2-2022

The Data Analytics and the Science Revolution

Leila Halawi
Embry-Riddle Aeronautical University, halawil@erau.edu

Amal Clarke
Embry-Riddle Aeronautical University, Amal.SbeitiClarke@erau.edu

Kelly George
Embry-Riddle Aeronautical University, georged8@erau.edu

Scholarly Commons Citation
Halawi, L., Clarke, A., & George, K. (2022). The Data Analytics and the Science Revolution. *Harnessing the Power of Analytics,* (). https://doi.org/10.1007/978-3-030-89712-3

Leila Halawi
Amal Clarke
Kelly George

Harnessing the Power of Analytics

Springer

Harnessing the Power of Analytics

Leila Halawi • Amal Clarke • Kelly George

Harnessing the Power of Analytics

Springer

Leila Halawi
College of Aeronautics
Embry–Riddle Aeronautical University
Clearwater Beach, FL, USA

Amal Clarke
College of Arts and Science
Embry–Riddle Aeronautical University
Frimley, Surrey, UK

Kelly George
College of Aeronautics
Embry–Riddle Aeronautical University
South Riding, VA, USA

ISBN 978-3-030-89711-6 ISBN 978-3-030-89712-3 (eBook)
https://doi.org/10.1007/978-3-030-89712-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Introduction to Analytics and Data Science

Learning Objectives

- Grasp the difference between and need for analytics and data science
- Recognize the difference between the data analyst and data scientist
- Identify and describe the different types of analytics
- Classification of different applications and vendors
- Identify the many different methods of prediction
- Introducing SAS Viya
- Road plan for the book

1.1 The Data Analytics and Data Science Revolution

The future pauses for no one. Change nowadays is more complicated, quick, and tough to predict than ever. Data have and will continue to transform businesses such as FedEx, Google, Intel, Apple, Tesla, Uber, and Amazon. Innovation driven by data, powered by venture capital, is restructuring the world. The Internet provides instantaneous access to just about all types of information, and we should anticipate similar urgency for all kinds of solutions within the workplace. Data operationalization and manipulation enhance business functioning to best define competitive advantage well into the future. This voracious desire for data is a cultural change and paradigm shift witnessed on a global stage (Kiron and Shockley 2011).

Data analytics and data science are now being promoted in online media, journals, and even at conferences, thus surpassing the limited time to a publication that printed books present. Data science and analytics have the power to predict political races in real time, expose buying habits, and predict, with succinct results, many of our pressing research questions today. According to a study by Gartner (2018), a worldwide survey (n = 196) established that 91% of companies

L. Halawi et al., *Harnessing the Power of Analytics*,
https://doi.org/10.1007/978-3-030-89712-3_1

have not yet achieved a "transformational" maturity level in data science and analytics. Therefore, data science and analytics are now the number 1 investment priority of CIOs in recent years (Meulen and McCall 2018). Historically, novel technological advances initially emerged in technical and academic periodicals. The knowledge and synthesis presently seeped into other publications, many in book format.

Data science and the related fields of business intelligence and analytics are becoming progressively central to academic and business communities, as seen in recent literature (Chen et al. 2012). However, there is still a great deal of confusion surrounding the meaning of data science and analytics among practitioners and academia.

To dispel this confusion, especially for those seeking a career in these fields, one needs a clear definition of data analytics and data science and then a clear occupational description of the job of a data analyst as opposed to the data scientist, including the duties, tasks, knowledge, skills, and traits.

1.2 The Difference Between Data Analytics and Data Science

While data analytics and data science are often used as synonyms, there are differences between the two disciplines, most specifically in terms of knowledge and skills.

The phrase data analytics is used in place of business intelligence (BI). Many practitioners and consultants describe analytics differently. For the Institute for Operation Research and Management Science (INFORMS), analytics denotes the mixture of computer technology, management science techniques, and statistics to resolve original problems. Of course, other enterprises propose their interpretation and motivation for analytics. For SAS Institute Inc., analytics follows the data everywhere, and analytics is more than algorithms; the value generated is the focus point (Schabenberger 2020). Throughout this book, "data analytics" will generally refer to the analysis of data sets to derive meaningful trends and develop visual displays of existing data to help businesses answer questions or problems.

