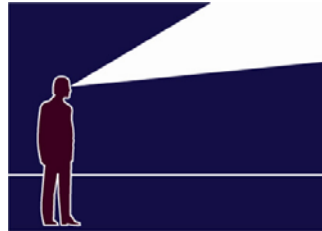# The Data Provocateur's Boot Camp: Step One: Getting Started DAMA Phoenix

Thomas C. Redman, Ph.D.

"the Data Doc"

Data Quality Solutions
tomredman@dataqualitysolutions.com

On twitter @thedatadoc1

Objectives of Today's Session

- The complete Data Provocateur's Boot Camp aims to provide you the essential material to become an effective data provocateur.

- Today's focus is on step 1 of a 4 step series

- "Getting Started" Answers the question, "Do I (we) have a data quality problem?"

- It also aims to empower and embolden you to move forward.
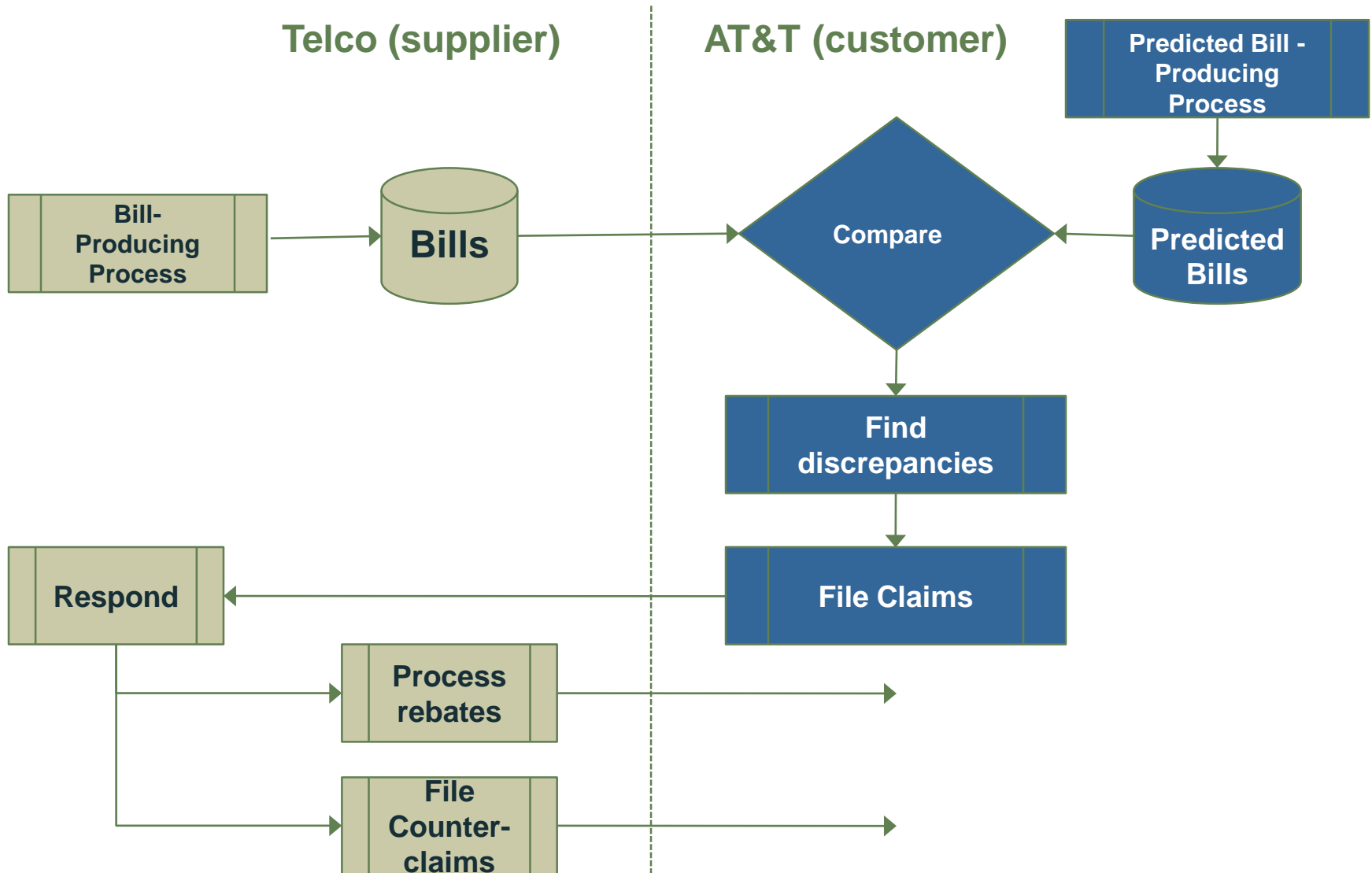
Today's Outline

◻ Example: Bob Pautke at AT&T

◻ What is a "data provocateur?"

  ◼ The hidden data factory

  ◼ Provocateur Personas

◻ Step 1: Answer "Do I (we) have a data quality problem?"

  ◼ The Friday Afternoon Measurement

  ◼ The Rule of Ten

  ◼ Summing up

◻ Wrap Up and Next Steps

  ◼ A Promise to yourself

Bob Pautke provokes AT&T's Access Management Department

The Prequel: Access Bill Verification at AT&T

Telco (supplier)   **AT&T (customer)**

Predicted Bill - Producing Process

Bill-Producing Process → Bills → Compare ← Predicted Bills

Compare → Find discrepancies → File Claims → Respond

Respond → Process rebates

Respond → File Counter-claims

An early discussion

"What is the size
sample need
to calculate,…

"How does the end-to-end
process work?

Gee, I don't know.
How can we find out?

