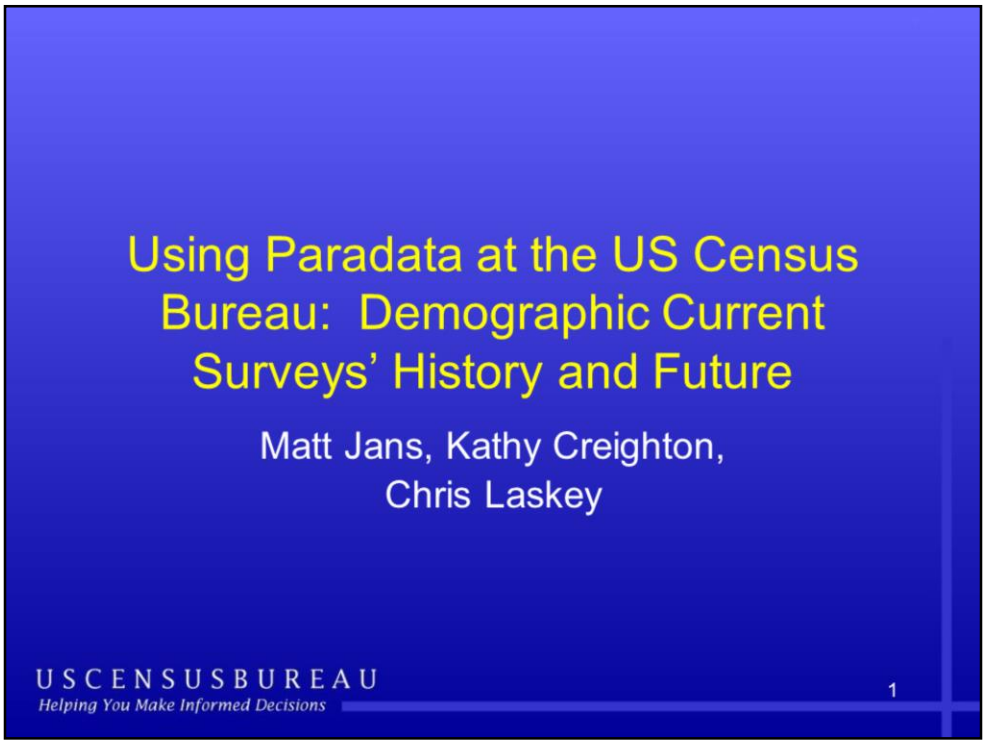
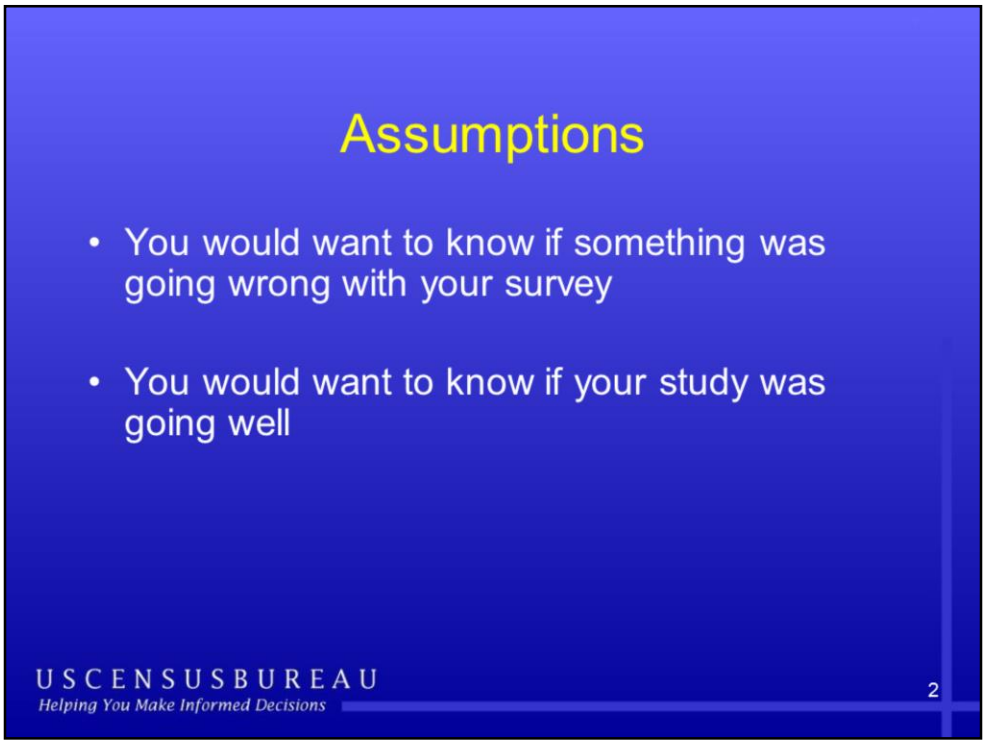


I'm going to talk today about some current efforts at the Census Bureau to integrate paradata sources and provide more complete project information to managers on a real-time basis.

I'll focus on how the **Census Bureau's paradata structures** affect the accessibility and usability of paradata, but I'll try to present this to you in a way that is **also generalizable to other research agencies** and survey research firms.

I should say at the outset that in addition to the authors listed here, there are **several others who are part of the paradata integration project**. There are also **scores of individuals more** working on different aspects of paradata at the Census Bureau. This should become obvious as I discuss the breadth and depth of the data we have.

