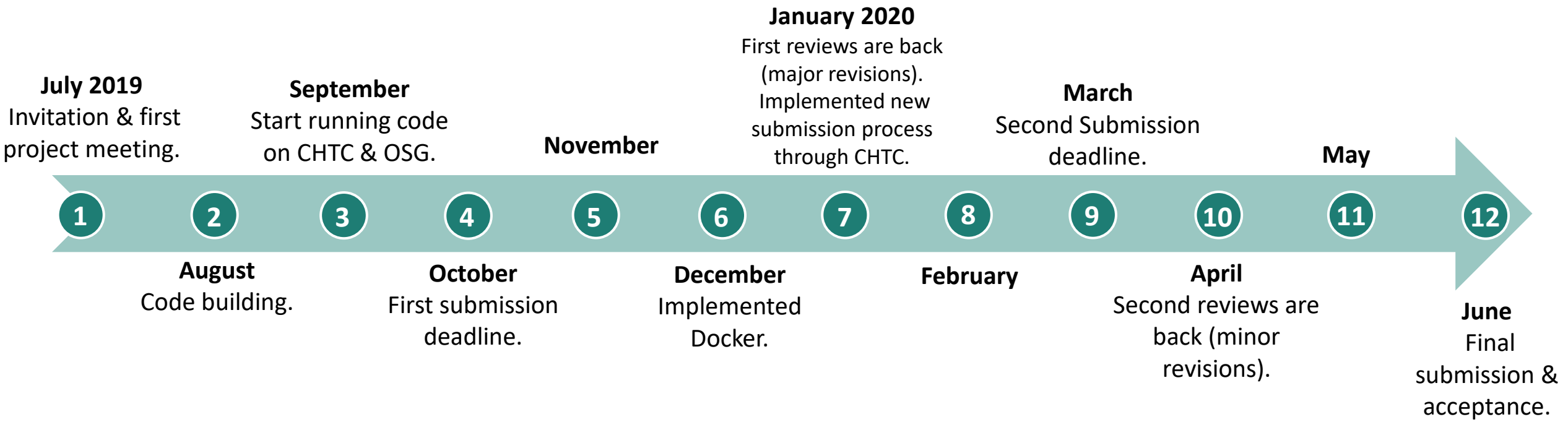
OSG resources enable optimization of field experimental designs for agriculture

OSG All-Hands Meeting 2020

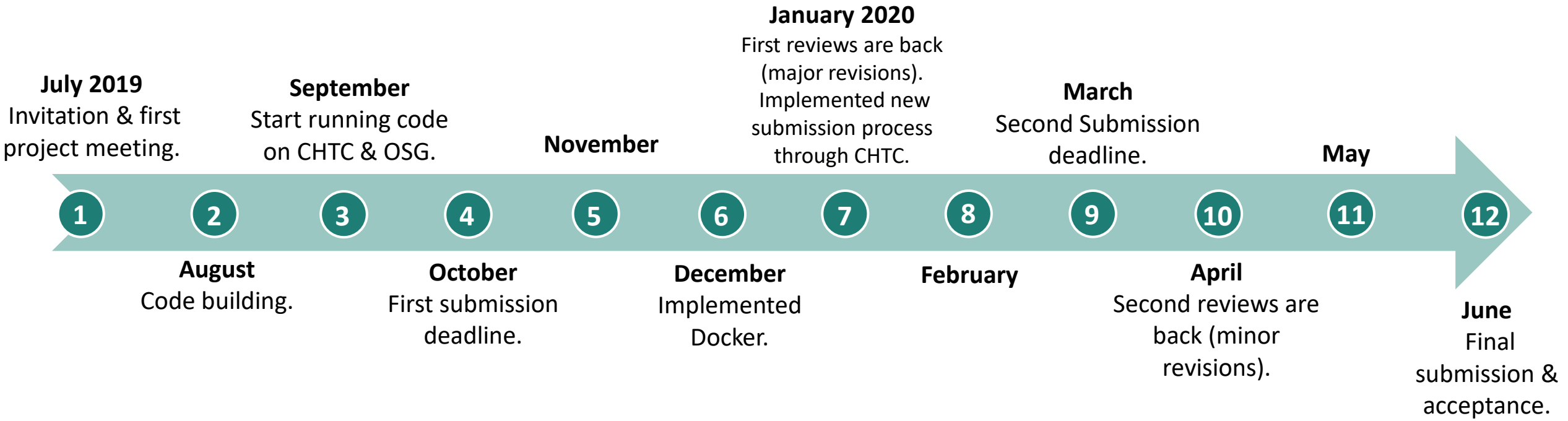
Cereals Breeding and Quantitative Genetics
University of Wisconsin-Madison

Raegan Hoefler, Pablo Gonzalez-Barrios, Madhav Bhatta, Jose A. R. Nunes, Ines Berro, Rafael S. Nalin, Alejandra Borges, Eduardo Covarrubias, Luis Diaz-Garcia, Martin Quincke, Lucia Gutierrez

Overview

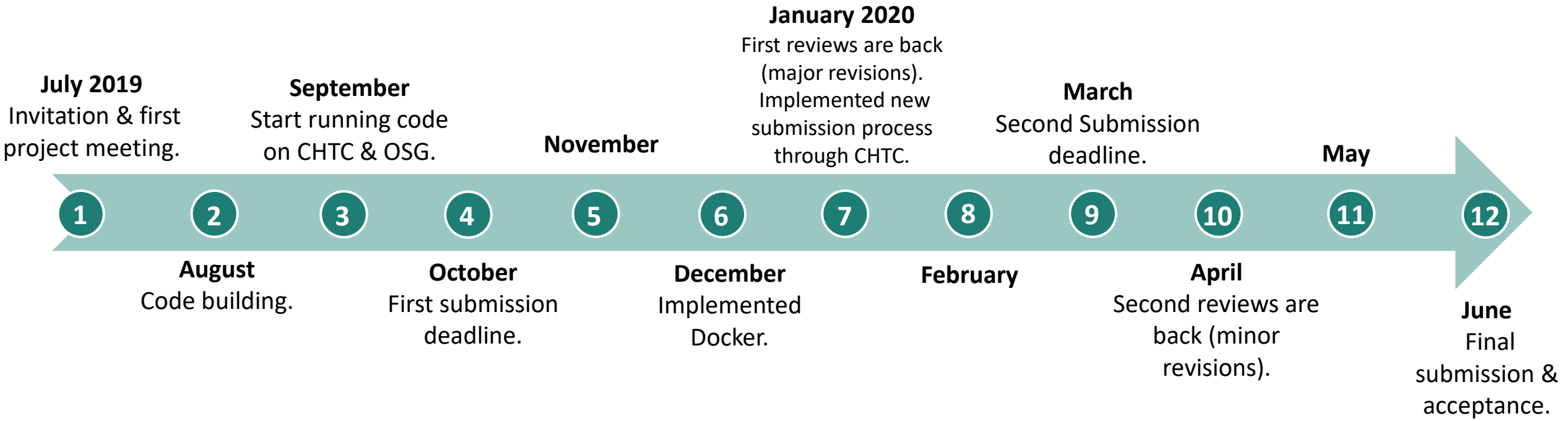

July 2019
Invitation & first project meeting.
August
Code building.
September
Start running code on CHTC & OSG.
October
First submission deadline.
November
December
Implemented Docker.
January 2020
First reviews are back (major revisions). Implemented new submission process through CHTC.
February
March
Second Submission deadline.
April
Second reviews are back (minor revisions).
May
June
Final submission & acceptance.
1
2
3
4
5
6
7
8
9
10
11
12

Overview



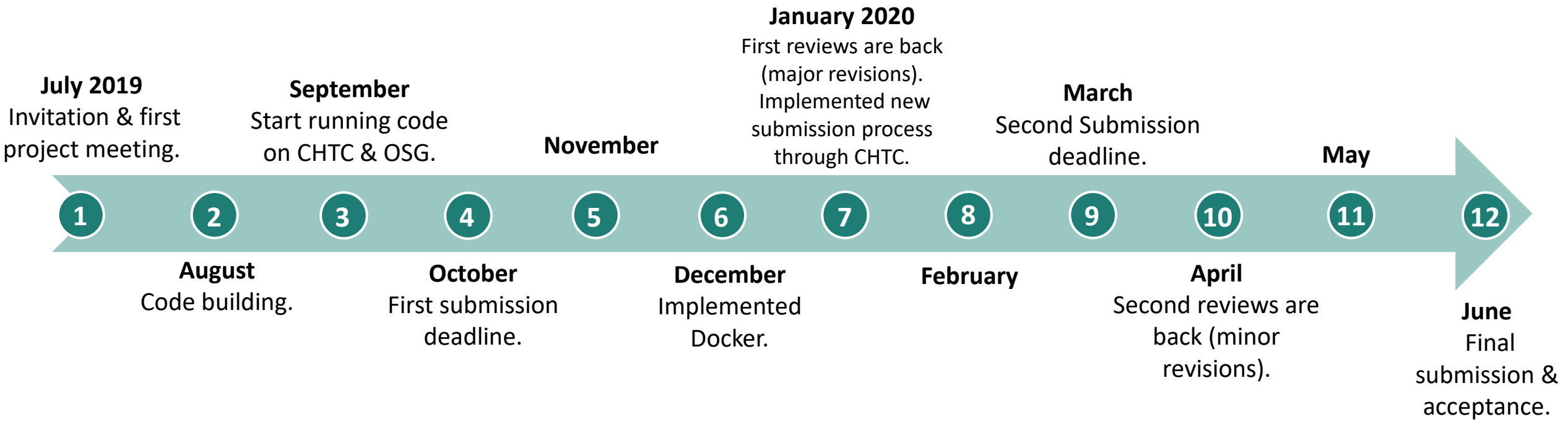
3.73 M hours through OSG Connect

Overview



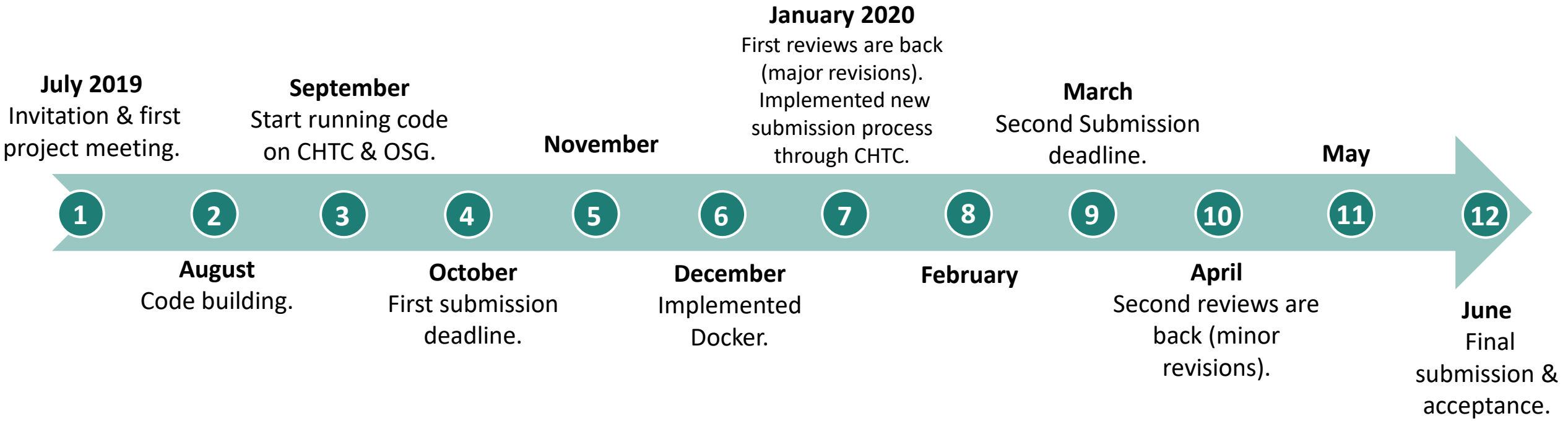
3.73 M hours through OSG Connect
+ 533,000 hours through OSG via CHTC

