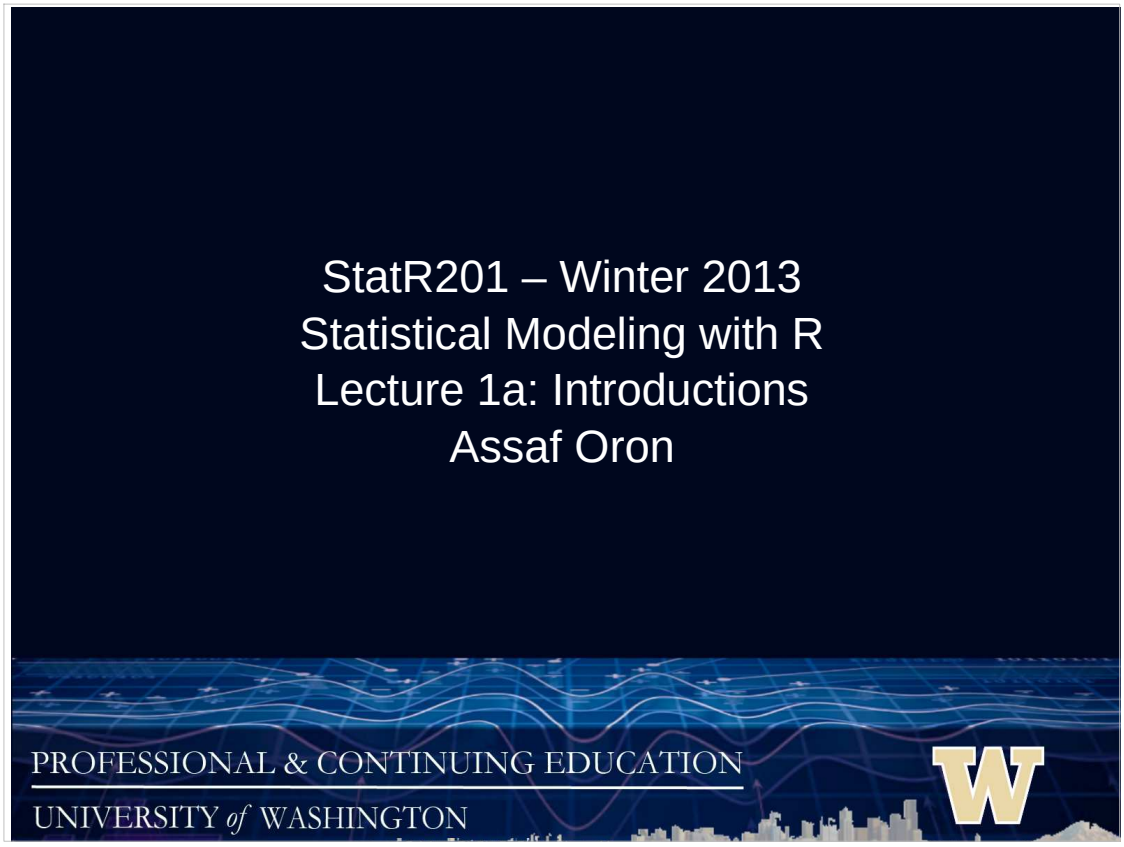

About Myself

Been here & there, done this & that

- Originally from Israel
- B.Sc in physics, M.Sc. in environmental science
- Worked as Intel engineer – my first glimpse of the power of statistics in action
- A couple of years doing mathematical R&D at an Israeli biotech startup *(it's a wild scene)*
- Came to Seattle 10.5 years ago (w/family), for the Ph.D. Program at the UW statistics dept.
- Graduated 5 years ago. Since then, 1 year @Hutch, 3.5 years @UW, and now at Seattle Children's Research Institute.

About Myself

My foremost methodological interest is design and estimation of dose-finding experiments, esp. clinical trials.

As a statistician, I've also dabbled in: *(partial list)*

- Analysis of game-theory experiments
- Methods for microarray data analysis
- Spatial and spatio-temporal modeling of air pollution data, and related data-quality issues
- Methods and metrics for prediction-oriented model selection
- Masking/Blinding analysis for randomized trials
- And what about you? Twice a lecture or so, I'd like to learn a bit about 2-3 of the students – so be prepared...

R and Me

I've been using R since Year 1 of my Ph.D. Studies. It quickly became my nearly-exclusive tool for doing statistics.

I've co-authored a couple of R packages. I had a solo one on CRAN for a couple of years, but took it off (need to re-do some of the theory).

I am definitely more statistician than programmer. Not a programming wizard by any means. *(I have a baby brother who is, so I know the difference)*

Therefore, I expect to learn a lot of programming tricks with/from **YOU.** Hopefully, I will teach you a few as well :)

This Sequence and Me

I enjoy teaching *(except maybe the pressure of facing a class in a few hour, and things not in order yet ;).*

Therefore, when offered this opportunity by UWEO, I hesitated little before agreeing.

I also hoped, that teaching this material will encourage me to update myself about latest developments in the rapidly evolving R universe.

This last wish has already been more than granted. Thanks Eli for discovering `knitr,` and thanks Yihui Xie for inventing it!

