50 Things a Data Scientist should know

Mike Meyer

Me

Education and Professional Path

- BA, Math/Stat University of Western Australia
  - Lots of summer internships
  - Mining, Railways, Egg Board, …
- PhD, Stat, University of Minnesota
- Visiting Faculty Carnegie Mellon, 1yr
- Asst Prof University of Wisconsin, 4yr
- Faculty Carnegie Mellon, 12yr
- Move to Seattle
- Statistician, Boeing, 1yr
- Statistician, Amazon, 1yr
- Founder and Chief Scientist, Intelligent Results, 5yr

- Statistician/Quantitative Analyst/Data Scientist, Google, 11yrs
  - All in “Ads Metrics”
  - Position Normalizers
  - Experiment infrastructure (we do **lots** of experiments)
  - Measure, Monitor, Experiment (but not much Model)
  - Principal/Director

50 Things you should know

Really only 40, so far

1. Simpson's paradox, in all ways, especially "mix" vs "metric" changes
2. Time series, autocorrelation
3. Actually communicating with people.
4. Causality thinking (Anyone's formulation, but Rubin's version is helpful)
5. Counterfactual thinking (and logging when appropriate)
6. Logistic regression + boosting, bagging, etc
7. t-test
8. Shrinkage, L1, L2, lasso, etc.
9. Statistical significance vs practical significance. (with enough data everything is statistically significant, but you might not care)
10. snippet of XX code
11. snippet of XX code
12. snippet of XX code
13. Poisson process
14. Hashing

1. Factorial experiments, fractional factorial would be nice
2. Randomization, blocking, replication (who knows what a Balanced Incomplete Block Design is?)
3. Delta method
4. Jackknife, bootstrap
5. Simulation, when appropriate
6. Cox model, hazard function, etc.
7. MH as an example of quick and dirty approximate solutions
8. How to calculate a variance in one pass, and why rounding error still matters.
9. Draw a graph (in R), with slicing (ggplot, trellis/lattice)
10. Multiple comparisons!
11. Scripting language, Python mostly, but any will do.
12. Quick "data-sense". How many people in the US? How many households. What is Apple yearly revenue. What is Apple global revenue / global population? Does the output of your analysis make sense, if not, iterate.

more

1. sample size calculation, power analysis.   At least back-of-the-envelope.
2. Why second order methods (based on inverting the information matrix) won't work for you, and what can you actually do.
3. Compare two distributions (they won't be the same, but at least try)
4. Selection bias, and why it will kill you.
5. some version of length-bias
6. Within group variance and between group variance, and why it matters.
7. overfitting, cross-validation. Training vs validation, etc.
8. business data sense. Does your data make sense? have you made an error pulling the data. Check before you spend weeks building a model?
9. so what? If you can't answer the "so what" question, then why are you doing the analysis.
10. Survey Sampling. Sampling weights.   Re-weighting observational data.

1. Aggregation, get variance from aggregate data.
2. Data cleaning.
3. Map-reduce.
4. Uniformity trials, no-op experiments, variance in general
5. Learn how to triangulate -- when you generate new data also generate something you already know. Look at other data sources, etc
6. Know the basics for Survey sampling, and Biases.
7. Unit of Analysis, Unit of Experimental Diversion and all that stuff.

What am I going to Teach you?

Not much -- or maybe nothing.

This presentation will be a list of things you should know. It is up to you to encourage your instructors to actually teach you these things.

This is an interactive talk. If you don’t interrupt and ask questions it will go really quickly.

1. So what?

If you can’t answer the “so what?” question, then why are you doing the analysis in the first place???

No amount of fancy methodology or “smarts” will help you if you can’t explain what it means or what action you can take based on the analysis.

Always be prepared to do the “next thing”, as in “OK, that is interesting, what do we do now”?

2. Simpson’s paradox, aka mix shifts

Without looking it up, who remembers Simpson’s paradox?

Here is how it might bite you.

Queries per user is going up in the US, good.

Queries per user is going up outside the US, good.

But overall (combining US and outside US), queries per user is going down.

WHAT? How can that be?

Mix Shift example

	Last Year			This Year			
	Queries	Users	Q/U	Queries	Users	Q/U	
US	100	10	10	121	11	11	good
Outside	80	20	4	800	160	5	good
Total	180	30	6	921	161	5.7	BAD!

3. Correlation, and Autocorrelation

You know that if variables are correlated then they aren’t independent.

But what about time series?

