IEEE potentials
THE MAGAZINE FOR HIGH-TECH INNOVATORS
July/August 2018, Vol. 37 No. 4

Analyze This

In this issue

- Distributed computing
- Analytics on the cloud
- Data visualization
- Machine learning for data-driven control of robots

IEEE

IEEE Collabratec™

Bright Minds. Bright Ideas.

Introducing IEEE Collabratec™

The premier networking and collaboration site for technology professionals around the world.

IEEE Collabratec is a new, integrated online community where IEEE members, researchers, authors, and technology professionals with similar fields of interest can **network** and **collaborate**, as well as **create** and manage content.

Featuring a suite of powerful online networking and collaboration tools, IEEE Collabratec allows you to connect according to geographic location, technical interests, or career pursuits.

You can also create and share a professional identity that showcases key accomplishments and participate in groups focused around mutual interests, actively learning from and contributing to knowledgeable communities. All in one place!

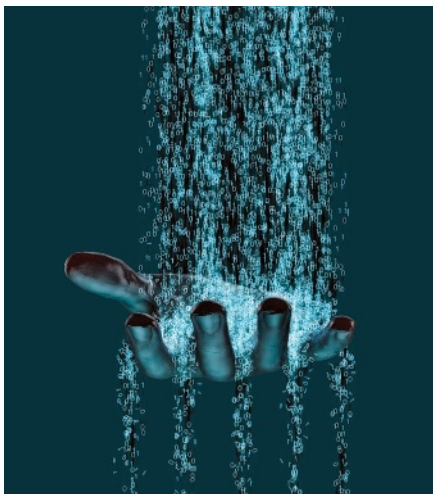
Network.
Collaborate.
Create.

Learn about IEEE Collabratec at
ieee-collabratec.ieee.org

IEEE

IEEE potentials
THE MAGAZINE FOR HIGH-TECH INNOVATORS

July/August 2018
Vol. 37 No. 4



ON THE COVER:
Diving deep into data analytics.

COVER IMAGE: ©ISTOCKPHOTO.COM/GRANDEDUC

THEME: DATA ANALYTICS

DEPARTMENTS & COLUMNS

MISSION STATEMENT: *IEEE Potentials* is the magazine dedicated to undergraduate and graduate students and young professionals. *IEEE Potentials* explores career strategies, the latest in research, and important technical developments. Through its articles, it also relates theories to practical applications, highlights technology's global impact, and generates international forums that foster the sharing of diverse ideas about the profession.

IEEE

Digital Object Identifier 10.1109/MPOT.2017.2783532

IEEE Potentials (ISSN 0278-6648) (IEPTDF) is published bimonthly by The Institute of Electrical and Electronics Engineers, Inc. **Headquarters address:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. Phone: +1 212 705 7900. **Change of address** must be received by the first of a month to be effective for the following issue. Please send to IEEE Operations Center, 445 Hoes Lane, Piscataway, NJ 08854. **Annual Subscription**, for IEEE Student members, first subscription US$5 included in dues for U.S. and Canadian Student members (optional for other Student members). Prices for members, nonmembers, and additional member subscriptions are available upon request. **Editorial correspondence** should be addressed to IEEE Potentials, 445 Hoes Lane, Piscataway, NJ 08854. Responsibility for contents of papers published rests upon authors, and not the IEEE or its members. Unless otherwise specified, the IEEE neither endorses or sanctions any positions or actions espoused in *IEEE Potentials*. All republication rights including translations are reserved by the IEEE. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. copyright law, for private use of patrons, articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For other copying, reprint, or republication permission, write to *IEEE Potentials* at Piscataway, NJ. All rights reserved. Copyright © 2018 by the Institute of Electrical and Electronics Engineers, Inc. Printed in U.S.A. **Subscriptions, orders, address changes:** IEEE Operations Center, 445 Hoes Lane, Piscataway, NJ 08854, Phone: +1 732 981 0060. Other publications: IEEE also publishes more than 30 specialized publications. **Advertising Representative:** IEEE Potentials, 445 Hoes Lane, Piscataway, NJ 08854, Phone: +1 732 562 3946. **IEEE Departments:** IEEE Operations Center (for orders, subscriptions, address changes, and Educational/Technical/ Standards/ Publishing/Regional/Section/Branch Services) 445 Hoes Lane, Piscataway, NJ 08854, USA. Operations Center +1 732 981 0060; Washington Office/Professional Services +1 202 785 0017. Headquarters: Telecopier +1 212 752 4929, Telex 236-411.

Periodicals postage paid at New York, NY, and at additional mailing offices. **Postmaster:** Send address changes to IEEE Potentials, IEEE, 445 Hoes Lane, Piscataway, NJ 08854, USA. Canadian Publications Agreement Number 40030962. Return Undeliverable Canadian Addresses to: Fort Erie, ON L2A 6C7 Canada. Canadian GTS #125634188.

PRINTED IN THE U.S.A.

Digital Object Identifier 10.1109/MPOT.2017.2783533

Pie in the sky

by Cristian Quintero

This issue of *IEEE Potentials* focuses on data analytics. I would like to begin with a story that exemplifies the topic.

"Hello, Pizza Planet?"

"No sir. This is Google Pizzeria."

"Excuse me. I must have dialed incorrectly."

"No sir, you dialed correctly. Google bought the Pizza Planet chain."

"Ah, well, can you please take my order?"

"The usual?"

"How do you know what I usually order?"

"We have a caller ID, and, according to your phone number, we know that the last 53 times you called, you ordered a large Neapolitan pizza with ham."

"Yes, that is what I want."

"Can I suggest a pizza without salt, with ricotta, arugula, and sun-dried tomatoes?"

"No thanks. I hate vegetables!"

"Your cholesterol is not good, sir."

"What?!"

"Our company has the largest database on the planet. We crossed data with your health insurance and have the results of your last seven blood tests. I see that your triglycerides have a value of 180 mg/dL, and your LDL is …"

"All right—enough! I want the Neapolitan! I take my medication!"

"Excuse me, sir, but according to our database, you don't take it often. The last time you bought your cholesterol medication was three months ago, and the box has 30 tablets."

"But I bought more at another pharmacy!"

"Your credit card statements do not show it."

"I paid in cash!"

"The details of your withdrawals with your debit card do not show it."

"I have another source of cash!"

"Your latest income statement does not prove it. We do not want you to have problems with the IRS, sir."

"Stop it. shut up!"

"Excuse me, sir, we just want to help you."

"Help me? I'm tired of Google, Facebook, Twitter, WhatsApp, Instagram! Do I have to move to an island without Internet, cable, or a cell phone to get some privacy?"

"I understand, sir, but that will be difficult for you."

"Why?"

"It turns out that your passport expired five months ago."

"Fine. Just send me the vegetarian pizza."

In this issue, you will find some interesting articles focusing on how data is analyzed and managed.

Currently, data is of the utmost importance, and how to manage it is a wide field of study. Our privacy is at stake, but is this necessarily bad? Well, it varies from person to person, but in this issue, you will find some interesting articles focusing on how data is analyzed and managed. Just remember to always read the conditions the next time before clicking on or signing the "terms of agreement."

About the author

Cristian Quintero (cristianquintero@ieee.org) is the student editor of *IEEE Potentials*.

Digital Object Identifier 10.1109/MPOT.2018.2829320
Date of publication: 11 July 2018

IEEE ResumeLab—A suite of career management tools

by J. Patrick Donohoe

If a job search is in your near future, you should take advantage of the IEEE ResumeLab. Much more than a highly effective resume builder, the IEEE ResumeLab is actually a diverse suite of career management tools. In addition to the professional resume builder, the suite contains a portfolio builder; training in interview skills; tools to assist in writing effective letters, developing a video resume, and skills assessment; and a website builder for the online presentation of proficiencies and abilities.

Starting from scratch, the resume builder provides hints on what sections should be included, how to order sections, how to promote experience, and various ways to style your resume. You can browse sample resumes with numerous section arrangements, and many useful tips are provided. The resume builder provides a share button to quickly and easily exchange your resume with career counselors or potential employers. The video resume builder provides detailed information on generating a professional resume in video form.

The portfolio builder helps you create an online gallery of your work that can be downloaded as a zipped file. Online portfolios document your education, work samples, and skills. It is useful when applying for jobs and training programs, demonstrating your transferable skills, and tracking your professional development. The portfolio builder provides information on organization, project tools, image displays, and how to share your portfolio.

The letter builder aspect combines expert advice and high-quality samples to help create focused, targeted letters for a wide variety of opportunities. Specific details are provided on how to develop, style, edit, and share the letter. IEEE ResumeLab will help you develop the various letters required when searching for a job and assist you in communicating optimally throughout the hiring process.

IEEE ResumeLab's mock interview module provides a tool for preparing for one of the most important components in the employment process. The module has more than 900 typical questions to select from and provides a variety of interview types. Additionally, built-in tips are provided to help you with answering questions. You can record your interview to share with a mentor to gather feedback.

The content generated in the other IEEE ResumeLab modules is easily added to your website, including resumes, letters, portfolios, skills, and video resumes.

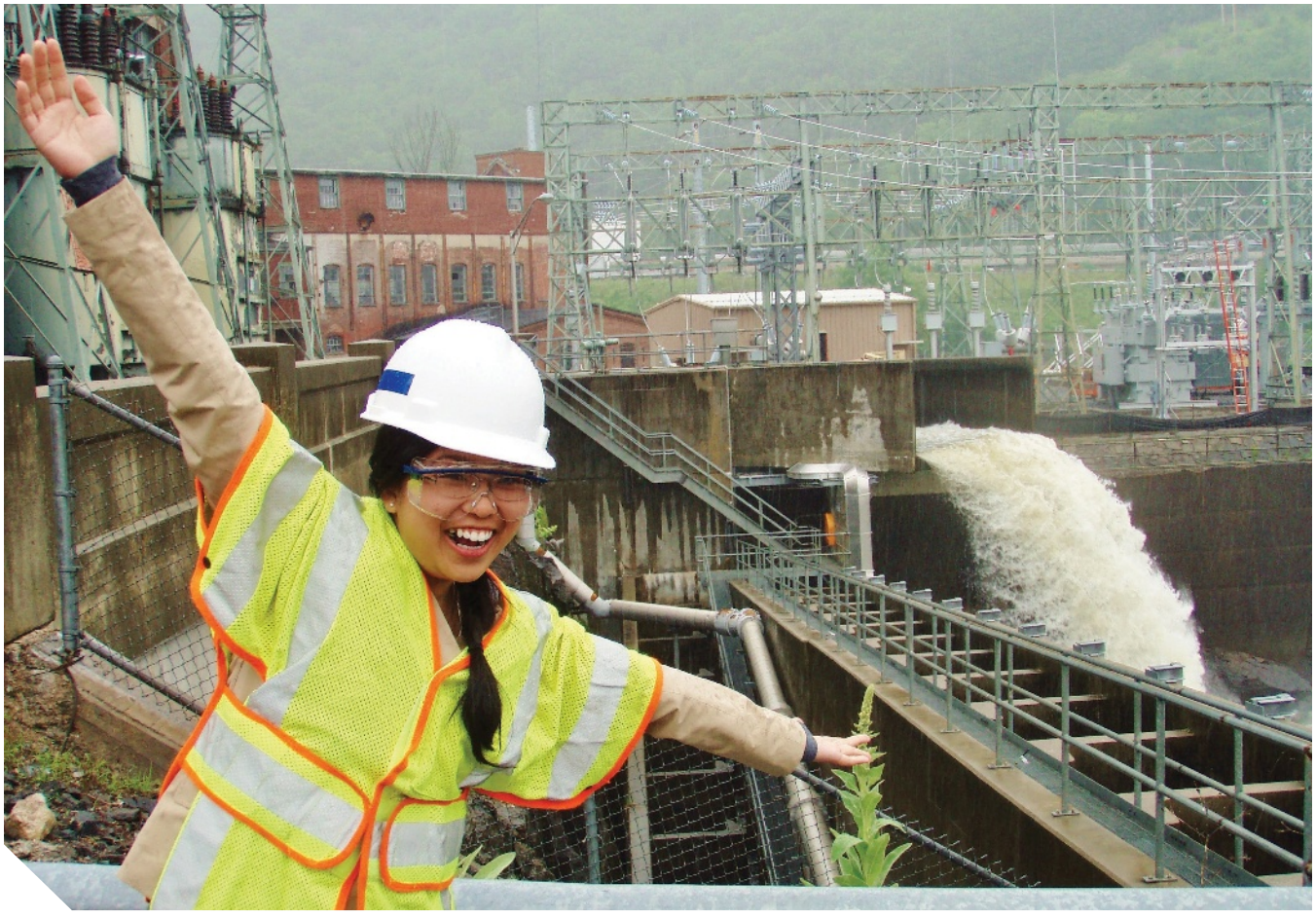
If you are thinking of changing careers or are entering the workforce for the first time, the skills assessment module is a great place to start. Using this module, you can identify and present your transferable skills and abilities to prospective employers. This tool provides information on developing and sharing a skills inventory, an accomplishment sheet, and a proficiency list.

The website builder provides dozens of templates to choose from when starting the build. The content generated in the other IEEE ResumeLab modules is easily added to your website, including resumes, letters, portfolios, skills, and video resumes. You can learn more about the IEEE ResumeLab, and sign in using your IEEE account, by visiting https://www.ieee.org/membership/products/resumelab.html.

About the author

J. Patrick Donohoe (p.donohoe@ieee.org) is the IEEE Member and Geographic Activities chair–Student Activities.

Digital Object Identifier 10.1109/MPOT.2018.2829018
Date of publication: 11 July 2018



Be the force behind change

Bring the promise of technology — and the knowledge and power to leverage it, to people around the globe. **Donate now to the IEEE Foundation and make a positive impact on humanity.**

- **Inspire technology education**
- **Enable innovative solutions for social impact**
- **Preserve the heritage of technology**
- **Recognize engineering excellence**

IEEE Foundation

Discover how you can do a world of good today.

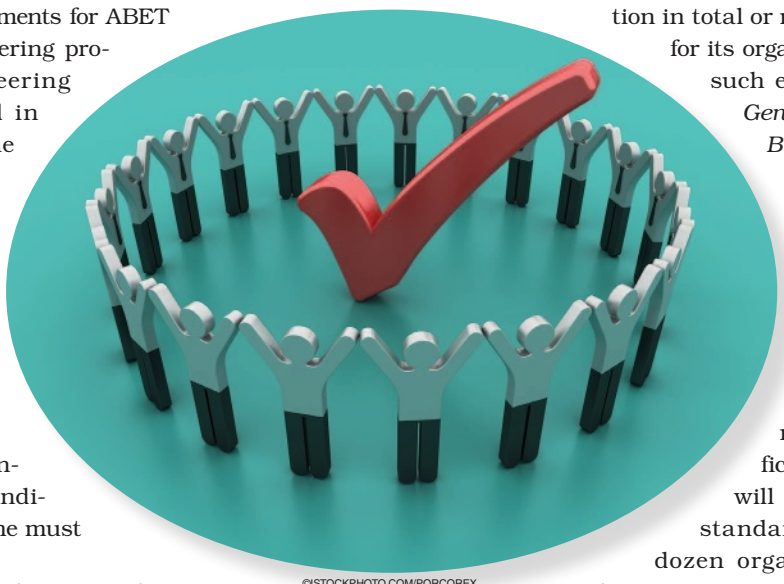
Learn more about the IEEE Foundation at **ieeefoundation.org**.
To make a donation now, go to **ieeefoundation.org/donate.**

IEEE

Standards and you

by Raymond E. Floyd

There are literally dozens of standards associations worldwide (almost every nation has some type of regulating unit for standards). Some relate only to national equipment requirements, while several are recognized on a global basis, where their standards must be adhered to for product acceptance. Unfortunately, most engineering students will get, at best, limited exposure to standards during their academic experience. If discussed at all, the topic will most often be covered as part of the capstone project during their senior year of college. One of the requirements for ABET accreditation of engineering programs is that engineering standards be covered in some manner, but the depth of the coverage is not prescribed. Regardless of the particular field of engineering the student pursues, there will be standards that apply. Whether a person is designing a new television or planning an installation of an air-conditioning system, he or she must adhere to standards.



©ISTOCKPHOTO.COM/PORCOREX

As previously noted, there are a large number of standards organizations. Of that large number, there are a few that are most often noted as being applicable to a product or process in particular. If you look at an appliance, computer, battery pack, or most any power tool, a series of symbols and numbers can be found. These symbols identify a particular organization's trademark, while the numbers correspond to the specific applicable standard appropriate to that product. Some of the more recognizable organizations are listed in Table 1.

Standards provide specific requirements for products and installations. For example, a search of UL standards with the lookup being "appliances" will return more than 50 standards applicable to household and/or commercial product standards ranging from the installation of heater vents to solid-state controls for kitchen appliances. In some cases a standard written by an organization may also be codified by another organization in total or modified appropriately for its organizational needs. One such example is UL-2595—*General Requirements for Battery-Operated Appliances* has a counterpart released by CSA Group—CSA C.22.2 No. 0.23-2015, *General Requirements for Battery-Operated Appliances*. Similarly, a search for radio-frequency identification (RFID) standards will reveal more than 270 standards released by some dozen organizations (including the ISO, IEEE, ANSI, ASTM, and Deutsches Institut fur Normung), with many of the released standards covering the same topical area.

All standards do not relate to products or systems implementations. Perhaps one of the most recognizable standards is published by the ISO (ISO-9000). This standard provides the requirements for a quality management system. It does not furnish requirements for products but is more closely related to processes and/or procedures that, if followed, will facilitate an environment for an organization to produce a quality product. ISO was established in 1946 and now consists of more than

Digital Object Identifier 10.1109/MPOT.2016.2549988
Date of publication: 11 July 2018

160 national country members, supported by approximately 3,500 technical bodies, and has published more than 20,000 standards across a broad spectrum of applications and/or products.

An engineering guaduate will cross paths with a multitude of standards that will impact his or her day-to-day work, whether in product development, manufacturing, field engineering, product test, or other assignments.

What's the point?

You might question the purpose, or need, for such a multiplicity of standards. Why not just one simple standard for household appliances? If one thinks about the great variety of devices that can be classed as an "appliance," it becomes more obvious as to why so many standards are required. A standard describing requirements for a gas-powered stove has little relevance to an electric-powered convection oven.

Another reason for standards is the expectation level of the user community. Without standards, the product marketed could contain extreme differences, even in the same product line. Consider a television set: Where would we be if there were no standards, much less governmental regulations, concerning such systems? It could be so extreme that some sets worked and some would not, which is total chaos for the consumer.

A couple of examples of how standards are developed and how the different standards communities work together to develop a general standard to address different requirements for some particular product are helpful, especially in the international community. The American Association of Railroads (AAR) had a major problem of tracking assets throughout the vast rail network. At times, a particular asset might disappear from one railroad's tracks for months, or even years, at a time. This, coupled with foreign rail assets being on their tracks, made for difficult and expensive tracking. The AAR developed a standard method of identification using RFID tags on every rail vehicle. It took time to develop the agreements between railroads (where to place the tag, what data would be included, how would the data be captured, and how would the data be shared). All of the railroad companies had to agree on the approach and cost of implementing the project. Once that challenge was complete, the AAR had to work with RFID suppliers to ensure that the tags being supplied met all of the AAR requirements—thus a standard rail tag was agreed upon, developed, and implemented and is currently in use across the United States and Canada.

Another example can be found in the all-terrain vehicle (ATV) market. As originally designed and sold, the ATV was for backwoods entertainment and transportation. Even in its early introduction, the Specialty Vehicle Institute of America (SVIA) and ANSI worked together to fashion a standard to best describe the safety features required to protect the public user (ANSI/SVIA-1-2010). As the use of ATVs expanded and some began to traverse public roadways, the U.S. Consumer Product Safety Commission (CPSC) and the National Highway Safety Administration (NHSTA) began to work with both ANSI and SVIA to extend the safety standards required for ATVs to ensure the best possible safety features for that product line. Note that none of the organizations involved are manufacturers of ATVs. The most recent release of ANSI/SVIA-1-2010 included all of the requirements originally established by ANSI and SVIA and have been expanded according to the needs of CPSC and NHSTA.

Standards are not limited to one nation or adjoining nations. ISO has many standards that are recognized and adhered to worldwide. Again, an example from the transportation industry is the identification tagging of shipping containers. This application of RFID technology involves mounting RFID tags on shipping containers. In this particular instance, ISO had to get international agreements on the application as well as determine such things as tag location, tag data content, and, most difficult, a common carrier frequency for communications with the tag. The ISO did not get involved with the data collection or distribution, as that was left to the particular user requirements.

It should be noted in all of these discussions that none of the standards were written by a particular manufacturer. It's obvious that such an approach would not necessarily result in the best standard—except as it applied to that particular manufacturer's product line. (Not necessarily endearing confidence on the part of the user!)

TABLE 1. Common standards groups.

ANSI – American National Standards Institute
API – American Petroleum Institute
ASTM – American Society of Test and Materials
CEN – European Committee for Standardization
CSA – Canadian Standards Association (now CSA Group)
ICC – International Code Council
IEEE – Institute of Electrical and Electronic Engineers
ISO – International Organization for Standardization
NEMA – National Electrical Manufacturers Association
NIST – National Institute of Standards and Technology
UL – Underwriters Laboratory

A multitude of standards

An engineering graduate will cross paths with a multitude of standards that will impact his or her day-to-day work, whether in product development, manufacturing, field engineering, product test, or other assignments. Some affect design efforts, while others assist in product evaluations and manufacturing. There will also be opportunities to assist in the generation of standards and/or modifications to existing ones as they need updating as technology changes. Standards are important to product development, manufacturing, and test, but most essential is their effect on consumers. In alignment with that, the United States established the U.S. Consumer Product Safety Commission, which provides voluntary and mandatory standards as well as bans applicable to products. Students will find that it can be very expensive to purchase a particular standard or set of standards. Fortunately, most college libraries have working agreements with multiple libraries and can obtain copies without charge to the student, or, at worst, for a very nominal service charge. Understand applicable standards in your career activities—they are important to your success as well as the success of any product with which you become involved.