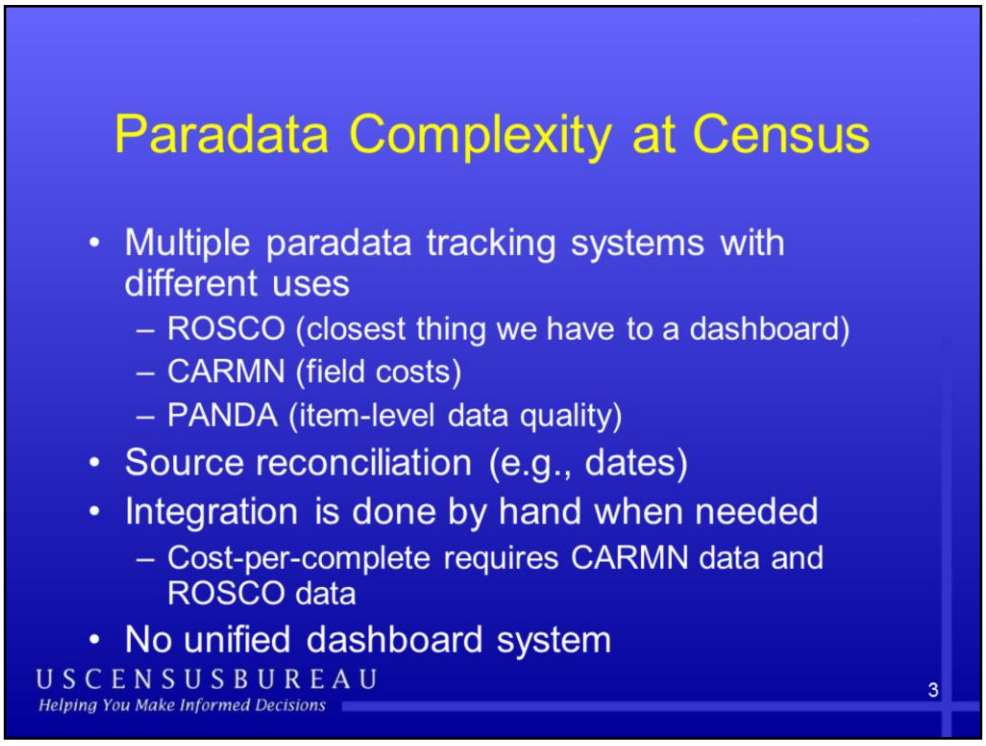
TERMINOLOGY



I'm going to be upfront about my assumptions about why you might be interested in paradata at all.

I assume that if you could, you'd want to know if something was going wrong with your study. Is production moving slowly? Are you getting particularly high item nonresponse rates from certain interviewers or in certain geographical regions? Are cost being incurred in unexpected ways or at unexpected times.

I'll talk about the ability to meet these goals through the integration and presentation of paradata



The discussion today will be about efforts in the **Demographic Directorate at Census** that focus on our Current **(or Reimbursable) Surveys like CPS**. Decennial has similar operations underway for tracking cost and scheduling.

Currently at Census we have a number of different paradata systems, each of which performs a specific task for a specific segment of survey operations.

The three listed here are just examples of three of the primary sources and represent general classes of paradata.

ROSCO is our Regional Office Sample Control program which **consolidates several data sources**, and serves as a **case management and progress tracking** system for the **field division.** From what I can tell, it may be the **closest thing to an integrated system we currently have.**

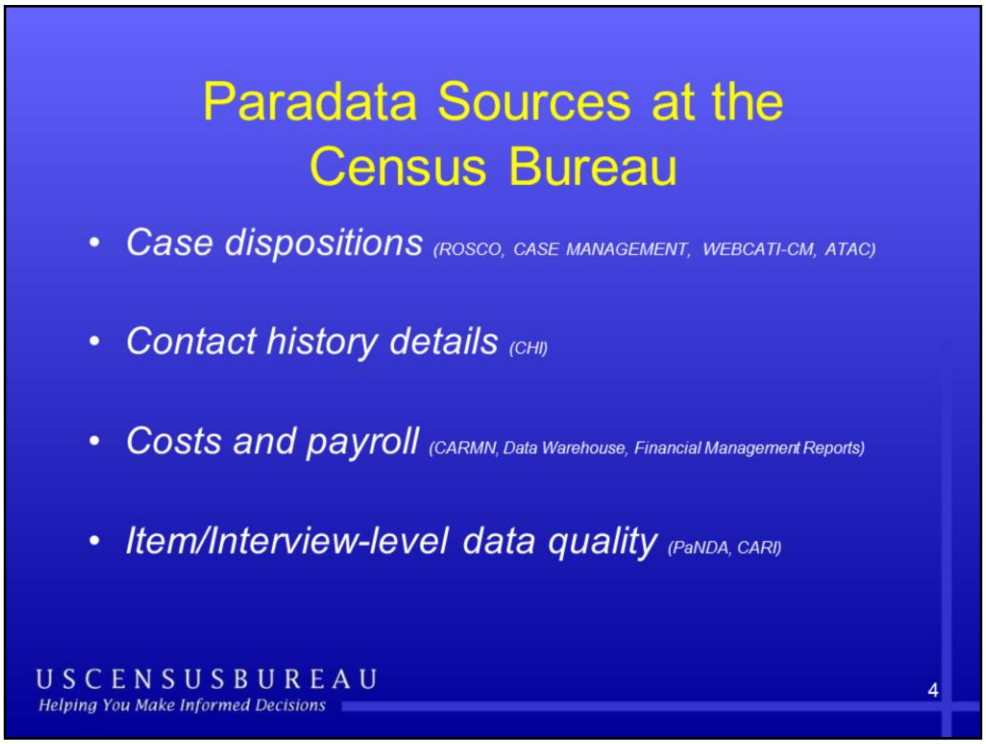
CARMN serves as a **cost tracking system for field** as well

PANDA is used to track **individual respondent performance** on such measures as **item nonresponse rates and speed of completion**.

These systems **serve their individual purposes** well, but when **ad hoc reports are needed**, integration is often **done by hand**. CPS uses **Excel to produce daily cost and**

production reports that combine data from other reports produced by **ROSCO and CARMN**.

Currently at Census, there is **no unified dashboard system (or data access system)** that allows a manager or data analyst to look at **different types of paradata in one place**, **drill-down to different levels of specificity**, and create **ad hoc reports or run statistical analyses**.