Overview



3.73 M hours through OSG Connect
+ 533,000 hours through OSG via CHTC
+ 1.8 M hours through CHTC

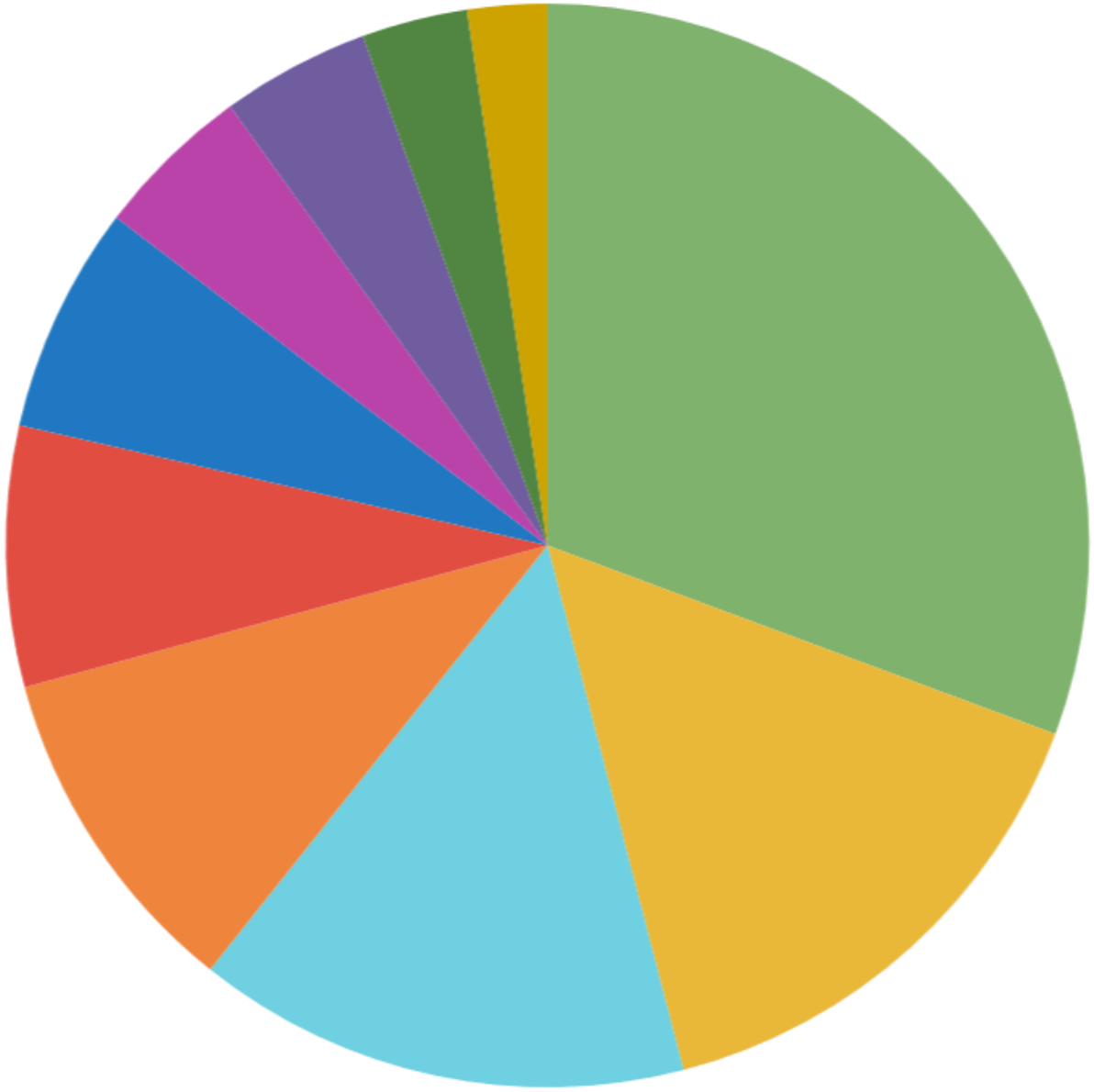
Overview



3.73 M hours through OSG Connect
+ 533,000 hours through OSG via CHTC
+ 1.8 M hours through CHTC

= 6 M hours (many triumphs, many fails)

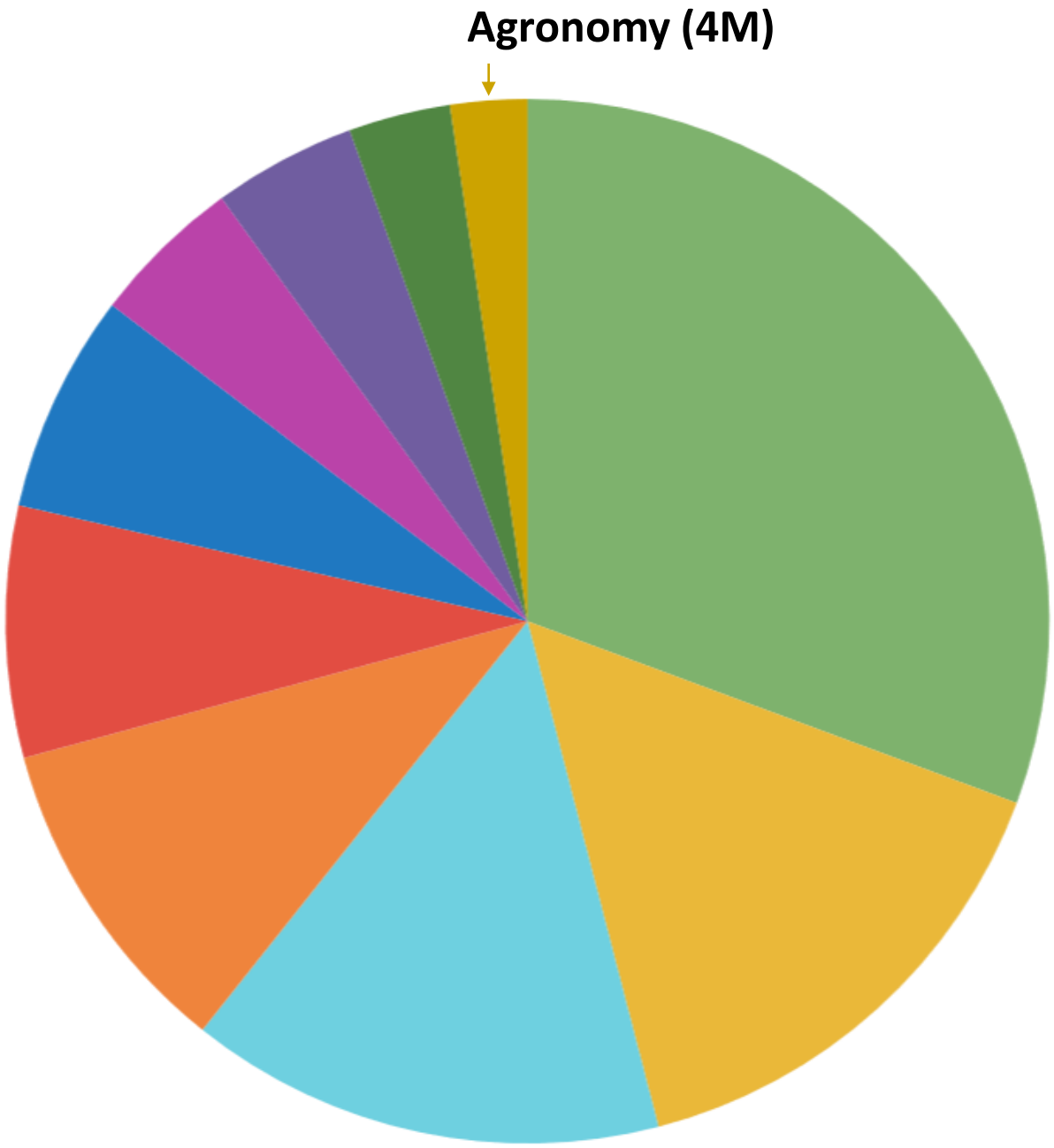
Overview



OSG Connect Summary:
Total Wall Hours for Top 10 Projects
July 1, 2019 - June 30, 2020

https://gracc.opensciencegrid.org/d/000000099/osg-connect-summary-osgconnect-net-submit-hosts-only?orgId=1&from=1561939200000&to=1593561599000&var-interval=1d&var-project=All&var-user=All&var-Filter=ProbeName%7C%3D~%7C.*osgconnect%5C.net

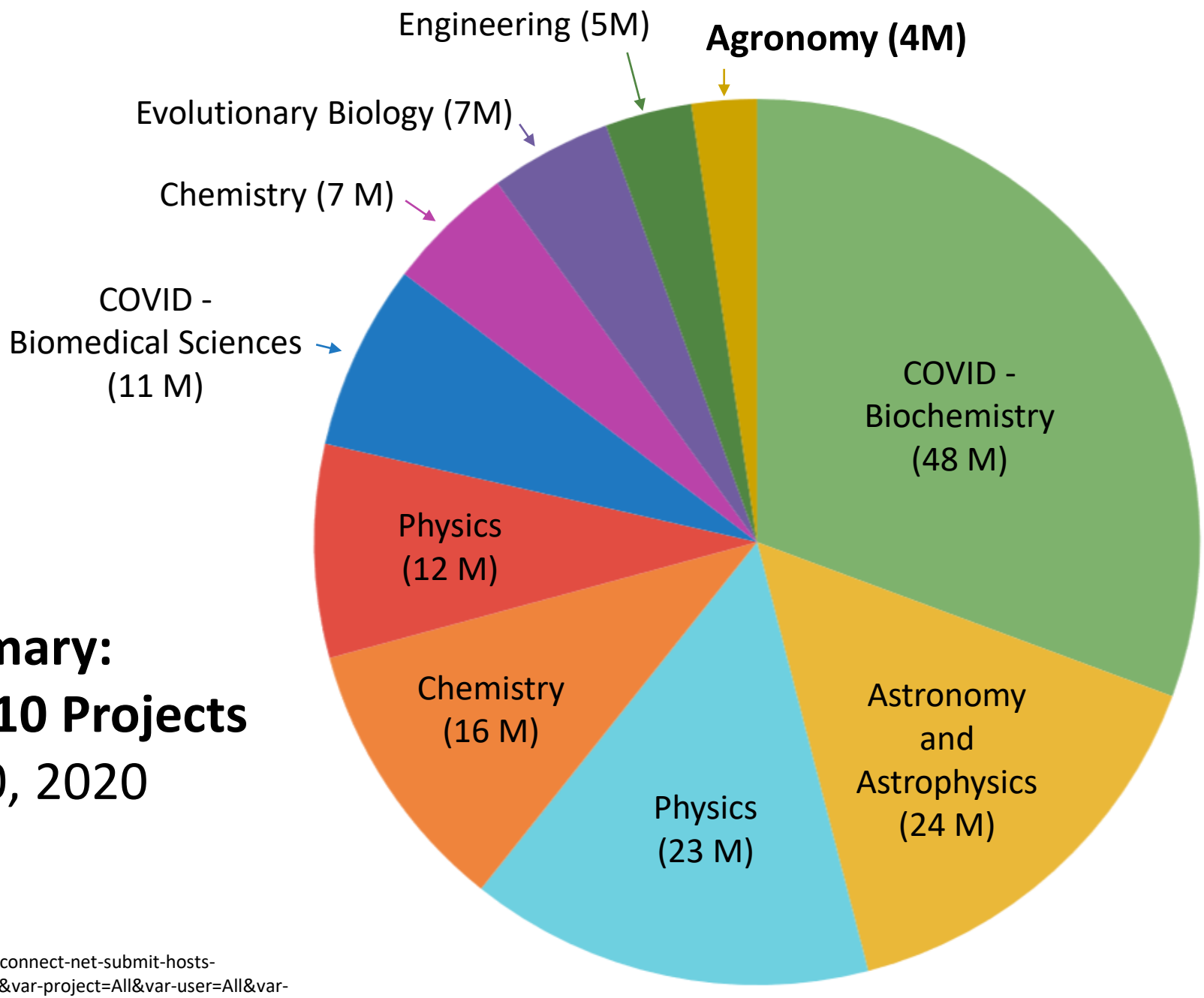
Overview



OSG Connect Summary:
Total Wall Hours for Top 10 Projects
July 1, 2019 - June 30, 2020

https://gracc.opensciencegrid.org/d/000000099/osg-connect-summary-osgconnect-net-submit-hosts-only?orgId=1&from=1561939200000&to=1593561599000&var-interval=1d&var-project=All&var-user=All&var-Filter=ProbeName%7C%3D~%7C.*osgconnect%5C.net

Overview



OSG Connect Summary:
Total Wall Hours for Top 10 Projects
July 1, 2019 - June 30, 2020

https://gracc.opensciencegrid.org/d/000000099/osg-connect-summary-osgconnect-net-submit-hosts-only?orgId=1&from=1561939200000&to=1593561599000&var-interval=1d&var-project=All&var-user=All&var-Filter=ProbeName%7C%3D~%7C.*osgconnect%5C.net

July 2019

Invitation & first project meeting.

July 2019
Invitation & first project meeting.

- Our group: Quantitative Geneticists, Statisticians, and Breeders

July 2019
Invitation & first project meeting.

- Our group: Quantitative Geneticists, Statisticians, and Breeders
- Our areas of research: experimental design optimization for genomic studies and modeling genotype by environment interactions

July 2019
Invitation & first project meeting.

