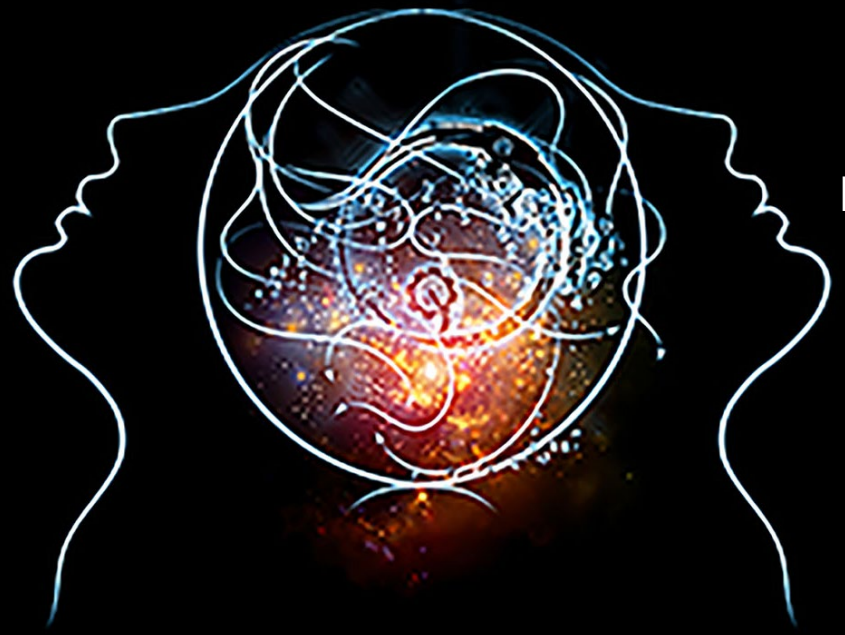
Mining GitHub to Identify Open-Source Software Health in Blockchain Projects



Jeff Nijsse

Postgraduate Research Symposium
November 26, 2021
Auckland, New Zealand

AUT



Notepad++

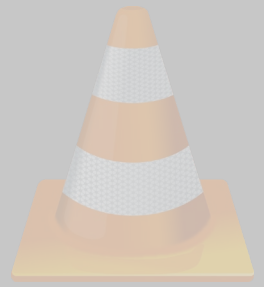
LATEX





POWERED BY
APACHE


R



Why do people contribute to open source projects?

- Trust
- Experimentation
- Community
- Productivity
- Permissionless

56 M+
total developers on GitHub

72 %
of Fortune 50 companies use GitHub Enterprise

60 M+
new repositories created in the last year

1.9 B+
contributions added in the last year

(GitHub, 2020)

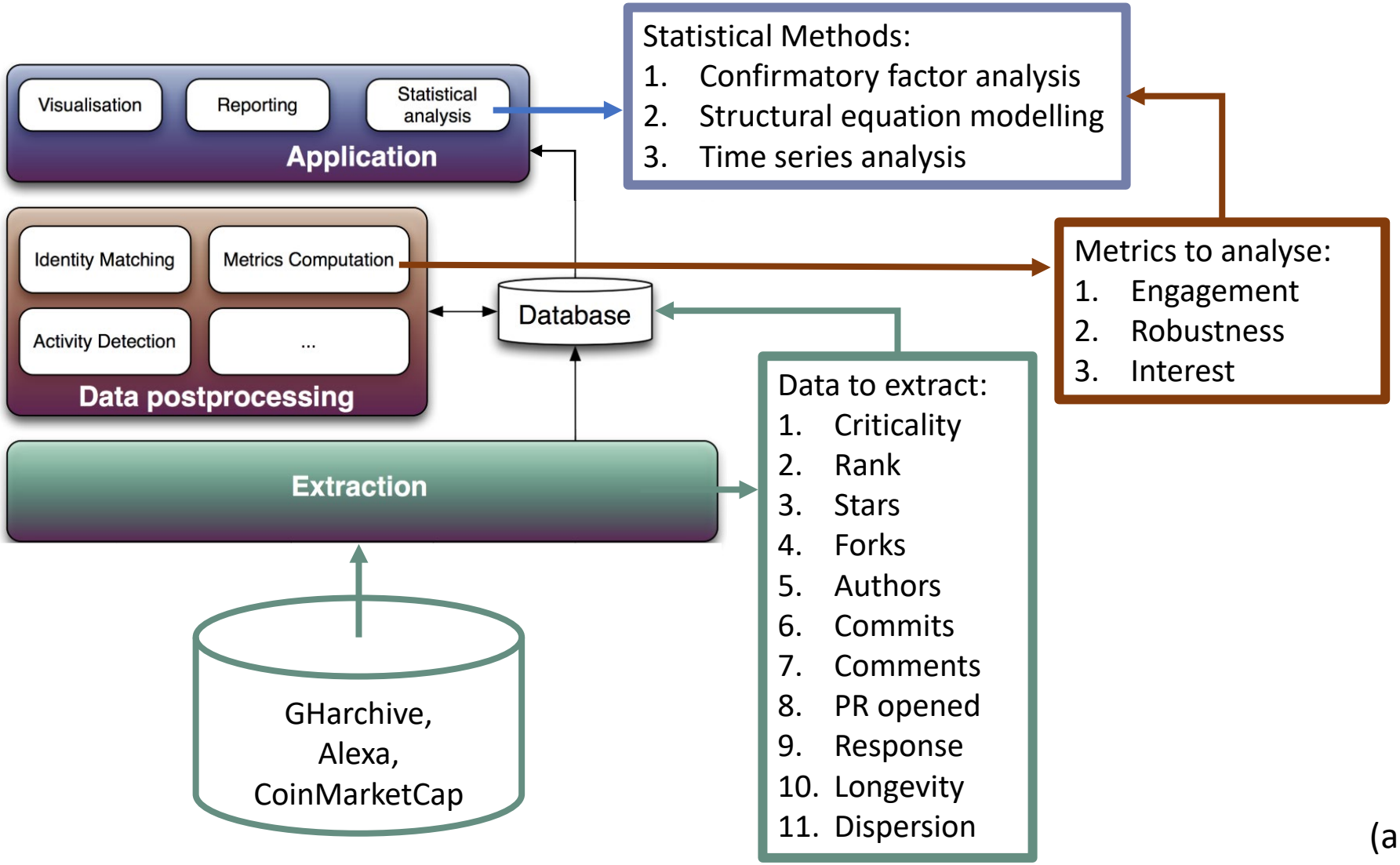
Open-source data can be mined to determine software health

How can factors that influence health of an open-source blockchain project be identified?

