Science AMA Series: I'm Andrey Rzhetsky, professor at the University of Chicago. I study big datasets—like 150 million patient records to find links between autism and environment, or all of PubMed to find diseases that we should be investing more resources in. AMA!

Andrey$_R$zhetsky[1]andr/ScienceAMAs[1]

[1]Affiliation not available

April 17, 2023

Abstract

Hi Reddit, My name is Andrey Rzhetsky, professor of medicine and director of the Conte Center for Computational Neuropsychiatric Genomics at the University of Chicago. I'm interested in how genetic and environmental factors combine to make people sick. To study this, I use mathematical modeling to look for insights in very large collections of data. For example, I've analyzed 150 million U.S. electronic medical records (all anonymous, of course) to examine the links between autism and environment. We found hotspots across the U.S., which we hope will give us insight into causes and prevention. We are now looking at specific environmental factors that might have correlations to autism, as well as in other diseases like bipolar disorder. We also apply large-scale data modeling to other topics. For example, my colleagues and I recently text-mined all of PubMed to identify research strategies that might maximize scientific discovery and return on public investment. You can read about some examples of our work here: http://www.cbsnews.com/news/more-evidence-environmental-exposures-contribute-to-autism/ http://sciencelife.uchospitals.edu/2015/09/15/improving-the-allocation-of-biomedical-research-resources-with-big-data/ http://www.ci.uchicago.edu/press-releases/scientific-research-conservative-could-be-accelerated Thank you very much for wonderful questions. Apologies for not answering all of them! I have to go now, but thanks Reddit, it's been a lot of fun!!"

the WINNOWER

REDDIT

Science AMA Series: I'm Andrey Rzhetsky, professor at the University of Chicago. I study big datasets—like 150 million patient records to find links between autism and environment, or all of PubMed to

ANDREY_RZHETSKY R/SCIENCE

Hi Reddit,
My name is Andrey Rzhetsky, professor of medicine and director of the Conte Center for Computational Neuropsychiatric Genomics at the University of Chicago. I'm interested in how genetic and environmental factors combine to make people sick. To study this, I use mathematical modeling to look for insights in very large collections of data.
For example, I've analyzed 150 million U.S. electronic medical records (all anonymous, of course) to examine the links between autism and environment. We found hotspots across the U.S., which we hope will give us insight into causes and prevention. We are now looking at specific environmental factors that might have correlations to autism, as well as in other diseases like bipolar disorder. We also apply large-scale data modeling to other topics. For example, my colleagues and I recently text-mined all of PubMed to identify research strategies that might maximize scientific discovery and return on public investment.
You can read about some examples of our work here:
http://www.cbsnews.com/news/more-evidence-environmental-exposures-contribute-to-autism/
http://sciencelife.uchospitals.edu/2015/09/15/improving-the-allocation-of-biomedical-research-resources-with-big-data/
http://www.ci.uchicago.edu/press-releases/scientific-research-conservative-could-be-accelerated
Thank you very much for wonderful questions. Apologies for not answering all of them! I have to go now, but thanks Reddit, it's been a lot of fun!!"

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

DATE RECEIVED:
April 16, 2016

DOI:
10.15200/winn.146072.24664

ARCHIVED:
April 15, 2016

CITATION:
Andrey_Rzhetsky , r/Science , Science AMA Series: I'm Andrey Rzhetsky, professor at the University of Chicago. I study big datasets—like 150 million patient records to find links between autism and environment, or all of PubMed to , *The Winnower*

Here's the article (open access) that Dr. Rzgetsky published, which found 'autism hotspots' across the U.S.

The biggest risk factors for autism in a region seem to be:

- congenital malformations of the reproductive system in males (an increase in ASD incidence by 283% for every per cent of increase in the incidence of malformations)
- non-reproductive congenital malformations (31.8% ASD rate increase)
- viral infections in males (19% ASD rate increase)

Dr. Rzhetsky uses congenital malformations as a proxy marker for environmental risk factors. The idea is, although it is not known what causes these malformations, there presence hints at some systemic environmental risk factor(s)?

My question for Dr. Rzhetsky: now that you have had time to reflect on this data more, can you speculate as to what some of these 'environmental factors' may be (fully recognizing this is still

3:e146072.24664 , 2016 , DOI: 10.15200/winn.146072.24664

speculation)? How much do you think these environmental factors may differ from region to region?

