

Big Data – Bad Data

Rayna Stamboliyska, PhD

Paris Descartes University
RS Strategy

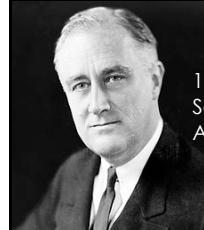


XXXL
Issues

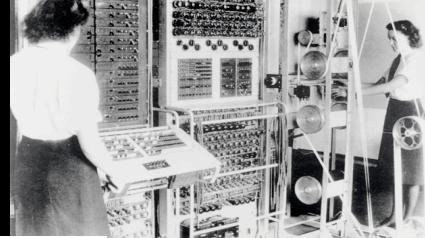
XXXL
Data



"Is there anywhere on earth exempt from these swarms of new books?"



1935-1937,
Social Security
Act



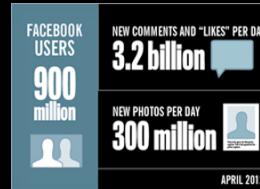
1943-1945,
'Colossus' &
Enigma

1 Introduction

Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (in

1997: first mention of 'big data' in
Application-Controlled Demand Paging
for Out-of-Core Visualization (NASA)

...



No, Big Data isn't new... but

What is it?



Maximizing computation power and algorithmic accuracy

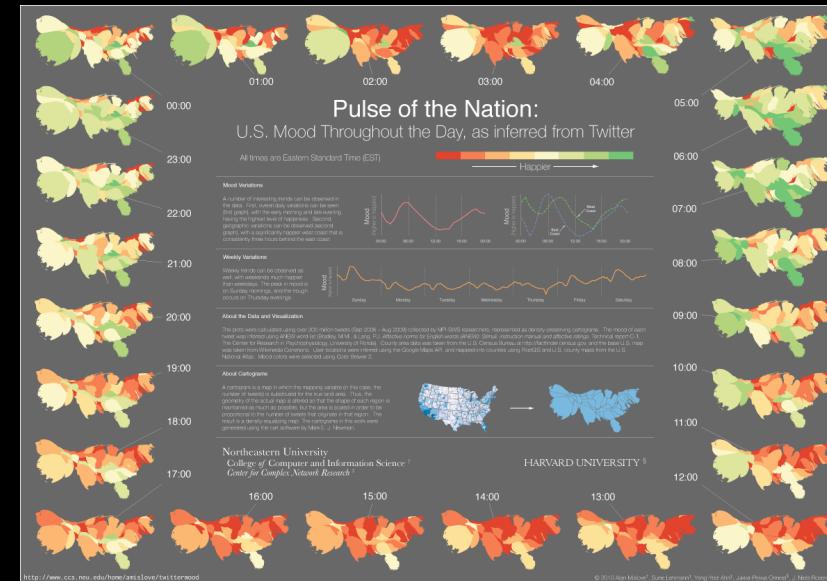


Drawing on a range of tools to analyse and compare large datasets

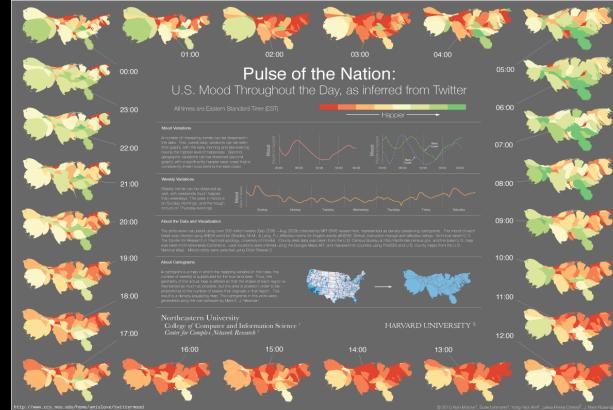


Believing large datasets give greater objectivity and accuracy

No, Data doesn't speak of itself



Numbers never lie... o rly?