Modified framework for analysing open-source software
(Goeminne, 2013)

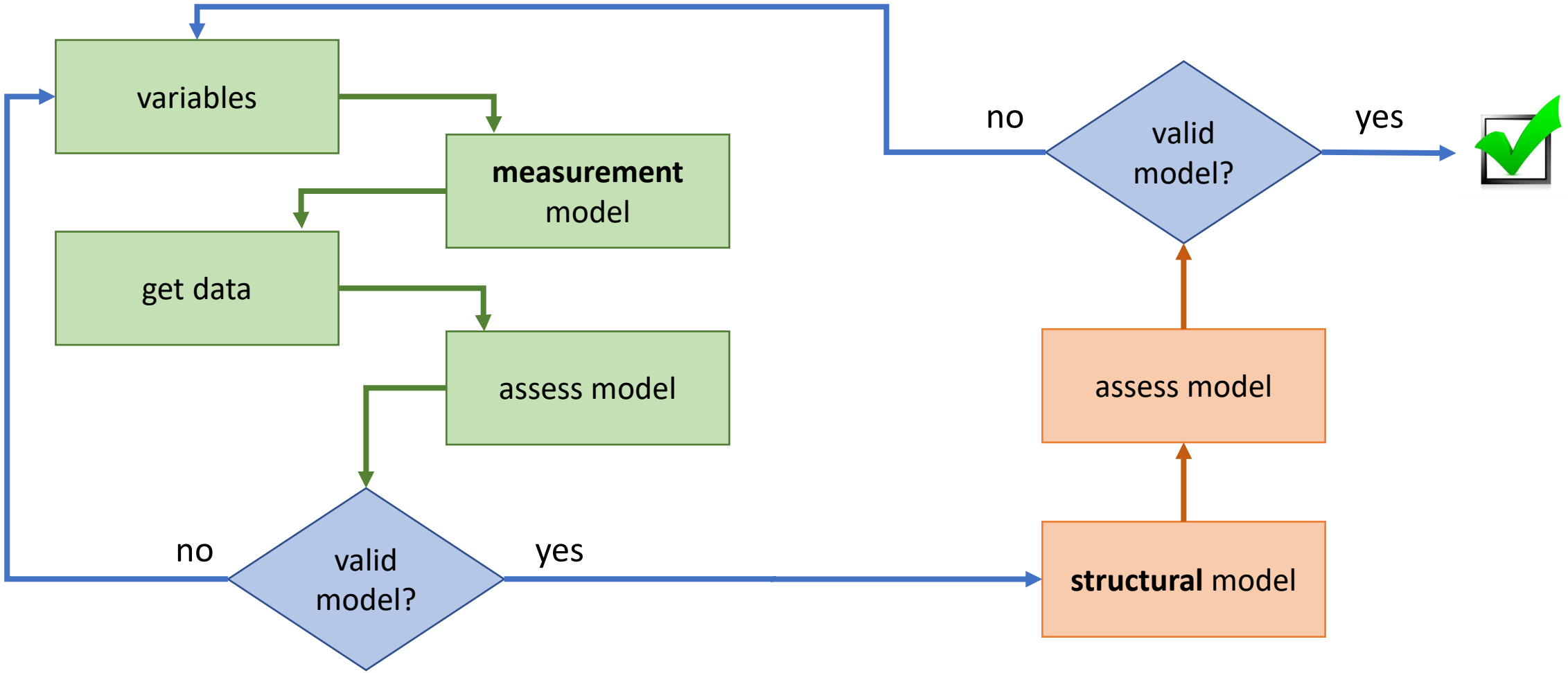
Blockchain health assessment tool (app)

Methodology



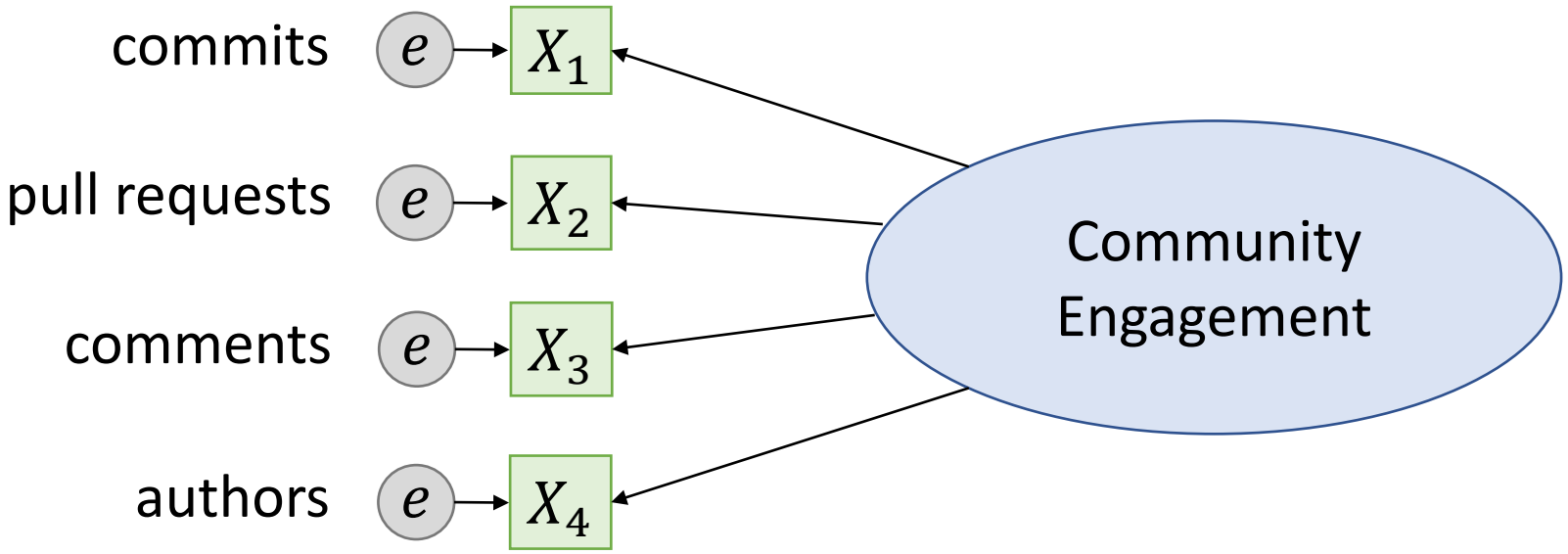
(adapted from Goeminne & Mens, 2013)

