Science AMA Series: I'm Dr. Elad Yom-Tov, a Principal Researcher at Microsoft Research. I use Internet data to learn about health and medicine. AMA!

Elad$_Yom - Tov^1andr/ScienceAMAs^1$

[1]Affiliation not available

April 17, 2023

Abstract

Hello Redditors! I'm Elad Yom-Tov, a Principal Researcher at Microsoft Research. I am a Machine Learning and Information Retrieval researcher, and for the past few years my work has focused on using Internet data to study our health. Internet data are all those things that we create while browsing the web: posts on Facebook and Twitter, queries on Google and Bing, blogs, and other content. These data can teach us about aspects of medicine that are hard to learn about in other ways. A few examples include measuring the effect of mainstream media on the development of eating disorders, estimating the effectiveness of flu vaccines, detecting new side effects of medical drugs, and discovering how visiting a dating site can lead to catching an STD. My book on these topics, Crowdsourced Health: How What You Do on the Internet Will Improve Medicine (MIT Press) was published earlier this year. AMA, including questions you are interested in, and perhaps we can research together! I will be back at 11 am EDT (8 am PDT) to answer your questions, AMA! Edit: Folks, thank you for your being interested in this work, and for your questions. It was a real pleasure discussing my work with you. I'll check in later to see if there are additional questions.

the WINNOWER

REDDIT

Science AMA Series: I'm Dr. Elad Yom-Tov, a Principal Researcher at Microsoft Research. I use Internet data to learn about health and medicine. AMA!

ELAD_YOM-TOV R/SCIENCE

Hello Redditors!
I'm Elad Yom-Tov, a Principal Researcher at Microsoft Research. I am a Machine Learning and Information Retrieval researcher, and for the past few years my work has focused on using Internet data to study our health. Internet data are all those things that we create while browsing the web: posts on Facebook and Twitter, queries on Google and Bing, blogs, and other content. These data can teach us about aspects of medicine that are hard to learn about in other ways. A few examples include measuring the effect of mainstream media on the development of eating disorders, estimating the effectiveness of flu vaccines, detecting new side effects of medical drugs, and discovering how visiting a dating site can lead to catching an STD. My book on these topics, Crowdsourced Health: How What You Do on the Internet Will Improve Medicine (MIT Press) was published earlier this year. AMA, including questions you are interested in, and perhaps we can research together!
I will be back at 11 am EDT (8 am PDT) to answer your questions, AMA!
Edit: Folks, thank you for your being interested in this work, and for your questions. It was a real pleasure discussing my work with you. I'll check in later to see if there are additional questions.

CORRESPONDENCE:

DATE RECEIVED:
September 06, 2016

DOI:
10.15200/winn.147307.79837

ARCHIVED:
September 05, 2016

CITATION:
Elad_Yom-Tov , r/Science , Science AMA Series: I'm Dr. Elad Yom-Tov, a Principal Researcher at Microsoft Research. I use Internet data to learn about health and medicine. AMA!, *The Winnower* 3:e147307.79837 , 2016 , DOI: 10.15200/winn.147307.79837

Hi Elad!

I am happy to see you do this AMA. We have similar research interests. I used to work in Computational Social Science and now, I have shifted to Computational Life Sciences. I am particularly interested in this part of your study.

detecting new side effects of medical drugs

Here are my specific questions on these methods.

1. We know that Internet data - particularly social media - is noisy. How did you filter out the noise?
2. Are there limitations on types of drugs where you could find side effects? For example, I can guess it is much more likelier to find side effects for popular (regular consumption) drugs rather than unpopular (non regular consumption) ones.
3. How good are the methods to find side effects? In the sense, were the side effects found to have a meaningful impact? If so, do you have some quantification of this impact?
4. How do/did you validate your results?

Thanks!

PS - I plan to order a copy of your very interesting book. :)

denzil_correa

To your questions: 1. Internet data is noisy for several reasons, not least because we don't know the clinical state of each person. Social media has additional noise from the fact that people might not be entirely truthful if they can be identified by friends or family. For this reason, we sometimes prefer to use search engine query logs, which are less prone to this kind of noise (or bias). That said, in our experience people tend to be truthful on specific media and for certain conditions, so we should pick our data sources carefully. We also sometimes use text filtering algorithms to identify those posts that are of interest to us, namely, truthful and accurate. 2. Indeed, the popularity of drugs affects our ability to accurately identify side effects. Also, more acute side effects are less likely to appear in our data: People who experience a heart attack are (thankfully) more likely to go to the emergency room that they are to a search engine. 3. A lot of times we find a "ground truth" dataset, and show that our results correlate well with it. In the case of side effects, those can be the side effects listed by the FDA or the manufacturer. For influenza it might be the (retrospective) reports on the number of cases per week made available by CDC or other health authorities. Other times, when we figure out an interesting effect we might run a survey among people with the condition, and ask them to validate our results.

