GEMS

Data Sharing and Analysis Enabling Data Driven Agricultural Innovation While Respecting IP

Kevin Silverstein, PhD, GEMS Operations Manager

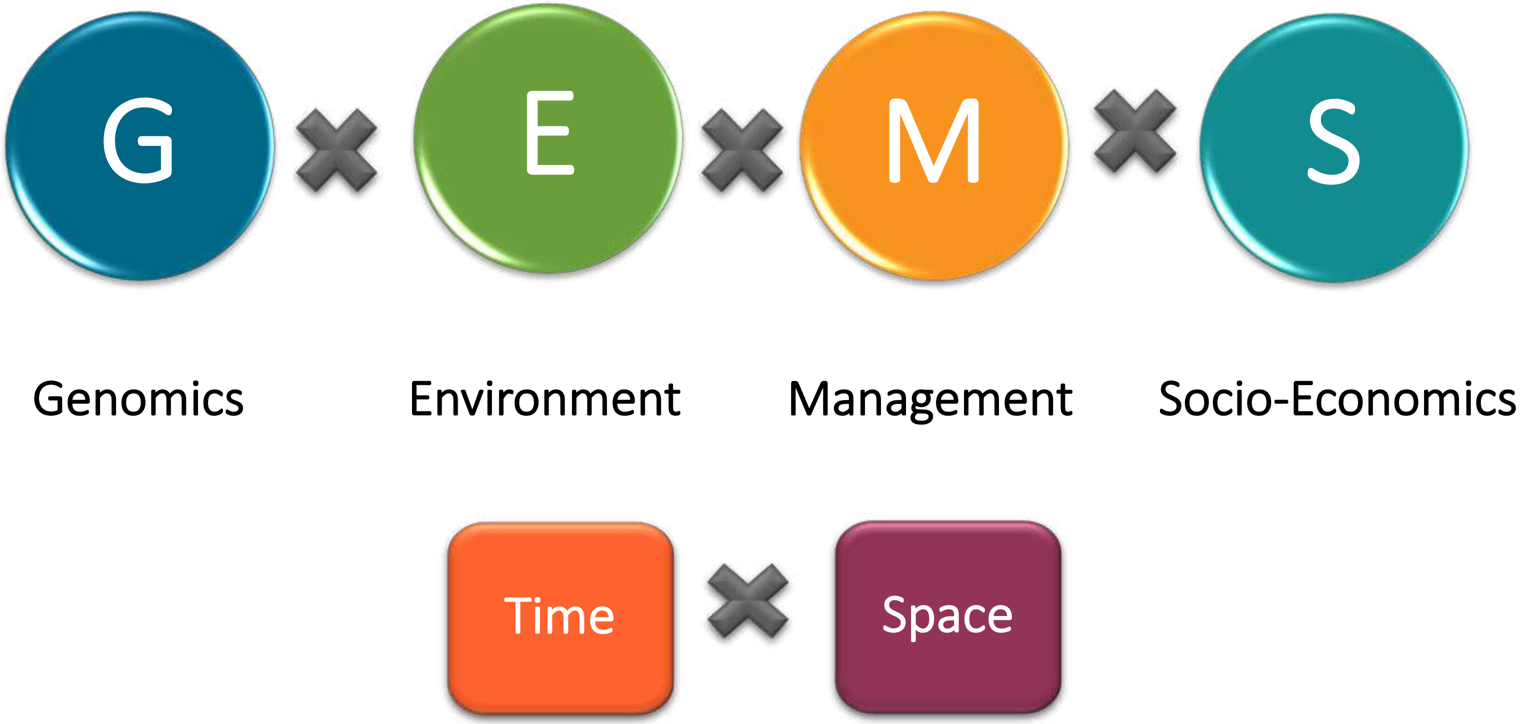
GEMS led by: Philip Pardey, Jim Wilgenbusch and Kevin Silverstein
College of Food Agricultural and Natural Resource Science, CFANS
Minnesota Supercomputing Institute, MSI
University of Minnesota

PAG XXVII Workshop Connecting Crop Phenotype and Genotype Data

January 16, 2019

What is G.E.M.S?

A novel data sharing and big data analytical platform that enables public-private research collaborations for innovation in food and agricultural production, and other domain areas



G × E × M × S

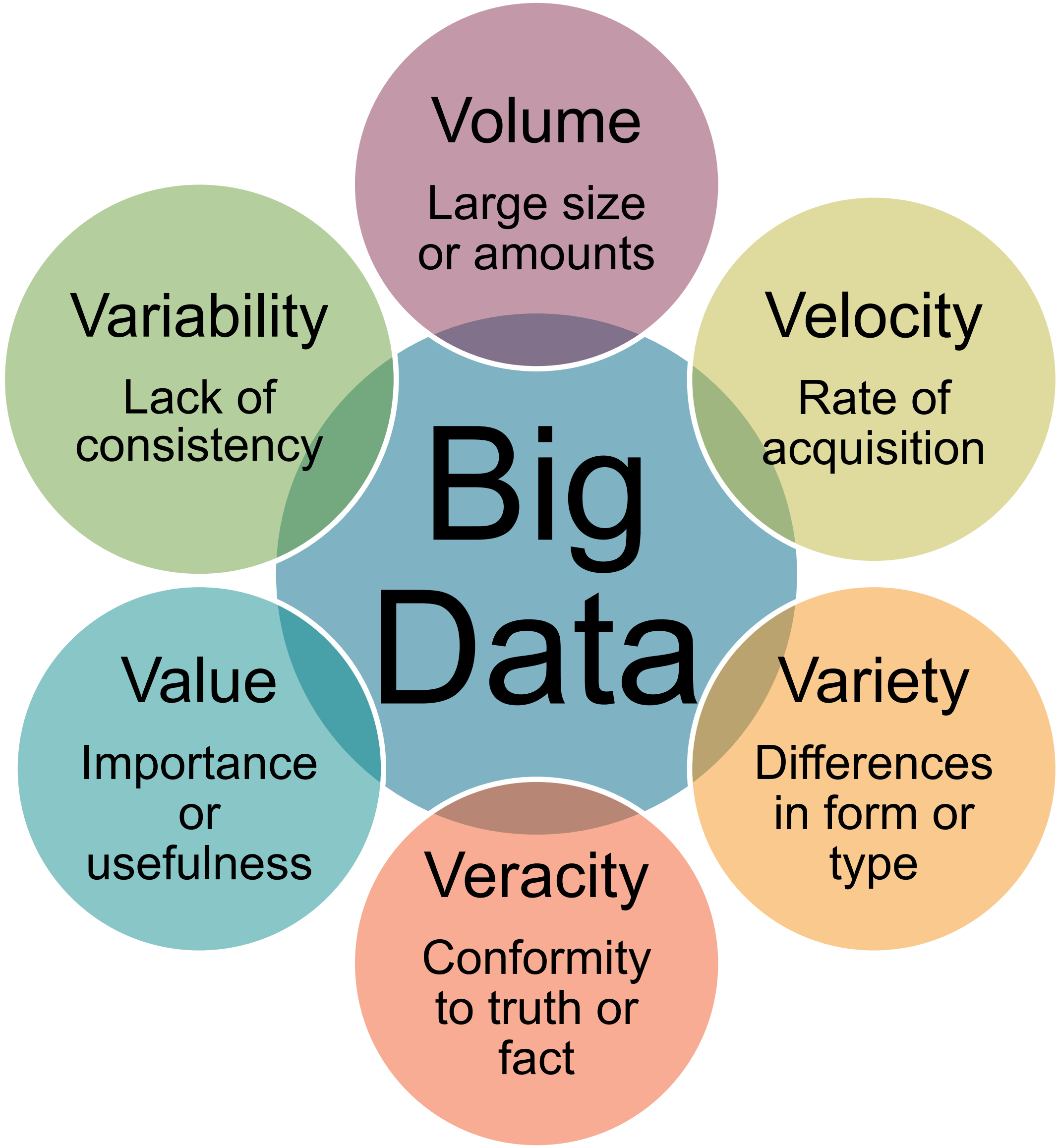
Genomics Environment Management Socio-Economics

Time × Space

The G.E.M.S Team (more than 20 brains strong!)

