VisBIG

Proposta FCT C492000616_00085511

1 - Parametrização

Localização do Projeto (NUTS II):

Lisboa	100%

1.1 - Domínios Prioritários

DOMÍNIOS PRIORITÁRIOS

ENEI

Domínio Prioritário	Principal Área de Atuação	Fundamentação (Português)
Automóvel, aeronáutica e espaço	**TIC Aplicadas ao Automóvel, aeroespacial e espaço**	É cada vez mais comum, nos dias de hoje, que os automóveis tenham a capacidade de, continuamente ou em intervalos de tempo regulares, transmitir dados de telemetria ao fabricante ou outras partes interessadas. É o caso, por exemplo, dos automóveis da marca Tesla. Consiste, na realidade, numa grande quantidade de dados que têm de ser geridos em tempo real. Os resultados deste projeto fornecerão diretamente as ferramentas necessárias para a análise desse tipo de dados.
Energia	**Cidades Inteligentes**	As Cidades Inteligentes são estruturas vivas que respondem em tempo real às necessidades dos seus cidadãos e infraestruturas. Em última análise, essas cidades dependerão (mais ainda do que hoje em dia) da recolha e análise de dados em tempo real face a essas necessidades. Esta informação não se trata apenas de *Big Data* mas dados heterogéneos em natureza e contexto. Um analista que pretenda criar uma imagem holística clara do que se passa numa cidade necessita ferramentas que não só agreguem os dados correspondentes, mas também torne evidentes tendências e situações. O subsistema de análise automatizada que se planeia investigar neste projeto será diretamente aplicado num cenário com estas características.

Energia	**Eficiência energética e utilização final de energia**	Os medidores inteligentes são hoje uma realidade. Estão a ser rapidamente adotados em Portugal e em todo o mundo e desempenham um papel importante no consumo racional de energia, tanto por consumidores finais, para que possam ajustar o seu consumo, como pelas companhias energéticas que pretendam ter uma noção em tempo real das necessidades que têm de conseguir suportar. Estes conjuntos de dados, que podem e devem ser analisados geograficamente, permitirão um consumo energético mais eficiente. No entanto, para que os consumidores e analistas possam tirar partido dos mesmos, a utilização de técnicas de visualização adequadas é de extrema importância. Este projeto pretende desenvolver este tipo de técnicas.
TIC	**Internet das Coisas**	A Internet das Coisas tem vindo a tornar-se uma realidade em que mesmo os dispositivos mais reduzidos estão ligados em rede. Isto significa não só que estes podem ser controlados remotamente, mas também que os dados que recolhem ou criam podem ser transmitido ou armazenado para análise. Como resultado, surgem dois cenários de Big Data em *streaming*: analisar todas as *streams* de dispositivos da mesma família, e de todas as *streams* de dispositivos co-localizados (e.g., dentro de uma casa). Em ambos os casos são necessárias técnicas avançadas de visualização que permitam ligar com Big Data heterogéneos, tais como as que serão desenvolvidas no âmbito deste projeto.
TIC	**TIC aplicadas à Saúde**	Um dos domínios com maior potencial para gerar (e tirar partido de) Big Data em *streaming* é a saúde. O surgimento de dispositivos como *trackers* de atividade física e relógios inteligentes tornou ainda mais comum a monitorização constante de métricas fisiológicas através de *hardware* especializado. A análise desses dados, não só de um paciente, mas tendo em conta um maior número amostras, poderá ter um impacto significativo não só na saúde individual, mas também na saúde pública em geral (se for possível detetar determinadas tendências na população). A análise destes dados é, atualmente, feita por médicos e outros profissionais que não têm formação na área da análise de dados. Como tal, uma forma de lidar com a grande quantidade de dados existente é de extrema importância. A visualização e

		técnicas de análise a criar neste projeto serão capazes de dar resposta a estes cenários.
Transportes, mobilidade e logística	**Transportes e logística Inteligentes**	PuOs transportes públicos encontram-se em evolução. Este é um dos domínios onde Big Data em *streaming* já existe. Cada autocarro transmite informação, de forma contínua, sobre a sua localização, portas a abrir e fechar, passageiros a bordo (através das máquinas de validação de bilhetes), etc. Hoje em dia, estes dados são subutilizados, principalmente devido à falta de ferramentas de análise de transportes que permitam tirar partido destes dados em tempo real. Este tipo de ferramentas serão disponibilizadas. Algo semelhante se passa no contexto dos transportes e logísticas, em que são continuamente capturados dados telemétricos dos veículos e informação relacionada. Em ambos os casos, decisões atempadas poderão implicar reduções significativas de custos e melhoria da qualidade do serviço.

EREI Lisboa

Domínio de Especialização	Domínio Prioritário	Principal Área de Atuação	Fundamentação (Português)
Investigação, Tecnologias e Serviços de Saúde	Investigação	Promoção de projetos com alinhamento estratégico entre academia e empresas	O consórcio concorrente a este projeto inclui não só academia (universidade e instituto de investigação) como uma empresa líder de mercado na área da Análise dos Dados. A participação conjunta nesses moldes não só possibilita este tipo de investigação graças à visão abrangente e de produto da empresa, como pressupõe um veículo privilegiado para transferência de conhecimento entre academia e o tecido empresarial.
Investigação, Tecnologias e Serviços de Saúde	Investigação	Fomentar a harmonização entre a investigação fundamental e a investigação aplicada e promover o diálogo entre academia e empresas	O consórcio concorrente a este projeto inclui não só academia (universidade e instituto de investigação) como uma empresa líder de mercado na área da Análise dos Dados. Este projeto versa um tema que é neste momento importante tanto em termos de investigação como de produto, nessa área.

			Estrategicamente esta colaboração permitirá viabilizar a exploração de ambas vertentes, tomando partido da inovação científica, mais abstrata e teórica, produzida pelos investigadores e pelo conhecimento técnico aplicado e de produto da empresa..
Conhecimento, Prospeção e Valorização de Recursos Marinhos	Conhecimento e Transformação de Conhecimento	Investigação em áreas de interesse para a indústria	A Big Data e Streaming Big Data em particular surgem cada vez mais como relevantes nas mais diversas indústrias relacionadas com o mar. Onde quer que haja sensores é produzida Big Data. No entanto, as ferramentas de recolha e análise tradicionais estão vocacionadas para dados estáticos e de pequena dimensão. Iremos produzir soluções de ampla aplicabilidade nos vários domínios relevantes para a indústria marítima (pescas, meteorologia, etc.) para colmatar a grave falta de ferramentas acessíveis para lidar com Streaming Big Data.
Mobilidade e Transportes	Apoiar o desenvolvimento e teste de soluções inovadoras	Promoção de soluções inovadoras de mobilidade e sustentabilidade	Tam como referido anteriormente num dos Domínios do ENEI, lidar com Big Data em tempo real gerada por sistemas de transportes permitirá a tomada de decisões atempadas que podem levar a novos modelos de transporte. Coisas como transportes flexíveis, com rotas ajustadas às necessidades a cada momento, alocação de material circulante e preços diferenciados são possíveis apenas com base numa análise sólida, informada, em tempo real.
Mobilidade e Transportes	Apoiar o desenvolvimento e teste de soluções inovadoras	Disponibilização de ferramentas avançadas, alimentada em tempo real...	Esta Área de Atuação é exatamente o que propomos conseguir com o projeto: o desenvolvimento de ferramentas para análise de Big Data produzida em tempo real.

2 - Declarações

3 - Caracterização da Instituição Proponente

4 - Contratos Públicos
5 - Demonstrações de Resultados
6 - Balanços
7 - Taxa de Incentivo das Entidades NE SI&I

8. Lista de Instituições Participantes

Código	NIF	Designação Social da Entidade
1	504547593	INESC-ID - Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa
2	508528283	Webdetails - Consulting Unipessoal, Lda.
3	509830072	IST-ID, ASSOCIAÇÃO DO INSTITUTO SUPERIOR TÉCNICO PARA A INVESTIGAÇÃO E O DESENVOLVIMENTO

Página 8.1 - Copromoção - Caracterização da Instituição
Página 8.2 - Copromoção - Caracterização da Instituição (cont.)
Página 8.3 - Copromoção - Contratos Públicos
Página 8.4 - Copromoção - Demonstração de Resultados
Página 8.5 - Copromoção - Balanços
Página 8.6 - Copromoção - Taxa de Incentivo das Entidades NE SI&I

9. Caracterização do Projecto

Descrição e Tipologia do Projecto

Acrónimo: VisBig
Título do projecto (português) (250): Visualização em Tempo Real de Big Data em Streaming
Título do projecto (inglês) (250): Real-Time Visualization of Streaming Big Data

Breve descrição do projecto (1000):
Big Data, and especially Streaming Big Data (SBD) is often analysed by automated processes, not amenable to real-time decisions by human analysis. There, timely decision making must involve the quick and easy understanding of the data. Making that possible is the purview of Visual Analytics, but existing solutions are tailored for mostly static (comparatively) small datasets and will not scale well for SBD. We will fill that gap by researching solutions that separate the wheat from the chaff and highlight real-time relevant changes in the data streams. We will develop novel visualization techniques capable of displaying important features in SBD. These will be supported by machine-learning for the identification of relevant changes in the streams, to be highlighted in the visualizations. Underlying it we will study how best to cope with

SBD in terms of collection, filtering and (where necessary) storage, in ways that make the other components feasible.