To look at it another way, here are some types of paradata available at Census and their respective sources…and I think I've left some off this list.

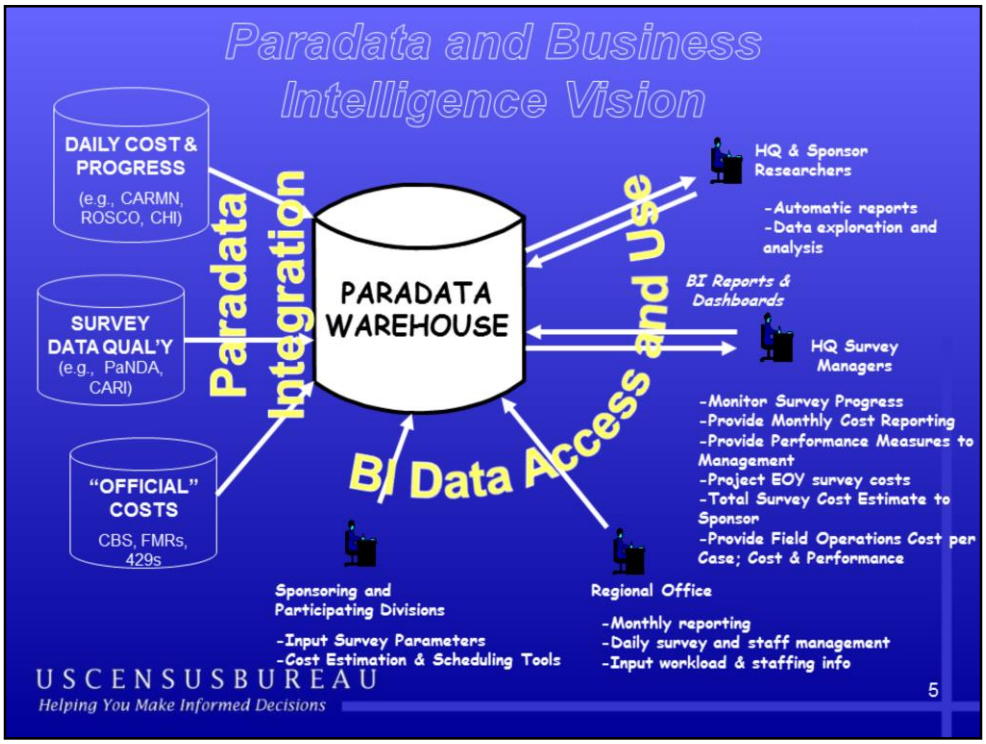
Not only are there **several different sources**, but **some types of data can be gathered from more than one source**.

Integration of paradata sources requires detailed knowledge of the contents of each of these sources. This is something we're gaining on a daily basis.

Its clear that **paradata repositories** at Census were **not developed with integration or data analysis in** mind. They were developed to **serve a specific task** of a **specific area of project management**, such as budgeting and cost reporting, sample management, or staff management. This situation isn't unique to Census, but shear volume of data and size of the organization makes their integration a unique challenge.

****

ASIDE: Each of these represents an independent system, database, staff, etc. The lack of integration is not just technological (i.e., not just the data) but organizational too.



This is a **major over-simplification** of the situation but describes relationships between data and users under and idealized situation that we are trying to attain.

The paradata sources just discussed are on the **left side.** If you think of these **canisters as databases**, there would be **several smaller canisters** around each of the 3. would be integrated through a paradata warehouse. This could be a physical consolidation of the files from each of the data sources, or some sort of relational query system.

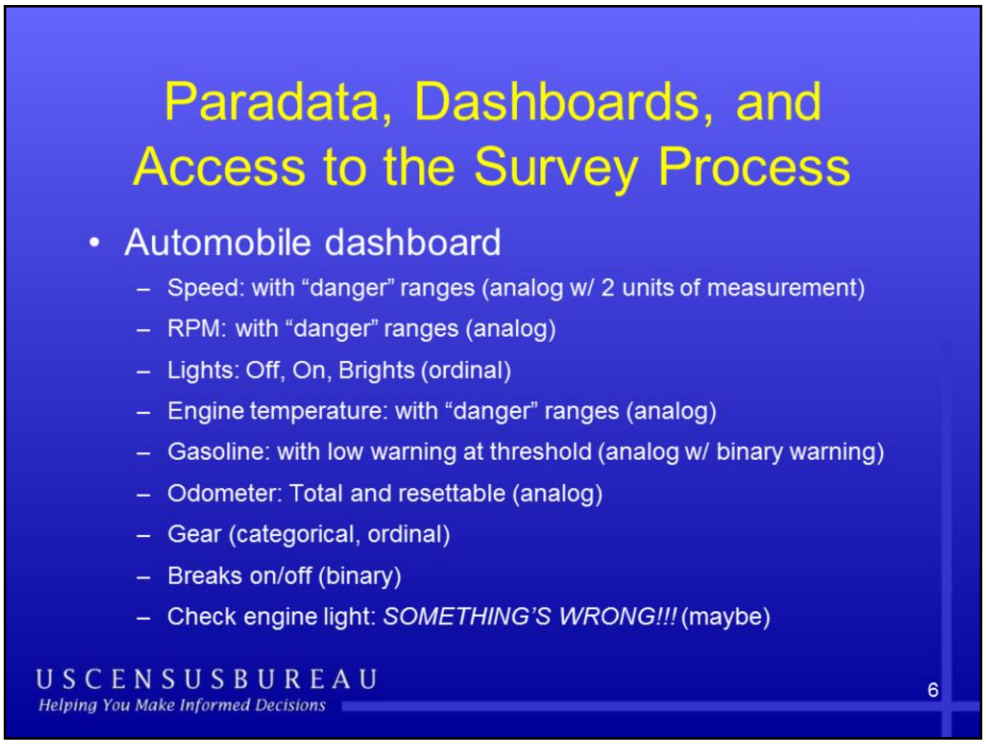
On the **right side** of this figure you can see all the individual data users and their unique roles. Some will simply be **submitting data or requirements** to the systems. Some will also **passively receive reports** from the system. Yet **other users (managers and analysts)** will be interacting directly with the data.

