Bibliometric data and actual development in technology life cycles: flaws in assumptions

BIBLIOMETRIC DATA AND ACTUAL DEVELOPMENT IN TECHNOLOGY LIFE CYCLES: FLAWS IN ASSUMPTIONS

STRUCTURED ABSTRACT

Purpose
Motivated with the ever growing number of bibliometric trend extrapolation studies, we demonstrate through two technologies how the selection of an upper limit of growth affects the correlation and causality of technology development measured with bibliometric data

Design/methodology/approach
We use Gompertz and Fisher-Pry curves to model the technological development of white light emitting diodes and flash memory, and show with extrapolation results from several bibliometric sources how a typical bias is caused in trend extrapolations.

Findings
We show how drastic an effect the decision to set an upper bound has on trend extrapolations, to be used as a reference for applications. We recommend carefully to examining the interconnection of actual development and bibliometric activity.

Originality/value
We are motivated by the fact that despite increasing interest in modelling technological data using this method, reports rarely discuss basic assumptions and their effects on outcomes. Since trend extrapolations are applied more widely in different disciplines, the basic limitations of methods should be explicitly expressed.

KEYWORDS: *forecasting, trend extrapolation, growth, Fisher-Pry, Gompertz, case, white light emitting diodes, flash memory*

1. Introduction

Bibliometric trending and trend extrapolations have been widely adopted in technology forecasting within the last two decades. Empirical evidence gathered from tangible technological advancements, modeled with significant accuracy[1] by using extrapolations, has made extending the approach to bibliometric data attractive. Due to the simplicity of data gathering and validated methodological options, bibliometric trending has become an approachable method. However, discussion of the underlying assumption that trends based on bibliometric data could model actual development is virtually absent from the growing number of bibliometric studies.

Motivated by the growing number of bibliometric studies, our aim is to develop an understanding of the correlation and causality of bibliometric data and actual, measurable, technical development. We focus on the following research question: *How do Technological Life Cycle (TLC) indicators compare with technical advancements*? The theoretical background of the study focuses on technology forecasting, trend extrapolations, and bibliometrics. Technology forecasting, as defined by Ayres (1969), answers (1) where is it possible to go from where we are now, (2) where we intend to go, and (3) where we expect to go. These questions have been

[1] Refer to work from Ayres (1969) Martino (1993) and Porter et al. (2011) for application examples.

approached using different foresight methods. Defining the methodical options available in forecasting, Popper (2008) conceptualized the Foresight Diamond. Taking advantage of earlier work by Cameron et al. (1996), who defined the triangular structure of methodologies, Popper (2008) defined a diamond with four types of knowledge sources: expertise, interaction, evidence, and creativity. These are attained using methods defined as quantitative, semi-quantitative, or qualitative. In this study, we focus on quantitative methods in which trend extrapolations, referred to as the "workhorse of technological forecasting," are perhaps the methods most frequently used (Lenz and Lanford Jr, 1973). Based on an analysis of time series data with selected parameters, extrapolations are used to forecast a development trend in the future. In modeling the complex socioeconomic system of technological development, these extrapolations are often based on different S-shaped growth curve models. These growth models, such as the Fisher-Pry (1971) and Gompertz (1825) models, have been validated by an abundance of empirical case studies using actual development data.

In contrast to using actual development data, bibliometrics takes advantage of the quantifiable information within databases, such as the number of articles in science databases that are directed toward a specific topic and uses this information as the basis for evaluating technological development. Evaluations are made as to the extent of the current state as trends or to the future as extrapolations.

Current bibliometric approaches, however, make the underlying assumption that growth in a quantified bibliometric time series would coincide with actual technological development; they also assume that different stages of TLC would be visible within the bibliometric series. However, the empirical evaluations made to validate the growth models have been based on modeling tangible developments (Ayres, 1969; Roper et al., 2011; Martino, 1993), such as Moore's Law. With the growth of accessible databases, the possibility of extending the same approach to bibliometric information and using this to model technological development has been suggested (Daim et al., 2006; Bengisu and Nekhili, 2006). The quantified information is then also used as a basis for trend extrapolations in the future. These approaches accept the underlying assumption that the bibliometric data models for technological development is a more concrete basis for evaluation than the assumption based on using the lumen/watt efficiency of white LEDs to model advancements in LED technology.

In this paper, we demonstrate through the findings of a study of two technologies---white light emitting diodes and flash memory---the potential problems in bibliometrical technology forecasting. The study takes advantage of bibliometric data made available by different databases as measures of technological advancements. The technologies are modeled through their current TLC by using databases, such as the Science Citation Index, Compendex, and the US Patent and Trademark Organization and News Services. The bibliometric series is measured against actual development data found in scientific or professional literature. We conclude the paper with guidelines for researchers to avoid potential problems in further studies.

2. Theoretical Background

Technology forecasting focuses on providing timely information about the prospects of a technology (Watts and Porter, 1997). Forecasting can be performed using different methods, some of which are qualitative while others rely on quantifying information embedded in databases. The latter often refers to analyzing textual databases with quantitative methods, which is referred to as bibliometrics (Borgman and Furner, 2002). Bibliometric methods are tools that extract information from large databases, uncovering the underlying structure of the databases and producing information from the apparently unstructured dataset (Daim et al., 2006). The data gathered can then be used to model the current state of a technology (Chao, Yang and Jen, 2007; Woon, Zeineldin and Madnick, 2011; Motoyama and Eisler, 2011) or to serve as the basis for extrapolations for future development (Marinakis, 2011).

Bibliometrics is defined as a method of analyzing textual databases with quantitative methods (Borgman and Furner, 2002). Closely related to scientometrics, informetrics, and technometrics, the aforementioned definition

provides a wide scope for understanding bibliometrics. A more narrow focus would be the definition given by Broadus (1987), for example, who defines bibliometrics as "the quantitative study of physical published units, or of bibliographic units, or of surrogates of either." When we then define scientometrics as the "quantitative study of science and technology" (van Raan, 1998) and informetrics as "the study of quantitative aspects of information in any form," we would understand informetrics as the overall term used when studying a broader set of data. Technometrics, on the other hand, focuses on mathematical modeling of technological advancements (e.g., (Sahal, 1984)), which differs from the aforementioned definition in that it applies a more tangible dataset as the basis for modeling. The terminological confusion related to different metrics is, to some extent, indifferent because all of the metrics focus on modeling databases with quantitative methods, as suggested by Borgman and Furner (2002). In this, the term bibliometrics is most often used (Hood and Wilson, 2001).

The significant increase in different quantitative approaches, whether we call them bibliometric, technometric, or informetric, is driven by the availability of different databases. However, while more and more databases and electronically accessible sources are available for bibliometric analysis, most technology-oriented analyses use only one database as a source (Kostoff et al., 2001; Chao, Yang and Jen, 2007; Kajikawa et al., 2008). In bibliometrics studies, the Science Citation Index (SCI) or Compendex indexes are commonly used sources. The Science Citation Index is considered one of the best sources for bibliometrical publication data (Kostoff, Koytcheff and Lau, 2007; Kajikawa, Takeda and Matsushima, 2010) and is often used as a database of information about emerging technologies (Kostoff et al., 2005).

Watts and Porter (Watts and Porter, 1997) introduced the concept of TLC indicators that attempt to understand how different kinds of databases mine information about technological innovations. The indicators represent each stage of R&D (Table 1). The TLC indicators---and technological forecasting overall---rely on a degree of orderliness or linearity in the innovation process (Watts and Porter, 1997).

Table 1: Stages of technology growth and sources of TLC data (Martino, 2003).

The linear model has been widely criticized, and this criticism has been targeted mostly on an overly simplified modeling of nature. (Balconi, Brusoni and Orsenigo, 2009) Several competing models have been developed (e.g. the chain-linked model (Kline and Rosenberg, 1986), multi-channel interactive learning model (Lundvall and Johnson, 1994)) and evolutionary model (Basalla, 1988). The evolutionary perspective has been further elaborated, for instance, by Arthur (2009) who explained that all technologies are passed down from earlier ones, and best performing and more efficient than others will be selected for future development. Further, Arthur argues that novel technologies arise by combining existing technologies. However, the linear view defends its position because of its intuitive simplicity and ease to use (Balconi, Brusoni and Orsenigo, 2009) (Godin, 2006). The linear model may well survive and be useful in science-based industries for example, or complement "broader, more general theories which recognize more clearly the dynamic interactive nature of the innovative process" (Balconi, Brusoni and Orsenigo, 2009). We employ linear model view in this paper to examine specifically how to avoid some of pitfalls with the linear modeling.