My second question is about the robustness of this methodology (i.e. linking congenital malformations to environmental risk factors). Based on your data, you presumably have a map of congenital malformation hotspots in the U.S.. How well does this map correlate to diseases and disorders with known environmental risk factors? How well does it correlate to things like Type I Diabetes, Crohn's Disease -- diseases which are increasingly speculated as having environmental risk factors?

SirT6

1. There are a number of known factors increasing birth malformations, such as plasticizers, estrogen analogs, certain organic solvents, environmental lead, etc. Every year there are uncountable new chemical compounds introduced to the consumer market, for example, to ensure that your stove or bathroom tiles are cleaned with the minimum effort. Unlike medications, these new chemicals are not subject to mandatory safety testing.

2. This is a fair question. The ultimate test of the methodology would be finding and confirming THE actual causal environmental factor, such as an environmental chemical. Regarding type I diabetes and Crohn's disease, they could be environmentally driven, but the causal factors need not be the same as for autism. So I am not convinced that this is the best test of the autism-environment association.

Dr. Rzhetsky,

What platform(s) do you use in your research and is there a specific one that is most ubiquitous in the realm of big data analysis?

SpudFlaps

There is a whole toolbox of hardware and software tools that are useful for different aspects of applications: supercomputer or distributed clusters on computing side, Hadoop for large dataset processing, Python and R for exploratory analyses of data, and C++ for heavy-duty computations. Some applications require other languages and commercial software (such as STATA, SAS, SPSS).

Chronic fatigue syndrome / myalgic encephalomyelitis (CFS/ME) is said to have impact on American economy (due to lost wages and social welfare) measured in billions of dollars, with estimates of people being severly affected ranging from 800,000 to 2,000,000 in US alone. And yet it has been receiving one of the lowest research funding among all conditions (less than male pattern baldness, for example).

Do you think your research can help allocate research funds according to financial and disease burden, rather than popularity of the illness?

Soktee

Allocating research funds is not my job! :) Although, I do believe that modeling should guide funders towards more efficient reduction of disease burden. For example, see our recent paper:

1. Yao LX, Li Y, Ghosh S, Evans JA, Rzhetsky A. Health ROI as a measure of misalignment of biomedical needs and resources (vol 33, pg 807, 2015). Nature biotechnology. 2015;33(8). PubMed PMID: WOS:000359274900013.

Thanks for doing this AMA. I'm curious, when you say you've identified hotspots regarding autism, is

that a link between the diagnosis and place of birth or their current location? I work with families of children with autism and know that many will try relocate to different states/regions that offer more comprehensive services for their kids.

StudentII

For the majority of kids the place of birth is not very far from their current location. With the insurance claims, in most (but not all) cases, we can trace both.

Hi Andrey

What data should be collected in medical records that currently is not?

Thanks!

napolitp

Ideally, we would know as much about individual history (food, places visited, life habits) and family history as possible. Honest answer is that we don't know which additional data types would be critical in the near future. Think of asbestos, DDT, or lead paint -- all of them were considered absolutely safe a few decades ago.

Dr. R - fascinating research! Assuming that you can correlate these hotspots to environmental conditions (nearby factories, mines, chemical plants, etc.) What happens next?

As a father of a child with autism caused by a particular genetic anomaly (15th chromosome), I've often wondered if his mother or I were exposed to something early on that affected his development or if this was truly a denovo presentation.

Anyway, I look forward to seeing the work develop!

kashakesh

Ideally, we (as a community) would try to develop approaches minimizing the harmful environmental influence, such as banning DDT and asbestos.

On the second point, nearly everybody in the research community agrees that the disease is a joint action of genetic predisposition and environmental exposures. It is quite possible that, even give an unfavorable genetic predisposition, one can avoid the environmental stimuli that would trigger the disease. Of course, we should first identify these stimuli.

Given the level of data that you're analyzing, what methods and equipment are you employing? Specifically, are there any commercially available programs being used or is this a 'home-grown' application?

Also, do you think the growth/development of AI and deep-learning will aid you in your future research?

Thanks!!!

broadnax

I commented on tools above. About AI and deep learning -- absolutely! I am thinking of these tools as analogs of powertools in home improvement.

