



A Handbook for Portfolio Analysis

Paul Sheehy
Division of Extramural Activities, NIGMS
sheehyp@nigms.nih.gov
301.594.4499

OPA/DPCPSI Portfolio Analysis Workshop
Natcher Conference Center, NIH
Bethesda, MD
6 February 2012



Background

Why Perform Portfolio Analysis?

Science to be efficiently and accountably administered must be understood.

Nature to be commanded must be obeyed.

Francis Bacon, *Novum Organum*

Why Use Automation for Portfolio Analysis?

The gold standard is manual curation. Software may (will) miss nuances of context. However IT provides:

- **Independence** – no inter rater variability
- **Replication** – no intra rater variability
- **Transparency** – algorithm can be shared
- **Tunable** – usually quantitative parameters that can be adjusted
- **Scaling** – can be applied to vast numbers of documents and variables

These establish **credibility**.

“It is hard to believe that a man is telling the truth when you know that you would lie if you were in his place.” (Mencken)

What is Portfolio Analysis?

- A specific instance of the general activity of Data (or Text) Mining which is used in
 - market analysis,
 - scientific and engineering process analysis,
 - bioinformatics,
 - homeland security
- Most familiar usage of Portfolio Analysis is in the financial sector:
 - An analysis of elements of a company's product mix to determine the optimum allocation of its resources.

What is a Portfolio?

- Aspects of the Enterprise that are linked by a theme
 - Note that this does not mean that the objects themselves are similar but it does imply the inclusion of the associated properties of those objects (*e.g.* car or house and price)
- Portfolio definitions: “Lumpers” or “Splitters”
 - Specific programmatic or process focus (*e.g.* PCC, translational)
 - Summary (*e.g.* Disease, technology, organizations)
 - Scale and approach vary accordingly
- Objects are associated with other data (*e.g.* geospatial data, collaborators, publications)

- Association, correlation, and frequent pattern analysis.
- Classification.
- Cluster and Outlier Analysis.
- Time-Series and Sequence Data.

- Reporting
 - Descriptive or comparative analysis
 - Content (subject matter) vs. feature (PI characteristics, score)
 - Formalized (GPRA, Congressional) or *ad hoc*
- Management
 - Optimal allocation of resources
- Planning
 - Identification of gaps and overlaps
- Evaluation
 - Limited by lack of definition of “success” or “effectiveness” and validated surrogates/indices

Fundamentals

Things to think about before you start

- Database and Data Management Issues
- Data Preprocessing
- Choice of Model and Statistical Inference
- Metrics
- Analysis
- Visualization and Understandability
- Model

Database and Data Management Issues

- Where do the data reside?
- How is it to be accessed?
- What forms of sampling are needed? are possible? are appropriate?
- What are the implications of the database or data warehouse structure and constraints on data movement and data preparation?

Data Preprocessing:

- Are data transformations required for the chosen analytic approach?
- Should the data dimensionality be reduced to improve algorithm efficiency? What methods are available?
- What about missing data?
- What transformations properly encode *a priori* knowledge of the problem?