I heard recently that medicines often have a warning not to take them if pregnant or breastfeeding, not because they're inherently dangerous but because there are (for good reason) almost no drug tests done on pregnant or nursing women. Could your research program help fill in this gap?

iorgfeflkd

That's a really interesting idea. I'm no expert on the use of medicines during pregnancy, but I think you are right in that there is a knowledge gap, as well as a preference to err on the side of caution when prescribing drugs for pregnant (or breastfeeding) women. Unfortunately, I don't think our data can help filling the knowledge gap. If the drug is not prescribed to women, they will not take it, and so will not experience its side effects and search for them online. Our advantage is when people take drugs for a long time, begin to experience a side effect, and don't even realize that it's related to the drug. Yet, because we see so many people searching, if a sufficiently large number search for the drug (ostensibly, when it is prescribed to them) and them for the side effect, we can suggest that there is a link even though they don't realize one exists.

Dear Elad,

I'm super curious about how you manage with all the ethical compliance rules that limit research nowadays. Basically, in a hospital, we're forbid to run any analysis on any database, for any reason, even on public data. Are you free to collect internet data (which are public anyway), or do you have limitation?

I've read on a couple of IEEE spectrum articles that allowing free data mining on the databases we have could revolutionize medicine!

lucaxx85

Ethical compliance, and ethics in general, is a major concern in our work. We approve our studies with an Institutional Review Board (IRB). Beyond that, I would not want to compromise the trust of our users by doing something unethical. We take specific steps to minimize harm, for example, by hiding the identity of users. That said, I think that, as a profession, we're still grappling with what should and shouldn't be allowed. For example, most people will say that it's OK for a drug company to run an ads campaign online to market their drug, even without IRB approval. Should we be allowed to do the same for our research? I would say "no", but other people will differ.

What's the most interesting set of data that you've been able to glean through your research so far? How about the weirdest?

blanketswithsmallpox

It is hard for me to decide on a single most interesting dataset. The nice thing about my work is that every (successful) project has its own interesting dataset. There were quite a few weird ones, though: My colleague Dan Pelleg found us a dataset where teenagers stated their age and said that they were going to have intercourse for the first time, and wanted to get tips. This allowed us to learn about the age of first intercourse in the US population. More recently, I found a survey of about 200 questions that a person posted on a pro-anorexia site, and more than 800 people answered. I didn't know what kinds of research questions it would answer, but it was too good to skip, so I downloaded it and found a group of eating disorders specialists with whom I worked with to figure out how people on this site became better over time. Sometimes it's not weird, but perhaps scary: I wanted to measure how news affects people's intention to do certain things, both positive and negative. I found that many tens of thousands of people use search engines to ask how to commit suicide :-(

How do you sort through all the self diagnosis of aspbergers, autism, depression, eating disorders and other possible illness that people often assign to themselves without a true clinical diagnosis?

eatonsht

People tend to self-diagnose for a variety of conditions, not just those you mentioned. We have looked at a few of these cases. My colleagues (White and Horvitz) have called some forms of Internet self-diagnosis "cyberchondria" – that tendency people have to assume they have the worst possible condition for their (usually benign) symptoms.

We're currently trying to build tools that will help with self-diagnosis for some conditions, but there is a long way to go until these tools will be ready, even in the lab.

Hi elad! I am currently studying software engineering in ben gurion University in be'er sheva. I was wondering if you could give me some advice as to build my career after I finish studying and reach an r&d department like yours.

Thank you for the AMA

yosayoran

That's a hard one: I guess it is a combination of academic work and luck. My career thus far has been at industrial research labs, but I always kept publishing, and was very fortunate to work with some very smart, nice people. I think it's less what you work on then showing an interesting approach to the problem. I know that doesn't completely answer your question, but I've seen people with very different research backgrounds in our labs, so I guess there's more than one way of gong about it.

Hello Dr. Yom-Tov,

Are there any interesting differences you have noticed (in the Twitter data, for example) between diseases such as the flu, which develop quickly in a person and are transmitted quickly, and diseases that take longer to develop, such as HIV, or non-communicable health issues such as obesity?

I have read that, for example, that obesity can sort of act like a virus, such that if a person has obese friends, they are more likely to become obese. But I am interested in what you have found.

IndianSurveyDrone

That's a great question: In general, it is hard for us to see longer-term developments. Partly this is because many of the platforms (Twitter, Facebook, etc.) have not been in use for very long, and even then have morphed considerably over time. Partly it is because people change their identity, so it is hard to follow them over long periods. What we do know, though, is that there are big differences in the amount of information we have for conditions that have a stigma attached to them (for example, sexually transmitted diseases) compared to ones which do not (e.g., flu).

