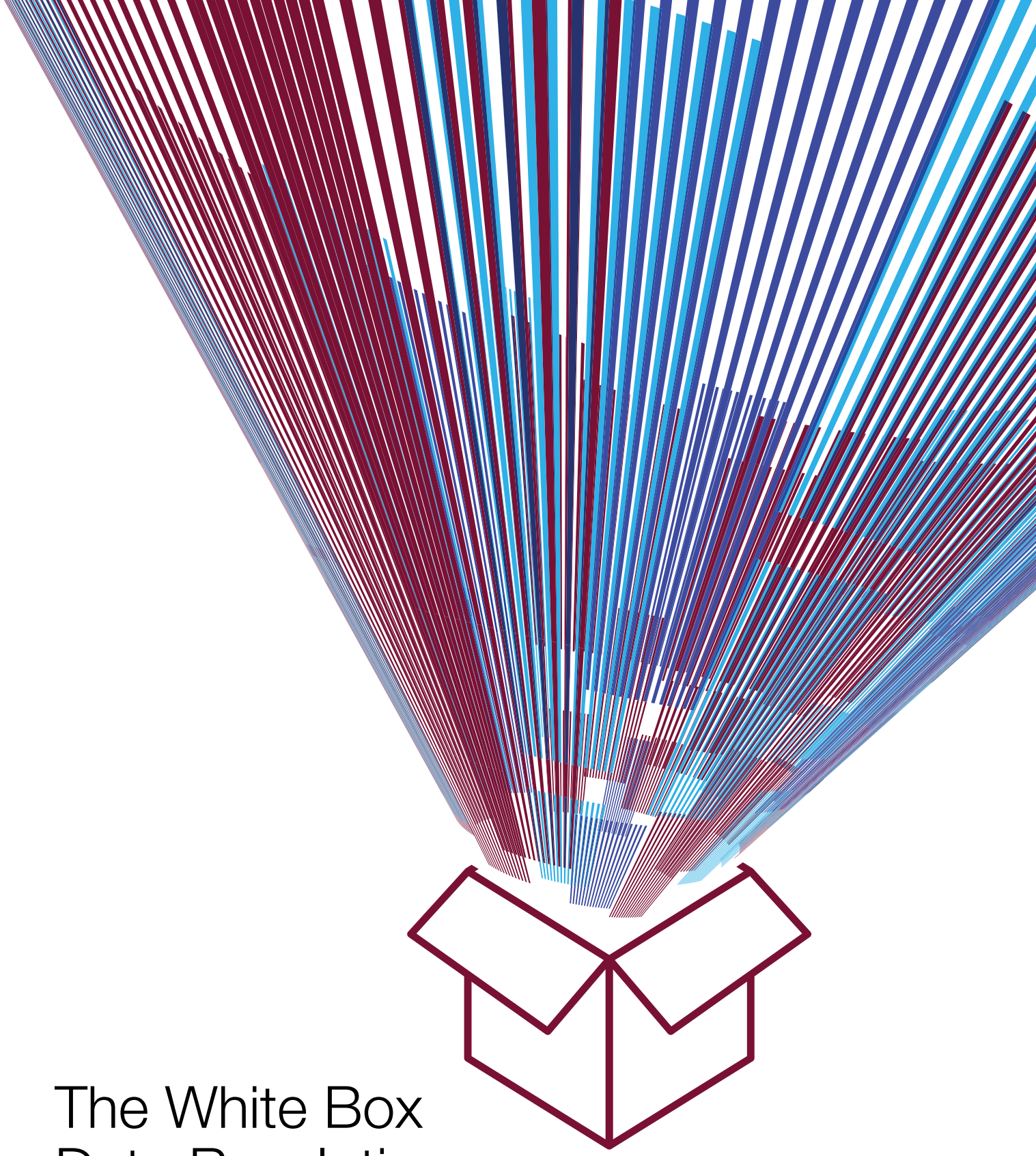


The White Box Data Revolution

opencorporates WHITE PAPER

The White Box Data Revolution

WHITE PAPER

What is White Box data?
Black Box data is vague about what exactly it represents, White Box followed a well-defined and transparent data model which is clear about what it represents. With Black Box you have no information about where the data came from, with White Box you get that provenance. Black Box does not tell you when it was retrieved, White Box tells you how fresh it is. Black Box is a hint or a lead; White Box gives you a trail you can follow. For journalists and law enforcement, Black Box is a dead end but White Box gives them an evidential trail.

Chapter 1: **Black and White**

For the past 50 years, we have lived in a world dominated by "Black Box data" – data that is opaque, not well-defined, uses proprietary identifiers, and has poor data-quality feedback loops. This is rapidly changing – and nowhere more so than in the field of company data – as it is replaced by data fit for use in the modern world: so-called White Box data.

This white paper focuses on this seismic shift – what White Box data means, the myriad benefits of it, how it is being used by some large companies, and how its importance will increase in the future, especially in company data. A dozen data experts were interviewed for this white paper and a clear theme emerged: White Box data can transform the way companies, regulators, investigators and journalists work, while those that stick to Black Box risk becoming extinct.

