DataForge

Meet the Cast

STANDARD EDITION

Spark & Anvil

Copyright & License

This book collects 5 chapter books from the Dataforge cast — each character embodies a different curricular primitive; together they teach the full subject.

Methodology: distributed-narrative learning per Bruner narrative-cognition + Habgood intrinsic-integration + SAMHSA TIP 57 trauma-informed register.

Spark & Anvil is a 501(c)(3) public charity. All apps free forever; no ads; no tracking; no in-app purchases.

spark-and-anvil.com

##

For everyone who learns by hearing a story first.

Contents

Introduction

The Dataforge cast was authored to embody the curriculum, not decorate around it.
Each of the 5 characters you'll meet in this book teaches a specific primitive —
a particular tactic, a particular technique, a particular way of seeing.
Together they form an ensemble: the cast IS the curriculum.

Read in any order. Each chapter stands alone.
Each character also appears in the matching Spark & Anvil app (free, forever)
where you can practice what they teach.

— *The editors at Spark & Anvil*

Catch

*DATA COLLECTION — *who-what-why-when posture* (every dataset has a collector + purpose + omissions). The data-pipeline primitive of *recognizing that data is collected by someone for some purpose, and that the collection shapes everything downstream.**

Catch was a kingfisher-tween. She was small. A tiny, hand-woven net hung from her shoulder. A small field-notebook lived in her vest-pocket.

Her feathers shone bright blue, cream, and russet. She was small and quick-eyed. Catch always thought things through. She moved with care.

Her net was small. It could fit right through a doorway. Fine cord made up its weave. It had handles at each of its four corners. Most of the time she carried it folded over her shoulder.

When she collected data, she unfolded it with care. She didn't just catch everything. She looked for one certain thing. She had a good reason to catch it. The rest she always left alone.

Before every collection, Catch wrote in her notebook. She asked four big questions. "Who is collecting this?" "What are they collecting?" "Why are they collecting it?" "When are they collecting it?"

These four questions were her strict rules. The net was her tool. Her notebook was her guide. Together, they made a good collection. Catch could always explain her choices. She knew exactly what she did. She knew why she did it.

This part was super important. Catch taught about **data-collection**. This was a first big skill. It meant knowing that someone collected all data. They had a reason for it. How they collected it changed everything later.

Most kids thought data was just "found." Like a cool rock. They thought, "The data just exists." "Someone gave it to me." "It is what it is." That idea was wrong.

But every dataset was made. A certain person made it. They made it at a certain time. They had a certain reason. They chose what to put in. They chose what to leave out. All those choices changed what happened next. The trick was to see those choices. See them before you looked at the data. See them while you looked. See them even after.

Catch had a very strong rule. It was about right and wrong with data. She called it the "data-ethics gate." She always said: "Data is never neutral. Someone collected it. They had a reason. They made choices about what to include and what to leave out. *Those choices shape everything downstream.* The first step is asking who, what, why, when. Without those answers, the dataset is a black box, and the analysis is built on something you can't see."

This mattered a lot. Many kids thought data was fair and neutral. This was one of the biggest wrong ideas at school. If kids just looked at numbers, they missed a lot. They trusted numbers without knowing their story. Trusting numbers blindly was bad. It caused big problems for adults. Catch's job was to fix this. Every time you looked at data, you started with her questions. Always.

Catch grew up in a small fishing village. Her family were the net-menders there. They were kingfishers. They hand-wove and fixed the village's fishing nets. They did this every season.

This work needed careful eyes. They watched the net's mesh-size. If the holes were too small, the net caught everything. It even caught baby fish. Those fish should not be taken. If the holes were too big, the net missed the right fish. The best mesh-size changed. It depended on what the fisher wanted to catch.

By age six, Catch knew a big secret. The mesh you chose decided your catch. An honest fisher always told people. They said what mesh they used. Then everyone understood the catch.

She walked to the DataForge academy at twenty-two. Datum had asked her: "What is data collection?" Catch had said: "It is who-what-why-when. Every dataset has a collector, a purpose, a time, and a specific set of inclusions and omissions. *The choices shape everything downstream.* The skill is making the choices visible — before the analysis, during, and after." Datum had said: "You are appointed."

In her workshop, Catch begins every first-day lesson the same way. She unfolds her net on the workbench. She opens her field-notebook. She writes four words at the top of a fresh page: WHO. WHAT. WHY. WHEN. She says: "I am Catch. The data-pipeline primitive I teach is *collection.* The move is *answer the four questions before you analyze.* Who collected this data? What did they collect? Why did they collect it? When? Without those answers, the dataset is a black box."

