aarnet

CONFESSIONS OF A DATA WRANGLER

Dr Carina Kemp

Director of eResearch

Australian Academic and Research Network

IS A JOURNEY (OF DISCOVERY)
NOT A DESTINATION

What is a data wrangler?

Why are we forced to be wrangers?

Technology?
Friend or enemy…

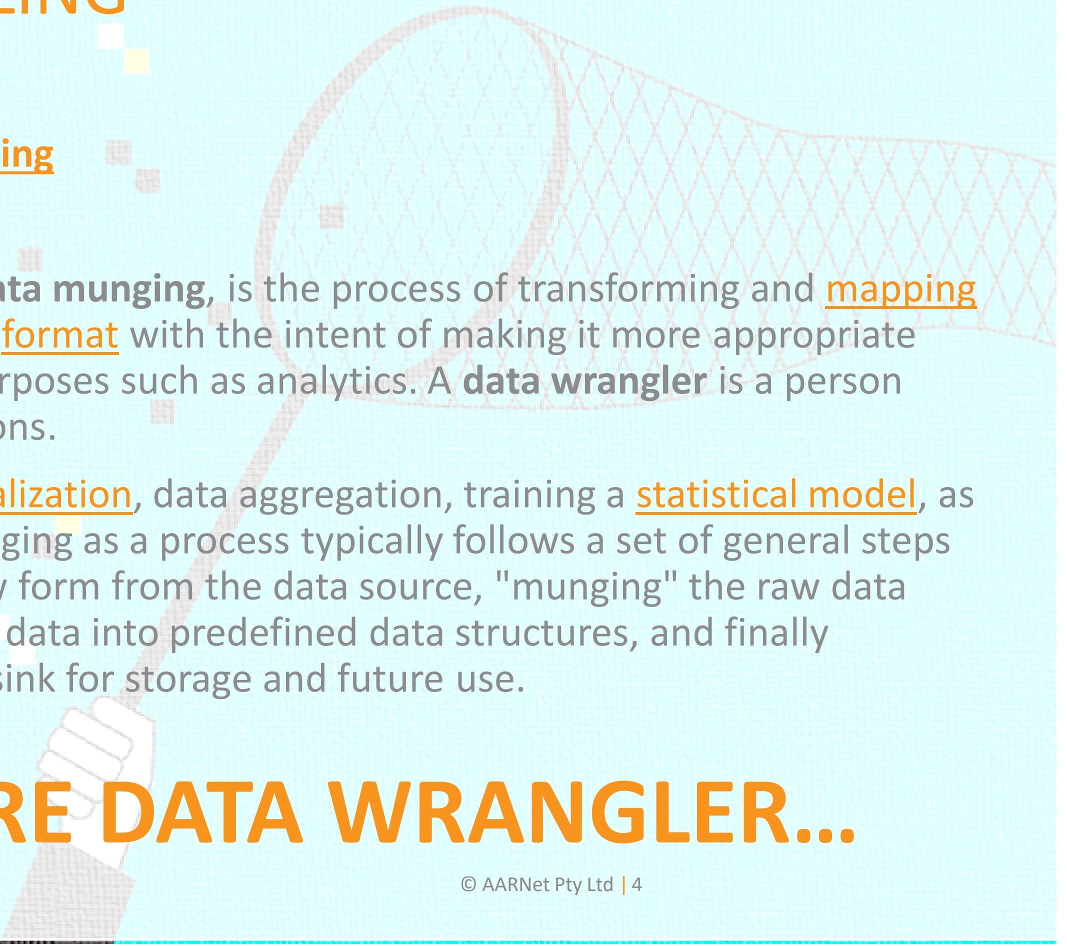
WHAT does this all mean?

WHAT IS A DATA WRANGLING

https://en.wikipedia.org/wiki/Data_wrangling

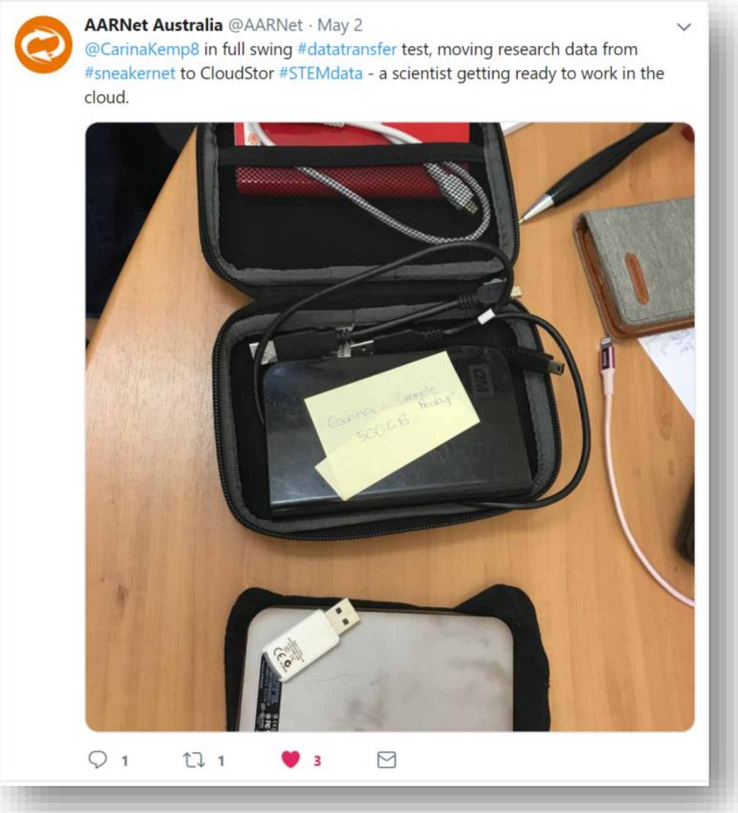
Data wrangling, sometimes referred to as **data munging**, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. A **data wrangler** is a person who performs these transformation operations.

This may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.



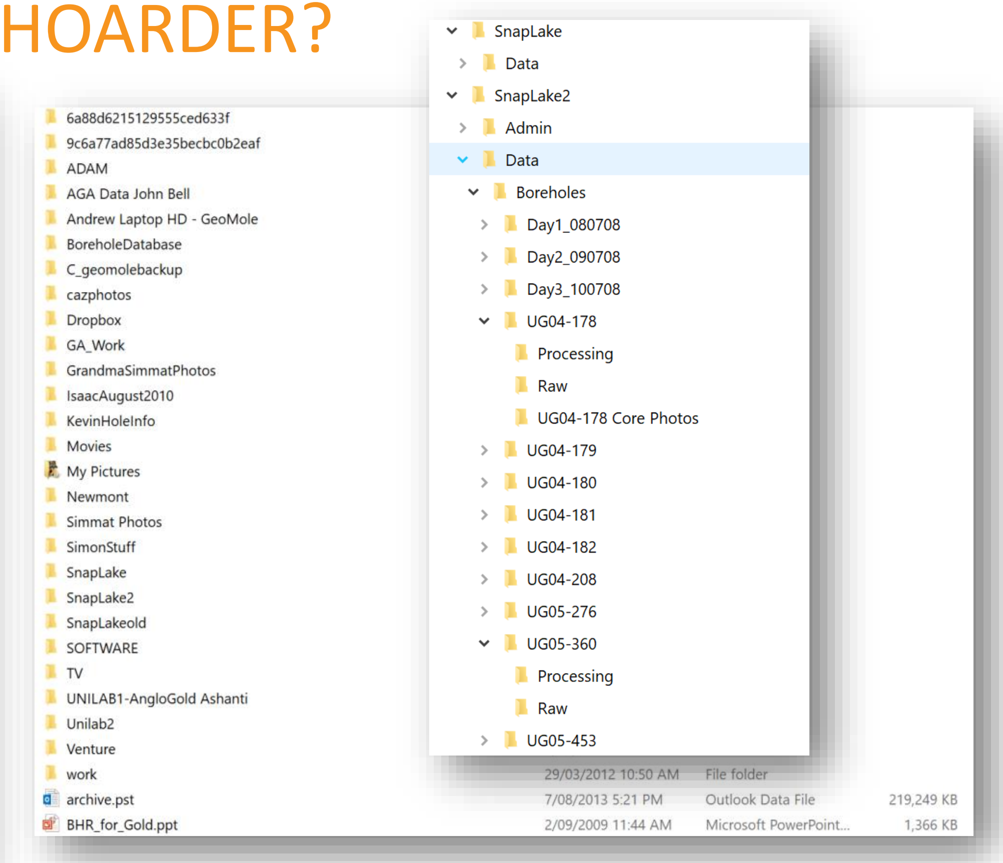
ALL SCIENTISTS ARE DATA WRANGLER…

CONFESSIONS OF A DATA HOARDER?



AARNet Australia @AARNet · May 2

@CarinaKemp8 in full swing #datatransfer test, moving research data from #sneakernet to CloudStor #STEMdata - a scientist getting ready to work in the cloud.



