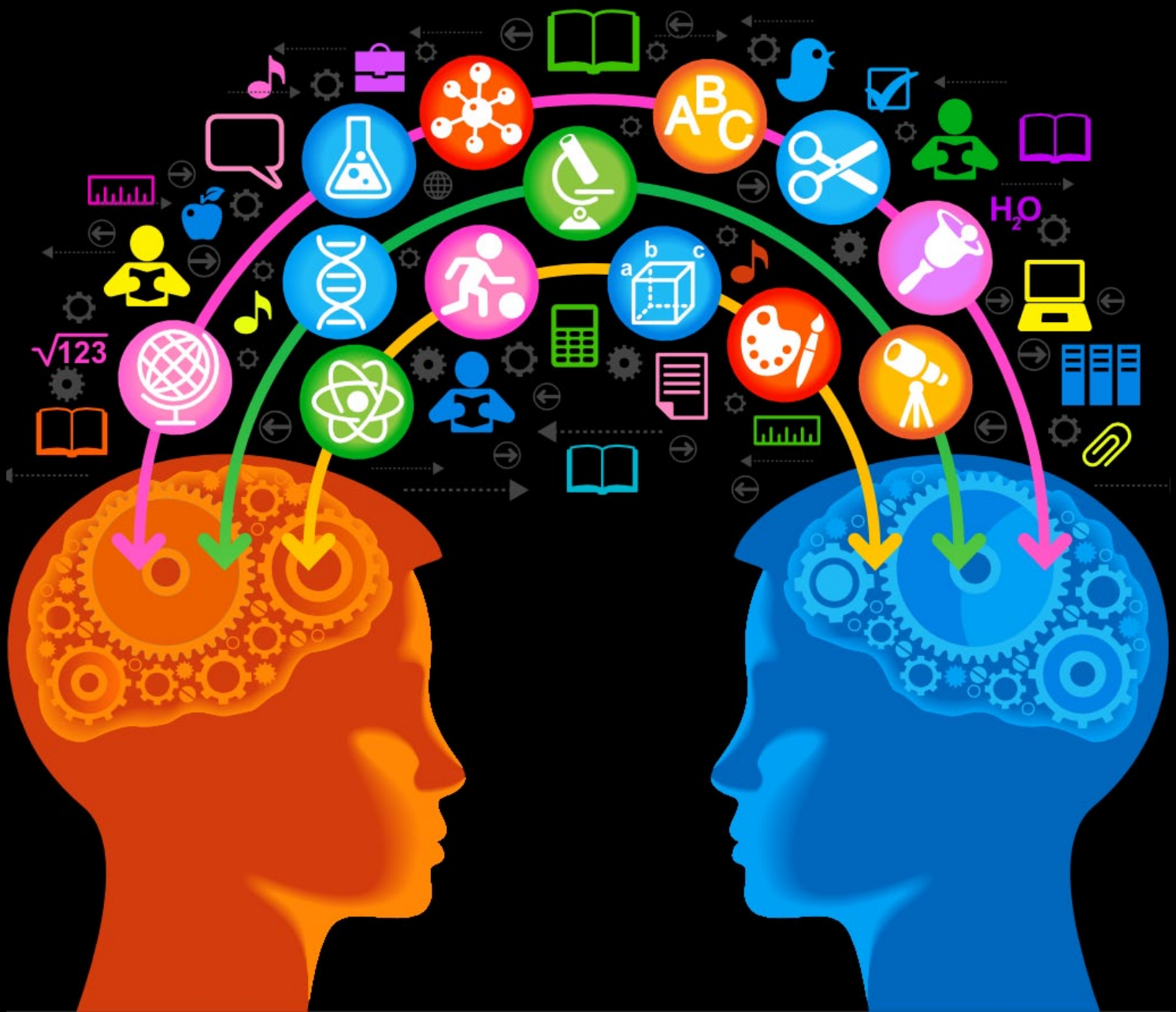
[BIO IN DOCKER]

9 - 10 NOVEMBER 2015,

WELLCOME BUILDING, LONDON



INTRODUCTION

Docker is now establishing itself as the de-facto solution for containerization across a wide range of domains. The advantages are attractive, from reproducible research to simplifying deployment of complex code. Several bioinformatics groups are now utilizing this for various purposes: we would like to bring together some notable cases to discuss how adavantage of this new technology can best be acheived.