The research you mentioned (pioneered by Christakis and Fowler) is certainly something we should try and validate with our data, though we haven't tried it yet.

Hi Elad, thanks for the AMA!

Putting aside how interesting this sounds, what value does such research have to a software company like Microsoft?

Also, isn't there a conflict of interest between you publishing a book about your research while it's performed under Microsoft?

Thanks again!

jeepon

As our vice president, Harry Shum wrote, "At Microsoft, we believe in the value of research, for our own work and for the world we live in". I'm fortunate to work for a company that believes in long-term research that advances science, even if it has no immediate bearing on Microsoft products.

In the same vein, Microsoft is supportive of me publishing my book, as is of the publication of my research papers. Indeed, every year people at Microsoft Research publish hundreds (if not thousands) of scientific papers, and quite a few books.

How do you address the sampling issue with social media? Not everyone makes their posts public. Not every talks about their health on social media, even if the posts are private. These differences aren't random.

publishandperish

You're pointing to a very important issue in our research. If we want to say something about the general population, we have to de-bias our data. It's not every time that we have to do this, because sometimes all you want to do is to talk about the specific population you've analyzed. Frequently, de-biasing is an important part of the work. Moreover, as you say, there are things that people won't talk about, especially on identifiable social media. In those cases we have to turn to the places where they do discuss them, such as anonymous social media (for example, some medical forums) or search engines.

This seems like it depends quite a bit on invasion of privacy, how do you gain access to all this info in the first place?

What statistical modeling are you starting with?

auiovhe

Access depends on the type of data: Some public posts on social media (for example, specific forums, Yahoo Answers, and Twitter) are publicly available. This means you can easily obtain them (though check the terms of service first).

Search engine data is usually much harder to get at: Some companies buy their data from internet users and sell it. We have access to data from Bing, after appropriate anonymization to ensure privacy.

Hi Elad,

I'm submitting my PhD thesis in computational linguistics and healthcare in the coming weeks. In it, I point out that natural language processing, especially of online text, can have downstream clinical applications, and ultimately improve health outcomes. One of my supervisors, from a linguistic/qualitative healthcare communication background, finds my argument unconvincing and farfetched. It's fair enough---this supervisor simply isn't a computer person.

Part of the reason for the skepticism is that the (few) studies demonstrating the potential are very new, and sometimes a bit speculative or optimistic. Also, many are conference proceedings, which qualitative researchers find less trustworthy than books and peer-reviewed articles (though this is not really the case for computer science).

My question: I'm sure you've needed to convince a skeptical academic that machine learning, information retrieval and/or natural language processing can be used to improve health outcomes. How on earth did you do it?

I'm happy with either some rhetorical nuggets or references :)

Thanks for the AMA!

_normalverbraucher

I feel your pain...

As I've mentioned above, there are very few high quality journals that are read by both medical professionals and computer science people. There are also cultural differences: You pointed out to the conference\book division (CS people view a conference paper as good as a journal paper, whereas some fields only appreciate books).

I think the proof is in the pudding: If you show the medical people that you can reconstruct things they know are true, they are more likely to trust you when you show them things that are new to them. For example, in our study on adverse reactions of drugs, we had to show that we get a good correlation between our data and FDA data, before we could more further and show a new class of side effects. For those, we had to supply circumstantial evidence to suggest that they are real. Even then, the pharma company we worked with wanted to test our results for themselves before they felt comfortable with our data.

Good day, Dr. Yom-Tov ;)

With regards to tech usage and health, my mind quickly goes to the obesity problem in the US. What are you finding with regards to eating habits and exercise? What's holding us back and what does the

internet lead us to believe caused the overwhelming obesity of America?

Ookitarepanda

I personally haven't looked at obesity, but my friend Ingmar Weber (of QCRI) and his colleagues have begun looking at these issues. I think you'll find they have interesting insights into these issues.

Hi, Dr. Yom-Tov, reading one of your papers on vaccination (information is in the eye of the beholder) - I am in a similar field but approaching health misinformation from a health policy perspective. Our predisposition to opinions on health information may dictate our feed, our web queries, and subsequently exposure to information (false and not false)...in a nutshell, what future research do you recommend public health policymakers look into more to curb the spread of false information? Thank you.

justavg1

One of the takeaways from our work was that people will read the same piece of information and react to it differently, depending on their prior stance. Other researchers have shown that understanding doesn't equal likelihood to vaccinate.

I think it would be interesting to see if public health officials can find a way to write suitable content for different stances, and provide it in a way that matches content to stance.

Thanks for doing this AMA!