6a88d6215129555ced633f
9c6a77ad85d3e35becbc0b2eaf
ADAM
AGA Data John Bell
Andrew Laptop HD - GeoMole
BoreholeDatabase
C_geomolebackup
cazphotos
Dropbox
GA_Work
GrandmaSimmatPhotos
IsaacAugust2010
KevinHoleInfo
Movies
My Pictures
Newmont
Simmat Photos
SimonStuff
SnapLake
SnapLake2
SnapLakeold
SOFTWARE
TV
UNILAB1-AngloGold Ashanti
Unilab2
Venture
work 29/03/2012 10:50 AM File folder
archive.pst 7/08/2013 5:21 PM Outlook Data File 219,249 KB
BHR_for_Gold.ppt 2/09/2009 11:44 AM Microsoft PowerPoint... 1,366 KB

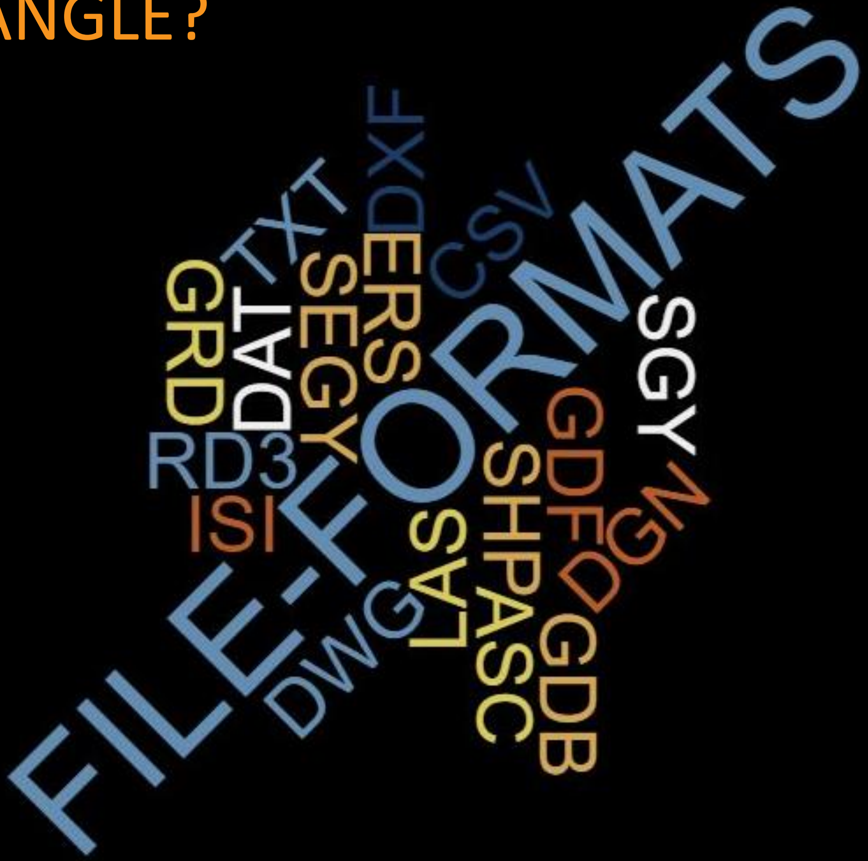
```
SnapLake
  Data
SnapLake2
  Admin
  Data
    Boreholes
      Day1_080708
      Day2_090708
      Day3_100708
      UG04-178
        Processing
        Raw
        UG04-178 Core Photos
      UG04-179
      UG04-180
      UG04-181
      UG04-182
      UG04-208
      UG05-276
      UG05-360
        Processing
        Raw
      UG05-453
```

WHY DO I WRANGLE?

data-swamp
interoperability
collaborate
flexibility
usable
workflow
software
Wrangling?
lazy
proprietary
formats
open-source
longtail
reusable
exchange

Source: https://www.wordclouds.com/

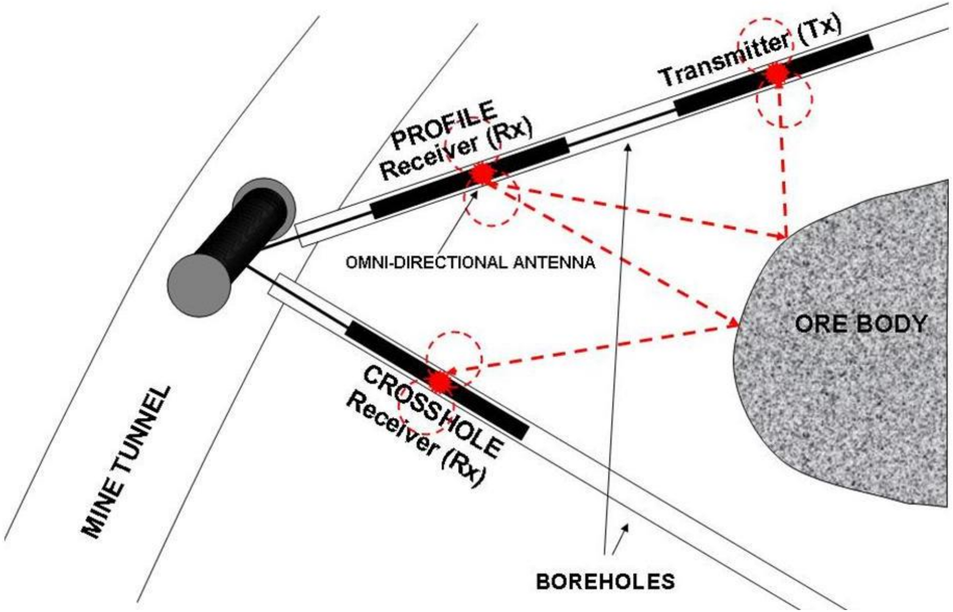
WHAT DO I WRANGLE?

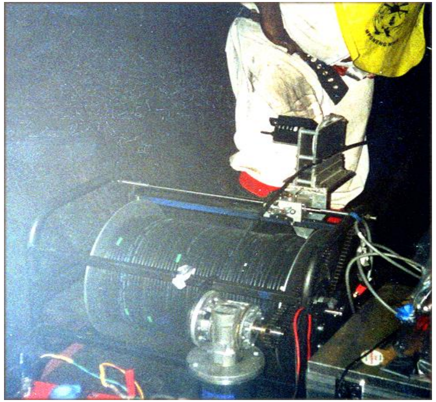


Source: https://www.wordclouds.com/

THE DATA WRANGLING BEGAN…

"MINE-SCALE THREE DIMENSIONAL BOREHOLE RADAR (BHR) IMAGING"





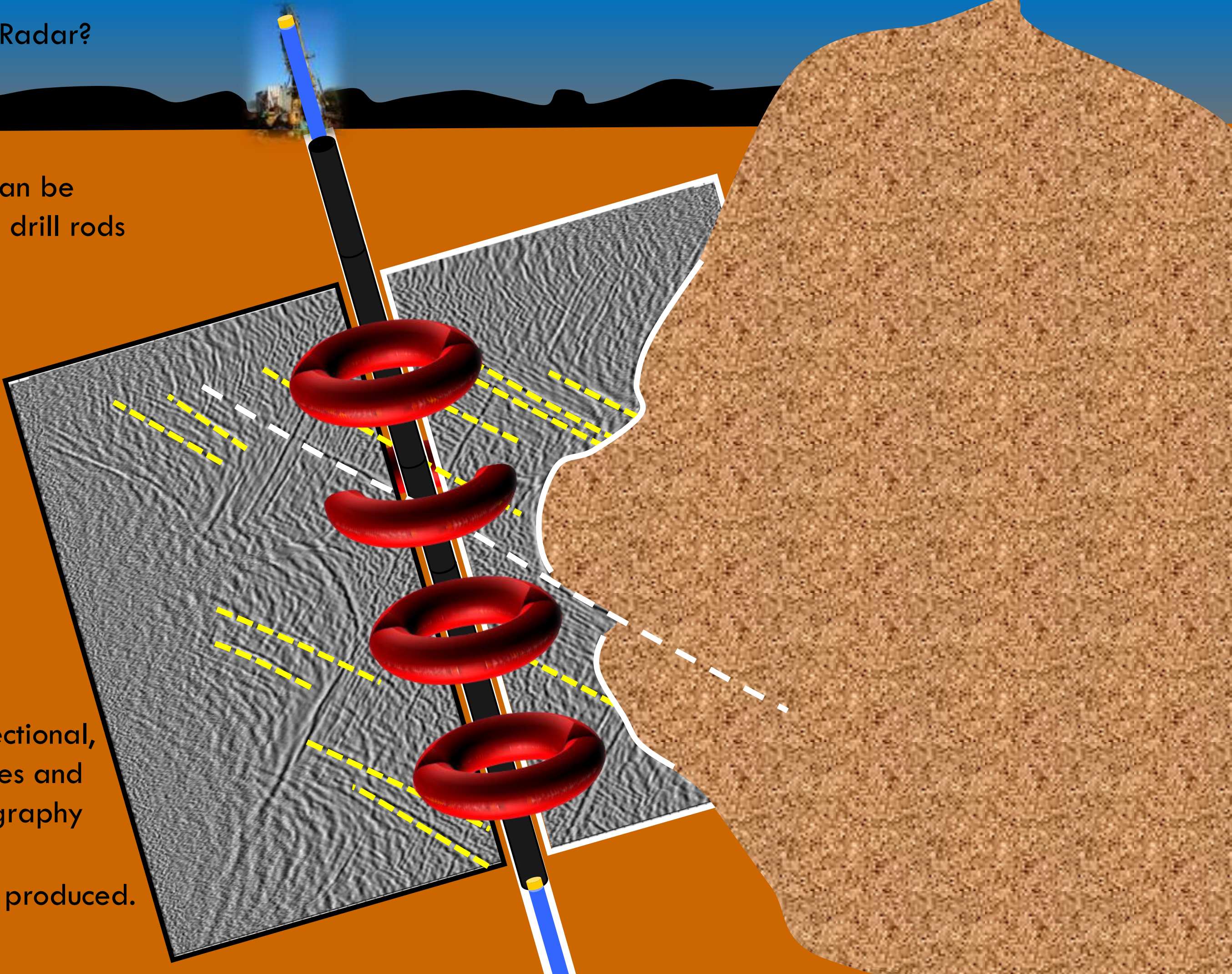
What is Borehole Radar?

The borehole radar system can be deployed by winch or on the drill rods similar to a gyro survey

Data is acquired continuously as the rods are pulled and the radar ascends the drillhole

Signal is sent radially outwards into the surrounding rock

The radar images the rock surrounding the drillhole.

