Grand Valley State University
ScholarWorks@GVSU

2014 Presentations Big Data Conference

4-2014

High Performance Statistical Computing

David Zeitler
Grand Valley State University, zeitlerd@gvsu.edu

Follow this and additional works at: http://scholarworks.gvsu.edu/bigdata_conference2014

Recommended Citation
Zeitler, David, "High Performance Statistical Computing" (2014). *2014 Presentations.* Paper 7.
http://scholarworks.gvsu.edu/bigdata_conference2014/7

This Article is brought to you for free and open access by the Big Data Conference at ScholarWorks@GVSU. It has been accepted for inclusion in 2014 Presentations by an authorized administrator of ScholarWorks@GVSU. For more information, please contact scholarworks@gvsu.edu.

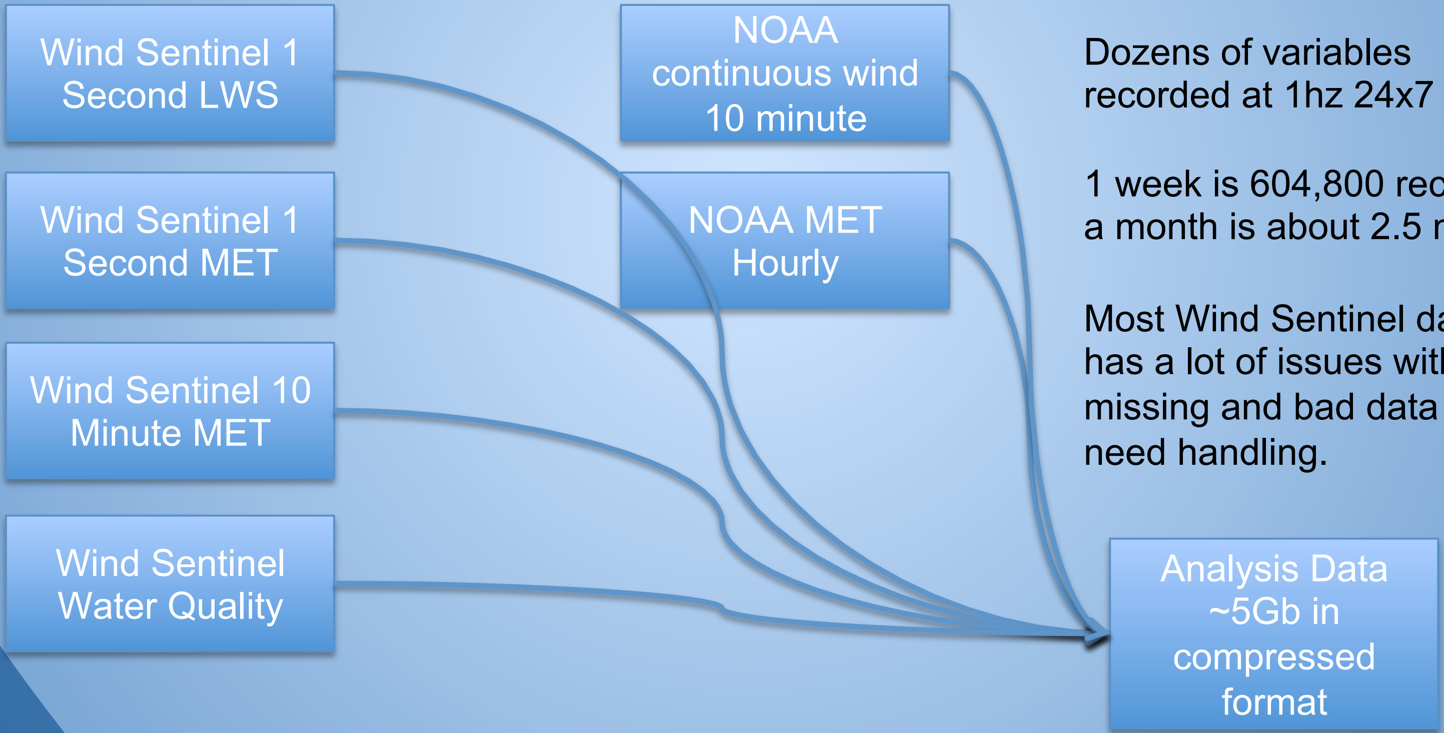
Lake Michigan Wind Data Quality and Analysis

Working with rather big data.

