Scientific Creativity as a Combinatorial Process

The Chance Baseline
Goal

- Formulate a theory of scientific creativity
  - that uses
    - Parsimonious assumptions and
    - Logical derivations
  - to obtain
    - Comprehensive explanations and
    - Precise predictions
  - with respect to the most secure empirical results
- In other words, getting the most with the least
Argument: Part One

- Combinatorial models
  - currently get the most with the least relative to any alternative theory.
    - That is, such models
    - make the fewest assumptions,
    - and by logical inferences
    - explain the widest range of established facts
    - and make the most precise predictions with respect to those data
Argument: Part Two

- Even if combinatorial models are incomplete from the standpoint of one or more criteria, such models must still provide the baseline for comparing all alternative theories.
- That is, rival theories must account for whatever cannot be accounted for by chance alone, or what exceeds the chance baseline (cf. “null” hypothesis; research on the “hot hand” or parapsychology; etc.)
Creativity in Science: Two Critical Research Sites

- Scientific Careers:
  - Publications

- Scientific Communities:
  - Multiples
Publications

- Individual Variation
- Longitudinal Change
Individual Variation

- Skewed Cross-sectional Distribution
Individual Variation

- Skewed Cross-sectional Distribution
  - 10% → 50% / 50% → 15%
Individual Variation

- Skewed Cross-sectional Distribution
  – Lotka’s Law
Individual Variation

- Skewed Cross-sectional Distribution
  - Lotka’s Law:
    - \( f(T) = kT^{-2} \) or \( \log f(T) = \log k - 2 \log T \)
Individual Variation

- Skewed Cross-sectional Distribution
  - Lotka’s Law:
    • $f(T) = k T^{-2}$ or $\log f(T) = \log k - 2 \log T$
    • where $T$ is total lifetime output
Individual Variation

■ Skewed Cross-sectional Distribution
  – Lotka’s Law:
  – Price’s Law:
    • $N^{1/2} \rightarrow 50\%$ of total field output
Individual Variation

- Skewed Cross-sectional Distribution
  - Lotka’s Law:
  - Price’s Law:
    - $N^{1/2} \rightarrow 50\%$ of total field output
    - where $N$ is size of field
Individual Variation

- Skewed Cross-sectional Distribution
- Quantity-Quality Relation
Individual Variation

- Skewed Cross-sectional Distribution
- Quantity-Quality Relation
  - Equal-Odds Baseline: $H_i = \rho_1 T_i + u_i$
Individual Variation

- Skewed Cross-sectional Distribution
- Quantity-Quality Relation
  - Equal-Odds Baseline: \( H_i = \rho_1 T_i + u_i \)
  - where \( \rho_1 \) is the overall “hit rate” \( 0 < \rho_1 < 1 \)
    for individuals in a given domain
Individual Variation

- Skewed Cross-sectional Distribution

- Quantity-Quality Relation
  - Equal-Odds Baseline: $H_i = \rho_1 T_i + u_i$
  - where $\rho_1$ is the overall “hit rate” ($0 < \rho_1 < 1$) for individuals in a given domain,
  - $H_i$ is the number of “hits” (e.g., high-impact publications) for individual $i$, and
  - the random shock $0 \leq u_i \leq T_i (1 - \rho_1)$
Productivity (Quantity) vs. Citations (Quality)

- Perfectionists
- Mass Producers
- Prolific
- Silent
- Mass Producers

The graph shows a positive correlation between productivity and citations, with a trend line indicating an increase in citations as productivity increases.
Individual Variation

- Skewed Cross-sectional Distribution
- Quantity-Quality Relation
  - Equal-Odds Baseline: \( H_i = \rho_1 T_i + u_i \)
  - where \( \rho_1 \) is the overall “hit rate” (0 < \( \rho_1 < 1 \)) for individuals in a given domain,
  - \( H_i \) is the number of “hits” (e.g., high-impact publications) for individual \( I \), and
  - the random shock \( 0 \leq u_i \leq T_i (1 - \rho_1) \)
  - N.B.: If \( \rho_1 \) were a linear function of \( T_i \), then the overall function would be quadratic, not linear
Longitudinal Change