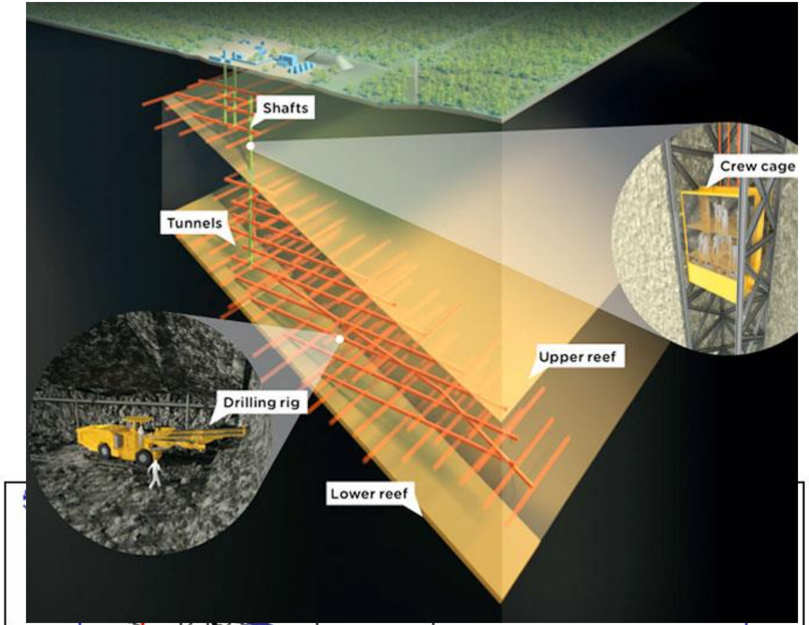


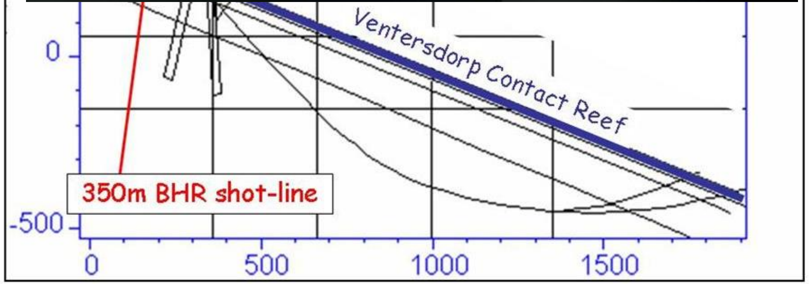
The radar is not directional, Neighboring drillholes and knowledge of stratigraphy aids interpretation.

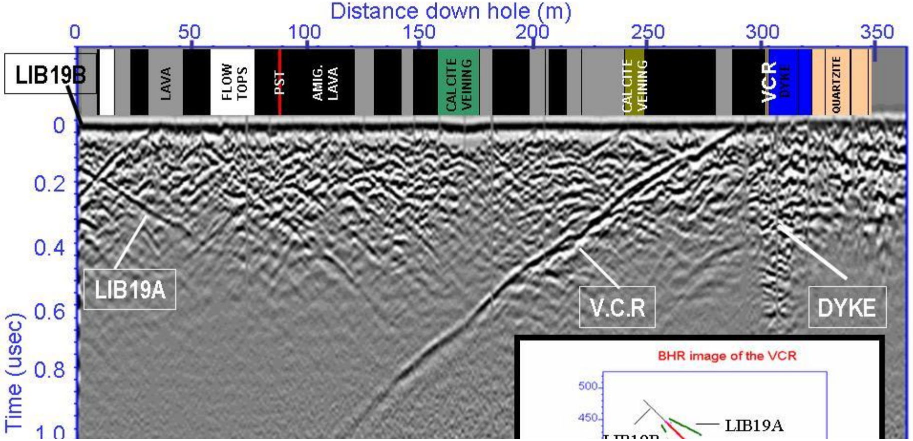
Final interpretation is produced.

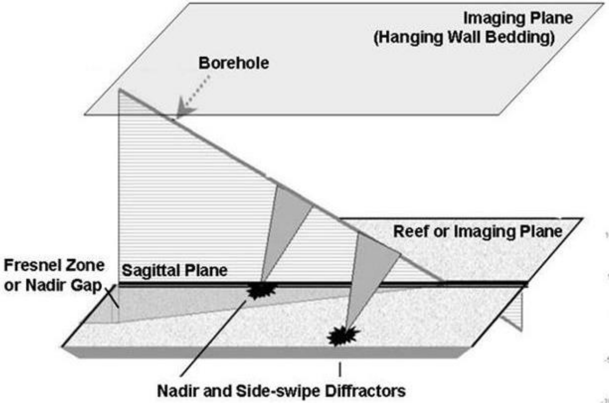
https://www.wired.com/2011/02/st_ultradeepmines/