- How do you see this form of research affecting the traditional clinical trial method?
- Are you currently working with any pharmaceutical companies with this research?

joshmaxd

The best outcome would be if we could help improve clinical trials by working together where it's appropriate. For example, we recently helped Public Health England to validate that a pilot vaccination campaign they ran was successful, even though traditional means of measurement couldn't do this.

We have a few other ideas on such integration, but the gist of it is that we can provide a view that's rarely available to researchers using traditional data, and this view could certainly improve things like monitoring of adverse reactions.

We're working with one pharmaceutical on discovering adverse reactions of drugs. They approached us after we published our paper, and we were happy to collaborate. I won't mention their name because I'm not sure they would want me to, but you can look up our papers and figure it out…

What do you think about the recent study where researchers were able to diagnose depression from a person's Instagram stream?

Detaineee

I thought it was a really interesting paper. It seems very reasonable in light of Munmun de Choudhury's work with Facebook and Twitter data a few years ago.

NLP researcher here. I have been working on social media for a while and in particular I'm focusing a part of my work on the detection of irony and emotions in text. Irony is important because it can completely revert the meaning of something being said in the text. Have you ever been concerned by the use of irony in social media like Facebook, Twitter (and Reddit, of course)?

davidebus

Irony is indeed one of the problems we face when analyzing text. There are other similar issues: An example given to me by people at Treato is when someone states "I'm worried that drug X will cause me muscle pain". If you're not careful with your algorithm, you might deduce that drug X causes muscle pain.

Since you can only see and take what you read as face value how do you extract relevant data from the very superficial internet persona of your samples ?

Do you follow people as some kind of subjects ? Or are your studies tied more to a community as a whole ?

Razzmot

This really depends on the topic we are trying to research: Sometimes we really want to look at the biased sample of a specific group, for example, parents who are considering vaccinating their kids. Other times we look at data from the entire population. Even then, we still need to de-bias the data, because internet users are (still) not a representative sample of the entire population. Please also see my response above about the subject of anonymity, which plays an important role. Data on anonymous media is often more truthful than on media where people are easily identified. With regards to taking things at face value: We do have to filter the data, and we usually do so by training an automatic classifier to filter out those things that are of interest to us.

Dear Mr. Yom-Tov,

Are you planning on holding some kind of MOOC (massive online open course)?

redwarden

I haven't thought about it, but perhaps I should...

Are your works currently focused on English-speaking worlds only, or does it extend to other languages? If yes, how did you proceed? What challenges/interesting facts did you encounter?

helmstif

I've focused almost exclusively on English, and even more specifically the US. There are a few reasons for this: First, we have very good tools for processing English, and less so for other languages. Second, it is easier when you can read the texts you are working with, and I only speak English, Hebrew, and some Arabic. Thirdly, in the US market we have a lot of auxiliary data such as education, income, and measures of health which can be linked to users through their physical location.

Any thoughts or comments on the statement "Computers and coding/programming are the future of science and technology"?

And would you recommend that somebody specialize in their desired field *before* or *after* they've learned to code?

Sandusson

Programming is an important tool, but a tool nonetheless. Is it more important to learn statistics before or after somebody specializes in their field?

I'd say do it in parallel: Find a simple problem in your field that needs programming to solve it, and then learn to program while you solve it.

Hi Elad, thanks for doing this AMA. How can we avoid learning biases and arriving at unfair conclusions and solutions using such data as you described? How can we detect that we might not be accounting for certain biases inherent in internet data that could lead to us sacrificing performance for a particular demographic by favoring the more predominant demographics present in our data?

onto_something

I think that the only solution we have right now is to be very careful and very honest about the limits of our analysis. Several people at Microsoft Research (for example, Kate Crawford at our New York Lab) are thinking (and doing) important work in this area.

How do you differentiates facts tyat are not conequential.but are always together from actually consequential ones Example: when ice cream sales goes up drowning also goes up Probably because it's summer and people buy ice cream and drown during summer not because ice cream causes drowining

ThatBlackAndWhiteGuy

A lot of times we can't: All we can say is that there is a correlation, but not causation. At other times, there are "natural experiments". For example, we found that sometimes there is a spike in the searches for celebrities that are perceived as suffering from an eating disorder, and in parallel there are similar waves of mentions of these celebrities on social media. When we look at the people who searched for these celebrities, we find that some of them later developed an eating disorder. This could be just a correlation, except that the reason for these "waves" are usually newspaper articles, which are the trigger for searches and mentions. This kind of process (a news article, followed by searches for the person mentioned in them, and the later development of an eating disorder in a few of the people who looked at that article) suggests more than a simple correlation, and is a step towards causation.

How do you get what you know into the hands of the healthcare industry? Publications can be both slow and specific - what do you do if/when you learn something really important? What hill do you stand on, and which megaphone do you use?

kreel

As you write, we publish papers and try to get them in front of the healthcare industry. Frankly, there aren't too many venues that both Computer Scientists and medical professionals read.

