Stat 215A: Lecture 2
Questions and data

Office hours: Bin, 2 pm Tuesday/Thursday, 409 Evans
Karl: 11-1, Monday; 12-2 Wed, 444 Evans

Logistics

Questions about the handouts?

Course materials are posted on bcourses.berkeley.edu

If you are not enrolled yet, please write your email address on the sheet of paper going around so Karl can add you to bcourses …

Reminder: speak or not to speak

- Pros

Practice communication skills – for your work to be appreciated, to be commented and improved.

Helping to think, clarify your own thoughts

Getting people to know you and you to know others

Concentrating better if you engage so you learn more.

- Cons:

Embarrassment

Self-deception, confusing people

10 Sets of questions to build data wisdom

- See handout…

We will be going back to the 10 sets of questions often in the class so have the handout handy.

The Celniker Lab at LBNL

https://www.youtube.com/watch?v=xIW9lx-E6gM&feature=youtu.be

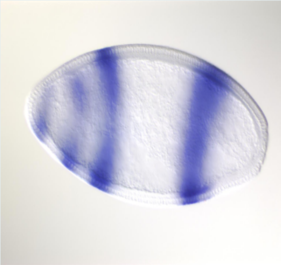
Thanks to Karl…

Co-lead: Dr. Erwin Frise

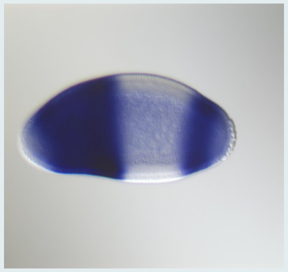
Seeking a specific question to answer

- Could we take the "optimal" approach here?
- Exploratory data analysis
- Often people do things that they have skills for
- Uncontrolled inspirations – embed oneself in a wet-lab maximize the opportunities, also reading literature.

