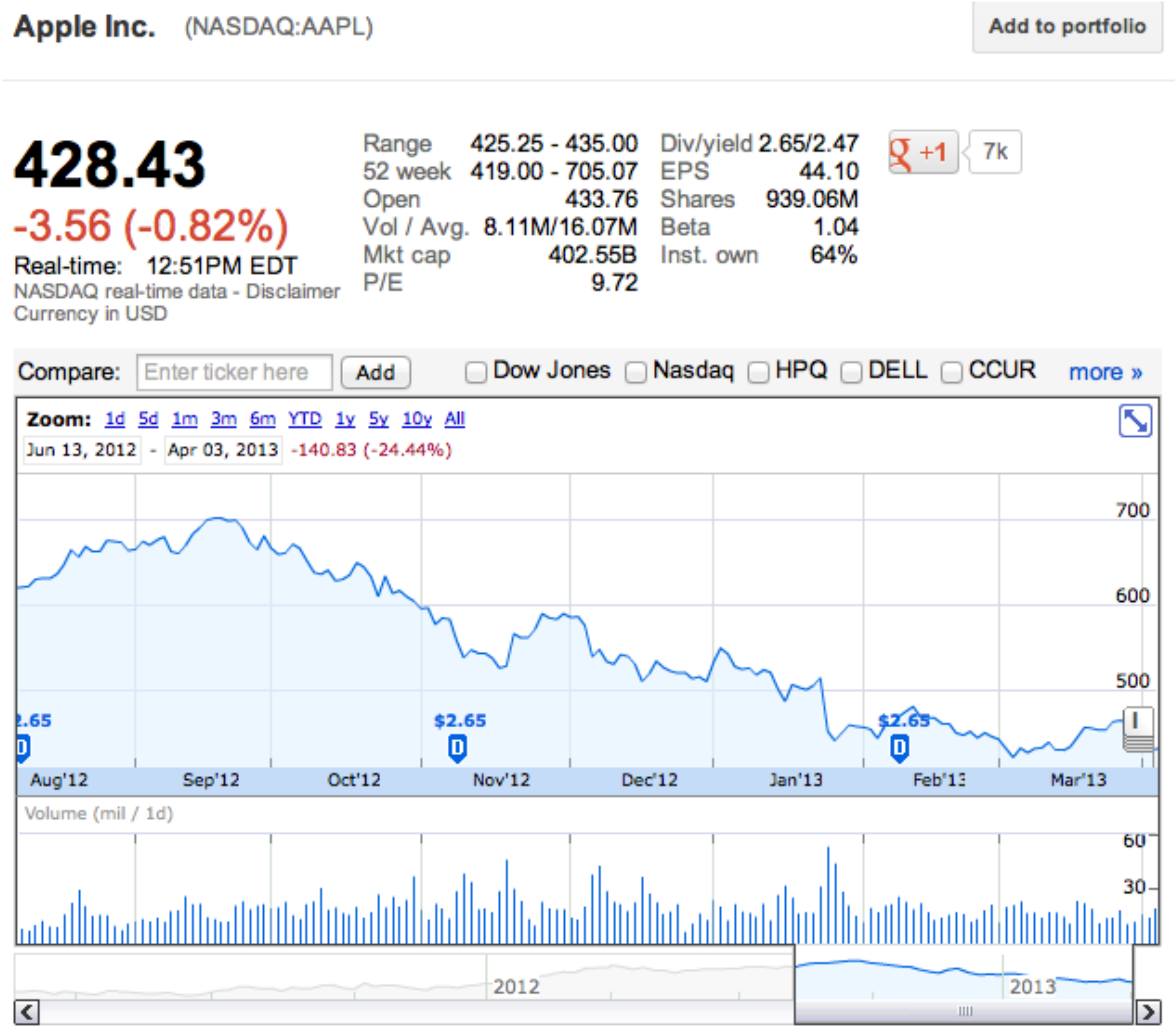
Overall conclusions

- Similarity search: **Euclidean**/time-warping; **feature extraction** and **SAMs**
- Linear Forecasting: **AR** (Box-Jenkins) methodology;
- Non-linear forecasting: **lag-plots** (Takens)