This event has been organised by **Dr. Stephen J. Newhouse** and **Dr. Amos Folarin**, both senior members of the NIHR Maudsley Biomedical Research Centre (BRC), Bioinformatics Core and made possible by the expert organisation and project management of **Lucy O'Neill**. Thanks to **Tanya Hardy** from **Mindwave Ventures** for lending us her time and project management skills. The agenda and poster have been designed by **Jessica Morgan**.

The symposium is generously supported by **Genomics England** and the **Medical Research Council**.

Join us for the launch of the F1000 channel, **'Container Virtualisation in informatics'** at 9:45 on the second day of the conference on November 10.
See it at: f1000research.com/channels/containers

From 13:40, we'll be conducting an **afternoon mini-hackday** to introduce, demonstrate, and invite participation using Docker on some interesting and well scoped problems

Sign up, clone our repo and get involved with our **github**.

Follow the action throughout the day at **@bioindocker15** on Twitter and using the hashtag **#bioindocker15**.

TABLE OF CONTENTS

MONDAY, NOV. 9

9:00	**Networking breakfast**	Dale Room, where breaks and lunch will take place.
9:45	**Welcome**	Franks/Steel room, where all talks will take place.
10:00	**Peter Belmann** Bioboxes	Evaluating and ranking bioinformatics software using docker containers and an overview of the BioBoxes project.
10:40	**Nebojsa Tijanic** Seven Bridges Genomics	Portable workflow and tool descriptions with Common Workflow Language and Rabix
11:20	**Break**	
11:40	**Paolo Di Tommaso** Nextflow, CRG	Manage reproducibility in genomics pipelines with Nextflow and Docker containers
12:20	**Amos Folarin & Stephen Newhouse** NGSeasy, King's College London	Next generation sequencing pipeines in Docker
13:00	**Lunch**	
13:40	**Tim Hubbard** Genomics England, King's College London	Pipelines to analysis data from the 100,000 genomes project as part of the Genomics England Clinical Interpretation Partnership (GeCIP)
14:20	**Fabien Campagne** Weill Cornell Medicine	MetaR and the Nextflow Workbench: application of Docker and language workbench technology to simplify bioinformatics training and data analysis.
15:00	**Break**	
15:20	**Brad Chapman** Harvard T. H. Chan School of Public Health	Improving support and distribution of validated analysis tools using Docker
16:00	**Elijah Charles** Intel	Bioinformatics and the packaging melee
16:25	**Kai Davenport** ClusterHQ	Data, Volumes and portability with Flocker
17:05	Day ends	Meet for drinks.

TUESDAY, NOV. 10

Join us at 9:45 for the launch of the F1000 channel, **'Container Virtualisation in informatics'**.

From 13:40, we'll be conducting an **afternoon mini-hackday** to introduce, demonstrate, and invite participation using Docker on some interesting and well scoped problems

Sign up, clone our repo and get involved with our **github**.

9:00	**Networking breakfast**	Dale Room, where breaks and lunch will take place.
9:45	**Thomas Ingraham & Michael Markie** F1000	F1000Research - a publishing platform for the Docker community
10:00	**Alfonso Acosta** Weaveworks	Weaving Containers in Amazon's ECS
10:40	**Aanand Prasad** Docker	Orchestrating Containers with Docker Compose
11:05	**Matthew Bates** Jetstack	Manage your infrastructure like Google with Kubernetes.
11:30	**Break**	
11:50	**Clive Stringer & Adam Hatherly** King's College London & HSCIC	Docker and real world problems
12:15	**Yannick Wurm** Queen Mary University of London	OSwitch: One-line access to other operating systems
12:25	**Introduction to the mini-hack**	Prizes for the mini-hack will be announced.
13:00	**Lunch**	
13:40	**Mini-hack**	Sign up, clone our repo and get involved with our **github**.
16:15	**Discussion session and wrap-up**	
17:05	Day ends: drink	Meet for drinks.

WITH THANKS TO

Gold sponsers

Genomics England

Genomics England was set up to deliver the 100,000 Genomes Project. This flagship project will sequence 100,000 whole genomes from NHS patients and their families.

Medical Research Council

The Medical Research Council improves human health through world-class medical research. We fund research across the biomedical spectrum, from fundamental lab-based science to clinical trials, and in all major disease areas.

Group affiliates

King's College London

King's College London was founded by King George IV and the Duke of Wellington (then Prime Minister) in 1829 and is dedicated to the advancement of knowledge, learning and understanding in the service of society.

Farr Institute of Health Infomatics Research

The Farr Institute of Health Informatics Research brings together researchers, clinicians and those with an interest in e-health records research in an environment to foster collaborations and to establish a centre of excellence in innovative health informatics research.

