1. Aleksey Bilogur - Data Analyst at Kaggle and a civic open data tinkerer outside of work  
  
Open Data Visualization and Learning by Doing  
  
Aleksey will highlight a few of the viz projects he has done on his personal blog: <http://www.residentmar.io/> . Going a step further, he will pose the methodology of blogging on visualizing civic data on the web as being a great way of starting out in the data visualization and data analysis space.

He works on the open data platform recently added to Kaggle

Hobby is building viz w/ open data.

David Robinson – good data sci blogger at stack overflow. Aleksey endorses his piece on starting a blog.

Aleksey has a post on mapping every Starbucks in Manhattan to determine avg distance to a starbucks, and to other stores.

How to write blog posts: Question(s) -> find data -> analysis -> answer -> explanation to other people

When you see a viz you like: comprehend it -> codify what it’s doing -> try to reproduce it -> figure out how to present an updated version to an audience

Why blog?

-Learn skills and techniques. This is a fun way!

-Learn how to present your work to an audience. This is hard to do while at a co – expl to coworkers and expl to general public are diff skills

-Good way to build a portfolio!

Has some further-reading advice in his slides

2. Jessie Henshaw - Natural scientist, semi-retired, studies the natural behavior of complex systems of all kinds.  
  
Exposing systems hidden by local noise  
  
Jessie will show the power of smoothing the 3rd derivative (suppressing the ‘jerk’) of timeseries data. It helps expose hidden dynamics in the behavior of underlying systems and exposes a great error often made in interpreting data as scatter plots.

It is easy to confuse highly variable data with noise. There is often hidden continuity that you can see with smoothing methods.

Has example using photons from GRBs. Can see peaks, but it looks like a v noisy scatter plot, not smooth curves. Can use *derivative reconstruction* to get clear plots out of that.

Set the ‘jerk’ to 0 at the middle of every 5 points. This removes the jitters from the shape!

Interpolation method

Smoothing kernels:

Gausian

Linear

3rd deriv is like Gaussian and linear kernels, except that those are peaked in the middle and 3rd deriv is *not* peaked in the middle

Gives an example smoothing crime rates in NYC over a couple decades. Shows how peaks can be misleading; makes decline in 90s look more sudden than it was

Crime rate dropped greatly starting 1990 – before broken windows policing start (1996)

She argues that growth of music industry contributed to decline in murder rate

3. Joyce Robbins – Lecturer in statistics and data visualization at Columbia Univ.  
  
Let's Talk About Boxplots  
  
Joyce will discuss the history of boxplots, variations in how they are displayed, their strengths and weaknesses, and how to be a better reader of boxplots.

(Joyce’s mom organizes this meetup – Naomi Robbins)

Box plots are divided into quarters. Bottom 1/2, and top ½ in the box, bottom ¼ and top ¼ in the whiskers.

Half the data’s in the box. *Not* true that most of the data is in the box.

Line thru the box is the median, not the mean.

Outliers in boxplots have a formal def, unlike outliers in many kinds of graph. It is defined as anything 1.5x [width of box] beyond the limits of the box. This leads many people to think that if something shows up as an outlier on a boxplot, it is “really” an outlier and should be disregarded. This is not necessarily the case! You need to think about your data. Sometimes there is good reason for something to be out there.

A boxplot w/ no outliers is telling you 5 numbers. Min, max, limit of top half, limit of bottom half, median.

Discrete data does not belong in a boxplot. Use a stacked bar chart instead. Boxplots don’t display frequency counts, don’t show whole distribution; they are misleading for discrete data like things that come in integer units.

Whiskers can jump around a lot depending on the extrema. Can be deceptive.

A boxplot w/ really long box and really short whiskers – [didn’t really explain this].

4. Ben Dilday – Data Scientist at Gallup  
  
TILT: A Mobile Datavis Experiment  
  
Ben will show a couple experiments he has done with using tilt-of-phone to interact with visualizations in d3. The best example of the concept is this - it allows a user to simulate mouse-over brushing by tilting their phone, and to "jump" between categories by snapping their wrist.  
  
<http://bdilday.github.io/mlb_viz/lineups-viz-brushing-orientation-switch/>

bdilday.github.io/tiltdemo

Tilt is interesting interaction modality for mobile devices and for VR

3 axes of tilt possible. They’re called alpha, beta, and gamma.