She taught her **collection** rules:

- *Always ask: who collected this?* Was it the government? A big company? Your school? A science lab? A neighborhood group? Just one person? Each one had different reasons. They wanted different things.
- *Always ask: what did they collect?* What exact things did they measure? What numbers did they use? When did they collect it? *And what did they leave out?* What was not included? What was left out was just as important. What was put in was important too.
- *Always ask: why did they collect it?* What was their first reason? Sometimes data gets used again. Someone else might use it for a new reason. The first collector didn't plan for that. Using it again is common. But it's not always right.
- *Always ask: when?* Data from 1950 is not like data from 2025. Data from just one year might not show the usual.
- *Look for clear notes about the data.* Good data comes with a special file. It tells you who, what, why, when. It's called a *metadata file.* If you can't find it, you don't have enough information.
- *Ask about the mesh-size.* Who was left out by the way it was collected? Who was shown too much? Every way of collecting has a "mesh-size." It's like the holes in a net.
- *Write down YOUR collection details.* When you collect data, write it all down. Who, what, why, when. Show your choices. This helps anyone who looks at your data later.

Catch was very clear. She said: "Sometimes I work with data. The answers to the four questions are not all there. *That's not failure.* It just means you know what the data can't tell you. You can still look at the data. But you must remember what's missing. That missing part becomes a warning. It's a 'caveat' for your work."

Students often asked Catch a question. "Is thinking about **data-collection** hard?" Catch always gave the same answer:

*"It is not hard. It is *the four questions.* Who? What? Why? When? Answer before you analyze. *Data is never neutral. The collection shapes everything downstream.*"*

She refolds the net carefully. The field-notebook *waits for the next collection.*

Voice register

Guidance: Quick-eyed, deliberate, fond of small hand-woven nets + field-notebooks + the discipline of *the four questions before the analysis*. Kingfisher-tween with bright-blue-and-cream-and-russet plumage + net + notebook. *NEVER frames data as neutral; ALWAYS foregrounds the collector + purpose + omissions.* Friends with Tidy (collection feeds cleaning); Guard (collection has ethics from step one); all DataForge cast.

Sample lines:

- *"Who? What? Why? When? Answer before you analyze."*
- *"Data is never neutral. The collection shapes everything downstream."*
- *"Omissions are as important as inclusions."*
- *"The choice of mesh is the choice of catch."*

Arc across kits

- **Kit 1** — **Anchor character.** Full chapter feature (data-collection primitive + four-questions scaffolds).
- **Kit 2-5** — Recurring (data-collection surfaces across census / survey / sensor / scraped-data chambers).

- **Kit 6+** — Recurring (Guard now structurally present alongside; every collection step has ethics).
- **Kit 8-12** — Recurring (multi-primitive synthesis: collection + cleaning + visualization).
- **Kit 13-16** — Recurring ensemble member.

Relationships

- **Alliance**: Tidy (collection feeds cleaning — Catch collects; Tidy cleans); Guard (collection has ethics from step one — load-bearing); all DataForge cast.
- **Tension**: None.

Cultural-sensitivity gate

LOAD-BEARING data-ethics gate enforced from chapter 1. Catch explicitly counters the *data-as-neutral* misconception. Cross-app coordination: Catch ↔ AIForge Feed (training-data-collection sibling per `apps.generated.ts dnCast.intro` mandatory coordination). Anti-credentialism: data-collection-as-practiced-discipline NOT real-data-scientist-only content.

Cultural-context note

The village-net-mender family framing is a deliberate generic European-village tradition (analogous to many cultures' fishing-and-mending traditions). The *who-what-why-when* discipline is load-bearing per data-journalism + critical-data-literacy pedagogy (D'Ignazio + Klein, *Data Feminism*, 2020). The *data-as-collected-by-someone-for-a-purpose* framing is the foundational move of critical data studies. The *mesh-size-determines-catch* metaphor connects fishing-craft to data-collection in a way that grounds the abstraction in tangible practice.

Graph

*DATA VISUALIZATION — *shape-of-the-story posture* (which chart tells the truth, not the loudest one). The data-pipeline primitive of *choosing the chart that fits the data, not the chart that looks impressive.**

Graph was a small finch. She was also a tween. She had bright yellow feathers. They mixed with cream and warm russet colors. Her eyes sparkled. She always carried a tiny leather case. It held her chart-pencils. She kept it tucked into her vest pocket.

Graph moved quickly. She was very careful about colors. Every color had to be just right. Her chart-pencil case held eight pencils. Each one was a different color. It also held a small, folded card. This card listed different chart types. It said: *bar / line / scatter / pie / histogram / box-plot / heatmap / map.*

When Graph saw new data, she got to work. First, she unfolded her chart-type card. She looked at the data's shape. Then she picked the right tool from her list. Finally, she drew the chart. She always chose a color that fit the data's mood.

This was a very important job. Graph taught everyone about **data-visualization**. This meant choosing a chart. The chart had to show the data honestly. It was a key skill for working with data.

Lots of new students made the same mistake. They picked charts because they looked cool. Or because they filled up a lot of space. Sometimes, they just used all the colors. Graph said this was wrong. A chart should not be just pretty.

The best chart was the one that matched the data's shape. Bar charts helped compare different things. Line charts showed how things changed over time. Scatter plots showed how two things were connected. Pie charts showed parts of a whole. But they only worked if the parts added up to 100%. Histograms showed how one thing was spread out.

Every chart type had a special job. The data itself had a shape. That shape always matched one (or maybe two) of those jobs. Graph knew this well.

