UNIVERSITÀ DEGLI STUDI DI TORINO

CULTURE POLITICA SOCIETÀ

METIS RICERCHE

COREP

La Statistica, il ritorno.

Dalle schede perforate ai big data

roberto leombruni
università di torino e laboratorio revelli

Analisi dei dati: cosa serve alle imprese, cosa serve alla pubblica amministrazione

Machine Learning, Intelligenza Artificiale, Mining, Statistica, Modelli di simulazione: cosa c'è di vecchio nel nuovo, cosa c'è di nuovo nel vecchio

Torino
Campus Luigi Einaudi
10 gennaio 2019

indice della presentazione

- ✓ i dati, dai tempi delle schede perforate ai big data…
- ✓ …e quando non ci sono dati!
- ✓ dati sì, ma di qualità: vecchi e nuovi *epic fails*
- ✓ cosa c'è di vecchio nel nuovo: dati amministrativi e "found data"
- ✓ understanding *vs* forecasting

DATO, in italiano

i dati servono per
prendere decisioni

- …
- (2) Ciascuno degli elementi di cui si dispone per formulare un giudizio o per risolvere un problema.
  - "raccogliere tutti i d."
  - *Dato di fatto*, elemento certo.

DATO, in italiano, in statistica

- …
- (2) Ciascuno degli elementi di c
  giudizio o per risolvere un prob
  - "raccogliere tutti i d."
  - *Dato di fatto*, elemento certo.
  - *Dato statistico*, la **misura** di un fenomeno collettivo risultante dalla **rilevazione** di fatti singoli della stessa specie.
  - *Dati sensibili*, vedi sensibile.

i dati sono il risultato di un **processo**

DATO, in italiano, in statistica, in informatica

- …
- (2) Ciascuno degli elementi di cui si dispone per formulare un giudizio o per risolvere un problema.
  - "raccogliere tutti i d."
  - *Dato di fatto*, elemento certo.
  - *Dato statistico*, la **misura** di un fenomeno collettivo risultante dalla **rilevazione** di fatti singoli della stessa specie.
  - *Dati sensibili*, vedi sensibile.
- (3) **Qualunque** informazione rappresentata in modo da poter essere trattata da un calcolatore

1492

stolisco, e tèrra per *tersa *terra*, ossia *la parte disseccata* (cfr. *Terra* e *Torrido*). Confr. lo *zendo* tasta *tazza* e *guscio*, che altri però riferisce alla *rad. sser.* TAK-Ś-*sgrossare, comporre, fare* (v. *Tecnico*).

Vaso di terra cotta dove si coltivano le piante; Stoviglia di terra cotta, rotonda e alquanto cupa, colla quale si copre la pentola; Sorta di stoviglia di terra cotta, piatta, per uso di cuocervi sopra alcuna cosa.

Cfr. *Teschio; Testa; Testaccio; Testaceo; Testicolo; Testuggine; Stoviglia.*

2. *fr.* texte: = *lat.* TÈXTUM da TÈXERE *tessere:* propr. *tessuto*, quindi *intreccio*, e traslat. *discorso continuato* (v. *Tessere*).

Ciò che è contenuto parola per parola in uno scritto; Scritto di un autore, considerato in rapporto ai commenti e alle note, che vi sono state fatte sopra.

« Testo d'Aldo » Nome di un carattere presso gli stampatori.

Deriv. *Testino* nome di un piccolo carattere nelle Stamperie; *Testuale.* Cfr. *Contesto; Pretesto.*

32%

Esempi di dati: cos'hanno in comune?

i dati sono il risultato
di un **processo**…

Esempi di dati: cos'hanno in comune?

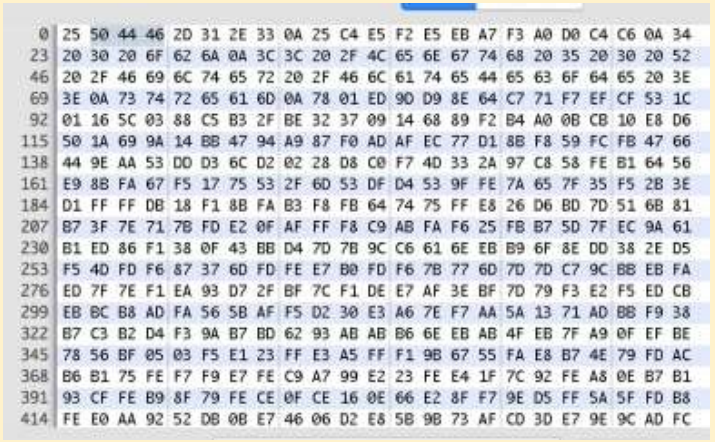
i dati sono il risultato di un **processo**…

…che produce **numeri**

Esempi di dati: cos'hanno in comune?

Quindi un dato è un numero, ma…

Non tutti i numeri sono dati: quale di questi due è un dato?!

