Experimental Approaches in Computer Science

Dror Feitelson
Hebrew University

Lecture 5 – Data analysis

Given measured data, we can

- Describe it
  - Mean, median
  - Range, standard deviation
  - Histogram, scatter plot, empirical distribution
- Model it
  - Fit to a distribution function
  - Apply regression
  - Find a generative mechanism

Fitting a Distribution

- Data is given
- Assume data was created by sampling from a distribution
- What distribution was this?
  - Use a set of predefined candidates
  - Estimate parameter values
  - Check goodness of fit
- Process can be automated
- Limited to those predefined distributions
  - Cannot handle mixtures

Parameter estimation

- Moments matching method
- Maximum likelihood method

Example: moments matching for gamma distribution

- Distribution has two parameters: α and β
- Mean is $\bar{x} = \alpha\beta$
- Variance is $var(x) = \alpha\beta^2$
- These can be inverted to find parameters based on (estimated) mean and variance:

$$\alpha = \bar{x}^2 / var(x)$$

$$\beta = var(x) / \bar{x}$$

Warning:

- With *k* parameters, need to use *k* moments
- High moments are very sensitive to outliers
- Especially troublesome for distributions with a heavy tail
- Example: 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 15
  - 5^{th} moment is 63,676
  - 99.4% of this is due to the outlier 15
  - If outlier was 16, 5^{th} moment would be 87,776
  - an increase of 38%

Maximum likelihood estimation: find the parameter values that are most likely to have led to the observed samples

- Likelihood is product of probability to observe each one
- Differentiate and equate to 0 to find max
- Done in log-space to turn product into sum

Example: exponential distribution

- Likelihood is product of probability to observe these samples

$$L(X_{1,} X_{2,} \ldots X_n; \theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathrm{e}^{-X_i/\theta}$$

- Take log to get a sum

$$\ln(L(X_{1,} X_{2,} \ldots X_n; \theta)) = \sum_{i=1}^{n} \left(\ln\left(\frac{1}{\theta}\right) - X_i/\theta \right)$$

$$= n \ln\left(\frac{1}{\theta}\right) - \sum_{i=1}^{n} (X_i/\theta)$$

- Differentiate by θ

$$\frac{\partial}{\partial \theta}\left[n \ln\left(\frac{1}{\theta}\right) - \sum_{i=1}^{n}\left(X_i/\theta\right)\right] = -n\frac{1}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i$$

