THE ART OF DATA

Listen to the podcast: http://bit.ly/1eHIug2
Subscribe on iTunes: apple.co/1FNoeDl

Audio Bumper

You're listening to the Slice of MIT podcast, a production of the MIT Alumni Association.

> **SINAN ARAL:** "Frankly, I think that MIT has always been this creative anarchy type culture, where there are very few boundaries, very few borders, you get a lot of mixing between departments, between types of people, and that's one of the things I love most about it. This unstructured creative intellectual anarchy that merges technological artifacts and solutions and theories. And that is part of what I love about MIT so much."

CHAPTER: Prologue

NARRATOR: Data is everywhere—nearly anything can be represented by a number. And in its simple form, data is pure—a collection of measured information that, when analyzed and processed, tells a story backed by numerical truth.

But data is rarely simple or pure. And thanks to platforms like advanced computing systems, social media, and millions of online consumer reviews, society has access to more data than any time in history.

So how can we make sense of this never-ending wave? And how can we better understand data and use it solve real-world problems? In this Slice of MIT podcast, we'll hear from five MIT alumni…

Produced in association with MIT Comparative Media Studies/Writing.

> **SINAN ARAL:** "Sinan Aral, PhD at MIT, graduated in 2007"
> **DENISE CHENG:** "Denise Cheng, I graduated in 2014"
> **TIFFANY CHU:** "Tiffany Chu, class of 2010"
> **JACQUELYN MARTINO:** "Jacquelyn Martino—PhD in 2006"
> **MATT STEMPECK:** "Matt Stempeck—I graduated in 2013"

NARRATOR: …whose work and research are tackling these questions in innovative ways. We'll hear how five-star ratings online are driven by social identity; how designers are mapping data to improve major U.S. cities; and how a *Jeopardy!*-winning computer uses cognitive computing to discover new recipes like Italian-Pumpkin Cheesecake.

These conversations took place at the 2015 South by Southwest Interactive, an emerging technology forum in Austin, Texas, that attracted more than 30,000 attendees and featured more than 100 MIT alumni who spoke at the festival. In a few of these interviews, you'll notice a little bit of background noise. That's just the non-stop organized chaos that is South by Southwest.

CHAPTER: Trusting Data

NARRATOR: Data is nothing without credibility. How much trust can we put in data if we don't know where it came from, and the reviewer's motives, like a review on Etsy or five-star rating on Uber? Who exactly do these reviews benefit—the consumer, the brand, or no one?

Matt Stempeck is the director of civic technology for Microsoft and a graduate of the Media Lab's Center for Civic Media, where he studied how technology is encoding subjective aspects of data into platforms in ways that might not be helpful. At South by Southwest, Matt presented with Denise Cheng, a peer economy expert whose research focuses on civic engagement. Matt and Denise discussed the idea of confidence in digital data, especially in peer-to-peer online marketplaces. How much should we trust this information—and how accurate is it? And, in some cases, what's the point of writing these reviews in the first place? They explain:

> **STEMPECK:** "On Yelp, if you look, people review prisons. They don't generally get well-reviewed but it's an interesting kind-of societal feedback loop, right? Because the prison doesn't have to say yes to get reviewed, right? Just like the restaurant didn't have to agree to get reviewed."

> **CHENG:** "You also have the opposite thing going on Amazon, right, where you have like the horse mask as one of the most popular selling items and then everyone who comments on it—it's actually a game. People who comment on it because it's really funny. And create these really wild stories that they want to post."

NARRATOR: The five-star rating might be the most standard review system, but it's far from the most accurate. Denise says it's dominated by social cues and motivated by politeness more than accuracy.

> **CHENG:** "The five-star rating is ultimately not objective because it's a number, you can aggregate, you can average things out, but at the end of the day, these are based on cultural biases. Because if these ratings are so blunt and unhelpful, they really become acts of courtesy, so when we really get out of a lift or Uber, you give a five-star rating because it's an act of courtesy. It's a way to ensure that those people are going to continue to have work."

NARRATOR: So, think of these flawed systems as a small part of a larger reputation. And if a subjective review doesn't seem helpful, the overall transaction history might be, like the number of times someone stays at an Airbnb listing. Matt and Denise explain:

> **STEMPECK:** "I think it's interesting that we're using all of these numbers, and five stars, and data science really, to encode our subjective opinions, right? The cultural influence on Airbnb is subjective. Even though there are a lot of benefits to quantifying these feelings, it's important to know that these algorithms were designed by biased human beings and we don't even know how they work."
>
> **CHENG:** "There's also this idea that if the ratings themselves aren't useful, the transaction history that you see of people is really useful…going back to peer-to-peer marketplaces, you can see how people stayed at the host's apartment for example, and that host can see how many times you've stayed at other places. Even if the ratings themselves aren't necessarily useful, you know enough that the host is attached to his or her profile."

