Conference Report
November 18-19, 2013

NORTHWESTERN
UNIVERSITY
IN QATAR

Northwestern University in Qatar Symposium

BIGD8TA

Big Data, Smart Media?

Connecting Content, Audience and Information


Questions about Events
What is the effect of Hurricane Sandy on New York City?
It's dark.There is minor price gouging. There are restaurants selling hot food through their bay windows. The Police are doing an amazing job with traffic concerns and directing traffic. Many stores have set up recharging stations for people to recharge their electronics. People are meeting other people in various parts of the city and offering showers. Bars are cash only. Banks have waived ATM fees. Stores with non-perishables are open and selling goods. The buses are running great, just like any other day. Everyone is walking around with flashlights or phone camera lights on.
Northwestern University in Qatar Symposium
BIGD8TA
Big Data, Smart Media? Connecting Content, Audience and Information
November 18-19, 2013
Follow #BIGD8TA on Social Media
@NUQatar
@NUQwildcats
BIGD8TA
BIGD8TA
BIGD8TA

Contents

Big Data:

What is it, what kind of data and how much?

VIDEO

- Streaming video takes up **more than 1/3** of the Internet traffic during normal television watching hours
- **72 hours** of video are added to YouTube every minute
- **864,000 hours** of YouTube video are uploaded each day
- **22 million hours** of TV and movies are watched on Netflix each day
- Zynga processes **1 petabyte** —1,000 terabytes—of video game content per day

SOCIAL MEDIA

- **More than 1.4 billion** online consumers spend 22 percent of their time on social platforms
- **172 million** individuals visit Facebook each day
- They spend **4.7 billion** minutes on Facebook each day and update their statuses **532 million** times daily
- **250 million photos** are uploaded to Facebook each day
- **30+ billion** pieces of data are added to Facebook each month
- **40 million** individual users log on to Twitter each day
- They generate **50 million tweets** per day
- **32 billion searches** are performed on Twitter per month
- **22 million** individuals use LinkedIn each day

OTHER DIGITAL PLATFORMS

- **1.3 exabytes of data** are sent and received by mobile Internet users each month
- The average teenager sends **4,762 text messages** per month
- More iPhones are sold than babies born each day
- **294 billion emails** are sent each day
- **72.9 products** are ordered per second on Amazon
- **18.7 million hours** of music is streamed on Pandora each day
- **1,288 new apps** are available to download each day
- More than **35 million apps** are downloaded each day

By 2018, there will be a demand for about **450,000** data scientists in the US, leaving a shortage of more than **150,000** positions

Program

NOVEMBER 18, 2013

LOCATION:
W Hotel, West Bay

6:30pm
Reception

8:00pm
Keynote Speakers and Welcome Dinner

Kenneth Neil Cukier, Data Editor, *The Economist*
"Big Data: A Revolution in How We Live, Work and Play"

Khalifa Al Haroon, Founder, ILoveQatar.net
"Social Media Metrics and Building Online Audience and Revenue"

NOVEMBER 19, 2013

LOCATION:
3069 CMU Building, Education City

9-9:30am
Registration

9:30-10:00am
Welcome & Introduction
Everette E. Dennis, Dean and CEO, NU-Q

Defining Big Data
John Pavlik, Associate Dean for Research, NU-Q

10:00-11:30am
Panel: Driving Media Innovation With Big Data
Moderator: Justin Martin, NU-Q

Larry Birnbaum, Northwestern University
"Telling Stories at Internet Scale" *Birnbaum's pioneering research developing data-driven storytelling using computer algorithms*

Mohammed Haddad, Al Jazeera Media Networks
"The Translation of Data Into Visual Presentations" *How Al Jazeera makes sure it uses the right data to tell the right story*

Kathy McKeown, Columbia University
"Social Media and Big Data Analytics: Why You Should Care" *Research on the new media implications of big data*

11:30am to 12:00pm
Luncheon

12:00pm-1:30pm
Panel: Tracking People Through Big Data
Moderator: Starling Hunter, Visiting Associate Teaching Professor, Carnegie Mellon University Qatar

Martha Stone, University of Oxford Reuters Institute
"Surveillance and Big Data" *What big data means for citizens' digital communications and online behaviors in the advent of a surveillance society*

Greg Bergida, NU-Q
"Web Analytics and Predictive Power" *How social media can help develop predictions about media consumer behavior*

Darnel Moore, PerceptiAn Consulting LLC "Media, Sports and Big Data" *Using data to do "customer journey mapping," particularly in the context of sports and the media*

1:30-1:45pm
Concluding Remarks

1:45pm
Adjournment

Introduction

The waves of data produced in just one day by myriad sources affect every area of life.