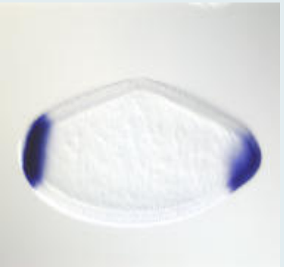
Gap Gene Network



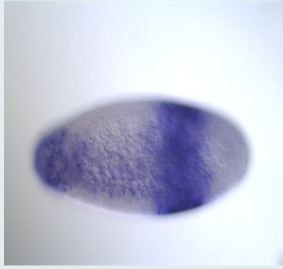
gt



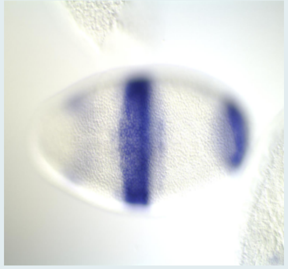
hb



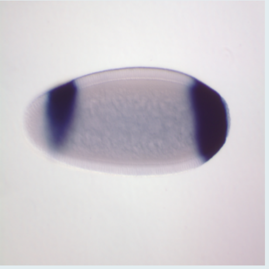
hkb



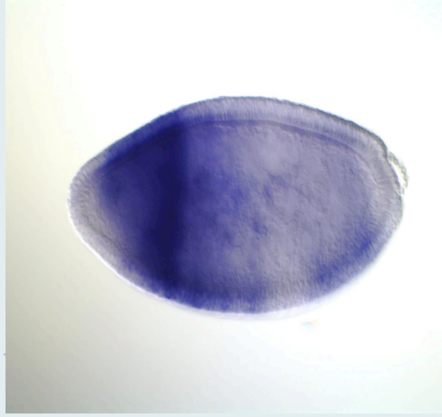
kni



kr

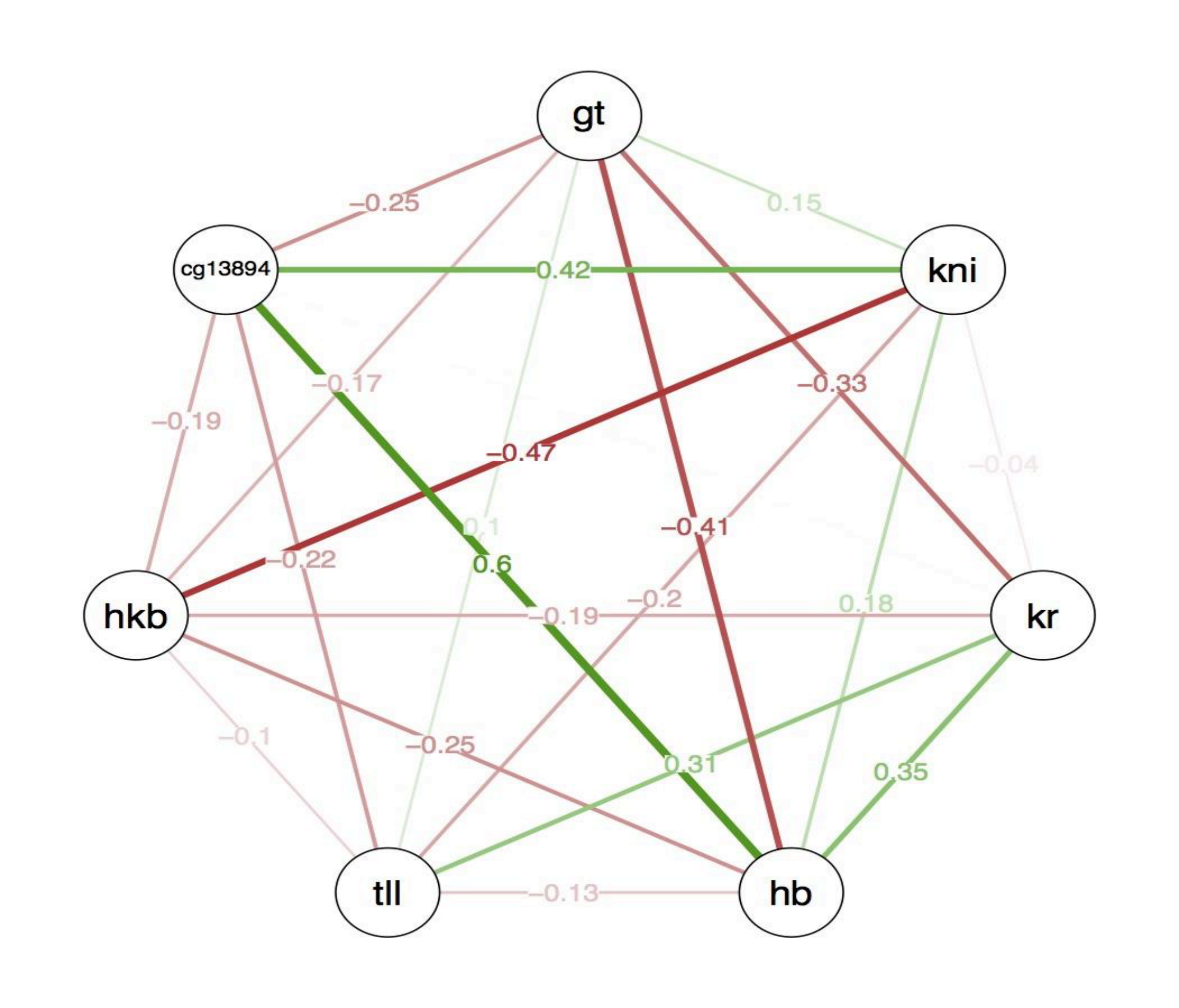


tll



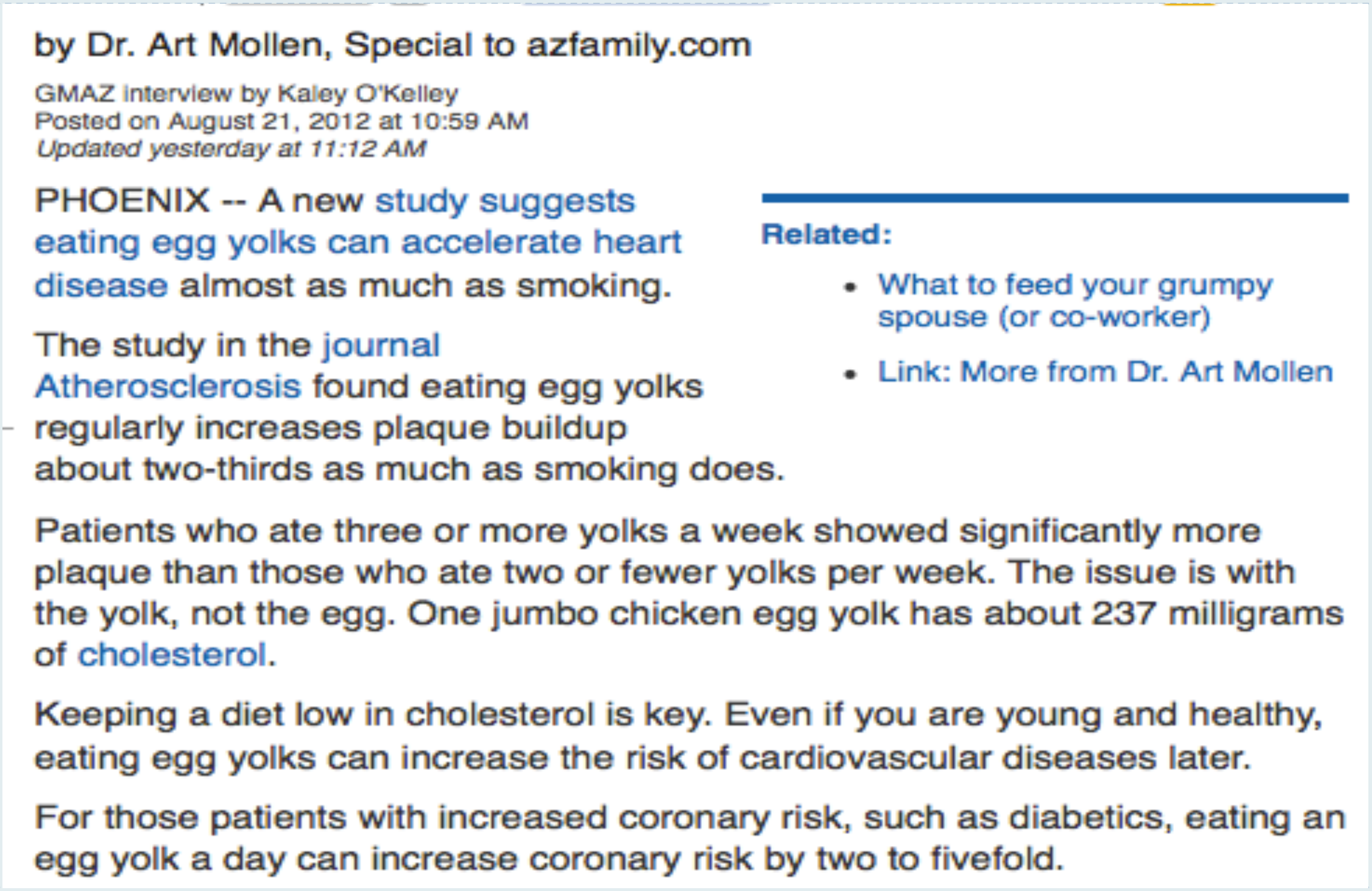
candidate gene: CG13894

Global Correlations with CG



Suggestions from class

Age of quantitative information



by Dr. Art Mollen, Special to azfamily.com

GMAZ interview by Kaley O'Kelley
Posted on August 21, 2012 at 10:59 AM
Updated yesterday at 11:12 AM

PHOENIX -- A new study suggests eating egg yolks can accelerate heart disease almost as much as smoking.

