Capítulo 2-2º | Chapter 2-2º

La ciencia de datos | Data science

Fig. I A. C2.2.1-Ciclo de exageración de Gartner Crédito imág (De IOTpreneur - Trabajo propio, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=42665925). URL: https://upload.wikimedia.org/wikipedia/commons/7/75/Ciclo_de_sobreexpectacion_de_Gartner.png

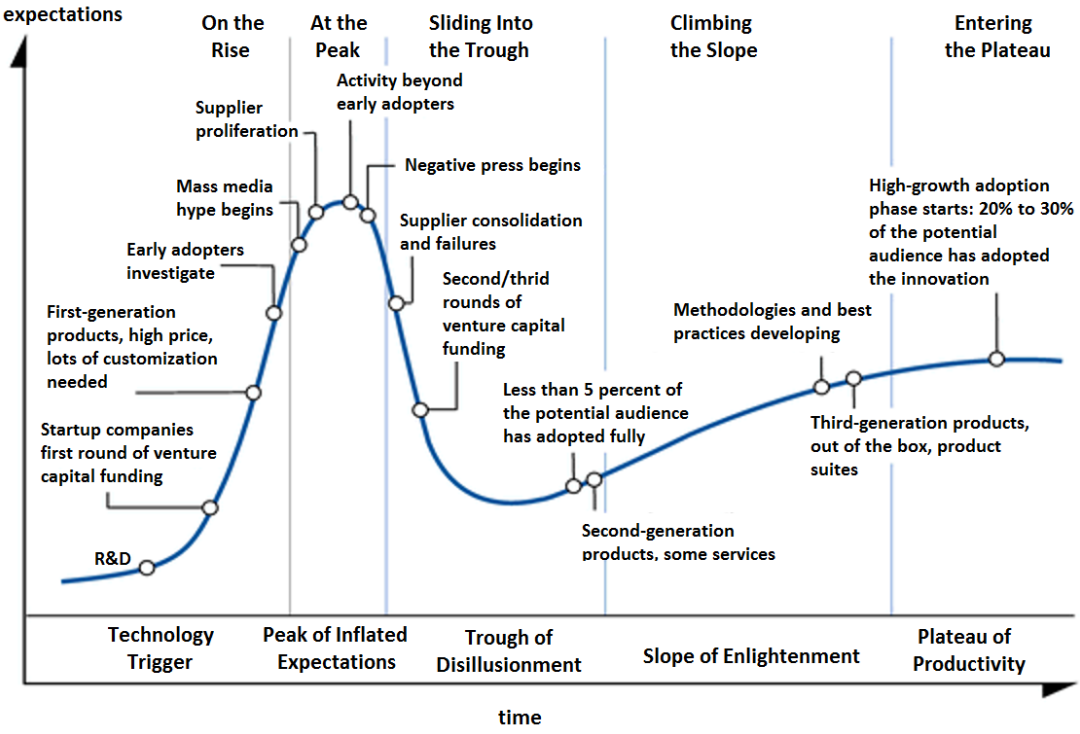

Fig. I A. C2.2.2- Gartner hype cycle. Credit imag. (From *Data science is different now*). URL: https://veekaybee.github.io/2019/02/13/data-science-is-different/

⇦ **Ir** al índice principal del libro	⇦ **Go** to the main index of the book
Contenidos	Contents
4.4-1.- Introducción	4.4-1.- Introduction
4.4-2.- El exceso de nuevos científicos de datos	4.4-2.- The excess of new data scientists
4.4-3.- La ciencia de datos como un trabajo engañoso requerido	4.4-3.- Data science as a deceptive job required
4.4-4.- Éstas son las claves	4.4-4.- Those are the keys
4.4-5.- Curso completo para principiantes de cienda de datos	4.4-5.- Full Course for Beginners Data Science

| Autor / Author: Juan Antonio Lloret Egea | Miembro de la Alianza Europea para la IA / Member to the European AI Alliance |https://orcid.org/0000-0002-6634-3351| |ISSN 2695-3803|| Escrito / Writed: 18/08/2019. Actualizado / Updated: 18/08/2019 |

> (Extracto del artículo publicado por Vicki Boykis *Data science is different now*. 13 de febrero, 2019) | (Excerpt from the article published by Vicki Boykis *Data science is different now* . February 13, 2019).