The new tools of the digital age not only capture, curate and store, but also search, transfer, visualize and analyze. That's why we are here: to gain an understanding of big data, why it matters and what we ought to do about it.

— Everette Dennis

Everette E. Dennis,
Dean and CEO, NU-Q

At no time in history has so much information been so readily available. As with any shift so monumental, this phenomenon presents both opportunities and challenges. Northwestern University in Qatar's big data symposium on November 18-19, 2013, brought together experts in statistics, journalism, marketing and computer science to discuss the impact of this new paradigm on the media and communications worlds.

"Big ideas and complex ideas often acquire monikers that are simple and straightforward, if not fully satisfying," NU-Q Dean and CEO Everette E. Dennis said in his introductory remarks. "That, of course, is the case with big data, which is a marker for meaning to explain the interface between a torrent of information and our capacity to manage it. People are always looking for the next big thing, and many commentators and experts believe that [big data] is that and that it's already arrived on the scene."

Dennis noted the rapid but under-the-radar nature of the big data revolution: Its explosive overnight growth happened without much notice from beyond the tech world. Only recently, when new media tools, digital devices and social media sites became ubiquitous, did that paradigm shift become mainstream. This symposium represents a step toward understanding the new implications this change brings for media, communication and journalism.

"The new tools of the digital age can not only capture, curate and store, but also search, transfer, visualize and analyze," he said. "That's why we are here: to gain an understanding of big data, why it matters and what we ought to do about it."

As far back as 1968, journalists were figuring out the power of data analytics, explained NU-Q Associate Dean for Research John Pavlik. That year, *Detroit Free Press* reporter Philip Meyer won a Pulitzer Prize for his work on the city's race riots—work that centered on data analysis.

Tools for accessing and analyzing big data are quickly becoming essential to doing quality investigative reporting, Pavlik said. "Job opportunities in journalism and media increasingly require skills in data-driven reporting and storytelling."

The symposium also acknowledged the limits and potential pitfalls of big data from conclusions drawn from bad data to growing concerns over surveillance. "We are not only here as cheerleaders of big data," Dennis said, reiterating big data's formidable and growing influence on all areas of life.

"It is becoming crucial to understand and know how to work with big data if one is to succeed in communication and journalism," he concluded. "Big data can paralyze us or liberate us, and the choice is ours."

The Big Data Symposium presentations touched on how massive amounts of data are changing daily life, marketing trends and, of particular interest to the university, the field of journalism. The following coverage highlights relevant and pivotal notions raised by panelists and audience members.

John Pavlik, Associate Dean for Research, NU-Q

"Job opportunities in journalism and media increasingly require skills in data-driven reporting and storytelling."

— John Pavlik

Keynote Speaker:

Kenneth Cukier

Kenneth Cukier, data editor of *The Economist* and co-author of *Big Data: A Revolution That Will Transform How We Live, Work and Think*, started off the symposium with a discussion of big data as an amorphous concept that defies its simple name.

The core analytical techniques required to make sense of big data have been around since before computers, he explained. But the breakout moment of multimedia and widespread data creation has driven the evolution of these methods, with massive computers crunching incomprehensible amounts of data in shorter periods of time.

Cukier explained that while infrastructure is at the core of understanding how data came to be formidable across industries, the character of the digital world changed as it broke the barriers of military and governmental uses into mainstream multimedia content.

About a decade ago, scientists at universities and throughout Silicone Valley noted that data sets were so large that they were outstripping technology. Thus began the competitive evolution of analytical methods versus data generation across all sectors—inaugurating the era of big data.

Cukier cautioned, however, that the origin of big data can no longer be the only way to frame it: "The term has been diluted, and the dilution is to say that big data now refers to using statistics to apply data to areas that we have never applied it to before."

Improvements in the analysis of data have driven the other side of the evolutionary competition—now that it's easier to make sense of huge data sets, people are coming up with more and more sources of data.

"For example, we are all sitting down right now," he said. "If I were able to put a sensor under your chair, I could create an index that's very unique that describes how you sit and the way you sit and your posture. It would be something like a fingerprint that would identify it to be you."

Such a "fingerprint" could be used in automobile technology to identify the driver and prevent theft in case someone else sits in the driver's seat, for example.

For journalists and media professionals, the increase in readily available data brings the need to learn new technologies and methods. "If you do want to go into journalism, marketing or communications, you do have to understand data," Cukier said. "What that means is you are going to have to learn the hard stuff that you otherwise might have tried to avoid 20 years ago—like stats, mathematics and computer programming."