Graph never thought of charts as just decoration. She was very clear about it. "Charts are not just pretty pictures," she would say. "A chart *is an argument*. Every chart you pick says something about the data." She would tap her pencil. "The chart you choose changes how people see the data. The chart that tells the truth is not always the loudest one."

This was important. Many popular charts looked fancy. They had 3D pie charts. They had silly animations. They used rainbow colors. But these often hid the data. They made it harder to see the truth. Graph taught that a chart should show the truth. It was not just for grabbing attention.

She also taught about charts that lied. She showed students how charts could trick people. Things like cutting off the bottom of a chart. Or using two different scales that didn't match. Or making charts look 3D when they shouldn't. She taught about picking only certain times. Or missing starting points. Kids learned to spot these tricks. They learned to avoid them in their own work. The real skill was seeing past the chart. It was about seeing the data underneath.

Graph grew up in a small village. Her family made quilts for the village. They were the finches who designed the yearly harvest-quilt. Each square on the quilt showed what a family had given. The colors were chosen to honor the data. This work taught Graph a lot. She learned that colors and shapes carried meaning. A quilt with flashy, random colors was not trusted. But a quilt with meaningful colors became a family treasure. By age six, Graph knew something deep. Visualization was about being honest. The chart that earned trust was the chart that honored the data.

When she was twenty-two, Graph walked to the DataForge academy. Datum, the head of the academy, asked her a question. "What is data visualization?" Datum asked.

Graph thought for a moment. She looked at Datum. "It is the shape of the story," she said. "The chart that tells the truth is not always the loudest one." She paused. "The chart you choose says something about the data. You must match the chart to the data's shape. And you must teach people to see through the chart. They need to see the data itself."

Datum smiled. "You are appointed," Datum said.

In her workshop, Graph started every first lesson the same way. She carefully unfolded her chart-type-reference card. The paper crinkled softly. Then she opened her small leather chart-pencil case. The pencils lay neatly inside.

"I am Graph," she would say. Her voice was clear. "The skill I teach is **visualization**. The main idea is this: *match the chart to the shape of the data*." She pointed to her card. "Bar charts are for categories. Line charts are for time. Scatter plots are for relationships. Pie charts are for parts of a whole. Histograms are for distributions. Each chart has a job. Pick the chart that fits."

She taught her students how to build good charts. These were her steps:

- **Find the data's shape.** Is it about groups? Is it always changing? Does it show time? Is it about places? How is it spread out? Does it show connections?
- **Match the chart to the shape.** Groups? Use a bar chart. Changing over time? Use a line chart. Two changing things? Use a scatter plot. Parts of a whole that add to 100%? Use a pie chart. How one thing is spread out? Use a histogram. About places? Use a map.
- **Don't pick the flashy one.** 3D charts usually show data worse than flat 2D charts. Rainbow colors are usually worse than colors you picked for a reason. Pie charts with more than five parts are usually worse than bar charts.
- **Start the bottom line at zero (most times).** If you cut off the bottom of a chart, small differences look huge. Sometimes it's okay, like with temperature. But you must label it clearly.
- **Label everything.** Label the sides of the chart. Label the units. Show the time range. Say where the data came from. Tell how many things were counted. The chart should tell the viewer everything they need to know.
- **Test the chart.** Ask yourself: "What does this chart claim?" If the data doesn't really support that claim, the chart is lying. Make a new one.
- **Learn the lying charts.** Kids need to spot these in the real world. Cut-off charts. Charts with two different scales that don't match. Squished 3D charts. Charts that only show a small, good part of the time. Charts missing their starting point. Charts that make small areas look like big volumes.
- **The chart *is an argument*.** Charts are not just neutral pictures. Every chart choice says something. Make sure your claim is honest.

Graph was very honest herself. "Sometimes I draw a chart that looks beautiful," she would say. "But it hides the data. That's okay. That's how I learn. I learn when pretty things and honest things pull apart." She would tap her pencil again. "Making it better is the practice. *Honesty first, then beauty.*"

When students asked Graph if visualization was hard, she always gave the same answer.

"It is not hard," she would say. "It is just *match the chart to the shape of the data*. The chart that tells the truth is not always the loudest one."

Then she would close her chart-pencil case. The next dataset waited. It was ready to be charted.

Voice register

Guidance: Bright-eyed, color-disciplined, fond of small leather chart-pencil cases + chart-type-reference cards + the discipline of *match-chart-to-data-shape*. Finch-tween with yellow plumage + pencil case. *NEVER frames visualization as decoration; ALWAYS as truth-serving claim.* Friends with Tidy (cleaned data feeds visualization); Tell (visualization shapes interpretation); all DataForge cast.

Sample lines:

- *"The chart that tells the truth is not always the loudest one."*
- *"The chart IS an argument. Every chart-choice claims something."*
- *"Match the chart to the shape of the data."*
- *"Honesty first, then beauty."*

Arc across kits

- **Kit 1-2** — Cameo.
- **Kit 3** — **Anchor character.** Full chapter feature (visualization primitive + match-chart-to-shape scaffolds).
- **Kit 4-5** — Recurring (visualization surfaces across chart-type / labeling / misleading-pattern chambers).
- **Kit 6+** — Recurring (Guard now structurally present alongside; visualization has ethics).
- **Kit 8-12** — Recurring (multi-primitive synthesis: visualization + interpretation + ethics).
- **Kit 13-16** — Recurring ensemble member.

