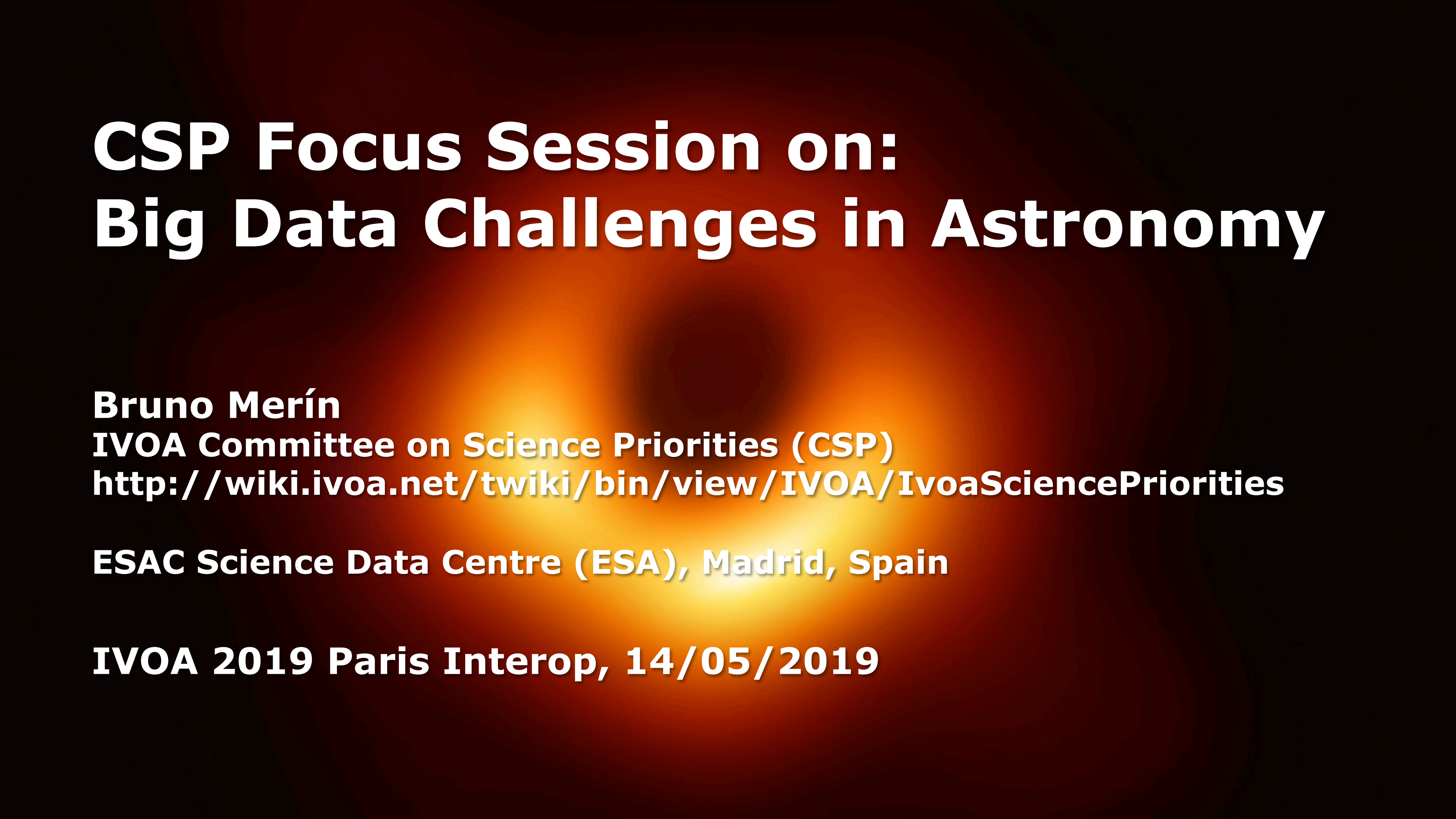


CSP Focus Session on: Big Data Challenges in Astronomy

Bruno Merín
IVOA Committee on Science Priorities (CSP)
http://wiki.ivoa.net/twiki/bin/view/IVOA/IvoaSciencePriorities