But even as developers and companies celebrate advances in the use of data in new and imaginative ways, Cukier noted that the risks and challenges are showing up in the realms of privacy and ownership. Privacy regulations around data are constantly changing due to the rapid advances in the ways data can be mined and analyzed by both private and government entities.

Cukier warned attendees about the "deification of data," i.e., when data are given more meaning than they deserve despite their inherent imperfections. "It's only a simulacrum of reality, in the same way that a map is not a real territory," Cukier said. "We have to remember that while we embrace the idea of big data, we also hold on to our common sense and our humanity and our humility. This is about real people and real information, not just bits and bytes and spreadsheets."

Keynote Speaker:

Khalifa Al Haroon

Big data's impact has been a major boon to marketing and communication professionals and organizations. Keynote speaker Khalifa Al Haroon, founder and CEO of the website iloveQatar.net and Haroon United Group, gave insights into the ways big data is helping him track the growth of his media channel. His talk helped provide context for big data's effect on Qatar and the nation's current ability to tap into the information flow.

"In Qatar we don't have a lot of information that's shareable, because, regardless of the fact that some say we don't have big data in Qatar, that's rubbish," Al Haroon said. "We have data all around us, everywhere, and what we need is someone to help us get that information and share it with everyone else."

He placed special emphasis on the market for Arabic content, which he argues is relatively underserved. "Eighty percent of YouTube creators who create in Arabic have gone on to be millionaires—.01 percent of all content on YouTube today is in Arabic," he said.

Scientists in Qatar are coming up with innovative ways to harness the country's disparate sources of data. For example, Al Haroon pointed out that Qatar's Stars of Science competition winner was a man who developed a shoe for camels that analyzes their walking patterns and identifies illnesses and injuries.

He also said that 42 percent of markets surveyed in early 2013 invested in big data technology in the hopes of benefiting from the seemingly endless possibilities for monitoring, tracking and measuring consumer data. These efforts can be used to better target customers, but inevitably come with concerns about privacy and security.

"Let me tell you what I know about my users: I know where you are from, what you like, how likely you are to click on a link, how you move your mouse across my browser and, using eye tracking, if you give permission, I can see how your eyes move on my site," he said.

"We have data all around us, everywhere, and what we need is someone to help us get that information and share it with everyone else."

Panel 1

STORYTELLING 2.0

Panel 1

Few, if any, industries haven't been transformed by the rise of big data. Journalism is currently undergoing a massive—albeit incremental—shift in the way stories are reported and told. Through the use of artificial intelligence, specifically natural language processing, huge amounts of information are being mined at rapid-fire pace for highly customizable bits of information that are then automatically woven into story form.

Larry Birnbaum, professor of electrical engineering and computer science and journalism at Northwestern University, and co-founder and chief scientific advisor at Narrative Science, elaborated on automated storytelling in the opening panel.

Birnbaum said effective data analysis boils down to training systems to detect and arrange themes and patterns from information sets. The amount of data in many fields precludes usage unless machines are tasked with generating stories.

"We don't consider ourselves [replacing journalists] at all. These are stories that could never have otherwise been written ... most of these stories are going to be aimed at you and to you, and this is about giving you stories that are relevant to you right now."

— Larry Birnbaum

"What is interesting is that you can look at data, and you can see these patterns, and you can see that they are actual patterns," he said. "They give you advice. [The resulting] stories highlight and summarize the critical aspects of a situation. They convey trends and causes and, in some cases, make recommendations."

The computer-generated stories produced by companies such as Narrative Science are both consumer-facing and used internally by companies in a range of fields, including sports, business and medicine. The data in these fields is massive, and the meaning would be buried if it were not selected and ordered. Birnbaum asserted that stories, rather than raw data and graphs, are the best way to get points across.

"Stories can pull out and make explicit really critical elements in thinking about a data set—what matters to the reader, what they need to be paying attention to and why they need to be paying attention to it," he said. "This is what a good writer does. The machine is a good writer and does a reasonable job of pulling stuff out."

Further exploring the concept of automated storytelling, Kathleen McKeown, professor of computer science at Columbia University and director of the school's Institute for Data Sciences and Engineering, discussed how useful stories can be constructed from informal text.

Panel 1

My own work is on detection of sentiment, and much of what you see out in the real world works on the basis of words looking for context—positive sentiment within a negative context.