Forecasts---or trend extrapolations as the forecasts in this context are often called---are often done by using an S-shaped growth curve model. S-shaped growth curves fit well in modeling technological growth processes (Roper et al., 2011; Martino, 2003), although other forecast models, such as the ARIMA (Christodoulos, Michalakelis and Varoutas, 2010) and Richards model (Marinakis, 2011) have also been suggested. The application of trend extrapolations to quantitative information embedded in databases could be argued to be an extension of their use in modeling concrete technological development. Trend extrapolations have been based on the notion that "a specific technical approach to solving a problem will be limited by a maximum level of performance that cannot be exceeded" (Martino, 1993). Trend extrapolation results from modeling the S-shaped growth of a technical approach to a specific maximum level. The availability of information in databases has expanded the use of trend extrapolation to model the quantitative number of database entries. This would embed the underlying assumption

that when a specific maximum number of database entries is reached this would coincide with the "maximum level of performance."

The process of trend extrapolation involves fitting a chosen growth curve to a dataset and is seen as modeling the technological development. This fitted model is then extrapolated into the future. This process, in most cases, includes the acceptance of several assumptions:

1) "The upper bound of growth is known,"
2) "The chosen growth curve to be fitted to the historical data is the correct one," and
3) "The historical data gives the coefficients of the chosen growth curve formula correctly" (Martino, 1993).

To analyze the first assumption, we should note that using historical data as the only source for setting the upper bound of growth is considered bad practice (Martino, 1993). Although we often see the "goodness of fit" of historical data presented as a significant variable in making trend extrapolations (Huang, Guo and Porter, 2010), using historical data poses several challenges when making trend extrapolations. When using bibliometric quantities as the historical data on which the trend extrapolations are based, the researcher lacks a practical point of reference for the analysis. This demands that the database captures the development of a technology by emphasizing the use of several sources of information to validate the results. Alternatively, the database needs to be able to anchor the development trend to some other practical point of reference. However, recent studies have only used one database as a source (Chao, Yang and Jen, 2007; Kajikawa et al., 2008; Kajikawa, Takeda and Matsushima, 2010).

The second assumption focuses on the growth curve model used. Scholars are seen to use two distinct S-shaped growth models, the Fisher-Pry model and the Gompertz model, to forecast growth. In addition to the previously mentioned models, several other growth models have also been suggested and analyzed (Young, 1993). Both growth models produce an S-shaped growth curve, which, in addition to technological development, model several natural phenomena. These growth curves have a relatively slow early growth period, followed by a steep growth period that then turns into a saturation period in which the growth approaches the limit set. However, the Fisher-Pry and Gompertz models used in this study describe technological development quite differently.

The Fisher-Pry model, named after its originators Fisher and Pry, was described by its authors as "a substitution model of technological change." Fisher and Pry (1971) explained that the model would, for example, be powerful in forecasting technological opportunities. The Fisher-Pry model depends on the fraction of the technology penetration as well as on the fraction still being penetrated. This is loosely analogous to a situation in which the initial sales of a product will make subsequent sales easier by familiarizing prospective customers with the product. In contrast, the Gompertz model is most applicable in situations where "equipment replacement is driven by equipment deteriorations rather than technological innovation" (Roper et al., 2011). Sometimes referred to as the mortality rate, the Gompertz model fits a situation in which increased activity does not affect the future. This is analogous to a situation where "initial sales do not make subsequent sales easier" (Roper et al., 2011).

The underlying assumption made in both models is, however, that the dynamics of the developing technology would fit that of a growth curve. In this, we could argue that the "goodness of fit" would be insufficient for analyzing whether the growth curve fits the dynamics of technological development. Assumptions, especially when using short periods of historical data, should be based on empirical evidence for similar developments.

The third assumption focuses on making the statistical fit of the actual data available. It is commonly approached with a least squares fit. In the case of S-shaped growth, a transformation to a linear form is often used. After a linear regression, the least squares approach is used to fit the transformed values, the statistical fit can be

evaluated with the linear regression fit. In most cases, the fit between the actual values and the fitted values should also be evaluated.

3. Methodology and Data

In response to our objective, we employ the following research strategy here. First, we have already described three assumptions that are often made in bibliometric studies. In this chapter, we describe and justify why we have selected our two example technologies (white LED, flash memory). Next, we briefly describe extrapolation methods and their use, focusing on Fisher-Pry and Gomperz curves. This is followed by representing the quantitative data of the selected technologies. Throughout the illustrative case examples, we attempt to demonstrate the potential outcomes of implicit general assumptions in bibliometric studies.

3.1. Case technologies

The research question has been approached by selecting two different technologies: white light emitting diodes (LEDs) and flash memory. The technologies have been selected with hindsight, based on three criteria found after several attempts. First, development activities need to fit the availability of sources. Thus, only technologies developed since Science Citations Index (SCI) coverage began (1974) were possible. Second, the vocabulary of technologies, especially high-level naming, must be quite non-ambiguous since we used bibliometric methods. Third, the scope of the technology should represent that particular technology, not only the given "marketing name" (e.g., Bluetooth). A short description of each technology follows.

White LEDs

LED technology is a practical application of semiconductor technology that has been advantageously used for several decades. As an electronic component, LEDs have been available since the 1960s but have been restricted to wavelengths that enabled only small indicator lights. The first LED presented in 1962 (Holonyak and Bevacqua, 1962) had the luminous efficiency of 0.1 lm/W. More recent developments have led to white LEDs, which have a greater luminous efficiency, enabling LEDs to be used for lighting.

LED is a semiconductor diode which, through a process of electrons recombining with holes, releases energy as photons. An LED consists of a structure called a p-n junction. Electrons are injected in the p-type region of the junction while holes are injected in the n-type area. The recombination process at the junction leads to the emission of light. The wavelength, or, in practical terms, the color of the light, is determined by the band gap of the semiconductor, which is determined by the materials used. Although several materials have been used, for high-powered LEDs to become efficient and reliable, suitable semiconductor materials had to be fabricated.

While working toward the widespread use of LED technology, early increases in the efficiency of LEDs can be credited to the development of semiconductor technology. The practicality of the invention---it had already been used as an indicator during the late '60s---and the rapid developments in semiconductor technology resulted in a near order of magnitude development in the lm/W efficiency of LEDs (Craford, 1997). However, this resulted only in red, yellow and green LEDs becoming more efficient. Materials enabling efficient white light still depended on the development of a blue LED.

White light has been produced by either combining red, green, and blue LEDs or by using phosphorous material to convert blue or UV LED to a white light–emitting one (Yam and Hassan, 2005). The technological breakthrough produced by Nakamura, which enabled a gallium nitride–based blue and green LED (Nakamura, Mukai and Senoh, 1991; Nakamura, Senoh and Mukai, 1993), had a significant effect on the development of white LEDs. The invention propelled the development of white LEDs and has to date enabled LEDs to replace traditional lighting systems. Nakamura's invention enabled the development of practical white LEDs.

Because of this development cycle, LED technology has advantages when used as a light source. LEDs are highly efficient, reliable and rugged light sources. Although LEDs as such have been used for decades, Nakamura's invention enabled the further development of efficient and practical white LEDs. The rapid development of LED efficiency is often referred to as Moore's Law, which is similar to Haitz's Law, which forecasted an exponential rate of development in lumen/watt efficiency of LEDs with doubling occurring every 36 months (Anonymous, 2007). Since 2010, LEDs have matured into the general lighting market, and as its technology develops, it is taking an ever larger share of the lighting market.

Flash memory

The invention of flash memory was a continuum in the development of memory cells. The need for memory in different solutions ranging from personal computers to portable devices has increased the need for different memory cells. Semiconductor memory cells can be easily divided into two main categories: volatile and non-volatile memory. Volatile memory, such as SRAM and DRAM, enables fast reading and writing but loses its data when the power supply is turned off. Non-volatile memory, such as flash, on the other hand, can sustain data even without a power supply. Non-volatile memory has several applications mainly due to this characteristic.

Semiconductor-based memory technologies have taken up a significant portion of the whole semiconductor market. In particular, the explosive growth of the flash memory market has been a significant change. Flash memory demand has been driven by portable electronic devices, which use flash because it offers the best compromise between size and flexibility. Flash memory is used for two major applications: code storage and data storage. As the need for these applications increases, the demand for flash memory also increases (Bez et al., 2003).