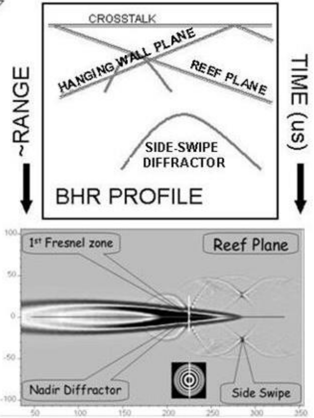
THE DATA…



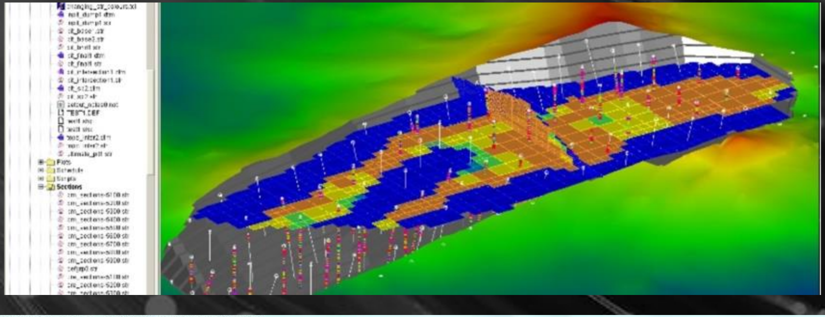


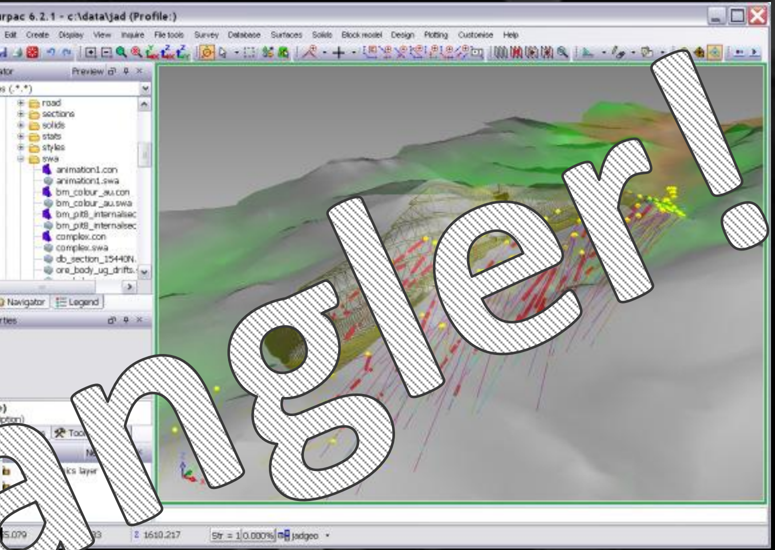


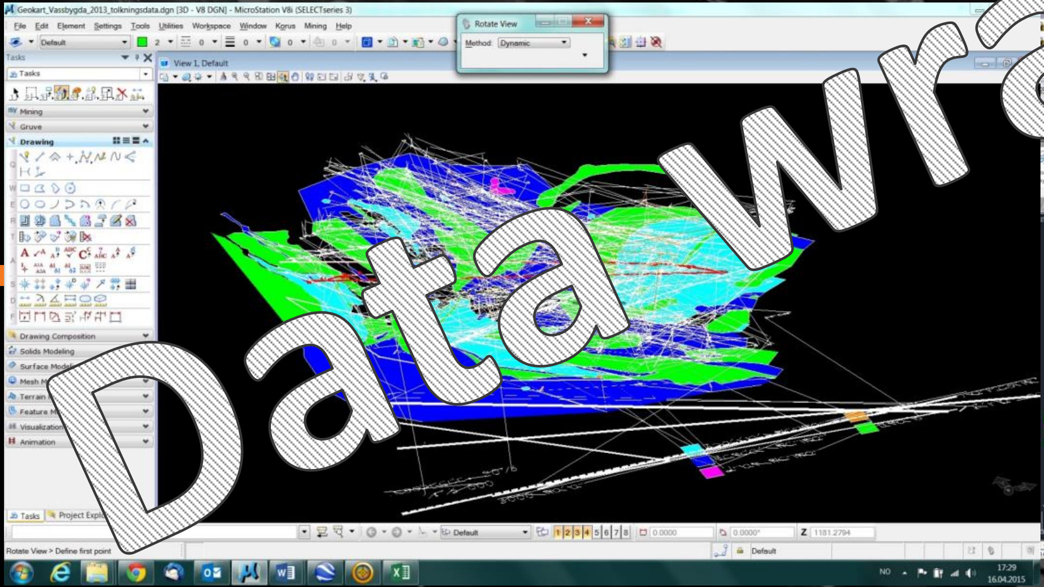


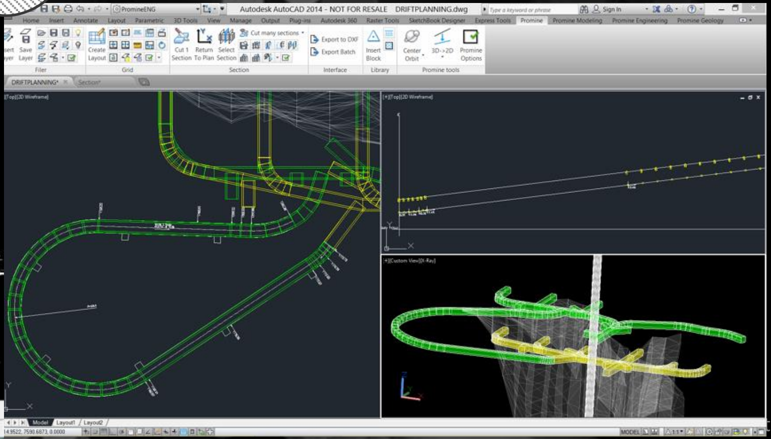


THE SOFTWARE…







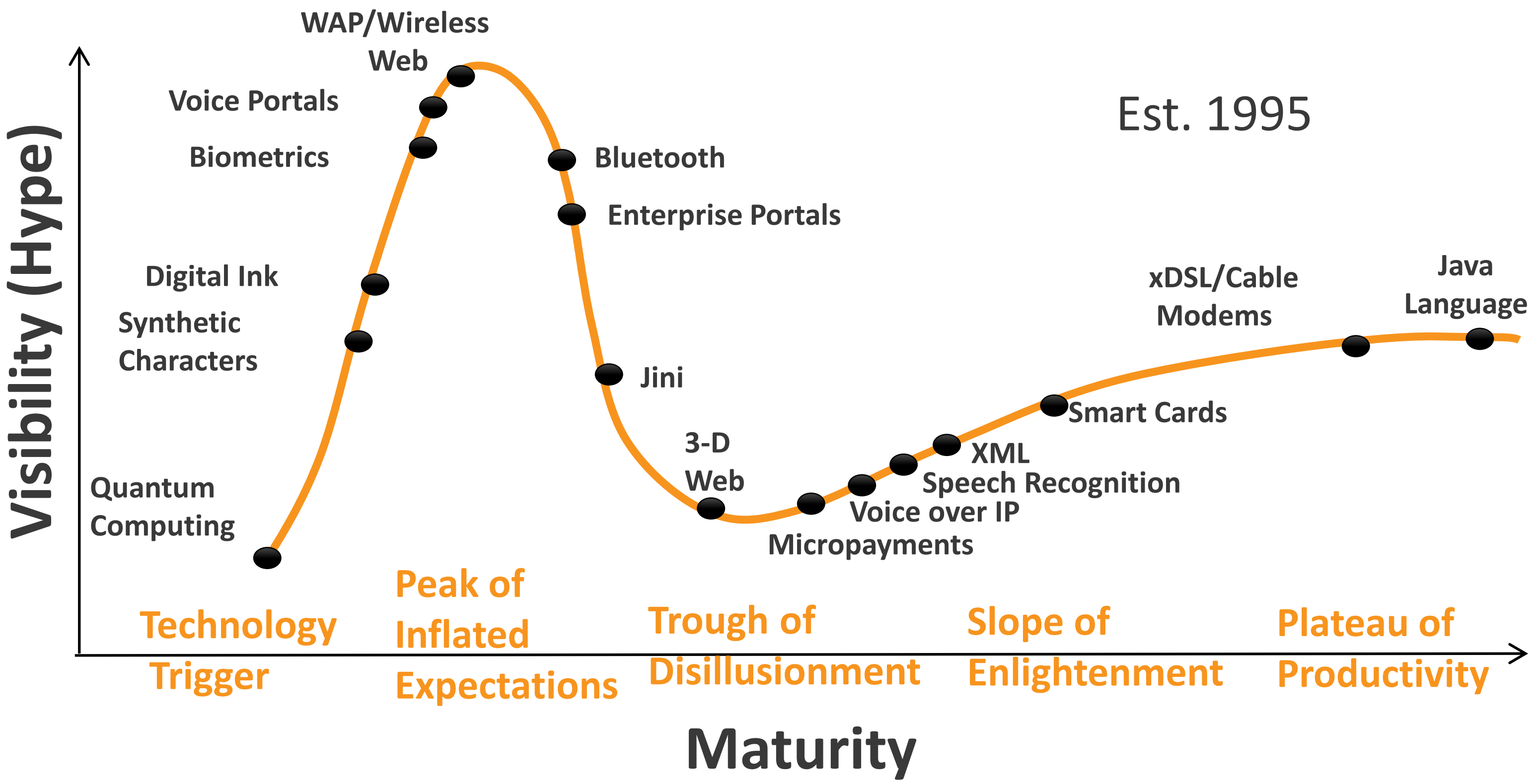

Autodesk AutoCAD 2014 - NOT FOR RESALE DRIFTPLANNING.dwg
Home Insert Annotate Layout Parametric 3D Tools View Manage Output Plug-ins Autodesk 360 Raster Tools SketchBook Designer Express Tools Promine Promine Modeling Promine Engineering Promine Geology
Cut many sections
Export to DXF
Export Batch
Create Layout
Cut 1 Section
Return To Plan
Select Section
Insert Block
Center Orbit
3D->2D
Promine Options
Filer
Grid
Section
Interface
Library
Promine tools
DRIFTPLANNING*
Model
Layout1
Layout2

Data wrangler!

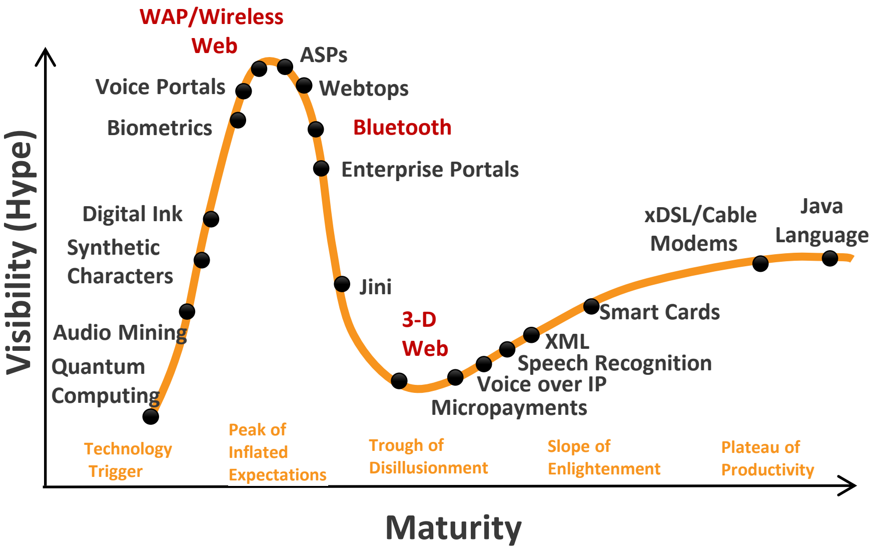
TECH JOURNEY: FRIEND OR ENEMY…

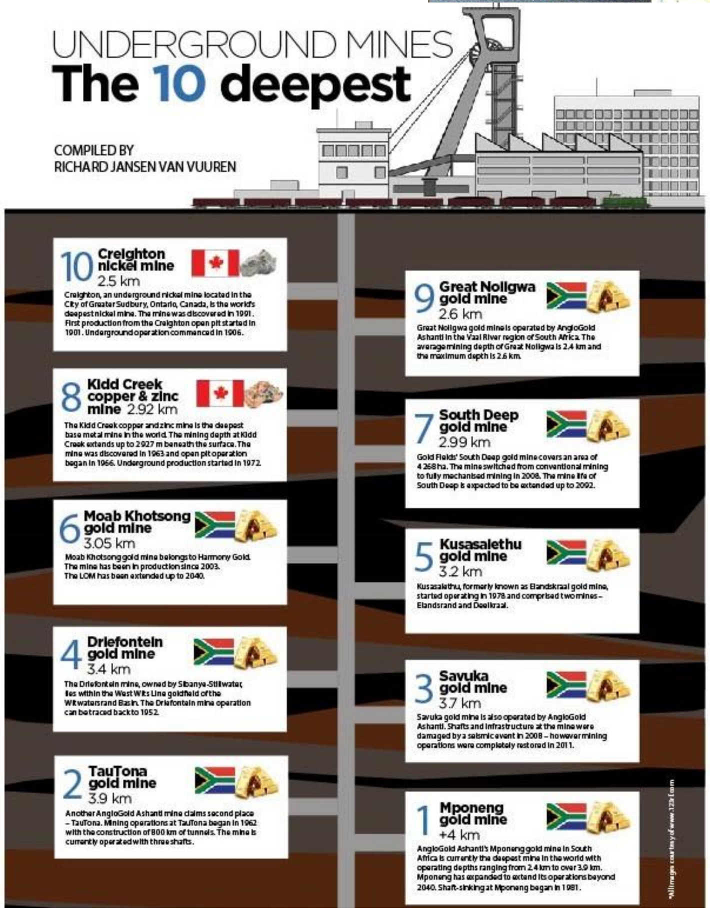
aarnet

THE GARTNER HYPE CYCLE



GARTNER HYPE CYCLE 2000





UNDERGROUND MINES
The 10 deepest

COMPILED BY
RICHARD JANSEN VAN VUUREN

10 Creighton nickel mine 2.5 km

Creighton, an underground nickel mine located in the City of Greater Sudbury, Ontario, Canada, is the world's deepest nickel mine. The mine was discovered in 1991. First production from the Creighton open pit started in 1901. Underground operation commenced in 1906.

9 Great Noligwa gold mine 2.6 km

Great Noligwa gold mine is operated by AngloGold Ashanti in the Vaal River region of South Africa. The average mining depth of Great Noligwa is 2.4 km and the maximum depth is 2.6 km.

8 Kidd Creek copper & zinc mine 2.92 km

The Kidd Creek copper and zinc mine is the deepest base metal mine in the world. The mining depth at Kidd Creek extends up to 2927 m beneath the surface. The mine was discovered in 1963 and open pit operation began in 1966. Underground production started in 1972.

7 South Deep gold mine 2.99 km

Gold Fields' South Deep gold mine covers an area of 4 268 ha. The mine switched from conventional mining to fully mechanised mining in 2008. The mine life of South Deep is expected to be extended up to 2092.

6 Moab Khotsong gold mine 3.05 km

Moab Khotsong gold mine belongs to Harmony Gold. The mine has been in production since 2003. The LOM has been extended up to 2040.

