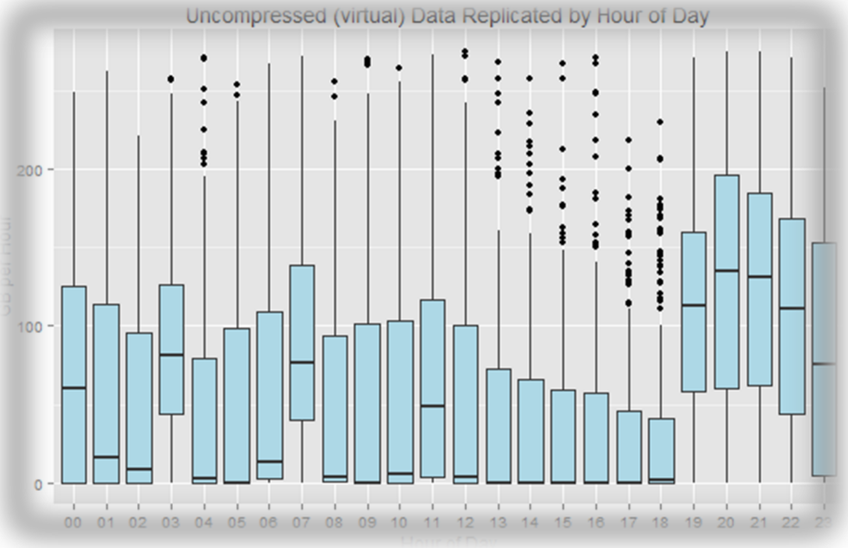


Mark Chamness

Data Scientist

EMC Corporation

Summer 2014

EMC Corporation

- Fortune “100” Corporation
- 61,000 employees in 86 countries
- Products:
  - Hardware: Data Storage systems (Petabyte scale)
  - Software/Security “data protection”
- $24 Billion revenue in 2013

My Background

Academic:

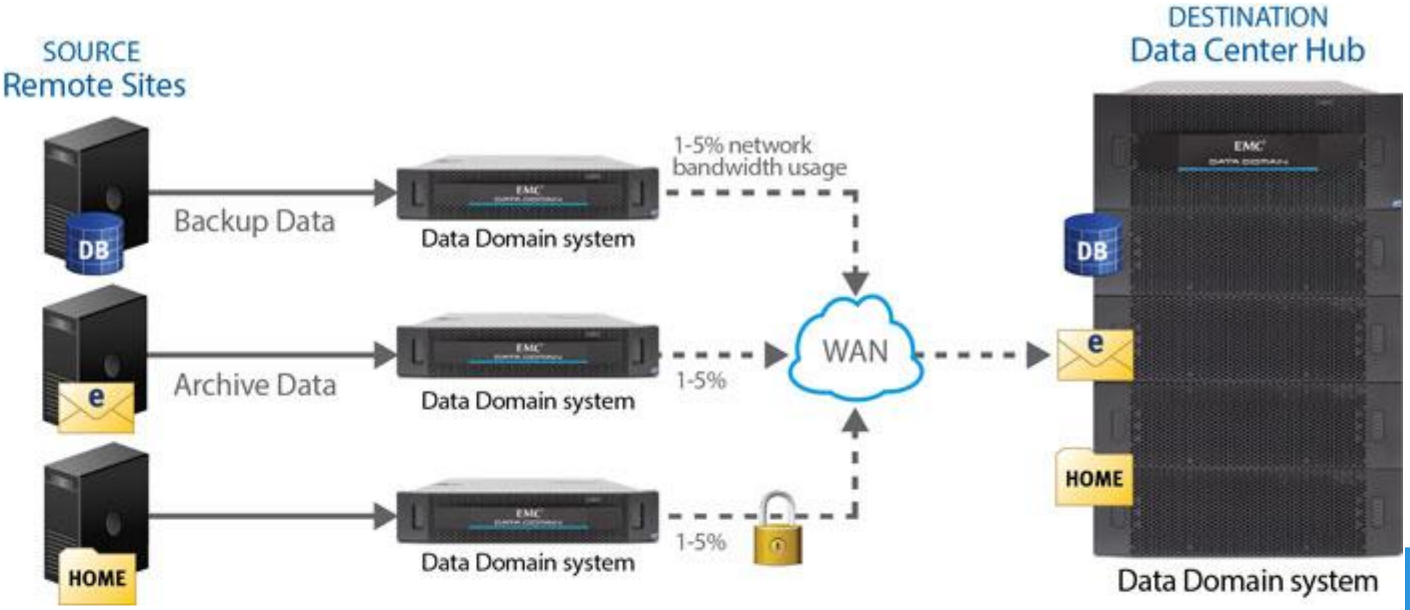
- B.S. Physics, California Institute of Technology
- Ph.D. Candidate, Physics, Brown University
- M.S. Statistics, San Jose State.
  - (Expected 2015)

