Getting Started

People analytics is the evidence-based practice of surfacing actionable insights from data to help people and organizations thrive. Relative to other functions such as Finance and Marketing, data-informed decisions in the talent domain is a recent concept. People analytics has the power to transform organizations by surfacing subtle barriers to success, optimizing conditions for sustained effectiveness, and increasing shareholder value.

The importance of the future of work, employee well-being, candidate and employee experience, worker productivity and collaboration, diversity, equity, inclusion, and belonging (DEIB), and retention of critical talent has resulted in companies making growing investments in people analytics capabilities. The nature of workforce challenges is increasingly too complex and nuanced for traditional HR skills, and the ability to extract intelligence from workforce data to inform strategic talent decisions is critically important.

Many organizations struggle to progress beyond basic operational reporting and dashboards, but reports and dashboards are not people analytics; they merely help inform what questions to ask based on unanticipated observations (e.g., surprising changes and trajectories) which are often the impetus for people analytics projects. The ability to think critically and reason through the available data to properly frame problems and theoretical explanations are perhaps the most essential skills for success in people analytics.

This book will cut through the fluff and teach you how to do stuff. Knowledge of concepts is futile without an understanding of how to apply them to people analytics use cases. The goal of this book is not to boil the ocean by implementing an exhaustive set of analysis methods in every tool. This book is guided by a goal of optimizing for the *fewest* number of concepts and applications required to successfully design and execute the *majority* of projects in the people analytics domain. While this is a technical book, it indexes more heavily on concepts and practical applications to people analytics than on mathematical underpinnings.

C. Starbuck, *The Fundamentals of People Analytics*,
https://doi.org/10.1007/978-3-031-28674-2_1

Whether you are a people leader, individual contributor, or aspiring analytics practitioner, this book is for you. This book will serve as a guide through the analytics lifecycle, curating the key concepts and applications germane to common questions and hypotheses within people analytics and providing a repeatable framework for successful analytics projects.

Guiding Principles

Among the many principles guiding how analytics teams operate, there are three that I have found to be universally applicable and fundamental to the success of an analytics capability.

Pro-Employee Thinking

With great power comes great responsibility.

"Pro-employee" thinking is addressed first and for good reason. People analytics has the power to improve the lives of people in meaningful ways. Whether we are shedding light on an area of the business struggling with work-life balance or identifying developmental areas of which a group of leaders may be unaware, people analytics ideally improves employee well-being and effectively, the success of the business. It is important to embrace a pro-employee philosophy, as newfound knowledge could also have damaging repercussions if shared with the wrong people or if findings are disseminated without proper instruction on how to interpret and act.

One way to error on the side of caution when considering whether to disseminate insights is to ask the following: "With this knowledge, could the recipient act in a manner that is inconsistent with our pro-employee philosophy?" If the answer to this question is not a clear *no*, discuss with your HR, legal, and privacy partners and together determine how best to proceed. The decision may be to not share the findings with the intended audience at all or to develop a proper communication and training plan to ensure there is consistency in how recipients interpret the insights and act in response. Employment Law and Data Privacy Counsel are our friends, and it is important to build strong relationships with these critical partners.

Quality

Garbage in, garbage out.

Never compromise quality for greater velocity. If quality falls to the bottom of the priority list, all other efforts are pointless. It is unlikely that requestors of data and analytics will ever ask us to take longer to prepare the information. The onus is on us

as analytics professionals to level set on a reasonable timeline for analyses based on many factors that can impact the quality of analyses and insights. A single instance of compromised quality can have lasting damage on the reputation of the analytics function and cause consumers of insights to view all findings as suspect. Be sure quality is consistently a top value and guard your team's reputation at all costs. If stakeholders lose trust, there will likely be additional data requests for validation; this is wasteful to both you and your user community and detracts from the bigger story that needs to be communicated.

Trustworthy results are highly dependent on the quality of data in source systems. If tight controls do not exist within source applications to support data integrity, downstream data cleaning efforts can only go so far in delivering reliable and valid findings. It is often the analysts who identify data integrity issues due to the nature of their work; therefore, close relationships should be formed with source application owners to put into place validation rules that proactively prevent the entry of erroneous data or at the very least, exception/audit reports to identify and address the issues once they are observed.

Prioritization

If everything is a priority, nothing is a priority.

Analyses should always have a strong value proposition—a clear expectation of how an analysis will support a General Manager, People Partner, Salesperson, or other member of the organization. Nothing in this book will increase the value of an analysis no one needs. There should be a clear business need and commitment to action before implementing an analysis. While curiosity is important, it is not a justification for an analysis.