Domínio Científico Principal: Exact Sciences
Área Científica Principal: Computer and Information Sciences
Subárea da Área Cientifica Principal: Information Sciences
Dominio Cientifico Secundário: Exact Sciences
Área Científica Secundária: Computer and Information Sciences
Subárea da Area Cientifica Secundaria: Informatics

Palavra-chave (1): processamento de fluxos de dados
Palavra-chave / inglês (1): stream processing

Palavra-chave (2): Visualização de Informação
Palavra-chave / inglês (2): Information Visualization

Palavra-chave (3): Deteção de mudança
Palavra-chave / inglês (3): Change detection

Palavra-chave (4): data analytics / análise não supervisionada de dados
Palavra-chave / inglês (4): unsupervised data analytics

Investigador Responsável (IR) do projecto:

Chave de Associação IR: J013170P7FH E-mail: daniel.goncalves@inesc-id.pt
Nome: Daniel Jorge Viegas Gonçalves
Função do beneficiário: Researcher
Vínculo contratual com a Instituição Proponente: Affiliated researcher to the beneficiary institution

Co-investigador Responsável (co-IR) do projecto:

Chave de Associação co-IR:J526339861TS E-mail: sandra.gama@tecnico.ulisboa.pt
Nome: Sandra Pereira Gama
Função do beneficiário: Researcher.
Vínculo contratual com a Instituição Proponente: None

Name	FCT ~~Username~~ Public Key	Nationality	NIF	Qualification Level	ORCID ID
Muhammad Amin Khan	J678976Jca7A	Paquistanês	283880325	8	0000-0002-8103-4944
Cláudia Antunes	J0282074F59	Portuguesa	204625750	8	0000-0002-9467-2515
Pedro Vale	J687274wc91m	Portuguesa	199464880	7	
Daniel Gonçalves	J013170P7FH	Portuguesa	194888002	8	0000-0002-5121-6296
Sandra Gama	J526339861TS	Portuguesa	214727610	8	0000-0002-9679-7004

Actividade Económica do Projecto

CAE	Designação	%
80300	Atividades de Investigação	80
62010	Atividades de Programação Informática	20

Declaração de Ética

Questões de ética:
Fundamentação **1000**: No ethical questions are raised by this project.

Calendarização e Investimento

Data de Inicio: 1 Jan 2018
Data de Fim: 31 Dez 2020
N meses: 36 meses
Investimento Elegível:
Investimento Total:

10. Abstracts

Abstract(Português) **3000**

Grande parte da investigação em *Big Data* tem-se focado no processamento automático (machine-learning, etc.), levando a lacunas, de conhecimento e tecnológicas, relativamente a ferramentas e métodos para a sua utilização em cenários analíticos por responsáveis pela tomada de decisões. De facto, embora várias técnicas de Análise Visual possam ser usadas em conjuntos de dados comuns, a sua aplicabilidade a grandes volumes de dados é discutível. A adicionar a este problema, *Big Data* em *Streaming* (SBD) representa um conjunto de outros desafios: uma solução adequada deve lidar com os dados e evidenciar tendências relevantes e alterações atempadamente, de forma a potenciar a tomada de decisões.

A escala de SBD, com o volume e velocidade inerentes, coloca novos desafios. Surge a necessidade de técnicas de visualização em tempo real, flexíveis, interativas e dinâmicas, colmatando as limitações das abordagens existentes. No projeto VisBig vamos investigar e desenvolver soluções para a Análise Visual de SBD.

Serão investigadas técnicas e métodos que permitirão a visualização clara e compreensível de SBD, o que incluirá: (a) processamento eficiente e consumo de dados em streaming; (b) deteção automática de alterações relevantes nos dados, destacando entidades que sugiram uma análise detalhada; (c) seleção e desenvolvimento de idiomas de visualização adequados a SBD; (d) uso apropriado de transformações de idiomas de forma a permitir visualizar alterações em tempo real.

Ao nível mais fundamental, o sistema deve lidar com os dados gerados continuamente. Depois disso, os pontos de interesse nos dados serão detetados e evidenciados, o que poderá acontecer quando a sua natureza é alterada, possivelmente devido a eventos externos. Serão aplicados algoritmos de *machine-learning* para operacionalizar esta vertente e estudado o impacto do uso de conhecimento específico de domínio no estudo de soluções mais robustas. Depois de identificar alterações relevantes na *stream* de dados, permanece a questão de como visualizar os mesmos. Num cenário típico de Inteligência Empresarial, o conjunto de dados e as análises relevantes são conhecidos antecipadamente, permitindo a criação de *dashboards* especialmente concebidos tendo em conta os dados e as questões para as quais se pretendem respostas. Com SBD, especialmente em domínios complexos com rápidas mudanças, podem surgir situações em que seja relevante disponibilizar diferentes idiomas de visualização, surgindo a questão de como adaptar (semi-) automaticamente os idiomas aos dados em cada momento do tempo, evidenciando facetas relevantes para a análise. Como fazer a transição entre diferentes estados da visualização sem perder o contexto é uma questão aberta a que este projeto permitirá responder.

Os resultados do BigVis, tanto científicos quanto como um produto, terão um impacto nas comunidades de Análise Visual e de Inteligência Empresarial, em que a necessidade de soluções para análise de SBD é cada vez mais crucial.

Abstract (English) **3000**

Most research on Big Data has focused on its automated processing (machine-learning, etc.), leading to a knowledge and technological gap in tools and methods for its use in analytical scenarios, by human decision makers. Indeed, while many well established techniques from Visual Analytics can be used for regular datasets, their applicability to large volumes of data is questionable. Compounding on this problem, Streaming Big Data (SBD), constantly updating in real time, poses additional challenges: an adequate solution must not only cope with the incoming barrage of data, but also be able to highlight relevant trends and changes in the data stream in a timely fashion, in such a way as to allow relevant decisions to be taken.

The scale of SBD, with its volume and velocity, poses new challenges and the need for real-time flexible, interactive, and dynamic visualization techniques, beyond the limitations of existing approaches. In VisBig we will research and develop solutions for the problem of SBD Visual Analytics.

We will investigate techniques and methods that allow the clear and understandable visualization of SBD. This will include: (a) efficient processing and consumption of streaming data; (b) automated detection of relevant changes in the data stream, highlighting entities that merit a detailed analysis; (c) selection and development of visualization idioms adequate for SBD; (d) appropriate use of idiom transformations to allow for the real-time visualization changes in the stream.

At the most fundamental level the system must cope with the data that keeps pouring in. Then, we will detect of points of interest in the data. An uninteresting stream can suddenly become relevant when the data therein changes in nature, due perhaps to some external event. Those changes must be highlighted to the analyst lest they go unnoticed. We will employ machine-learning to do this and address how the use of domain-specific knowledge can lead to more robust solutions. Having identified interesting changes to the data stream, the question of how to visualize them remains. In a typical Business Intelligence scenario, the dataset and relevant analyses are known beforehand, allowing the creation of custom-tailor dashboards matching the data and the questions analysts have. With SBD, especially in rapidly evolving, complex, domains, different situations may arise where different visualization idioms may be more relevant, raising the question of how to (semi-)automatically adapt visualization idioms depending on the data and its properties at each moment, to highlight facets relevant for analysis. How to transition between different states of the visualization while maintaining the context of the analysis is also an open question that this project will solve.

VisBig's results, both scientific and as a product, will have an impact on the Visual Analytics and Business Intelligence communities, where the need for solutions for SBD Analytics is increasingly dire.

11. Abstracts for publication

Sumário para publicação (Português) **3000**

Grande parte da investigação em *Big Data* tem-se focado no processamento automático (machine-learning, etc.), levando a lacunas, de conhecimento e tecnológicas, relativamente a ferramentas e métodos para a sua utilização em cenários analíticos por responsáveis pela tomada de decisões. De facto, embora várias técnicas de Análise Visual possam ser usadas em conjuntos de dados comuns, a sua aplicabilidade a grandes volumes de dados é discutível. A adicionar a este problema, *Big Data* em *Streaming* (SBD) representa um conjunto de outros desafios: uma solução adequada deve lidar com os dados e evidenciar tendências relevantes e alterações atempadamente, de forma a potenciar a tomada de decisões.

A escala de SBD, com o volume e velocidade inerentes, coloca novos desafios. Surge a necessidade de técnicas de visualização em tempo real, flexíveis, interativas e dinâmicas, colmatando as limitações das abordagens existentes. No projeto VisBig vamos investigar e desenvolver soluções para a Análise Visual de SBD.

Serão investigadas técnicas e métodos que permitirão a visualização clara e compreensível de SBD, o que incluirá: (a) processamento eficiente e consumo de dados em streaming; (b) deteção automática de alterações relevantes nos dados, destacando entidades que sugiram uma análise detalhada; (c) seleção e desenvolvimento de idiomas de visualização adequados a SBD; (d) uso apropriado de transformações de idiomas de forma a permitir visualizar alterações em tempo real.

Ao nível mais fundamental, o sistema deve lidar com os dados gerados continuamente. Depois disso, os pontos de interesse nos dados serão detetados e evidenciados, o que poderá acontecer quando a sua natureza é alterada, possivelmente devido a eventos externos. Serão aplicados algoritmos de *machine-learning* para operacionalizar esta vertente e estudado o impacto do uso de conhecimento específico de domínio no estudo de soluções mais robustas. Depois de identificar alterações relevantes na *stream* de dados, permanece a questão de como visualizar os mesmos. Num cenário típico de Inteligência Empresarial, o conjunto de dados e as análises relevantes são conhecidos antecipadamente, permitindo a criação de *dashboards* especialmente concebidos tendo em conta os dados e as questões para as quais se pretendem respostas. Com SBD, especialmente em domínios complexos com rápidas mudanças, podem

surgir situações em que seja relevante disponibilizar diferentes idiomas de visualização, surgindo a questão de como adaptar (semi-) automaticamente os idiomas aos dados em cada momento do tempo, evidenciando facetas relevantes para a análise. Como fazer a transição entre diferentes estados da visualização sem perder o contexto é uma questão aberta a que este projeto permitirá responder.