- Randomness of Annual Output
  - No “runs”
  - Poisson Distribution
    \[ P(j) = \mu^j e^{-\mu} / j! \]
    \[ e = 2.718... \text{ and } j! = 1 \times 2 \times 3 \times ... \times j \]
Representative Productivity Distributions for 10 Hypothetical Scientists

| Scientist | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|-----------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| 1         | 1 | 1 | 0 | 2 | 2 | 1 | 3 | 2 | 0 | 1 | 0 | 3 | 0 | 1 | 2 | 1 | 1 | 0 | 2 | 2 |
| 2         | 2 | 2 | 0 | 1 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 2 | 1 | 1 | 1 | 2 | 0 | 1 |
| 3         | 2 | 1 | 1 | 0 | 2 | 0 | 1 | 3 | 0 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 |
| 4         | 0 | 1 | 2 | 0 | 2 | 0 | 1 | 3 | 1 | 4 | 0 | 0 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 2 |
| 5         | 2 | 1 | 0 | 1 | 0 | 1 | 1 | 3 | 2 | 1 | 1 | 2 | 3 | 2 | 1 | 1 | 2 | 1 | 0 | 0 |
| 6         | 0 | 0 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 0 | 1 | 1 | 2 | 0 | 1 | 3 | 2 | 2 | 2 | 2 |
| 7         | 2 | 2 | 0 | 1 | 2 | 0 | 1 | 1 | 2 | 3 | 1 | 2 | 0 | 3 | 1 | 2 | 1 | 0 | 1 | 0 |
| 8         | 1 | 2 | 0 | 2 | 2 | 1 | 3 | 0 | 1 | 1 | 3 | 2 | 1 | 0 | 0 | 1 | 0 | 1 | 2 | 2 |
| 9         | 2 | 1 | 0 | 2 | 1 | 1 | 2 | 4 | 0 | 0 | 2 | 1 | 3 | 0 | 1 | 1 | 0 | 2 | 1 | 1 |
| 10        | 1 | 1 | 2 | 1 | 2 | 1 | 0 | 3 | 2 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 0 | 0 | 1 | 0 |

*Note.* Each scientist is presumed to produce 25 contributions randomly distributed over 20 career years, with a Poisson distribution for the number contributions per yearly unit (where $\mu = 1.25$).
Longitudinal Change

- Randomness of Annual Output
- Quantity-Quality Relation
  - Random Fluctuation around a Quality Ratio Baseline
  - Hence, the Equal-Odds Baseline:
    \[ H_{it} = \rho_2 T_{it} + u_{it} \quad (\rho_2 = \rho_1 \text{ if estimated from the same cross-sectional sample}) \]
  - for the \( it \)th individual in career year \( t \),
  - and where \( 0 \leq u_{it} \leq T_{it} (1 - \rho_2) \)
Multiples

- Distribution of Multiple Grades
Multiple Grade

0 1 2 3 4 5 6 7 8 9 10

Frequency

Simonton
Merton
Ogburn and Thomas

- Ogburn and Thomas
- Merton
- Simonton

Multiple Grade

Frequency

Multiple Grade
Multiples

- Distribution of Multiple Grades
- Temporal Separation of Multiple Discoveries
Multiples

- Distribution of Multiple Grades
- Temporal Separation of Multiple Discoveries
- Individual Variation in Multiple Participation
Multiples

- Distribution of Multiple Grades
- Temporal Separation of Multiple Discoveries
- Individual Variation in Multiple Participation
- Degree of Multiple Identity
Combinatorial Processes

- Definitions
- Assumptions
- Implications
- Elaboration
- Integration
Definitions

- Individual
- Domain
- Field
Assumptions

- Individual Samples from Domain Ideas
Assumptions

- Individual Samples from Domain Ideas
  - Assume samples random or quasi-random
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
  - Postulate a normal distribution
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
- Quasi-Random Combination of Ideas
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
- Quasi-Random Combination of Ideas
  - Variable degrees of constraint depending on nature of domain
    - Scientific revolutionaries vs. normal scientists
    - Paradigmatic vs. nonparadigmatic scientists
    - Scientists vs. artists
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
- Quasi-Random Combination of Ideas
- Variation in Quality of Combinations
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
- Quasi-Random Combination of Ideas
- Variation in Quality of Combinations
  - Differential fitness with respect to scientific criteria (facts, logic, etc.)
  - Small proportion publishable, an even smaller proportion high impact
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
- Quasi-Random Combination of Ideas
- Variation in Quality of Combinations
- Variation in Size of Fields
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
- Quasi-Random Combination of Ideas
- Variation in Quality of Combinations
- Variation in Size of Fields
- Communication of Ideational Combinations
Assumptions

- Individual Samples from Domain Ideas
- Within-Field Variation in Sample Size
- Quasi-Random Combination of Ideas
- Variation in Quality of Combinations
- Variation in Size of Fields
- Communication of Ideational Combinations
  - If accepted, then incorporation into the domain pool, completing the cycle
Communication-Incorporation:

- Rate increases with speed of
  - Communication practices (journals vs. books; least-publishable units)
  - Gate-keeping procedures (peer review; editorial policies)
  - Publication lags (1st- vs. 2nd-tier journals)
  - Diffusion to secondary sources (introductory texts, popularizations, etc.)