But given that White Box data has so many benefits, the question might reasonably be asked: why do we have Black Box data in the first place? Why has, for example, the Dun & Bradstreet "business data" identified by their proprietary DUNS number been successful in the US, despite its many shortcomings? Why do firms use the DUNS number so heavily, despite the fact that it cannot tell them whether it refers to a business, a corporate address or an individual, nor where or how the associated data came from.

Black Box data has been the default because, in the past, it was sufficient to meet the basic data needs of businesses, governments and banks. Data was typically looked at on a pay-per view, record-by-record basis. The uses of data were very limited, by a small number of clients, and for those specific uses Black Box was just about good enough.

Moreover regulation – for example, relating to Anti-Money Laundering – was mostly a 'tick-box' affair, blindly trusting the claims of third parties. This allowed systemic risk and money laundering to proliferate en masse. Today, the regulatory compliance landscape has changed and regulators

say this approach is insufficient. But prior to this shift, if you were getting data from just one source, it mattered less that the data was of unknown provenance and uncertain quality.

Black Box data has also been propped up by a business model which is based on restricting access to the data – the pay-per-view model – and the metadata surrounding it. The data providers did not want to disclose the source of the data – otherwise, users could bypass them and go directly to the source. Disclosing the source could show they had used mechanisms or sources that were less than reliable (typically other black-box data providers), or had even just modelled the data silently and opaquely.

This restricted-access business model also had the consequence of shutting out large groups of users who would like to access the data. This in turn led to poor data quality feedback loops. By contrast, White Box data is accessed by hundreds of thousands of users, which in turn creates a high-quality feedback loop.

These data quality problems will be familiar to business intelligence companies, compliance officers, journalists, regulators, consulting and accountancy firms, and fintech innovators. But a change in the way these organisations use data means that the Black Box, proprietary business model is quickly becoming obsolete.

Chris Taggart, Co-Founder of OpenCorporates, has held hundreds of conversations with myriad data users over the last few years, and observed a clear trend. “All of these people – whether they are in business insight, journalists, regulators, people using AI, Machine Learning (ML) or blockchain – have an instinctive understanding of the problem of closed and opaque data, and the importance of accuracy and provenance,” he said. “Talk to them briefly and seriously about their problems and it is clear that they realise their current data provider is no longer fit for purpose. The solution for them is to replace their Black Box data with what we are calling ‘White Box’.”

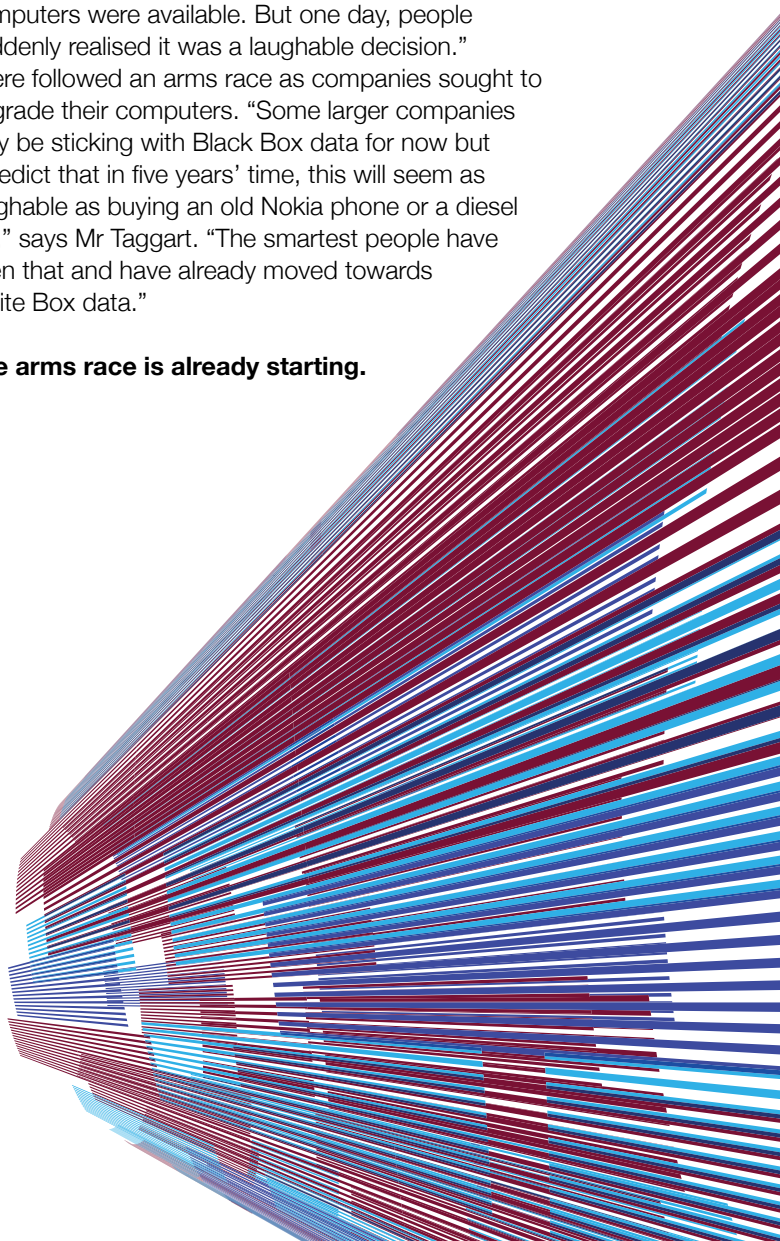
The shift away from opaque, proprietary data is even happening at governmental level. After two decades of using the DUNS number, the US General Services Administration recently decided to switch from the DUNS to a new, open identifier when validating its business with contractors and other third parties.