- Bi-Weekly build meetings
- Weekly technical meetings
- Numerous ad hoc consultations in the Cargill Branary & MSI

Big Data Challenges in Food & Agriculture



Big Data

Volume
Large size or amounts

Velocity
Rate of acquisition

Variety
Differences in form or type

Veracity
Conformity to truth or fact

Value
Importance or usefulness

Variability
Lack of consistency

GEMS

Realizing the Big Data Revolution

Get the data to the tool or get the tool to the data

Data Transfer

Reconcile file formats, units, vocabularies, languages, and ontologies

Data Interoperability

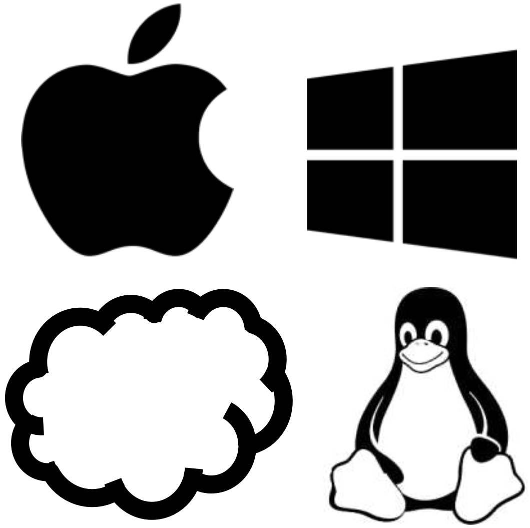
Access to complex software and ability to replicate analyses

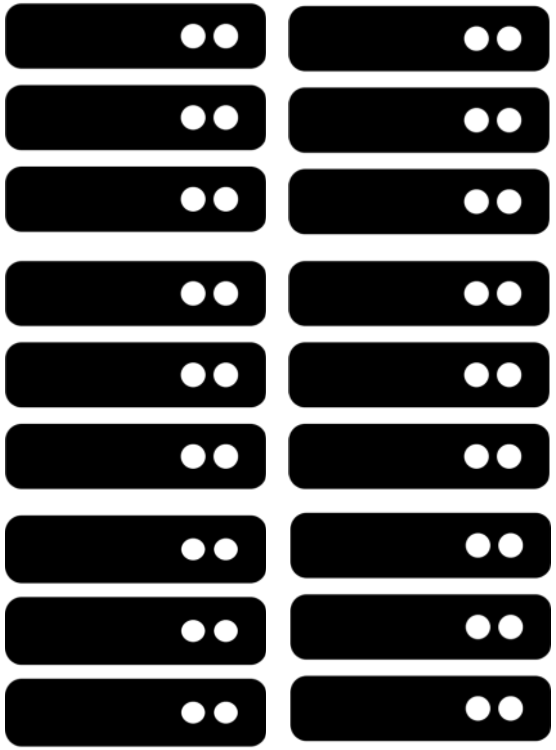
Data Analysis

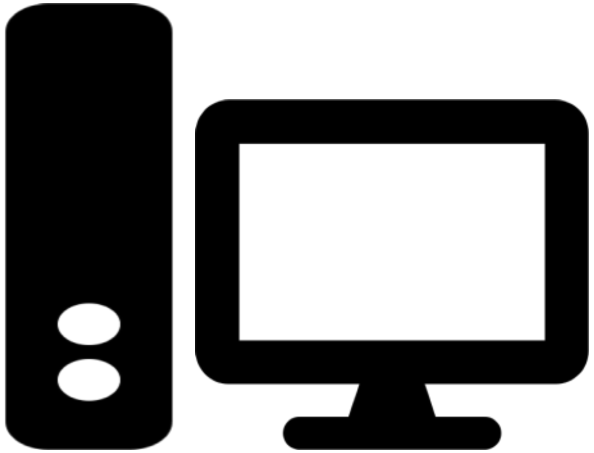
Facilitate complex partnerships and respecting data ownership and privacy