5 Kusasalethu gold mine 3.2 km

Kusasalethu, formerly known as Elandskraal gold mine, started operating in 1978 and comprised two mines – Elandsrand and Deelkraal.

4 Driefontein gold mine 3.4 km

The Driefontein mine, owned by Sibanye-Stillwater, lies within the West Wits Line goldfield of the Witwatersrand Basin. The Driefontein mine operation can be traced back to 1952.

3 Savuka gold mine 3.7 km

Savuka gold mine is also operated by AngloGold Ashanti. Shafts and infrastructure at the mine were damaged by a seismic event in 2008 – however mining operations were completely restored in 2011.

2 TauTona gold mine 3.9 km

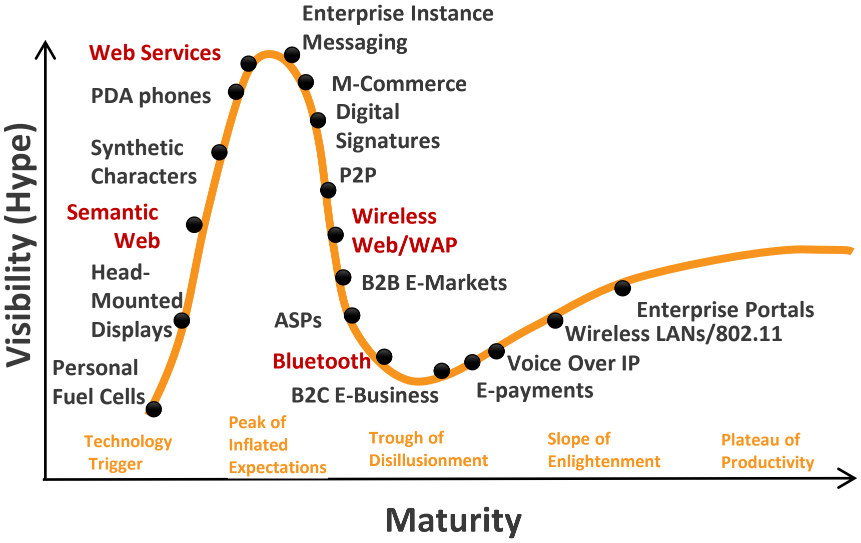
Another AngloGold Ashanti mine claims second place – TauTona. Mining operations at TauTona began in 1962 with the construction of 800 km of tunnels. The mine is currently operated with three shafts.

1 Mponeng gold mine +4 km

AngloGold Ashanti's Mponeng gold mine in South Africa is currently the deepest mine in the world with operating depths ranging from 2.4 km to over 3.9 km. Mponeng has expanded to extend its operations beyond 2040. Shaft-sinking at Mponeng began in 1981.

*All images courtesy of www.123rf.com

GARTNER HYPE CYCLE 2001



Tech

Semantic Web / Web Services / WAP / Bluetooth

Me

Started my PhD – data collection

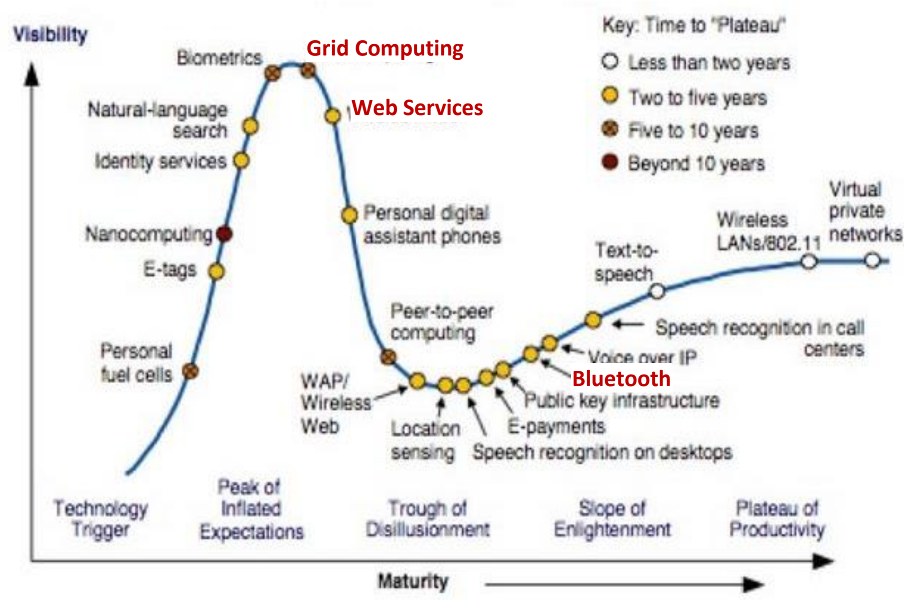
Some coding – Fortran / Matlab / C

BHR /Geophysics

Begin Testing wireless technologies

Data Standard…?

GARTNER HYPE CYCLE 2002



Tech

Grid Computing

Web Services maturing

Bluetooth stabilising

Me

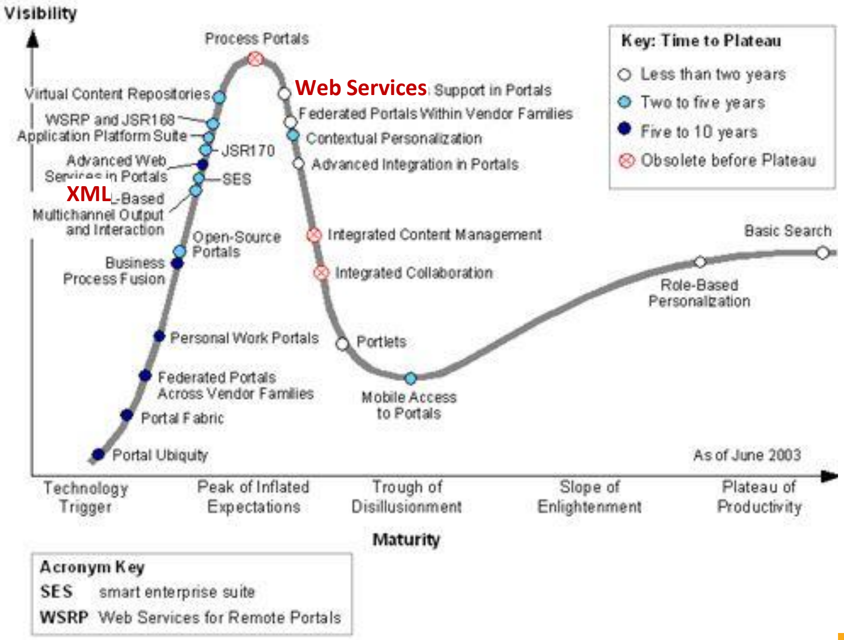
PhD Year 2 – more data collection

coding – Matlab / C

BHR / Geophysics

Flash memory / Bluetooth tests

GARTNER HYPE CYCLE 2003



Tech

XML Based Multi-channel Output and Interaction

Web Services

Me PhD….

BHR

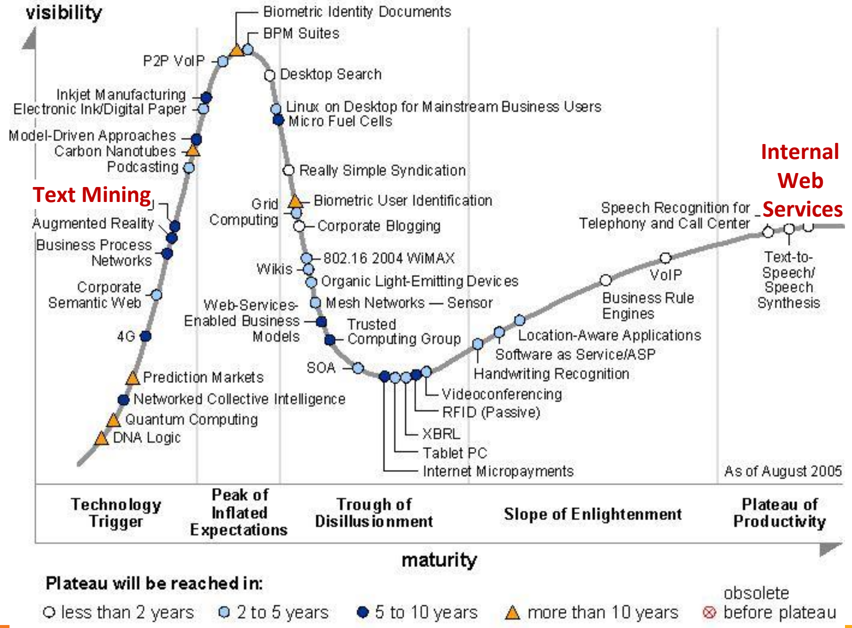
Single stick wireless radar tests

Geophysics

ASEG-GDF2 published – A standard for point located data exchange

Australian Society of Exploration Geophysicists

...GARTNER HYPE CYCLE 2005



Tech

Text Mining...

