SAMPLE

Building Tomorrow's Statisticians

The Signal's Keeper in the Machine Era

Aiman JoeHanz

Free sample: Introduction & Chapter 1

putrabooks.github.io

Buy the full book:

Carousell carousell.sg/p/building-tomorrow-s-statisticians-ebook-pdf-1471376304

Ko-fi

Payhip

Contents

Introduction

Chapter 1: The Definition

Chapter 2: The Territories

Chapter 3: The Path

Chapter 4: The Question

Chapter 5: The Model and the Honest Error

Chapter 6: The Record

Chapter 7: The Conversation

Chapter 8: The Machine's Audit

Chapter 9: The Working Program

Chapter 10: The Practitioner's Path

Introduction

Ask a room of strangers what a statistician is, and two caricatures will answer before any living member of the trade can. The first is the soothsayer: the figure who gazes into a scatter of points the way an oracle once gazed into entrails or tea leaves, speaks in significance rather than sentences, and issues pronouncements that the uninitiated must either trust or dismiss, with nothing in between. The second is the automaton: the functionary imagined as a row of buttons, whose work the machine era has supposedly finished, since the software now fits the lines, tests the differences, and prints the answers at a click, leaving behind a person to press the click. The soothsayer is too mysterious to be managed; the automaton is too mechanical to be paid; and between the two of them, the actual statistician has been caricatured out of the story entirely. This book begins by clearing both figures from the room, because the working reality they obscure is worth more than either fantasy: a trade whose product is not a prophecy and not a printout but the honest number, a figure that arrives at a decision with its provenance named, its method open, and its error priced, carried there by a practitioner whose whole craft is keeping the signal distinguishable from the noise.

The book is written for four readers at once, and the chapters serve all four altitudes. The student weighing the trade can read it as a reconnaissance of the ground before committing years to it: what the work actually is, where it is done, what the road from a degree to a first post looks like, and what the first decade will ask of the person walking it. The analyst already inside the trade can read it as a mirror and a repair manual, a chance to check the working habits against the craft's own standards before the habits harden into doctrine. The data scientist and the machine-learning engineer can read it as an answer to a question their own work keeps raising: what does the statistician's training still add, now that the fitting is automated, and the honest answer runs both ways, since the credential adds exactly the habits this book describes, and the automation adds exactly the need for them. And the reader outside the trade, the manager, the public servant, the citizen who consumes numbers every morning with the news, can read it as a consumer's kit: the working knowledge of what keeps a figure trustworthy, so that the next confident number to arrive on a desk can be received with the questions a statistician would ask it before it is believed.

A word on the book's discipline, which the reader will meet in every chapter and should meet once up front. This is a companion of plain language: no formula is paraded, no equation stands where a sentence can stand, and

every technique is taught by what it is for rather than by how it is written on a blackboard. The people who appear in these pages are composites: the first-year analyst at the statistics board, the industry team's senior reviewer, the ministry's skeptical director are all invented figures assembled from the trade's ordinary furniture, and no real case, named scandal, dated incident, or identifiable organization appears anywhere in the book. Where the chapters touch the real landscape of a real place, they identify the sorts of institutions the reader can consult directly, and then teach the fetching habit instead of quoting the figures: the reader who wants the count of statisticians employed by the national office, the pass rates of the professional examinations, or the fee schedules of the universities is being taught, deliberately, to fetch those numbers from the institutions that own them, because a book about honest numbers does not borrow them, and a reader being trained to trust figures should start by trusting the primary sources.

A note on the book's devices, so the reader knows what is invented and what is not. Every chapter builds its teaching from invented furniture: the composite analyst followed through a working week, the invented ministry and its invented director, the composite bank team and its composite reviewer. The composite method is deliberate, and it does two jobs at once. It keeps the book free of borrowed text, since a companion series about honest numbers cannot quote figures it has not verified, and a composite week teaches the craft better than a named scandal, because the reader is invited to argue with the composite analyst's judgment rather than to gawk at someone's misfortune. Where the chapters name the real world, they name the kinds of institutions the reader can look up, and stop there; the reader will find no dated case, no identifiable firm, no quoted statistic anywhere in these pages, and the absence is the book's own practice of what it preaches. The reader who wants real numbers is sent to the owners of the numbers, on the theory that the first honest number anyone fetches should be fetched by the person who intends to use it.

