Machine Learning Applications to z/TPF Systems

Robert Blackburn Ph.D.

IBM **z/TPF**

April 3rd, 2017

Machine Learning (ML)

- Gives computers the ability to learn without being explicitly programmed
- Artificial Intelligence (AI) has had a long history
    - often with optimistic promises
- In the 1990s ML changed its goal from achieving AI to tackling solvable problems of a practical nature
- It shifted away from the symbolic approaches from AI
    - to methods and models used in statistics and probability theory

Machine Learning

- Many methods developed in 70s and 80s by statisticians and mathematicians
    - Generally linear methods
- In 90s very large data sets became available
- Combined with increase in computational power
    - Fitting non-linear relationships no longer computationally infeasible
        - Potentially more accurate prediction
        - trade-off between prediction accuracy and model interpretability
- Led to new research in computer science as well as statistics

Extreme Thoughts of 2 Kinds

- ML can solve any problem effortlessly
    - just throw data at it
    - by ‘magic’
    - Spend time, effort and $$
        - Unrealistic expectations
        - Disappointing results
- ML can never help in my environment
    - missing out on competitive advantage

Supervised versus Not

- Supervised learning
  - For given inputs the desired outputs are supplied
  - Fit a model that relates response to predictors(covariates)
    - Prediction
      - Aim of accurately predicting response for future observations
    - Inference
      - Better understanding relationship between response and predictors
- Unsupervised learning – challenging situation
  - We observe vector of measurements x_i
  - But no associated response y_i
    - e.g. Cluster analysis
      - Group data in clusters that are similar to each other

Supervised Learning Example

- Let $X = (X_1, X_2, \ldots X_n)$
  - X_i might be characteristics in patient's blood sample
- Y = variable encoding patient's risk for severe adverse drug reaction
- Predict $Y = f(X) + \text{error}$
  - reducible error
    - Improve accuracy by better ML
  - irreducible error
    - unmeasured variables
    - unmeasurable variation

Prediction vs Inference---TPF example

- Prediction
  - Say we had a function f(x)
    - Used 172 variables
      - e.g. NVP + owners + other things
  - Assume f(x) could predict magnitude of spikes in TPF customer utilization throughout the day
    - clearly of significant value
    - highly non-linear model
      - not very interpretable
      - a black box

Prediction vs Inference---TPF example

- Inference
  - Interested in understanding the way Y =f(X) is affected by the 172 variables
    - now f cannot be treated as a black box
    - we need to know its exact form
  - Might try to develop a simpler linear model
    - capture most of black box f predictive ability
      - may not be easy if true relationship is complicated
    - easier inference
      - identify the fewer important predictors
  - Customers potentially could control some input variables