— Kathleen McKeown

"There's a lot of information out there in social media every day," she explained, "and so what we are interested in is creating systems that can make sense of that data for people. At the same time, we're also looking at how we can exploit that data as we build systems."

McKeown described a system developed through Columbia called Newsblaster that has been sifting through news data since 2001 to identify discrete events and generate reports and summaries.

She and her colleagues are using data from social media to generate summaries and answer questions. "The first question that we ask is, 'How does online discussion differ from online news?'" she said. "Online discussion provides unedited perspectives from the point of view of the everyday perspective. While news is an article, a monologue, online social media is often in the form of dialogue, so as we build systems that's something we have to take into account. It presents opinions, viewpoints and quite a bit of emotion on the part of the everyday person."

One challenge lies in determining the actual meanings of the informal and idiomatic word combinations particularly common in social media. Another challenge comes in training the system how to know what to look for.

"The question never appears in the answer," McKeown said. "You can Google or Bing keywords and retrieve matching items, but we have to find relevant sentences and we don't always get these [from a straightforward query]."

Nevertheless, mining social-media data for stories can provide more illustrative first-person perspectives. For example, McKeown's group used on-the-ground data from police, restaurants, traffic, stores and recharging stations in its coverage of Hurricane Sandy.

In the end, she envisions the next generation of storytelling moving along a continuum of both time and narrative, incorporating local and big-picture perspectives and opinions gathered from

Panel 1

those experiencing events firsthand.

To round out the discussion on journalistic applications of big data, Mohammed Haddad, interactive journalist at Al Jazeera English, explored the translation of data into visual presentations. He addressed students in particular, sharing key lessons he has learned at Al Jazeera.

His first lesson was to humanize data. This means looking at the people behind the data—in other words, how raw numbers actually reflect the real world. For example, one Al Jazeera English project tried to map the hundreds of rebel groups in Syria, including their size, location and composition.

"When you are taking data and trying to abstract it, eventually you are going to see it as a visualization," Haddad said. "There are real people behind data. Data does not just exist in isolation."

He stressed that media professionals must take time to probe data deeper and wider for more context. "Oftentimes we think that because data is sometimes easily available that it's the best data and we should use it," he said. "In fact, in cases like the Syria project, you have to go out and find the data that is actually adding value to the whole process of journalism."

Haddad's second lesson was that data does not automatically equal truth. He emphasized the importance of analyzing the quality of datasets in terms of who put them together and when. For example, his team had to take into account out-of-date data sets when covering the United Nations conference on climate change in Doha.

"If we draw conclusions based on certain data points [in the older set], then we may be misleading our readers," he said. "The convenience of the data should not make the decision for us that this is the best data set out there. We need to do our due diligence in interrogating the data in looking at its full effect."

> "There is a fine line between forming data and narrative and then letting the data speak for itself. As a storyteller, you want to surface what you believe is the most important part of the story."
>
> ***— Mohammed Haddad***

Panel 2

The Whole World is Watching

The conference's second panel explored the variety of ways in which big data is used in the public and market spheres, from the innocuous—movie recommendations, for example—to the controversial, such as government surveillance.

Panel 2

Big Data and Surveillance

Martha Stone, CEO of research firm World Newsmedia Network, explored how big data is increasingly subject to surveillance, corruption and leaks, and the effect of those vulnerabilities on both journalists and everyday citizens.

She began with the case of Edward Snowden, who leaked US National Security Agency (NSA) documents about its surveillance of phone and online activity of millions of Americans. That lead to The Guardian's exposure of PRISM, a surveillance program that collects data from phone companies and major tech companies such as Google, Yahoo, Facebook, YouTube, Skype, Apple and more. While the US government sought to quell the resulting criticism—citing national security—even more outrage emerged after it was revealed that NSA was also running surveillance on many foreign leaders.

The extent of the domestic surveillance revealed caused even those who favored such measures in the past to balk, Stone explained. "The bottom line in all of this is that we must balance the need for security with the right to privacy," she said.

Although tech companies denied that they supply their information to PRISM, Stone presented budgetary allowances that argue otherwise—US$20 million of the NSA's budget is allocated to compliance costs incurred by tech companies.

Stone touched on the power of some of the information gathered through PRISM to give security officials advance warning of potential crimes; in some cases, however, this information is not applied, as was the case in the Mumbai terrorist attacks of 2008.

Highlighting the case of Hao Jian, a Chinese government critic, Stone asked the audience about the possible implications of a system like PRISM. Hao is periodically called in to have a "cup of tea" with government officials to discuss his latest critical statements.