The other way is to give talks at universities or companies, and (rarely) to call up the relevant

healthcare people and tell them what we have.

It's one of the reasons I wrote my book -- my hope is that it'll show people in the medical profession what we might be able to do together, and get them to collaborate with us.

Hi Elad! As a Bioinformatics student starting to learn about Machine Learning, could you give me some advice?

On the other hand, how usefull is health related search data from google and other platforms assessing trends on a population?

Hjorvik

You'll have to focus the question re:advice a bit more. One of the things I love about ML is that you can approach it in so many ways, not least from both practice and theory.

We are only beginning to gain useful insights from Internet data for public health, so the potential is still unrealized. I think it'll be a boon to medical research.

What social media outlet provides you with the largest wealth of information?

Also thanks for the AMA!

Tehrula

The social media I work with most are Twitter and Yahoo Answers. This is more because of availability than other reasons.

Hi Dr. Yom-Tov, thanks for taking this AMA.

- It seems you can find some interesting links between some activities and health issue Internet data. How do you know there are causality rather than correlation?
- I read a article which got a brunch time by analyzing the tweets. Someone make a [comment] about the article,

  My only issue is you don't know WHY people are tweeting "brunch". If you have more people tweeting "Heading out for #brunch!" it will skew early whereas if people are posting after the meal something like a picture or a "had a great time at #brunch" it will skew later. I guess some people do literally tweet during a meal.

  How do you deal with a similar issue like this?

JimballoonX

To your questions:

- Regarding causation, please see my reply to "ThatBlackAndWhiteGuy".
- Frequently, what we do is build an automatic classifier to distinguish between what we're interested in, and all the rest. For example, when we wanted to find tweets of people who said they had the flu (and were not posting general information about it), we first manually labeled a set of tweets that contained words about the flu, and then built at automatic classifier to do the same task. Once we do that, we can apply the classifier to as many tweets as we want.

Hi Elad!

Thanks for taking the time out of your busy schedule to do this AMA!

I am currently a PhD student and have been contemplating the importance of machine learning in big data analysis. Specifically, I have found that a lot of behavioral researchers in places such as Microsoft research, Googles research team etc. are using machine learning. How critical is it today to have a good understanding of ML to be able to join such research teams? If it is not critical yet, do you see it being a key tool to have in the next 5-10 years?

Behavioral_Econ

Because of the volume of the data that we have, it is usually impossible to manually process these data. For this reason, ML has become hugely important.

That said, it isn't hard to begin studying ML: There are excellent books and online courses, and lots of tools to try.

Hello Dr! I'm a recent physics PhD graduate looking for work and I'm seeing a lot of postings involving machine learning. Unfortunately, I missed that boat during my undergrad and grad work, so I'm wondering if you could suggest some starting points to learn about this exciting field?

Also, unrelated but in your role in information retrieval, what's your stance on open internet issues?

Many thanks!

jockey_tofu

I'd start with the MOOCs on ML. I haven't watched them myself, but several people have told me that they are a good intro. Also, there are several great books. My personal favorite (also because I've worked with one of the authors), is Duda, Hart and Stork's "Pattern Classification".

What's your GitHub account for your code?

mikaelhg

Don't have one. Sorry.

Hi Elad,

Is there any need to consider that data taken from the internet may be inherently biased? Or have you not seen that to much effect? The reason I ask is that I assume most internet data comes from a younger generation or users who are more likely to discuss health issues online (hypochondriacs for example). If so, how do you control for data biases like this?

postcollegethrowaway

Bias is certainly an issue with these data. Please see some of my responses above to this issue.

Though younger people feel more comfortable online than older ones, we find that over time this gap is closing, and we see a larger representation of the older generation.

However, when we want to say something general about a population, we still have to de-bias (see my response to "publishandperish" above).

Hello Elad,

Do you have any experience in working with autoimmune diseases? How might information retrieval from the internet relate to autoimmune diseases like Sjogren's syndrome and ankylosing spondylitis?

pkScary

I have not looked at these specific conditions. Is there something unique about these conditions that Internet data might possibly tell us?

Do you work with UW researchers? The work they're doing is interesting...

NoxInvictus

They are doing fantastic work, but currently I'm not working with then. Some of my colleagues at Microsoft Research are, though.

Hi Elad Yom-Tov!

Thanks for doing this AMA. I was wondering if you could answer the following:

1. Your research seems to involve lots of data from many different sources. How do you acquire the data (by scraping?), and are there any legal issues involved?
2. Could you elaborate on the machine learning methods you guys are currently using to derive predictions from your data?
3. I think that internet data can be (a) very noisy and (b) very difficult to extrapolate any sort of patterns on account of the variable interpretive nature of language, and the difficulties of assigning appropriate weights and targets. How do you overcome these trials?
4. What advice would you give a college student studying Machine Learning? Any course recommendations? Any other cool research topics?
5. What do you like to do in your free time?

