Capturing Value from Big Data through Data-Driven Business Models

Patterns from the Start-up world

Philipp Hartman,
Dr Mohamed Zaki and Prof Duncan McFarlane
Cambridge Service Alliance
University of Cambridge
“Data is the new oil”\(^1\)

\(^1\) various authors, e.g. Clive Humby

IDC’s Digital Universe Study, December 2012

**Top Priority:**
“How to get value from big data” \(^3\)

\(^3\) Gartner “Big Data Study” 2013
Two general areas can be identified where big data creates value.

How to get value from Big Data?

- Optimization of existing business
  - Chesbrough, Rosenbloom (2002): Business model to capture value from an innovation
  - Crisculo (2012): New technologies often first commercialized through start-up companies

- New Business Models

---

Based on this motivation the research question was developed.

**Research Question**

What types of business models that rely on data as a key resource (i.e. data-driven business models) can be found in start up companies?

**Sub-questions**

- How to analyse data-driven business models?
- How to identify patterns?

**Data-driven business model framework**

**Clustering**
The research was done in five steps:

1. **Literature Review**
2. **Build the framework**
3. **Data collection & coding**
4. **Finding Patterns**
5. **Case studies**

- How to analyse data-driven business models?
- How to identify patterns?
The first step was a literature review with three different topics:

- **Literature Review**
  - Big Data
  - Business Model
  - Related Work

- **Build the framework**
- **Data collection & coding**
- **Finding Patterns**
- **Case studies**
The first step was a literature review with three different topics.
Literature review: Business Model

- Literature Review
- Build the framework
- Data collection & coding
- Finding Patterns
- Case studies

- Big Data
  - Definition
  - Business Model Frameworks
- Business Model
  - Related Work

- Data driven business models
- Cloud business models
Business model key components were synthesized from existing frameworks.

**Business Model Definition**
- No universally accepted definition of the concept (Weill, Malone et al. 2011)
- Most definitions refer to value creation & value capturing

**Existing Business Model Frameworks**
- Chesbrough & Rosenbloom 2002
- Hedman & Kaling 2003
- Osterwalder 2004
- Morris 2005
- Johnson, Christensen et. al. 2008
- Al-Debei 2010
- Burkhart 2012

**Business Model Key Components**
- Value Creation
  - Key Resources
  - Key Activities
  - Value Proposition
- Value Capturing
  - Customer Segment
  - Revenue Model
  - Cost structure
Only a few papers are available in this field.
The review was extend to cloud business models.

- Literature Review
  - Build the framework
  - Data collection & coding
  - Finding Patterns
  - Case studies

- Literature Review
  - Business Model
  - Related Work
    - Data driven business Models
    - Cloud business models
The literature review identified several gaps:

- Little academic research on big data and value creation – mostly whitepapers
- Gap in literature: data-driven business models
- Otto, Aier (2013) interesting paper but limited to specific industry > no generalization possible
- Similar research for cloud business models (cf. Labes, Erek et. Al. 2013)
The framework was build from literature starting from the key components.
Synthesizing the different sources leads to the taxonomy

**Data Sources**
- **Internal**
  - existing data
  - Self-generated Data
- **Acquired Data**
  - Customer provided
  - Free available
- **External**
  - Open Data
  - Social Media data
  - Web Crawled Data
Dimension: Activities

Key Activity

- Data Generation
  - Crowdsourcing
  - Tracking & Other
- Data Acquisition
- Processing
- Aggregation
- Analytics
  - descriptive
  - predictive
  - prescriptive
- Visualization
- Distribution
Dimension: Revenue Model

- Asset Sale
- Lending/Renting/Leasing
- Licensing
- Usage fee
- Subscription fee
- Advertising
Dimension: Target Customer

- Target Customer
  - B2B
  - B2C
The final framework

Literature Review → Build the framework → Data collection & coding → Finding Patterns → Case studies

Data Sources:
- Internal
- External
- Acquired Data
- Customer-provided
- Open Data
- Free available
- Social Media data
- Web Crawled Data