Must-Read Material

- Byong-Kee Yi, Nikolaos D. Sidiropoulos, Theodore Johnson, H.V. Jagadish, Christos Faloutsos and Alex Biliris, *Online Data Mining for Co-Evolving Time Sequences*, ICDE, Feb 2000.
- Chungmin Melvin Chen and Nick Roussopoulos, *Adaptive Selectivity Estimation Using Query Feedbacks*, SIGMOD 1994

Time Series Visualization + Applications



Apple Inc. (NASDAQ:AAPL)

Add to portfolio

428.43

-3.56 (-0.82%)

Real-time: 12:51PM EDT

NASDAQ real-time data - Disclaimer

Currency in USD

Range	425.25 - 435.00	Div/yield	2.65/2.47
52 week	419.00 - 705.07	EPS	44.10
Open	433.76	Shares	939.06M
Vol / Avg.	8.11M/16.07M	Beta	1.04
Mkt cap	402.55B	Inst. own	64%
P/E	9.72		

+1 7k

Compare: Enter ticker here Add □ Dow Jones □ Nasdaq □ HPQ □ DELL □ CCUR more »

Zoom: 1d 5d 1m 3m 6m YTD 1y 5y 10y All

Jun 13, 2012 - Apr 03, 2013 -140.83 (-24.44%)

Why Time Series Visualization?

Time series is the **most common data type**

- But why is **time series** so common?

How to build time series visualization?

Easy way: use existing tools, libraries

- **Google Public Data Explorer** (Gapminder)
  http://goo.gl/HmrH
- **Google acquired Gapminder**

  http://goo.gl/43avY

  (Hans Rosling's **TED talk** http://goo.gl/tKV7)
- **Google Annotated Time Line**
  http://goo.gl/Upm5W
- **Timeline**, from MIT's SIMILE project
  http://simile-widgets.org/timeline/
- **Timeplot**, also from SIMILE
  http://simile-widgets.org/timeplot/
- **Excel,** of course

How to build time series visualization?

The harder way:

- R (ggplot2)
- Matlab
- gnuplot
- ...

The even harder way:

- D3, for web
- JFreeChart (Java)
- ...

Time Series Visualization

Why is it useful?

When is visualization useful?

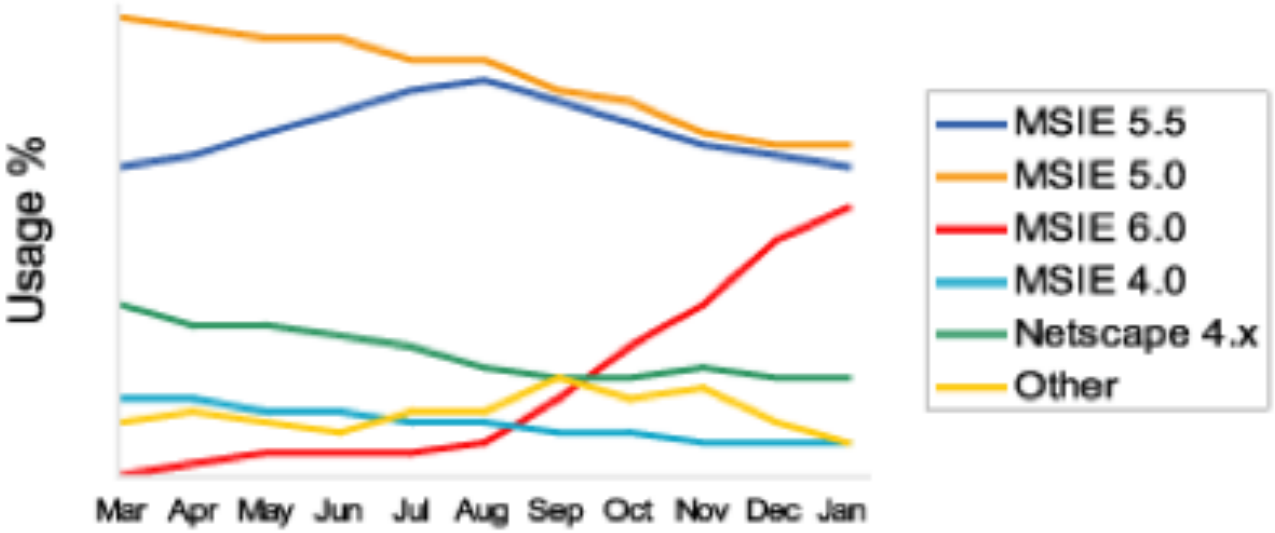
(Why not automate everything? Like using the forecasting techniques you learned last time.)

