http://poloclub.gatech.edu/cse6242

CSE6242 / CX4242: Data & Visual Analytics

Time Series

Non-linear Forecasting

Duen Horng (Polo) Chau
Professor
Associate Director, MS Analytics
Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos, Parishit Ram (GT PhD alum; SkyTree), Alex Gray

Chaos & non-linear forecasting

Reference:

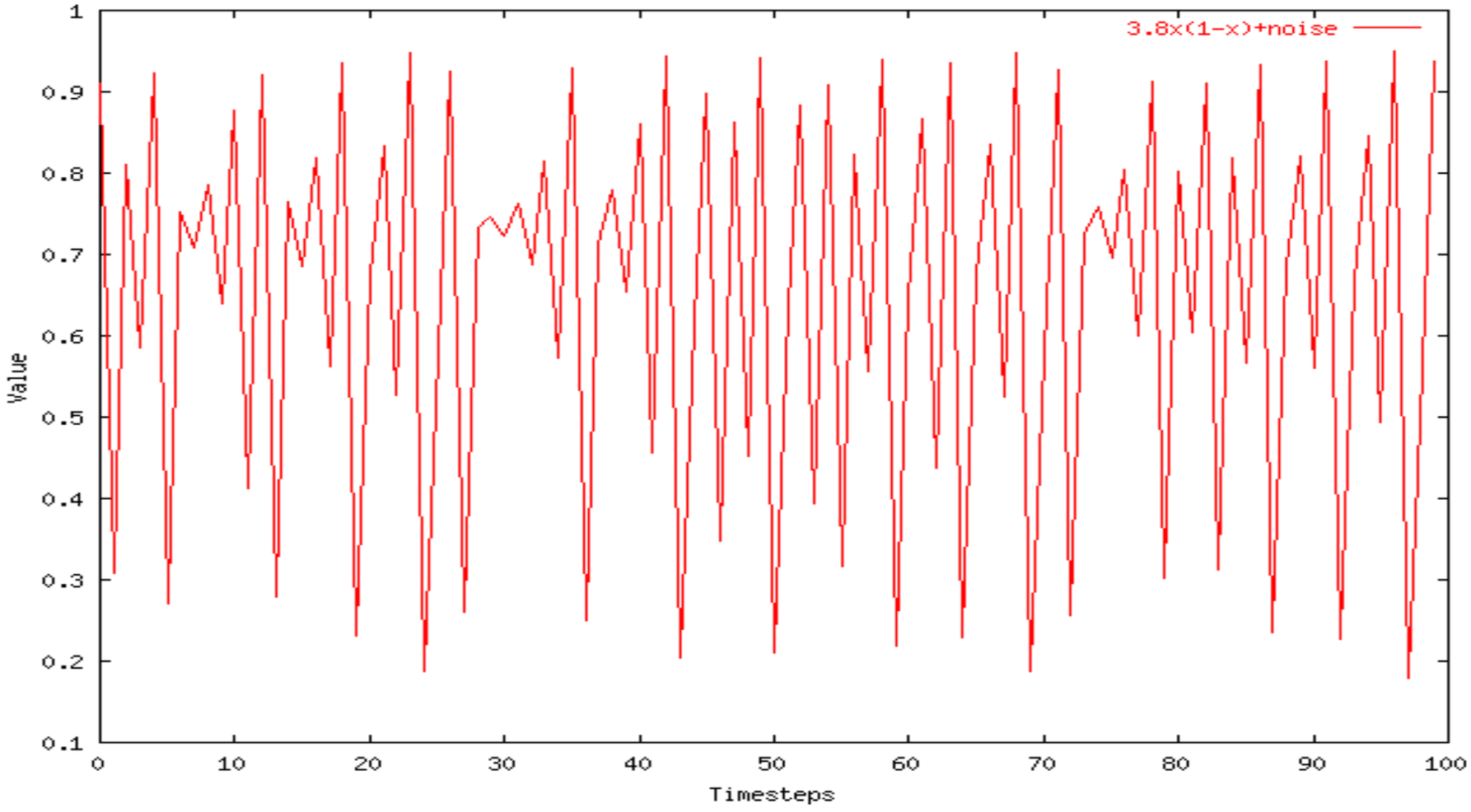
[Deepay Chakrabarti and Christos Faloutsos *F4: Large-Scale Automated Forecasting using Fractals* CIKM 2002, Washington DC, Nov. 2002.]

Detailed Outline

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

Recall: Problem #1

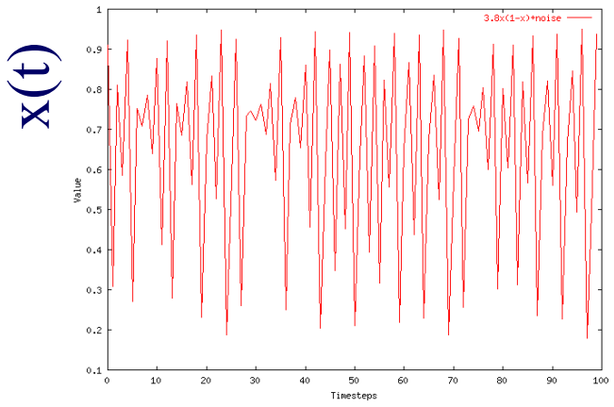
Value



Time

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

Datasets

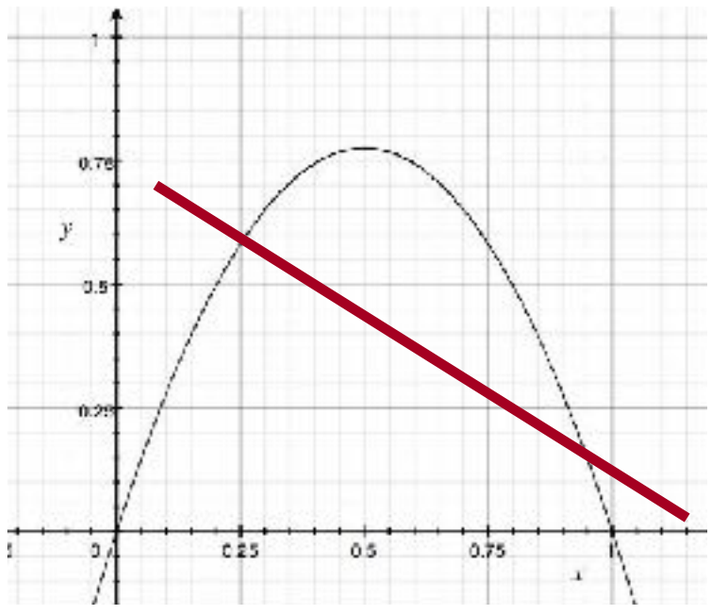


time

Logistic Parabola:

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

Models population of flies [R. May/1976]

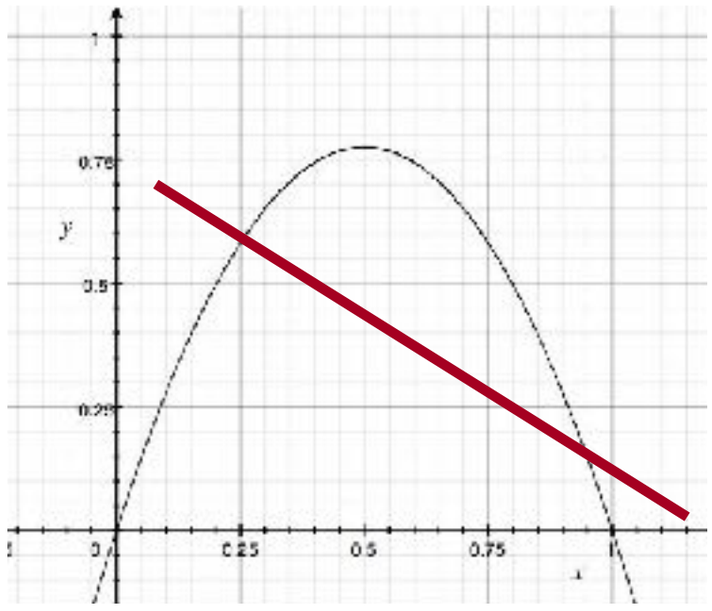


Lag-plot

ARIMA: fails

How to forecast?

- ARIMA - but: linearity assumption



Lag-plot

ARIMA: fails

How to forecast?

- ARIMA - but: linearity assumption

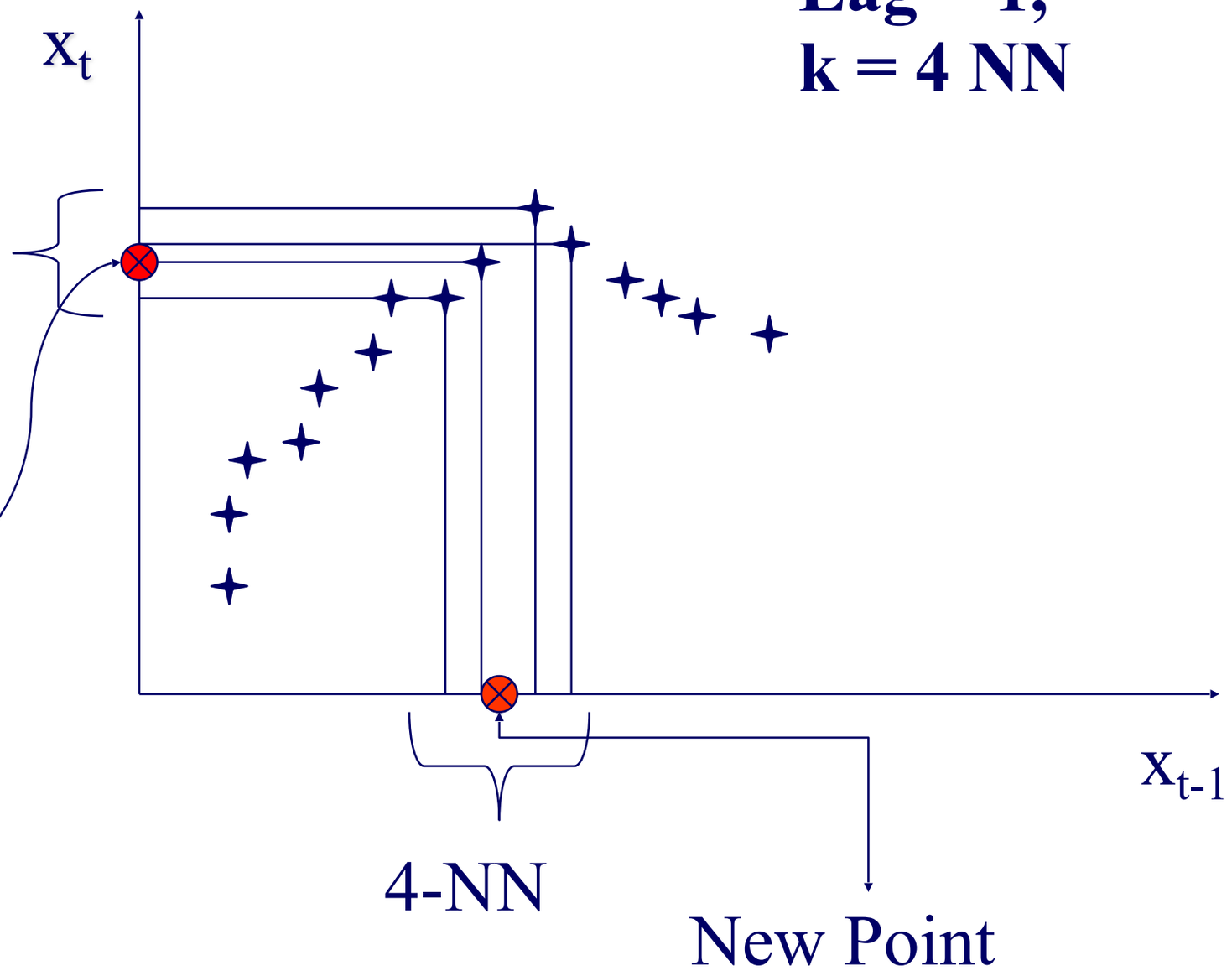
- ANSWER: ‘Delayed Coordinate Embedding’
  = Lag Plots [Sauer92]
  ~ nearest-neighbor search, for past incidents