It is crucial to be relentless about prioritizing strategically important projects with "measurable" impact over merely interesting questions that few care to answer. According to the Pareto Principle, 80% of outcomes (or outputs) result from 20% of causes (or inputs). In analytics, it is important to be laser focused on identifying the 20% of inputs that will result in disproportionate value creation for stakeholders. There are some general customer-oriented questions I have found to be helpful in the intake process to optimize the allocation of time and resources:

1. Does this support a company or departmental objective? If not, why should this be prioritized over something else?
2. Who is the executive sponsor? Really important projects will have an executive-level sponsor.
3. What quantitative and/or qualitative data can be provided as a rationale for this request? Is there data to support doing this, or is the problem statement rooted merely in theories and anecdotes?
4. Will this mitigate risk or enable opportunities?
5. What actions can or will be taken as a result of this analysis?

6. What is the *scale* of impact (# of impacted people)?
7. What is the *depth* of impact (minimum—> significant)?
8. Is this a dependency or blocker for another important deliverable?
9. What is the impact of not doing (or delaying) this?
10. What is the request date? Is there flexibility in this date and/or scope of the request (e.g., what does MVP look like)?

These questions can be weighted and scored as well to support a more automated and algorithmic approach to prioritization.

Tooling

Applications in this book are demonstrated in R, which is open-sourced statistical and data visualization software that can be downloaded free of charge. It is incredibly powerful, and there is a package (or at least the ability to easily create one) for every conceivable statistical method and data visualization. R is also widely used in highly regulated environments (e.g., clinical trials).

As of this writing, R Markdown—the dynamic document creator in which this book is written—allows for coding in 56 different languages! Therefore, debating whether to use Python, Julia, or other software is unproductive; we need not sacrifice the advantages of other languages by choosing one. All code in this book is fully reproducible, and it is recommended that you implement the analysis methods on your machine as you progress.

Software such as R allows analysts to organize and annotate steps of the analytical process in a manner that is both logical and reproducible. End-to-end analytics workflows (data extraction—> wrangling—> cleaning—> analysis—> visualization) can be fully automated and executed in R without opening the script, mitigating the risk of inadvertently modifying data or formulas. This is one of the many reasons tools like Excel will not be covered in this book.

Please note that while R basics are covered, this is not a book on how to code. An introductory programming course is highly recommended, and this is one of the best investments you can make for a successful career in analytics. The ability to write code is now table stakes for anyone in an analytics-oriented field, as this is the best way to develop reproducible analyses. Coding is to analytics professionals what typing was for Baby Boomers decades ago; a lack of coding proficiency is a major limiting factor on one's potential in this field.

The goal of the code provided in this book is not to represent the most performant, succinct, or productionalizable approaches. The code herein is intended only to facilitate understanding and demonstrate how concepts can be implemented in people analytics settings. The most performant approaches are often at odds with more intuitive alternatives. Programming expertise is important for optimizing these approaches for production applications.

Data Sets

Several data sets are leveraged throughout this book to demonstrate how to implement various analysis methods. All data sets are available in the R package named `peopleanalytics`. Instructions on installing this package and loading the data sets will be covered in chapter "Introduction to R".

Employees

The primary data set used in this book is `employees`, which contains information on active and terminated employees. Fields are defined in the data dictionary provided below:

- `employee_id`: Unique identifier for each employee
- `active`: Flag set to *Yes* for active employees and *No* for inactive employees
- `stock_opt_lvl`: Stock option level
- `trainings`: Number of trainings completed within the past year
- `age`: Employee age in years
- `commute_dist`: Commute distance in miles
- `ed_lvl`: Education level, where 1 = *High School*, 2 = *Associate Degree*, 3 = *Bachelor's Degree*, 4 = *Master's Degree*, and 5 = *Doctoral Degree*
- `ed_field`: Education field associated with most recent degree
- `gender`: Gender self-identification
- `marital_sts`: Marital status
- `dept`: Department of which an employee is a member
- `engagement`: Employee engagement score measured on a 4-point Likert scale, where 1 = *Highly Disengaged* and 4 = *Highly Engaged*
- `job_lvl`: Job level, where 1 = *Junior* and 5 = *Senior*
- `job_title`: Job title
- `overtime`: Flag set to *Yes* if the employee is nonexempt and works overtime and *No* if the employee does not work overtime
- `business_travel`: Business travel frequency
- `hourly_rate`: Hourly rate calculated irrespective of hourly/salaried employees
- `daily_comp`: Hourly rate * 8
- `monthly_comp`: Hourly rate * 2080 / 12
- `annual_comp`: Hourly rate * 2080
- `ytd_leads`: Year-to-date (YTD) number of leads generated for employees in Sales Executive and Sales Representative positions
- `ytd_sales`: Year-to-date (YTD) sales measured in USD for employees in Sales Executive and Sales Representative positions
- `standard_hrs`: Expected working hours over a two-week payroll cycle