Class Formalities and Expectations

Essentially, the same as StatR101, with minor tweaks:

- Lectures on Thursdays thru March 21 (1 week off during Spring break). Office hours remain on Tuesdays.
- Homework grading remains the same, but not necessarily an assignment every week. Need 75% cumulative.
- To clarify a misunderstanding from Fall: unless told otherwise, the final project **is** part of the completion requirements.
- I am at assaf@uw.edu.
- Questions on expectations?

Less Formal Expectations

Expect lecture notes to be not as pretty; you will soon discover how lazy I am.

Seriously: this sequence is a brand-new work in progress. The syllabus and content had to be adjusted until the last moment, to match the material delivered in Fall.

The amount of material in Winter can easily become staggering (more on that, soon). Active learning retention beats passive learning by >10:1. If forced to choose (and I am), I prefer to focus lectures on interactive and hands-on active learning, and not on elegantly-finished material for passive learning.

Also, I am falling in love with the brand-new tool of `knitr,` in conjunction with `pandoc` and `slidy.` I plan for us to learn to use this combination and others together. As Eli started showing you, these tools are very useful for fast reproducible, analysis-derived publishing, and for active learning.

Ok, Let's hear from 2-3 Students

Whatever you want to share about yourself....

…. and then, a micro-break?

Class Content

In this certificate program, we are in a bind. The format is roughly equivalent to a sequence of three 3-credit classes. Teaching a truckload of R content and trickery stand-alone, *as if you already knew the statistics*, is probably feasible within these limits. However:

- It would be ***wildly*** irresponsible. All of us – instructors, the advisory board and UWEO – want to adhere to the program's name and spirit, i.e., Statistical Analysis with R. Statistics first, R second. And remember: I'm saying that as someone for whom R has been my data-analysis eyes, nose, ears and hands for 10 years.
- This is especially true from Winter onwards, when we gain access to a wide variety of easy-to-operate, even-easier-to-wreak-havoc-with, black-box methods.
- This means that we have to teach a **lot** of statistical material.

Class Content

During Fall, in the space of a 3-credit course you learned the equivalent of

- Introduction to probability and statistics, advanced level (5 credits), *PLUS*
- Introduction to R (2-3 credits)

In Winter, you are due to pick up the equivalent of a

- Prediction-oriented Regression course (3-4 credits, if you can find such a course), *PLUS*
- Introduction to Machine Learning from a statistical perspective (3-4 credits), *PLUS*
- Publishing and Advanced Graphics with R (2-3 credits, if you can find it)

The two statistical parts will be in series; the R bit is in parallel.

Class Content

On the positive side, you are grown-up and more focused than undergraduate students, and I have less to worry about grade-haggling, babysitting, etc. So we can get more real work done.

That said, given the amount of material we will not delve as deeply into mathematical derivation. My focus is on a reasonably deep understanding of the concepts, and on using them responsibly.

There will still be a fair bit of math, including quite a few matrix-vector equations.

Which reminds me: brush up on your linear algebra, if you have it. If not, please pick up some by next week. You can't really understand the inner workings of regression and many machine-learning tools without it.

Class Content

My approach to content in this class is the 90:10 rule.

- Roughly speaking, on a regular basis I use <10% of the R tools I know (probably much less) to do >90% of my analysis work.
- So instead of filling your brain with a laundry-list of R trees *(proverbial trees, but also classification trees)* – I'd rather teach you the essence of the forest.
- The hope is to give you the solid basis and the right mindset to be able to acquire new tools, even if they have barely (or not at all) been discussed during the course itself.

… ok, let's take a look at the syllabus...

...and the textbooks...

Class Content

There's an additional, “wild card” content element not mentioned in the syllabus. Every week, I plan to introduce...

- ...A “Cool R Trick of the Week”, something relatively simple but not very-well-known, that has been really useful for me.
- ...an “R Annoyance of the Week and How to Avoid it”. No language is perfect, and R – being the hodgepodge that it is – surely ain't. I will highlight one (again less-well-known) minor but annoying drawback of R, and show ways to work around it.

- When either of these is introduced, there will be a (not too taxing) exercise in the homework to reinforce them.

Fine Print: A Word on GNU Etiquette

I've heard that some people outside the course told the UW “it violates open-source rules” by not opening the contents of this course to everyone, all the time, for free. Now, I am no expert on GNU etc., but here's what I have gleaned:

- R is a free GNU software, and – to quote GNU – free indicates *FREEDOM*, not free of payment.
- Any code excerpts, manipulations, tips, etc. you get here – you are free to distribute while citing their source (and that includes LaTex, knitr, slidy, etc.)
- However, the training itself is a service separate from the software, and the UW, to my understanding, has the right not only to charge for it (Duh) - but also to restrict it to those who pay.

There is another side to this: as you graduate here and return to work in business, you should adhere to the GNU principles when using R.

- To my understanding: if you embed R code in a commercial software product, then the code of this product needs to be freely shared as well.
- But check it out yourself – and make sure you adhere to the principles. A lot of people donated their goodwill to make this amazing, free (in both senses) world of statistical tools.

Questions?

Thank you!

After the break, we will do some real statistics.

(speaking of which: any ideas regarding break policy?)