Overview

• **Active Data Canvas** presents big data in familiar interactive contexts to promote data exploration.

• **Active Data Canvas** facilitates fluid, versioned, collaborative analysis.

Introduction

• Requiring computational acumen limits participation in data exploration.

• Visual analytic tools foster interaction, and greatly improve the speed and quality of analysis.

• Current tools lack the ability to link data to external knowledge-bases and to effectively share analysis.

Interactive Data Views

• Users visually browse data.

• Heatmap view condenses Big Data into patterns of expression (Fig 2a).

• Heatmap nodes can be clicked for statistical significant tests.

• Pathway viewer overlays data on a familiar biological context (Fig 2b).

The Canvas – a thought space

• Pinning data (Fig. 3) adds it to the Canvas, a space for organizing and contemplating data.

• Recommenders identify relevant information from external knowledgebases.

Collaborative Research

• Big Data analysis requires teams with diverse skills.

• Each user maintains a personal canvas.

• Multiple users collaborate via GitHub (Fig. 4).

Methods

• Multiple data views implemented as web applications.

• Software assistants proactively research external knowledge sources associated with pinned items.

• Canvas state is saved as JSON and versioned via GitHub.

Figure 2 – Interactive Data Views. Images are dynamic portals for data browsing and access, not a static image. (a) Heatmap view supports on-the-fly statistical evaluation. (b) Pathway view layers data on top of KEGG.

Figure 3 – The Canvas. Users pin data from the interactive viewers, webpages, or even Excel through RESTful APIs. Data entities appear as a green node. Recommendations for relevant data/knowledge (grey boxes) are automatically surfaced for PubMed, NCBI, KEGG, and so on.

Figure 4 – Projects are repositories on GitHub. Active Data Canvas versions the data and each user’s analysis.

Conclusions

• View data in biological contexts

• Facilitates more effective collaborations

• Dramatically reduces the time to explore the external resources

• Speeds scientific discovery from large-scale omics data

Acknowledgements

This work was funded by the Laboratory Directed Research and Development Program (LDRD) at Pacific Northwest National Laboratory (PNNL), and U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Early Career Research Program and the Panomics Program. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under contract DE-AC05-76RL01830.

Contact: Samuel Payne, Ph.D.
Biological Sciences Division
Pacific Northwest National Laboratory
E-mail: samuel.payne@pnnl.gov