NARRATOR: We know that data can be inexact and sometimes unreliable. But what happen when data is organized and thoughtful? And, say, programmed inside the world's smartest chef?

CHAPTER: Cognitive Cooking

> **JACQUELYN MARTINO:** "We're getting to that wonderful space where it is actually a collaboration, where there's no way I can read everything. I don't even think the best MITer could read 23 million research articles and retain the amount of information that would be necessary to identify patterns and promote innovating solutions.
>
> "We do love our firehose, but I think that's a pretty big chunk!"

NARRATOR: That's Jacquelyn Martino, a designer in IBM's Watson Group. Watson, of course, is the computer best known for its success on *Jeopardy!*, beating champions Ken Jennings and Brad Rutter in 2011. The computer visited MIT later that year, and for good measure, defeated a group of MIT Sloan students in a *Jeopardy!* exhibition by a little more than $50,000.

In 2015, Watson added a new title: chef. The computer is a sort-of personal chef that can develop new recipes based on the ingredients you'd like to use or exclude. Chef Watson uses a database of more than 10,000 recipes from *Bon Appétit* magazine and can sift through more than one quintillion ingredient combinations. The idea, Jacquelyn says, is not to tell us what to eat or how to cook, but to use cognitive data to perhaps create a dish we might have never considered.

> **MARTINO**: "These aren't answers. People who are using the Watson app have a goal in mind, they have a hypothesis, they have an educated guess about where they may be able to go with their expertise, but there is no answer yet about what it is they're working on.
>
> "So in this case of recipe, if I'm going to put in my chicken and my garlic, I wouldn't have thought about the strawberry and mushroom—that answer wasn't knowable before. And now the Watson application can help me push my understanding, push me into looking at patterns I couldn't have realized on my own and to see a context that I wouldn't have come up with."

NARRATOR: Jacquelyn's work at IBM intersects art, design, and computation. She says this new field of cognitive computing—or cognitive cooking—is a collaboration between data and human inquiry. Information that in the past was too massive to consume can now be digested—pun intended.

> **MARTINO:** "It's the opportunity to be collaborative with your computing environment in a way that is—we have vast quantitative space that we can help people explore, we also the qualitative benefits that we bring to the situations, that we bring the inquiry as human being.
>
> **"**When you have so much information to consume literally, and you have your hypothesis and your educated guess about how to you can impact a domain or industry or find a solution, that's how you can impact the cognitive computing era. This ability to bring together what everything does very well. What computation does really well, what data does very well, the quantitative and the qualitative and get to this place that heretofore hasn't existed."

NARRATOR: Chef Watson shows us how a computer can learn from human expertise and extend what people can do on their own. So what if we took a similar approach to redesigning U.S. cities?

CHAPTER: New Approach to Open Data

NARRATOR: Urban planner have been designing cities the same way for decades—using a spreadsheet, a pencil, and not much else. Tiffany Chu is rethinking urban planning by harnessing the mass of open transit data made available by nearly all large U.S. cities, and moving away from outdated city-building tools.

> **CHU:** "You be surprised by how cities are being planned with paper, pens, and Excel spreadsheets. When you're planning a new bus route, for example, what you do today, often times, is you take a paper map, sketch out routes with markers and pens, and take that into earth to get the distance measurements. Then you take that into Excel to get your cost analysis. And this process is long and laborious, and if you wanted to experiment or try out a new tweak in the bus route—or if you wanted to say what would happen if you went down this street and served this community—it would take hours if you would have to start from step one."

NARRATOR: Chu is a former fellow at the San Francisco non-profit Code for America. In 2014, as part of the Code for America fellowship, she spent a year working with city officials in Charlotte, North Carolina, helping the city launch their open data policy and initiatives. While at Code for America, she and a team of fellows used San Francisco's open transit data to reimagine the city's transit system.

> **CHU:** "We thought it would be really fun to draw lines on a map and suggest new bus routes the city of San Francisco. We released it quietly online about June of last year and the next day our server went down because so many transit planners came out of the woodwork all over the world and started making maps in their own city. It was so overwhelming and incredible for us to see that we had tapped into this niche community of planners who needed better tools."

NARRATOR: That project bore Remix, a city planning tool that uses open data to help redesign transit networks. Since its launch, more than 80,000 new bus routes have been proposed by using the tool. Remix is now used by city planning agencies in the U.S. and worldwide.