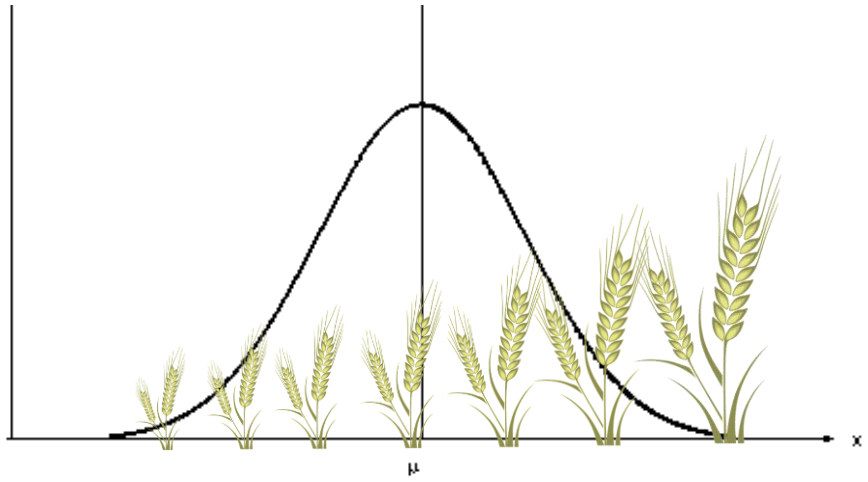
- Our group: Quantitative Geneticists, Statisticians, and Breeders
- Our areas of research: experimental design optimization for genomic studies and modeling genotype by environment interactions
- Invited paper: Journal of Agricultural Biological and Environmental Statistics (JABES)

July 2019
Invitation & first project meeting.

- Our group: Quantitative Geneticists, Statisticians, and Breeders
- Our areas of research: experimental design optimization for genomic studies and modeling genotype by environment interactions
- Invited paper: Journal of Agricultural Biological and Environmental Statistics (JABES)
  - Questions of interest:
    1. Do spatial designs outcompete randomization-based designs?
    2. Do additional methods of spatial correction help?