Ver también | See too: Procesamiento de lenguaje natural (PLN) y minería de texto (NLTK) | Natural language processing (PLN) and text mining (NLTK).

Relacionado | Related: Procesamiento del Lenguaje Natural (PLN) | Natural Language Processing (NLP)

4.4-1.- Introducción | Introduction

Lo que sigue son las palabras de un especialista en ciencia de datos que, desde la inconsistencia e inexistencia de esa materia hace años, ha aprendido y errado y superado los obstáculos que ha encontrado por el camino. La consecuencia ha sido una experiencia muy importante y que puede ser transferida a futuros candidatos que quieran adentrarse en esta materia.

No dispare para un trabajo de ciencia de datos. Prepárese para que la mayoría de su trabajo de científico de datos no sea ciencia de datos. Ajuste su conjunto de habilidades para eso[C2.2-1].

> No se meta en la cienca de datos y aprenda las habilidades necesarias para la ciencia de datos hoy

---

Aquí hay algunos problemas con los que realmente tendrá que lidiar en el espacio de datos. (Aunque hay muchos problemas estadísticos interesantes en los que pensar en la ciencia de datos, ninguno de estos enlaces de abajo se ocupa de ellos. Si bien los modelos de ajuste, la visualización y el análisis constituyen un componente de su tiempo como científico de datos, la ciencia de datos es y siempre ha sido principalmente para obtener datos limpios en un solo lugar para ser utilizados para la interpolación).

---

1. Crear paquetes en Python.
2. Poner R en producción.
3. Optimizar los trabajos de Spark para que se ejecuten de manera más eficiente.
4. Datos de control de versiones.
5. Hacer modelos y datos reproducibles.
6. SQL.
7. Construir y mantener datos limpios en lagos de datos.
8. Herramientas para pronósticos de series de tiempo a escala.
9. Uso compartido de escala de portátiles Jupyter.
10. Pensar en sistemas para datos limpios.
11. JSON.

[English]

What follows are the words of a data science specialist who has learned, erred and overcome obstacles he has encountered along the way since the inconsistency and absence of this matter. The consequence has been a very important experience and that can be transferred to future candidates who want to get into this subject.

Do not shoot for a data science job. Prepare yourself so that most of your data scientist work is not data science. Adjust your skill set for that.

> Don't get into data science and learn the skills necessary for data science today

---

Here are some problems that you will really have to deal with in the data space. (Although there are many interesting statistical problems to think about in data science, none of these links down deal with them. While adjustment models, visualization and analysis constitute a component of your time as a data scientist, data science is and always has been primarily to obtain clean data in one place to be used for interpolation).

---

1. Creating Python packages
2. Putting R in production
3. Optimizing Spark jobs so they run more efficiently
4. Version controlling data
5. Making models and data reproducible
6. Version controlling SQL
7. Building and maintaining clean data in data lakes
8. Tooling for time series forecasting at scale
9. Scaling sharing of Jupyter notebooks
10. Thinking about systems for clean data
11. Lots of JSON

Desde 2012, la industria de la ciencia de datos se ha movido extremadamente rápido. Ha pasado por casi todas las etapas del ciclo de exageración de Gartner.

> Hemos pasado por la fase de adopción temprana, la presión negativa en torno a la inteligencia artificial y el sesgo, la segunda y tercera ronda de capital de riesgo para empresas como Facebook, y ahora estamos en el punto de adopción de alto crecimiento: donde los bancos, las compañías de atención médica y otras compañías de Fortune 100 que llevan cinco años detrás del mercado también están contratando ciencia de datos en aprendizaje automático.