General Intuition (Lag Plot)

Lag = 1,
k = 4 NN

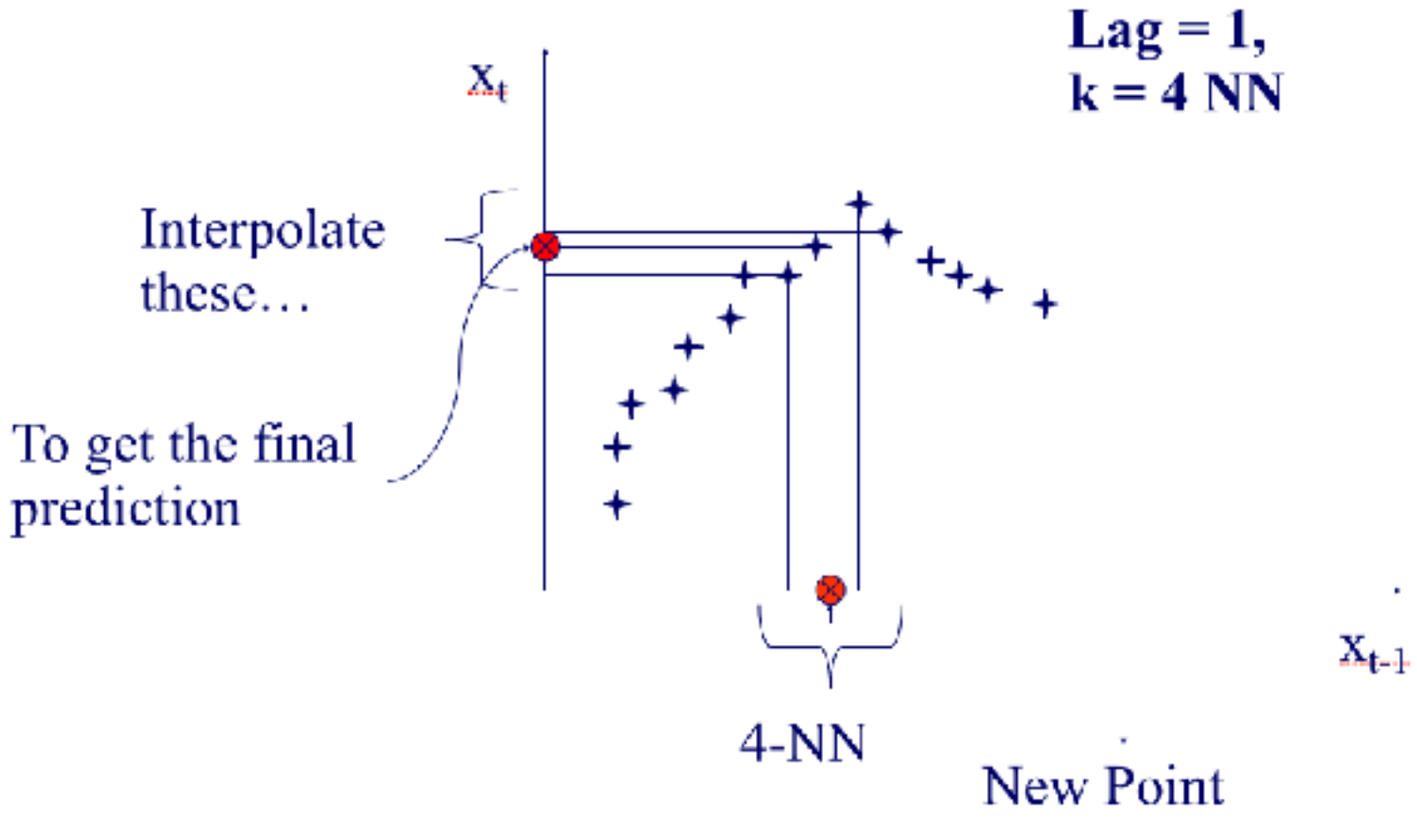
Interpolate these…

To get the final prediction



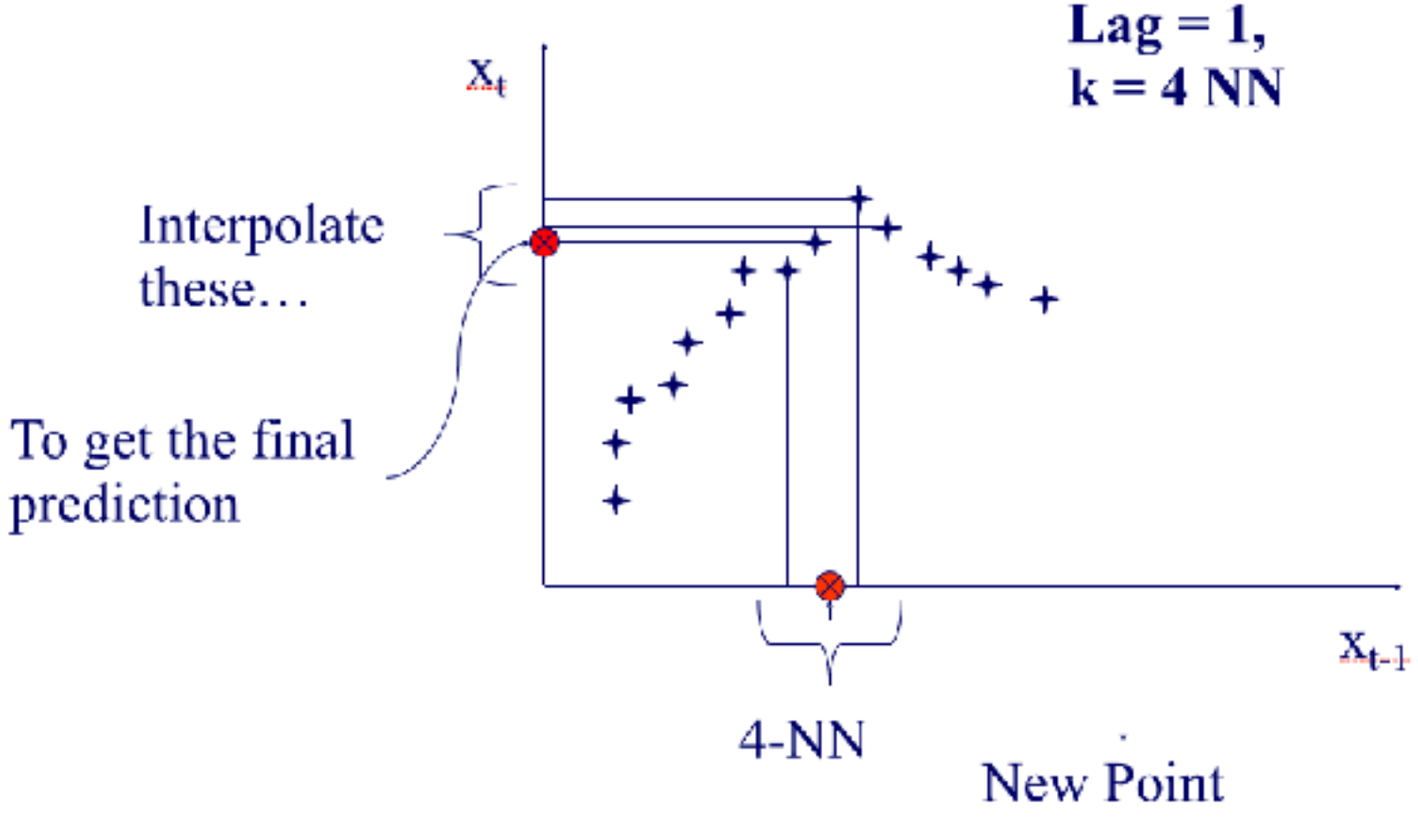
Questions:

- 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?



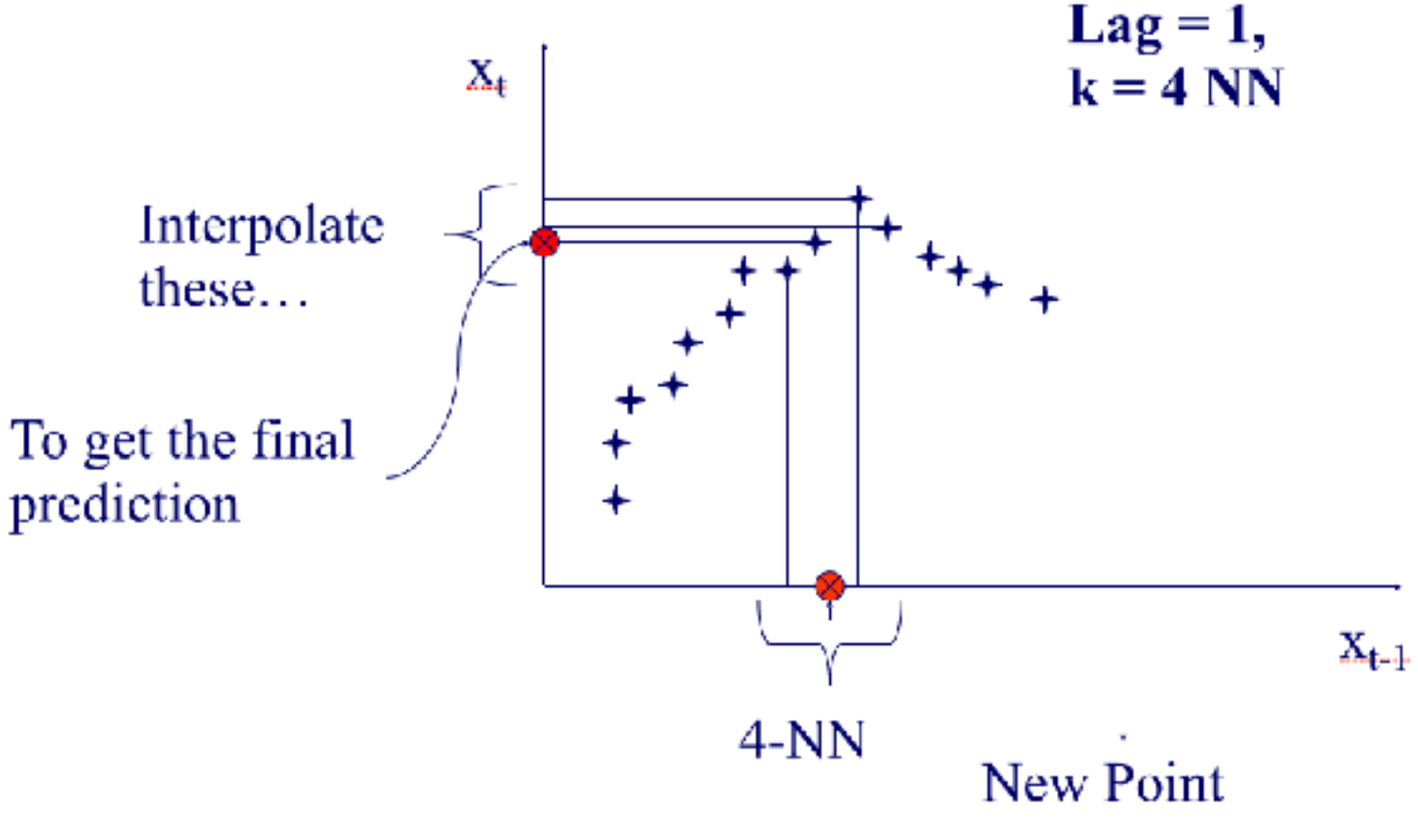
Q1: Choosing lag *L*

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



Q2: Choosing number of neighbors *k*

- Manually (typically ~ 1-10)

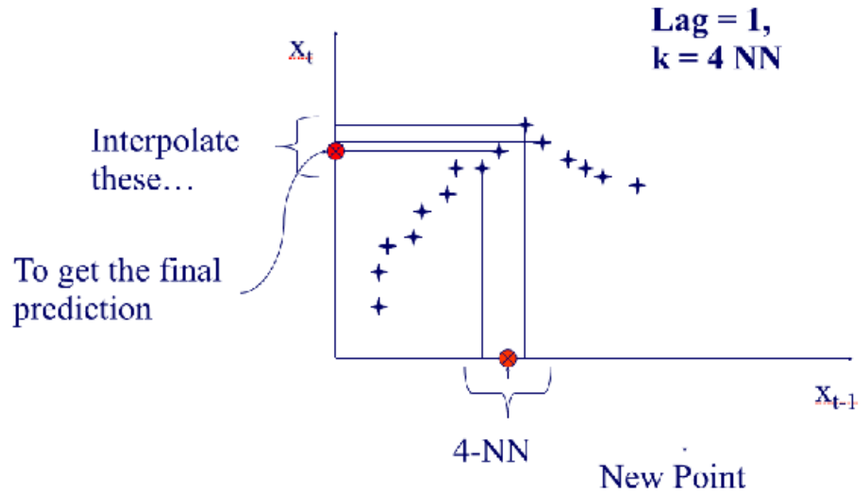


Q3: How to interpolate?

