Harvard-MIT Division of Health Sciences and Technology
HST.512: Genomic Medicine
Prof. Alvin T.Kho

DANA-FARBER CANCER INSTITUTE

Children's Hospital Boston

Genomic Medicine HT 512

Data representation, transformation & modeling in genomics

Lecture 11, Mar 18, 2004

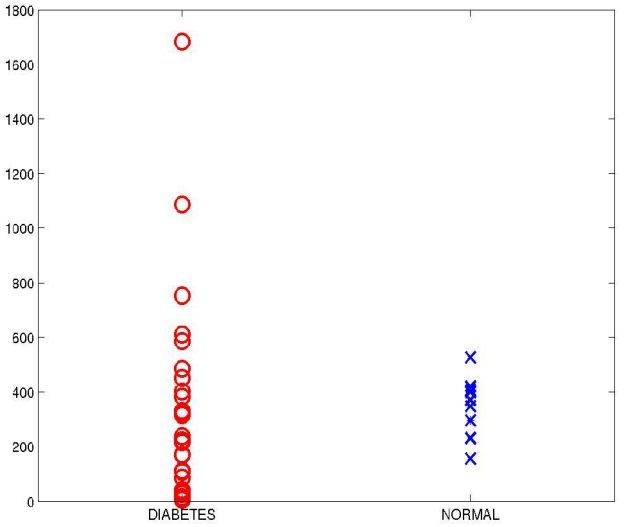
Alvin T. Kho
Children's Hospital Boston
Dana-Farber Cancer Institute

Lecture Outline

- 2 prototypical study design
  - 2-way comparison
  - Time series
- Data representation (DR)
  - What is DR ?
  - Measurement device to spreadsheet
  - Dimensionality - scales
  - Transformations / Changes-of-coordinates
- Background concepts
  - “Noise”
  - “Replicates”, reproducibility
  - Normalization
  - “Fold”
- Miscellany

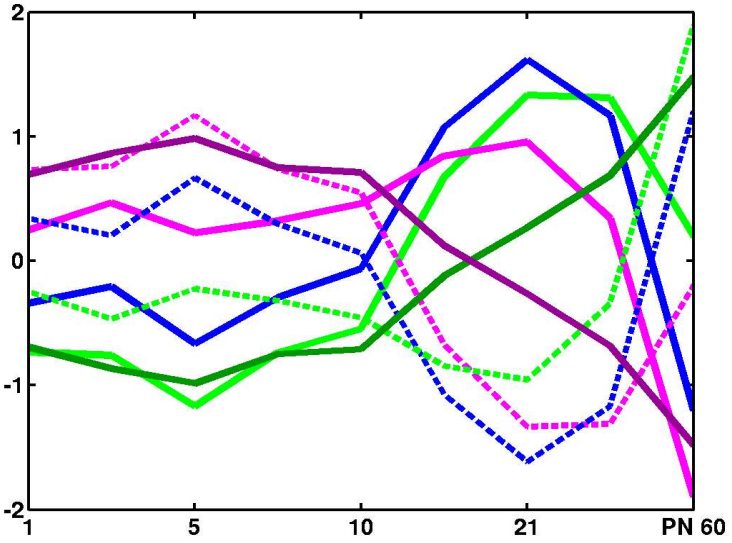
Prototype study 1: 2-way comparison

- Molecular differences in adipocytes of type-2 Diabetes vs Normal humans. 27 type-2 diabetes patients. 11 normal patients. ∝arrays.
- Pre-study reality check: Stratification – clinical phenotype.
- Partial math formulation: Let D_j = chip data of j-th Diabetic. N_j = j-th Normal. D_j , N_j are vectors/matrices whose dimensions depend upon # genes/parameters/variables measured per sample.
- Next steps ...



Prototype study 2: Time series

- RNA expression profile of a developing whole mouse pancreas at time points P1 to P60.
- Pre-study reality check: Stratification – histology, regulation.
- Partial math formulation: Let T_j = chip data of j-th developmental stage are vectors/matrices whose dimensions depend upon # genes/parameters measured per sample.
- Next steps ... patterns



What is DR?

- A mathematical reformulation of a scientific, real-world problem.
- Mapping observations and measurements to a set of symbols (typically, real numbers) that can be acted upon by an algebra.
- The form of representing the data values in the integrated dataset, including any numerical type conversion or re-classification that was performed. Usually this will involve only type conversions, although occasionally actual numerical changes may have been necessary. The units and precision are also indicated. [www.ngdc.noaa.gov]
- Something to do with database annotation and standards.
- Multimedia: Charts, graphs, plots.

Device to Spreadsheet: Know measuring device

- Understand general principles of measurement
  - Relevance of scanner / device settings
  - Phosphor imager mechanism
  - 1-/2-channel (competitive hybridization) ?
- Image to Device to Numbers / Spreadsheet
  - Internal image analytic software? Pre-processing?
  - “Spot” evaluation: Statistics, microscopic diffusion/thermodynamics
  - Units / dimensions of the device output

DR: Dimensionality/scales; Transformations

- Typical dimensionality/scale
  - 2-channel readout is fold/ratio = dimensionless
  - 1-channel = absolute or relative intensity units
- Importance of dimensionality/scale in large datasets
  - Different math techniques apply
  - Guides formulation of null hypothetic distribution / state
    - Gamma distributions for radiation measurements
    - Power/scaling laws for gross error detection, e.g., Zipf's
- Why perform DR transformations?

  1. Simplify mathematical manipulation.

  2. Reveal existing intrinsic "geometries" in the data.

DR: Why transform data 1

1. Simplify mathematical manipulation

- "Writing data/problem on paper, apply basic math rules"
- All spreadsheets are essentially matrix - subject to formal math operations/theorems (linear algebra).

	Exp 1	Exp 2	Exp 3		Exp M
Gene 1	$G_{1,1}$	$G_{1,2}$	$G_{1,3}$	...	$G_{1,M}$
Gene 2	$G_{2,1}$	$G_{2,2}$	$G_{2,3}$	...	$G_{2,M}$
Gene 3	$G_{3,1}$	$G_{3,2}$	$G_{3,3}$	...	$G_{3,M}$
:	:	:	:	:	:
Gene N	$G_{N,1}$	$G_{N,2}$	$G_{N,3}$	...	$G_{N,M}$

Warning: Heterogeneity of matrix entries. Blind application of math, e.g., adding apples and oranges.