Structural Equation Modelling

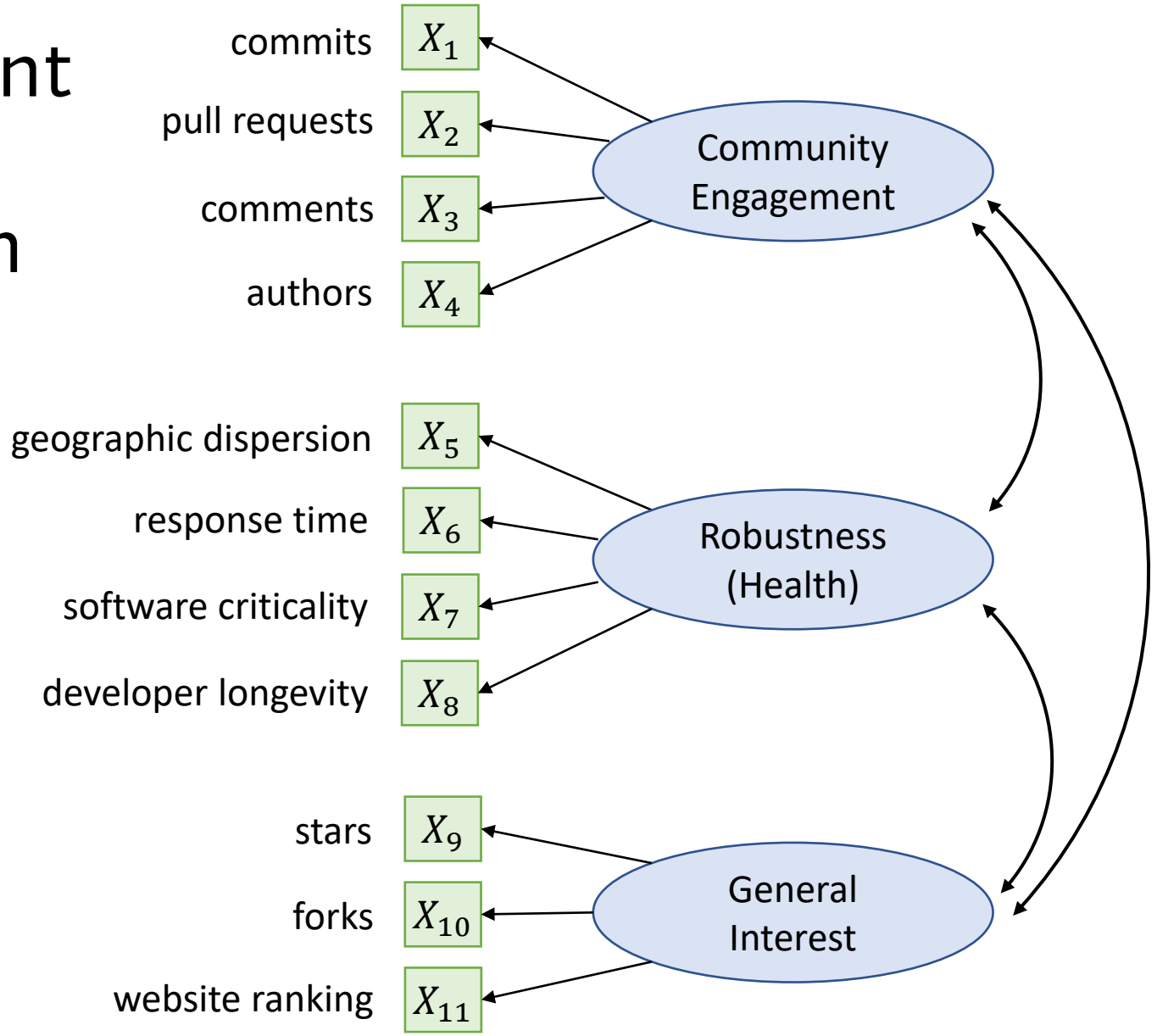


(Hair et al, 2014)