Data science is not a specific discipline in itself but yet spans across all disciplines and is a broader term than Data Analytics. Data science is a ubiquitous term designating an interdisciplinary field regarding processes and systems to gain knowledge and insights from data (Bichler et al. 2017). Data science generally refers to generating information from large unstructured and structured data sets. Figure 1.1 from Northeastern University highlights the difference between the disciplines, the methods, and the overlap (Burnham 2019).

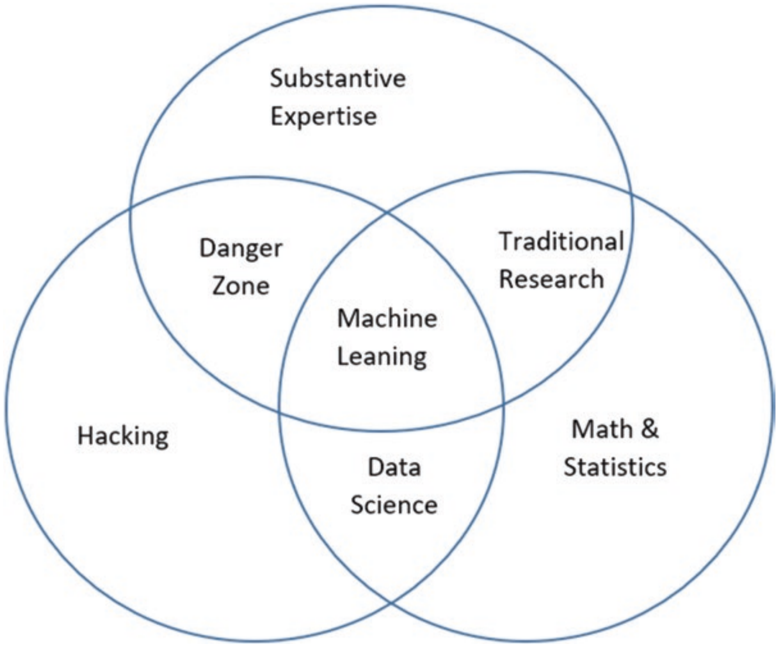

Fig. 1.1 Data analytics and data science skills

1.3 Data Analyst Versus Data Scientist

Data analytics and data science fields offer a range of duties, tasks, knowledge, skills, and traits. There is a clear overlap between the data scientist and the data analyst. However, the breadth and depth of the skills necessary differ, making each role complementary but different in focus, viewpoint, and expertise.

A data analyst is a professional who used to do “business intelligence” (now called analytics) in data compilation, cleaning, reporting, and visualization. Some of these professionals may have a more in-depth knowledge of programming to code for data cleaning and analysis.

A “talented” data scientist would generally be supposed to have rigorous business skills to assess the significance of generated insights and address significant business problems (Provost and Fawcett 2013).

According to NAP (2014), extracting meaning from data requires considerable skills that include:

- Statistics
- Machine learning
- Optimization
- Software engineering

- Product sense
- Careful experimentation

While a profound appreciation of numbers and mathematics is imperative to be successful in this field, a data scientist must also possess outstanding communication skills, be an eminent system thinker, have keen visualization skills, and above all possess critical thinking skills concerning how data can be utilized in decision-making and the impact this type of analysis can have upon people's lives.

1.4 Example: Data Science and Analytics in Aviation

The airline industry has grown approximately 5% per year over the last 30 years. The data science revolution is also transforming aviation. From airlines to air navigation service providers or airports, the capacity to gather information across separate physical data sensors is expanding exponentially. Data science in aviation offers countless opportunities to enhance products, processes, and imagine new means to develop safer and more efficient aviation systems. A significant amount of unstructured, varied data from distinct stakeholders and various types are collected and stored within the aviation sector, comprising safety data and reports, flight plans, navigation data, airport data, and radar tracks, among other data types.

Airlines and airports have limited capability to process this haul of data and employ advanced analytics and artificial intelligence to inform operations and maintenance and rarely in real time (Maire and Spafford 2017). Airlines can use data to fine-tune their fuel loads, beverage inventory, and personnel requirements to save money and increase efficiency. To bridge the gap between supply and demand, training the next generation, and retooling the current aviation workforce is necessary for the long run. Yet, it will not suffice long term; it is not sufficient.