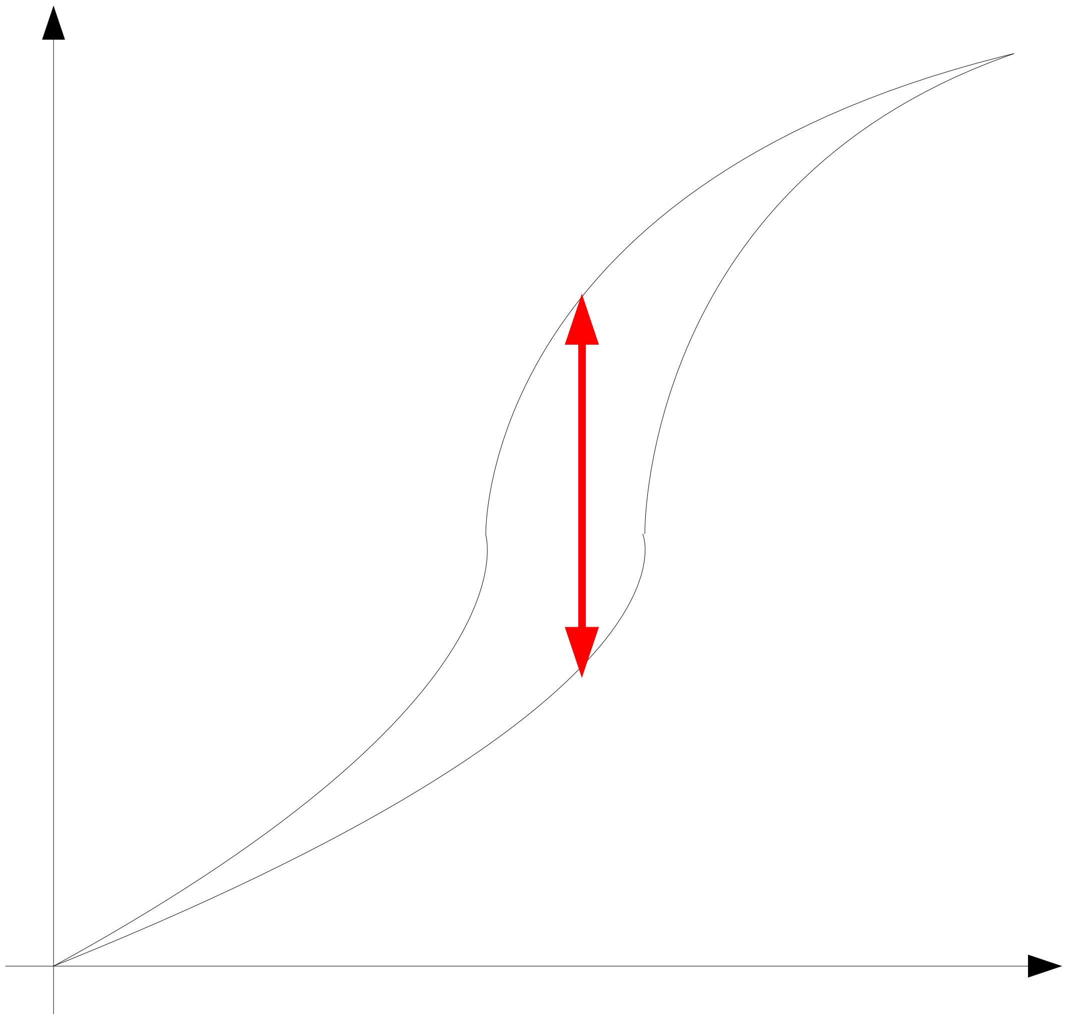
- Equate this to 0, giving

$$n\frac{1}{\theta} = \frac{1}{\theta^2}\sum_{i=1}^{n} X_i$$

$$\theta = \frac{1}{n}\sum_{i=1}^{n} X_i$$

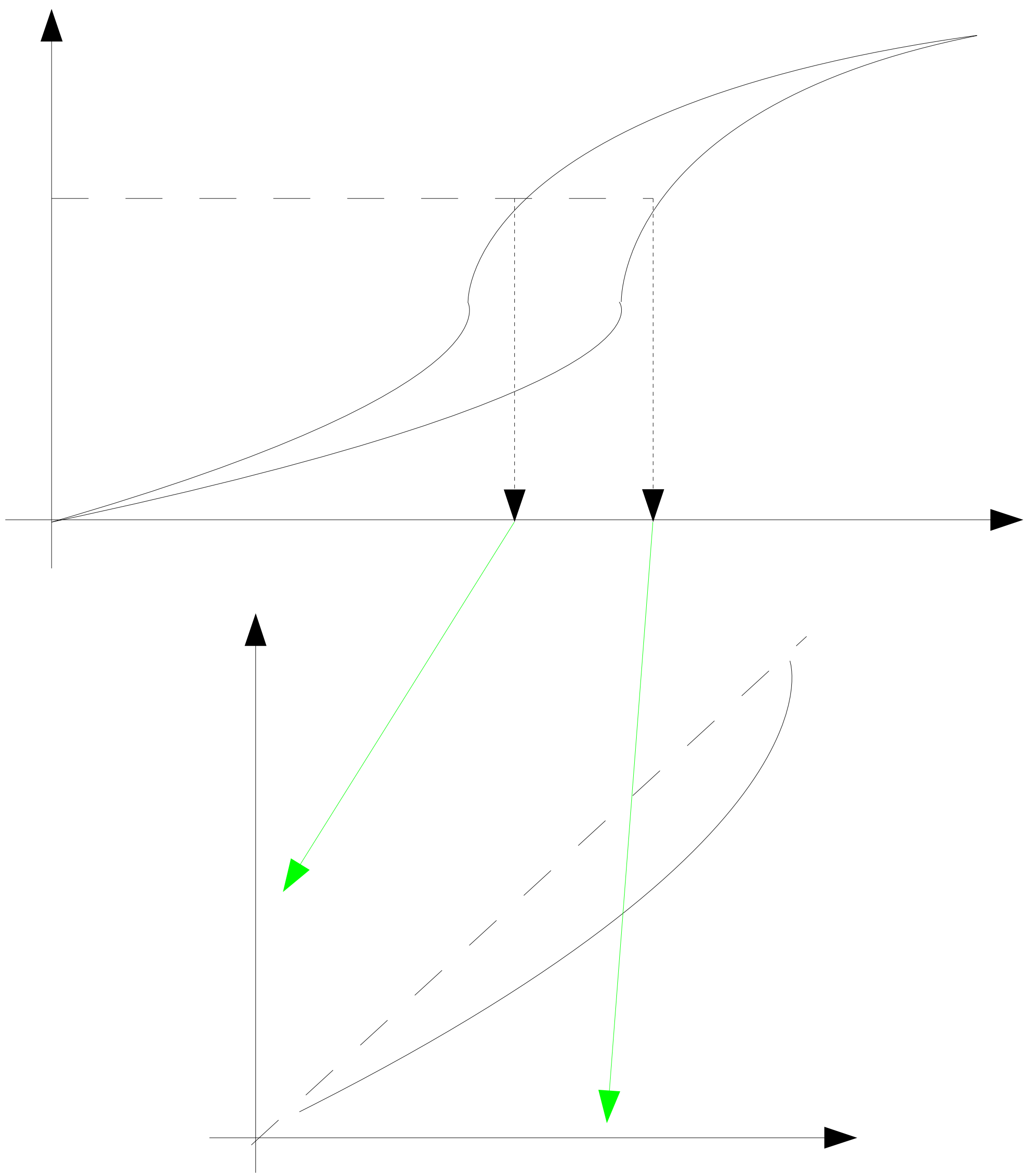
- Note: finding the best parameters for an assumed distribution does not necessarily imply a good fit!
- Check goodness of fit using statistical tests
  - chi-square
  - Kolmogorov-Smirnov
  - Anderson-Darling
- Or simple visual tests
  - Q-Q plot