Flash technology is based on a floating-gate transistor memory cell. Masuoka et al. presented the structure of flash memory in a 1984 paper in which they argued flash memory could overcome the problems of conventional EPROM memory cells and thus be more reliable. After its invention, in the early 1980s, the first papers expecting a "market burst" were presented in 1988 (Lineback, 1988; Cole, 1988). However, due to the early reliability problems with flash memory, the technology had a relatively low market penetration (Pavan et al., 1997). It was suggested in the mid-1990s that flash memory would have a market share of six percent by 2000, as dynamic random access memory would dominate the market. Since then, two dominant flash architectures have emerged: NOR flash designed for code and data storage and NAND flash for data storage (Bez et al., 2003).

The benefits of NOR and NAND flash memory have increased the market size of the technology. Flash memory, for example, is currently used in solid-state disks which take advantage of its small dimensions, low power consumption, and lack of mobile parts. With an increase in the number of applications taking advantage of the benefits of flash memory, flash memory's market share has increased. Since 2000, flash memory technology has been seen as a mature technology, which has increased rapidly in market size (Bez et al., 2003).

3.2. Trend extrapolation and limit curves

The trend extrapolation method often relies on basic time series analysis. Using regression techniques to fit nonlinear relationships is seen as suitable for technological forecasting. The use of the methods is derived from the historical understanding that a specific nonlinear model would describe the complex system of technological development. This has been the case with limit curve models such as Fisher-Pry and Gompertz, which have been validated by the vast number empirical studies in which they have been used.

To effectively model these non-linear relationships, we tend to use the data as a linear function of time. This requires the transformation seen in Table 2 for the Fisher-Pry and Gompertz curves. In both transformations, L, the upper limit of growth, affects the model fit. By selecting an appropriate upper level of growth, we can use linear regression in estimating the values of the constants *a* (intercept) and *b* (slope) in the linear model equation:

$$Y = a + bX + e.$$

The model validity is often evaluated statistically by selecting constants "*a*" and "*b*," that minimizes the sum of squares errors "e" between the value of Y and the value predicted by the linear model. This straightforward statistical analysis rests heavily on the assumptions that (1) the upper bound of growth is known, and (2) the environment of the past will continue in the future. In this type of modeling, the researcher is forced to assume the development is a static process without discontinuities and thus affects the model only through the selection of an upper bound of growth.

Table 2 Linear transformation with Fisher-Pry and Gompertz models adopted **(Roper et al., 2011)**.

In addition to using the least squares approach, the fitted values are evaluated by using Mean Absolute Percentage Error (MAPE) to analyze model fit (Young, 1993). By setting the upper bound of growth to minimize the MAPE, a statistical evaluation of the overall model fit is analyzed.

3.3. Quantitative data of the case

The databases used for this study were selected according to the Stages of Technology Growth and Sources of TLC data presented in Table 1. Thus, the SCI was selected to represent fundamental research, Compendex was used to represent applied research, patents from the US Patent and Trademark Office were used to represent development, and Newspaper Abstracts Daily were used to represent application. In the context of this study, the social impact of the technology has been omitted.

Table 3 shows the summary of results on the cumulative document frequency of white LEDs. Table 4 shows the summary of results on flash memory.

Table 3: Cumulative document frequency of white LEDs and the actual development of LED efficiency.

The databases were analyzed by using "white led," "white leds," "white light emitting diode" and "white light emitting diodes" as search algorithm for LEDs. Similarly, "flash memory" or "flash memories" were used for flash memories. These were, through a process of trial and error, seen to find the relevant database entries. The first entries found in each database were further checked by an expert to make sure that the starting point of each dataset was set correctly.

Table 4: Cumulative document frequency of flash memory and the actual development of NAND flash memory cell size.

The bibliometric data in Table 3 and Table 4 is given as a cumulative document frequency. These cumulative data series are then used for the trend extrapolation. It should be pointed out that the data series, excluding patent data series for both technologies, have a growing number of documents appearing yearly. The number of patent documents has, however, peaked for both technologies and is now decreasing. The tables also show the Pearson correlation of each of the bibliometric data series compared to the reported actual development. This shows that the bibliometric data series are significantly correlated with the actual technological development. This shows

that there is a correlation between each of the bibliometric data series and actual development. In the case of flash memory the correlation is naturally negative as the technology gets smaller.

4. Results

Figure 1 summarizes the search results. This shows an early increase in applied research in Compendex, which is not supported by the theory of linear development (Järvenpää, Mäkinen and Seppänen, 2011).

Figure 1: Non-cumulative summary of Table 3 and Table 4, excluding actual development.

The historical data was transformed into linear form using the equations in Table 2. Then the upper bound of growth that would result in the highest fraction of the total variance of the dependent variable explained by the model was selected. This is described by the coefficient determination, R^2. In Table 5, the linear regression of the indicators is shown.

Table 5: Upper bound and model fit based on R^2 values.

However, the indicator development forecasted by the highest R^2 does not seem plausible. By relying on the analysis, the described TLC of LEDs does not seem practical. In the graphical representations given in Figure 2 and Figure 3, LED development is described in normalized form throughout the TLC. The Fisher-Pry model suggests that basic research is lagging overall LED development by several years and that the first indicators, "development" and "application," would reach the upper bound of growth within a few years. This forecast seems implausible based on the order of development or by the upper bound of growth.

Figure 2: Summary of Fisher-Pry model fit from Table 5.

In comparison, in the Gompertz model, the least squares model fit resulted in implausible upper bounds of growth for basic and applied research seen in Table 5, while retaining a similar development path for the two following indicators. Again, these do not seem practical either by the order of development or by the upper bound of growth.

Figure 3: Summary of Gompertz model fit from Table 5.

The data was further analyzed with MAPE between the fitted value and the historical data as described by Young (1993). The upper bound of growth resulting in the smallest MAPE was selected for each R&D stage. This was done while accepting the lower R^2 value of for the models, but by minimizing the MAPE for each dataset. Table 6 shows the upper bounds of growth resulting from the analysis as well as the MAPE values for the models.

Table 6: Upper bound and model fit based on MAPE.

The results of the MAPE fitted forecast were extrapolated to the future and can be seen in normalized Figure 4 and Figure 5.

Figure 4: Summary of Fisher-Pry model fit from Table 6.

Figure 5: Summary of Gompertz model fit from Table 6.

Table 7 summarizes how upper bounds vary depending on the statistical fit of the model and the model used. As we see with Fisher-Pry the values, the upper bounds are within the same range, but with Gompertz models, the

values of the upper bounds vary significantly. This underlines the effect of setting an upper bound in the trend extrapolations.

Table 7: Summary of different upper bounds in the above runs.

The upper limit of the growth curves were also evaluated against the values of R^2. Testing several values of the upper limit growth against a constrained value of R^2, we found that with several of the curves the result was out of bounds of a practical number of iterations. By expecting R^2 value of over 0,90, the upper limit of growth can range from the largest known (actual) value to a value several times the optimal value fitted by R^2 or MAPE. This should be of particular interest when setting the upper bound of growth based on expert opinion and then using curve estimation to evaluate the result.

5. Discussion

Motivated by the paucity of explicitly expressed assumptions in bibliometric studies, we have presented two examples of technological development and demonstrated the effects of assumption on trend extrapolation outcomes. The results above seem to be promising on the surface. Looking at the statistical analysis, we found that the trend extrapolations modeled the developments with a good statistical fit. The variance in the data was explained---in most cases even with an R^2 value of 0.9. There are, of course, inherent differences with the results produced by different growth curves, such as the Fisher-Pry compared to Gompertz, but these were known prior to the analysis. For example, in the case studies used we expected that compared to Gompertz, Fisher-Pry would produce a stronger upward trend resulting in an earlier saturation point.

However, looking below the surface, the studies emphasize several key factors affecting the practical analysis in the trend extrapolation. First, the statistically fitted model lacks the ability to produce a practically plausible model fit throughout the TLC. Second, the possibility of varying the upper bound of growth while still having a statistically good model was large. Third, the ability of different data sources to model stages or model a phase of development is questionable. These factors lead to several considerations and limitations, which are suggested as guidelines for trend extrapolation studies.

First, considering the data used for the studies, researchers should consider if the sample is a representative sample. We often assume that the data sources used in bibliometric studies are representative of the whole population. We do not sufficiently question what biases are made by the nature of the data source that are independent of the actual development. For example, we might argue if the USPTO, most often used in bibliometric studies, is a representative sample for patent development---and turn a blind eye to the Chinese Patent Office database or European Patent Office that would lead to a more representative sample. Researchers should make the effort to validate the sample used or accept the limitations set by a biased sample.