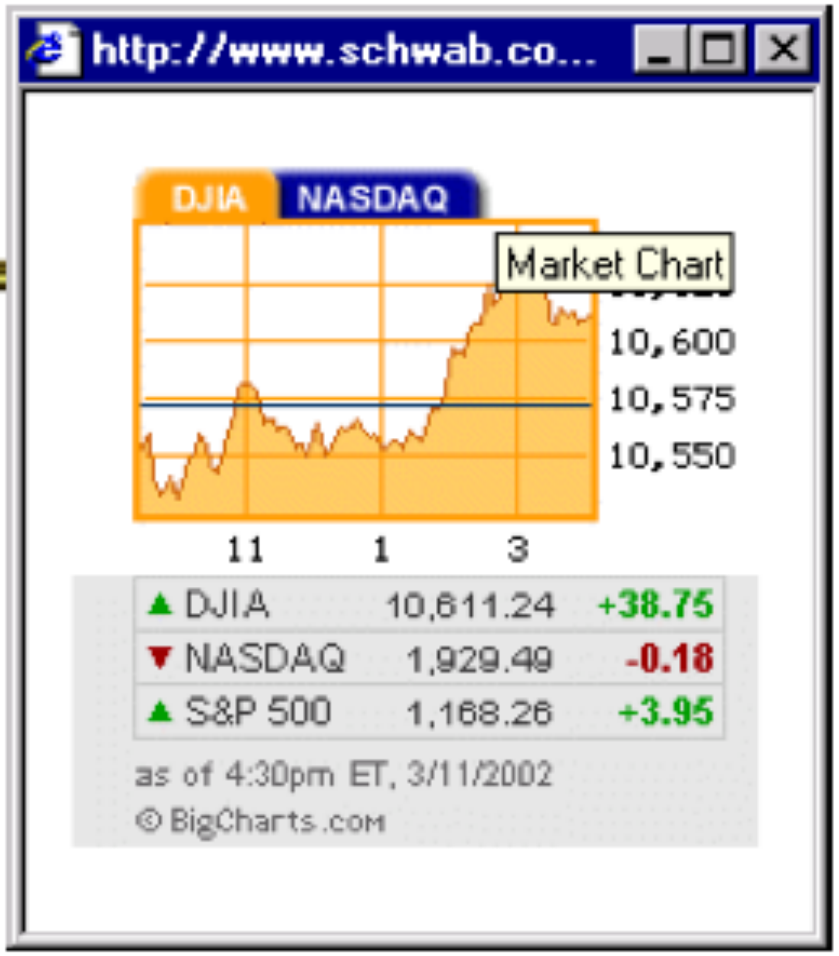
Time Series User Tasks

- When was something greatest/least?
- Is there a pattern?
- Are two series similar?
- Do any of the series match a pattern?
- Provide simpler, faster access to the series
- Does data element exist at time t ?
- When does a data element exist?
- How long does a data element exist?
- How often does a data element occur?
- How fast are data elements changing?
- In what order do data elements appear?
- Do data elements exist together?