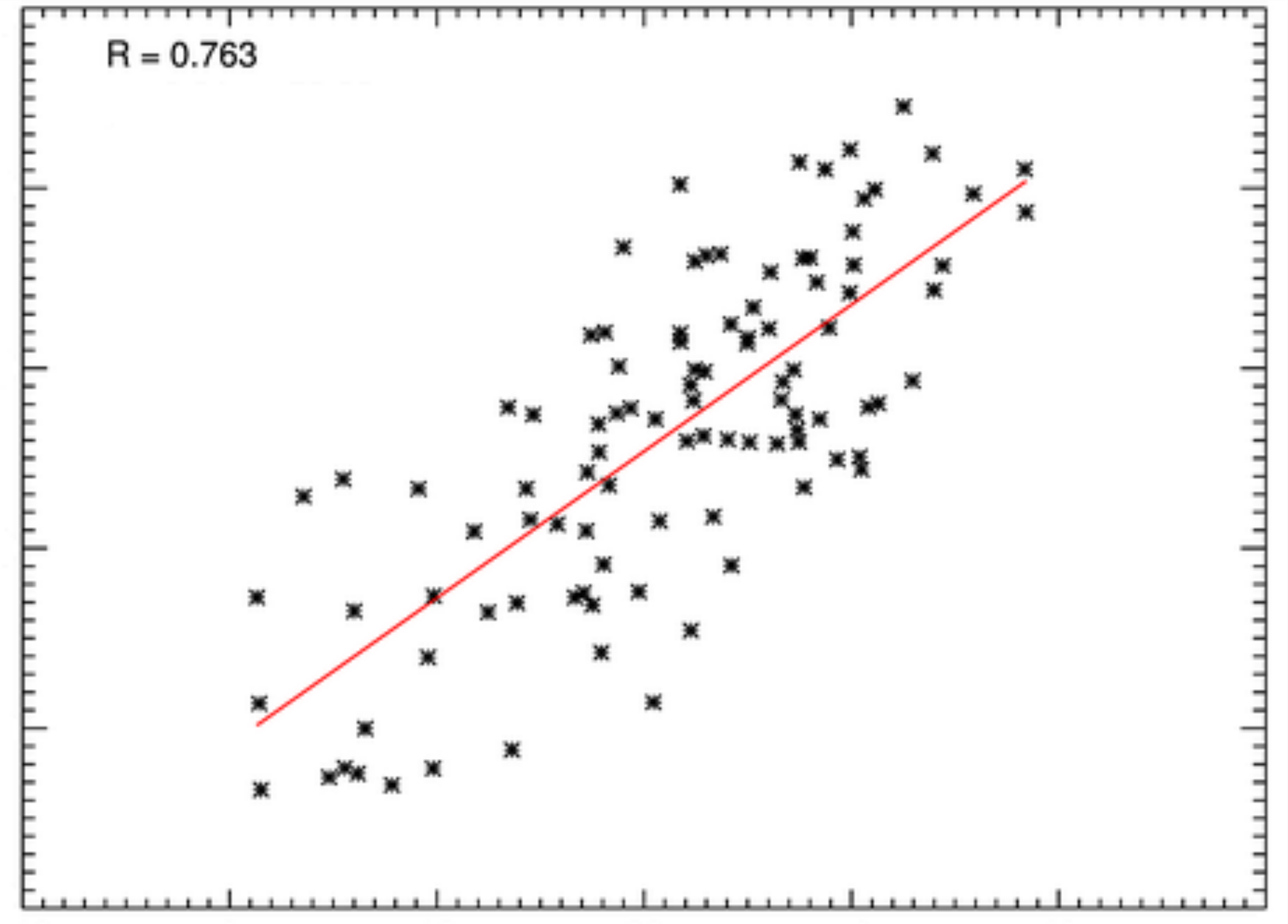
Shoe Leather - Understanding

- Prof. David Freedman
  - Text – ‘Statistical Models and Causal Inference’ – Ch. 3
  - Very readable – no deep mathematics
- Snow’s analysis on cholera used logic and shoe leather
  - Statistics - simple comparison of rates
  - But clever and convincing argument
- Regression models make it too easy to substitute technique for work
  - Asbestos in water vs lung cancer
  - Huge changes in water concentration
    - 5% increase lung cancer
  - Non control for smoking
    - Unconvincing study