This integration, combined with the access of paradata would **create an environment oriented around Business Intelligence**, This term that has gained popularity in the private sector to describe business practices that are **driven by analysis of data** and **decision rules**.

Rather than a trendy buzz word that won't do much to change project management practice, we're **thinking of Business Intelligence as a matter of technological affordances (e.g., access to data)** that make **managers jobs more interactive and**

individualized, and thus lead to more **proactive and efficient management practices**. Further, data are more accessible to methodologists and analysts for scientific research.

So you might wonder how these users would access integrated paradata.



This brings us to the notion of **dashboards as one way to disseminate data in reports that can be both fixed and flexible**. You can think of dashboards as **automated combinations of reports that are displayed to data users and managers**.

Take a minute to think about that term "dashboard". What's on the dashboard of your car?

There's something that tells you how fast **you're going**

There's something that tells you how fast **you *engine* is going**, and if it's in a danger zone.

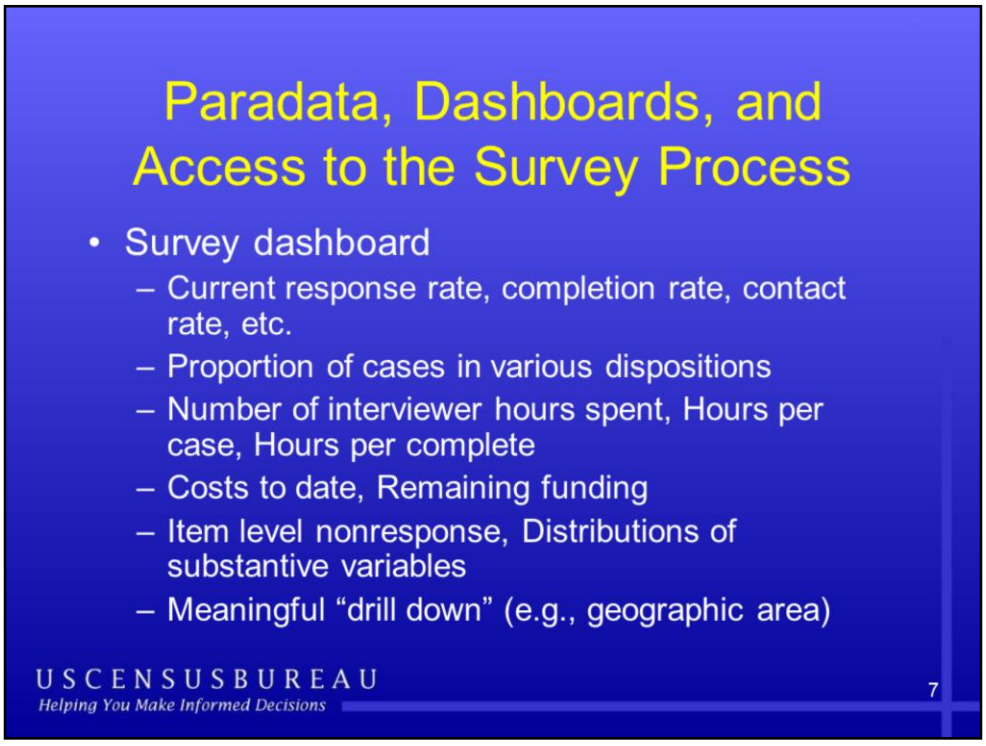
There's a little light that comes on when **something is wrong.**

If you think about it, there are a lot of data contained in this relatively small dashboard.

****
****

Note: For those who commute…On Metro there is a train schedule, a real-time updated table tells you when the next train is arriving and where it's going, lights

blink as the train is about to arrive, as-needed announcements tell you of problems and delays, pre-recorded announcements remind you of how to be a “good rider”



Now think about what you would want to see to **make your SURVEY runs** the way you want it to run.

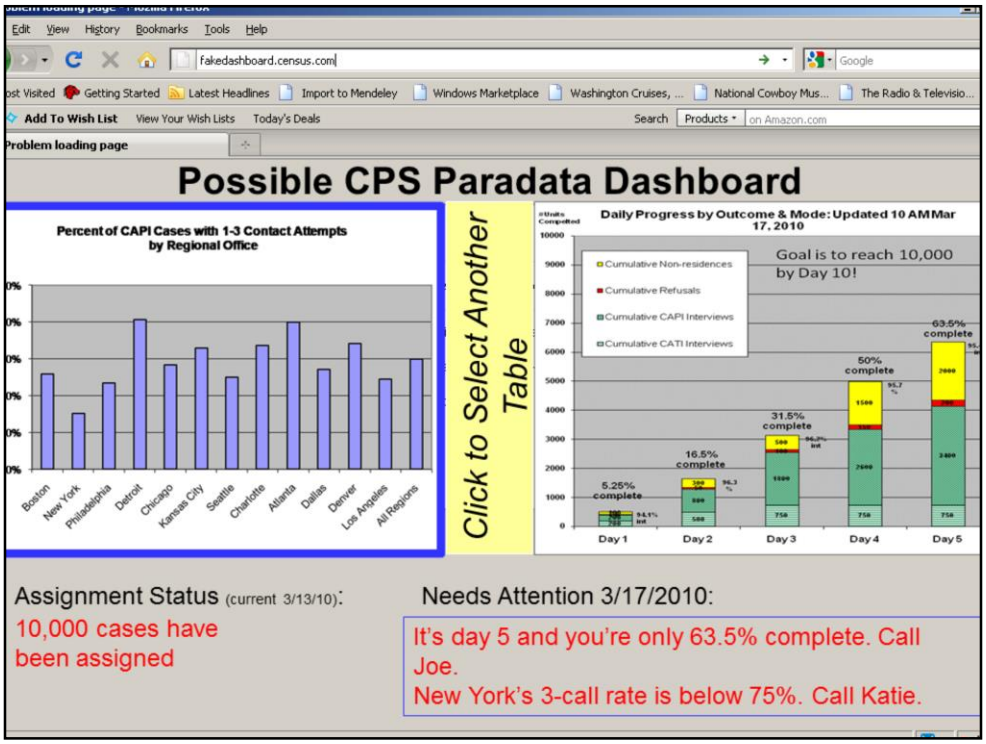
You might want to see **current response and cooperation rates**

