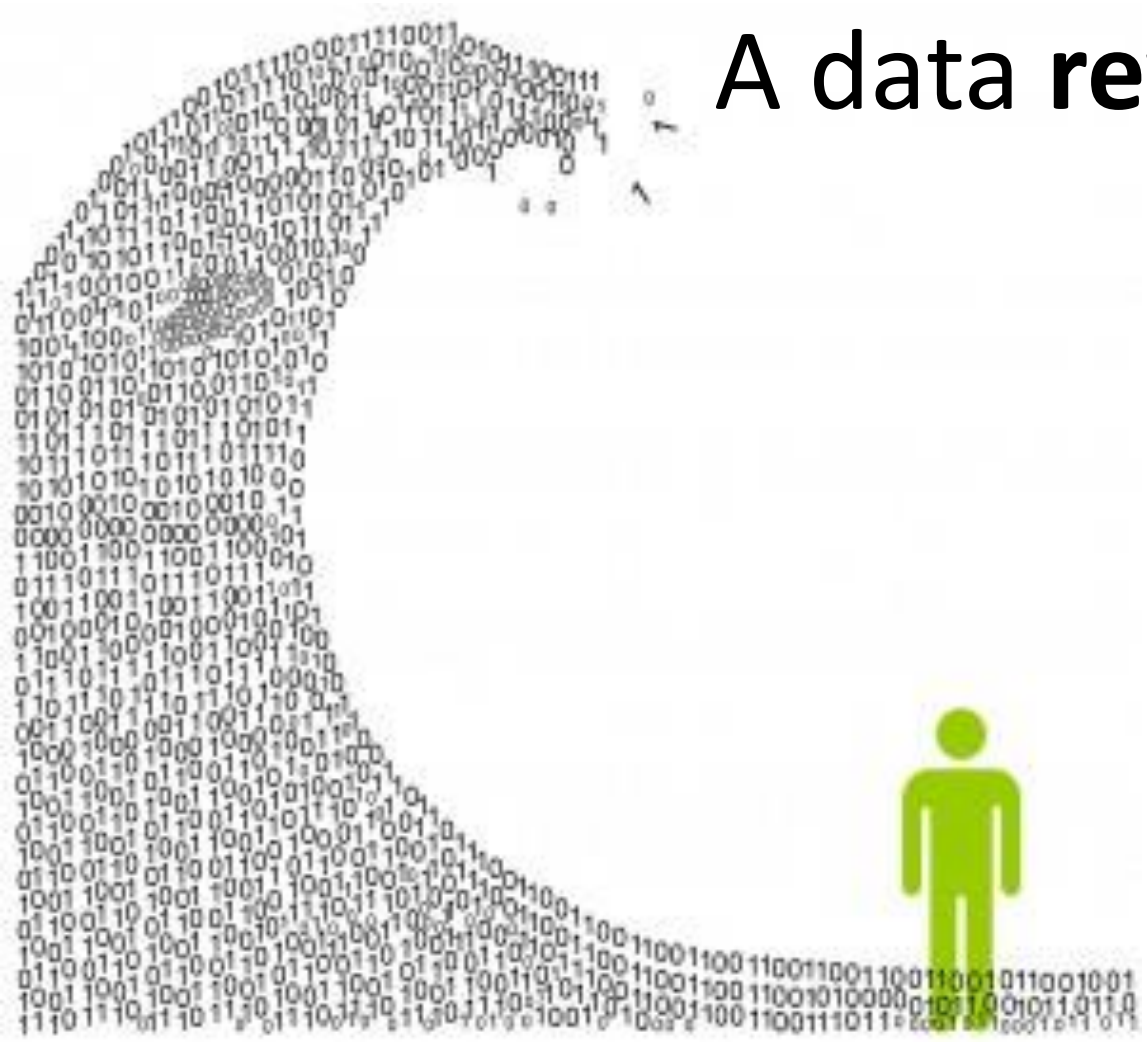
The more data….

Information Management evolution…

Once a time…

- When we had no data: “**design me a survey**”
- When data were not computerized: “**build me a database**”
- When electronic devices started spreading: “**configure me a mobile data collection system**”
- When data reached a good level of quality: “**build me a dashboard**”
- When the data started being used externally: “**build me a data visualization**”