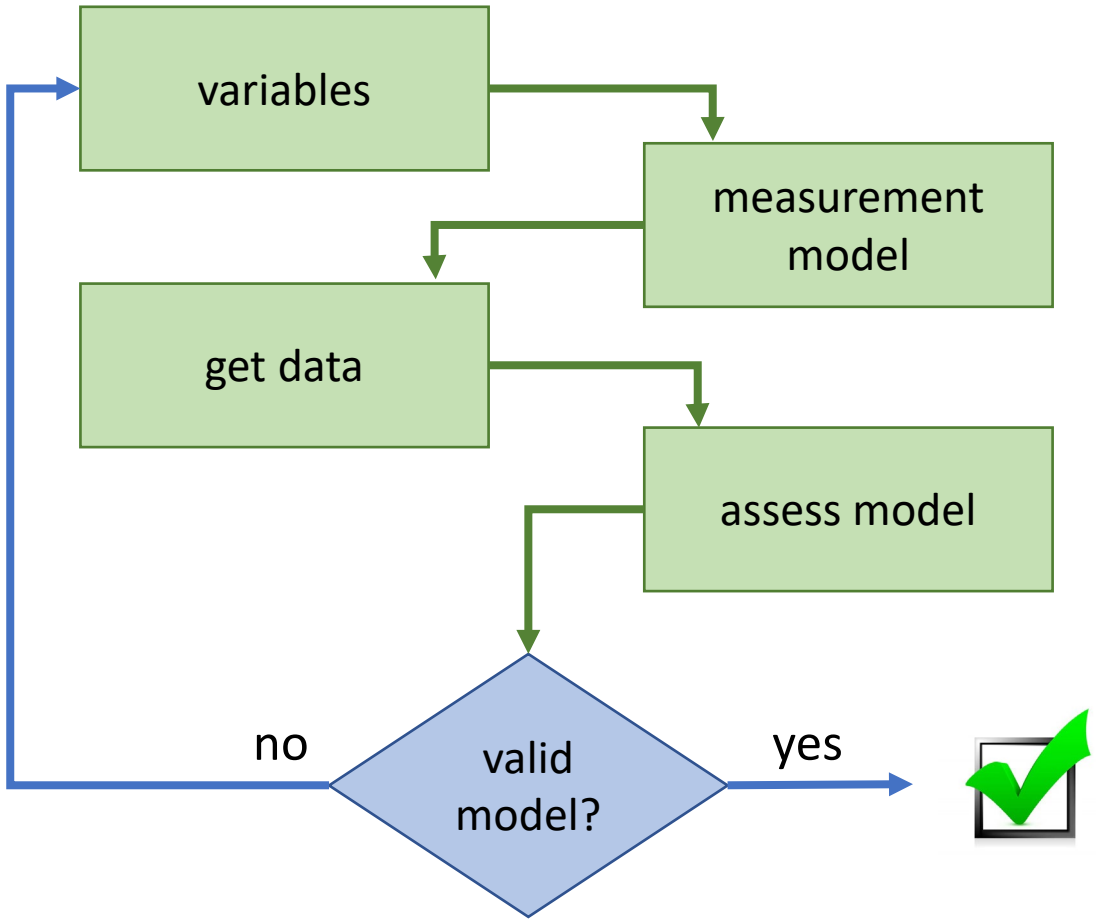
“Engagement” – a Latent factor

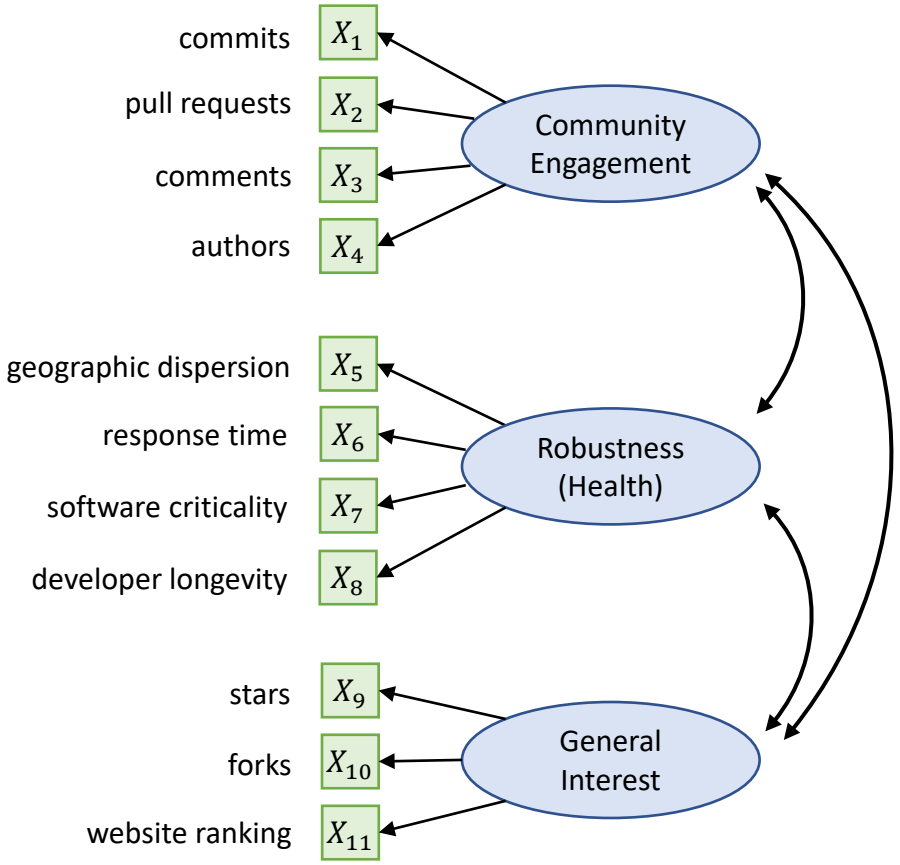


Measurement model specification

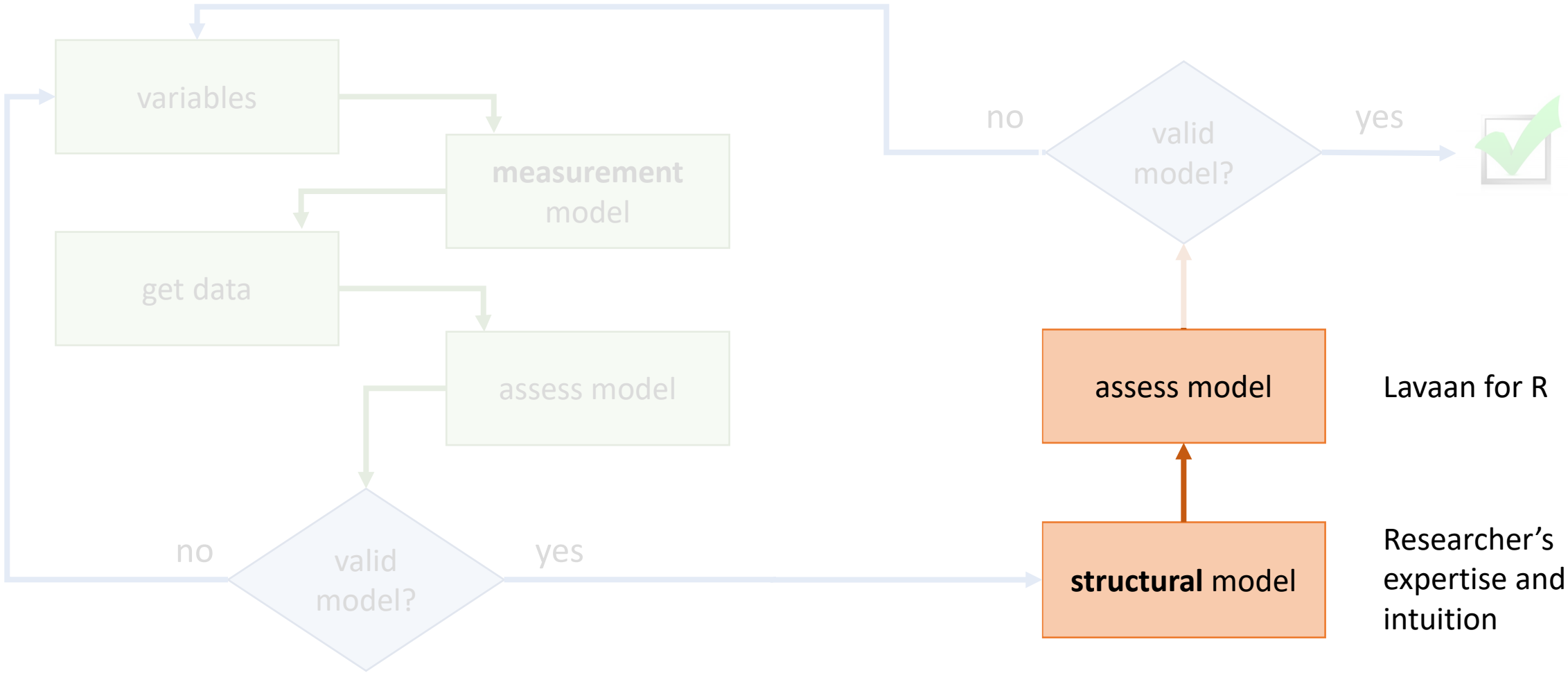