About the author

Raymond E. Floyd (r.floyd@ieee.org) earned his B.S.E.E. degree from the Florida Institute of Technology in 1970, his M.S.E.E. degree from Florida Atlantic University in 1977, and his Ph.D. degree in industrial management from California Coast University in 2009. He spent 26 years with IBM, retiring in 1992 as a senior engineer. He is a Life Senior Member of the IEEE, a life senior member of the Society of Manufacturing Engineers, and holds four patents. He has served as a program evaluator for the Engineering Technology Accreditation Commission of ABET (ETAC/ABET) for 20 years and is an associate editor of *IEEE Potentials*.

Solution #1: Monk Spot

Instead of one monk climbing the hill one day and returning the next, consider two men: one going up and one coming down on the same day (both on the same path and walking at the same speed as the monk). Where they meet is the place that the monk passed at the same time going both ways.

NUMBERS—© CAN STOCK PHOTO/123DARTIST, ANDROID—© CAN STOCK PHOTO/KIRSTYPARGETER

Solution #2: The Full Monty

The obvious answer is that it doesn't matter which of the two closed doors you choose. The probability that the Ferrari is behind either one of them is 1/2, right? You can get this result by intuition or applying Bayes' theorem. Unfortunately, it's wrong. Here is the correct solution: If your strategy is not to switch, it doesn't matter whether the host opens a door or not, and your probability of winning the car is 1/3. If you decide to switch, you'll lose only if you originally choose the door concealing the Ferrari. If you choose either of the other two doors, you'll win. So, your probability of winning is 2/3. The best bet is to switch your choice.

Solution #3: The Fuller Monty

Without switching, you win if you choose the door concealing the Ferrari and lose if you choose another door. So if there are n doors, your probability of winning is $1/n$. With switching, you lose if you originally choose the door concealing the Ferrari and win if you originally choose any of the other doors. So, your probability of winning is $(n - 1)/n$.

Solution #4: Lock, Stock, and Sinking Barrels

When some of the cargo falls off of a ship in a closed canal lock and sinks, it lowers the water level. When cargo was on the ship, it displaced a volume of water that weighed the same as the cargo. When in the water, the cargo displaces a lesser volume of water, namely its own volume.

Solution #5: By Hook or By Crook

Only one politician is honest, and 99 are crooked.

Digital Object Identifier 10.1109/MPOT.2018.2828179
Date of publication: 11 July 2018

Data analytics implications

Paul C. Hershey

This special theme issue of *IEEE Potentials* introduces the rapidly evolving area of data analytics, which encompasses data analysis, data fusion, data storage, data sources, infrastructure and technology, screening and filtering algorithms, machine learning, and complexity. An emergent technology from data analytics is big-data analytics, defined by the Gartner IT Glossary as "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." Big data definitely changes the way we approach data analytics. For example, we favor very large volumes of data over smaller volumes, even if the smaller volumes are more accurate. Also, our big-data analytics is more focused on probabilities and correlations over causality and certainty. In fact, *The Fourth Paradigm: Data-Intensive Scientific Discovery* states that "increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets."

Clearly, big-data analytics has great implications for engineering students preparing to enter the work force, regardless of their specific engineering discipline. A data scientist is a person specifically employed to analyze and interpret complex digital data. Data scientist roles have grown over 650% since 2012. Job growth in the next decade is expected to create 11.5 million jobs by 2026, according to the U.S. Bureau of Labor Statistics.

In this issue of *IEEE Potentials*, we present six theme articles describing aspects of data analysis that will be of interest to students who wish to become data scientists as well as those considering work in operations centers who may be responsible for the processing, exploitation, and dissemination of large amounts of data collected from a wide variety of sensors. Likewise, systems engineering students focused on the analysis of complex systems, mission support, and automated decision aides are included in the target audience.

- Distributed computing (DC) is a foundational technology that undergirds much of the modern global economy. Hosted by massive data centers humming with thousands (or hundreds of thousands) of servers, DC facilitates ecommerce, social media, logistics,

Digital Object Identifier 10.1109/MPOT.2018.2824361
Date of publication: 11 July 2018

 0278-6648/18©2018IEEE

government, finance, gaming, and other mass-digital phenomena for hundreds of millions of people every day. Without it, many of today's online services would either not exist or be greatly diminished in usefulness. Note that while the DC field covers much ground across multiple domains, the focus in Daryl Nelson's article is on data-intensive processing at scale that is typically in a data center, sometimes referred to as *big-data analytics*.

- Mobile communications capabilities have been demonstrated to be invaluable with respect to transfering large amounts of data without the need for an existing infrastructure or, in emergency situations, where power is lost or extra network capacity is needed. Aerial platforms, such as unmanned aerial systems, have been shown to be ideal for use in maintaining mobile networks. They enable deployment in areas impossible for other vehicles to occupy while maintaining the necessary mobility to provide coverage to highly dynamic or widely dispersed networks. Paul C. Hershey, Mu-Cheng Wang, and Steven Davidson present an agent-based bandwidth reservation technique for data-intensive mobile networks that allocates bandwidth based on requirements and currently available resources.
- Cloud computing provides a single platform for organizations to consolidate data from all channels at a big-data scale—for data with huge volume, variety, and velocity. The key take away from this is that data are continuously collected in an endless stream, but data are useless without analytics. The infrastructure and computing power required to store big data and run advanced analytics algorithms can kill the thought of big-data analytics for many firms. Cloud computing and analytics provide a cost-effective solution for businesses to take advantage of the billions of bytes of data generated by their services and customers to maximize their businesses and profits. Ankita Christine Victor and Shrisha Rao explore the topic of analytics on the cloud.
- Visual analytics describes the use of data analysis and visualization for design decisions where each step has consequences for how different aspects of the data are emphasized, obscured, or contextualized. Analyzing numerical data is an essential phase of most smart grid research. We will explore trends, search for patterns, check relationships, and inspect distributions in our increasingly large, interconnected data set. Yet it can be rather challenging to produce graphical representations of data that are legible, honest, and attractive. It is all too easy to confuse ourselves and deceive others. Paul Cuffe, Harold Kirkham, Chris Dent, and Amy Wilson tackle the ins and outs of visual analytics.
- Machine learning (ML) takes advantage of data analytics to support autonomous decision-making strategies for a number of disciplines, ranging from biology to economics to engineering. Biologists have successfully applied ML to solve problems such as the analysis of protein cell interactions, while economists have used it to predict financial market performance. In engineering, ML has been used for classification, signal processing, system identification, and control systems. Sidney Givigi and Peter Travis Jardine delve into ML as it relates to the data-driven control of robots.
- Video analytics is used to support tactical military operations and related applications. Prompted in part by technology; advances in low-cost, high-resolution cameras; wireless and wired networking; and computer vision, video surveillance has become ubiquitous in homeland protection, tactical military operations, and business and home owner security applications. As a result, homeland security and military forces are being overwhelmed by a glut of live and archived video streams. These video streams can help personnel quickly and accurately access a situation and make appropriate decisions but do not lend themselves to rapid search and/or review. Susan Gottschlich, Brian Stone, and Bill Gerecke review research related to video surveillance analytics and describe how ongoing advances may be leveraged and expanded upon to address video surveillance stream searchability and discovery. They discuss how this can help address the video glut problem for military tactical applications and closely related security and defense applications.

Our big-data analytics is more focused on probabilities and correlations over causality and certainty.

Read more about it

- Adapted from Gartner IT glossary, (2018). [Online]. Available: http://www.gartner.com/it-glossary/big-data/
- K. Cukier and V. Mayer-Schoenberger, "The rise of big data," *Foreign Affairs*, vol. 92, no. 3, p. 29, May/June 2013.
- T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research, 2009.
- L. Columbus. (2017, Dec. 11). LinkedIn's fastest-growing jobs today are in data science and machine learning. [Online]. https://www.forbes.com/sites/louiscolumbus/2017/12/11/linkedins-fastest-growing-jobs-today-are-in-data-science-machine-learning/

About the author

Paul C. Hershey (Paul_C_Hershey@raytheon.com) is a principal engineering fellow at Raytheon Company, Intelligence, Information, and Services in Dulles, Virginia.

Distributed computing: The unsung hero of the modern global economy

Darryl Nelson

Invisible to all but small groups of elite practitioners, distributed computing (DC) is a foundational technology that undergirds much of the modern global economy. Hosted by massive data centers humming with thousands (or hundreds of thousands) of servers, DC facilitates e-commerce, social media, logistics, government, finance, gaming, and other mass-digital phenomena for hundreds of millions of people every day. Without it, many of today's online services would either not exist or be greatly diminished in usefulness.

Once the domain of a very small group of experts, DC began seeing widespread adoption with the increasing availability of large data sets, a generation of new software frameworks, and ever-cheaper and more powerful hardware. This article will discuss what DC is, why it matters, its challenges, two prominent architecture patterns, trends, and resources to learn more. Note that, while the DC field covers much ground across multiple domains, the focus here will be on data-intensive processing at a scale that is typically in a data center, sometimes referred to as *big-data analytics*.

DC and why it matters

DC handles a large problem (e.g., indexing millions of documents) by splitting it into many small workloads that are dispersed and processed across a set of networked computers. DC employs *distributed systems*, which is defined as a computer system "in which hardware or software components located at networked computers communicate and coordinate their actions only by passing messages" (Coulouris et al.). The system achieves a common goal (e.g., Internet-scale query) through the interaction of these components.

DC matters because it is the predominant technology to handle the modern "data deluge," where petabyte-scale+ volumes are increasingly the norm. Single computers cannot store and process the continuous generation of vast amounts of data. By using a "divide and conquer" approach, multiple computers can decompose a massive computational or storage problem in ways that can be effectively managed. The implication for software is profound, as systems must be explicitly designed to operate in this "the data center as a computer" model (Barroso et al.). Properly

Digital Object Identifier 10.1109/MPOT.2018.2824378
Date of publication: 11 July 2018

 0278-6648/18©2018IEEE

engineered distributed systems will exhibit the essential nonfunctional attributes of both scalability and availability.

Scalability and availability

Through DC, software systems have both scalability and availability. *Scalability* is the ability to add more computing resources (servers) when the demand for storage or processing outgrows the current footprint of servers. DC allows a system to *partition for scale*, splitting up a large computational problem into small tasks. Scalable systems are said to be *elastic* when you can easily add or remove servers in an automated way that does not disrupt operations. One of the value propositions of cloud computing is elastic scalability.

There are essentially two types of scalability: vertical and horizontal. With vertical scalability, a system handles need resource demands by getting a better computer [e.g., a faster central processing unit (CPU), more random-access memory, or bigger hard drives]. This type remains useful, but you quickly hit a scalability brick wall when the demands on your system overwhelm the available resources. Many Internet companies initially took the vertical scalability approach during the dot-com bubble of the late 1990s/early 2000s and quickly concluded that even the most powerful computers were not up to the task. They turned to horizontal scalability out of necessity, and the tech world has not looked back since.

In contrast, the horizontal model scales by adding a new server instances to bring more storage and/or processing online to meet demand, as shown in Fig 1. DC is all about horizontal scalability. Servers are either added or taken away depending on the current scaling requirements. Typically deployed on commodity hardware, horizontally scalable software knows how to operate across multiple computers in a coordinated way. This model places very specific demands on software; more on that later.

Availability provides fault tolerance in the face of failure, including hardware and software (as well as human-induced errors, which is the most common type). For example, a highly available system will remain operational despite a hard drive or network failure (or both simultaneously). DC replicates software, hardware, and data to achieve a certain level of availability that is typically described in terms of "the Nines": 99.999% ("five nines") of availability has approximately 5.26 min of downtime per year.

The challenges of DC

Everything in engineering involves tradeoffs and DC is no exception. Despite many advances in software frameworks, architecture models, and infrastructure, DC remains a challenging endeavor. The reason is omnipresent failure in software, hardware, wetware (human), and networking (see Fig. 2). Omnipresent failure pervades DC for two fundamental reasons:

1) independent things fail independently (Takada)
2) the network is inherently unreliable (Rotem-Gal-Oz).

Failure is pervasive because there are many components that need to coordinate among themselves. Likewise, each individual component has its own potential for failure. Failure modes at multiple granularities include in software, hardware, wetware (human), and networking. The cumulative effect of software bugs, hard-drive crashes, ongoing maintenance, router failure, overheating,

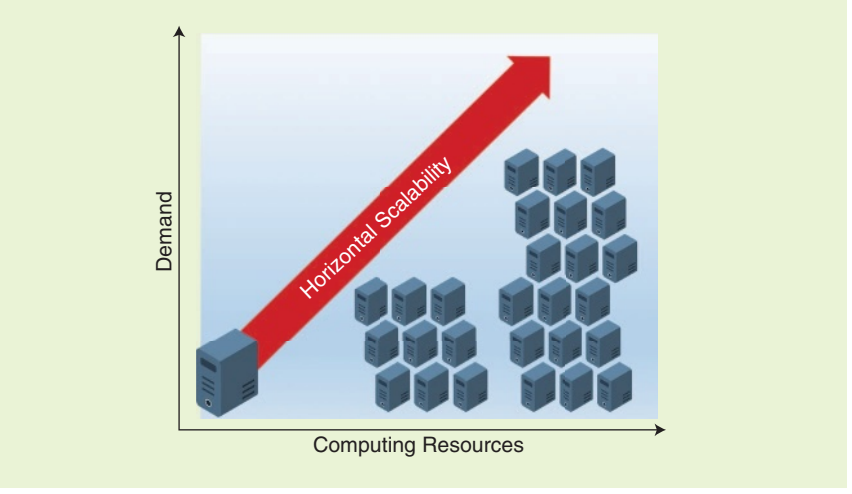


FIG1 Horizontal scalability.

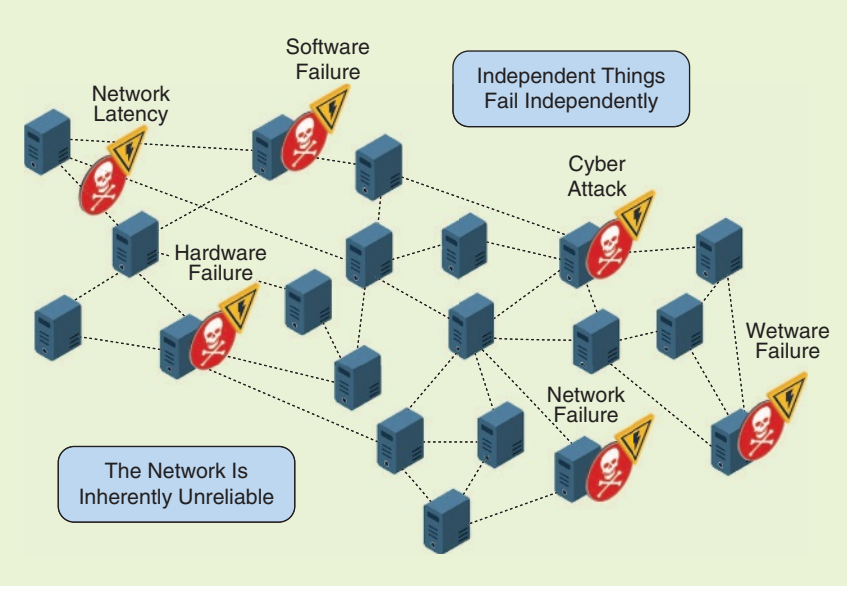


FIG2 Failure modes in DC.

unintended power shutdowns, data corruption, and cyberattack (to name just a few) all contribute to a challenging technological problem. In this environment, poorly designed software does not have a chance.

Architecting software for DC

Software suitable for DC must be designed for it from the outset; that is, it is built to be distributed from the very beginning. "Bolting on" scalability and availability to a program originally built for a single computer is to invite frustration and near-certain failure. Software that congenitally "bakes in" a distributed capability is essential. Properly designed software expects all types of failure and is inherently fault tolerant. Likewise, a DC program must pass the scalability test: You have a scalable architecture when, under pressure to scale, you need a new instance, not a new architecture. In other words, a scalable software system does not need to be re-designed when the demands on the system exceed the current capacity. New software servers for computation, storage, or both can be started seamlessly to meet the new demand.

The advent of Apache Hadoop in 2006 ushered in the big-data era by dramatically lowering the barriers to entry for DC. Hadoop advanced the state of the art by hiding the complexity of DC through a simplified application programming interface (API) to allow developers to focus on business problems, not lower-level details of distributing computing. Based on a seminal *MapReduce* paper by Google engineers (Dean and Ghemawat), Hadoop allowed a much broader audience to leverage the power of distributing processing system. Since then, more open-source frameworks have emerged as the field learned hard-won lessons in designing and building distributed systems.

One increasingly popular DC software stack for data-driven applications is the SMACK stack (McFadin). This complementary set of technologies was coined *SMACK* due to the choice of the open source frameworks Spark, Mesos, Akka, Cassandra, and Kafka. It allows large-scale systems for data collecting, processing, and storage. The SMACK stack offers a comprehensive set of capabilities to cover many common data processing use cases. Table 1 describes the components of the SMACK stack.

Despite the increasing capability and maturity of DC software offerings in processing and storage, systems still require significant forethought and consideration in software architecture. Two architecture models, Lambda and Kappa, have emerged to structure the elements of distributed systems, particularly for big-data analytics applications.

TABLE 1. The SMACK stack.

COMPONENT		DESCRIPTION	
Apache Spark		General-purpose engine for big-data processing, with built-in capabilities for streaming, Structured Query Language (SQL), machine learning, and graph operations	https://spark.apache.org/
Apache Mesos		Resource management to deploy and manage service, disk allocation, CPU, and memory	https://mesos.apache.org/
Akka		Actor-based concurrency for data processing	https://akka.io/
Apache Cassandra		Highly available, horizontally scalable NoSQL database-management system	http://cassandra.apache.org/
Apache Kafka		Stream processing platform	https://kafka.apache.org/

The Lambda and Kappa architecture models

In 2015, Nathan Marz and James Warren published the book *Big Data* that describes the Lambda architecture, a way to combine batch and real-time processing to handle massive volumes of data (Marz and Warren). Lambda is an attempt at a generic, scalable, and fault-tolerant data-processing architecture. In this model, batch processing provides comprehensive and accurate views of the data in aggregate and real-time stream processing to provide views of online data while both view outputs join together for client requests.

Lambda has three layers (batch, serving, and speed) each with a distinct function. Figure 3 illustrates the different layers and interactions among them. The batch layer literately re-computes all the data stored in the master data set. This data is safeguarded against loss or corruption through redundancy or replication. The master data set allows you recover from catastrophic failure (by recomputing the raw data), as well as extract future, undiscovered value in the data.

The output of the batch later is stored in the serving layer and is completely replaced with each processing iteration. Incoming data into a Lambda system simultaneously

enters both the batch layer (in the master data set) as well as the speed layer. The speed layer incrementally processes the incoming data missed by an iteration of the batch-layer processing. Data in the speed layer is transitory and deleted once its data is processed in the batch layer. Note that the computation functions (e.g., analytics) in the batch and speed layers are the same but may likely be implemented with different programming languages or frameworks. Requests for the data products are aggregations of the speed and serving layers.

Critiques of the Lambda model have emphasized the potential maintenance footprint and the inherent complexity of layers with duplicated functionality. Regardless, the Lambda model reintroduced much-needed architectural rigor into the field and, for many organizations, served as a step function to facilitate business value from DC.

While at the professionally oriented social media company LinkedIn, Jay Kreps published his views on the Lambda approach as well as an alternative model, the Kappa architecture (and later referred to as the *Stream Data Platform*). Kappa is a stream-centric approach where continuous flows of data in time-stamped, append-only data structures (logs) is the organizing principle of this architecture. Beyond the obvious traditional stream processing, Kappa also handles batch modes by streaming repositories of historic data and then processing the data as a stream. In Kappa, everything is a data stream.

This Stream Data Platform (see Fig. 4) serves as a highly decoupled, central data pipeline where each system component (analytics engines, search indexes, data stores, monitoring apps) can feed or be fed by the universal pipeline. Furthermore, the system components can read from the pipeline and then derive new streams (e.g., generated metadata). Subsequently, the new feed is available for consumption by other components or systems. These features allow Kappa to promote decoupling through a stream platform, allowing development teams to build a system on a single platform. Both Lambda and Kappa have tradeoffs, strengths, and weaknesses that have stimulated significant discussion on their merits as distributed architectures (Lin).

Current trends in DC

The distributed systems field, like almost all computer technology, continues to evolve at a pace that even dedicated practitioners struggle to keep current. A constant stream of new hyperscalable databases, analytics engines, messaging protocols, and infrastructure advancements can overwhelm, but several key trends have emerged in DC. These include serverless computing in the cloud, machine learning, and chaos engineering.

Serverless computing

Like Hadoop, cloud computing is a democratizing force for DC, allowing a much broader audience to solve large-scale data storage and processing challenges. Many more organizations have on-demand access to computing resources through cloud services. One no longer needs to fund, build, staff, and manage a large data center to process mass volumes of data; compute, memory, and storage can now be consumed as a service. This approach has been extended, where cloud vendors now offer a serverless model, including Amazon, Microsoft, and Google. It adds an additional complexity-hiding abstraction layer to expose analytics, messaging, database, and other functional services.

