iRODS UGM 2019
Utrecht University
Netherlands
Jun 25-28, 2019

Bristol-Myers Squibb iRODS Journey

Employing iRODS to manage petabytes of genomics data on cloud

Oleg Moiseyenko
Sr. Scientific Cloud Engineer
Bristol-Myers Squibb Company

WORKING TOGETHER FOR *Patients*™

Bristol-Myers Squibb

COMPANY OVERVIEW

"If you're going to fight the battle of your life, you've got to stay positive – in the midst of any storm, there's always something to be grateful for."

Carol Willis
Renal cell carcinoma patient, benefiting from an *Opdivo-Yervoy* combination

Latest update - April 2019

Our Mission

To **discover**, **develop** and **deliver** innovative **medicines** that help patients prevail over serious diseases.

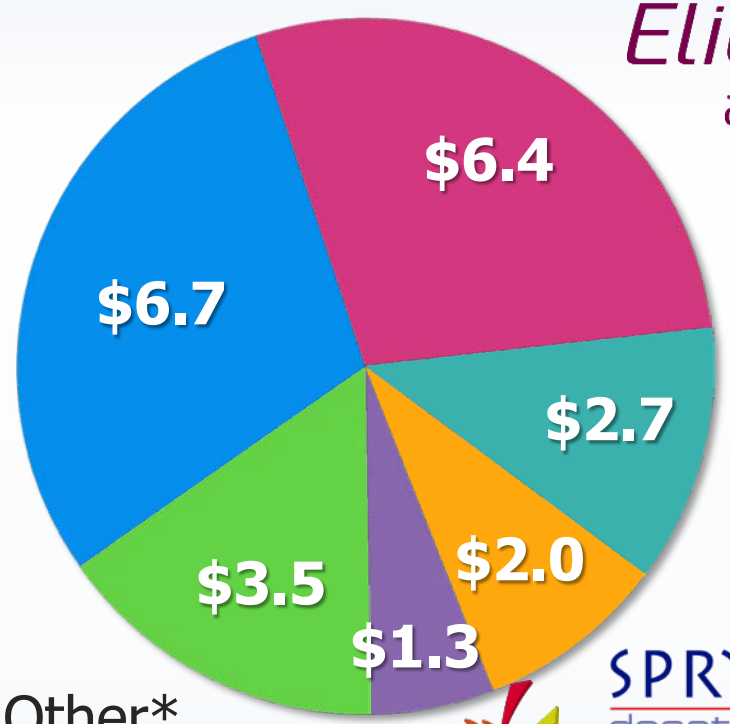
Bristol-Myers Squibb Delivering in 2018

DELIVERING by the **NUMBERS**

$22.6
BILLION in Revenue

9%
Revenue Growth VS. 2017

PRODUCT REVENUES
$ BILLIONS



Eliquis apixaban

OPDIVO (nivolumab) INJECTION FOR INTRAVENOUS USE 10 mg/mL

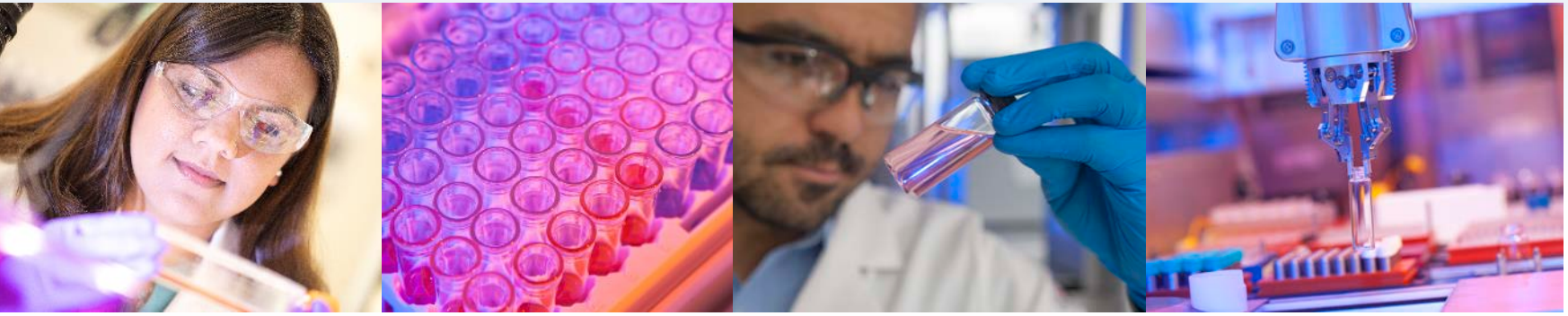
ORENCIA (abatacept)

