571

THE BIG DATA REVOLUTION

VIKTOR MAYER-SCHÖNBERGER

@ Imperial College London

Introduction by Aija Leiponen

MANY data providers
Relatively UNKNOWN

acxiom ← 1BN $
CoreLogic
Datalogix
eBureau
id:analytics
INTELIUS

DATA BROKERS
Report by FTC 2014

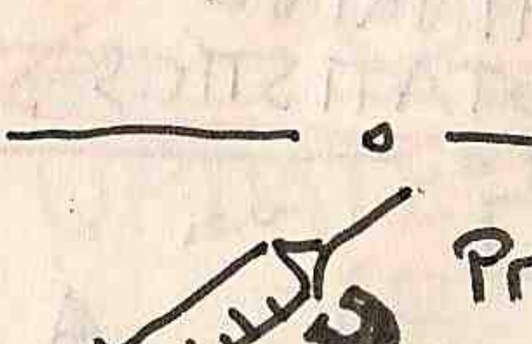
by Viktor
BIG DATA: A REVOLUTION
. . . . .
DELETE: THE VIRTUE OF FORGETTING IN THE DIGITAL AGE

THE 

Prediction case

→ TRADITIONAL APPROACH ←  2 WEEKS VS REAL TIME

→ GOOGLE APPROACH ← REAL TIME
(FOLLOW SEARCH TERMS)