- * 51% of web traffic is non-human
- * there are 30 million fake Twitter accounts

Yes, Big Data discriminates between social groups

The image shows two screenshots side-by-side. The left screenshot is from the International Journal of Communication (IJOC) website. It features a colorful logo 'I J O C' and the text 'International Journal of Communication'. Below the logo is a banner with a red and black circuit board pattern. The main content area shows a journal article titled 'Data Mining Difference in the Age of Big Data: Communication and the Social Shaping of Genome Technologies from 1998 to 2007' by Peter A. Chow-White and Sandy Green, Jr. The right screenshot is from a Google search results page. It shows a patent listing for 'Dynamic Pricing Models for Digital Content' (US 20080154798 A1). The patent details include the publication number US20080154798 A1, the publication date 29 jun 2008, and the priority date 22 dec 2006. Inventors listed are Duane R. Vaz and Yahoo! Inc.

International Journal of Communication

HOME ABOUT LOGIN REGISTER SEARCH CURRENT ARCHIVES SUBMIT

EDITORIAL BOARD FOUNDING EDITORS

Manuel Castells
Larry Gross
EDITOR
Larry Gross
Larry Gross, International Journal of Communication, Inc.

Data Mining Difference in the Age of Big Data: Communication and the Social Shaping of Genome Technologies from 1998 to 2007
Peter A. Chow-White, Sandy Green, Jr.

ARTICLE TOOLS

Google

Brevets

Dynamic Pricing Models for Digital Content
US 20080154798 A1

RÉSUMÉ

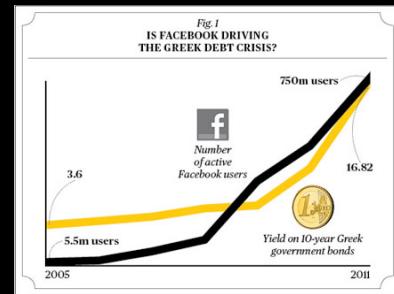
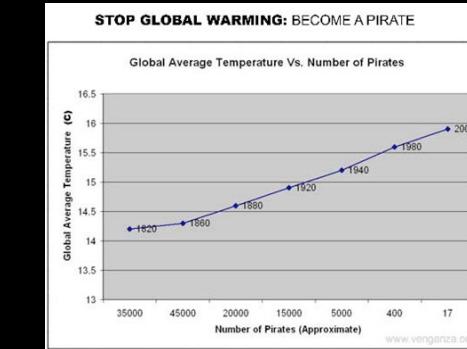
Dynamic pricing models which facilitate efficient distribution of digital content online. Particular implementations of the invention dynamically base pricing for digital content on relatively current, aggregated information regarding Internet user behavior and preferences, such as search query and/or page hit logs. Some implementations of the present invention are directed to pricing digital content based on the inherent properties of digital content and the mechanics how electronic files are typically distributed on the Internet.

Numéro de publication: US20080154798 A1
Type de publication: Demande
Numéro de demande: US 11/615,602
Date de publication: 29 jun 2008
Date de dépôt: 22 dec 2006
Date de priorité: 22 dec. 2006
Autre référence de publication: EP209821A4, WO2008079817A1
Inventeurs: Duane R. Vaz
Cessionnaire d'origine: Yahoo! Inc.
Référencé par (7), Classifications (10), Legal Events (1)
Liens externes: USPTO, Cession USPTO, Espacenet

Scientists are allowing their assumptions about race to shape their big-data genomics research

If your past buying history indicates you are more likely to pay top euro for trousers, your starting price the next time you shop for fooclothes online might be considerably higher.

Big Data Fundamentalism?



There's a word for that: **apophenia**!

... Does big data change the erroneous 'correlation = causation' dynamic?

Math majors, rejoice!

*"With big data techniques, you can get
much closer to being able to predict
causal relations."*

I beg to disagree

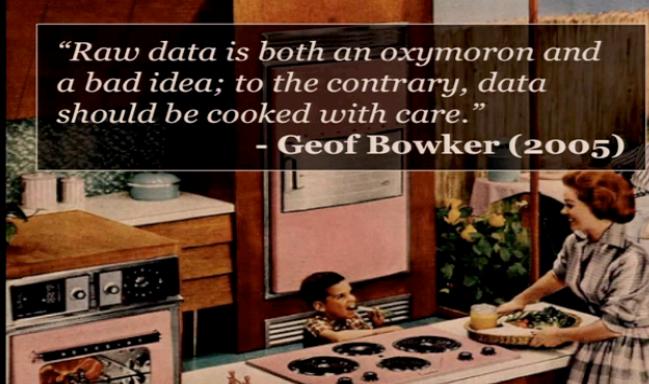
Your question and your data are not
random entities.