I know that larger sample sizes are generally better, but I've heard that one problem with extremely huge samples is that you'll always be able to find statistically significant correlations, even if they're spurious. Is this true? If so, what sort of statistical adjustments can be done to control for this fact?

DNASnatcher

It is true, except "spurious" means here that you can see real significant trends, some of which are not very interesting. For example, the apparent rate of disease reports drops several times per year -- in the vicinity of holidays. This is a real effect, but not biomedically interesting.

Can you tell me how you arranged access to so many records? Is this all EMR? What did you do to anonymize the data?

Wastedmindman

The datasets are mostly electronic medical records (EMRs), but some of them include genotypes and other information.

Accessing data could be a monograph (or novel!) on its own. For many datasets, especially in the other countries, the analyses were done "blindly" -- I have never seen the actual data, but researchers, say, in Denmark used my programs to run analyses on their data formatted according to a standard hat is shared by all collaborators.

The US insurance claims were anonymized by a company that licensed the data to us.

Thanks for doing this AMA, Dr. Rzhetsky. I have two questions

1) Working with big data like this, did you have a specific hypothesis to test or was it a discovery project at first from which arose questions and hypotheses related to environment and autism?

2) As you are examining research strategies across all the publications on PubMed, what is your take on the value placed on p-values by researchers? How do you account for over interpretation of results and therefore, the success of an applied research strategy?

bowlofpetuniass

1. Typically, my colleagues and I start with a specific hypothesis. However, the alternative (large-scale scanning for factors affecting a disease) is also a legitimate approach. It is used, for example, in genetics where scientist scan whole genomes for association with a given disease (such as autism or diabetes), comparing healthy people (controls) with people with disease (cases).
2. Well, the simple model that we used does not account for these nuances. It looks at*choices* of subjects to study. In this naive model the success is defined as discovery of true relation (such as chemical reaction) between two chemicals. The assumption is that the peer review process does its job in filtering out spurious associations.

Do you think there would be a benefit in creating one central patient database and means of access for all healthcare facilities and providers public and private in the US, instead of having dozens of different poorly made EMRs and a plethora of locally hosted databases.

The research potential seems limitless.

underwatr_cheestrain

Oh, yes, but it is very hard to get there. This would probably require decisions on the government level.

What are the implications of wearable technology (step counters, heart rate monitors, SpO2, sleep trackers, etc...) on the kind of data focused research that you do?

johnmflores

Unfortunately, right now, the datastreams (clinical records and output of wearable gadgets) are disconnected. Potentially it is a fantastic opportunity to record high-resolution behavioral patterns of millions of people, which can be then modeled for causal effects.

Dr. Rzhetsky, Thank you for doing this AMA. I have two questions for you. 1. Autism spectrum disorders are almost certainly over diagnosed. Even if an incorrect diagnosis is later corrected, it still shows up in medical records alongside correct ASD diagnoses. How do you manage this source of error? 2. I've long hypothesized that some sociodevelopmental factors (parental availability, emotional accessibility of parents) must relate to ASD risk, but I've never seen good studies on this. What kind of work is being done on this front?

d_paulson

1. Keep in mind that we can see a "trajectory" of diagnoses. While false-positives are possible, it should be possible to distinguish the "corrected" diagnosis from the confirmed one. But I do acknowledge that this is a real problem.
2. There are numerous studies on the subject -- the oldest theory of autism (which is largely rejected) is a "refrigerator mother" -- emotionally cold parenting.

Junior researcher from Norway here, PhD in economics with a focus on applied statistics and currently self-studying Andrew Ng's on-line Machine Learning class. I'm interested in studying the interaction between nature and human health. My question is, is there a way for me to join the research efforts at your lab?

ent0

:) Sure -- please e-mail me directly.

I have a general question about your experience making a career in the sciences.

How did you decide on what field you wanted to end up in and for what type of purpose (government, industry, academia, etc). I am currently a undergraduate student studying microbiology and molecular, cellular, and developmental biology as my majors wrestling with trying to find a career path. I've found plenty of areas that are interesting and that I would be okay with investigating, but I am looking for insight on finding a "passion" to drive my career like everyone on a college campus speaks about. Hearing stories from experienced scientists about their career paths is helpful to me to get a better idea of the sorts of possibilities available.

Ry2D2

My two penny-advice (apologies if it is trivial!): I would use the following guidelines in selecting a field.