SPRYCEL dasatinib 100 mg tablets

YERVOY (ipilimumab) Injection for intravenous infusion

* Includes *Empliciti, Baraclude, Sustiva, Reyataz,* Hepatitis C franchise and Other Brands

R&D: Delivering Innovative Medicines to Patients



40 compounds in development

12 new medicines for Patients since **2011**

~5,700 R&D Colleagues Worldwide

R&D Investment IN 2018

$5.1 BILLION on a non-GAAP basis*

5 PERCENT Increase over 2017.

*This non-GAAP amount excludes significant upfront and milestone payments for business development transactions and other specified R&D items. A reconciliation of GAAP to non-GAAP measures can be found on our website at www.bms.com. The GAAP amount is $6.3B.

Data as of January, 2019

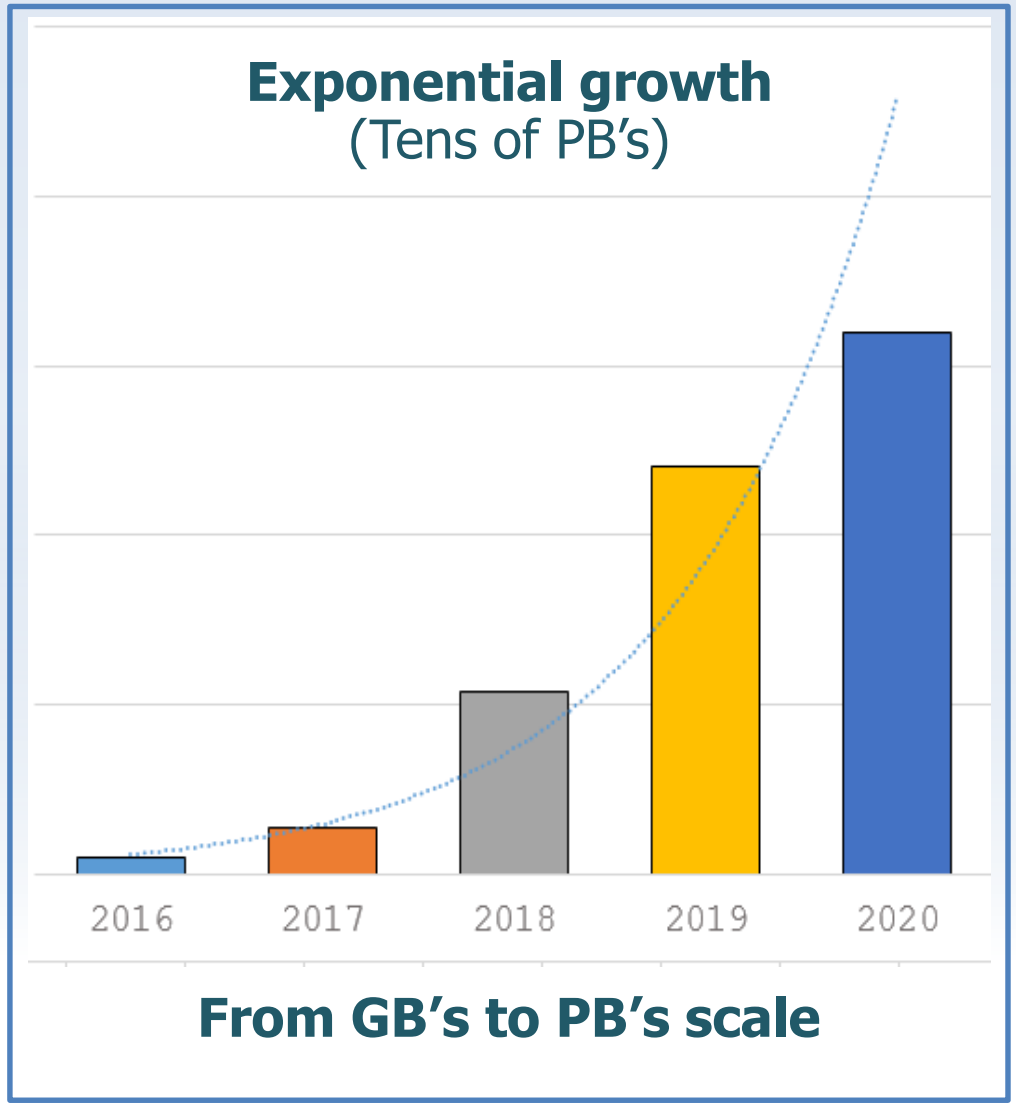
It's all about data, Big Data!