Big Data needs several steps:

- of preparation

And...

If data are somehow subject to us, we are also
subject to data



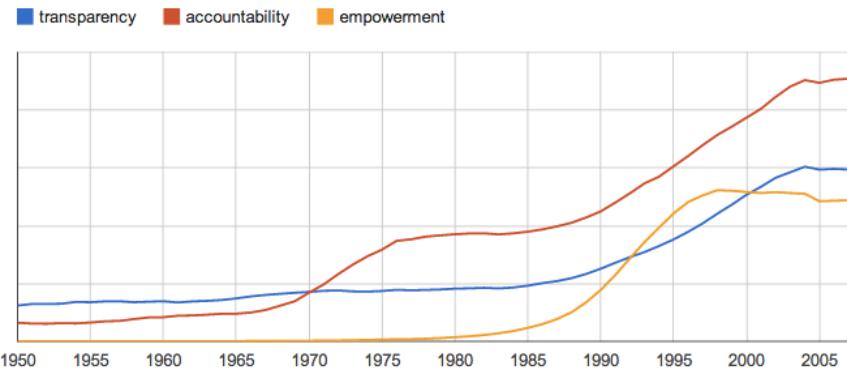
Open Data as sovereignty.

Romain Lacombe
Etalab (data.gouv.fr)

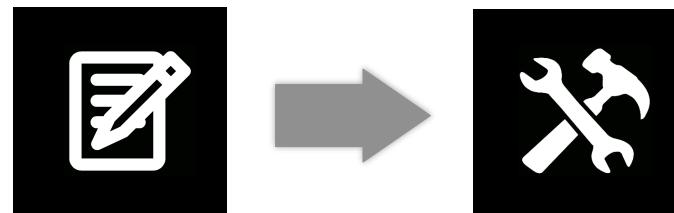
Code is **law**
Architecture is
politics