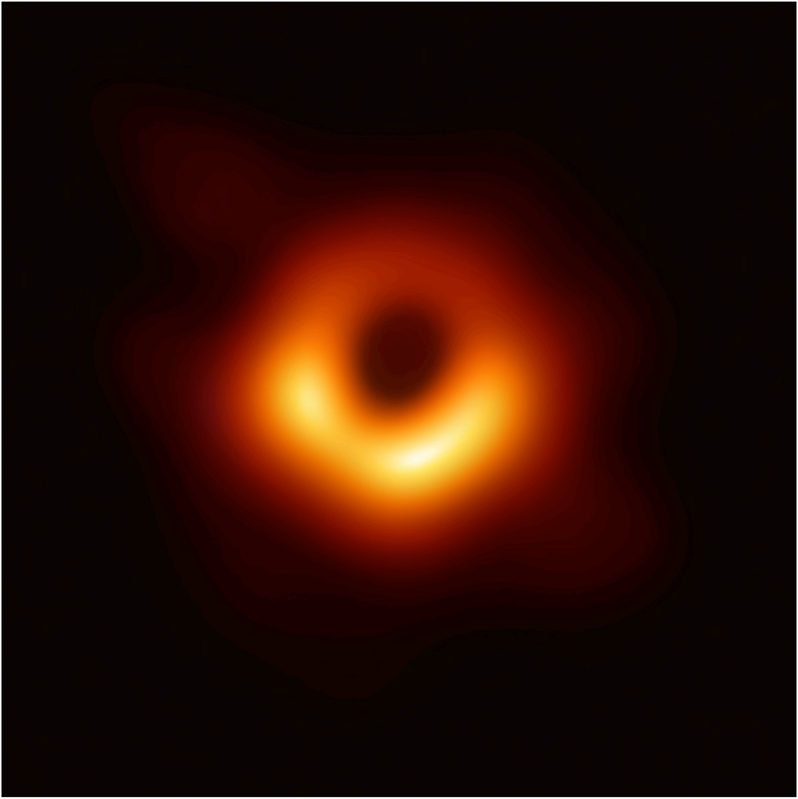
ESAC Science Data Centre (ESA), Madrid, Spain