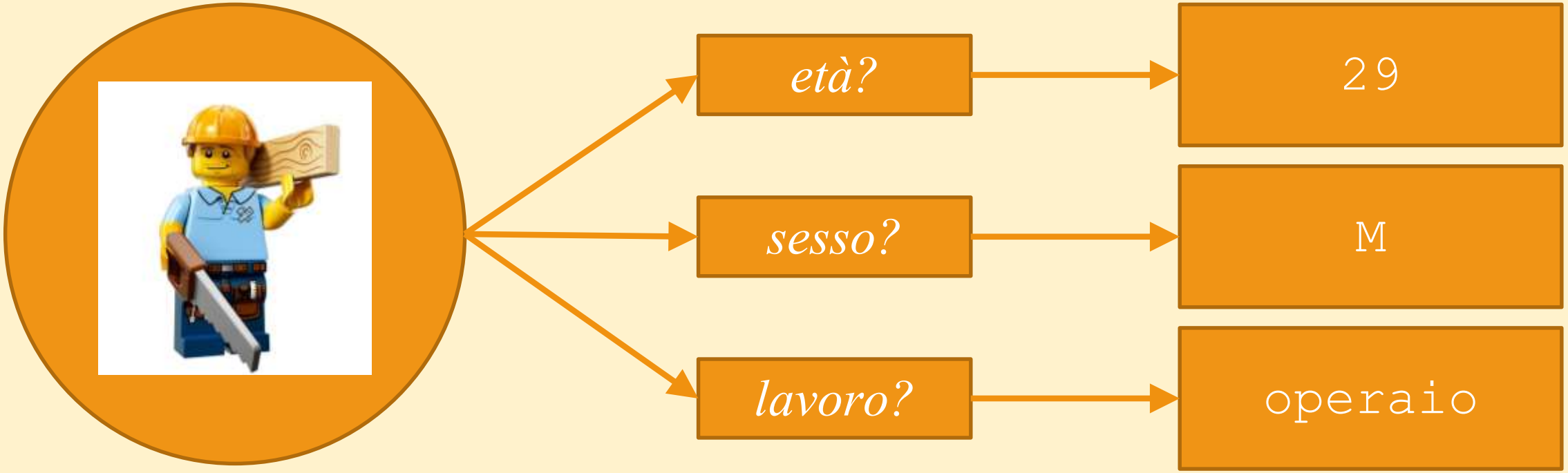


Offset	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00000000	74	65	6D	70	6C	61	74	65	20	22	42	6F	6F	74	20	53
00000010	65	63	74	6F	72	20	4E	54	46	53	22	0D	0A	0D	0A	2F
00000020	2F	20	54	65	6D	70	6C	61	74	65	20	62	79	20	50	61
00000030	75	6C	20	4D	75	6C	6C	65	6E	2C	20	70	63	67	75	72
00000040	75	40	74	68	65	2D	61	6E	73	77	65	72	2E	63	6F	6D
00000050	0D	0A	2F	2F	20	6C	61	73	74	20	6D	6F	64	69	66	69
00000060	65	64	20	4A	75	6C	20	31	36	2C	20	32	30	30	30	0D
00000070	0A	0D	0A	2F	2F	20	54	6F	20	62	65	20	61	70	70	6C
00000080	69	65	64	20	74	6F	20	73	65	63	74	6F	72	20	30	20
00000090	6F	66	20	61	6E	20	4E	54	46	53	2D	66	6F	72	6D	61
000000A0	74	74	65	64	0D	0A	2F	2F	20	6C	6F	67	69	63	61	6C
000000B0	20	64	72	69	76	65	20	6F	72	20	74	6F	20	74	68	65
000000C0	20	6D	69	72	72	6F	72	20	63	6F	70	79	20	6F	66	20
000000D0	74	68	65	20	62	6F	6F	74	0D	0A	2F	2F	20	73	65	63
000000E0	74	6F	72	2C	20	77	68	69	63	68	20	77	69	6C	6C	20
000000F0	62	65	20	6C	6F	63	61	74	65	64	20	6E	65	61	72	20