The study in the journal Atherosclerosis found eating egg yolks regularly increases plaque buildup about two-thirds as much as smoking does.

Patients who ate three or more yolks a week showed significantly more plaque than those who ate two or fewer yolks per week. The issue is with the yolk, not the egg. One jumbo chicken egg yolk has about 237 milligrams of cholesterol.

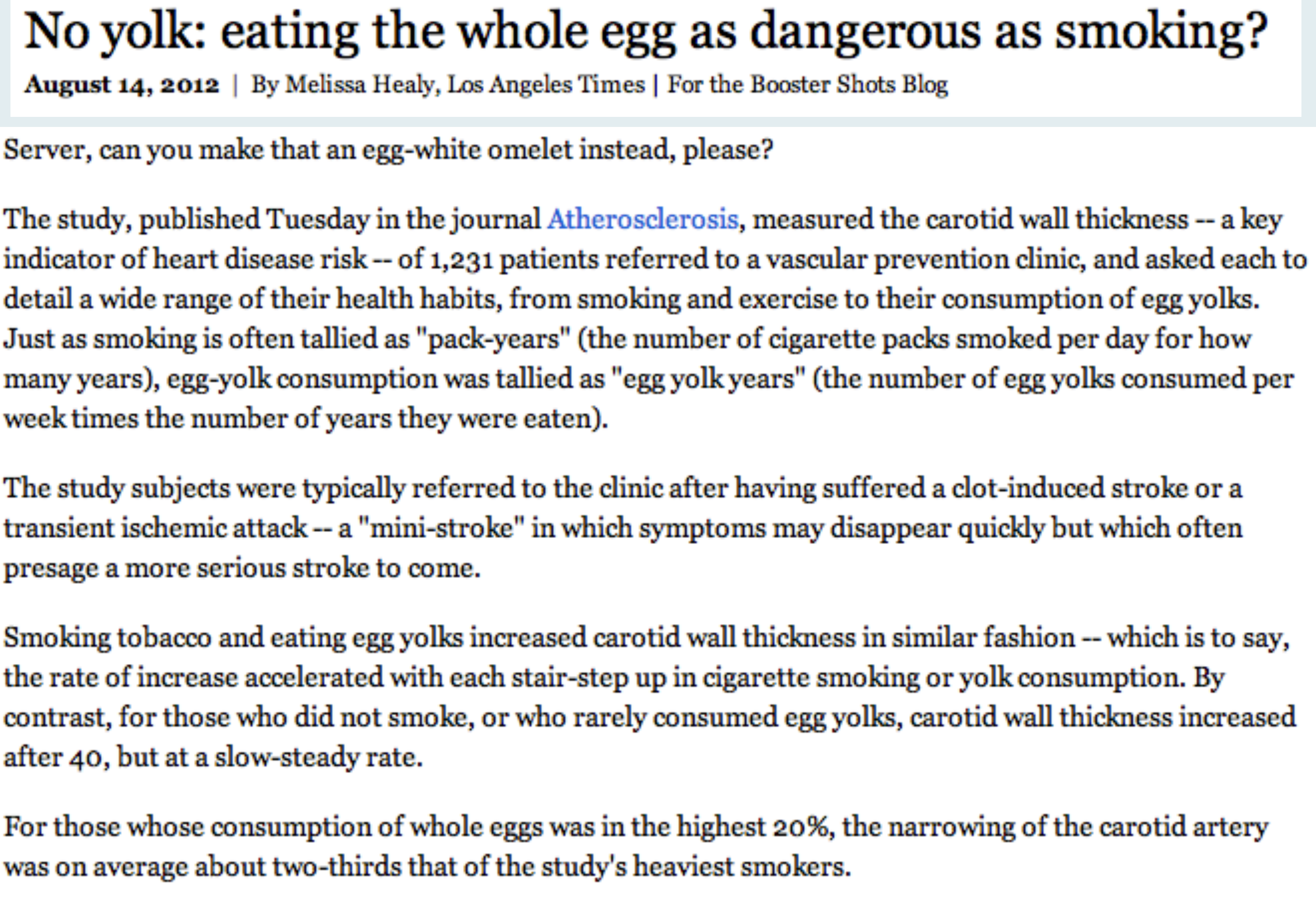
Keeping a diet low in cholesterol is key. Even if you are young and healthy, eating egg yolks can increase the risk of cardiovascular diseases later.

For those patients with increased coronary risk, such as diabetics, eating an egg yolk a day can increase coronary risk by two to fivefold.

Related:

- What to feed your grumpy spouse (or co-worker)
- Link: More from Dr. Art Mollen

Age of quantitative information



No yolk: eating the whole egg as dangerous as smoking?

August 14, 2012 | By Melissa Healy, Los Angeles Times | For the Booster Shots Blog

Server, can you make that an egg-white omelet instead, please?

