Science AMA Series: I'm Steve Gardner, CEO of RowAnalytics. I develop tools to find relationships across massive amounts of data (mostly HealthSci but not exclusively) and use A.I. machine learning to do the heavy lifting for me. AMA!

Steve-Gardner [1] and r/Science AMAs[1]

[1]Affiliation not available

April 17, 2023

Abstract

Our current focus is to solve the data challenge to model diseases the way they actually happen – where multiple factors from genes to lifestyle (how much you drink and smoke) work in parallel to protect or harm. The challenge is to analyze both structured and unstructured data in parallel, from genetic sequences to MRI (or other clinical scans) and to do so across large populations of diseased patients plus healthy controls to inform treatment for an individual based on his/her own unique profile. The computational challenge is massive given the exponential number of data combinations that need to be analyzed. I've also been applying artificial intelligence & machine-learning to develop knowledge graphs and semantic search tools to enable automated discovery of related concepts versus the traditional approach of applying keywords to search queries. I have designed, built and brought to market a number of innovative and commercially successful products to support drug discovery and clinical informatics. I am a serial entrepreneur with numerous patents and occasional angel investor specializing in informatics with a strong track record of building world-class companies, teams and products. I'd welcome your question on any of these topics above or to simply lend my experience and best practices for your professional development. AMA! My technical expertise is in: Hyper-combinatorial Multi-omics analytics Bioinformatics and Computational Biology Precision Medicine Semantic Search and Knowledge Graphs Artificial intelligence (AI) and Data Science Software-as-a-Service (SaaS) Internet of Things (IoT)

the WINNOWER

REDDIT

Science AMA Series: I'm Steve Gardner, CEO of RowAnalytics. I develop tools to find relationships across massive amounts of data (mostly HealthSci but not exclusively) and use A.I. machine learning to do the heavy lifting for me. AMA!

STEVE-GARDNER R/SCIENCE

Our current focus is to solve the data challenge to model diseases the way they actually happen – where multiple factors from genes to lifestyle (how much you drink and smoke) work in parallel to protect or harm. The challenge is to analyze both structured and unstructured data in parallel, from genetic sequences to MRI (or other clinical scans) and to do so across large populations of diseased patients plus healthy controls to inform treatment for an individual based on his/her own unique profile. The computational challenge is massive given the exponential number of data combinations that need to be analyzed. I've also been applying artificial intelligence & machine-learning to develop knowledge graphs and semantic search tools to enable automated discovery of related concepts versus the traditional approach of applying keywords to search queries.
I have designed, built and brought to market a number of innovative and commercially successful products to support drug discovery and clinical informatics. I am a serial entrepreneur with numerous patents and occasional angel investor specializing in informatics with a strong track record of building world-class companies, teams and products. I'd welcome your question on any of these topics above or to simply lend my experience and best practices for your professional development. AMA!
My technical expertise is in:
Hyper-combinatorial Multi-omics analytics
Bioinformatics and Computational Biology
Precision Medicine
Semantic Search and Knowledge Graphs
Artificial intelligence (AI) and Data Science
Software-as-a-Service (SaaS)
Internet of Things (IoT)

CORRESPONDENCE:

DATE RECEIVED:
March 02, 2018

DOI:
10.15200/winn.151990.08689

ARCHIVED:
March 01, 2018

CITATION:

What's the major technological bottleneck?

adenovato

There are a number of bottlenecks but most approaches struggle with three issues: * the curse of dimensionality and combinatorial complexity * establishing accurately the identity of objects in the space * the imprecision and variability of input data

The first problem is problem dependent and it boils down to how many dimensions of data you need to consider in order to find good solutions to your problem. As the number of dimensions goes up there is a combinatorial (Cartesian) explosion in the number of possible states that a space can be in. If we're

Steve-Gardner , r/Science , Science AMA Series: I'm Steve Gardner, CEO of RowAnalytics. I develop tools to find relationships across massive amounts of data (mostly HealthSci but not exclusively) and use A.I. machine learning to do the heavy lifting for me. AMA!, *The Winnower* 5:e151990.08689 , 2018 , DOI: 10.15200/winn.151990.08689

trying to find specific combinations that satisfy some criteria that have specified there are two standard approaches: we can do a search through the space using graph traversal or gradient descent type methods, or we can use dimensionality reduction methods such as PCA. Both of these require quite a lot of CPU and RAM as they need to represent at least a subset of the whole information space.