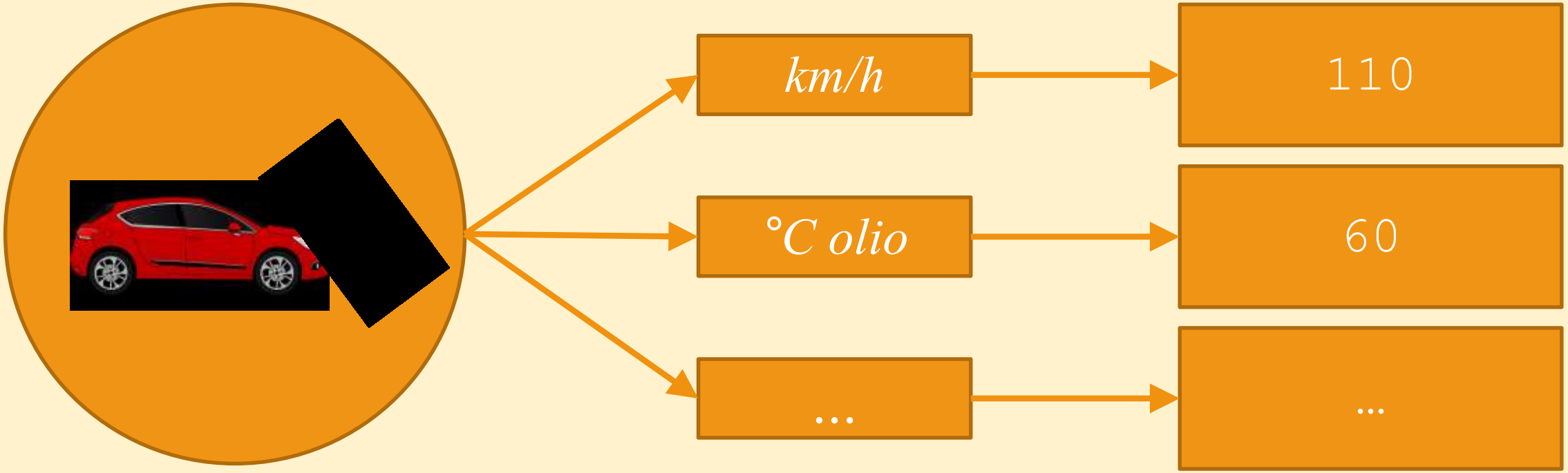
Un dato, se non conosciamo il processo che l'ha generato, non è più un dato.
Che dato è «1492»?

Anno scoperta dell'America	1492
Prezzo bicicletta in fibra di carbonio	1492
Articolo codice civile sugli effetti della garanzia di un prodotto	1492
…	

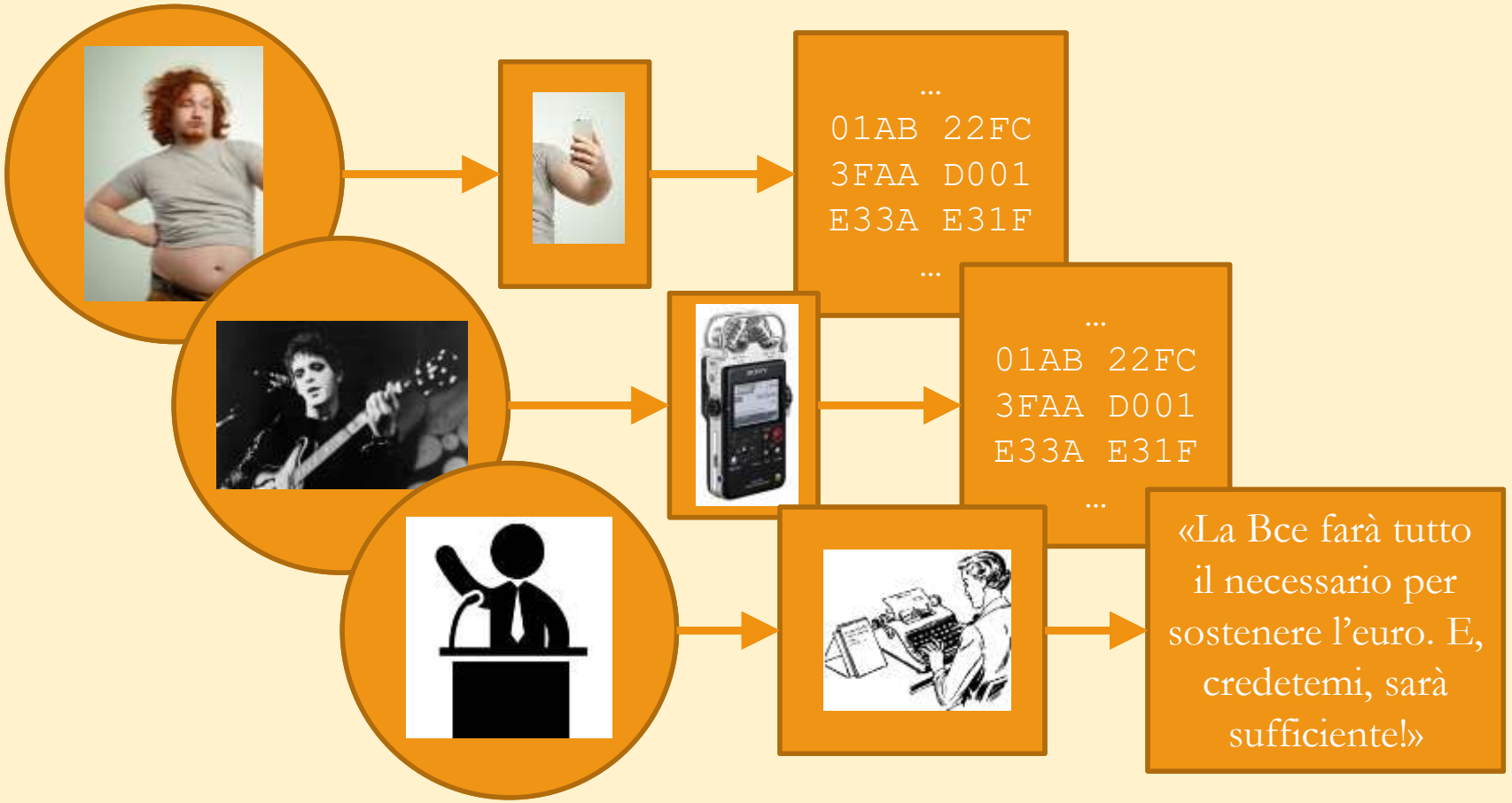
Esempi di PROCESSO: una rilevazione statistica…