For example, instead of installing, configuring, and maintaining the SMACK stack, developers can focus on business logic by coding against cloud APIs without worrying about provisioning or administering servers/software. The potential benefits include significant productivity gains whereas the possible tradeoffs include vendor lock-in and cost containment, among others.

Machine learning

The driving force behind the current "artifical intelligence spring" is machine learning, where a computer program is "said to learn if its performance of a certain task, as measured by a computable score, improves with experience" (Brink et al.). The key here is *learn*; unlike traditional rules-based analytics, machine learning applications do not need to be explicitly programmed. Machine learning software learns from the software

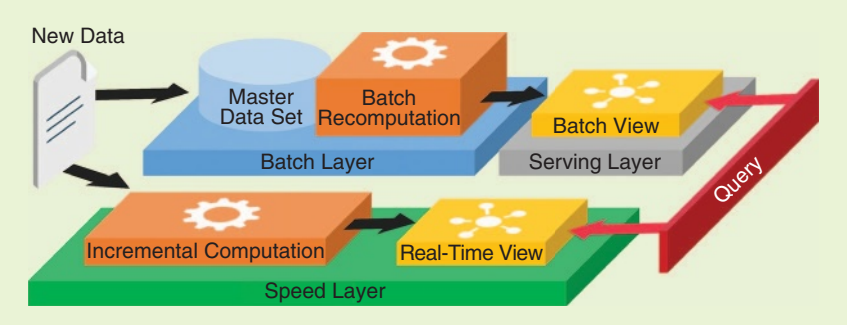


FIG3 The Lambda architecture.

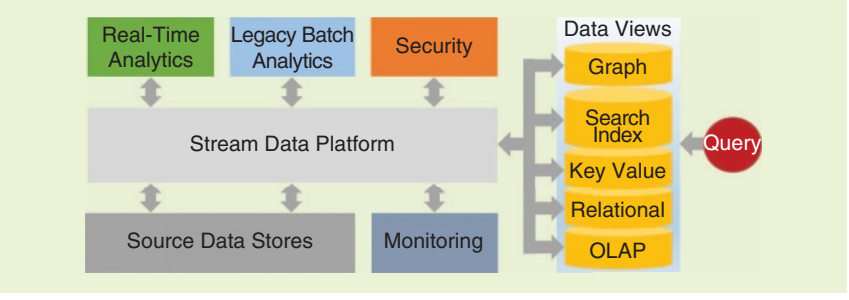


FIG4 The Kappa architecture, the Stream Data Platform.

fed to it (known as *training*). Machine learning applications inherently require data for training and the more data, the better. Therefore, DC is increasingly a key enabling technology for machine-learning systems. Distributed systems using machine learning must handle both model training and serving (deploying a model in product to make predictions on live data feeds) at scale. An example is Google's TensorFlow Extended, a scalable platform design for machine learning. (Baylor et al.). Similarly, the machine-learning era has spawned a vigorous investment and adoption of advanced specialized hardware including graphical processing units and tensor processing units. It is a safe bet that machine learning will increasingly dominate DC analytics projects today and in the future. Those in the software field would be wise to study and embrace machine learning as a key tool to solve tomorrow's computing challenges.

Chaos engineering

The sheer potential complexity and voluminous failure vectors of DC have shown that traditional testing methods are inadequate. Necessity being the mother of invention, chaos engineering emerged as "a method of experimentation on infrastructure that brings systemic weaknesses to light" (Rosenthal et al.). The goal of chaos engineering is resilient systems that can operate under conditions of pervasive failure. It injects disorder into a system to expose fragile characteristics.

Netflix has open-sourced the dominant tool set for chaos engineering, Simian Army. It includes applications like Chaos Monkey, which randomly shuts off one or more key components of a system (e.g., an application server) to observe resulting system behavior. Chaos engineering is an increasingly mandatory endeavor for testing and verification of distributed systems and is a rich opportunity for innovation.

Resources

For those seeking a deeper dive into DC, an excellent start is Marz and Warren's book, *Big Data,* about the Lambda architecture. Regardless of one's opinion on the Lambda architecture, the book is an excellent discussion of the various aspects of DC. It is a good primer on the tradeoffs, challenges, and reasoned solutions in the field. Likewise, *Distributed Systems: Concepts and Design (5th Ed.)* is a comprehensive survey of the DC landscape (Coulouris et al.). On the web, the High Scalability website (http://highscalability.com/) is a rich source of DC information as well as other advanced technology topics. Finally, Roland Kuhn's *Reactive Design Patterns* is an outstanding resource for building highly resilient and response software.

Read more about it

- L. A. Barroso, J. Clidaras, and U. Hölzle. (2013). The datacenter as a computer: An introduction to the design of warehouse-scale machines, 2nd ed. Morgan & Claypool Publishers. [Online]. Available: https://research.google.com/pubs/pub41606.html
- D. Baylor, E. Breck, H. Cheng, N. Fiedel, C. Foo, Z. Haque, S. Haykal, M. Ispir, V. Jain, L. Koc, C. Koo, L. Lew, C. Mewald, A. Modi, N. Polyzotis, S. Ramesh, S. Roy, S. Whang, M. Wicke, J. Wilkiewicz, X. Zhang, and M. Zinkevich, "TFX: A tensorflow-based production-scale machine learning platform," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Halifax, Nova Scotia, Canada, 2017, pp. 1307–1395.
- H. Brink, J. Richards, and M. Fetherolf, *Real World Machine Learning*. Shelter Island, NY: Manning, 2017.
- G. Coulouris, J. Dollimore, T. Kindberg, and G. Blair, *Distributed Systems: Concepts and Design*, 5th ed. Boston, MA: Addison-Wesley, 2011.
- J. Dean and S. Ghemawat. (2004). MapReduce: Simplified data processing on large clusters. [Online]. Available: https://research.google.com/archive/mapreduce-osdi04.pdf
- J. Kreps. (2014). Questioning the lambda architecture O'Reilly. [Online]. Available: https://www.oreilly.com/ideas/questioning-the-lambda-architecture
- R. Kuhn, B. Hanafee, and J. Allen, *Reactive Design Patterns*. Shelter Island, NY: Manning, 2017.
- J. Lin, "The lambda and the kappa," *IEEE Internet Computing*, vol. 21, no. 5, pp. 60–66, Sept./Oct. 2017.
- N. Marz and J. Warren, *Big Data: Principles and Best Practices of Scalable Real Time Data Systems*. Shelter Island, NY: Manning, 2015.
- P. McFadin, (2017, July 10). The SMACK stack: A new architecture for today's data-rich modern applications O'Reilly ideas. [Online]. Available: https://www.oreilly.com/ideas/the-smack-stack
- C. Bennett. (2015). Netflix Simian Army Github. [Online]. Available: https://github.com/Netflix/simianarmy/wiki
- C. Rosenthal, L. Hochstein, A. Blohowiak, N. Jones, and A. Basiri, *Chaos Engineering: Building Confidence in System Behavior through Experiments*. Sebastopol, CA: O'Reilly Media, 2017.
- A. Rotem-Gal-Oz. (2008). Fallacies of distributed computing explained. [Online]. Available: http://www.rgoarchitects.com/Files/fallacies.pdf
- M. Takada. (2018). Distributed systems for fun and profit. [Online]. Available: http://book.mixu.net/distsys/single-page.html

About the author

Darryl Nelson (darryl.nelson@raytheon.com) earned his M.E. degree from Texas Tech University. He is a Raytheon Engineering fellow and currently leads a software engineering research and development center specializing in the next generation of advanced analytics and software architectures. His areas of interest include distributed computing for large-scale software systems, analytics including machine learning, the human-intelligent machine interface, and the operational art and tradecraft of software development. He is a U.S. Army veteran.

P

Real-time communications bandwidth allocator

Paul C. Hershey, Mu-Cheng Wang, and Steven A. Davidson

Mobile communications capabilities have been demonstrated to be invaluable with respect to establishing communications without an existing infrastructure, especially in emergency situations where power is lost or extra network capacity is needed. Aerial platforms, such as unmanned aerial systems (UASs), have been shown to be ideal for use in maintaining mobile networks. They enable deployment in areas impossible for other vehicles to occupy while maintaining the necessary mobility to provide coverage to highly dynamic or widely dispersed networks. UASs have also been used for surveillance and reconnaissance operations.

Multiple UASs can also work together to perform cooperative missions. Here the interactions between UASs are not only information exchanges but also physical couplings required to cooperate in the joint

Digital Object Identifier 10.1109/MPOT.2018.2824406
Date of publication: 11 July 2018

transportation of a single load. As missions employ multiple diverse sensors to collect data, it becomes very critical to deliver the time-sensitive information to the ground systems operators and analysts promptly so that the data can be processed, fused together, exploited, and disseminated to commanders and warfighters in the form of actionable information. How to design a communication network to satisfy mission requirements plays a very important role to ensure the success of a mission.

Due to the frequent changes in network topology and link quality, a wireless link may not always sustain its data rate. Instead of using peer-to-peer communication between sensors and ground stations, an alternative solution is to leverage the multihop mobile ad-hoc network (MANET) to opportunistically exploit, for data delivery, nodes' mobility and contacts with other nodes/networks. The MANET can provide sufficient and reliable end-to-end communications to sensors at the run time. When properly embedded into the architecture of a communications network, in addition to sensing and actuation capabilities, multiple UASs can also form a multilayered hierarchical MANET such that the UAS-aided network provides transport of flows that span longer distances, yield better reliability, and achieve higher throughput.

Although MANET can potentially offer multiple routes for each given source and destination pair, it is important for each network node to select an "appropriate" path that can satisfy the individual mission requirements. To achieve this outcome, mission planers should determine mission objectives and relative priorities during the premission or planning phase. Then the planers analyze the mission objectives and determine the resource constraints, such as latency, jitter, and the desired and minimum bandwidth requirements, among others, to accomplish these goals before the mission starts. The network nodes can then work together to determine the best route that satisfies individual mission requirements.

Given the dynamically changing nature of wireless networks, a link may not always sustain its data rate. When a link is degraded, a network node should dynamically redistribute its bandwidth to data streams if needed at run time. The Real-Time Communications Bandwidth Allocator (CBA) agent was proposed to support bandwidth reservation and dynamic reallocation. The CBA provides a process, architecture, and algorithm to effectively manage network resources.

The CBA agent allocates bandwidth to a mission based on both its resource requirements and the available network resources before the mission starts. During a mission, the CBA can dynamically reallocate bandwidth, if necessary, based on the latest link assessment, individual mission requirements, and priority. Because each CBA agent makes bandwidth allocation decisions autonomously and cooperatively, this architecture is highly scalable.

Related ideas

The goals of the Networked UAS C3 project are to provide the necessary command, control, and communications functions to a group of UASs to maintain a purely autonomous flock in support of ground communications through an ad-hoc network. This system uses an existing ad-hoc network to demonstrate the ability of UASs to make mission-level decisions autonomously based upon network metrics and specified operating conditions. However, this work does not address how to reserve bandwidth and dynamically redistribute bandwidth among missions.

Providing quality of service (QoS) in MANET is not an easy task due to its broadcast and dynamic nature. Bandwidth reservation schemes in MANET have been extensively studied in the literature. For example, Verma, Lanka, and Patel proposed a protocol, the Preemption and Bandwidth Reservation Scheme, which adds more functionality with automatic repeat request protocol and added with ad-hoc on-demand distance vector routing protocol. In addition to reserving bandwidth, it also provides a preemption scheme. It will minimize number of preemption and will assure that preemption is being done fairly.

Jawhar and Wu presented a dynamic range bandwidth reservation protocol for time-division-multiple-access-based MANETs. The intermediate nodes along the path try to reserve a number of slots, b_{cur}, which is equal to the maximum number of slots that are "available" within this range ($b_{min} \leq b_{cur} \leq b_{max}$). The protocol also permits intermediate nodes to dynamically "downgrade" existing paths that are functioning above their minimum requirements to allow the successful reservation for the maximum number of requested paths. When the network traffic load is later decreased, the existing paths are able to be "upgraded" to function with higher bandwidth requirements that are close or equal to the maximum desired level (b_{max}). This allows the network to admit new QoS paths instead of denying such requests by allowing for graceful degradation of other paths.

Similar to the CBA, Sharma and Bhadauria proposed an agent-based bandwidth reservation technique for MANETs. The mobile agent from the source starts forwarding the data packets through the path, which has minimum cost, congestion, and bandwidth. The status of every node is collected, which includes the bottleneck bandwidth field, and the intermediate node computes the available bandwidth on the link. At the destination, after updating the new bottleneck bandwidth field, the data packet is feedback to the source. In the resource reservation technique, if the available bandwidth is greater than bottleneck bandwidth, then bandwidth reservation for the flow is done. Using rate monitoring and adjustment methodologies, rate control is performed for the congested flows. Unlike the CBA approach, this article does not address how to dynamically reallocate bandwidth to missions when a link is unable to sustain its data rate.

Idea description

A communication node in the proposed architecture consists of a

routing device [e.g., a commercial off-the-shelf (COTS) router] and a real-time CBA agent, as shown in Fig. 1. The COTS router performs packet forwarding according to the packet's destination address and enforces the QoS policies at the egress interfaces. The CBA agent manages the bandwidth allocation for the attached COTS router and monitors the quality of the router's links and performs resource redistribution if necessary. Figure 2 shows a typical wireless network consisting of one or more communication nodes.

Mission requirements and resource allocation

One of the most critical components of a UAS system is the data link between the UAS and the ground control station. The data-link system must be able to collect the data and transmit them to the ground station. Hence, the success of the UAS's mission is extremely dependent on the availability of robust and high-performance data links.

In addition, using high-capacity data links on UASs has become an inevitable need because of developments in the optical and electronic sensor field. This requirement is met by using digital-based data links through satellite or in the line of sight (LOS). These digital links provide the capabilities of high data rate, jam resistance, the ability to transfer various traffic types (e.g., image, data, and voice), robust link protocols, and secure communication.

Even with the high capacity and most robust links, the optimal resource allocation to satisfy most mission requirements is still very critical. Two methods were deployed when designing the allocation algorithm, i.e., the mission planning and reinforcement learning (RL). To select an appropriate communications path to meet the mission requirements, the mission planners determine the objectives and their relative priorities for each mission during the premission or planning phase. The planers then analyze the mission objectives and determine the resource constraints (e.g., latency, jitter) and the desired and minimum bandwidths to accomplish these goals before the mission starts.

During the launch and recovery phase, a communication path from the source to the destination is selected, and each node in the path must satisfy the mission resource requirements as determined during the mission-planning phase. To accomplish this goal, two steps are required: 1) select a path that can satisfy the constraints, such as jitter and latency, and 2) reserve at least the minimum bandwidth on each link in the candidate path. Because each mission can have its own unique requirements, given the same source and destination pair, data, video, and voice traffic may choose different routes.

The fundamental issue with the existing standard-based routing protocols is that only a single cost function is used to build the routing table, and then the packet forwarding decision is made according to this table without considering individual traffic needs. Thus, an identical route is used by all traffic to that destination. An innovative route selection scheme that includes traffic characteristics in the route selection process has previously been proposed to address this issue and thus will not be discussed here.

In addition to selecting the optimal routes, the routes need to be adaptable to changes in the plan before and during the mission. The objective of this research is to focus on how to manage the bandwidth usage to satisfy the mission requirements and dynamically adjust the bandwidth allocation during the link failure or bandwidth degradation.

Because of the dynamically changing nature of wireless networks, it is unlikely for all of the communication nodes to receive the latest situational awareness of the entire network before the routing decision is made (and, as a result, it is made under uncertainty). The proposed route selection algorithm leverages the concept of the value-function-based reinforcement learning (RL) method. The agent learns an intermediate data structure known as a *value function*, which maps states (or state-action pairs) to the expected long term reward, i.e., learning by trial and error to perform sequential decision making until the decision is confirmed. The process is a method for modeling sequential problems, model uncertainty, state uncertainty, and cooperative decision making involving multiple interacting CBA agents.

Because a network may support multiple missions with various resource

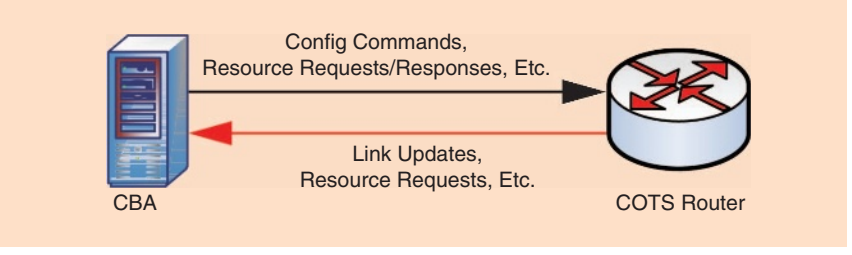


FIG1 A communication node.

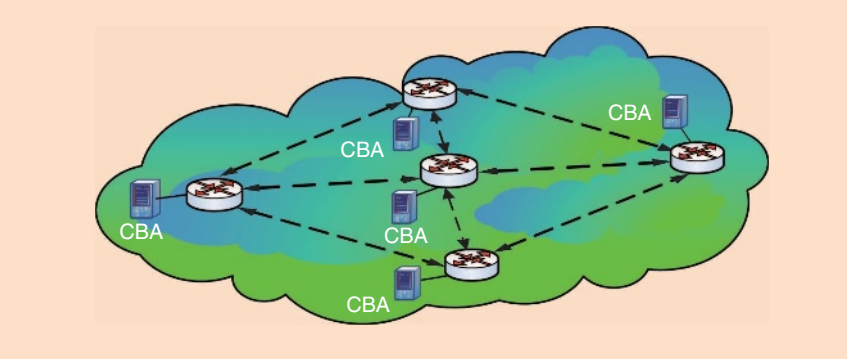


FIG2 A typical wireless communication network.

needs and priorities concurrently and the network situation changes dynamically, the allocation algorithm needs to provide an adaptable method with which to determine the probability of success (*Ps*) for object/technique matches applied to a complex set of time-based events for courses of action taken within multiple, simultaneous mission areas. The detailed algorithm is described next.

Route selection and bandwidth reservation

For a given a mission, a communication node becomes an eligible node if it can reach the destination directly or indirectly and satisfy the constraints and bandwidth requirement of the mission. Define the offered rate to be the bandwidth which all the eligible nodes in the path will reserve for this mission. Initially the offered rate is set to be the desired bandwidth of the mission. This rate can be modified during the route selection process.

When receiving a resource reservation request, the CBA first checks if it is directly connected to the destination. If the answer is yes, then the CBA sends a reservation success notification, including the latest offered rate, to the sender. Otherwise, the CBA stores the resource constraints of this request into its local database and then selects an eligible node from its neighbors to forward this reservation request as shown in Fig. 3. For each neighbor that can reach the destination directly or indirectly, the CBA compares the quality of the link connecting these two nodes and its available bandwidth against the mission minimum bandwidth requirement to determine whether this node is eligible for this mission.

If none of its neighbors is eligible, then the CBA will reject this reservation request and send a reservation failure notification to its predecessor node. Otherwise, among all the eligible neighbors, the CBA selects the one that offers the largest bandwidth to this mission (i.e., the one that is able to achieve high *Ps* potentially). After the next node is determined, the CBA records the corresponding bandwidth, which it temporarily reserves for this mission along with the mission constraints in its local database. If the reserved bandwidth is less than the offered rate currently specified in the request, then the CBA resets the offered rate to the currently reserved rate. This bandwidth reservation will be finalized once a reservation success notification is received. After the reservation request is processed, the CBA forwards this reservation request to the selected neighbor and waits for the notification to come back.

When a reservation success notification is received, the CBA checks the offered rate included in the message. If the current reservation is larger than the offered rate, the CBA adjusts its reserved bandwidth accordingly and forwards this notification to its predecessor until the message arrives at the original requester.