Desafortunadamente, lo que no ha cambiado es la exageración de los medios de comunicación en el campo de la ciencia de datos, que ha proclamado al científico de datos como la 'carrera más sexy del siglo XXI' tantas veces que ahora hay un problema importante. Ese problema es una sobreoferta de científicos de datos *junior* que esperan ingresar a la industria y expectativas dispares sobre lo que pueden esperar encontrar una vez que obtengan ese codiciado título de 'científico de datos'.

[English]

Since 2012, the data science industry has moved extremely fast. It has gone through almost every stage of Gartner's cycle of exaggeration.

> We have gone through the early adoption phase, the negative pressure around artificial intelligence and bias, the second and third round of venture capital for companies like Facebook, and now we are at the point of high growth adoption: where the Banks, health care companies and other Fortune 100 companies that have been behind the market for five years are also hiring data science in machine learning.

Unfortunately, what has not changed is the exaggeration of the media in the field of data science, which has proclaimed the data scientist as the 'sexiest career of the 21st century' so many times, that there is now a major problem. That problem is an oversupply of junior data scientists who expect to enter the industry and disparate expectations about what they can expect to find once they get that coveted title of 'data scientist'.

4.4-2.- El exceso de nuevos científicos de datos | The excess of new data scientists

---

> Existe una verdadera burbuja de suministro de ciencia de datos

LinkedIn dice que hay 151,717 personas con habilidades de ciencia de datos que faltan en el mercado Aunque no está claro si esto significa directamente científicos de datos o sólo personas con algún subconjunto de esas habilidades, supongamos que es lo primero. Entonces, hay 150,000 vacantes para científicos de datos en el país. Dado que hay 100,000 que han comenzado un curso de ciencia de datos, supongamos nuevamente que 7,000 de éstos terminan. Pero ninguno de esos números está tomando en cuenta todos los programas y vías para crear nuevos candidatos de ciencia de datos: MOOC fuera de fast.ai como Coursera, más de 10 campamentos de arranque a nivel nacional como Metis y Asamblea General que tienen cohortes de 25 personas cada 12 semanas. títulos remotos de lugares como UCLA, títulos de pregrado *in situ* en análisis y ciencia de datos, YouTube y más. También hay una gran cantidad de doctores que, incapaces de encontrar trabajo en un mercado laboral extremadamente apretado, están migrando de la academia a la ciencia de datos.

> Como resultado, el mercado puede ser muy difícil y muy desalentador para la avalancha de principiantes

---

[English]

> There is a true data science supply bubble

LinkedIn says there are 151,717 people with missing data science skills in the market. Although it is not clear if this directly means data scientists or just people with some subset of those skills, suppose it comes first. So, there are 150,000 vacancies for data scientists in the country. Since there are 100,000 who have started a data science course, suppose again that 7,000 of these end. But none of those numbers is taking into account all the programs and ways to create new data science candidates: MOOC out of fast.ai as Coursera, more than 10 nationally-initiated start-up camps like Metis and General Assembly that have cohorts of 25 people every 12 weeks. remote titles from places like UCLA, undergraduate degrees in situ in data analysis and science, YouTube and more. There are also a large number of doctors who, unable to find work in an extremely tight labor market, are migrating from the academy to data science.

> As a result, the market can be very difficult and very daunting for the avalanche of beginners

---

4.4-3.- La ciencia de datos como un trabajo engañoso requerido | Data science as a deceptive job required

---

El segundo problema es que una vez que estas personas jóvenes llegan al mercado, entran con un conjunto de expectativas poco realistas sobre cómo será el trabajo de ciencia de datos. Todos piensan que van a hacer aprendizaje automático, aprendizaje profundo y simulaciones bayesianas. Esto no es su culpa; esto es lo que enfatizan los currículos de ciencia de datos y los medios tecnológicos. La realidad es que la 'ciencia de datos' nunca ha tenido tanto que ver con el aprendizaje automático como con la limpieza, la configuración de datos y el traslado de un lugar a otro.