“This was probably the most egregious example of the use of proprietary data so it is the culmination of a trend,” says Hudson Hollister, who founded the highly influential Data Coalition (and now runs HData). “This decision is probably the largest ever move away from Black Box data towards White Box data. This data ought to be freely available to those who paid for it as taxpayers.”

> **[The move away from the DUNS number] is probably the largest ever move away from Black Box data towards White Box data**
>
> **Hudson Hollister, Founder of the Data Coalition**

Change will not happen overnight, but this white paper will show how some large companies have moved from Black to White Box data with impressive results. Mr Taggart compares it to the revolution in computing 30, 40 years ago. “There used to be an idea that nobody would get fired for buying IBM,” he says. “Companies saw it as a safe and uncontroversial bet, even when smaller and faster computers were available. But one day, people suddenly realised it was a laughable decision.” There followed an arms race as companies sought to upgrade their computers. “Some larger companies may be sticking with Black Box data for now but I predict that in five years’ time, this will seem as laughable as buying an old Nokia phone or a diesel car,” says Mr Taggart. “The smartest people have seen that and have already moved towards White Box data.”

The arms race is already starting.



Chapter 2:
Why White Box is replacing Black Box

This white paper focuses on data on companies and business information, but the problems of Black Box data are relevant in many other sectors. These drawbacks are obvious to their users, but they are also long-standing. So why is this shift towards well-defined and provenance White Box data only happening now?

1. Regulation
Regulatory requirements have become more stringent in a wide range of areas relating to companies. A good example is the drastic change in regulation about Know-Your-Customer (Anti-Money Laundering) requirements. Regulators now expect financial services companies to provide an audit trail and proof that they have done all they can to establish the facts about a customer, including the people associated with it. If a bank can only point to some opaque, proprietary information, that will not be enough to satisfy regulatory expectations and may lead to enforcement action and a fine. There are similar legal risks when dealing with personal data that has no provenance – for example under GDPR.

2. Connectedness of data
Sir Tim Berners-Lee once said: "In an extreme view, the world can be seen as only connections, nothing else." That view is no longer extreme, but a reality. And in this world of globally connected data, "Black Box data" is just not fit for purpose. Metadata – especially provenance – is essential context to making sense of these connections.

The days of using data in a siloed way – one source at a time – are over. Today, firms need to take datasets from different sources and combine them to build a data ecosystem. The ecosystem should not only refresh when new data comes in but allow the user to make connections between the different types of data.

3. Contractualisation
For nearly two centuries, business has relied on contracts between legal entities underpinned by the rule of law. But in the past ten years, the world has become 'contractualised' at an unimaginable scale. Whenever you click to sign up to a website or install a new app or use Google Maps, even, you are entering into a contractual agreement – which in turn is backed by hundreds of other agreements. It's vital that we – citizens, but also business too – understand this new world: who we're buying from, working for, selling to. In this new 'data world', we also need clarity on who has our data and who it has been given to. It has never been more important to have well-defined data on the many third parties with whom you are contracting.

4. Globalisation
Black Box company data was less problematic when business tended to be done locally between two entities in the same jurisdiction – when the village shop bought food from the local farm and did its books with the accountant down the street. You knew exactly who you were dealing with, so had less need for well-defined data about that entity. Today it is normal for transactions to be cross-jurisdictional, so companies need to know who they are doing business with in a range of jurisdictions, and what third parties those companies are connected to.

"One of the struggles organisations have when doing business internationally is that the availability of legal entity data differs so much from jurisdiction to jurisdiction," says Caryn McEwen, Head of Global Licensing & Content Operations at LexisNexis. "As the world becomes more globalised and companies need to rely on supply chains in other areas, this becomes more critical."

Scott Taylor, The Data Whisperer and Principal Consultant at MetaMeta Consulting who used to work for Dun & Bradstreet, agrees. "One of the holes in the Black Box structure is the lack of a true global commercial view of companies," he says. "Black Box data is a legacy way of doing things, like taking old Kodak photos."

5. A data world

The scale and speed of our world of data is increasing, and companies are adopting technology like Artificial Intelligence to try to make sense of this new world. “Digitisation of the world is driving tremendous changes in the way business operates, and individual companies are using data at scale to do real-time decision-making,” says Mike Olson, Co-Founder at Cloudera.

“We are going to see corporations being created digitally, exist briefly and disappear for legitimate and illegitimate business purposes. They will be registered and dissolved by machines. The data volumes to deal with in this case will be absolutely staggering and we will need algorithmic ways to monitor that activity and spot good and bad companies.”