When a reservation failure message is received, the CBA will try the next eligible node from the candidate

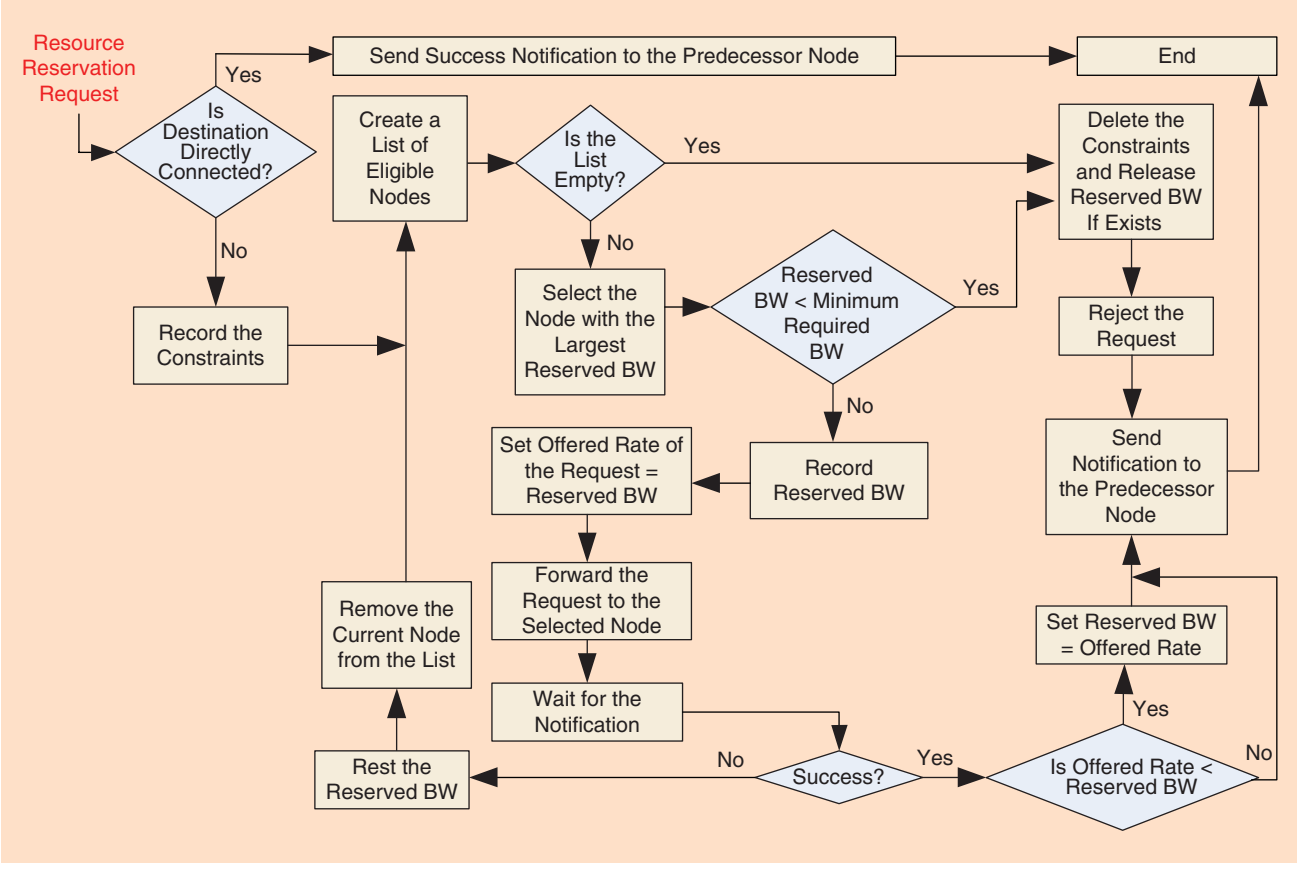


FIG3 The resource reservation process-flow diagram.

list and adjust its bandwidth reservation accordingly. The above steps are repeated until either a reservation success notification is received or none of the eligible nodes is able to complete the bandwidth reservation process successfully. In the latter case, the CBA releases the previously reserved bandwidth, deletes the corresponding resource constraints from its database, and forwards the failure notification to its predecessor. Once the resource reservation process is completed successfully, all nodes in the selected path must guarantee the offered bandwidth to this mission regardless of its priority level.

Bandwidth reallocation occurs only when the link situation changes, such as link failure or quality degradation, and it can no longer meet the mission's requirements. The proposed reservation process both helps the communication networks serve the missions better, i.e., selecting a path to match the mission requirements the best and improving the overall network utilization (e.g., reduce the packet drops due to the insufficient bandwidth down the path).

Although beneficial to the mission, the proposed resource reservation process is not mandatory. Missions are allowed to deliver traffic without performing the resource reservation in advance. A communications node can still forward the traffic received as long as there is unused bandwidth available. However, the priority is given to the missions that have completed the resource reservation. Missions that do not complete the reservation will compete for the remaining unclaimed bandwidth. Without knowing the mission's bandwidth requirement, a communication node will perform the route selection according to the destination Internet Protocol address when the standard-based protocols, such as Open Shortest Path First and Border Gateway Protocol, are used. Or, the decision is made based on the Differentiated Services Code Point (DSCP) marking and the destination address if the previously mentioned innovative scheme is adopted. Thus, the resulting path may not fully satisfy the mission's requirements.

Link monitoring and bandwidth reallocation

In addition to resource reservation, the CBA also monitors bandwidth usage and senses link quality change above a predefined threshold or a link failure. Figure 4 describes the process for this detection and response. Note that the bandwidth reallocation process applies only to the missions that have previously completed the resource reservation. Because the CBA has recorded the resource constraints and bandwidth requirements of a mission, after the re-evaluation, it will determine whether it can continue to support the mission or not based on the current network situation and resource requirements. Missions without the prior resource reservation will share the remaining unclaimed bandwidth after the bandwidth reallocation according to their priority.

As described previously, if the bandwidth degradation is detected and exceeds the predefined threshold, then the CBA will re-evaluate its current resource allocation. To avoid the packet delivery interruption due to the path change and save the path re-establishment time, the preferred choice is to reduce the reserved bandwidth to keep the traffic flow on the same link.

For each mission, the revised bandwidth reservation must still meet the mission's minimum bandwidth requirement. The actual value is depended on the current link capacity, sum of bandwidth requirements from all the missions that have completed the resource reservation, and their priority. Note that if multiple missions are involved in the reevaluation process, then the missions with the strict priority will be served first and then others after that, according to their priority setting. If no sufficient bandwidth is available for a mission, then it will be rerouted. Only the minimum bandwidth requirement specified in the mission profile is allocated to a mission initially. If there is remaining bandwidth available after the allocation, then it will be distributed to all the missions according to their weight and need.

Following the bandwidth reallocation, for each impacted mission, the CBA will update its bandwidth reservation in its local database and send a bandwidth reservation notification, including the revised offered rate, to its predecessor and successor

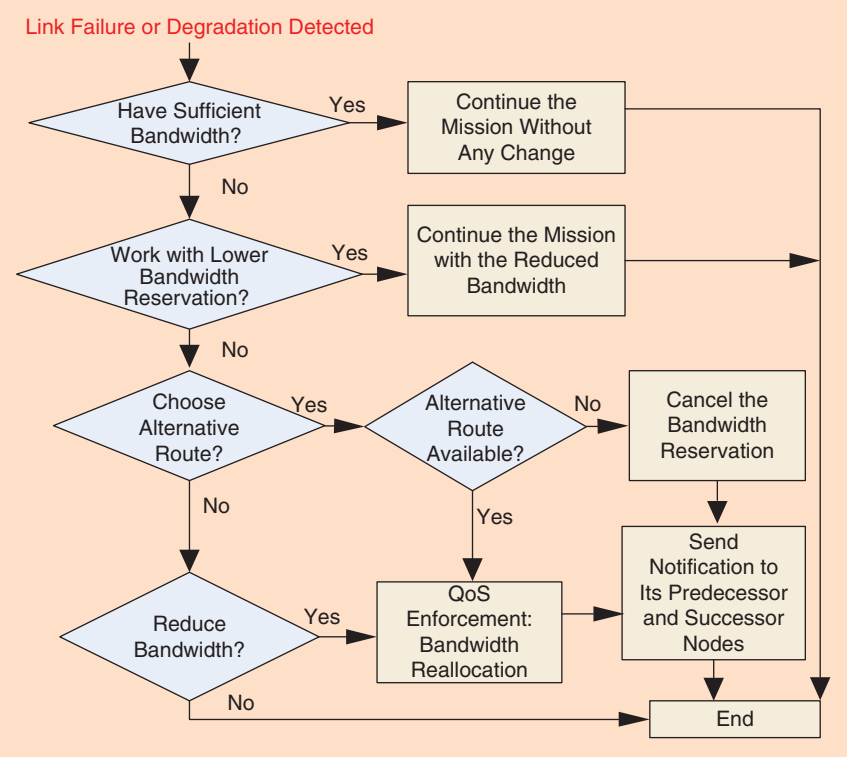


FIG4 Link monitoring and resource reallocation.

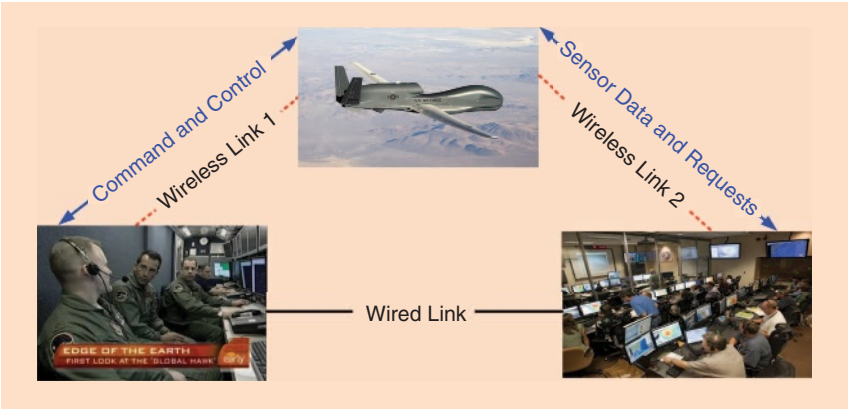


FIG5 A UAS with two communication links.

nodes in the path. After receiving a reservation notification, a communication node will update its local database accordingly and then forward it to its predecessor or successor node in the path depending on from where it receives this notification.

Following the re-evaluation, if the CBA determines that it can no longer satisfy the minimum bandwidth requirement of a mission, then it will cancel its bandwidth reservation (i.e., release the bandwidth reserved for this mission, remove the corresponding mission requirements information from its local database, and send a cancellation notification to its predecessor and successor nodes in the path).

Upon receiving a reservation cancellation from its predecessor node in the path, the CBA will release the bandwidth reserved for this mission, remove the corresponding mission requirements information from its local database, and send a cancellation notification to its successor nodes in the path. After obtaining a reservation cancellation from its successor node in the path, the CBA will

- Search alternative eligible neighbors based on link quality, present mission constraints, and bandwidth requirements.
- Similar to the resource reservation process described in Fig. 3, if at least one eligible neighboring node can satisfy the mission requirements, this mission will continue.
- The CBA selects the link with the highest offered bandwidth and sends a reservation request to the selected neighbor on behalf of the source to establish a new path.
- If none of the eligible neighbors can satisfy the mission requirements, the mission will be canceled. Under such a circumstance, the mission maintenance operators are notified.

When a link failure is detected, the CBA enters contingency operations (i.e., search alternative routes for all traffic traveling through this link and having previously completed the resource reservation). The decision is made based on the available bandwidth from all candidate routes and the mission requirements (e.g., the minimum bandwidth, latency constraint, and security).

For each alternative link being considered, both the data on the failed or saturated link and the data on the alternative link are assessed. The total capability must at least satisfy the sum of the minimum bandwidth requirement of all traffic flows that go through this link. Once the decision is made, the QoS enforcement engine on the impacted link(s) will adjust the bandwidth allocation for each stream according to its priority. When the CBA process ends, the mission maintenance operators are notified.

TABLE 1. Bandwidth reallocation after C2 link failure.

	MISSION PLAN		
TRAFFIC	MAXIMUM BANDWIDTH	MINIMUM BANDWIDTH	STRICT PRIORITY/WEIGHT
C2	2 Mb/s	2 Mb/s	Yes/1
Sensor #1	4 Mb/s	2 Mb/s	No/2
Sensor #2	6 Mb/s	4 Mb/s	No/4

Implementation and demonstration

A test bed was constructed to prove this concept. Only one UAS with two communication links was considered, as shown in Fig. 5.

Drools, an open-source and Java business rules engine, was adopted to implement the decision engine of the CBA. In this test bed:

- A command and control (C2) link is primarily used to deliver the C2 messages.
- A sensor data (SD) link is primarily used to deliver the SD and requests.
- Link capacity is 6 Mb/s for the C2 link and 10 Mb/s for the SD link.
- C2 messages have the highest priority, i.e., strict priority—no bandwidth constraints.
- For the remaining bandwidth, deliver SD according to its weight.

C2 link failure detected

When C2 link failure is detected, the CBA performs the traffic reroute according to the principle described in the previous section. Consequently, the following steps are executed.

- Reroute C2 messages via the SD link after the failure is detected.
- Keep C2 messages as the highest priority traffic, i.e., strict priority—no bandwidth constraints.
- For the remaining bandwidth: deliver SD according to the weight.
  - First allocate the minimum required bandwidth for each traffic stream if possible.
  - Allocate the remaining bandwidth according to the weight, e.g., 2/6 and 4/6.

The resulting bandwidth reallocations before and after C2 link failure are described in Table 1.

TABLE 2. Bandwidth reallocation after SD link failure.

	BEFORE FAILURE		AFTER FAILURE	
TRAFFIC	LINK	ALLOCATED BANDWIDTH	LINK	ALLOCATED BANDWIDTH
C2	C2	2 Mb/s	C2	2 Mb/s
Sensor #1	SD	4 Mb/s	N/A	0
Sensor #2	SD	6 Mb/s	C2	4 Mb/s

SD link failure detected

Similar to the C2 link failure case, the CBA reroutes the SD streams via the C2 link and executes the following steps:

- Keep C2 messages as the highest priority traffic, i.e., strict priority—no bandwidth constraints.
- For the remaining bandwidth: deliver sensor data according to its weight.
  - Allocate the minimum required bandwidth for each message, sensor request, and or SD packet if possible.
  - If needed, drop traffic according to the weight.

The resulting bandwidth reallocation after SD link failure is described in Table 2.

Conclusion and future works

In this work, the CBA was proposed to addresses the network resource allocation problem by providing a process, architecture, and algorithm to distribute bandwidth to tasks. Two methods were employed when designing the allocation algorithm: the mission planning and the value-function-based reinforcement learning. A test bed was constructed to prove this concept and test results demonstrated that the CBA agent can reserve and dynamically reallocate bandwidth to missions based on the mission requirements, priority, and link assessment at the run time.

During these tests, only one CBA node was used for evaluation. In the future, we plan to extend the test bed to include multiple nodes to evaluate the performance of route discovery and resource reservation and traffic reroute when link conditions change in a mobile ad-hoc network.

Read more about it

- J. Elston, E. Frew, and B. Argrow, "Networked UAV command, control and communication," in *Proc. AIAA Guidance Navigation and Control Conf.*, Aug. 2006, pp. 1–9.
- S. S. Verma, S. K. Lanka, and R. B. Patel, "Precedence based preemption and bandwidth reservation scheme in MANET," *Int. J. Comput. Sci. Issues*, vol. 9, no. 6, pp. 407–412, Nov. 2012.
- I. Jawhar and J. Wu, "A dynamic range resource reservation protocol for QoS support in wireless networks," in *Proc. 3rd ACS/IEEE Int. Conf. Computer Systems and Applications*, 2005, pp. 65–72.
- V. K. Sharma and S. S. Bhadauria, "Agent based bandwidth reservation routing technique in mobile Ad Hoc networks," *Int. J. Adv. Comput. Sci. Applicat.*, vol. 2, no. 12, pp. 134–139, 2011.
- M.-C. Wang, S. A. Davidson, and S. Chuang, "A design method to select optimal routes and balance load in wireless communication networks," in *Proc. IEEE Military Communications Conf.*, Nov. 2013, pp. 916–921.
- M.-C. Wang, S. Chuang, S. A. Davidson, and B. Liu, "In the design of tactical communication systems," in *Proc. 12th Annu. IEEE Int. Systems Conf.*, to be published.
- R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.
- M. J. Kochenderfer, *Decision Making under Uncertainty: Theory and Application*, 1st ed. Lexington, MA: MIT Lincoln Laboratory Series, July 2015.
- P. C. Hershey, "Analytics and simulation for decision support good results achieved by teaming the two," *IEEE Syst., Man, Cybern. Mag.*, vol. 4, no. 1, pp. 32–40, Jan. 2018.
- J. Tyrrell. (2014, Sept. 4). Drools: The Business Logic Integration Platform. [Online]. Available: http://www.jboss.org/drools

About the authors

Paul C. Hershey (Paul_C_Hershey@raytheon.com) is with Raytheon Company, Inelligence, Information, and Services, Dulles, Virginia, where he is a Principal Engineering fellow focusing on data analytics, autonomous systems, cloud computing, and cybersecurity. He earned his A.B. degree in mathematics from the College of William and Mary and his M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park. He is an IEEE Senior Member and a Distinguished Lecturer for the IEEE Systems Council in the area of data analytics.

Mu-Cheng Wang (mu-cheng.wang@raytheon.com) has worked for Raytheon Space and Airborne Systems for more than seven years and is currently a senior principal systems engineer. He earned his B.S. degree in electrical engineering from National Cheng-Kung University, Taiwan; his M.S. degree in computer science from the University of Iowa; and his Ph.D. degree from the School of Electrical and Computer Engineering at Purdue University. He has five U.S.-issued and seven pending patents.

Steven A. Davidson (steve_davidson@raytheon.com) is the director of product family development and open system architecture at Raytheon Space and Airborne Systems (SAS). He is an engineering fellow and a Certified Architect with nearly three decades in the industry as technologist, systems engineer, and architect. In his current role, he is leading the transformation of SAS to a product family-centric organization that leverages the advantages of open systems architectures for the benefit of Raytheon customers.

P

Analytics on the cloud

Ankita Christine Victor and Shrisha Rao

Much of today's business and research is driven by data and the analysis of large volumes of it. *Big-data analytics* has become a key phrase in computing. When it comes to business, it plays a crucial role in strategizing and decision making, and it provides insight on performance and usage. Areas of research are driven by the need to create new models to better understand and analyze data.

The International Data Corporation (IDC) predicts that by 2025, the sum of all data created, captured, and replicated—what they call the *global datasphere*—will grow to 163 ZB or 1.63×10^{23} bytes. That is more than a hundred billion of the 1-TB hard disks thare are included in typical computers. The size of the global datasphere is only increasing, and handling such volumes of raw data can cause serious management problems for human engineers. Analytics becomes a necessary counterpart.

Data is the footprint of today's meteoric, digital existence. The growing global datasphere provides large, medium, and small organizations the opportunity to use this data to create meaningful impact within their businesses and for their customers. Once upon a time, computing and analysis were localized processes carried out within the organization. Today, these have moved onto the cloud and are often delivered over the Internet as services. The cloud primarily offers infrastructure and software as services, giving organizations the opportunity to make use of the vast amounts of data generated by users and machines without large investments in local infrastructure and computing resources. Analytics on the cloud provides organizations the opportunity to take advantage of the volumes of data generated by their operations; many organizations would not be able to use advanced analytics if they had to set up their own in-house resources. Analytics are integrated into their business processes using cloud services, which is far more convenient than setting up

Digital Object Identifier 10.1109/MPOT.2018.2824358
Date of publication: 11 July 2018

 0278-6648/18©2018IEEE

a local system for analytics and maintaining the hardware, software, and trained personnel, among others.

Keeping information technology (IT) systems running 24/7, and scaling them up in response to demand spikes, is not a trivial task. From the perspective of small- and medium-sized organizations, cloud services offer a solution using the pay-as-you-go concept, so an organization can draw as much from the cloud as it needs at a given time, without being concerned with either the availability or scaling of the service. Cloud analytics is cost-effective, reduces time to insight, integrates with decision making, can be self-provisioned, and gives organizations a competitive advantage at reasonable cost.

For organizations with seasonal business needs, the cloud offers rapid scaling and readjustment of resources with few hassles. Given that organizations are progressively realizing the advantages of using data to guide their businesses and are turning to cloud analytics to provide it, understanding the buzz around data, analytics and the cloud is of significant consequence.

Big data

Data and data analytics typically (and in this article) refer to big data. "Big" in the context of data is itself a subjective term. How big is big? Gandomi and Haider say that size is the first, and at times, the only dimension that leaps out at the mention of big data, but is only one of several dimensions. They attribute the current hype to promotional initiatives undertaken by IBM and other leading technology companies that invested in building the niche analytics market. As a consequence, *big data* is often used in the context of predictive analysis and structured data, which only form a subset of the entire global datasphere.

The Gartner IT glossary defines *big data* as "high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

Cloud analytics is cost-effective, reduces time to insight, integrates with decision making, can be self-provisioned, and gives organizations a competitive advantage at reasonable cost.

Volume, velocity, and variety—the three Vs of big data—are the dimensions along which data is classified. *Volume* refers to the magnitude of data and can be a very subjective dimension as capacities change with time. What is large in volume today might not be large ten years down the line with the global datasphere expanding as it is; therefore, no hard threshold can be placed on volume.

Variety refers to the structural heterogeneity in a data set—data can be structured, semistructured, and unstructured. Structured data refers to tabulated data; semi-structured data, while not tabulated, contains markers and tags that indicate hierarchy (like XML); and unstructured data is all other text and multimedia content. A high level of variety is a defining characteristic of big data.

Velocity refers to the rate at which data is generated and the speed at which it should be analyzed and acted upon. Big data is generated in large volumes, high frequency, and in real time.

Big data analytics is where advanced analytic techniques operate on big data. The sheer volume, variety, and velocity of generation of big data require advanced analytics techniques that are optimized and designed to operate on big data. Analytics can be practiced on any data, but big data requires efficient techniques that can provide timely results on huge volumes.

Why cloud analytics?

Big data, though potentially valuable, is not useful unless it can be analyzed and leveraged to contribute to tactical decision making in an organization. A survey conducted by *MIT Sloan Management Review* found that organizations that "strongly agreed that the use of business information and analytics differentiates them within their industry" were more likely to be top performers. Such organizations use analytics in a wide spectrum of decision making, both to guide future strategies and for day-to-day operations.

Data is an asset when used to drive operations. Organizations increasingly acknowledge the value that can be derived from collecting and analyzing huge volumes of data related to their business. Top players are those who are able to gain insights from their data and apply it in their business processes. Unfortunately, owing to the high costs of computing resources required to run complex algorithms that return timely results, few organizations are in a position to deploy such resources to leverage their data to gain a competitive advantage.