GRAND VALLEY STATE UNIVERSITY
MICHIGAN ALTERNATIVE AND RENEWABLE ENERGY CENTER

Acknowledgement: This material is based upon work supported by the Department of Energy; Grand Valley State University; Michigan Public Service Commission; We Energies; Sierra Club of the Great Lakes; Grand Valley State University; Michigan State University, Michigan Natural Features Inventory; and the University of Michigan, under Award Number DE-EE0000294. And the following organizations for technical support: National Oceanic and Atmospheric Administration/Great Lakes Environmental Research Laboratory; National Data Buoy Center; Pacific Northwest National Laboratory; United States Coast Guard; United States Army Corp of Engineers; Michigan Department of Natural Resources; and West Michigan Energy Partners.

Multiple data sources


Wind Sentinel 1 Second LWS
Wind Sentinel 1 Second MET
Wind Sentinel 10 Minute MET
Wind Sentinel Water Quality
NOAA continuous wind 10 minute
NOAA MET Hourly
Analysis Data ~5Gb in compressed format

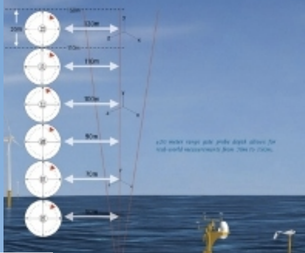
Data from several sources recorded at different rates and in different formats.

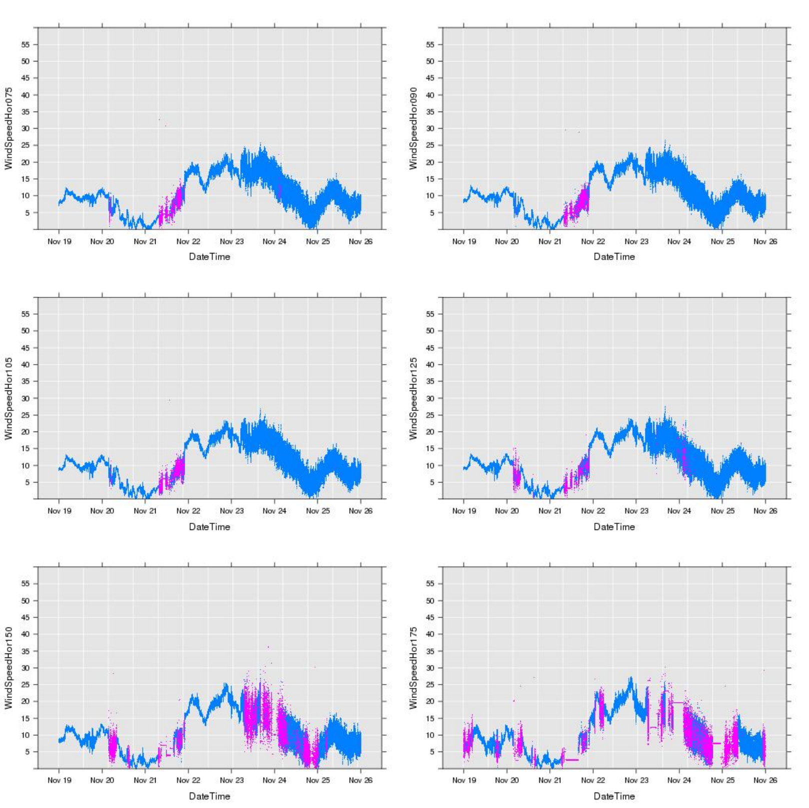
Dozens of variables recorded at 1hz 24x7

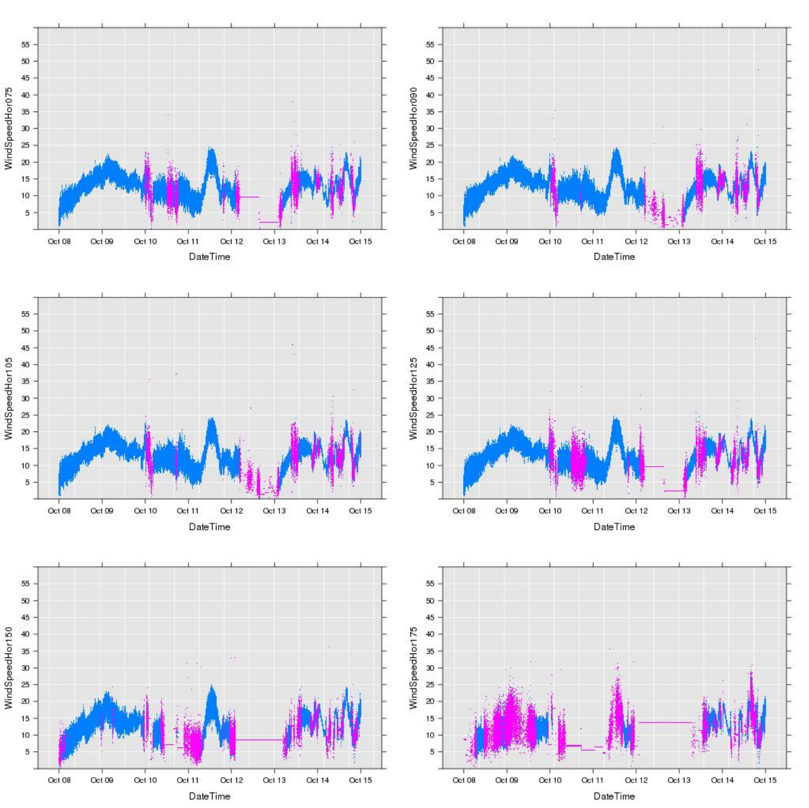
1 week is 604,800 records, a month is about 2.5 million