Assume you do an experiment and look at Queries/User over a number of days. Q/U is up on day 1, and day 2, …, and day 7. So by a simple sign test it is UP (good!)

But wait. It is the same users every day, so the results are (auto)correlated.

Calculating significance is obviously not that easy.

4. Communicating

Obviously, this should be No. 1.

The best analysis is useless if you can’t communicate **clearly.**

You need to explain the technical details -- there will be someone in the room who follows the details.

And you need to be able to explain in a non-technical anecdotal manner for everyone else.

Think about mix shifts. For some “mix shift” is enough, for the rest you need more.

4a. Communicating

… and of course, the usual business communication material

- Much of what you do will be confidential, so be careful
- Much of what you do might be subject to litigation, so be **very** careful what you put in writing
- Be really careful with what you write. A misplaced _not_ can change the entire meaning of what you write, and that might cause late night panic amongst VPs. You don’t want that.
- …….

5.Causality

Obvious (to you). Correlation is not causation. Beer and Diapers. Correlation may still be useful.

Less Obvious. You need a framework to think about causality. There are several --- Don Rubin’s formulation is a good one.

Mostly you need to twist your brain to think in “counterfactual” terms. What could have been?

Example. “I took a pill and now feel better”, but we don’t have the counterfactual of “I didn’t take the pill”.

6. Experiments and Logging and Counterfactuals

Obvious. Measuring is hugely important. You’ll spend a lot of time measuring things properly

Obvious. Randomized Experiments are hugely important.   Why? (What about observational studies??)

Less Obvious. If you have good logging and counterfactual logging, then life is much better --- in the sense that variance is less.

Example?

Counterfactual Example

Experiment to increase Q/U.

Control and Experimental arms, run worldwide.

Treatment only affects users in the US.

If we compare worldwide stats then the noise from the “outside US” traffic might overwhelm any US signal.

Solution. Log Queries in Experiment as within the US and ditto with Control.

There are much more subtle versions of this, e.g. Maps example.

7. Logistic Regression

You need to be able to perform a logistic regression (what is that?) on a large to very large data set.

Linear LR is a fine start, but some version of

GAM (Generalized Additive Models) is useful too.

Bagging and Boosting are methods of fitting a LR that lead to something like a GAM.

At the very least know what all these things are.

8. T-test and z-test

You need to know how to do these, even when slightly disguised.

Example.

Z1 and Z1 are Normal RVs with known means and variances $(\mu_1, \mu_2, \sigma_1, \sigma_2)$

How do you calculate Pr(z1 > z2)?

??

9. Simulation

What if I take the previous problem and say.

I have 10 RVs, Z1,...,Z10. What is

Pr(z1 > max(z2, …, z10))?

Theoretically really hard, but you know everything in sight, so Simulate.

Simulation is cheap and very powerful.

10. Significance

You just found a statistically significant difference,   W00t.

Does it pass the “so what?” test.

Statistical significance isn’t the same as practically significant, or meaningful.

Screw this up a few times and you have lost credibility, and that is about the only currency that matters for a Data Scientist.

11. What code is this?

```
func classifyMobileDevice(browser, os, yearClass string, height, width int32) string {
 switch {
    case browser == "iPhone/iPod":
        // Label only portrait cases for clean data, which account for ~99%.
        // If want landscape also, can take the max of height and width as height, and min as width.
        // Only consider the most frequent height without url bar for clean data.
        switch {
        case width == 320 && height == 460:
            return "iPhone4/5"
        case width == 375 && height == 559:
            return "iPhone6"
```

12. What code is this?

```
SELECT Age, Gender, AVERAGE(IF(Dead, 1, 0)) as FatalityRate

FROM FatalityData

GROUP BY Age, Gender;
```

13. What code is this

```
percentage <- function(x) {

  if (is.logical(x)) {

    return(sum(x, na.rm=T) / sum(!is.na(x)))

  }

  return(x / sum(as.numeric(x), na.rm=T))

}
```

14. Experiments, what are the three core tenets?

Randomization, why?

Replication, why?

Blocking, why?

15. Fancy experiments

You should know.

- Simple A/B test
- Randomized block design
- Factorial Design
- Fractional Factorial Design

You can probably forget

- Balanced Incomplete Block Designs, but bonus points if you actually know it
- Ditto, Latin Squares

16. Simulation

We already talked about this, but Simulation is very powerful and often inexpensive.

It is sometimes useful to know about “variance reduction” methods in simulation, things like antithetic variates, control variates, importance sampling and stratified sampling.