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

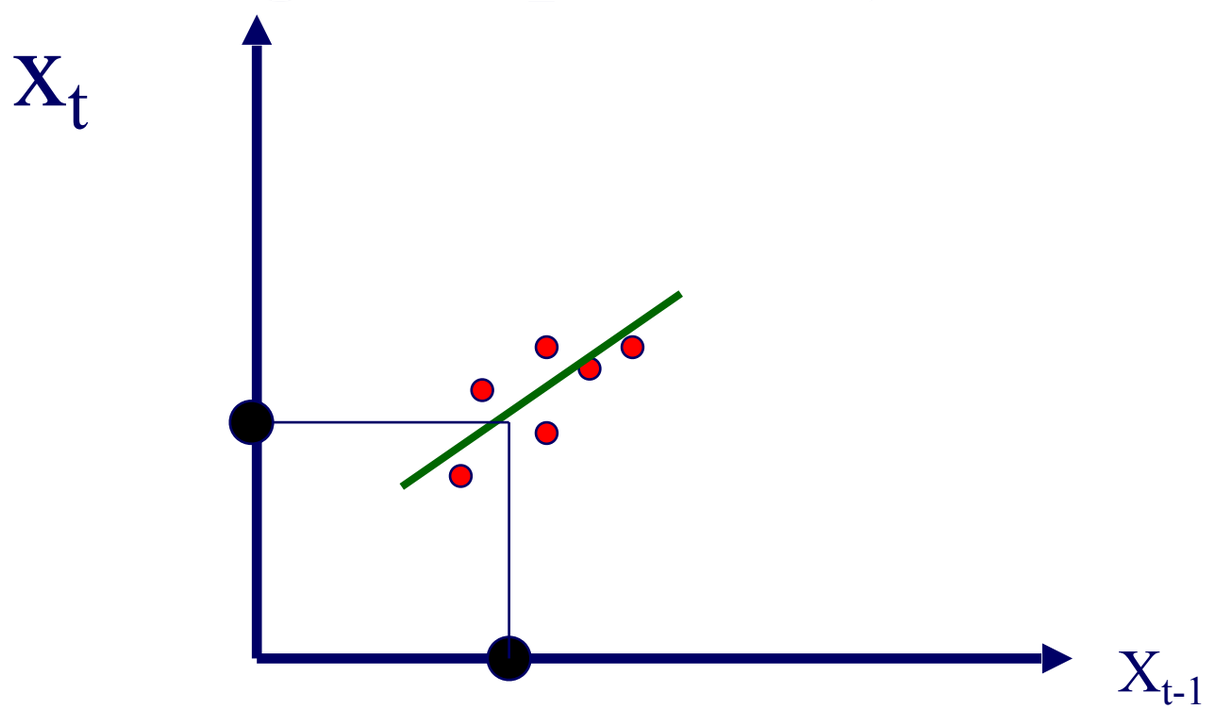
A3.1: Average

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)



Q4: Any theory behind it?

A4: YES!

Theoretical foundation

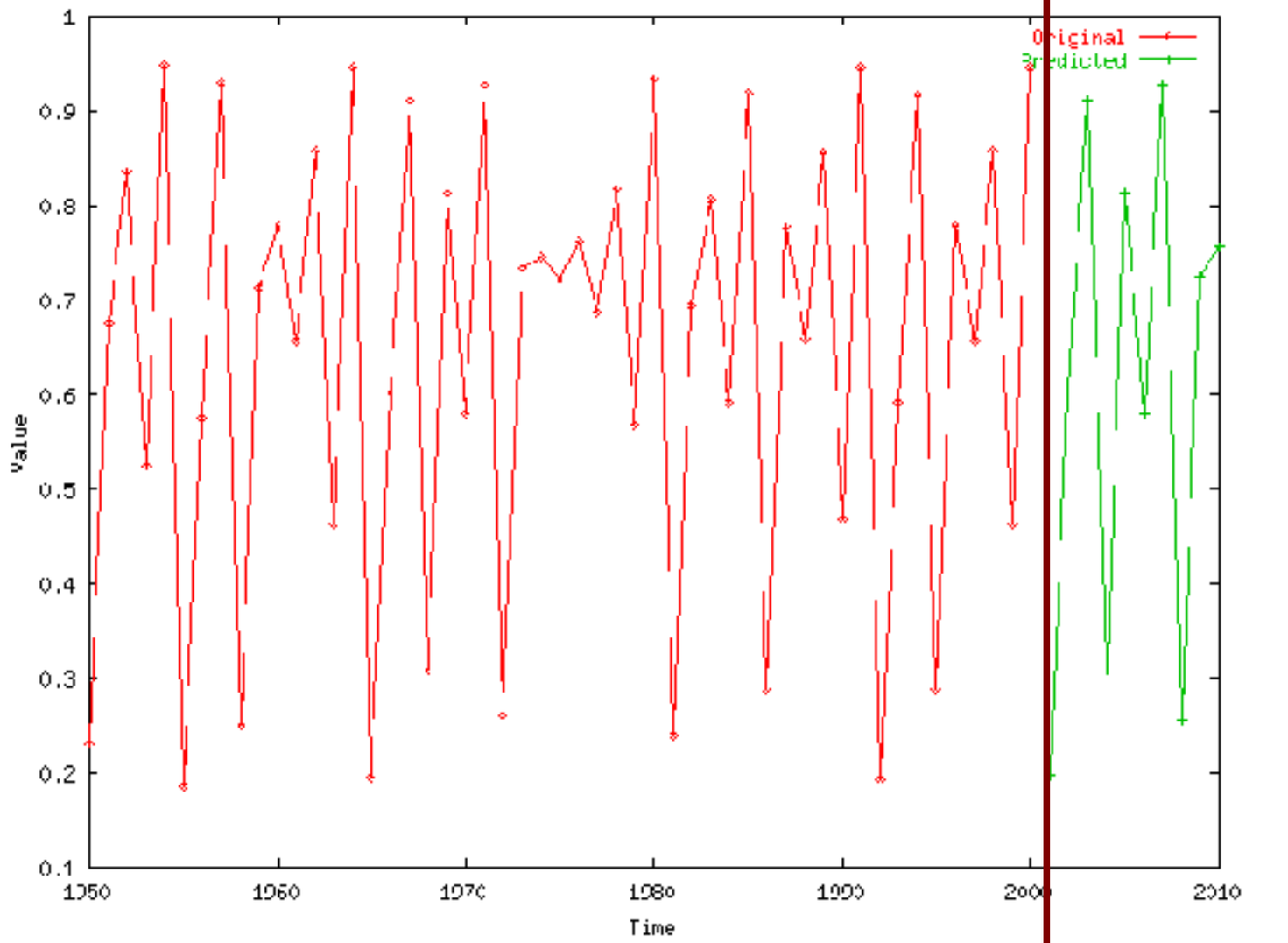
- Based on the ‘Takens theorem’ [Takens81]
- which says that long enough delay vectors can do prediction, even if there are unobserved variables in the dynamical system (= diff. equations)