Data Sharing

Data Transfer – Platform Portability





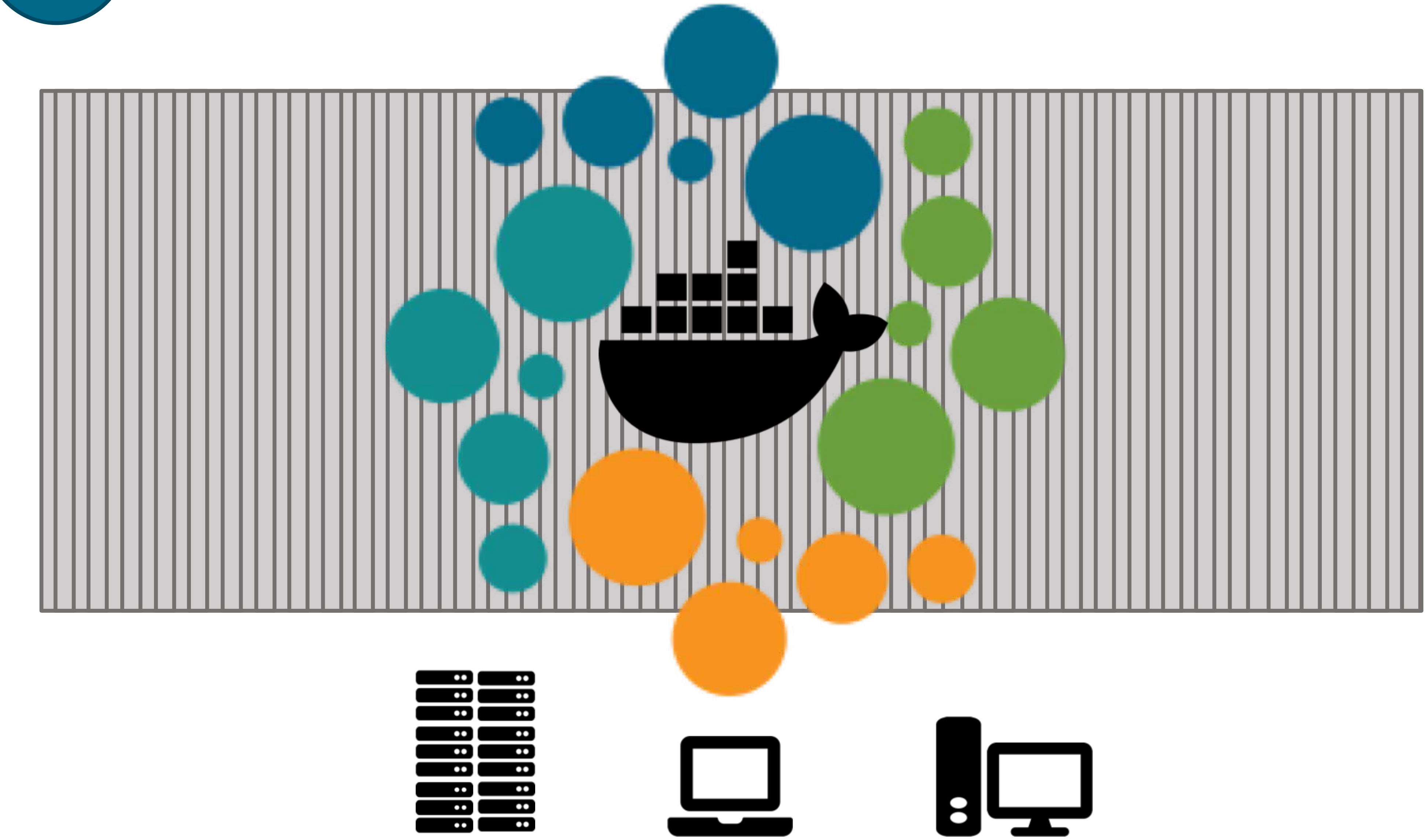


GEMs on Clusters

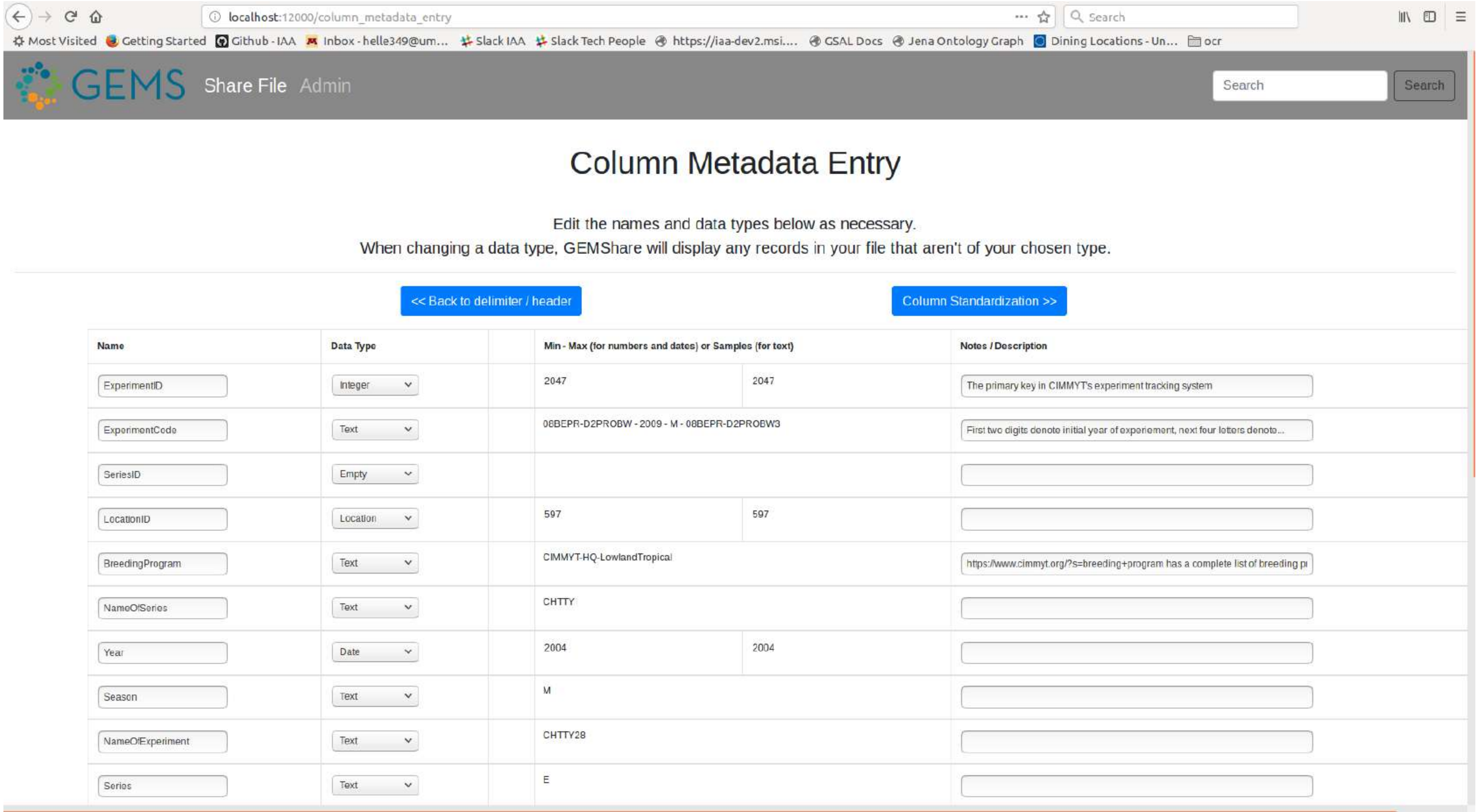
GEMs on Workstations

GEMs on Laptops

Data Transfer – Containerize the Platform



Data interoperability: column metadata

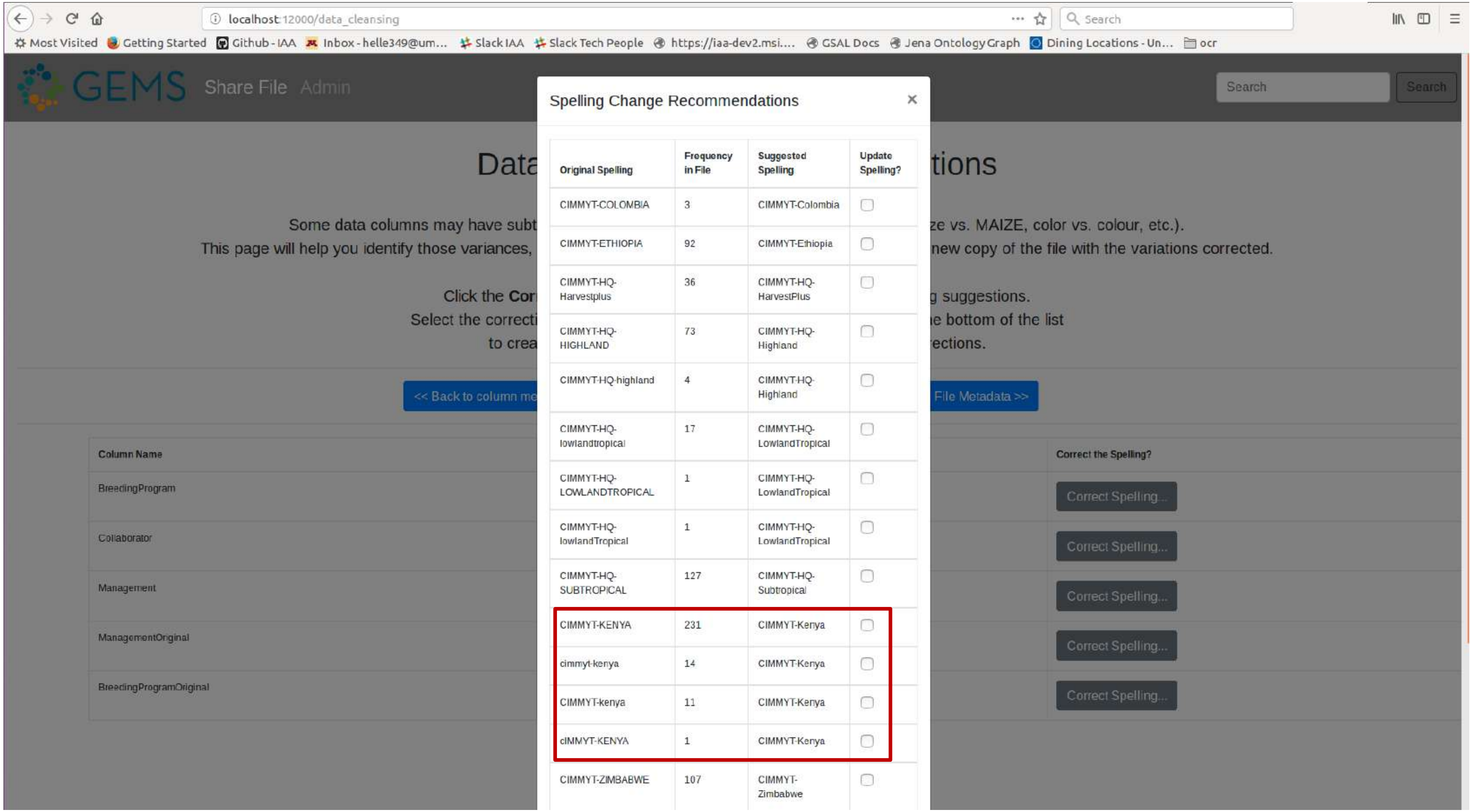


Column Metadata Entry

Edit the names and data types below as necessary.
When changing a data type, GEMShare will display any records in your file that aren't of your chosen type.