Most Wind Sentinel data has a lot of issues with missing and bad data that need handling.

Laser Wind Sensor Data



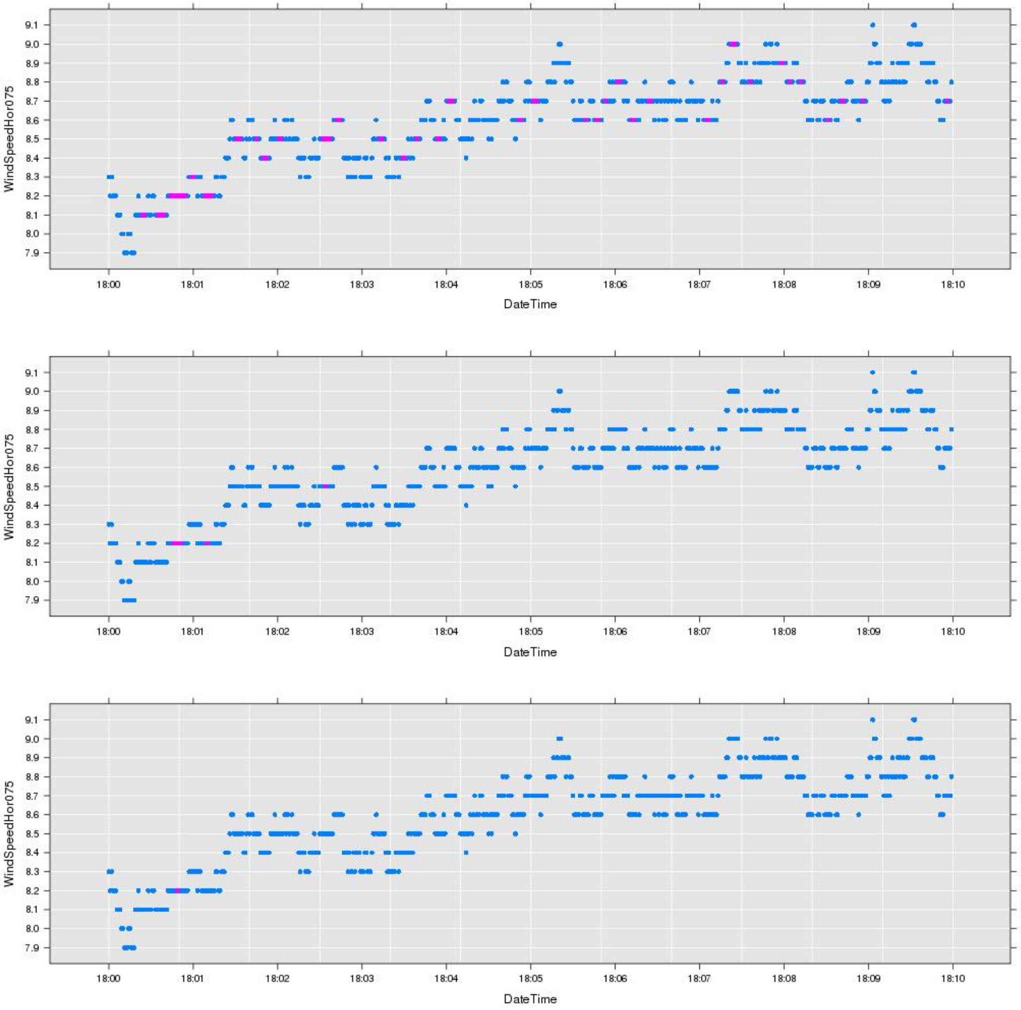




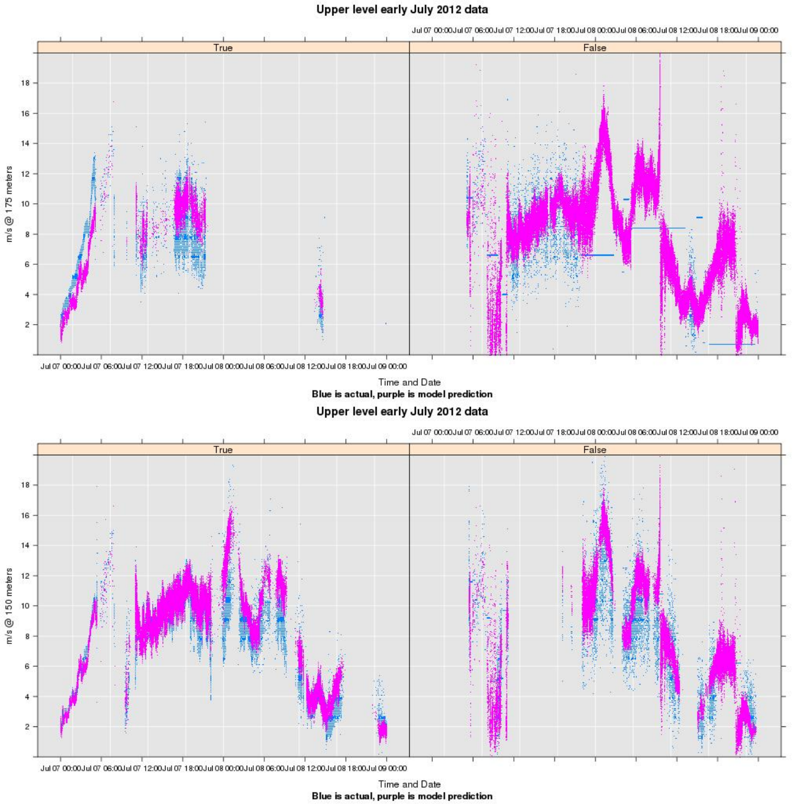
Early attempt

An early attempt to regain data marked as bad was to use moving standard deviations. It was hypothesized that zero standard deviations over several seconds would indicate a ‘stuck’ sensor condition like the flat regions shown in the previous graphic. However a