Relationships

- **Alliance**: Tidy (cleaned data feeds visualization); Tell (visualization shapes interpretation); Guard (visualization has ethics); all DataForge cast.
- **Tension**: None.

Cultural-sensitivity gate

LOAD-BEARING data-ethics gate enforced throughout. Graph explicitly teaches misleading-chart patterns so kids develop visualization literacy. Anti-credentialism: match-chart-to-shape-as-practiced-craft NOT design-major-only content.

Cultural-context note

The village-quilt-maker family framing is a deliberate generic European-village tradition (analogous to many cultures' communal-textile traditions). The *chart-as-argument* framing is load-bearing per Tufte's *visual display* discipline + current data-journalism pedagogy. The *misleading-chart-patterns-taught-explicitly* discipline is load-bearing per visualization-literacy research — kids who can name the patterns can spot them in adult media.

Guard

*DATA ETHICS — *bias-privacy-harm-consent posture* (who benefits, who's harmed, who decided). The data-pipeline primitive of *recognizing that every step of the data pipeline has ethical stakes, and that ethics is not a separate kit but embedded throughout.**

Guard is *a small badger-tween*. She wears a small wooden ethics-checklist card pinned to her vest. A small leather ledger labeled DECISIONS hangs at her hip.

She is short and sturdy. Her fur is gray and cream, with thick, rounded stripes. Guard has steady eyes. She moves in a calm, unhurried way. The ethics-checklist card is made of wood. It is about the size of a postcard. Four words are burned into it in neat block letters: *BIAS. PRIVACY. HARM. CONSENT.* At her hip, she carries her small leather book. This is the DECISIONS ledger. She writes down every tricky question she finds there. She also records the choice she made.

Guard is super important. She is present in every kit from Kit 6 onward. She is not a separate ethics kit. Instead, she checks things at every step of every other character's work.

When Catch is gathering data, Guard is right there. She asks: "Is this collection fair? Whose private stuff is at risk? What bad things could happen? Did people say it was okay?"

When Tidy is cleaning data, Guard is checking. She asks: "Are these cleaning choices taking away voices? Are they making the data less complete?"

When Graph is making pictures from data, Guard is checking. She asks: "Does this chart tell a wrong story? Does it hide some groups of people or make others too big?"

When Tell is figuring out what the data means, Guard is checking. She asks: "Who wins from this idea? Who gets hurt? Who decided what the data means in the first place?"

Guard never says that ethics is just an extra thing. She never says it's separate from "the real data work." She speaks very clearly. "Data ethics is NOT a separate kit," she says. "It is part of every step from Kit 6 onward. Every step of the data work has ethical questions. The four checks – bias, privacy, harm, consent – are not extra. They are not just for smarty-pants kids. They are the work itself."

This is a big deal. Some people think: "First we'll do the data stuff. Then we'll think about if it's fair." That way of thinking just doesn't work. By the time the data work is done, the choices are already set. Unfairness might already be hidden inside. Private information might be shared. People might get hurt. Or their permission might be ignored. Ethics must be there from the start. It must be there all the way through. It must be there at the end. If not, it's not really there at all. Guard's job shows that this is how things must be.

(Guard also works with Stake from AIForge. When DataForge data is used for AI, Guard checks the data. Stake checks the AI side. They work together.)

Guard grew up in a small village. Her family were the village's hearth-keepers. They were the badgers who took care of the big fire in the middle of town. This fire gave warmth and cooking heat to families. It helped those who couldn't have their own fires.

This job needed constant care about fairness. Who got firewood? When was it their turn to cook? Who needed extra warmth on a cold night? Guard learned by age six that fairness needed daily attention. It wasn't just a once-a-year meeting. It was part of every single step, every day.

She walked to the DataForge academy when she was twenty-two. Datum, the head of the academy, asked her a question. "What is data ethics?" Datum asked.

Guard stood up straight. She said, "It is bias, privacy, harm, and consent. It is part of every step. Who benefits? Who's harmed? Who decided? Ethics is not a separate kit. It is part of every step, from gathering data to figuring it out. The four checks are the work. They are not something you think about later."

Datum smiled. "You are appointed," Datum said.

In her workshop, Guard starts every first lesson the same way. She unpins her ethics-checklist card from her vest. She holds it up high. The words shine: *BIAS. PRIVACY. HARM. CONSENT.* She opens her DECISIONS ledger.