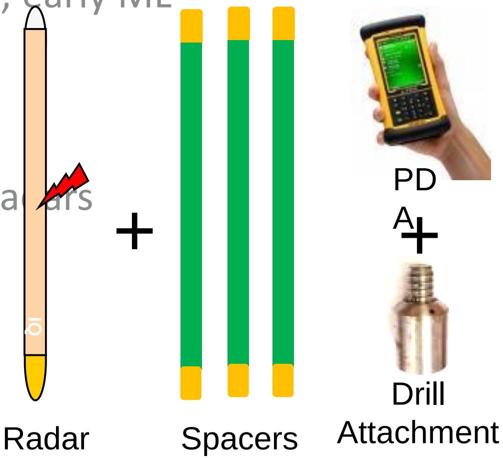
Me

Started Postdoc

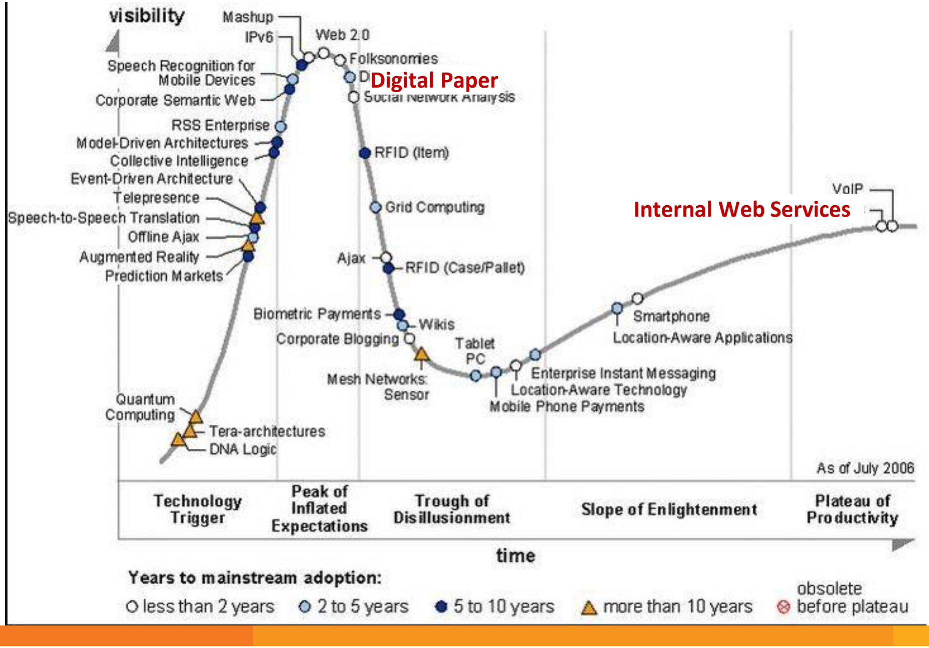
Data Fusion Project, early ML

BHR / Geophysics

Mature Bluetooth radars



GARTNER HYPE CYCLE 2006



Digital Paper

Internal Web Services

Te

Di

M

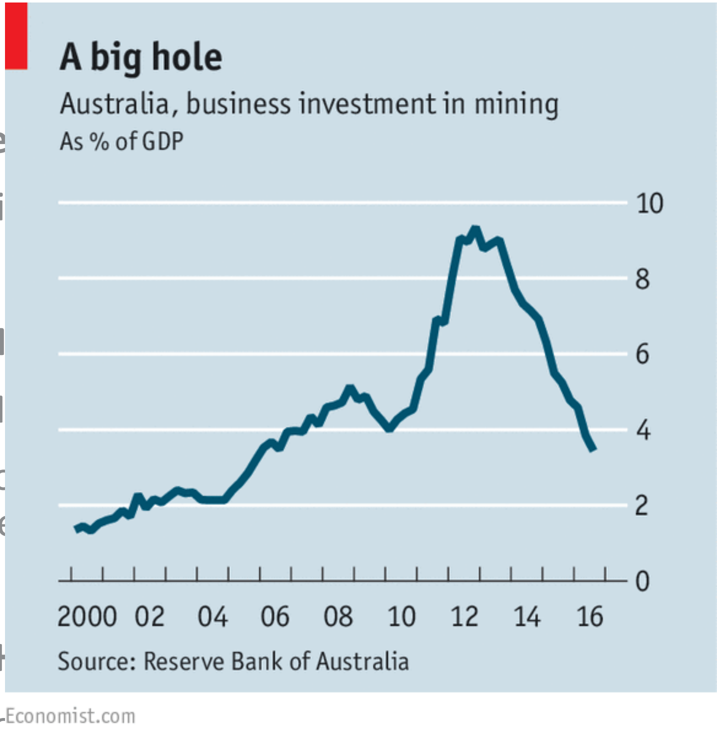
M

Co

Ge

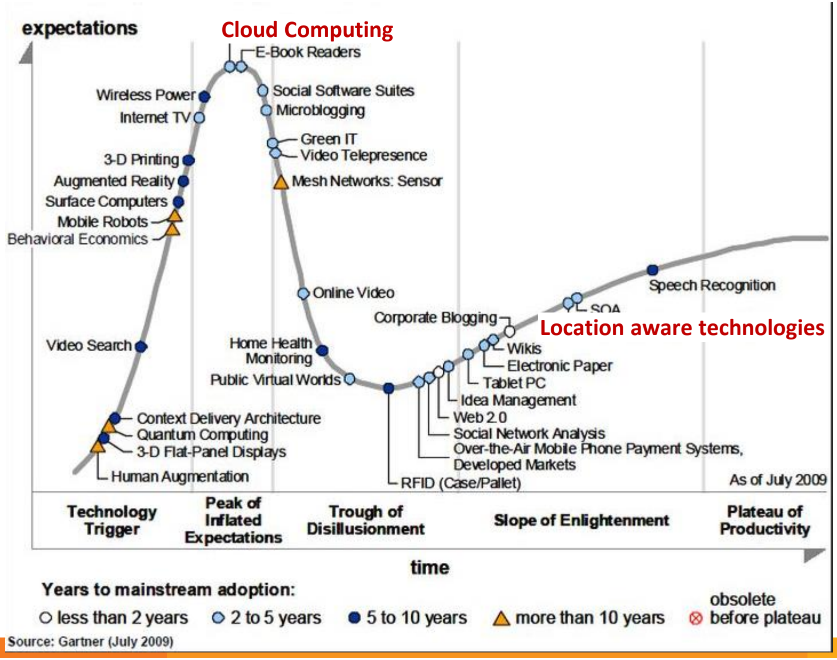
Bl

Cc



https://www.economist.com/asia/2017/03/09/the-end-of-a-mining-boom-leaves-australias-economy-surprisingly-intact

…GARTNER HYPE CYCLE 2009



Cloud Computing

Location aware technologies

Source: Gartner (July 2009)

https://www.adafruit.com

Tech

Cloud Computing –

Me

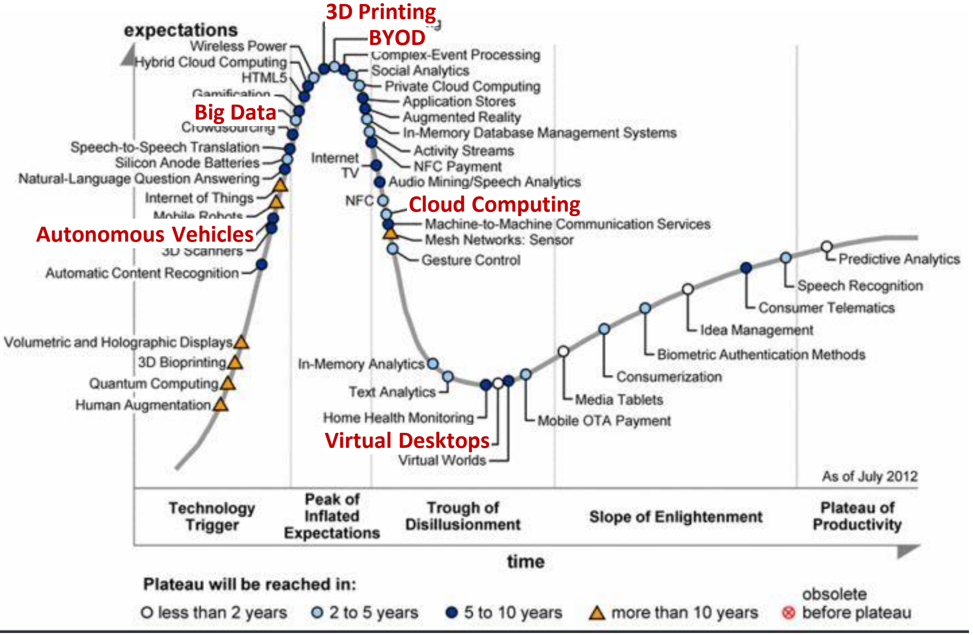
BHR Business Development

GOLD / PLATINUM / NICKEL / DIAMONDS

BHR / Geophysics

Downhole surveying tools with accelerometers

....GARTNER HYPE CYCLE 2012



Tech

Big Data is emerging

Virtual Desktops…

Me

Moved to Canberra

Started playing with BIGish Geophysics Data

Virtual Geophysics Exploration Laboratory (VEGL)