Muller & Schumann 03
citing MacEachern 95

Classic Views

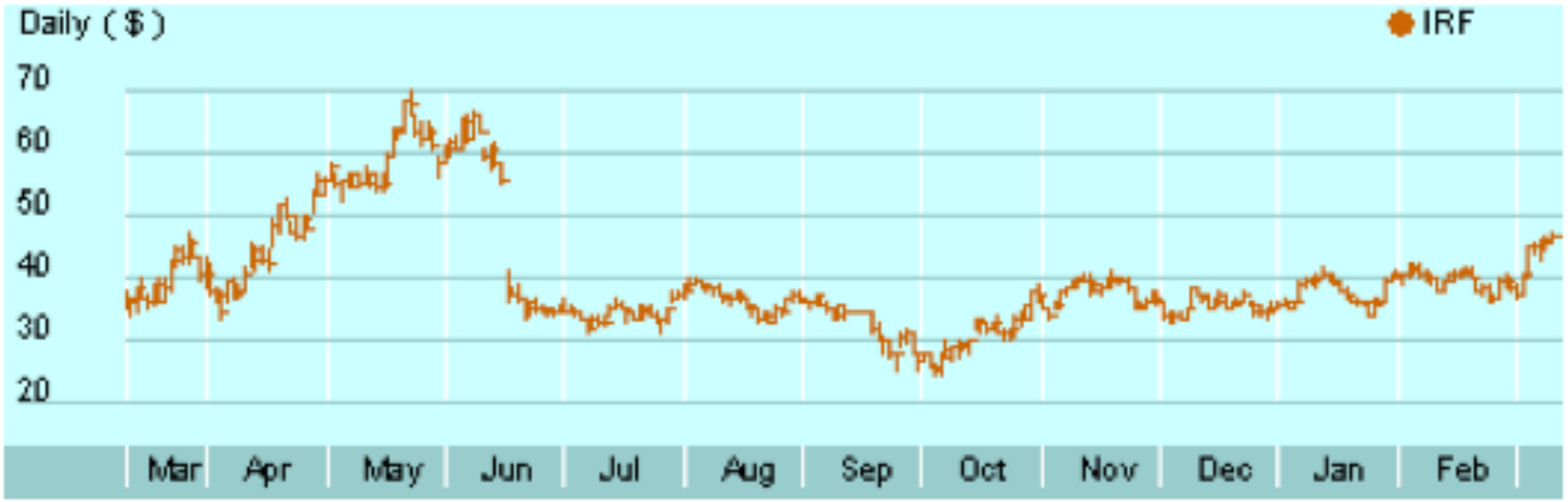