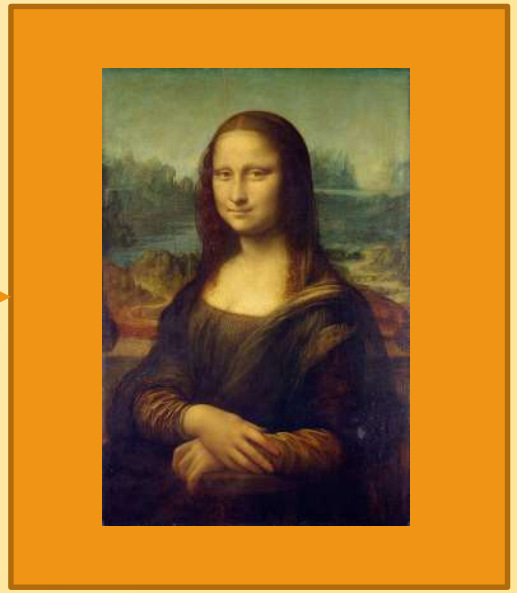
…un *device* della Internet of things…



…o qualunque altro mezzo di codifica!



Non tutti i processi generano numeri



Non tutte le evidenze empiriche hanno la forma di un dato!

- Nella ricerca empirica, il ricercatore “classico” sa che a volte gli serva una rilevazione statistica, altre volte una analisi qualitativa.
- La sovrabbondanza di dati può portare al pregiudizio che qualunque decisione **può** essere supportata soprattutto da analisi quantitative.
- O addirittura che **deve** essere basata sulla elaborazione automatica dei *byte di dati immagazzinati.

data driven *vs* evidence driven

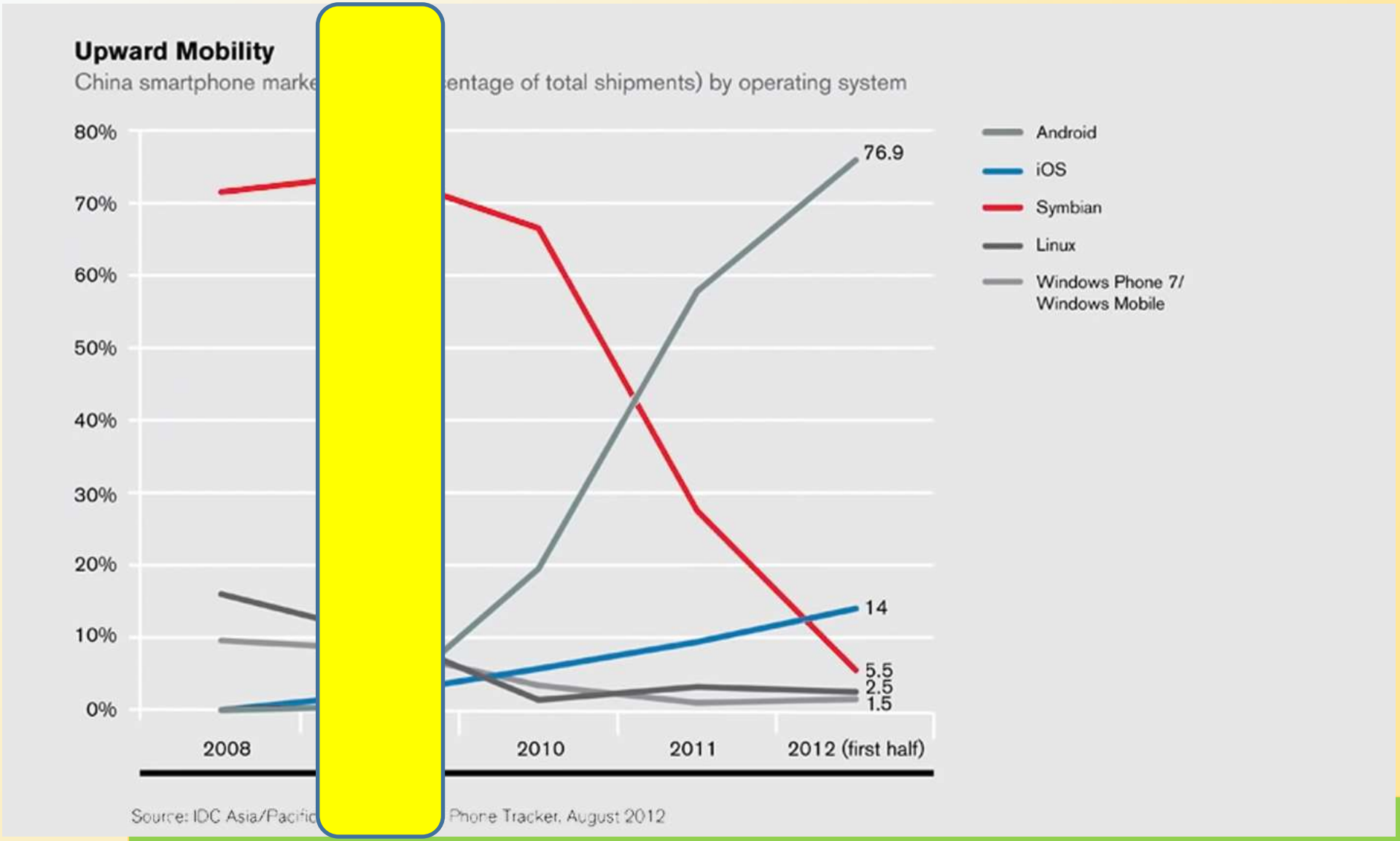
“The Cost of Missing Something”