A Tracked Data Record

"You had me at Hello"

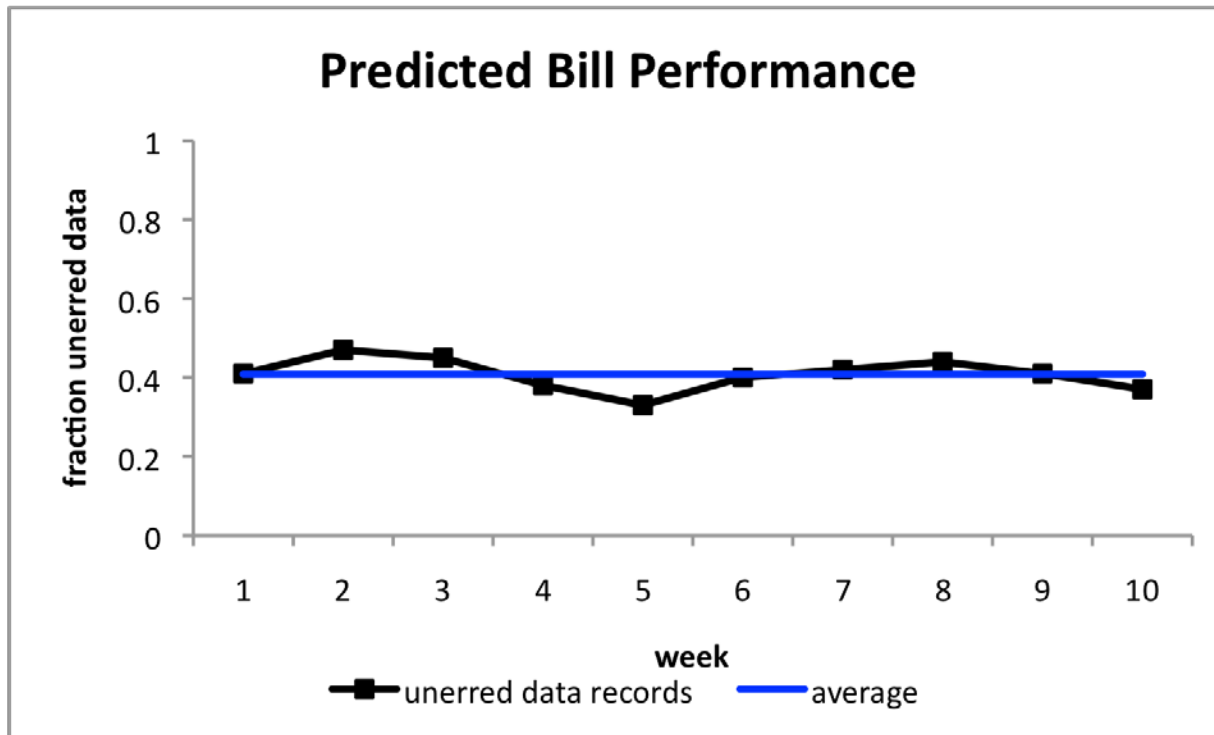The only change here involves format. Not of concern

Attribute	Step of Process				
	A	B	C	D	E
Name	XYZ.1234	XYZ-1234	XYZ-1234	XYZ-1234	XYZ-1234
Billing Number			272-791-2424	272-791-9100	272-791-9100
Bill Code	1	A	A	A	A
Office	408727	408727	408727	408927	408970
...					

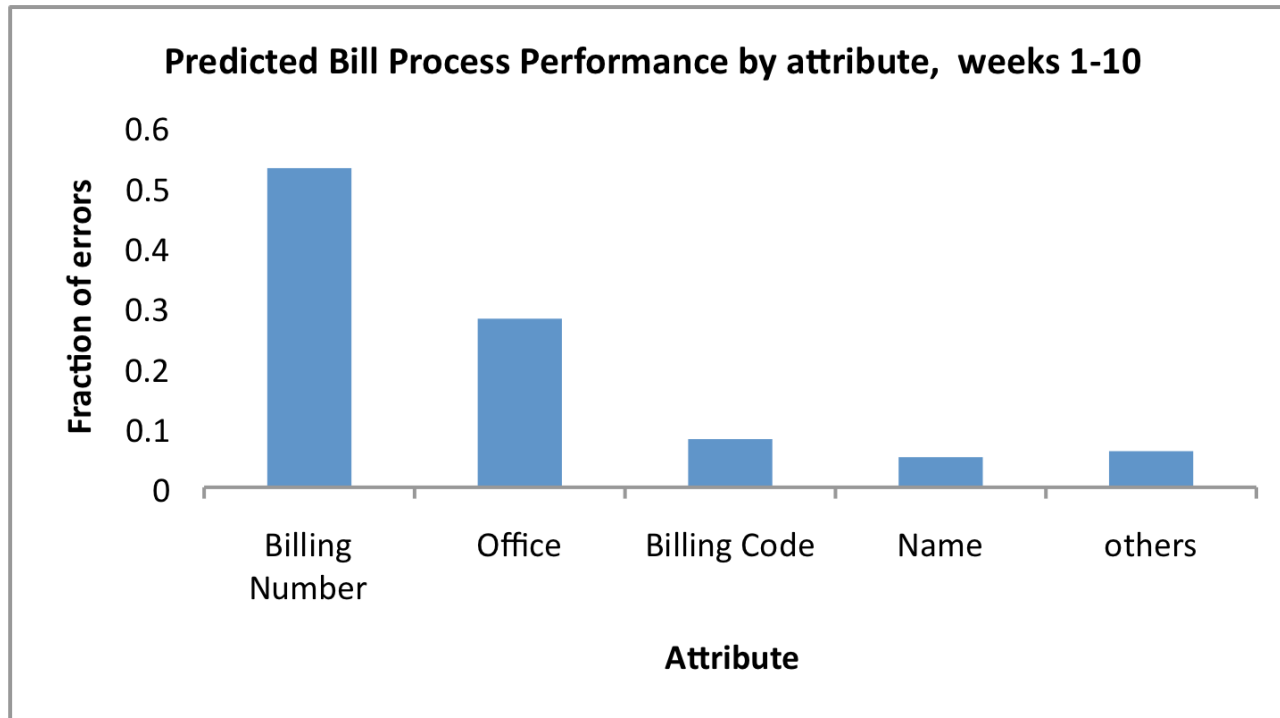In this case, "1" and "A" both mean yes. It reflects poor architecture, but is not a data error

Changes in data values. Errors of serious concern. Note we cannot be certain where the errors actually occurred.