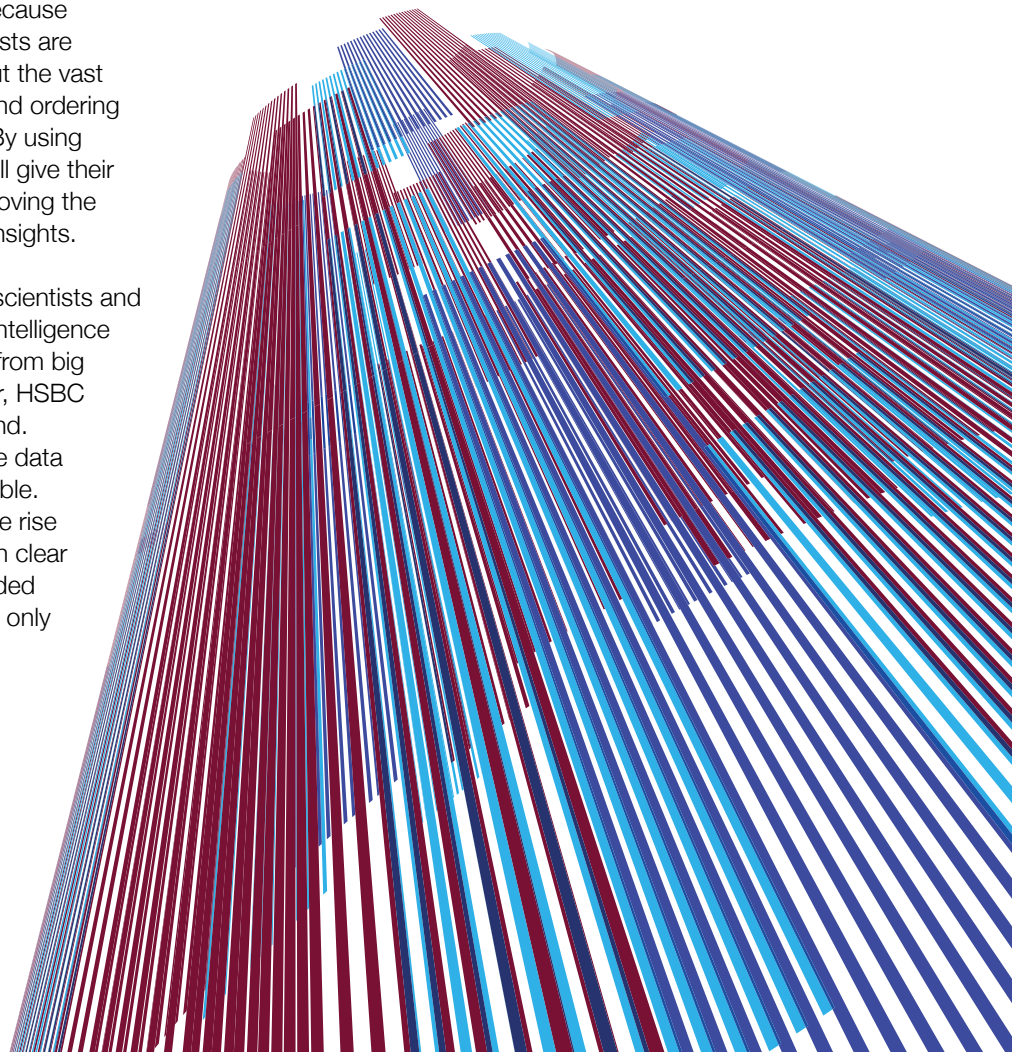
Scott Taylor says that as companies experiment with these new technologies, they are realising they will only be effective if they fuel them with high-quality data. “There are two big buckets – data management and Business Intelligence (BI) – and companies are focused on BI of which the sexiest version is AI,” he says.

“But the two go together – data management is determining the truth, BI is deriving the meaning, and companies are now understanding they need to focus more on data management because their machines don’t work. Data scientists are very highly paid and a lot in demand but the vast majority spend their time on cleaning and ordering the data – this is data janitorial work.” By using White Box data sources, companies will give their data scientists the best chance of improving the business’ operations and finding new insights.

Companies are racing to acquire data scientists and technological capacity to use Artificial Intelligence and Machine Learning to gain insights from big data and solve problems. Only last year, HSBC recruited 1,000 digital experts to this end. For these insights to have any utility, the data needs to be accurate and understandable. Blockchain is another technology on the rise that needs data that is well-defined with clear provenance. No matter how well-regarded these technologies are, their results will only be as good as the data feeding them.