Data is
democracy

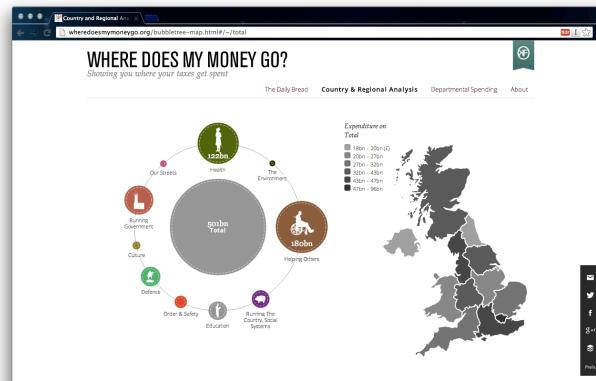
Answer to rising **aspirations**



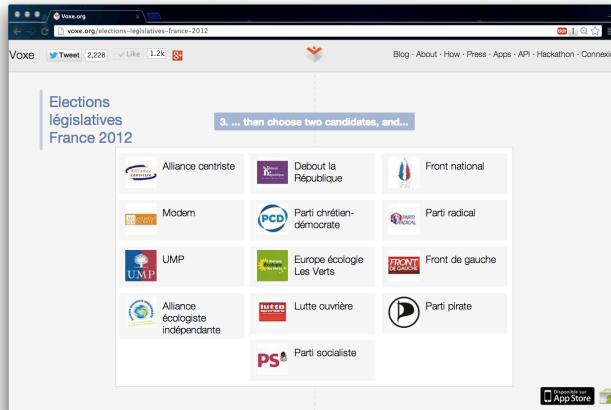
Tool for **empowerment**



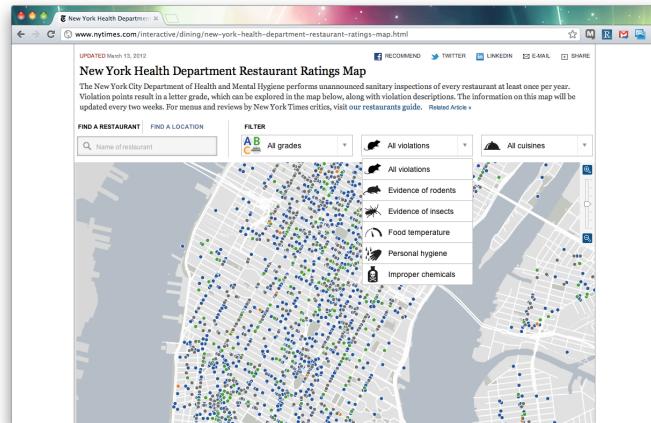
Perspective on collective action



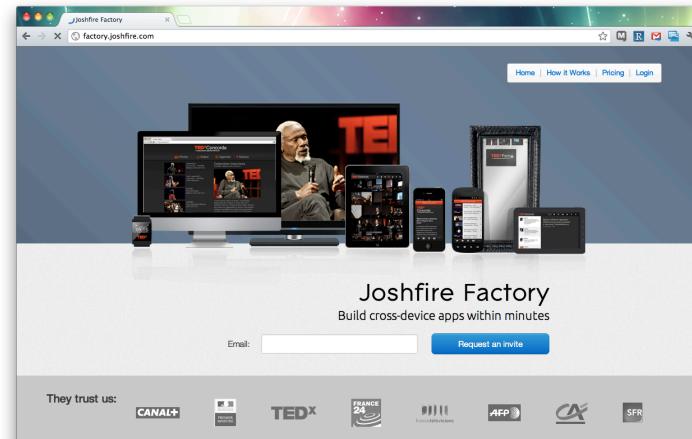
Informing individual choices



Transparency as regulation



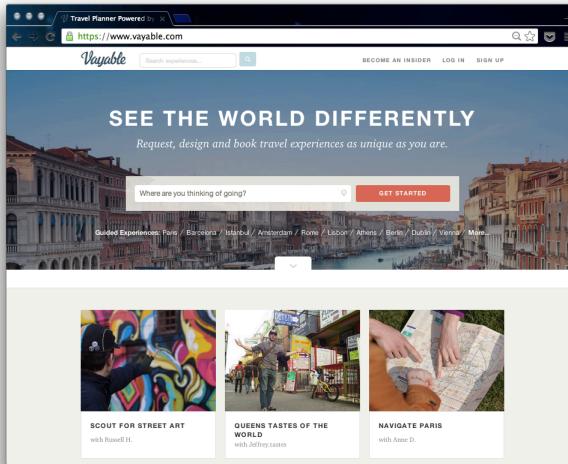
Innovation as reinvention



Sharing economy



Shift to experience

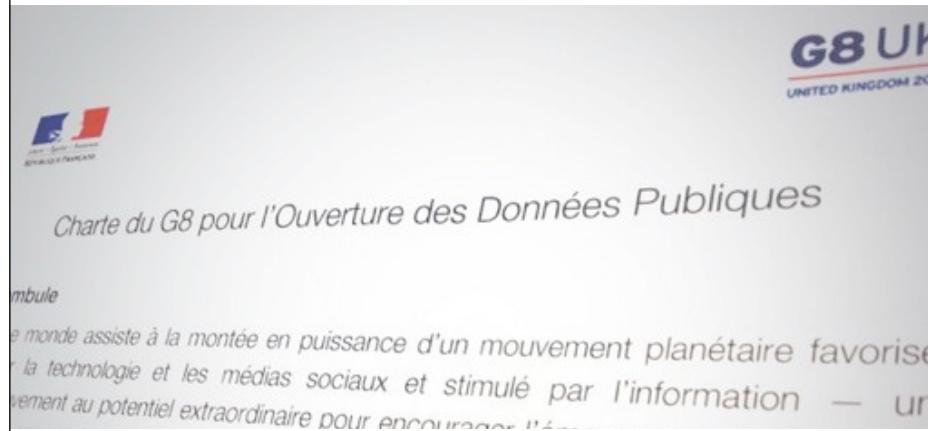


Connected cities and policies



Governing
data?