> Lo que está quedando claro es que, en la etapa tardía del ciclo exagerado, la ciencia de datos se está acercando asintóticamente a la ingeniería, y las habilidades que los científicos de datos necesitan para avanzar están menos basadas en la visualización y en las estadísticas, y más en línea con la informática tradicional currículo.

Esto ha llevado a varias cosas. En primer lugar, el aumento del título de trabajo de 'ingeniero de aprendizaje automático' como uno que tiene más prestigio y un mayor potencial de ganancias en los últimos 3-4 años, En segundo lugar, ha llevado a un grado severo de deflación del título de trabajo para los científicos de datos, donde debido al prestigio del título de trabajo de científico de datos, compañías como Lyft contratarán para títulos de trabajo de ciencia de datos, pero con conjuntos de habilidades de analistas de datos, lo que resulta en un mayor imagen sesgada de lo que constituye un trabajo de 'ciencia de datos', y exactamente cuántos de ellos están disponibles para los nuevos

participantes.

[English]

The second problem is that once these young people reach the market, they enter with a set of unrealistic expectations about what data science work will be like. Everyone thinks they will do machine learning, deep learning and Bayesian simulations. This is not your fault; this is what the data science curricula and technological media emphasize. The reality is that 'data science' has never had so much to do with machine learning as with cleaning, data configuration and moving from one place to another.

> What is becoming clear is that, at the late stage of the exaggerated cycle, data science is asymptotically approaching engineering, and the skills that data scientists need to advance are less based on visualization and statistics, and more in line with traditional computer science curriculum.

This has led to several things. First, the increase in the job title of 'machine learning engineer' as one that has more prestige and greater earning potential in the last 3-4 years. Secondly, it has led to a severe degree of deflation of the job title for data scientists, where due to the prestige of the data scientist job title, companies like Lyft will hire for data science job titles, but with data analyst skill sets, resulting in a greater biased image of what constitutes a 'data science' work, and exactly how many of them are available to new participants.

> Lo realmente clave de todas estas habilidades es que también son fundamentales y críticas para el desarrollo de *software* fuera de la ciencia de datos, lo que significa que, en caso de que no pueda encontrar un trabajo de ciencia de datos, puede hacer la transición rápidamente al desarrollo de *software* o *devops*.

> The really key thing about all these skills is that they are also critical and critical for software development outside of data science, which means that, in case you can't find a data science job, you can quickly transition to software development or devops.

Fig. I A. C2.2.3- Crédito imág (Imagen de Pete Linforth en Pixabay). URL:
https://pixabay.com/es/illustrations/inteligencia-artificial-robot-cyborg-4117070/

4.4-4.- Éstas son las claves | Those are the keys

> Aprenda SQL / Learn SQL

SQL no es atractivo, y no es una solución a la lista de problemas que acabamos de enumerar. Pero para todos los efectos, para comprender cómo acceder a los datos, es muy probable que encuentre una base de datos en algún lugar que requiera que escriba algunas consultas SQL y obtenga una respuesta. SQL es tan bueno y tan popular que incluso NoSQL y las soluciones de almacenamiento de valores clave lo están reimplementando. El siguiente paso, después de que aprenda bien SQL, es comprender un poco sobre cómo funcionan las bases de datos para que pueda aprender a optimizar sus consultas.

[English]

SQL is not attractive, and it is not a solution to the list of problems just listed. But for all intents and purposes, to understand how to access the data, it is very likely that you will find a database somewhere that requires you to write some SQL queries and get an answer. SQL is so good and so popular that even NoSQL and key value storage solutions are reimplementing it. The next step, after you learn SQL well, is to understand a little about how databases work so you can learn to optimize your queries.