For some decision support or personalization problems we use a different approach - tensor algebras - essentially using geometrical operations running across logics that have been set up to describe the multidimensional system. We construct the state (problem) space by defining all the variables and relationships between them, and then apply the constraints/rules that define valid solutions. This effectively defines a hyperplane through the hyperdimensional problem space where the plane touches only states that contain 'valid' solutions. This can be done quickly in a away that is provably complete and consistent from a mathematical perspective, and it provides a very quick way of reducing the dimensional complexity. It can be used directly as a decision support platform or to reduce the training time for other AI approaches.

In other problems we're looking for combinatorial features - picking combinations of 20 genetic variations (SNPs or mutations) out of a set of 2.5 million or more, that in a specific combination are aossciated with disease risk. This gets big computationally very quickly - each additional order of combinations (3 SNP patterns instead of 2 SNP patterns) requires 10,000x the computation. These analyses can quickly get up to > 10^{100} combinations, which isn't tractable with existing methods. We combine similar mathematical approaches as described above in a vectorizable, parallelizable and distributable fashion across multiple GPUs to bring this high order combinatorial SNP association analysis down to hours or days.

The second and third problems are somewhat interlinked - how can you be sure you are talking about the same concepts with the same semantic meaning and that the data you have collected about them are truly intercomparable. Traditionally ontologies and other categorization systems play a role here but they are expensive to build and maintain. You can use deep semantic indexing as a way of defining the pattern of relationships between objects in a space to determine identity (contextual disambiguation). This enables you to build semantically normalized knowledge graphs automatically, which helps massively to determine identity. It also lets us automatically generate and/or curate ontologies which can be massively useful for reasoning and graph traversal problems. On the final point, choose your data carefully - we always try to work with datasets that have been collected using harmonized methods as far as possible for the latter. I'm sure there will be more on this later!

Hello Mr Gardner. What software tools, comp-languages, in particular, you(and people in your organization) use?

wavey_particle

We tend to be pretty technology agnostic - we use a bunch of tools depending on what we're trying to do. We prefer to match the tool to the problem rather than have a single hammer to beat on all the problems we're tackling. We don't use standard AI frameworks much but that's because we're usually trying to solve multi-dimensional and hypercombinatorial problems where they don't perform so well.

On the front end we mainly use node.js, AngularJS and NGiNX for our UI with some d3 for visualization.

Our code tends to be held together with Python, but we also use a lot of C++ and CUDA (mainly for linear algebra). Even though we don't use a lot of neural nets, both our deep semantic indexing/searching and combinatorial data mining tools take full advantage of GPUs.

On the back end we've used redis, Aerospike and MongoDB as well as graph DBs such as Neo4J and OrientDB. I think we even have a couple of relational DBs in the mix for some simpler apps - MySQL

and PostgreSQL.

Our platform compiler is written in portable C++ on Windows and Linux with our Runtime implemented in: Java, (Eclipse IDE) and JavaScript for client side web browsers, autogenerated from Java source using jSweet, plus C#/.Net and C++. From a decision support perspective we deploy on fairly low power devices such as mobile, wearables and IOT, so we take advantage of whatever runs most effectively on those - we have some Java, some .Net and Python implementations, but can also get much closer to the metal on things like FPGAs. On top of the basic C++ API, we have APIs for Java, which is provided by a Java Native Interface layer and .Net, which is provided by a C++/CLI layer.

Our personalized digital health mobile apps use the Java API for RESTful Services (JAX-RS) with the Jersey service executed on a Linux server by Apache Tomcat. API documentation is automatically generated by Enunciate and Swagger. IDEs are mainly MS Visual Studio 2015, Anaconda and GCC to personal preference Git for version control system and GitHub as a common repository, Jenkins for Continuous Integration server, which compiles and tests builds as changes are committed to GitHub. For our team, the work tool solutions we rely on are: JIRA/Confluence for project management, JIRA Service Desk for user support / defect management, Zoom / Skype / Dropbox / AWS for comms, compute & file share.