Os resultados do BigVis, tanto científicos quanto como um produto, terão um impacto nas comunidades de Análise Visual e de Inteligência Empresarial, em que a necessidade de soluções para análise de SBD é cada vez mais crucial.

Sumário para publicação (Inglês) **3000**

Most research on Big Data has focused on its automated processing (machine-learning, etc.), leading to a knowledge and technological gap in tools and methods for its use in analytical scenarios, by human decision makers. Indeed, while many well established techniques from Visual Analytics can be used for regular datasets, their applicability to large volumes of data is questionable. Compounding on this problem, Streaming Big Data (SBD), constantly updating in real time, poses additional challenges: an adequate solution must not only cope with the incoming barrage of data, but also be able to highlight relevant trends and changes in the data stream in a timely fashion, in such a way as to allow relevant decisions to be taken.

The scale of SBD, with its volume and velocity, poses new challenges and the need for real-time flexible, interactive, and dynamic visualization techniques, beyond the limitations of existing approaches. In VisBig we will research and develop solutions for the problem of SBD Visual Analytics.

We will investigate techniques and methods that allow the clear and understandable visualization of SBD. This will include: (a) efficient processing and consumption of streaming data; (b) automated detection of relevant changes in the data stream, highlighting entities that merit a detailed analysis; (c) selection and development of visualization idioms adequate for SBD; (d) appropriate use of idiom transformations to allow for the real-time visualization changes in the stream.

At the most fundamental level the system must cope with the data that keeps pouring in. Then, we will detect of points of interest in the data. An uninteresting stream can suddenly become relevant when the data therein changes in nature, due perhaps to some external event. Those changes must be highlighted to the analyst lest they go unnoticed. We will employ machine-learning to do this and address how the use of domain-specific knowledge can lead to more robust solutions. Having identified interesting changes to the data stream, the question of how to visualize them remains. In a typical Business Intelligence scenario, the dataset and relevant analyses are known beforehand, allowing the creation of custom-tailor dashboards

matching the data and the questions analysts have. With SBD, especially in rapidly evolving, complex, domains, different situations may arise where different visualization idioms may be more relevant, raising the question of how to (semi-)automatically adapt visualization idioms depending on the data and its properties at each moment, to highlight facets relevant for analysis. How to transition between different states of the visualization while maintaining the context of the analysis is also an open question that this project will solve.

VisBig's results, both scientific and as a product, will have an impact on the Visual Analytics and Business Intelligence communities, where the need for solutions for SBD Analytics is increasingly dire.

12. Technical Description

State of the Art **6000**

Nowadays, there is wide access to a wealth of information. While information generation potential continues to grow, storage capacity had also increased [Eaton2011]. Big Data consists of massive datasets with large, varied and complex structure with the difficulties of storing, analysing and visualizing for further processes or results [Sagiroglu2013]. Alas, getting insights out of such data has become increasingly difficult; many datasets are so large that not only are they virtually impossible to read, but visually representing them would lead to incomprehensible clutter.
Automated processing of Big Data has been the focus of current research in machine-learning and data-mining. However, the use of such data in analytical, decision-making scenarios leads to the need for higher knowledge, often difficult to obtain due to size and complexity. Although Information Visualization is an excellent means to display large quantities of data while alleviating cognitive load associated with interpretation [Munzner2014], while many well-established techniques from Visual Analytics can be used for regular datasets, their applicability to large volumes of data is not straightforward. Traditionally, complexity is dealt with by reducing items and attributes, often through filtering or aggregation [Munzner2014].

Several authors have approached Big Data visualization using such mechanisms, such as data reduction with binned aggregation and sampling, as well as interactive querying and parallel query processing [Liu2013], or exploratory browsing through dynamic prefetching of portions of data for interactive analysis of large datasets [BCS16]. An iteration-level interactive visualization for Big Data has also been proposed to allow the user to view intermediate results and interact with such results in real time through data reduction by the means of clustering algorithms [Choo2013].

There are several data analytics and visualization products in the industry, including Pentaho, Tableau, Talend, MicroStrategy, Informatica, Elasticsearch's Kibana, along with Business

Intelligence (BI) products from Microsoft, Oracle, SAS, IBM. However, their majority are geared towards traditional data warehousing and Online Analytical Processing (OLAP), and run into significant limitations at the scale of Big Data. Most are batch processing systems and lack real-time analytics. To handle the volume of Big Data in a timely manner, one approach, MapD [Mat17;Tal13], relies on massively parallel in-memory databases and GPUs for quick interactive visualizations. Another approach is to generate approximate visualizations, sacrificing accuracy for speed, but preserving visual properties of the data features [Kim15]. Other works have explored visualization recommendation engines to assist in fast visual analysis [Var15] or related problems like preserving privacy when visually exploring datasets [HRM16], interactive data-driven visual query interfaces [BCD16], and visualization-oriented data aggregation and dimensionality reduction [JJH14].

Nevertheless, Streaming Big Data (SBD), which is constantly updating in real time, leads to additional challenges. The volume, heterogeneity and time restrictions associated with SBD lead to the need for techniques that go beyond existing approaches. A meaningful solution must be able to deal with the incoming data, often large and unprocessed, and to make important patterns immediately perceivable, highlighting relevant changes in the data stream. It must also provide a real-time response to allow proper data interpretation, relevant for decision making. Hence, effective mechanisms must be provided to overcome such challenges.

The field of data science has reached a considerable maturity level. Supervised techniques developed under machine learning research have plenty of successful applications, but the last decade had also provided effective and efficient methods to deal with data streams [LSB14, GBEB17].
Nevertheless, the success of unsupervised techniques over data streams, where the goal is to gain some insight into unlabelled data, is much more modest, despite the existence of a significant number of algorithms, both for clustering [GLA16] and pattern mining [NASNG14]. Some of this lack of success derives from the huge amount of patterns discovered, and research continues to look for effective ways to reduce the number of patterns to present. Several approaches have been pursued, such as finding top-k patterns [Chen14] and new formulations like crucial patterns [DZ16] or creating condensed representations [SA14].

An important issue when dealing with data streams, either through supervised or unsupervised techniques, is change detection, usually known as concept-drift. Despite its importance, research has been centred on the identification of changes over the statistical distribution of data ([HSRC14], [SK17]) and little have been done on understanding the changes happening on the discovered patterns along time.

In fact, one of the greatest challenges in SBD, stemming from the detection of interesting challenges in the data stream, is to visualize such changes, allowing a meaningful analysis. The combination of re-encoding and reconfiguration mechanisms[Yi2007] may cope with the existing challenges by allowing the detection of points of interest in the changing data. Additionally,

animated transitions, which smoothly transition from one state to another, may help users maintain the context [Munzner2014].

Even though such mechanisms have been profusely used [Stolte2002, Spenke2000, Heer2005], a meaningful, real-time visualization for SBD had not yet been created which takes advantage of such visualization techniques for highlighting relevant changes. Hence, a SBD visualization that operates in real-time, using meaningful techniques to allow the user to keep track of changes in the stream, may be an effective tool for analysis, interpretation and decision-making.

Objectives / Research Plan and Methods **10000**

Nowadays, big data is all around us. It is produced by automated sensors, deployed in our homes (smart meters) streets (transit and traffic), industries (sensor monitoring of the production chain) and ourselves (fitness and health trackers). Even if individual sources are not “big” data, in many domains the proliferation of sensors leads to large overall amounts. This trend is accelerating, as more and more cheaper sensors and connected devices (“Internet-of-Things”) become a reality. In particular, many big data sources, being automated, continuously produce a stream of data. We are facing, thus, Streaming Big Data (SBD). Here, not only must we be concerned with its large amount, but also with the fact that it is continuously changing, with new data ever present in the stream. Coping with such a deluge poses new challenges.

Data in itself is not easy to understand, and there are several ways to try to make sense of it. From statistical analyses to machine learning, techniques abound. Once such approach is the use of Information Visualization (Infovis). Indeed, through visualization we are able to easily perceive trends, patterns and salient features in datasets. This is the basis for the fields of Visual Analytics (VA), where a visually supported analysis is used to gain relevant insights. Business Intelligence also traditionally uses VA and Infovis, in the guise of Information Dashboards, to help a human analyst make decisions.

Unfortunately, traditional Infovis and VA techniques and solutions are usually tailored for datasets of small dimensions. As an extreme but easy to understand example, consider the humble barchart or scatterplot: while it is feasible to use them to display dozens or a few hundred items, they will not scale for larger orders of magnitude. What is more, most visualization techniques and idioms assume a static dataset, known beforehand, and are unable to properly cope with data streams. Again using a barchart as an example, even if (for low update rates) we could continuously replace the bars with new ones for new datum, the overall trends and history are lost. This loss of context hinders appropriate decision-making.

The objective of this project is, thus, to research and develop new Information Visualization techniques that are able to deal with Streaming Big Data, in ways that allow timely decision-making by including not only the latest data but also historical patterns and trends, while at the same time automatically highlighting possibly relevant changes in the stream.

In order to accomplish this, we will support the techniques on novel machine learning pattern discovery techniques. While most machine learning, like InfoVis, are based on pre-existing datasets, of comparatively small sizes, some efficient algorithms for mining patterns over data streams (for example the FP-stream algorithm [Gia+03]) exist. However, applying them to SBD will quickly render solutions unusable, due to the sheer number of patterns detected and stored. We will strive to develop better management, reducing their number without losing information, by compacting sets equivalent patterns and abstracting similar ones. This will need to be done within a reasonable algorithmic complexity to use in real-time. Also we will develop change detection algorithms that can spot differences in patterns (and, thus on the nature of the data in the stream) hinting at important analysis sites.