Scientific data sets

- NGS data
- Proteomics
- Flow Cytometry
- Imaging data
- High-Throughput screening
- Mass spectrometry
- Databases

Data governance

- 25 years of retention
- Backups



Major data sources

- Raw data from labs
- Scratch space
- Results data
- External collaborations
- Public & government agencies
- R&D

Key considerations for data management system

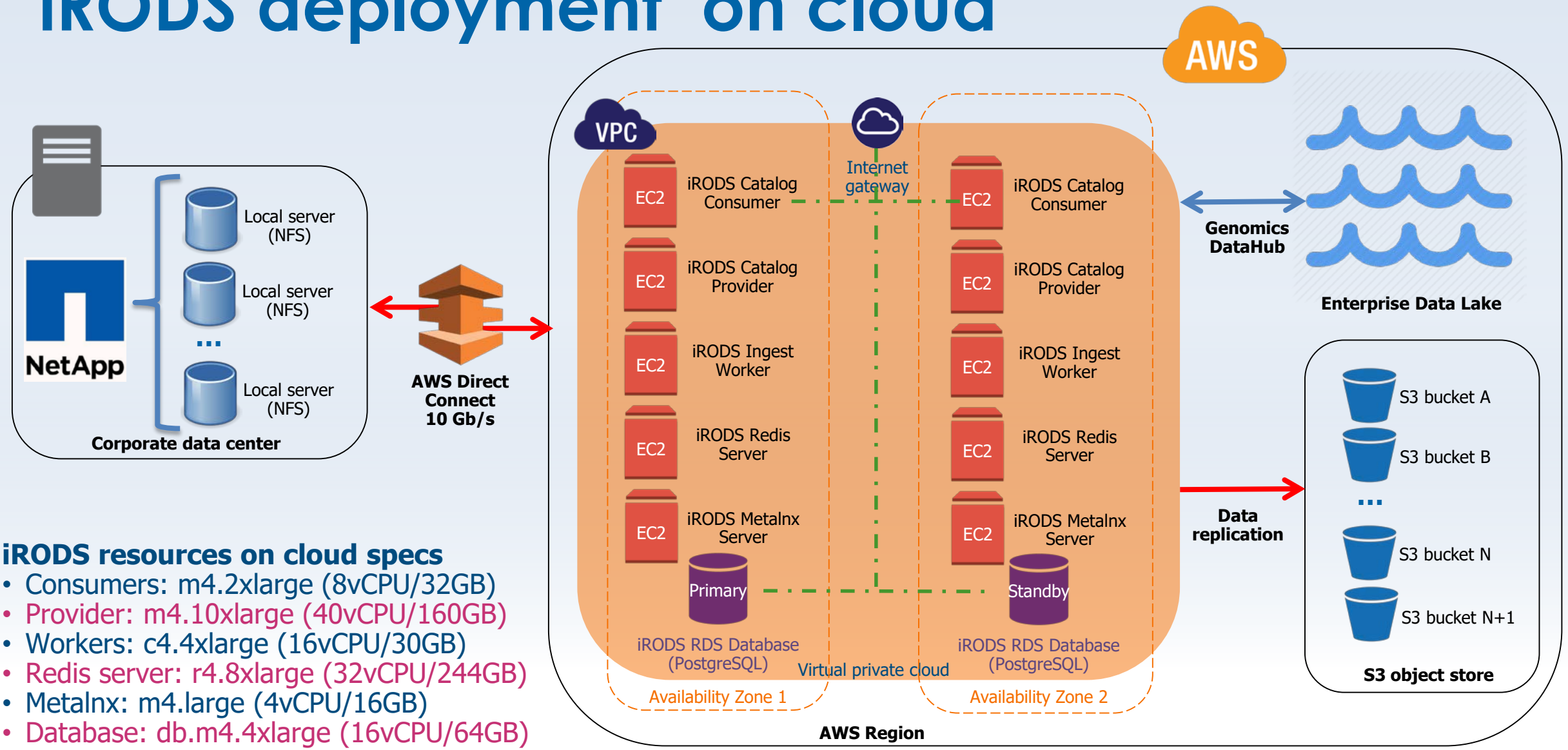
BMS acceptance criteria

- **Cloud integration**
- **Petabyte scalable**
- CLI interface
- Rich API
- **Metadata driven**
- NFS – S3 connectivity
- User's access management
- **Security**
- Low price tag
- Low administrative efforts
- **Established presence in life science & healthcare**
- Support

Cloud advantages

- S3 object store
  - Unlimited size
  - Data protection: 99.999999999% durability
  - Build-in data distribution & replication
  - Easy integration with other cloud micro services
- No hardware / storage technology lock-in
- Cloud elasticity: vertical & horizontal
- Backups (versioning, snapshots, lifecycle rules)
- PaaS platform for database technologies
- High data security
- Low cost

iRODS deployment on cloud



iRODS resources on cloud specs

- Consumers: m4.2xlarge (8vCPU/32GB)
- Provider: m4.10xlarge (40vCPU/160GB)
- Workers: c4.4xlarge (16vCPU/30GB)
- Redis server: r4.8xlarge (32vCPU/244GB)
- Metalnx: m4.large (4vCPU/16GB)
- Database: db.m4.4xlarge (16vCPU/64GB)

iRODS use cases

Data management

- Moving data from labs to cloud
- Managing various scientific datasets
- Providing access to clinical data sets

NFS/S3 data sync

- Sync S3 object store with on-prem data stores (NFS)
- Confirm no deltas left
- Provide logs for audits
- Unmount local storage