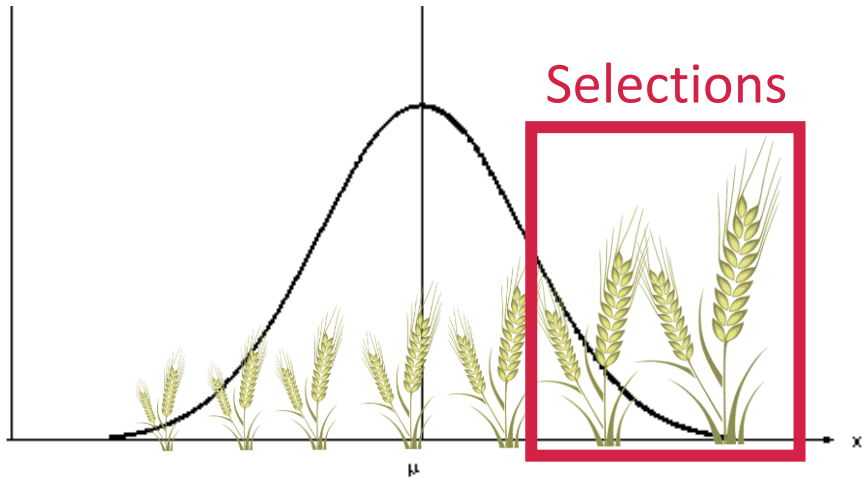
Background: Response to Selection

Parents

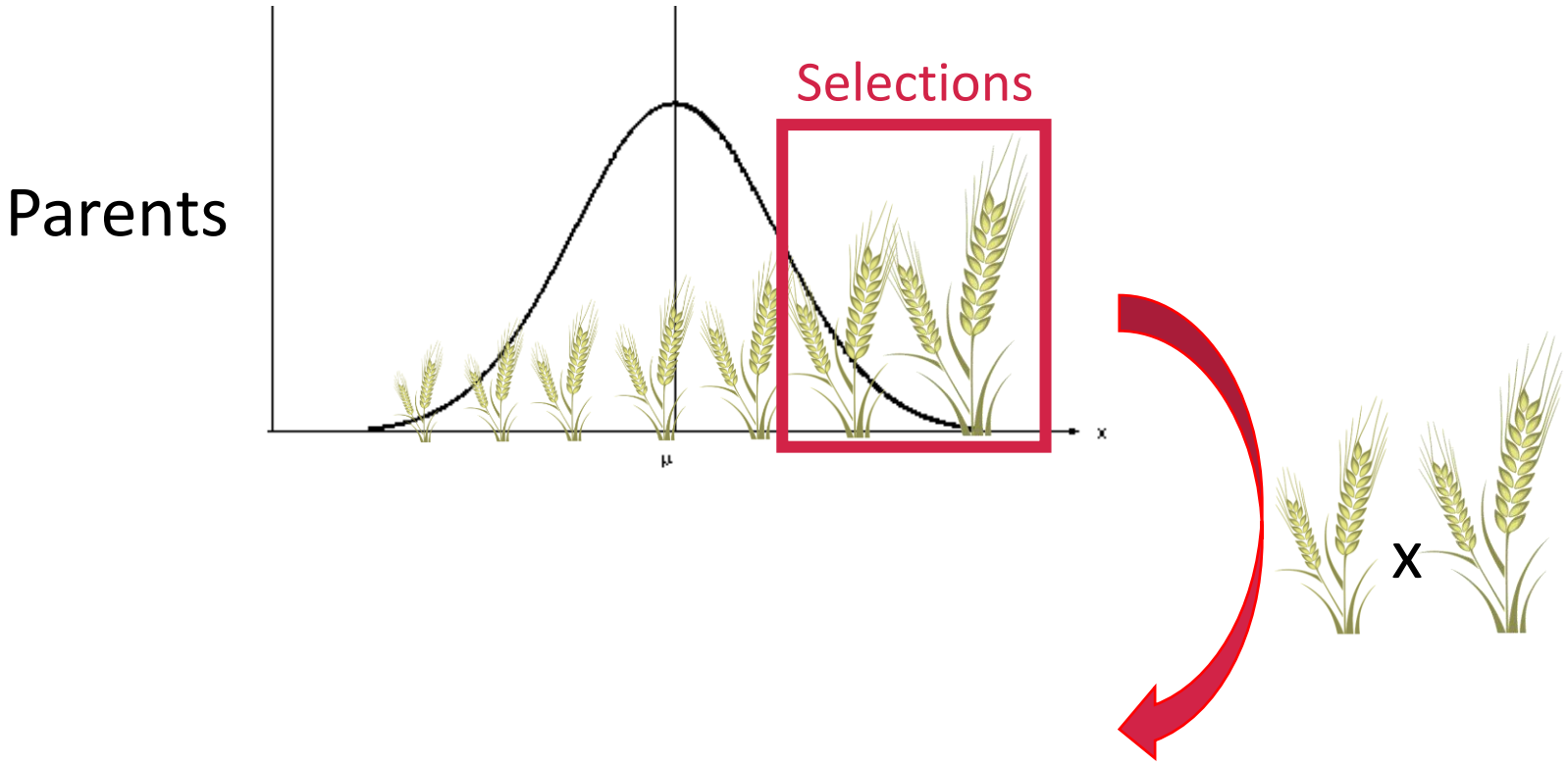


Background: Response to Selection

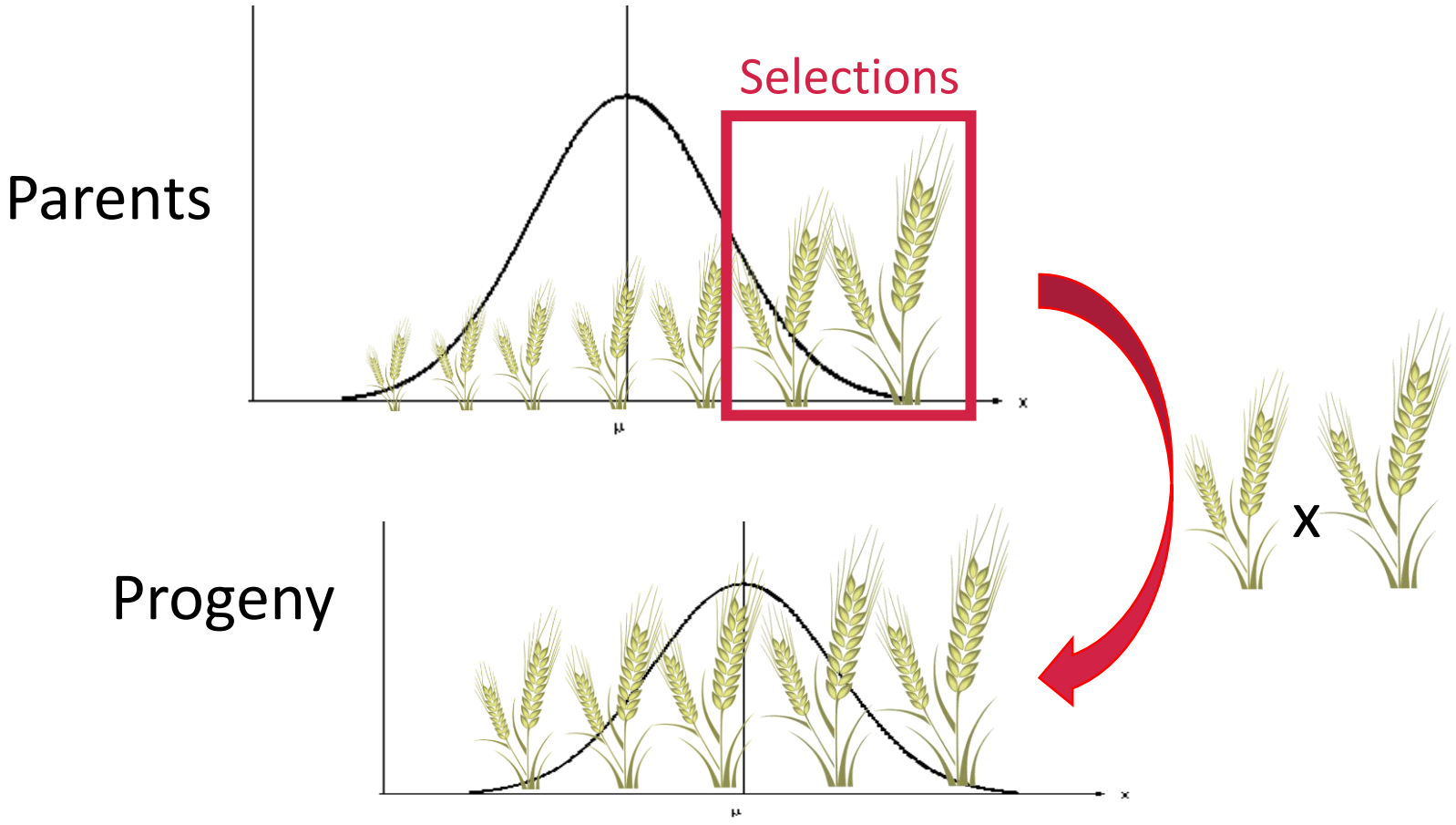
Parents



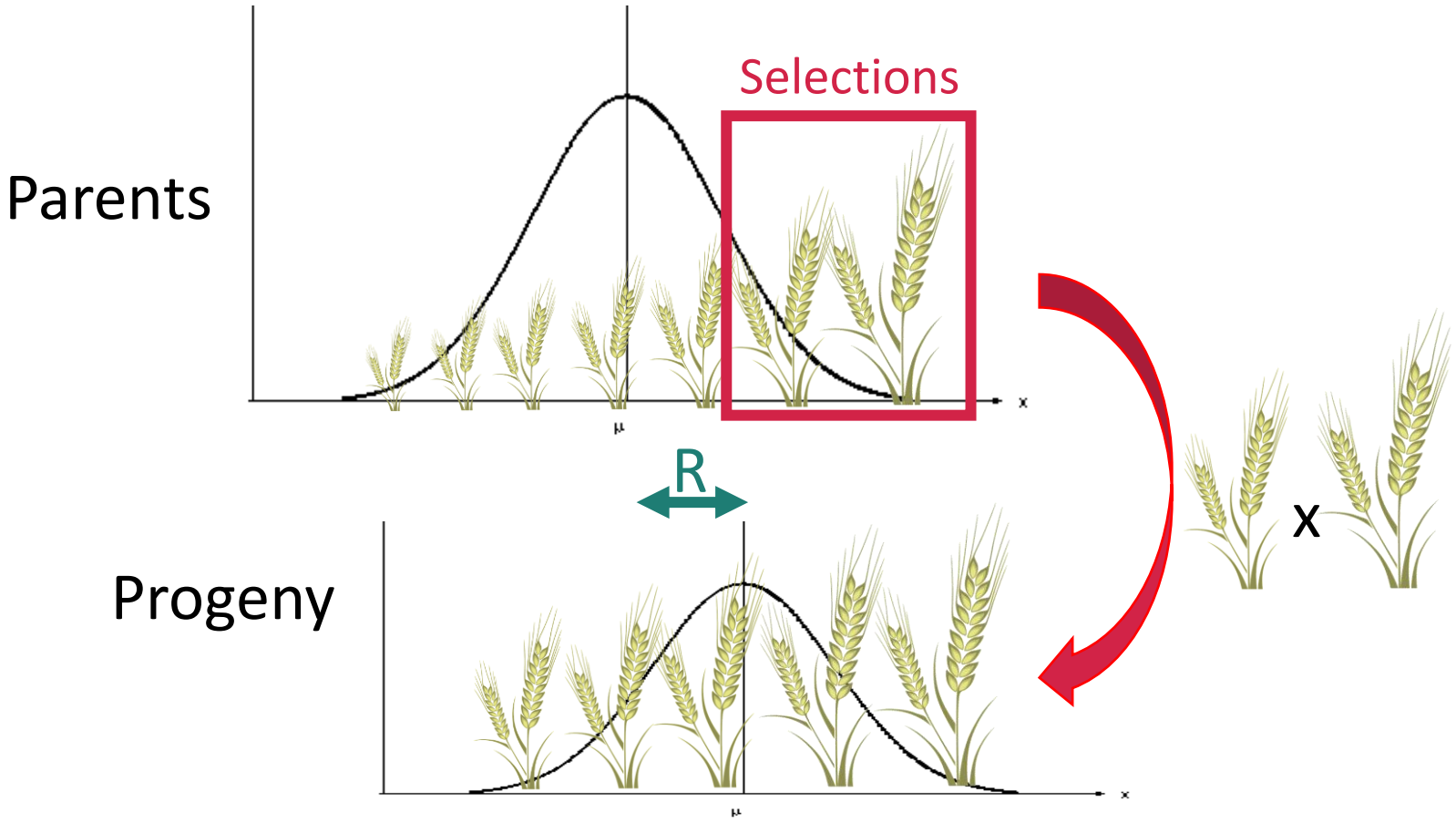
Background: Response to Selection



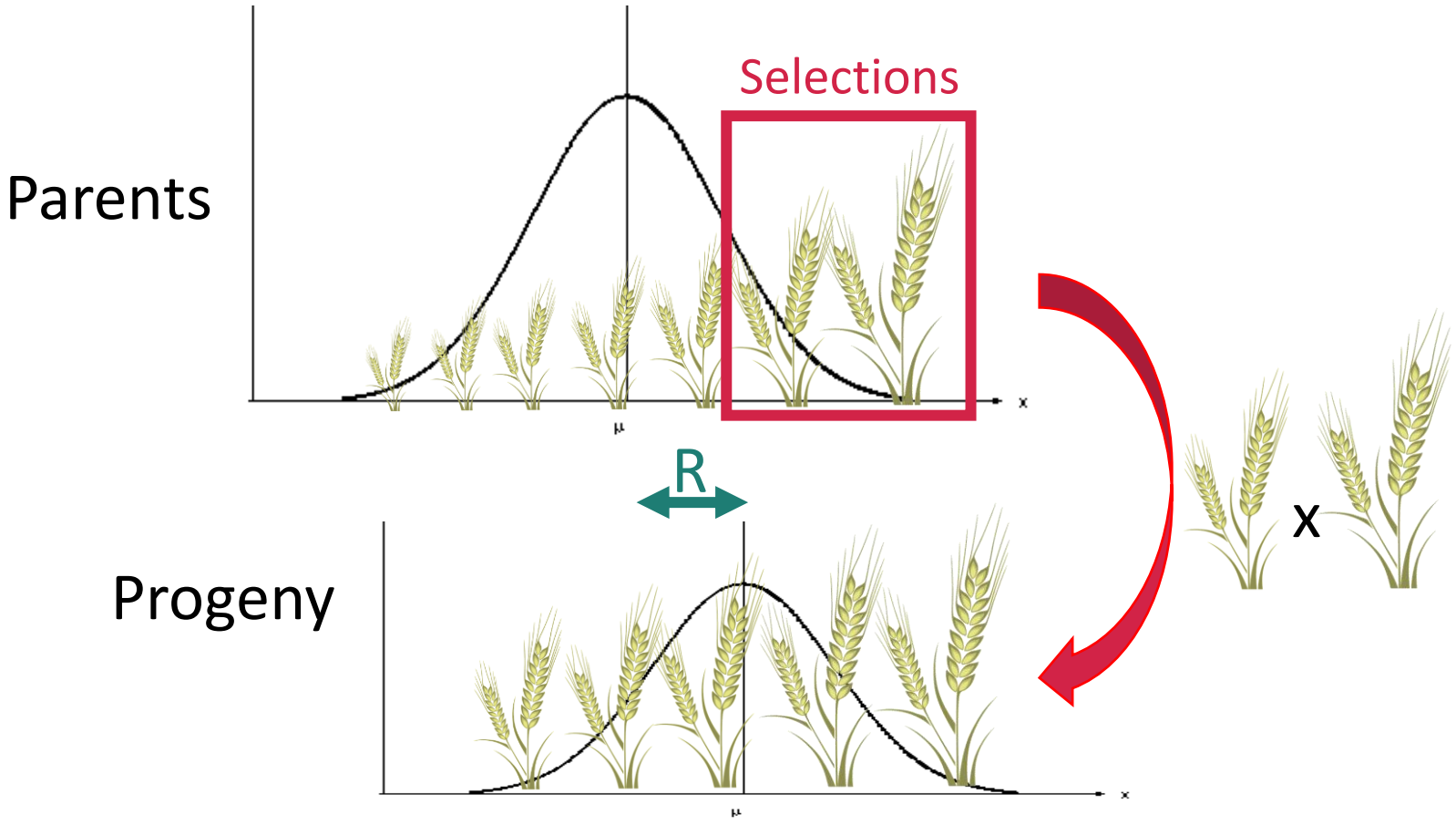
Background: Response to Selection



Background: Response to Selection



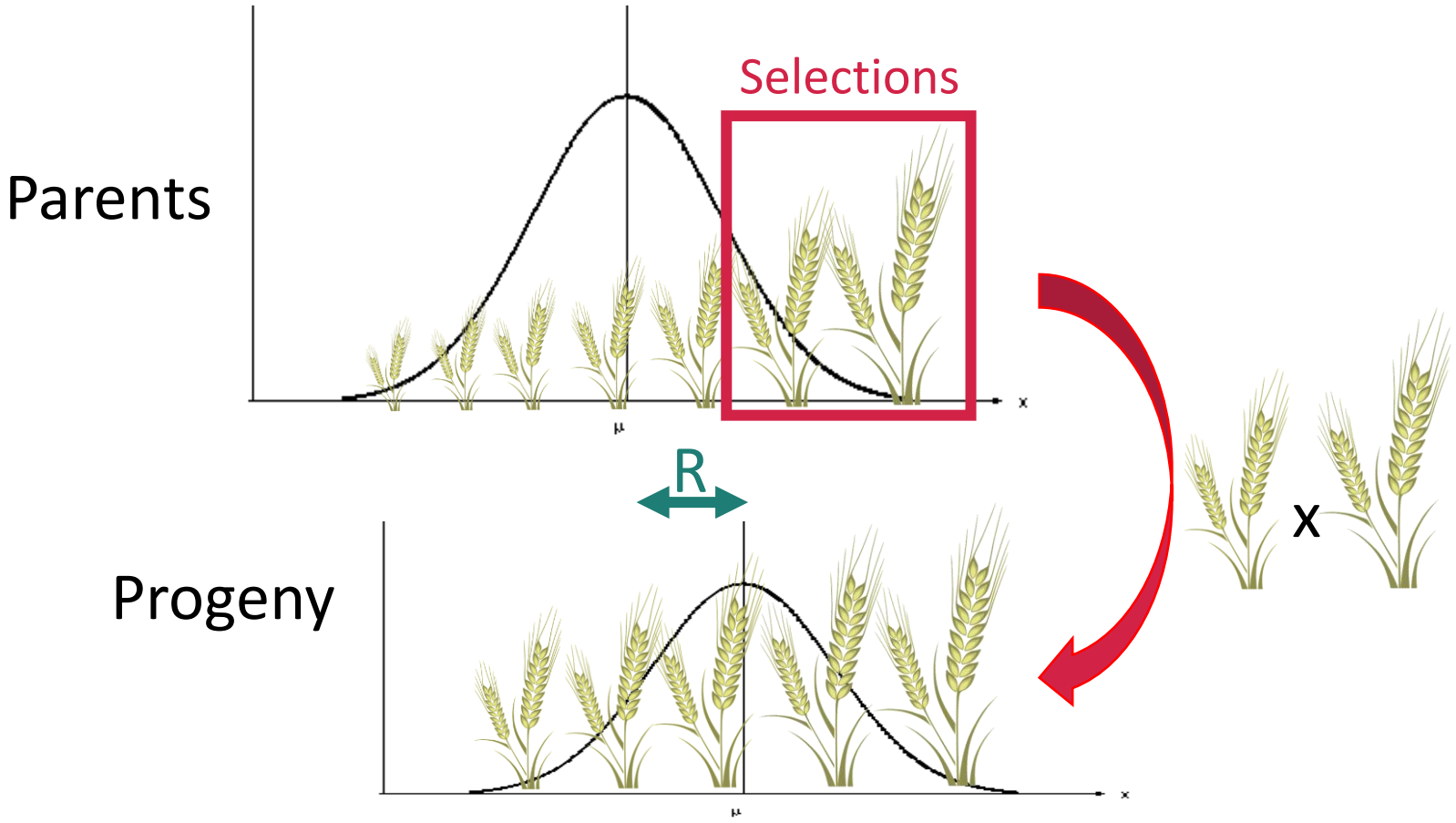
Background: Response to Selection



"Breeder's Equation"

$$R = ir_{g,\hat{g}}\sigma_g$$

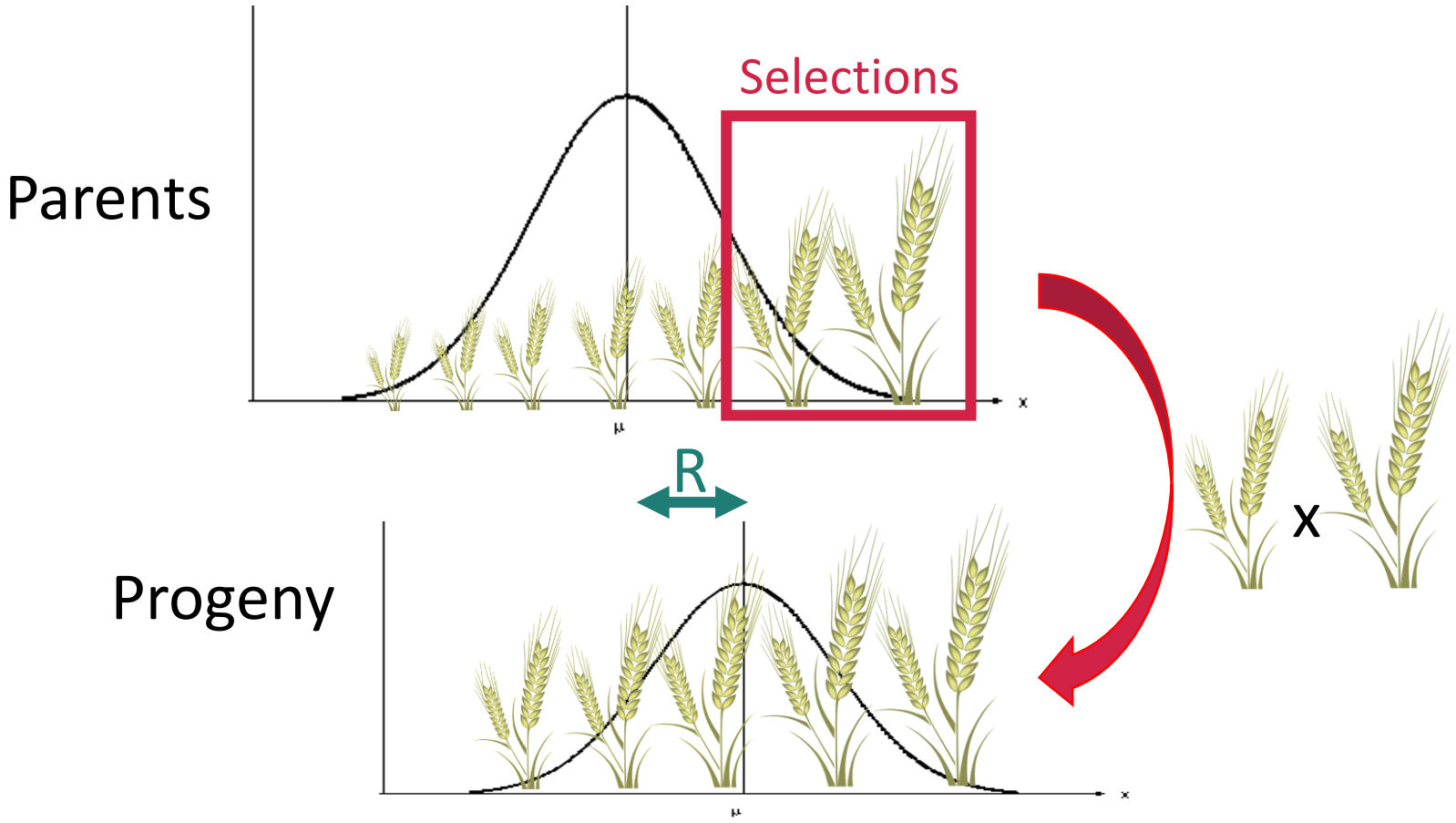
Background: Response to Selection



"Breeder's Equation"