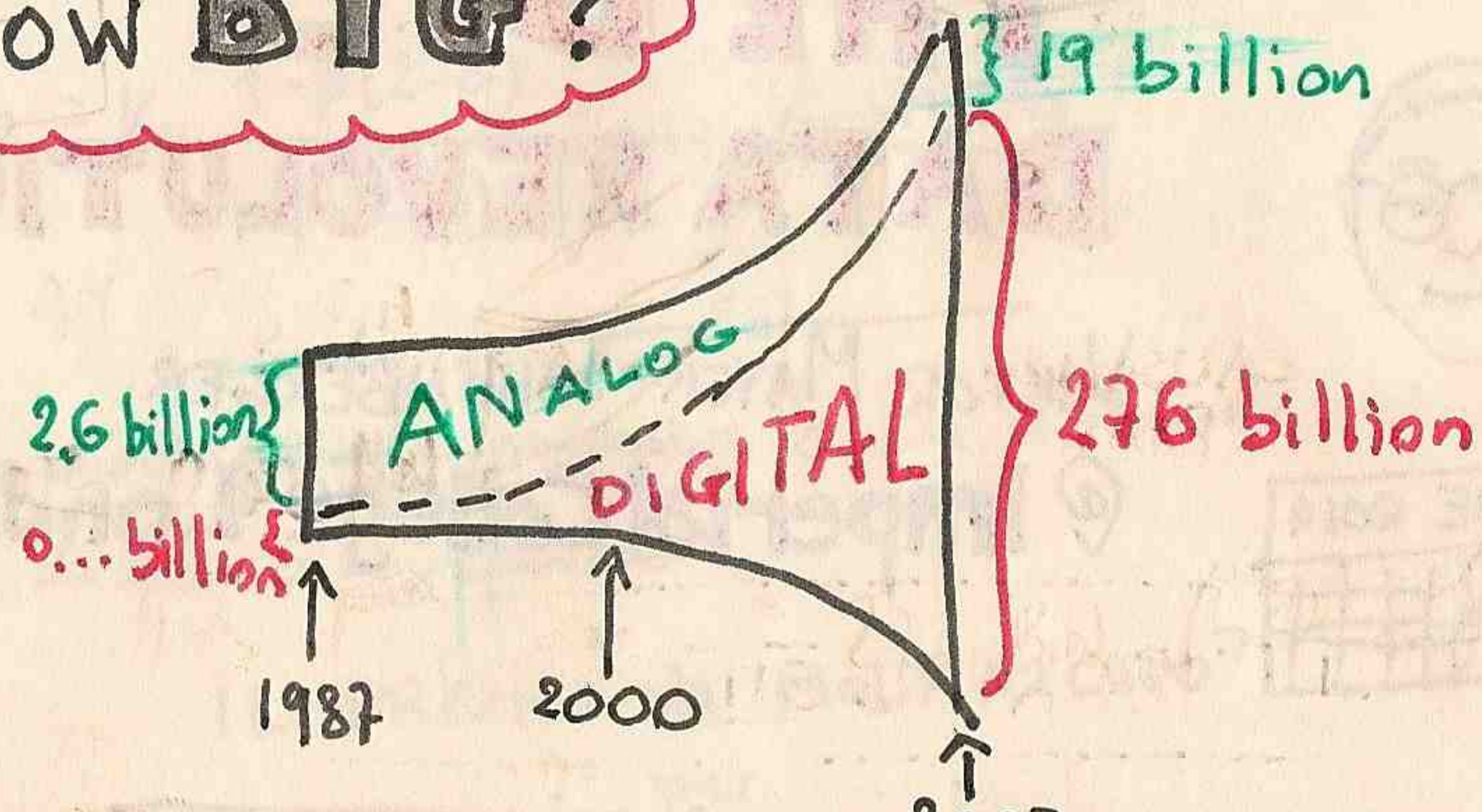


CDC* data over 10 years

*CDC = Center for Disease Control

@ Maurizio Naldi

How BIG?



THEN and NOW

DATA COLLECTION WAS EXPENSIVE

$ $ $

COLLECT DATA JUST AS NEEDED

01001

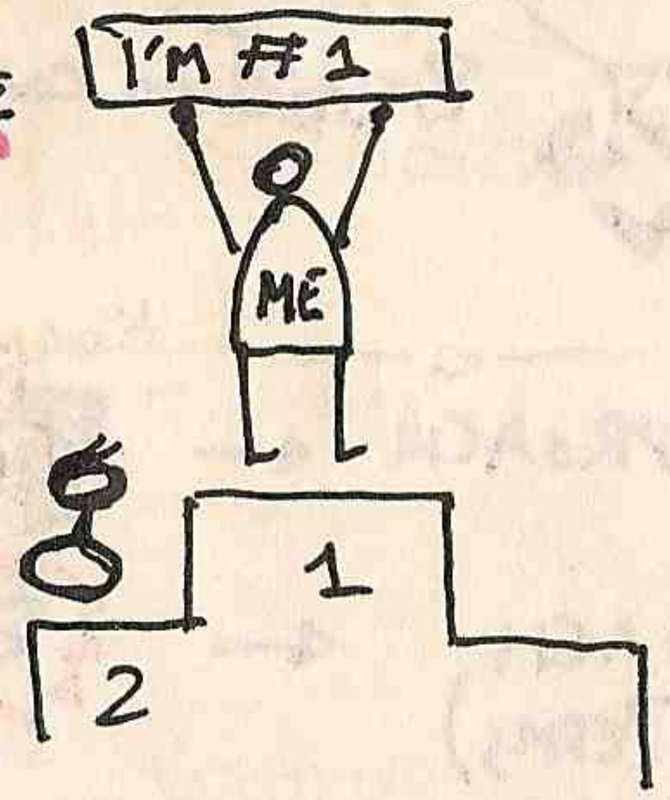
WHAT IF DATA COLLECTION IS CHEAP?

01100101101
10100011100
00100.....

COLLECT ALL YOU CAN

MAYBE RETHINKING STATISTICS

BACK TO THE FLU CASE

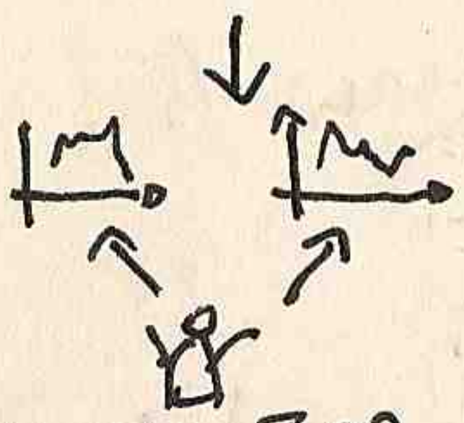


A CANADIAN RESEARCHER CLAIMS TO HAVE BEEN FIRST

BUT.... THE APPROACH WAS DIFFERENT!