The Bureau of Aircraft Accident Archives (B3A) reports that, on average, 230 fatal aviation accidents occurred, and 1709 passages perished annually in the last half-century. The best estimate we found from the literature states that "somewhere between 60 and 80% of aviation accidents are due, at least in part, to human error" (Shappell 2006). Such a preliminary estimate can be and should be improved at the current age of data science and the state of artificial intelligence (AI) technology. However, pinning down the exact measures for reducing such tragic accident rates and annoying flight delays need thousands of data points to analyze the ground truth of enormous aviation data.

The subject matter expertise in aviation is extensive enough for a non-aviation background person to understand and interpret most aviation data sets. The aviation industry mostly hires students with an aviation background paired with a data management perspective.

Data science and analytics in aviation offer countless opportunities to enhance products, processes, and imagine new means to develop safer and more efficient aviation systems. Within the aviation sector, a significant amount of unstructured, varied data from distinct stakeholders and various types is collected and stored, comprising safety data and reports, flight plans, navigation data, airport data, and radar tracks, among other data types. A recent *Aviation Week* article highlights the characteristics of the new Pratt & Whitney Geared TurboFan (GTF) engines that employ 5000 sensors and are capable of generating up to 10 GB of data per second. By comparison, current engines only have 250 sensors at most. This results in a twin-engine aircraft equipped with the new Pratt & Whitney GTF, with an average of 12 h flight, generating up to 844 TB of data. With thousands of engines to be built, zettabyte (ZB) of data will be available.

Consequently, the infrastructure needed to handle such data needed significantly upgraded. That infrastructure will be required to be put in place to benefit from the wealth of information the engine data could provide. From this point of view, engine health monitoring is getting an entirely new perspective. The INNAXIS research institute deems the application of data science principles to the aviation sector as an open gate to substantial improvements in numerous main aspects of aviation, such as safety enhancement, flight efficiency, environmental impact mitigation, or delay reduction. According to Maire and Spafford (2017), the flight-related data amount is increasing significantly, and this increase enables making a more profound analysis; however, airlines and airports have limited capability to process this haul of data and employ advanced analytics and AI to inform operations and maintenance and rarely in real time.

SAS has been successfully used by several airlines for various use cases. See the SAS website for more information (www.sas.com/fi_fi/customers/scandinavian-airlines.html).

1.5 Analytics Methods

INFORMs, The Institute for Operations Research and the Management Sciences, proposed three levels of analytics. These levels are identified as descriptive, predictive, and prescriptive (Informs 2014).

Figure 1.2 presents a graphical overview of the analytics process with the three different analytics levels.

While descriptive analytics focuses primarily on what has already happened in the past, and predictive analytics tries to find correlations to make forward-looking projections, prescriptive analytics looks to determine what to do or give you an answer as to how to proceed on the information provided to the data model.

1. **Descriptive analytics—What happened and what is happening?** Descriptive analytics is the entry-level in the analytics world. It frequently entails working with queries, looking at descriptive statistics, data visualization including dashboards, and generating reports as most of the analytics actions at this level involve creating a report to summarize business activities and answer the question, "What happened or what is happening?" For example, a query to a database