Confirmatory Factor Analysis

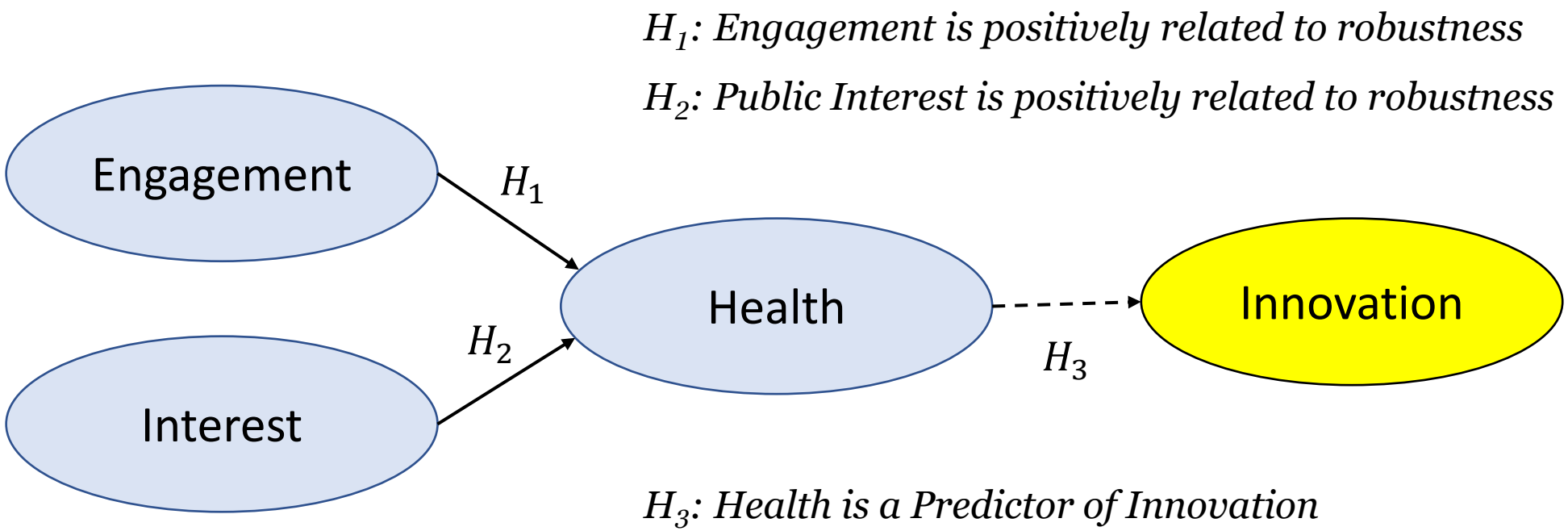




Structural Equation Modelling

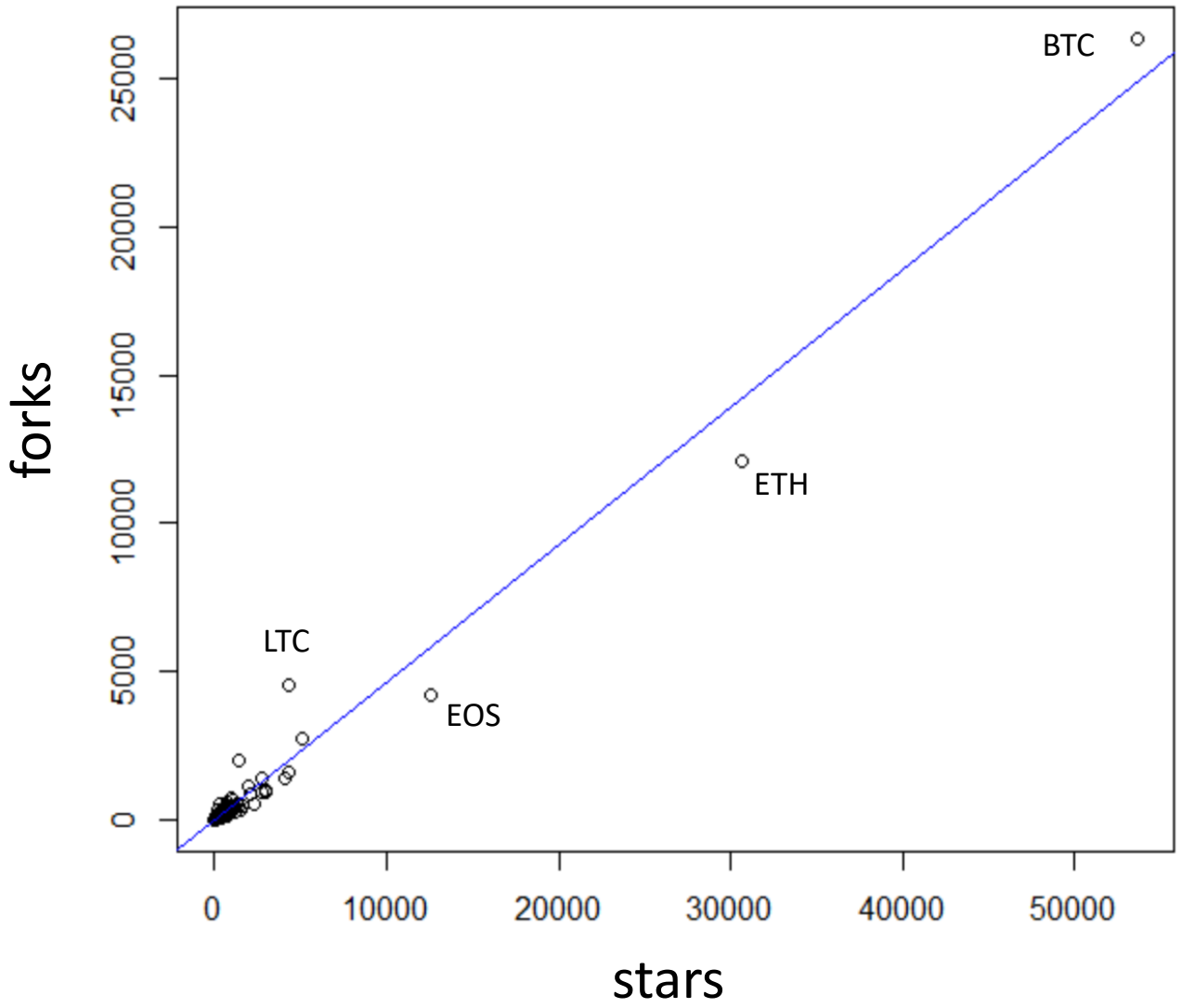


Proposed Structural Model