Whenever such change is detected, we will make it visibly salient. In its simplest form, this can be accomplished by a change in color or other visual property. We foresee, however, that some changes may require a more in-depth modification. Indeed different visualization idioms are more well suited to display certain types of information and patterns. It makes sense, thus, to reconfigure the visualization, transitioning to a different idiom as the data requires. This poses a risk, as it may lead to a loss of context, causing the analyst to become lost in his analysis. Thus, we will study how better to do this minimizing that risk. We will do this by researching which idiom transitions are visually effective and non-distracting (it may be possible to transition from a barchart to a line chart in meaningful ways but not to a scatterplot, for instance). We will follow a User-Centered Design methodology, based on extensive user testing, to assess the cognitive demands of each alternative, in order to select the most adequate.

Overall, we will go beyond the state of the art in two areas: Information Visualization and Pattern Discovery, leaping beyond existing techniques and algorithms applicable mostly to “traditional” datasets, and developing new ones that can be used with Streaming Big Data. Partial solutions exist in the literature, but not ones that represent and take into account both the latest data and historical trends synergistically combining InfoVis and machine learning, to make the SBD amenable not only to automated but also human analysis. Also, most visualizations, once designed, are static in terms of the techniques they employ. We will create guidelines for the automated selection and effecting of transitioning of visualization idioms, using the most adequate for the nature of the data in the stream at all times.

To accomplish our goal, we will follow a three-tiered approach (see figure attached to this proposal). The bottommost layer will be a Data Access Layer (DAL), to provide utility functions for facilitating access, consumption, integration and analysis of streaming data from a variety of data sources. The goal is to provide a streamlined and convenient data access interface, and sophisticated and powerful data manipulation functions.building on existing stream processing

frameworks, and customizing them for the primary goal of BigViz platform: to provide real-time interactive intelligent data visualization. We will build on top of existing open source stream processing tools such as Apache Flink or Twitter's Heron. The second tier will be the Analysis Engine Layer (AEL), where data from the bottommost tier will mined for patterns, as described above. The third layer is the InfoVis layer (IVL), where data provided by the DAL will be visually rendered in such a way as to make both historical and recent data equally visible, in an integrated way. Furthermore, emphasis will be placed on highlighting change, as detected by the AEL.

The project is divided into five Activities. The first concerns management and is not relevant in this discussion. Activities 2-4, however, map directly onto the three layers described above. At a first stage, for all layers, preliminary work needs to be done until the first results are relevant. This distribution into Activities will allow it to proceed independently and in parallel for a more efficient collaboration. Once the first results from each Activity are in place, they will start to be used by the other activities. Our management structure, that foresees regular meetings and shared requirements, knowledge and code repositories, ensures all results are interoperable and compatible.

Here is where the complementary nature of the project partners and their knowledge and skills is made apparent: Webdetails, led by Pedro Vale, has an indisputable track record on the development of data wrangling, cleaning, access and manipulation tools, with a tradition of embracing collaborative and open-source projects. They will lead Activity 2, the DAL. The IST-ID team is led by Cláudia Antunes whose area of specialization is temporal pattern detection in streams. Daniel Gonçalves (the PI) leads INESC-ID’s team, helped by the Co-PI (Sandra Gama), both of which have a track record in InfoVis. As such, we are in the presence of the exact mix of teams to lead this project to fruition.

Activities 2 to 4 will focus on the research of new techniques in each of their individual areas. While there will be, as mentioned, extensive collaboration between the tasks, in Activity 5 we will do an explicit integration of the tools and solutions developed in the other tasks, creating a proof-of-concept system, for a test (domain to be decided). This will serve both as a way to fine tune any inconsistencies that may still exist and as validation of the project’s overall goals.

In Activity 5, we will use as a basis Webdetails’ (Pentahos’) tools. This will have the added value of ensuring the compatibility of our solutions with those tools, which will be an important factor towards the lasting impact of this project. Indeed, while we intend that the resulting tools will be made open-source and freely available, making them compatible with a toolchain in use in thousands of institutions worldwide will ensure they will remain in use beyond the scope of this project. What is more, given the increasing need for VA tools for big data, the timely availability of VisBig’s results have the potential to represent a lasting economic impact, by saving in-house development costs and providing a ready-made, adaptable solution. Indeed, we are in line with H2020 Societal Challenges and hope to play our part in solving them. The quest for Safe, Non-Polluting, Clean Energy, in particular reducing energy consumption is to be aided by a

more efficient consumption based on real-time knowledge of available energy from different sources and the awareness of real-time consumption. Similarly this is the basis for a next-generation intelligent power grid. Real-time analysis of data, made possible by VisBig, can help make this a reality. Likewise, better data about public transit usage and the profiles of those that do, another source of big data, will benefit from our results, as will clearly, relating to the Environmental challenges, the analysis of data from extensive or global monitoring systems.

Management Structure Description **3000**

The Principal Investigator (PI), from INESC-ID, will manage and coordinate the whole project execution, assisted by the Co-PI. The institution named INESC-ID, as promoter, will financially manage the project.

Each of the other partners will have a privileged senior representative that will be actively involved in project management, and serve as the contact point for the transmission of strategic decisions to the rest of their team. The PI will coordinate all management activity, and he will represent the team (academic and industrial partners included) in all interactions with FCT.

A kick-off meeting in the first days of the project will ensure a shared mental model and clearly define the boundaries and responsibilities for the first six months. With three partners, a risk we will explicitly address is that of ensuring there is no drift in the development of each activity that may put into question their integration at a later stage. To that end, we will set-up online collaboration tools for continued, transparent, collaboration. This will include not only tools such as Trello or Slack for communication, but a shared private Github repository, for synchronizing development and sharing of actual code, datasets and results. We will define regular reporting cycles of one month, at the end of which a meeting between the three partners' representatives (plus any relevant team members) will be held. Physical project meetings with all team members from all partners will be held every six months.

We will set-up an internal review process for all deliverables of the project. Each deliverable will be need to be be ready before the actual deadline/milestone so that at least one member from each partner not involved in its creation can review it, offering comments and suggestions for the final version. In cases where additional input is felt as necessary, we will engage the help of domain experts for further reviewing.

For administrative work, such as project tracking, the PI and the senior members will mostly rely on the administrative staff at their respective institutions. INESC-ID offers its researchers exceptional conditions to perform research, by adopting a decentralized structure whereby the management team oversees the top-down running of the organization from a hierarchical

viewpoint, while allowing researchers to aggregate into teams and groups on a per project basis. Moreover, the institutional management structure is composed by a highly qualified science interface team (projects support office and accountancy services), supporting not only the proposals, but also the accomplishment, reporting and management of the projects, besides ensuring organization of events and dissemination activities.

The prototypes developed in each activity during the course of the project, will be made available as open-source software. We will be using publicly available source-code hosting platforms (e.g., GitHub) to disseminate the prototypes.

13. Activity List

N	Name	Classification	Start	End	n.months
1	Project Management [inesc-id]	Investigação Industrial	1	36	36
2	Data Access [webdetails]	Investigação Industrial	1	24	24
3	Streaming Analysis Engine [ist]	Investigação Industrial	1	30	30
4	Adaptive Big Data Visualization [inesc-id]	Investigação Industrial	1	30	30
5	Integration and Validation [webdetails]	Investigação Industrial	19	36	18

Activity 1 - Project Management

Activity Description **3000**

The aim of this task is to deal with all management-related activities within the project, including the coordination of the three research teams, fulfilment of the goals of the project, within the

time and budget constraints; project planning and scheduling; logistic and contractual tasks, the production of (non-technical) project reports for the duration of the project, organization of kick-off- and regular meetings of the project; as well as the organization of two project workshops.

The project will start with the kick-off meeting, where the project activities plan will be presented and clarified as per the application proposal. This includes the definition of the deliverables for each single project milestone, their acceptance and quality criteria. It will also ensure (at this point and throughout the project) the project and its outcomes will not deviate from the Priority Actions and Societal Challenges the project is relevant to.

This task will also ensure that project team members comply with the timely delivery of results, assuring quality of work executed with the enforcement of an internal review mechanism for all deliverables.

Two yearly progress reports and one final report, will produced within this task in months M12, M24 and M36, respectively. It is also the purview of this task to create and maintain the project website, set-up online collaboration tools (github, slack, etc.), and the dissemination of project results in social media.

The organization of two Workshops about the research performed in the project will be undertaken in this task. The first, to take place at M18, will be a technical/scientific workshop. Project participants will invite members of the research community to present their provisional results. This will work both as a dissemination action and as a way to validate the work and collect comments and ideas useful for the remainder of the project. The second workshop, at M36, will gather not only researchers but also developers and potential users of the framework resulting from the project and present it to them. This workshop has the goal of bridging the gap between the project's team and the community, making the project's results available to a wide range of users, ensuring its continued use and usefulness..

This task will be the main responsibility of the PI and Co-PI, from INESC-ID, with the participation of one senior representative from each partner, and span the entire duration of the project.

Milestones **max. 6**

Data	Designação	Descrição
31-01-2018	Online Collaborative Tools	Set-up of the necessary online collaborative tools
28-02-2018	Project Website	Website for the dissemination

		of project results
31-12-2018	Year 1 Report	Report on the technical and financial status of the project
31-12-2019	Year 2 Report	Report on the technical and financial status of the project
31-12-2020	Final Project Report	Report on the technical and financial status of the project

Activity 2 - Data Access

Activity Description **3000**

The purpose of Real-time Data Access and Integration Layer (DAI) layer is to provide utility functions for facilitating access, consumption, integration and analysis of streaming data from a variety of data sources, and making the streaming data available for Analysis Engine and Visualization Engine. The DAI layer will provide functions for the Analysis and Visualization Engines to access the raw streaming data, and also allow access data in reduced or aggregate form in order for other components to better cope with the scale of the data. The goal is to provide Analysis and Visualization Engines with a streamlined and convenient data access interface, and sophisticated and powerful data manipulation functions.