You might want to see **how many cases are in each disposition**

You might also want to see information about **interviewers and costs**, as well as **interview-level data** or **survey estimates to date**

You might want to compare these estimates to **previous values**, **budget expectations**, or **statistical tolerance limits.**

Benchmarking in short



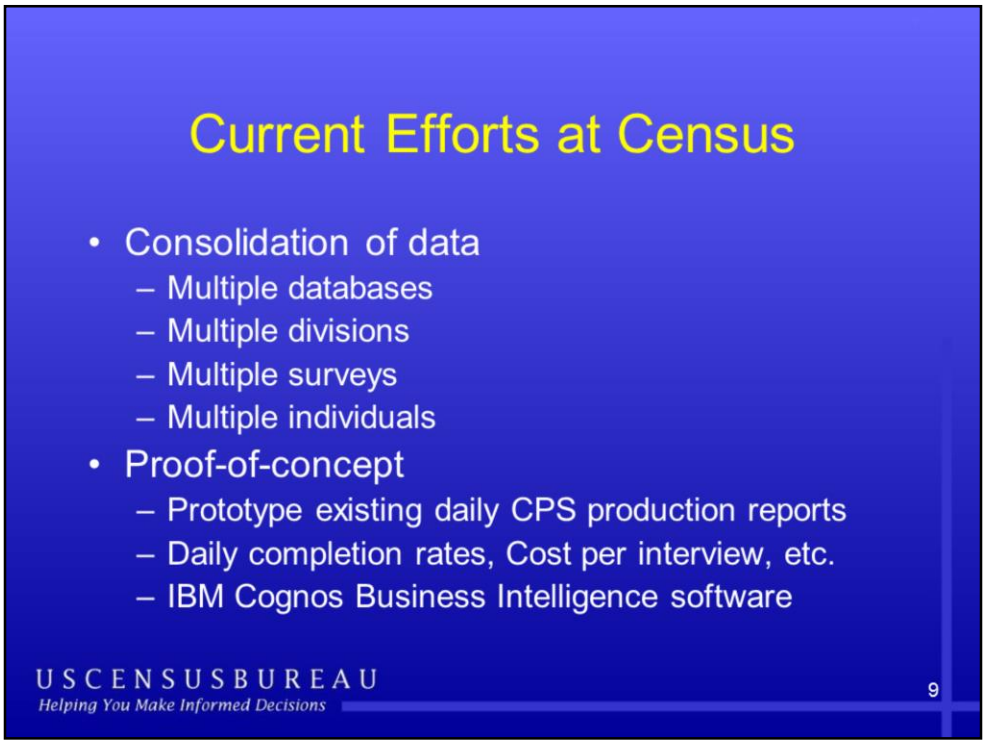
This is a very unattractive mock-up of a potential dashboard.

Two points should be noted…**One**, that we you would want to **display charts that would be updated as frequently** as data become available.

But we can also update **project status milestones, and include messages** (to our staff or to ourselves) that would be automatically generated based on the progress data.

Different graphs and figures could also be selectable or created ad hoc.

Something that isn't shown here would be the **ability to comment back to a manager or co-worker** when a task has been completed, or to discuss some component of a tasks. It seems like only a technical issue to link such a dashboard to an email or IM system.

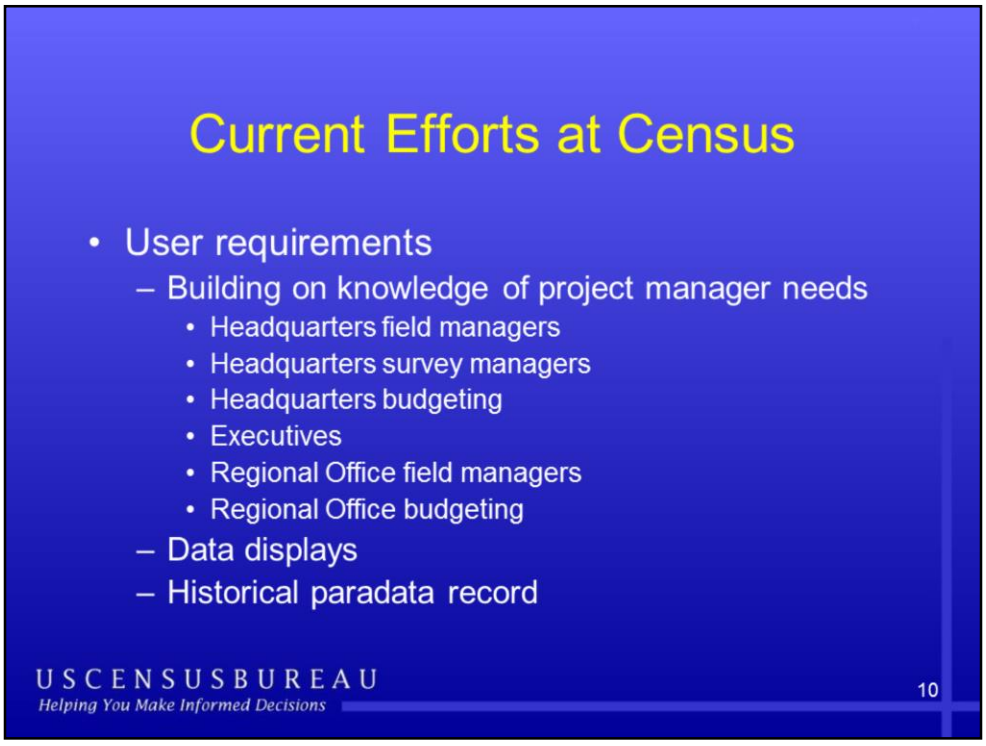


To summarize or progress and current work, we're actively involved in the **consolidation and integration of paradata** with the goal of producing some of the **daily progress reports produced** by our field division for CPS.