Kolmogorov-Smirnov



- Metric is maximal vertical distance between model distributions
- Reflects maximal error of model relative to data
- Not sensitive to tails

Q-Q Plots



- For each quantile, find its value in the data and in the model, and plot
- Deviations from straight line indicate lack of fit
- More sensitive to tails

Generative Models

How does one choose a candidate distribution when fitting parameters?

- Try all of them and see which gives a good fit
- Select a distribution that matches your understanding of what is going on

Example 1: arrivals

- Assume arrivals occur at a constant rate (at each instant, there is an equal probability of a new arrival)
- And they are independent (arrivals at one instant have no relation to those at other instants)
- And they do not come in bursts (at each instant, there is at most one arrival)
- Then arrivals are a Poisson process
- And interarrival times are exponentially distributed

Example 2: file sizes

- Assume files are created as derivatives of existing files
  - editing a files produces a new file and leaves the old one as a backup
  - compiling a program file creates an executable file
- And the process is multiplicative (the new file size is derived by multiplying the original file size by a random number)
- And repetitive (file sizes are the result of many such multiplications)
- Then the distribution of file sizes is lognormal

Handling Censored Data

- When measuring the duration of an event, some events may not have finished yet
- This provides partial information: we don't know what the duration is, but we know it is longer than t
- This is called right-censored data
- Question is, how to incorporate it into the empirical distribution

Examples:

- Trying to find the distribution of session durations based on a log of sessions – ongoing sessions are censored
- Trying to find the distribution of process lifetimes from a trace of processes that ran on the system – processes that were killed are censored
- Trying to find how long users are willing to wait from the distribution of wait times – cases where the user received service are censored

Finding the empirical distribution function

- Let x_i denote a sampled value
- Let d_i denote the number of real samples with value x_i (excluding censored samples with this value)
- Let n_i denote all samples larger than or equal to x_i (both real and censored)
- Then the hazard at x_i (risk of surviving up to x_i and then dying) is d_i / n_i
- And the probability of surviving x_i is $1 - d_i / n_i$