DAI layer will build an initial proof-of-concept prototype for data access component that will be pivotal in integrating the Analysis and Visualization Engines in BigViz platform. This requires building on existing stream processing frameworks, and customizing them for the primary goal of BigViz platform to provide real-time interactive intelligent data visualization. This will involve evaluating popular open source stream processing tools - for instance, Apache Flink, Apache Spark's Structured Streaming, Apache Storm, Twitter's Heron, among others - for their suitability to BigViz use case. The selected stream processing tools will be then used for implementing the proof-of-concept prototype for working with the available data sets for BigViz use cases, and integrating visualization and analysis engine to build the overall BigViz platform.

Activity 2.1: Ingestion and Storage of Streaming Data

The DAI layer will provide tools for importing and exporting streaming data from the external sources and third-party systems. DAI will also manage storage of streaming data, ensuring access to historical data, allowing for analysis using "temporal shift" for example.

Activity 2.2: API for Streaming Analysis Engine

The DAI layer will also support Streaming Analysis Engine by providing necessary functions for data access and manipulation. Analysis engine would require specific data functions for facilitating machine learning and other data operations. DAI layer will provide the necessary functions to integrate data insights gained from Analysis engine into visualization idioms.

Activity 2.3: API for Visualization Engine

The DAI layer will also support Visualization Engine by providing necessary functions for data access and manipulation. Interactive data visualization requires providing data querying and processing functions in the Visualization engine, which use relevant API from the DAI layer. Visualization idioms may require functions for data aggregation, caching, and dimensionality reduction etc. at the data access level in DAI layer, which also satisfy the performance requirements in the Visualization engine.

This task will be led by Webdetails and managed by Pedro Vale, with the collaboration of Amin Khan.

Milestones **max. 6**

Data	Designação	Descrição
M12	Ingestion and Storage of Streaming Data	In this milestone, we have the data access system for streaming data.
M18	API for Streaming Analysis Engine	In this milestone, we have the integration for Analysis Engine to work with streaming data through DAI layer.
M24	API for Visualization Engine	In this milestone, we have the integration for Visualization Engine to work with streaming data through DAI layer.

Activity 3 - Streaming Analysis Engine

Activity Description **3000**

Despite the advances in the area of data science, and the plethora of data analysis tools, unsupervised analysis techniques continue to pose some challenges, which are even harder when dealing with data streams.

Activity 3 aims for defining the Streaming Analysis Engine, which comprises the discovery of patterns over data streams and the detection of changes on those patterns along time.
Since there are already efficient algorithms for mining patterns over data streams (for example the FP-stream algorithm [Gia+03]), the activity focuses on addressing the major issues on managing the discovered patterns.

The first task in the activity will address the adaptation of the FP-stream algorithm for mining temporal patterns over data streams, such as cyclic and calendric patterns, but also convergent and divergent ones, as proposed by [BA14] in the context of sequential data.

The second task concerns to the management of the discovered patterns, usually stored in a FP-tree. In this case, the goal is to reduce the number of stored patterns, without loosing information, by compacting set of patterns to equivalent ones involving variables. This approach will follow and extend our previous work [SA14], since the data stream to address should be a flux of multivariable records. The multivariable nature of the data allows for the detection of correlations among variables as usual, but also among variables valuations. For numerical variables we will follow the regression approach as before, but for nominal ones we should consider the new tendencies such as the work by Pedro Domingos, on Markov Logic Networks and their approach for dealing with the different concepts [RD06].

The third and final task, should address change detection directly over the stored patterns. This detector may consider the advances on the reduction task, and adapting change detection techniques to manage the discovered patterns along time.

The output of these three tasks will be directly used in Activity 4, allowing the visualization engine to know when relevant changes in the data stream have taken place (and also which and their nature) so that those changes can be visually highlighted.

This task will be the purview of IST-ID's team, led by Prof. Cláudia Antunes and supported by MSc Scholarship students. The PhD Hire, from INESC-ID will also have a small participation

mainly to ensure the solutions developed in this Activity conform to the requirements of Activity 4, thus facilitating their use.

Milestones **max. 6**

Data	Designação	Descrição
M18	Algorithms for mining temporal patterns	At this milestone, we will have proposed new algorithms for mining temporal patterns over data streams, as well as a technical report describing and validating them
M24	Algorithms for pattern reduction	At this milestone, we will have proposed new algorithms for reducing the number of patterns stored, as well as a technical report describing and validating them
M30	Algorithms for change detection	At this milestone, we will have proposed new algorithms for detecting changes over the stored patterns along time, as well as a technical report describing and validating them

Activity 4 - Adaptive Big Data Visualization

Activity Description **3000**

This activity will focus on the the development of Information Visualization techniques appropriate for SBD. This will require a two-pronged approach: dealing with quantity, and dealing with change, as follows:

Activity 4.1 - SBD Vis

Datasets handled by typical InfoVis idioms are relatively limited in size, when compared to Big Data. Techniques to deal with large amounts of information include item reduction (clustering, filtering, etc.) and attribute reduction (projections, factor analysis, etc.). These can hide the data itself, be too computationally expensive for large datasets, and usually work with the entire dataset, unsuited for a streaming context. We will research how to efficiently effect data reduction in streaming contexts, keeping at every moment the most relevant facets of the data. We will accomplish this through simplified algorithms, and the use of incremental algorithms that use the results from previous iterations to very efficiently provide solutions for the latest data in the stream.

Another approach is the development of visualization techniques that show the data without reduction, giving analysts a complete view of potentially relevant information. There are minimalistic approaches that, in the extreme, represent each datum with a single pixel. Alas, this is a comparatively unexplored area, that we plan to further research. Single pixel representations usually encode a maximum of three attributes (based on color and position). We'll go beyond this state of the art by studying how more attributes can be encoding at the minimal possible cost. For instance, color blending, the use of different color properties (hue, saturation, value, etc.) and motion will all be studied. We'll also ponder on how "big" Streaming Big Data really is, for InfoVis? A stream may, over time, produce huge amounts of data. However, it is likely that recent data is more important than historical data. Thus, recent data may be shown in its entirety and older data can be abstracted, displayed in other ways, with smaller screen real-estate and cognitive needs. We will strive for single idioms able to display recent and old data with visual and semantic continuity..

Activity 4.2 - Highlighting Change

As data in the stream changes, certain attributes, correlations or other aspects may become relevant. Activity 3 will flag those moments and relevance changes. When the visualization changes, however, this can cause disorientation and loss of context, especially in extreme cases when using completely different visualization idioms is necessary to better highlight what is important. We will thus study: (1) which changes work the best to represent particular kinds of differences in the stream; (2) which of those do not imply large cognitive loads or disorientation; and (3) possible "transformation paths", where intermediate representations are used to maintain continuity while still providing an efficient, real-time, reactive solution.

Milestones **max. 6**

Data	Designação	Descrição
M18	Streaming Big Data Visualization Idioms	At this milestone, we will have available prototype visualization techniques that deal with Big Data, as well as a technical report on the underlying principles, so that they can be used in Activity 5.
M30	Highlighting Change	At this milestone, we will have developed appropriate techniques to mutate visualizations to deal with changes in the data. A prototype and documentation will be available for their use in Activity 5.

Activity 5 - Integration and Validation

Activity Description **3000**

The integrated framework brings together the three components, DAI, Analysis, and Visualization engines, to provide a rich interactive real-time data visualization experience for streaming data. Analysis and Visualization engines use the API provided by DAI layer to access data, and with this integration the focus is on developing a comprehensive platform for streaming data analytics and visualization.

Integration is important as different optimizations are necessary across the layers for achieving the goals of real-time interactive analytics. For instance, visualization focused data aggregation and dimensionality reduction may be required at the DAI layer for streaming data. This requires better integration and feedback loop between visualization engine and DAI layer. Similarly, for Analysis engine to drive a compelling visualization experience requires machine learning to build on features in data sets, and not only help with feature selection, but also recommend the appropriate data visualization idioms. Therefore, the Analysis engine has to be closely integrated with Visualization engine, where data operations are closely coordinated by the DAI layer between the other components.

In order to demonstrate the utility and effectiveness of the BigViz platform, we will develop a demonstration for a specific use case from the industry, taking advantage of the integration of all

the three main components, DAI, Analysis and Visualization engines, and applies further customizations necessary for the use case. This demonstration validates the advancement in state-of-the-art in streaming analytics and visualizations for this specific scenario, and also shows how the framework can be utilized for other Big Data and IoT applications.

Lastly, the Pentaho BI suite provides industry leading comprehensive reporting, OLAP (Online Analytical Processing) analysis, dashboards, which are ideally suited and widely deployed for data warehousing. We will, explore possible integrations of the BigViz platform with Pentaho's tools, which would allow Pentaho platform to use the streaming data functionality of BigViz platform.

Activity 5.1 - Integrated Framework

All the three components, DAI, Analysis, and Visualization engines, are integrated into a cohesive framework, taking advantage of the data access API. The integration builds a complete platform for analysis and visualization of streaming data.

Activity 5.2 - Use Case Demonstration

The framework is used for a particular use case from a relevant domain, like industrial IoT, to demonstrate the usefulness and potential of BigViz platform. This demonstration can also act as benchmark, by using common data sets, in order to validate the improvement achieved by the framework in handling common data analytics scenarios.