Organizations can collect data from customer behavior, social media, "the clickstream" of their users, and performance metrics such as production and sales. Collecting, consolidating, and analyzing all of this data is cost-prohibitive for most small- and medium-sized organizations. Cloud analytics offers scalable, cost-effective, and reliable services to organizations that cannot afford to set up their own facilities but want to leverage the value of their data.

Cloud analytics is cost-effective—economies of scale operate on cloud networks with a large base of organizations driving end costs down further. It improves productivity, as the behind-the-scenes jobs of setting up and maintaining the systems performing the analytics are taken care of by the cloud vendor, leaving more time for decision making and actual leveraging of results.

Providers of cloud analytics offer it as a key service and invest in their infrastructure to provide high performance to customers. As a firm, everything is pretty much taken care of all under built-in cost. Backups and failure recovery are built-in components

Data and analytics can be accessed via the cloud, computing power can be added on the fly in a pay-as-you-go system, and all this is done without any sunk costs.

of cloud analytics, which provides continuity to businesses.

A typical cloud analytics model

A cloud service architecture consists of a front-end receiving device, a back-end platform that includes servers and storage, and a delivery mechanism—the Internet (Fig. 1).

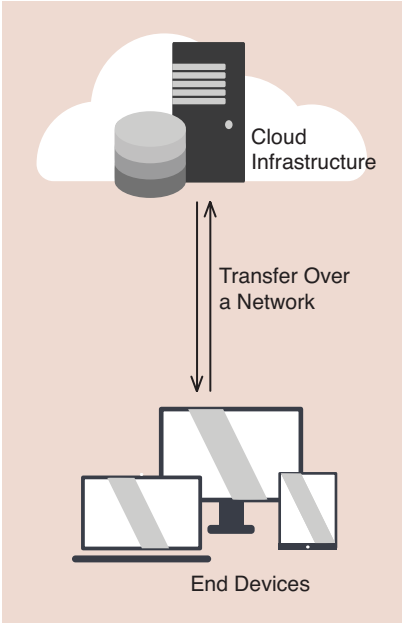


FIG1 Cloud architecture.

The National Institute of Standards and Technology (NIST) defines cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. These resources can be networks, servers, storage, applications, and services. Cloud platforms provide access to IT systems without requiring up-front investments in hardware, software, and maintenance by the user of the service. Cloud analytics is simply analytics provided as a service on the cloud.

NIST says that the five specific qualities that define cloud computing are on-demand self-service, broad network access, resource pooling, rapid elasticity or expansion, and measured service. The three service models are software as a service, platform as a service, and infrastructure as a service. The four deployment models are private cloud, community cloud, public cloud, and hybrid cloud.

According to Gartner, the six key elements of analytics are data sources, data models, processing applications, computing power, analytic models, and sharing or storage of results (Fig. 2). Any service in which one or more of these elements is implemented qualifies as a cloud analytics service. It may be simply data housing or involve complex analytics services.

Advantages and disadvantages

Business intelligence (BI) is an analytics-driven process that enables improved and optimized decisions and performance. It simply leverages analytics software to transform data into intelligence that can be acted upon and used in decision making. BI is a primary use case of cloud analytics. Deploying cloud analytics to enable BI is easier than attempting to develop the capacity in-house.

Cloud analytics offers cost advantages. Data and analytics can be accessed via the cloud, computing power can be added on the fly in a pay-as-you-go system, and all this is done without any sunk costs. That analytics are not handled by a local on-site team but available on the cloud for easy access reduces communications costs between analysis and operations.

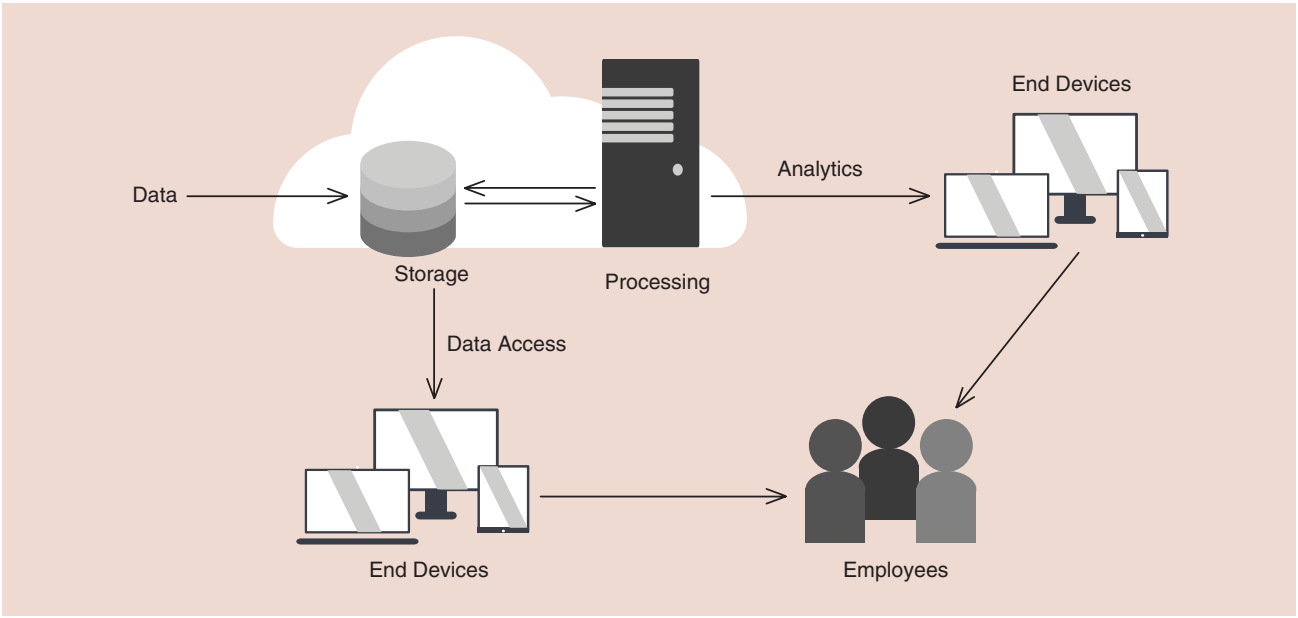


FIG2 Service delivery.

Process separation increases the time needed to convey insights. Insights, unless delivered in a timely fashion, often are not leveraged as business strategies but are filed away or wasted.

Cloud analytics reduces the time to insight, particularly within a hierarchical business structure where it takes time for reports to travel to the top. The cloud can help perform analytics of data from multiple channels and provide an improved overall picture, allowing for better decision making.

A business that relies on information can benefit from an accessible service that provides timely results. Cloud analytics integrates what would otherwise be independent units. The cloud can be accessed anytime, anywhere, and, potentially, on multiple devices. Cloud services can be self-provisioned with a simple sign-up and start, thereby providing quick set up as well as termination of vast computing resources.

Cloud analytics is not entirely rosy. Access to cloud services is fully dependent on a connection to the Internet. A service provided over a network is susceptible to outages and network failure. The underlying hardware can also fail. In both cases access to the service is interrupted.

Another barrier to the adoption of the cloud for data storage and analytics is the security and privacy of data. When organizations place sensitive data on the cloud, they are no longer in complete control, as their data is not stored in the safety of local computers but on storage provided and maintained by a third party. Moreover, since cloud analytics is accessed over the Internet, it is subject to attacks and security breaches.

Vendor lock-in is also a problem in cloud computing. This occurs when organizations become dependent on a single cloud technology and cannot easily switch to a different one without substantial costs. In cloud computing, vendor lock-in takes place due to the lack of well-defined open standards that would enable interoperability. Organizations can be hesitant to trust services without interoperability.

Cloud analytics makes it possible for organizations lacking the necessary wherewithal to integrate analytics with their business processes.

Conclusion

Cloud analytics enables businesses to carry out analytics through an integration of hosted data warehouses, business intelligence, and other analytics. It eliminates upfront investments in infrastructure and other fixed and variable investments into maintenance and personnel. Data is processed on the cloud and insights are presented over a network with few hassles. The disconnect between the decision makers and data is reduced. Organizations can focus on their core product rather than invest in efforts for IT systems and analysis.

Data is generated in a continuous stream and has value when it can be transformed into actionable intelligence. Cloud analytics makes it possible for organizations lacking the necessary wherewithal to integrate analytics with their business processes. Despite the disadvantages, cloud analytics has potential for organizations. Once open standards are in place, it should become an easier choice to consider cloud analytics.

Read more about it

- R. D. Reinsel and J. Gantz. (2017). "Data age 2025: The evolution of data to life-critical don't focus on big data, Focus on the data that's big sponsored by seagate the evolution of data to life-critical." Available: https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf
- SAP. (2012). "Small and mid-size companies look to make big gains with 'big data,' according to recent poll conducted on behalf of sap." [Online]. Available: http://global.sap.com/news-reader/index.epx?PressID=19188
- M. Haider and A. Gandomi, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.
- Gartner. (2018). "Big data." [Online]. Available: https://www.gartner.com/it-glossary/big-data/
- R. Shockley, M. S. Hopkins, N. Kruschwitz, S. LaValle, and E. Lesser, "Big data, analytics and the path from insights to value," *MIT Sloan Manage. Rev.*, vol. 52, no. 2, p. 21, 2011.
- T. Grance and P. Mell. (2011). "The NIST definition of cloud computing." [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-145/final
- NIST. (2018). "NIST cloud computing program–NCCP." [Online]. Available: https://www.nist.gov/programs-projects/nist-cloud-computing-program-nccp
- M. Rouse. (2012). "Cloud analytics." [Online]. Available: http://searchbusinessanalytics.techtarget.com/definition/cloud-analytics

About the authors

Ankita Christine Victor (Ankita.Victor@iiitb.org) is a graduate student at the International Institute of Information Technology, Bangalore. Her main area of expertise is computer graphics, and her research interests include virtual and augmented reality and autonomous systems. She is currently examining areas that combine deep learning and graphics.

Shrisha Rao (shrao@ieee.org) earned his M.S. degree in logic and computation from Carnegie Mellon University and his Ph.D. degree in computer science from the University of Iowa. He is a professor at International Institute of Information Technology, Bangalore. He is a Senior Member of the IEEE, an Association for Computing Machinery distinguished speaker, and a life member of the American Mathematical Society and the Computer Society of India.

Data visualization: The signal and the noise

Paul Cuffe, Harold Kirkham, Chris Dent, and Amy Wilson

Analyzing numerical data is an essential part of modern engineering practice. We explore trends, search for patterns, check relationships, and inspect distributions. Yet it can be rather challenging to consistently produce graphical representations of data that are legible, honest, and attractive. It is all too easy to confuse ourselves and mislead others.

Turning numbers into insights is no automatic process, and it still needs an analyst who will roll up his/her sleeves and become immersed in the data. To find that elusive signal amidst the noise, the human visual system is hard to beat. It just needs clean, clear, and honest visualizations to peer at, and producing those graphics requires deliberate, yet learnable, design decisions. The human visual system offers the highest bandwidth of any of our senses, and it is no coincidence that when we come to understand something, we say *"I see."* We perform data analytics to gain in*sight*.

This article is not a comprehensive graphic design style guide nor a software how-to manual. Rather, we wish to point out the differences between data and information and between opaque numbers and useful insights. Modern software can

Digital Object Identifier 10.1109/MPOT.2018.2824359
Date of publication: 11 July 2018

 0278-6648/18©2018IEEE

make data visualization almost too easy: with just a few clicks your dull numbers can dance to life as a colorful graph. That sort of convenience can suggest that there's no need to stop and think about what you're producing or to criticize the resulting graph. An effective engineer should be equipped with many questions to ask about any graphic they encounter or create. Has the appropriate type of plot been chosen? What might be obscured by the choice of axes? Is there other data that should be included within or alongside the figure?

There's no need to summarize rich data sets with a few stingy numbers: pixels are cheap, so why not show as much of your data as possible?

Why graphs matter

Many data sets have striking or distinguishing features that remain hidden by their summary statistics. Consider the 13 distinct data sets represented as scatterplots in Fig. 1, which are taken from a recent paper by Justin Matejka and George Fitzmaurice. These data sets were craftily constructed to have matching means and standard deviations in both their vertical and horizontal ordinates as well as equal coefficients for their linear regressions. The scatterplots make clear how very different the relationships are with each data set. The plots in Fig. 1 make a strong case for Francis Anscomb's injunction that "Graphs are essential to good statistical analysis." There's

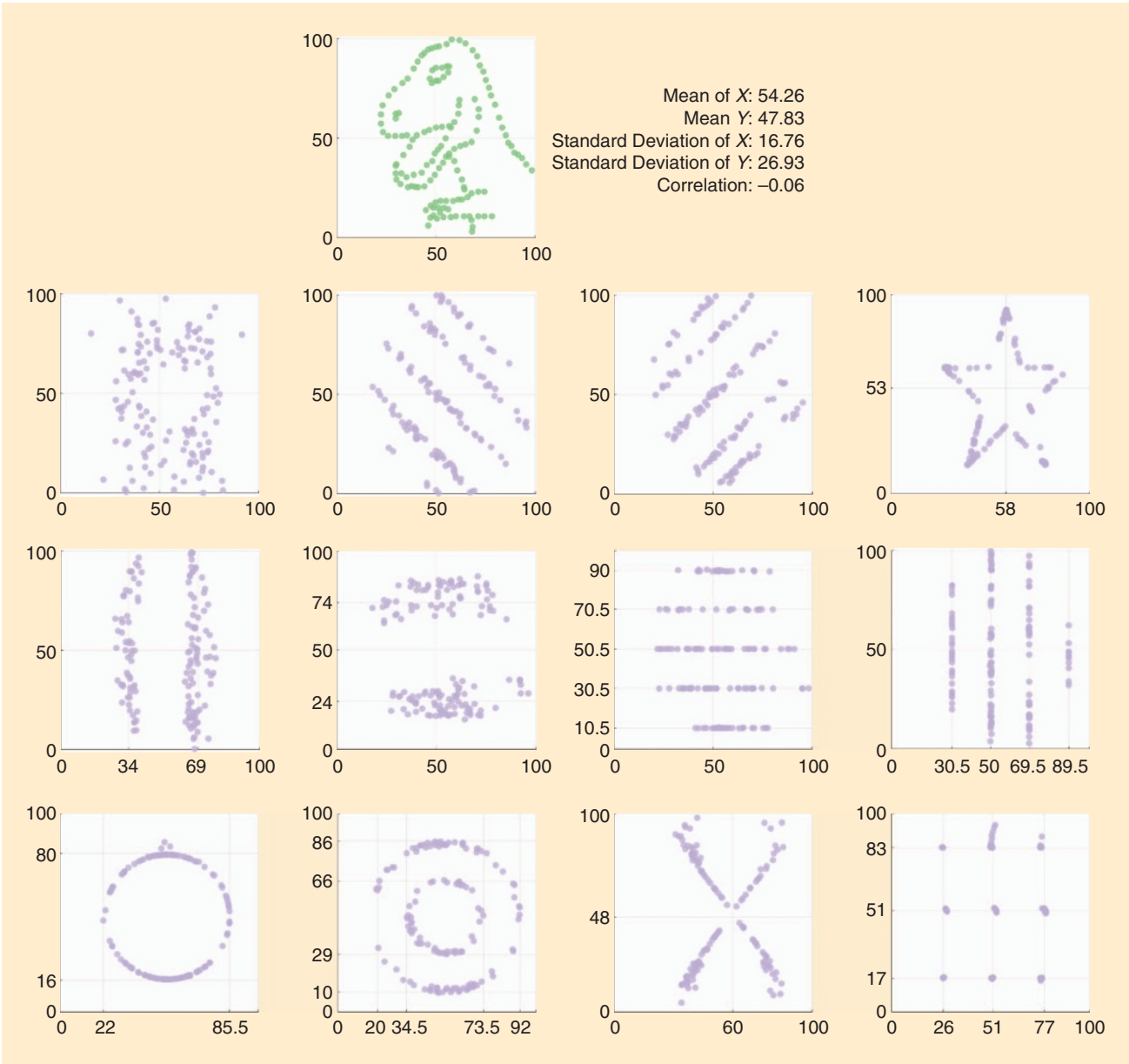


FIG1 Scatterplots showing 13 distinct data sets that share certain summary statistics. The figure was redrawn using the publicly available numerical data from Matejka and Fitzmaurice (2017).

Systematically examine every single graphical element in your visualization and demand that each justifies its existence.

no need to summarize rich data sets with a few stingy numbers: pixels are cheap, so why not show as much of your data as possible?

The eyeball test

The design of the ensemble of scatterplots in Fig. 1 is quite simple: they simply provide a unobstrusive window onto the underlying data. Like a fair referee or a well-made toupée, the mark of good design decisions is passing unnoticed. For instance, in this example, the axis ranges are the same for each data set, allowing them to be compared in a fair and even-handed way. By repeating the axes in a consistent fashion, a display of *small multiples* is achieved. Most software will automatically pick axis ranges based on each individual data set, so an equal treatment requires the designer to deliberately enforce consistency. The tick mark positions in Fig. 1 are also selected carefully to clearly label thresholds that are relevant to each particular data set. The ensemble seeks to emphasize the data itself, while deemphasising other scaffolding as far as possible.

Another design decision imposed on each scatterplot pane in Fig. 1 is their absolute squareness. That is, not only are the axis ranges uniformly square [0 to 100], but each pane is also square in physical dimensions. Enforcing this ensures that the neatly circular feature to the bottom left can be identified by the eye as can the directly proportional 45° angle relationship shown in the middle of the first full row of panes.

Thinking more about graphical details

As the discussion of Fig. 1 has illustrated, the difference between an effective and unclear graph hinges on a sequence of design decisions. Here are some of the more common missteps found in engineering graphics:

- Inappropriate, inconsistent, and poorly divided axes. Inappropriate suppression of axis zero points.
- Fussy or unnecessary symbol markers and overly busy line dashings.
- Graphs that are so reduced in size as to be unreadable when inserted into a single column of a two-column paper. Likewise, insufficient resolution and inappropriately small fonts.
- Clumsy use of color coding, such as perceptually nonlinear mappings and graphs that do not reproduce in grayscale.
- Counterintuitive plot aspect ratios and inconsistent scaling between comparable graphics.
- A tendency to use needlessly elaborate graphics, such as three-dimensional bar charts in which the depth dimension has no meaning, or stacked bar charts rather than a line plot, or colored surface plots rather than heat maps.

The virtue of minimalism

The legendary English rockers Motörhead would ask for a live mix that had "everything louder than everything else." This type of maximalist approach is not recommended when creating effective data visualizations. The graphic should reserve the greatest emphasis only for those elements that encode the underlying data, and it should downplay the prominence of other components as far as is practical.

It is easy to inadvertently create visualizations that contain lots of graphical clutter and scaffolding that distract the eye and deaden the figure's impact. This hides the actual data and makes it hard to extract meaningful information and insights. Be suspicious of every supporting element in your graph. Every pixel or drop of ink given to lines, labels, or legends will pull the reader's attention away from the interesting trends in the data itself.

So, how do you de-emphasize supporting elements to streamline your graph? The motto here is "remove to improve." Systematically examine every single graphical element in your visualization and demand that each justifies its existence. Extraneous components should be summarily deleted, and those that can't be purged outright should be toned down. Every graphical component must either 1) directly encode numerical variation or 2) provide the context that allows the proper interpretation of that data. To reduce the visual prominence of supporting elements, try making them thinner, less colorful, less central, and more translucent. Once you've tamed the clutter, you can amp up the prominence of your data-ink by making it brighter, bigger, and more attention-grabbing.

Consider the two graphs in Fig. 2. Note how much extra ink is used to produce Fig. 2(a). Various dubious design decisions include:

- the ambiguous horizontal axis tick marks on Fig. 2(a)
- unintuitive vertical axis range and divisions, labeled without a percentage symbol
- a needlessly wordy vertical axis label (a % symbol would be more easily understood)
- The heavy gray background of the plot area is visual deadweight
- The grid lines feature very prominently (shorter tick marks would suffice)
- no relationship between trace color and category is denoted
- the legend box is distracting and visually heavy (adding a legend requires that the reader work, and you should never make your reader do extra work)
- the data markers are too large (and they are not really needed at all).

Overall, the Fig. 2(a) display creates an impression that is dour, dense, and off-putting. It takes only moments to thin out certain components and smarten up the main points to create the more inviting display in Fig. 2(b). Note, for instance, the

directly labeled traces that avoid the need for a legend.

A careless choice of plot in a report can give a misleading impression or even lead research completely down the wrong track.

The importance of good axes

In a classic ballroom dance, the ladies wear bright, colorful dresses, while the gentlemen wear elegant, dark tuxedos, to provide a suitable frame to their partner's glamour. The same principle holds for constructing the axes that frame a data visualization: all eyes on the numbers, please!