much closer look at the details of the data illustrated by zooming in show that ‘flat’ spots are even more common in very good data such as shown at the left. This is due to the limited resolution of the data.

This graphic shows a very good section of data in high detail with 5, 10 and 15 second moving standard deviations equal to zero being marked in magenta.



The beginning of a solution

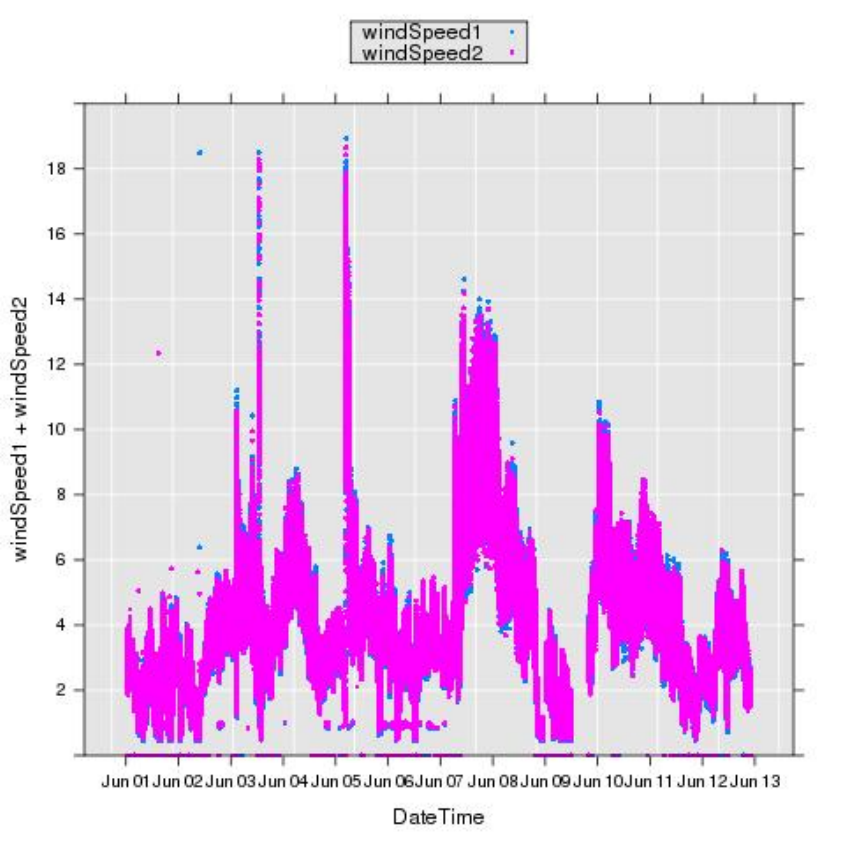


This graphic shows a simple solution.