- TED Conference di Tricia Wang, consulente di NOKIA nel 2009.

- In ambito big data si usano spesso metafore tipo l’ago nel pagliaio, o il granello di sabbia nella spiaggia.
- E gli analytics in NOKIA dicevano che gli smartphone non avevano mercato: troppo pesanti e costosi.

https://www.youtube.com/watch?v=WaOUJa9fjXU

l’analisi qualitativa in action

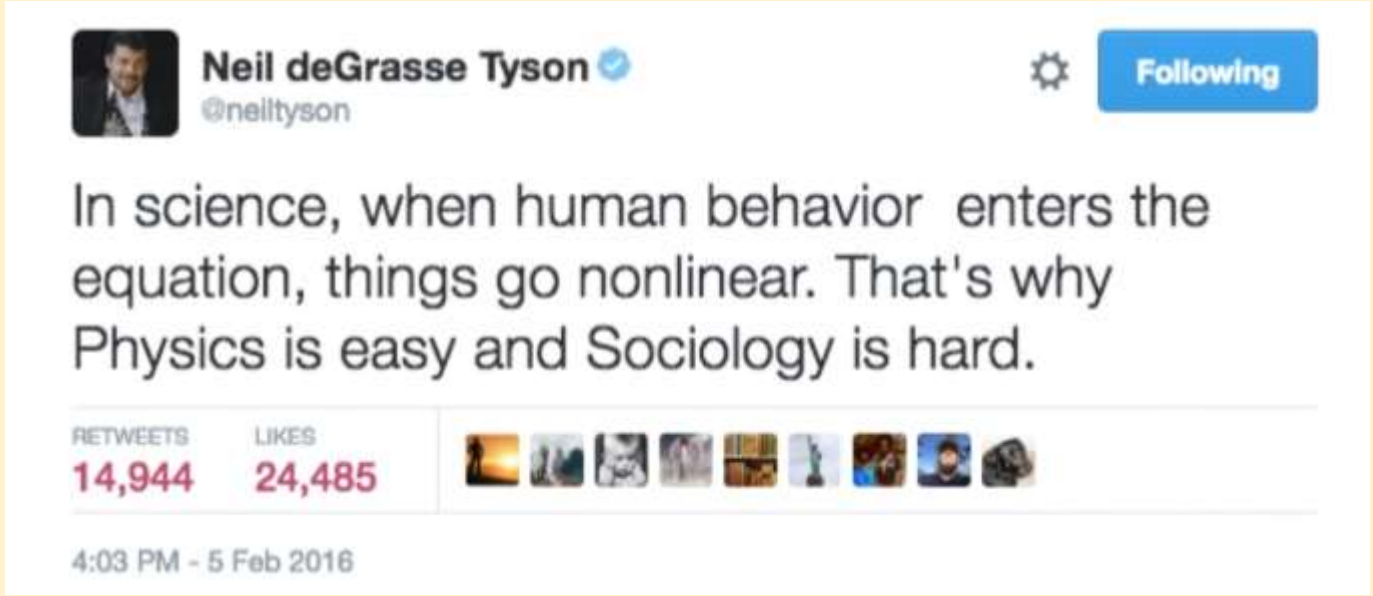


NOKIA prima e dopo la *strategic decision*

Il pagliaio allargando lo sguardo…

altre due conclusioni

1. Dati e analisi qualitativi sono fondamentali. Quelli che, con termine “nuovo”, Patricia Wang chiama “THICK DATA”.
2. Sono tanto più importanti quanto più si ha a che fare con le persone e le loro scelte.



Tornando ai dati quantitativi

- La caratteristica fondamentale che determina la qualità di un dato statistico risiede tutta nel **processo** che ha portato a quel dato.
- Processo che in ambito statistico si concretizza in una **raccolta sistematica** di misure del fenomeno.