DR: Why transform data 2

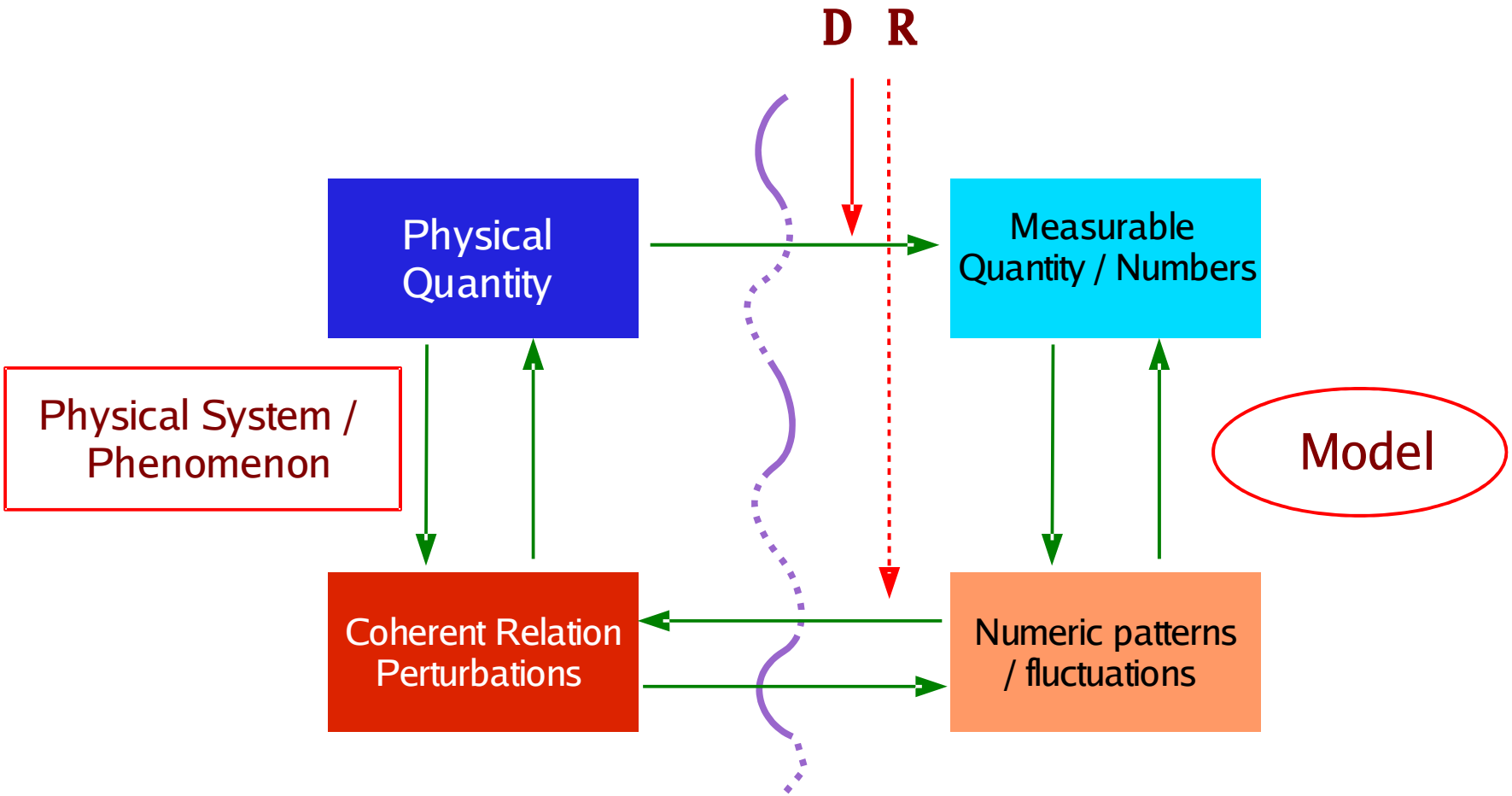
2. Reveal intrinsic "geometries" in the data

- **Q:** What's meant by intrinsic? "geometries"?
- **A:** Data = set of numbers that may contain internal orderings or structures, explicit or implicit
  - Explicit: *A priori* gene, sample labels / relations.
  - Implicit: Higher-order gene-sample relations.
- Clues to existence of relational structures:
  - Numeric changes, patterning,
  - that in a graphical/geometric setting may become more obvious/intuitive.
  - High dimensional space.

Warning: Know prior assumptions. Explicit & Implicit. No study / analysis is ever "hypothesis-free".

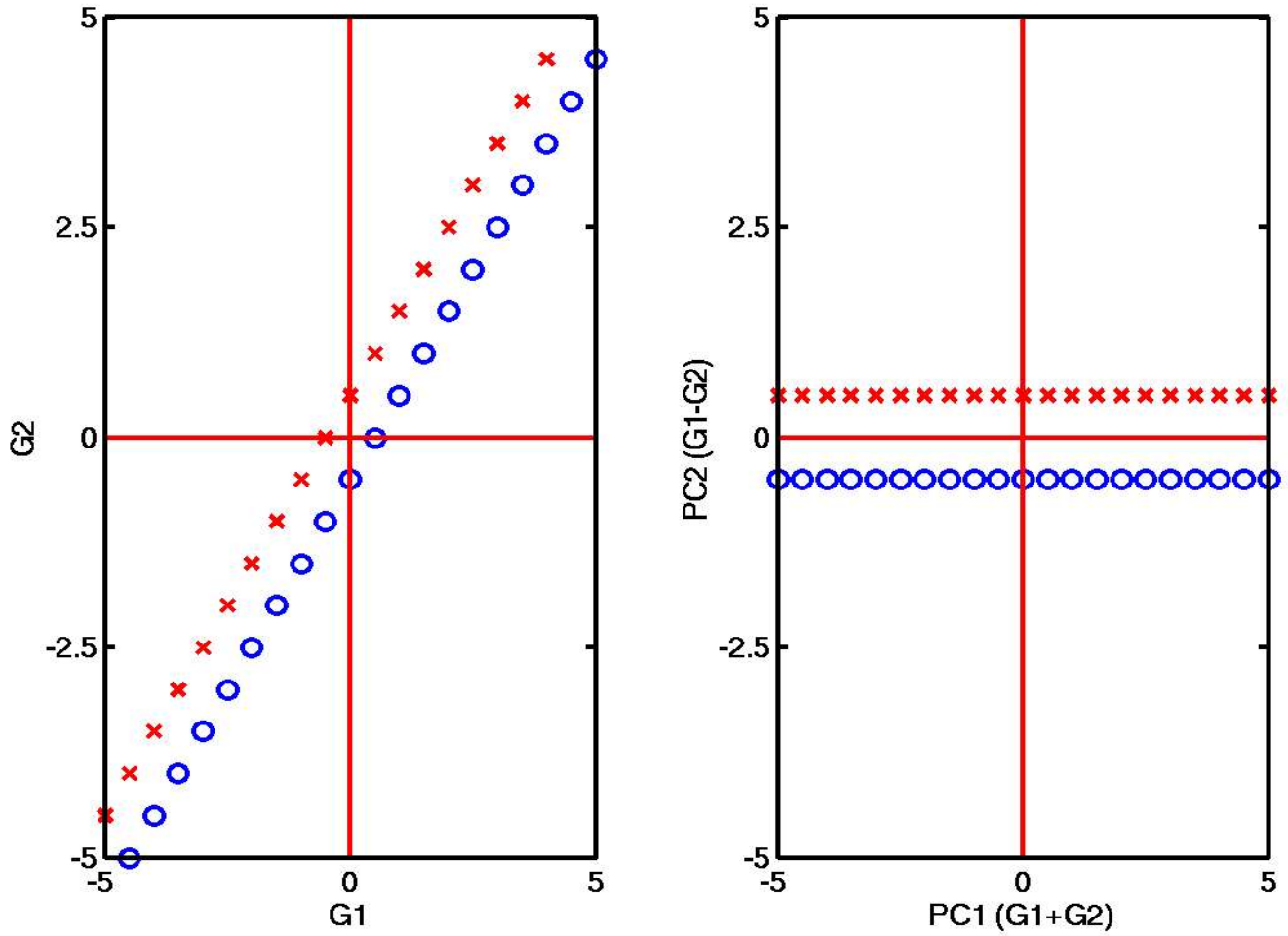
DR: Model idea

- Modeling summary diagram:



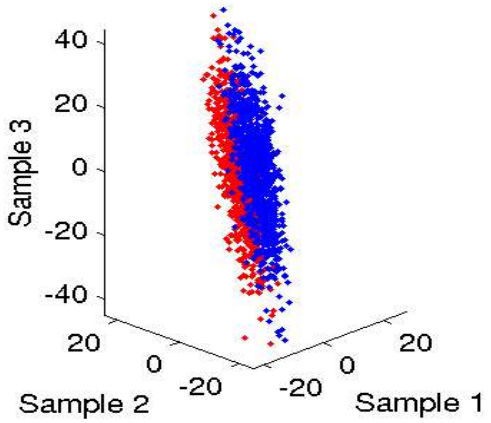
DR: Transformations example 0

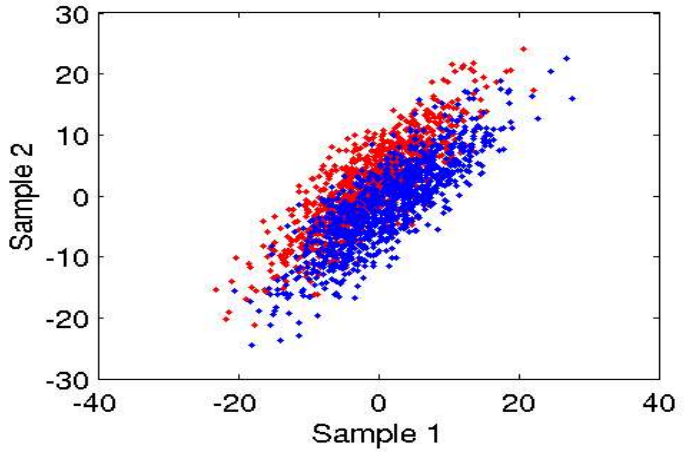
- Example 0: 2 diff patient populations **X** **O**. Each patient has 2 gene measurements: G1, G2. Principal component analysis (PCA) -> G1-G2 is discriminant

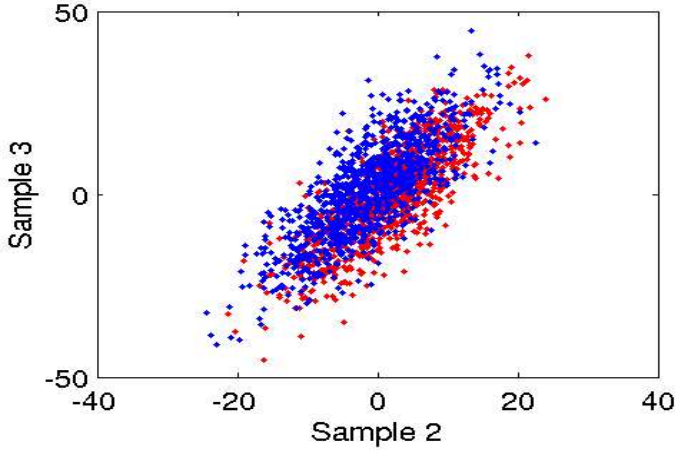


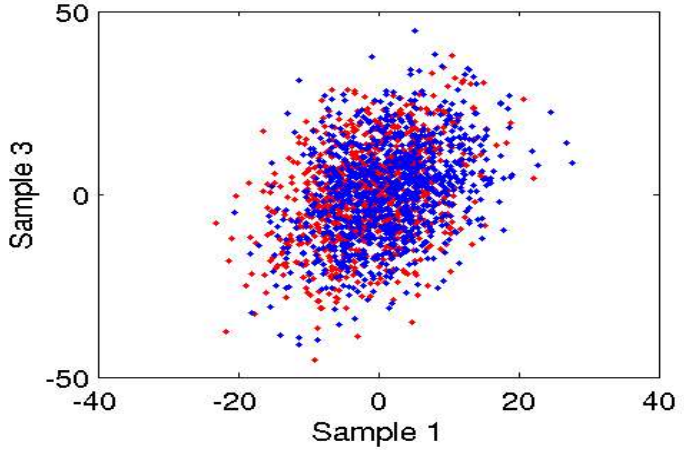
DR: Transformations example 1

- Example 1: Simulated 2 diff gene populations R, B in 3 sample conditions. 5000 genes in each population.