Black Box data is a legacy way of doing things, like taking old Kodak photos

Scott Taylor, The Data Whisperer

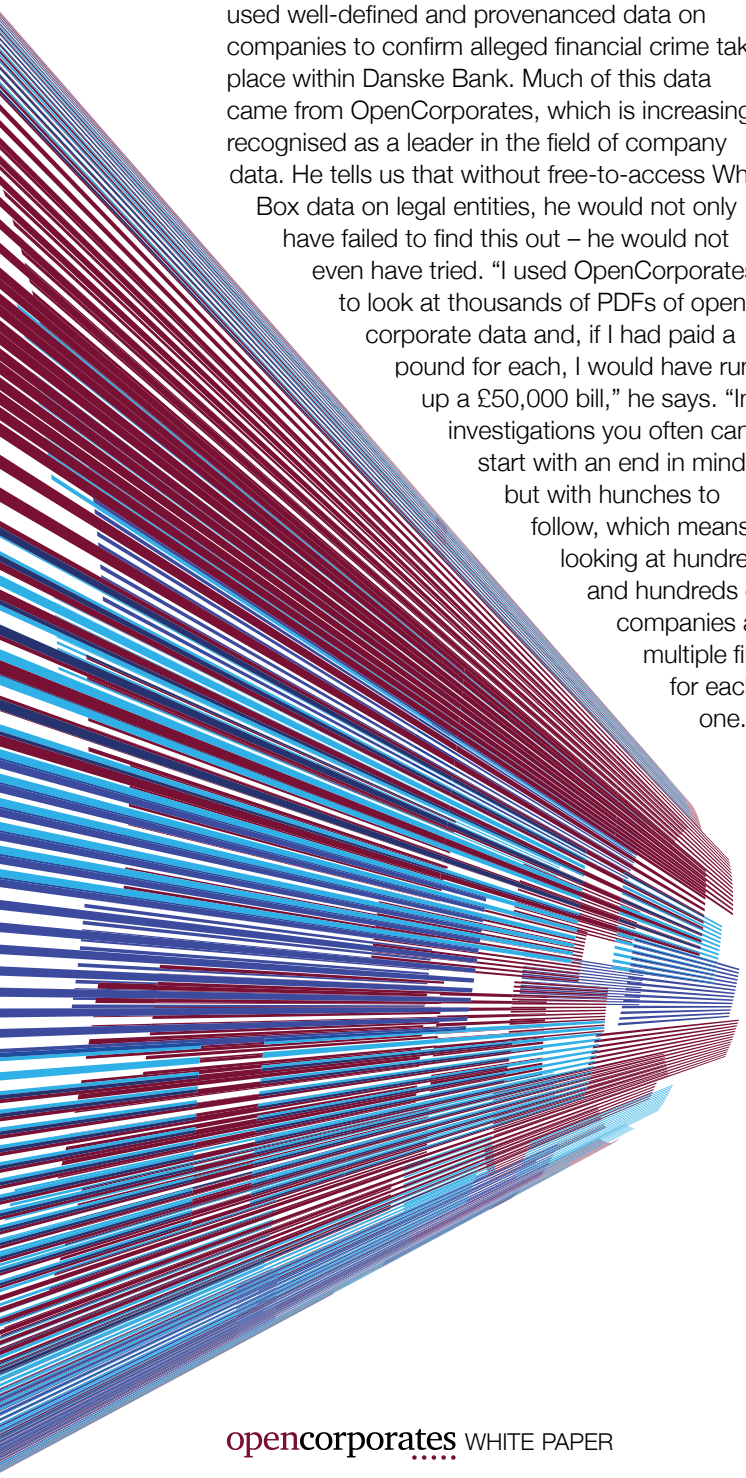


Chapter 3:
Opening the White Box

White box data may be the future, but in many cases it is already transforming the way companies, regulators and journalists work:

A. Investigations

Graham Barrow is a writer and investigator who used well-defined and provenanced data on companies to confirm alleged financial crime taking place within Danske Bank. Much of this data came from OpenCorporates, which is increasingly recognised as a leader in the field of company data. He tells us that without free-to-access White Box data on legal entities, he would not only have failed to find this out – he would not even have tried. "I used OpenCorporates to look at thousands of PDFs of open corporate data and, if I had paid a pound for each, I would have run up a £50,000 bill," he says. "In investigations you often cannot start with an end in mind but with hunches to follow, which means looking at hundreds and hundreds of companies and multiple filings for each one.



White box data is the only way investigators like me can do that."

Mr Barrow likens company data that is closed and Black Box to "a world where the only potential to report crime was by the person who suffered it, or proactive action by the police themselves." "It's impossible," he says, "and yet, this appears to be the argument for those who would require corporate registers to be available solely to those who inhabit them and those who have responsibility for policing them."

Investigations have become more complicated with the globalisation of business. "You can now have a UK LLP that has a director located as a legal entity in the Seychelles, a trading address in Moscow, a beneficial owner in Ukraine and a bank account in Estonia," he says. So it has never been more important that regulators, journalists and compliance officers are able to access White Box data on legal entities across multiple jurisdictions, including connections between firms and their directors. The stakes are high. "Without a shadow of a doubt, enforcement agencies do not have resources or manpower to discover all illegal activity," says Mr Barrow. "Without globally accessible corporate data which is aggregated in a consistent way, criminals will always find a way to hide their ownership of a company."

Following his revelations on Danske Bank, Mr Barrow was invited by Denmark's business minister to make recommendations on how Denmark's banks should be regulated. "Part of that advice," he said, "was to make their data White Box."

> **Without globally accessible corporate data which is aggregated in a consistent way, criminals will always find a way to hide their ownership of a company**
>
> **Graham Barrow, investigator**

B. Verification
Verifying that a customer, client or third party is who they say they are is crucial to the work of financial services firms, consulting companies, tech companies and compliance officers. Black Box data made this very difficult not least because of opacity of the data model. For example, the proprietary DUNS number does not distinguish between a company, an address or a director. But firms who are using well-modelled White Box data are already enjoying efficiency savings and more accurate results.