We've chosen **IBM Cognos for this**.

The reports themselves are fairly simple, **completion rates**, **costs per case, and miles per case by day in tabular form**. They also benchmark to the same point in the project last month and the same point in the project from the previous year.

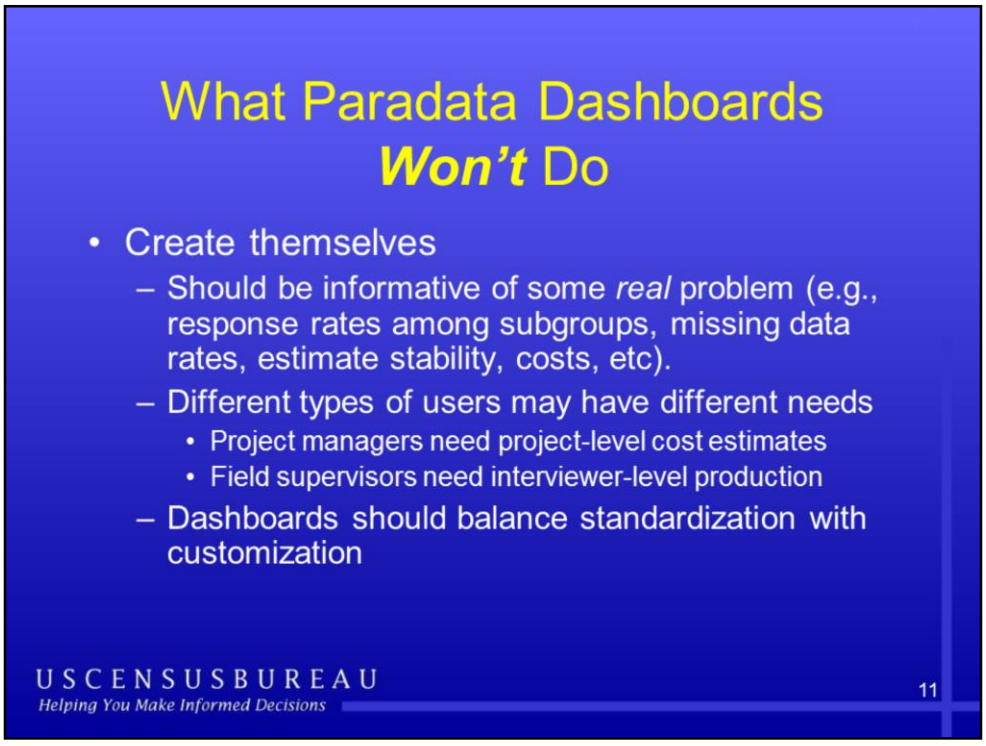
Automation is part of the proof of concept, which for us means being to create charts daily with current data. This proof of concept test is of the entire dashboard process.



We're taking what we know about the survey process and **relevant management and planning needs**, which include some current reporting practices, and **moving toward a more streamlined reporting practice**. This is our **long game**.

Our knowledge base will likely grow as we make progress. **Different users have different data needs**, and will likely want to see **different standardized reports and displays**.

As we build these reports we'll be retaining a **historical paradata record as well**.

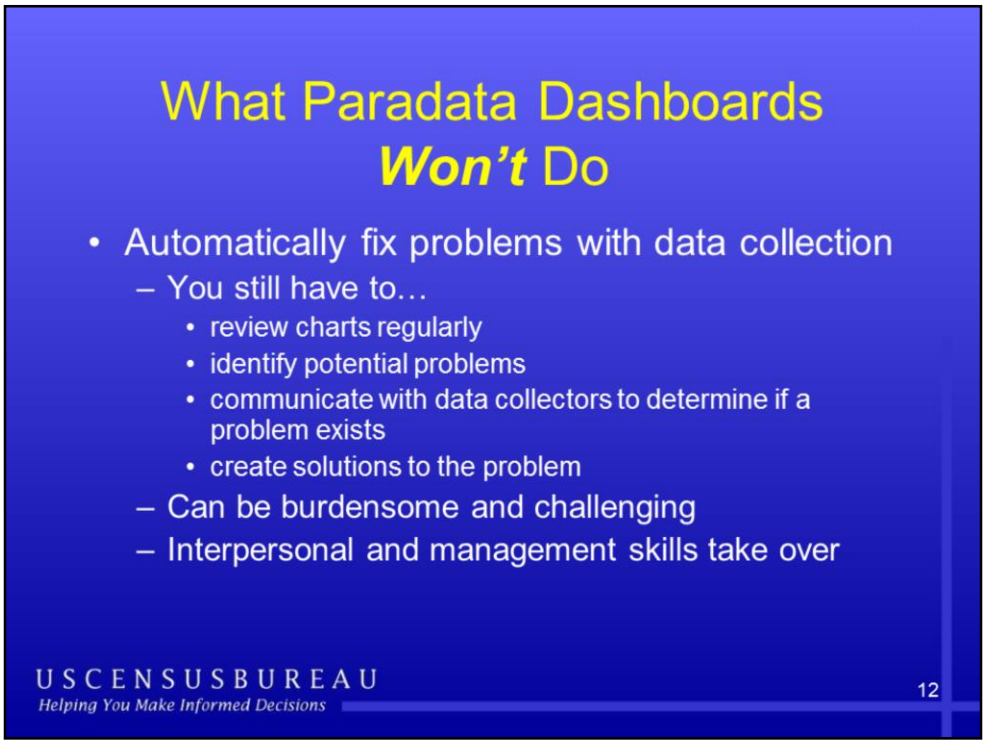


But so that we don't leave with our view too rosy, let me just make a few comments about the limitations of dashboards based on integrated paradata.