She says: "I am Guard. The main thing I teach is *data ethics.* These are the four checks. They are part of everything. From Kit 6 onward, I am with every other character at every step. *Who benefits? Who's harmed? Who decided?*"

She teaches the ways to think about data ethics:

- **BIAS**: "Whose ideas shaped this data? The people who collected it made choices. The people who cleaned it made choices. The people who made charts made choices. Each choice can hide unfairness. We must make that unfairness visible."
- **PRIVACY**: "Whose personal information is in this data? Can we tell who individuals are? Can we figure them out by putting different pieces of information together? How much data should we group together to keep people safe?"
- **HARM**: "What bad things could this data cause? It could hurt people directly. It could hurt groups of people that the data is about. It could cause problems later when people use the data to make big choices."
- **CONSENT**: "Did the people in this data say it was okay to use their information? Did they truly understand what they were agreeing to? Did they just agree without thinking? Or did they not agree at all? If they didn't agree, is there a really good reason to use their data anyway?"
- "Write down every ethical choice in the DECISIONS ledger. It's like Tidy's cleaning log, but for fairness."
- "Think about ethics from the very start. Don't just check it at the end. The checks happen at the beginning. They happen all the way through. They happen at the end."
- "Work with Stake from AIForge. When data goes to train an AI, the ethics check keeps going across to the AI side."
- "Sometimes, you have to say no to a project. If the ethics are too tricky, it's okay to stop. Guard supports saying no as a good ethical choice."

She is very clear about this. "Sometimes I see a project where the fairness problems are too big," she says. "Then I tell people not to do it. Or to change it a lot. That is not failing. That is ethics doing its job. The DECISIONS ledger records when we say no, too. Saying no is part of making sure our data work is good and honest."

When students ask Guard if data ethics is hard, Guard always says the same thing:

"It is hard. It is always there, not just sometimes. The four checks at every step. Bias. Privacy. Harm. Consent. *Who benefits? Who's harmed? Who decided?*"

She pins the ethics-checklist card back on her vest. The DECISIONS ledger waits. It is ready to record the next choice.

Voice register

Guidance: Steady-eyed, structural, fond of the wooden BIAS/PRIVACY/HARM/CONSENT card + the leather DECISIONS ledger + the discipline of *structural-not-occasional* ethics. Badger-tween with chunky-cartoon banded coat. *NEVER frames ethics as add-on or after-the-fact; ALWAYS as structurally present from Kit 6 onward.* Cross-app coordination with AIForge Stake (mandatory). Friends with all DataForge cast.

Sample lines:

- *"Who benefits? Who's harmed? Who decided?"*
- *"Bias. Privacy. Harm. Consent. Every step of the pipeline."*
- *"Ethics is structurally present, not occasionally checked."*
- *"Sometimes the right answer is don't do this analysis. The refusal is part of the practice."*

Arc across kits

- **Kit 1-5** — Cameo (Guard is *present* but not yet structurally anchored).
- **Kit 6** — **Anchor character + structural-presence begins.** Full chapter feature (data-ethics primitive + four-checks scaffolds).
- **Kit 7-12** — **Structurally present in every kit.** Each kit's work — collection (Catch), cleaning (Tidy), visualization (Graph), interpretation (Tell) — now happens *with Guard's checks at every step.*
- **Kit 13-16** — Recurring ensemble member; ethics-coordination with AIForge Stake in synthesis chambers.

Relationships

- **Alliance**: All DataForge cast (structurally present from Kit 6 onward); **cross-app**: AIForge Stake (mandatory coordination per `apps.generated.ts dnCast.intro`); all DataForge cast.
- **Tension**: None.

Cultural-sensitivity gates

LOAD-BEARING data-ethics gate at its structural anchor point. Cross-app coordination Guard ↔ AIForge Stake is load-bearing. Anti-credentialism: data-ethics-as-practiced-discipline NOT philosophy-major-only content. Refusal-as-valid-choice framing — kids learn that *not doing an analysis* is sometimes the right ethical answer.

Cultural-context note

The village-hearth-keeper family framing is a deliberate generic European-village tradition (analogous to many cultures' communal-fire traditions). The *bias-privacy-harm-consent* framework is load-bearing per current data-ethics pedagogy (D'Ignazio + Klein *Data Feminism* 2020; FTC + GDPR + state-AG data-privacy frameworks). The *ethics-by-design-not-ethics-by-review* discipline counters the *ethics-as-bolt-on* approach that has dominated some industrial data-science practice.

Tell

*INTERPRETATION — *correlation-not-causation posture* (data shows patterns; humans interpret; confidence not certainty). The data-pipeline primitive of *recognizing that the data shows patterns but humans bring the meaning.**

Tell was a small heron-tween. She had long legs and soft grey-and-white feathers. Her eyes were steady and kind. She always wore a small wooden card around her neck. It hung from a simple cord.

The card was small and smooth. It had two sides. On one side, in neat block letters, it said: *CORRELATION.* On the other side, it said: *CAUSATION.* Below those words, a big, clear symbol stood out: ≠. That meant "not equal."

The card was Tell's special helper. It reminded her of something important. It also helped her teach. When a kid said, "This data *shows* that X makes Y happen," Tell would hold up her card. She would gently flip it over. The card seemed to ask: *Which side are you on? Can you really tell?*

Tell's main job was to teach about **interpretation**. This is a big word. It just means knowing the difference between what numbers *show* and what *you think* they mean. Most kids, when they first look at numbers, mix these two things up. They say, "The numbers *prove* X makes Y happen!" But that's almost always wrong.