Name	Data Type	Min - Max (for numbers and dates) or Samples (for text)		Notes / Description
ExperimentID	Integer	2047	2047	The primary key in CIMMYT's experiment tracking system
ExperimentCode	Text	06BEPR-D2PROBW - 2009 - M - 08BEPR-D2PROBW3		First two digits denote initial year of experiement, next four letters denote...
SeriesID	Empty			
LocationID	Location	597	597	
BreedingProgram	Text	CIMMYT-HQ-LowlandTropical		https://www.cimmyt.org/?s=breeding+program has a complete list of breeding pı
NameOfSeries	Text	CHTTY		
Year	Date	2004	2004	
Season	Text	M		
NameOfExperiment	Text	CHTTY28		
Series	Text	E		

Data interoperability: spelling correction



Data ... tions

Some data columns may have subt ... ze vs. MAIZE, color vs. colour, etc.).
This page will help you identify those variances, ... new copy of the file with the variations corrected.

Click the Cor ... g suggestions.
Select the correcti ... e bottom of the list
to crea ... ections.

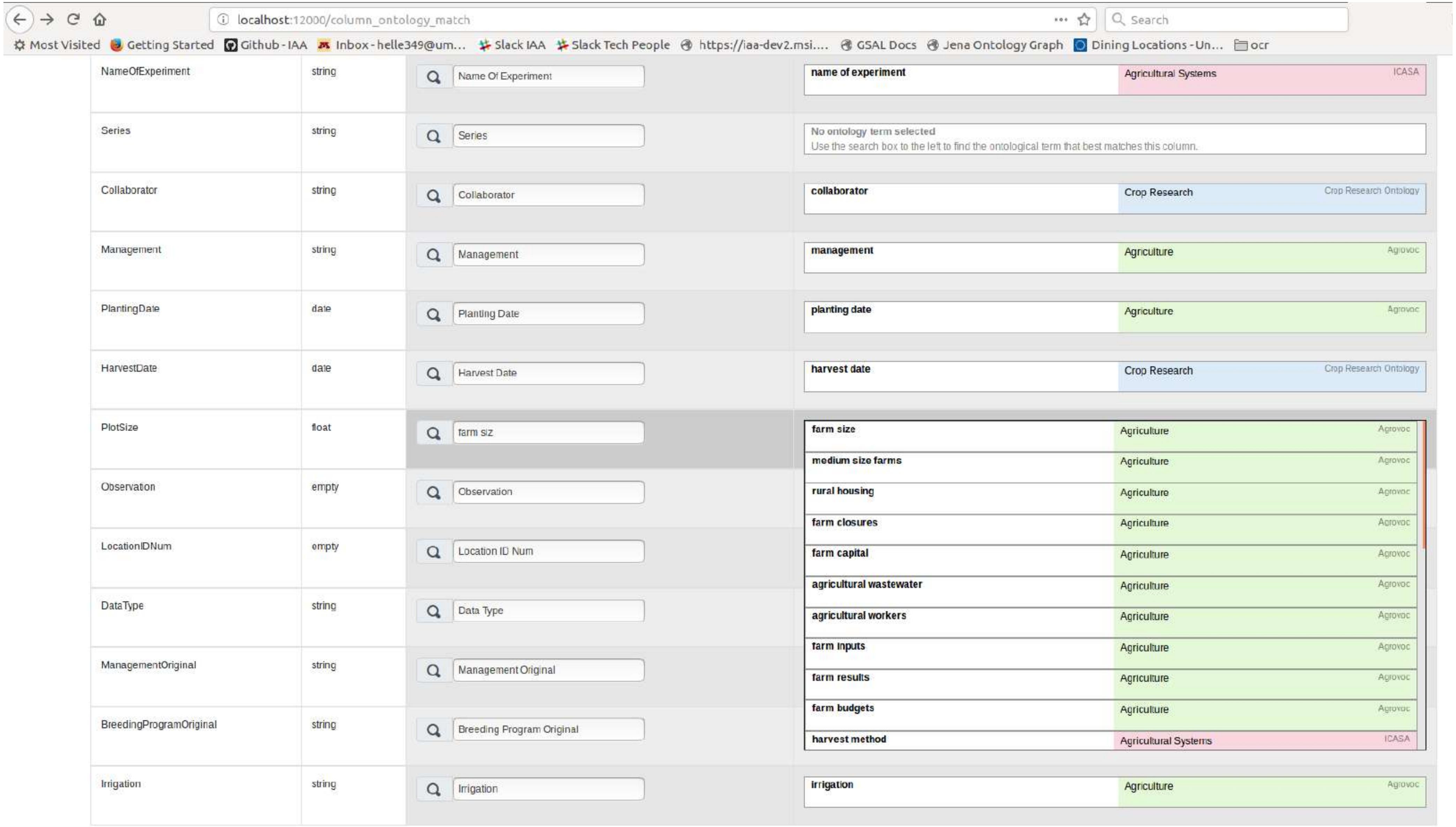
Column Name	Correct the Spelling?
BreedingProgram	Correct Spelling...
Collaborator	Correct Spelling...
Management	Correct Spelling...
ManagementOriginal	Correct Spelling...
BreedingProgramOriginal	Correct Spelling...

Spelling Change Recommendations ×

Original Spelling	Frequency in File	Suggested Spelling	Update Spelling?
CIMMYT-COLOMBIA	3	CIMMYT-Colombia	☐
CIMMYT-ETHIOPIA	92	CIMMYT-Ethiopia	☐
CIMMYT-HQ-Harvestplus	36	CIMMYT-HQ-HarvestPlus	☐
CIMMYT-HQ-HIGHLAND	73	CIMMYT-HQ-Highland	☐
CIMMYT-HQ-highland	4	CIMMYT-HQ-Highland	☐
CIMMYT-HQ-lowlandtropical	17	CIMMYT-HQ-LowlandTropical	☐
CIMMYT-HQ-LOWLANDTROPICAL	1	CIMMYT-HQ-LowlandTropical	☐
CIMMYT-HQ-lowlandTropical	1	CIMMYT-HQ-LowlandTropical	☐
CIMMYT-HQ-SUBTROPICAL	127	CIMMYT-HQ-Subtropical	☐
CIMMYT-KENYA	231	CIMMYT-Kenya	☐
cimmyt-kenya	14	CIMMYT-Kenya	☐
CIMMYT-kenya	11	CIMMYT-Kenya	☐
cIMMYT-KENYA	1	CIMMYT-Kenya	☐
CIMMYT-ZIMBABWE	107	CIMMYT-Zimbabwe	☐

GEMS

Data interoperability - ontology matching



NameOfExperiment	string	Name Of Experiment	name of experiment	Agricultural Systems	ICASA
Series	string	Series	No ontology term selected Use the search box to the left to find the ontological term that best matches this column.		
Collaborator	string	Collaborator	collaborator	Crop Research	Crop Research Ontology
Management	string	Management	management	Agriculture	Agrovoc
PlantingDate	date	Planting Date	planting date	Agriculture	Agrovoc
HarvestDate	date	Harvest Date	harvest date	Crop Research	Crop Research Ontology
PlotSize	float	farm siz	farm size medium size farms rural housing farm closures farm capital agricultural wastewater agricultural workers farm inputs farm results farm budgets harvest method	Agriculture Agriculture Agriculture Agriculture Agriculture Agriculture Agriculture Agriculture Agriculture Agriculture Agricultural Systems	Agrovoc Agrovoc Agrovoc Agrovoc Agrovoc Agrovoc Agrovoc Agrovoc Agrovoc Agrovoc ICASA
Observation	empty	Observation			
LocationIDNum	empty	Location ID Num			
DataType	string	Data Type			
ManagementOriginal	string	Management Original			
BreedingProgramOriginal	string	Breeding Program Original			
Irrigation	string	Irrigation	irrigation	Agriculture	Agrovoc

Data Interoperability via Automated (units & terms) Standardizations

Total Phosphorus

207 lb/A
46lb per acre
46 pound/A
22 lbs
68 kg/ha
54lbs/acre
55.5 lbs P per Acre
80lb/acre
None
40 pounds
none applied
192 lbs;17-Apr-14
...