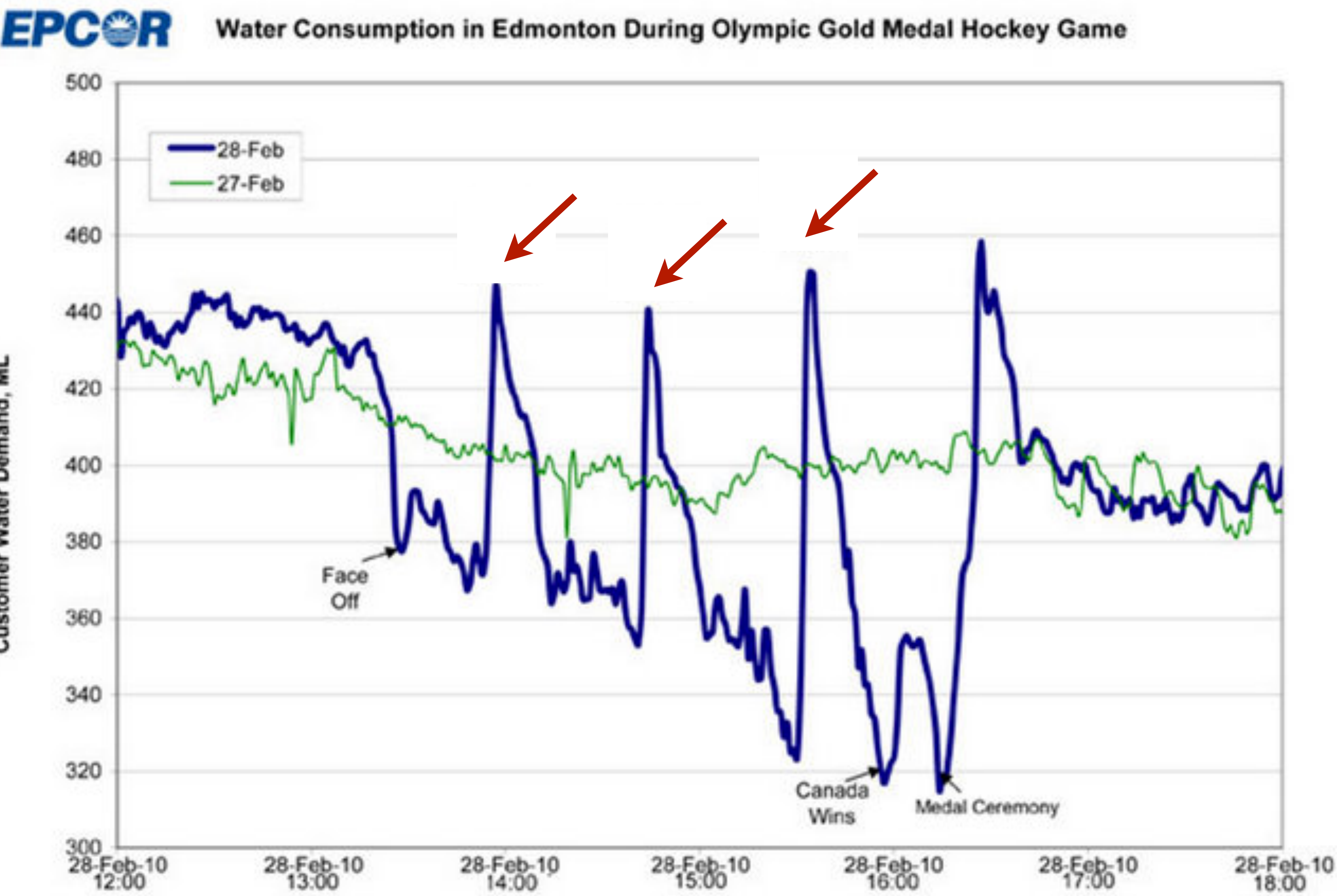
▲ DJIA	10,611.24	+38.75
▼ NASDAQ	1,929.49	-0.18
▲ S&P 500	1,168.26	+3.95