Second, researchers should consider if the data is valid for analysis “as it is” or if there are some underlying factors influencing the explanatory power of the dataset. The data in a bibliometric trend extrapolation is a dataset measuring activity (1995). These are subject to variations by causal forces derived from several factors, such as growth, decay, supporting, and opposing forces (Armstrong and Collopy, 1993). It is evident that the impacts of causal forces should seriously be taken into consideration prior to the analysis. For example, if a technology is heavily supported by policy and therefore research funding, we might argue about the extent to which there might be supporting causal forces present in the data.

In addition to the causal forces, researchers are restricted by the period of the data used for the analysis. Curve fitting is often used ambitiously to demonstrate how novel technologies develop to maturity; however, there is severe danger that the result might be very different depending on single data points. Despite the fact that it is possible to model development with a fairly good statistical accuracy using only a few data points (for instance

with the Fisher-Pry), a broad set of acceptable end results are produced. Thus, keeping in mind the rule of thumb---not to extend a forecast more into the future than the same period of historical data might be able to prove---should also be a valuable strategy for trend extrapolations. With short historical datasets, this might lead to a situation in which a different methodological approach, such as a moving average or an autoregressive integrated moving average, might be worth trying.

Third, the result of a bibliometric extrapolation needs to be connected with actual development. By approaching technology forecasting with bibliometric trend extrapolation, the most significant assumption made by researchers is that the bibliometric data actually models technological development. Thus, an evaluation of the extent to which this holds true to the case at hand needs to be done. This should eventually lead to the use of expert opinion or rigorous analysis of the correlation between actual development data and bibliometric data.

The final notion appears in the flash memory data (Table 4) which show that the development of Flash memory has been visible three years before the first patent document materialized. This notion contradicts the traditional belief that the interest of mass media is the last phase of TLC indicators, and for some technologies, it would be possible to detect the development through news pieces earlier than patent documents. Thus, further studies could investigate why some technologies may receive such interest from media, and in addition, if there are any shared factors among the technologies that may create identifiable patterns.

6. Conclusion

We have demonstrated above that using trend extrapolations with bibliometric datasets has challenges. Agreeing on a valid upper bound of growth is a well-known factor that affects the validity of the model. This is not news for statistically oriented researchers. However, many studies published do not seem to consider the effects of varying the upper bound. Thus, creating a practical context for the data is a significant research factor for validating trend extrapolation results. In addition, understanding the nature of the bibliometric data is relevant. Evaluating the limitations of the data used, such as limitations of data sources and causal forces, will help to prevent researchers from coming to wrong decisions. We recommend carefully examining the interconnection of actual development and bibliometric activity that is illustrated in Figure 6 below.

Figure 6 Figure 1 in the context of actual development, seen in Table 3 and Table 4. Percentages of bibliometric data calculated from cumulative data. Actual development for NAND Flash Memory Cell Size has been inverted.

In Figure 6 actual development data is included with the quantitative data extracted from the databases. Both of the technologies are either mature or rapidly maturing (See Section 3.1). Visual inspection suggests that the bibliometric data resembles the actual development to some extent. LED technical development seems to follow a slope similar to the bibliometric dataset. Naturally, technological development of flash memory (measured with cell size in μm) has a negative slope in comparison to the upward slope of bibliometric data. Similarities and differences are in the eye of beholder, thus, it is recommended that some correlation analyses for facts is made. In these cases, the Pearson correlation seen in Tables 3 and 4 suggests a clear correlation with bibliometric data series and the actual development. These fairly strong correlations do not serve as evidence of causality and successful extrapolation results.

Further studies need to gain an understanding of the limitations of the methodology used. Trend extrapolations can, by selecting the database, growth curve, and upper bound of growth, be made to produce significantly different results. This calls for broader sensitivity analysis on the limits of extrapolation results with varying bibliometric datasets. Thus, researchers should also make the assumptions clear to readers when extrapolating with bibliometric datasets.

Acknowledgements
To be added to the final version

References

Anonymous (2007) Haitz´s law, *Nat Photon*, vol. 1, Jan, pp. 23-23.
Armstrong, S.J. and Collopy, F. (1993) Causal forces: Structuring knowledge for time-series extrapolation, *Journal of Forecasting*, vol. 12, pp. 103-115.
Arthur, Brian (2009). The nature of technology: What it is and how it evolves. Penguin Books London, UK
Ayres, R.U. (1969) *Technological forecasting and long-range planning*, McGraw-Hill New York.
Balconi, M., Brusoni, S. and Orsenigo, L. (2009) In defence of the linear model: An essay, *Research Policy*, vol. 39, no. 1, pp. 1-13.
Basalla, George.(1988). The evolution of technology. Cambridge University Press. Cambridge, UK
Bengisu, M. and Nekhili, R. (2006) Forecasting emerging technologies with the aid of science and technology databases, *Technological Forecasting and Social Change*, vol. 73, no. 7, sep, pp. 835--844, Available: 0040-1625.
Bez, R., Camerlenghi, E., Modelli, A. and Visconti, A. (2003) Introduction to flash memory, *Proceedings of the IEEE*, vol. 91, pp. 489-502.
Borgman, C.L. and Furner, J. (2002) Scholarly communication and bibliometrics, *Annual Review of Information Science and Technology*, vol. 36, pp. 3-72.
Broadus, R.N. (1987) Toward a definition of "bibliometrics", *Scientometrics*, vol. 12, no. 5-6, pp. 373--379, Available: 0138-9130.
Cameron, H., Loveridge, D., Cabrera, J., Castanier, L., Presmanes, B. and Vasquez, L. (1996) *Technology foresight: perspectives for European and international co-operation*, Manchester: PREST: Mimeo.
Chao, C., Yang, J. and Jen, W. (2007) Determining technology trends and forecasts of RFID by a historical review and bibliometric analysis from 1991 to 2005, *Technovation*, vol. 27, no. 5, pp. 268-279, Available: 0166-4972.
Christodoulos, C., Michalakelis, C. and Varoutas, D. (2010) Forecasting with limited data: Combining ARIMA and diffusion models, *Technological Forecasting and Social Change*, vol. 77, no. 4, pp. 558-565.
Cole, B.C. (1988) Flash - theres more than one road to dense nonvolatile memory, *Electronics*, vol. 61, p. 108, Available: 0883-4989.
Craford, M.G. (1997) Overview of Device Issues in High-Brightness Light-Emitting Diodes, *Semiconductors and Semimetals*, vol. 48, pp. 47-63.