Finding the empirical distribution function

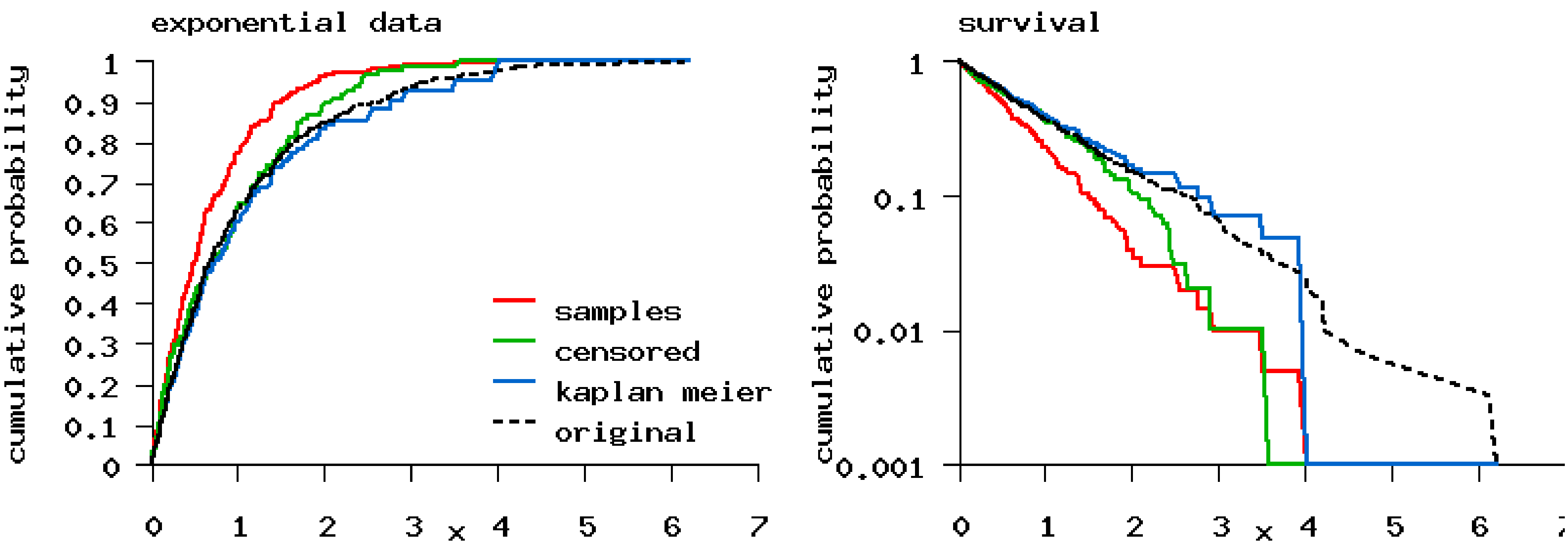
- The probability of surviving x_i is $1 - d_i / n_i$
- Then the probability of surviving an arbitrary x is the probability of surviving all smaller x_i:

$$Pr(X \geq x) = \prod_{x_i < x} (1 - d_i / n_i)$$

(This is the Kaplan-Meier formula)

- Note: this assumes censoring is random
  - population is homogeneous
  - long events have a higher probability of being censored