What the book promises, stated once so the reader can hold it to account. By the final chapter, the student will be able to read a job posting in this trade and hear the question beneath it; the analyst will be able to price an error honestly in a room that would rather not hear it; the data scientist will be able to say which parts of a fitted model the machine vouches for and which parts remain on the practitioner's own account; and the citizen will be able to take a confident figure from a news feed and ask it the four questions that keep numbers honest, in under a minute, in plain language. The promise is modest

by design: no reader becomes a statistician by reading a companion, but every reader can acquire the trade's reflexes, and the reflexes are the profession's real gift to the public, since a population that asks the four questions receives better numbers than a population that applauds or boos. The chapters are the reflexes, arranged in the order a career meets them, and the reader who wants the whole path in one sentence already has it: the keeper keeps the signal, prices the error, keeps the record, and stands behind the number in the room where it is used.

The volume belongs to a series of companions, one craft to a book, and the reader who has walked others in the series will recognize the shared spine: a trade defined against its caricatures, its ground mapped, its road walked from study to practice, its craft taught as habit rather than as trivia, its machinery audited honestly, and its long game taken seriously. This book keeps the spine and changes the instrument. Where the companion volumes describe professions whose product is a thing or a service, this one describes a profession whose product is a claim about the world, which is why the chapters circle one question the way the others circle a workflow: what has to be true about a person's habits for a number to survive the journey from a raw reading to a public decision with its honesty intact. The reader can enter here without the others, and the reader who continues through the series will find this volume's habits travel: the four questions of the honest number are, in different clothes, the working questions of every profession this series has kept.

How to use the companion, in one practical paragraph. The reader dipping by need should start at the chapter that names the week's actual problem, the question in chapter four, the error in chapter five, the conversation in chapter seven, since each chapter is built to stand alone, and the composite analyst's week reappears often enough that a dip lands in familiar furniture. The reader walking straight through should let the chapters argue with each other, because they were written to: the record chapter defends the habit the conversation chapter strains against, the machine chapter audits the tools the method chapter uses, and the practitioner's path at the end revises the path mapped at the start. Whichever way the book is read, the four questions should travel with it, and the reader is invited to test every example against them in the margin, including the composite analyst's own decisions, which are not always right on the first telling, and are argued with in the later chapters on purpose.

A last word on how the chapters behave, since the reader will notice the pattern inside the first three: every chapter ends its teaching in the same invented working week, the composite week, because a definition that never touches a Tuesday is a definition this book does not trust itself to have written, and the reader who wants the book's method in one sentence has it here: concept, then cost, then week, which is to say what the thing is, what it takes from a working life, and how it behaves on an ordinary Tuesday at an ordinary desk in the composite rooms this book has built. The chapters can be read in order or raided singly, and the reader who raids is asked to keep one habit from the introduction: whenever a number arrives in these pages, ask it the four questions, provenance, method, error, revisability, and if the reader can answer them, the book has done its work whether it was read whole or in pieces.

One definition, then, carried forward through everything that follows. The statistician is the signal's keeper: the practitioner who receives a question, chooses the reading that can answer it, prices the error honestly, keeps the record that lets a stranger check the work, and stands in the room where the number is used, saying plainly what the number knows and what it does not. Nothing in that sentence requires prophecy, and nothing in it can be pressed into a button. The chapters that follow treat the keeping as a craft: examinable, teachable, schedulable, and daily, practiced by ordinary people in ordinary offices who would not recognize either caricature as their reflection, and who hold, between the mystery and the machinery, one of the quiet load-bearing professions a modern state runs on.

Chapter 1: The Definition

The soothsayer caricature deserves the first inspection, because it is the older, the more romantic, and the more quietly destructive of the two. In the caricature's theater, the statistician is a seer: the data is an oracle's smoke, the model is a trance, and the deliverable is a pronouncement delivered from a stage the audience cannot approach. The caricature flatters everyone it touches. It flatters the analyst, who would rather be mysterious than accountable; it flatters the vendor, whose software sells better as a magic lantern than as a filing system; it flatters the journalist, whose headline reads better over a prophecy than over an interval; and it flatters the decision-maker, who would rather receive a verdict than an argument, since a verdict can be delegated and an argument must be owned. But the corrosion follows the flattery in a straight line. A trade understood as divination cannot be checked, because smoke has no method section; it cannot be taught, because a trance cannot be scheduled into a syllabus; it cannot be corrected, because a prophecy that fails was, by the oracle's own rules, never really wrong. Worst of all, the caricature teaches the number's consumer to be a believer or a cynic, and both stances abandon the one posture that makes numbers useful, which is the reader who asks where the figure came from, what was measured, and what the error costs. The seer's stage is a bad place to keep anything, and the first move of an honest statistician is to come down off it.