Webdetails, under the lead of Amin Khan, will lead this Activity, with significant participation from the other partners. This will ensure the adequate set of skills and expertise are available.

Milestones **max. 6**

Data	Designação	Descrição
M30	Integrated Framework	At this milestone, we will combines all the three components (from activities 2,3 and 4) in one cohesive framework.
M36	Use Case Demonstration	At this milestone, we will have a use case demonstrated and validated through the integrated framework.

14. Project Characterization

Bibliography

Publicações citadas na descrição técnica da proposta. Esta lista não está limitada a publicações dos membros da equipa.

Num	Reference	Year	URL	Publication
1	Spenke2000	2000		Spenke, M. and Beilken, C., InfoZoom - Analysing Formula One Racing Results With An Interactive Data Mining And Visualisation Tool, in WIT Transactions on Information and Communication Technologies, Vol. 25, pp. 10. Doi: 10.2495/DATA000441
2	Stolte2002	2002		C. Stolte, D. Tang and P. Hanrahan, "Polaris: a system for query, analysis, and visualization of multidimensional relational databases," in IEEE Transactions on Visualization and Computer Graphics, vol. 8, no. 1, pp. 52-65, Jan/Mar 2002. doi: 10.1109/2945.981851
3	Heer2005	2005		J. Heer and D. Boyd, "Vizster: visualizing online social networks," IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., Minneapolis, MN, 2005, pp. 32-39. doi: 10.1109/INFVIS.2005.1532126
4	Yi2007	2007		J. S. Yi, Y. a. Kang and J. Stasko, "Toward a Deeper Understanding of the Role of Interaction in Information Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, pp. 1224-1231, Nov.-Dec. 2007. doi: 10.1109/TVCG.2007.70515
5	Eaton2011	2011		Eaton, C., Deroos, D., Deutsh, T., Lapis, G., Understanding Big Data (2011) IBM Report.
6	Choo2013	2013		Choo, J., Park, H. Customizing Computational Methods for Visual Analytics with Big Data, IEEE Comput. Graph. Appl. (2013) 33(4):22-8. doi: 10.1109/MCG.2013.39.
7	Liu2013	2013		Liu, Z., Jiang, B. and Heer, J. (2013), imMens: Real-time Visual Querying of Big Data. Computer Graphics Forum, 32: 421–430. doi:10.1111/cgf.12129
8	Sagiroglu2013	2013		S. Sagiroglu and D. Sinanc, "Big data: A review," 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, 2013, pp. 42-47. doi: 10.1109/CTS.2013.6567202

9	Tal13	2013		Talbot, David. 2013. "Graphics Chips Help Process Big Data Sets in Milliseconds." MIT Technology Review, October.
10	Chen14	2014		H. Chen, Mining top-k frequent patterns over data streams sliding window. J Intell Inf Syst (2014) 42: 111. doi:10.1007/s10844-013-0265-4
11	HSRC14	2014		Maayan Harel, Mannor Shie, El-Yaniv Ran, and Koby Crammer. "Concept Drift Detection Through Resampling." In ICML, pp. 1009-1017. 2014.
12	JJH14	2014		Jugel, Uwe, Zbigniew Jerzak, and Gregor Hackenbroich. 2014. "M4: A Visualization-Oriented Time Series Data Aggregation." In VLDB, 7:797–808. 10. doi:10.14778/2732951.2732953.
13	LSB14	2014		Vincent Lemaire, Christophe Salperwyck, Alexis Bondu.A Survey on Supervised Classification on Data Streams. eBISS 2014: 88-125
14	Munzner2014	2014		Munzner, Tamara. Visualization analysis and design. CRC Press, 2014.
15	NASNG14	2014		Shamila Nasreen, Muhammad Awais Azam, Khurram Shehzad, Usman Naeem, Mustansar Ali Ghazanfar, Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey, Procedia Computer Science, Volume 37, 2014, Pages 109-116, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2014.08.019
17	BCD16	2016		Bhowmick, Sourav S, Byron Choi, and Curtis Dyreson. 2016. "Data-driven Visual Graph Query Interface Construction and Maintenance : Challenges and Opportunities." In VLDB, 984–92.
18	BCS16	2016		Battle, Leilani, Remco Chang, and Michael Stonebraker. 2016. "Dynamic Prefetching of Data Tiles for Interactive Visualization." In SIGMOD, 1363–75. New York, New York, USA: ACM Press. doi:10.1145/2882903.2882919.
19	DZ16	2016		Ariyam Das and Carlo Zaniolo Fast Lossless Frequent Itemset Mining in Data Streams using Crucial Patterns Proc 2016 SIAM International Conference on Data Mining. 2016, 576-584
20	GLA16	2016		Mohammed Ghesmoune, Mustapha Lebbah and Hanene Azzag, State-of-the-art on clustering data streams, Big Data Analytics 2016(1):13 DOI: 10.1186/s41044-016-0011-3
21	GBEB17	2017		Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. 2017. A Survey on Ensemble Learning for Data Stream Classification. ACM Comput. Surv. 50, 2, Article 23 (March 2017), 36 pages. DOI: https://doi.org/10.1145/3054925

22	Mat17	2017		Matheson, Rob. 2017. “Split-second data mapping.” February 9.
23	SK17	2017		Tegjyot Singh Sethi and Mehmed Kantardzic. "On the Reliable Detection of Concept Drift from Streaming Unlabeled Data." Expert Systems with Applications. 2017.

Publicações anteriores - Incluir as cinco publicações mais representativas do trabalho da equipa no âmbito desta proposta. (máx. 5)

N	Reference	Year	URL	Publication
1	[SilA14]	2014	http://web.ist.utl.pt/claudia.antunes/artigos/silva2015dmkd.pdf	Andreia Silva, Cláudia Antunes. Multi-Relational Pattern Mining over Data Streams. In Data Mining and Knowledge Discovery, vol. 29, nr. 6, pp. 1783-1814. Springer. ISSN: 1384-5810. DOI: 10.1007/s10618-014-0394-6. November 2014.
2	[Jordão16]	2016	http://web.ist.utl.pt/~daniel.j.goncalves/publications/2016/EduVis-IJCICG.pdf	Vilma Jordão and Sandra Gama and Daniel Jorge Viegas Gonçalves, Visualizing Sequencial Datamining Patterns, International Journal of Creative Interfaces and Computer Graphics, 7(1), pp. 1-18, Jul. 2016, IGI Global
3	[Gama14]	2014	http://web.ist.utl.pt/~daniel.j.goncalves/publications/2014/WIBEuroVis.pdf	Sandra Gama and Daniel Gonçalves. Studying Color Blending Perception for Data Visualization. In Proceedings EuroVis 2014. pp 121-125. Swansea, Wales, UK. 9-13 Jun 2014. Eurographics. ISBN: 978-3-905674-69-9. DOI=10.2312/eurovisshort.20141168
4	[SA14]	2014	http://dl.acm.org/citation.cfm?id=2628243	Daniel Serrano, Cláudia Antunes. Condensed Representation of Frequent Itemsets. In 18th International Database Engineering & Applications Symposium (IDEAS 2014), pp. 168-175. Porto, Portugal. July 2014. ISBN: 978-1-4503-2627-8. DOI: 10.1145/2628194.2628243
5	[KGL17]	2017	https://zenodo.org/record/580317	Amin M. Khan, Daniel Gonçalves, Duarte C. Leão. Towards an Adaptive Framework for Real-Time Visualization of Streaming Big Data. In *19th Eurographics/IEEE VGTC Conference on Visualization (EuroVis 2017).*

Projectos financiados em que o IR ou Co-IR participaram nos últimos 5 anos **(máx 5)**

N	IR ou Co-IR	Referência	Título	Inicio	Fim	Programa Financiador
1	IR	AAL/014/2009	PAELife	01/03/2012	31/10/2014	EU AAL
2	IR	FCT PTDC/EIA-EIA/110058/2009	EDUCARE	14/02/2011	15/2/2014	FCT

Papel do IR/Co-IR no projecto 1000:

PAELife:

The IR was the leader of INESC-ID's participation in this project, managing that team's work and ensuring the quality of the results, both internal and overall, through direct coordination with other project partners.

EDUCARE:

The IR was the leader of INESC-ID's participation in this project, managing that team's work and ensuring the quality of the results, both internal and overall, through direct coordination with other project partners. The Co-IR was also a part of this project, where she had a more hands-on role, directly coordinating the technical team and also doing some development herself. That is part of the reason why we are confident we possess the skills required to lead VisBig to fruition, even on the technical implementation side.

Principais resultados relevantes para a candidatura 2000:

PAELife:

The goal of Project PAELife was the development of a multimodal Personal Assistant for the Elderly, that helped minimize social exclusion. Indeed, many of our everyday contacts and interactions with other people are through technology or mediated by technology. However, older users have a hard time understanding and using some of that technology, not only due to cultural issues but also to the fact that existing solutions do not properly address accessibility questions. Withing PAELife a personal assistant, AALfred, was developed that serves as a broker to several services (news, calendar, faceboo, twitter, email, etc.) but using an interface that is properly tailored for older users. This is a multimodal interface where speech, gestures,

and other interaction modalities can be used to interact. It was envisioned to be used in a living room scenario, thus minimizing adoption barriers. INESC-ID's team participated in the development of novel modalities. This required us to perform extensive user testing, in the same way as it will be necessary for VisBig. Our success in PAELife is proof of our ability to conduct such testing, procuring users, even if of particular profiles (in PAELife, we established cooperations with senior centers to ensure we'd get real, representative users). This implies not only technical expertise in designing the tests and doing the statistical analysis of their results, but also also shows we have ample expertise in managing the team and dealing with users in fruitful ways. We are, thus, completely convinced we will be able to develop, in VisBig, solutions that cater to the real needs of real Visual Analysts.