The axes are where the analyst must provide context, insight and structure to allow the data to be interpreted, but this should be done with deft minimalism, so that attention remains on the data itself. Default software-generated axes are too often a tell-tale sign of a careless visualization. There are many considerations that go into crafting useful axes, and each must be weighed carefully. For instance:

- Axis ranges: Suppose that you're creating a line graph showing the power flows into a power network that contains renewable generators. What is the most meaningful range for the axis? Left to its own devices, your software might find the maximum power export to be 95 MW, and it might use that as the axis maximum. Humans like round numbers, so there is always a better choice when possible: even if your data only varies between +95 MW and −75 MW, your axis should still run from ± 100 MW. If this leaves a lot of white space in the bottom half of your figure, that's a feature not a bug. This negative space lets us infer, at a glance, that power flows are usually positive.
- Axis tick marks: By default, software might carve up your axes into a half-dozen equal increments. So with our chosen axis range, the vertical axis ticks might march along as [−100, −60, −20, +20, +60, +100]. Are these the ideal intervals, though? For sure, zero deserves a place on every axis and certainly warrants a tick mark and corresponding grid line. Are 20 and 60 relevant thresholds? Can they be removed? Perhaps reactive power support only becomes available above +25 MW or curtailment becomes a possibility above +80 MW. All of these should be added and labeled accordingly. Remember: the axis exists to contextualize and frame your data, and only the analyst has the domain expertise to select the interesting thresholds.
- Axis units: What choice of unit will allow your axes to give the most succinct context for the data? The internal format of your working data might be quite inappropriate. For instance, it is common to see time-series plots where the horizontal axis is labeled *Minutes*, and runs up into the thousands. This is unnatural: no one measures the time of day by specifying the number of minutes that have elapsed since midnight. Moving to hours would help but only marginally. If I see a big upramp in the output of a windfarm, my natural curiosity wants to know when this happened, in human terms. So, "shortly after lunch" is meaningful information, whereas "800 min after midnight" is an abstract and opaque factoid. For time-stamped data, the horizontal axis should be labeled in the 24-h clock format, such as "14:00."
- Axis shape: If you are interested in spotting temporal patterns in cyclical data, using circular axes might be the most appropriate, to keep 23:59 and 00:01 appropriately close! While these complex graphs require care to execute well, they are a useful tool for time-series analysis: consider the colorful example in Fig. 3.

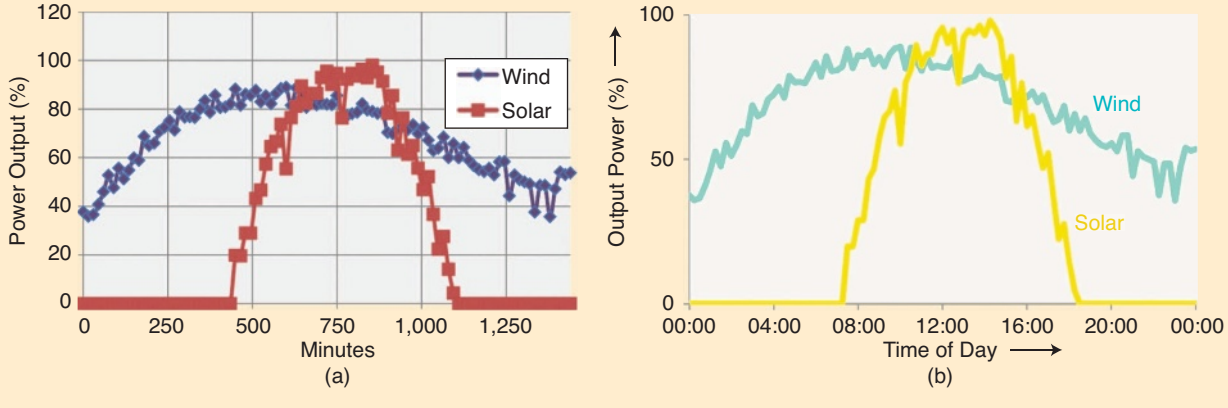


FIG2 Two views of the same synthetic data set, where (a) is bogged down with distracting clutter, whereas (b) is simpler, cleaner, and more attractive for a general audience.

Relationships between variables in a plot should be subject to a healthy dose of suspicion, with additional analysis or further scientific investigation used to confirm any conclusions drawn.

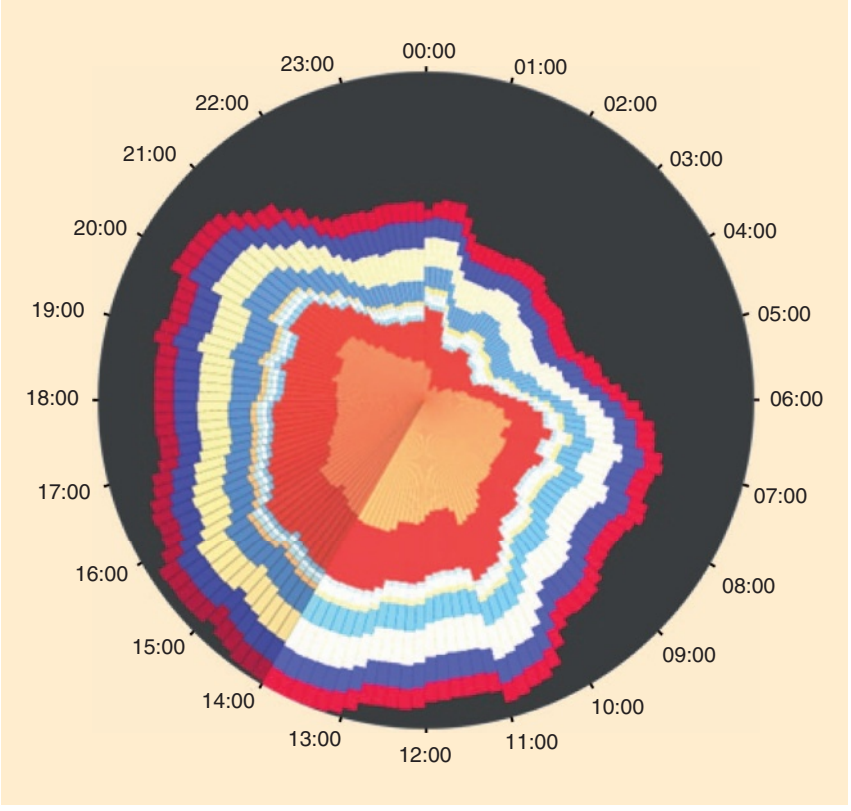


FIG3 Electricity generation over a 24-h period from hydroelectric dams in the northwest United States. The shading indicates the age of the data presented. Different colors are used for different dams. Plot used with permission from (Huang, 2017).

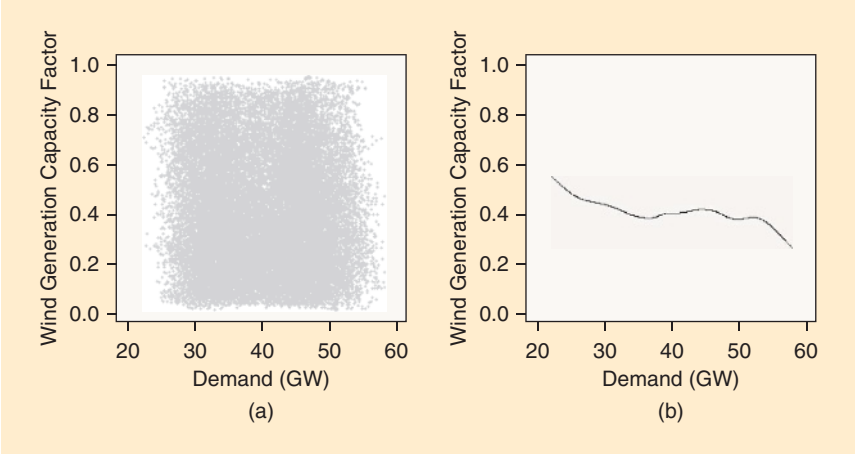


FIG4 (a) A scatter plot of hourly wind capacity factors against the prevailing hourly demand. (b) A smoothed curve of the same hourly wind capacity factors against hourly demand. Both plots use data from nine winter seasons in Great Britain. Data from (Staffel and Pfenninger, 2016; Staffel and Pfenninger, 2017).

Give the right message deliberately, not the wrong one accidentally

As we saw in Fig. 1, a data visualization can draw out details that are obscured by simple summary statistics. Unfortunately, just as a summary statistic can mislead, so can a badly thought-out plot. In his book *How to Lie with Statistics*, Darrell Huff provides many examples of plots that deceive the reader. These include line graphs that exaggerate trends by cutting off axes or altering the unit of measurement and bar charts that use area rather than height to exaggerate differences between categories. Another example is given in (Kirkham and Dumas, 2009), where the case is made that the improper use of a line graph, rather than a bar chart, helped to erroneously establish a spurious link between the measels, mumps, and rubella vaccine and autism. To avoid traps such as these, the author of any data visualization must check carefully that the data are being represented fairly. A careless choice of plot in a report can give a misleading impression or even lead research completely down the wrong track.

Too much and too little information

Consider the plots in Fig. 4. In Fig. 4(a), nine years of hourly wind capacity factors during winters in Great Britain have been plotted against the corresponding demand in that hour. From this plot, there does not seem to be any obvious relationship between instantaneous wind and demand level. In Fig 4(b), a cubic spline has been used to smooth the relationship between wind and demand. Considering this plot alone, there might seem to be a strong relationship between the two variables. The wind capacity factor at high demand seems to be around 0.15 lower than at low demand, which is a potentially worrying situation. Despite both plots in Fig. 4 being based on the same data, the reader could draw very different conclusions from each.

The problem here is that too much data has been included in Fig. 4(a). Any relationship between wind and demand has been obscured; perhaps the use of market transparency, or marginal distributions, could help to articulate the varying density of data here. Conversely, the plot in Fig 4(b) is misleading because it does not contain enough information. Given the huge spread in data seen in the plot in Fig. 4(a), a single smoothed line does not fully represent the relationship between the two variables. At a particular level of demand, the associated wind capacity factor has substantially more uncertainty than that implied by the use of a single deterministic line in Fig. 4(b). As shown in the plot, any level of wind output is possible at any demand.

Relevance of the plot to the message

Even if a plot perfectly represents the data it can still be misleading if the data are not relevant to the question being asked. In (Huff, 1954), accident statistics for different types of transportation are discussed as an example of how data can be irrelevant and misleading. The number of deaths occurring on train tracks in one year is given as a worrying 4,712. However, it turns out that this number includes those hit by trains in their car. Only 132 of the 4,712 deaths were actually train passengers.

The plots in Fig. 4 could be used to discuss the risk of power generation shortfall at times of high demands (i.e., will the wind power be there when we need it?). The problem is that the majority of the data points in these plots are irrelevant to the calculation of reliability metrics: power shortfalls only occur in Great Britain at times of very high demand and very low wind levels. Including corresponding thresholds on the axes could help to guide the reader here. Limited data in the extremes of a range can make it difficult to extrapolate relationships seen in the bulk of the data. When producing plots like these, it is important to ensure that any data presented are relevant to the broader research questions.

Limited data in the extremes of a range can make it difficult to extrapolate relationships seen in the bulk of the data.

Spurious correlations

Spurious correlation occurs when the correlation between two variables can be explained by a third variable. Problems arise if this correlation is mistaken for a causal relationship when no such relationship exists. Relationships between variables in a plot should be subject to a healthy dose of suspicion, with additional analysis or further scientific investigation used to confirm any conclusions drawn.

In Fig. 5(a), nine years of hourly solar capacity factors in winter in Great Britain have been plotted against the corresponding demand in that hour. The plot clearly shows that the solar capacity factor seems to be minimal at both high and low demand levels, but it can be high for medium demand levels. This plot alone does not, however, give us the full story. In Fig. 5(b), the average solar capacity factor in each hour over the winter season (and the associated 5% and 95% quantiles) is shown. In Fig. 5(c), the average demand in each hour over the winter season (with 5% and 95% quantiles) is shown. By looking at Fig. 5(c), it is clear why the relationship between

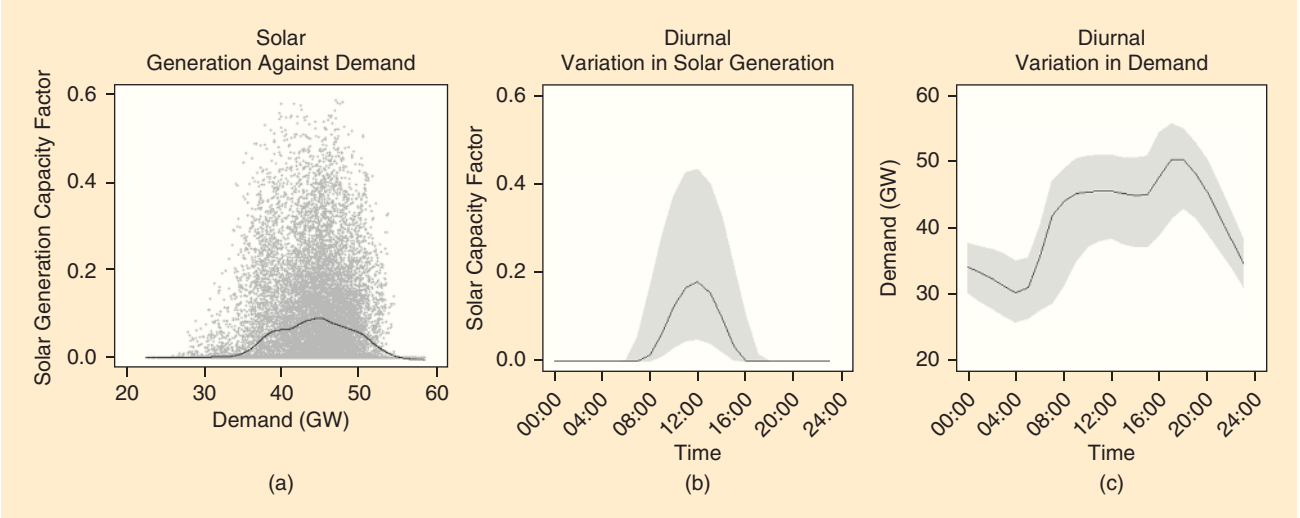


FIG5 (a) A scatter plot of hourly Great Britain solar capacity factors against hourly demand. Data are from nine winter seasons. The trend has been removed from the demand data so that different years are comparable. A smoothed curve is drawn using cubic soothing spline. (b) The mean diurnal variation in Great Britain solar capacity factor over the winter season (solid line) with 5% and 95% quantiles (shaded region). (c) The mean diurnal variation in Great Britain demand over the winter season (solid line) with 5% and 95% quantiles (shaded region). Data are from (Staffel and Pfenninger, 2016; Staffel and Pfenninger, 2017).

The insights and impressions we infer from a visualization crucially depend on the way it is designed and presented, and these skills should be embraced by all ambitious engineers.

demand and solar in Fig. 5(a) looks as it does. At low demands (overnight) and at high demands (early evening), there is no solar generation. At medium-sized demands (during the day), the solar generation is highest. Using Fig. 5(a) alone, and without explanation, could be misleading—it is not that demand is dependent on solar but rather both solar and demand are dependent on time of day.

Conclusion

Though we speak and hear thousands of words of every day, it can still be challenging to write convincing dialog. Likewise, even though we may have plenty of experience working with data, presenting it in a tangible way without distortions and distractions can be tricky. The creation of any type of data visualization involves a sequence of deliberate design decisions. Chefs know that the way food is plated truly affects the way it tastes: we eat first with the eyes. Likewise, the insights and impressions we infer from a visualization crucially depend on the way it is designed and presented, and these skills should be embraced by all ambitious engineers.

Read more about it

- J. Matejka and G. Fitzmaurice, "Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing," in *Proc. CHI Conf. Human Factors Computing Systems*, 2017, pp. 1290–1294.
- F. J. Anscombe, "Graphs in statistical analysis," *Amer. Statist.*, vol. 27, no. 1, pp. 17–21, 1973.
- H. Huang, "Big data access, analytics and sense-making," presented at the Power and Energy Society General Meeting, July 2017.
- D. Huff, *How to Lie with Statistics*. New York: Norton, 1954.
- H. Kirkham and R. Dumas, *The Right Graph: A Manual for Technical and Scientific Authors*. Hoboken, NJ: Wiley, 2009.
- T. Vigen. (2018). Spurious correlations. [Online]. Available: www.tylervigen.com/spurious-correlations
- I. Staffell and S. Pfenninger. (2018). Renewables ninja datasets. [Online]. Available: www.renewables.ninja/
- I. Staffell and S. Pfenninger, "Using bias-corrected reanalysis to simulate current and future wind power output," *Energy*, vol. 114, pp. 1224–1239, Nov. 2016.
- S. Pfenninger and I. Staffell, "Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data," *Energy*, vol. 114, pp. 1251–1265, Nov. 2016.
- S. Few, *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Oakland, CA: Analytics Press, 2012.
- H. Kirkham and R. Dumas, *The Right Graph*. Hoboken, NJ: Wiley, 2009.

About the authors

Paul Cuffe (paul.cuffe@ucd.ie) is an assistant professor in electrical power systems within the University College Dublin (UCD) School of Electrical and Electronic Engineering. He is a member of the UCD Energy Institute. His research interests include the integration of renewable generators, structural analysis of power systems, visualization of technical data, and blockchain technologies.

Harold Kirkham (harold.kirkham@pnnl.gov) earned his Ph.D. degree in electrical engineering from Drexel University, Philadelphia, in 1973. He worked for many years at NASA's Jet Propulsion Laboratory, Pasadena, California, where he managed a U.S. Department of Energy project that was a forerunner of smart-grid work today. He has been with the Pacific Northwest National Laboratory Richland, Washington, since 2009.

Chris Dent (Chris.Dent@ed.ac.uk) is a reader (associate professor) in industrial mathematics at the University of Edinburgh, United Kingdom, and Turing Fellow at the Alan Turing Institute, United Kingdom. He earned his M.A. degree in mathematics from the University of Cambridge, his Ph.D. degree in physics from Loughborough University, and his M.Sc. degree in operational research from the University of Edinburgh. He is a Chartered Engineer, a fellow of the OR Society, and a Senior Member of the IEEE.

Amy Wilson (Amy.L.Wilson@ed.ac.uk) is a research associate in statistics at the University of Edinburgh. She holds an M.Math. degree from the University of Cambridge and a Ph.D. degree in statistics from the University of Edinburgh. She currently works on a project building time-series models of demand and variable generation for security of supply calculations.

P

Machine learning for data-driven control of robots

Sidney Givigi and Peter Travis Jardine

Machine learning has become a popular area of research and development in many domains over the past decade. Researchers in a number of disciplines, ranging from biology to economics, have successfully applied machine learning to solve a diverse range of problems including the analysis of protein cell interactions to the prediction of financial market performance. In engineering, machine learning has been used for classification, signal processing, system identification, and control systems. This article provides a discussion of how machine-learning techniques can use data acquired through simulations and experiments to derive more effective sensory abilities, controllers, and decision-making strategies for robotic autonomous systems.

Autonomous robotic systems may be seen as a conjunction of different subsystems as shown in Fig. 1: sensing, control, and decision making. Traditionally, these subsystems have been modeled from first principles, and solutions were implemented based on these models. However, machine learning can be used to assist in sensing (e.g., deep convolutional networks used for image

Digital Object Identifier 10.1109/MPOT.2018.2824398
Date of publication: 11 July 2018

processing), control (e.g., iterative learning used in the adaptation of controllers for autonomous vehicles), and decision making (e.g., reinforcement learning used in the derivation of decision-making policies).

There are several reasons for using a model-based approach and among them is risk reduction. Designers need to guarantee some safety boundaries for robots (or any other physical dynamic system). If we know the model of a system (the plant in Fig. 1), we can design sensing, control, and decision-making subsystems to perform the task or mission we want with reasonable guarantees of safety.

When models cannot be found, or desired system behaviors are difficult to specify, one may decide to use alternative methods based on data collection, as shown in Fig. 2. However, when this approach is implemented with machine learning, accidents may happen when the data is noisy or if the system needs to explore new solutions. Therefore, the acceptable risk level needs to be taken into account when designing a data-driven system. An example involving human behavior may help explain this idea.

When humans are born, we do not know how to walk. As a parent, one wants to assist kids in learning to perform this task. Since we evaluate the risk associated with falling down as being low, we allow children to make several mistakes, knowing that they will not be seriously injured. A few years later, when a child learns to ride a bicycle, parents become a little bit more conservative, as the risks are also higher. Therefore, parents buy training wheels and helmets to safeguard their child. A few more years and the child, now a teenager, decides to learn how to drive a car. As vehicle collisions are potentially much more harmful, parents usually pay specialized teachers to train their kids and reduce the risk of harm.

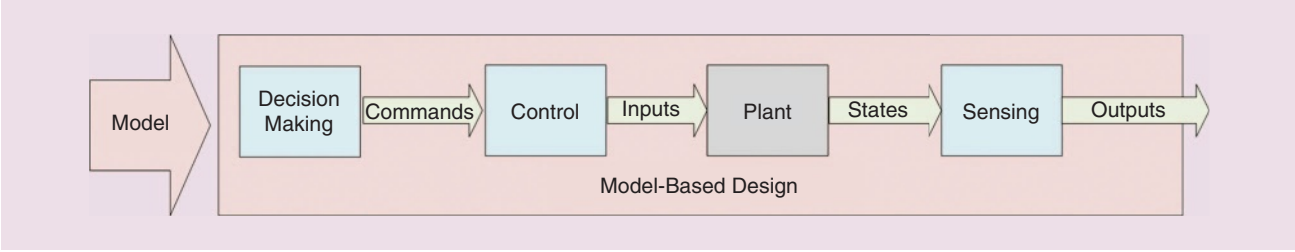


FIG1 A robotic systems block diagram.