Detailed Outline

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

Logistic Parabola

Our Prediction from here

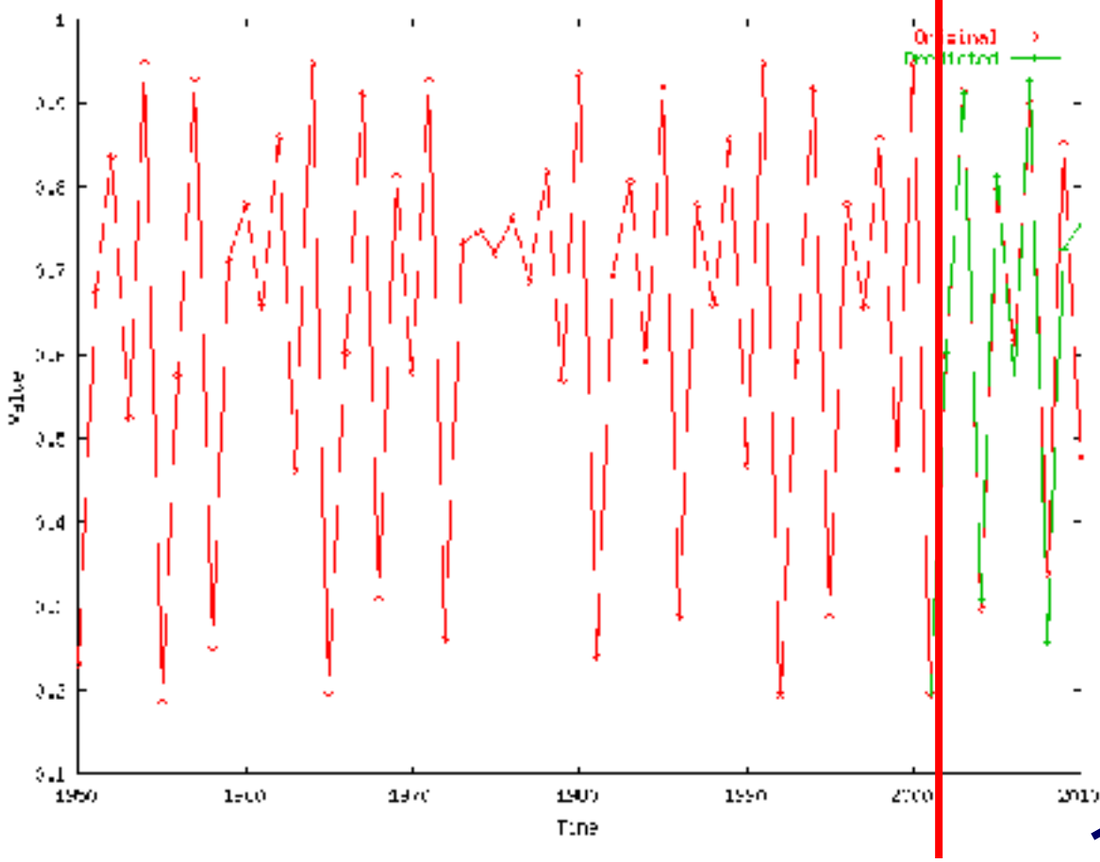


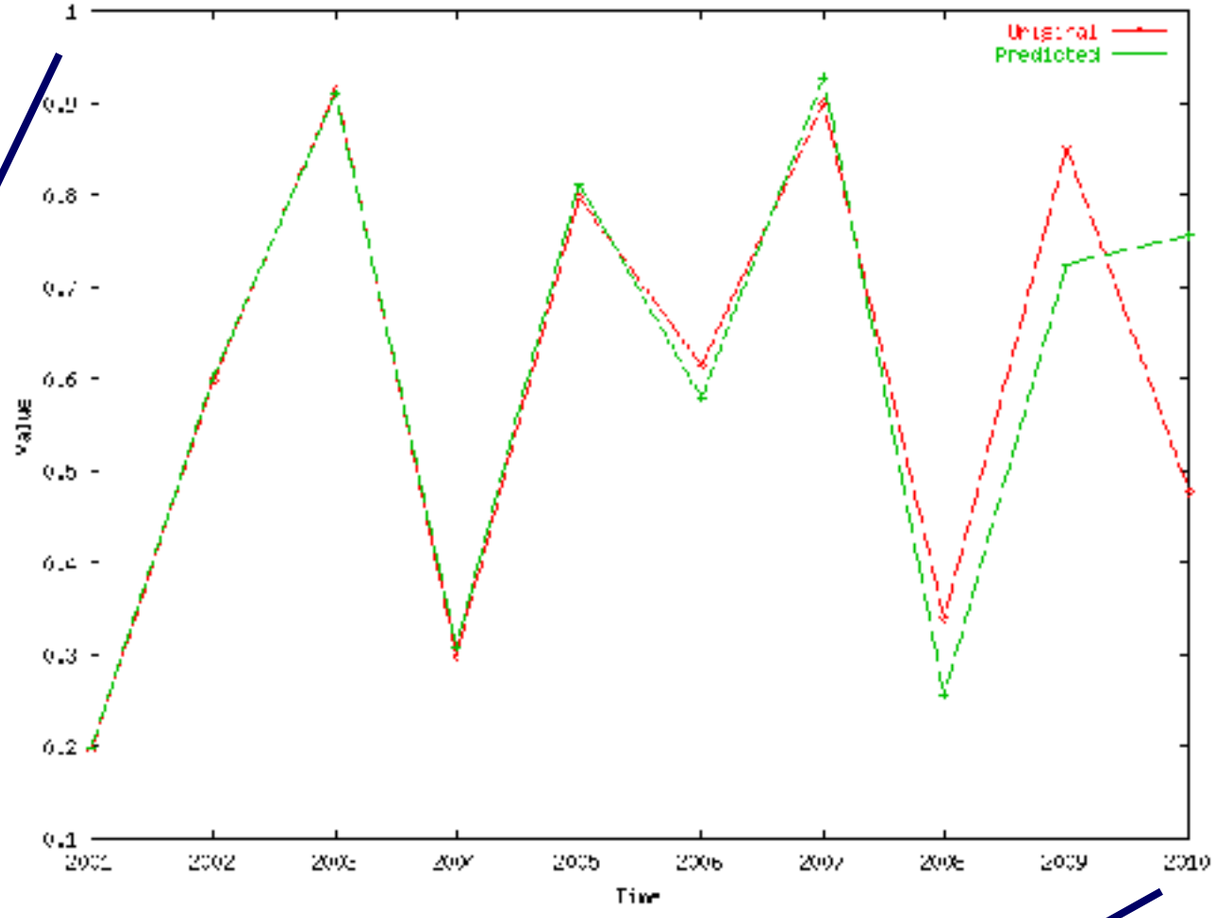
Value

Timesteps

Logistic Parabola

Comparison of prediction to correct values





Value

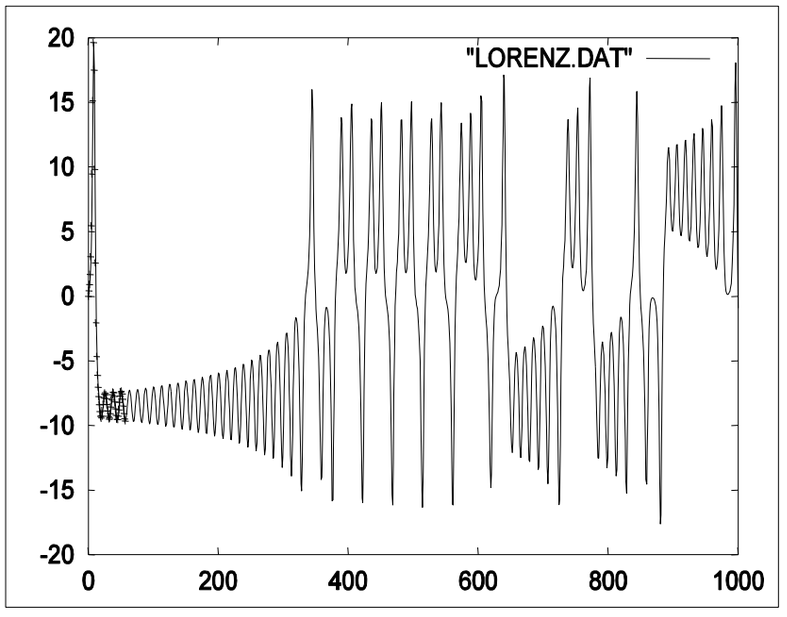
Timesteps

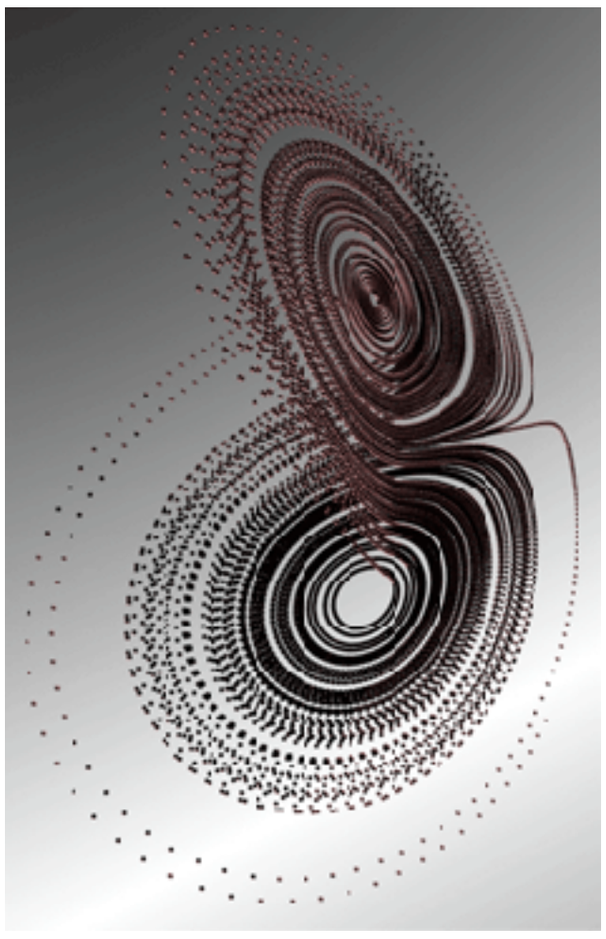
Value

Datasets

LORENZ: Models convection currents in the air

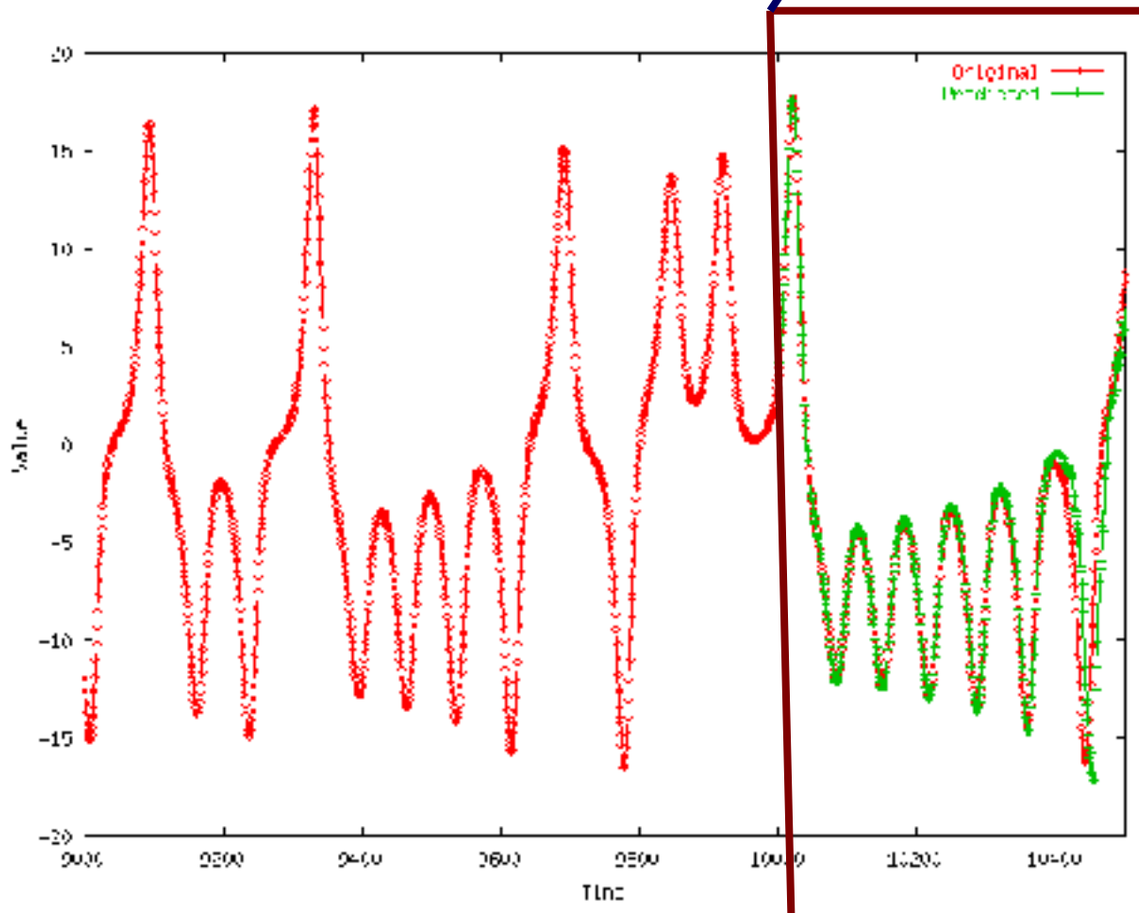
$$dx / dt = a (y - x)$$
$$dy / dt = x (b - z) - y$$
$$dz / dt = xy - c z$$