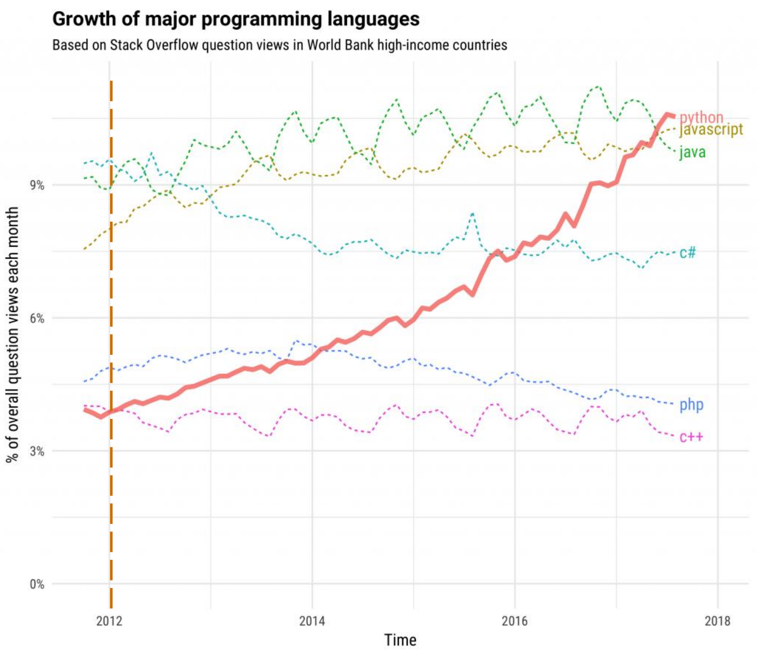
Python

~~BHR~~ / Geophysics

ASEG-GDF2 still the same since 2003…

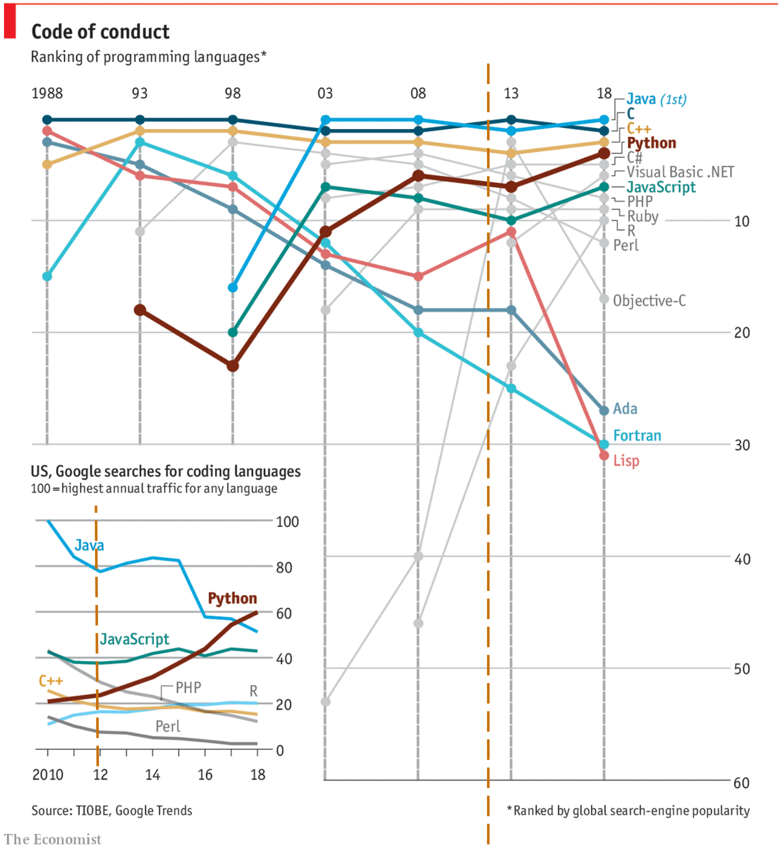
ASEG-ESF for Electrical Surveys published

ASIDE – PYTHON GROWTH…



Growth of major programming languages

Based on Stack Overflow question views in World Bank high-income countries



Code of conduct

Ranking of programming languages*

US, Google searches for coding languages

100=highest annual traffic for any language

Source: TIOBE, Google Trends

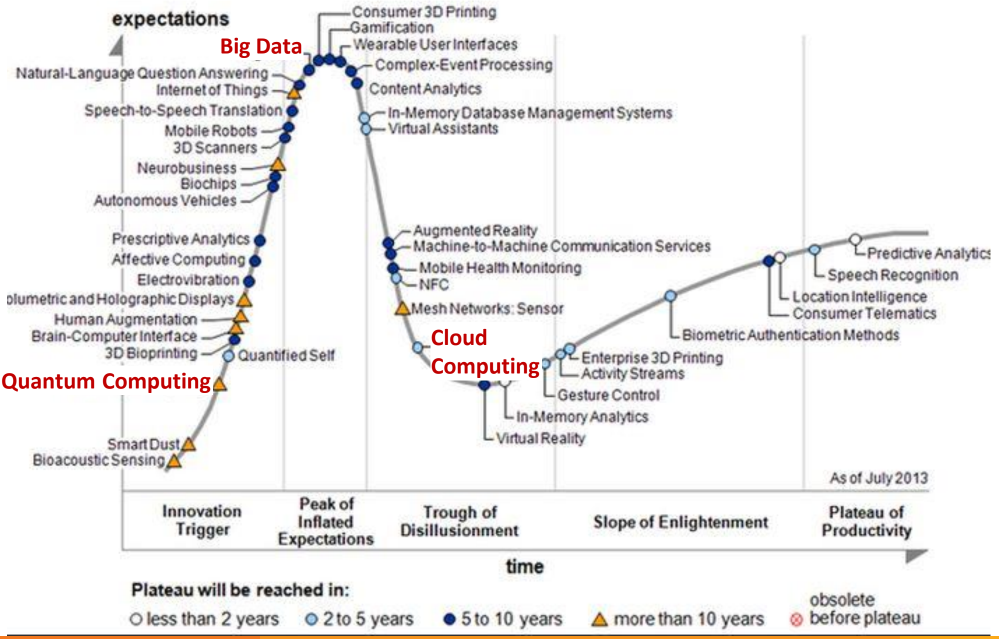
*Ranked by global search-engine popularity

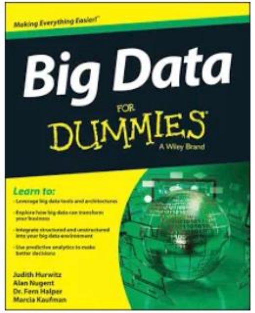
The Economist

https://stackoverflow.blog/2017/09/06/incredible-growth-python/

https://www.economist.com/graphic-detail/2018/07/26/python-is-becoming-the-worlds-most-popular-coding-language

GARTNER HYPE CYCLE 2013





Tech

Big Data hits its peak of HYPE…

Quantum Computing still emerging

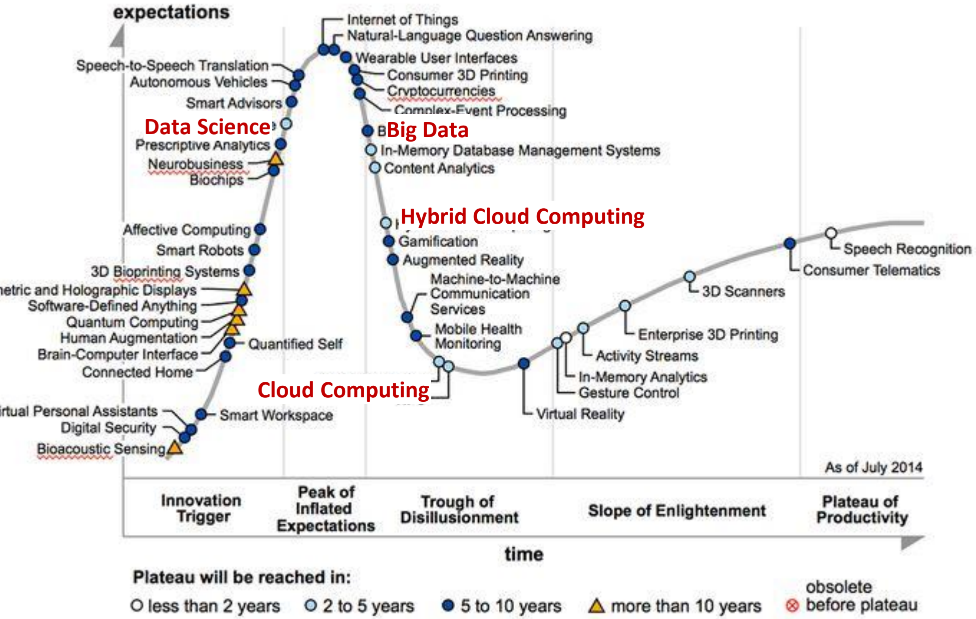
Cloud Computing in the dumps..

Me

Maternity Leave

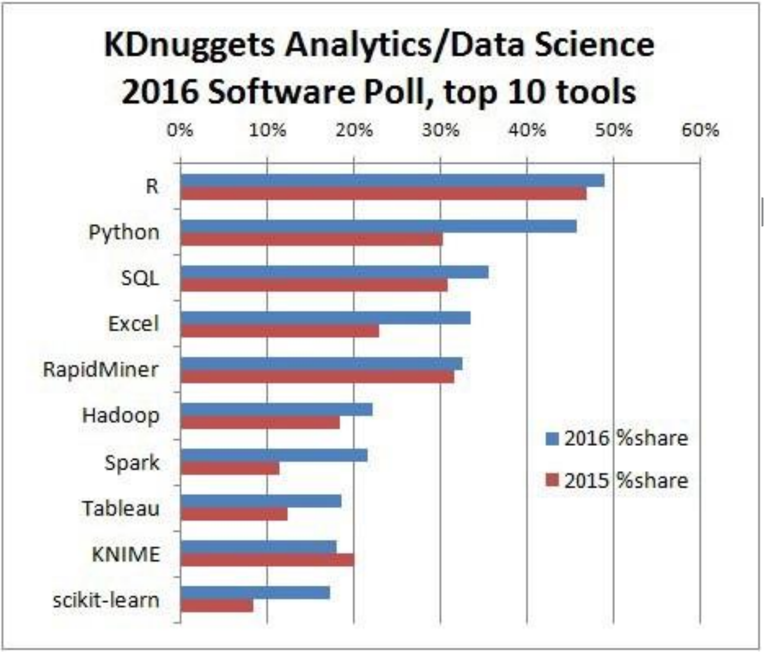
Virtual Geophysics Laboratory (VGL)