Example: samples from an exponential distribution, about half are censored



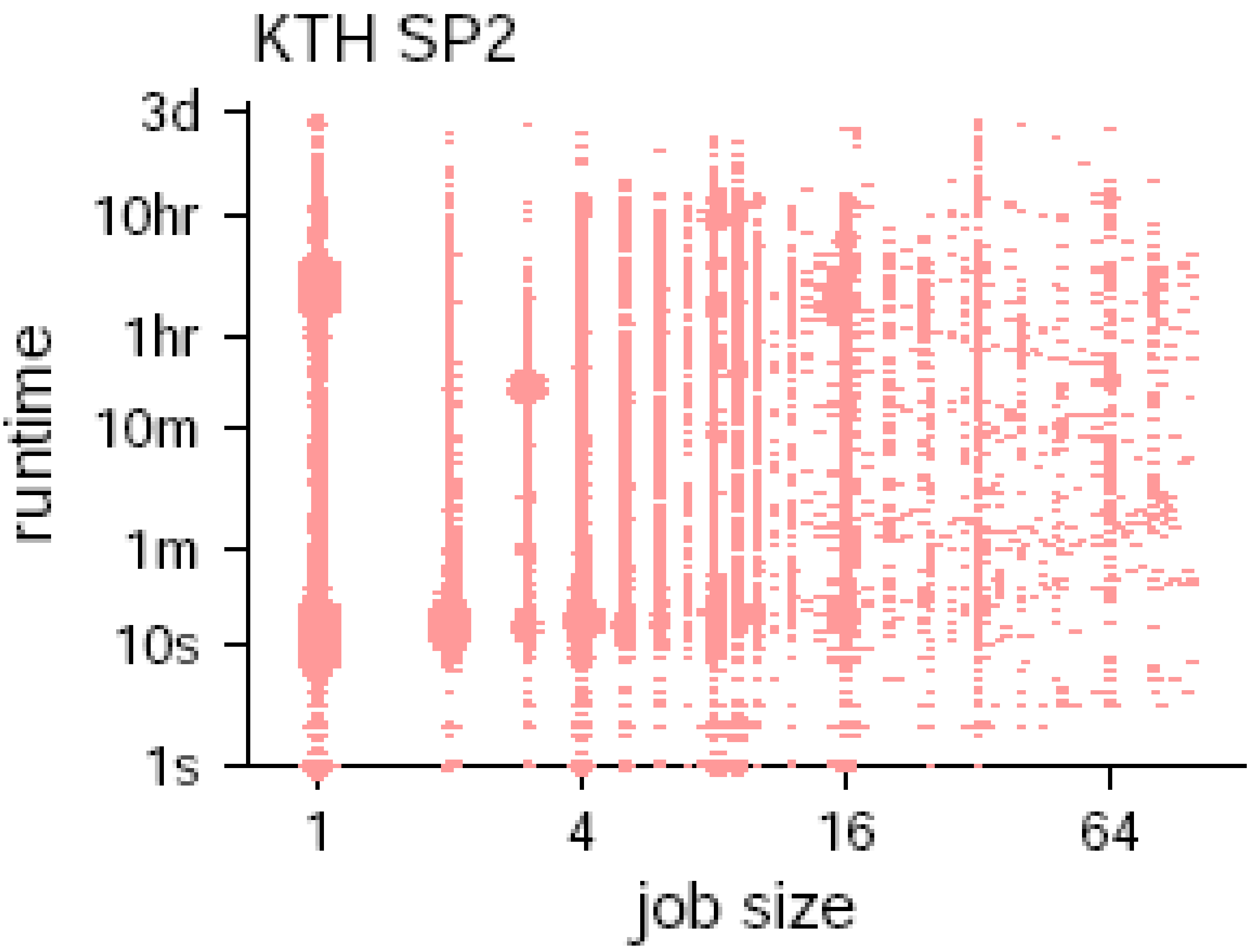
2D Data

- Data may come in 2 (or more) dimensions
- The question is then whether one may be used to predict the other
  - Do processes that use more memory run longer?
  - Do systems with more users also experience higher levels of activity?
  - Are short files accessed more often?