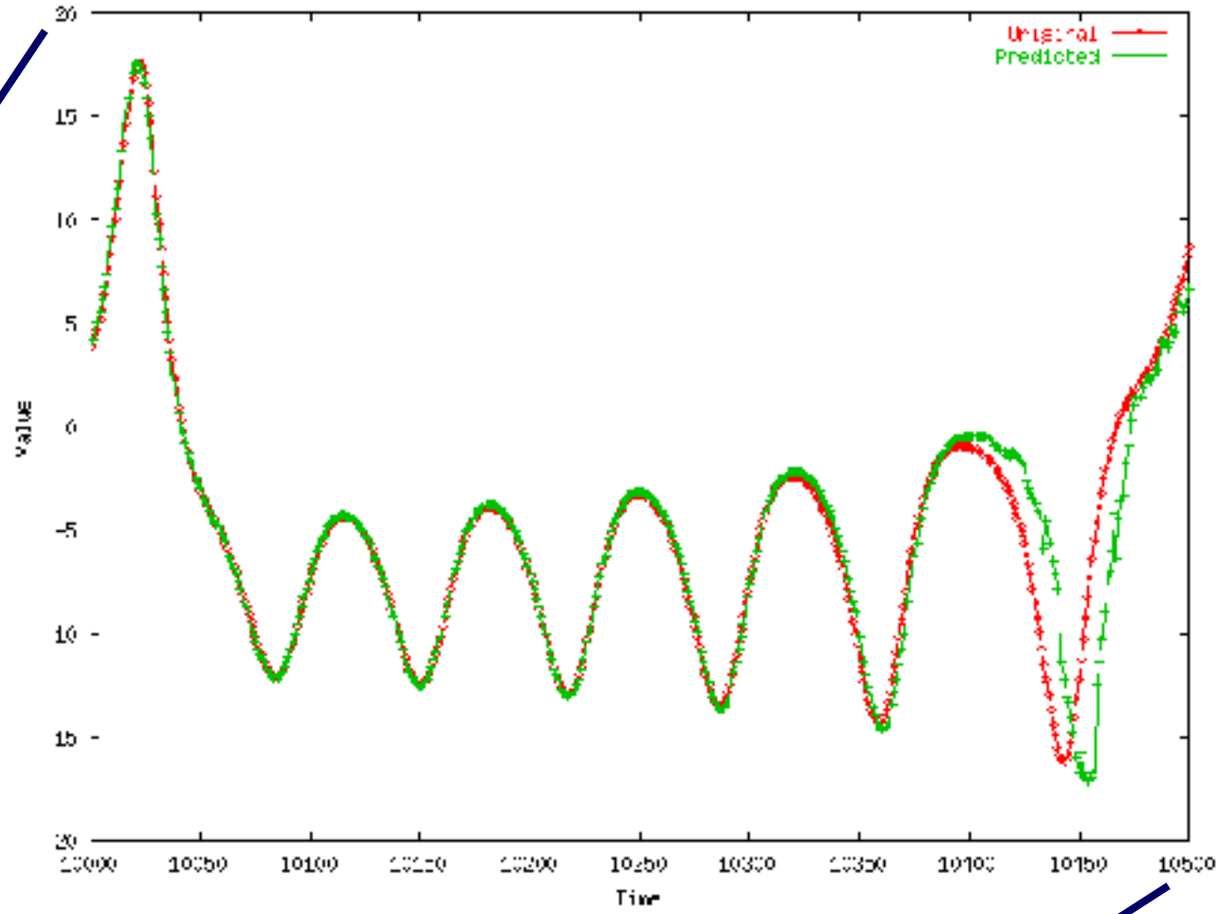




LORENZ

Comparison of prediction to correct values

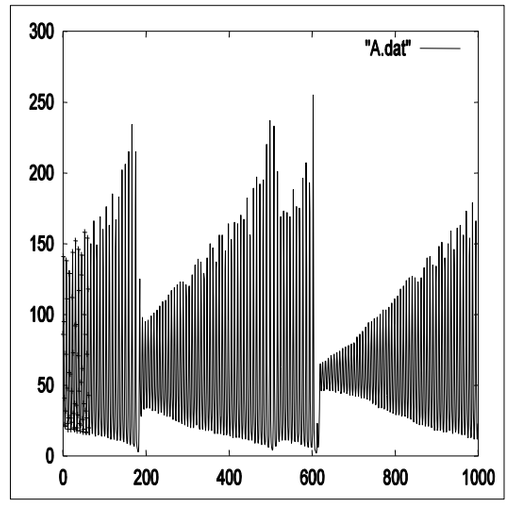




Value

Timesteps

Value



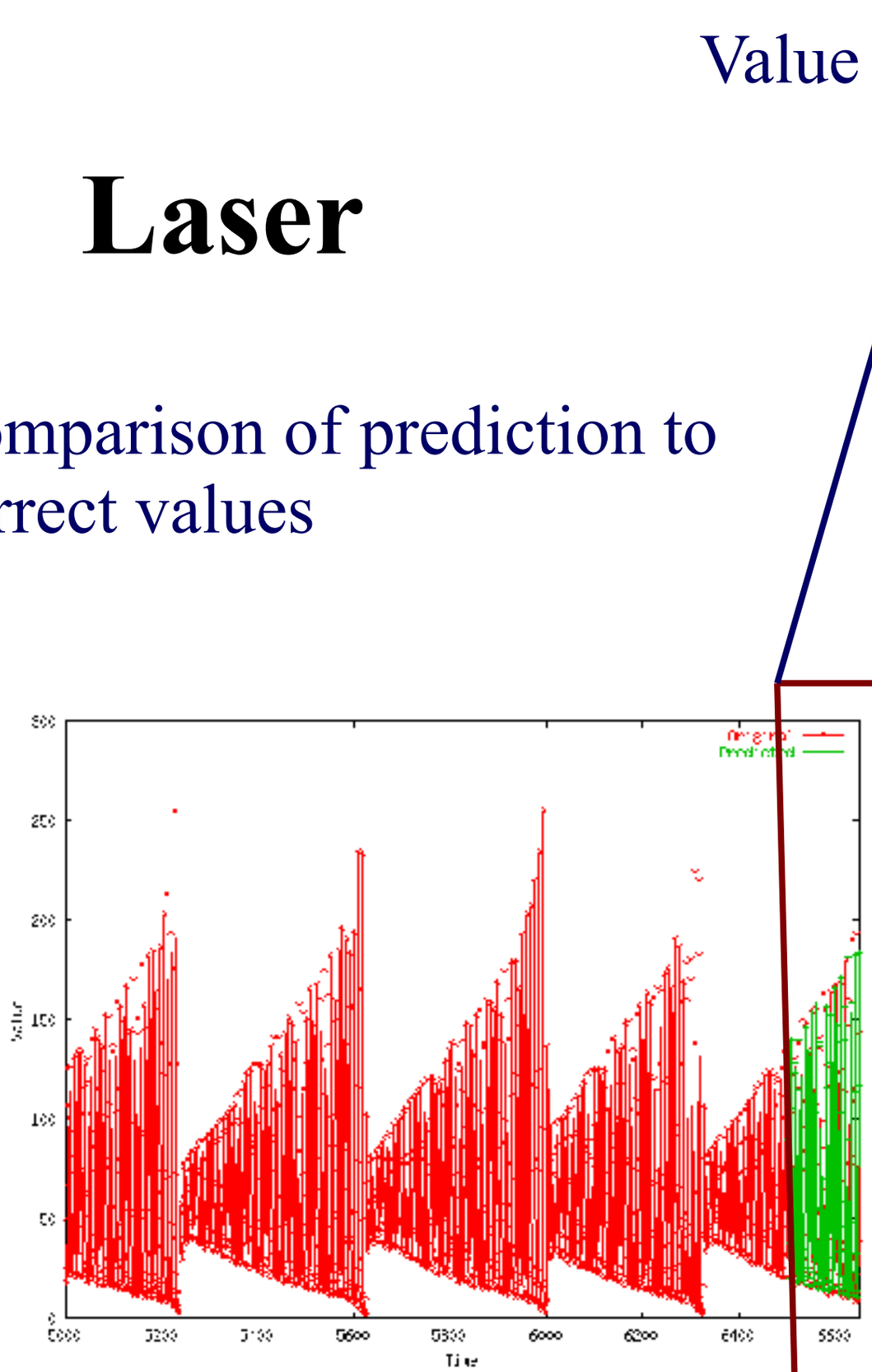
Datasets

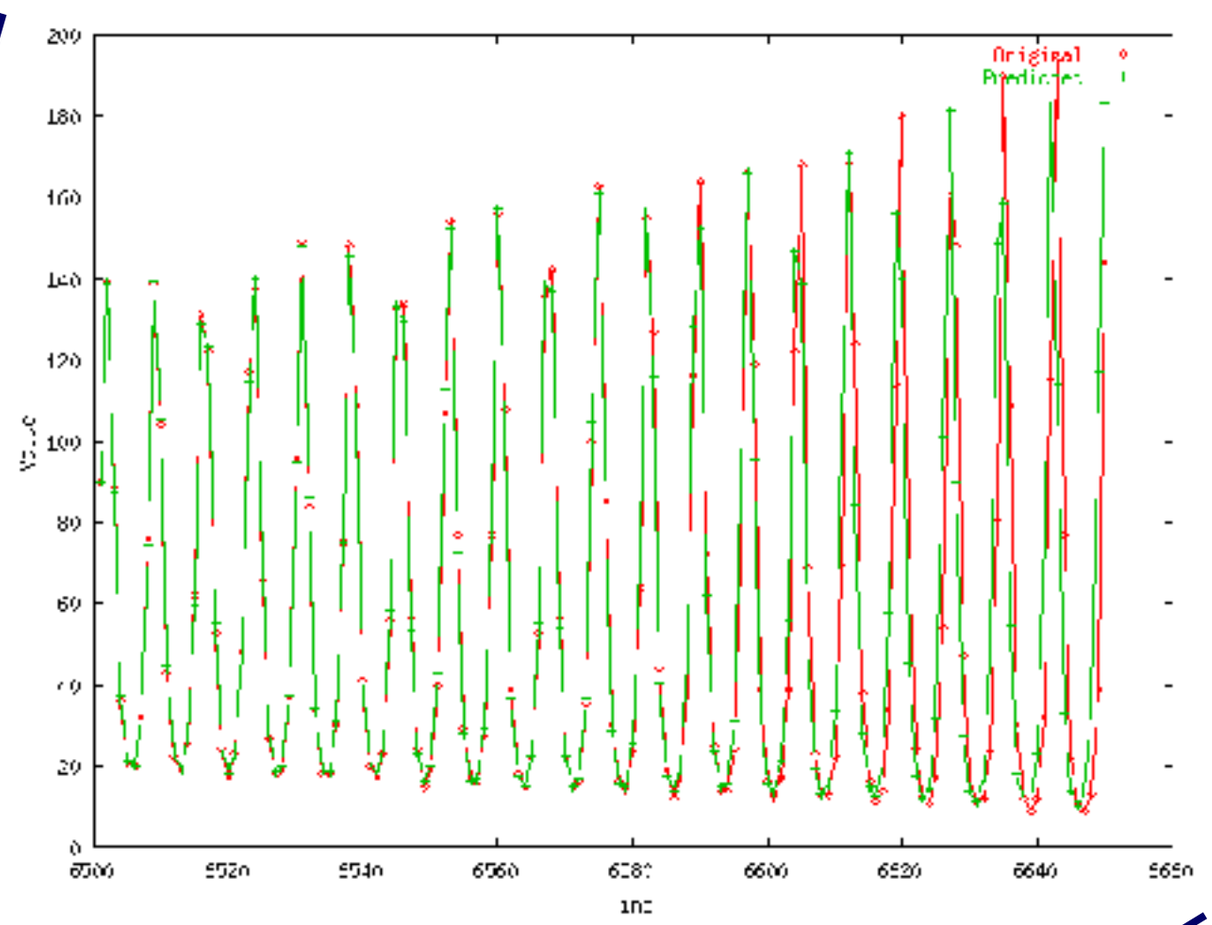
- LASER: fluctuations in a Laser over time (used in Santa Fe competition)