Like any other **statistical graph,** dashboards must be **informative of some *real* problem of interest** (e.g., response rates among subgroups, missing data rates, estimate stability, costs, etc).

They **won't create themselves,** nor will **"off the shelf"** templates likely work well. That doesn't mean you can't use off-the-shelf software, but you will probably have to build your own reports.

Needs of users may also vary, and as a result dashboards should be created to be **both standard (in some minimum way) but also maximally flexible**



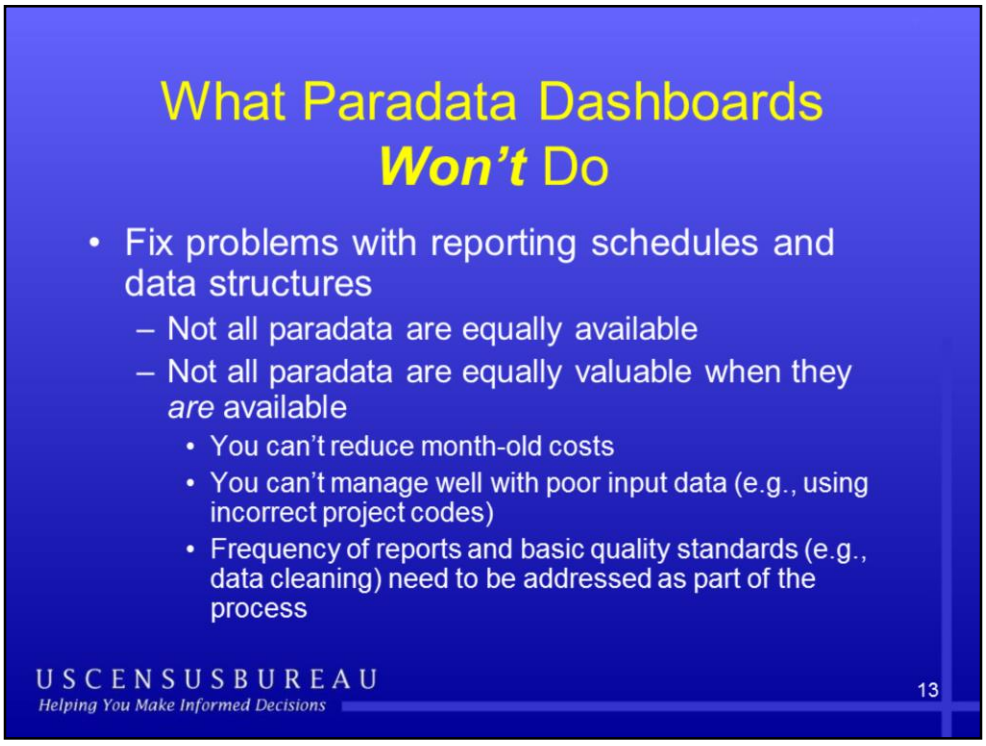
While dashboards can automatically deliver data, they **don't automatically solve** problems.

MENTION BOB'S CHARTS AND NSFG HERE

-DAILY PRODUCTION AND REVIEW: BOB'S OFFICE WITH 50+ FIGURES POSTED ON THE WALL DAILY.

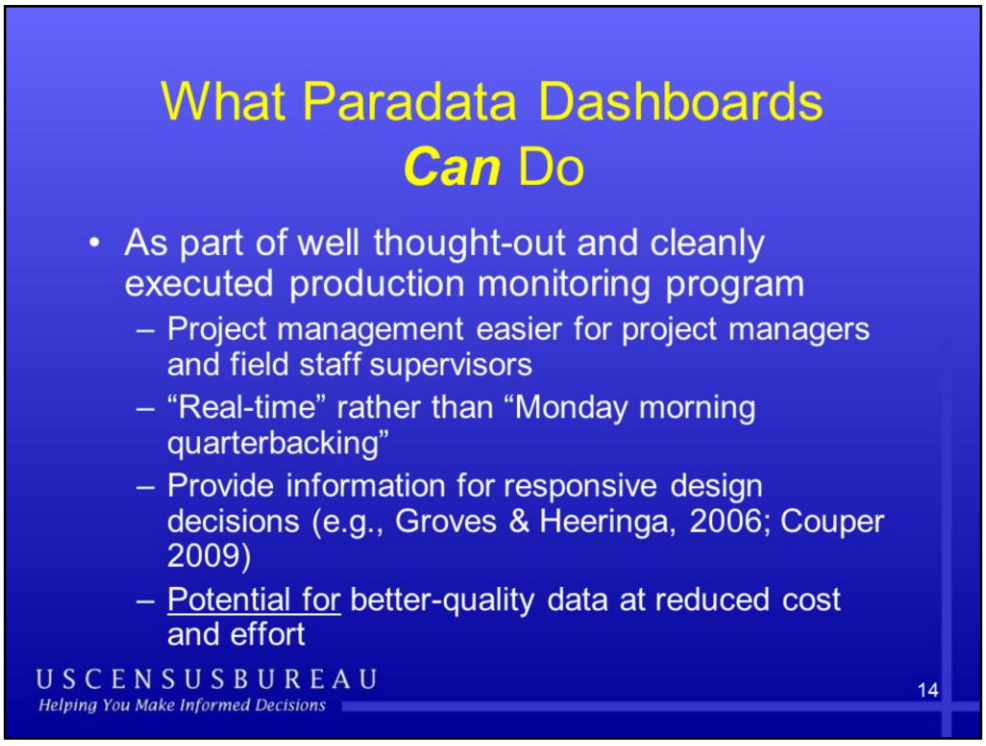
-2 YEARS PREP BEFORE PRODUCTION

Where the data leave off, **good old-fashioned project management** takes over.