EDUCARE:
dew

16. Investment

Fundamentação do Investimento **3000**

Publishers charge open-access fees, for papers to be published in that fashion. This is an average conservative value for a typical journal in our areas of expertise.

In the knowledge domains of this project, Conferences play an exceptionally important role. In fact, most citations and impact often comes from conference papers, which undergo rigorous peer-reviewing (often two-round). This value will conservatively cover the publication and presentation of a paper at an European conference.

In the knowledge domains of this project, Conferences play an exceptionally important role. In fact, most citations and impact often comes from conference papers, which undergo rigorous peer-reviewing (often two-round). Most important conferences have a worldwide scope and often take place outside Europe. This value will conservatively cover the publication and presentation of a paper at a conference outside our continent.

National conferences are important as a way to disseminate our results within the national community and are important both as scientific and dissemination activities.

This laptop will be necessary for activities related to the management of the project, as well as Activity 4, where mobility will be important during meetings, conferences, etc.

The PhD hire needs material to work. This will be his PC, where he will undertake work for this project.

To be used in Activity 3. Given the highly demanding algorithms to be developed, it will ned to be a machine with better specs.

To cover costs for the organization of the mid project workshop, from poster printing to coffee breaks and lunch expenses.

To cover costs for the organization of the final project workshop, from poster printing to coffee breaks and lunch expenses.

It is important to keep up to date with the latest literature in the field. This value covers the purchase of books and other similar materials (papers, journal issues, etc.) necessary to keep the research team up to date.

No project can be conducted without a myriad of supporting material for day-to-day operations such as paper, toner, and even pens and pencils.

18. Indicators

Fundamentação dos indicadores 3000

The team from INESC-ID and IST-ID is led by professors. As such, we often work with our students in our research, both by paying them scholarships and as their supervisors. Thus, we plan to integrate their work into the project. That is why we promise six MSc dissertations. From INESC-ID's side, we will have one MSc student per school year undergoing research related to VisBig, even if not officially part of the project's core team (unnecessary, since we will hire a full-time PhD researcher for the project). IST-ID's three MSc are part of the project team, "MSc Student" one to three in the personnelle list.
The four Computational Applications promised directly match the outputs of Activities 2-5: a Data Access framework, the Streaming Analysis Framework (with support for change detection), the Adaptive Visualization Framework and, finally, the Integrated Framework. Each Activity will produce a self-contained, documented, piece of software, to increase the usefulness of the project's results: each of those components can conceivably be used in other application domains. The Prototype will come from Activity 5 and will be the instantiation of the Integrated Framework for a particular example domain: a demonstrator.
The PI of this project is the responsible for the teaching of Information Visualization at Instituto Superior Técnico. As such, he will integrate the scientific results of the project into the syllabus of relevant courses, further increasing their dissemination and usefulness in the community. This is how the "Integration of knowledge into higher learning activities" will be done (beyond the MSc works in themselves).

Finally, the remaining indicators all have to do with scientific publication, both in journals and conferences, especially important in our area. Indeed, conferences in the knowledge areas of the project are the first and foremost places where to publish to achieve high impact and citation factors (in some cases far above those of journals). They require full-paper submissions, which are peer-reviewed, often in a two-step process (rebuttal) and in many cases have acceptance rates of 20% or less. Thus, we will publish our results in the main conferences and journals in the area. We foresee that some of the papers will be multi-disciplinary, co-authored by members from several partners. We will focus on the topmost journals and conferences in the area, as described in the “Dissemination Plan” section.

Plano de acções de disseminação de resultados e promoção do cimento e divulgação da cultura científicas **3000**

- Acções de divulgação de cultura científica;
- Acções Promoção e disseminação do conhecimento;
- Publicações técnicas / cientificas;
- Conferencias, seminarios ou forums;
- Acções junto dos sectores alvo;
- Outros (especificar)

This project will have two main types of outcomes: new scientific results, going beyond the current state of the art in the relevant areas; and new software tools that embody those results, making them practical and usable by a wider audience. Thus, our dissemination will take place in a two-pronged way.

Scientific results will be disseminated, as expected, mainly through scientific publications. Reaching a broader audience and with more impact we will target the main conferences and journals in the area: IEEE InfoVis, EG Eurovis, ACM CHI, ACM UIST, the IEEE Transactions on Visualization and Computer Graphics, the International Journal of Human-Machine Studies, ACM Multimedia; Data Mining and Knowledge Discovery, ACM KDD; ACM SIGMOD international Conference on Management of Data (SIGMOD); IEEE International Congress on Big Data (BigData Congress); International Conference on Very Large Data Bases (VLDB).

Software tools and products will be made open-source and available to all. They will be disseminated through a public GitHub repository, one of the best known open-source repository and collaboration platforms, for higher visibility.

Also, we will organize two workshops, one at the middle of the project and the other at the end. Both the academic community and industry will be invited to attend (we have access to both given the nature of consortium members). Both will help make the project better known in both communities. The first will also work as a place to gather comments and suggestions from both

sides of the aisle to produce sounder results and, at the same time, increase the chances that the end results will be useful and used by a wider community.

We will leverage on the existing, market leading tools and applications made available by Webdetails/Pentaho, by making all the software stemming from the project compatible. Again this will allow for an easier adoption of the project's results, given that those tools and applications are already in use by thousands of companies and individuals worldwide.

Overall, we will have since the beginning of the project a Project Website where all news, results and relevant information will be available.

All this will be facilitated by INESC-ID, that takes special care of the research projects carried out within the institution and dedicates a significant effort to dissemination activities that allow general public to acknowledge and take part in the developed research, including the support in organizing conferences, seminars, lectures and other training activities like summer schools and workshops that cover a wide range of public interested in the topics we work with. Activities are disseminated through specific channels and general media. Specific human resources working within the institution are dedicated to the interface between researchers and the rest of the world, in order to give science an open access.

17. Societal Challenges

O projecto dá resposta a desafios societais? Quais?

N	Desafio Societal	Principal Linha de Actuação	Justificação
1	Energia Segura, Não Poluente e Eficiente	Redução do consumo de energia e da pegada de carbono mediante uma utilização inteligente e sustentável;	One of the areas primed for the production of Big Data (and its subsequent analysis) is energy production. New technologies such as smart meters are a reality, and generate a continuous stream of data on energy consumption. VisBig will provide the tools for a more efficient analysis and visualization of such data,facilitating the optimization of energy

			production and distribution, better matching real-time demand, in a more efficient and sustainable way. Similarly it will assist in the identification of atypical conditions and in identifying the most ecological way to address them.
2	Energia Segura, Não Poluente e Eficiente	Uma rede europeia de eletricidade única e inteligente;	A smart grid presupposes the existence of data that gets assessed and upon which decisions regarding production and distribution are based.The technology researched in VisBig will assist in the collection of the data, in real-time; its visualization by human operators; and, more importantly, the machine-learning algorithms that identify changes in the data-stream can be the basis for automated decision processes.
3	Energia Segura, Não Poluente e Eficiente	Processo decisório sólido e envolvimento do público (compreensão das tendências e perspetivas socioeconómicas relacionadas com a energia);	For the public to be involved in decision making it needs to understand the matters at hand. A solid, non-confusing way of visualizing energy consumption data, in real-time, will help not only with more rational and environment friendly behaviour change at the individual level, but also with a collective understanding of the circumstances and even the real-time effect of

			policies and events.
4	Transportes Inteligentes, Ecológicos e Integrados	Investigação socioeconómica e comportamental e atividades prospetivas para a definição de políticas (compreensão dos impactes socioeconómicos, tendência e perspetivas relacionadas com os transportes);	Information such as the real-time routes, speeds and position of vehicles, as well as data regarding when passengers get onboard vehicles and the trips they perform are crucial for the understanding of the state of a transportation network and its usage. This is a classical Big Data scenario for which real-time understanding of the data streams, as proposed by VisBig, will enable better decision making.
5	Ação Climática, Ambiente, Eficiência de Recursos e Matérias-Primas	Desenvolver sistemas de observação e informação globais abrangentes e sustentados (observação e monitorização da Terra);	Long are the days of isolated probes or sensors measured manually every few days. Nowadays (and increasingly so) networks of sensors scattered throughout the environment monitor the it and produce real-time streams of data that need to be analysed. VisBig will provide the tools for that analysis.
6	Sociedades Seguras – Defender a Liberdade e a Segurança da Europa e dos seus Cidadãos	Proteger e melhorar a resiliência das infraestruturas críticas, das cadeias de fornecimentos e dos meios de transporte	This is an area where rapid and accurate decision making is of the utmost importance. VisBig aims to enable exactly that: the means for the real-time collection of Big Data (from transportation, infrastructure, utilities, etc.), detect abnormal situations and allow an analyst to make sense of what is happening. This

			domain offers an almost perfect match for the aims of the project.

18. Documentation to Deliver

1. Cronograma
2. Protocolo de colaboração acordado entre os copromotores.
3. Balanço e demonstração de resultados.
4. Declaração do responsável da entidade assegurando a inscrição orçamental do projecto e as necessárias condições financeiras e orçamentais para a sua realização (entidades públicas)
5. ~~Acordo escrito entre o IR e a IP Ponto 6.1 i) do Aviso para Apresentação de Candidaturas.~~
6. ~~Carta de elegibilidade da FAPESP~~
7. ~~Candidatura da equipa brasileira …~~
8. ~~Carta de elegibilidade da FUNCAP~~
9. ~~Candidatura da equipa brasileira a FUNCAP ….~~
10. ~~Outros~~

Appendix: Partner Profile Webdetails

The following section provides detail about Webdetails, move the text to relevant sections in the document above.