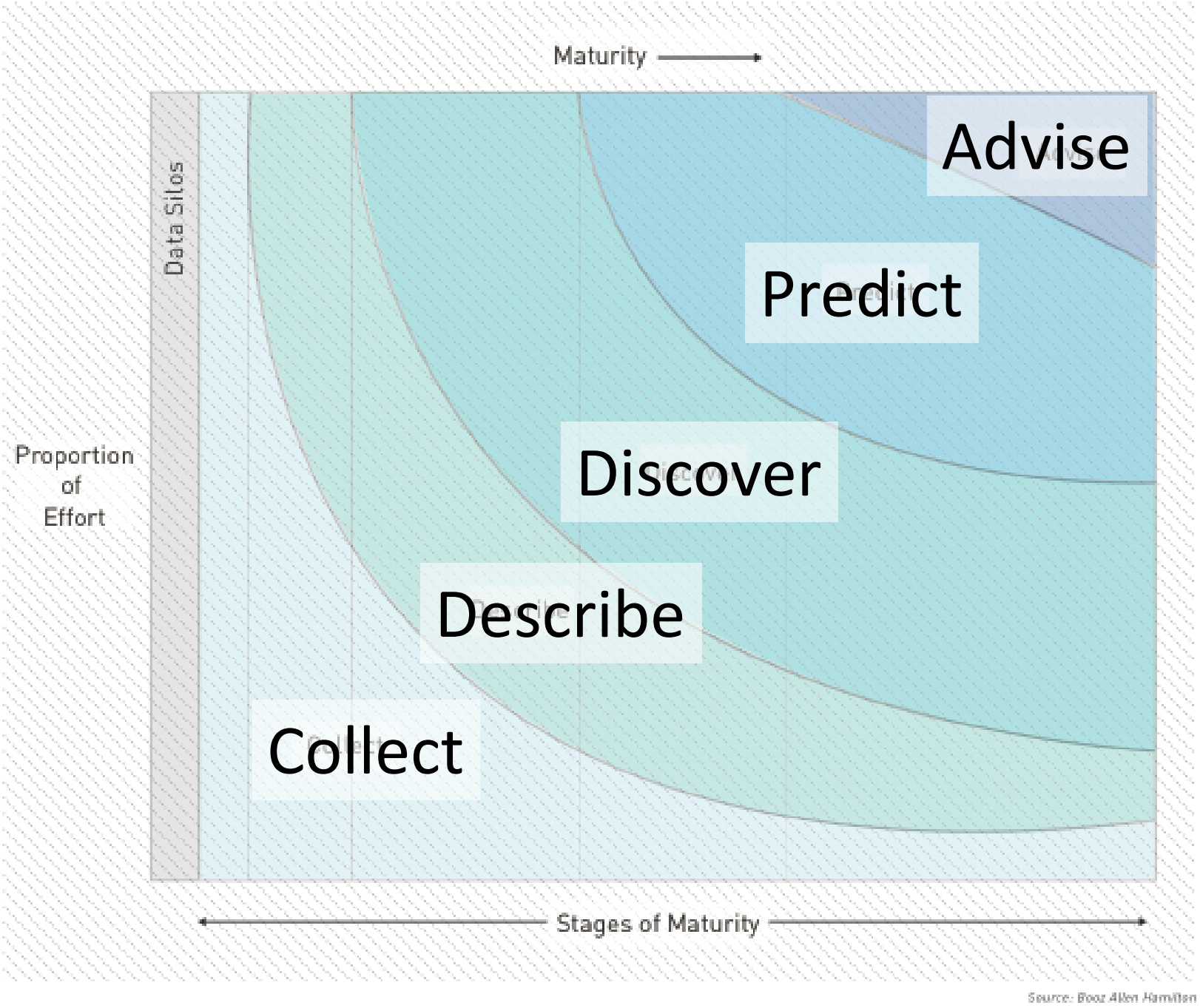
What's the next bottleneck?



A data **revolution** or a data **deluge…**

70,000 household interviews with 700 variables!!!

Maturity & Effort



Discover

Hidden patterns : Statistical clusters of individuals within a large population group

Better Beneficiaries Group definition!

Hierarchical clustering, KNN clustering, Multiple Correspondence analysis…

Predict

"Proxy Mean Testing": Predicting poverty level based on demographic profile

Can we predict more? Risk of **Early Marriage**, **Documentation**, **Out of School**, **Child labour**…

Linear Regression, Logistic Regression, Decision Tree, Random Forest, Naive Bayes ...

Advise

How to allocate the Budget ? What's the vulnerability profile of my population?

What is the most efficient combination of activities to reduce the vulnerability of this population?