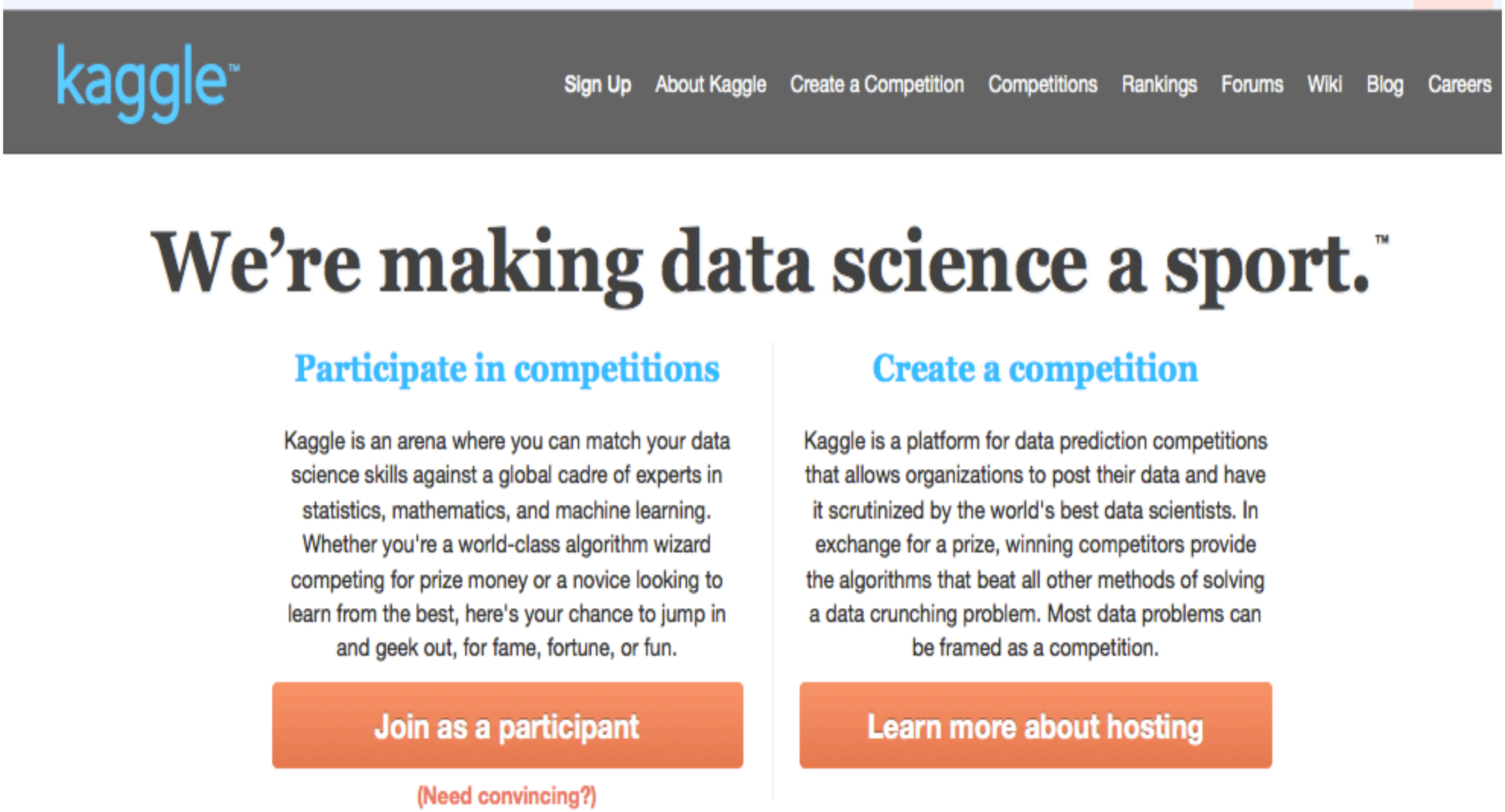
The study, published Tuesday in the journal Atherosclerosis, measured the carotid wall thickness -- a key indicator of heart disease risk -- of 1,231 patients referred to a vascular prevention clinic, and asked each to detail a wide range of their health habits, from smoking and exercise to their consumption of egg yolks. Just as smoking is often tallied as "pack-years" (the number of cigarette packs smoked per day for how many years), egg-yolk consumption was tallied as "egg yolk years" (the number of egg yolks consumed per week times the number of years they were eaten).

The study subjects were typically referred to the clinic after having suffered a clot-induced stroke or a transient ischemic attack -- a "mini-stroke" in which symptoms may disappear quickly but which often presage a more serious stroke to come.

Smoking tobacco and eating egg yolks increased carotid wall thickness in similar fashion -- which is to say, the rate of increase accelerated with each stair-step up in cigarette smoking or yolk consumption. By contrast, for those who did not smoke, or who rarely consumed egg yolks, carotid wall thickness increased after 40, but at a slow-steady rate.

For those whose consumption of whole eggs was in the highest 20%, the narrowing of the carotid artery was on average about two-thirds that of the study's heaviest smokers.