The automaton caricature is younger, mechanically minded, and wears the costume of common sense. In its theater, the statistician is a button-presser: the software loads the data, fits the model, tests the difference, and prints the conclusion, and the person between the chair and the keyboard is a toll the automation has not yet finished removing. The caricature is plausible precisely because the automation is real; the software genuinely does, in seconds, work that once took rooms of clerks and a month of arithmetic; and the honest response to the plausibility is not denial but precision. The machine answers; it does not know what the question was. The machine fits; it does not know what was measured, or whether the instrument was calibrated, or whether the sample walked through the door with the population's shape intact or conveniently altered by the very process that gathered it. The machine reports a difference; it does not know whether the difference matters, to whom, or against what decision. Every one of those judgments moved when the arithmetic was automated, and none of them vanished: the judgment about what to measure now hides inside the data's provenance, the judgment about what to compare hides inside the specification, the judgment about what to

report hides inside the default that came out of the box. A trade that pretends its remaining work is button-pressing will be managed, budgeted, and eventually replaced as a trade of button-pressers, and the pretense is the only part of that sequence that is false.

The working reality between the caricatures is best entered, as most working realities are, through a single ordinary week of one honest analyst, and this book will use one invented analyst as its composite guide: a statistician a few years into the trade, working in the analyst room of a large public body, and referred to throughout simply as the analyst, since every figure in this book is a composite and no real person is described. Monday morning, a question arrives from a director: the public wants to know whether the new service is working, and the analyst spends the day on the question's phrasing, because the sentence as it arrived cannot be answered by any data that exists, and the trade's first craft is finding the answerable question the director actually needs answered, then negotiating the difference between the two in plain words. Tuesday is the reading: the data is fetched, and the morning goes to its flaws, the missing months, the duplicate records, the definition that changed midway through the collection, because the analyst knows that the week's honesty is decided here, at the intake, before any model runs. Wednesday is the estimate and its error: the figure is computed, and then, in the half of the day that separates the statistician from the button-presser, the uncertainty is computed with it, and the two are glued together so firmly that no later slide can separate them. Thursday is the record: the analysis is scripted from intake to output, so that a stranger, or the same analyst three years older, could re-walk the whole road. Friday is the conversation: the director receives the figure, its interval, its caveats, and its comparison, asks the skeptical questions the analyst has already asked privately, and makes a decision the number can actually support. Nothing in the week was oracular. Nothing in the week was automatic. The week's product is one honest number, carried to the room where it was needed, and the person who carried it is not a seer and not a toll booth but a keeper.

Defining the product sharpens the definition of the trade. The statistician's product is the honest number: a figure with four properties that no other profession reliably assembles in one place. It has provenance, meaning the reader can trace where the reading came from, what was counted, and under what definitions; the dishonest number, by contrast, arrives naked, or dressed only in confidence. It has method, meaning the road from reading to figure is published, examinable, and open to a rival's re-walking; the dishonest number

offers only the figure, and calls the road a technicality. It has error, meaning the uncertainty is not an embarrassment appended in small type but a part of the number's meaning, glued to it, priced, and carried into every use; the bare figure is the dishonest number's favorite costume, since a point estimate without its interval looks decisive precisely where it is ignorant. And it has revisability, meaning the honest number's owner will publish the correction when the error is found, through the same door the original left, with the correction's own provenance attached; the dishonest number never apologizes, and a trade that never corrects is not confident but unequipped for truth. The keeper's craft is the assembly and defense of those four properties, and the rest of this book is the craft's curriculum: the chapters that follow are, in order, the places the craft is practiced, the road into it, and the habits that hold it together.

The soothsayer has a modern wardrobe, and the reader should learn to recognize the costume changes, because the caricature updates faster than the criticism. The contemporary soothsayer sells insight rather than prophecy: the dashboard is presented as a crystal ball, the word predictive does the work the word mystical used to do, and the deliverable is styled as a revelation, the hidden pattern, the thing the data secretly knows, delivered to a room invited to marvel rather than to check. The costume reaches its highest stage in the press release and the keynote, where the interval has been quietly discarded because the interval does not photograph well, and the finding has been inflated from an association measured in a sample to a law of human behavior. None of this is a caricature of statistics; it is a caricature wearing statistics as a coat, and the working trade pays the caricature's bills, since every disillusioned consumer of an overclaimed insight arrives at the next honest number already believing less. The practitioner's defense is not embarrassment but contrast: the honest analyst's deliverable deliberately photographs badly, arriving with its error, its definitions, and its list of things it cannot say, and the contrast, maintained over years, is the only thing that separates the keeper from the seer in the public's eye.