$$R = i\underbrace{r_{g,\hat{g}}}_{\text{Correlation}}\sigma_g$$

Background: Response to Selection

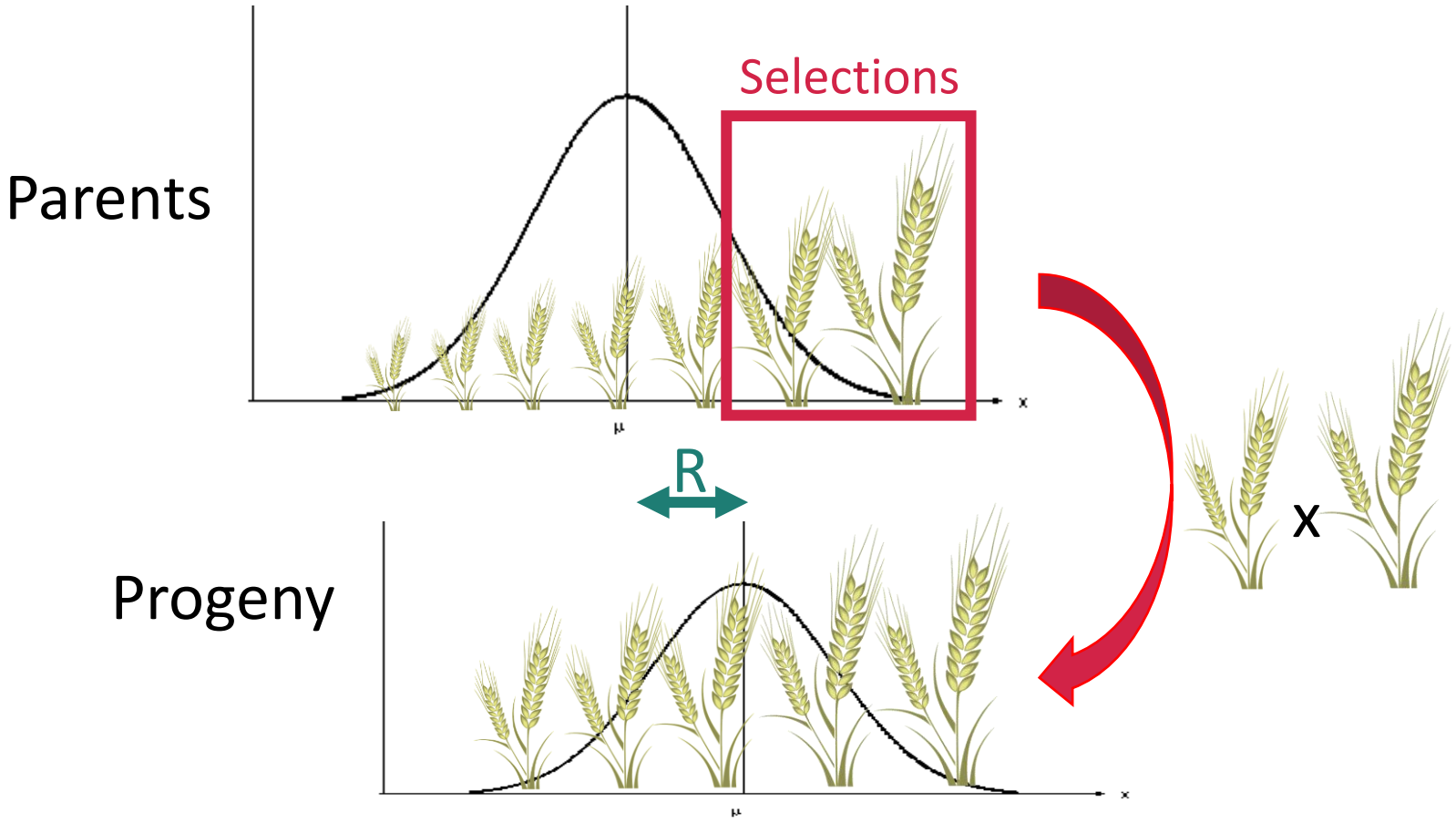


"Breeder's Equation"

$$R = i\underbrace{r_{g,\hat{g}}}_{\text{Correlation}}\sigma_g$$

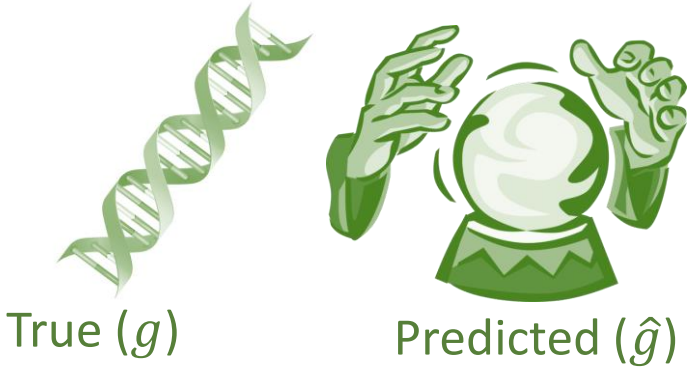
True (g)

Background: Response to Selection



"Breeder's Equation"

$$R = i\underbrace{r_{g,\hat{g}}}_{\text{Correlation}}\sigma_g$$

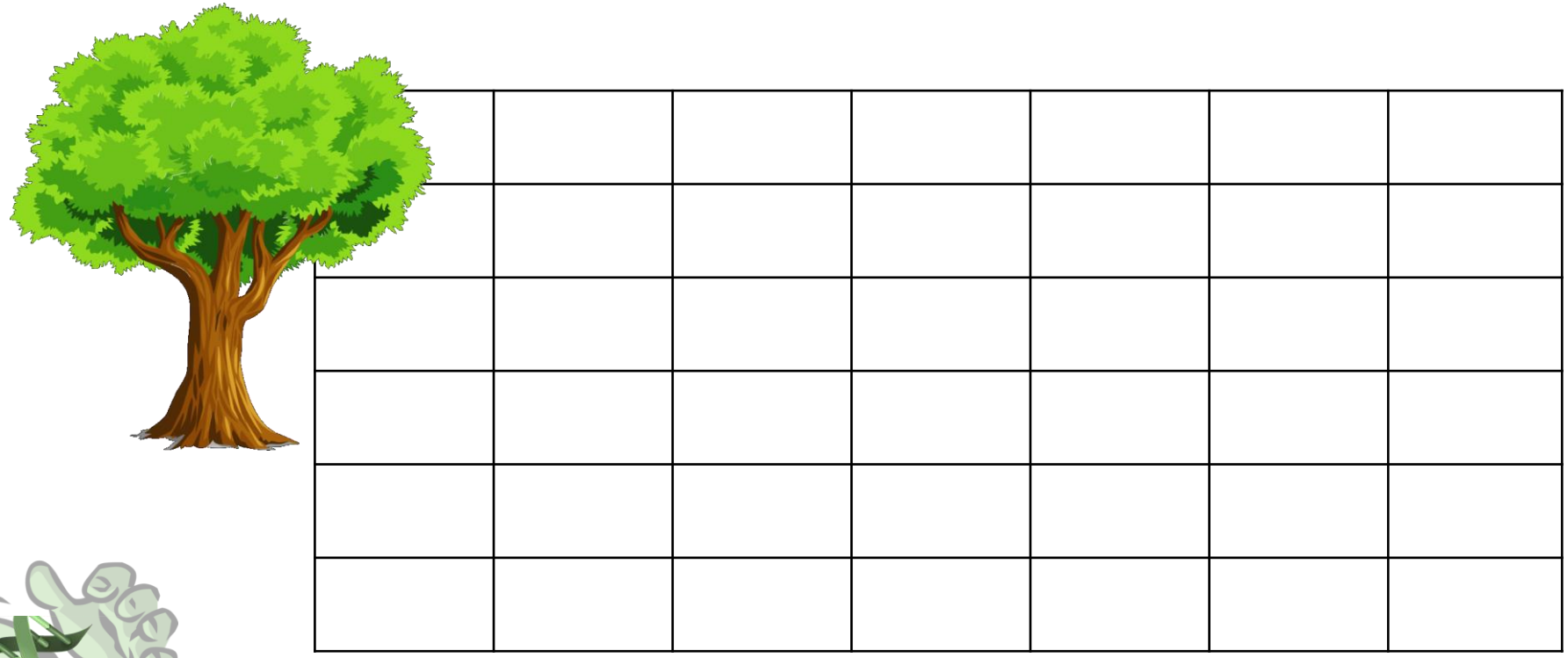


Background: Spatial Variation

Background: Spatial Variation

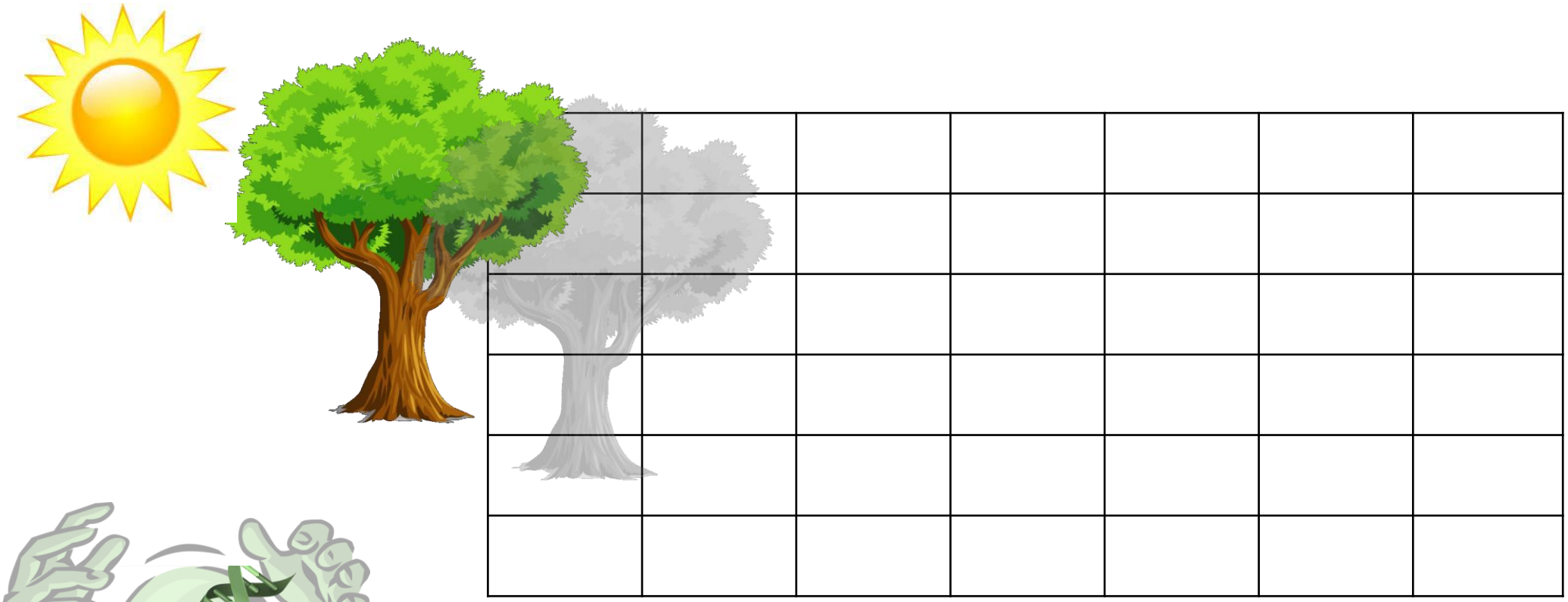
Must control spatial variation!