Time

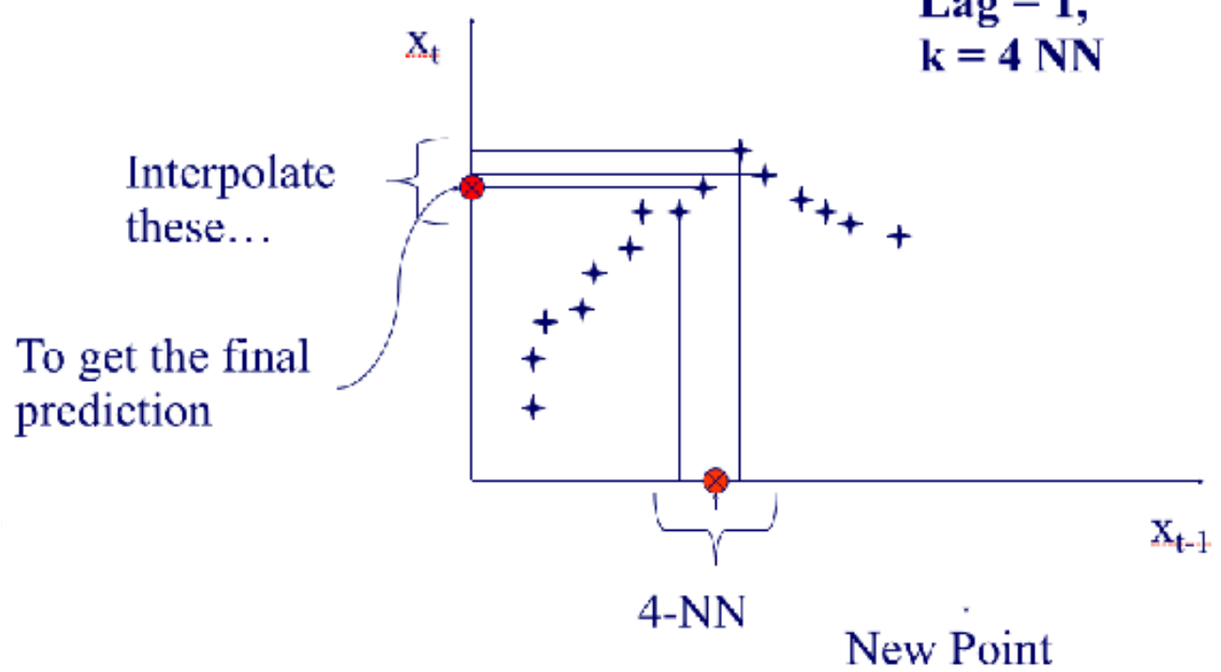
Laser

Comparison of prediction to correct values





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

171.10 +2.02 (1.19%)
After Hours: 171.16 +0.06 (0.04%)
Nov 16, 4:20PM EST
NASDAQ real-time data - Disclaimer
Currency in USD

Range	170.30 - 171.87	Div/yield	0.63/1.47
52 week	106.60 - 176.24	EPS	9.19
Open	171.18	Shares	5.13B
Vol / Avg.	23.52M/26.43M	Beta	1.25
Mkt cap	878.48B	Inst. own	61%
P/E	18.63		

Compare: Enter ticker here Add ☐Dow Jones ☐Nasdaq ☐MSFT ☐VZ ☐HPQ ☐GOOGL ☐SNE more »

Settings | Technicals | Link to this view

Volume delayed by 15 mins.

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:

- Cross filter. http://square.github.io/crossfilter/
- R (ggplot2)
- Matlab
- gnuplot
- **seaborn** https://seaborn.pydata.org

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.)

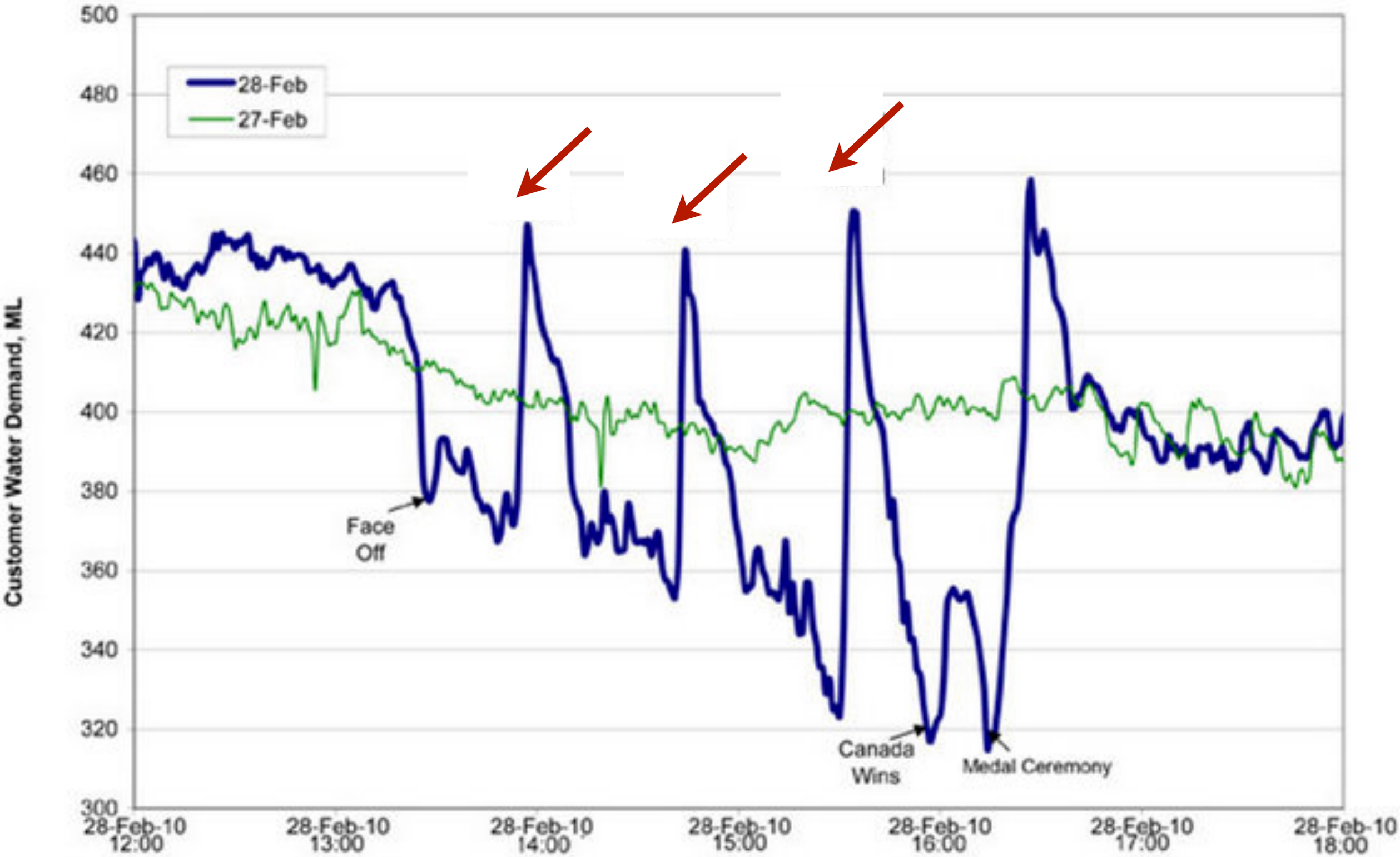
Time Series User Tasks

- 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

EPCOR

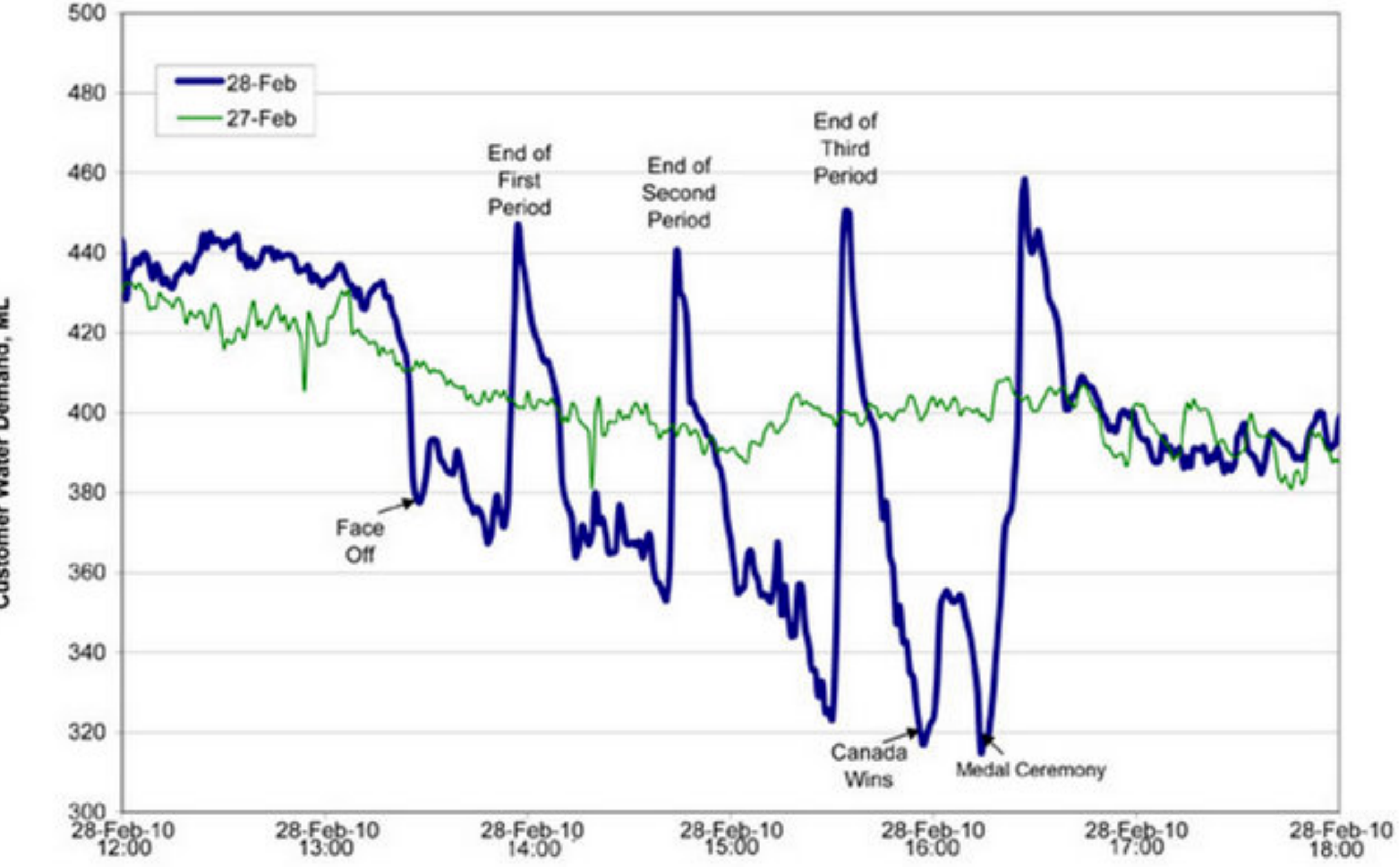
Water Consumption in Edmonton During Olympic Gold Medal Hockey Game