Daim, T.U., Rueda, G., Martin, H. and Gerdsri, P. (2006) Forecasting emerging technologies: Use of bibliometrics and patent analysis, *Technological Forecasting & Social Change*, vol. 73, no. 8, pp. 981-1012.
Fisher, J. and Pry, R. (1971) A simple substitution model of technological change, *Technological Forecasting and Social Change*, vol. 3, pp. 75--88, Available: 00401625.
Godin, B. (2006) The Linear Model of Innovation - The historical construction of an analytical framework, *Science, Technology & Human Values*, vol. 31, no. 6, pp. 639-667.
Gompertz, B. (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies, *Philosophical transactions of the Royal Society of London*, vol. 115, pp. 513--583.
Haitz, R. and Tsao, J.Y. (2011) Solid state lighting: The case 10 years after and future prospects, *physica status solidi (a)*, vol. 208, no. 1, jan, pp. 17-29, Available: 1862-6319.
Holonyak, N. and Bevacqua, S.F. (1962) Coherent (visible) light emission from GA(AS1-XPX) junctions, *Applied Physics Letters*, vol. 1, no. 4, p. 82.
Hood, W.W. and Wilson, C.S. (2001) The literature of bibliometrics, scientometrics, and informetrics, *Scientometrics*, vol. 52, no. 2, oct, pp. 291-314, Available: 0138-9130.
Huang, L., Guo, Y. and Porter, A.L. (2010) Identifying Emerging Roles of Nanoparticles in Biosensors, *Journal of Business Chemistry*, no. 1.
Järvenpää, H., Mäkinen, S. and Seppänen, M. (2011) Patent and publishing activity sequence over a technology´s life cycle, *Technological forecasting and Social Change*, vol. 78, no. 2, pp. 283-293.
Kajikawa, Y., Takeda, Y. and Matsushima, K. (2010) Computer-assisted roadmapping: a case study in energy research, *Foresight*, vol. 12, no. 2, pp. 4-15.
Kajikawa, Y., Yoshikawa, J., Takeda, Y. and Matsushima, K. (2008) Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy, *Technological Forecasting and Social Change*, vol. 75, no. 6, 0, pp. 771-782.
Kline, S.J. and Rosenberg, N. (1986) An overview of innovation, in Landau, R. and Rosenberg, N. *The Positive Sum Strategy: Harnessing Technology for Economic Growth*, The National Academy of Sciences.
Kostoff, R.N., Antonio del Rio, J., Cortes, H.D., Smith, C., Smith, A., Wagner, C., Leydesdorff, L., Karypis, G., Malpohl, G. and Tshiteya, R. (2005) The structure and infrastructure of Mexicos science and technology, *Technological Forecasting and Social Change*, vol. 72, no. 7, pp. 798-814.
Kostoff, R.N., Koytcheff, R.G. and Lau, C.G. (2007) Global nanotechnology research metrics, *Scientometrics*, vol. 70, mar, pp. 565--601, Available: 0138-9130, 1588-2861.
Kostoff, R.N., Toothman, D.R., Eberhart, H.J. and Humenik, J.A. (2001) Text mining using database tomography and bibliometrics: A review, *Technological Forecasting and Social Change*, vol. 68, no. 3, Nov, pp. 223-253.
Lenz, R.C. and Lanford Jr, H.W. (1973) Trend extrapolation: Workhorse of technological forecasting, *Industrial Marketing Management*, vol. 3, 0, pp. 57--65.
Lineback, J.R. (1988) High-density flash eeproms are about to burst on the memory market, *Electronics*, vol. 61, Mar, pp. 47-48, Available: 0883-4989.
Lundvall, B. and Johnson, B. (1994) The Learning Economy, *Journal of Industry Studies*, vol. 1, no. 2, pp. 23-42.
Marinakis, Y.D. (2011) Forecasting technology diffusion with the Richards model, *Technological Forecasting and Social Change*, vol. 79, no. 1, pp. 172-179, Available: 0040-1625.
Martino, J.P. (1993) *Technological Forecasting for Decision Making 3rd ed.*, McGraw-Hill Inc.
Martino, J.P. (2003) A review of selected recent advances in technological forecasting, *Technological Forecasting and Social Change*, vol. 70, no. 8, pp. 719-734.
Moed, H.F., Bruin, R.E. and Leeuwen, T.N. (1995) New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications, *Scientometrics*, vol. 33, no. 3, pp. 381-422, Available: 0138-9130, 1588-2861.
Motoyama, Y. and Eisler, M.N. (2011) Bibliometry and nanotechnology: A meta-analysis, *Technological Forecasting and Social Change*, vol. 78, no. 7, sep, pp. 1174-1182, Available: 0040-1625.
Nakamura, S., Mukai, T. and Senoh, M. (1991) High-power GaN pn junction blue-light-emitting diodes, *Japanese journal of applied physics*, vol. 30, pp. 1998-2001.
Nakamura, S., Senoh, M. and Mukai, T. (1993) High-power InGaN/GaN double-heterostructure violet light emitting diodes, *Applied physics letters*, vol. 62, no. 19, pp. 2390-2392.
Pavan, P., Bez, R., Olivo, P. and Zanoni, E. (1997) Flash memory cells - An overview, *Proceedings of the IEEE*, vol. 85, aug, pp. 1248-1271, Available: 0018-9219.
Popper, R. (2008) Foresight methodology, in Georghiou, L., Cassigena Harper, J., Keenan, M., Miles, I. and Popper, R. *The handbook of technology foresight*, Edwar Elgar Publishing.

Roper, A.T., Cunningham, S.W., Porter, A.L., Mason, T.W., Rossini, F. and Banks, J. (2011) *Forecasting and management of technology*, 2nd edition, New York: John-Wiley.
Sahal, D. (1984) The innovation dynamics and technology cycles in the computer industry, *Omega*, vol. 12, no. 2, pp. 153-163, Available: 0305-0483.
Shin, Y. (2005) Non-volatile memory technologies for beyond 2010, *VLSI Circuits - Digest of Technical Papers 2005* , pp. 156-159.
van Raan, A.F.J. (1998) Special topic issue: Science and technology indicators - Introduction, *Journal of the American Society for Information Science*, vol. 49, jan, pp. 5--6, Available: 0002-8231.
Watts, R.J. and Porter, A.L. (1997) Innovation forecasting, *Technological Forecasting and Social Change*, vol. 56, sep, pp. 25--47, Available: 0040-1625.
Woon, W.L., Zeineldin, H. and Madnick, S. (2011) Bibliometric analysis of distributed generation, *Technological Forecasting and Social Change*, vol. 78, no. 3, mar, pp. 408--420, Available: 0040-1625.
Yam, F.K. and Hassan, Z. (2005) Innovative advances in LED technology, *Microelectronics Journal*, vol. 36, no. 2, pp. 129-137.
Young, P. (1993) Technological growth curves: A competition of forecasting models, *Technological Forecasting and Social Change*, vol. 44, no. 4, pp. 375-389.

Table 1: Stages of Technology Growth and Sources of TLC Data (Martino, 2003).

Stages of technology growth	R&D stages	Typical sources of TLC data
Scientific Findings and Demonstration of Laboratory Feasibility	Basic Research	Science Citation Index
Operating full-scale prototype or field trial	Applied Research	Engineering Index
Commercial introduction and/or operational use	Development	Patent databases
Widespread adoption/Proliferation and diffusion to other uses	Application	Newspaper Abstracts
Societal effect and/or significant economical involvement	Social Impacts	Business and Popular press

Table 2 Linear Transformation of Fisher-Pry and Gompertz Models, adopted (Roper et al., 2011)

Growth Model	***Transformation***
Fisher-Pry	$Z = \ln[(L - Y) / Y]$
Gompertz	$Z= \ln[\ln(L / Y)]$

Table 3: Cumulative Document Frequency of White LEDs and the actual development of LED efficiency .

Year	*SCI*	*Compendex*	*Patents*	*News*	*lm/lamp approximate value*
1991		1			
1992		1			
1993		1			
1994		2			
1995		2			
1996	1	3			
1997	7	6	2	5	
1998	9	7	5	6	
1999	17	10	15	15	
2000	27	18	25	22	10
2001	39	38	37	41	
2002	60	60	44	86	60
2003	87	92	67	144	
2004	139	149	85	208	110
2005	205	229	108	285	
2006	293	316	121	341	800
2007	414	417	128	410	900
2008	587	570	130	473	
2009	823	764	130	565	3000

Actual development based on (Haitz and Tsao, 2011)
Pearson correlations between actual development and each bibliometric dataserie: SCI 0.981, Compendex 0.969, Patents 0.697 and News 0.887

Table 4: Cumulative Document Frequency of Flash Memory and the actual development of NAND Flash Memory Cell Size.

Year	*SCI*	*Compendex*	*Patents*	*News*	*NAND Flash Memory Cell Size (µm2)*
1988	2	4		1	
1989	9	9		3	
1990	9	21		10	
1991	11	28	1	27	
1992	20	32	6	108	
1993	36	53	23	151	
1994	65	104	55	230	
1995	100	190	99	283	
1996	135	273	200	407	1
1997	173	360	353	578	
1998	199	439	574	833	0,6
1999	248	519	840	1018	0,3
2000	294	619	1166	1304	
2001	353	771	1511	1572	0,15
2002	414	916	1914	1842	
2003	484	1130	2342	2176	0,07
2004	573	1453	2740	2509	0,04
2005	657	1855	3124	3015	0,02
2006	790	2273	3546	3671	0,015
2007	939	2743	3950	4407	
2008	1109	3207	4208	5104	
2009	1302	3692	4268	5785	

Actual development based on (Shin, 2005)
Pearson correlations between actual development and each bibliometric dataserie: SCI -0.693, Compendex -0.628, Patents -0.724 and News -0.694

Table 5: Upper Bound and Model Fit Based on R^2 values

Model Fit			*SCI*	*Compendex*	*Patent*	*News*
Fisher-Pry	White LED	(R^2)	0,975	0,987	0,987	0,993
		Upper Bound	1148	2878	133	632
	Flash	(R^2)	0,934	0,979	0,965	0,933
		Upper Bound	5786	4305	4330	6043
Gompertz	White LED	(R^2)	0,990	0,985	0,983	0,993
		Upper Bound	21144	$11*10^{14}$	154	1229
	Flash	(R^2)	0,991	0,994	0,999	0,991
		Upper Bound	5637	21373	6076	12466

Table 6: Upper Bound and Model Fit based on MAPE.