Background: Spatial Variation



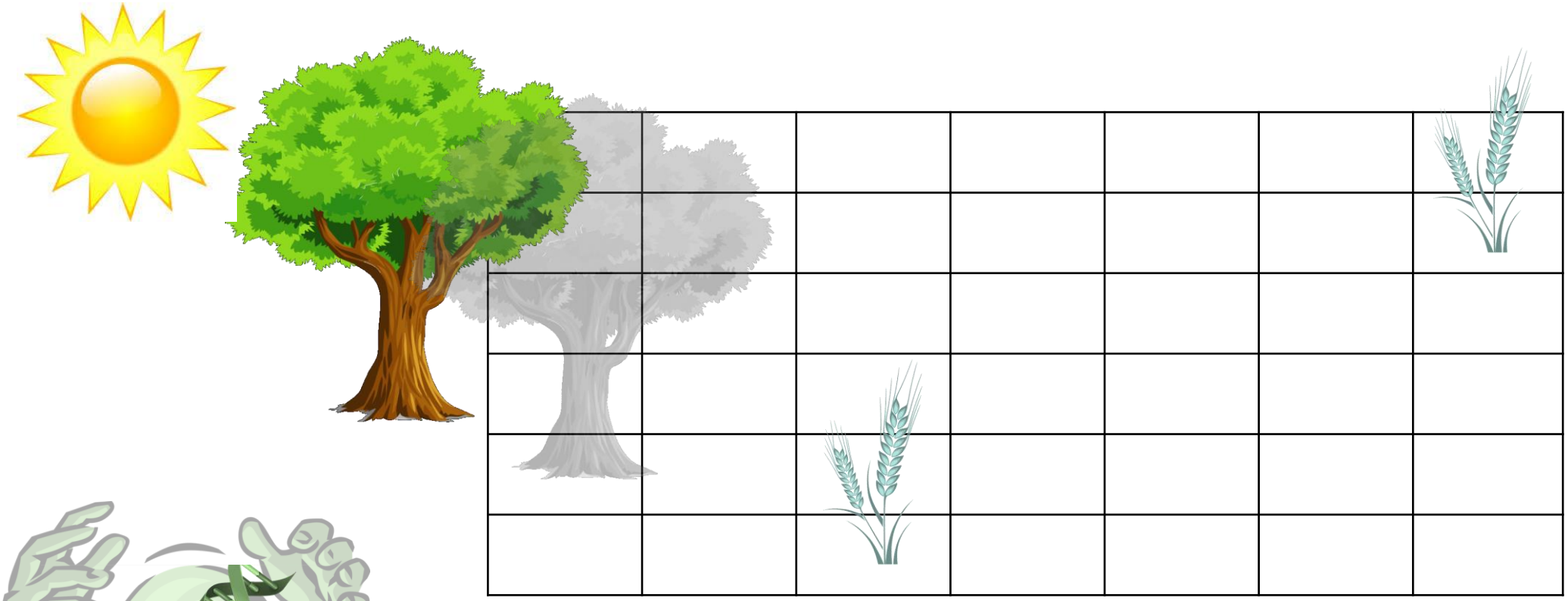
Must control spatial variation!

Background: Spatial Variation



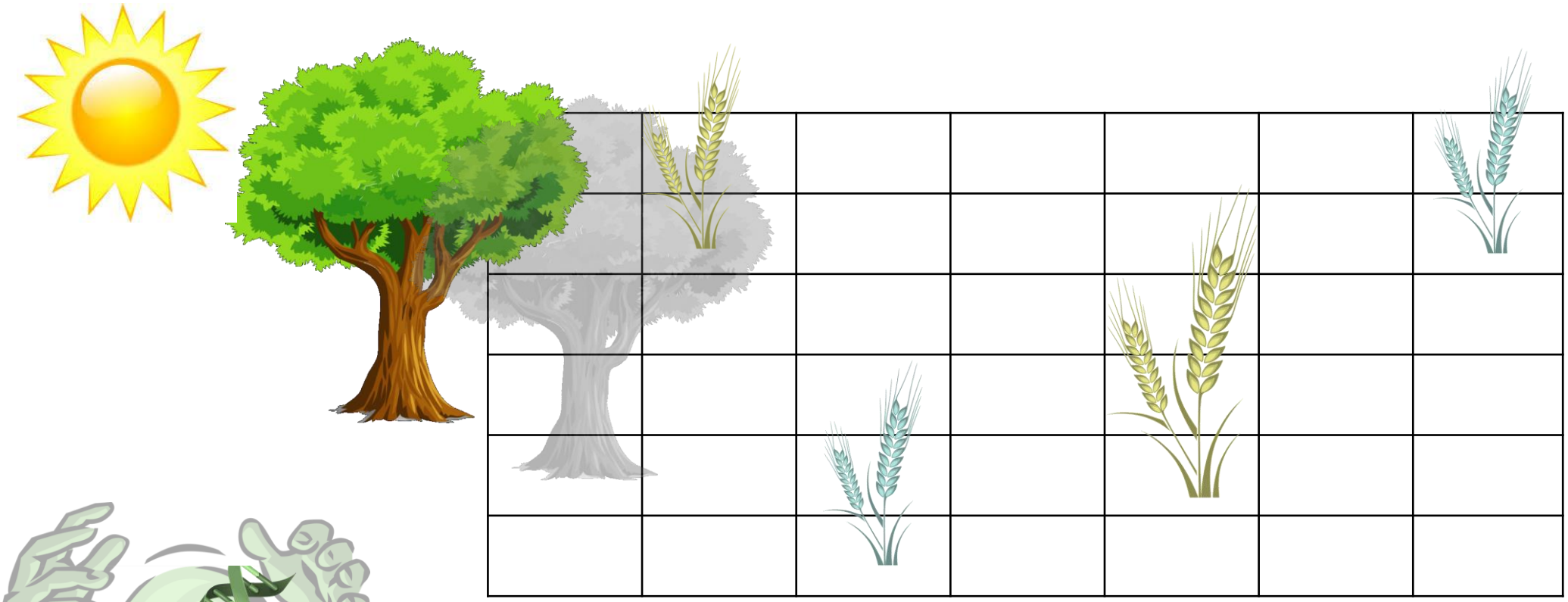
Must control spatial variation!

Background: Spatial Variation



Must control spatial variation!

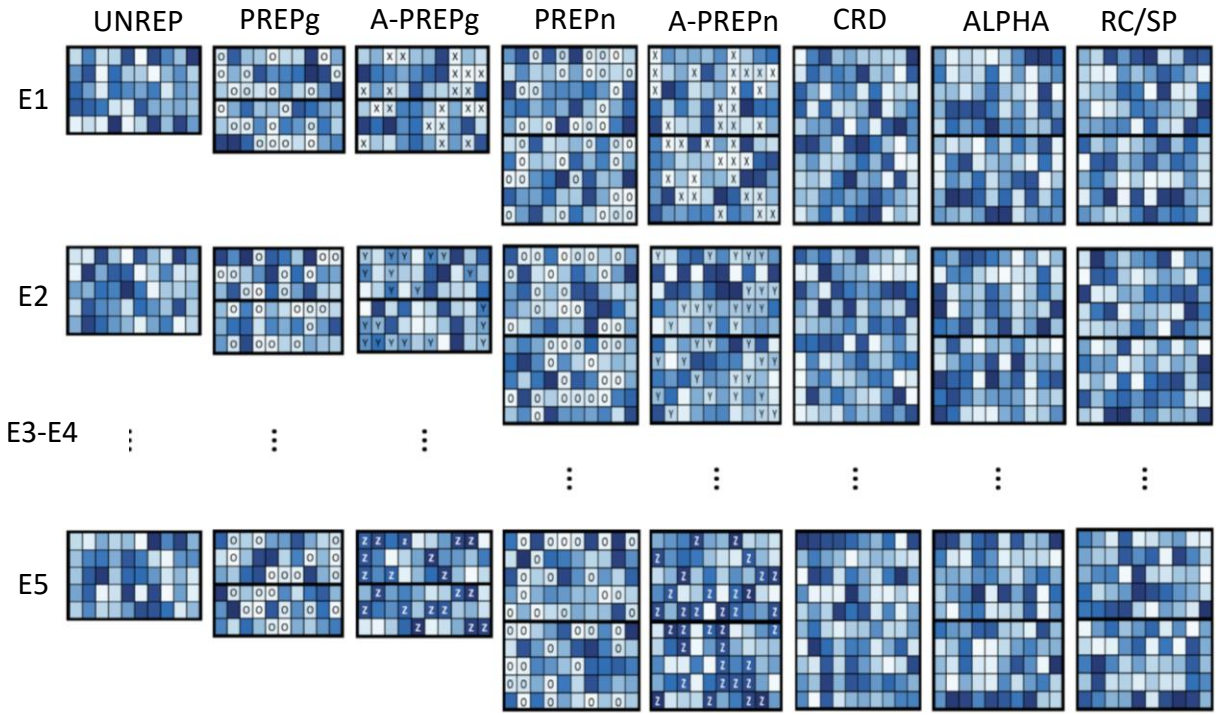
Background: Spatial Variation



Must control spatial variation!

Background: Experimental Design

Design

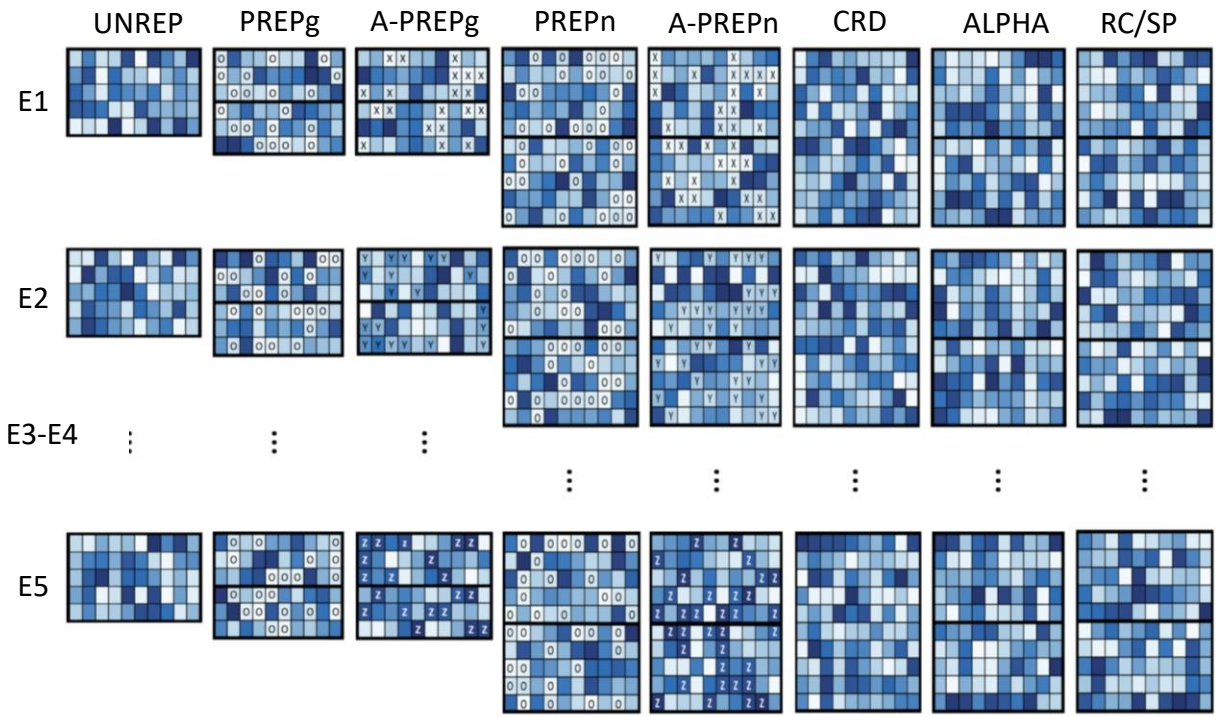


-How many genotypes?
-How many plots?

-How many replicated?
-How many environments?