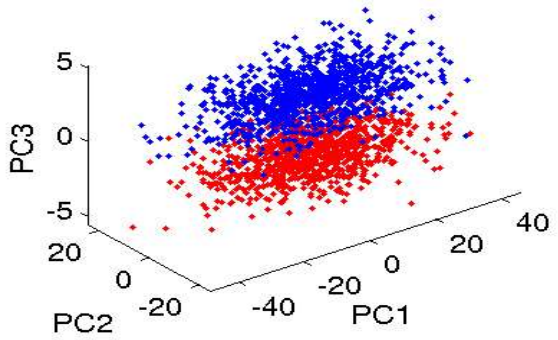


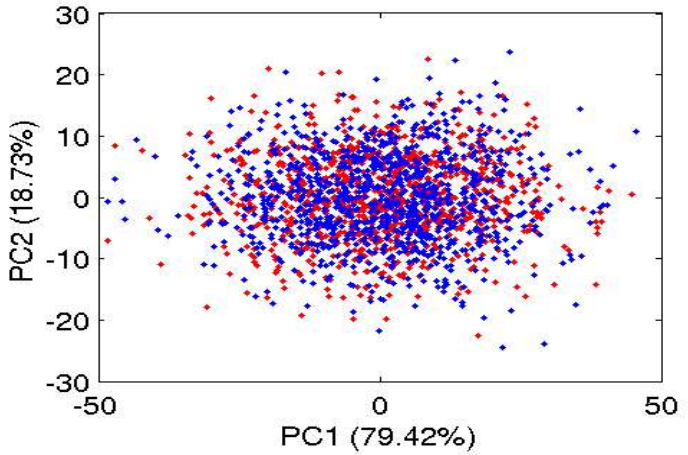


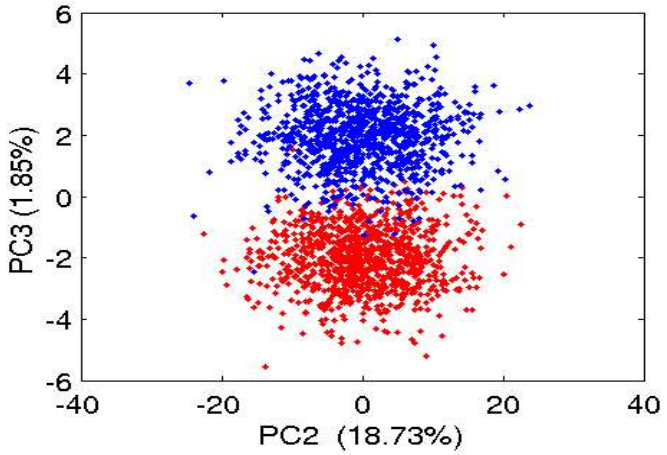


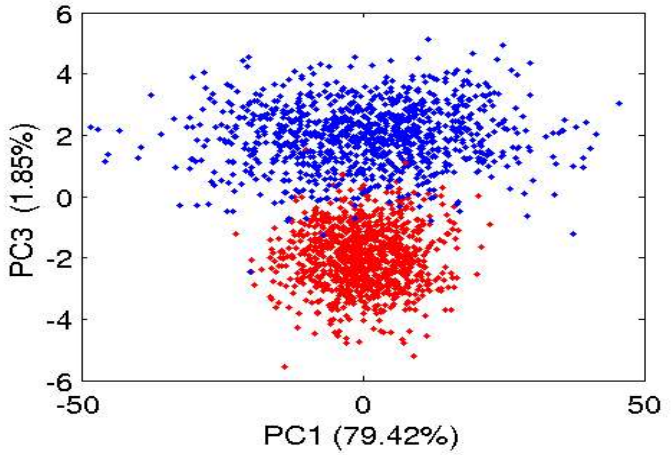
DR: Transformations example 1

- Example 1: R, B in new coordinates after PCA of condition axes.







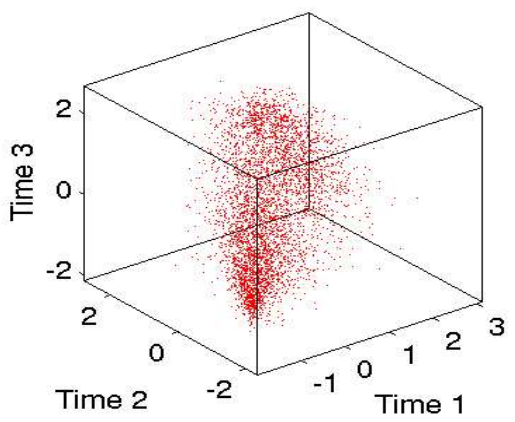


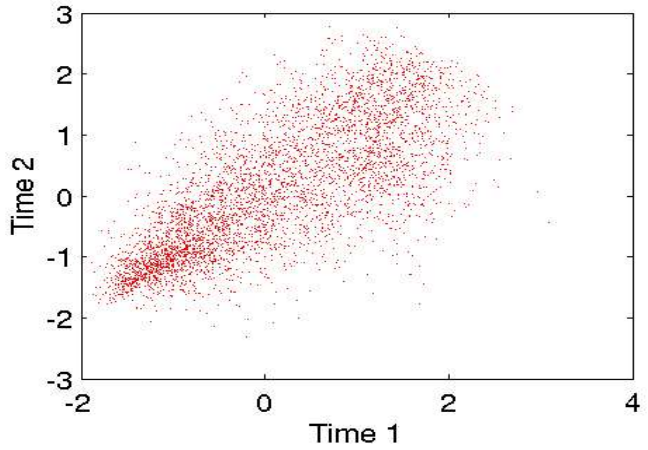
DR: Transformations example 2

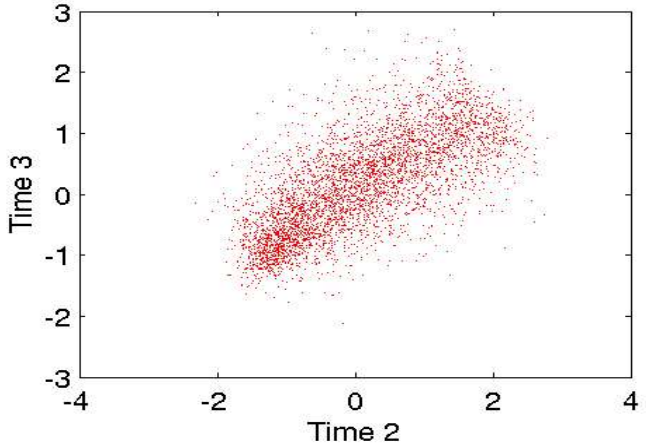
- Example 2:
  - Prototype study 2 pancreas development time series.
  - Principal component (PC) analysis = an affine change-of-coordinate. Similarity/distance in Euclidean/L^2 sense.
    - Sample-wise (as is): CLT-scaled genes
    - Sample-wise (PC): CLT-scaled genes
    - Gene-wise (as is): CLT-scaled samples
    - Gene-wise (PC): CLT-scaled samples
    - CLT ~ Central Limit theorem normalization -> mean 0, var 1. Later.