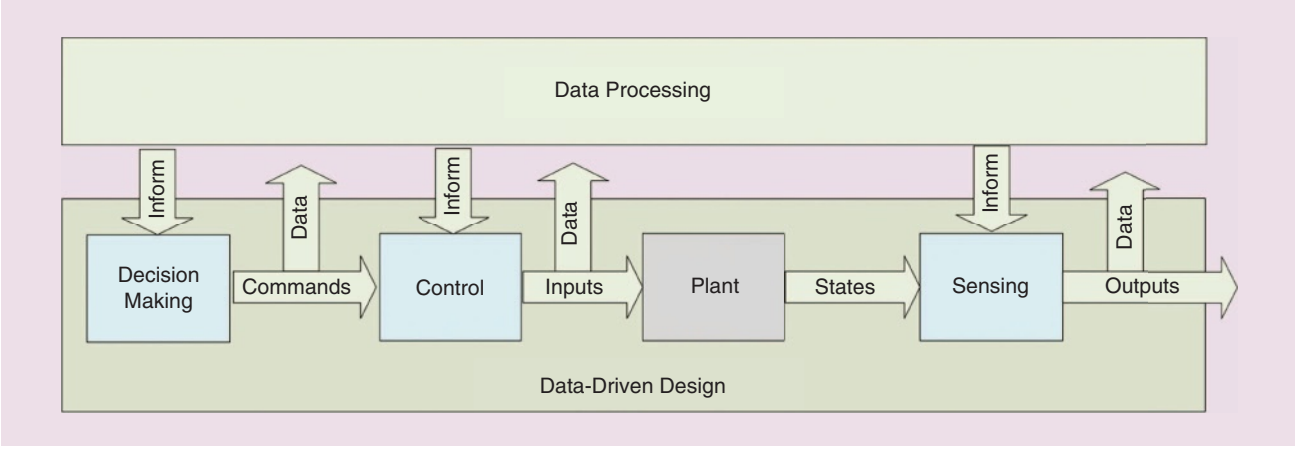


FIG2 Data-driven design.

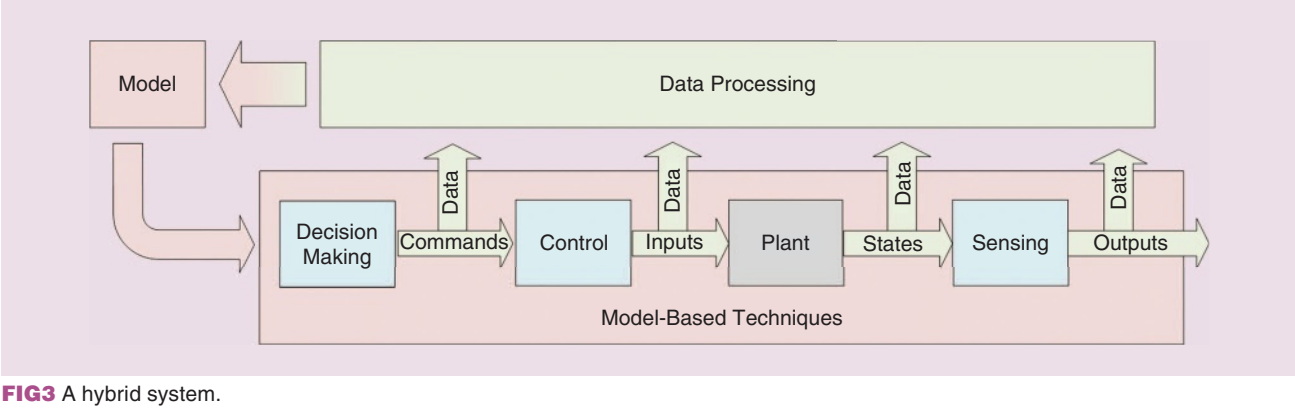


FIG3 A hybrid system.

Interestingly, risk intolerance grows proportionally to how easily we can represent the systems. A car is a very stable system, and its model can be easily derived by an undergraduate student, while models for walking are very difficult to be derived even by senior graduate students. Also, as modeling difficulty increases, so does the difficulty in the mathematical specification of what the desired behavior is. In these cases, data-driven approaches (and machine learning more specifically) are desirable.

One way to reduce the risk is by coupling model-based and data-driven approaches as shown in Fig. 3. By constraining the behavior of robots with models, even if they are uncertain, we can guarantee a baseline for safety of the system. This is the main topic of this article.

In the next sections, we will outline how machine learning and data science can be used to help in designing efficient mobile ground and aerial robots including sensing, decision making, and control.

Sensors

Mobile robots use several different types of sensors for navigation. The most common of them, as shown in Fig. 4, are an inertial measurement unit (IMU), the Global Positioning System (GPS), cameras, and lidars. All of these sensors generate a large quantity of data that is used to find the state of a robot, such as its position and velocity. However, the way that data is merged (or fused) varies depending on the source. Whereas it is relatively easy to fuse data from an IMU and a GPS, it is more challenging to fuse a camera or lidar with inertial sensors.

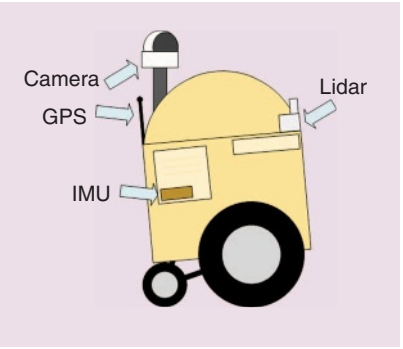


FIG4 A robot with common sensors.

Traditionally, when a camera or lidar is used, data is preprocessed to reduce the number of points (or features) that need to be analyzed. For example, there are several machine-vision algorithms that extract features from different images and perform comparison and integration so common features may be found and matched.

Alternatively, other methods, based on data analytics and machine learning, can be used for extraction of features and/or data fusion. One such method is convolutional neural networks (CNNs) with deep structures, popularly known as *deep-learning artificial neural networks* (*DLANNs*). These networks have multiple layers of different types (convolution, pooling, etc.) in sequence, hence the use of the term *deep* (Fig. 5). However, for these methods to be useful, a large number of data points need to be available, and the algorithms cannot usually be run in real time.

Some improvements may be added to the data-driven machine-learning approach by adding the model robot to the sensing system. In this case, the output of the DLANN could be constrained by the model of the robot such that the states could be better estimated. In this way, spurious results may be filtered out of the measurements and sensing improved.

Control

In the context of this article, a controller may be defined as a device that is used to drive a robot to a desired state, usually a position and/or velocity reference. If we know the model of the robot, we can derive controllers to stabilize the system as well as present a desired response (move faster or slower) with very small (or zero) error. When we do not know the model of the robot or when the environment interferes with its behavior, we can use a system identification approach to find the parameters that define the motion of the vehicle.

Several system identification approaches are commonly used for this task: from simple (least squares approximations) to those that are complex (identification of Volterra series models). Machine-learning approaches can also be used, the most of which are are fuzzy systems and, again, neural networks.

Neural network approaches are usually referred to as *black-box*

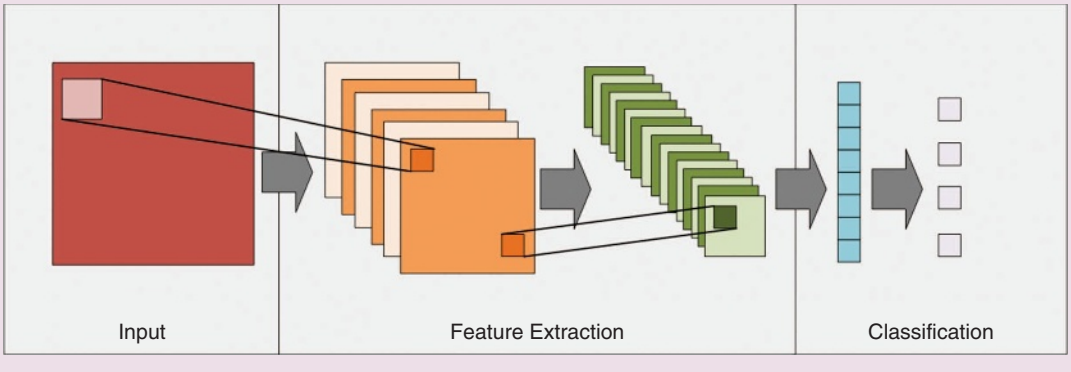


FIG5 A DLANN.

approaches, for the parameters do not have any physical meaning as parameters of a system. With CNNs, however, one can introduce meaning to the parameters, as, in determining the model of a robot, one usually ends up with a convolution. Recent approaches use CNNs to find finite impulse response or infinite impulse response linear models for robots and unmanned aerial vehicles (UAVs). The same approach can be used to find more sophisticated models such as Volterra series and generalized orthonormal basis functions. The disadvantage of using neural networks is that training is usually performed offline.

Fuzzy systems, on the other hand, can be used in real time. Fuzzy systems are used to model systems using what is called *linguistic variables*, shown in Fig. 6. In this way, the system can be approached by using normal language such as *near* or *far*. By allowing the definition of the terms to change based on feedback from the environment, controllers may change over time. For example, one can use reinforcement learning to change the membership functions related to the states of a robot. In this way, new control can be achieved from identification of the plant.

Beyond uncertainty of the model, it is not always clear what the control parameters for a robot or UAV would be. For example, since the model of a robot can change depending on the surface in which it navigates, what should be the gains of a proportional–integral–derivative controller for smooth navigation? In the same manner, when using newer and more advanced approaches such as model predictive control (MPC), what should be the parameters used in the cost function?

MPC is an advanced technique that treats control as an optimization problem. Rather than simply choosing one control action for the next time step, MPC produces a sequence of control inputs that minimizes some cost function over a prediction horizon. The cost function is determined by the designer and reflects the relative importance of certain parameters.

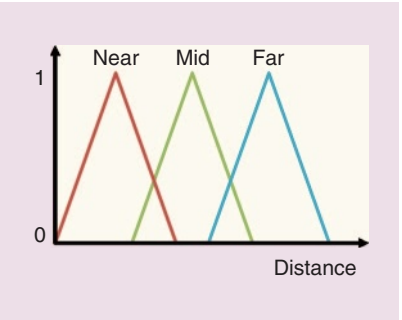


FIG6 Fuzzy linguistic variables for distance.

For example, let us say we wish to reduce the distance between the robot and a target (Δx) while also minimizing the input changes (Δu). We want to N plan time steps into the future (our *horizon*). The inputs changes are related to the control effort over this horizon (i.e., how much fuel or battery power the robot must consume). A cost function involving the distance to the target (Δx) and the control effort (Δu) is then designed. However, since the importance of these two parameters is not necessarily the same, weights are used to identify which cost is more important. If we decide that the control effort is twice more important for the robot than tracking the target, we could select a weight 1 to be multiplied by the term that includes (Δx) and a weight of 2 to be multiplied by the term that includes (Δu).

Typically, we only execute the first input in the optimized sequence. We rerun the optimization repeatedly (always executing only the first input) as the robot moves. This allows the sequence to adapt to any changes in the environment. The trade-off is that larger prediction horizons are more difficult to compute.

The benefit of MPC is that we produce an optimal sequence of inputs that take into account our predictions of how we think the robot will behave in the future. To produce these predictions, we must rely on mathematical models. The better the models, the more accurate our predictions will be. Perhaps most importantly, the optimization can be constrained to take into account things like input bounds and obstacles. MPC is unique in that it is the only control technique that provides optimal solutions that take into account these constraints.

Going back to the example of tracking a target while minimizing the control effort, machine learning can be used to find the values of the parameters of the cost function 1 and 2 in the example. This is especially important when the number of states and inputs are large. One of the machine-learning techniques that can be used is reinforcement learning with the cost being used as the reward of the controller. This technique has been successfully used for ground and aerial robots (see the "Read More About It" section). When coupled with MPC, this technique uses real data to choose the parameters while keeping the system stable by constraining the range of allowed values to only those quantities that guarantee stability of the system.

Decision making

A controller, as discussed in the previous section, is a decision-making device. As depicted in Fig. 7, a controller takes a reference command and, by measuring the outputs of the robot, provides commands that guarantee that the robot execute the desired task.

However, in this article, we use the subsystem "decision making" to mean the high-level choice of strategies to be made by a robot. The main distinction is that the choice of strategies is not easily specified, since the desired outcome of the system is not clear, whereas the performance of the controller is well stipulated in terms of transient and steady-state parameters. In a very broad sense, one may say that the functions of decision making and control in robotic systems are analogous to the human nervous system. The decision-making system's function would be similar to the function of the central nervous system whereas the controllers' functions would be similar to the peripheral nervous system. The first is responsible for the conscious decisions while the second is responsible for the automatic response. Both are extremely important and in

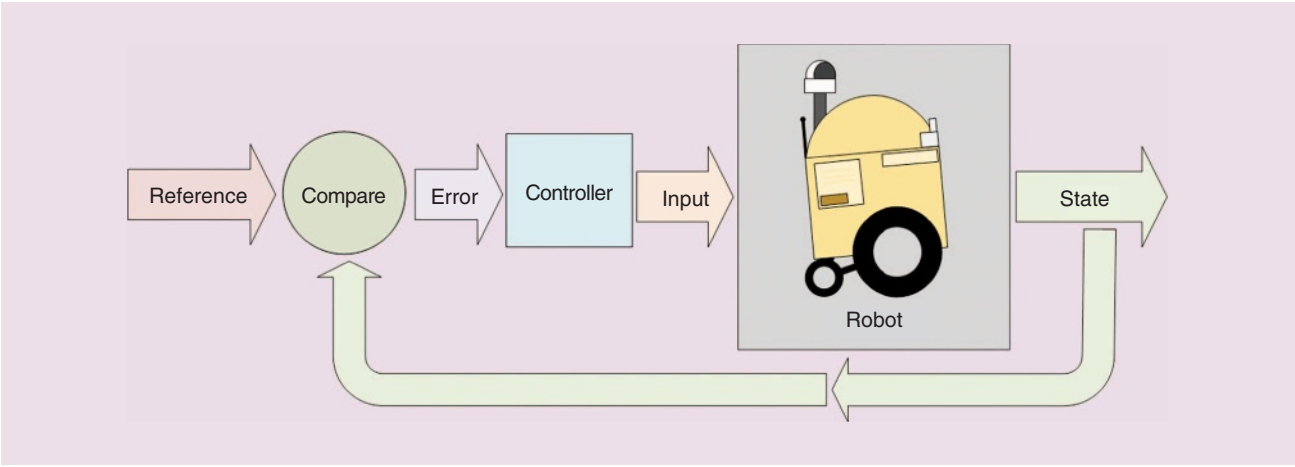


FIG7 A block diagram for the control of a robot.

conjunction provide a smooth behavior for the whole system.

For designers or robotic systems, the question then becomes how these strategies may be defined. Traditionally, one of the most popular ways to derive strategies for robots is through the use of Markov decision processes (MDPs) in several different flavors (partially observable MDPs, mixed observability MDPs, etc.). But to apply this type of method, a model of the environment (including, especially, its probability distribution) needs to be available. When the environment's probability distribution or the cost function that determines the behavior of the robot are not readily accessible, it is complicated to use MDP-based solutions. In this case, another method that can automatically and implicitly extract the probability distribution may be successfully used. For example, reinforcement learning has been used to approximate the probability distribution of unknown environments. Also, automated derivation of rewards and cost functions have been proposed in the context of multiple agent environments. Both solutions depend on the availability of data, but they do not necessarily rely upon preacquired data. Therefore, these methods may be executed in real time.

It must be noted that sometimes the design of machine-learning solutions is not straightforward. One common criticism of these methods is that they only displace the complexity of the design from one subsystem (the model of the environment) to another (the machine-learning algorithm). A rule of thumb to measure the effectiveness of any approach may be the number of parameters necessary to define the solutions. We always aim to have a smaller number of parameters to tune. It can be shown that in several applications, machine learning (or other big-data approaches) is successful in satisfying this criterion.

Conclusion

An autonomous robotic system consists of several subsystems that, together, determine the behavior of the robot or autonomous vehicle. Sensors, controllers, actuators, and other components, such as communication networks, are all important devices in a robot. Each provides a large amount of data that can be used by machine-learning algorithms to improve the behavior (or behaviors) of the system. Of special interest is the derivation of strategies for high-level decision making. For this reason, in the near future, we expect to see an even greater influence of machine-learning and big-data approaches in robotic systems.

Acknowledgment

We acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada.

Read more about it

- K. Pereida, M. K. Helwa, and A. P. Schoellig, "Data-efficient multirobot, multitask transfer learning for trajectory tracking," *IEEE Robot. Automat. Lett.*, vol. 3, no. 2, pp. 1260–1267, Apr. 2018.
- S. M. Hung and S. N. Givigi, "A Q-learning approach to flocking with UAVs in a stochastic environment," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 186–197, Jan. 2017.
- P. T. Jardine, M. Kogan, S. Givigi, and S. Yousefi, "Adaptive predictive control of a differential drive robot tuned with reinforcement learning," *Int. J. Adapt. Control Signal Process.*, to be published.
- H. Shi, Z. Lin, K. S. Hwang, S. Yang, and J. Chen, "An adaptive strategy se lection method with reinforcement learning for robotic soccer games," *IEEE Access*, vol. 6, pp. 8376–8386, Feb. 2018.

About the authors

Sidney Givigi (sidney.givigi@rmc.ca) is an associate professor at the Royal Military College of Canada in Kingston, Ontario. His research interests are focused on the use of machine learning in robots and autonomous vehicles.

Peter Travis Jardine (Peter.Jardine@rmc.ca) is a Ph.D. degree candidate at Queen's University in Kingston, Ontario, Canada. His research focuses on the application of machine learning to improve the performance of autonomous vehicles.

Exploiting advances in video analytics to support military operations and related applications

Susan Gottschlich, Brian L. Stone, and Bill Gerecke

As the prime contractor for the U.S. Army's Persistent Surveillance and Dissemination System of Systems (PSDS2) program, we have witnessed firsthand the U.S. military's video glut problem. Established in 2005 to address an urgent requirement for a video processing, exploitation, and dissemination (PED) capability, the PSDS2 program has provided the core video dissemination backbone used by the U.S. Central Command (CENTCOM) in Iraq and by the International Security Assistance Force (ISAF) in Afghanistan, where it is still in active use. PSDS2 integrates real-time sensor data from multiple sensor types—including hundreds of video sensors—and makes it available on-demand simultaneously to thousands of users both in near real time and streamed from forensic archives. Given the large number of stationary and mobile video and other sensors that are being integrated, the criticality of the services provided, and the austerity—in terms of available communications bandwidth, equipment size, weight, power, cost and cooling (SWaP+C2) limitations, and military personnel availability—we are motivated to investigate the use of analytics to address the video glut problem.

MAN—©ISTOCKPHOTO.COM/ALEXS84
EYE—©ISTOCKPHOTO.COM/PETERHOWELL

Just as there are many different kinds of sensors used by the military, there are also many different uses of them. Some examples include:

- Aerial vehicle surveillance: Whether manned or unmanned, aerial system surveillance of the ground is almost always prescripted via a mission plan. Commanders and analysts operating in the conflict zone may actively monitor

Digital Object Identifier 10.1109/MPOT.2018.2824858
Date of publication: 11 July 2018

 0278-6648/18©2018IEEE

video transmitted from an aerial vehicle to help coordinate the mission. Unmanned aerial system (UAS) ground teams, typically operating in remote locations, may monitor video to both control the vehicle and its sensors as noted by (Gregory, 2011). However, there is always the possibility that the UAS sensors register activity or footage that is unrelated to the mission and initially unnoticed but is of use to other ongoing or future missions.

- Base protection monitoring: Similar to any other government or commercial facility, video cameras are often used to help protect military bases. Generally speaking, no day-to-day preplanning is conducted for such monitoring. The extent of preplanning is usually limited to the selection of stationary locations for camera placement. Camera placement decisions are as likely to consider easy access to power and or network connections as it is camera viewpoints.
- Cameras of convenience: Governments, businesses, and other persons or organizations collect and disseminate video from many cameras in some of the most-densely populated areas today. These include traffic, surveillance, cell phone, and news organization cameras. Particularly when supporting security for a major event or forensic analysis after an attack, live and/or archived video from these cameras of convenience can be made available to military forces and homeland security agencies.

At a high level, video surveillance analytics help address two fundamental military issues:

1) Searchability. The goal of searchability is to perform simple keyword, temporal, geospatial, and more complicated contextual searches on all available video streams to find and review relevant video in support of an underlying military goal. Furthermore, it is desirable to perform such searches forensically (on archived video streams), in real time on live video streams, and even on future streams to the extent that predictive analytics can be brought to bear on the military operation.
2) Discovery. The goal of discovery is to avoid surprises. Searchability is a powerful tool when a commander or analyst knows what to look for but may not help address unforeseen events in a timely fashion. Analytics that discover and/or predict unanticipated events are of great interest.

Overview of recent research

Solving the searchability problem is currently an active area of research. It generally involves developing metadata or "tags" that are associated temporally and spatially with a video stream. While there are no widely accepted video metadata standards to support searchability in current use, various standards bodies have developed related standards.

The National Geospatial Agency (NGA) formed the Motion Imagery Standards Board (MISB) to help develop standards for video sources for use by the U.S. Department of Defense, intelligence community, and National System for Geospatial Intelligence. The North Atlantic Treaty Organization (NATO) has developed many standard agreements (STANAG) that define metadata standards that are related to video analytics. These include STANAG 4607, a metadata standard for a ground moving-target indicator and a digital motion imagery standard that references the MISB standard. The International Telecommunication Union has led video coding for generic audio-visual service recommendation efforts including H.264 and H.265.

Nevertheless, the standardization of metadata schemas is likely being hampered by the evolving nature of the technologies that can create metadata and use cases for its utilization. In our work, we have found that the following attributes should be included in any evolving metadata standard to appropriately support both searchability and discovery:

- keyword(s) or phrase(s)
- timestamp(s)—may be a single point in time or a time span
- physical location(s)—may be a single point, track, or region
- frame coordinate(s)—a point or region in the frame (for wide fields of view, it may be necessary to specify the target's location in the frame)
- confidence—supports tag filtering
- provenance—defines who/how the metadata was created (it may reflect a chain of creation steps)
- contextual information—can frequently be derived by cross referencing multiple data sources, such as nearby landmarks or events (however, it may be more efficient to record it directly into the analytic metadata).