IVOA 2019 Paris Interop, 14/05/2019

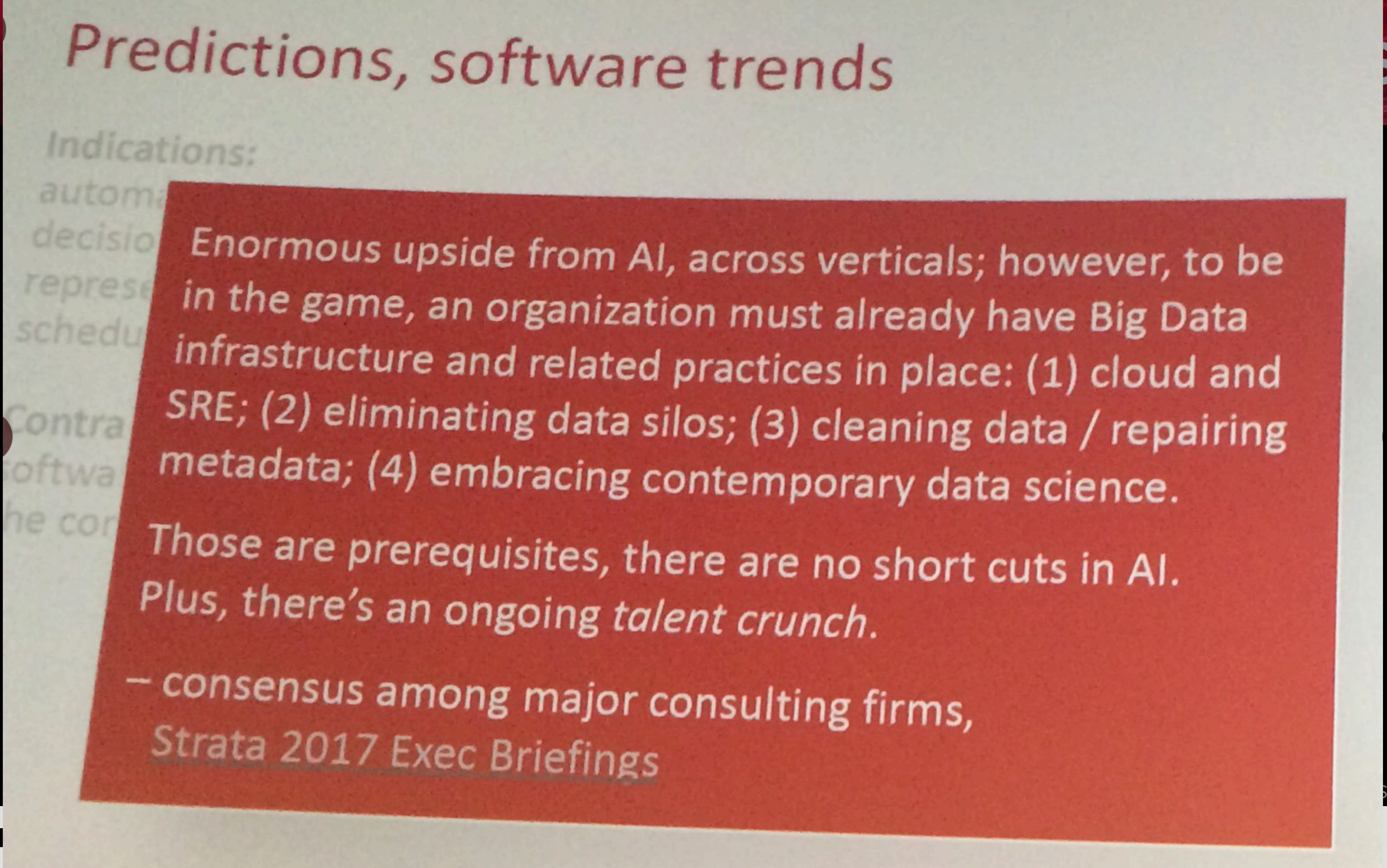
Outline

1. Motivation
2. Session contents
3. What happens next

Motivation: to enable more science !



- 5 PBs of data, collected by plane
- Months of data processing



Predictions, software trends

Indications:

autom

decisio

represe

schedu

Contra

oftwa

he cor

Enormous upside from AI, across verticals; however, to be in the game, an organization must already have Big Data infrastructure and related practices in place: (1) cloud and SRE; (2) eliminating data silos; (3) cleaning data / repairing metadata; (4) embracing contemporary data science.

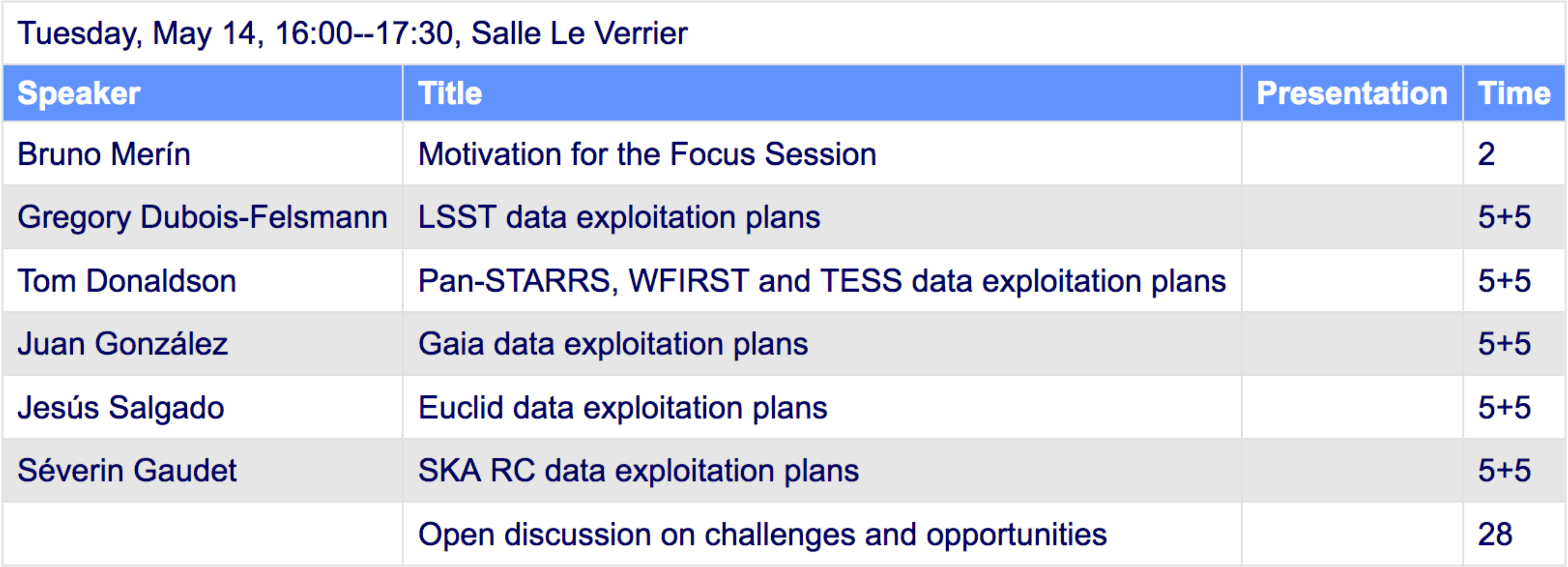
Those are prerequisites, there are no short cuts in AI. Plus, there's an ongoing *talent crunch.*