Thanks!

LearningByTeaching

Thank you for these questions. I think you'll find answers to (1), and (3) above.

To your number 2: This is very much problem dependent. We've used everything from simple linear classifiers to random forests and Bayesian Networks.

1. I'd say work on problems that are of interest to you, and will benefit people (if you can afford it :-))
2. I love to hike, and have just returned from the Tour du Mont Blanc.

Good day to you too!

What do you think will be the most significant change to medicine caused by the Internet of things and/or new data systems and methods?

FILTHY_GOBSHITE

If I knew, I'd invest in the stock of these companies. More seriously, it's hard to predict the future, though I've pointed out to what I think are interesting directions below.

Thank you for your time. I have noticed the effect of the internet on homes and classrooms for children with Autism, did you happen to see any trends in your research? I teach 6-9 year olds with Autism, and as the internet has become more available to lower income people (the school I work at receives Title 1 federal money due to an extremely high number of low income families) I have seen an increase in knowledge amongst my parents over the last 10 years. I also mentor new teachers, and in the last 10 years we have gone from 9 classrooms for children with Autism, pre-k to grade 8, to 39 classrooms now, and we need even more as the ones we have are all full (10-14+ kids.) The amount of I formation available about teaching children with Autism has increased but as we hire these new teachers, they are not any better prepared. I think it is interesting.

ChompyGator

Thank you for this question. It's something that I haven't looked at, but sounds like a really good question.

We do know that many parents that suspect their children have autism go online to try to diagnose them (and we've looked at the quality of the answers they get). I do wonder how much Internet is helping in early diagnosis of this condition. It would be interesting to test it.

Shalom Dr Elad!

Where do you think we should put the line between allowing medicine research companies to sell new drugs at very high prizes to fund their researches and regulating the price so the drugs will be available to all the social classes?

Pokeputin

That's a hard one to answer, especially given the excesses we've seen recently. I think the arguments for and against are clear, but drawing the line is tough.

How did you get into this sort of job? Was medicine something you have always been interested in?

ClassyBrainiac

I guess I always had a soft spot for medicine, even though I don't think I have the skills to be a medical professional (I'm too squeamish).

My Ph.D. is in Electrical Engineering, and my thesis was on building Brain Computer Interfaces: machines which allow people to communicate using their EEG (“brain waves”). For a few years I after my Ph.D. I worked in other areas (but using the same tools). Then, when I joined Yahoo Research (and later Microsoft Research) and was told I could research pretty much whatever I was interested in, I was happy to go back to medicine. I really like the fact that my work might help people get better.

Do you believe the internet will become a much larger tool and indicator towards the general health and wellbeing of a populace?

HectorZeronie

I think the Internet will become a very important tool for medical research. It won't replace other forms of such research, but we now know that Internet data has specific advantages that we can increasingly put to good use. For example, whereas (if you're healthy) you see your family doctor once or twice a year, Internet data is collected 24/7, so it's a much more immediate sensor of your condition.

One of the reasons I wrote my book is to expose medical professionals to this tool, in the hope that they will use it more often.

How will your work impact evidence-based medicine?

leaveworldbetter

I can suggest a few directions. For example, if we think that certain information can cause people to vaccinate, it's easy to run a trial online and test if it's working.

Another idea is some of the amazing work John Brownstein and his researchers at Boston Children's hospital have done, using interventions on social media to detect gastrointestinal diseases.

I recently installed desktop BOINC in my computer. It claims to use the distributed power of al compjuters in which it's installed, when they're idle, to advance simulations regarding things like SETI, solving ancient Enigma coded messages, climate change... and a lot of molecular simulations regarding disease.

I know it's different than what you're doing, but... what would be your opinion on that? Do you think it's actually useful?

MarsNirgal

I love this idea, and as far as I know it's shown tremendous benefit. We frequently use distributed computation because Internet data is so voluminous, even though our computers are inside the company, so I appreciate the power of such approaches. For problems such as SETI, "crowdsourcing" computer power to people is a marvelous idea.

What's a discovery you made that you still find unbelievable and inconceivable, albeit you have proved otherwise via your unorthodox research?

MrKennethTong

A few years ago we showed that we could screen about 300 thousand behaviors, locations, and activities for their likelihood to precede disease. Some of these precursors that we identified seems pretty strange: For example, it turned out that people were more likely to browse certain dating sites before they developed an STD (and they were heterosexual dating sites when the STD was herpes and homosexual sites when the STD was HIV). We also found that people who had a heart attack were more likely to have visited fast-food joints before their attack.