> **CHU:** "Basically it's a mapping tool for planners to plan bus routes but it's also a communications tools to show the public and explain why certain planning decisions need to be made in very real constraints.
>
> "Because every city has different needs, we adapt to them through open data, so every city, or most cities, have their transit data in the same format…we develop our infrastructure to be able to pull in exactly that data format. Once we have that, we can visualize their network in all different ways and the next step is more than just planning, the next step is to really operate those buses in a more optimized way, and get them out on the roads and serving people."

CHAPTER: Data by Design

NARRATOR: Sinan Aral is a professor of management at MIT Sloan and chief scientist at Humin, a company whose app that combines data from phone and social media to predict your most important contacts, and who you're most likely to connect with at a given time. Or, as they put it, to use data to conceptualize human relationships in context. He’s also the voice that you heard at the beginning of this podcast.

Sinan has worked with Facebook, Yahoo!, the *New York Times*, and Nike to help make sense of the big data these companies control and design a better user experience for customers. He explains his role:

> **ARAL:** “My role as chief scientist is to build and run the data science group. The mission of any data science group is to derive insight from data—data being produced by a platform that we are building, the users of that platform, but also bringing in outside data to drive insight in the company to open that data back up to users so that it can be useful to them as sort of a quantified self-service to the user.”
>
> “Our job as data scientists is to make sense of that data both for the organization but also to provide that data back to the user in a way that is useful to them.”

He says that the idea of merging data and design to help solve real-world problems is still relatively new.

> **ARAL: “**That’s fairly recent in the sense that today data scientists are sort of rock stars. A few years ago, not the case. It wasn’t very cool to be a computer programmer. But today it’s the hottest thing going.”
>
> “To be honest, I think we’re very early days. I think we’ve barely scratched the surface of what’s possible. The rate of change is increasing—the amount of new things that we see per year, per day, per week, is increasing and we've only scratched the surface.”

NARRATOR: But a massive amount of data should come with an even larger amount of responsibility. And any fears from the public are not groundless. Not only from a privacy aspect, but the idea that relying more and more on artificial intelligence to make decisions could stagnate the human mind.

> **ARAL:** “In terms of what we can expect or fear or hope for algorithmic thinking, a lot of these fears are real and well-founded and we need to be very careful about privacy rights. We need to be very careful about how automation might drive wage inequality. Very careful about sort of losing our society as automation and artificial intelligence takes greater and greater hold.”

NARRATOR: But when harnessed responsibly, and combined with human ingenuity, Sinan says data can map culture in a way that elevates creative thought and insight. It can create a collaborative environment that can help alleviate real-world problems.

> **ARAL:** “There’s a ton of opportunity in the technology that’s being created today. There’s sort of limitless potential and possibilities to create life-saving new

> medicine and create new business models and new processes, to reach greater and greater proportions of people in the world, to address poverty and HIV and violence, and that's what we're working on and I think that the more people work on things like that with technology, the more we'll realize the good and address some of the bad."

NARRATOR: What's your take on the growing connection between data and cognitive thinking? How can big data make society a better place, and what should we be concerned about? Tweet your thoughts on this episode to @mit_alumni—that's at-mit-underscore-alumni. And if you want to hear more surprising, insightful, and quirky stories about MIT, subscribe to the Slice of MIT podcast on iTunes. Please rate the podcast and leave a review—tell us what you liked, and didn't like, about this episode. And let us know what type of stories you'd enjoy hearing in the future.

Special thank you to all of the MIT alumni who took part in the 2015 South by Southwest Interactive, especially Sinan Aral, Denise Cheng, Tiffany Chu, Jacquelyn Martino, and Matt Stempeck.

Subscribe on iTunes to automatically receive next month's episode of the Slice of MIT podcast. For more stories about MIT at South by Southwest, visit the *Slice of MIT* blog at slice.mit.edu and search s-x-s-w-15.

CHAPTER: CMS/W

CMS/W (ANDREW WHITACRE): "Hello, this is Andrew Whitacre from MIT's Comparative Media Studies/Writing, where we reinvent the humanities to engage the day's big challenges. Enjoy what you hear in this episode? Check out all our work on data. We've got Denise Cheng and the sharing economy, breakthroughs with journalism, digital humanities, and data-driven games. You can find all of that, and our podcast at cmsw.mit.edu/data."

NARRATOR

Jay London
Marketing Strategist and Multimedia Writer | MIT Alumni Association
londonj@mit.edu

MUSIC

"Carefree"
"Go Cart"
"Pamgaea"

All songs by Kevin MacLeod (incompetech.com)
Licensed under Creative Commons: By Attribution 3.0
http://creativecommons.org/licenses/by/3.0/