- Hence, variation across time and discipline
Implications

- Research Publications
  - Cross-sectional Variation
  - Longitudinal Change
Implications

- Multiple Discoveries
  - Multiple Grades
Multiple Grade

Frequency

- Observed
- Predicted

Multiple Grade

0 1 2 3 4 5 6 7 8 9 10
Implications

- Multiple Discoveries
  - Multiple Grades
    • Variation across time and discipline
  - Temporal Separation
    • Variation across time and discipline
Implications

- Multiple Discoveries
  - Multiple Grades
  - Temporal Separation
  - Multiples Participation
Implications

- Multiple Discoveries
  - Multiple Grades
  - Temporal Separation
  - Multiples Participation
    - Number of ideational combinations
    - Number of overlapping domain samples
Implications

- Multiple Discoveries
  - Multiple Grades
  - Temporal Separation
  - Multiples Participation
  - Multiple Identity
Elaboration

- Aggregated Data on Career Output
  - Aggregated Across Time Units
  - Aggregated Across Scientists

- Cognitive Combinatorial Model
  - Two-step process
    - Ideation generates combinations
    - Elaboration generates communications
  - Individual differences in
    - Domain sample
    - Career onset
\[ p(t) = abm(b - a)^{-1}(e^{-at} - e^{-bt}) \]

where \( p(t) \) is ideational output at career age \( t \) (in years),
- \( e \) is the exponential constant (~ 2.718),
- \( a \) the typical ideation rate for the domain \((0 < a < 1)\),
- \( b \) the typical elaboration rate for the domain \((0 < b < 1)\),
- \( m \) the individual’s creative potential (i.e. maximum number of ideational combinations in indefinite lifetime).

If \( a = b \), then \( p(t) = a^2 m t e^{-at} \)

Number of communications \( T_{it} \) is proportional to \( p \).

Individual differences in
- Creative potential \((m)\)
- Age at career onset (i.e., chronological age at \( t = 0 \))
Implications

- Typical Career Trajectories
Annual Productivity ($p$)

\[ \frac{dp}{dt} = 0 \]

\[ \frac{d^2p}{dt^2} = 0 \]

\[ p(t) = 61 \left( e^{-0.04t} - e^{-0.05t} \right) \]
N.B.

- The above curve has been shown to correlate in the mid- to upper-.90s for numerous data sets in which output information has been aggregated across many individual careers.

- Yet even in the case of highly productive individuals, the predicted curve does reasonably well.
e.g., the career of Thomas Edison

\[ C_{Edison}(t) = 2595(e^{-0.044t} - e^{-0.058t}) \]

\[ r = 0.74 \]
Implications

- Typical Career Trajectories
- Individual Differences in Trajectories
  - Fourfold Typology
    - High versus Low Creative Potential
    - Early versus Late Age at Career Onset
High Creative Early Bloomers

Low Creative Early Bloomers

High Creative Late Bloomers

Low Creative Late Bloomers
Specific Prediction

- Individual differences in output across consecutive age periods (5- or 10-year units) for scientists with same age at career onset yields a specific pattern of correlations across those units, namely one most consistent with
  - a single-factor model, rather than
  - an autoregressive (simplex or quasi-simplex) model.
Single-Factor Model

$\begin{align*} m \end{align*}$

20s → 30s → 40s → 50s → 60s

Autoregressive Model

20s → 30s → 40s → 50s → 60s
Former single-factor model already confirmed on distinct data sets (e.g., there is no tendency for the correlations between two age periods to decline as a function of the temporal separation between the two periods; i.e., no decline with distance from matrix diagonal)
Implications

- Typical Career Trajectories
- Individual Differences in Trajectories
- Domain Variation in Trajectories
The diagram illustrates the percent of total lifetime output by ARTISTS, SCIENTISTS, and SCHOLARS across different age decades. The x-axis represents age in decades, ranging from 10 to 80, and the y-axis represents the percent of total lifetime output, ranging from 0 to 30. The graph shows distinct peaks and trends for each category, indicating variations in productivity across different ages.
Annual Productivity ($p$)

- $p(t) = 30(e^{-0.05t} - e^{-0.06t})$
- $p(t) = 20(e^{-0.04t} - e^{-0.05t})$

Age

0 20 30 40 50 60 70 80
### Estimates for Three Disciplines

<table>
<thead>
<tr>
<th>Domain</th>
<th>a</th>
<th>b</th>
<th>Career</th>
<th>Chrono-logical</th>
<th>Half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemists</td>
<td>.042</td>
<td>.057</td>
<td>20.4</td>
<td>40.4</td>
<td>16.5</td>
</tr>
<tr>
<td>Biologists</td>
<td>.033</td>
<td>.052</td>
<td>23.9</td>
<td>43.9</td>
<td>21.0</td>
</tr>
<tr>
<td>Geologists</td>
<td>.024</td>
<td>.036</td>
<td>33.8</td>
<td>53.8</td>
<td>28.9</td>
</tr>
</tbody>
</table>
Implications