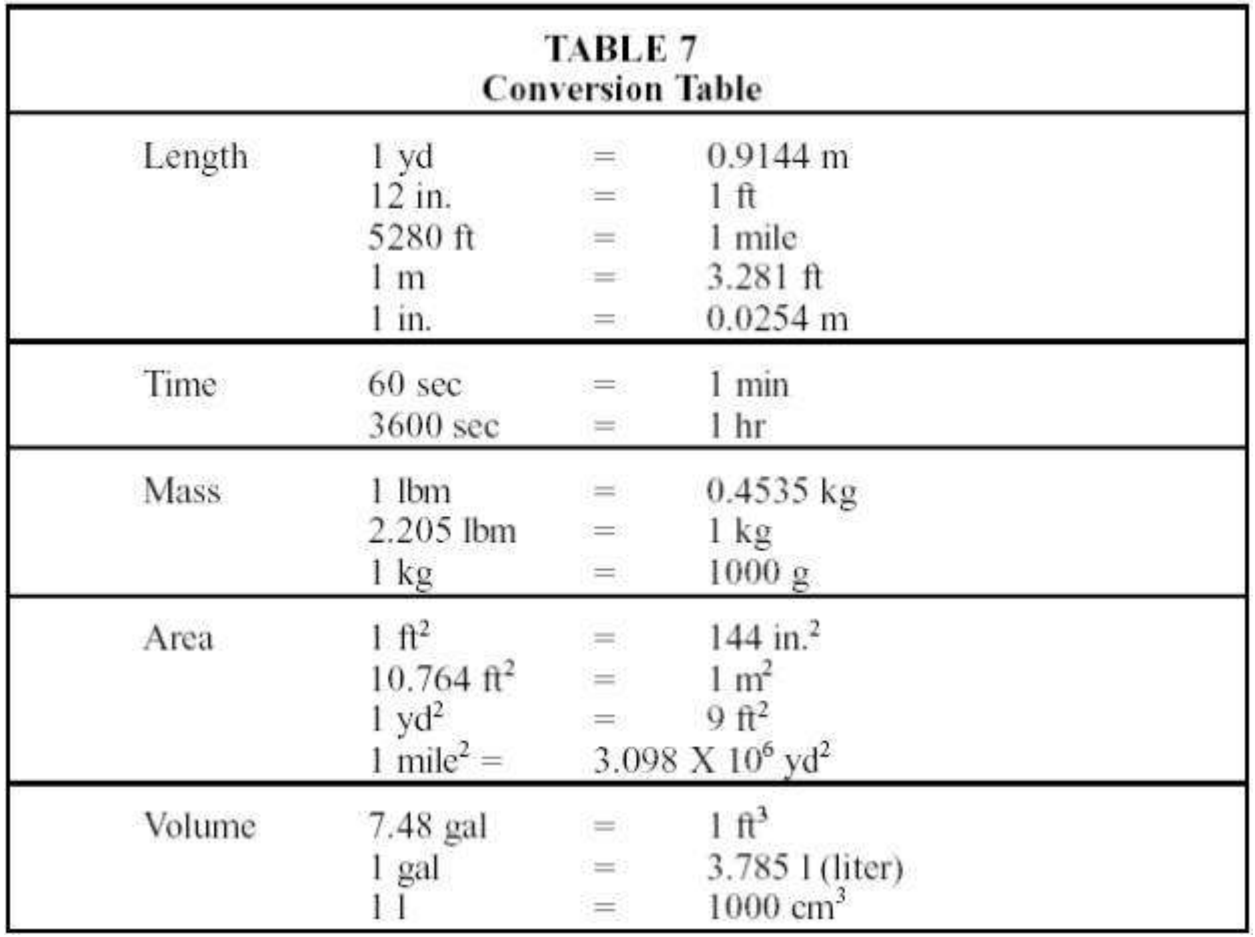
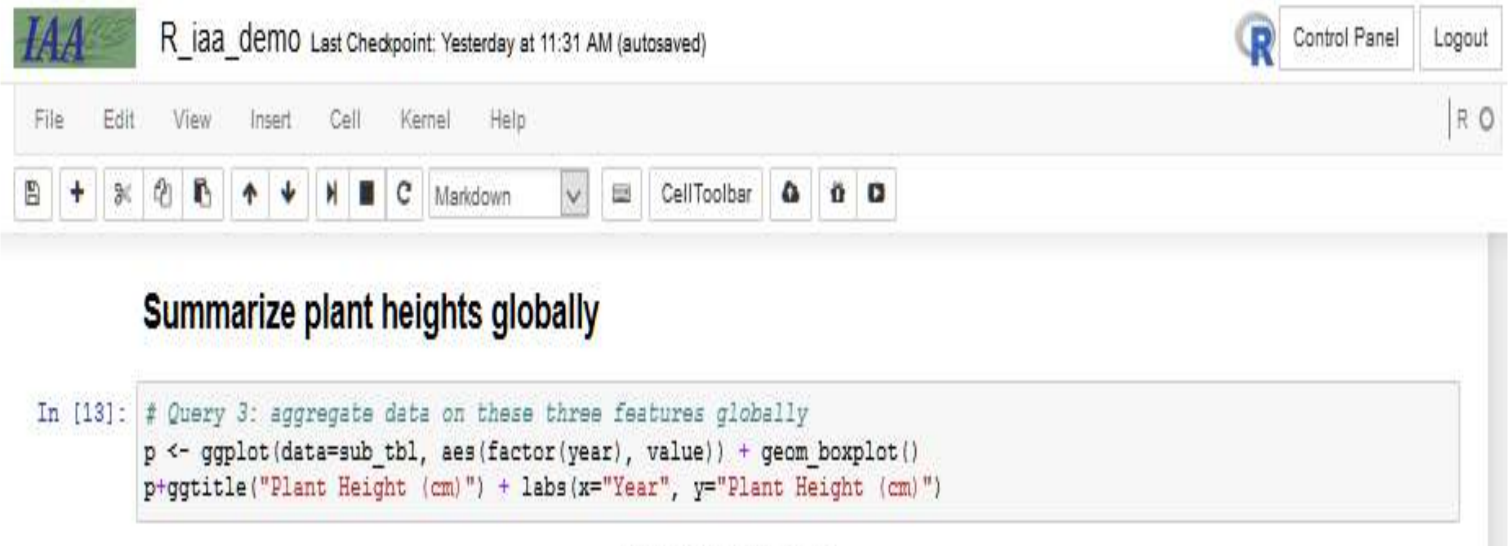


TABLE 7
Conversion Table

Length	1 yd	=	0.9144 m
	12 in.	=	1 ft
	5280 ft	=	1 mile
	1 m	=	3.281 ft
	1 in.	=	0.0254 m
Time	60 sec	=	1 min
	3600 sec	=	1 hr
Mass	1 lbm	=	0.4535 kg
	2.205 lbm	=	1 kg
	1 kg	=	1000 g
Area	1 ft^2	=	144 in.^{2}
	10.764 ft^2	=	1 m^2
	1 yd^2	=	9 ft^2
	1 mile^2 =		3.098 X 10^6 yd^2
Volume	7.48 gal	=	1 ft^3
	1 gal	=	3.785 l (liter)
	1 l	=	1000 cm^3