Open government data



From repository to open platform



Openness drives innovation



enigma

TOURISME
CONNEXION'04



KelQuartier
Trouver le quartier où habiter

TRANQUILIEN

CubicWeb

Innovation is **sovereignty**



Code is **law**
Architecture is **politics**
Data is **democracy**
Openness is
sovereignty

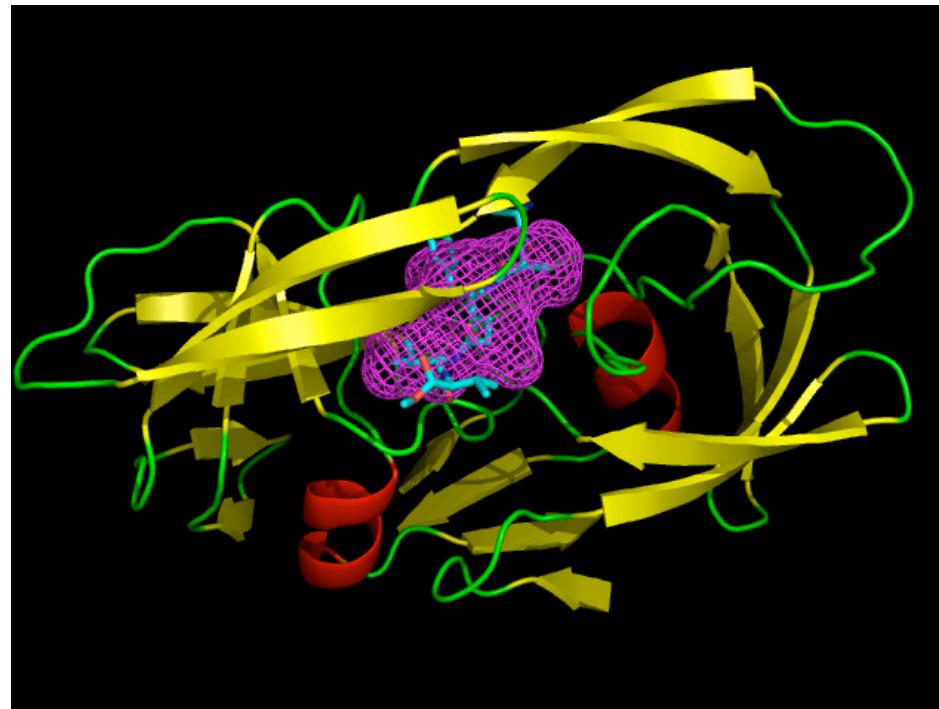
Thank you.

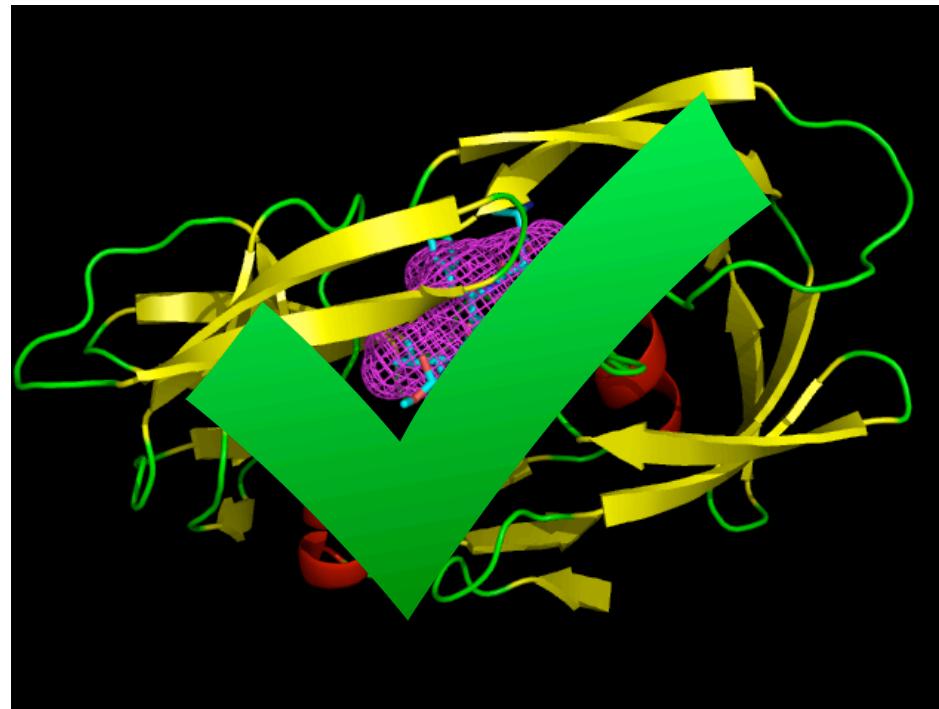
Romain Lacombe
Etalab (data.gouv.fr)