Description of the legal entity

Webdetails - Consulting Unipessoal, Lda. was founded in Cascais, Portugal in 2008, focused on providing data analytics and visualizations solutions. It all started with the purpose of delivering tailor-made Business Analytics solutions and services for the Pentaho platform. Its 50-member team provides services all across the globe, remotely from its headquarters just outside of Lisbon, in Cascais in the beginning and now in Oeiras, with most of its business coming from the U.S. and Europe. Webdetails has been designing plugins for Pentaho for several years, with special emphasis on its open source Community Tools or CTools suite for creating and managing Dashboards and Reports. In 2013, Webdetails was acquired by the Pentaho Corporation, and in 2015, Pentaho Corporation was acquired by Hitachi. Hence, since 2013, Webdetails operates as a Pentaho company in Portugal, and represents the largest engineering setup of Pentaho Corporation outside US.

Profile and expertise

Pentaho, a Hitachi Group company, is a leading data integration and business analytics company with an enterprise-class, open source-based platform for diverse big data deployments. Its mission is to help organizations across industries harness the value from all their data, including big data and Internet of Things (IoT), enabling them to find new revenue streams, operate more efficiently, deliver outstanding service and minimize risk.

Pentaho Corporation is the commercial open source alternative for Business Intelligence (BI). Pentaho BI Suite Enterprise Edition provides comprehensive reporting, OLAP analysis, dashboards, data integration, data mining and a BI platform that have made it the world’s leading and most widely deployed open source BI suite. Pentaho’s commercial open source business model eliminates software license fees, providing support, services, and product enhancements via an annual subscription. In the years since Pentaho’s inception as the pioneer in commercial open source BI, Pentaho's products have been downloaded more than three million times, with production deployments at companies ranging from small organizations to The Global 2000.

Publications & Software, relevant to the call content

1. **Pentaho Data Integration**, also known as Kettle is a powerful open-source Extraction, Transformation and Loading (ETL) engine, using a groundbreaking, metadata-driven approach.
    a. Casters, Matt, Roland Bouman, and Jos Van Dongen. "Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration". John Wiley & Sons, 2010.
2. **CTools Dashboards** are a set of community-driven open-source tools which are installed as a stack on top of the Pentaho Server. CTools provides its users with the ability to utilize Web technologies and data visualization concepts, and make the most of best practices to create a huge visual impact.
    a. Gaspar, Miguel. "Learning Pentaho CTools". Packt Publishing Limited, May 2016.
3. Dixon, James, Doug Moran, and Marc Batchelor. "Business intelligence system and methods." U.S. Patent Application No. 11/498,943.
4. Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.
5. Frank, Eibe, Mark Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H. Witten, and Len Trigg. "Weka." In Data Mining and Knowledge Discovery Handbook, pp. 1305-1314. Springer US, 2005.

Previous projects or activities, relevant to the proposal

1. Pentaho, developed an industrial IoT project for **Caterpillar Marine Asset Intelligence**, which uses the data generated by sensors from shipping vessels for predictive maintenance analytics using machine learning models.
2. Pentaho worked with the city of Copenhagen to develop **Copenhagen City Data Exchange**, a Smart City project, where the Pentaho platform was used to build the analytics dashboards made available to the citizens of Copenhagen.
3. Pentaho is currently involved in an industrial IoT project with **Hitachi Rails Europe** that uses predictive maintenance analytics for building 'self-diagnosing' trains.
4. **Pentaho MQTT Project** is a Pentaho plugin for Message Queuing Telemetry Transport (MQTT), which is a popular choice for Machine-to-Machine (M2M) communication and other Internet of Things (IoT) applications and solutions.

Dissemination Plans

Webdetails is part of a leading international Big Data and IoT Analytics company, with millions of downloads of its open-source software, and with production deployments at companies ranging from SMEs to The Global 2000. As part of Hitachi Insight Group, Webdetails plays a central role in Hitachi's upcoming core platform for Industrial and Enterprise IoT, Lumada. Webdetails is also part of The European Alliance of IoT Innovation (AIOTI). Pentaho has a vibrant open-source community in Europe and USA involving SMEs working with Data Analytics, Big Data and IoT, and Webdetails can play an important role in dissemination of the project results and various Pentaho events, like Annual Pentaho Community Meeting, Pentaho World conference, and numerous Pentaho User Group Meetups. Webdetails will also involve their Marketing Department to publish white papers and press releases, as well as promote the project at various industry forums and events.

Exploitation Plans

Webdetails develops Pentaho Business Intelligence (BI) platform, which is world's leading and most widely deployed open source BI suite. Webdetails is investing extensively in R&D of Big Data enhancements, as well as IoT Analytics as part of Hitachi's Lumada platform. The technologies and tools developed by the VizBig project are of critical significance for Webdetails, to enhance its portfolio with real-time data processing, analytics and visualization of streaming data. The techniques and solutions from VizBig benefit Webdetails' offerings to existing customers like Caterpillar Marine Asset Intelligence and Hitachi Rails Europe, and offer Webdetails competitive advantage in Big Data and IoT Analytics market for future projects. Webdetails can quickly transfer the results from VizBig to market, with the innovations from VizBig directly impacting numerous SMEs in Europe who use Pentaho's open-source suite.

Curriculum vitae of the Persons

Amin Khan (male)

is Software Engineer at Pentaho Portugal. His work focuses on Big Data Security features of Pentaho Data Integration and Analytics platform. He has over ten years experience working on research projects and enterprise applications in academia and industry. He is author and co-author of various articles appearing in leading international refereed journals and conferences. Amin completed his Ph.D in Distributed Computing from UPC BarcelonaTech and IST Lisbon in 2016. He received his Masters degree in Informatics from Edinburgh and Trento in 2007.

Pedro Vale (male)

is a Msc. in Artificial Intelligence who found himself early in his career helping launch a software development startup in the Business Process Management field. Realizing developing real software was great fun, his professional journey has been focused on architecting and building software platforms, managing developer teams and generally being concerned with the ability to systematically deliver high-quality software.

Needing a breather from large enterprise environments, he joined Webdetails in 2011 as the Development Team Lead, getting acquainted at that time with the Pentaho platform and the Business Intelligence/Big Data field. With the acquisition of Webdetails by Pentaho in 2013, he's now part of the engineering management at Pentaho, helping deliver the next groundbreaking versions of the platform, finding it amusing that his Artificial Intelligence skills might actually be useful in building some of the pieces of the platform.

References

- [BA14] A. Barreto, C. Antunes. Finding Periodic Regularities on Sequential Data: Converging, Diverging and Cyclic Patterns. In International C* Conference on Computer Science & Software Engineering (C3S2E 2014), pp. 19.1-19.4. Montreal, Canada. August 2014. ISBN: 978-1-4503-2712-1. DOI: 10.1145/2641483.2641519
- [BCS16] Battle, Leilani, Remco Chang, and Michael Stonebraker. 2016. “Dynamic Prefetching of Data Tiles for Interactive Visualization.” In *SIGMOD*, 1363–75. New York, New York, USA: ACM Press. doi:10.1145/2882903.2882919.
- [BCD16] Bhowmick, Sourav S, Byron Choi, and Curtis Dyreson. 2016. “Data-driven Visual Graph Query Interface Construction and Maintenance : Challenges and Opportunities.” In *VLDB*, 984–92.
- [CBD10] Casters, Matt, Roland Bouman, and Jos Van Dongen. 2010. Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration. John Wiley & Sons.
- [Gia+03] Giannella, Chris, et al. "Mining frequent patterns in data streams at multiple time granularities." *Next generation data mining* 212 (2003): 191-212.
- [HRM16] He, Xi, Nisarg Raval, and Ashwin Machanavajjhala. 2016. “A Demonstration of VisDPT : Visual Exploration of Differentially Private Trajectories.” In *VLDB*, 9:1489–92. 13.
- [JJH14] Jugel, Uwe, Zbigniew Jerzak, and Gregor Hackenbroich. 2014. “M4: A Visualization-Oriented Time Series Data Aggregation.” In *VLDB*, 7:797–808. 10. doi:10.14778/2732951.2732953.
- [Kim+15] Kim, Albert, Eric Blais, Aditya Parameswaran, Piotr Indyk, Sam Madden, and Ronitt Rubinfeld. 2015. “Rapid sampling for visualizations with ordering guarantees.” In *VLDB*, 8:521–32. 5. doi:10.14778/2735479.2735485.
- [Mat17] Matheson, Rob. 2017. “Split-second data mapping.” February 9. http://news.mit.edu/2017/startup-mapd-fast-big-data-mapping-0111.
- [RD06] Richardson, Matthew; Domingos, Pedro (2006). "Markov Logic Networks" (PDF). *Machine Learning*. **62** (1-2): 107–136. doi:10.1007/s10994-006-5833-1
- [Tal13] Talbot, David. 2013. “Graphics Chips Help Process Big Data Sets in Milliseconds.” *MIT Technology Review*, October. https://www.technologyreview.com/s/520021/graphics-chips-help-process-big-data-sets-in-milliseconds/.
- [TSG+17] Traub, Jonas, Nikolaas Steenbergen, Philipp Grulich, Tilmann Rabl, and Volker Markl. 2017. “I2: Interactive Real-Time Visualization for Streaming Data.” In 20th International Conference on Extending Database Technology (EDBT’17), 526–29.
- [Var+15] Vartak, Manasi, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2015. “SeeDB: Efficient data-driven visualization recommendations to support visual analytics.” In *VLDB*, 8:2182–93. 13. doi:10.14778/2831360.2831371.