Key Activity:
- Data Generation
- Data Acquisition
- Processing
- Aggregation
- Analytics
- Visualization
- Distribution
- Descriptive
- Predictive
- Prescriptive

Data-Driven Business Model:
- Data
- Information/Knowledge
- Non-Data Product/Service

Target Customer:
- B2B
- B2C
- Asset Sale
- Selling/Renting/Leasing
- Licensing
- Usage Fee
- Subscription Fee
- Advertising

Revenue Model:
- Asset Sale
- Lending/Renting/Leasing
- Licensing
- Usage Fee
- Subscription Fee
- Advertising

Specific cost advantage
Data collection and coding

Literature Review → Build the framework → Data collection & coding → Finding Patterns → Case studies

Sampling → Data collection → Data analysis
The data was generated using public available sources.

**AngelList**
- Tag: “big data”
- “big data analytics”
- **1329 companies**

**Data collection**
- Company information
  - Company websites
  - Press releases
- **Public sources**
  - The New York Times
  - TechCrunch
  - Forbes
- **299 Sources**
  - ~3 sources/comp

**Sampling**
- Random sample
- Cleaning
- **100 Companies**

**Data analysis**
- Coding of sources using data driven business model framework
- Nvivo
- **100 binary feature vectors**
Overall Analysis: Data Source

- >50% of companies rely on free available data
- >50% of companies use data provided by customers/partners

Note: Sum > 100% as companies might use multiple data sources
Overall Analysis: Key Activities

- Aggregation
- Analytics
- Descriptive Analytics
- Predictive Analytics
- Prescriptive Analytics
- Data acquisition
- Data generation
- Data processing
- Distribution
- Visualization

Note: Sum > 100% as some companies rely on multiple revenue models

• >70% of companies use analytics - mostly descriptive
Overall Analysis: Revenue Model

- Majority of revenue models based on subscription and/or usage fee
- No information about the revenue model as many companies are in an early stage

Note: Sum > 100% as some companies rely on multiple revenue models
Overall Analysis: Target Customer

- There seems to be a noteworthy predominance of B2B business models
- But no reference data found
BM patterns were identified using a clustering approach


Han, Jiawei; Kamber, Micheline (2011): Data mining. Concepts and techniques.


7 Business Model Cluster were identified

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquired Data</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Customer-provided Data</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Free available</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CrowdSourced</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tracked, Generated &amp; other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Key Activity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggregation</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Analytics</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Data acquisition</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Data generation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Number of companies</strong></td>
<td>17</td>
<td>28</td>
<td>5</td>
<td>16</td>
<td>14</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
</tbody>
</table>
6 significant Business Model types were identified

Type A: “Free Data Collector & Aggregator”

Type B: “Analytics-as-a-Service”

Type C: “Data generation & Analytics”

Type D: “Free Data Knowledge Discovery”

Type E: “Data Aggregation-as-a-Service”

Type F: “Multi-Source data mashup and analysis”
The 6 BM types are characterised by the key activities and key data sources.
Type D: “Free Data Knowledge Discovery”

Companies
1. DealAngel
2. Gild
3. Insightpool
4. Juristat
5. Market Prophit
6. MixRank
7. Numberfire
8. Olery
9. PeerIndex
10. PolyGraph
11. Review Signal
12. Tellagence
13. traackr
14. Trendspottr

Key Data Source
- Free available
  - Social Media
  - Open Data
  - Web Crawled

Key Activities
- Analytics

Revenue Model

Target Customer

- B2B
- B2C
Type D: Examples

“Using patent-pending technology, Gild evaluates the work of millions of developers so companies using Gild’s talent acquisition tools know who’s good and can target the right candidates.”

- **Key Data:** Free available websites (GitHub, Google Codes)
- **Key Activities:** Analytics
- **Revenue Model:** Monthly subscription
- **Target Customer:** B2B

“Our goal is to provide the most accurate and honest reviews possible by using the data consumers create. We listen to the conversations, analyze them and visualize them for consumers.”