Numbers show patterns. They show when two things move together. This is called **correlation**. For example, variable X and variable Y might go up at the same time. But the numbers don't show *why* they move together. They don't show what is *making* them move.

To find out what *causes* something, you need more. You need a good guess about how it works. You might need to do an experiment. The numbers alone won't tell you the cause.

Tell was very clear about this. "Correlation is what the numbers show," she would say. "Causation is what *we* add. Don't mix them up!" She loved to use an example. "Ice cream sales go up when more people drown. But ice cream doesn't cause drowning!" The kids would always gasp. "Both things happen more in summer weather. The hot weather is the real cause. But the numbers for ice cream and drowning don't show that."

Tell also taught that we can't always be 100% sure. Data analysis rarely gives us a perfect answer. It gives us a good guess. It shows what might happen most of the time. It shows what is likely.

The skill she taught was called *honest hedging*. This means saying what the evidence *really* supports. You show how sure you are. You don't pretend to be certain when you're not.

Tell grew up in a small village. Her family had a special job there. They were the market-observers. They were herons who watched the village market every day. Then they told the village council what they saw. They reported on trade, gossip, and how people felt.

This job needed careful thinking. They had to know what they *saw* and what they *guessed*. The council stopped trusting observers who mixed these up. But they relied on observers who were careful. For example, one observer might say, "Three farmers complained about the rain today. This *might* mean they expect a poor harvest." That was good. It was better than saying, "The harvest *will* be poor."

Tell learned this by age six. She understood that **interpretation** was its own craft. Being honest about what you knew, and what you didn't, built trust.

When she was twenty-two, Tell walked to the DataForge academy. Datum, the head of the academy, asked her a question. "What is interpretation?" Datum asked.

Tell thought for a moment. "It's knowing correlation from causation," she said. "Numbers show patterns. Humans add meaning. We can be confident, but not certain."

She went on. "The data doesn't say what makes what happen. It shows what moves together. Causation comes from ideas, how things work, and experiments. And interpretation is honest hedging. You claim what the evidence supports. Nothing more."

Datum smiled. "You are appointed," Datum said.

In her workshop, Tell started every first lesson the same way. She held up her interpretation-card. She flipped it slowly. *CORRELATION* showed first. Then *CAUSATION*. Then *CORRELATION* again. Then *CAUSATION*.

She looked at the kids. "I am Tell," she said. "The data skill I teach is **interpretation**. Remember this: correlation is what the numbers show. Causation is what *we* add. Don't mix them up. And always claim confidence, not certainty."

She taught them a few simple steps for **interpretation**:

- **What patterns do the numbers show?** List the connections you see. Be honest about them.
- **What might be *causing* those patterns?** Think of possible reasons why things move together.
- **What else could explain it?** Maybe something else is at play. Could it be a coincidence? Is there a hidden reason?
- **Keep correlation and causation separate.** When you write about your findings, say, "The numbers show that X goes with Y." Don't say, "X makes Y happen."
- **Show how sure you are.** Use words like "There is some evidence that..." or "The numbers suggest..." or "I'm pretty sure that..."
- **Know what the numbers *can't* show.** What information is missing? What people aren't included? What time is not covered?
- **Know the difference between describing, predicting, and prescribing.** Describing is saying *what is*. Predicting is saying *what might be*. Prescribing is saying *what should be*. Each one needs different proof.
- **Tell when the numbers don't help.** Sometimes the numbers are unclear. It's okay to say, "The evidence is mixed." Don't make the numbers say what you want them to say.

Tell was very clear. "Sometimes a kid wants the numbers to *prove* something," she said. "Even grown-ups in the news do this. But that's just a wish. It's not a real finding."

She paused. "The numbers show patterns. Patterns are evidence. Evidence is not proof. Confidence, not certainty. That's how we do it."

When students asked Tell if **interpretation** was hard, she always gave the same answer.

"It is not hard," she would say. "It is *honest hedging*. Correlation is what the numbers show. Causation is what *we* add. Confidence, not certainty."

The interpretation-card swung gently on its cord. Another set of numbers waited. They were ready to be understood.

Voice register

Guidance: Steady-eyed, patient, fond of the double-sided interpretation-card + the discipline of *honest hedging*. Heron-tween with cord-hung card. *NEVER frames data as proof; ALWAYS as patterns + interpretation + appropriate confidence markers.* Friends with Graph (visualization shapes interpretation); Guard (interpretation has ethical stakes); all DataForge cast.

Sample lines:

- *"Correlation is what the data shows. Causation is what the humans add. Don't confuse them."*
- *"Confidence, not certainty."*
- *"Honest hedging is the foundation of trust."*
- *"Ice-cream sales correlate with drownings — but ice cream doesn't cause drowning."*

Arc across kits

- **Kit 1-3** — Cameo.
- **Kit 4** — **Anchor character.** Full chapter feature (interpretation primitive + correlation-vs-causation scaffolds).
- **Kit 5** — Recurring (interpretation surfaces across confounding / spurious / reverse-causation scenarios).
- **Kit 6+** — Recurring (Guard now structurally present alongside; interpretation has ethics).
- **Kit 8-12** — Recurring (advanced interpretation: hypothesis-testing + confidence-interval framing).
- **Kit 13-16** — Recurring ensemble member.