The automaton has a modern wardrobe too, and it is more dangerous because it flatters the competent. The contemporary automaton is not a clerk but a pipeline: the data arrives through a connector, the model retrains itself on a schedule, the scores flow to the decision system, and the whole road runs without a human touching it, which is presented as maturity and is often the opposite. An unattended pipeline is a set of decisions nobody is currently making: the definitions inside the connector were chosen once, by someone,

and are now enforced forever; the defaults inside the fitting were chosen by a software engineer for a generic user and now decide this organization's thresholds; the drift between the world the model was trained on and the world it now scores is nobody's standing assignment. The honest statistician inside such an organization is not the person who fits the model, since the pipeline does that, but the person who owns the questions the pipeline's builders never asked: what was the reading, has the population shifted, does the error cost the same this quarter as last, and who corrects the pipeline when the road turns out wrong. The machine era did not remove the keeper; it removed the keeper's excuse for being elsewhere while the arithmetic ran.

The definition needs its neighbors, because the trade shares a building with several professions the public blends, and the keeper's own territory is clearest at the boundaries. The data engineer builds the roads the readings travel on, and is not responsible for what the readings mean; the statistician's territory begins where the road ends, at the question of what the reading measures and what may be concluded from it. The data scientist, in the contemporary sense, fits models to predict, and is accountable for the prediction's accuracy; the statistician is accountable for the inference's honesty, the narrower and harder claim about what the data supports, and the two trades increasingly share rooms, tools, and job titles, which makes the boundary habit more rather than less important. The actuary prices risk under contract and under a professional discipline of standards; the statistician working beside an actuary supplies the inference and the design, and the two disciplines check each other in the rooms where they overlap. The economist argues mechanism, the auditor examines after the fact, the manager decides; the statistician keeps the passage from reading to justified claim, and every neighboring trade crosses that passage daily, usually with the statistician's help, occasionally over the statistician's objection, and the objection is part of the job description.

The credential question deserves a direct answer, because the machine era reopened it and the honest answer is neither defensive nor triumphant. What the degree and the professional examinations certify is not the arithmetic, which the software does, but the four properties and the judgment that assembles them: the training to know what a reading measures before running it, to price an error before reporting a figure, to design a comparison that can fail, and to say in public which parts of a claim the data carries and which parts the claimant is carrying alone. The examinations exist because those judgments are examinable, and the credential exists because the judgments

are expensive to learn on the public's time: an untrained analyst produces figures faster than a trained one, and the difference arrives later, in the decisions the figures quietly misinformed. The machine era raised the credential's value for a simple reason the reader can verify in any hiring cycle: the cheaper the fitting becomes, the more figures circulate, and the scarcity has moved from producing numbers to producing numbers that can be defended. The warning runs the other way too, and the trade polices itself with it: a credential without the habits is a laminated caricature, and the reader carrying one should expect the room, eventually, to notice.

The definition ends with the keeper's ordinary virtues, which no syllabus lists and every working room relies on. The keeper is patient with the unglamorous hours, because the trade's honesty is decided at the intake, in the duplicate records and the shifted definitions, long before any model runs; the keeper is comfortable being wrong in public, because a professional whose estimates never miss has either done nothing worth estimating or is hiding the misses, and the willingness to publish the correction is the trade's visible conscience; the keeper takes a quiet pleasure in the residual, the part of the world the model did not explain, because the residual is where the next question lives, and the profession's best work has always started there. The reader who wants prophecy should not enter this trade, and the reader who wants buttons should not either; the reader who wants to be the person a room trusts when the room must know what a figure actually says, and who will accept that this trust is built slowly, defended publicly, and occasionally resented, has the temperament the chapters ahead are written for.

The definition earns its keep as a daily practice, and one composite teaching case shows the whole craft in miniature. A press officer arrives at the analyst's desk with a sentence that has already been drafted for the public: satisfaction with the new service is up eight points, and the sentence needs the analyst's blessing before it travels. The analyst asks the four questions the blessing requires, and the morning becomes a walk. Where did the reading come from: a survey of users, conducted by the service's own front desk, to which the happy were likelier to respond, and the eight points ride on that skewed frame. How was the road walked: the same instrument as last year, but the scale changed midway, and a point on the new scale is not a point on the old one. What does the error cost: the sample is small enough that the eight points sit inside the band the survey's own noise occupies, and the honest finding is a cannot-tell, which the drafted sentence converts into a trend. Will the owner correct it: the press officer offers to soften the verb instead of

correcting the claim, and the analyst declines, because softened is still wrong, and the sentence the analyst will bless says what the reading actually supports, in the officer's own register, without the points that are not there. The trade's whole definition is in that morning, and the reader who can walk the four questions across a drafted sentence has already practiced the craft further than the reader who can recite the definitions.