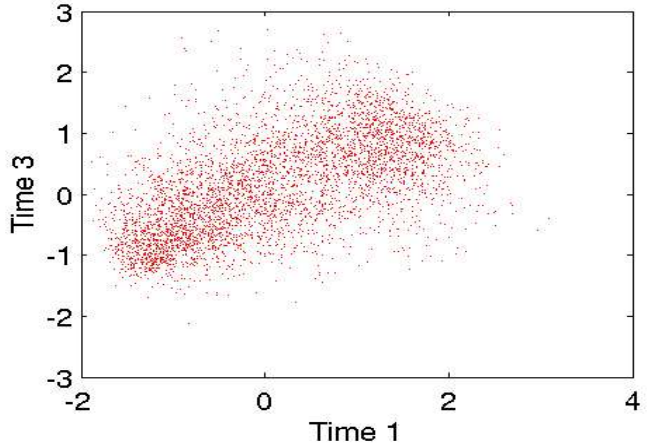
DR: Transformations example 2

- Sample-wise (as is): CLT-scaled genes



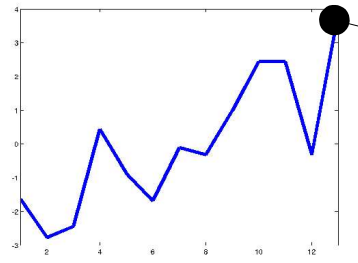




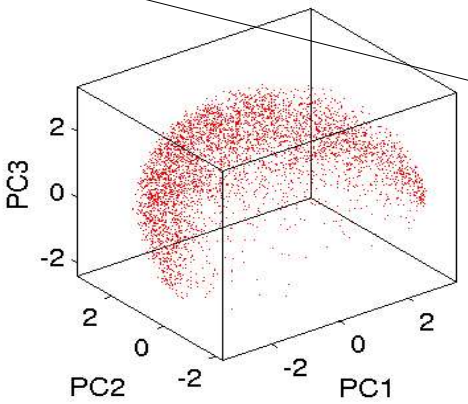


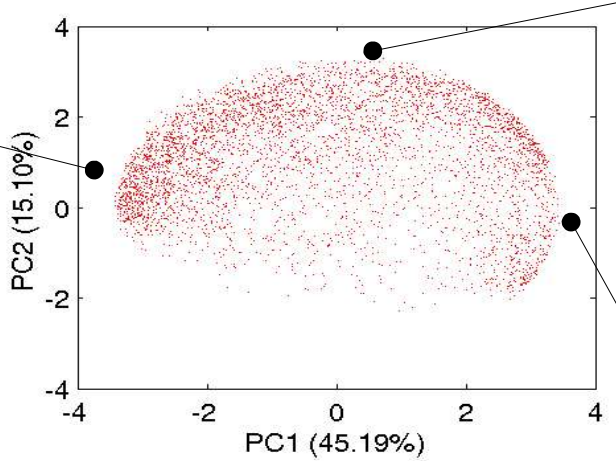
DR: Transformations example 2

- Sample-wise (PC): CLT-scaled genes

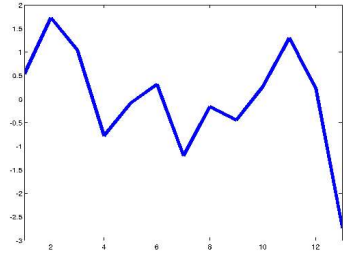


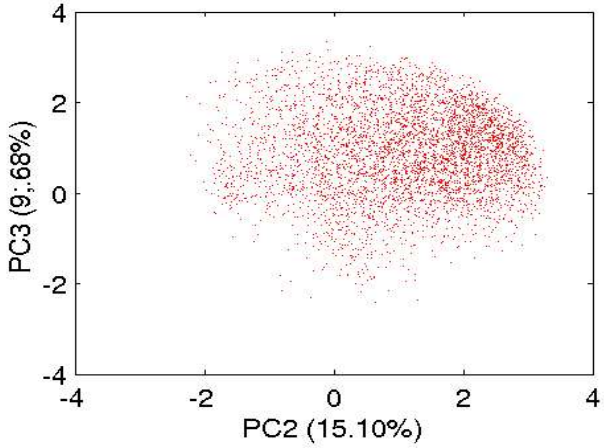
Profile

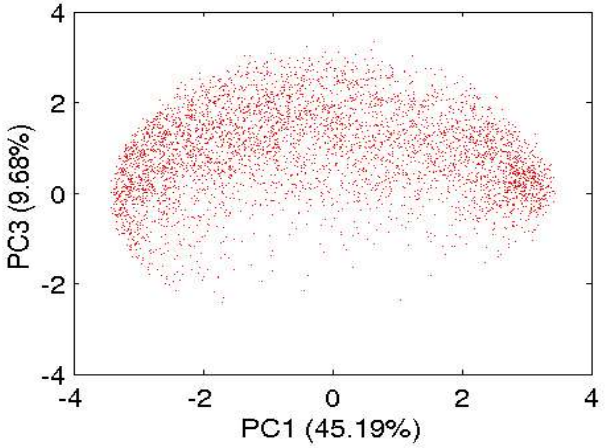


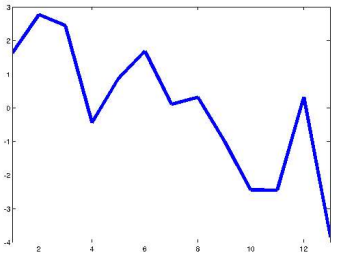


Profile





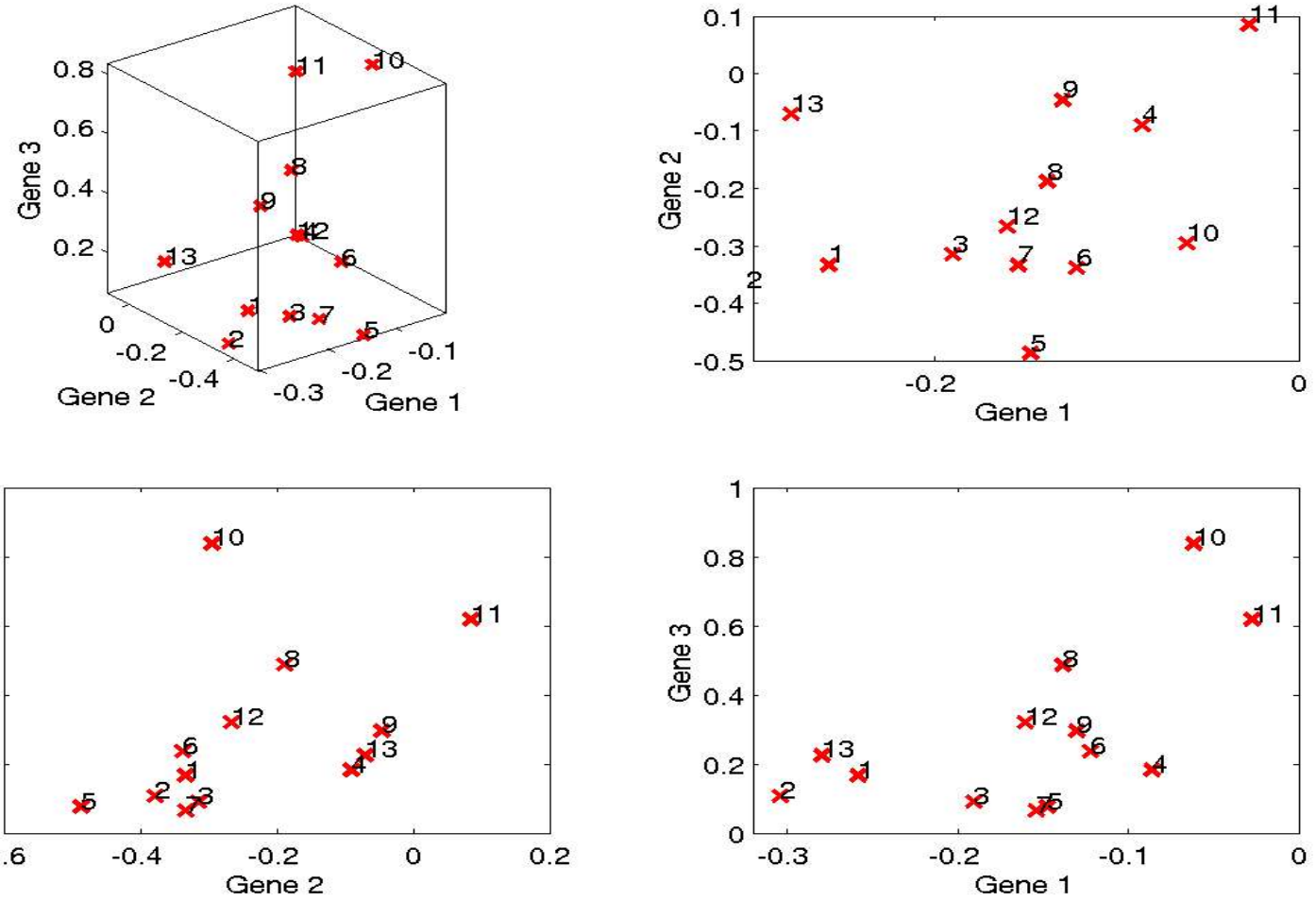




Profile

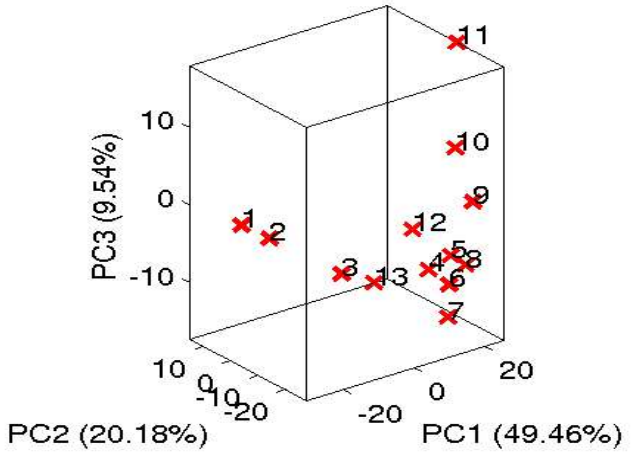
DR: Transformations example 2

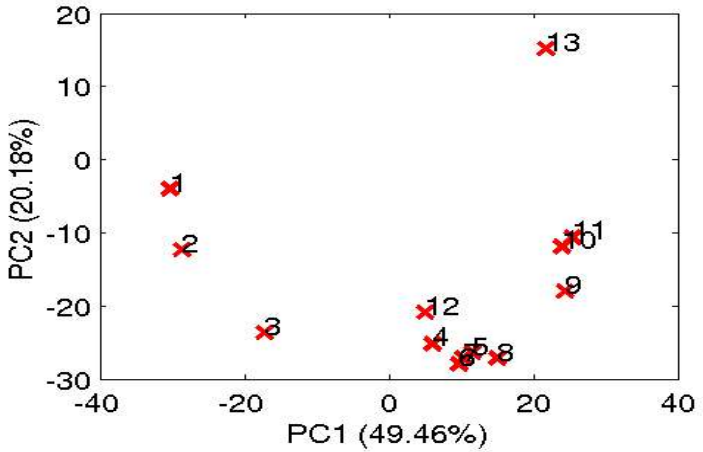
- Gene-wise (as is): CLT-scaled samples

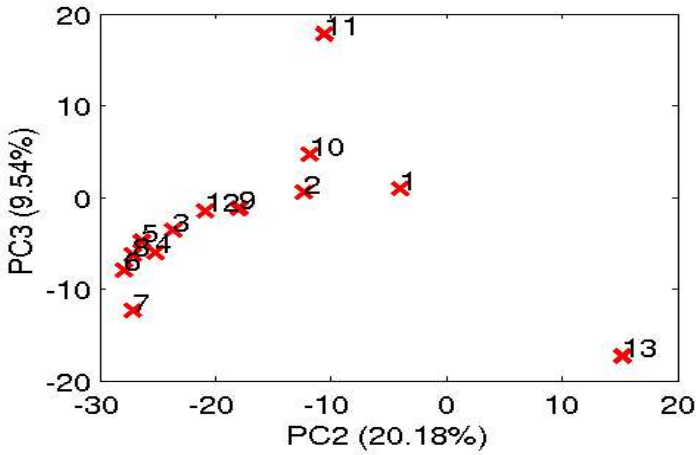


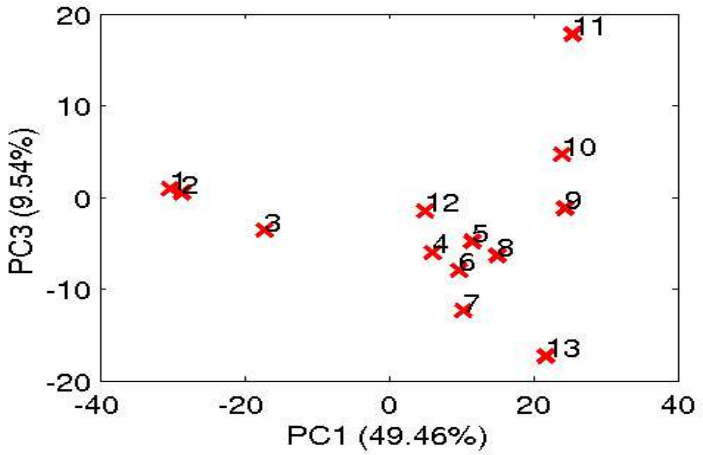
DR: Transformations example 2

- Gene-wise (PC): CLT-scaled samples





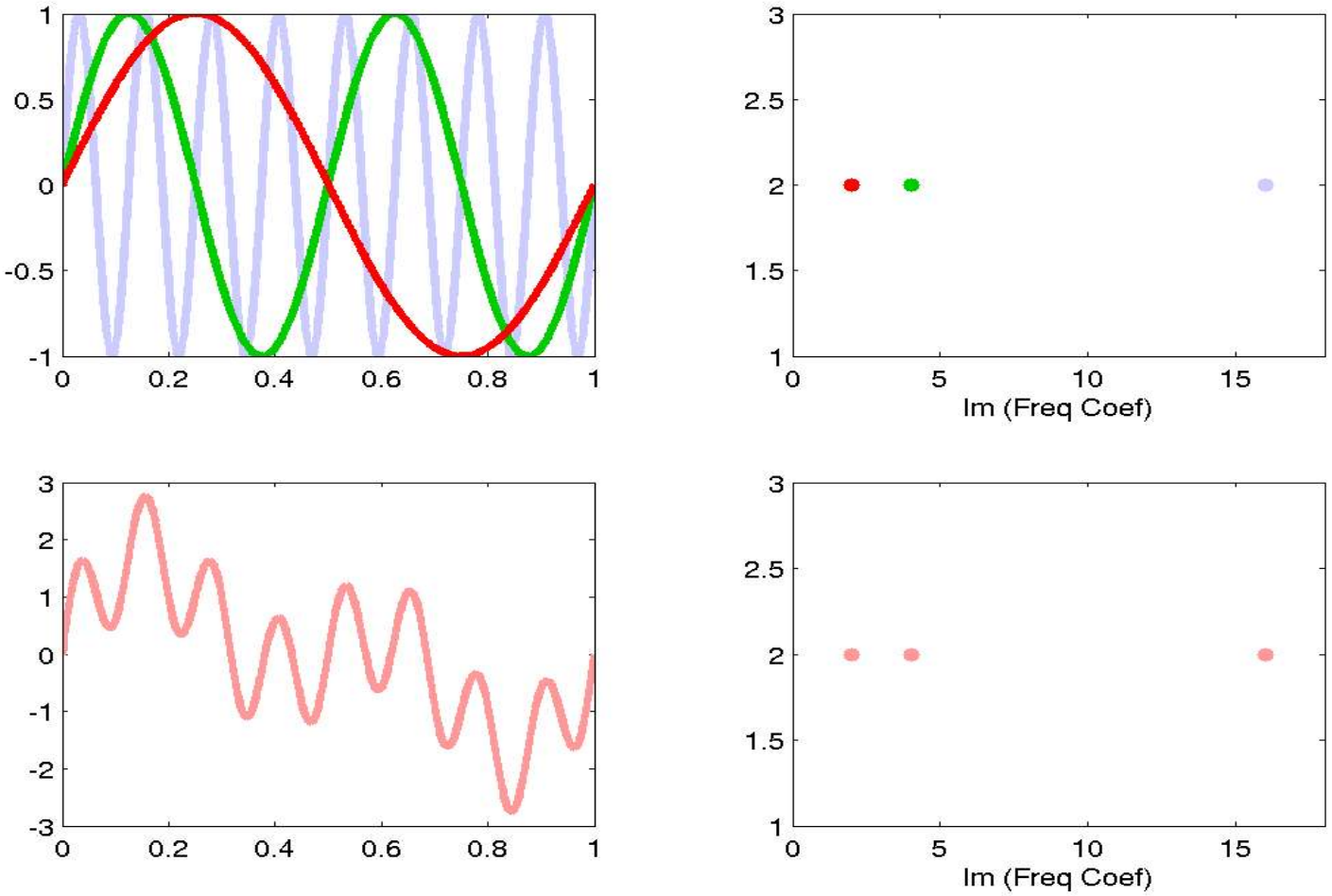