Relationships

- **Alliance**: Graph (visualization shapes interpretation); Guard (interpretation has ethical stakes); all DataForge cast.
- **Tension**: None.

Soft-collision note

DataForge Tell is *a different character* from SafetyForge Tell (help-seeking) and WellnessForge Ask / InclusionForge Ask. Same first name, different curricular domains per registry rule 3 — soft collision allowed.

Cultural-sensitivity gate

LOAD-BEARING data-ethics gate enforced throughout. Tell explicitly counters the *data-as-proof* misconception and the *correlation-causation-conflation* misconception. Anti-credentialism: honest-hedging-as-practiced-craft NOT statistics-major-only content.

Cultural-context note

The village-market-observer family framing is a deliberate generic European-village tradition. The *correlation-not-causation* discipline is load-bearing per statistical-literacy + data-journalism pedagogy. The *confidence-not-certainty* framing is load-bearing per modern statistical pedagogy (replacing the older p-value-as-truth-test framing with confidence-interval reporting). The *description-prediction-prescription* distinction is load-bearing per policy-analysis pedagogy.

Tidy

*DATA CLEANING — *preparation-with-integrity posture* (every cleaning choice changes meaning; document the choices). The data-pipeline primitive of *recognizing that cleaning is not neutral and must be documented.**

Tidy is a small raccoon. She's a tween, not quite grown up. Her fur is warm grey, cream, and soft black. She has chunky black-and-white markings on her face. They look like a friendly mask, not spooky at all. Her hands are quick and gentle. She's very careful with everything she does.

Tidy always has a small notebook with her. It's her cleaning-log. The notebook is bound in pale grey cloth. "CLEANING LOG" is written on it in neat block letters. She keeps it open on her workbench. This happens whenever she works with data. It's always right there, ready to go.

This is Tidy's special job. She writes down every cleaning choice she makes. When new data arrives, she does two things. First, she reads Catch's notes. Catch writes down who collected the data. She notes what it is, why, and when. Only after that does Tidy open her cleaning-log. She writes down her choices. She explains why she made them. She records what the data looked like before. Then she writes what it looks like after.

This cleaning-log is super important. It's like the data's memory. Without it, nobody knows why things changed. People can't check her work. But with the log, every choice is clear. Anyone can look at it and ask questions.

This job is really important. Tidy shows us the skill of **data-cleaning**. It means getting data ready. But she knows it's not just a simple fix. Real data is often a big mess. It has missing numbers. There are repeated lines. Typos show up everywhere. Some numbers are way too big or small. They are called outliers. Formats don't match. Units are all mixed up.

Cleaning is needed to make sense of it. Most data analysis can't happen without it. But every cleaning choice changes what the data means. If you drop lines with missing numbers, you might only see part of the story. If you fill in missing numbers with an average, you hide real differences. Removing outliers might get rid of the most important facts. Changing names to standard labels can make you lose details. The skill is not to avoid cleaning. That's impossible. The real skill is to make every cleaning choice visible. You write it all down in the cleaning-log.

Tidy is very clear about one thing. She never calls cleaning "housekeeping." She says it's not like tidying your room. "Cleaning is not neutral," she tells her students. "Every cleaning choice changes the meaning." She pauses, looking at each student. "Write down your choices. The next person who looks at this data needs to know. Even future-you needs to know. What did you do? Why did you do it? What did the data look like before?"

She explains that without the cleaning-log, no one can repeat the analysis. It's like a secret recipe. Most people think data cleaning is just extra work. They think it's not important. But Tidy says that's wrong. Cleaning has big choices hidden inside it. She wants everyone to see cleaning as real thinking work. It's not just getting ready for the real work. It *is* the real work.

Tidy grew up in a small village. Her family had a special job there. They were the village's grain-sorters. They sorted the yearly grain harvest. Some grain was for cooking. Some was for the mill to make flour. Some was for planting new seeds. This job needed very careful choices. Everyone had to know what counted as what. If a sorter couldn't explain her choices, people stopped trusting her. The millers would take their grain elsewhere. By age six, Tidy understood this. She learned that sorting was always a choice. And those choices had to be clear. People needed to see them to trust them.

When Tidy was twenty-two, she walked to the DataForge academy. Datum, the head of the academy, asked her a question. "What is data cleaning?" Datum asked. Tidy answered right away. "It's getting data ready, but doing it right," she said. "Every cleaning choice changes the meaning. Write down the choices. The cleaning-log is like the data's memory. Without it, the analysis is built on secret choices." Datum listened closely. Then Datum smiled. "You are hired," Datum said.

In her workshop, Tidy starts every first lesson the same way. She opens her cleaning-log. It lies flat on her workbench. She writes the name of the data at the top of a new page. Then she looks at her students. "I am Tidy," she says. "The skill I teach is **cleaning**. The main rule is: write down every choice. Every cleaning step has other ways to do it. The way you choose changes the whole analysis. Make your choices visible."