(1) Ideally, it would be a young field -- it is much, much harder to succeed in the old/mature fields and to find an important problem that was not attacked millions of times before; (2) It should be something that you personally care about and enjoy working with, (3) It would help if the subject of the study could be potentially (positively) impactful for the society.

Hi Dr. Rzhetsky, What Large-scale data models do you use to analyze the datasets? Also what framework/languages do you use for computation (R, Hadoop, etc.)? I am currently studying Computer Science and would really like to get into Big Data.

DaddyDays

All of the above -- see my earlier reply.

Dr. Rzhetsky – How can donors, small and large, support your work?

With the Illinois budget crisis affecting hospital reimbursements and slowed market returns on endowed funds, have you had to spend more time applying for grants?

Here's an article from St. Louis Public Radio for more context.

b9918

Well, this is another sensitive subject. Indeed, federal funding is becoming tougher to get and the university faculty are spending much more time writing proposals (and getting frequent rejections). On the positive side, this situation forces us to think through research design more carefully, so the time is not wasted.

When you say 150 million electronic medical records, do you mean inpatient discharge data? Or do you also have outpatient data as well? What year(s) of data are you using?

Jrizzler

Both inpatient and outpatient. 2003 to present.

Thanks for doing this AMA Andrey. I have two questions:

1) How likely do you think it is that genome wide association studies can help identify therapeutic targets for all illnesses, from to developmental to psychiatric to intestinal diseases?

2) As a med student, I was wondering how you got into the world of big data and computational modelling? Do you still have opportunities to work on wards and practice as a physician?

benjaminreddit

1. GWAS is extremely important technique and corresponding studies made enormous contribution to our understanding of disease-predisposing variation. However, GWAS cannot answer all questions about disease genetics because (a) it is focusing only on common polymorphisms, rare variants are invisible for GWAS, and (b) it essentially focuses on one locus at a time. So additional synergistic approaches (such as whole-genome sequencing) will have a lot to contribute in the near future.
2. I had a strange research trajectory :) -- being a PhD, I am, unfortunately, unqualified to see patients.

Dr. Rzhetsky, my deepest thanks for taking the time to do this AMA and to you and your team for this particular area of study. My sister was autistic and died of a seizure over a decade ago. We donated her brain for research in hopes that one day, someone like you may discover something. It gives meaning and purpose to our heartbreak, which my family and I are so grateful to have.

A question from my mother: Is Las Vegas (or southern Nevada) a hotspot?

Two questions from me: With such rapid technological advancements we've made as a species, could Autism Spectrum Disorders be an evolutionary change? What are my genetic risks/responsibilities having a sister with ASD (I'm over 30 and have avoided having children for this very reason).

Thank you again!!!!

Laboitedepute

1. (I had to check the map.) The rate in Las Vegas is increased, but not the highest in the country.
2. I don't think that this is an evolutionary change. The changes are happening too quickly for genetics -- evolution takes numerous generations. It is possible that our non-specific genetic burden grows, but environment clearly plays an important role in change of prevalence of many diseases.
3. All that we have for autism is heritability estimates (which are high, but depend on numerous assumptions and a specific genetic model). I don't think anybody can responsibly predict risks in your specific case (the risks depend on your and your spouse's genetics and live style). On the personal level, having children is so large part of personal and family happiness!

Thanks for doing this AMA. If you were a novice and only had time to take one or two data science classes (and had a minimal background in math-basic stats and calc), what would you take to get the most bang for your buck?

horizoner

This is a hard one. The more computational classes you get, the bigger is your toolbox for attacking the real-life problems. At the very least, I would take something like R and Python programming, data mining and probability theory.

Have you ever seen a link between vector-borne infections like Lyme disease or Bartonella and Autism? Do the hot spots correlate with the hot spots for these vector-borne infections?

liketosee

To my knowledge, these infections are under-documented; would be exciting to test if better data happen to be available.

With such large data sets, what method do you use to determine which observable factors deserve your attention?

the_one_54321

With huge datasets it is feasible to use in modeling *all* observable factors; what keeps me awake at night is *unobservable confounders*.

How do you aggregate that much data? Do you have an army of students who comb through research articles? Do you make your own software to scrape the relevant data from databases?

I'm in awe of the sheer volume.

Frodo_Fragg1ns

It is combination of several factors: (1) there are companies that make their business aggregating and selling data; (2) scientific community working with hospital databases and other diverse data is open to collaborations and typically pleasure to interact with.