Start with a Basic Time-Series Plot

Predicted Bill Performance
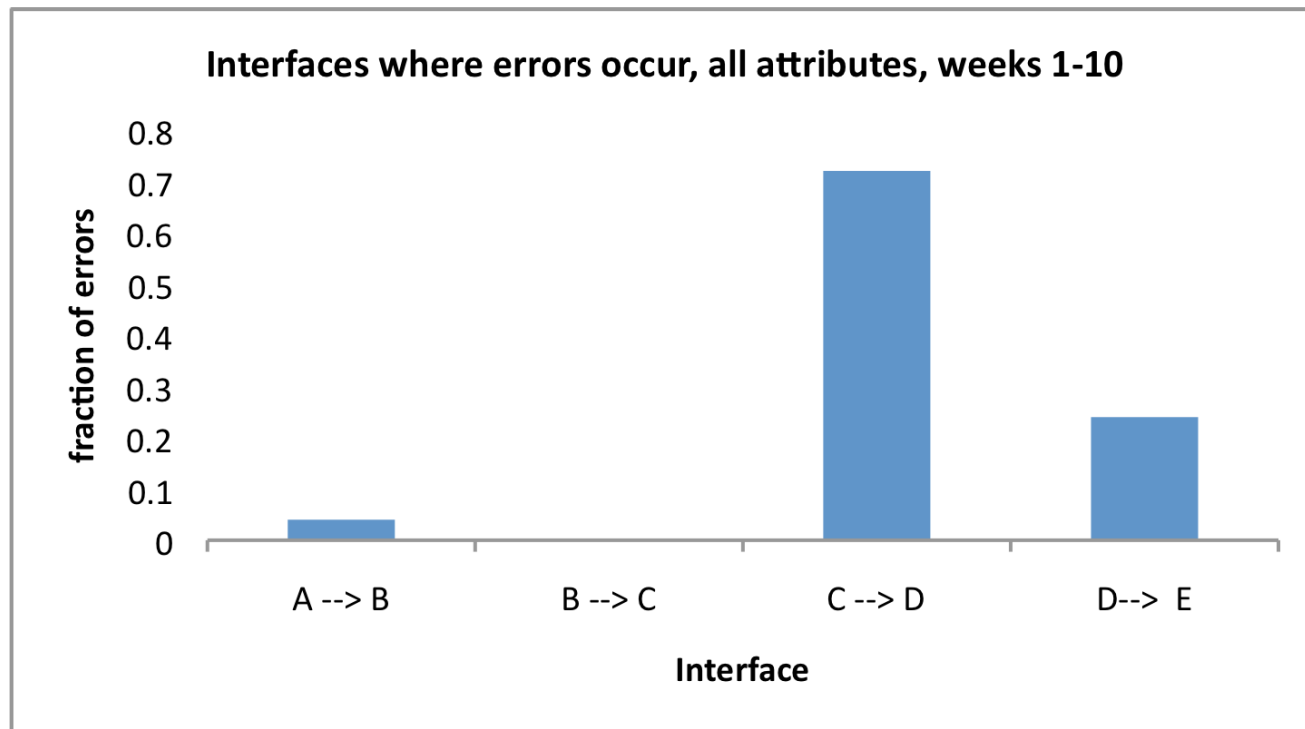
Clearly, this process is broken!

The Search for Root Causes

Predicted Bill Process Performance by attribute, weeks 1-10

The Search for Root Causes, Another Perspective

Interfaces where errors occur, all attributes, weeks 1-10

Combining the Two Perspectives

Predicted Process Performance by Interface and Attribute, weeks 1-10
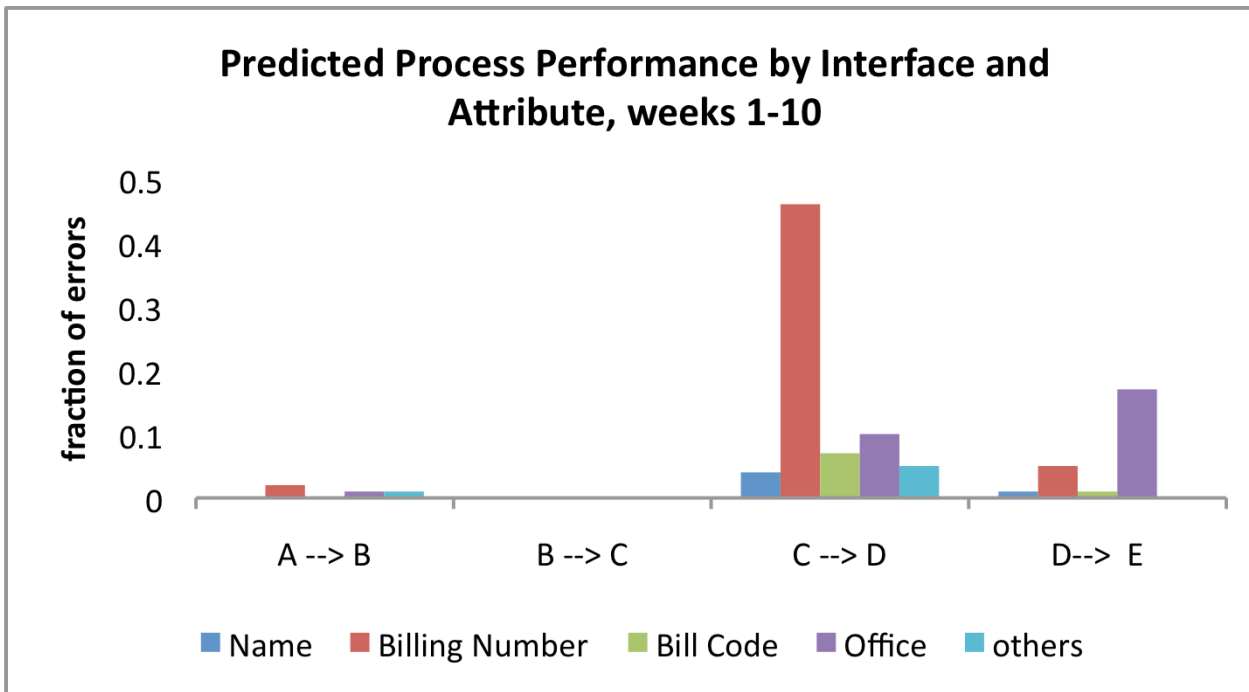
With this plot, Bob has identified potential root causes well enough to charter specific improvement projects.

Eliminating the top three will reduce the overall error rate by 80!