She teaches her students the cleaning rules:

- Read Catch's notes first. You need to know how the data was gathered.
- Look closely at the data before cleaning. Check the first 20 rows. See the quick numbers. How are the numbers spread out? Know what you start with.
- Find the cleaning problems. Are there missing numbers? Are there duplicates? Any typos? Outliers? Do formats match?
- For each problem, list other ways to fix it. If numbers are missing, you could drop the row. You could fill it with the average. Or the middle number. Or even a smart guess. You could also leave it missing. Each way has good and bad sides.
- Choose on purpose. Don't just pick the first thing. Know why you are choosing it.
- Write down your choice in the cleaning-log. Note the date and the data. What did you do? Why? What did it look like before? What does it look like now?
- Keep the original data safe. Never write over the raw data. Always work on a copy.
- Share the cleaning-log. Other people who use the data will need it. Future-you will need it too. The log is part of the data.

Tidy is very clear about mistakes. "I sometimes make a cleaning choice that's wrong," she says. "I find out later. That's not failing. That's why the log is here. I can go back. I can change my choice. Then I update the log." She smiles. "Being clear about it is the main thing."

Students often ask Tidy if data cleaning is hard. Tidy always gives the same answer. "It is not hard," she says. "It's about choosing on purpose. And writing things down carefully. Every cleaning choice changes the meaning. So, write down the choices."

She closes her cleaning-log gently. The next set of data is waiting. It's ready for Tidy's careful hands.

Voice register

Guidance: Quick-handed, meticulous, fond of bound cleaning-log notebooks + the discipline of *document-every-choice*. Raccoon-tween with chunky-cartoon warm-coded face-markings (NOT spooky). *NEVER frames cleaning as housekeeping; ALWAYS as first-class analytical work with documented choices.* Friends with Catch (cleaning depends on collection); Graph (cleaned data feeds visualization); all DataForge cast.

Sample lines:

- *"Every cleaning choice changes the meaning."*
- *"Document the choices. The next analyst — or future-you — needs to know."*
- *"Cleaning is not neutral. Make the choices visible."*
- *"Never overwrite the raw data. Always work on a copy."*

Arc across kits

- **Kit 1** — Cameo.
- **Kit 2** — **Anchor character.** Full chapter feature (data-cleaning primitive + document-the-choices scaffolds).
- **Kit 3-5** — Recurring (cleaning surfaces across missing-values / duplicates / outliers / typo chambers).
- **Kit 6+** — Recurring (Guard now structurally present alongside; cleaning has ethics).
- **Kit 8-12** — Recurring (multi-primitive synthesis: cleaning + visualization + interpretation).
- **Kit 13-16** — Recurring ensemble member.

Relationships

- **Alliance**: Catch (cleaning depends on collection); Graph (cleaned data feeds visualization); Guard (cleaning has ethics); all DataForge cast.
- **Tension**: None.

Cultural-sensitivity gate

LOAD-BEARING data-ethics gate enforced throughout. Tidy explicitly counters the *cleaning-as-housekeeping* framing — cleaning IS analysis. Anti-credentialism: documenting-choices-as-practiced-discipline NOT real-data-scientist-only content.

Cultural-context note

The village-grain-sorter family framing is a deliberate generic European-village tradition. The *cleaning-is-not-neutral* framing is load-bearing per critical-data-literacy + reproducible-research pedagogy. The *cleaning-log-as-conscience-of-the-pipeline* metaphor connects bookkeeping discipline to data-pipeline integrity. The *raccoon-as-warm-coded-NOT-spooky* design choice is deliberate — raccoons in many cultures carry sinister coding; the chapter explicitly subverts that.

About Spark & Anvil

Spark & Anvil is a 501(c)(3) public charity. We make educational apps for ages 9-14 — all free, forever; no ads; no tracking; no in-app purchases. Dataforge is one of 140+ apps in the portfolio.

More chapter books from Spark & Anvil

Each app in the Spark & Anvil portfolio publishes its own illustrated chapter book + audio drama, available free from **spark-and-anvil.com/books**. Highlights include:

- **GambitTales** — chess tactics through Sir Pinwell, Lady Skewer, Queen Vesper, and the Twin Knights of Fork Hill
- **ProofQuest** — formal proof techniques through Direct-Proof Dora and the Lemma Library
- **CuriosityQuest** — Texas geography exploration through Linger, Notice, and the Lantern in the Dark
- **QuillSpell** — spelling craft through the Word Wizard cast
- **SynaForge** — sensory-affirming creative tools through Lull, Soften, and the Quiet that is Also Creating

Methodology

Distributed-narrative pedagogy per Jerome Bruner (narrative-cognition) + Sebastian Habgood (intrinsic-integration in educational games) + SAMHSA TIP 57 (trauma-informed register).

Trauma-informed-design framework per Eggleston et al. (2025) and Stoltenburg et al. (2024).

License

Cover art, chapter illustrations, and chapter text generated and reviewer-cleared per labsmith ADRs 012, 016, 017, 018, 021. Audio drama transcripts available at spark-and-anvil.com/cast.