The keeper's lineage is worth one paragraph, because the definition this chapter settles is older than the software that now surrounds it. The trade's oldest photograph is the census: the state counting itself, on a fixed night, with a method, because a country that cannot count its households cannot govern its households, and the counter's discipline, the definitions, the schedules, the published road, is this chapter's honest number in its earliest uniform. The trade's second uniform came from the factories: the quality engineers who read a production line's drift from small samples and kept the machines honest, and whose control charts were the signal's keeping drawn in ink. The third came from the experimenters: the designers who learned that a comparison is only as honest as its randomization, and that the most expensive number in any institution is the one from the study that could not fail. The uniforms changed, the arithmetic industrialized, the instruments migrated from counting houses to cloud platforms, and the product never moved: an estimate with its error priced, delivered to a decision, by a person who will answer for it. The lineage is the credential's real content, and the reader entering the trade is joining a service that every modern state and most modern firms quietly runs on, and has since before the machines that now carry the arithmetic.

The definition has public moments, and the keeper meets them in the news before ever meeting them at work. The inflation reading argued about at the market; the polling night's margins narrated on the broadcast; the health study that contradicts last year's health study; the weather's confidence bands drawn on the national map: these are the trade's product performing in public, and the definition explains both the performance and its miseries. The honest number in public carries its error openly, which reads, to a public trained on verdicts, like weakness, and the keeper's public voice exists to translate: the interval is not uncertainty about whether to act, the correction is not a scandal but the system working, the margin of error is not a license to dismiss the reading but a fence telling the reader how far the reading may be trusted. The reader who has internalized this chapter watches those moments differently, hears the soothsayer's costume on the pundit and the automaton's shrug in

the spokesperson, and reserves respect for the figure that arrives dressed as this chapter has dressed it: provenance, method, error, and the standing promise of the correction. The public moments are also the trade's recruiting ground, since half the profession's future members first fell in love with the craft from the outside, watching a number behave honestly in an argument and wanting to be the person who kept it that way.

The temperament question belongs in the definition because readers choose trades with their hearts as much as their marks, and the keeper's temperament is specific enough to describe: a curiosity that survives cleaning, meaning the analyst who still wants the answer after the third hour of the pantry's mess, since most of the trade's hours are spent there and the glamour of the question is what the recruitment posters show; a comfort with being wrong in company, since the band guarantees public error at the claimed rate and the keeper who needs to be right always will corrupt the interval to buy peace; a tolerance for the unfinished, since every honest number is provisional by design and the temperament that requires finality will manufacture it; and the quiet pleasure of small certainties, the explained residual, the found leak, the table that finally balances, which is the trade's actual daily joy and never makes the posters either. The trade's honest advice to the reader who lacks these and knows it is neither harsh nor flattering: the data estate has many rooms, and the rooms nearest this one, the engineering, the product, the communication of numbers, are honourable and pay better, and the reader who wants the keeper's exact temperament knows it within the first working year.

The definition settles, as good definitions should, into a test the reader can apply on any working day. When a figure arrives at a decision, the four questions follow it: where did the reading come from, how was the road walked, what does the error cost, and will the owner correct it when the road turns out wrong. A statistician is a professional whose training exists to answer those four questions before anyone else has to ask them, and the credential's value in the machine era reduces to exactly that: the analyst who cannot answer them is running the software without the supervision the software assumes, and the analyst who can is the one the machine era has actually made more necessary, not less, since the cheaper the arithmetic becomes, the more figures are produced, and the more figures are produced, the scarcer the honest ones become. The keeper, in one sentence: the practitioner who keeps the signal distinguishable from the noise, and keeps the number distinguishable from the wish, all the way from the question to the decision,

and who will say, in the room where the number is used, the sentence this book will keep returning to: this is what the figure knows, this is what it does not, and this is what it would cost to know more.

ABOUT THE AUTHOR

The question this series keeps asking is what a profession actually protects, and for the statistician the answer is the signal itself: the difference between a pattern and a coincidence. Aiman JoeHanz built the Building Tomorrow's People volumes around that single organising idea, one trade at a time, and the twenty-fifth of them belongs to the people who decide how much of the world gets to count. The book takes the sample, the confidence interval, the machine that now proposes its own model, and the honest errand of stating uncertainty in a language a decision maker will act on.