Recent regulatory developments around Know-Your-Customer (KYC) and Anti-Money Laundering requirements mean that companies are now expected to take a risk-based approach to financial crime, using relevant data sources to accurately assess the level of risk posed by a prospective customer then acting on that assessment. That assessment will only be as good as the data on which it is based.

"KYC is a great example of Machine Learning (ML) being applied to great effect in financial services, and the better your data, the more reliable and predictive your model is going to be," says Mike Olson, Co-Founder of Cloudera. "We can collect more information about customers, including their corporate registration data and directors and officers and relationships among them. We are then able to use algorithms to export those networks and use ML to recognise patterns of fraudulent companies set up to crunch credit card transactions then fold up and disappear."

Justin Fitzpatrick is CEO and Co-Founder of DueDil, a company intelligence platform that banks use in their Know-Your-Business and KYC processes. "From a business perspective, a manual compliance process just does not scale, particularly if you are a bank or insurer with SME business lines where you are having

The better your data, the more reliable and predictive your model is going to be

Mike Olson, Co-Founder of Cloudera

Case study: Quantifind

Adam Mulliken, VP of Analytics at Quantifind, says: "More and more people know they need to use White Box data on companies and we made an investment to integrate this data, including data from OpenCorporates. It was a strategic decision for us and it has proved a lot of value in the market, we feel like it has been a competitive advantage. We build software products to help investigators of financial crime and our White Box data on legal entities gets a disproportionately high amount of attention from investigators. It's the very factual, credible data they have wanted to get their hands on but it hasn't been easy before this data became available. Investigators of money laundering and fraud have lots of internal data from banks which is very White Box and factual, but they need similar data that helps them to understand the external profile of individuals and businesses they are investigating. The more credible, reliable and trustworthy data is, the more actionable it is from an enforcement standpoint."

to onboard hundreds or thousands of customers every month," he says. "If a regulator such as the FCA comes to your door asking for evidence of how you went through a process to satisfy guidance on KYB and KYC, it does not really cut it if you give them a stack of papers. Companies need to use technologies to bring together these different datasets that often don't interact with each other. But right now far too much of companies' onboarding processes are completed manually by people who are scraping around for information."

Scott Bell, Managing Director at Deutsche Bank, says there is a "high cost" to a bank that fails to establish good compliance and KYC systems. "Banks have incurred billions of fines because they had not worked out who their customers were and where their money came from, and frontline bankers are under individual requirements and face criminal cases if they do not do compliance properly," he said. "A bank has to take all reasonable professional efforts to figure out the source of money, ownership of companies and domicile of companies, so the higher quality and quicker data it can access, the better."

C. Data aggregation

The way firms use data has changed dramatically in recent decades, and it has made the business model of Black Box data providers redundant and in turn made White Box data essential. "The old method by traditional data providers was to go away and collect data on a company from a single source – usually a phone call to that company – then deliver that to a paying customer," says David Clarke, formerly Chief Data Officer at D&B. "We are now in a world of aggregation where we cannot just go to one place. A multi-sourcing or data aggregation solution is far more appropriate for today's business world, and the kind of data provided by OpenCorporates has an advantage here."

Anyone trying to establish a clear picture of a legal entity must aggregate different sources of data. "I would match legal entity data with web data, financial accounts and compliance data to get as total and complete a view as I can of the commercial entities I am interested in," says Mr Clarke. "Open data which does not impose a standard or format or parochial numbering system makes it so much easier for me to tie that data in with other sources."