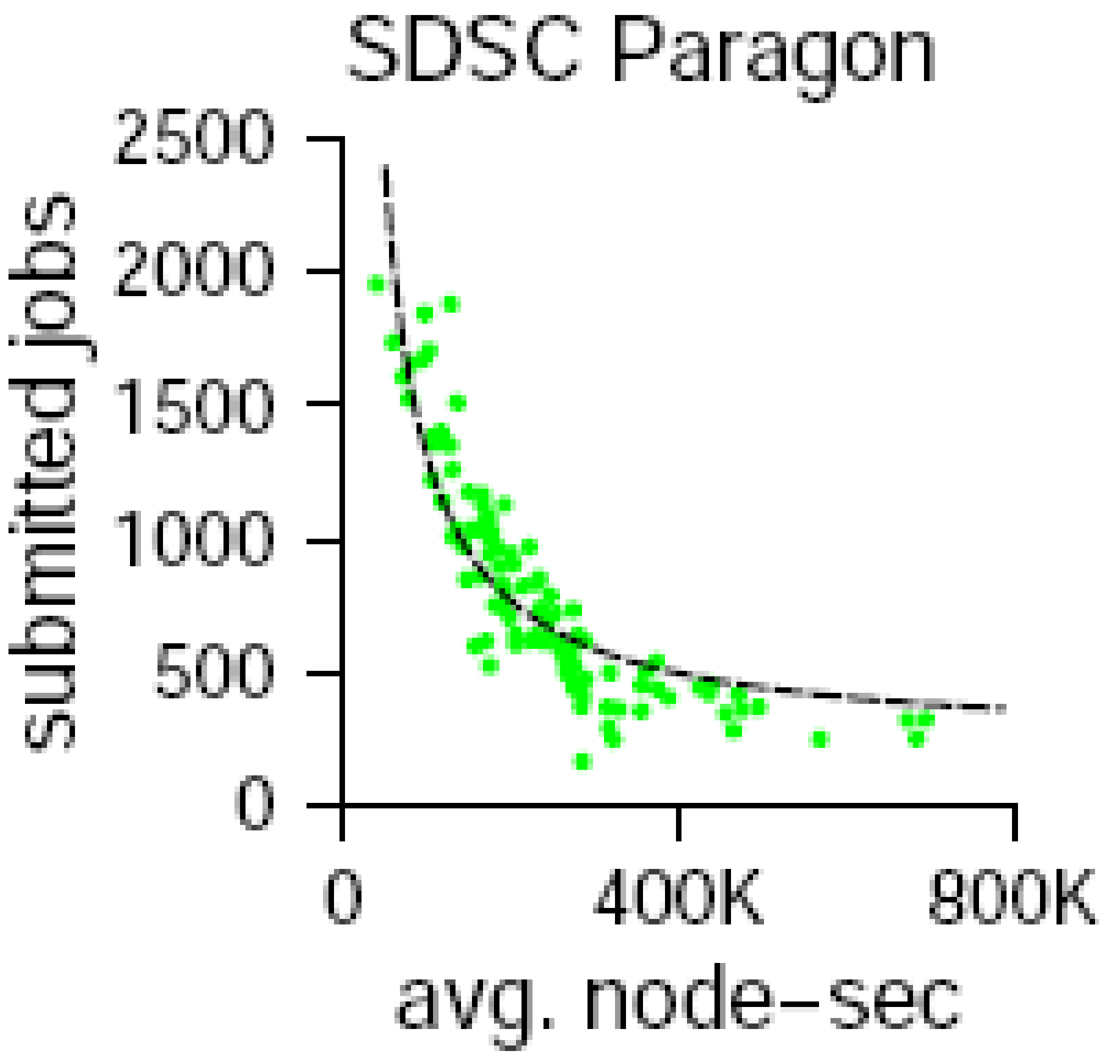
FIRST, LOOK AT THE DATA

- Draw a scatter plot
- Verify that you see a well-defined pattern
- Then try to find an equation that models it
  - may require a transformation
- Problematic when more than 2 dimensions are involved

- Runtime vs. size of parallel jobs
- No appreciable correlation



- Average job size vs. number of jobs in a week
- Looks like an inverse relationship



Regression

- Simplest model is a linear one:

$$Y = aX + b$$

- Note asymmetry between X and Y: X is given and used to predict Y
- The quality of the model is assessed by the quality of the predictions: given a data point (X,Y), how close is Y to $aX+b$?

Finding a and b

- Goal: minimize vertical distances between Y_i and corresponding prediction aX_i+b
- Method: differentiate and equate to 0

$$\frac{\partial}{\partial a}\left[\sum_{i=1}^{n}\left(Y_i-(aX_i+b)\right)^2\right]=0$$

$$\frac{\partial}{\partial b}\left[\sum_{i=1}^{n}\left(Y_i-(aX_i+b)\right)^2\right]=0$$

The solution (after some algebra):

$$a = \frac{\sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n \bar{X}^2}$$

$$b = \bar{Y} - a \bar{X}$$

Quantifying the quality of the regression

- If we didn't know anything, our best prediction would be that every Y_i is like the mean Y

  The variation is then $SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

- But given the model predictions, we can explain away much of this variation: it results from having different X_is