:snips

Rand Hindi, PhD
@randhindi









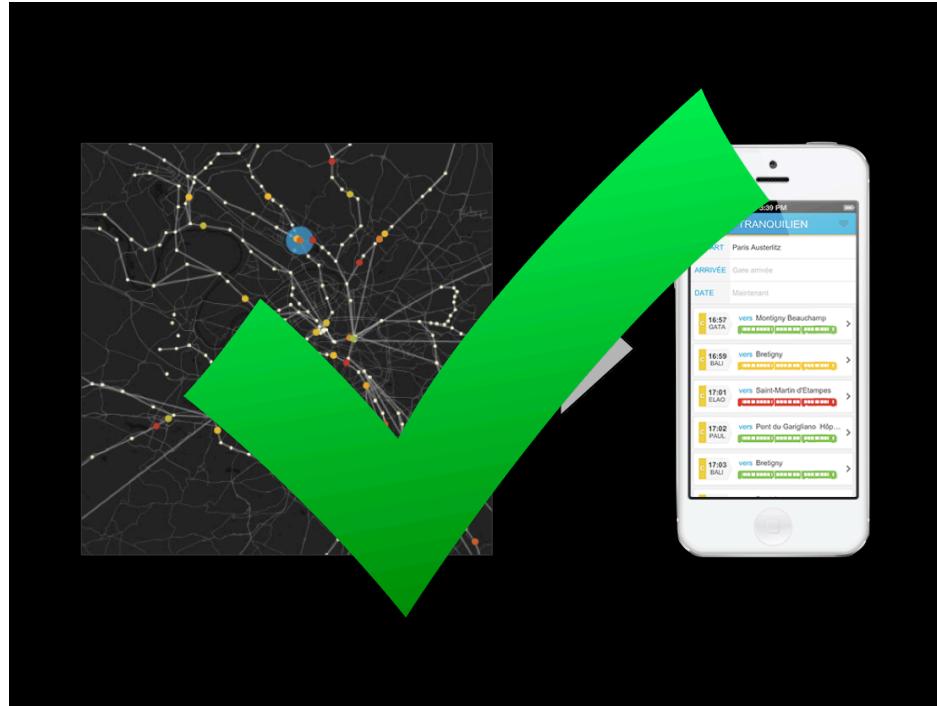


BEACON







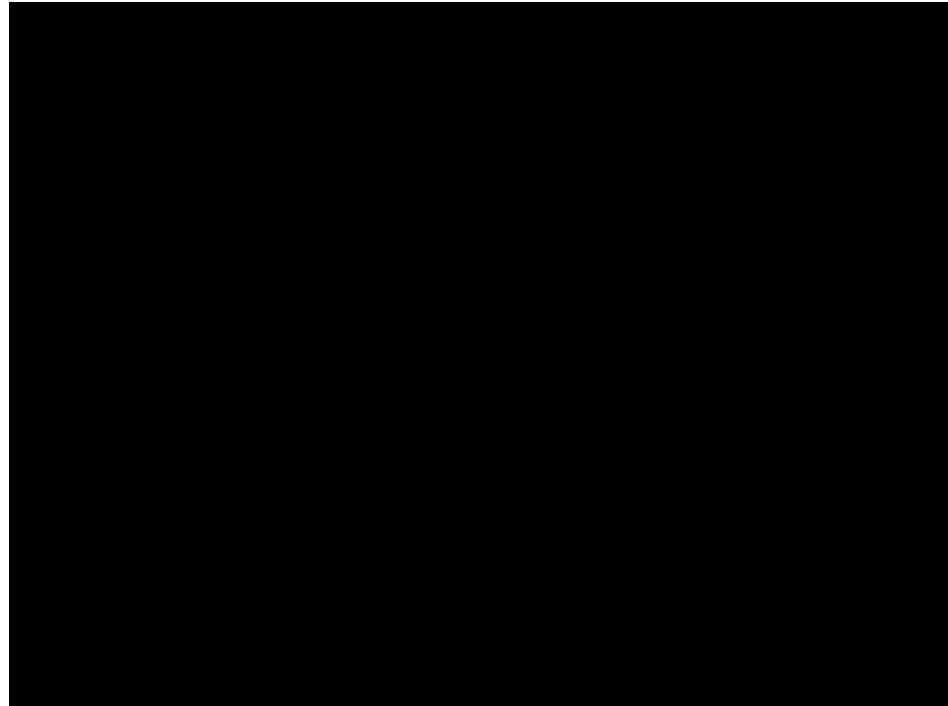


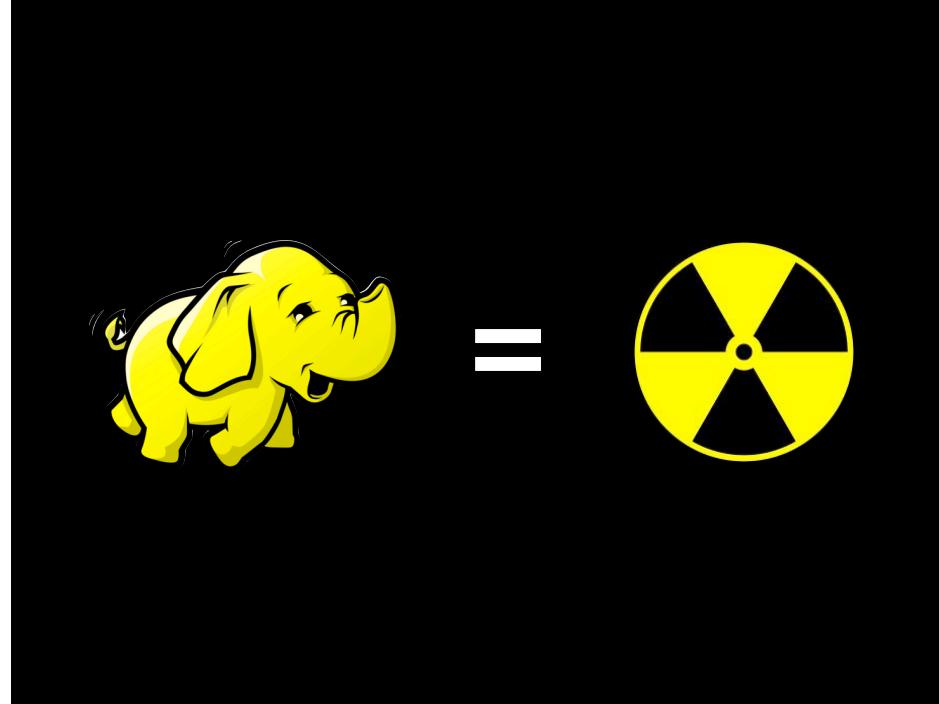












1. AGGREGATE

- 1. AGGREGATE**
- 2. ANONYMIZE**

- 1. AGGREGATE**
- 2. ANONYMIZE**
- 3. OPT-OUT / IN**

**WOULD YOU LIKE IT
IF SOMEONE ELSE
DID THAT TO **YOU** ?**



HAVE
FUN AND
DON'T
FUCK UP