Finding the Position

- A previous manager contacted me
- After ~1 year, you create your own role
- Be proactive, not reactive in career goals

Why Study Replication?

- Most common customer query: “Replication”
- Long time to resolve



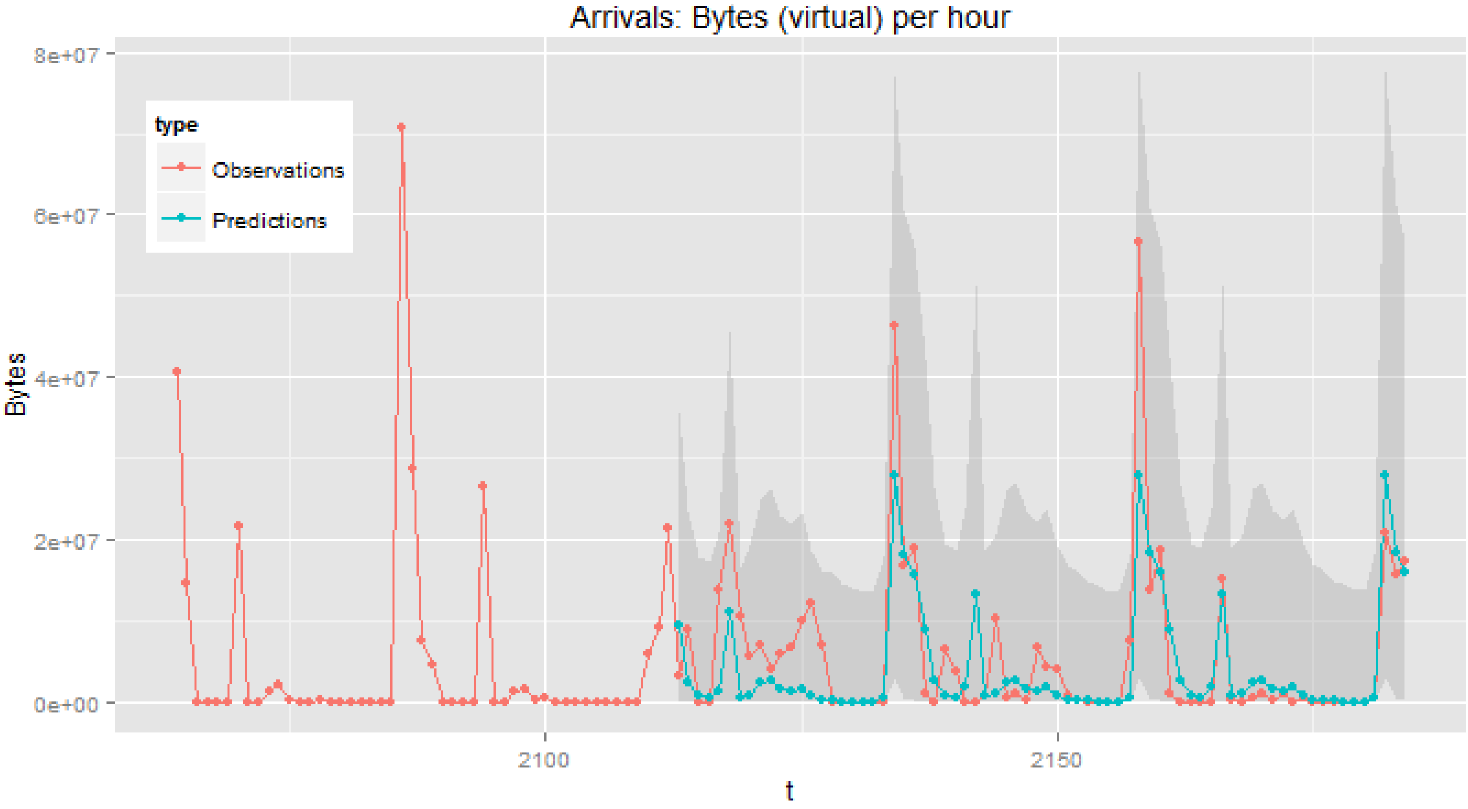
Research Goals

- Define, validate, & document the data set
  - Mountain of undocumented data
- Algorithm to indicate the probability and/or severity of replication lag