"China is blanketed with surveillance cameras installed on most streets, in supermarkets and in classrooms," she said. "The main purpose of the surveillance of course is control and intimidation in order to maintain the Communist Party's hold on power."

Finally, Stone talked about the current and future impact of government data-mining on the field of journalism, saying that the practice is already making potential sources less willing to talk when they know they could be recorded without their knowledge.

Panel 2

Big Data and the Custom Fit

Not all data collection is done for such morally (and legally) ambiguous purposes, of course. Many companies use big data to help their customers—and give themselves a competitive edge.

Greg Bergida, NU-Q's director of student affairs and former business consultant, presented the topic of predictive analytics, a variety of statistical and analytical techniques used to develop models for predicting future events or behaviors.

Distinguishing between data and information is critical, he argued. Data are raw, quantitative values, whereas information is contextualized to add insightful detail.

With specific analytical tools, data can be mined and transformed into information so that consumer trends and preferences, as well as popular ideas, can be targeted and used to provide market insights.

Several media companies, including Netflix and Worldwide Motion Picture Group, use data analysis to create products that are more likely to sell. Netflix, for example, greenlighted an entire season of House of Cards based on data analysis that told the company that the show would be successful.

In that same vein, Darnel Moore, managing partner at PerceptiAn Consulting LLC, stressed the importance of putting the individual at the center of any marketing strategy. He urged organizations to create an empathic customer experience, noting that "empathy creates dialogue, messages and themes that energize and trigger 'moments of truth' with fans."

Big data comes into play because it allows companies to monitor consumer activity at every stage of the purchasing and usage process through "consumer journey mapping."

Moore explored how this has already proven valuable to sports teams as they try to tightly define fans in order to predict and accommodate their behavior.

Research into this area, Moore explained, will help offer a personalized experience to sports fans into the future, including personalized merchandise offers, trivia games, mobile snack orders, mobile check-ins to games and many more services.

Top: **Greg Bergida**
Bottom: **Darnel Moore**

#BIGD8TA TWEETS

NU-Q at a Glance

Quick Facts

- Northwestern's 12th school and only overseas campus
- Classes began in 2008
- 2,500-acre Education City campus
- Four video production studios*
- Two 150-person lecture halls*
- One black box theater*
- Multimedia newsroom*
- Diverse student activities

**In progress*

Current Numbers

157
Students enrolled

28
Nationalities

30
Full-time faculty

Northwestern University in Qatar

Northwestern University brings to Qatar everything one expects from a top-ranked university—a distinguished history, famous programs and a world-class faculty. NU-Q provides a framework through which students explore the world and, ultimately, shape its future. NU-Q recruits students of demonstrated academic achievement from diverse social backgrounds.

Faculty

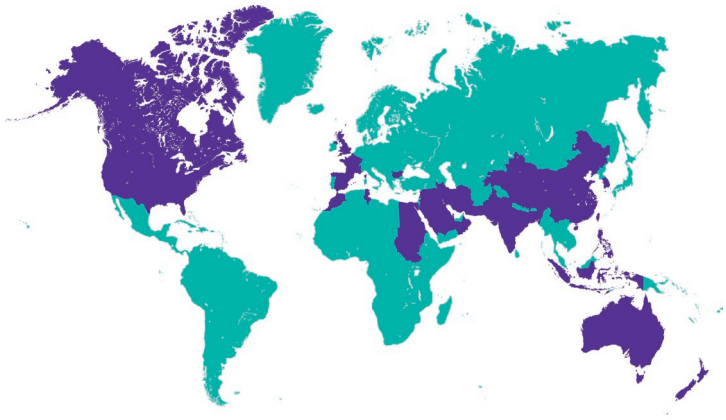
NU-Q's faculty upholds the commitment to excellence that has won the university international recognition for over 150 years. In addition to challenging students in the classroom, they lead in their fields through research and innovation.

Students

NU-Q's Wildcats are a living and diverse group of students who enjoy not only a quality education, but also fun and social learning activities, clubs and organizations, supportive career guidance, and international travel. The diversity of the 188-member student body allows each student to learn about other cultures, traditions and countries.

Student Nationalities

Australia Bahrain Bangladesh Bulgaria Canada China Egypt France India Indonesia Iran Iraq Jordan Korea Kuwait Lebanon Morocco New Zealand Oman Pakistan Palestinian Territory Philippines Qatar Saudi Arabia Spain Sudan Syria Tunisia United Kingdom United States



Northwestern University in Qatar
P.O. Box 34102
Education City
Doha, Qatar

www.qatar.northwestern.edu