Fascinating research. I was wondering how you take into account different regional awareness among clinicians and parents that might lead to different diagnosis. I would have thought this would be a big problem with a disease that has varied definitions and such a major public interest.

redteddy23

Ideally, we would have this captured with a real-valued variable for each county... We don't have such variable measured and use state-mandated level of diagnostic rigor as a proxy.

What open source tools do you use during the course of your work?

ANTIVAX_JUGGALETTE

Too many to name all of them -- including Hadoop, numerous Python modules and R packages, open access systems, etc.

How much experience do you have with spatial epidemiology? I'm a senior in college studying Geographic Information Science, and I'm leading a team on a capstone project studying the spatial, temporal, and socioeconomic patterning of a certain social disease in a major urban area. We're working for the county Health & Human Services department, and it's pretty exciting.

We're scrubbing the personally identifiable information from the data by performing analysis at the Census Tract level, and trying to develop a predictive model for the disease based on census data (race, income, education, population density, employment, etc).

I'm curious about what demographic factors you'd look at in an epidemiological study, and which methods of analysis you find most useful. I'm using Getis's G and Moran's I clustering analysis, and Multifactor ANOVA and Multiple Regression for our quantitative analysis.

And finally, how useful do you find spatial data that's not mapped at the case-level? Dancing around HIPAA feels like it's having a large impact on our analysis (e.g. no point clustering or useful network analysis).

chiliedogg

It sounds like we should discuss this offline -- you are welcome to contact me directly. The full answer would be long!

Thanks for the AMA, Dr. Rzhetsky. Really interesting research.

I am a graduate student specializing in analytics and after undergrad I worked in an immunology research lab that was studying the potential link between autoimmune disease and neuropsychiatric disorders. What are your thoughts regarding pediatric autoimmune neuropsychiatric syndrome (PANS) and its manifestation of disruptive behaviors? Do you suspect there might be a some misdiagnosis of psychiatric disorders when the actual cause may be an autoimmune disorder?

The jury seems pretty divided on this, but the idea that your child might have a treatable condition caused by some immunological malfunction is pretty seductive to a lot of parents whose children have autism.

mean-sharky

Very important research topic; I have not done much on this subject and therefore should not express opinions, but I would keep an open mind.

Thank you for doing this AMA! I have a question about Zika and Microcephaly, which I know isn't your specialty. There have been recent outcries from doctors in Brazil that we need to give more funding and research into the possible connection between increased rates of Microcephaly and the excess use of the insecticide Pyriproxyfen in water treatment. Do you think it's possible to do such an analysis and is it worthwhile? I've researched into how Pyriproxyfen functions and it looks like it is completely possible for it to be the cause of Microcephaly due to its ability to inhibit hormonal releases in mosquito larvae during gestation. Should there be more research into this topic or is it completely unfounded?

octaviousprime

I think, such analysis is possible and absolutely necessary (actually, it is being done, I believe).

How much consideration is given to return on investment when considering disease states to invest more resources into?

How much consideration is given to patient quality of life when deciding which disease states to invest more resources into?

I_AM_BISON

Are you asking about the current practice or how it should be? There only limited number of diseases with estimates of loss of quality of life (the World Health Organization conducting such studies, look up DALY and related measures). Currently, dynamic estimates of disease burden are rarely used in funding decisions; research trends and "hot" topics play significant role.

Hello doctor I always dreamed of a master modeling system that could take disease rates, crime rates, socio-economic demographics, and mental health funding to see if some forecasting can be made in future crime rates for locations in our country. With all the cuts to that division of medicine and the increase in crime and imprisonment I'd hope it could be used to show policy makers the importance of mental health. Have you ever thought or worked on such a modeling system. Great work btw, thank you for sharing.

djb85511

Yes, I did think of it, but never worked in this direction. Very hard and fascinating problem.

What is your opinion on the biopsychosocial model and the interpretation that Gabor Maté has of autism? Basically that autism is for the most part an emotional problem caused by interaction between parents and their children.

His interpretation on autism: http://drgabormate.com/article/autism-is-the-child-of-social-disconnection/

Splitje

As I mentioned above, this resembles the older "refrigerator mother" theory. But I am not familiar with this theory enough to pronounce judgements.