It turns out that these associations were suggested before, but were hard to validate. Our algorithms picked them up without us knowing that they were suspected...

Is this more big data/data science or more health-oriented?

penatbater

It's both. One the one hand, I'd like to develop new data science (or machine learning, or information retrieval) tools, and on the other hand, I want our results to have impact on medicine. That's the nice thing about this area.

Can you speculate as to where this technology will be in five years, ten years and twenty years? And what applications exist now or might exist in the future. What's the next step?

theantirobot

That's a hard question. If you'd have asked me 5 years ago, I would never have guessed the state of our research today. There are some really interesting directions, though, that might show the way: I've pointed above to the use of Internet data for more personal medicine. In terms of public health, I think we will be using Internet data much more to get daily insights on public health and to monitor for outbreaks.

Hi Elad! Thank you for doing this.

How do you feel Microsoft Research is placed when competing with the likes of Google's Brain and DeepMind, as well as Facebook's FAIR programs?

Basically, do you feel Microsoft if well placed to bring out ML based commercial products to the public in the near future?

PulsingQuasar

My feeling is that Microsoft Research is a unique place to work, as one of the last remaining industrial research labs that do basic research.

As for products, I'm not really able to comment on this.

How does the process of getting health and medicine information from internet data work?

SaltySalads

There are some details in my responses above (and in my book), but in general, there are Internet data that are publicly available (using an interface or through crawling), and these data can then be filtered for their medical content using things like term matching or automatic classifiers.

Have you ventured into psychological effects of the constant connection between humans? I'd love to know your thoughts and/or findings on the matter: Does the constant connection between us and our friends via Facebook cause there to be less and less true friendships.

I've noticed that there are less plans being made between friends and more drama in young adolescent friendships than ever before!

Kem1zt

I have not looked at these questions, but I've read some research that tried to quantify these effects. I'll have to find the references, though.

How exactly would you figure out if a user visited a dating site and caught an STD?

I get that you could build an API for OKC or Tinder, but health records are always behind a HIPPA firewall.

Thepumpndump

The fact that someone has an STD can be directly identified if they queried, for example, "I have HIV. What are my options?", or in less direct ways, by inferring from their searches about drugs and symptoms. We can then go back in time and ask, how did this person's behavior change just before they developed the STD? If many people visited this dating site in the week or two before they developed the STD, it means that there is an added risk associated with this site.

This statistical method (known as the self-controlled case series) was originally developed to test the side effects of vaccines. We applied it to Internet data.

Can you layout new career pathways available to those trained in bio tech?

I have an MD and degree in biomedical engineering with comfort programming. I have a big data storage system in my basement, but I am repeatedly frustrated that my doctor colleagues, engineering colleagues, and software colleagues all operate in separate silos. I want to be the one that brings them together, but I have not found the best way to do this.

Thanks for this AMA!

fallsdownsometimes

As I've commented above, I feel very similarly that we need to break down those silos, but that's it's not an easy task. Partly it is because we all find it easier to live in our familiar environment, and partly because we have to show that these new data have real value. Computational biology, for example, is one area where this has already happened.

My limited experience is that we can collaborate when we build credibility with the medical researchers, thus making them more willing to test new ideas.

Hi Mr Yom-Tov, i will finish this year my master's degree in datascience and machine learning. Would you recommend to students like me to go on with a PhD or to start working now ? How much do fear that computers can (maybe ?) soon do our jobs ? Anyhow, it's awesome to hear about your research! I hope you all the best!

Kayjoue

My father (who is a university professor) always said that you should go for a Ph.D. only if you feel the need "burning inside of you". I tend to agree. However, the other option is to go work for a company for a few years, and then go back to academia.

To your other question: I (perhaps naively) don't fear computers so much. They will most likely change the workplace, but I don't think we will be out of jobs.

Do you see a point in the near future when a database that can dynamically respond to basic medical questions and ascertain contraindications to a medicinal combination will be available for free to the public? The internet has eliminated the scarcity of knowledge, and the data you are collecting is only made possible from the countless individuals that you gather it from. When will this come full circle? Or will it?

As for me, I would be highly interested to see if there could be endocrinological ramifications from the prescription of amphetamines or methylphenidate drugs during childhood and adolescence. I think there are some profound implications that haven't been properly explored that you would be put in the prime position of elucidating.

Debonaire_Death

It depends on your definition of "basic". Search engines are already pretty good at giving reasonable answers to very simple medical questions, or, as you mention, find contraindications of drugs.

However, they are still limited, for example, by the description given to them by people. One of the things medical personnel are trained to do is to ask the right questions. I'm therefore less convinced of the ability of computers to replace doctors.

