**There are 3 O’Reilly conferences coming up:**

**Strata, Velocity, Jupyter. NY data viz can give 20% off discounts for these.**

**Data Visualization in Translation Bioinformatics**

Scott D. Chasalow – director of translational bioinformatics methodology, Bristol-Meyers Squibb

Goals of his group: nominate novel drug targets, predict pt response to therapy. Precision medicine etc.

Scott loves data viz, and collects many examples of bad data viz.

# Common types of bad data viz

Spaghetti plot. Lots of differently colored line plots that overlap, can’t tell well diffs b/w.

Here, one line for each pt.

*Does* successfully show that increase over time in most of them

Need to go from here to showing what’s going on in average

What to do instead: *lattice plots*. Facets of different combinations. Here, faceting by median, mean, types of cells.

However, note that plotting raw means/ medians looks quite different than plotting mean/ median change from baseline. (That is, means and medians after subtracting baseline). Inconsistent as to whether increasing or decreasing.

This is bc missing data! Some pts have 1 or 2 timepoints, not all 3

Impt to think ab what aspects of the data you choose to show, not just how to show it

So, what to do instead instead: lattice of fitted means and confidence intervals from a model

Improves over visualizing raw summary statistics

Here, estimated means from fitting a model, rather than raw means

Projecting multiple data dimensions into a single graphical dimension

Bar plot where each bar is one of 24 categories – diff categories of mean, median, doses, time points

Incredibly hard for viewer to decipher. Must read each label on axis very carefully.

Also fuck pie charts and Venn diagrams.

## Summary

WHAT you display matters!

Raw summary stats?

Measures of variability?

Model-estimated stats?

HOW you display it matters

Plots generally better than tables

Map data dims to graphical dims!

# Common viz in bioinfo

## Underused

Lattice plots (faceting). Gives examples.

Scatterplot lattice – one panel per gene example

Lattice plot w/ one panel per pt, sorted by response strength

Lattice plot w/ one panel per treatment arm (e.g. dose), showing slope as dose increases

Graphical listing

A “listing” in clinical study reports is usually just a table dumping raw data for every pt

Better to plot them. Much nicer to review.

Scatterplots

Rather than overly busy line plots

Gives example w/ 4 quadrants, w/ change in percentage naïve T cells and change in percent activated T cells. Almost all pts are in quadrant w/ decreased % of naïve and increased % of activated, showing that approach has worked.

“Manhattan plot” – chromosome x axis, -log10(polymorphisms) is y-axis. Tons of polymorphisms associated w/ phenotype of interest at particular chromosomes. Shows which chromosome mutations are relevant for the disease.

High-density scatterplots with alpha are good

Can use contours, like topographical maps, to make density clearer

3D scatter plots are rarely but sometimes useful. Likes it for doing first three principal components of a big gene expression dataset, here with red and blue to show clusters, and showing clusters merging to make it clear that a particular data manip did what it’s supposed to.

Ellipse plots

Scatterplot matrix to show correlations is useful for 5 or so patients. For thousands of things, not gonna be useful. Ellipse plots are a solution for when you need to do a matrix of lots of things.

It’s sort of like a heat map! Red show neg corrs, blue show pos corrs, color intensity and shape of ellipse at each grid intersection indicates strength of corr

Can quickly detect patterns – certain areas of genes all neg corr w/ others

Sort the things you’re doing so can see stuff like this at a glance.

## Overused

Venn diagrams for things other than formal logic.

Gives example w/ Venn diagram of p-values that 2 different approaches give

Lose info, bc dichotomizes things! Where you draw the line between what goes in each category makes all the difference. Scatter plot is much more informative.

e.g. here, lots of points near but not in the intersection of the Venn diagram

Bar plots

Ones w/ just 2 numbers: just write the numbers. Not worth the space. Do a table instead.

Bar plot + arbitrary thresholds -> bad way to visualize correlations

Gives example where corr w/ one gene is y-axis, all others are listed on x axis

Tons of colored bars put together, unclear what pattern you’re supposed to see.

Anytime the reader has to keep going back and forth between key and plot, you fucked up.

**Do faceted dot plot instead!**

**Always try to find a way to make the plot not need a key.**

## Used about the right amount

“Target plots”

4 quadrants, each point is one patient. 1st quadrant shows decrease in 1st time point and increase in 2nd; 2nd quadrant shows increase in both time points; 3rd shows decrease in both time points; 4th shows increase at 1st time point and decrease at 2nd.

Tricky to initially decode, but can be useful.

Here, it summarizes how many patients had what trajectory of response pretty efficiently.

I would do this with facets rather than quadrants.