GEMS

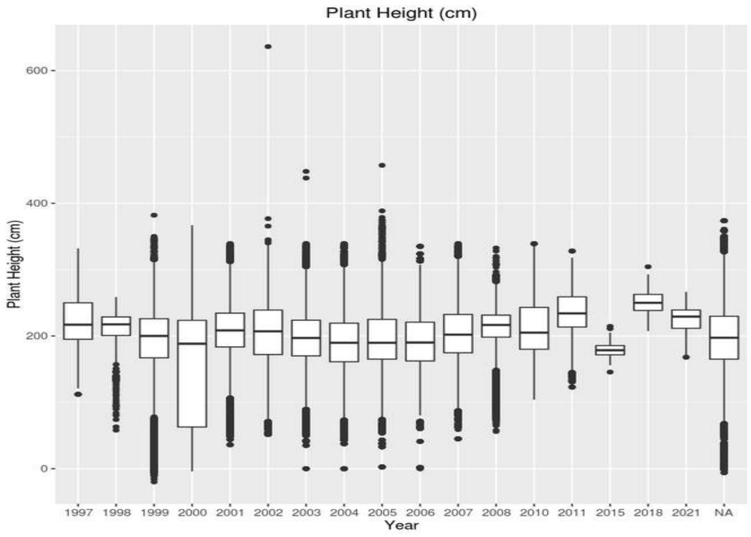
Data analysis – ad hoc investigation



Summarize plant heights globally

In [13]:

```
# Query 3: aggregate data on these three features globally
p <- ggplot(data=sub_tbl, aes(factor(year), value)) + geom_boxplot()
p+ggtitle("Plant Height (cm)") + labs(x="Year", y="Plant Height (cm)")
```



Data and Tools Sharing

GEMShare™

- Smart sharing -- Enables data providers to control who sees what, and when
- Supports open, private and pooled data
- Beyond data -- Enables sharing of data, tools and workflows

GEMSTools™ is an ever-expanding suite of analytical tools designed to

- Cleanup messy (meta-)data
- Intelligently impute missing data
- Enable data interoperability
- Apply advanced analytic methods to genomic, environmental, management and socio-economic data

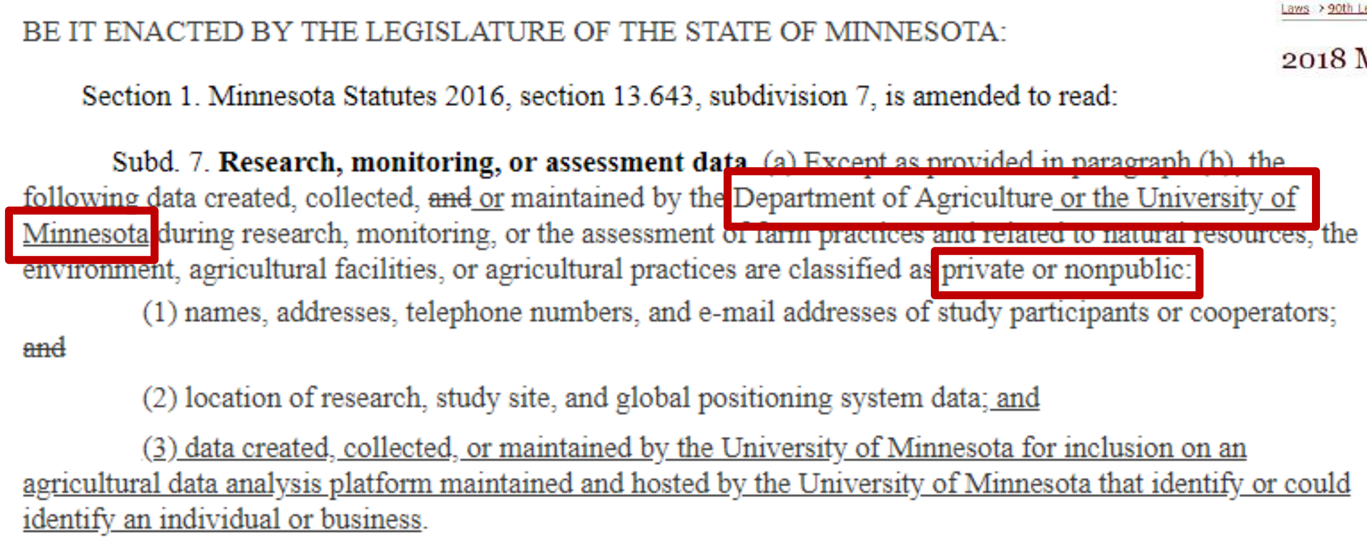
Your Data, Your Tools, Your Choice!

Technical Security

- Staff trained to handle sensitive data
- Ability to move the platform to the data
- Analyses run in isolated containers
- Servers hosted in a robust and secure data center
- Data encrypted at rest and in flight
- Systems constantly monitored

Legislative/Legal Privacy

2018 Minnesota Session Laws



BE IT ENACTED BY THE LEGISLATURE OF THE STATE OF MINNESOTA:

Section 1. Minnesota Statutes 2016, section 13.643, subdivision 7, is amended to read:

Subd. 7. **Research, monitoring, or assessment data.** (a) Except as provided in paragraph (b), the following data created, collected, ~~and~~ or maintained by the Department of Agriculture or the University of Minnesota during research, monitoring, or the assessment of farm practices and related to natural resources, the environment, agricultural facilities, or agricultural practices are classified as private or nonpublic:

(1) names, addresses, telephone numbers, and e-mail addresses of study participants or cooperators; ~~and~~

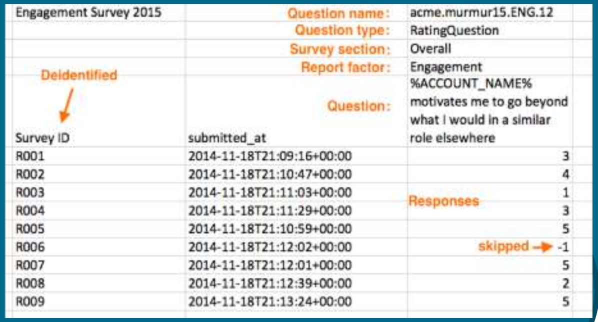
(2) location of research, study site, and global positioning system data; and

(3) data created, collected, or maintained by the University of Minnesota for inclusion on an agricultural data analysis platform maintained and hosted by the University of Minnesota that identify or could identify an individual or business.

Law came into effect August 1, 2018

GEMS

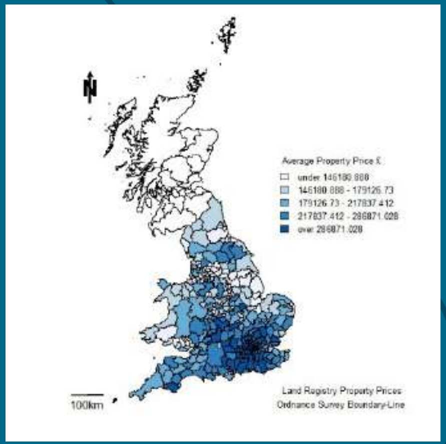
Data Sharing: Data Fuzzing



Engagement Survey 2015	Question name:	acme.murmur15.ENG.12
	Question type:	RatingQuestion
	Survey section:	Overall
Deidentified	Report factor:	Engagement
Survey ID	Question: submitted_at	%ACCOUNT_NAME% motivates me to go beyond what I would in a similar role elsewhere
R001	2014-11-18T21:09:16+00:00	3
R002	2014-11-18T21:10:47+00:00	4
R003	2014-11-18T21:11:03+00:00	1
R004	2014-11-18T21:11:29+00:00	3
R005	2014-11-18T21:10:59+00:00	5
R006	2014-11-18T21:12:02+00:00	skipped → -1
R007	2014-11-18T21:12:01+00:00	5
R008	2014-11-18T21:12:39+00:00	2
R009	2014-11-18T21:13:24+00:00	5