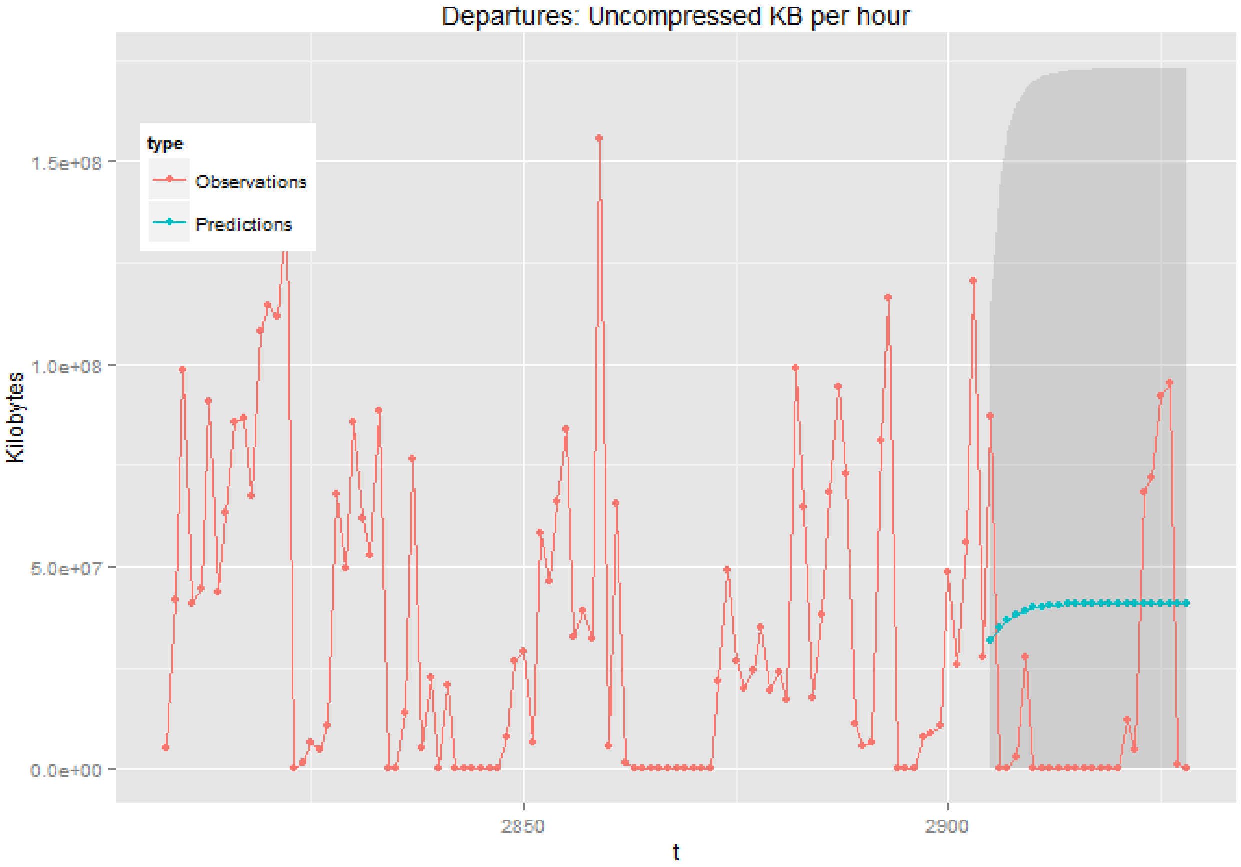
Challenges: Big Bad Data

- Data origin: computers all over world
- Hundreds of variables
- Billions of observations
- Not controlled experiment
- Significant Data Scrubbing