Volcano plots

A kind of specialized scatter plot

MvA plots

Axis-rotated from a scatter plot

It’s a way of showing principal components, I think

Shows example w/ facets of these, showing something weird was happening w/ 1 particular pt

# Scott questions

In genetics, recommends always showing two things: direction and strength of association; and strength of evidence for association.

How do we avoid falling for misleading data by coworkers?

Having standard analyses to do helps

Review process for any impt results to try to catch everything

Often discover things from visualization, put rigor and confidence intervals on them through further analysis and summary stats.

Good to try out a whole lot of diff viz methods before publishing

# Isaac Neuhaus

Also from BMS

canvasXpress – an R library for interactive data viz for reproducible research

Isaac comes from bio perspective, rather than stats perspective.

Got into data viz as way of analyzing. Tools for exploring data, rather than as final analyses to present.

Has lecture notes for this presentation online.

BMS bioinformatics moved from point and click to scripting tools, to make research reproducible.

Use R!

Knitr reports

Shiny web apps to complement a static report. Scientists prefer being able to mess with the data before they’ll believe your claims. Use Shiny to create a bounded tool to explore the analysis.

Shiny has out of the box limits.

Zoom, mouseovers, clicking on data pts, drag to select data pts, drag and drop… Tableau and such have these, and can’t reproduce how stuff done with that was generated.

Htmlwidgets – an R library to bridge R and JavaScript. Lets you make much more flexible and powerful user interfaces in a Shiny app.

canvasXpress started development 6 years ago. Goal: simple to put in a web page, works with any dataset.

In a viz, the landscape is critical. Often there’s a lot of text, hard to get at right scale.

Give example of a circus plot from genomics. Each arc segment is a chromosome, each line is a location in that chromosome mapped to a particular gene. Each circle has collection of samples, each sample has particular value for particular gene.

In canvasXpress, can expand the data and explore in color-coded table form to make it a lot easier to understand than a static circus plot. Can filter, etc.

A data-rich viz can be pretty but really hard to understand. If you don’t wanna zoom in and stare, you want a way to filter out parts you don’t want to look at.

Another problem: when you select some pts in one of multiple plots on a page, can you tell how they relate to others?

Gives example of diff kinds of scatter plots, made w/ random data. Profile, expression, volcano, and contrast plots.

Want a way to explore the data such that if you mess w/ points in one graph, it affects those pts in the other graphs.

Can set that up on canvasXpress

Per Isaac, scientists love heatmaps.

Fast and responsive handling of big datasets is a must in genomics. In canvasXpress, can load heatmap with 1 mil data pts in ~8 seconds.

Axes – “overlays” w/ genes on one axis and samples on the other.

Can add script with popups to guide user.

The unique thing a/b canvasXpress is that can track everything the user does to the plots.

User manips can be retrieved programmatically, saved to database, embedded as metadata in printed report.

Ctrl+r to reproduce how you got from starting point to what you were on.

And can save your final thing as a png. Which saves metadata such that if you open that png elsewhere in canvasXpress, you can get back to the original data it was generated from!

Summary: lets you add very rich interfaces, while keeping track of everything the user did to the data.

# Isaac questions

There are ways to embed canvasXpress reports on another webpage. There’s some CSS code you can use and just point it to a CSV.

Isaac says feel free to email him if need advice doing so.

BMS practices: everything on Amazon cloud is encrypted at rest and in motion; and everything is de-identified before can upload it.

Bandwidth limits depend how much money you have. They’re doing a dataset that takes $12,000 a day in cloud computation to work on.

When is it appropriate to do interactive vs static? He does interactive for scientists to make their lives easy and guide them thru it, bc otherwise they’re gonna download it anyway and it may be more of a pain for them to figure out what to do.

Scott advocates that need limits. The allowed parameters of an interactive are important! Need to make it hard for people to “find out” stuff that missing data or other limits *do not* allow them to truly find out.

Domino Data Labs – company that provides framework for running R in cloud environment.

Azure has that, doesn’t it?

In corporate America, it’s great to have a number you can call to get support, rather than have to support the infrastructure yourself.

How to handle missing data? Depends on what you’re trying to do

Impute/ interpolate to mean of non-missing values, if required

Do nothing, for analysis that doesn’t req complete data

Multiple imputation is good: don’t fill in missing values just once, do it multiple times to account for variation of possible values

There aren’t enough people w/ stats skills, and data sci skills broadly. Need better understanding of uncertainty, in particular. Need more people w/ subject matter expertise too, to apply data analytics to. Finding someone w/ both good data analytics and good bio skills can be hard.

Isaac says his department used to have 1 bioinformatician per 15-20 scientists. Now that ratio is flipped! V rapid progress, and v rapid *change*, in what’s technically possible. So generate v large amt of info, and takes a lot of time and thought to go thru it. Thinks will be growing demand for bioinformaticians for quite some time.