Dashboard reports are only as good as the input data. If costs or progress are out of sync with each other or out of date, decisions need to be made about their integration.

Users should be notified of recency and limitations of data

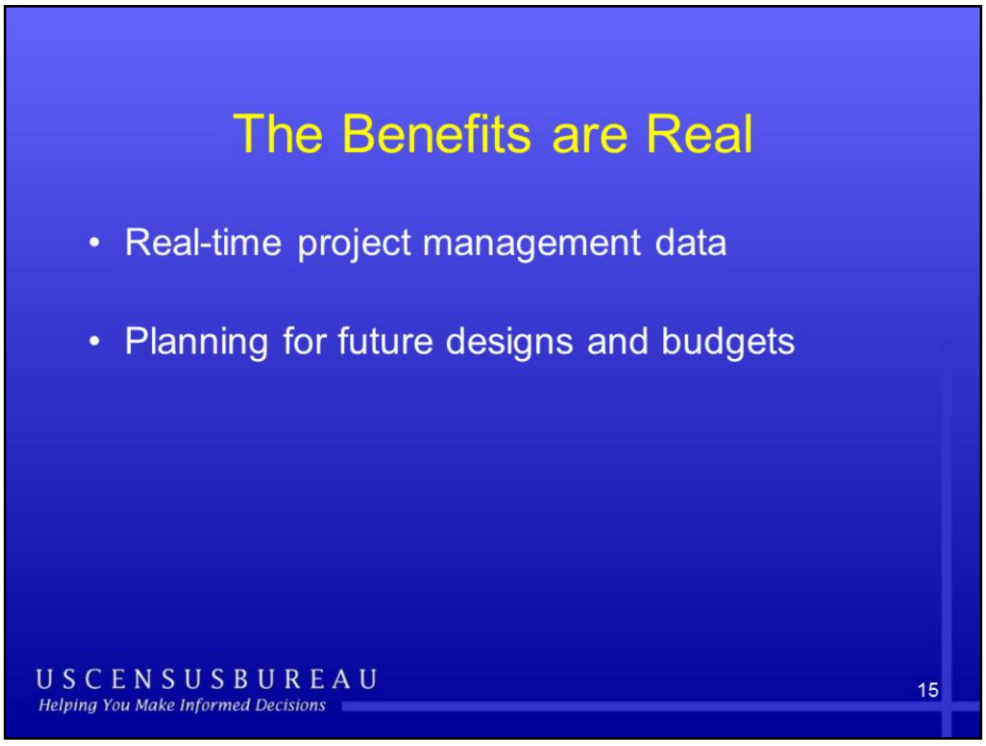


However, distributing paradata through dashboards has vast potential to improve project management practices if it's taken as part of a comprehensive project management and survey error research program.

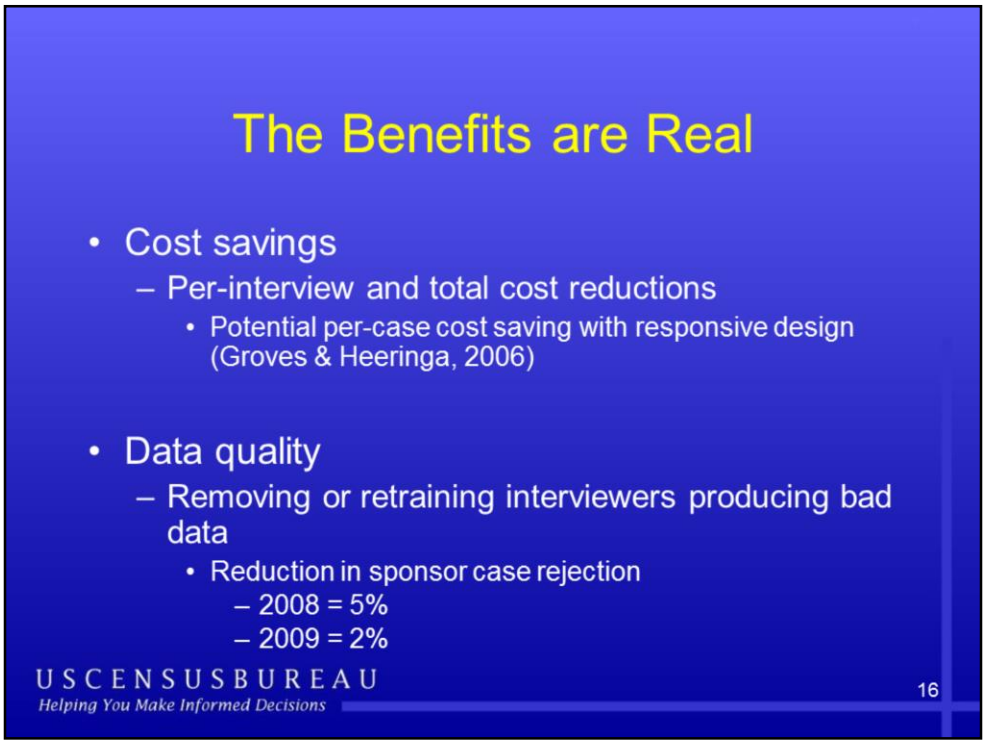
Real-time, automated data access gives managers **more time to manage** and **makes management proactive**

This type of paradata access can also **provide information for more intensive and explicit responsive design techniques**

Of course the **ultimate goal** of this approach is improve upon **data quality while reducing cost**.



More up-to-date and flexible data for project management alone is a benefit, as well as access to these data for project planning and budgeting.



Research out of Michigan on NSFG has demonstrated cost savings when a responsive design is used.

Work in-house has shown that proactively looking at item-level data quality produces better data for sponsors. In this case the rate at which cases are rejected as being unusable.

Thank you

matthew.e.jans@census.gov
kathleen.p.creighton@census.gov
christopher.j.laskey@census.gov

USCENSUSBUREAU
Helping You Make Informed Decisions

17