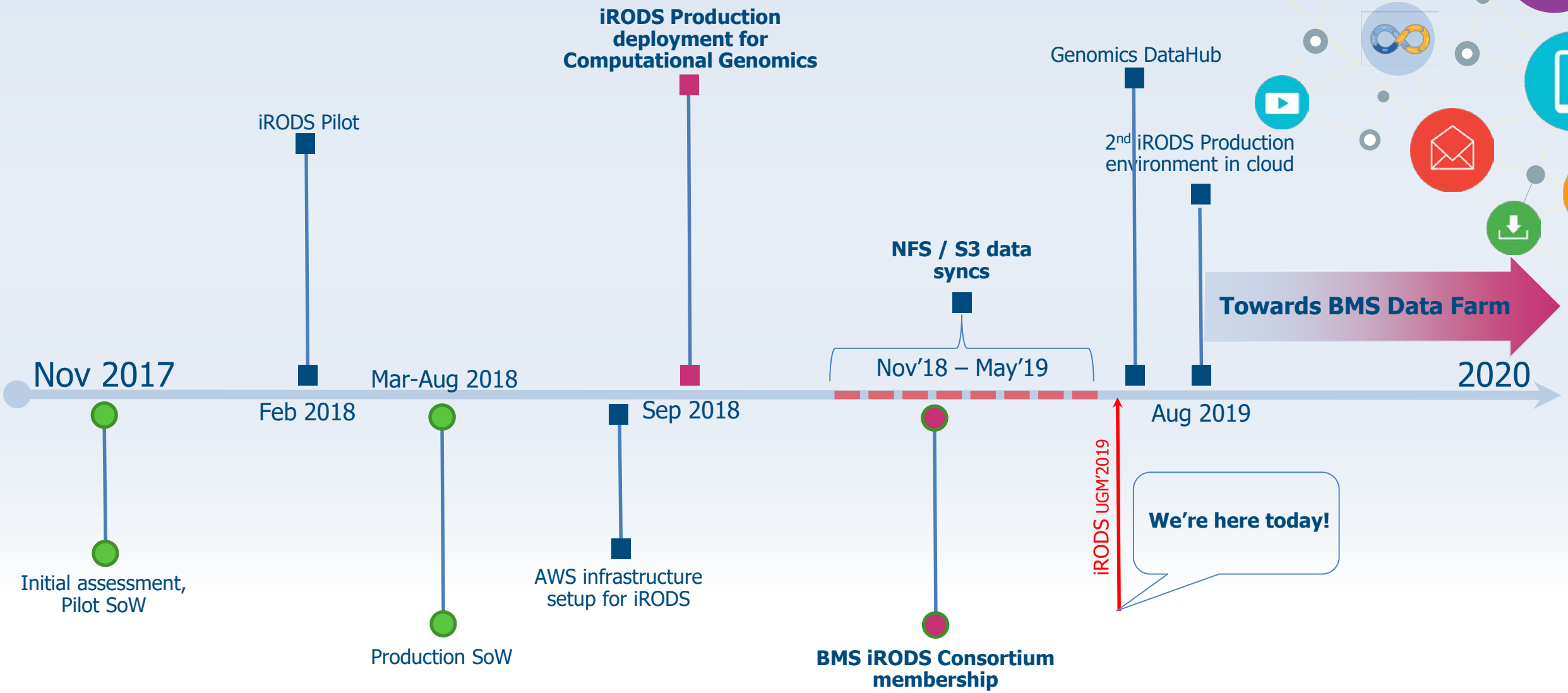
ML based data enrichment

- ML and deep learning algorithms classify image data
- iRODS catalog is updated with tags with classification information

Data Lake integration

- Integrate iRODS meta data catalog with Clinical data lake
- Enterprise Data lake ingestion tools use iSQL to read iRODS meta data catalog

Our journey to BMS iRODS Data Farm



iRODS: Pros & Cons

Pros

- Easy to deploy
- Metadata driven
- Flexible rule engine
- Same names for logical/physical file paths
- Established presence in life science
- Rich API
- Data virtualization
- Flexible & PB-scalable system
- High data retention requirements (10-25 years)
- ACL's and permissions support
- Secure data sharing
- Workflows automation & data replication

Cons

- Higher complexity level
- Requires advance development
- No mechanism to enforce good metadata system ("garbage in, garbage out")
- No user-friendly front-end interface

Challenges

- MD5 checksums
- Scanning speed: every million files on S3 takes about two hours to scan on the NFS side
- Data replication speed
- Non-readable characters in file names
- Permission issues
- Redis cache issue (once)
- Verification upon data sync process completion

BMS Wishlist