Model Fit			*SCI*	*Compendex*	*Patent*	*News*
Fisher-Pry	White LED	MAPE	24,35 %	20,59 %	10,49 %	40,73 %
		Upper Bound	1387	2294	139	592
	Flash	MAPE	21,14 %	24,57 %	39,17 %	54,42 %
		Upper Bound	1345	4904	4269	5786
Gompertz	White LED	MAPE	17,16 %	23,81 %	7,04 %	12,94 %
		Upper Bound	56621	$3*10^{11}$	160	1014
	Flash	MAPE	13,41 %	12,68 %	5,82 %	15,69 %
		Upper Bound	2756	17468	6143	8162

Table 7: Summary of Different Upper Bounds in the above runs.

Model Fit			*SCI*	*Compendex*	*Patent*	*News*
Fisher-Pry	White LED	R^2	1148	2878	133	632
		MAPE	1387	2294	139	592
	Flash	R^2	5786	4305	4330	6043
		MAPE	1345	4904	4269	5786
Gompertz	White LED	R^2	21144	$11*10^{14}$	154	1229
		MAPE	56621	$3*10^{11}$	160	1014
	Flash	R^2	5637	21373	6076	12466
		MAPE	2756	17468	6143	8162

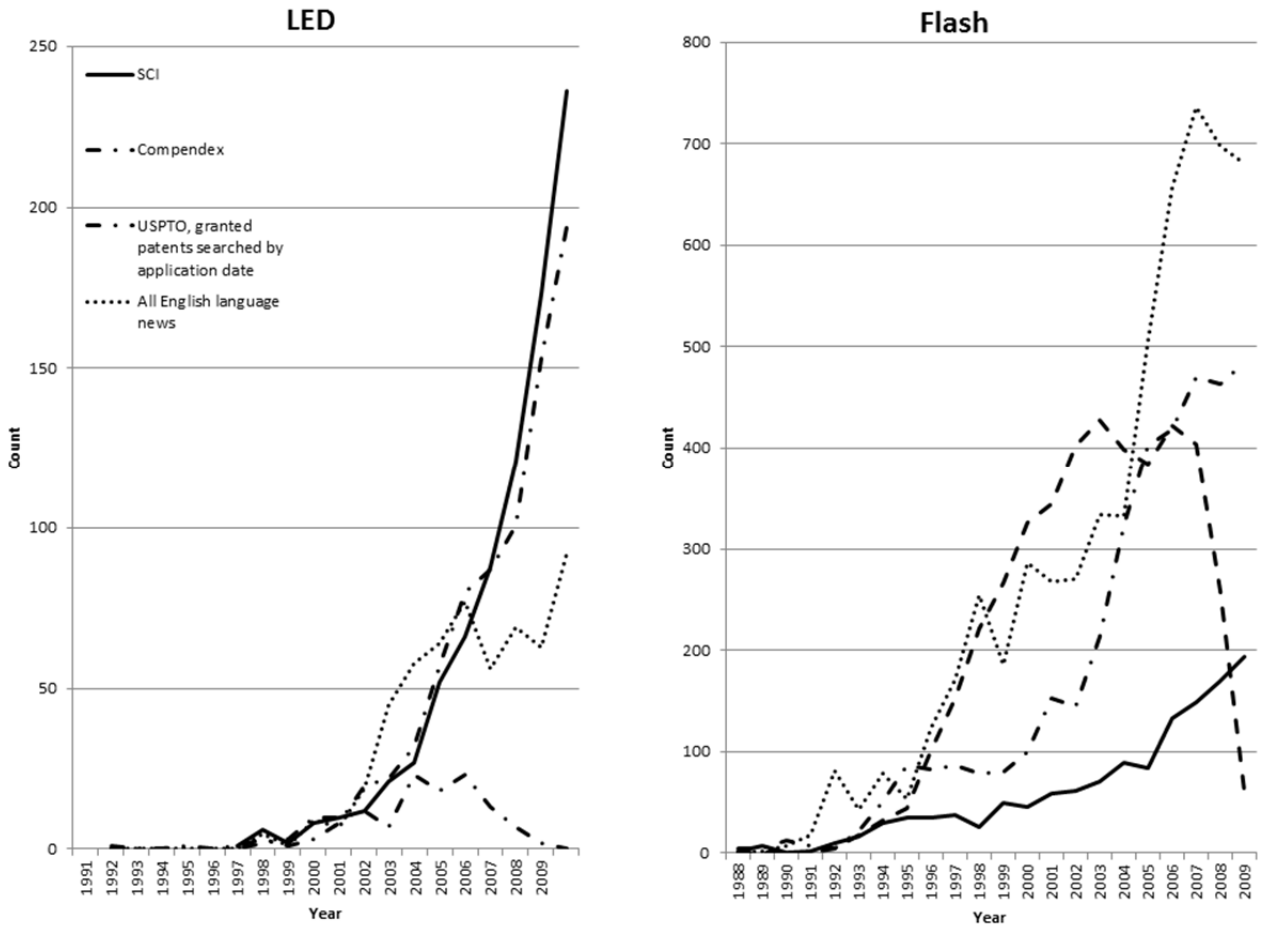

Figure 7: Non-cumulative summary of Table 3 and Table 4, excluding actual development.

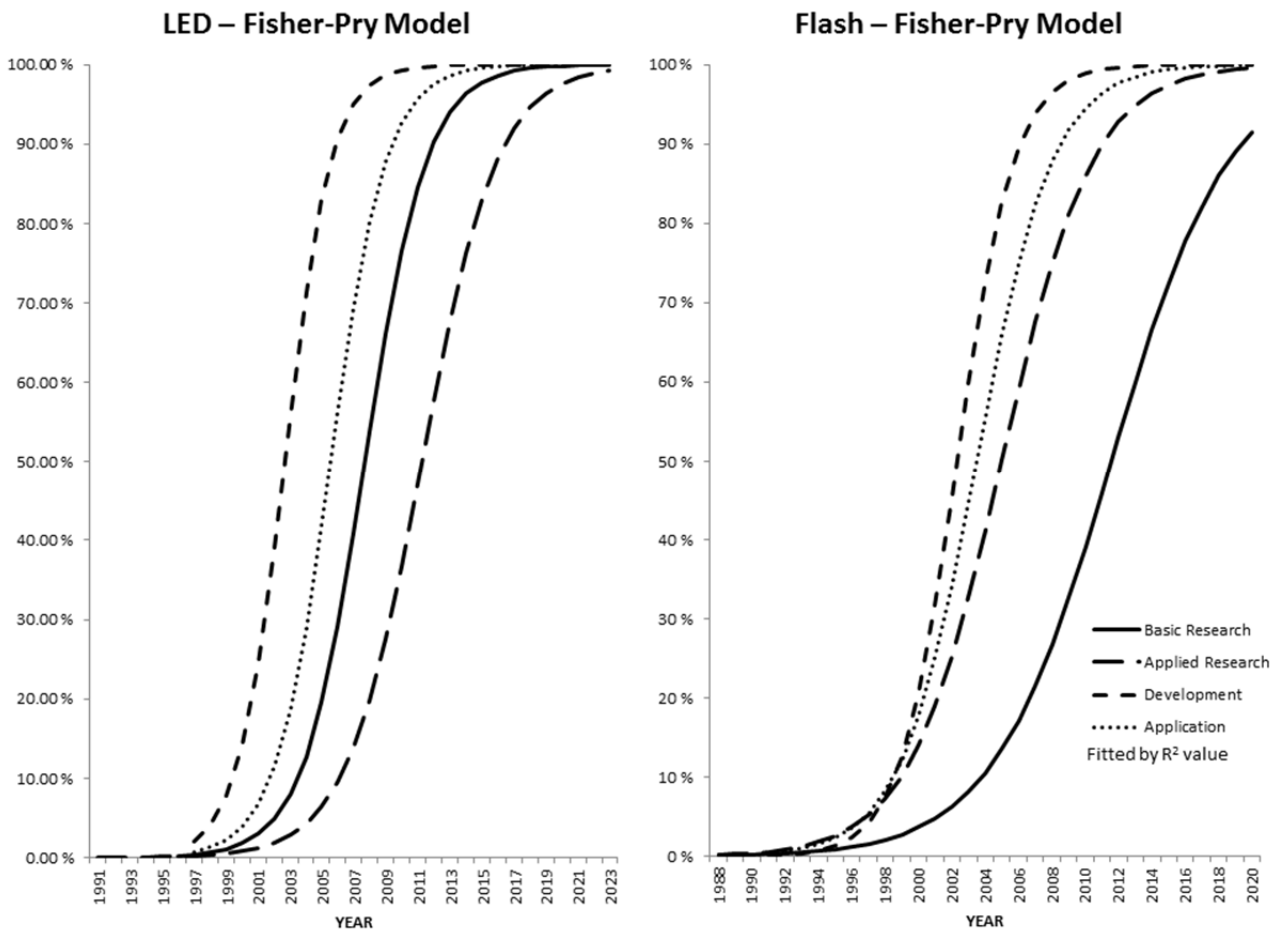

Figure 8: Summary of Fisher-Pry model fit from Table 5.

