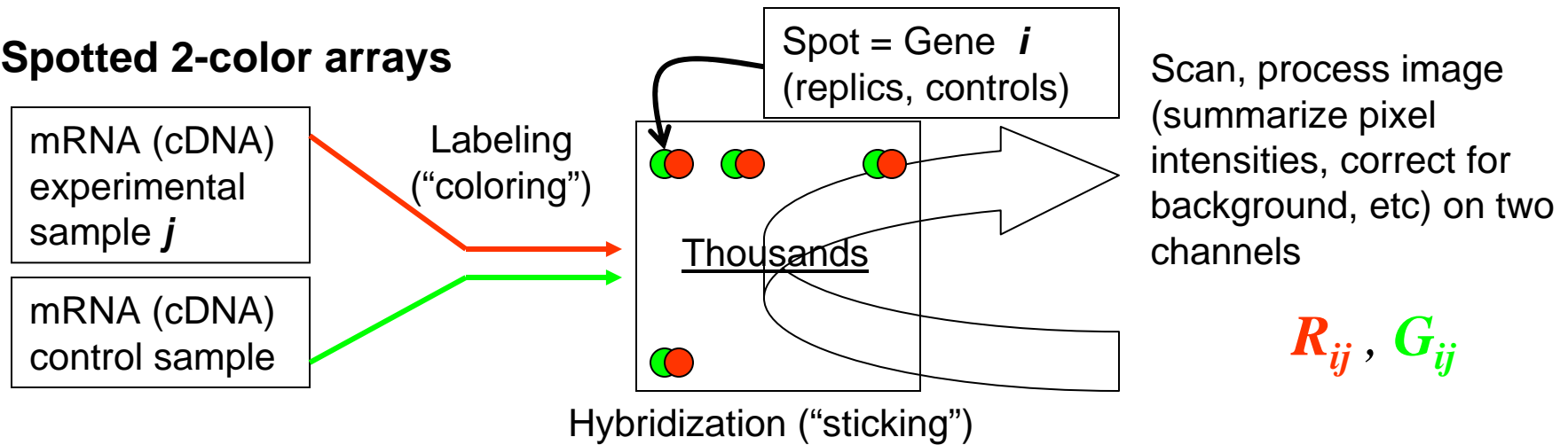
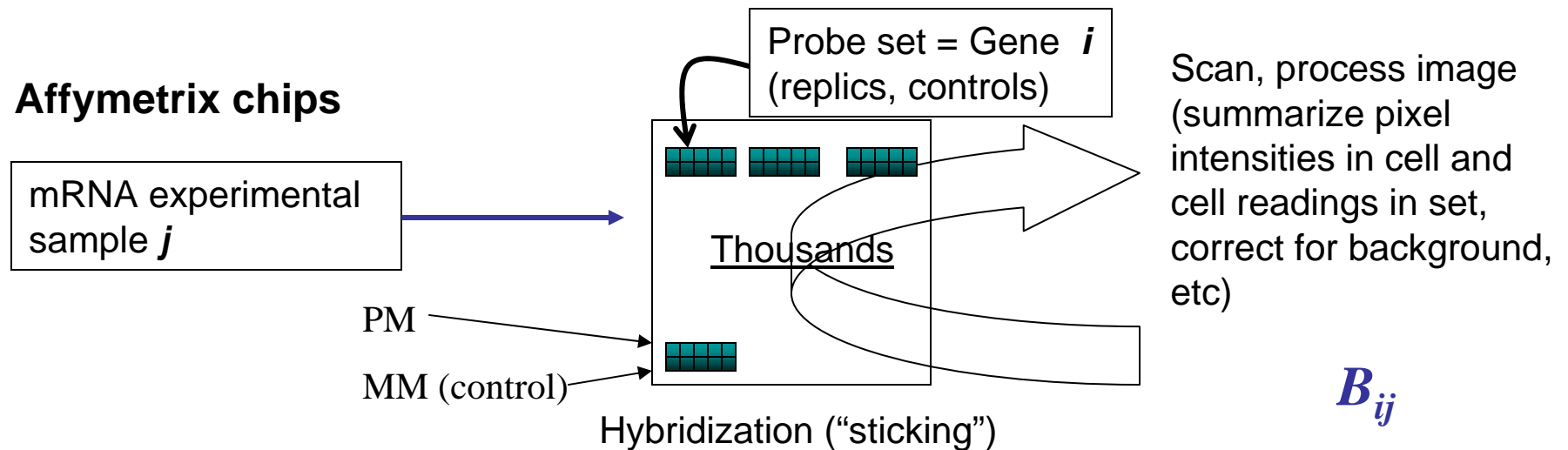