Silver sponser

Software Sustainability Institute

Bronze sponsers

intel

F1000Research Open for Science.

docker

Our speakers

bioboxes

ClusterHQ

CRG Centre for Genomic Regulation

NHS National Institute for Health Research

Weill Cornell Medical College

King's College Hospital NHS

hscic Health & Social Care Information Centre

Queen Mary University of London

JETSTACK

nextflow

SevenBridges

weaveworks

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Our host

wellcome trust

Wellcome Trust

The Wellcome Trust is an independent global charitable foundation dedicated to improving health, because good health makes life better.

PETER BELMANN

Bioboxes

Monday, November 9 at 10:00

2009 – 2012 Bachelor of Science in Bioinformatics and Genome Research at the Bielefeld University

2012 - 2015/7 Master of Science in Computer Science in the Natural Sciences at the Bielefeld University

Since 2015/8 Phd student at Alex Sczyrba's Computational Metagenomics lab at Bielefeld University

Github

Software has proliferated in bioinformatics and so have the problems associated with it: missing or unobtainable code, difficult to install dependencies, unreproducible workflows, all with terrible user experiences.

We have created bioboxes with the aim to make accessing and using bioinformatics software more simple and more easy. Bioboxes is a standard for creating interchangeable bioinformatics software containers for the bioinformatics community.

The Docker platform, which bioboxes uses, makes it much simpler for developers to share their their bioinformatics tools. Evaluating software is one use case of Bioboxes. Projects like CAMI and Nucleotid.es are involved in the Bioboxes project in order to benchmark software in an automated way and provide reproducible results.

bioboxes

NEBOJSA TIJANIC

Seven Bridges Genomics

Monday, November 9 at 10:40

Focusing on development and adaptation of bioinformatics tools and pipelines for a data processing platform built on top of AWS.

Github

The Common Workflow Language (CWL) is an informal, multi-vendor working group consisting of various organizations and individuals that have an interest in portability of data analysis workflows. Our goal is to create specifications that enable data scientists to describe analysis tools and workflows that are powerful, easy to use, portable, and support reproducibility. CWL builds on technologies such as JSON-LD and Apache Avro for data modeling and Docker for portable runtime environments. Rabix is an open-source project that aims to simplify creation and execution of CWL tools and workflows.

SevenBridges genomics

PAOLO DI TOMMASO

Nextflow, CRG

Monday, November 9 at 11:40

Paolo Di Tommaso is Research Software engineer at CRG. He has developed algorithms and software architectures for more than 20 years and his main specialities are parallel programming, HPC and cloud computing. Paolo is B.Sc. in Computer Science and M.Sc. in Bioinformatics He is the creator and project leader of the Nextflow pipeline framework.

@PaoloDiTommaso
nextflow.io

Genomics pipeline usually rely on a combination of several pieces of third party research software. This software is normally difficult to install, configure and to deploy. Moreover their experimental nature can result in frequent updates that can raise serious reproducibility issues.

In this presentation I will show how we managed to tackle the reproducibility issue with Nextflow and Docker containers. I will give a brief introduction to the main concepts in the Nextflow programming environment and I will show how this tool can used together with Docker to deploy and run genomics pipelines on multiple platforms in a repeatable manner.

nextflow CRG Centre for Genomic Regulation

STEPHEN NEWHOUSE

NGSeasy, KCL

AMOS FOLARIN

NGSeasy, KCL

Monday, November 19 at 12:20

Bioinformatics pipelines often use large numbers of components and deploying them incurs substantial configuration and maintenance burden (a significant barrier to reproducible research). Our aim is to define a new paradigm and best practices for developing, distributing and running pipelines encapsulated in Docker containers (lightweight virtualization), with a focus on Next Generation Sequencing (NGS) workflows. This approach provides several advantages, namely: efficiency, portability, versioning and reproducibility. Using the NGSeasy pipeline, a user can quickly deploy any pipeline version in any environment (e.g. operating systems, workstations, clusters, clouds). While this might also be achieved with a virtual machine (VM); VMs lack portability, have substantial overhead (disk, CPU, RAM), and require allocated resources to be provisioned statically -- Docker, to a large extent, solves these issues.

KING'S College LONDON

Amos is the Senior Software Development Group Leader at the KCL BRC-MH. He qualified as a biochemist/molecular biologist (Bristol), and has since worked as a bioinformatician over the last 13 years (Inpharmatica, Birkbeck, UCL, KCL/SLaM).