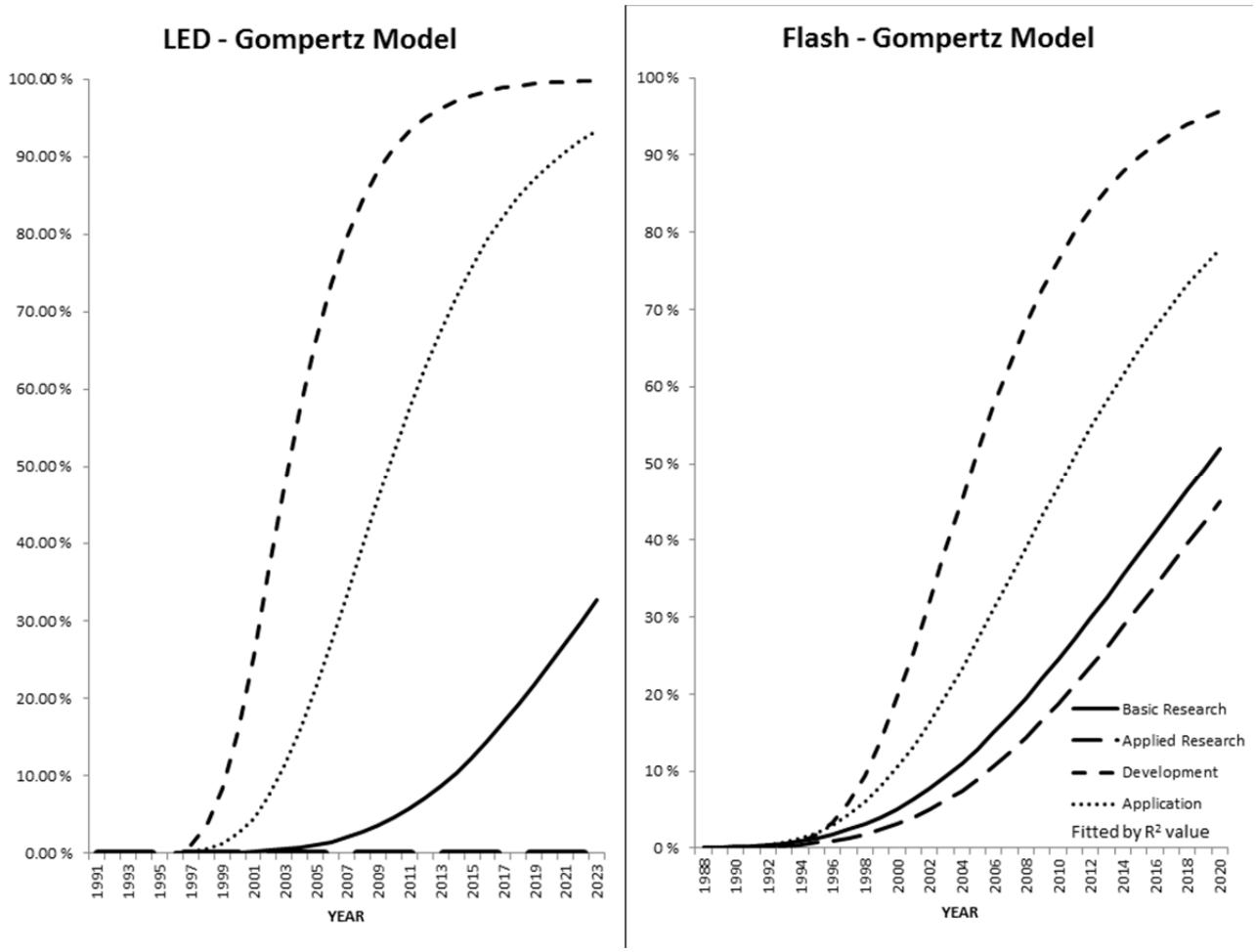

Figure 9: Summary of Gompertz model fit from Table 5.

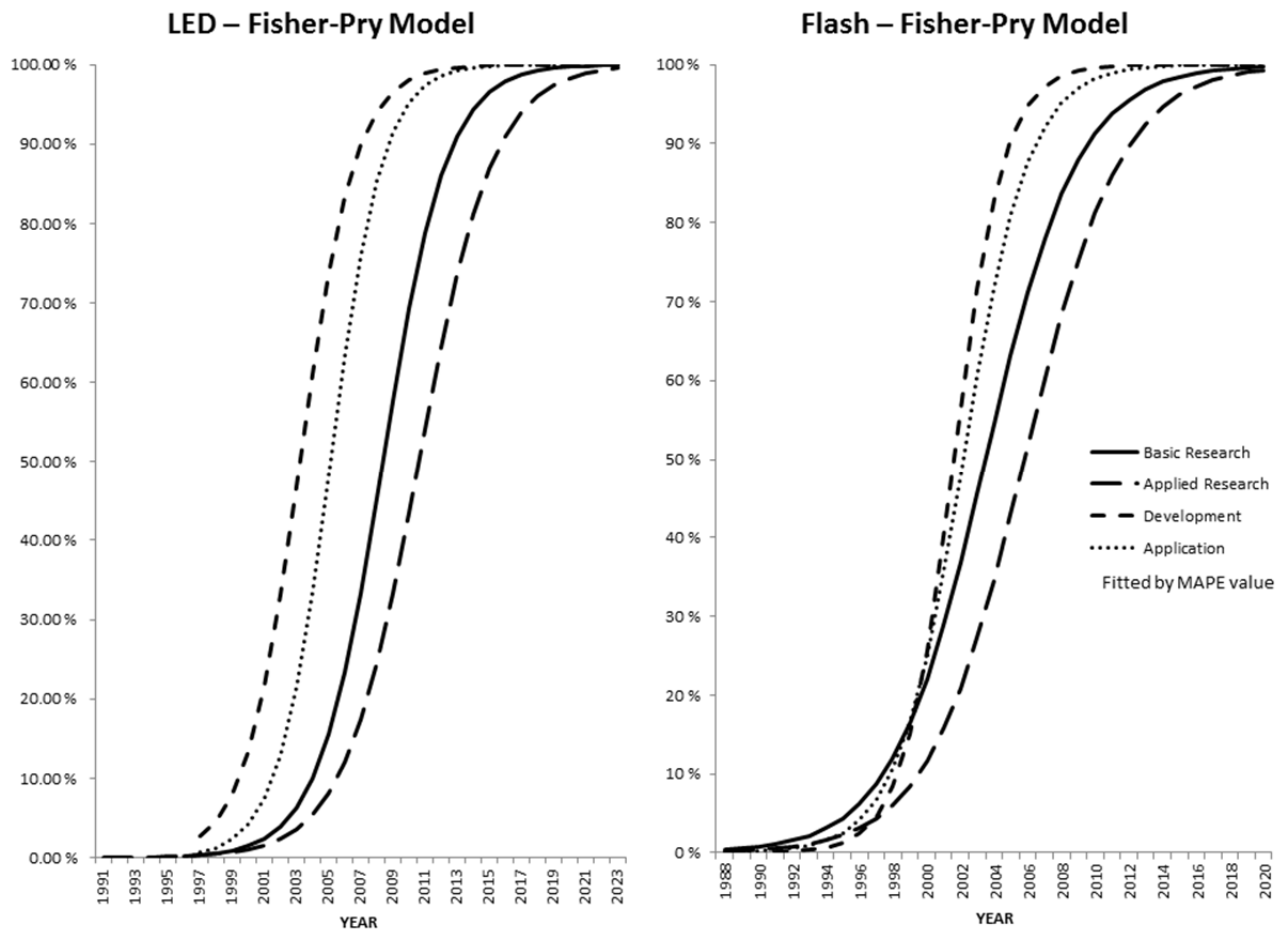

Figure 10: Summary of Fisher-Pry model fit from Table 6.