- NFSRODS integration
- Minio iRODS Gateway
- Better LDAP/AD integration
- Metadata templates
- iRODS catalog structure specs
- Advance SQL support
- Push notifications instead of polling
- Database performance optimization
- User-friendly front end
- Improved documentation
- AWS EC2 spot instances for workers

BMS – iRODS: Next steps

- Capture, manage, apply metadata to data collections
- Deliver continuous data scans for S3 store
- Unify access to metadata; metadata enrichment
- Unify the governance approach for iRODS
- Advance development: rules, policies, etc.
- Genomics DataHub (gateway to BMS data lake)
- LDAP integration for user's authentication
- Dashboarding/system health (iRODS audit plugin)
- Towards BMS Data Farm (zones federation)

Acknowledgements

BMS iRODS Core Team

- Mohammad Shaikh
- Isaac Neuhaus
- Carlos Rios
- Mark Russo
- Oleg Moiseyenko

iRODS Consortium Team

- Jason Coposky
- Terrell Russell

BMS iRODS Cross Team

- Dan Huston
- Valerie Williams
- Eric Sison
- Dmitry Khavich
- Paul O'Malley
- Gopal Prakriya

- Sponsor, Business Partner: Ajay Shah

WORKING TOGETHER FOR Patients™

Bristol-Myers Squibb

Produced by Bristol-Myers Squibb, Corporate Affairs & Scientific Computing Services.

www.bms.com