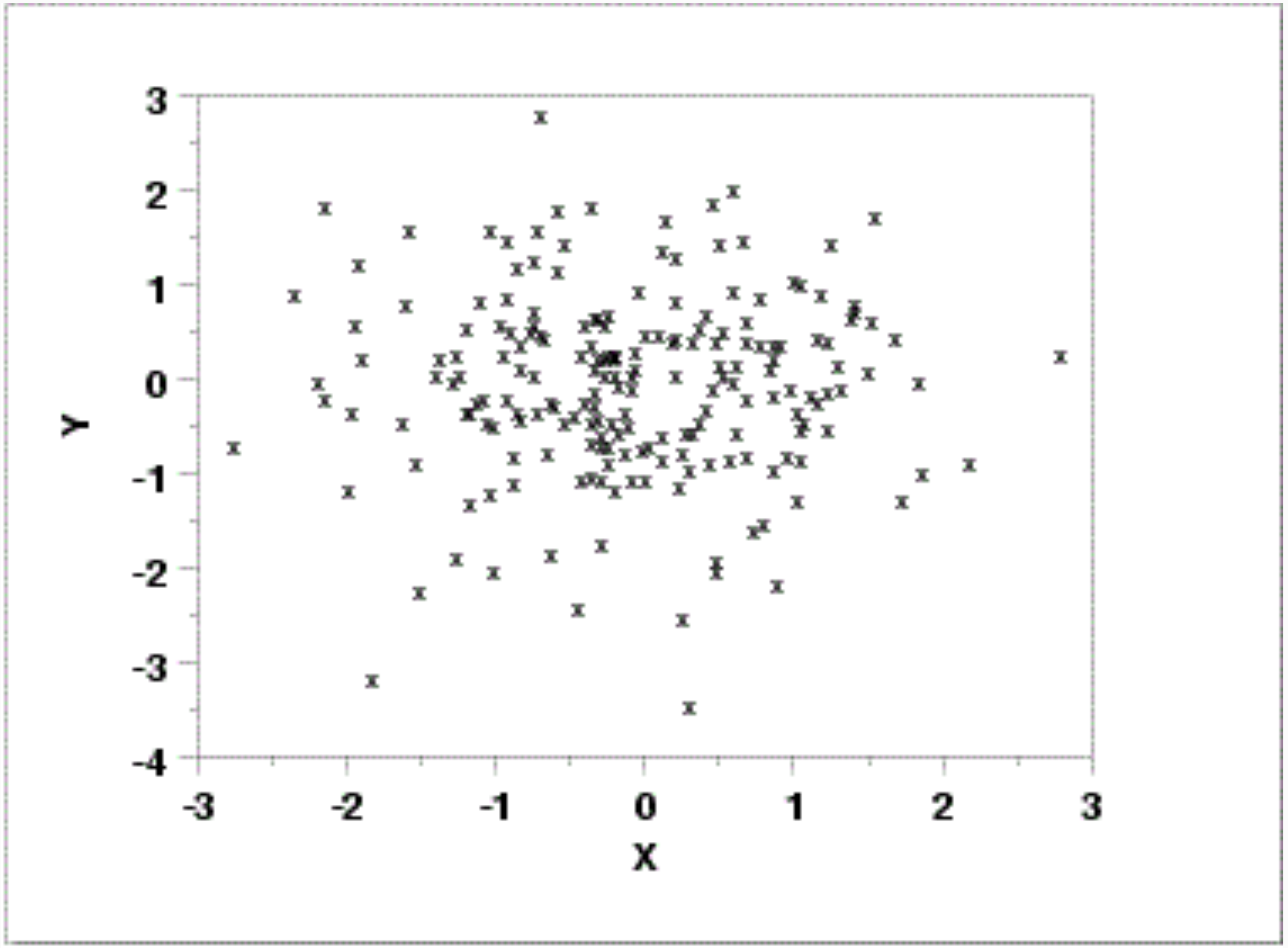
Correlation

- For possibly distinct random variables *X*(*s*) and *Y*(*t*) at different time points the correlation function is
  - $C(s,t) = corr(X(s),Y(t))$
- "Correlation is not causation but it sure is a hint.“
- In a z/TPF system if we had an event/failure
  - Look at covariance matrix
    - Say (3,25) matrix
      - 3 events
      - 25 predictors
  - Examine correlation changes

Height and Weight Correlation for Humans



Stretched Height and Weight Correlation Near 0



z/TPF Customer Sample Event 1

- Critical hash function by mistake had small width
    - bad program loaded
    - large synonym chains
    - major system loss of service
        - CPU utilization near 100% for long period of time
        - ML could have noticed a very significant deviation in CPU usage density
    - various distance metrics between functions
- ML – incorporate z/TPF system changes into algorithms

z/TPF Customer Sample Event 2

- z/TPF periodically going into shutdown for ~10 minutes duration
  - Applications put work (ECBs) on defer list
    - Defer list size got huge and shutdown input list
    - CPU at ~ 100% utilization
      - Usual CPU utilization ~ 70%
- Low rate very expensive searches were consuming all remaining MIPS
- Eventual solution: shut off relevant message ports
  - after month+ of investigation!
- ML combined with Name Value Pair Collection (NVPC) and ECB Owner Names
  - could have found the offered load deviation
  - correlate utilization and message port rate

Sample Space of z/TPF System Events

- Lab and customers (jointly)
  - know z/TPF usages better than anyone else
- Together identify goals for ML
  - many clever algorithms and methods
- Work with ML package
  - Solve z/TPF specific problems

Sample Space of z/TPF System Events

- Problem
  - not enough data (actual failures)
  - Concerned with selection bias
    - Have a few events and ML seemingly applies
- We need to expand to include
  - ‘near miss’ events
- Use our knowledge and intellect to
  - estimate if ML could have helped

Detection of Stricken CEC in a z/TPF Loosely Coupled (LC) Environment

- ML could have LC heartbeat rates
- Deactivate dead CEC
  - build trust in ML
- Take DASD mods offline
  - build more trust in ML

z/TPF Interface with Outside Systems

- ML needs inputs from all systems involving in processing transactions
  - not just z/TPF
  - careful choice of variables
    - not easy
    - iterative approach
- Critical need for
  - common time source
- Covariance matrix to discover interesting relationships

Executive summary

- ML's time has come
    - 25 years of growth
- Significant business value
    - predict future events given training data
- Need to find specific TPF problems for ML to work on
    - supervised learning

Customer Recommendations

- Sign up as sponsor users
- Provide IBM with production data to feed into ML
  - exploit NVPC and ECB Owners
  - consoles
- Document failures/outages
  - with some detail
  - With a view ‘could ML have assisted?’
- Add appropriate context information to your business events to enable business analytics