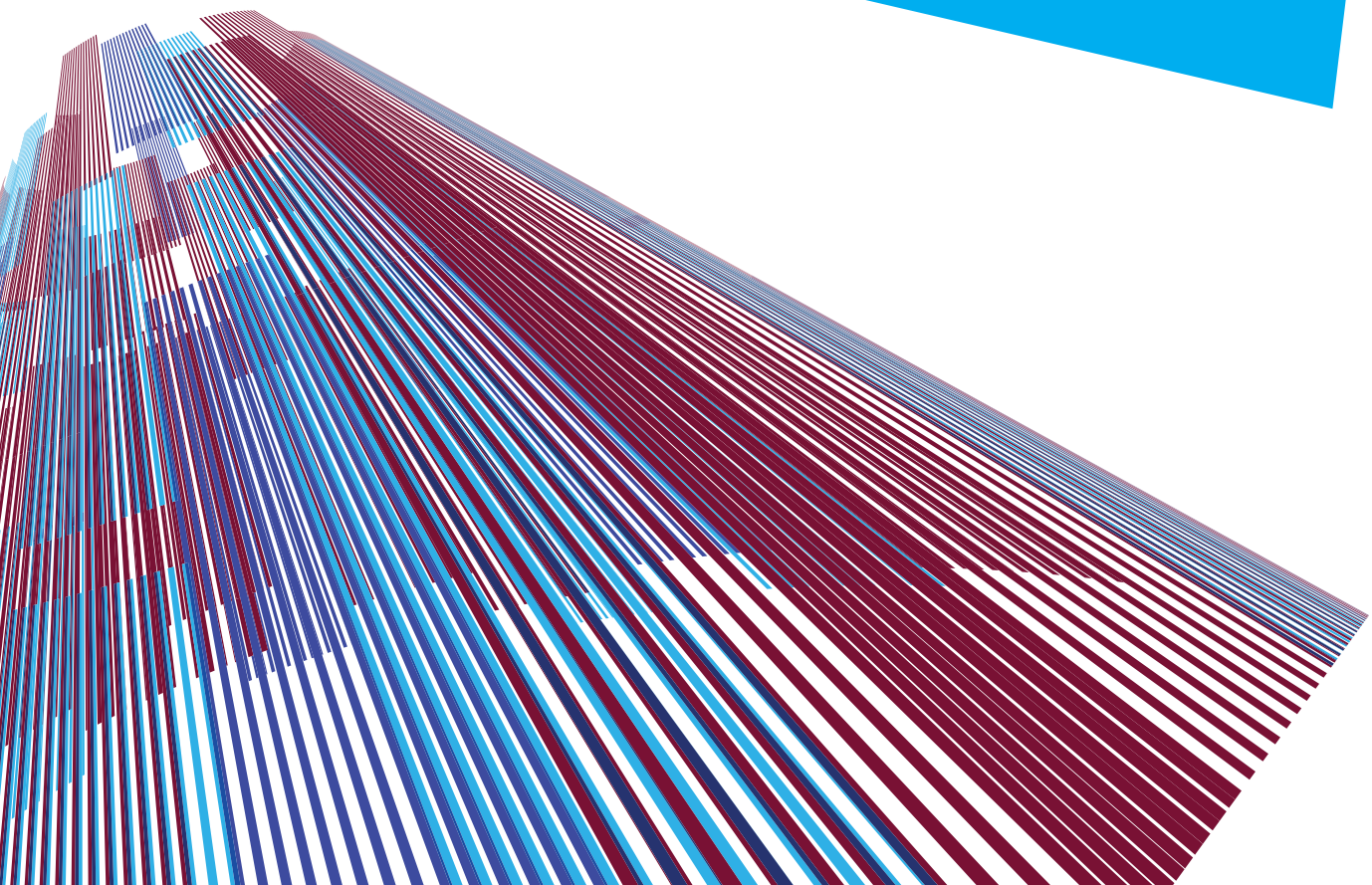
Scott Bell says there has been a "significant change in understanding" of what banks can do with data over the last 20 to 30 years. "Banks are learning to use their own data and other sources of data for compliance purposes and also for revenue generating purposes," he says. "Banks now have large data teams which aggregate information from multiple sources and use the data for multiple objectives."

Marc DaCosta, Co-Founder and Chairman of Enigma, says many banks have a lot of work to do to upgrade their data aggregation capacity. "The institutional legacy of banks like Barclays or Chase is that they may have dozens of different database systems in the background that describe and track all their accounts – one bank may have over 40 codes to refer to a checking account," he says. "So when there may be some suspicious activity in a financial network, it can take an investigator half a day to go through dozens of legal systems within the bank before they can check external public data to understand what is going on."

But he says it will be well worth the effort for a bank to improve their data quality and ensure it can be linked together: "If a bank or financial institution goes in and cleans up their internal data and links it together so that they have a comprehensive and 360 degree view of who they are doing business with, that unlocks all sorts of other opportunities: for compliance work, for figuring out how much credit they should give someone, and for onboarding credit card applications much more quickly." At the core of this system, a bank needs data that clearly defines the individual or company with which it is considering doing business, rather than a vague proprietary identifier.

Case study: Mastercard Track

Chapin Flynn, a senior vice president at Mastercard, says: “Over the last two and a half years we have been developing Track, which has used a core set of base business data including OpenCorporates data to pull these different files together and offer a single point of access for this information, rather than companies needing to use multiple fragmented solutions to cobble together their data solution. We have used the White Box model of sourcing data from known and verified entities – one example is global registrars to which companies can refer when asked to provide the provenance of their data. Track, which represents an important and innovative expansion of the core Mastercard model, is currently rolling out in Asia, the US and Europe and aims to simplify and enhance how companies around the world do business with each other.”



Chapter 4:
Can you "go clean"?

This white paper has demonstrated that a move from Black Box data to White Box is well underway. Chris Taggart estimates that in only five years, companies that rely on Black Box data risk becoming extinct. "Even fifteen years ago, when people started using data en masse and companies became more global, it was not viable to use Black Box data," he says. "But since then, the requirements we have for our data have increased dramatically yet Black Box data has remained more or less consistent and has not evolved with the needs of the market." The gulf between Black Box data and the data users' need is widening every year.

This gulf is also widening because White Box data gets better and more reliable over time because its wide range of users create a positive feedback loop. Millions of people use the OpenCorporates database to access company data, which ensures the data is more accurate and constantly refreshed. Restricted access and poorly-defined Black Box data is unable to receive any feedback on its data because its users are not informed of how it was compiled or how any assessments on the data were made.