- Typical Career Trajectories
- Individual Differences in Trajectories
- Domain Variation in Trajectories
- Placement of Career Landmarks  
  – Across domains
Annual Productivity vs. Career Age

- "Best" Work
- First "Hit"  Last "Hit"

Graph showing the relationship between annual productivity and career age, highlighting key milestones such as "Best" Work, First "Hit," and Last "Hit" during a career.
- First Major Contribution
- Best Contribution
+ Last Major Contribution
Implications

- Typical Career Trajectories
- Individual Differences in Trajectories
- Domain Variation in Trajectories
- Placement of Career Landmarks
  - Across domains
  - Across individuals
Specific Predictions

Given the above, it is possible to derive predictions regarding the pattern of correlations among

- the ages of the three career landmarks ($F$, $B$, $L$),
- the age at maximum output rate ($x$),
- final lifetime productivity ($T$),
- the maximum output rate ($X$), and
- the time lapse or delay ($d$) between career onset and first career landmark (i.e., preparation period)

In particular …
Specific Predictions

1A: Total lifetime productivity correlates

– negatively with the chronological age of the first contribution ($r_{TF} < 0$) and
– positively with the chronological age of the last contribution ($r_{TL} > 0$).
Specific Predictions

1B: Maximum output rate correlates

– negatively with the chronological age of the first contribution ($r_{XF} < 0$) and
– positively with the chronological age of the last contribution ($r_{XL} > 0$).
Specific Predictions

- **2A: Total lifetime productivity correlates**
  - zero with the chronological age at the maximum output rate \( (r_{Tx} = 0) \) and
  - zero with the chronological age at the best contribution \( (r_{TB} = 0) \).
Specific Predictions

- **2B: Maximum output rate correlates**
  - zero with the chronological age at the maximum output rate \((r_{XX} = 0)\) and
  - zero with the chronological age at the best contribution \((r_{XB} = 0)\).
Specific Predictions

- 3A: The chronological age at the maximum output rate correlates positively with both
  - the chronological age at the first contribution ($r_{xF} > 0$) and
  - the chronological age at the last contribution ($r_{xL} > 0$).
Specific Predictions

- 3B: The chronological age of the best contribution correlates positively with both
  - the chronological age at the first contribution \( (r_{FB} > 0) \) and
  - the chronological age at the last contribution \( (r_{BL} > 0) \).
Specific Predictions

4: The first-order partial correlation between the ages of first and last contribution is negative after partialling out either

- the chronological age at the best contribution \((r_{FL,B} = r_{FL} - r_{FB}r_{LB} < 0)\) or
- the chronological age at the maximum output rate \((r_{FL,x} = r_{FL} - r_{Fx}r_{Lx} < 0)\)
Specific Predictions

5: The time interval between the chronological age at career onset and the chronological age at first contribution is negatively correlated with both

- total lifetime productivity ($r_{Td} < 0$) and
- the maximum output rate ($r_{Xd} < 0$).
Discussion

- Foregoing predictions unique to the combinatorial model
  - That is, they cannot be generated by alternative theories (e.g., cumulative advantage, human capital)

- Furthermore, all predictions have been confirmed on several independent data sets
Discussion

Moreover, if we assume that eminence (E) is highly correlated with lifetime productivity ($r_{ET} >> 0$), then we obtain additional predictions:

- *Eminence correlates*
  - negatively with the age of the first contribution ($r_{EF} < 0$),
  - positively with the age of the last contribution ($r_{EL} > 0$),
  - zero with the age at the maximum output rate ($r_{Ex} = 0$),
  - zero with the age at the best contribution ($r_{EB} = 0$), and
  - negatively with the time interval between the age at career onset and the age at first contribution ($r_{Ed} < 0$)

These predictions also empirically confirmed
Integration: Combinatorial Process Emerges from …

- Creative Scientists
- Research Programs
- Research Collaborations
- Peer Review
- Professional Activities
- Individual-Field-Domain Effects

\[ \frac{dI}{dt} = \gamma I/N \]
Conclusion

- Because combinatorial models work so well with respect to scientific creativity
- (and because they have been extended successfully to non-scientific creativity),
- they seem to provide a valid baseline for gauging other explanations.
- Hence the next question: What other processes or variables add an increment to the variance already explained by combinatorial models?