What credentials are the best to get into this field? I'm currently studying bioinformatics/comp bio as an undergraduate. I'm trying to either get into medical school or do more grad school, but I'm also seriously interested in getting into either a medical or genomic startup.

youflow

As an old structural bioinformatician I would be bound to say that you're on exactly the right track!

What I would say is that in the old days bioinformatician meant someone who pulled together all the available types of data and found ways to make it useful to answer a question. In the genomic age it has often been used to describe someone who can run a genome assembly/variant calling workflow and analyze the output. This in my view at least is not nearly the same.

At the same time, there is not very much bioinformatics or AI taught in most medical degrees, so you would have to proactively strike out yourself to get involved in such projects.

Either route can get you there - our COO Serdar is a medic, who switched first to biomedical engineering and then on to comp. sci and AI, but always with an interest in complex systems health problem (sometimes on rockets and sometimes on people)! I took the more scientific, drug discovery route getting a PhD in molecular modelling and then lots of practical project related experience.

Hello Steve!

What advice would you give to a medical student (in Europe) who would like to get into this field? Are there even any opportunities for MDs in the data science, analytics and AI field?

For example, do you recommend any particular tools or courses to learn during and/or after the medicine degree? Is a specialisation (residency) in Public Health useful, or would you say it's better to skip residency and take that time to learn CS or Stats or any other subject instead?

Thanks!

Executer13

It's a really good question. I think that you have to find a way in that suits your interest, whether that is

more about the people and populations or the technology.

I was speaking with a very smart guy the other day who has had a serious ambition to deliver personalized health advice to his patients, and I realized that he was taking a very epidemiological approach to gathering and analyzing the data from which to make his recommendations. I mentioned this and he turned out to have been a public health epidemiologist for 20 years. If you go down that route you will have to get the stats training as well as the field is massively statistical.

If building tools and methods is more your thing you will absolutely need the comp. sci. and AI training. It is however (at least in my opinion) not enough just to have this - the insights that your medical training have given you into the practice, culture and biological complexities of medicine will be very useful for you and any companies you might work for in the field. A true combination of the two is fairly rare and very useful.

Are you taking outside investments? Who have you taken investments in the past? Where are you getting funding for? What series of funding rounds are you in? Can I buy shares?

outsurfing

We've just completed a Seed round from angel and family office sources and we're not looking for other investors at the moment thanks.

What is the oddest or most interesting relationship you've found analyzing data?

CasuallyNotCasual

There are a number of particularly interesting observations that we find universally (so far) when we look at large disease populations - people who nominally share the same disease diagnosis. Across all of the populations we have studied, even those that have been molecular sub-typed, we are finding genetically non-overlapping cohorts, where the patients in the cohorts share specific combinations of disease associated genetic variants (SNPs or mutations) that are non-intersecting between cohorts.

This is counterintuitive because we expect that people with the same disease will share many of the same mutations - in other words there should be very few separate clusters of patients who have the same disease. We tend to think of many diseases as being spectrum disorders - we expect there to be some big shared genetic causes such as BRCA1/2 in breast, ovarian and prostate cancer for example, and then we expect individual patients to have a few more random mutations as well. Even within a population of people with BRCA2 mutations it turns out this is not entirely or even usually true - we have segregated that population into distinct non- or minimally-overlapping clusters as well.

The structures of these clusters generates interesting insights into the disease. We tend to find that cohorts with lower order combination patterns (e.g. sharing 3 or 4-SNP patterns) are homozygous for rare variants, which we infer may be in directly disease causing (e.g. protein coding) regions of the genome, whereas higher order patterns (e.g. 10 or 15 SNPs in combination) tend to include heterozygous individuals with more common variants or even some wild type features. This is supportive of the Omnigenic principles laid down by Jonathan Pritchard at Stanford in 2017 when commenting on the limitations of GWAS.