- Top graphic is 175 m, bottom 150 m
- Left side is data marked good
- Right is data marked as bad
- Blue is actual measurements
- Magenta is modeled data

Note that the modeled data does a good job of filling in the good and bad data with reduced variation while also filling in sections entirely missing.

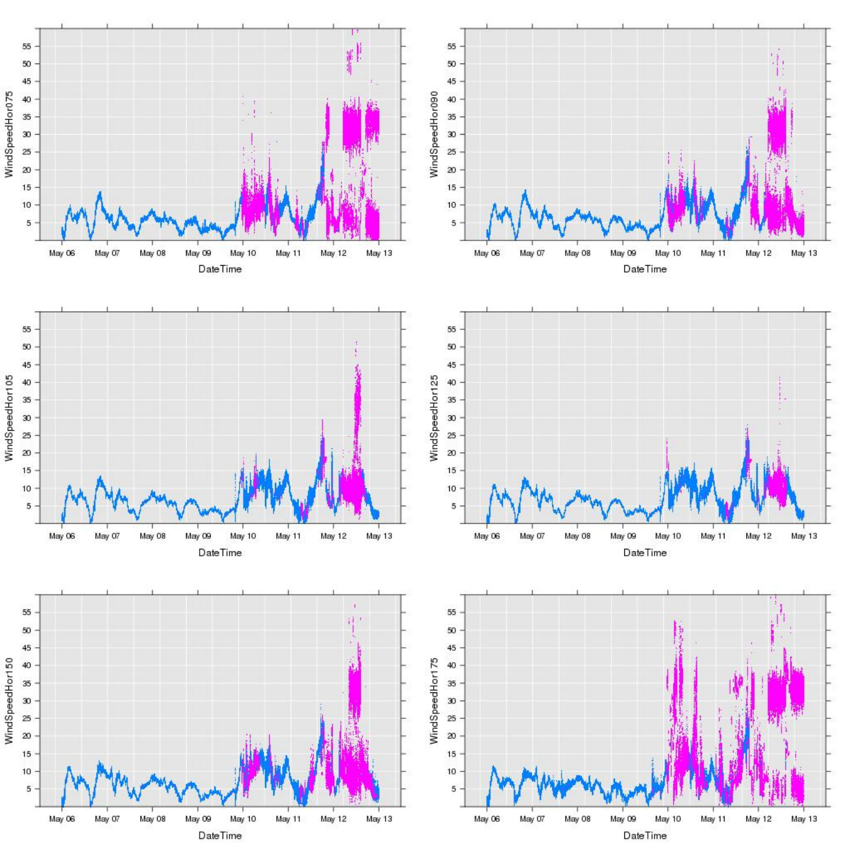
Anemometer data

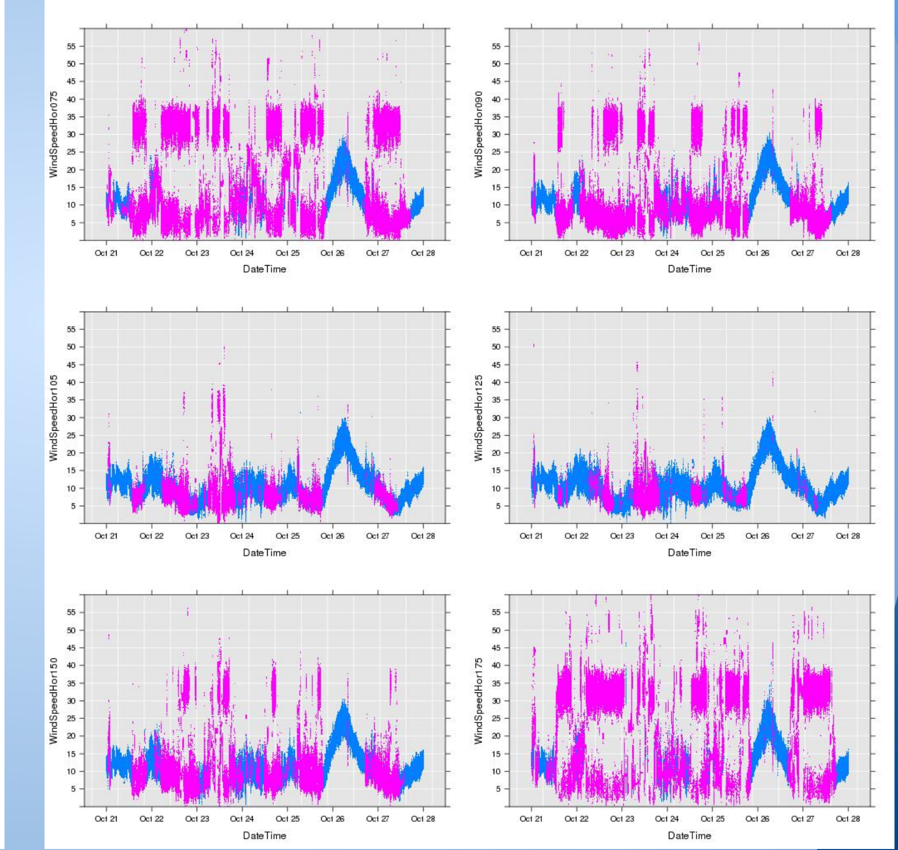


Dual anemometers but VERY unreliable. Note:

- Zero readings across the bottom
- Ghost readings near 1 m/s
- High stray readings (up to 160 m/s)
- Spurious peaks.
- Complete dropout ~ the 9th.

2013 data
Nothing stays the same





meijer

Meijer Marketing Analytics

Working with REALLY big data.

sas

THE POWER TO KNOW

Windows 7

TERADATA

Purpose

- Learn more about the what and why of marketing analytics at Meijer.
- Familiarize myself with the kind of data available.
- Work with the actual tools used by the team.
- Determine what should be changed in my teaching to better prepare students.

What they do

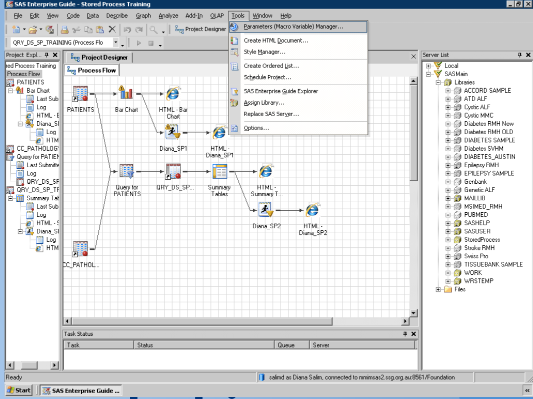
- Try to understand their customers based on data collected from POS systems.
- Determine the effectiveness of incentive programs. Not as easy as it sounds ☺
- Provide marketing personnel insight into customer behavior based on data.

The Data

- every item purchased at any of over 200 Meijer facilities goes into the database.
- ~30 million transactions per month, about 10 records per second 24x7 constantly flowing into just one table of the database.
- over 1 billion records over the last 4 years in this one table alone.
- tie that to data tables for product information, pricing, transaction details, etc. and we have trillions of pieces of information.

Big data tools

- Teradata database system
- Teradata 720 appliance for SAS analytics
  - Grid computing platform, two 8-core machines, 800Gb of memory,
- SAS Enterprise Guide, Visual Analytics, Enterprise Miner and ETS



Ready

salimd as Diana Salim, connected to mmimsas2.ssg.org.au:8561/Foundation

Learning by doing

- Building models for gross sales with the goal of being able to identify unusual variation.
- Used an iterative process of modeling off sources of variation (peeling the onion).
- Even with 700Gb of space to work with, queries had to be broken down into shorter periods.
- Boiled 4 years of data using a series of SQL queries down to ~84,000 records for analysis.

Learned

- Students need to see more:
  - SQL and get database practice.
  - Batch mode processing.
  - Work with a wider set of tools.
    - Advances SAS tools (EG, EM, ETS)
    - SQL and basic database concepts
  - Reinforced the importance of domain knowledge in statistical modeling