Tilt in 1 axis is a good way to scroll through lots of available years for a dataset that has a lot of graphs available for each year. His example uses baseball data.

His implementation: gamma in -90 to 90 degrees. Compute sign(gamma) x (90 - abs(gamma))

Also tries using a snap motion to jump b/w categories in the data set. Uses this to switch decades in his tilt-to-scroll-by-year data.

Logic for switching: keep running avg of gamma over N=20 frames. When gamma changes by D=50 degrees, do switch. Then freeze switching for M=50 frames, so can bring your phone back up after the snap.

5. Ying He - Computational designer and researcher at the Metropolitan Museum of Art  
  
Understand China through Visualization  
  
Ying will try to answer the following questions: How can data visualization design help the Chinese see the unseen faces of Tai chi? How can we use open data to explore the differences of design and art colleges around the world? And how can we use data visualization to unlock the secrets of winning art awards in China?  
  
Three projects that she will show include:  
  
1.Tai-chi Motion (<http://heyinging.com/Taichi>)

Mapping tai chi motions to geometry and color. Diff b/w transformation and translation (not in the geometric sense, in the sense of changing data vs making it clear)

2.The Structural Formula of Design Colleges (<http://heyinging.com/JELLYFISH>)

3.Award Puzzle (<http://awardpuzzle.strikingly.com/>)

Looking for trends in the winners of an art competition. E.g. lots of harvest scenes did well.

Dataset of 2276 images dating back to start of competition in 1984. Visualized them by color; mostly-yellow images did disproportionately well compared to mostly-blue and mostly-yellow

Canvas size of winners steadily increased over time

Repeat winners, when they exist, won 2nd time w/ an image v similar to their first

Avg age of winners mostly quite young

Made an interactive for searching through the paintings, incl mapping by color, and map of China that displays locations of winners w/ transparent circles for number of winners from a given city.

Key words in descriptions that tend to win.

6. Michail Xyntarakis - Transportation Engineer with Cambridge Systematics. Principal investigator in data analytics and mobility related projects sponsored by the Federal Highway Administration.  
  
Visualizing Transit Movements and Delays in New York  
  
The space-time diagram that visualizes the movement of trains across space and time is on the cover page of one of the most well known and best selling visualization books. However, it is rarely used in practice even though it can tell a good story, sort of an X-ray on system performance on any given day. Using publicly available MTA bus data we will use the space-time diagram to show bus movements, stop dwells, and delays for some routes and days. We will experiment with color to encode additional information such as speed or delay from schedule. Finally, we will enhance the plot with percentile distributions to obtain additional insights about system performance.

Starts w/ diagram of visualizing train schedules from Tufte’s “Visual Display of Quantitative Information”

Distance vs arrival times; slopes give you the speed of each train

Made an updated version, visualizing vehicle movements. Flat when vehicle not moving, steep slope when moving fast.

Grey lines for schedules, red lines for actual positions.

“Stair” where there’s a delay at a specific stop

“Hill” where there is backup of several buses

Another plot, w/ delay color-coded in minutes. Red for running behind schedule, blue for ahead. Delays become more significant later in the day. Can see correspondence from slowdown in several buses in previous graph to subsequent delays here.

Another, w/ avg trajectory as red line inside grey 75th percentile geom. Can see how 75th percentile is tightest early in the day, spreads out later as more variation in arrival times occurs.

Some routes have more variability than others

7. Seth Surchin – founder at Chair Four Development Group LLC  
  
An Internet of Things problem: data presentation to Machine Operators on the Factory Floor  
  
Seth’s startup collects data from sensors that they mount on machines and combine that data ('facts' in the parlance of business intelligence) with human input data ('dimensions') to provide real-time feedback on operations to everyone in the business – from the people that run machines to the CEO.  
  
An important data visualization challenge for them is how to effectively present information back to machine operators. Seth will describe their problems and ask the audience for advice.

Puts photocells and other sensors on industrial machines. In particular, mail-stuffing and mail-sorting machines.

Trying to figure out how to best give feedback to users.

So, I would inform them of avg time of their coworkers, and offer them a goal a little before that.

Temporal data (changes every few minutes), low resolution data, needs to be visible from a distance (bc the machines are large).

Column order – why is it Shift, Job, Hour, Site? Should flow from broadest to most specific. (He says that is the order)

Also, has a bunch of diff time scales displayed on same axis.

Column labels should be all caps