Over time, a revelation unfolds,…

"We clean up our data every other year. And its good for me—I get a big bonus. But somehow it doesn't feel right for the company!"

Twain: "A man with a watch knows what time it is. A man with two is never sure."

"Bill verification doesn't work (this is stupid!)"

"Perhaps we should go to a self-billing arrangement."

A Chance (but timely) Meeting

- Cincinnati Bell and AT&T share a building
- After a chance meeting between two former colleagues, the two set up a larger meeting.

"You know Bob, we don't come in every morning thinking of ways to foul you up. What does 'timely and accurate' mean?"
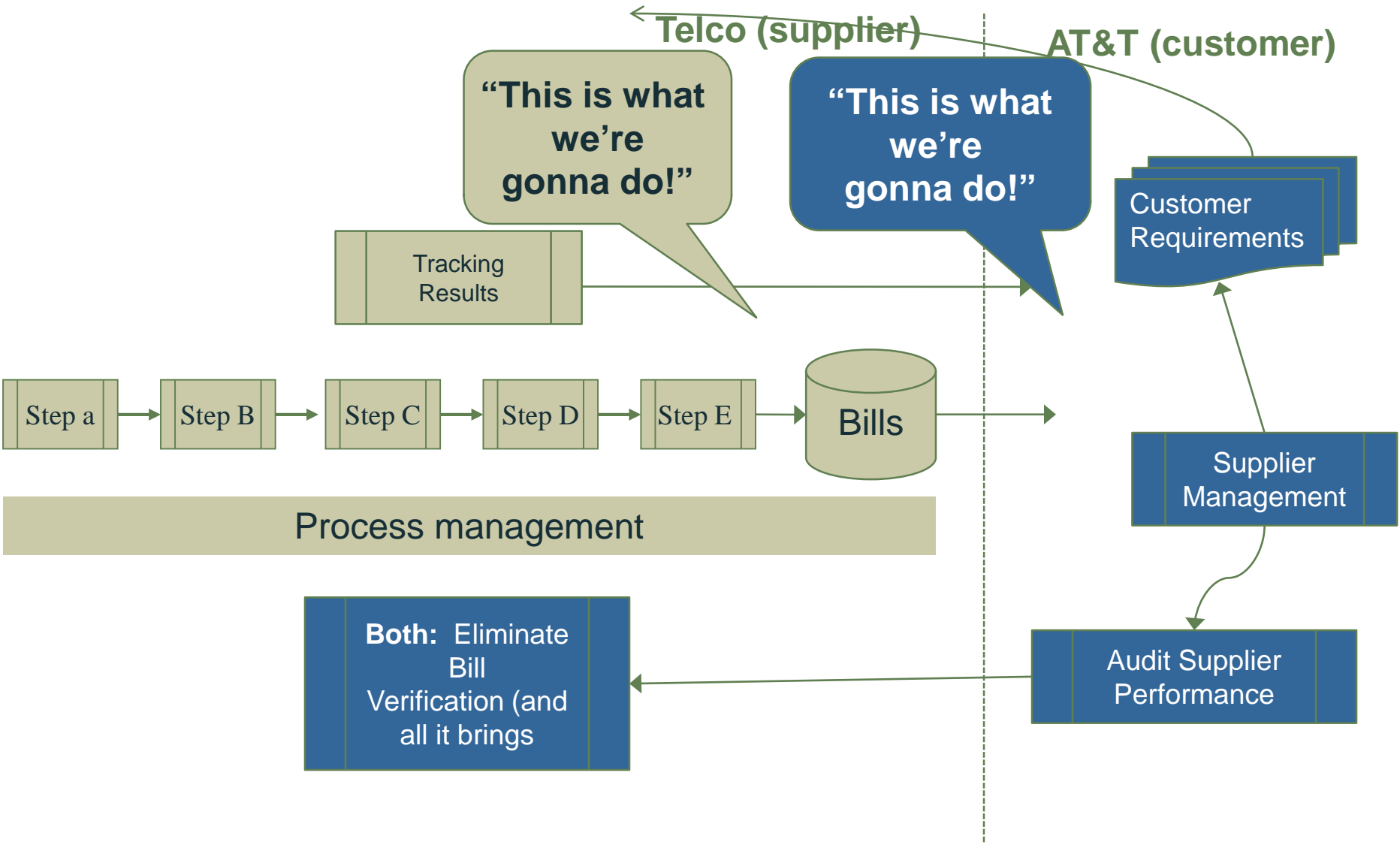
"Hmmm. That's A good question?"

This has to be end-to-end

- Joint tracking continues to reveal that most problems occur on interfaces, this time between companies.

- After a few improvement projects, Bob can claim a "real result."

**HE HAS NOW SATISIFIED ALL REQUIREMENTS
AS A PROVOCATEUR
AND HE HAS "WRITTEN A SCRIPT"
FOR THE REST OF AT&T**

New Access Financial Assurance Process

Business Results

- Data accuracy errors reduced 90%
- Billing errors reduced 98%
- Cycle time (bill period closure) reduced 67%
- Customer costs reduced 73%
- Costs across supplier base reduced 20%
- Achieved Financial Assurance
- Created Financial Predictability

> "Tom, I really do appreciate the $100M.
> What I appreciate even more is that now we can manage the business. That is worth even more."

What is a "data provocateur?"

Rising Middle Manager

The Hidden Data Factory

We hide a lot of non-value-added work in our processes:

Input data → Check → Fix errors → Complete our Value-added work →
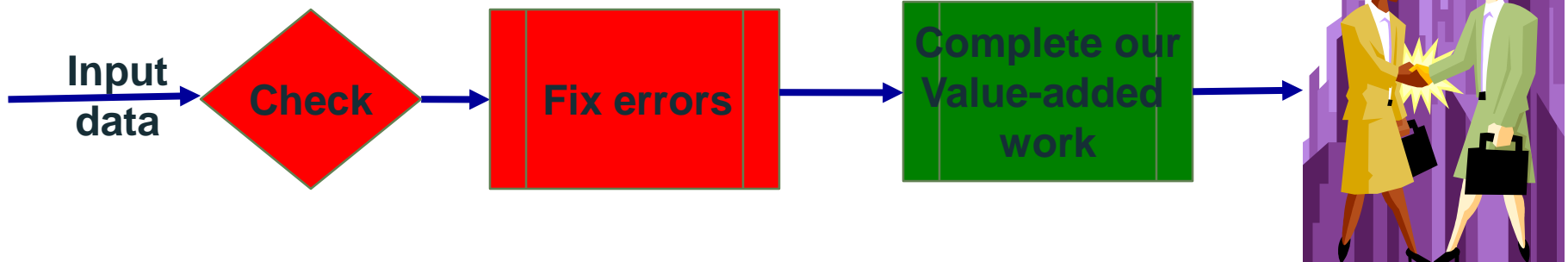
- We are too busy to reach out to those who provide data to explain our requirements, provide feedback, or ask them to correct their errors.