Google

5 SEARCH TERMS



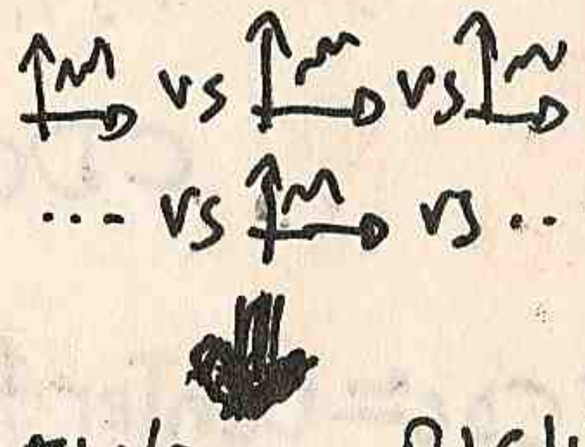
LOOK FOR CORRELATION

CLASSICAL APPROACH

THEORY
↓
DEFINE 1 HYPOTHESIS
↓
TEST HYPOTHESIS THROUGH DATA

NEW APPROACH

THEORY
↓
GENERATE MANY HYPOTHESES
↓
TEST ALL THROUGH DATA



FIND THE RIGHT TERMS

(45)

SMALL DATA

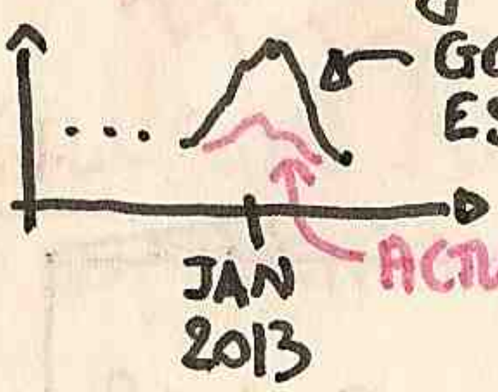
CAUSALITY BIAS

BIG DATA

SPURIOUS SIGNALS

THEORY IS STILL ALIVE!

BUT Google was WRONG in winter 2012-2013



⇒ OVERESTIMATE!

WHY?

...NO UPDATE...

UPDATE THE MODEL! (BAYES)

SMALL DATA

DATA COLLECTION was TARGETED

⇩

UNUSABLE AFTERWARDS

BIG DATA

REUSE DATA

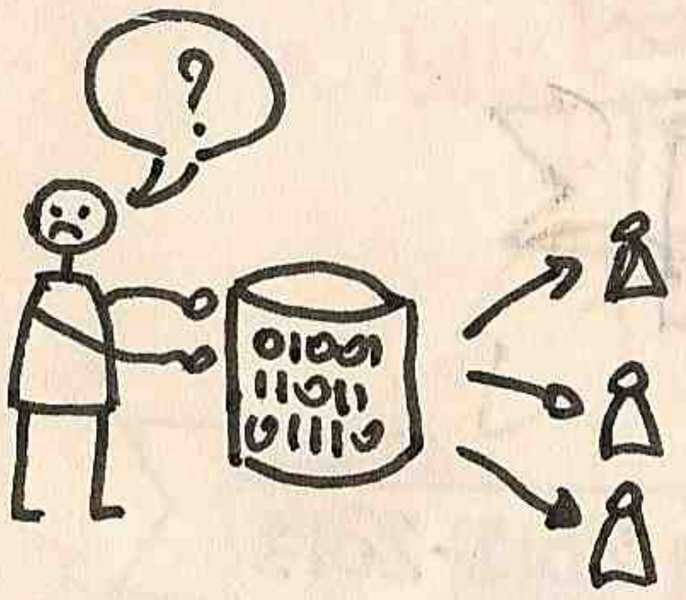
DATA ≠ CONTENT

REUSE MUST BE INCENTIVIZED

REUSE INCENTIVIZED by INTELLECTUAL PROPERTY LAW

PROTECTION ≡ EXCLUSION ⟹ MARKET $$

NO EXCLUSION ⟹ NO INCENTIVE TO PRODUCE

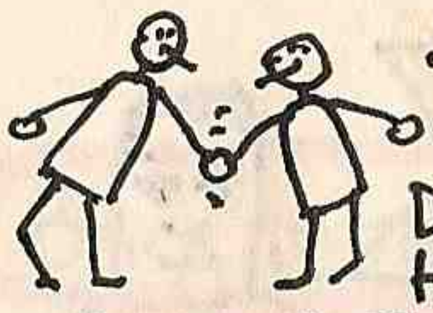


THE DATA HOLDER'S DILEMMA

TO SHARE OR NOT TO SHARE?