Joining the NIHR BRC-MH Bioinformatics group in 2012 he took up the role of Bioinformatician/Software Developer to work on a number of big data projects. His research interests principally include developing bioinformatics and clinical genomics analyses.

His software development interests include virtualization platforms, cloud technologies, mobile applications. Current work includes developing a mobile platform for monitoring patient behaviour and environment using small wearable devices and mobile phone sensors.

TIM HUBBARD

Genomics England, KCL

Monday, November 9 at 13:40

Timothy John Phillip Hubbard is a Professor of Bioinformatics at King's College London, Head of Bioinformatics at Genomics England and Honorary Faculty at the Wellcome Trust Sanger Institute in Cambridge, UK..

Hubbard was educated at the University of Cambridge where he was awarded a Bachelor of Arts degree in Natural Sciences (Biochemistry) in 1985. He went on to do research in protein design in the Department of Crystallography, Birkbeck College, London where he was awarded a PhD in 1988 for research supervised by Tom Blundell.

Hubbard's research interests are in Bioinformatics, Computational biology and Genome Informatics. During his tenure at WTSI he supervised several successful PhD students to completion in these areas of research.

@timjph

FABIEN CAMPAGNE

Weill Cornell Medicine

Monday, November 9 at 14:20

Fabien Campagne is an Assistant Professor at the Weill Cornell Medical College, Weill Cornell Medicine, New York, NY, USA.

He received his PhD in computational chemistry in 1998 from the University of Nancy I, France and furthered his training in the field of Bioinformatics at the Mount Sinai School of Medicine (NY, USA, 1998-2003).

His laboratory focuses on translational bioinformatics. Current areas of interest research include allogenomics of kidney transplantation and development of biomarkers for Chronic Fatigue Syndrome/Myalgic Encephalopathy (CFS/ME).

The Campagne laboratory was an early adopter of Language Workbench Technology in bioinformatics and data analysis and has demonstrated the advantages of this technology with fully functional prototypes, including MetaR and the NextflowWorkbench.

campagnelab.org

fac2003@campagnelab.org

@FabienCampagne

Our group was drawn to Docker containers for their ability to provide a consistent and predictable deployment environment to facilitate training sessions for MetaR (http://metaR.campagnelab.org). While R and bioconductor packages may appear installed and functioning properly on a given machine, CRAN and Bioconductor provide no guarantee that installations of the same package(s) on other computers will be possible at a future time. As we experienced acutely during some training sessions for MetaR, the challenge of installing software on a diversity of end-user laptops can consume a large amount of the time that would be better spent explaining how to use the tool. To address these challenges, we have developed Docker images and the ability to run MetaR analyses directly and seamlessly inside a Docker container. Advantages and limitations of this approach will be discussed for data analysis with MetaR and the R language. In a second project, we are developing an interactive workbench for Nextflow (http://www.nextflow.io/), where we provide an interactive environment to facilitate the development of reproducible analysis workflows (such as those developed for analysis of high-throughput sequencing data, see [1]). In addition to the features provided by Nextflow, the Nextflow Workbench can act as a Docker WYSIWYG integrated development environment (http://campagnelab.org/software/nextflow-workbench/). For instance, scripts written for Nextflow with the Workbench provide auto-completion for files and software installed in the container associated with a Nextflow Process. Furthermore, domain specific languages can be developed and extended to facilitate the creation of families of Dockerfiles and associated images (e.g., for specific applications, such as those that require the installation of R packages in the container). These two projects were developed with language workbench technology [2-5] which supports some novel capabilities when working with containers, including WYSIWYG editing and seamless configuration and execution from the workbench.

1. Dorff, K. C. et al. GobyWeb: Simplified Management and Analysis of Gene Expression and DNA Methylation Sequencing Data. PLoS One 8, e69666 (2013).
2. Simi, M. & Campagne, F. Composable languages for bioinformatics: the NYoSh experiment. PeerJ 2, e241 (2014).
3. Benson, V. M. & Campagne, F. Language workbench user interfaces for data analysis. PeerJ 3, e800 (2015).
4. Campagne, F. The MPS Language Workbench: Volume I. (Fabien Campagne, 2014).
5. Campagne, F. The MPS Language Workbench: Volume II. (Fabien Campagne, 2015).