Age of data with many analysts



We're making data science a sport.™

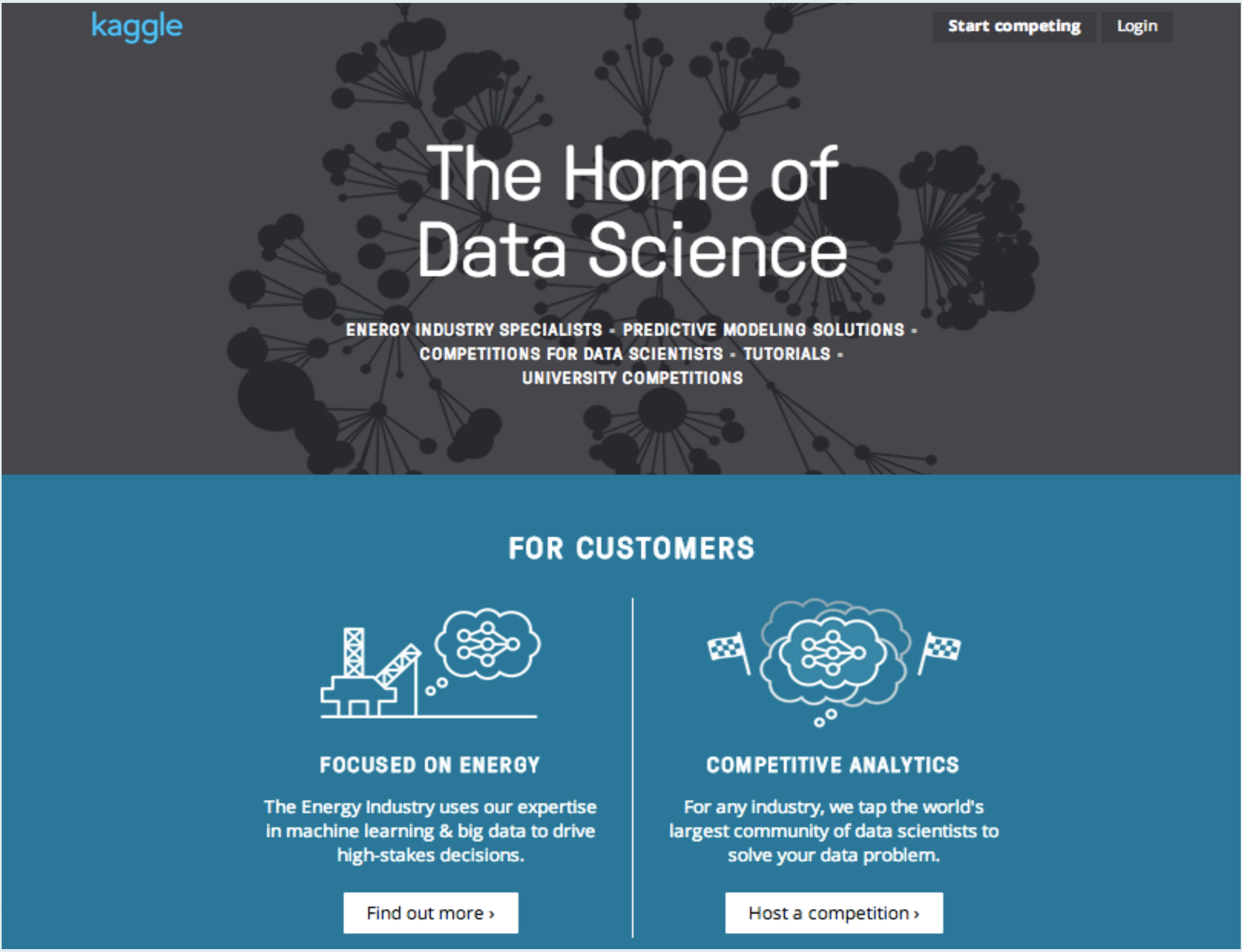
Participate in competitions

Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.

Create a competition

Kaggle is a platform for data prediction competitions that allows organizations to post their data and have it scrutinized by the world's best data scientists. In exchange for a prize, winning competitors provide the algorithms that beat all other methods of solving a data crunching problem. Most data problems can be framed as a competition.

Kaggle in 2014



The Home of Data Science

ENERGY INDUSTRY SPECIALISTS - PREDICTIVE MODELING SOLUTIONS - COMPETITIONS FOR DATA SCIENTISTS - TUTORIALS - UNIVERSITY COMPETITIONS

FOR CUSTOMERS

FOCUSED ON ENERGY

The Energy Industry uses our expertise in machine learning & big data to drive high-stakes decisions.

COMPETITIVE ANALYTICS

For any industry, we tap the world's largest community of data scientists to solve your data problem.

Statistics inference is inductive inference from data

- From part to whole
- From now to future

Future is always implicit for impact.

Condition of generalizability has to be satisfied to "generalize" from part to whole, from past to future.

Humans make the call based on prior knowledge.

Quotes from a Science article (2011):

Statistician Rob Hyndman of Monash University, … recently used Kaggle to lure 57 teams, … into improving the prediction of how much money tourists spend in different regions of the world. “The results were amazing … They quickly beat our best methods,” he says.

http://promo.aaas.org/ema/booklets/data.pdf

Do you think the results from the training data would hold up

with future tourists? Why or why not?