GARTNER HYPE CYCLE 2014



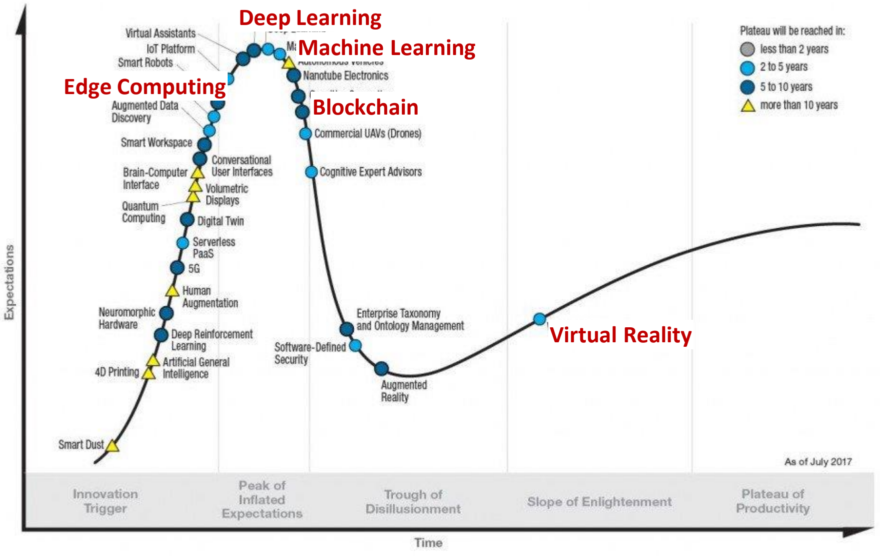
Tech

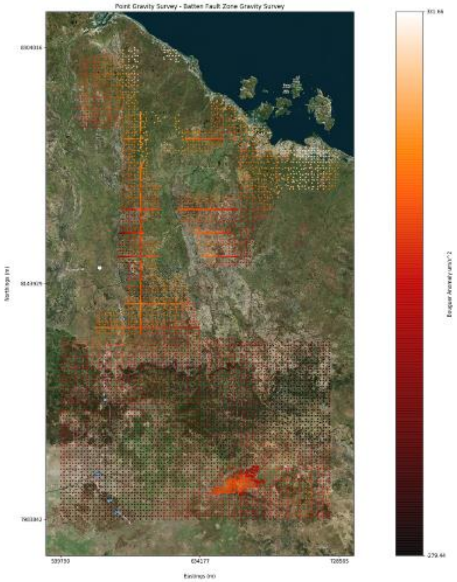
Data Science Appears



Source: Alex Ip, Geoscience Australia

GARTNER HYPE CYCLE 2017





Tech

Deep Learning Hype

Machine Learning Hype

Blockchain heading for the tr

Semantics and Web Services I

Me

ANVGL in the Cloud in 2015

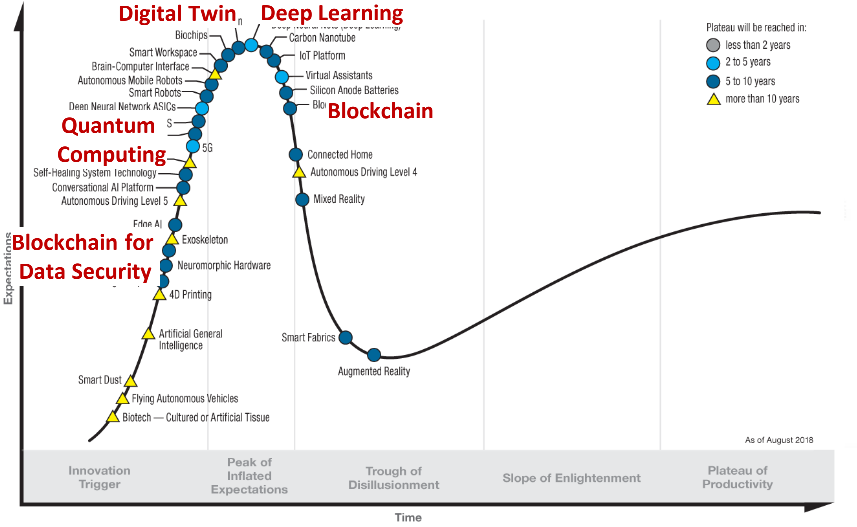
National Geophysical Collection Maturity – NetCDF-CF

Geophysics

ASEG-GDF2 still the same since 2003…

ASEG-ESF still the same since 2012…

GARTNER HYPE CYCLE 2018



Tech

Blockchain for Data Security

Quantum Computing is getting closer

Digital Twins??

Deep Learning

Semantics and Web Services Matured??

Me

Management… Acting CIO…

Digital Transformation

Geophysics

ASEG-GDF2 still the same since 2003…

ASEG-ESF still the same since 2012…

WHAT CAN WE LEARN FROM THE HYPE?

CONFESSION 1.

I am a bad data manager naturally

▶

But good practice can be learned over time and with good collaborators..

CONFESSION 2.

I was a Data Wrangler and now I enable Data Wranglers..

But leveraging technologies can make it easier..

Efficient.. Repeatable..

FAIR…

CONFESSION 3.

Web-services to enable Machine actionable FAIR Data has been possible since 200#

But where are we now?

The Geophysical Community standardized early for interoperability but not machine to machine actions…

CONFESSION 4.

Change is hard

▶

But we must be patient and work together across our communities to promote the benefits of change.

CONFESSION 5.

Change is hard
(especially in Silos)
Hindsight is easy

▶

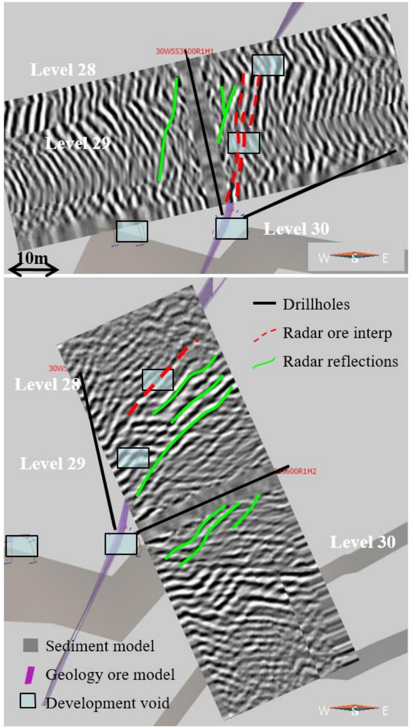
But we must be patient and work together to promote the benefits.

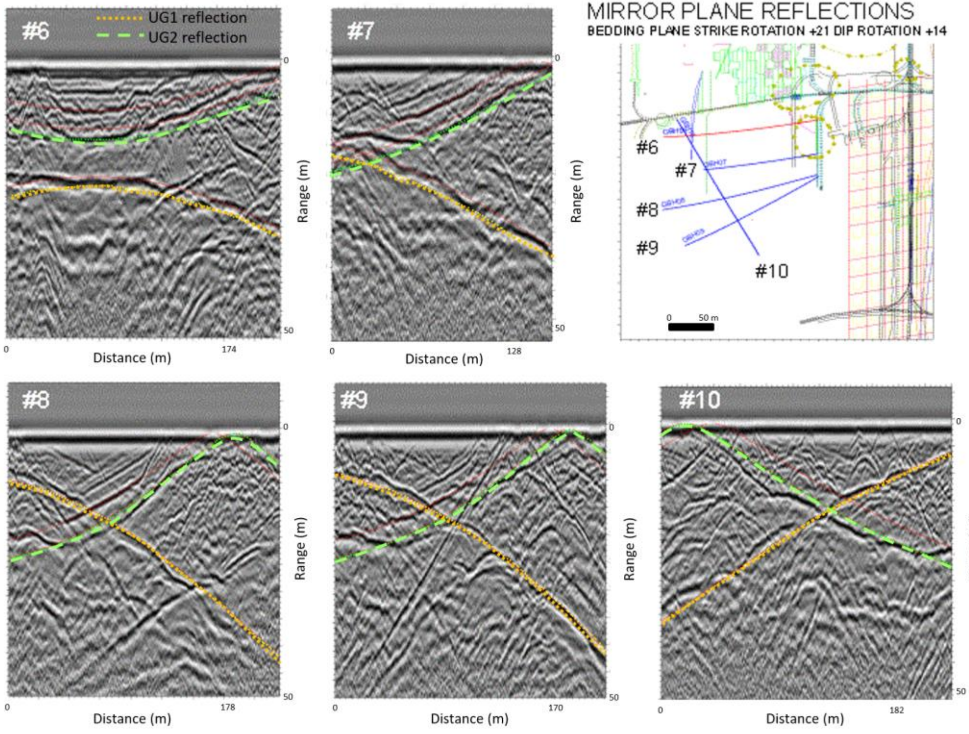
MORE RESCUED DATA…

aarnet

GOLD

PLATINUM

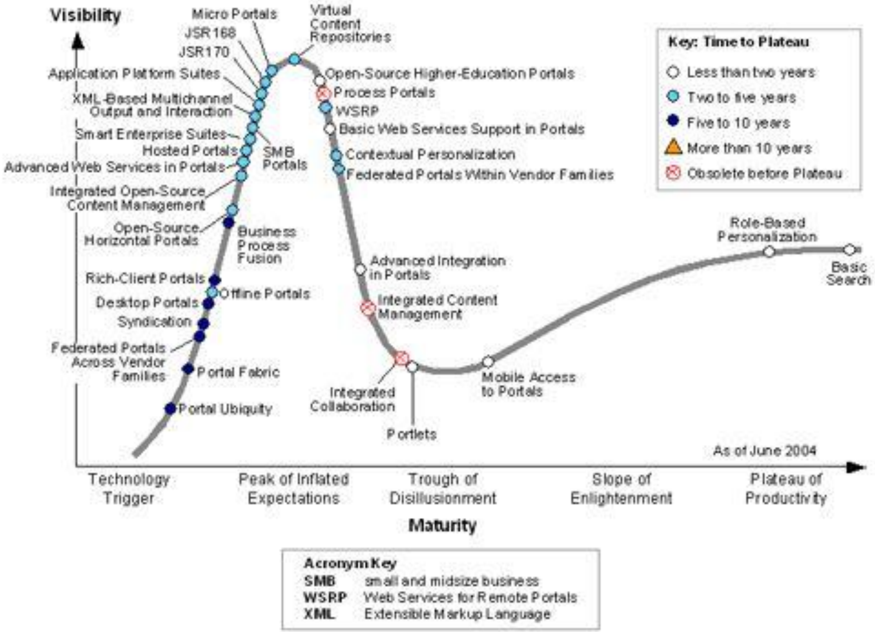




THANK YOU

GARTNER HYPE CYCLE 2004

Gartner Hype Cycle 2004



Me

BHR

Geophysics