> Aprenda un lenguaje de programación extremadamente bien y aprenda conceptos de programación | Learn an extremely good programming language and learn programming concepts

Hay mucho debate sobre qué lenguaje elegir para la ciencia de datos. Python me ha servido extremadamente bien. Es bastante fácil comenzar como principiante y nos sirve desde armar un modelo en *scikit learn* hasta acceder al AWS API para crear una aplicación web, para limpiar datos, para crear modelos de aprendizaje profundo. Pero, una vez más, el consejo es no ir a profundidad estadística, sino a la programación general. Hay algunas tareas para las que Python no es ideal: aplicaciones a gran escala, dependencias de empaque y algunos trabajos numéricos específicos, particularmente series de tiempo y un montón de características que vienen en R de fábrica, similar a lo que ofrece statsmodels, pero a un nivel mucho más granular.

[English]

There is much debate about which language to choose for data science. Python has served me extremely well. It is quite easy to start as a beginner and it helps us to build a model in scikit learn to access the AWS API to create a web application, to clean data, to create deep learning models. But, once again, the advice is not to go to statistical depth, but to general programming. There are some tasks for which Python is not ideal: large-scale applications, packaging dependencies and some specific numerical jobs, particularly time series and a lot of features that come in factory R, similar to what statsmodels offers, but to A much more granular level.

> Aprenda a trabajar en la nube / Learn to work in the cloud

La nube está en todas partes en estos días, y es probable que tenga que trabajar en la nube en uno de sus próximos trabajos. Es mucho más fácil comenzar con ventaja, particularmente a medida que más y más paradigmas de aprendizaje automático se trasladan a los proveedores de la nube en estos días (SageMaker, Cloud AI y Azure Machine Learning), hay plantillas listas para usar para implementar algoritmos, y más de los datos de su empresa comienza a almacenarse allí. Los paradigmas de diseño de la nube son similares en el sentido de que debe comprender cómo unir los servicios, cómo dividir lógicamente su parte de la nube de otros servidores en la nube y cómo trabajar con muchos JSON.

> Es probable que trabaje con AWS, el líder de la industria, pero cada vez más lugares están adoptando Google Cloud, y muchas empresas más conservadoras y tradicionales que ya hacen negocios con Microsoft tendrán Azure.

A modo de resumen para los nuevos científicos de datos recojo el final del artículo de Vicki Boykis, que me parece una sabia directriz para los tiempos de indecisiones y nuevas tareas asociadas a la inteligencia artificial. Nos habla de un libro, *Bird by bird* (pájaro por pájaro), de Anne Lamott. Nos dice:

> “Hace treinta años, mi hermano mayor, que tenía diez años en ese momento, estaba tratando de obtener un informe sobre pájaros escrito que tenía tres meses para escribir... Estábamos en la cabaña de nuestra familia en Bolinas, y él estaba en la mesa de la cocina cerca de las lágrimas, rodeado de papel y lápices y libros sin poder escribir sobre pájaros, inmovilizado por la enorme tarea que tenía por delante. Entonces mi padre se sentó a su lado, puso su brazo alrededor del hombro de mi hermano y dijo. Pájaro por pájaro, amigo. Sólo tómalo pájaro por pájaro '”. Y lo consiguió.

> No dejes que el bombo te abrume. No te confundas con las palabras de moda o las imágenes de los *hipsters* con MacBooks. Concéntrate en un solo pájaro y construye desde allí.

[English]

The cloud is everywhere these days, and you probably have to work in the cloud in one of your next jobs. It is much easier to start with an advantage, particularly as more and more machine learning paradigms move to cloud providers these days (SageMaker, Cloud AI and Azure Machine Learning), there are ready-to-use templates to implement algorithms, and more of your company's data begins to be stored there. The cloud design paradigms are similar in that you must understand how to join services, how to logically divide your share of the cloud from other servers in the cloud and how to work with many JSONs.

> You are likely to work with AWS, the industry leader, but more and more places are adopting Google Cloud, and many more conservative and traditional companies that already do business with Microsoft will have Azure.