Weill Cornell Medical College

BRAD CHAPMAN

HARVARD T. H. CHAN SCHOOL OF PUBLIC HEALTH

Monday, November 9 at 15:20

Brad Chapman is a Research Scientist in the Department of Biostatistics. He has spent more than 15 years answering biological questions with computational approaches after switching over from a background in wet-lab research.He combines automated high-throughput analysis pipelines with custom visualization and processing tools.By utilizing a wide variety of languages, he strives to maximize code re-use while maintaining the flexibility to answer highly-specific collaborative questions.

Brad is involved in the open source community as a member of the Open Bioinformatics Foundation, bcbio, Biopython and CloudBioLinux, as well as contributing regularly to freely available GitHub and Bitbucket repositories. He posts about his research on Blue Collar Bioinformatics and can be found On Twitter as chapmanb. He has a PhD in plant biology from the University of Georgia.

Github

bcb.io

@chapmanb

Blue Collar Bioinformatics (bcbio; https://github.com/chapmanb/bcbio-nextgen) provides validated variant calling and RNA-seq analyses in a configuable, scalable implementation. Supporting installation of bcbio in a large variety of heterogeneous clusters is a major challenge. To mitigate this, we migrated to a Docker-based infrastructure that complements our standard distribution. Both versions use the same processing code, but the Dockerized application is dramatically easier to install with a single download of all tools and data. We used this to provide a ready to run implementation of bcbio on Amazon Web Services and plan to increasingly move to Docker-based distributions running on top of generalized workflow platforms implementing the Common Workflow Language (CWL). We look forward to discussing ways of working more closely with other Docker-enabled bioinformatics tools.

bcbio

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

ELIJAH CHARLES

Intel

Monday, November 9 at 16:00

Architect, Life Sciences and Analytics at Intel Corporation. Specialist in the championing and leading of enterprise wide technology change with strong emphasis on loosely coupled systems and expressive event sourcing.

elijah.charles@gmail.com

A talk highlighting the discrepancies in packaging and versioning of bioinformatic tooling. How this relates to the early days of packaging in the FOSS world and how docker helps get around many of the issues.

KAI DAVENPORT

Cluster HQ

Monday, November 9 at 16:25

Kai works on the developer relations team at ClusterHQ - the creators of Flocker. He has been busy working on the Docker Volume Plugin and previously was hard at work on Powerstrip - a prototyping tool for Docker extensions. In a previous life Kai was developing educational software and has been developing web-based software for 15 years.

@kai_davenport

Containers has taken the devops community by storm by providing a simple, standardized way to package applications between dev, staging and production environments. The one major missing piece, however, is support for databases which continue to be managed by external services.

We believe that databases should also be able to run on the same platform as the rest of the app and benefit from container's portability benefits.

In this presentation, Kai will discuss what data-focused container clustering could look like in a world where the entire application, including its data services, runs inside containers."

MICHAEL MARKIE

F1000

THOMAS INGRAHAM

F1000

Tuesday, November 10 at 9:45

Thomas Ingraham is Publishing Editor at the life science-orientated Open Science publisher, F1000Research. His main focus is on working with research communities to set up and develop their own publishing platforms, as well as managing various open data and software projects.

thomas.ingraham@f1000.com

@t_ingraham

Reproducible research is at the core of many of F1000Research's activities; it is one of the reasons why we require authors to make all data and code associated with their articles openly available. Such practices can allow others to try to reproduce published computational analyses, but not always and not easily. This is where Docker and other means of containerization hold great promise.

This talk will be used to announce the launch of the Container Virtualization Channel on F1000Research, which aims to help realize this potential. This channel has been created for those whose research involves containerization; a place to publish their work and have it collated in one centralised venue that is owned and shaped by the community. I will also take the opportunity to provide a brief overview of F1000Research's unique publication model and open science ethos.

F1000Research
Open for Science

ALFONSO ACOSTA

Weaveworks

Tuesday, November 10 at 10:00

I am a seasoned Software Engineer, interested in product management and thrilled by intellectually-appealing challenges related to computer science, especially when R&D, systems programming and/or functional programming are put into the mix.

I recently joined http://weave.works, the new startup of the RabbitMQ founders, where we develop open source software to help connect, observe and control Docker containers.

@2opremio

Weaveworks is the software company that develops Weave - the most productive way for developers to connect, observe and control Docker containers.