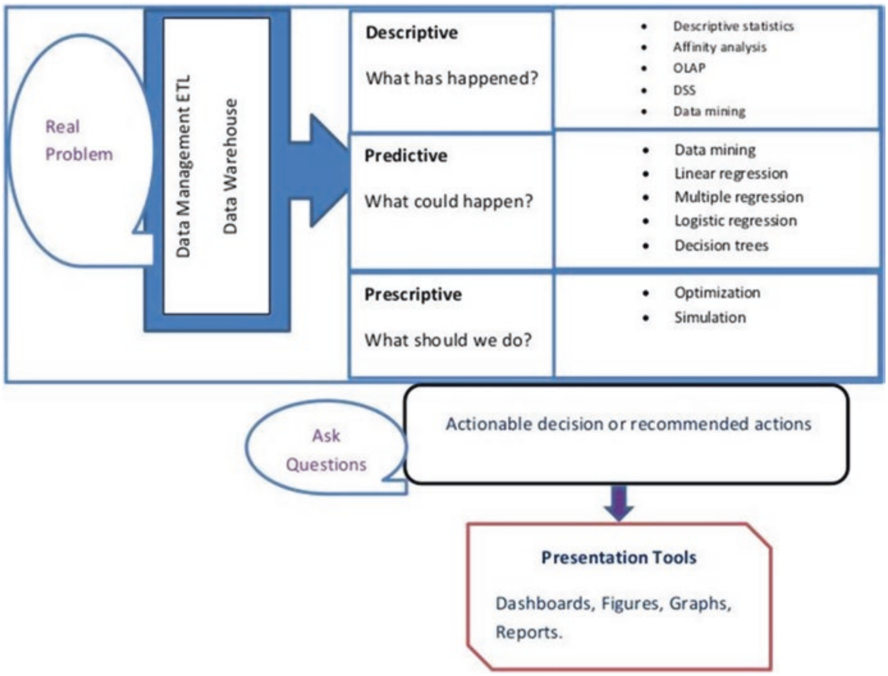

Fig. 1.2 Analytics process and levels

for a shipment facility for Amazon during April will provide descriptive information about all the shipments, the number of purchases, the dates for each delivery, and the type and quantities shipped per order, among other things. A report can later be generated with summary statistics such as mean, standard deviation, and some data visualizations to detect any patterns or relationships.

2. **Predictive analytics—What will happen? Why will it happen?** Organizations that matured in descriptive analytics often move into this level to look beyond what happened and try to predict or answer what will happen in the future. This form of analysis is usually based on statistical techniques and, data mining techniques. For example, in the financial sector, predictive models predict future financial performance and assess investments' risks. Another example used widely across all retail companies and services providers in developing predictive models to support pricing decisions they take, understand consumer behavior, increase loyalty, and customer satisfaction, among other uses.
3. **Prescriptive analytics**—What is the best outcome that will happen? Prescriptive analytics is the ultimate level in the analytics world or hierarchy. It is where the most excellent choice among many created and classified is determined using sophisticated mathematical models. This type of analysis answers the question: what should I do? It uses optimization, simulation, and heuristics-based decision-making modeling techniques. This group of methods has traditionally been studied under operation research (OR) or management sciences. For example, in

Table 1.1 The three forms of analytics

Descriptive analytics	Also known as exploratory descriptive analytics. Here you detect data inconsistency and formatting issues and execute statistical tests to establish which variables appear most closely related, complete, useful, and incomplete. Exploratory descriptive analytics is commonly aimed to provide insights into the variations, relationships, and patterns in data that can inform later analyses Descriptive analytics may also be focused on answering specific business questions. What were the sales last month? What was the shortage in a given product? It is important to note that the only thing that descriptive analytics requires is data (and a tool to analyze it)
Predictive analytics	The purpose of predictive analysis is to determine what is likely to happen in the future based on data mining techniques Running a regression to explain the impact of X1 of Y is descriptive. Building a predictive model with training and validation data sets to predict and score future Y data based on X1 is predictive In these models, we identify the who, what, when, where, and so on, which is vital
Prescriptive analytics	Its goal is to anticipate what is going well and the likely forecast to realize the best performance. For example, optimization tells you what to do (how many of each product to produce based on the expected demand of each product, which products to launch to market based on consumer survey responses, etc.) Identifying the *why* improves understanding the who, what, when, where, and so on, and the different ways they are connected

the sports industry, companies use prescriptive analytics to dynamically adjust their ticket prices throughout the season, reflecting each game relative attractiveness and potential demand.

Predictive and prescriptive analytics are collectively called advanced analytics. Prescriptive analytics may be considered the last step of business analytics. However, it is essential to note the difference between these two forms related to the outcome of the analysis. Also, predictive analytics offers you data in hard ways to help you make informed decisions.

For example, healthcare uses descriptive, predictive, and prescriptive analytics to improve the facility, staff, and patients' scheduling, provide more effective treatments, predict patient flow and diagnosis, and even treatments.

Table 1.1 summarizes the three forms of analytics.

1.6 Classification of Different Applications and Vendors

Gartner, Inc., formerly known as Gartner, is a worldwide research and advisory firm offering information, advice, and tools for leaders in IT, customer service and support, and supply chain functions, among others. The Magic Quadrant 2020 assesses data analytics and BI platform vendors and data science and machine learning (DSML) platforms.