Egg-yolk study vs Kaggle problem
Similarities and differences

- Differences

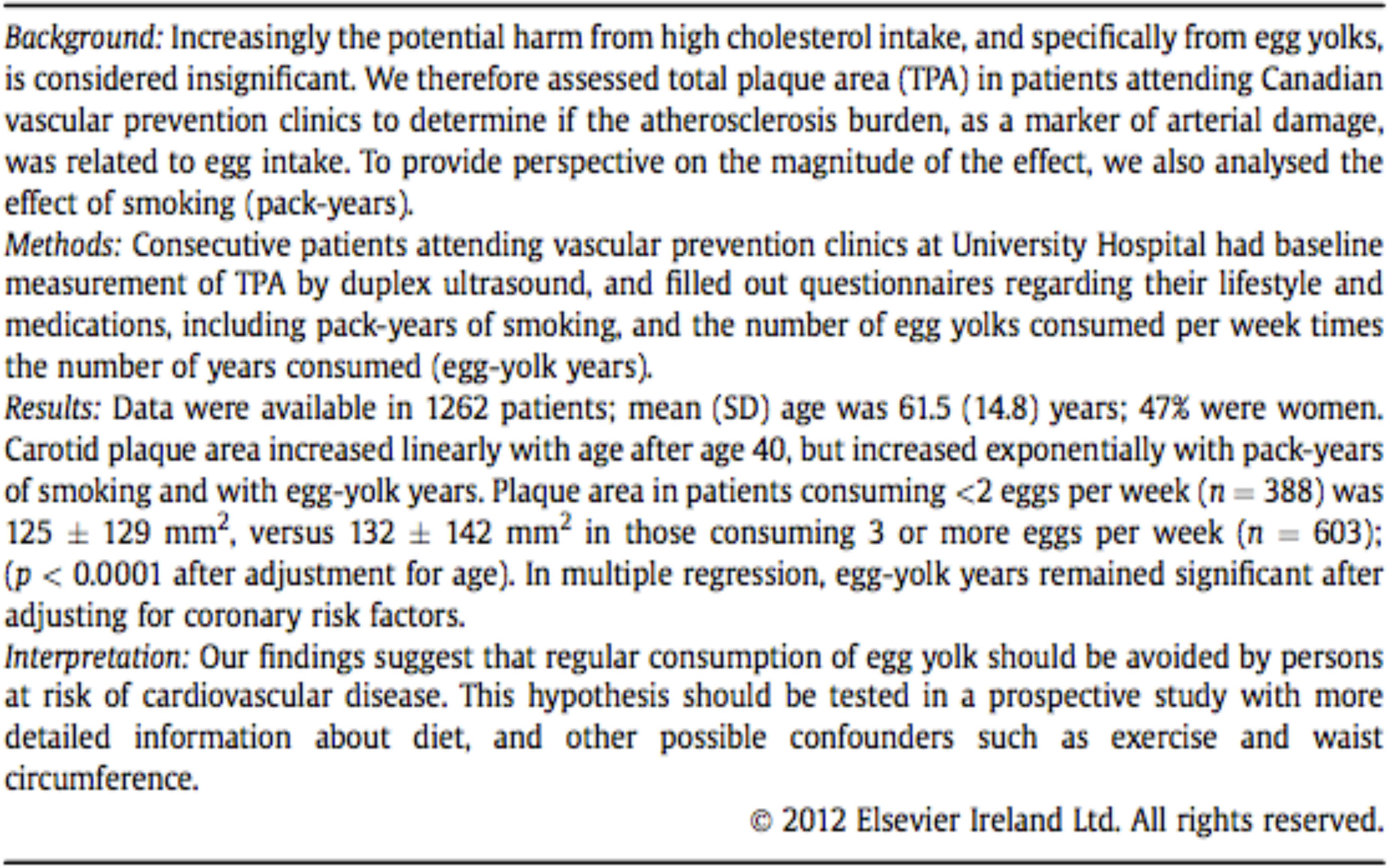
yolk: iterative process, science-driven, not enough data, causal inference needed, no real conclusion (but newspaper did), statistical inference called for

Kaggle: one hit, not driven by science so much, not reproducible (black box), crowd-sourcing, association/ correlation, decision-driven

Egg-yolk study

What are your questions and thoughts after reading the paper if you get to ask the authors?

Paper abstract from Spence et al (2012) in *Atherosclerosis*



Background: Increasingly the potential harm from high cholesterol intake, and specifically from egg yolks, is considered insignificant. We therefore assessed total plaque area (TPA) in patients attending Canadian vascular prevention clinics to determine if the atherosclerosis burden, as a marker of arterial damage, was related to egg intake. To provide perspective on the magnitude of the effect, we also analysed the effect of smoking (pack-years).

Methods: Consecutive patients attending vascular prevention clinics at University Hospital had baseline measurement of TPA by duplex ultrasound, and filled out questionnaires regarding their lifestyle and medications, including pack-years of smoking, and the number of egg yolks consumed per week times the number of years consumed (egg-yolk years).

Results: Data were available in 1262 patients; mean (SD) age was 61.5 (14.8) years; 47% were women. Carotid plaque area increased linearly with age after age 40, but increased exponentially with pack-years of smoking and with egg-yolk years. Plaque area in patients consuming <2 eggs per week ($n = 388$) was 125 ± 129 mm^2, versus 132 ± 142 mm^2 in those consuming 3 or more eggs per week ($n = 603$); ($p < 0.0001$ after adjustment for age). In multiple regression, egg-yolk years remained significant after adjusting for coronary risk factors.

Interpretation: Our findings suggest that regular consumption of egg yolk should be avoided by persons at risk of cardiovascular disease. This hypothesis should be tested in a prospective study with more detailed information about diet, and other possible confounders such as exercise and waist circumference.

Egg-yolk study: from Spence et al

"Plaque area in patients consuming <2 eggs per week (n = 388) was 125 ± 129 mm^2, versus 132 ± 142 mm^2 in those consuming 3 or more eggs per week (n = 603); (p < 0.0001 after adjustment for age). In multiple regression, egg-yolk years remained significant after adjusting for coronary risk factors."

pack-years of smoking - number of packs per day of cigarettes times the number of years of smoking

egg-yolk years - number of egg yolks per week times number of years consumed

Egg-yolk

Why significance testing relevant?