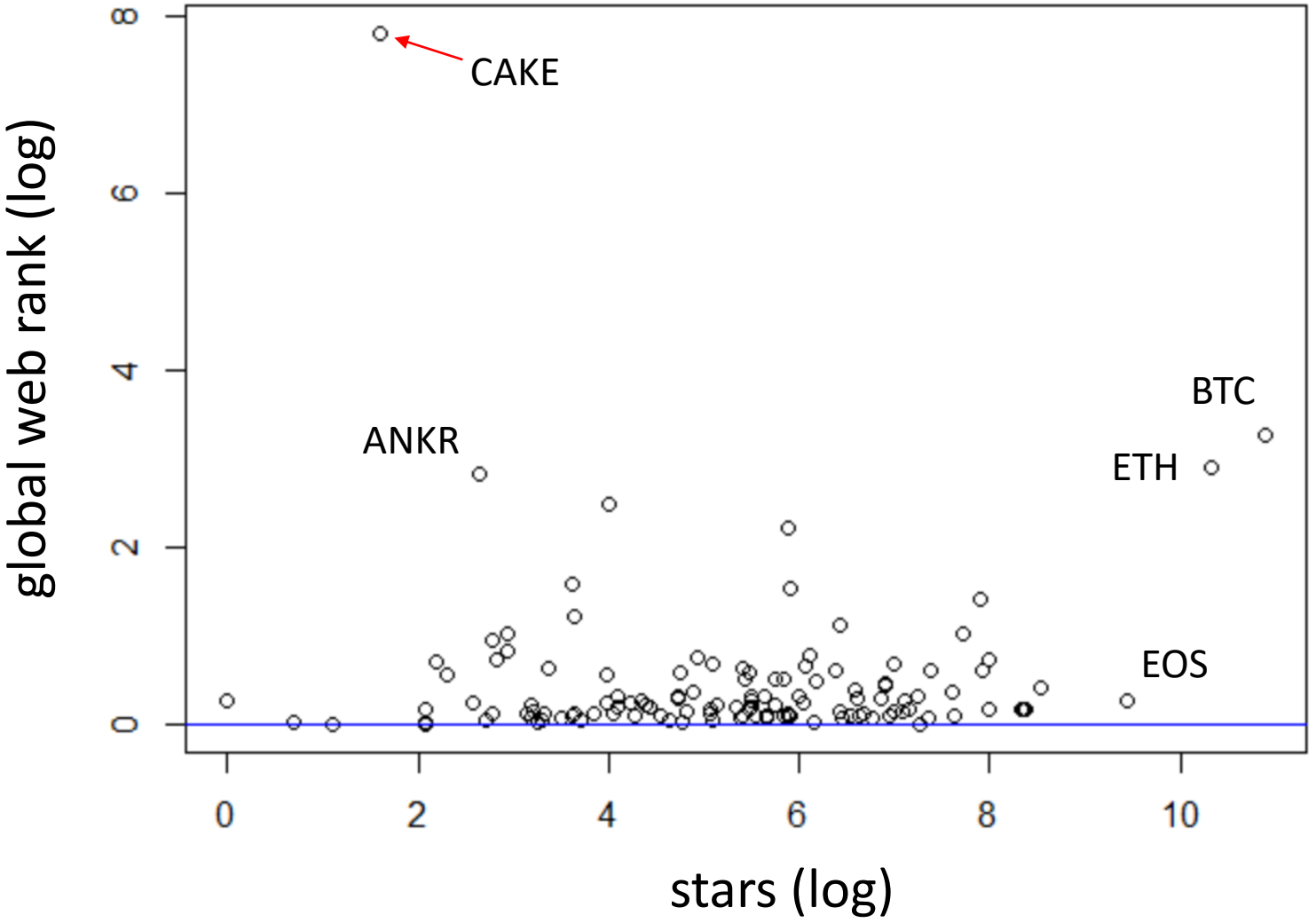


Sample of Preliminary Results

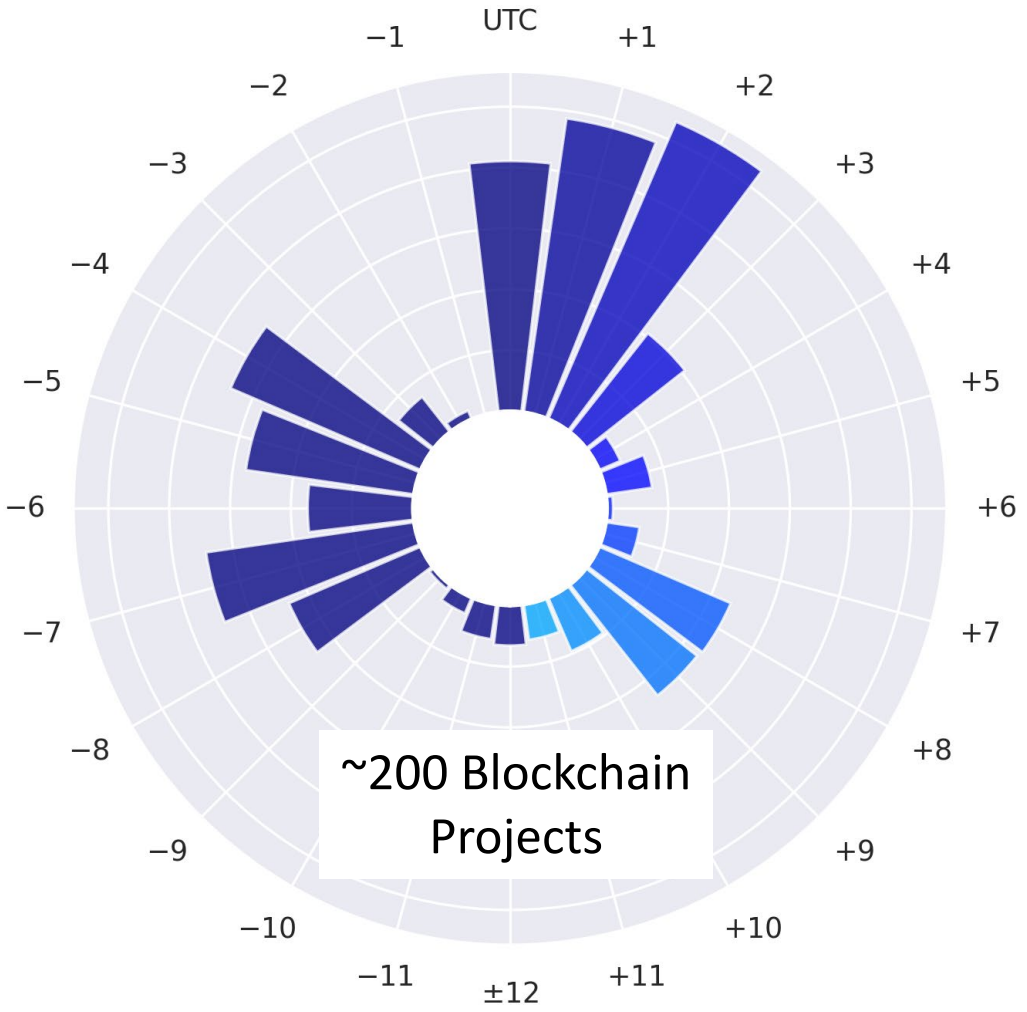
Stars and Forks are highly correlated (p<0.001)