- `salary_hike_pct`: The percent increase in salary for the employee's most recent compensation adjustment (whether due to a standard merit increase, off-cycle adjustment, or promotion)
- `perf_rating`: Most recent performance rating, where 1 = *Needs Improvement*, 2 = *Core Contributor*, 3 = *Noteworthy*, and 4 = *Exceptional*
- `prior_emplr_cnt`: Number of prior employers
- `env_sat`: Environment satisfaction score measured on a 4-point Likert scale, where 1 = *Highly Dissatisfied* and 4 = *Highly Satisfied*
- `job_sat`: Job satisfaction score measured on a 4-point Likert scale, where 1 = *Highly Dissatisfied* and 4 = *Highly Satisfied*
- `rel_sat`: Colleague relationship satisfaction score measured on a 4-point Likert scale, where 1 = *Highly Dissatisfied* and 4 = *Highly Satisfied*
- `wl_balance`: Work-life balance score measured on a 4-point Likert scale, where 1 = *Poor Balance* and 4 = *Excellent Balance*
- `work_exp`: Total years of work experience
- `org_tenure`: Years at current company
- `job_tenure`: Years in current job
- `last_promo`: Years since last promotion
- `mgr_tenure`: Years under current manager
- `interview_rating`: Average rating across the interview loop for the onsite stage of the employee's recruiting process, where 1 = *Definitely Not* and 5 = *Definitely Yes*

Most of these fields have also been broken into separate topical data sets to support data wrangling examples in chapter "Introduction to SQL": `benefits`, `demographics`, `job`, `payroll`, `performance`, `prior_employment`, `status`, `survey_response`, and `tenure`.

Turnover Trends

The `turnover_trends` data set contains turnover rates for each month across a five-year period. Fields are defined in the data dictionary provided below:

- `year`: Integer representing the year, which ranges from 1 (earliest) to 5 (most recent)
- `month`: Integer representing the month, which ranges from 1 (January) to 12 (December)
- `job`: Job title
- `level`: Job level, where 1 = *Junior* and 5 = *Senior*
- `remote`: Flag set to *Yes* for a remote worker and *No* for a non-remote worker
- `turnover_rate`: monthly turnover rate, calculated by dividing the termination count into the average headcount (beginning headcount + ending headcount / 2) for the respective month

Survey Responses

The `survey_responses` data set contains responses to various survey items. Each observation represents a unique anonymized survey respondent.

- `belong`: Belonging score measured on a 5-point Likert scale, where 1 = *Highly Unfavorable* and 5 = *Highly Favorable*
- `effort`: Discretionary Effort score measured on a 5-point Likert scale, where 1 = *Highly Unfavorable* and 5 = *Highly Favorable*
- `incl`: Inclusion score measured on a 5-point Likert scale, where 1 = *Highly Unfavorable* and 5 = *Highly Favorable*
- `eng_1`: Engagement score on item 1 of 3 measured on a 5-point Likert scale, where 1 = *Highly Disengaged* and 5 = *Highly Engaged*
- `eng_2`: Engagement score on item 2 of 3 measured on a 5-point Likert scale, where 1 = *Highly Disengaged* and 5 = *Highly Engaged*
- `eng_3`: Engagement score on item 3 of 3 measured on a 5-point Likert scale, where 1 = *Highly Disengaged* and 5 = *Highly Engaged*
- `happ`: Happiness score measured on a 5-point Likert scale, where 1 = *Highly Unfavorable* and 5 = *Highly Favorable*
- `psafety`: Psychological Safety score measured on a 7-point Likert scale, where 1 = *Highly Unfavorable* and 7 = *Highly Favorable*
- `ret_1`: Retention score on item 1 of 3 measured on a 5-point Likert scale, where 1 = *Highly Unfavorable* and 5 = *Highly Favorable*
- `ret_2`: Retention score on item 2 of 3 measured on a 5-point Likert scale, where 1 = *Highly Unfavorable* and 5 = *Highly Favorable*
- `ret_3`: Retention score on item 3 of 3 measured on a 5-point Likert scale, where 1 = *Highly Unfavorable* and 5 = *Highly Favorable*
- `ldrshp`: Senior Leadership score measured on a 5-point Likert scale, where 1 = *Highly Unfavorable* and 5 = *Highly Favorable*

4D Framework

In practical analytics settings, we generally operate with respect to five primary constraints: timeliness, client expectation, accuracy, reliability, and cost (Bartlett, 2013). Adherence to a lightweight framework over hastily rushing into an analysis full of assumptions generally lends to better outcomes that respect these constraints. A framework ensures (a) the problem statement is understood and well-defined; (b) relevant literature and prior research are reviewed; (c) the measurement strategy is sound; (d) the analysis approach is suitable for the hypotheses being tested; and (e) results and conclusions are valid and communicated in a way that resonates with the target audience. This chapter will outline a recommended framework as well as other important considerations that should be reviewed early in the project.