- As the next slide illustrates, this non-value-added work is expensive.

  We're remarkably forgiving, of both bad data suppliers and ourselves!

"Provocateurs" are people who disrupt the dynamic that leads to hidden data factories

- Dissatisfied with the status quo.
- Courage to try something new.
- Great corporate citizens.
- Achieve "real results" within their spans of control.
- At all levels!

THERE IS A LITTLE PROVOCATEUR IN ALL OF US AND IT IS TIME TO LET IT OUT!!

To qualify, provocateurs must Obtain a Real Result

1. Rqmts defined

First-time, on-time results

2. First Meas

3. Improvements

4. Control

Program start

Fraction Perfect Records

Month

━━◆ Accuracy Rate ━━ ave ━━ lower control limit ━━ upper control limit ━━ target
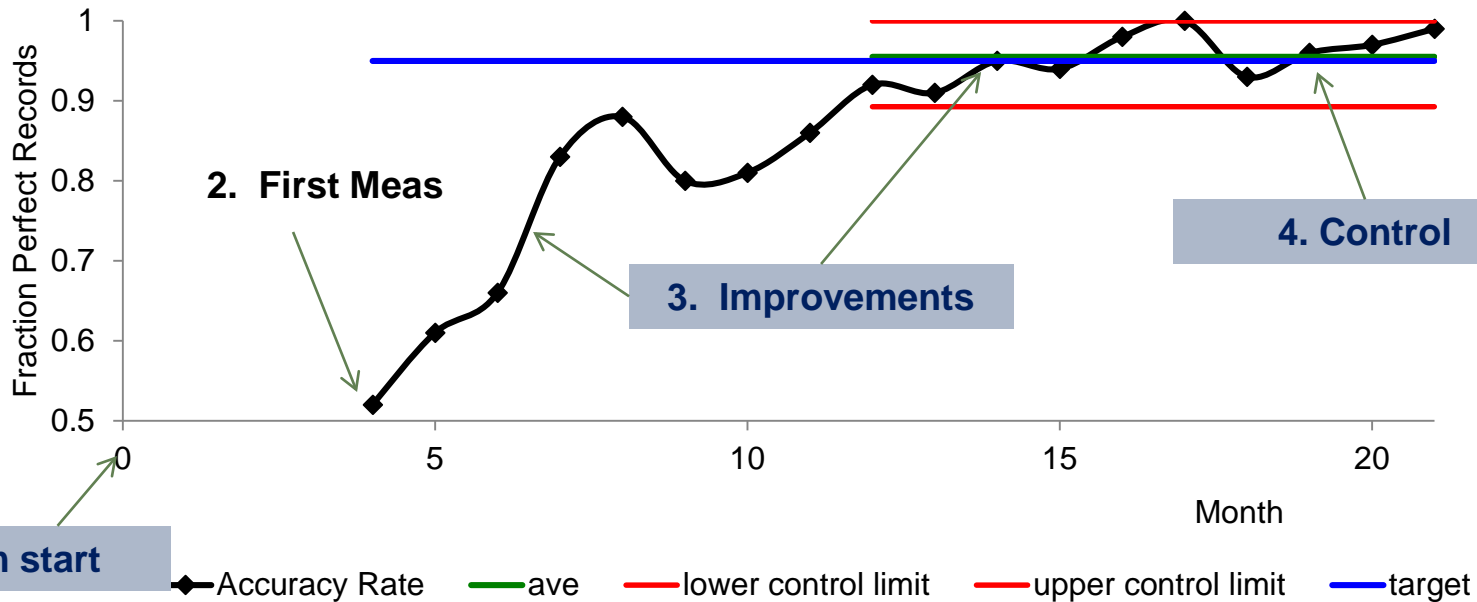
Each error not made saves an average of $500. Quickly millions. But the improved competitive position may be worth even more!

Provocateurs Can Go Only So Far

Progress of a Typical Data Quality Program

- penetration of DQ across organization (y-axis)
- time (x-axis)
- traction
- real results
- plateau
- next level = more data
- Order-of-magnitude improvement on some data