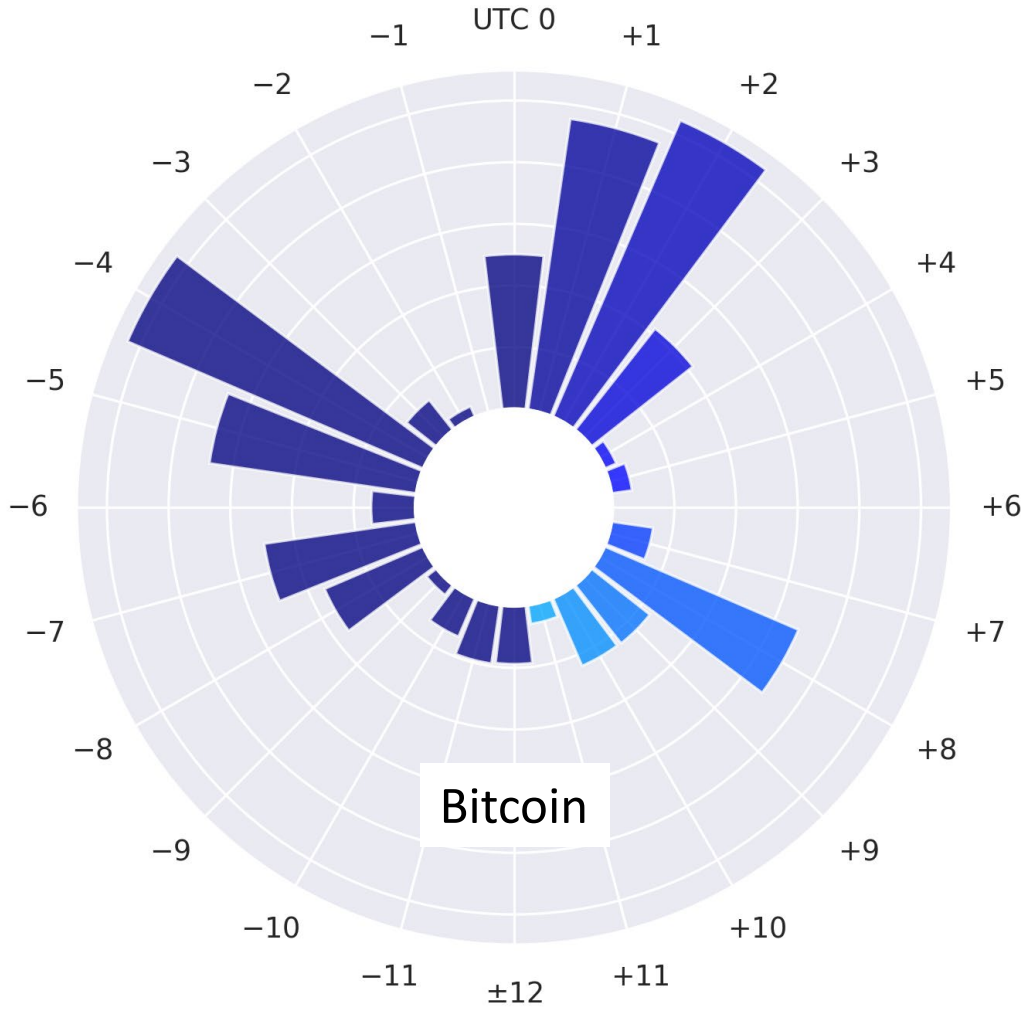


No correlation between ranking and stars



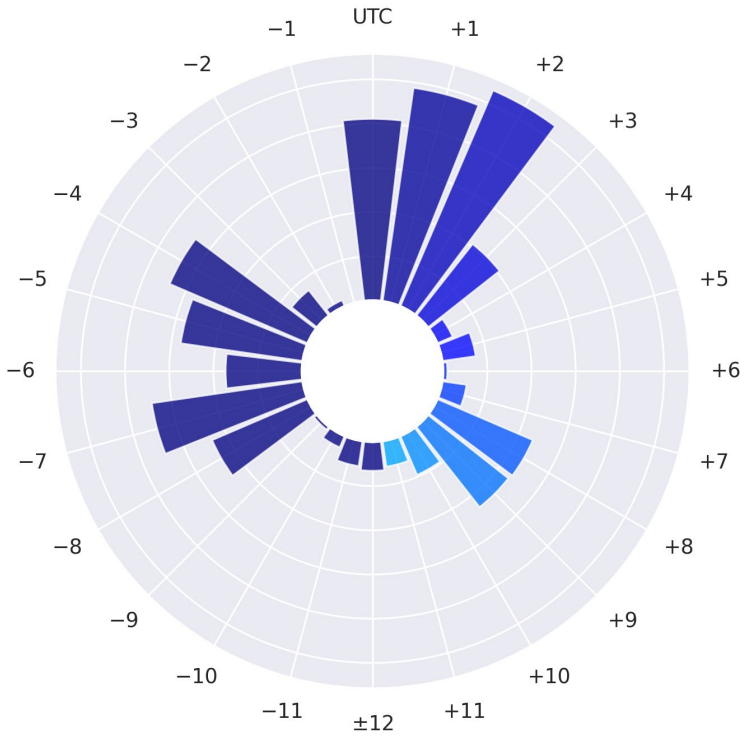
Geographic Dispersion over 48 months



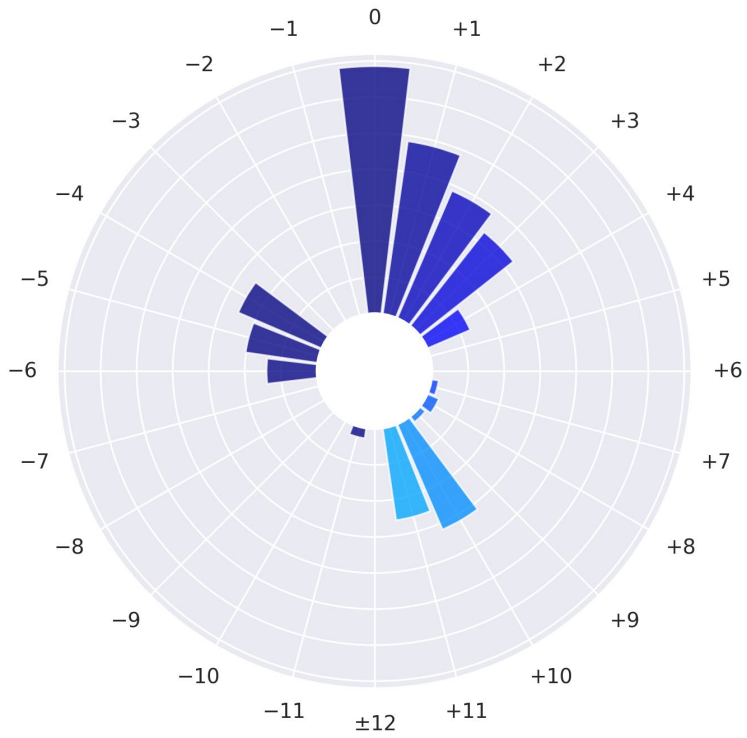


UTC-1 (Atlantic ocean: Azores, Cape Verde) has no activity
UTC+6 (Kazakhstan, Kyrgyzstan, Bangladesh, Bhutan) almost no activity
UTC-9 (Alaska, Marquesas Islands) also with minimal activity