Survival analysis, Portfolio Optimization…

The traditional approach…

Messing it up…

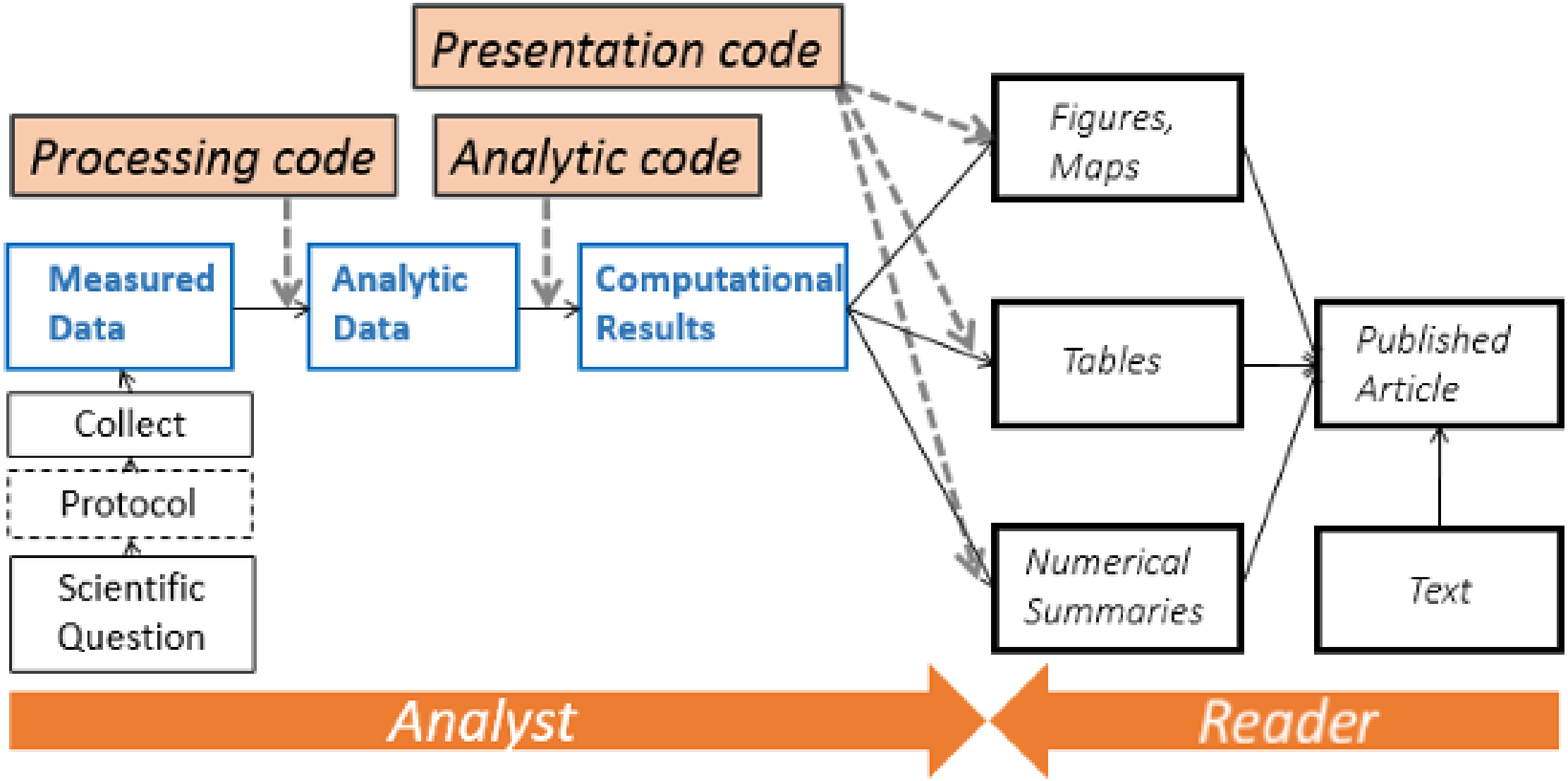
“Point-and-click" user interfaces!

Data constantly exported / imported…

Reproducible analysis

- What analysis is **behind the figure**?
- Were **outliers** identified?
- Which **dataset** version?
- Can I **rerun the analysis?**
- How to understand the **research process?**

From "click" to "script"



Tool…?


IBM SPSS
Science
sas
Business
Proprietary
STATA
Social Science
OpenSource
MATLAB
Engineer
python
Developper
R
Statistician

We need to speak the same language

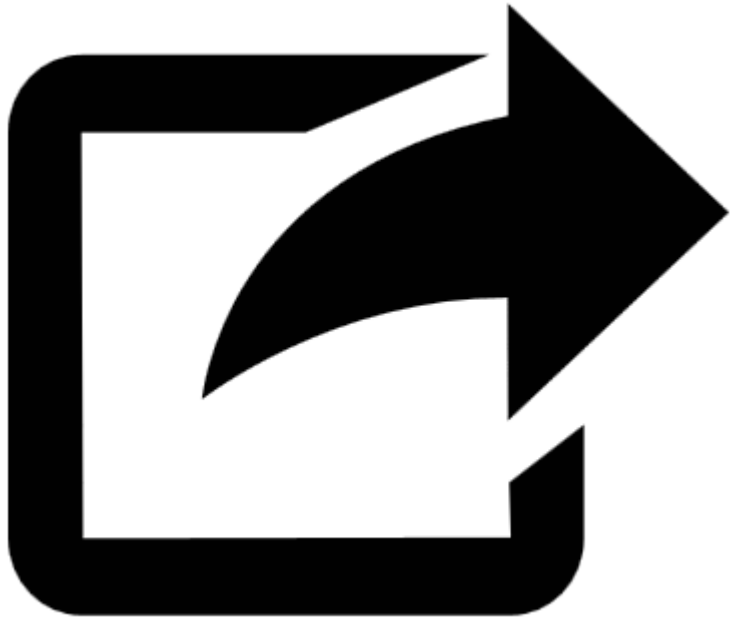
- No barrier to enter – Open Source
- Open to contribution – Open System
- With a strong community – and even Microsoft behind…
- Plenty of packages…
- That can do data manipulation, plotting, mapping, classification, clustering & prediction

Partnering

Contracting

Building capacity

And Share!



A first induction to R: http://edouard-legoupil.github.io/humanitaRian-data-science/slides

@edouard_lgp