Science Will Win | Season 4
Episode 1 | Part 1 – The Data Revolution
FINAL

Jeremiah:
Have you ever thought about the 37 trillion cells quietly at work inside of your body?

[Curious music]

To enter a cell, you'd need to shrink down to be 0.002% the width of a human hair.
Yet once you slip past the membrane and go inside, it's like a bustling metropolis: mitochondria whir like power plants, proteins fold and unfold, and messages shoot through the cytoplasm like neon signs.

At the center lies the nucleus, a tightly coiled library holding the genome – 3 billion letters of DNA, the instructions for building you.

Sometimes, something happens with those instructions – a genetic mutation, a virus, or bacteria might do some damage to a part of a cell. That could lead to a disease or disorder.

Only a few decades ago, this code was invisible to us. But today, we're not only reading this genetic library, but appreciating the nuances of each individual cell.

We're uncovering new things we can learn about what makes us tick, which can help scientists discover new medicines and treatments.

[Theme Music]

Welcome to season four of Science Will Win. I'm your host, Jeremiah Owyang. I'm an entrepreneur, investor, and tech industry analyst. I'm passionate about emerging technologies and the ways they can shape our world.

Last season, we talked about how Artificial Intelligence can help the scientific community address the rise of antimicrobial resistance.

But this season, we're diving deeper into that nexus of biology and technology, and exploring the ways artificial intelligence has had, and will have, an impact on the potential future of drug discovery.

We're taking you on a tour through history to learn about the development of three key elements that have helped form the powerful AI systems we have today
– data, hardware, and software.

We'll be talking to historians and researchers across a variety of fields, to understand how each of these three elements contributed to making strides in drug discovery and development – and why each of them are an essential piece of the ongoing AI revolution in medicine.

Today, we’re talking about data.

[Theme ends]

Today, scientists have access to an unprecedented amount of data: There are molecular measurements from cells and genomes, and vital biological data collected through surveys, clinical trials, and the growing expanse of databases and biobanks.

But how did we get here? And how exactly does data about our biology contribute to drug discovery?

To understand why data is so vital to drug discovery, we need to go back to the early days of drug discovery itself.

[Music]

The oldest approach to drug discovery was… serendipity.

Last season, you may remember we talked about how penicillin came to be. The example applies here too – just imagine, a cluttered London laboratory in 1928.

[Creaky door shuts, curious music]

Alexander Fleming returned from vacation to find a petri dish he’d accidentally left uncovered.

The dish, smeared with Staph bacteria, now had a curious blotch of mold growing in one corner. Around the mold, the bacteria had retreated.
Fleming peered closer and realized something extraordinary: the mold was releasing a substance that killed the bacteria.

Fleming’s discovery of penicillin was not the result of analyzing reams of data, but rather a lucky encounter. And it wasn’t an anomaly.

This story is a glimpse into how early drugs were often discovered: chance observations.

Well, I'm a young scientist by drug discovery standards, so I want to be respectful to my elders, but the way that I think about it is, in the previous generation of drug discovery and development, you usually had to start first with a compound that you thought might have some activity and you frequently went looking for the biology that that activity was appropriate for.

That’s Kellie Kravarik, Associate Research Fellow and Cellular Genomics Group Leader for Systems Immunology at Pfizer.

Kellie:

> We didn't understand biology enough to ask the question the other way round, which is to say, “I have this biology, I think I know what's gone wrong and I want to develop a drug for that.” But the more we understand about the biology, the more we can actually ask those questions from both directions.

At Pfizer, Kellie helps gather a ton of cell data to help contribute to treatments in the areas of inflammation and immunology.

This is to say, she helps researchers at Pfizer ask the right questions from both directions, with a “biology-driven” approach – that’s where we use our understanding of the biological basis of disease to inform the treatments we try to create.

In Alexander Fleming’s day, drug discovery was “phenotype-driven” – that means, researchers would find a potentially useful compound with an effect on the body, and then find what it’s good for.

That usually results in treating the symptoms of disease more than the underlying cause.

So how did we get to the biology-driven approach that gave rise to jobs like Kellie’s?

It took decades of learning about the processes of the cell alongside advancing technology. But at the beginning, progress was slow, and data was limited.