Background: Experimental Design

Design



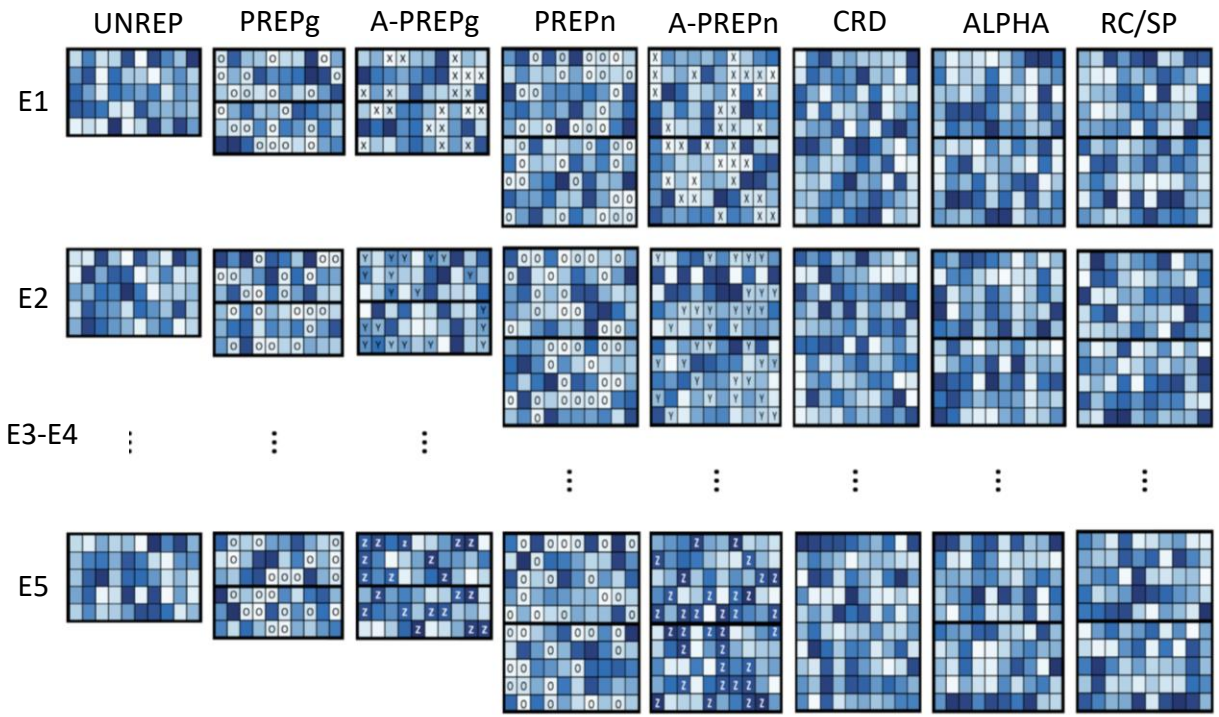
-How many genotypes?
-How many plots?
-How many replicated?
-How many environments?

Evaluate

- Labor
- Time

Background: Experimental Design

Design



-How many genotypes?
-How many plots?
-How many replicated?
-How many environments?

Evaluate

- Labor
- Time

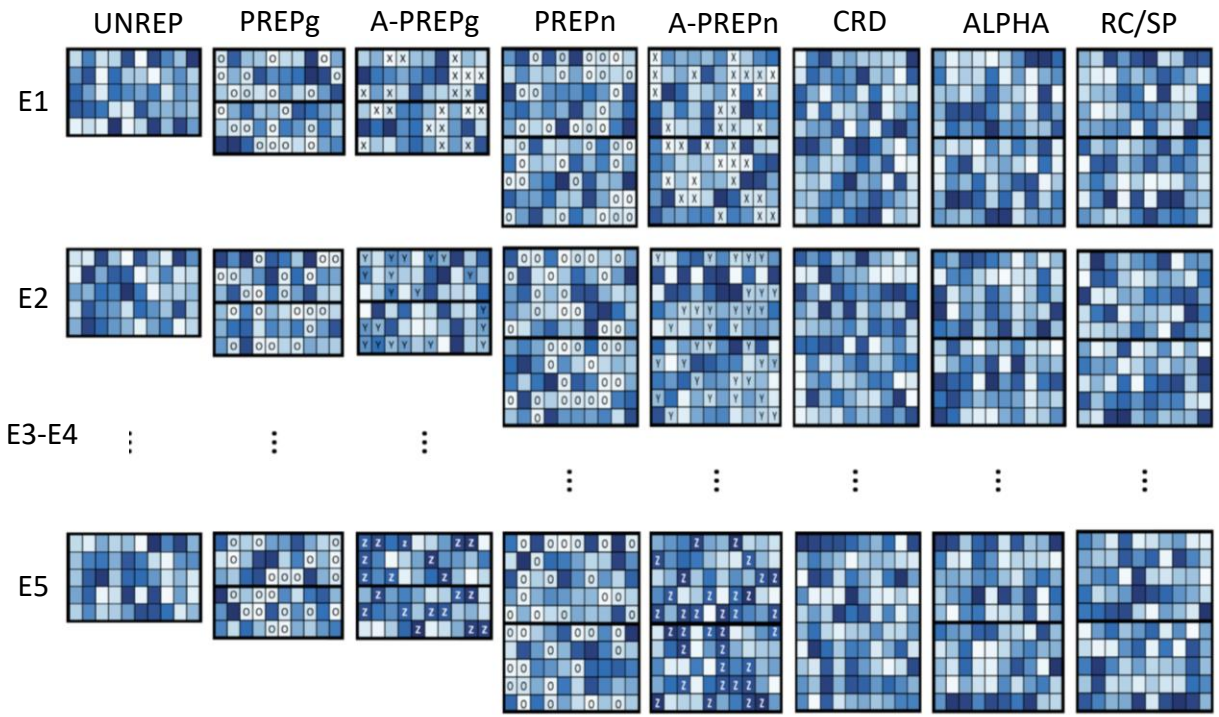
Analyze

$$y_{ijklm} = \mu + \hat{g}_i + e_j + ge_{ij} + D_{k(j)} + S_{lm(j)} + \varepsilon_{ijklm}$$

$$yield = mean + \hat{g}_i + \text{env} + \text{GE} + \text{design} + \text{spatial} + \text{error}$$

Background: Experimental Design

Design



-How many genotypes?
-How many plots?
-How many replicated?
-How many environments?

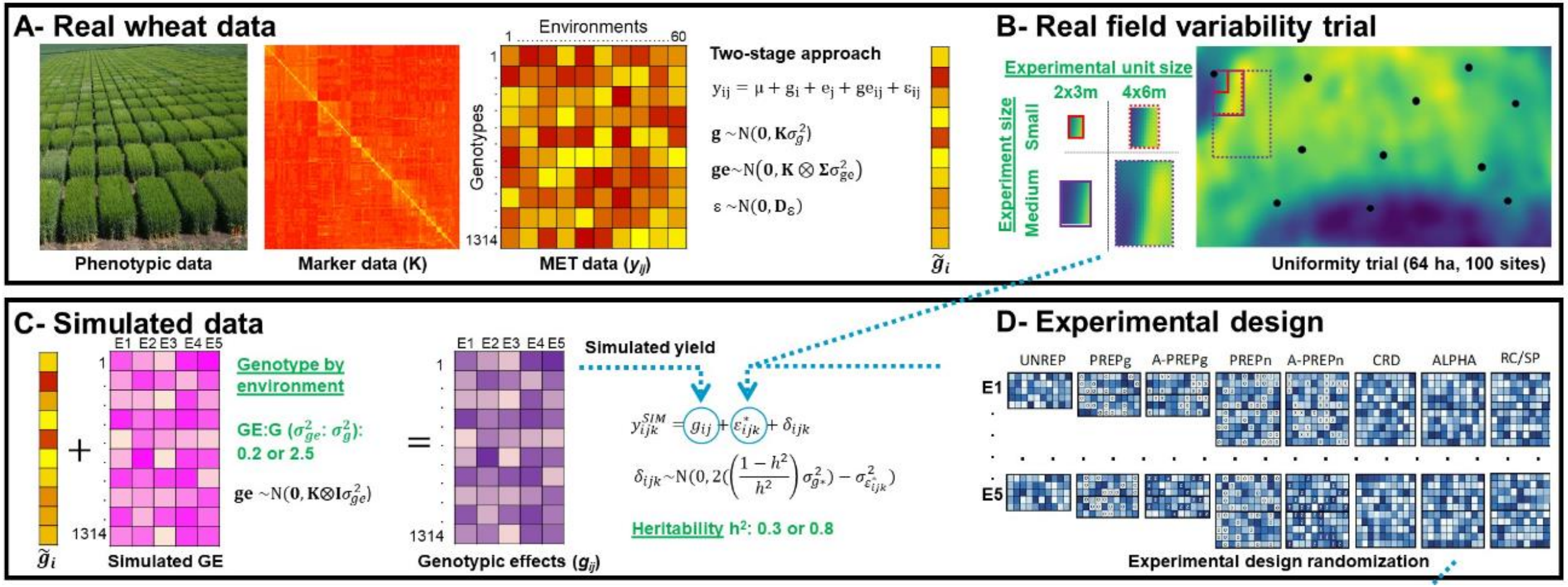
Evaluate

- Labor
- Time

Analyze

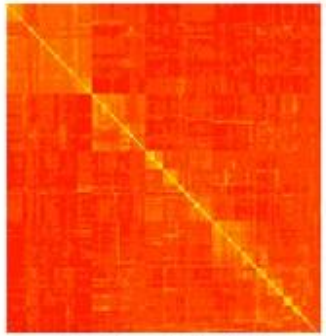
$$y_{ijklm} = \mu + \hat{g}_i + e_j + ge_{ij} + D_{k(j)} + S_{lm(j)} + \varepsilon_{ijklm}$$

$$yield = mean + \hat{g}_i + \text{env} + \text{GE} + \text{design} + \text{spatial} + \text{error}$$

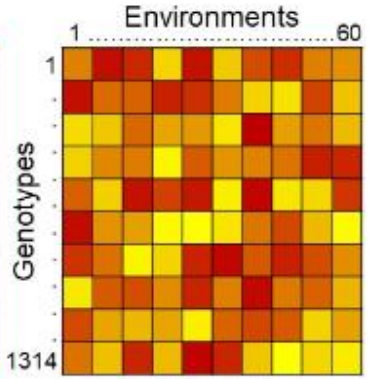

A- Real wheat data
Phenotypic data
Marker data (K)
Environments
1
60
Genotypes
1
1314
MET data (y_{ij})
Two-stage approach
$y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij}$
$g \sim N(0, K\sigma_g^2)$
$ge \sim N(0, K \otimes \Sigma\sigma_{ge}^2)$
$\varepsilon \sim N(0, D_\varepsilon)$
$\tilde{g}_i$
B- Real field variability trial
Experimental unit size
2x3m
4x6m
Experiment size
Small
Medium
Uniformity trial (64 ha, 100 sites)
C- Simulated data
E1 E2 E3 E4 E5
1
1314
$\tilde{g}_i$
Simulated GE
Genotype by environment
GE:G (σ_{ge}^2: σ_g^2): 0.2 or 2.5
$ge \sim N(0, K \otimes I\sigma_{ge}^2)$
Genotypic effects (g_{ij})
Simulated yield
$y_{ijk}^{SIM} = g_{ij} + \varepsilon_{ijk}^* + \delta_{ijk}$
$\delta_{ijk} \sim N(0, 2(\left(\frac{1-h^2}{h^2}\right)\sigma_{g*}^2) - \sigma_{\varepsilon_{ijk}^*}^2)$
Heritability h^2: 0.3 or 0.8
D- Experimental design
UNREP
PREPg
A-PREPg
PREPn
A-PREPn
CRD
ALPHA
RC/SP
E1
E5
Experimental design randomization

A- Real wheat data

Phenotypic data



Marker data (K)



MET data (y_{ij})

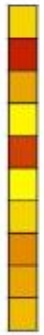
Two-stage approach

$$y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij}$$

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{K}\sigma_g^2)$$

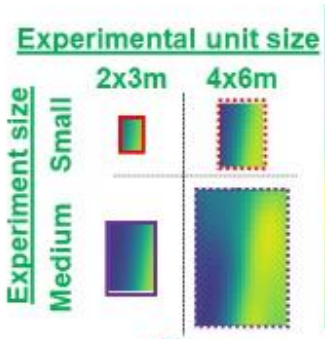
$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{K} \otimes \mathbf{\Sigma}\sigma_{ge}^2)$$

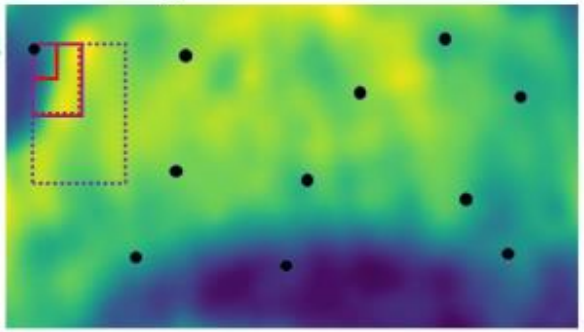
$$\varepsilon \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$$



$\tilde{g}_i$

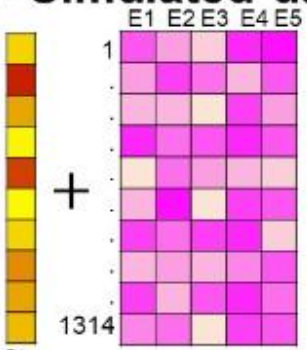
B- Real field variability trial





Uniformity trial (64 ha, 100 sites)

C- Simulated data

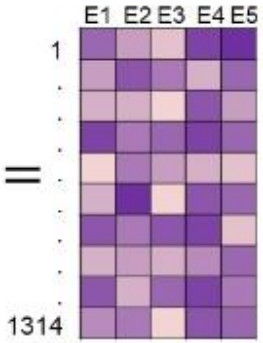


$\tilde{g}_i$ Simulated GE

Genotype by environment

GE:G (σ_{ge}^2: σ_g^2): 0.2 or 2.5

$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{K} \otimes \mathbf{I}\sigma_{ge}^2)$$