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

EPCOR

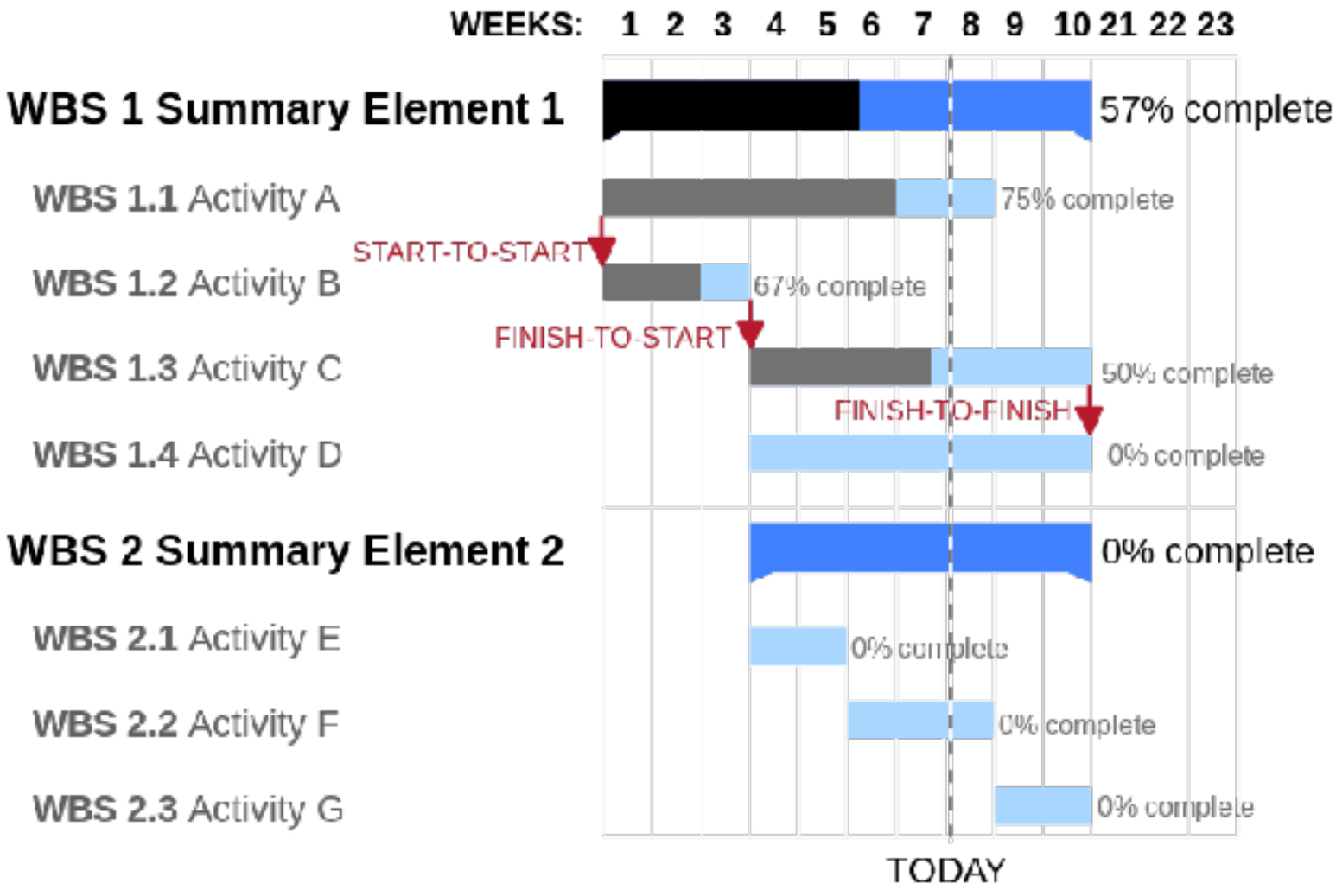
Water Consumption in Edmonton During Olympic Gold Medal Hockey Game



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

Gantt Chart

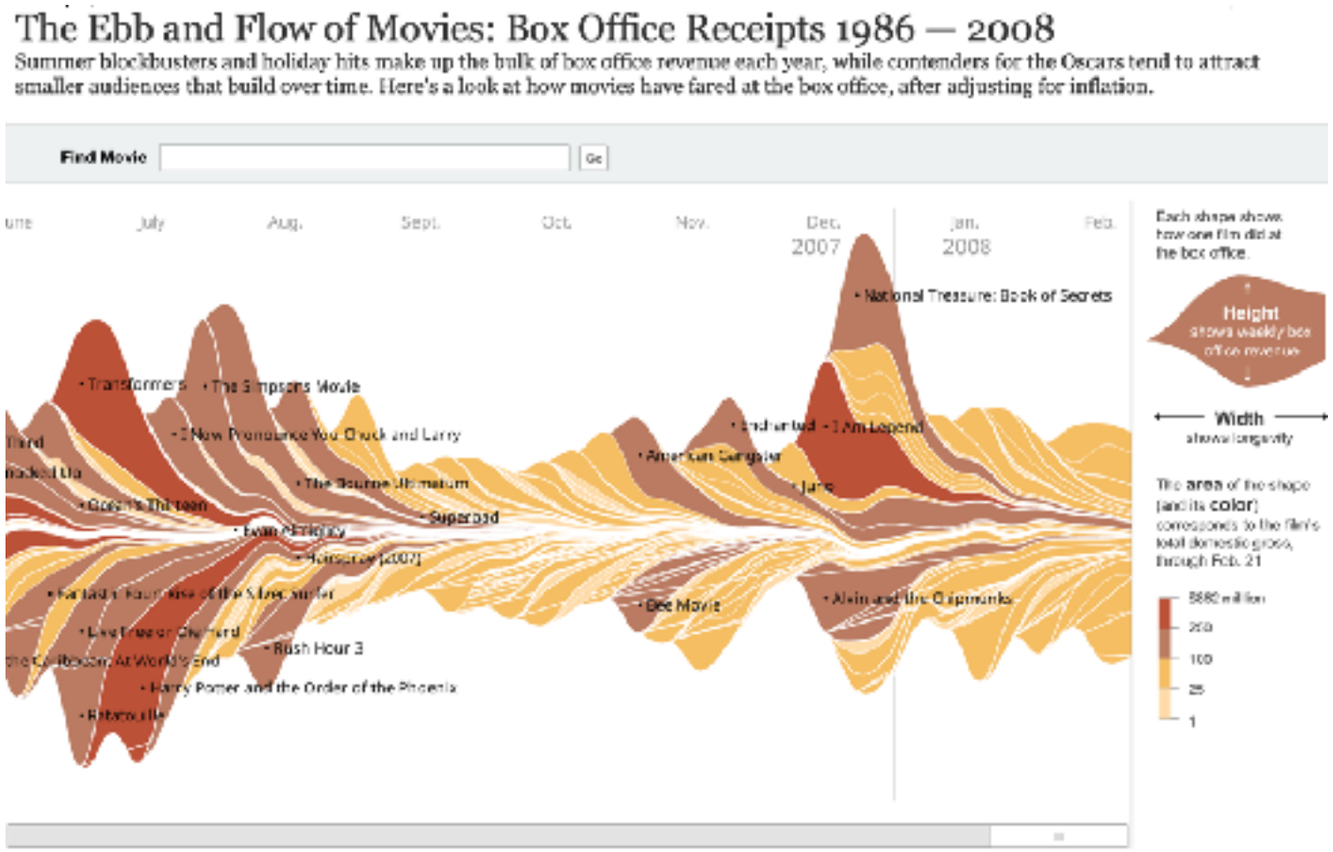
Useful for project

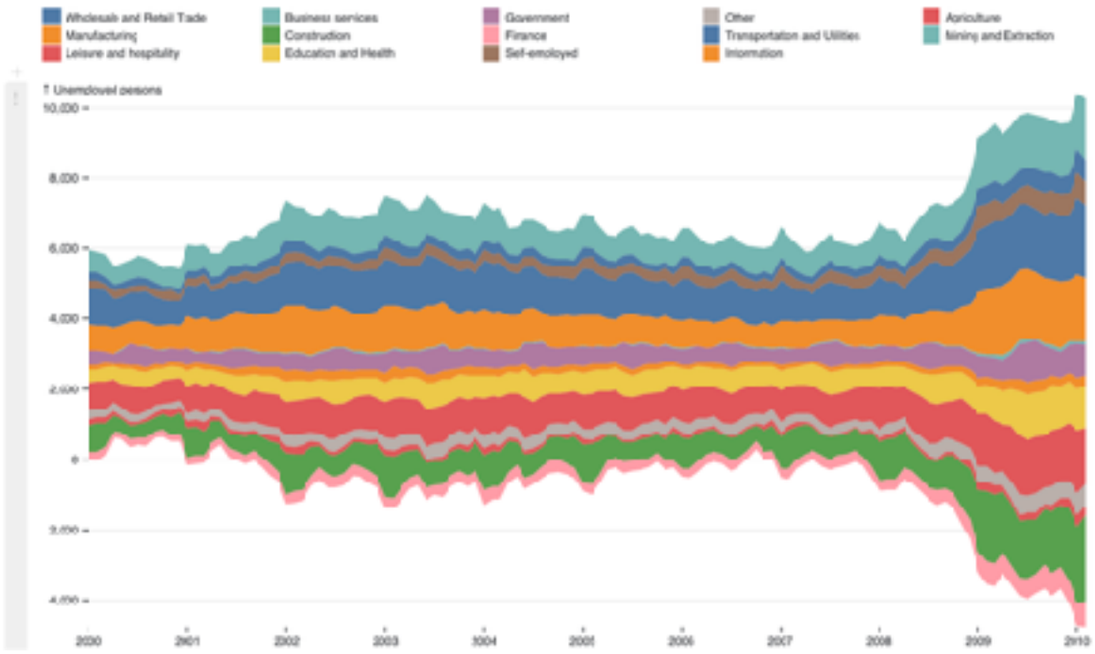


How to create in Excel:

https://create.microsoft.com/en-us/templates/gantt-charts

Streamgraph





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

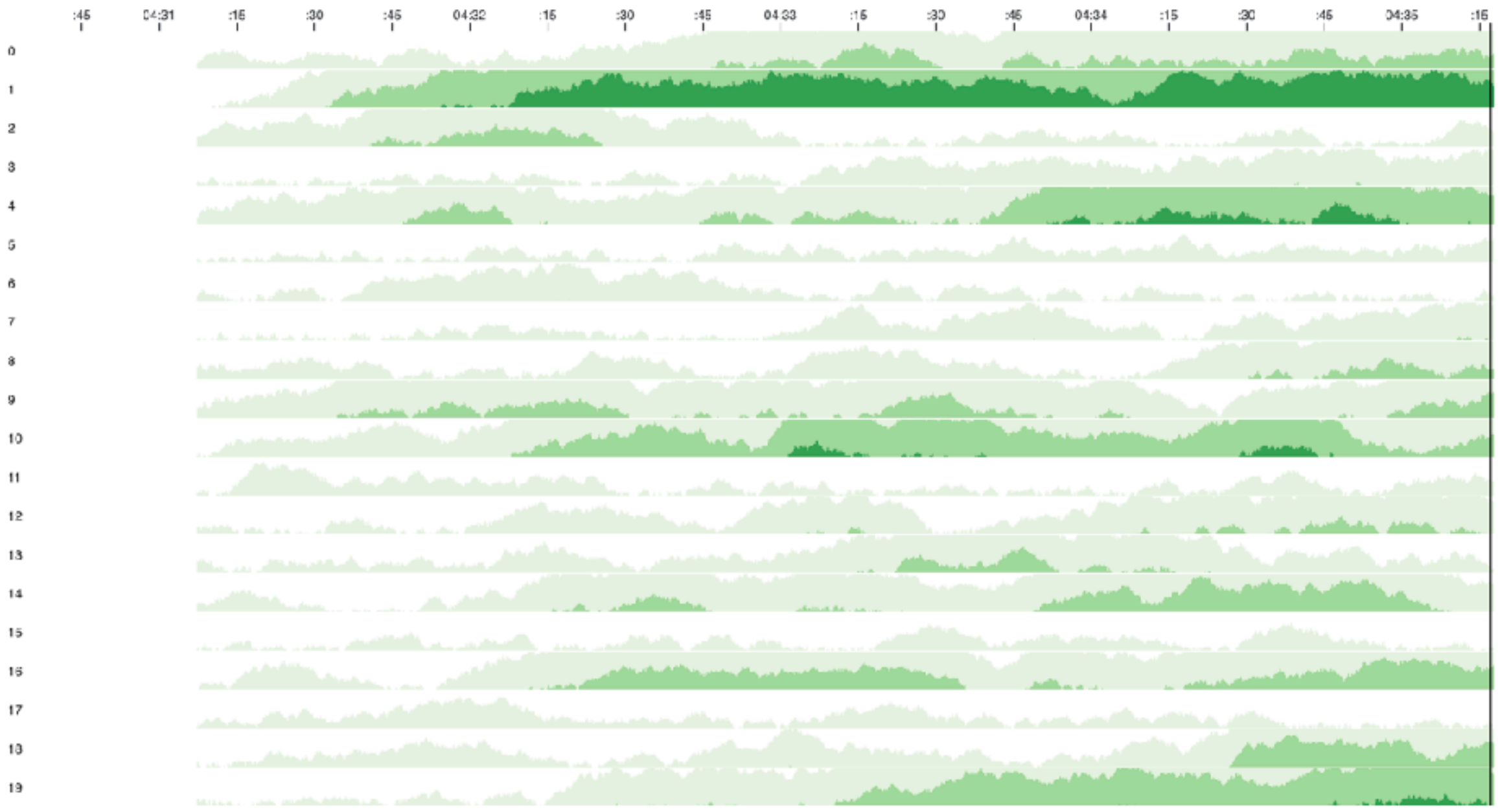
https://observablehq.com/@d3/streamgraph/2

Realtime horizon chart

This horizon chart shows realtime random walks. Only the new values are drawn on each tick, while the old values are shifted to the left.