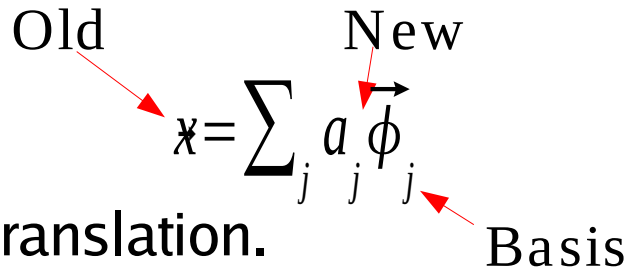
DR: Transformations example 3

- Example 3: Fourier decomposition. Individuals and sum of 3 sinosoids in freq space.

Application in sequence analysis: {A,T,C,G} -> {0,1,2,3}
-> Fourier



DR: Transformations summary



- Common transformations:
  - PCA (Euclidean) - finite bases. Rotation/translation.
  - Fourier (Euclidean) - infinite bases (localize “freq” domain). Signal decomposed into sinusoids. Periodic boundary conditions.
  - Wavelet (Euclidean) - infinite bases (localize “time” domain). Signal decomposed by discretized amplitudinal range.
- Different approaches emphasize different geometric/relational structures within the data. There is almost always a geometric interpretation.
- Secondary uses: Feature reduction, de-Noiseing, etc.

Noise/Reproducibility: What is "noise"?

- Axiom 1