> Kellie:
> Yeah, historically, in the forties, fifties and sixties, when I think a lot of old school drug development was initiated, data was something you wrote in a lab notebook.

The shape of DNA was discovered in 1953, but the idea of “sequencing” DNA, understanding what was inside of it, and how it actually worked, was still decades away.

> Miguel:
> And the thing is that sequencing, so was a technique that was first applied to proteins rather than DNA. So during the 1940s and fifties, it was a technique to determine the sequence of amino acids of proteins. And at that point, it was a technique that was very artisanal, very manual, and was just used very small scale, to determine, uh, relatively short sequences.

That’s Miguel Garcia-Sancho.

> Miguel:
> I am a lecturer and researcher at the University of Edinburgh in the UK. And I've been working for several years now on the history of genomics, biotechnology, and bioinformatics. And more recently, I'm interested as well in the forensic use of DNA data and in the history and the sociology of, uh, genetic and genomic data and databases more generally.

Miguel is the co-author of A History of Genomics across Species, Communities and Projects… along with James Lowe.

James:
I’m James Lowe, I'm a research fellow currently at the University of Exeter in the uk. Previously worked, uh, at the University of Edinburgh in Scotland, uh, working on a reasonably long-term project on the history and development of genomics. And I'm generally interested in, you know, variation of living things, be it in genetics, genomics, or, or my current project, which looks at, uh, AI and healthcare.

Miguel and James will help usher us through the history of how our understanding of the genome developed over time.

The man who first applied this particular sequencing technique to proteins was Frederik Sanger. He first sequenced the amino acids of insulin in 1954.

As Miguel said, amino acids are the building blocks of proteins. So, “sequencing” those amino acids means figuring out how proteins are constructed. Proteins, on the other hand, are the building blocks of cells, and work like Sanger’s began revealing specific biological molecules—like enzymes, hormones, or receptors—that played critical roles in disease.

This knowledge helped researchers treat certain cancers by inhibiting the enzymes involved in cancer cell growth. It was an early step toward a "biology-first" approach, but it was still limited by our incomplete understanding of biology and genetics.

Then, in 1977, Sanger went on to sequence DNA for the first time – starting with the DNA of a virus.

[Music]

That’s right – DNA sequencing didn’t start with the human genome. But understanding the DNA of bacteria and viruses can also help scientists understand diseases better, by breaking down the causes of disease.

Researchers moved on to sequencing segments of human DNA
– but the Sanger sequencing method remained too-time-and labor-intensive to scale.

Scientists had to gather a bunch of cells and mix them up, then “print” these long strands of DNA onto X-ray film and manually read each letter, recording them by hand.
With this method, sequencing the whole human genome would take… over 150,000 years.

To expand our knowledge enough to improve drug discovery, we would need to go a lot faster.

Miguel:
And at that point in the seventies and in the eighties, it was very important that they started converging with computing methods and with computer technologies. And especially with algorithms and other computer applications that were developed to process like strings of information. So, like, linear sequences of discrete units, for example, a text is a string of information. So the first computer applications that were used to to analyze DNA sequences were adapted from the first, uh, word and text processors.

By the mid 1980s, thanks to this adaptation, computers were able to read strings of DNA.

> Miguel:
> So these, uh, computer applications, allow, like to, to detect, uh, patterns in a sequence to, to, for example, put together different partial sequences into a broader sequence. And that was the first time in which sequences started growing. And it started scaling up.

In 1986, the first automated sequencer was released, significantly increasing efficiency and accuracy.

> [Bar ambience, murmuring]

That same year, in a Maryland bar, the geneticist Thomas H. Roderick sat around a table with nine of his colleagues. They had just come from a genetics lecture – and after a few beers, they started brainstorming a title for a brand new genetics journal.

> [Music picks up]

Nothing had quite the right ring to it. But deep into the second pitcher of beer, Thomas piped up – “Genomics?”

We know now just how well the name stuck. It may have been inspired by the word “genetics,” Thomas’s field – but most of all, the word suggested a completely new field of study involving holistic techniques.

There had been murmurings about the possibility of sequencing the entire human genome for some time. But with new automated methods available, that possibility was more real than ever – and genomics offered the field a name.