Cosa si intende per raccolta sistematica

Adrian Smith, *Mad Cows and Ecstasy: Chance and Choice in an Evidence-Based Society*, ci propone una lista dei passaggi tipici che sono necessari per produrre una SOLIDA evidenza statistica.

- ❑ the framing of questions;
- ❑ design of experiments or surveys;
- ❑ drawing up protocols for data collection
- ❑ collection of data
- ❑ monitoring compliance with protocols
- ❑ monitoring data quality
- ❑ data storage, summarization, presentation
- ❑ stochastic modelling
- ❑ statistical analysis
- ❑ model criticism and assumptions assessment
- ❑ inference reporting and the use of results for prediction, decision-making or hypothesis generation

raccolta dati

Un esempio di successo in ambito BD-A

- Il progetto ARTEMIS della University of Ontario, per il monitoraggio e la prevenzione delle infezioni in ospedale nelle unità di terapia intensiva per nati prematuri.
- Con un “Big Data approach”, è stata implementata una tecnologia per rilevare e registrare in real time più di 1,000 misure al secondo relative a electrocardiogrammi, battiti cardiaci, ritmi respiratori, saturazione dell’ossigeno nel sangue…
- Sono stati in grado di identificare “early warnings” di infezione 24 ore prima rispetto all’approccio tradizionale.

http://hir.uoit.ca/cms/?q=node/24

Un esempio di raccolta non-sistematica: Google flu trends

nature

Vol 457|19 February 2009|doi:10.103

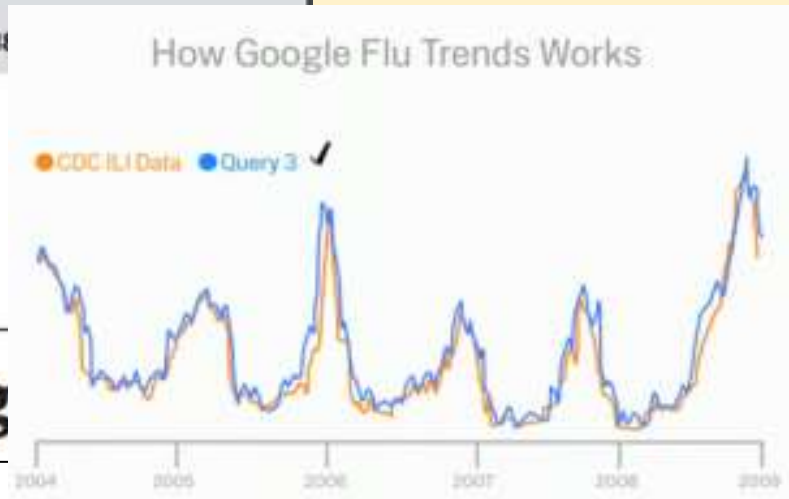
LETTERS

Detecting influenza epidemics using search eng query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. One way to improve early detection is to monitor

By aggregating historical logs o
between 2003 and 2008, we compu
50 million of the most common
Separate aggregate weekly count
state. No information about the i
time series was normalized by di
particular week by the total numb
in that location during the w
(Supplementary Fig. 1).



By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

Un esempio di raccolta non-sistematica: Google flu trends

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 494 > Issue 7436 > News > Article

NATURE | NEWS

عربي

When Google got flu wrong

US outbreak foxes a leading web-based method for tracking seasonal flu.

Declan Butler

13 February 2013

"When influenza hit early and hard in the United States this year, it quietly claimed an unacknowledged victim: one of the cutting-edge techniques being used to monitor the outbreak."

La differenza tra Artemis e GFT sta nella raccolta dati

- Nel caso di Artemis, è stata seguita tutta la checklist dello "statistician cook-book".

  → *L'aspetto nuovo rispetto a una indagine tradizionale sta nella mole dei dati da processare in real time, che richiede una **infrastruttura IT** dedicata e modelli in grado di **automatizzare le analisi**.*

- Nel caso di GFT i dati non sono stati prodotti per gli scopi dell'analista. Sono "**found data**", che a posteriori hanno rivelato una correlazione con il fenomeno di interesse.

  → *Non era il modello statistico ad essere sbagliato: più semplicemente, la variabile proxy utilizzata per predire la diffusione dell'epidemia era una proxy di **cattiva qualità***

Quality is KEY

- Non è perché i dati sono **BIG**, o il computer che usiamo è **POWERFUL**, o perché il posto dove registriamo i dati è addirittura una *NUVOLA*, che le evidenze che produciamo possono essere utilizzate a supporto di decisioni strategiche.

- L'ingrediente fondamentale è la **qualità dei dati di partenza**.

Non è perché i dati sono BIG

Un altro esempio di “smart forecast”, che nonostante la mole **BIG** di dati si è convertita in un *epic fail*.

Indagine su 10 milioni di cittadini americani per prevedere il futuro presidente degli Stati Uniti.