What effects do you see data science and big data technologies having on the medical industry in the next ten years? Will they become the goto technique when it comes to medical research?

ilia10000

"It is hard to predict. Especially the future." I see all necessary ingredients for this happening (talented people, necessary data, techniques and infrastructure).

I apologize if this question is a little technical or "inside baseball".

I work in autism research, so the work you're doing on autism is fascinating to me. However, my question pertains more to working with big data. I manage the data for my lab, though we work with a tiny fraction of what you do, only a few thousand records. I'm fascinated with data management and application, particularly in research and healthcare.

However, my university education was insufficient in teaching me much of anything about data management and database construction. Luckily, I'm a big fan of self-instruction.

Do you know of any good resources that I can use to learn about principles of big data management and proper database work? I'm also interested in learning how to code so I can build databases and analyze datasets without relying on paid service or finicky programs like SPSS or Microsoft Access. Any recommendations in this regard? Such as languages to use or software to learn?

Thanks for your time.

JetJaguar124

Well, "big data" is a generic buzz term, the real data and computational needs come in many shapes. To make sure that your needs are met, I would recommend talking to a CS faculty in this are who would help to narrow down database options: the proper answer depends on specific of your tasks, how much computation and data you are planning to handle.

I'm a noob getting into the data analytics sphere. How easy is this maths/stats method to pick up? Would one want to have a good background in a quantitative field, or is it mainly just plugging data into pre-built models?

Buttezvant

I think, there are currently open niches from pure method development to pure third-party-methods analysis.

Did you use GIS in your research?

Tombre

So far, we were getting by with Python packages handling GIS-like jobs. Doesn't mean that we won't in the near future!

How do you get access to such large volumes of patient data? How is this not prevented by privacy restrictions?

clearing

There are special protocols governing access to patient data. A researchers has to convince a special panel (Institutional Review Board) that the study is legitimate, it is potentially beneficial for the patients, it does not hurt the patients, and the investigator is trained in ethical conduct of such studies.

How do you justify that you aren't data fishing?

Hydro033

I don't know what that means. Is whole-genome association analysis is an association-fishing?

Hi Dr. Rzhetsky,

I'm a graduate student studying cancer immunology, but I have always been interested in bioinformatics. Are there any resources you could recommend to a graduate student who would like to learn how to perform big data analyses?

Thank you!

Jenna_bird

I would recommend to look at books (both technical and popular), such as "Automate this" by C. Steiner, "the signal and the noise" by Nate Silver, and "Super Crunchers" by I. Ayres -- there numerous others.

I'm a student of Pharmacoepidemiology and UChicago grad - very cool work, Andrey! My question is this: as the strengths of AI and machine learning are brought to bear on big data research, how will the role of researchers evolve? In my view, I believe programs will take on more and more of the analytical tasks, leaving researchers largely to think creatively and focus on patients, not on methods and statistics.

For instance, in my work with regression modeling in SAS, I am honestly struck with the huge potential for automation involved in my projects. In theory, I should be able to just tell a siri-like assistant which DV's and IV's to examine and then a program could spit out the final model, interactions, and so forth. Do you think this sort of program is in our future? I've heard of macros but why even are those necessary? I think Microsoft, Google or Apple could pretty easily automate such processes, with their knowledge in these areas.

IsomerSC

This is a bit depressing topic: yes, it looks like in computer-human collaboration humans are gradually

becoming less important. We had a lovely discussion on this topic in *Science* a few years ago:

1. Leonelli S. Machine science: the human side. Science. 2010;330(6002):317; author reply 8-20. Epub 2010/10/16. doi: 330/6002/317-a 10.1126/science.330.6002.317-a. PubMed PMID: 20947745.
2. Evans JA, Rzhetsky A. Machine Science: What's Missing Response. Science. 2010;330(6002):318-20. PubMed PMID: WOS:000282986700016.
3. Evans J, Rzhetsky A. Philosophy of science. Machine science. Science. 2010;329(5990):399-400. Epub 2010/07/24. doi: 10.1126/science.1189416. PubMed PMID: 20651141; PubMed Central PMCID: PMCPMC3647224.

I think this is a great concept, but my question is more of a, "rain on the parade" one. While identifying diseases that should be looked at more closely by researchers sounds great, will the money follow? Currently funding seems to come in two flavors; "popular diseases" such as as cancer, or chronic and manageable conditions, such as diabetes.