As we previously alluded, metadata may be developed manually, automatically, or via a hybrid approach. For military operations, manual "tagging" may be performed by analysts. However, various approaches have been patented for tagging broadcast video or other publicly available video. A system called *Videotator* [discussed in (Diakopoulos, 2006)] utilizes advanced human factors approaches to assist in the manual segmentation and tagging of video streams.

Hybrid and/or automated video analytics systems (see van der Kreeft, Macquarrie, Kemman, Kleppe, and McGuinness, 2014) use or integrate specialized analytics. Some examples of specialized analytics include:

- Visual "target" recognition. Target recognition analytics recognize specific targets in video streams. Face recognition [see (Faltemier, Bowyer, and Flynn, 2008)] and gait recognition [see (Man and Bhanu, 2006)] are used to recognize persons (targets). Vehicle recognition [see (Guo Rao, Samarasekera, Kim, Kumar, and Sawhney, 2008)] recognizes specific vehicles. While extremely capable, such analytics require video with clear views of the target being recognized. Pixels on target is one measure used to characterize the video quality. The rotation of the

target relative to the camera is another common measure.

- Voice, speech, and sound recognition analytics can be run against the audio stream embedded into a video transport stream. Voice recognition may extract gender, age, regional accent, or even the specific person. Sound recognition can be used to recognize nonspeech sounds such as the whirring of engines or small arms fire.
- Optical character recognition recognizes text such as on signage or uniforms.
- Human-sourced auxiliary video information can also be exploited to develop "tags." One system called *Broadcast News Analysis* (Maybury, Merlino, and Rayson, 1997) performs analysis on close (manually) captioned text. Another system [an approach discussed by Yao et. al (Yao, Mie, Ngo, and Li 2013)] mines user search behavior to infer annotations or tags in video streams. Content similarity to infer tags for one video from a similar video that was (manually) tagged is also used.

Emerging trends

As described in AI Index 2017, artificial intelligence, and highly related technologies such as machine learning and autonomous systems, have been rapidly evolving. This is driven by many technical and market factors including:

- troves of "labeled" data gathered from search engines and social media, among others
- the continued increase in compute power
- vast open-source software and tools offerings.

Some of these advances are already being used to support fully automated video analytics [see (Vandersmissen, Sterckx, Demeester, Jalalvand, De Neve, and Van de Walle, 2016)].

Two areas of emerging research support the discovery as well as searchability: knowledge discovery and predictive analytics. Both areas are broad and utilize myriad techniques. Further, they overlap somewhat as a discovery of a pattern might be considered to support predictions related to the pattern.

Knowledge discovery frequently employs algorithms that use lower-level analytics or labeled data to promote the discovery of patterns, trends, and relationships between persons and/or organizations. For instance, knowledge discovery might be used to uncover nonobvious collaboration between persons.

Predictive analytics may use similar techniques but also employ unsupervised learning techniques such as abnormal behavior or pattern discovery to predict events. For instance, atypical emerging rendezvous locations or facility reconnaissance might be a predictor of an impending attack.

Knowledge discovery and predictive analytics frequently utilize multiple sensors and thus are supported by video stream searchability. As searchability improves, then knowledge discovery and predictive analytic results should improve as well.

Video analytics to enhance military operations

Raytheon has invested research funds for several years in determining how video analytics may be used to support the war fighters, such as the ones backed in Afghanistan today. In the following sections, we report on the progress toward this goal and the recommendations for further advances.

Much of the emerging research, while generally supporting underlying goals and tool development, is not directly applicable to the military domain. Some reasons for this include:

- Military operations generally cannot exploit crowdsourcing because their primary video streams are not available for crowd viewing.
- Many maturing analytics are not practical for video captured from overhead (relatively distant) vantage points.
- Video sensors are relatively sparse in most complex battle spaces.
- Military organizations generally cannot support a large personnel footprint to maintain analytics tools in forward locations and war zones.
- Consumer industries have been able to fund large investments into analytics because they can amortize the cost across large consumer markets. Military organizations may not be able to make investments on the same scale.
- Military organizations are very sensitive to an analytics' receiver operating characteristic (ROC) curve.

To expand upon the last bullet point, the ROC curve is a concept from statistics that characterizes the true positive rate (TPR) and the false positive rate (FPR) of an analytic. The TPR, or *probability of detection*, characterizes the sensitivity of the analytic. The FPR, or *probability of false alarm*, characterizes its selectivity. If the TPR is low, then military forces may have a false sense of security that their analytics can protect them. When the FPR is high, military personnel may waste time and angst responding to false alarms.

Characteristics influencing video analytics usage

As described previously, many analytics will provide confidence values and may even yield multiple results. Military techniques, tactics and procedures (TTPs) can be defined around how software or personnel should react to analytic results that reach a given confidence level. Thus, it is important to tune an analytic to yield confidence values consistent with established TTPs. Typically, tuning an analytic for higher TPRs (high sensitivity) will also yield higher FPRs (lower selectivity). So fewer missed true events often coincides with more false alarms. A military commander may need to provide guidance on which is more important: selectivity or sensitivity. This may change frequently based on the battle rhythm and overall status of the war fighting effort.

We have also found analytics for military applications must not only recognize targets, they must aggregate target recognitions into tracks (Fig. 1). Moving target indicator standards describe data structures for associating target recognitions in consecutive video frames with a single track.

For low-resolution overhead video, it is further helpful to exploit the motion characteristics of a target object to improve selectivity and sensitivity. As shown in Fig. 2, it may be difficult for an analytic, or even a human, to distinguish between a vehicle, tumbleweed blowing along a highway, or stationary object when viewed from a single low-resolution frame. Optical flow or similar computer vision techniques make it possible to exploit object motion across frames improving overall results.

Augmenting video analytics with scene analytics

Unlike commercially available video sensors, most military sensors have an embedded metadata stream that complies with NATO's MISB-encoded video frame standard and 0601.X metadata standard (X denotes revision). The 0601.X metadata provide information that can be used directly or in conjunction with other analytics to generate useful tags. This metadata is measured directly by the sensor or the platform carrying the sensor. At a high level, 0601.X metadata contain the platform position, heading, and velocity, and the sensor field of view. From this metadata, "scene analytics" tags can be directly computed, including zoom in/zoom out, staring, field of view change, and trajectory change. These tags can help an analyst quickly find interesting video segments.

Making use of military ISR processes

We have also demonstrated the use of the Intelligence Preparation of the Operating Environment (IPOE) to automatically establish alerts and to guide analytics. The IPOE is the analytical process military intelligence organizations use to produce intelligence assessments, estimates, and other intelligence products in support of the commander's decision-making process.

There are byproducts of this process that can be exploited in establishing alerts and/or prioritizing analytics. Some examples include:

- A list of "persons of interest." Analytics that can recognize these persons can be given priority.
- Lists of geographic regions including target areas of interest and named areas of interest. Alerts can be raised while a video stream is traversing one of these regions.
- Intelligence products. Products developed by the process may be used as "labeled" data to periodically retrain analytics.

Applications to related domains

In this section, we will briefly describe significant differences and similarities to related domains. Table 1 summarizes these findings.

Over the past decade, we have been involved in border security (e.g., U.S. southern border), critical infrastructure protection (e.g., schools), and event security (e.g., the Boston Marathon) engagements. The primary difference that we observed between this domain and the military domain is the type and mix of video sensors that are employed. This heavily influences the analytics that are applicable.

FIG1 When searching for "green truck," a user will prefer the results aggregated into a track rather than individual frames. MTI standards retain all frame data to fully support geospatial searches.

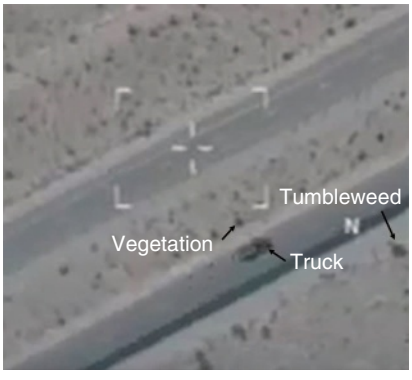


FIG2 The tumbleweed and truck are easily recognizable when seen in motion but are tough to distinguish in a single frame.

Our military work has focused on ground operations. Table 1 distinguishes this from military air or sea operations. Analytics for air or sea objects can be more powerful since the scenes are less cluttered and the variety of objects more limited.

Results and lessons learned

Civilian security and military organizations generally restrict specific information regarding how they are employing video analytics, what sensors are being used, and the effectiveness of their analytics. For this reason, the discussion here will be generalized.

In our engagements (Fig. 3), we have demonstrated and/or employed automated target-recognition analytics, text-recognition analytics, human-sourced tagging, and/or scene analytics. In some cases, video analytics have been used to augment other target-recognition technologies. These analytics may be used forensically (to find a person or vehicle in an archived stream) or in real time (to recognize a person, vehicle or event in live video).

We have seen TPRs nearing 100% and FPRs very close to 0% for specific video analytics. The factors mostly affecting these rates are quality of video (pixels on target), video viewpoint with respect to targets, and scene clutter. Even for video sensors with poor resolution or viewpoints or substantial scene clutter, the TPRs and FPRs are often good enough that our operators utilize the results.

While continuing to increase the sensitivity and selectivity of video analytics for all of those employed is important, especially in situations of poor video quality, it is probably just as important to focus on providing operators with context for all tags generated by an analytic (see Fig. 4).

Conclusion

In our work, we have found that video analytics is increasingly

TABLE 1. A summary of factors influencing video analytic usage in security applications.

	APPLICATION DOMAIN			
ANALYTIC FACTOR	MILITARY GROUND	OTHER MILITARY DOMAINS	HOMELAND SECURITY EVENT PROTECTION	HOMELAND SECURITY OTHER
Airborne overhead sensors	Heavily used. Air or sea scenes are generally much less cluttered than ground scenes.		Are less prevalent but will likely be used more heavily in the future.	
601.X metadata	Available with almost every video stream.		Generally not available on the video sensors used.	
Crowdsourcing	Generally not available. Military analysts can support some limited crowd sources.	Analyst footprint on platforms is generally very limited, so it's not practical.	Can be a very effective way of getting early warnings of aberrant events or behaviors.	Crowds are available but may not be as engaged as for event protection.
Cameras of convenience	Limited usage due to security and other restrictions.		Have been effectively used for forensic analysis. Simpler integration may make them more useful for real-time and predictive applications.	
Manpower	Limited availability of highly trained, focused analysts.	Extremely limited, especially on military aircraft.	Increasing manpower footprint to support events. Further IT advances will make them more effective in supporting analytics processes.	Several law enforcement and military reserve agencies and units coordinate. IT advances will make them more effective in supporting analytics processes.
Established ISR processes	Can be exploited to improve analytic results.		Not as mature as military processes but can support in similar ways.	Varies widely among organizations.
Ground-based stationary sensors	Limited availability and usefulness.		Predominant video streams available. Can be supportive of noncooperative face, gait, text, voice, and speech recognition.	Varies widely among organizations.
Labeled data sets	In general, data sets for conflict areas are much smaller than what is available for homelands.		Many labeled data sets exist.	

FIG3 A temporary incidence response command center that was established by Raytheon for the 2014 Boston Marathon. A running alert display helped responders focus their attention on potential threats. Video analytics further supported forensic analysis.

being used to address video glut as the cost/benefit ratio continues to decrease, where cost includes manpower considerations. As artificial intelligence, machine learning, and autonomous systems improve, the selectivity and sensitivity of analytics will increase, and they will be more widely employed. As the usage of UASs increases, advances in air-to-ground video analytics will emerge rapidly. As sensor technology improves, UASs will carry better video sensors inherently improving analytic results.

Acknowledgment

The work discussed in this article has been primarily funded by Raytheon internal investments dating back to 2008 and has been influenced by lessons learned from our PSDS2 program funded by the U.S. Army and the Mission Video Distribution System program funded by the U.S. Air Force.

Read more about it

- D. Gregory, "From a view to a kill: Drones and late modern war,"

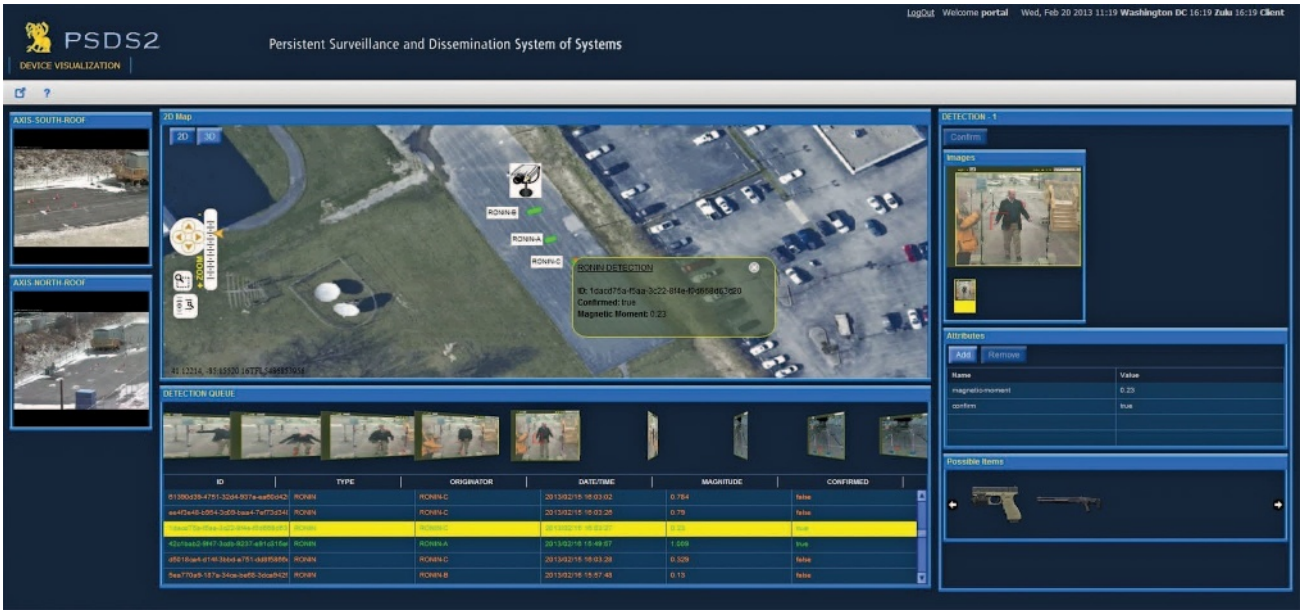


FIG4 This interface is intended to provide the user with the context of an alert so that he or she can quickly decide on an appropriate action, if any, to take in response.

Theory Culture Soc., vol. 28, no. 7–8, pp. 188–215, 2011.

• N. Diakopoulos and E. Irfan. "Videotater: An approach for pen-based digital video segmentation and tagging," in *Proc. 19th Annu. ACM Symp. User Interface Software and Technology*, 2006, pp. 221–224.

• P. van der Kreeft, M. Kay, K. Max, K. Martijn, and Kevin MGuinness. "AXES-RESEARCH: A user-oriented tool for enhanced multimodal search and retrieval in audiovisual libraries," in *Proc. IEEE 12th Int. Workshop Content-Based Multimedia Indexing*, 2014, pp. 1–4.

• T. C. Faltemier, K. W. Bowyer, and P. J. Flynn, "A region ensemble for 3-D face recognition," *IEEE Trans. Inform. Forensics Security*, vol. 3, no. 1, pp. 62–73, 2008.

• J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 2, pp. 316–322, 2006.

• Y. Guo, R. Cen, S. Supun, K. Janet, K. Rakesh, and S. Harpreet, "Matching vehicles under large pose transformations using approximate 3D models and piecewise mrf model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

• M. Maybury, M. Andrew, and James, R. "Segmentation, content extraction and visualization of broadcast news video using multistream analysis," in *Proc. ACM Multimedia Conf.*, 1997, pp. 102–112.

• T. Yao, M. Tao, C.-W. Ngo, and S. Li. "Annotation for free: Video tagging by mining user search behavior," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 977–986.

• S. Siersdorfer, J. S. Pedro, and M. Sanderson, "Automatic video tagging using content redundancy," in *Proc. 32nd Int. ACM SIGIR Conf. Research and Development Information Retrieval*, 2009, pp. 395–402.

• Stanford University. (2017, Nov.). AI INDEX. [Online]. Available: https://aiindex.org/2017-report.pdf

• B. Vandersmissen, S. Lucas, D. Thomas, J. Azarakhsh, D. N. Wesley, and R. Van de Walle, "An automated end-to-end pipeline for fine-grained video annotation using deep neural networks," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 409–412.

About the authors

Susan Gottschlich (Susan.Gottschlich@raytheon.com) earned her B.S. degree in computer and electrical engineering and her M.S. and Ph.D. degrees in electrical engineering from Purdue University. She is currently a senior engineer fellow at Raytheon Company, where she supports programs addressing video surveillance, military communications, and next-generation military platform mission equipment.

Brian L. Stone (Brian.L.Stone@raytheon.com) earned his B.S. degree in business economics from Southern Illinois University at Carbondale and graduated as a Distinguished Military Graduate, receiving a U.S. Army Reserve Officers Training Corps regular Army commission as an intelligence officer. He joined Raytheon Company in 2000, where he is currently the chief engineer for the Blue Eye Video Management System.

Bill Gerecke (W.L.Gerecke@raytheon.com) earned his B.S. degree in computer engineering from Clarkson University. He is currently a senior principal software engineer with Raytheon Company, where he leads a software team in development of systems to process and exploit video.

Technology insight on demand on IEEE.tv

Internet television gets a mobile makeover

A mobile version of IEEE.tv is now available for convenient viewing. Plus a new app for IEEE.tv can also be found in your app store. Bring an entire network of technology insight with you:

- Convenient access to generations of industry leaders.
- See the inner-workings of the newest innovations.
- Find the trends that are shaping the future.

IEEE Members receive exclusive access to award-winning programs that bring them face-to-face with the what, who, and how of technology today.

Available on the App Store

Available in Android Market

Tune in to where technology lives www.ieee.tv

IEEE
Advancing Technology for Humanity

IEEE POTENTIALS MAGAZINE REPRESENTATIVE
Mark David
Director, Business Development—Media & Advertising
Phone: +1 732 465 6473
Fax: +1 732 981 1855
m.david@ieee.org

445 Hoes Lane, Piscataway, NJ 08854

Digital Object Identifier 10.1109/MPOT.2017.2770667

NUMBERS—© CAN STOCK PHOTO/AGSANDREW,
ANDROID—© CAN STOCK PHOTO/KIRSTYPARGETER

Problem #1: Monk Spot

A monk started up a hill one morning and reached the top at sunset. He stayed overnight in the temple there, started back down the hill the next morning by the same path, and arrived at the bottom at sunset. Was there a place on the path that he passed at exactly the same time of day going both ways?

Problem #2: The Full Monty

The following is the Monty Hall problem, which is named after the host of the television game show *Let's Make a Deal*. On a game show, there are three doors. The host says, "Behind one door is a Ferrari, and behind the other two are goats. I know which door is which. Please point to a door. Then I'll open one of the other two doors to show a goat. After you've seen the goat, you may switch your choice to the third door or keep your original choice. You'll win whatever is behind the door of your final choice." Should you switch, stay with your original choice, or does it even matter?

Problem #3: The Fuller Monty

If you understand the solution to the Monty Hall problem, try this extension. Now there are n doors. Behind one is a Ferrari, and behind the other $n - 1$ are goats. The host lets you choose a door and then opens $n - 2$ of the others, revealing $n - 2$ goats. Again, you have the option of switching your choice. What is your probability of winning the Ferrari if you stay with your original choice or switch to the last door?

Problem #4: Lock, Stock, and Sinking Barrels

While a ship is floating in a closed canal lock, some of the cargo slides overboard and sinks to the bottom of the lock. Does the water level in the lock rise, fall, or remain the same?

Problem #5: By Hook or By Crook

There are 100 politicians at a political convention. Each one is either crooked or honest. At least one is honest. Given any two of the politicians, at least one is crooked. How many are honest, and how many are crooked?

P

If you have a problem for the Gamesman, please submit it along with the solution to potentials@ieee.org.
Solutions are on page 9.

Digital Object Identifier 10.1109/MPOT.2018.2828178
Date of publication: 11 July 2018

IEEE Access®

Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Best New Journal in STM 2015

Become a published author in 4 to 6 weeks.

Get on the fast track to publication with the multidisciplinary open access journal worthy of the IEEE.

IEEE journals are trusted, respected, and rank among the most highly cited publications in the industry. IEEE *Access* is no exception; the journal is included in Scopus, Web of Science, and has an Impact Factor.

Published online only, IEEE *Access* is ideal for authors who want to quickly announce recent developments, methods, or new products to a global audience.

Publishing in IEEE *Access* allows you to:

- Submit multidisciplinary articles that do not fit neatly in traditional journals
- Reach millions of global users through the IEEE *Xplore*® digital library with free access to all
- Establish yourself as an industry pioneer by contributing to trending, interdisciplinary topics in one of the Special Sections
- Integrate multimedia and track usage and citation data on each published article
- Connect with your readers through commenting
- Publish without a page limit for **only $1,750** per article

Learn more at:
ieeeaccess.ieee.org

IEEE

17-PUB-013 3/17