> James:
> What was novel about genomics was the ambition to move beyond particular genes for particular maybe clinical genetics purposes, uh, and move beyond that to, to the whole genome as an object, so all the DNA in a particular organism. So it's that sort of idea of comprehensiveness that moves beyond simply tackling one gene at a time.

While we used to look at very specific DNA segments, genomics is about looking at the whole picture.

That’s a big step up in the amount of data we could record.

Genomics became the first discipline within the umbrella term we call “omics” – since all the terms under the umbrella end with that suffix. For all “omics,” comprehensiveness is key.

> James:
> And, and this is one of the significant things that's come out, is it can tell you things about the evolution of organisms. It can tell you something about diversity and variation within organisms. It can also tell you about the differences and similarities between different species. So if we know the corresponding regions of DNA, for those particular organisms,

> we might be able to actually identify genes in their particular roles in physiological processes in humans. So it's this ideal of comprehensiveness that really characterizes omics.

As genomics advanced into the 90s, researchers began to identify the genes and proteins involved in specific diseases. **This enabled them to ask the question Kellie Kravarik proposed earlier:** "If this protein or pathway is causing the disease, can we design a drug to target it?"

In 1990, the famous Human Genome Project officially began. It was a collaboration between thousands of scientists across universities, disciplines, and at least five countries.

The publicly funded project used Sanger's sequencing method, elevated by automated sequencing techniques to enable a new level of efficiency.

The mission? Sequence all 3 billion base pairs of the human genome and identify all 20- to 25,000 human genes.

> Miguel:
> it was believed that by having the whole sequence of information of the human genome, basically drug discovery would be easier because it would be just a matter of, uh, picking the information of the region of the gene, uh, the drug was, uh, supposed to interact with. And uh, basically, uh, being able to, to inform the design of the molecule that the drug would be with that information.

By 2003, the scientists declared they had sequenced 99% of the genome with an accuracy of 99.99%, marking the project's formal conclusion.

By sequencing the human genome, scientists could systematically identify the genes, proteins, and pathways involved in disease.

This knowledge enabled a true biology-first approach to drug discovery.

> [Music]

However, it soon became clear that the genome sequenced in the
Human Genome Project was not enough by itself to make creating medicine as easy as originally hoped.

Here's Miguel again.

> Miguel:
> Genomic data is important as a reference, as a basic scaffolding on which you can anchor the particular uh variance that uh your drug uh is meant to address.
> (…)
> But having more local variants that are connected to the conditions that the drugs are supposed to address, is very important.

In other words, to make these new findings applicable for a lot of drug discovery, you need to be able to compare the so-called “normal” human genome to something.

And what’s more, it turned out that the minor differences in DNA between people could really add up when it came to understanding disease.

> James:
> 74% of the DNA used in the sort of original reference genome, came from one person, uh, an anonymous male who answered a newspaper advertisement in upstate New York in 1997. Uh And this was because it was believed that between any two humans anywhere in the world, the similarity of the DNA sequence would be 99.9%. So it didn't matter. We didn't need to have a representative sample, a sample that included people with known genetic diseases or not.

They also thought they didn’t necessarily need genetic samples from women, or people from other places other than America.

> James:
> And so there was a lot in that sort of DNA that was recorded in the initial reference genome that wasn't representative of particular genetic variants or, you know, that, that are associated with a disease that are prevalent or more prevalent in other populations. ... The reference genome is the standard. It's not a fixed thing. It's been updated all the time, and reference genomes get updated all the time according to the particular aims of the communities that are involved in managing them.

Ever since the Human Genome Project, the scientific community has prioritized generating and sharing as much genome data as possible.

This and the biology-first revolution allowed us to break barriers in drug discovery... but they also revealed how much we still had to learn.

In the mid 2000s, Pfizer’s Kellie Kravarik started her career in genomics.

> Kellie:
> I came to be a scientist at the end of what I think was the first kind of golden age of genomics, which is what biology calls this data-forward practice. so the human genome project was completed when I was in middle school and all of my training to become a baby scientist really happened when that was known and something you could download from the internet. So I've never known a world where you could not measure biology in high dimension and high resolution.