  "Nature makes no leaps" (Tissot-Coke-Leibniz). Continuity of physical phenomena at a macroscopic level.

- Example 4

  Make 100 separate measurements of room temperature within a 1-min interval at different locations in the room. High likelihood that measurements are not all identical.

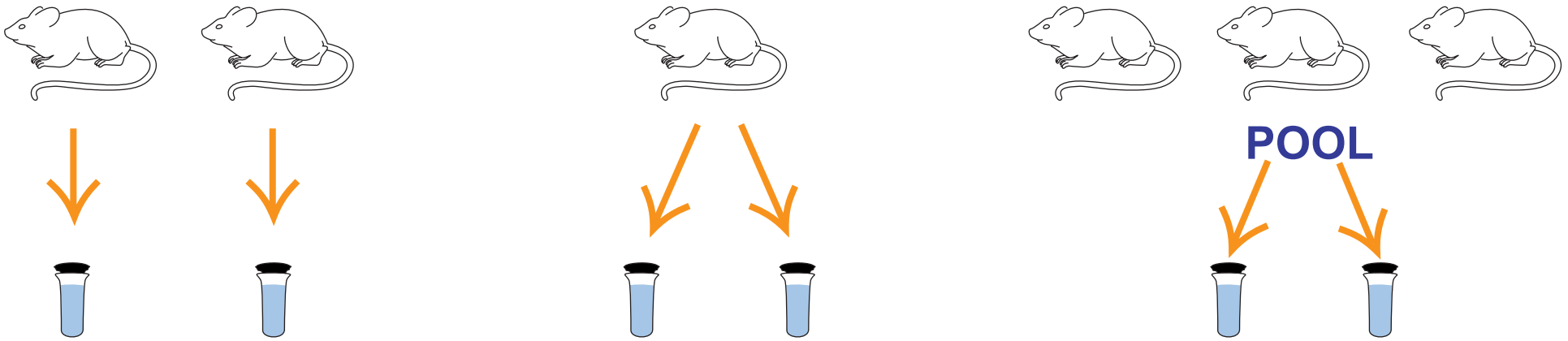
- Working definition of Noise

  In a narrow sense, noise is a/ny measurable divergence from Axiom 1, or more generally any applicable axiom, in a studied system.

- In ideal situations, math theorems apply: Central Limit, Large Numbers

Noise/Reproducibility: What is a "replicate"?

- What is a "replicate"? ... a repeated measurement?
- Need a reference system. Grades of being a replicate. 3 cases:
  - Separate RNA samples from pancreas of 2 ("normal") mice: Age, gender, weight.
  - RNA sample from 1 mouse pancreas, split into 2, aliquots.
  - RNA pooled from 3 ("normal") mice pancreas, and split into 2, aliquots.