- **Key Data:** Twitter
- **Key Activities:** Analytics
- **Revenue Model:** Advertising
- **Target Customer:** B2B (B2C)
The cases studies will be validated the framework and the clustering

Purpose:

1. Validate framework & clusters
2. Illustrate business model types through examples
3. Identify specific challenges

4 case studies with companies from the sample such as

- AGiLE customer insight
- GoSquared
- OpenSignal
- Gap in literature identified

- **RQ:** What types of business models that rely on data as a key resource (i.e. data-driven business models) can be found in start up companies?

  - 5 (7) different business model patterns identified
  
  - Data-driven business model framework created to enable analysis

  - Pattern identification through clustering

  - Validation through Case-Studies
## Limitations & Outlook

### Limitations

- Only 100 samples
- Only start up companies
- Bias of data source (AngelList)
- Statistical significance of clustering result
- Only public available sources used
- No statement about success of a particular business model

### Outlook/Next Steps

1. **Improve validity of findings**
   1. Increase sample size to test clusters
   2. More Case-studies to illustrate/validate clusters
2. **Include established organizations**
3. **Develop methodology to judge (financial) performance of different business models**
Appendix


Hagen, Christian; Khan, Khalid; Ciobo, Marco; Miller, Jason; Wall, Dan; Evans, Hugo; Yadava, Ajay (2013): Big Data and the Creative Destruction of Today's Business Models. ATKearney.


Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles; Hung Byres, Angela (2011): Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Morris, Michael; Schindehutte, Minet; Allen, Jeffrey (2005): The entrepreneur's business model: toward a unified perspective. In Special Section: The Nonprofit Marketing Landscape 58 (6), pp. 726–735.


Schroeck, Michael; Shockley, Rebecca; Smart, Janet; Romero-Morales, Dolores; Tufano, Peter (2012): Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data. IBM Institute for Business Value; Saïd Business School at the University of Oxford.


The Clustering Process

1. Clustering Variables
   - Variables relevant to determine clustering (Miligan 1996)
   - #Variables has to match #samples (Mooi 2011)
   - Avoid high correlation between variables (<0.9) (Mooi 2011)

2. Clustering method
   - 2 Dimensions: “Data source” & “Key Activity”

3. Number of Clusters
   - ~ $2^m$ samples for $m$ variables: 6-7 variables

4. Validate & Interpret C.
   - 9 variables
   - max. correlation: 0.5

---

Variables relevant to determine clustering (Miligan 1996)

#Variables has to match #samples (Mooi 2011)

Avoid high correlation between variables (<0.9) (Mooi 2011)
The Clustering Process

1. Clustering Variables
2. Clustering Method
3. Number of Clusters
4. Validate & Interpret C.

Clustering Method (Han 2011)
- Partitioning
  - Hierarchical
  - Density-based
  - Grid-based

Proximity Measure
- Include neg. match
- Exclude neg. match

K-Medoids
- Euclidean Distance
There is no “one right solution” for the number of clusters

1. Clustering Variables

2. Clustering method

3. Number of Clusters

4. Validate & Interpret C.

large to reflect specific differences

$k << n$

Several different approaches (Pham 2005, Mooi 2011, Han 2011, Everitt et. al. 2011):

1. Use a-priori knowledge to determine number of clusters ×

2. Visual approaches ×

3. Rule of thumb (Han 2011): $k \sim \sqrt{\frac{n}{2}} \rightarrow k \sim 7$

4. “Elbow” method

5. Statistical methods
"Elbow" method


1. Hierarchical clustering first
2. Plot agglomeration coefficient against number of clusters
3. Search for "elbows"
“Elbow” method

1. Clustering Variables
2. Clustering method
3. Number of Clusters
4. Validate & Interpret C.

Clustering Coefficient (distance)

Number of cluster k
Statistical Measure: Silhouette

1. Clustering Variables
2. Clustering method
3. Number of Clusters
4. Validate & Interpret C.

Silhouette Coefficient $s(i)$

For datum $i$:

Compares distance within its cluster to distance to nearest neighbouring cluster

$-1 \leq s(i) \leq 1$

The Clustering Process

1. Clustering Variables
2. Clustering method
3. Number of Clusters
4. Validate & Interpret C.

Silhouette Value
- no cluster
- good

Silhouette Value: 0.335