The former makes sense, if the public is all riled up about a disease of course funding will flow towards it. The latter is money coming from, or being used by, drug companies. It seems simply identifying new diseases that should be researched wouldn't guarantee the funding would be made available for that research. Particularly if it's a disease or condition that can be cured rather than managed. Do you think this will change as time goes on or is it a current concern for you as well?

lurpelis

This is a serious problem -- ideally, policy makers and funders would pay attention to what naive and not very naive modelers have to suggest. How to get there -- beyond my expertise.

As someone who routinely works with large data sets, it is becoming clear that managing, processing, and analyzing the "new wave" of big data is a unique challenge. On the software side of my field, common image processing toolsets don't really support the latest hardware or distributed computing methods. Also, institutions are struggling to find fiscal strategies for storing and preserving large datasets, and researchers are beginning to question whether they should be required to save raw data for the better part of a decade due to logistics. What is your view on big data, and what strategies have you employed to make working with these sorts of data sets manageable. Do you think that institutional monetization of data storage shares is a good solution, or is there a better alternative?

whiteknight521

I wouldn't pretend to have all answers. The institutions are struggling with new infrastructure challenges and my guesses about the best solution would not be well-informed at the moment.

Hi Professor Rzhetsky,

I was actually thinking about large datasets and connections that could be found in them last night! I've noticed that a lot of the work publicised in the media, in this field is either to help the medical field or for Internet giants to analyse their users.

But what I would like to ask you is, aside from the medical field, do you know of any other large data sets or situations where this kind of analysis can provide deeper insights but are not currently being researched as much as they could be by professionals like your self? Alternatively is there any interesting research like yours, but that isn't as well known, happening in fields that the general public would not know?

On mobile so apologies for grammar.

Thanks for taking the time to do this AMA!

bigpanda1

Of course: for example, the famous discovery of the Higgs Boson was associated with collection of enormous amounts of data and sophisticated analysis of these data. There are enormous amounts of data in text, in astronomical images, in sequenced genomes of humans and other organisms, in photographs and phone movies, the list could very long.

Hi Andrey, can you explain simply how does one collect such big volume of data sets which I assume are from multiple sites/sources into one single dataset in an efficient manner, has there been recent advances that have made such big data research Possible?

Also, one thing inherent difficulty about most research performed is the need to apply strict criterias for the study population yet maintain the quality of the data that can often be lost or incomplete during data collection. I'm imagining having a 150 million entries of data introduces higher likelihood of missing data, even non homogeneity with the data being analysed. How do you control for these factors and overcome these obstacles and still maintain validity of your study findings?

sheepdoc

For these 150 million data collection was done by a company which made it a business model (they sell access to the data). Yes, imperfections of data are always of huge concern.

Hi Dr Rzhetsky, thank you for doing the AMA.

My question for you might be a bit out of the scope of your job description, but I think it's equally as important.

Once you have come up with results and conclusions regarding links between autism and the environment, how do you go about convincing the general public to take the necessary actions to reduce their risks? And, perhaps more importantly, how can you go about removing non-existent links from the general public knowledge regarding autism?

ryanjrr27

The first question is easier to answer: look at the case of DDT and Rachel Carson; public awareness of a danger can help to find solutions. I don't have a good answer for the second question: ghosts of old theories and beliefs linger in the public culture indefinitely.

"Big Data" has major promises for epidemiology. My partner, who is an epidemiologist in a state health department, often laments the lack of data she has access to, and the privacy laws that prevent that access. I feel less sanguine about the promises of big data, and so it leads me to ask: Do you fear that it constitutes an ethical problem? What do you think would constitute an ethical or unethical use of it?

Thank you for your time.

cryopyre

Of course, there are ethical issues. I think, it is completely ethical to use the aggregated data to get closer to understanding (and treating) a disease condition. The history has numerous examples of

unethical data use (from unethical experiments on people to using private information for discrimination against certain groups of people). Sorry if this is obvious reply!

Hey Doc,

I'm of the opinion that all medical records the world over should be open source (all anonymous of course). What are your generalized big picture vision of what this could achieve? The benefits to humanity etc. I know this is an easy question compared to the other ones in this thread. But I love me a grand vision of the future.

Thanks.

wowlolcat

There is no such thing as completely anonymized medical record (or genome sequence). Until all information is erased, the record has a positive probability of being identified in conjunction with other public data.

