Predictability of Scientific Success

Raj Kumar Pan

Aalto University
School of Science and Technology
Helsinki, Finland
Is a scientist's "success" predictable?
How much is the predictability?
How it works in reality?
How it works in reality?
How it works in reality?
How it works in reality?

Assumption is 100% predictable.
How to quantify?
Assumptions:

- Success of scientists: total publications, total citations, h-index

**h-index:** is x if a scientist has x paper with more than x citations
What are we after?
What are we after?

\[ h(t) \rightarrow h(t + \Delta t) \]

\( h(t + \Delta t) \) depends on

- \( h(t) = \) H-index at career age \( t \)
- \( n_p(t) = \) number of publications (co)authored
- \( j(t) = \) number of distinct journals of publications
- \( q(t) = \) number of papers in high impact journals
- \( t = \) Career age of scientist
Quite predictable

Data: ~3000 Neuroscientist
D. E. Acuna, S. Allesina, K. P. Kording,
Different age cohort?

\[ h(t) \rightarrow h(t + \Delta t) \]

\( h(t + \Delta t) \) depends on

\[ h(t) = \text{H-index at career age } t \]
\[ n_p(t) = \text{number of publications (co)authored} \]
\[ j(t) = \text{number of distinct journals of publications} \]
\[ q(t) = \text{number of papers in high impact journals} \]
Young scientists are less predictable

200 Prolific authors of Physical Review Letters (PRL)
100 Prolific authors of Cell
Non-stationary time series

Distribution of annual increment of h-index

\[ h^s = \sum_{i=1}^{t} \Delta h^s_i \]
Non-stationary time series

\[ Cor[h^s(t + \Delta t), h^s(t)] = \sqrt{\frac{t}{t+\Delta t}} \]

Correlation \( \rightarrow \) 1 if \( t \) is large

or scientist’s of all ages are combined together
Non-stationary time series
Non-Cumulative measures

$$\Delta h(t, \Delta t) = h(t + \Delta t) - h(t)$$

$$\Delta h(t, \Delta t)$$ depends on

- \(h(t) = \text{H-index at career age } t\)
- \(n_p(t) = \text{number of publications (co)authored}\)
- \(j(t) = \text{number of distinct journals of publications}\)
- \(q(t) = \text{number of papers in high impact journals}\)
Non-Cumulative measures

Prominent physicists

Prominent biologists

Prominent mathematicians

Physics assistant prof

Biologists

Graphene researcher

Monday, November 18, 13
Non-Cumulative measures

Prominent physicists

Prominent biologists

Prominent mathematicians

Physics assistant prof

Biologists

Graphene researcher

Monday, November 18, 13
Non-Cumulative measures

Prominent physicists $t=3$

Prominent biologists $t=3$

C $t=10$

D $t=10$

Monday, November 18, 13
Correlating true past and true future
Prominent Physicist

$R^2 = 0.33$

$R^2 = 0.32$

$R^2 = 0.33$

$R^2 = 0.32$

$R^2 = 0.31$

$R^2 = 0.36$
Prominent Biologist

\[
\sqrt{n_c(t|\{T_{early}\})} = 0.34
\]

\[
\sqrt{n_c(t|\{T_{mid}\})} = 0.43
\]

\[
h(t|\{T_{early}\}) = 0.50
\]

\[
h(t|\{T_{mid}\}) = 0.50
\]
Asst. Prof. in Physics

$R^2 = 0.06$

$R^2 = 0.18$

$R^2 = 0.01$

$R^2 = 0.11$
Conclusions:

- Cumulative measure over-estimate prediction
- Correlation depends on career age
- Parameters of prediction also depend on age
- Focus on young scientist: unpredictable?

Divinations of academic success may be flawed

Method for predicting future $h$-index comes under fire.

Elizabeth Gibney

08 November 2013

Scientists reacted with curiosity — as well as disdain — to the launch last year of a nifty tool that, according to its architects, could predict a researcher's future scientific impact. A study published last week questions its power to do so.
Conclusions:

• Cumulative measure over-estimate prediction
• Correlation depends on career age
• Parameters of prediction also depend on age
• Focus on young scientist: unpredictable?