17. How to calculate the variance

You all know that Variance = mean((x_i - mean(x_i))^2)

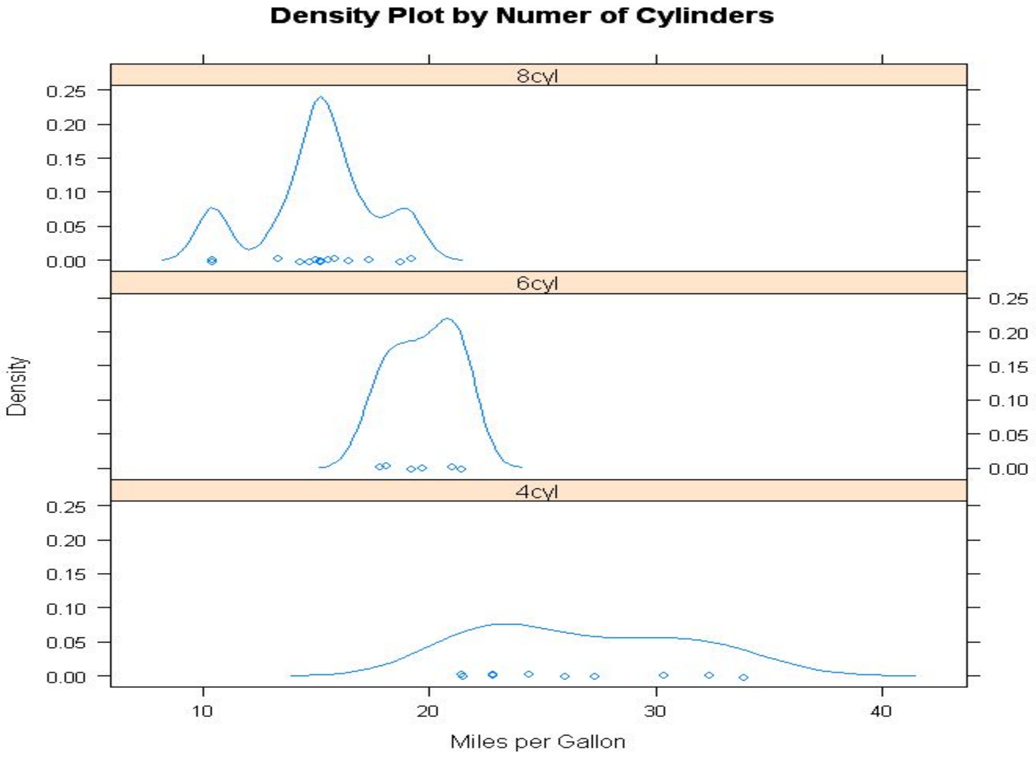
But that requires two passes over the data. One for the mean and then a second pass for the squares.

You can do it in one pass using

mean(x_i^2) - mean(x_i)^2

But that is poor because of rounding error.   How do you do it in one pass?

18. Lattice/Trellis/ggplot graph



Get good at some Graphics package

GGPlot seems to be the current vogue.

19. Multiple Comparisons

Assume you do 100 95% CIs.   Then about 5 of them will be significant even if there is no effect.   Be cautious about cherry picking.   Replicate whenever you can.

Learn about the various ways of correcting for multiple comparison. The Benjamini–Hochberg False discovery rate approach is standard and useful.

20. Uniformity Trials

An old idea from field experiments, but a new life in online experiments.

Assume that experiment diversion is easy (it often is online).

You can do 10, or 20, or 100, “null” (or A vs A) experiments and use those to

- Calibrate your intuition
- Make sure that your CIs really do have the correct false positive rate

21. Triangulate

REALLY obvious.

If your analysis says “doing this will improve Q/U by 10%”, then see if you have any other data that would verify such a claim.

- Have you ever seen a change that big?
- Does the new Q/U number make sense.
- Etc.

Ground your conclusion in reality.

22. Just business common sense

Headline. Google Pixel was more popular than the Iphone during Black Friday

Google’s Pixel phones proved very popular during the holiday shopping season for Black Friday and Cyber Monday, as device activations were up 112% compared to activations in the preceding weeks. A new report from Localytics also reveals that the iPhone activation rates during the same period were only up 13%. The Galaxy S7 proved popular among Black Friday shoppers as well, as activations for it were up 36%.

Does this make sense?

Conclusion

Skills you will need

- Some basic stats knowledge
- Some programming skills
- Good communication skills
- Good problem solving
- A lot of perseverance
- A **lot** of common sense

= Data Science