It is important to develop a clear understanding of the key elements of research. Scientific research is the systematic, controlled, empirical, and critical investigation

of natural phenomena guided by theory and hypotheses about the presumed relations among such phenomena (Kerlinger and Lee, 2000). In other words, research is an organized and systematic way of finding answers to questions. If you are in the business of analytics, I encourage you to think of yourself as a research scientist—regardless of whether you are wearing a lab coat or have plans to publish.

As we will discover when exploring the laws of probability in a later chapter, there is a 1 in 20 chance of finding a significant result when none exists. Therefore, it is important to remain disciplined and methodical to protect against backward research wherein the researcher mines data for interesting relationships or differences and then develops hypotheses which they know the data support. There have been many examples of bad research over the years, which often presents in the form of p-hacking or data dredging: the act of finding data to confirm what the researcher wants to prove. This can occur by running an exhaustive number of experiments to find one that supports the hypothesis or by using only a subset of data which features the expected patterning.

Academics at elite research institutions are often under immense pressure to publish in top-tier journals that have a track record of accepting new ground-breaking research over replication studies or unsupported hypotheses, and incentives have unfortunately influenced some to compromise integrity. As my PhD advisor told me many years ago, an unsupported hypothesis—while initially disappointing given the exhaustive literature review that precedes its development—is a meaningful empirical contribution given theory suggests the opposite should be true.

If you participated in a science fair as a child, you are likely already familiar with the scientific method. The scientific method is the standard scheme of organized and systematic inquiry, and this duly applies to people analytics practitioners in the promotion of robust analyses and recommendations.

An important feature of the Scientific Method, as reflected in Fig. 1, is that the process never ends. New knowledge resulting from hypothesis testing prompts a

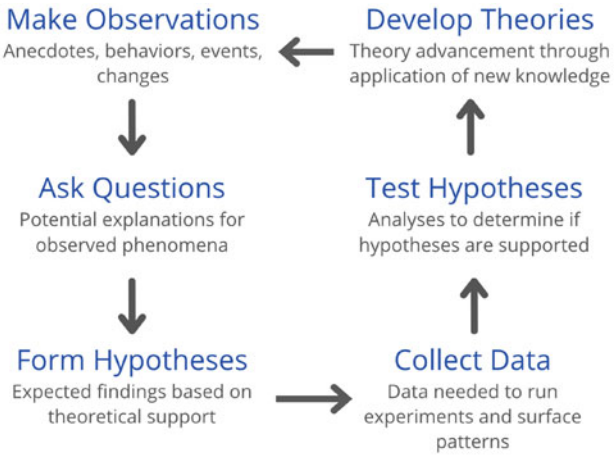

Fig. 1 The Scientific Method

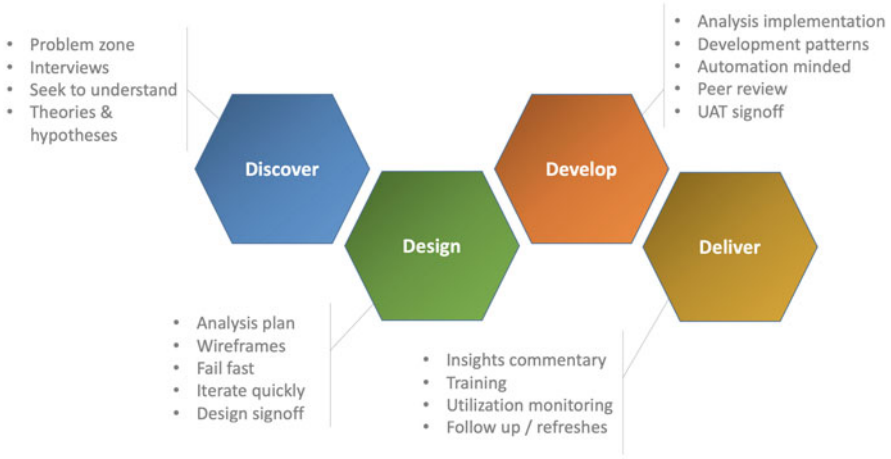

Fig. 2 4D framework

new set of questions and hypotheses, which initiates a new lifecycle of scientific inquiry.

Over the years, I have adapted the scientific method into a curtailed four-step framework to promote a rigorous and disciplined approach to the end-to-end analytical process. This framework is summarized in Fig. 2, and the four steps are (a) Discover, (b) Design, (c) Develop, and (d) Deliver.

The 4D framework, for which there is a detailed checklist in the Appendix, provides a method of structuring analytics projects to ensure they are anchored in well-defined problem statements and proactively consider the key elements at each stage of the analytics lifecycle to increase the likelihood of meaningful ROI.