In many ways thought this is good news - not only does it explain why the same drugs work differently for different people (they actually have different molecular etiologies and therefore different therapy responses and adverse drug reaction risks), but it also offers hope of finding new uses for existing drugs. Many of the clusters of SNPs associated with breast cancer, contain mutations in genes that have been reported to play a role in other diseases. We have 'sperm motor' proteins and 'brain'

proteins forming parts of breast cancer disease risk clusters. Many of these are druggable and therefore we may be able to repurpose drugs to design more personalised combinatorial therapies for those patients (subject to lots of regulatory work)!

Hi Steve! How do you model all those different factors -- genetics, lifestyle, and environmental factors -- in a realistic way? Where do you gather your data from and what is the cleaning process like? How do you incorporate patient metadata into your model and analysis?

poltergeixt

Great question, and a real challenge. It's complicated and time consuming.

We always bring in the best data we can obviously, and we will throw out data that cannot be relied upon. As mentioned earlier we try to establish the contextual identity of the concepts we are really dealing with (e.g. is this actually that SNP, does it affect the same gene/pathway as this other SNP etc.) We then expand our scope and try to fit the quantitative data onto the skeleton of the underlying ontology or knowledge graph.

Patient metadata can be some of the most useful and the most frustrating data to deal with. Where it comes direct from sensors, we have to characterise the sensors in terms of precision and accuracy. This can be really easy, like when we worked with data from an automated robotic blood analysis system that generated 90 analyte parameters from blood samples to help triage patients before they went into ER - this was really consistent and easy to analyze. On the other hand it can be really, and sometimes impossibly, messy. Where we have things like patient notes we have to go through and interpret this in a much more detailed and laborious way.

For the narrative aspects of electronic medical records I could tell you all about a ton about pattern/rule driven Natural Language Processing techniques and Entity Recognition and Extraction mining technologies, but it would be missing the fundamental point that the patient data are often intrinsically variable because of the way that they are captured. Not all clinicians record their findings the same way, they have different diagnostic technologies and resolutions available to them, and many are chasing reimbursement so they will manipulate the diagnosis to fit the payer (see above re the challenges of segregating patient populations!).

The best example of the collection of analysis friendly datasets are some of the new large disease population studies like Project MinE in ALS. They set out to capture harmonized data for genomics (WGS + exome + SNP chip), epigenetics, phenotype, clinical history, imaging and lifestyle. They worked hard ahead of time to standardize data acquisition platforms, processing platforms, imaging modalities and questionnaires, and they explictly designed in the collection of similar data using the same techniques from a large number of controls, which is joyful to us data analysts. It is much easier to analyze this than your local hospital's EMR data!

Hi Steve,

What do you consider the most promising avenue for research in semantic search and knowledge graphs? What about data science?

Thank you,

Toasty

mmm_toasty

Hi Toasty - I have spent the last 20 odd years of my career trying to find better ways to bring

information from disparate sources together to enable better decision making. Semantics (including search) and knowledge graphs (including ontologies) have a huge role to play here. That said, the challenge is as much philosophical as it is technical. When we seek to build semantically normalized knowledge graphs that represent our current state of knowledge, we have to recognise several things:

- there is no such thing as a universal truth - different data were collected for different reasons by people with different perspectives who interpreted it differently. The real question is more, is this useful and accurately enough described for your purpose?
- the language of science is very limiting - we report in papers using poorly constructed English (go read my papers if you don't believe me!). We use loose language 'CAUSES', 'INDUCES', 'LEADS TO', 'INCREASES' may all be used synonymously but are quite different.
- we (almost) NEVER put enough context in a paper to describe our methods accurately enough to enable truly quantitative meta-analysis
- there are several linguistic constructs - modifiers, modulators, negation, probabilities, list expansion, anaphora etc. which are currently very difficult to automatically capture and accurately assign.
- the final user should be able to have as much say in what is 'true' (relevant and sufficiently well-evidenced) as the person who builds the knowledge network.

This implies that you are building knowledge graphs that contain multiple assertions, which may be mutually exclusive or contradictory. This is messy and difficult to fully automate using lingusitic approaches. There is also a limit to what use you can make of this - if someone says 'CAUSES' and someone else says 'INDUCES' or 'AGONISES' where does this leave you from a symbolic reasoning perspective? How much confidence do you have that they actually mean the same thing.