Even though sequencing was automated at this time, researchers were still using similar analysis methods to the scientists on the Human Genome Project and even Frederik Sanger: measuring a mixture of many cells.

> Kellie:
> For much of my career through really my PhD, we were essentially measuring smoothies of biology. This isn't an analogy I invented, but it's one commonly used. You know, you would

> take samples from a tissue culture experiment or a human patient and, and you would blend them up and you would measure all of the blended components.

This smoothie method was leaps and bounds ahead of what was possible way back in the 1940s or 1950s, but it still had limitations. Blending all of this data was good for measuring entire genomes, but it made it impossible to understand what made each cell unique.

> Kellie:
> They have lineages just like we do. Cells birth other cells and they talk to one another just like we do. And when you blend things up, you reduce that fundamental unit of life.
>
> There's not very much material in a single cell. So if you blend them all together, you add all the material from all the different cells and you have more to measure, you don't need an instrument that's as sensitive.
>
> And so part of the innovations was to get chemistries that were more and more sensitive to measure individual cells. Part of it was getting more sophisticated in our ability to understand that information, and part of it was just trusting that this was something that was possible to do.

By the late 2000s and early 2010s, those sensitive instruments were developed, allowing researchers to break down information from within individual cells.

This is called single-cell sequencing. This technique isn’t used to sequence an entire human genome on a broad level, but rather to understand what’s going on within particular cells.

These new instruments and methods allowed both single-cell DNA and RNA sequencing. RNA helps turn the genetic code stored in DNA into action by telling the cell which proteins to produce.

Sequencing the RNA of cells is called transcript-omics, another discipline under this umbrella term of “omics.” Today, omics have been taken to another level, with many more new ways to examine the cell.

> Kellie:
> I think we've come to use omics as a shorthand for high dimension, high throughput because it can mean so many different things. And because specifically what we're measuring is the product of technology that's still updating. It's a little bit like we're listening to music on a computer and when I was young that meant a cd and then it meant a file on an MP3player, and today it means streamed from the cloud via an app. I think the measurements are still evolving, uh, but the point is that we now capture more properties of what it means to be a cell in more granular detail. And the practice of doing that systematically is the practice of omics.

Using omics data, researchers might study proteins in healthy human kidney tissue to understand how these proteins behave, interact, and carry out their jobs in the kidney.

Other researchers might study omics data to understand specific health outcomes,
like how long someone with prostate cancer might survive, the risk of breast cancer coming back, or how well a patient might respond to a certain treatment.

By using detailed omics data, researchers aim to create better tools to predict or understand diseases—tests that are more accurate than those based on traditional clinical methods.

> Kellie:
> The major advancement that's happened in the two thousands has been to move from that smoothie understanding of the world to actually measuring those cells one by one, letting them each tell us what is going on in them and letting us do the same computation that the human genome project enabled, but be able to do it on the millions and trillions of cells that are in and of our body.

The single-cell sequencing Kellie and her colleagues use today aren't reliant on blending cells together. She has another analogy for this new method:

> Kellie:
> (...) the analogy is a fruit salad format where you can actually say what came from the strawberry and what came from the grape. Uh, that has allowed us to understand this at a just a different level of complexity that we couldn't access before. And that means that some things really surprise us and some things really excite us. But either way, the closer we get to really understanding biology at this unit of life, the better we are able to develop drugs that help patients.

But there's still a lot of work to do, remember... There are a LOT of cells in the human body.

> Kellie:
> You cannot measure every cell in the world. It is not possible to do that. So we're still sampling very, very little, but because we've never been able to do it before, we're on this really exciting point in time where every measurement we make has a high chance of uncovering something we didn't know.

Since the advent of single-cell sequencing and omics data, researchers in drug discovery have already taken advantage of this new information.

At Pfizer, single cell transcript-omics contributes to huge datasets that allow researchers to train AI models and predict the best structure for medicines – including for diseases like COVID-19.

> Kellie:
> The measurements that we're now making in single cell dimensions are feeding machine learning and artificial intelligence models. The models are seeking to understand and basically describe disease biology and healthy biology as it really is. We know we will have achieved it when the models predict things that we can test in the lab. And so a lot of the day-to-day work is actually generating data to feed a model, training the model, and then asking the model to make some predictions that we then go back into the lab and test and then use that data to feed yet another model to improve it.
> (...)

when we talk about it being at an inflection point, I think what we mean is that we're getting better at the measurements that feed the models. We're getting better at the way we create the models, we're getting better at the way the models produce results and then we're getting better at training them with additional information to improve upon them.