La previsione, “Landon in a Landslide”:

Landon, 57.1%, Roosevelt, 42.9%

I risultati veri, ooops:

Roosevelt 60.8%, Landon 36.5%

Messy Matters

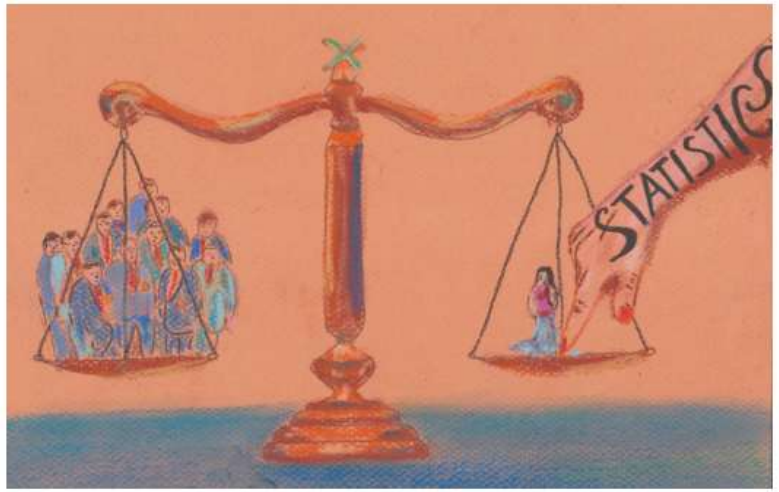
Bring your own data

« Google Glass(es) Yootles Postmortem »

Forecasting Elections with Dirty Data

Monday, September 30, 2013

By Sharad Goel



During the 1936 U.S. presidential campaign, the popular magazine *Literary Digest* conducted a mail-in election poll that attracted over two million responses, a huge sample even by today's standards. Unfortunately for them, size isn't the only thing that matters. *Literary Digest* notoriously and erroneously predicted a landslide victory for Republican candidate Alf Landon. In reality, the incumbent Franklin D. Roosevelt decisively won the

Made Data vs Found data (I)

«The “big data” that interests many companies is what we might call “found data”, the digital exhaust of web searches, credit card payments and mobiles pinging the nearest phone mast…»

Tim Harford, cit.

Le differenze (e i pericoli) rispetto a una indagine statistica tradizionale risiedono esattamente qua: I dati non provengono da un protocollo rigoroso di data procurement, ma sono stati “trovati” da qualche parte, li ha prodotti qualcun altro.

→ *c’è qualcosa di “già visto” anche in questo caso…*

Made Data vs Found data (II)

Anche nell'ambito della statistica ufficiale c'è un progressivo abbandono della raccolta dati con indagini campionarie (*made data*), a favore dell'utilizzo di **dati amministrativi**, cioè a dati raccolti da amministrazioni pubbliche nella loro normale attività gestionale (dati fiscali, contributivi, assicurativi…).

Il punto in comune a tutti i *found data*, che vale per i dati amministrativi così come per un tweet o una traccia GPS, è che sono generati/raccolti per scopi che non sono quelli dell'analista che li ha "trovati".

and the answer is...

Forecasting *vs* …

1) Forecasting vs understanding
2) Prediction vs explanation
   - black-box models *vs* y=f(x)
   - modelli strutturali *vs* forma ridotta
   - microsimulazioni statiche *vs* dinamiche
3) Correlation *vs* causation
   - modelli quasi sperimentali
   - social random experiments… *already on the go!*

   → Esperimento sociale randomizzato di Facebook nel 2012

   → Google per lo sviluppo di nuovi algoritmi di ricerca

Experimental evidence of massive-scale emotional contagion th

Adam D. I. Kramer[a,1], Jami

[a]Core Data Science Team, Faceboo
NY 14853

Edited by Susan T. Fiske, Princeton

Emotional states can be tra
contagion, leading people t
without their awareness. Emo
in laboratory experiments, wi
negative emotions to others.
network, collected over a 20-y
moods (e.g., depression, hap
networks [Fowler JH, Christa
though the results are controv
who use Facebook, we test w
outside of in-person interactio
the amount of emotional conte
expressions were reduced, pe
and more negative posts; wh
duced, the opposite pattern o
emotions expressed by other
emotions, constituting experi
contagion via social network
contrast to prevailing assumpt
verbal cues are not strictly nec
that the observation of othe
a positive experience for peop

computer-mediated communicatio

Editorial Expression of Concern and Correction

PSYCHOLOGICAL AND COGNITIVE SCIENCES

PNAS is publishing an Editorial Expression of Concern regarding the following article: "Experimental evidence of massive-scale emotional contagion through social networks," by Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, which appeared in issue 24, June 17, 2014, of *Proc Natl Acad Sci USA* (111:8788–8790; first published June 2, 2014; 10.1073/pnas.1320040111). This paper represents an important and emerging area of social science research that needs to be approached with sensitivity and with vigilance regarding personal privacy issues.