HOW TO INCENTIVIZE REUSE/SHARING?

LOW TRANSACTION COSTS ⟹



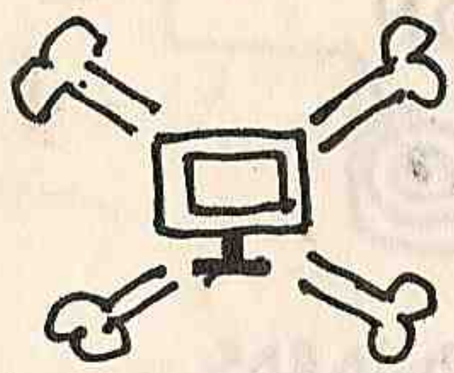
TRUST BY DATA HOLDERS (PERCEIVE FAIR SHARE) ⟹

CREATE DATA USER OPPORTUNITY

3 POSSIBILITIES

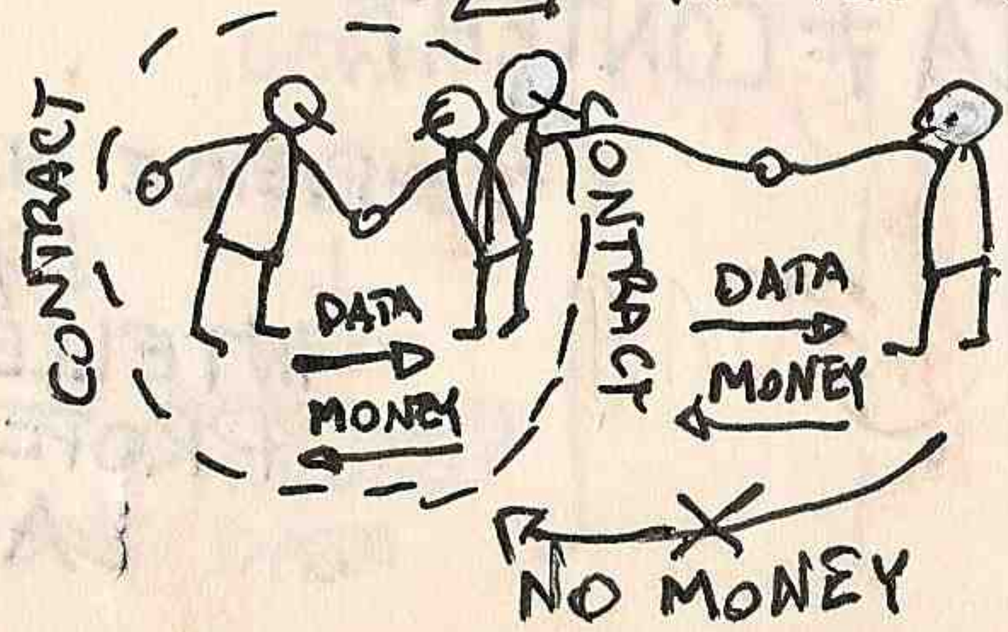
- CONTRACTUAL ARRANGEMENTS
- PROPERTIZATION
- EXCLUSION RIGHT SUI GENERIS

CONTRACTUAL ARRANGEMENTS ⟹



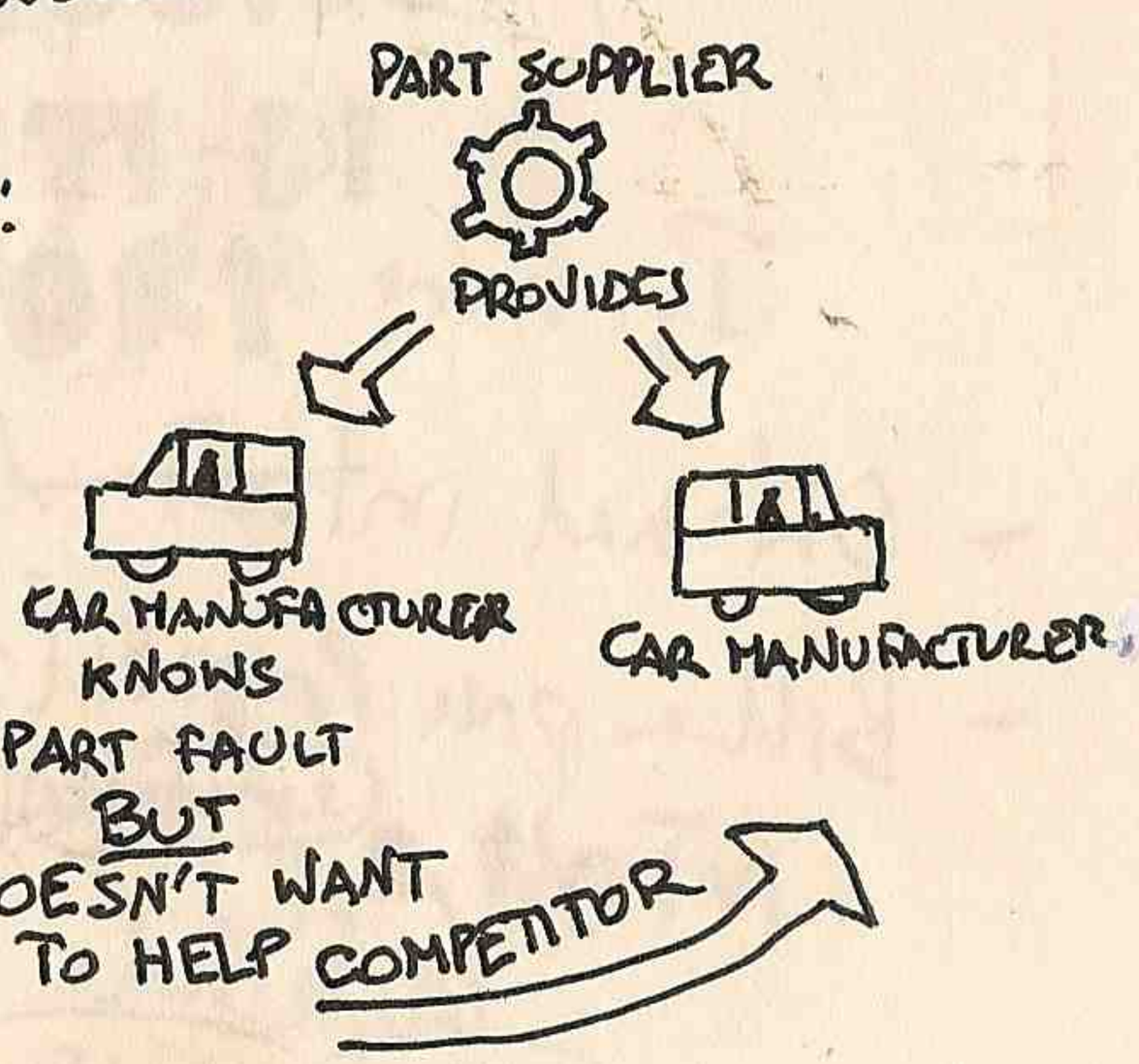
DATA PIRACY IS DIFFICULT TO DETECT

LIMITED TO CONTRACT PARTIES



PROPERTIZATION ⇒ NOT GRANULAR

⇒ EXAMPLE:

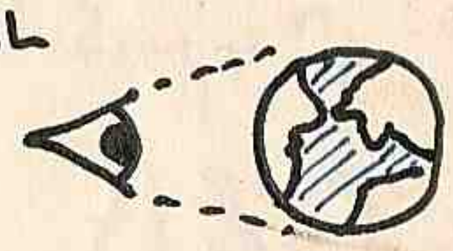


EXCLUSION RIGHT || HAVE PROBLEMS
SUI GENERIS

+ TECHNOLOGY ?

⇒ LABEL (CC)
ENABLE LAW TO WORK

⇒ ENFORCEMENT (DRM)
CREATE A GLOBAL SURVEILLANCE INFRASTRUCTURE



WHAT CAN WE DO?

1. SHARING OBLIGATIONS (FRAND) ⇒ SEE ITA-GOOGLE CASE
2. "OPEN DATA" MANDATES ⇒ DATA PRODUCED AFTER GOVERNMENT FUNDING MUST BE PUBLIC