Color scheme Greens

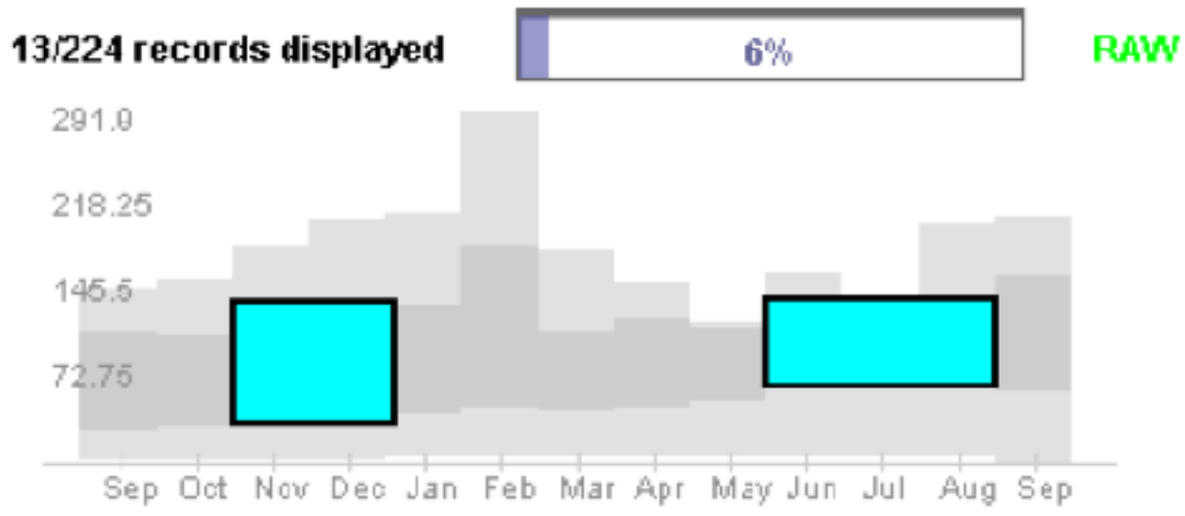
Bands 3



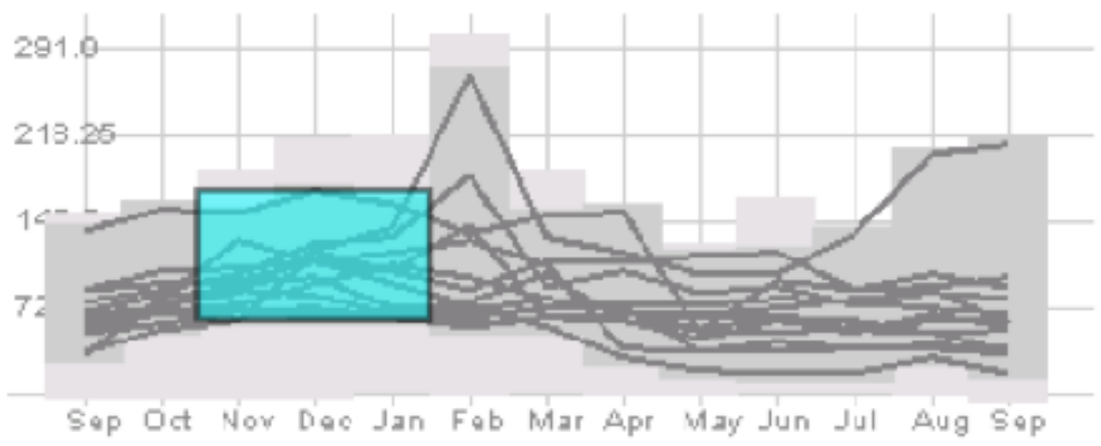
https://observablehq.com/@d3/realtime-horizon-chart

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

https://www.cs.umd.edu/hcil/timesearcher/
http://hcil2.cs.umd.edu/video/2005/2005_timesearcher2.mpg

TimeSearcher...13 Months of High and Low prices

File Edit View Transform Help

Search: Search Clear

Query Variable Low

HCIL

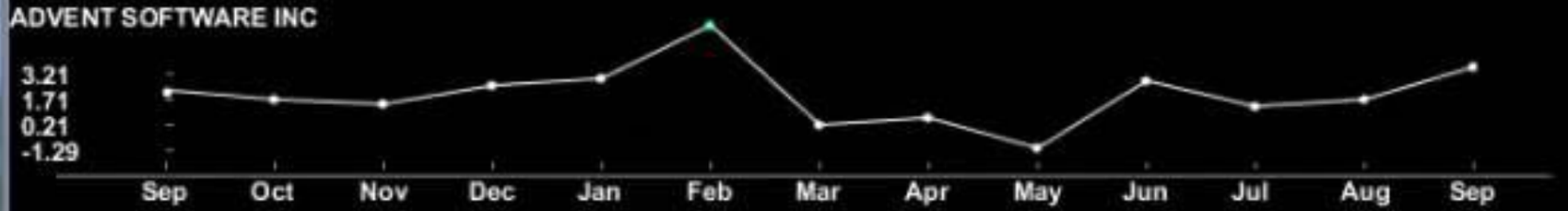
Low

Name:	ADTRAN INC	
	Low	High
September:	37.56	40.5
October:	36.44	37.63
November:	38.69	40.88
December:	51.25	52.75
January:	62.63	66.25
February:	72.06	75.88
March:	57.88	61.75
April:	65.5	67.75
May:	58.25	65.13
June:	57.38	60.63
July:	59.88	61.75
August:	50.63	54.75
September:	41.5	46.25

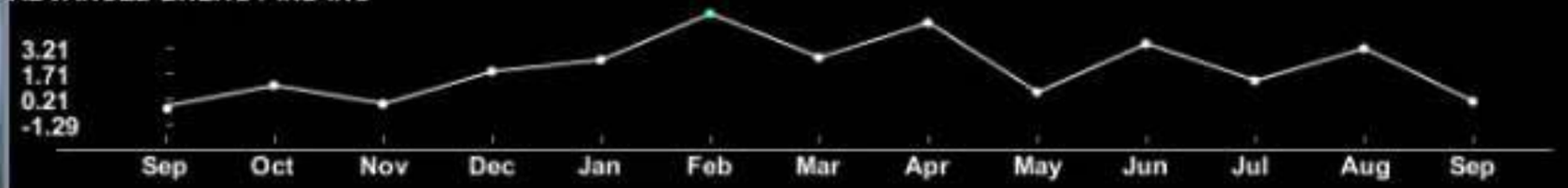
34/223 Records 15% DEVIATIONS

ADTRAN INC

ADVENT SOFTWARE INC



ADVANCED ENERGY IND INC



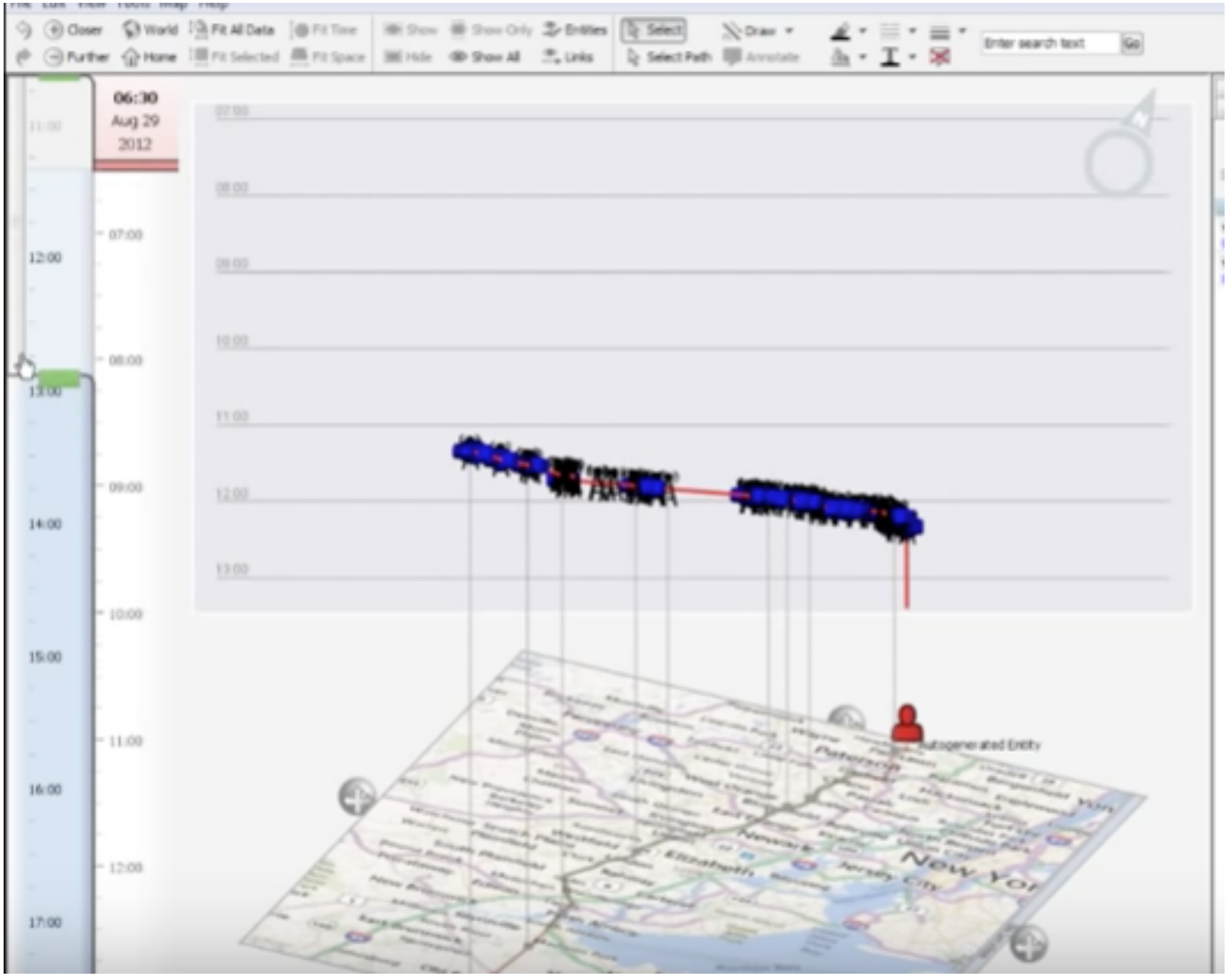
ACACIA RESEARCH CORP
ACTUATE CORPORATION
ADAPTIVE BROADBAND CORP
ANALOG DEVICES INC
ADVANCED DIGITAL INFO
AUTODESK INC
ADTRAN INC
ADVENT SOFTWARE INC
ADVANCED ENERGY IND INC
ADVANCED FIBRE COMMUN
ALPHA INDUSTRIES INC
ALKERMES INC
ALLIANCE PHARMACEUTICALS
ANALOGIC CORP
APPLIED MATERIALS INC
APPLIED MICRO CIRCUITS
AMERICAN SUPERCONDUCTOR
ANADIGICS INC
ARIBA INC
AREMISSOFT CORPORATION
ARM HOLDINGS PLC ADS
ARROW INTERNATIONAL INC
ART TECHNOLOGY GROUP INC
ASM LITHOGRAPHY HOLDINGS

February February

1.55 3.21

GeoTime

Infovis 2004



https://www.youtube.com/watch?v=XRyI3yaIIIs

https://web.archive.org/web/20100531063054id_/http://www.oculusinfo.com/papers/KaplerWright_GeoTime_InfoViz_Final_Conf.pdf

GeoTime Getting Started

Beginner Video #1

View Modes