  The fraction of the variation thus explained is

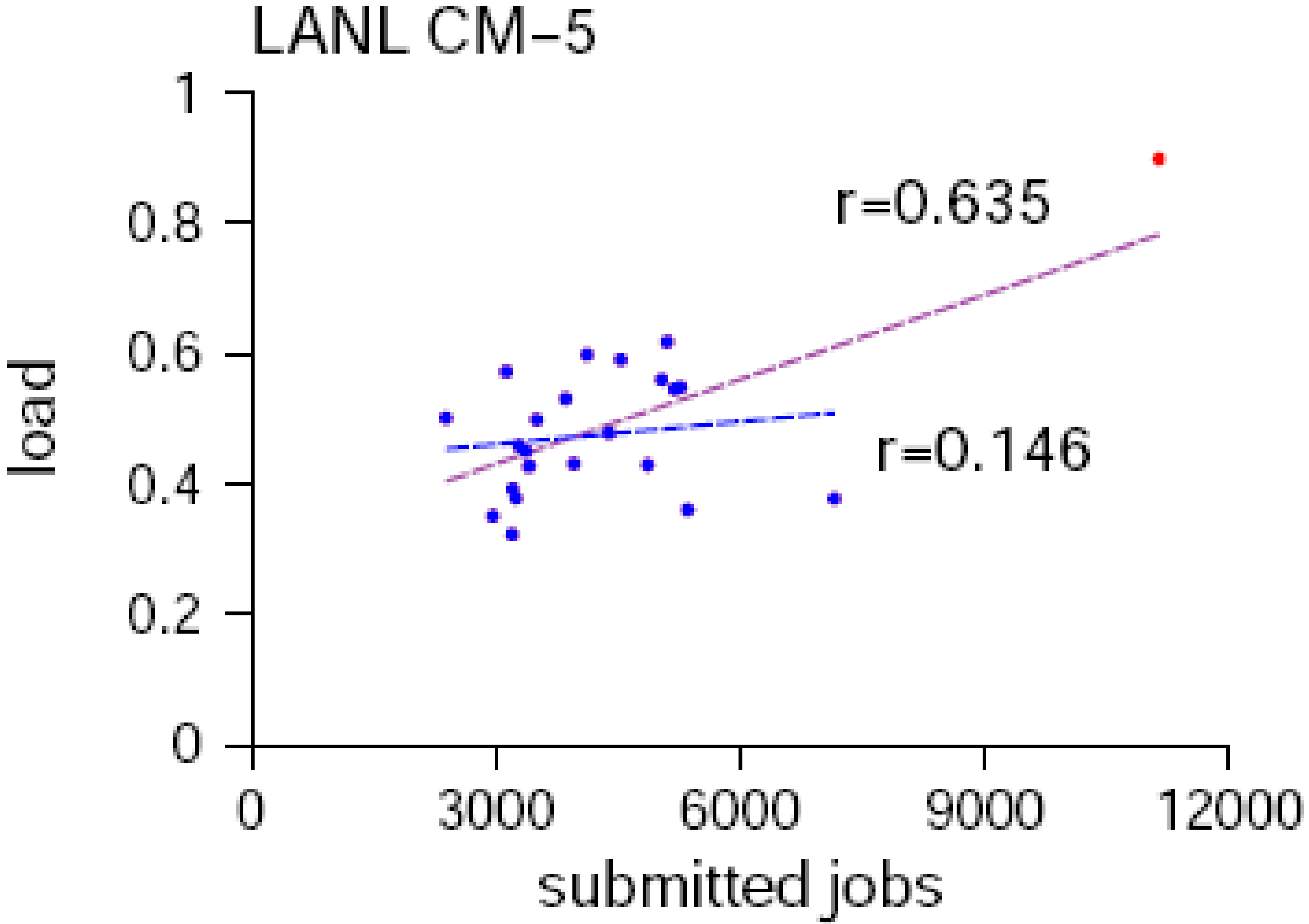
$$R^2 = \frac{\sum_{i=1}^{n} ((a X_i + b) - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

- If $R^2 \approx 1$ then the linear model explains most of the variation very well

Interpreting regression results

- $R^2 \approx 1$ implies that the relationship is near linear
  - as opposed to a diffuse cloud of points
- This is susceptible to strong effects by outliers
  - so does not necessarily look linear to humans
- It says nothing about the slope of the line
  - The slope is expressed by the *a* parameter

Example of the effect of an outlier



Not only for linear models:

- regression of Y with 1/X gives inverse relationship
- regression of Y with X^2 gives quadratic relationship
- regression of Y with $1\text{-}e^X$ gives exponential convergence