Provocateurs come from all levels, ages, departments, and …

Greg, Head of Analytics, age 33

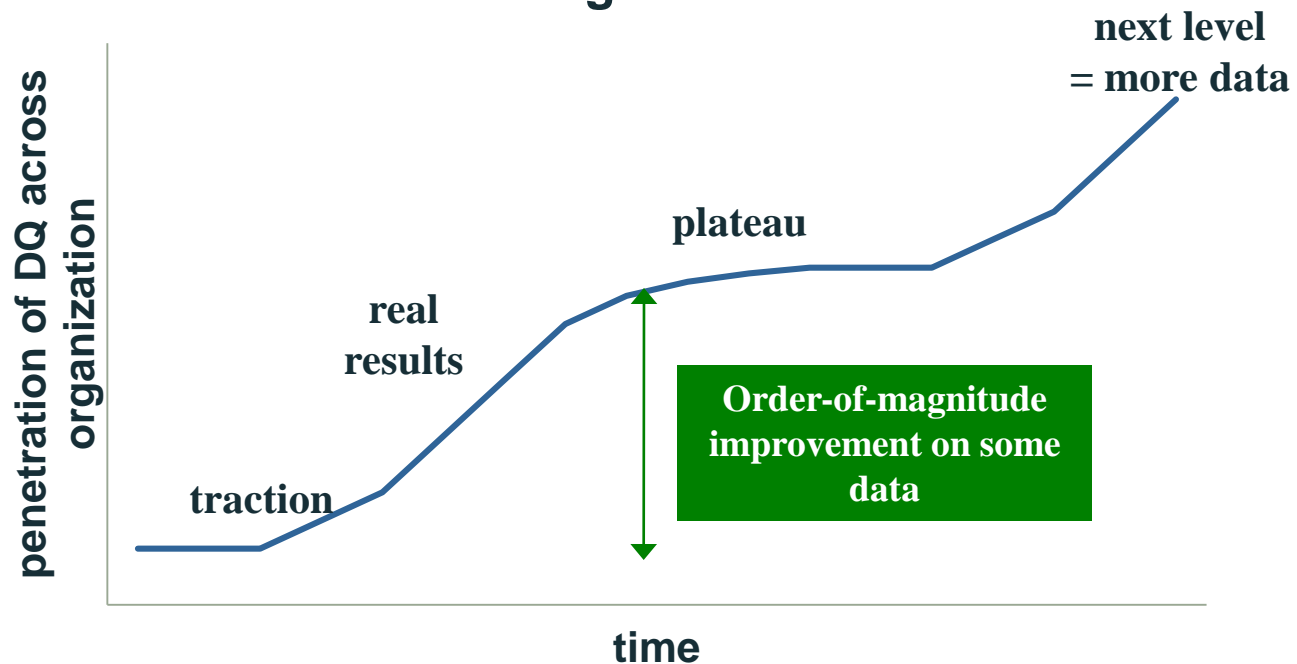"Twice last year we made really interesting discoveries. But management wouldn't even listen. They told us that we couldn't be right—that the data we used had too many problems, even though we spend more than half our time rooting out errors. Morale is horrible—two good data scientists left already. There has to be a better way to deal with these quality issues!"

"The first thing I realized was that we didn't approach quality in a structured way—we didn't spend enough time learning about our sources and so our clean-up routines were pretty ad hoc. So I asked a factory manager to help. Well, he couldn't have been nicer. He arranged a whole tour. Then he asked us to explain the problems we were having. Then he offered to set up a team to fix those problems. He told me to think of myself as a "customer" of his data.

"Looking back, I should have known better. I made a big mistake thinking we could do data quality on our own. When they do it at the source the data becomes ten times better than we could ever do. Management is listening to us now because I can explain why they should trust the data."

Huzaifa, Financial Analyst, age 27

"My team makes sure other departments provide the right data. When they don't, which is all the time, its our job to correct it. Budgets and other reports get fowled up when the data is wrong, so I see why we have to do it. But wouldn't it be better if those departments sent us good data? If I can figure this out, I'll get a big bonus!"

"My boss told me she didn't think I would have any luck, but she encouraged me to try. She even connected me with a good friend of hers in another department to get the ball rolling."

"It turned out that my bosses connection didn't even know we used their data. And they were pretty clear, right from the beginning, that they would be happy to help. They weren't just being friendly—bad data was slowing them down also. We started by working on budget data. Once I got everyone together, things went pretty well—a lot better than my boss expected. And by the way, I'm up for that promotion!"

Shelley, Marketing, age 43

"We license market data and it is cutthroat. All my competitors do is talk about how good their data is. Frankly, I don't think they are any better than we are, but I can't prove it. My colleague who runs Customer Service tells me that his team spends most of their time correcting simple errors, not providing real help. That can't be good. I wonder what we can do to distinguish ourselves???"
Shelley, Marketing, age 43

"I decided to measure quality as our customers saw it. When I plotted the data out I saw the problem—most of the time we are pretty good, but we had really bad weeks. So I reached out to the Operations' Group that created that data."

"When Barbara, who runs Operations, and I dug in, we found four problems with the market data. In a previous job she had learned a way to attack such problems and we applied it here. She got members of her team to lead the improvements. Three teams went really quickly. Even though the last one struggled a bit, within six months I had something to show prospective clients. And they really liked it! This helped us win two new deals."

Lwanga, Data Manager, age 50

"My team's job is to take data from all the transactional systems and faithfully load it into the data warehouse. Judging from the number of complaints we receive from the business, I'm pretty sure we are just moving junk around. I don't know how I'm going to keep up: We've got another big system coming on-line next year and there is talk of budget cuts." Lwanga, Data Manager,

"Maybe I could try something new. I'll focus on the data people use most often and one system. Maybe together we can get to the root of the problems."

"Wow, the last year has been the most amazing one of my career. The first thing we did was simply print out some of the data and look at it. I don't know why we didn't do that before. And it was awful. So we asked some people what needed to be fixed first. They suggested six areas and helped figure out where that data came from. It took a while to get the hang of working with people upstream. I'll never do it any other way!"

Liz, Head of 2000-person Business Unit, age 57

"I've always been pretty skeptical about data and it's getting worse-- people complain about it all the time. But last week was the final straw. Three of my best managers had a fight, right here in my office, about whose numbers were best. This is no way to run a business!"

"I decided to start with the weekly sales report. It's the best indicator of how we are doing. Right away I can see I am going to need to get everyone to agree on when we recognize a "sale" and I'll probably have to put someone in charge of the overall process."

"Frankly, it didn't take all that much effort. Once people understood I wasn't going to let up, they got on board pretty quickly. We did have to solve some process problems, but nothing my guys could not handle. For me, the biggest benefit is that, people are starting to see themselves as part of a team, not as individuals running their departments."

Discussion

What is your motivation?

Step 1: Do You Have a Data Quality Problem?

- The Friday Afternoon Measurement
- The Rule of Ten
- Other costs due to bad data
- A tantalizing possibility

The Friday afternoon measurement aims to help answer the question,

"Do I need to worry about data quality?"