– consensus among major consulting firms, Strata 2017 Exec Briefings

Goals of the session

- To review the future plans of surveys producing big data
- To identify the challenges to the data exploitation
- To identify the possible solutions to those challenges
- To identify which IVOA standards could help

Agenda for the session (all presentations online)



Tuesday, May 14, 16:00--17:30, Salle Le Verrier

Speaker	Title	Presentation	Time
Bruno Merín	Motivation for the Focus Session		2
Gregory Dubois-Felsmann	LSST data exploitation plans		5+5
Tom Donaldson	Pan-STARRS, WFIRST and TESS data exploitation plans		5+5
Juan González	Gaia data exploitation plans		5+5
Jesús Salgado	Euclid data exploitation plans		5+5
Séverin Gaudet	SKA RC data exploitation plans		5+5
	Open discussion on challenges and opportunities		28

Questions for the large surveys

Questions for large surveys :

Q1: Describe the data volumes and types of data expected from the mission/survey.
Q2: Describe your data dissemination/exploitation plan for users.
Q3: Are you looking at sending data to users or looking at a code to the data approaches?
Q4: How would you cross-correlate data with different surveys?
Q5: How and where does the IVOA fit into your plans?

Survey	Q1: max. data volumes & dates	Q2: plans	Q3: code to data?	Q4: X-correlate?	Q5: IVOA?
LSST	30e12 sources 15 PBs (2023 to 2033)	Portal + Notebook + Web APIs	Yes	At least w/ Gaia	TAP+ADQL, SIAv2, SODA, VOSpace, WebDAV, PyVO
Pan-STARRS	11e9 sources 1.4 PBs (Jan 2019)	MASP portal + astroquery	AWS ?	?	MAST API, TAP, Cone, likely DataLink, SODA
WFIRST	20 PBs (~2025 TBC)	MASP portal + astroquery	AWS ?	?	MAST API, TAP, Cone, likely DataLink, SODA
TESS	260 TBs (2018 – 2020+)	MASP portal + astroquery	AWS ?	?	MAST API, TAP, Cone, likely DataLink, SODA
Gaia	2 PB (2018 - 2027)	Gaia archive + astroquery	SEPP	Several all-sky cats.	TAP+, datalink, SODA, VOSpace
Euclid	20 PB (2022 - 2028)	Euclid archive + astroquery	SEPP	Several all-sky cats.	TAP+, datalink, SODA, VOSpace
SKA	600 PB/year (2028 -)	TBD	TBD	TBD	TBD
ZTF	63e9 sources 1.5 PBs (2018 -)	IRSA APIs	?	?	IRSA APIs

Challenges

- To distribute TBs of data to users
- Allow people discovering datasets from new surveys
- Provide sufficient computing resources for users (who pays?)

Possible solutions to the challenges

- To distribute TBs of data to users
  - Most users don’t need TBs, but just MBs or GBs
  - They will run jobs in a cloud and download less data
- Allow people discovering datasets from new surveys
  - Add data to registry?
- Provide sufficient computing resources for users (who pays?)
  - Clone data to commercial clouds?
- What is missing in the IVOA set of standards?
  - SODA for cut-outs, rebinning, resampling?
  - Interoperable Notebooks?
  - Hierarchical data structures for discoverability?
  - Standard for code-to-the-data?
  - New data distribution standards like torrent?