I spent many years trying to automate this using ontology driven NLP mining and other linguistic analysis tools and realised that this is only ever one facet of the solutions. Our domain changes all the time and the solutions we use must be able to cope with this change without us having to have an army of ontology curators at our disposal. I therefore think that the kind of deep semantic learning tools, analogous to the old latent semantic indexing technologies but vastly more efficient and scalable are really important here to enable emergent relationships in the data and new concepts to be incorporated fully without human intervention. We are using a deep semantic algorithm based loosely on random reflective indexing to give use the speed and scalability of keyword indexer combined with the deep semantic nuance of a full ontology driven NLP system.

The graphs we can build now combine symbolic, linguistically derived assertions with semantically discovered associations, and the combined graphs are all the richer and more powerful for it.

I also love some of the work that Michael N. Liebman is doing in biomedical analytics, especially around using the temporal nature of disease processes, and building ontologies that reflect more accurately the stages that a disease may be in when we measure parameters about it. This is crucial and important work.

Thanks for doing this AMA! I'm an undergrad in Bioinformatics and generally very excited about everything you are working on. I have two questions:

Considering the expanding number of metrices that you are able to track/leverage to find associations between populations, what kind of processes do you have to verify or validate your findings?

More generally, do you have any advice for someone graduating into the field within the next year or so?

nipple_king_

Great question - this is a real problem when looking for higher order combinations in populations.

The short answer is that we don't just run the analysis once. We usually run 1,000 fully randomized permutations where we randomly reallocate the cases and controls preserving the overall distribution of the population. This then allows us to look for the patterns (n-SNP combinations) across the random runs to see how often they appear. If they appear more than a specified level we throw them out as potentially random observations. Obviously we have to correct for multiple sampling, and we do this using a false discovery rate parameter using Benjamini-Hochberg corrections. Typically we use a pretty brutal FDR cutoff rate of between 1%-5%, which gives us very high coinfidence that the combinations we are seeing are genuine.

We also use the knowledge graphs to interpret the biological significance of the genes associated with the SNPs (or other phenotypic features) we are finding. Often these will be anchored with well-known genes strongly associated with the diseases but with many others that are more novel associations. We then of course go into the patient populations and see if the observations and predictions hold up in a clinical setting.

On advice - the biggest differentiator for me as an employer is more about attitude. The desire to get stuck in and contribute to projects to gather experience is really impressive to me. As a bioinformatician or data analyst there are loads of opportunities to get involved as an intern (we even have some slots this year), or with a disease charity or an industry group to get some real world experience.

Hi Steve

I think AI in health industry is incredibly exciting. I'd like to ask you how might one participate in such research? I'm a first year medical student with a very minor background in computer science/programming/bioinformatics in general, but I'd like to learn more about this stuff and make a contribution. Where should I start and are you accepting any students in your lab?

Thanks!

randomaccount4567

A great way to get involved is to find an internship as I said above, or perhaps find a way to one of the rare disease charities who have disease populations that they are studying. Other than that some of the insurance groups in the US (if that's where you're based) run big healthcare analytics programs which are beginning to apply AI to healthcare.

Do you ever like to look at things that correlate for fun? Like this http://www.tylervigen.com/spurious-correlations

powerlesshero111

Nice - like the examples!

We're obviously aware of some of this type of correlation - it's one of the side effects of working in multi-dimensional data, you're always likely to find something that correlates, but we tend to use these as cautionary tales about why you have to go the extra mile in establishing that these are reproducible and not just random observations.

Hi Steve, where do you get all your data from?

jo698

Obviously there is a lot of publicly available data across structured and unstructured sources, from literature, patents and regulatory filings to databases of genomics, chemistry and drug:drug/disease/food interactions among many others.

On the disease population side, there are a fair number of sources of data from academic studies, to pharma companies and health insurers. We particularly find that the disease charity studies are a great source of data for these kinds of studies. As I mentioned earlier with Project MinE, the best of these studies are really well thought through, everyone wants to deliver meaningful results and the patients are very supportive.

We also work with publishers and pharma companies to integrate their data into the publicly available datasets so that it appears as a single coherent whole.