Friday Afternoon Measurement Protocol

Assemble "last 100 records" → Assemble 2-3 experts

Mark "obviously-erred" data in red → Summarize and interpret results
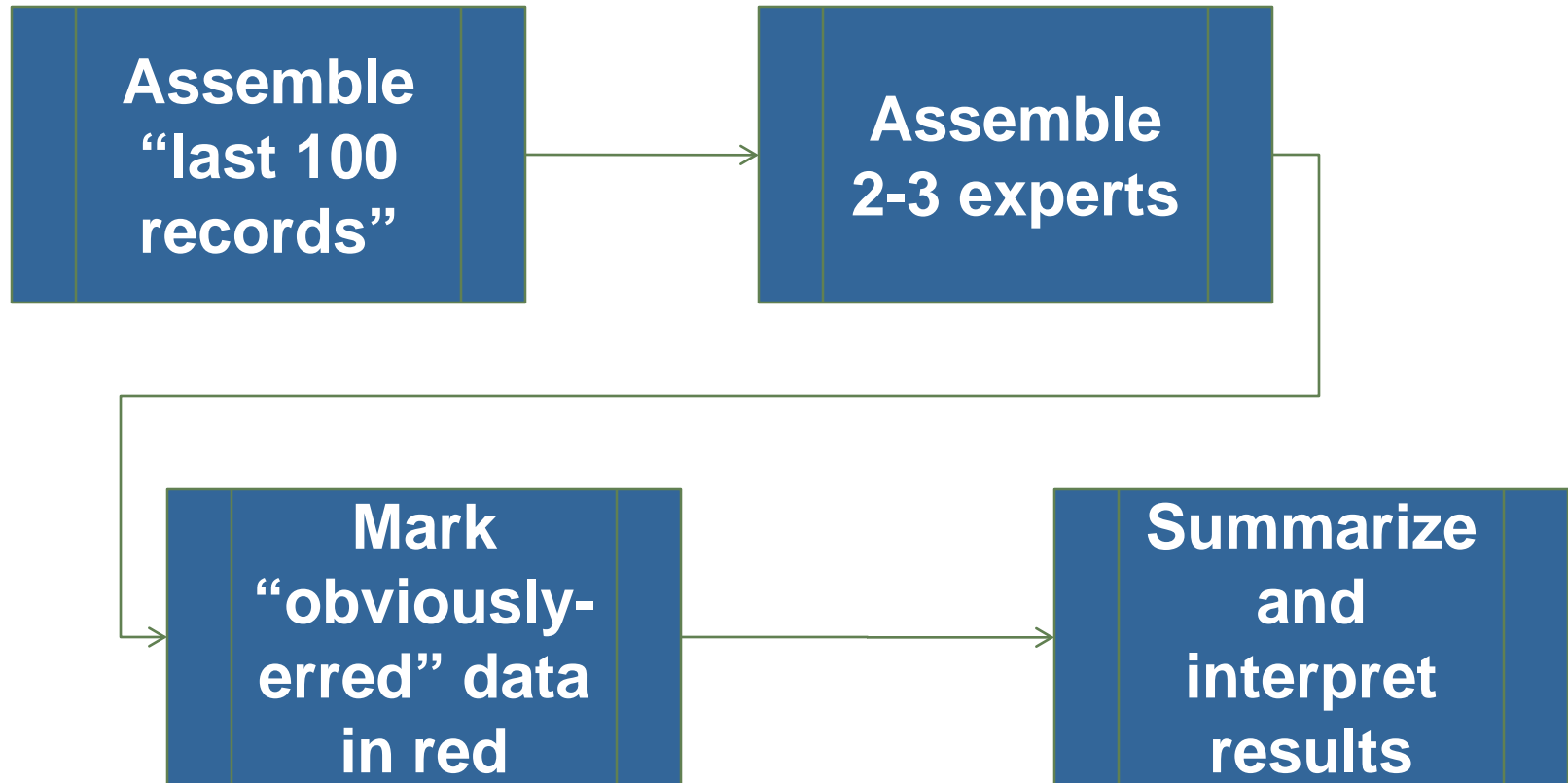
Assemble the last 100 records

Select 10 – 15 most important attributes

Upscale Sweater Company

Data collected on 1 October 2016				
Indicator	**Name**	**Size**	**Color**	**Amount**
Record A	Jane Doe	null	light blue	129.00€
B	John Smith	Med	blue	129.00€
C	Stuart Madnick	XXXL	red	129.00€
CV (100)	Alyson Heller	Med	blue	129.00€

After Fig 18.2, Redman, Data Quality: The
Field Guide

Now, working together, mark the obvious errors

Upscale Sweater Company

Data collected on 1 October 2016				
Indicator	**Name**	**Size**	**Color**	**Amount**
Record A	**Jane Doe**	**null**	**light blue**	**129.00€**
B	**John Smith**	**Med**	**blue**	**129.00€**
C	**Stuart Madnick**	**XXXL**	**red**	**129.00€**
CV (100)	**Alyson Heller**	**Med**	**blue**	**129.00€**

After Fig 18.2, Redman, *Data Quality: The Field Guide*

Rate the record as "perfect" or not

Upscale Sweater Company

Data collected on 1 October 2016					
Indicator	Name	Size	Color	Amount	record perfect? (y/n)
Record A	Jane Doe	null	light blue	129.00€	n
B	John Smith	Med	blue	129.00€	y
C	Stuart Madnick	XXXL	red	129.00€	n
CV (100)	Alyson Heller	Med	blue	129.00€	

After Fig 18.2, Redman, *Data Quality: The Field Guide*

Count the "perfects"

Upscale Sweater Company

Data collected on 1 October 2016					
Indicator	Name	Size	Color	Amount	record perfect? (y/n)
Record A	Jane Doe	null	light blue	129.00€	n
B	John Smith	Med	blue	129.00€	y
C	Stuart Madnick	XXXL	red	129.00€	n
CV (100)	Alyson Heller	Med	blue	129.00€	
Error Count	0	24	5	2	perfect = 67

After Fig 18.2, Redman, *Data Quality: The Field Guide*

Data Quality = 67%

Here the interpretation is a full third of recent customer orders had a serious DQ issue.

A worry indeed!

Prima facie **evidence of a problem.**

To calibrate

In a recent class with 15 people:

- ☐   Min: DQ = 8%
- ☐   Max: DQ = 92% (the highest I have seen)
- ☐   Ave: DQ = 53%

The 67% here is on "the high-end of typical."

Practice

- ☐ Team up, with one or two other people.

- ☐ Consider the last 100 customer orders at "Upscale Sweater Company."

- ☐ Follow the FAM protocol.

- ☐ Use the sheet provided to understand enough about sweaters to do so.