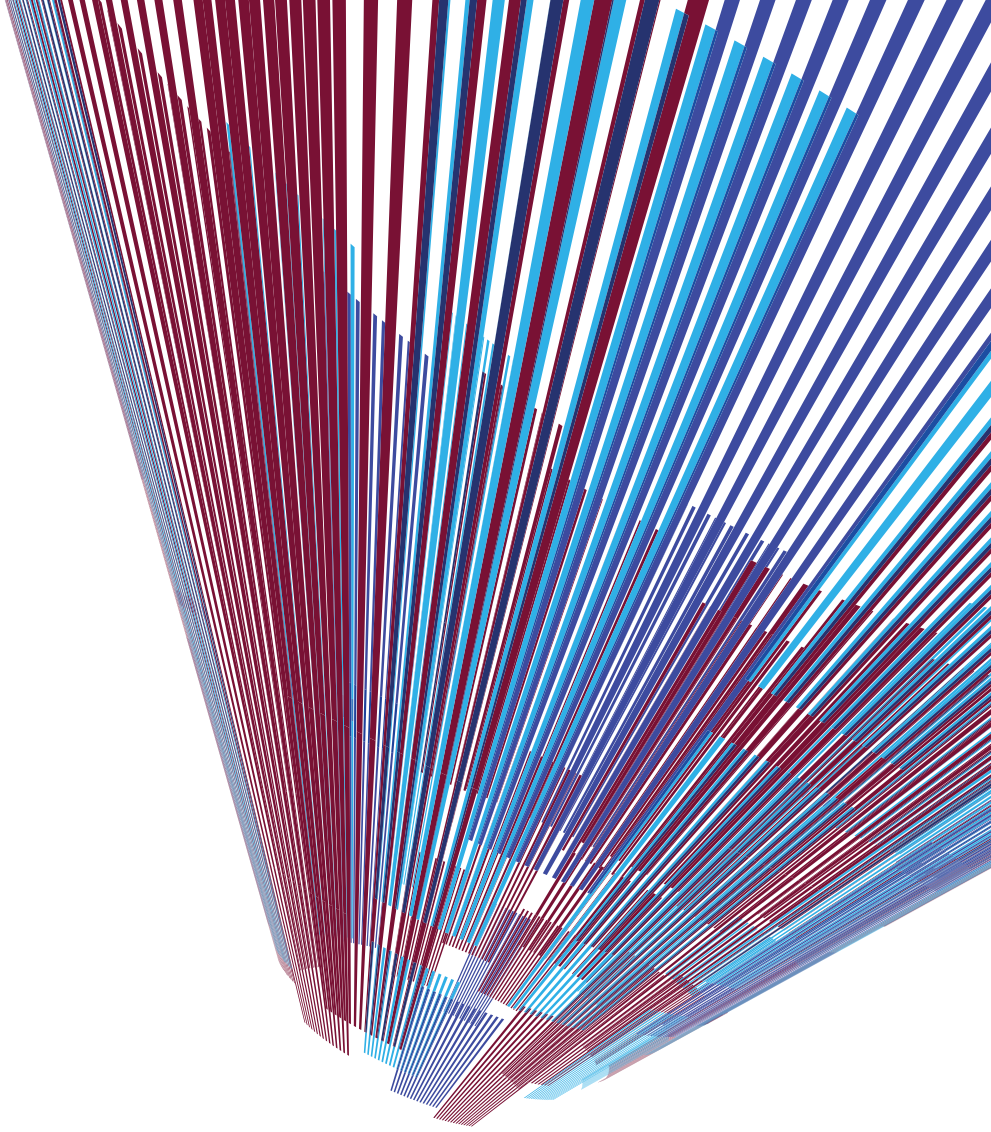
Today it has become essential that companies consider whether their data is White Box or Black Box, and take action accordingly. Making the change from using Black Box to White Box data will not be easy, especially for larger organisations. But a data overhaul does not need to happen overnight. Companies can gradually introduce White Box datasets into their operations. Each White Box dataset they adopt will not only remove some existing data quality problems, but highlight further problems the company was not aware of.

Companies can also make sure that any new products or databases they launch are based on a foundation of White Box data. This was the approach taken by Quantifind and Mastercard, who are quoted earlier in this white paper. "In all sectors, but particularly the tech world, we are seeing companies who have a ton of data starting from a clean slate," explains Mr Taggart. "If they are launching a new product, project or solution then they should try to base this on White Box data, because it guarantees quality and gives them the ability to iterate because the data is well-defined and boosted by a positive feedback loop."

Companies that continue to rely exclusively on Black Box data will continue to suffer from restrictions and data quality limitations. For as long as they use vague Black Box proprietary identifiers, they will remain unable to determine the true identity of a third party. "This approach will give your competitors a clear run and let them develop better products," says Mr Taggart. "Ultimately, this sort of stubbornness will put a company out of business."

Mr Taggart describes Black Box data as "crack cocaine". "It damages your health and distorts your view of the world, and there are two eventual outcomes – you die or you go clean. And the longer you stay on crack, the harder it is to get off."

There has never been a better time to start your own data revolution.



> In all sectors, but particularly the tech world, we are seeing companies who have a tonne of data starting from a clean slate

Chris Taggart, Founder of OpenCorporates

Is your data White Box?
OpenCorporates is a leading provider of White Box data on companies.

Visit **www.opencorporates.com** or to discuss how you could improve your data contact: sales@opencorporates.com

opencorporates