Richardson et al. (2021) identified many different vendors within the data analytics and BI platforms and compared them across many different functionalities and capabilities. These included security, cloud, data source connectivity, manageability, data preparation, data visualization, reporting, advanced analytics, and model complexity, to name a few. Microsoft, Tableau, SAS, SAP, Qlik, Microstrategy, Oracle, IBM, and others were identified, and advantages and disadvantages for each platform were listed.

The data science and machine learning (DSML) platforms, Gartner (2020a, b) reviewed and classified different vendors based on various features including data exploration, data preparation, model testing, deployment, maintenance, and collaboration. They also presented the distinct advantages and disadvantages for each vendor. This information is an integral part of the selection process for any organization. Vendors such as Alteryx, SAS, Databricks, MathWorks, Tibco Software, and Dataiku were classified as leaders in the field.

1.7 Why the Many Different Methods?

There are many prediction methods and one cannot but ask why they coexist and whether some are better than others. One should know that each method has advantages and disadvantages.

The data set's size may impact the usefulness of a particular method, the analysis goal, how messy the data are, the existing patterns, if any, in the data set, and whether the data meets the method's underlying assumptions. Figure 1.3 highlights the different techniques across the data analytics and data science field and the need for application tools.

1.8 What You Need to Know About SAS Viya

SAS has evolved a lot in the last 30 years and is still among the elites in the artificial intelligence (AI) and advanced analytics world. SAS® Viya® is an open analytics platform that may process any data type, volume, or speed. This cloud-enabled, in-memory analytics engine is adaptable, scalable, and fault-tolerant. It includes a standardized code base that adopts programming in SAS and other languages, such as Python, R, Java, and Lua. Furthermore, it may deploy flawlessly to any infrastructure or product ecosystem with support for cloud, on-site, or hybrid environments. The high-performance processing power of SAS Viya is offered by SAS Cloud Analytics Services (CAS), an in-memory engine that can radically hasten data management and analytics with SAS. SAS Viya is planned to synchronize with SAS 9.4 solutions and the SAS 9 environment. Some procedures are available in both SAS 9 and SAS Viya. While some existing SAS code can still run in SAS Viya,

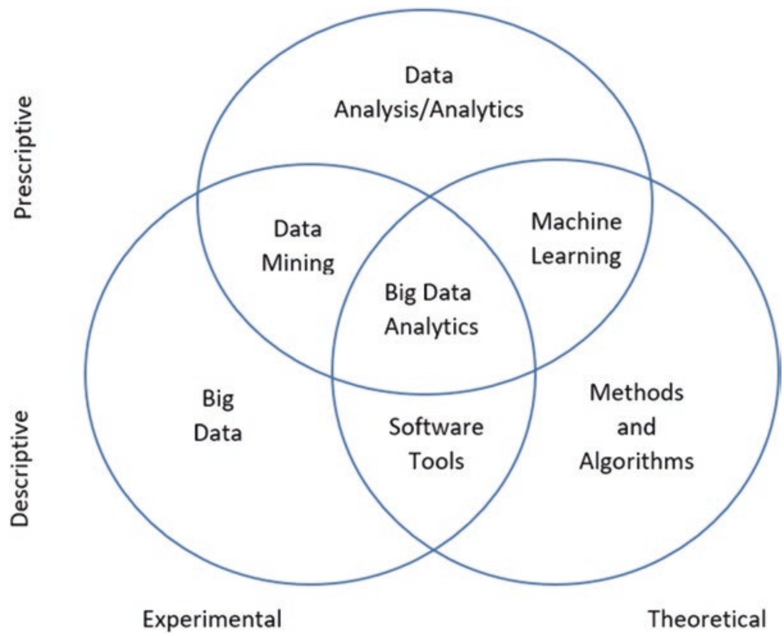

Fig. 1.3 The fields of data analytics and data science

SAS Viya also contains additional procedures that take advantage of the open, distributed environment.

1.9 Road Plan for the Book

The book covers many of the widely used predictive methods using SAS Viya. The book comprises eight chapters.

This introductory chapter discussed the difference and needs for data analytics and data science and highlighted the difference between the data analyst and data scientist. We identified and described the different analytics types and concluded the chapter with a discussion of SAS Viya, the leading platform used in this Book.