As a summary for the new data scientists, I collect the end of the article by Vicki Boykis, which seems to me to be a wise guideline for times of indecision and new tasks associated with artificial intelligence. He tells us about a book, *Bird by bird*, by Anne Lamott. Tells us:

> “Thirty years ago, my older brother, who was ten years old at the time, was trying to get a written bird report that he had three months to write ... We were in our family's cabin in Bolinas, and he was in the kitchen table near tears, surrounded by paper and pencils and books unable to write about birds, immobilized by the enormous task ahead. Then my father sat next to him, put his arm around my brother's shoulder and said. Bird by bird, friend. Just take it bird by bird. ' And he got it.

> Don't let the hype overwhelm you. Do not be confused with the buzzwords or images of hipsters with MacBooks. Focus on a single bird and build from there.

> “
>
> Conclusión al artículo / Conclusion to the article

En mi opinión el autor (Vicki Boykis) viene a decir (utilizando un ejemplo en línea con el autor del artículo y su *pájaro a pájaro*):

Que si los 'pollos habladores' es una industria que está de moda y que está siendo impulsada por una inconsistente base de información, aunque haya demanda, es conveniente no centrarnos en los pollos habladores. Y, sin embargo, centrarnos en el pienso que comen, las vacunas que se les pone, los hogares de los pollos, etc. Y nuestro beneficio será inmediato, apoyaremos a los criadores de pollos habladores con nuestros servicios auxiliares y en beneficio nuestro. Y como habrá tanto criador de pollos habladores, sin duda, trabajo no nos faltará y podremos elegir quién es el criador de pollos

Y su recomendación de no centrarnos demasiado en la estadística, a la hora del aprendizaje automático u otros, es tanto como decir que no tenemos que entender lo que dicen los pollos habladores para darles de comer o un hogar de pollos. (Aunque sea últil y un argumento bueno saber qué dicen). Y respecto al desarrollo de *software* o *devops*, pues quiere decir que si la industria de pollos habladores no tuviera espacio para nuestro futuro en él como procuradores de servicios auxiliares, entonces podemos fabricar pienso para otros animales, también vacunas para perros habladores y hogares para gatitos mudos. Sabio consejo.

[English]

In my opinion the author (Vicki Boykis) comes to say (using an example online with the author of the article and his *bird by bird*):

That if talkative chickens is an industry that is fashionable and that is being driven by an inconsistent information base, even if there is demand, it is convenient not to focus on talkative chickens. However focus on the feed they eat, the vaccines they are given, chicken homes, etc. And our benefit will be immediate, we will support the breeders of talkative chickens with our auxiliary services and for our benefit. And as there will be so much breeder of talkative chickens, without a doubt, work will not be lacking and we will be able to choose who is the breeder of talking chickens that interests us most for our work interest. And the mechanism is simple: we break the dynamics of demand and supply. And his recommendation not to focus too much on statistics, at the time of machine learning or others, is as much as saying that we don't have to understand what talkative chickens say to feed them or a chicken home. Although it is useful and a good argument to know what they say. And regarding the development of software or devops, it means that if the talkative chicken industry did not have space for our future in it as auxiliary services, then we can manufacture feed for other animals, also vaccines for talkative dogs and homes for dumb kittens. Wise advice.

4.4-5.- Curso completo para principiantes de cienda de datos | Full Course for Beginners Data Science

---

Fig. I A. C2.2.4- Curso completo para principiantes / Full Course for Beginners. Crédito imág (freeCodeCamp.org). URL: https://youtu.be/ua-CiDNNj30

Bibliografía | Bibliography

[C2.2-1] Boykis, V. (13 de febrero, 2019). Data science is different now. . *[Recuperado (18/08/2019) de* https://veekaybee.github.io/2019/02/13/data-science-is-different/]

 ISSN 2695-3803|

Revision #57
Created Sun, Aug 18, 2019 8:00 AM by Juan Antonio Lloret Egea
Updated Sun, Oct 27, 2019 2:11 PM by Juan Antonio Lloret Egea