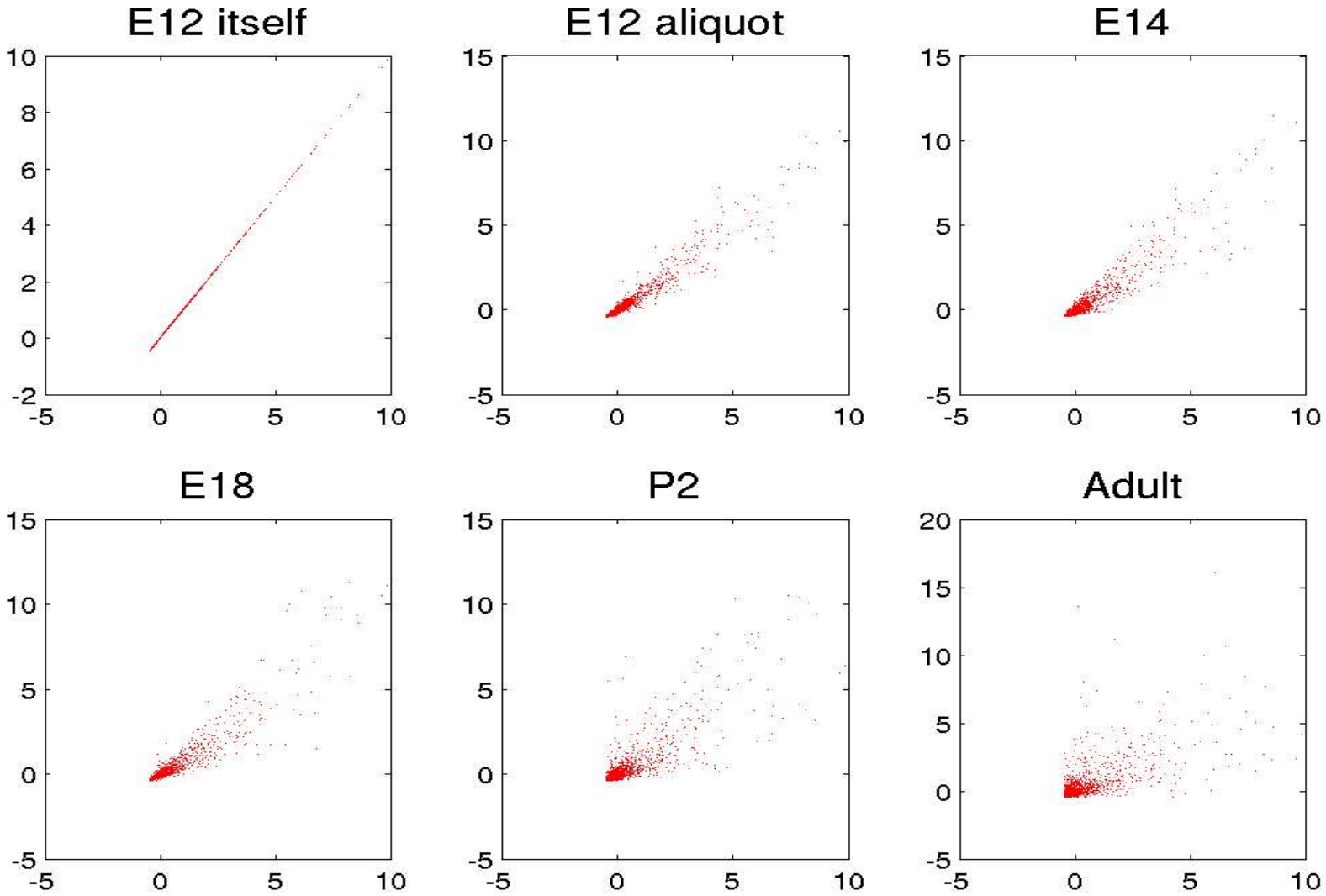
Noise/Reproducibility: Replicates & normalization

- Definition of replicate will have implication on noise definition. Biological versus measurement variation.

Warning: Over-restrictive definition of replicate may hinder generalizability of result to larger population.

- Despite taking all precautions, it is unlikely that replicate assays will be numerically identical. One never steps into the same river twice.
- Normalization: Informally, averaging out differences between replicate assays.

Noise/Reproducibility: Reproducibility example E12 versus ...

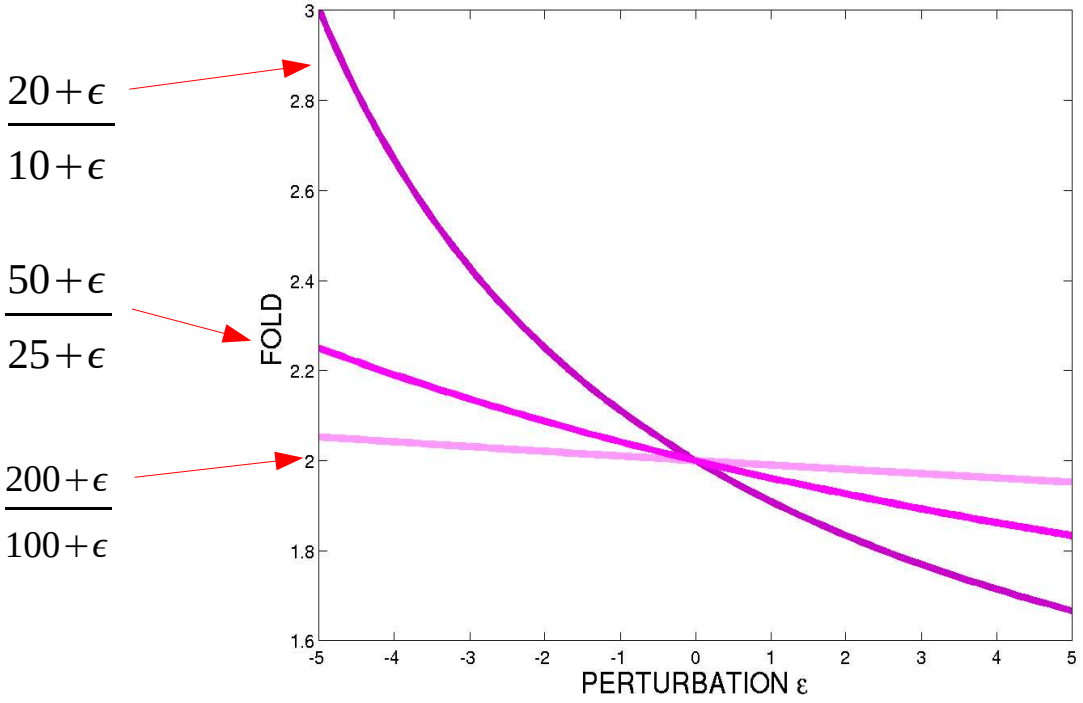


Noise/Reproducibility: Normalization

- To normalize or not ? *A priori* assumptions about how system behaves.
- Common normalization techniques, a vector x against reference r
  - "CLT" (central limit theorem) scaling: x -> (x – mean(x))/std(x)
    - What happens in normalized data/vector?
  - Linear regression: $x \rightarrow (x - a_0)/a_1$
    where a_0 is the y-intercept, and a_1 the slope of the regression of x against reference r.
    - What happens in normalized data/vector?

Fold

- Fold: While conventional in PCR/blots, the fold may not make sense in a 1-channel microarray setting. How to calculate? Limits?
- E.g., fold A = {-21.2, 14.9, -3.7} vs. B = {541.3, 596.6, 551.1}. Fold = ArithmeticAverage(A)/ArithmeticAverage(B); or GeometricAverage (A)/GeometricAverage(B) ?



Outline: Recall

- 2 prototypical study design
  - 2-way comparison
  - Time series
- Data representation (DR)
  - What is DR ?
  - Measurement device to spreadsheet
  - Dimensionality - scales
  - Transformations / Changes-of-coordinates
- Background
  - “Noise”
  - “Replicates”, reproducibility
  - Normalization
  - “Fold”
- Miscellany

A miscellany

- Our discussion so far makes nominal reference to biology. Approaches are general & apply in almost any setting. Math only provides the tools to biological discovery.
- Key point: The underlying biology is the point (at least our understanding of it). Lose this and the whole enterprise becomes purely technology/method driven. Biology dictates
  - Experiment design
  - Appropriate measure/similarity space to formulate the dual mathematical problem - representation, modeling.
  - Reading and making sense of the model outcome. Corroboration with observed/measured phenomenon?
- No study is “hypothesis free”. Know your prior assumptions.

References/Epilogue

Pattern Classification. 2nd ed. Duda, Hart & Stork. Wiley-Interscience, 2002 http://rii.ricoh.com/~stork/DHS.html

Applied Multivariate Statistical Analysis. Johnson & Wichern. Prentic-Hall, 1988.

The discoveries that one can make with the microscope amount to very little, for one sees with the mind's eye and without the microscope, the real existence of all these little beings.

George-Louis Leclerc
Comte de Buffon, 1707-1788