Chapter 2 focuses on the different data types and measurements, big data, data partitioning, honest assessment, an overview of the data preparation process, and the steps we take to connect data in SAS Viya. We will explore the SAS VIYA platform. The data set is a random sample subset of a more extensive database (over 600,000 records) for 2018 South West Flights (SWF) out of Florida on the Bureau Transportation Statistics (BTS) website (Bureau of Transportation Statistics 2020).

Chapter 3 focuses on the early stages of data exploration and highlights data visualization. We will also introduce the different descriptive measures and demonstrate how to load and explore data in SAS Viya.

Chapter 4 concentrates on evaluating predictive performance. Several measures are used and explained with their interpretations of the outcomes for assessing predictive performance.

Chapter 5 explains and demonstrates the use of a decision tree and ensemble for developing predictive models. Decision trees can model to predict a categorical variable or create a regression model for continuous variables.

Chapter 6 presents regression models, the math behind them, their uses, and then demonstrate regression modeling in SAS Viya using two different data sets: a classification example (a client who will subscribe to a term deposit or nonsubscriber) from the bank data set and a numerical variable from the concrete data set.

Chapter 7 presents neural networks. Neural networks are parametric nonlinear regression models. We create an interactive neural network and then create a NN through a pipeline demonstrating the different parameters and options and then compare the models. We show NN using the concrete data set. Concrete is an important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients used as our predictor variable.

Chapter 8 examines model deployment. It explains the deployment phase with its several tasks: Model assessment, model comparison, and monitoring model performance over time and updating as necessary.

Appendix A includes information about SAS VIYA for Learners and the many different offerings and materials for academics.

Appendix B includes a list of some open data sources.

Appendix C includes the data dictionary for the Aviation data set.

References

Bichler M, Heinzl A, van der Aalst WMP (2017) Business analytics and data science: once again? Business and information. Syst Eng 59(2):77–79. https://doi.org/10.1007/s12599-016-0461-1

Bureau of Transportation Statistics (BTS) 2020 Airline On-Time Statistics and Delay Causes. https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E

Burnham K (2019, Aug 29) Data analytics vs. data science: a breakdown. https://www.northeastern.edu/graduate/blog/data-analytics-vs-data-science/. Accessed 20 July 2020

Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36:1165–1188

Gartner (2018). https://www.gartner.com/doc/reprints?id=1-1XYUYQ3I&ct=191219&st=sb

Gartner (2020a) Magic quadrant for analytics and business intelligence platforms

Gartner (2020b) Magic quadrant for data science and machine learning platforms

Informs (2014, Dec 18) Defining analytics: a conceptual framework. https://www.informs.org/ORMS-Today/Public-Articles/June-Volume-43-Number-3/Defining-analytics-a-conceptual-framework

Kiron D, Shockley R (2011) Creating business value with analytics. MIT Sloan Manag Rev 53(1):57

Maire S, Spafford C (2017) The data science revolution that's transforming aviation. https://www.forbes.com/sites/oliverwyman/2017/06/16/the-data-science-revolution-transforming-aviation/

Meulen R, McCall T (2018) Gartner survey shows organizations are slow to advance in data and analytics. Gartner

NAP Report (2014) Training students to extract value from big data: summary of a workshop. The National Academies Press. ISBN 978-0-309-31437-4. https://doi.org/10.17226/18981. http://nap.edu/18981

Provost F, Fawcett T (2013) Data science and its relationship to big data and data-driven decision making. Big Data 1(1):51–59

Richardson J, Schlegel K, Sallam R, Kronz A, Sun J (2021). Magic Quadrant for Analytics and Business Intelligence Platforms. https://www.gartner.com/doc/reprints?id=1-1YOXON7Q&ct=200330&st=sb

Schabenberger O (2020, Feb 24) Digital transformation projects don't fail because of a shortage of 'tech'. https://thenextweb.com/podium/2020/02/24/digital-transformation-projects-dont-fail-because-of-a-shortage-of-tech/

Shappell S (2006) Human error and commercial aviation accidents: a comprehensive, fine-grained analysis using HFACS. www.faa.gov/library/reports/medical/oamtechreports/index.cfm