Practice continued

When called on note:

- ☐ DQ

- ☐ The total number of errors you found

- ☐ Any subtle errors you found
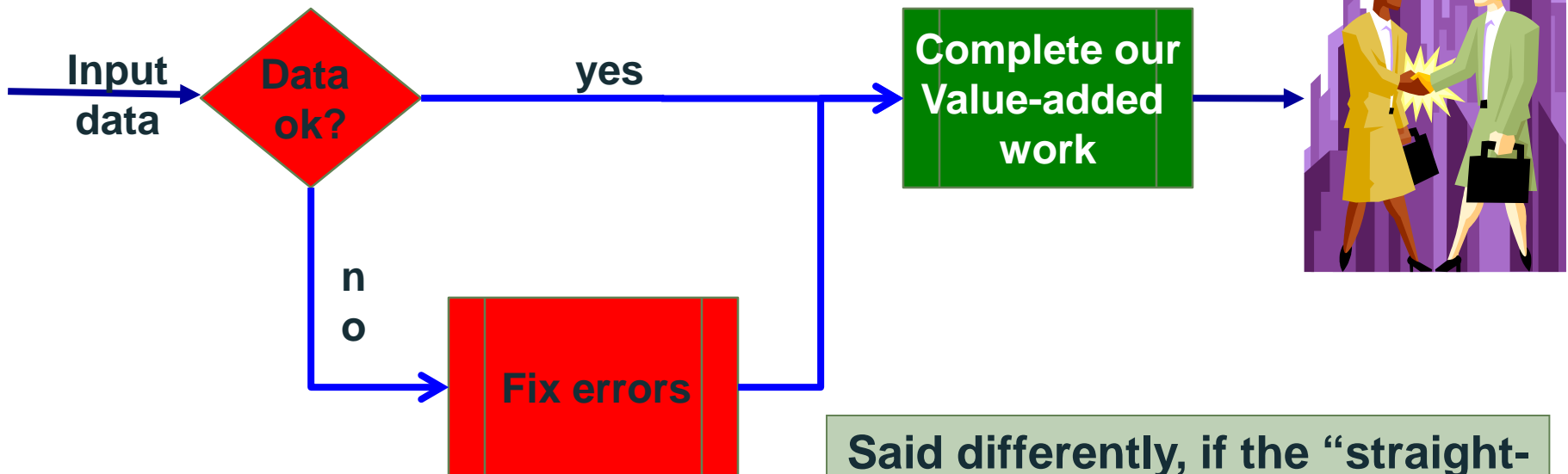
- ☐ Any questions or comments

Discussion: Can you make a FAM in your work?

Now, estimate the impact

Data Doc's Rule of Ten helps estimate the cost of the hidden data factory

"It costs ten times as much to complete a unit of work when the data is flawed in any way as it does when they are perfect!

Input data → **Data ok?**

yes → **Complete our Value-added work** →

no → **Fix errors**

Said differently, if the "straight-through path," costs a dollar, then the "fix errors path" costs ten.

Data Quality as the means to cost reduction

Cost-of-Poor-Data Quality (COPDQ) Calculator

Cost to complete 67 defect-free ($1 each)	=	$67.00
Cost to complete 33 defectives ($10 each)	=	$330.00
Total cost	=	$397.00
Value-added cost (all perfect)	=	$100.00
Non-value-added cost	=	$297.00
%COPDQ ($297/$397)	**=**	**75%**

The business cost is enormous!

There may be other costs as well

Pick something that "engages the heart"

COMPETITIVE POSITION

Strategy:
➢ Harder to set and execute strategy
➢ Fewer options to derive value from data
➢ Harder to align organization
➢ Distracts management attention

Operations:
➢ Higher operating cost
➢ Lower employee morale
➢ Lower customer satisfaction

Tactics/Decision Making:
➢ Lower trust between orgs
➢ Lost sales
➢ Poorer decisions
➢ Increased technology risk
➢ Harder to manage risk

Figure 2.1, Redman, *Data Driven: etc*

Now synthesize results

| The Provocateur's Opportunity		
	Current	**Opportunity**
Measured DQ	DQ = .67	DQ = .03
Associated Cost	Added cost = 300%	Reduce total cost by 67%
Other Benefit	Viewed as a "me-too" player	Improved competitive position

Discussion

- Is there any of this that you cannot do, in the context of your current job, within a very short amount of time?

- Describe an area in which to start?

- What's holding you back?

- Note a date by which you can commit to completing this step.

What comes next?

- Four-Steps for Becoming a Provocateur
  - Step 1: Answer "Do I (we) have a data quality problem?" (Which we worked on today)
  - Step 2: Clarify, document, and communicate customer needs
  - Step 3: Make improvements to close the gaps
  - Step 4: Take up your responsibilities for data quality/Be a role model

- Invite me back and we will discuss the remaining steps!

Questions?
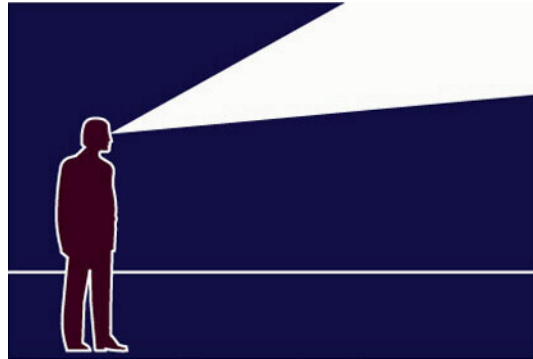
Thomas C. Redman, Ph.D.

"the Data Doc"

+1 732-933-4669

tomredman@dataqualitysolutions.com

www.dataqualitysolutions.com

on Twitter: @thedatadoc1

Thomas C. Redman, "the Data Doc"

- Ph.D., Statistics, Florida State, 1980.

- Conceived and led the Data Quality Lab at AT&T Bell Labs.

- Formed Data Quality Solutions in 1996.

- Latest and greatest: "Getting in Front on Data: Who Does What" *Technics Publications,* 2016.

- "Data's Credibility Problem," *HBR*, Dec. 2013.

- *Data Driven: Profiting from Your Most Important Business Asset,* Harvard Business School Press, 2008.

- Known bias: "Data are quite obviously the key asset of the Information Age. Yet today's organizations are unfit for data. Further, despite enormous potential, few are yet considering how they will compete with data. Finally, high-quality data is pre-requisite. These crystallize THE management challenges of the 21st century."