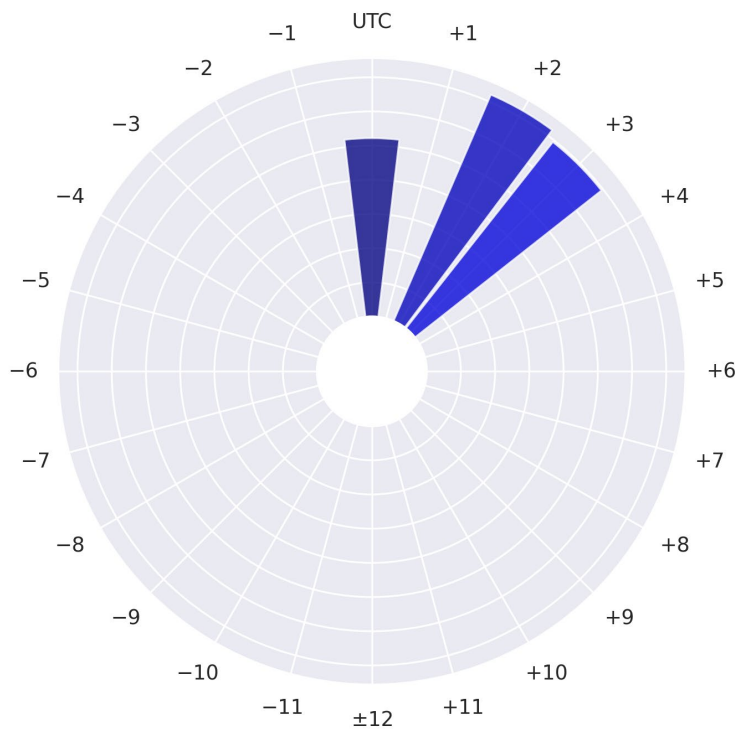
Which project is better?



standard



Cardano



Syntropy

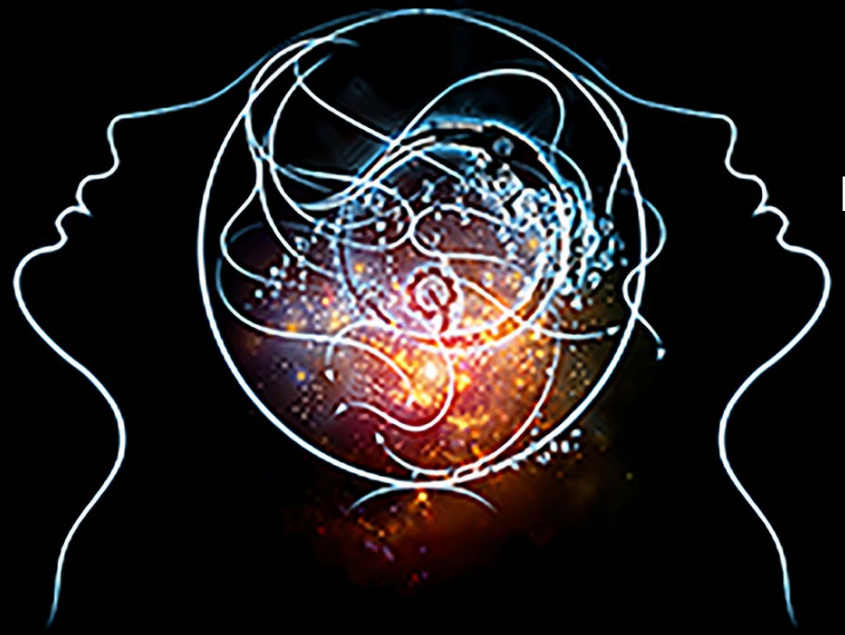
So What?

- Identify areas in need of improvement / successful
- People have finite ‘volunteer’ hours; find the innovation
- **These predictors may be able to *see* innovation**

Up Next:

- Polish (& automate) the data collection
- validate (or invalidate) the models

Thanks & Questions



Jeff Nijsse

Postgraduate Research Symposium
November 26, 2021
Auckland, New Zealand

AUT

References

- GitHub, “Empowering Healthy Communities,” 2020. [Online]. Available: https://octoverse.github.com
- M. Goeminne and T. Mens, “Analyzing ecosystems for open source software developer communities,” in Software Ecosystems: Analyzing and Managing Business Networks in the Software Industry, no. 2013, S. Jansen, Ed. 2013, pp. 247–275.
- J. F. Hair Jr., W. C. Black, B. J. Babin, and R. E. Anderson, “Multivariate Data Analysis”, Seventh. Essex: Pearson Education Limited, 2014.
- Y. Rosseel, “lavaan: An R Package for Structural Equation Modeling”. In Journal of Statistical Software (Vol. 48, Issue 2, pp. 1–36). 2012. http://www.jstatsoft.org/v48/i02/
- Y. Rosseel, “Structural equation modeling with lavaan,” Geneve, 2020. https://users.ugent.be/~yrosseel/lavaan/geneve2020/lavaan_oneday_geneve2020.pdf
- J.-F. Schrape, “Open Source Projects as Incubators of Innovation: From Niche Phenomenon to Integral Part of the Software Industry,” SSRN Electron. J., 2017, doi: 10.2139/ssrn.2977352.
- S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” 2008. https://bitcoin.org/bitcoin.pdf

Model fit evaluation

- Only a small discrepancy between the observed covariance matrix (data) and the model-predicted covariance matrix
- Chi-square statistic not significant indicates the model is a good fit
- Comparative Fit Index (CFI) ≥ 0.95
- Root Mean Square Error of Approximation (RMSEA) ≤ 0.08
- Standardized Root Mean Square Residual (SRMR) ≤ 0.08
- Be careful! These ‘rules’ are rather guidelines…

(Rosseel, 2020)