Is your big data messy? We're making an app for that

February 16 2017, by Cory Nealon

```
        ga.type = 'text/j
        ga.async = true;
        ga.src = ("https:" == document.location.protocol ? "https://ssl
        var s = document.getElementsByTagName("script")[0];
        s.parentNode.insertBefore(ga, s);
    })();
</script>
<?php
    if (is_singular() && get_option("thread_comments")) {
        wp_enqueue_script("comment-reply");
    }
?>
<?php wp_head(); ?>
</head>
<body <?php body_class(); ?>>
    <div id="header">
        <div class="wrapper">
            <h1>
            <?php if (is_front_page() && $paged < 2) : ?>
                <img src="<?php bloginfo("template_url"); ?>/images/logo.png" alt="<?php bloginfo() ?>" />
            <?php else : ?>
                <a href="/" title="Root"><img src="<?php bloginfo("template_url"); ?>/images/logo.png" alt=
            <?php endif;?>
            </h1>
            <form id="search" method="get" action="/">
                <div>
                    <input accesskey="s" type="text" id="s" name="s" />
                    <input type="submit" value="Find" />
```

Credit: University at Buffalo

Like a teenager's bedroom, big data is often messy.

Malfunctioning computers, data entry errors and other hard-to-spot problems can skew datasets and mislead people—everyone from data scientists to data hobbyists—trying to draw conclusions from raw data.Vizier, a software tool under development by a University at Buffalo-led research team, aims to proactively catch those errors.

The project, backed by a $2.7 million National Science Foundation grant, launched in January. Like Excel and other spreadsheet software, Vizier will allow users to interactively work with datasets. For example, it will help people explore, clean, curate and visualize data in meaningful ways, as well as spot errors and offer solutions.

But unlike spreadsheet software, Vizier is intended for much larger datasets; it will be used to examine millions or billions of data points, as opposed to hundreds or thousands typically plugged into spreadsheet software.

"We are creating a tool that'll let you work with the data you have, and also unobtrusively make helpful observations like 'Hmm... have you noticed that two out of a million records make a 10 percent difference in this average?'" says Oliver Kennedy, PhD, assistant professor of computer science and engineering at UB, and the grant's principal investigator.

Co-principal investigators include Juliana Freire, professor of computer science and engineering at New York University, and Boris Glavic, assistant professor in the Department of Computer Science at the Illinois Institute of Technology. The award is from NSF's Data Infrastructure Building Blocks (DIBBs) program.

For years, companies like Google, Microsoft and Apple have utilized big data to improve their products and services. That same power is now spreading to the masses as government agencies in the United States and elsewhere publish massive amounts of public data on the internet.

For example, New York City and the federal government have open data portals making it possible for anyone with an internet connection to download information and ask questions about their government. When properly used, these portals can shed light on issues relating to health code violations, discrimination, bias and other matters, Kennedy said.Vizier will be released as free, open-source software.

"We want to make it easier for data scientists—and eventually data hobbyists—to discover and communicate not only what the data says, but why the data says that," he said.

Provided by University at Buffalo

Citation: Is your big data messy? We're making an app for that (2017, February 16) retrieved 7 July 2024 from https://phys.org/news/2017-02-big-messy-app.html