Time series analysis

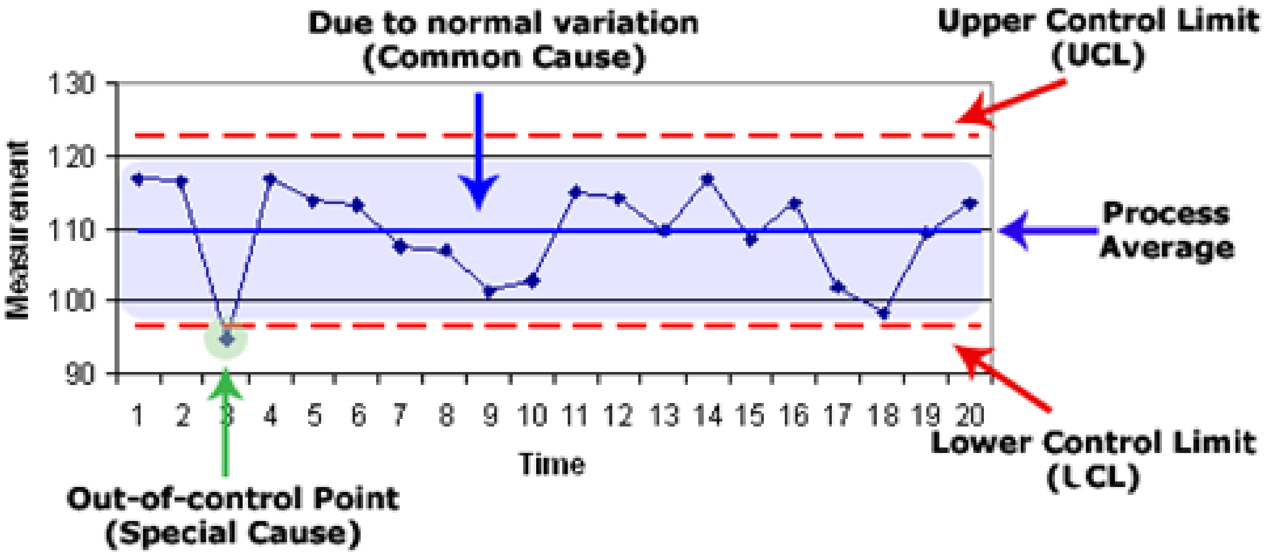


Some forecasts less useful

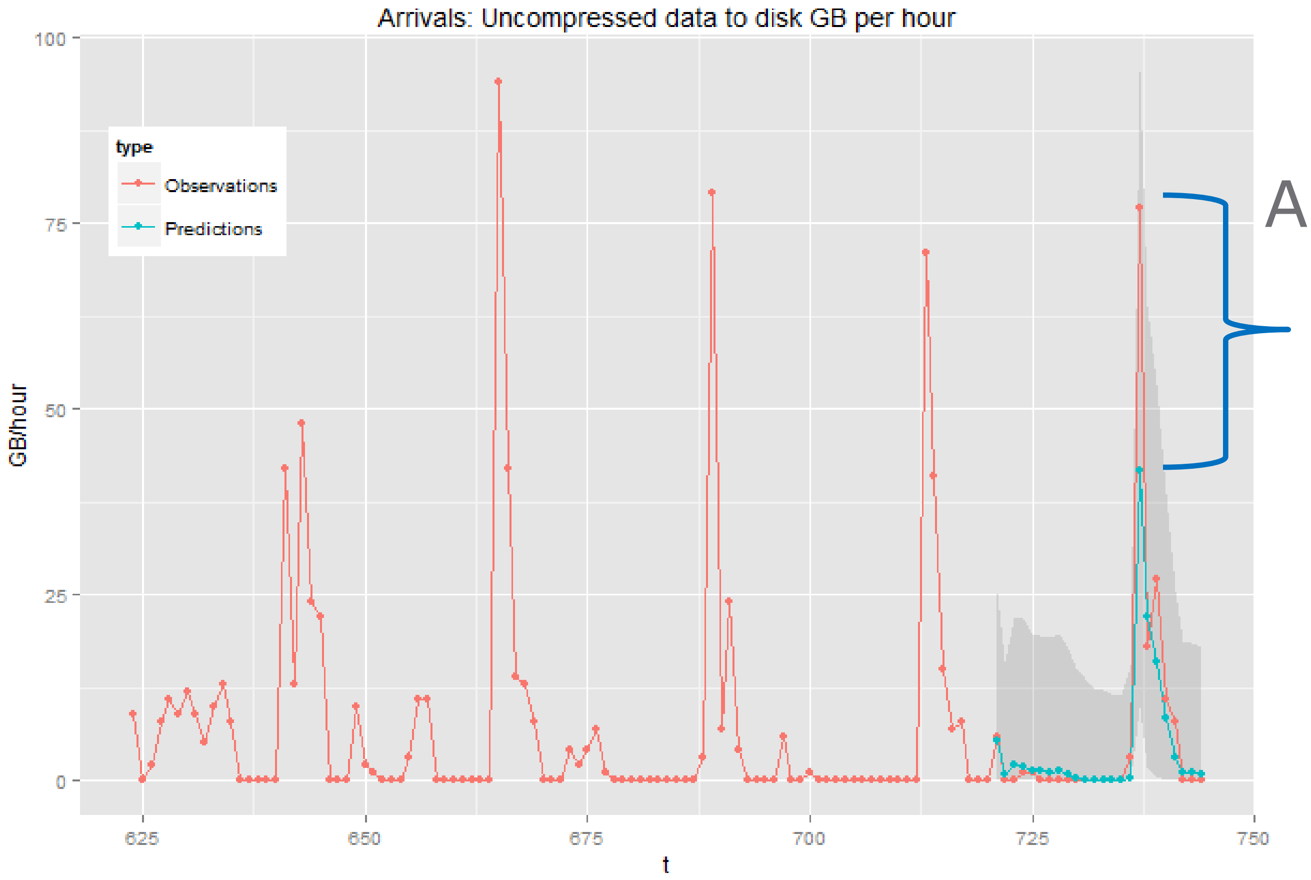


SPC - Statistical Process Control

- Define rules for “Out-of-Control” Conditions
- Test for rule match



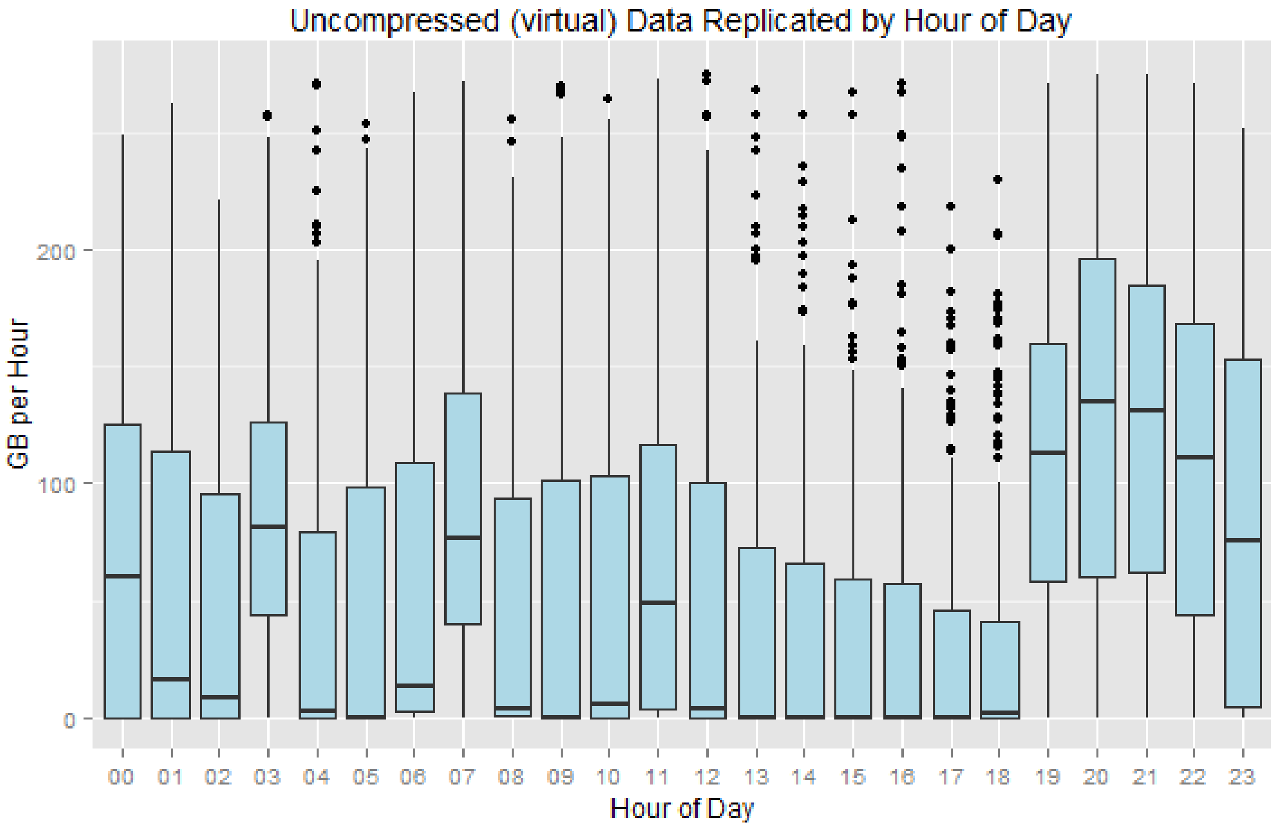
p-values are useful



Almost exceeds 95% confidence interval

~2σ

Replication has hourly correlation



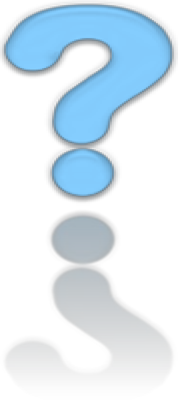
Visualization Helps!

Software/Statistical Methods used

- R/RStudio
  - ggplot
  - Time series analysis
- Database
  - SQL (Postgresql database)
  - pgAdmin
  - Madlib: “Big Data SQL for Data Scientists”
- Documentation via Excel/PowerPoint/Word

Recommendations for Finding Internship

- LinkedIn
  - Educate yourself: find profiles in target role
  - Find relevancy of your background/experience
  - Keep up-to-date & accurate
- Salary.com
  - Gather data on your worth
- Job sites: Dice/BrassRing/etc
- Apply to companies you like



Q&A