Empirical Spectral Test Research

WMU High Performance Computing Cluster

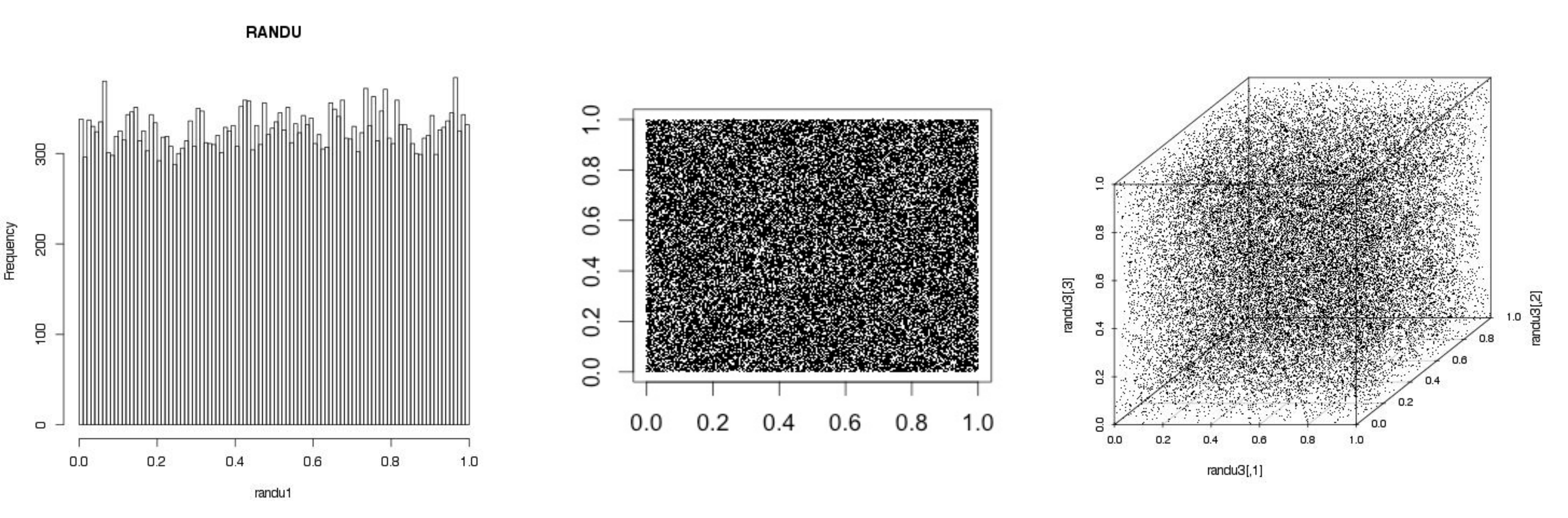
- 20+ nodes
- Up to 512GB ram
- Kepler GPU’s
- Terabyte HD’s

Even bigger data.

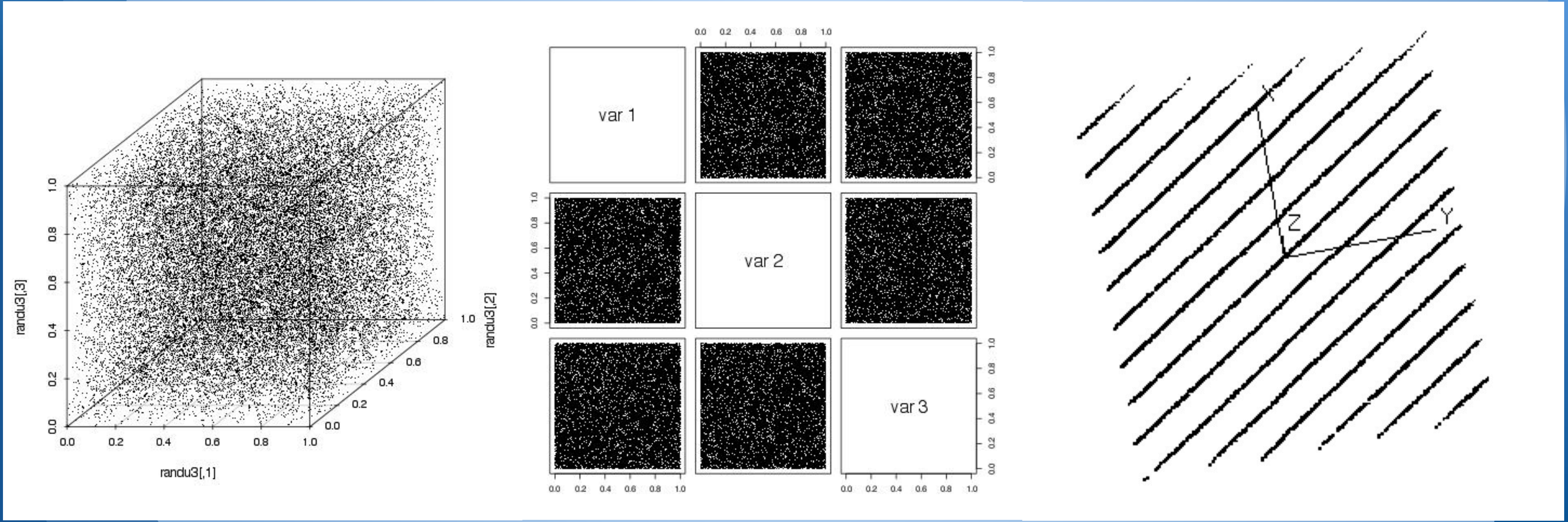
Dell T3500 Server

- Linux OS running SAS 9.4 and Revolutions R Enterprise
- Intel Xeon 6-core
- 24Gb ram
- 6.4Tb hard drives
- NVIDIA GTX 670 (4Gb ram and 1344 cores)