Questions have been raised about the principles of informed consent and opportunity to opt out in connection with the research in this paper. The authors noted in their paper, "[The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." When the authors prepared their paper for publication in PNAS, they stated that: "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program." This statement has since been confirmed by Cornell University.

PSYCHOLOGICAL AND COGNITIVE SCIENCES

Correction for "Experimental evidence of massive-scale emotional contagion through social networks," by Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, which appeared in issue 24, June 17, 2014, of *Proc Natl Acad Sci USA* (111:8788–8790; first published June 2, 2014; 10.1073/pnas.1320040111).

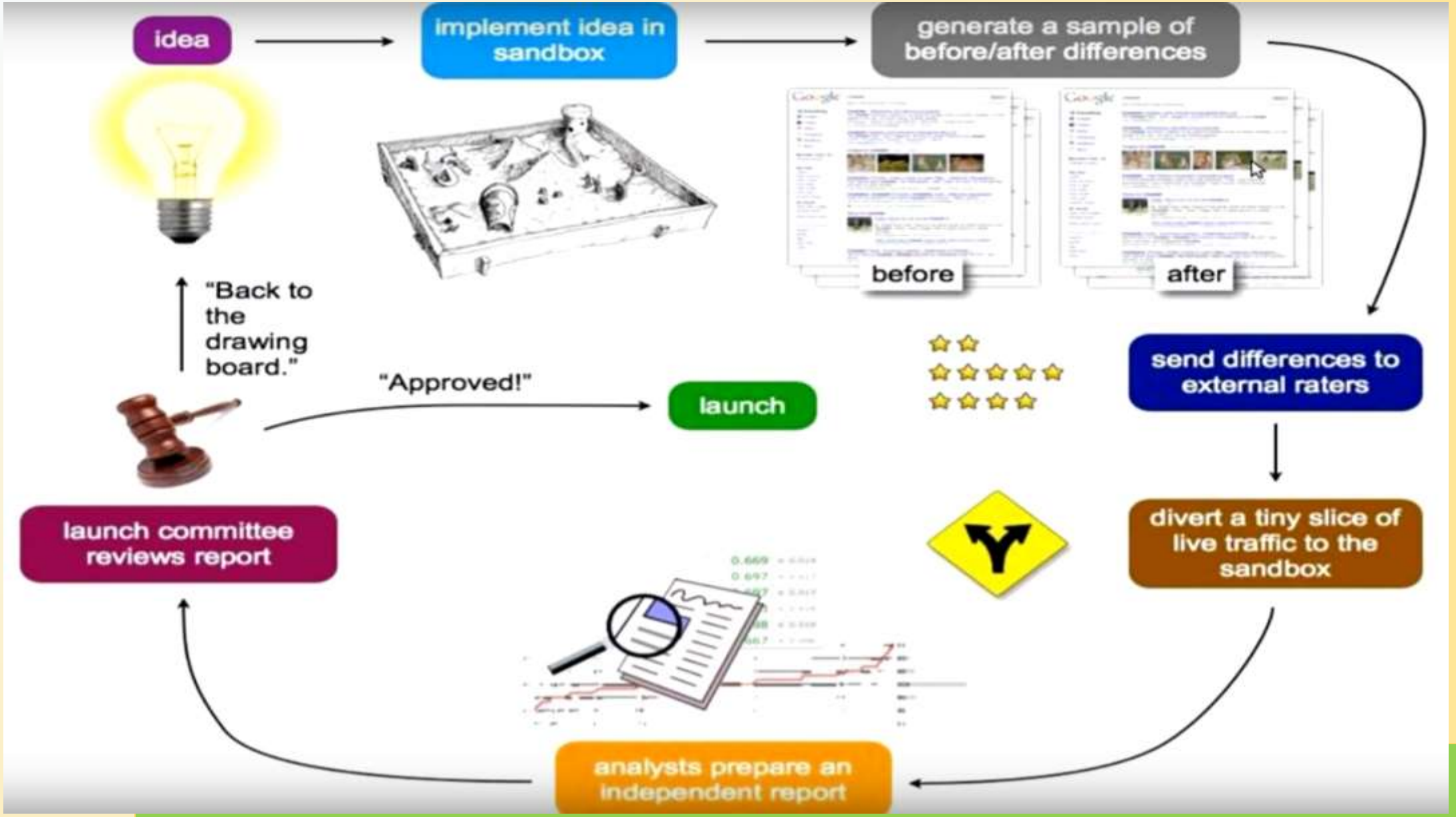
The authors note that, "At the time of the study, the middle author, Jamie E. Guillory, was a graduate student at Cornell University under the tutelage of senior author Jeffrey T. Hancock, also of Cornell University (Guillory is now a postdoctoral fellow at Center for Tobacco Control Research and Education, University of California, San Francisco, CA 94143)." The author and affiliation lines have been updated to reflect the above changes and a present address footnote has been added. The online version has been corrected.

The corrected author and affiliation lines appear below.

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell

https://www.pnas.org/content/pnas/111/24/8788.full.pdf



Matt Cutts, 2012 Webmaster Conference, Korea


THANK
YOU!