Responses

De-identifying

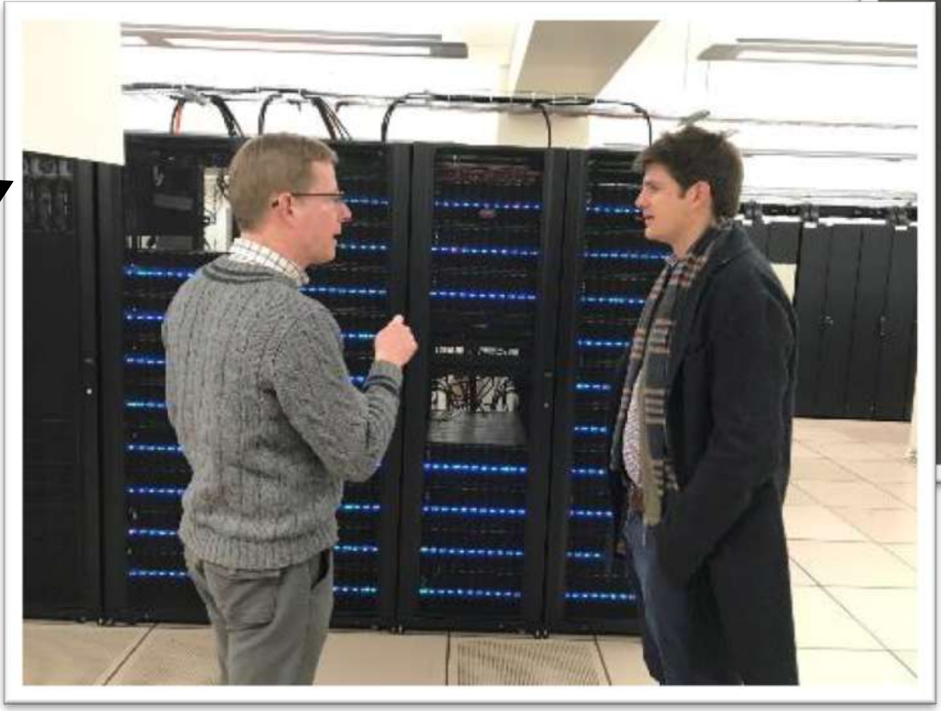


Aggregating

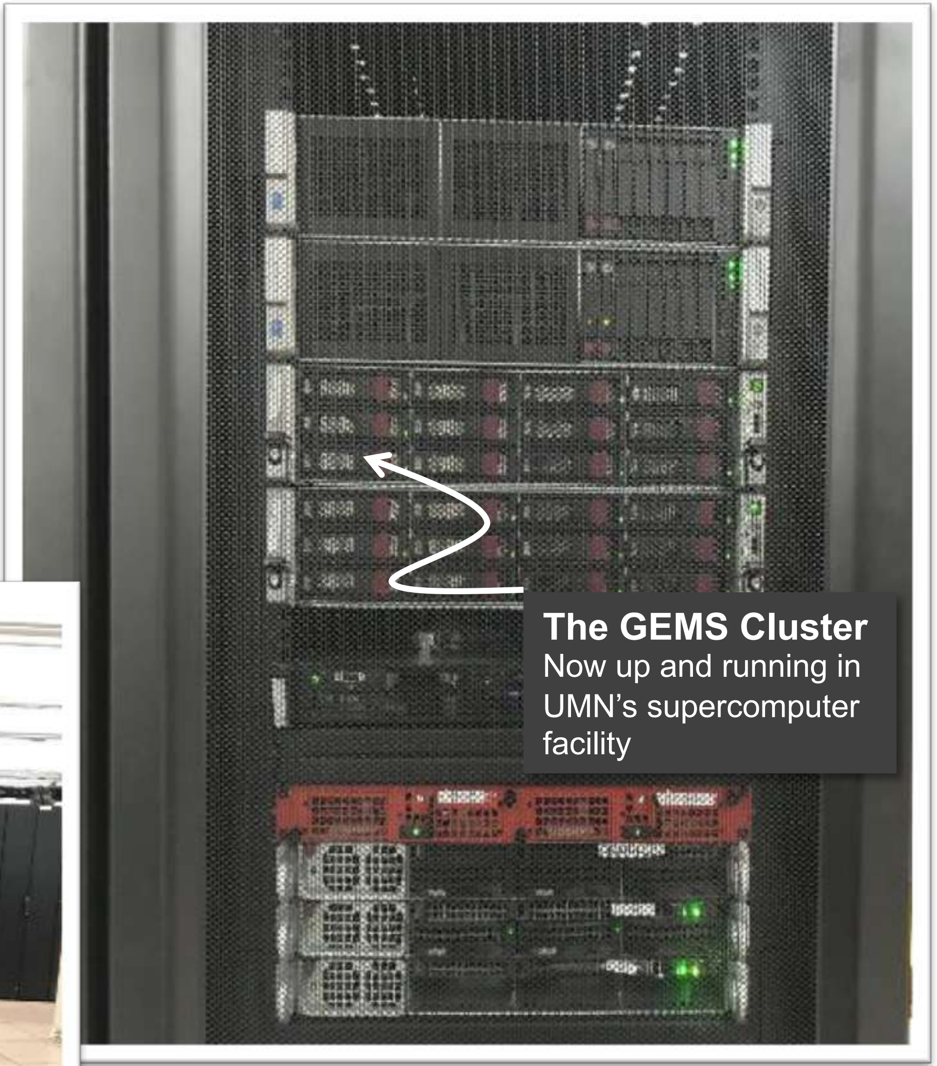
GEMS

Use Cases

Data to Actionable Information



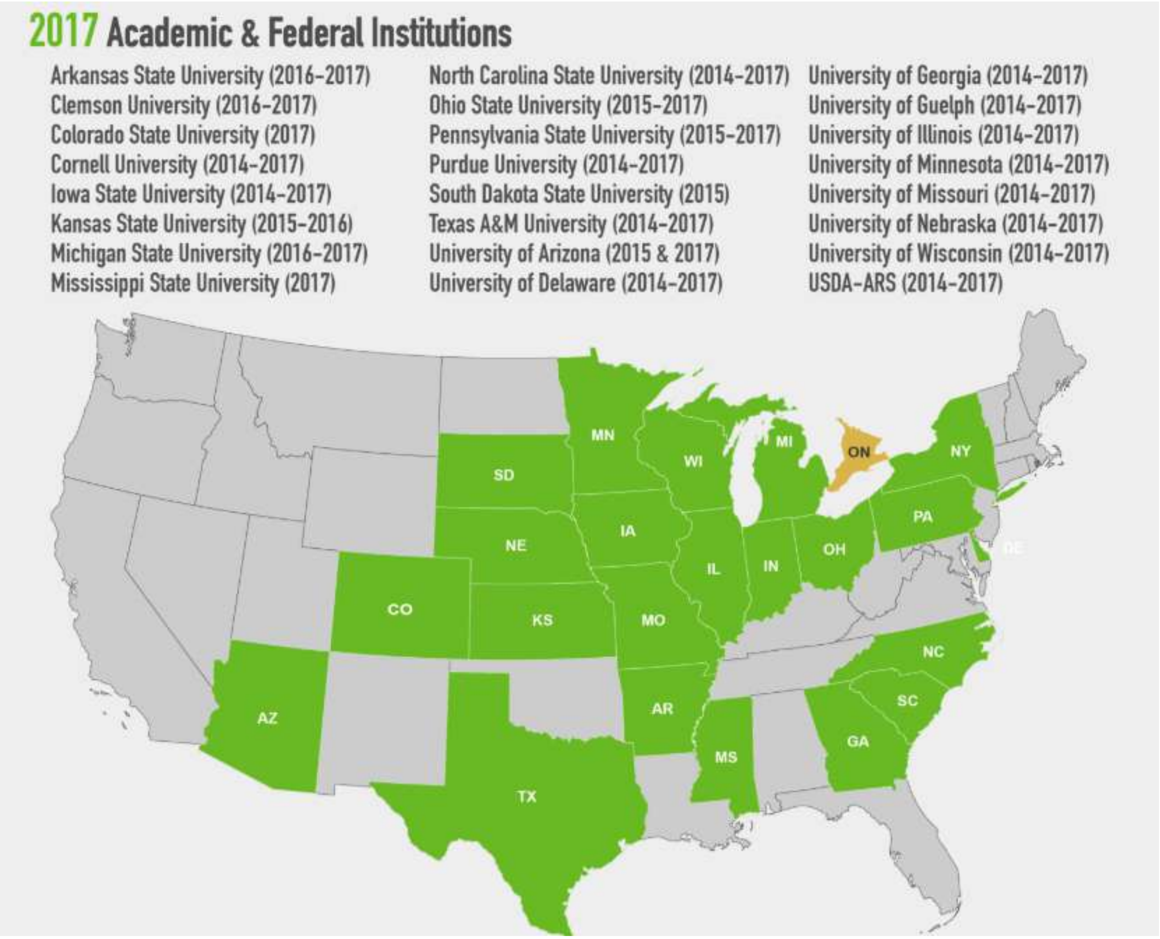
Jim Wilgenbusch , Director Minnesota Supercomputer Institute (left), chatting with Jan Greyling University of Stellenbosch, South Africa, local GEMS coordinator