What test was used?

What question does the test answer?

Significance testing is a statistical inference problem

Egg-yolk: when is the testing relevant?

- Statistical hypothesis testing is about parameters in a population from which a random sample is drawn.

  It does not answer questions about numerical characteristics of the data in hand. For the two groups, the difference is 132-125 = 7 – whether this 7 is a big enough a difference has to be decided based on medical knowledge, not on statistics.

Egg-yolk data sampling model

Stat testing is for population parameters, in this case, about population carotid wall thickness difference being zero or not --

What population?

How do data relate to the population

if there is a legitimate one?

Egg-yolk: when is the testing relevant?

Here is the text book situation for a two-sample t-test which I speculate was used in the paper for the results in the abstract:

Two populations:

Population 1: all Canadians over 40 who have had a minor stroke as people in the data set and who had eaten < 2 eggs in the same week as in the survey. Each person has a "ticket" in the population box and on the ticket are the value of his/hers carotid wall thickness in that week and other characteristics.

Population II: all Canadians over 40 who had had a minor stroke as people in the data set and how had eaten 3 or more egges in the same week as in the survey. Each person …

Egg-yolk: when is the testing relevant? (cont)

A simple random sample is taken from pop I to form the control group and a simple random sample is drawn from pop II to form the treatment group.

A simple random sample of size m from a population of size M puts a uniform distribution on all subsets of size m of the population. It can be obtained by drawing tickets one by one without replacement and uniformly from the remaining tickets in a population box.

When m<< M, the random variables on the draws are very close to independent and identically distributed.

(why?)

Egg-yolk: when is the testing relevant? (cont)

Obviously, no random sampling was done to form the treatment and control groups in Spence et al (2012).

And they didn't attempt at making any argument to convince a reader that their samples of "convenience" used in the paper could be viewed as simple random samples from two well-defined populations; or at least are "representative" of the populations. In fact, they didn't bother to discuss populations at all and hence didn't worry about what it means to reject a null hypothesis which was also not specified, even though p value < … was reported.

Null and alternative hypotheses are not symmetric

When an alternative is rejected, null is "accepted", which is different from "null" is proved.

Often null is accepted because there is not enough information to reject it. For example, with one toss of a coin, the null of the coin being fair is always accepted at 5%, but there is no proof that the coin is fair.

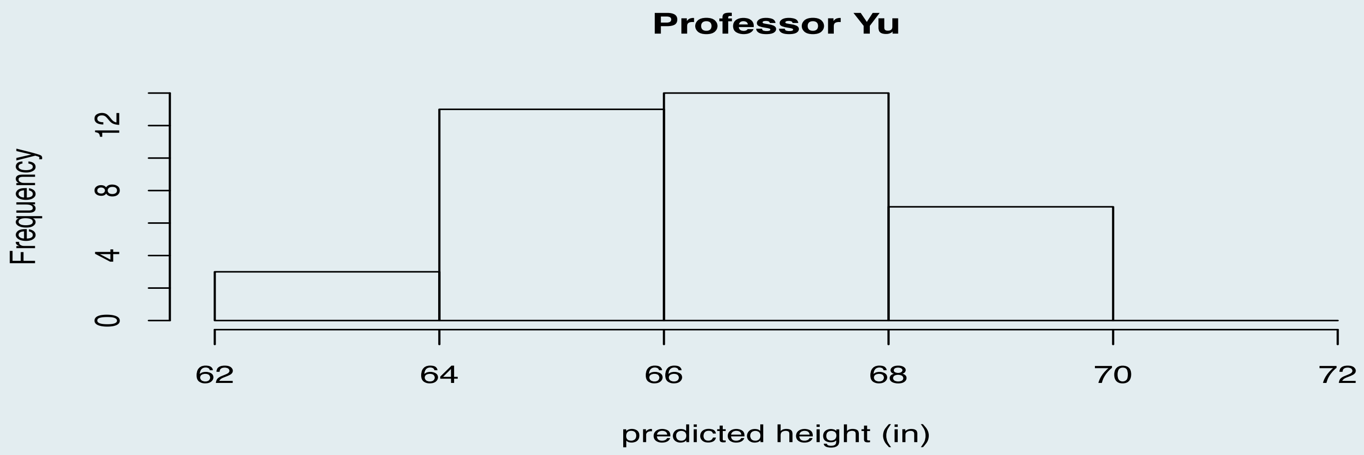
Bin’s data collection from last class

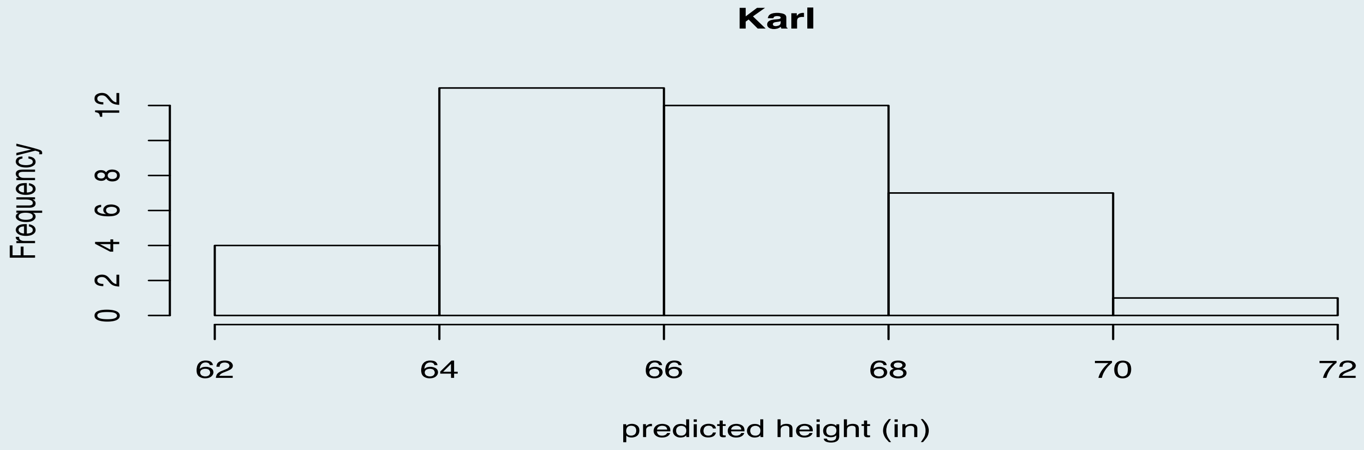
Question:

in a real-life event, Bin over-estimated the height of a distinguished white male colleague in another institution.

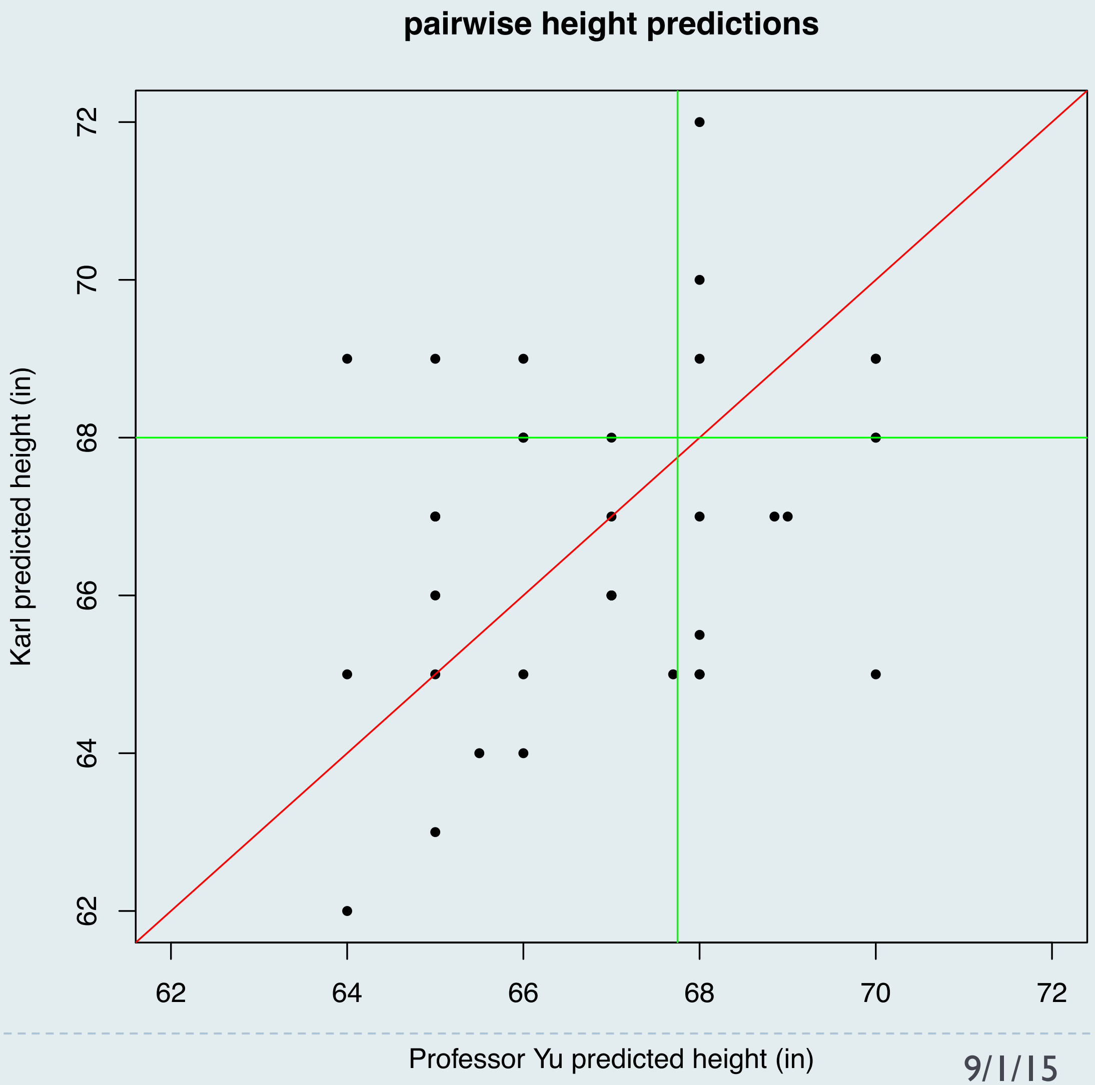
She wondered whether this is reproducible and a general phenomenon. So she collected the data in the first class on estimating her and Karl’s heights by the class.

Heights


Professor Yu
Frequency
0
4
8
12
62
64
66
68
70
72
predicted height (in)


Karl
Frequency
0
2
4
6
8
12
62
64
66
68
70
72
predicted height (in)

Heights in pairs



How to collect data to better answer the question?

- Class discussion

Reading

- Re-read Freedman Ch. 1, and Spench et al (2012), while keeping the 10 sets of questions in mind.
- Go over the slides, especially the ones not covered in class, and refresh memory on hypothesis testing