[Music]

New research techniques, more advanced computers, more sensitive instruments... all of these innovations have ushered in this modern data boom.

But to actually use that data and have it lead to potential breakthroughs, collaboration is key. Kellie called out one example.

Kellie:
So there is a database called Geo, where virtually every sequencing file ever published in a journal from an author who had NIH funding has deposited that sequencing data for you to access and reanalyze. There are also large consortia efforts, including public and private consortia, like the accelerating medicines partnerships, where companies and academia have come together to co-fund research where we're intentionally sharing data pre-competitively out into the public domain for everyone to use for commercial and non-commercial use.

Kellie:
And I think that that's something that, all of academia is really committed to, but I think many people don't appreciate how much industry is also committed to that.

Geo is far from the only data repository. The Sanger Institute, the European Bioinformatics Institute, the National Center for Biotechnology Information,
and more all have accessible databases that help advance scientific knowledge around the world.

[Transitional music]

Let's go all the way back to the cell from the beginning of the episode.

[Same curious music from the beginning]

The proteins, the DNA, the RNA, and all those messages traveling through the bustling metropolis. Today, science has decoded much of it, and harnessed that data to inform drug discovery.

But there's a lot more left to understand about that cell – today, we are just on the cusp of harnessing the understanding of a cell's relationship with the others around it, and the specifics of what makes every single cell unique.

That's millions more points of data for scientists to break down.

The story of biological data for drug discovery is an ongoing journey for knowledge. And as James Lowe put it, so often, gaining more knowledge reveals just as much about what we have yet to discover.

> James:
> I think one of the challenges of drug discovery – anything to do with living beings is, once you think you've climbed, you know, a mountain of you know, you're gonna capture the complexity finally and be able to intervene in it... You realize you've just climbed a foothill because you see something like Everest beyond. But even that's gonna be a foothill. So I think it's, yeah, that the story never ends.

> James:
> It's, it's, it's open and it's also about, you know, dealing with complexity and, and variation, But it can also confuse you. Unless you've got something, some way to guide your way through the complexity.

By now, I'm sure you are starting to realize that there are tons of complex data, and the analysis needed to bring this new information to drug discovery is also... very complex. Which brings us to artificial intelligence.

> Kellie:
> My field has really exploded from that place of kind of canonical science to something that looks a lot more like engineering and computer science a lot of the time, in part because we've benefited from the same technological evolutions, um, helping machine learning and artificial intelligence to be able to measure our biology in a lot more detail than we ever had before. So instead of measuring the color of a test tube, we're now able to measure every single cell in a tissue of a disease patient. We're able to ask what genes and proteins are present in that cell, um, and then we're able to ask how drugs and treatments for those diseases change those pathologies and, and help return patients to good health. And we're doing that now in not a test tube with color measurements, but really a thousand by thousand matrix of measurements that can only be understood by a computer or by algorithms crunching those measurements.

> [Music]

Next time on Science Will Win:

We're going to get into the hardware innovations that made today's AI-supported research possible.

Because even with a high volume of detailed omics data,
drug design and discovery is a time-intensive and complicated process.

Plus, more data requires A LOT of storage. To make use of big data and all its possibilities, scientists need new tools at their disposal.

> Tor:
> People like to use the phrase, you know, what's the killer app for GPU? And there was a lot of more traditional computer systems folks who are questioning whether the *GPU* would be good at anything important.

> Initially, like when I started doing research looking into applications, this was a question, so what, what is this good for?
> Why would you wanna do this? So the answer is that if you write new applications that just have lots of parallelism, and it turns out there's this super important one called machine learning.

Jeremiah:
Science Will Win is created by Pfizer and hosted by me, Jeremiah Owyang. It’s produced by Wonder Media Network. Please take a minute to rate, review and follow Science Will Win wherever you get your podcasts. It helps new listeners to find the show.

Special thanks to all of our guests and the Pfizer research & development teams. And thank you for listening!