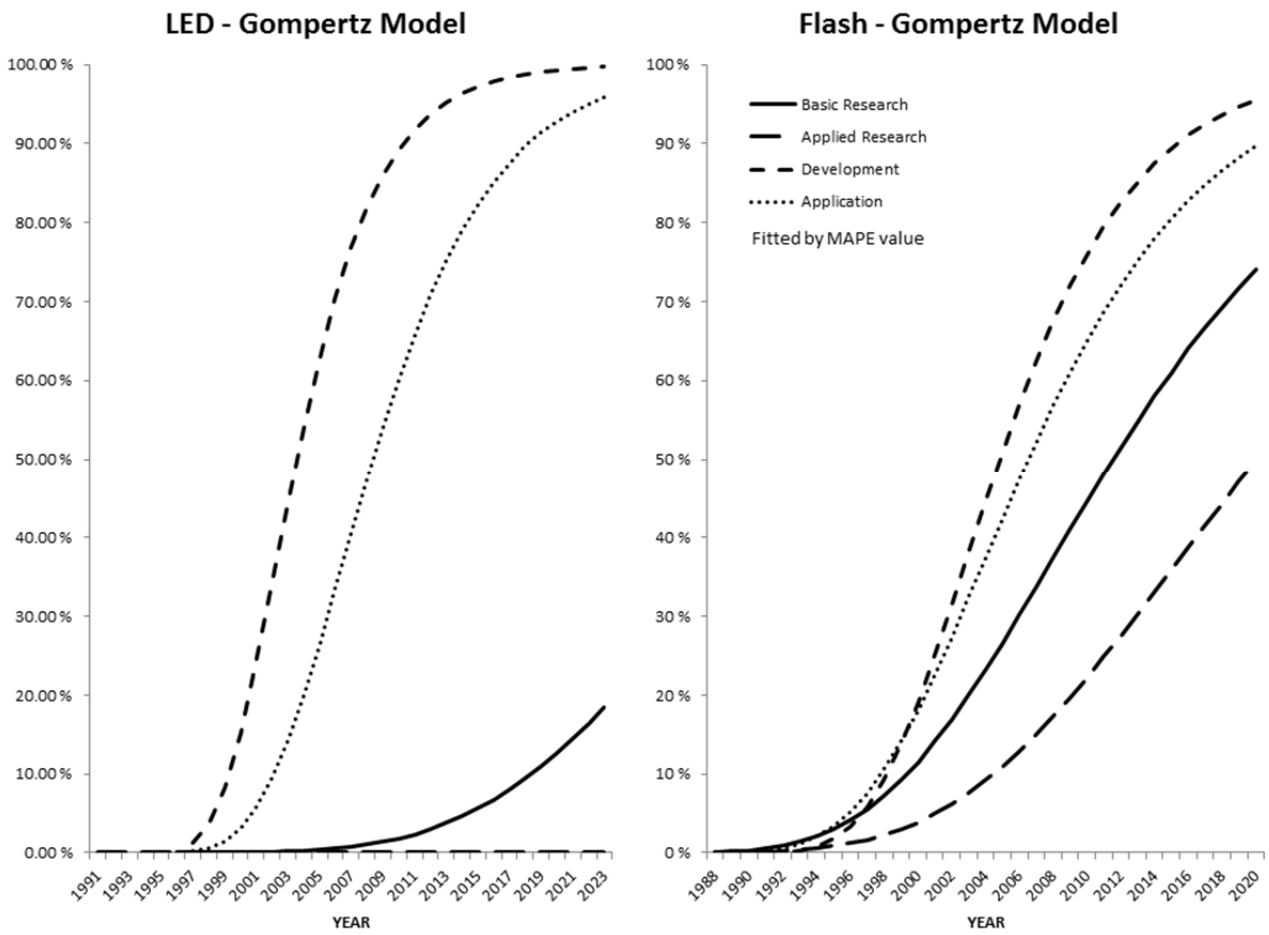

Figure 11: Summary of Gompertz model fit from Table 6.

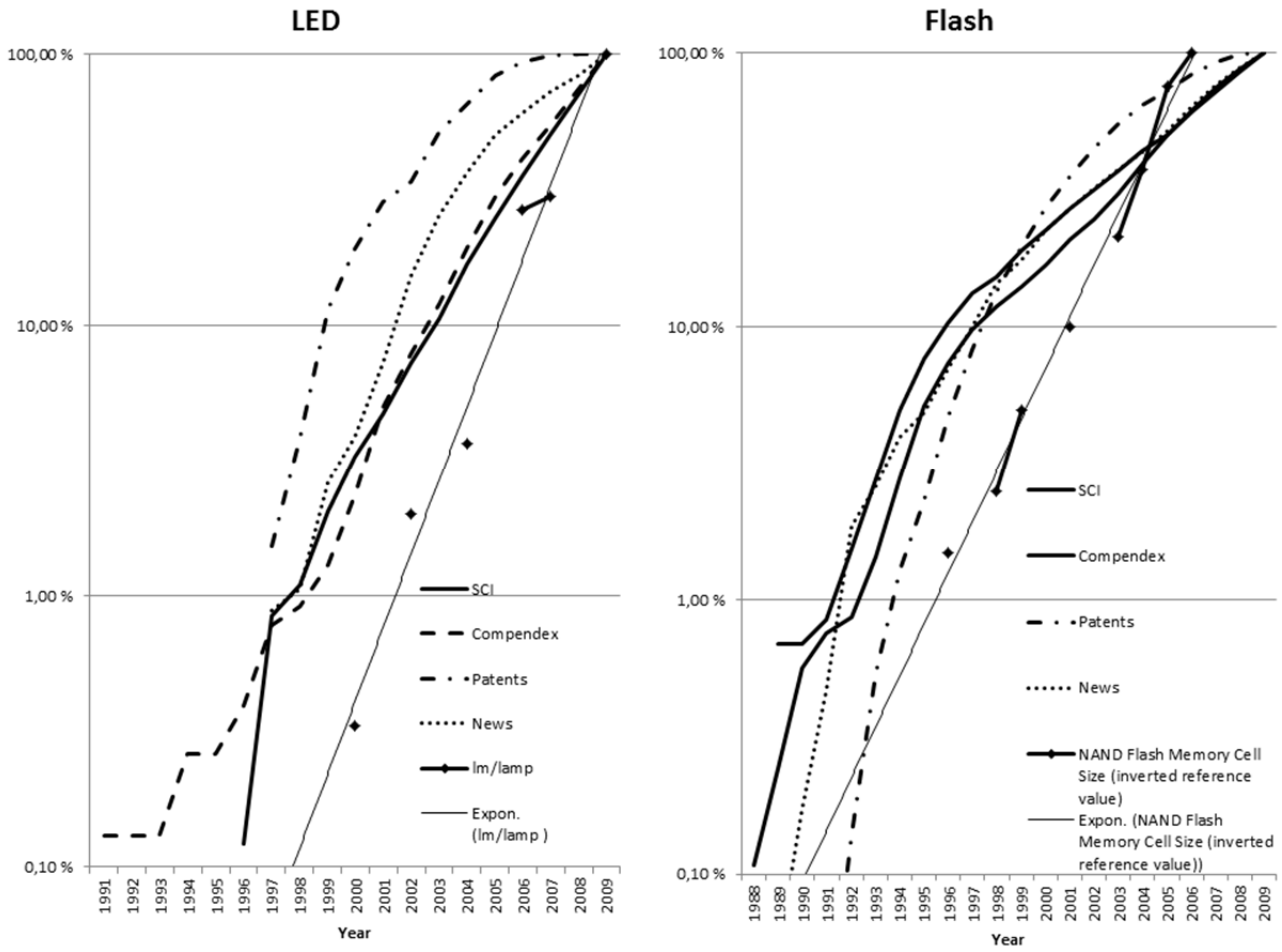
LED
100,00 %
10,00 %
1,00 %
0,10 %
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Year
SCI
Compendex
Patents
News
lm/lamp
Expon. (lm/lamp)
Flash
100,00 %
10,00 %
1,00 %
0,10 %
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Year
SCI
Compendex
Patents
News
NAND Flash Memory Cell Size (inverted reference value)
Expon. (NAND Flash Memory Cell Size (inverted reference value))

Figure 6