Have you found a significant difference between your learning techniques (e.g. did a neural network approach seem to predict better or worse than, say, a random forest)?

Also, say that someone (being me) was interested in this realm of research (currently working on MS in CS), do you have any tips in general?

Dandy-Lion

Typically it is not to hard to try numerous techniques -- no need to stick to one for all applications.

My main advice is to read as broadly outside of your immediate technical field as possible: the best research often comes from connecting subjects that appear to be completely unrelated, but connected in somebody's mind. I personally find high-quality popular books most useful in the search of research questions "out of the box."

"Computational Neuropsychiatric Genomics" sounds cool! What does it do? I am an almost-final-year undergrad (med student) who wants to go into research.

Second question, do you also work on epigenetics? As far as I understand, it is also a proxy to measure environmental effect, is it useful / reliable?

dillyia

:) It means that we are using computational "forensic" tools to get closer to disease etiology using the existing genomic and clinical data.

On epigenomic question: yes, one of the projects in the center, led by Nancy J. Cox addresses both genomics and epigenomics.

Thanks for doing this AMA! My first question is were there any other conditions that may have been misdiagnosed as autism?

And my second question is in these hotspots that you've discovered, have you gotten to a conclusion as to why they are hotspots?

Thanks in advance if you respond.

burtwart

Autism, most likely, is not a single homogeneous condition, so the answer is yes.

I don't know yet why the hotspots exist -- this is an open problem.

What's your preferred methodology to clarify whether a risk factor is causal versus correlated but non-causal? Eg in a large dataset we might see people buying bandages and also bleeding quite a bit, and come to the erroneous conclusion that bandages cause bleeding...

waymd

In this example there is a true causality link: bleeding causes people to use bandages. As you probably know, there is a whole battery of computational and experimental methods to get closer to causality, e.g. see

1. Morgan SL, Winship C. Counterfactuals and causal inference : methods and principles for social research. Second Edition ed. New York, NY: Cambridge University Press; 2015. xxiii, 499 pages p.

Where would one go to find out more about how data analytics can be used to evaluate the safety and efficacy of medical devices? Do data sets exist that could be used in this way?

(This is job-related question. I am researching this for my association.)

Kasper-X-Hauser

it can be done to some extent with the insurance claims: they do record use of medical devices and one can follow up (with some modeling) implications for health.

Does anyone you work with or anyone you know of in other institutions that study similar things have mild autism and do you guy's ever joke about the irony of an autistic studying autism?

Also I'm a person with mild autism studying genetics btw.

Ribozome8

Honestly, I don't think this is a joking matter; yes I know quite a few people, some of them in academia, with this condition.

Do you share my annoyance at celebrity disease awareness? By this I mean the idea that we should allocate limited research and treatment resources based on which celebrity is championing which cause, as opposed to what will do the greatest good.

GloomyClown

It is just a property of the world :)

A little off your topic, but I have wondered- if we can sequence the genome... is there a way,eventually, that we could make a computer simulation of the human body...?To turn off and on certain genetic

formatting and note the results?

zerowater

Yes, this is an active area of research, although, typically, not at the whole-genome level (yet). Look up DARPA Big Mechanism program, for example.

I have ASD Im from California. My hubby has ASD he is from New York. Our kids were all born in Kansas all ASD. How would environmental elements affect us since we are from 2 different regions and our children were born in a third?

Nagaempress

The claim is not that genetic side is unimportant (actually I started as a geneticist), but that environment can magnify or trigger disease given an existing genetic predisposition.

What are your thoughts on bid data versus good data?

I work in the oil and gas sector and there is a huge debate on bid versus good, which is better etc. the problem is that often bid data and good data lead to different results

chuckatx

I would argue that, unless data is fabricated, properly modeled "bad" data can be useful. The ideal, of course, to have "big good" data.

Hi and thanks for doing this AMA; I'm curious if you think we'll ever have/enable large corporations like Google with Deepmind-type tech look into overhauling EPIC and other EHRs to make them "smart"?

PHealthy

I don't see why not! The only barrier is computing power which grows remarkably fast.

Have you found any relations between Autism and Glyphosate? I've heard some studies make the connection and others discard it. I'd like to hear your personal opinion on this controversial matter.

Eze-Wong

Nope, never touched this subject.