Summary Scheme

Spotted 2-color arrays

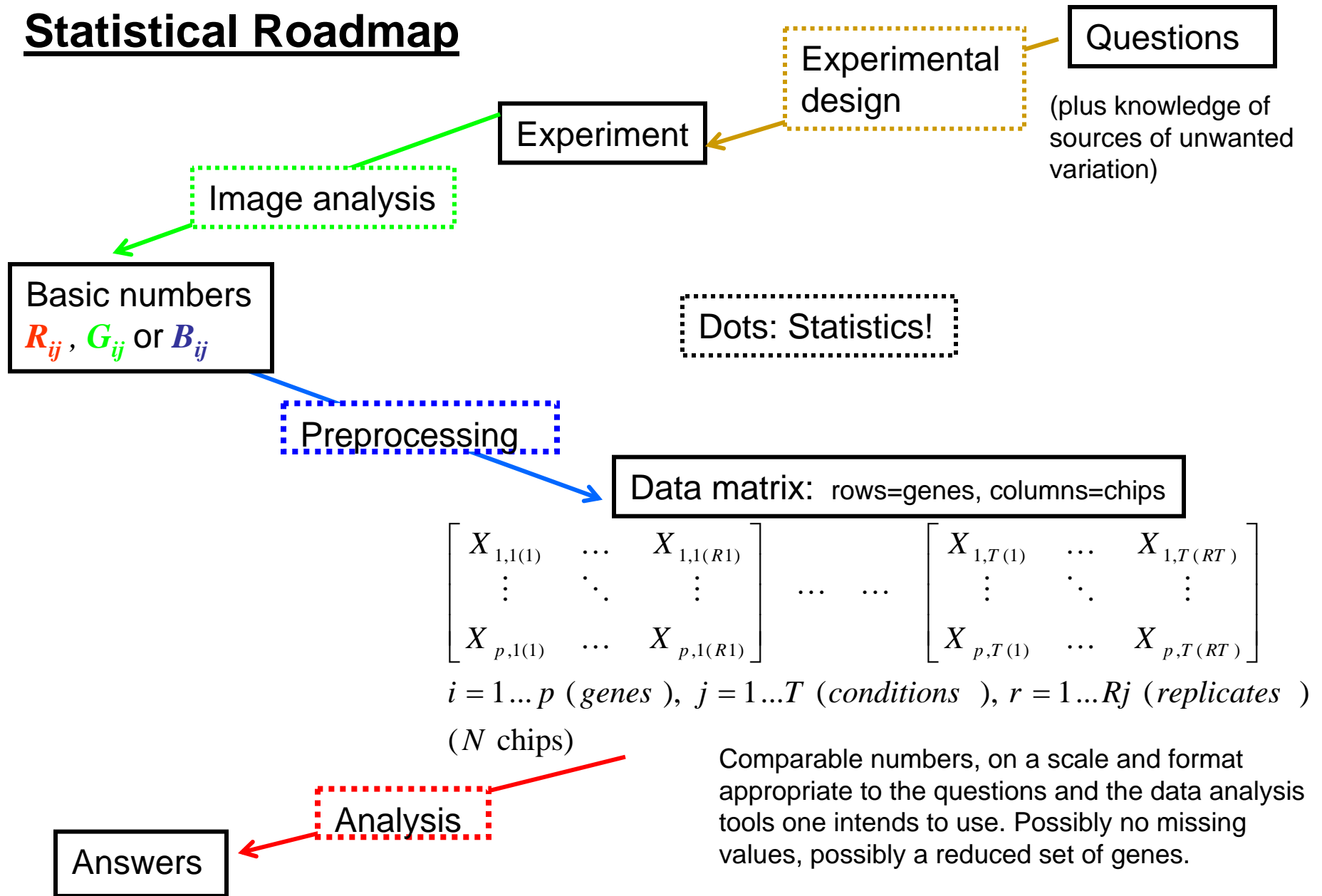


Affymetrix chips



j ranges over conditions (e.g. time points, treatments, types) and possibly replicates for each

Statistical Roadmap



Preprocessing the data:

- Normalization
- Imputing missing values
- Other preprocessing steps

Ensure comparability of different sets of expression measurements, and mitigate effects of unwanted variation. Sources, control with Design.

1. Further improve comparability of measurements across experimental conditions and/or across genes, e.g. centering and standardization.
2. Further decrease the effect of unwanted sources of variation by reducing noise and systematic errors, e.g. low-dimensional reconstructions.
3. Eliminate “inert bulk” that can affect detection of interesting signals by standard methods; filtering out genes with negligible expression in all conditions (e.g. absent calls in affy), or negligible variation across them.

Analyzing the data:

- Identifying differentially expressed genes
Computational approaches to [significance assessment](#)
Multiple comparisons
 - Investigating aggregate features of global expression data (constitutive or typical expression patterns, groups of genes with similar profiles, groups of conditions with similar signatures)
Dimension Reduction
Clustering (unsupervised classification)
 - Investigating expression correlates of sample classifications (cancer types, tissues), or quantitative responses (relapse times, levels of chemicals)
Classification (supervised) & Regression
 - Integrating expression information with other information; genomic sequences (e.g. transcription factor binding sites), genomic alignments (e.g. between species conservation), functional annotation (e.g. Gene Ontology terms), proteomics (e.g. protein-protein interactions)
 - Investigating gene networks
- (many others)