I think your questions about amphetamines is one we could try to answer. I'd start by looking at sites such as Yahoo Answers.

when will be he able to use data from our health to show what or when in the future we will become sick of.

endoredo

There are some beginning of this kind of work. My colleagues have recently published work where they showed they could identify pancreatic cancer according to people's searches before they knew they had the condition.

how can I break into the world of health informatics as a nurse?

wooder32

My answer would be, find a problem that's close to your heart, and that you think Internet data could solve. Then, collaborate with someone from the Computer Science side to try and answer it.

Hello! What kind of project or question would you recommend to someone who is just getting into machine learning as a discipline and wants to get some hands-on experience?

B12_D_Serotonin

Something that you find interesting. There's a lot of publicly available data that you can use, so it's more a question of finding a problem that interests you.

Hello! What kind of project or question would you recommend to someone who is just getting into machine learning as a discipline and wants to get some hands-on experience?

B12_D_Serotonin

Please see my reply on the AMA. Generally, I think you should find a problem that's interesting for you, and then find the data that's out there to solve it.

Picture this.

All US health records are stored in a central database run by a hypothetical national health organization. All Electronic Medical Records are standardized and link to this national database. For security reasons the patient information and clinical records are kept separate and can only be joined with a biometric key.

Would this method of clinical data storage spark an incredible age of healthcare advancement? The research potential of this kind of data system would be incredible in my opinion.

I work in a clinical setting and getting and maintaining records with today's emr options is just so incredibly backwards.

underwatr_cheestrain

I think it would. In Israel, datasets like this exist for the past 20-odd years for the two main HMOs, and even though Israel's population is small, compared to the US, these data are a huge boon to research.

measuring the effect of mainstream media on the development of eating disorders

discovering how visiting a dating site can lead to catching an STD

What do you think about the proverb 'correlation doesn't imply causation' ?

Charlemagneffxiv

Please see my replies above. We had to work quite hard to even begin to go beyond correlations. In the first of the cases you quoted, for example, it turned out that we had a natural experiment, but that's rarely the case.

Hey Elad,

I'm a recent biochemistry graduate with a strong belief that computer siencewill change the medical sciences forever. I am currently studying for a career in medicine tinged with some computing.

1) What were the steps you took to get into your position?

2) What are developments that you expect in your field in the near future that you think would surprise society?

3) How do you deal with the limitations of non-targeted data gathering, and of using data sets that are not your own?

Thank you for this AMA, much of it was very interesting to read!

arvendragon

I'll try and answer each:

1. Since my Ph.D. I worked in industrial labs, but kept publishing papers and being active in academic research. I was also very fortunate to work with some amazing people, who helped advance my

career. That said, there's a fair bit of luck involved.

2. When I hear a scientific talk, and the presenter states the problem, I like to think, how would I solve this? If the solution is different than the one they developed (or if I can't think of one), it's surprising to me. In that vein, I think people would be surprised that an app can aid in medical treatment, for example, by predicting a mood disorder event, or that we can identify that there are new strains of a flu virus going around just by the way people tweet.

3. There are two questions here: First, a lot of our work is filtering the data to get at the (small) subset that we're interested in. The second are the legal aspects of using other people's data. For that we have lawyers...

In other words you mine people's medical records for information you can sell to third parties? Right?

magicrat69

No and no. We don't mine people's medical records, only Internet data. And I don't sell what we find -- I publish scientific papers.

I don't like Microsoft. I have friends who work there who don't like Microsoft. Why did you choose them to develop your investigative project? Do they have something you can't find anywhere else?

juanjodic

For me, Microsoft Research brings together a superb mix of a great people, amazing data, and a company that's supportive of long-term research. However, YMMV :-)

I'm currently a Grad Student studying Applied Economics with a focus specifically in Health Services.

My question for you relates to the data and it's applications in areas such as outbreak detection. If we use the current Zika issue, is there any type of advanced analytics that could have possibly identified certain aspects of this earlier? For example, were people searching the symptom list more frequently in Bing / Google in the weeks leading up to the outbreak being identified through clinical processes?

It seems like there would be data indicators prior to the actual reporting mechanism that the healthcare community could aggregate.

thank you for your time!

cybersamurai

Great question! We've seen examples where Internet data can help in disease detection (for example, flu and gastrointestinal diseases) and others where it can't (I've looked at ebola). There are some things that can help us, for example, if it's very localized or it's common to people who were at the same place at the same time (we've looked at the stuff people contract at rock festivals...). Naturally, people also need Internet access for us to have any data.

Generalizing from these few examples is hard, but my feeling is that we can do better than traditional medical tools if the condition is relatively widespread and is relatively benign in the sense that not every person with the condition will see a medical facility. If it's a condition that very few people experience, it'll drown in the noise. If it's fatal, it'll quickly come up in medical records.

Therefore, I'm doubt we could have done much with Zika. However, I would like to see if we could

have detected the Flint lead crisis earlier.