as of 4:30pm ET, 3/11/2002

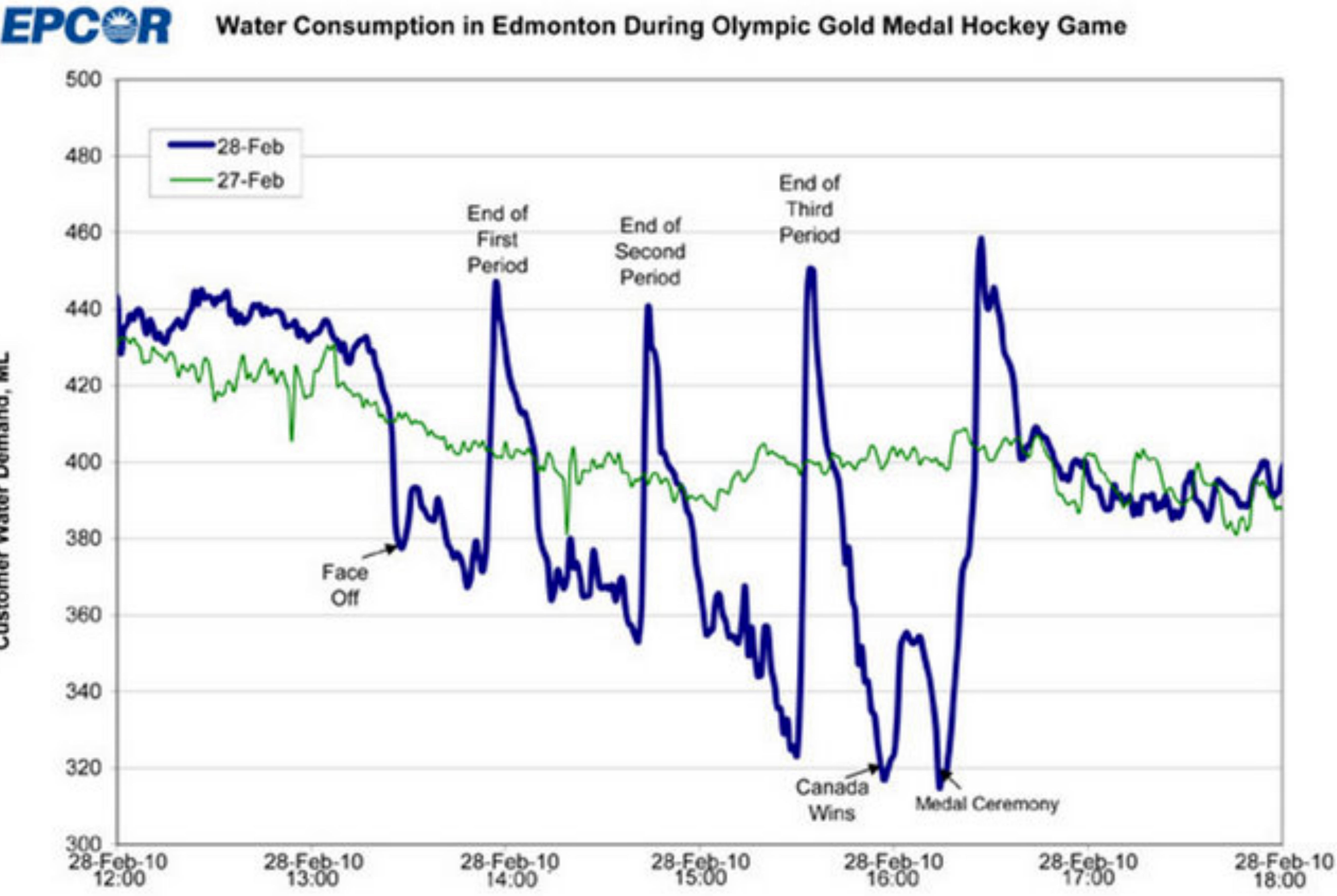
© BigCharts.com



horizontal axis is time



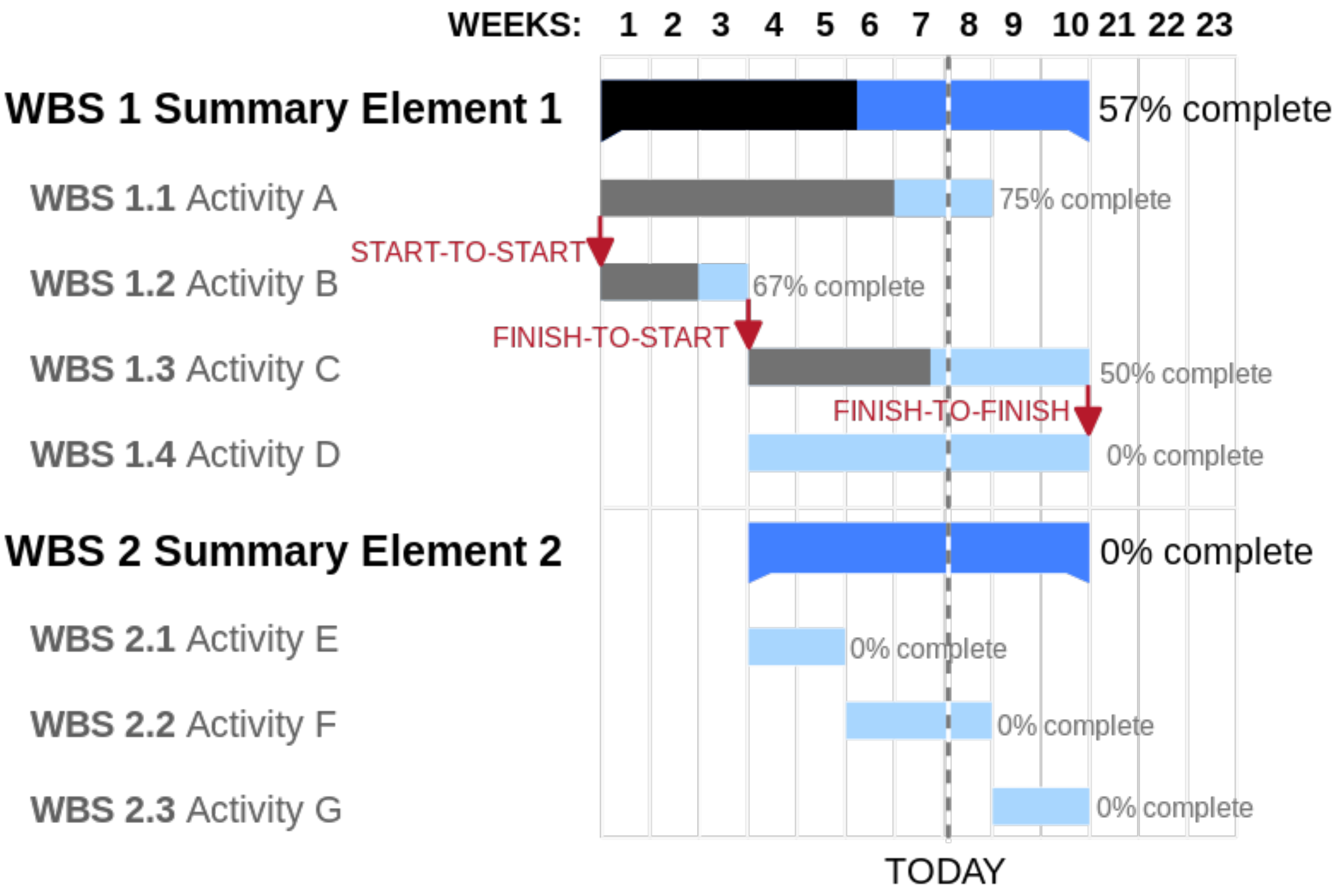
http://www.patspapers.com/blog/item/what_if_everybody_flushed_at_once_Edmonton_water_gold_medal_hockey_game/



http://www.patspapers.com/blog/item/what_if_everybody_flushed_at_once_Edmonton_water_gold_medal_hockey_game/

Gantt Chart

Useful for project



How to create in Excel:

http://www.youtube.com/watch?v=sA67g6zaKOE

ThemeRiver
Stacked graph
Streamgraph

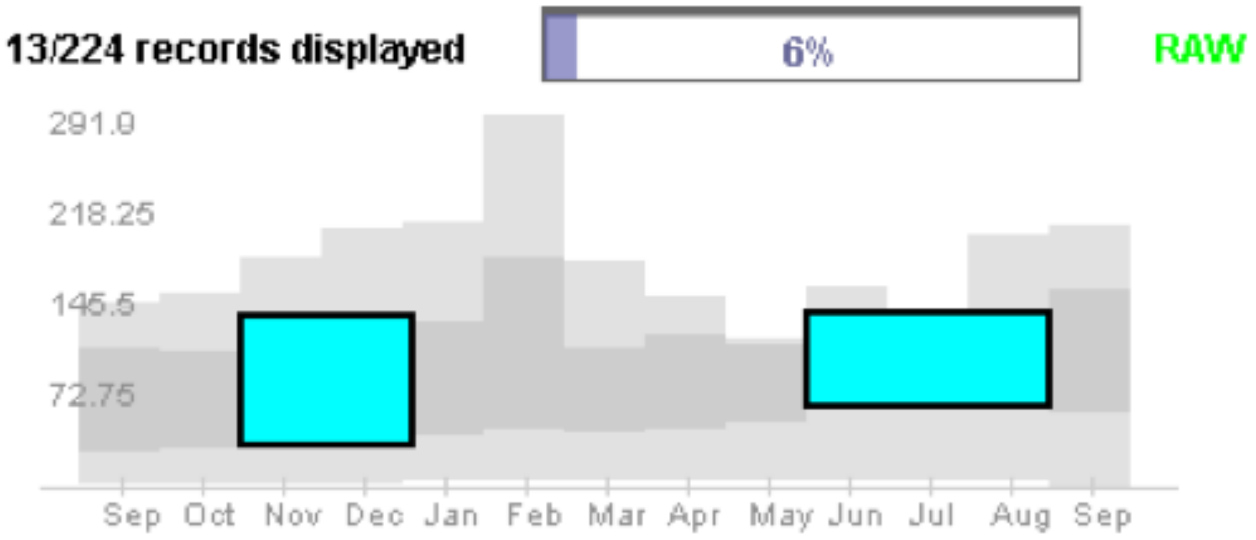
http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html

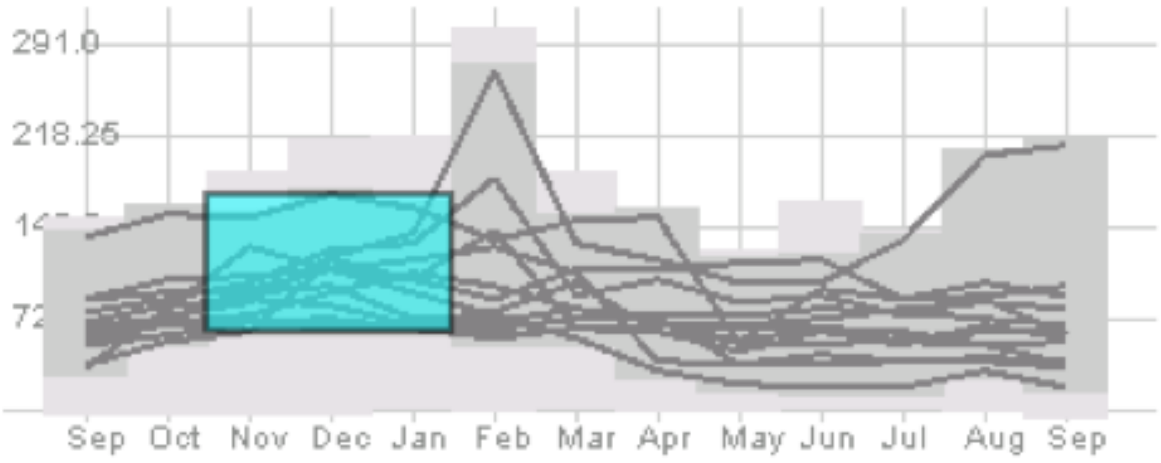
http://bl.ocks.org/mbostock/3943967

TimeSearcher

support queries

Can create rectangles that function as matching regions





Light gray is all data's extent

Darker grayed region is data envelope that shows extreme values of queries matching criteria

Multiple boxes are "anded"

Hochheiser & Shneiderman
Proc. Discovery Science '01

http://hcil2.cs.umd.edu/video/2005/2005_timesearcher2.mpg

GeoTime

Infovis 2004

http://www.youtube.com/watch?v=inkF86QJBdA

http://vadl.cc.gatech.edu/documents/55_Wright_KaplerWright_GeoTime_InfoViz_Jrnl_05_send.pdf