The GEMS Cluster
Now up and running in UMN's supercomputer facility

Agroinformatics Support for G2F

the Genomes to Fields initiative



2017 Academic & Federal Institutions

Arkansas State University (2016-2017)
Clemson University (2016-2017)
Colorado State University (2017)
Cornell University (2014-2017)
Iowa State University (2014-2017)
Kansas State University (2015-2016)
Michigan State University (2016-2017)
Mississippi State University (2017)
North Carolina State University (2014-2017)
Ohio State University (2015-2017)
Pennsylvania State University (2015-2017)
Purdue University (2014-2017)
South Dakota State University (2015)
Texas A&M University (2014-2017)
University of Arizona (2015 & 2017)
University of Delaware (2014-2017)
University of Georgia (2014-2017)
University of Guelph (2014-2017)
University of Illinois (2014-2017)
University of Minnesota (2014-2017)
University of Missouri (2014-2017)
University of Nebraska (2014-2017)
University of Wisconsin (2014-2017)
USDA-ARS (2014-2017)

- Standardizing nomenclature, units etc
- Outlier detection
- Data interoperability (weather, soil, management, phenotypic measurement)
- Pilot linkage of field measurements from tablet to GEMSTools
- Manage data distribution among G2F partners and to the world
- Data mining and other (predictive) analytics

23 States, 37 experimental sites

IOWA CORN

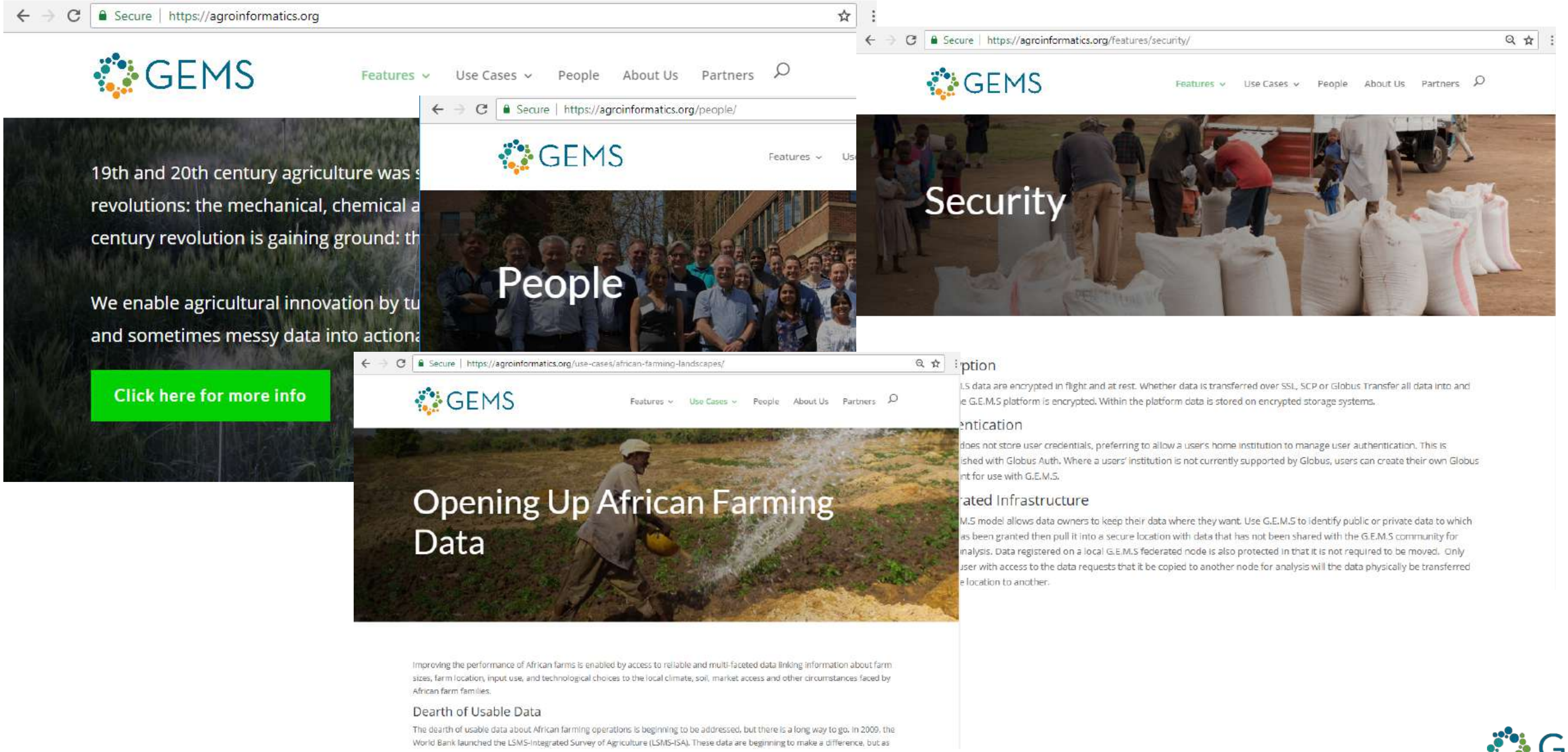
GEMS

International AgroInformatics Alliance

IAA 2.0 March 20-21, 2017, St. Paul MN

IAA 3.0 May 2-3, 2018, St. Paul MN

GEMS Web site: now online!


Secure | https://agroinformatics.org
GEMS
Features
Use Cases
People
About Us
Partners
Secure | https://agroinformatics.org/people/
Secure | https://agroinformatics.org/features/security/
Secure | https://agroinformatics.org/use-cases/african-farming-landscapes/
Click here for more info

19th and 20th century agriculture was s
revolutions: the mechanical, chemical a
century revolution is gaining ground: th

We enable agricultural innovation by tu
and sometimes messy data into actiona

People

Security

ption

.S data are encrypted in flight and at rest. Whether data is transferred over SSL, SCP or Globus Transfer all data into and
e G.E.M.S platform is encrypted. Within the platform data is stored on encrypted storage systems.

entication

does not store user credentials, preferring to allow a user's home institution to manage user authentication. This is
ished with Globus Auth. Where a users' institution is not currently supported by Globus, users can create their own Globus
nt for use with G.E.M.S.

ated Infrastructure

M.S model allows data owners to keep their data where they want. Use G.E.M.S to identify public or private data to which
as been granted then pull it into a secure location with data that has not been shared with the G.E.M.S community for
nalysis. Data registered on a local G.E.M.S federated node is also protected in that it is not required to be moved. Only
user with access to the data requests that it be copied to another node for analysis will the data physically be transferred
e location to another.

Opening Up African Farming Data

Improving the performance of African farms is enabled by access to reliable and multi-faceted data linking information about farm sizes, farm location, input use, and technological choices to the local climate, soil, market access and other circumstances faced by African farm families.

Dearth of Usable Data

The dearth of usable data about African farming operations is beginning to be addressed, but there is a long way to go. In 2009, the World Bank launched the LSMS-Integrated Survey of Agriculture (LSMS-ISA). These data are beginning to make a difference, but as

Dr. Norman E. Borlaug
1914 - 2009
University of Minnesota
B.S. Forestry 1937
M.S. Plant Pathology 1941
Ph.D. Plant Pathology 1942
"If you desire peace, cultivate justice, but at the same time cultivate the fields to produce more bread; otherwise there will be no peace."
Nobel Peace Prize (1970)
Presidential Medal of Freedom (1977)
National Academy of Sciences, the National Medal of Science (2006)
Congressional Gold Medal (2007)

Thanks

G.E.M.S:
https://agroinformatics.org

GEMS