In this presentation, Alfonso will introduce Weave Net/Run/Scope and will focus on how they integrate with Amazon's EC2 Container Service.

weaveworks

AANAND PRASAD

Docker

Tuesday, November 10 at 10:40

Github

aanand.prasad@gmail.com

@anand

Docker makes packaging and running software in individual containers easy, but most apps are made of more than one container. I'll show you how to use Docker Compose to effortlessly define and run applications made of multiple, interdependent containers. I'll also demonstrate how to use Compose in conjunction with two other Docker tools - Machine and Swarm - to run a multi-container app on a compute cluster using the same configuration and commands you use locally.

docker

MATTHEW BATES

Jetstack

November 10 at 11:05

Matthew is co-founder at Jetstack and looks after Engineering and Consulting. He has a background in solutions for the acquisition, management and exploitation of large-scale data, across a variety of industries. He was previously at MongoDB, Deutsche Telekom and Detica since graduating from University of Nottingham in Computer Science.

He uses Java, Python and Go and has in the previous year been working closely with the open source container management system Kubernetes, including code contributions and customer deployments.

@mattbates25

Kubernetes is an open source version of the Borg container cluster manager that has powered Google's infrastructure for over a decade. Google launch several billion containers a week, powering everything from GMail to Maps, Android Play and even their public cloud environment, Google Cloud Platform.Since first launching at OSCon in 2014, Kubernetes has matured to a production-ready 1.0 release, with contribution from 100s of developers from across all industries. At just over 10,000 stars on Github and almost 20,000 commits, it is one of the most active open source projects in development.

This talk will introduce Kubernetes and explain its core concepts. It will show how Google's secret-sauce is now open to us all for managing containerised applications. There will be demos of how it can be used to efficiently schedule and orchestrate mixed workloads in a cluster, including scalable web-based applications and NoSQL, as well as batch processing of large-scale data.

JETSTACK©

CLIVE STRINGER

King's College Hospital

Clive Stringer started work as a Microbiologist at King's College Hospital 35 years ago. He gradually moved into software development and IT. For 12 years he has been System Delivery Manager at the Trust, during which time King's rose to the top of the Clinical Digital Maturity Index in England. He has a particular interest in Genomics and interoperability.

ADAM HATHERLY

HSCIC

Adam Hatherly is a senior solution architect working for the Health and Social Care Information Centre. He has been working on technology solutions for health and social care informatics for around 7 years, including work on national systems as part of the NHS national programme for IT. He has also been involved in driving the strategy for interoperability, both within HSCIC and across health and social care by developing national resources and standards to support local delivery of solutions. A passionate advocate of open source technology, Adam is involved in a number of open-source integration initiatives, and still gets hands-on and contributes to open source projects whenever he can.

Tuesday, November 10 at 11:50

King's College Hospital are working with a number of partners on components designed to run in Docker containers which will improve the usability of health data captured at King's and to facilitate transfer of structured data between care providers.

The first is a range of components developed by the bioinformatics team in KCL which defines structured meaning from historical data . This should allow us to define Snomed CT codes from a rich source of data collected in free text documents since the Electronic patient record went live at King's 15 years ago. The first application of this technology will be to support the 100K genome project, both to find candidates for the project and to provide structured phenotypic data.

The second area of functionality is around inter-operability. King's College Hospital has an Open source integration engine which handles clinical messages within the hospital in a structured way using HL7 but there is little structured data passing between Acute Trusts and other care providers. We have been working closely with the HSCIC on components which will take our unstructured discharge summaries and turn them into a structured CDA document conforming to new standards for "Transfer of Care" documents.

King's College Hospital NHS Foundation Trust

hscic Health & Social Care Information Centre

YANNICK WURM

Queen Mary University of London

November 10 at 12:15

Lecturer in Bioinformatics at Queen Mary University of London

Github
y.wurm@qmul.ac.uk
@yannick__

Genomic analyses require jumping back and forth between many bioinformatics tools which are often difficult to install. Furthermore, it is challenging to keep different versions of software for different projects, yet changing versions can make analyses difficult to reproduce. To make matters worse, genomicists often lack the skills necessary to setup complex bioinformatics software, and systems administrators can be overwhelmed by large numbers of software installation requests.

We have developed oswitch to enable agile, seamless switching from one operating system to another – providing access to diverse ranges of tools. This project grew from our own need to rapidly access diverse pieces of specific versions of software including those distributed as part of BioLinux on our MacBooks and our university HPC system.

oswitch is a wrapper facilitating access to docker images. Importantly, when switching operating systems inside a shell, most things remain unchanged:

- Current working directory is maintained
- User name, uid and gid are maintained
- Login shell (bash/zsh/fish) is maintained
- Home directory is maintained (thus all .dotfiles and config files are maintained).
- read/write permissions are maintained
- Paths are maintained whenever possible.

Queen Mary University of London