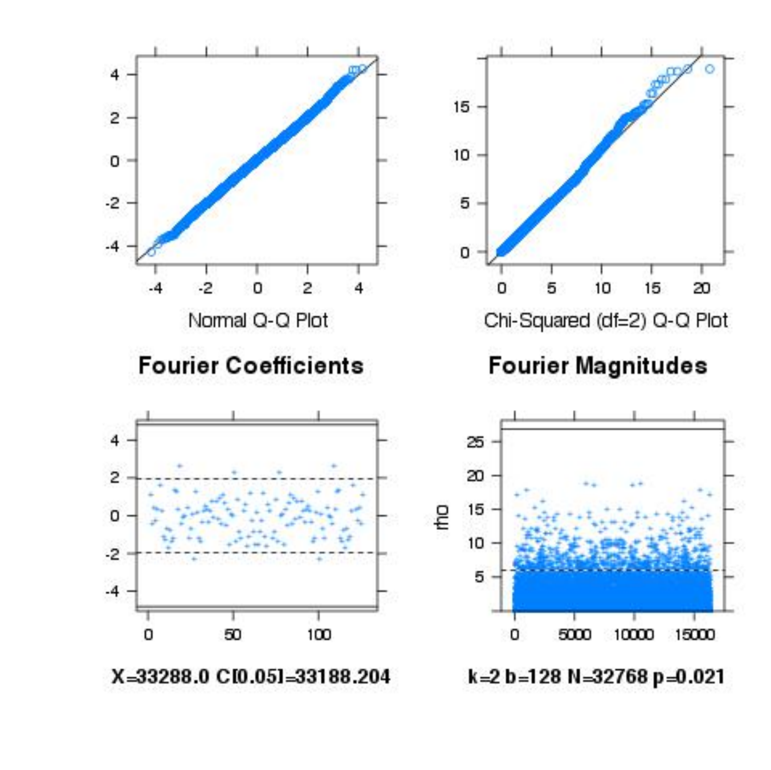
Random data

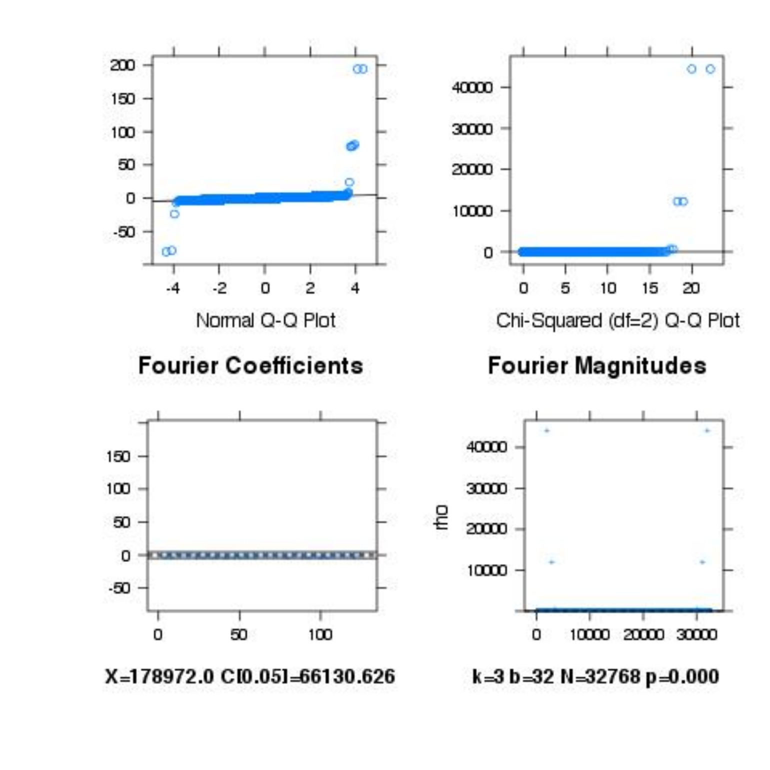

RANDU
Frequency
0
100
200
300
0.0
0.2
0.4
0.6
0.8
1.0
randu1
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
randu3[,3]
0.0
0.2
0.4
0.6
0.8
1.0
randu3[,1]
0.0
0.2
0.4
0.6
0.8
1.0
randu3[,2]

Random data in 3D



FFT of the 3D cell counts





The Data

- Testing random number sources, both pseudo and true random numbers.
- Tests quickly expand to utilize whatever hardware capacity is available, both memory and CPU.
- running a 5 dimensional test with 0.01 resolution does a 5d spatial FFT on 100^5=10 billion double precision complex cells and $3x10^9$ random numbers.

The Tools

- Prototype software being developed in R and RStudio.
- Linking to FFTW to use the cluster
- Writing done with Sweave and Lyx

Just getting started

- Theory has been pretty well hashed out and is being submitted for publication.
- Further developing the R version
- Summer work will include high performance FFT on the cluster
- Preliminary work is suggesting there is a problem with a well known generator used in R.