Genotypic effects (g_{ij})

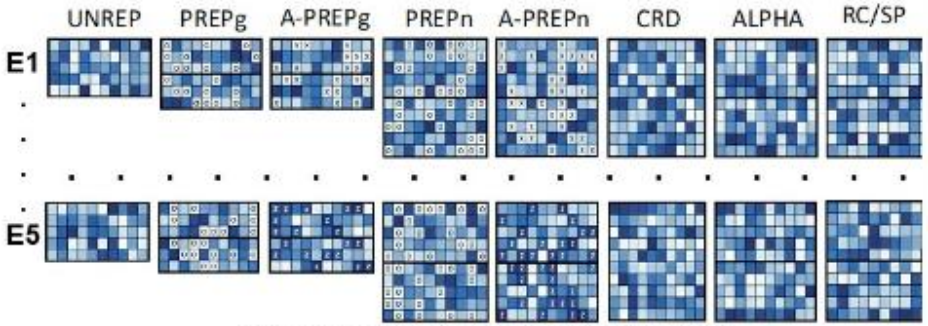
Simulated yield

$$y_{ijk}^{SIM} = g_{ij} + \varepsilon_{ijk}^{*} + \delta_{ijk}$$

$$\delta_{ijk} \sim N\left(0, 2\left(\left(\frac{1-h^2}{h^2}\right)\sigma_{g*}^2\right) - \sigma_{\varepsilon_{ijk}^{*}}^2\right)$$

Heritability h^2: 0.3 or 0.8

D- Experimental design



Experimental design randomization

E- Analysis model

$$y_{ijklm} = \mu + g_i + e_j + ge_{ij} + d_{k(j)} + s_{lm(j)} + \varepsilon_{ijklm}$$

$\tilde{g}_{ij}$

$$\varepsilon \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$$

Relationship among genotypes

G: I or K

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$$

$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{I}\sigma_{ge}^2)$$

Spatial corrections

NSC $s_{lm(j)} = \emptyset$

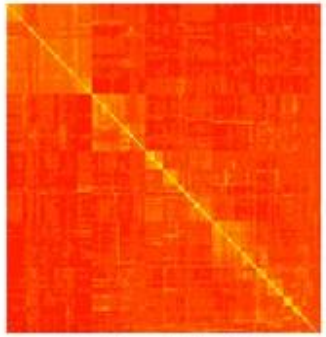
AR1xAR1 $s_{lm(j)} = r_{l(j)} + c_{m(j)}$

$$\mathbf{r} \sim N(\mathbf{0}, \mathbf{D}_R \otimes \mathbf{AR1})$$

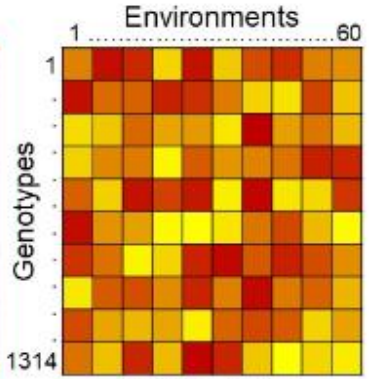
$$\mathbf{c} \sim N(\mathbf{0}, \mathbf{D}_C \otimes \mathbf{AR1})$$

A- Real wheat data

Phenotypic data



Marker data (K)



MET data (y_{ij})

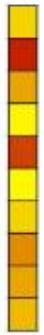
Two-stage approach

$$y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij}$$

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{K}\sigma_g^2)$$

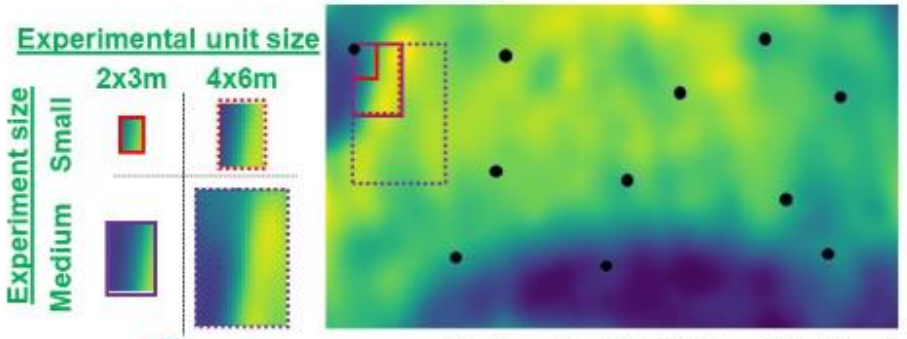
$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{K} \otimes \mathbf{\Sigma}\sigma_{ge}^2)$$

$$\varepsilon \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$$



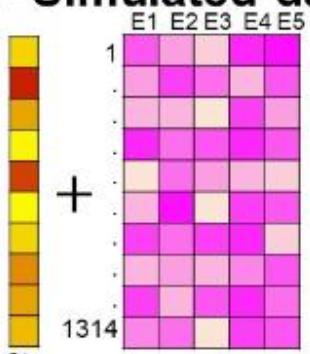
$\tilde{g}_i$

B- Real field variability trial



Uniformity trial (64 ha, 100 sites)

C- Simulated data

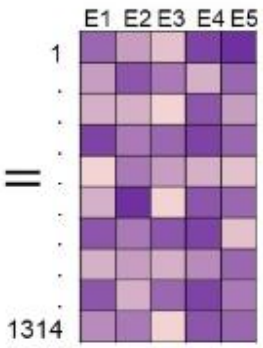


$\tilde{g}_i$ Simulated GE

Genotype by environment

GE:G (σ_{ge}^2: σ_g^2): 0.2 or 2.5

$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{K} \otimes \mathbf{I}\sigma_{ge}^2)$$



Genotypic effects (g_{ij})

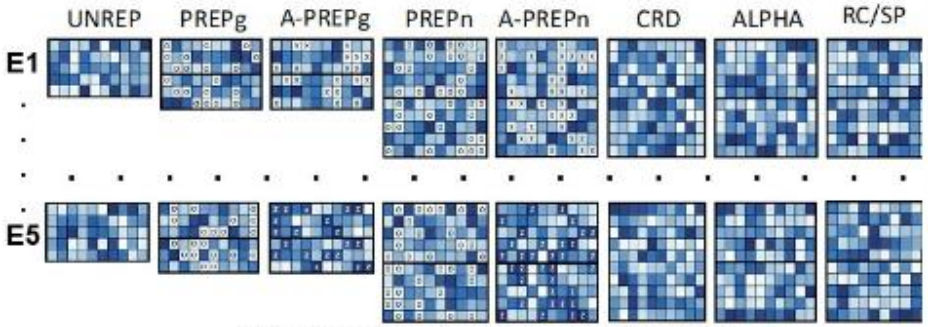
Simulated yield

$$y_{ijk}^{SIM} = g_{ij} + \varepsilon_{ijk}^{*} + \delta_{ijk}$$

$$\delta_{ijk} \sim N\left(0, 2\left(\left(\frac{1-h^2}{h^2}\right)\sigma_{g^*}^2\right) - \sigma_{\varepsilon_{ijk}^{*}}^2\right)$$

Heritability h^2: 0.3 or 0.8

D- Experimental design



Experimental design randomization

E- Analysis model

$$y_{ijklm} = \mu + g_i + e_j + ge_{ij} + d_{k(j)} + s_{lm(j)} + \varepsilon_{ijklm}$$

$\tilde{g}_{ij}$

$$\varepsilon \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$$

Relationship among genotypes

G: I or K

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$$

$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{I}\sigma_{ge}^2)$$

Spatial corrections

NSC $s_{lm(j)} = \emptyset$

AR1xAR1 $s_{lm(j)} = r_{l(j)} + c_{m(j)}$

$$\mathbf{r} \sim N(\mathbf{0}, \mathbf{D}_R \otimes \mathbf{AR1})$$

$$\mathbf{c} \sim N(\mathbf{0}, \mathbf{D}_C \otimes \mathbf{AR1})$$

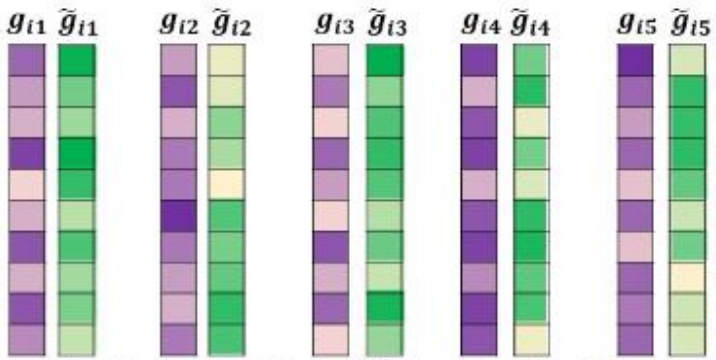
F- Performance evaluation

Accuracy

$$\mathrm{COR} = \mathrm{cor}(g_{ij}, \tilde{g}_{ij}) = r_{(g\tilde{g})}$$

Response to selection

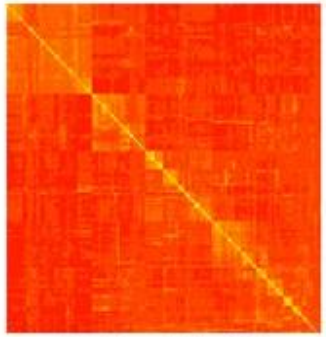
$$\mathrm{R} = i r_{(g\tilde{g})} \sigma_g$$



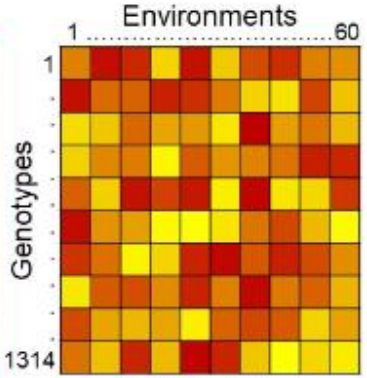
True (g_{ij}) vs. predicted ($\tilde{g}_{ij}$) genotypic effects

A- Real wheat data

Phenotypic data



Marker data (K)



MET data (y_{ij})

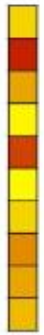
Two-stage approach

$$y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij}$$

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{K}\sigma_g^2)$$

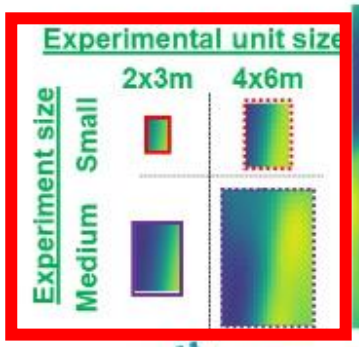
$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{K} \otimes \mathbf{\Sigma}\sigma_{ge}^2)$$

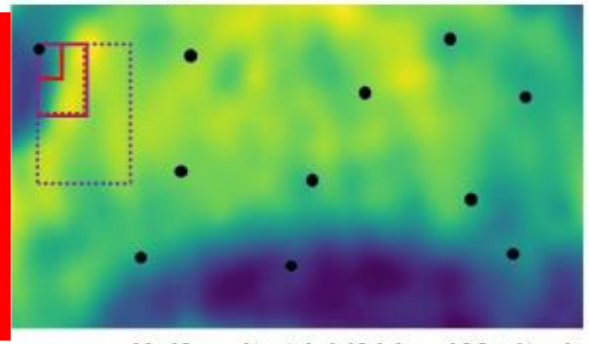
$$\varepsilon \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$$



$\tilde{g}_i$

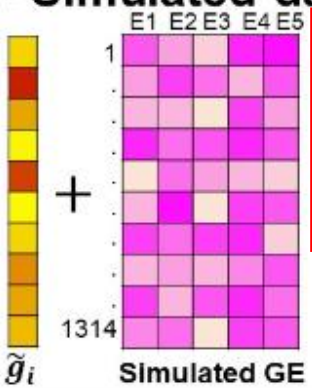
B- Real field variability trial





Uniformity trial (64 ha, 100 sites)

C- Simulated data

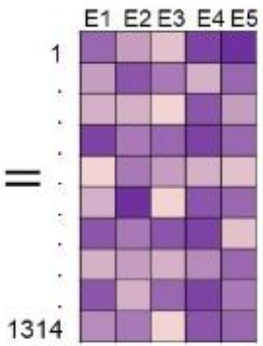


$\tilde{g}_i$ Simulated GE

Genotype by environment

GE:G (σ_{ge}^2: σ_g^2): 0.2 or 2.5

$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{K} \otimes \mathbf{I}\sigma_{ge}^2)$$



Genotypic effects (g_{ij})

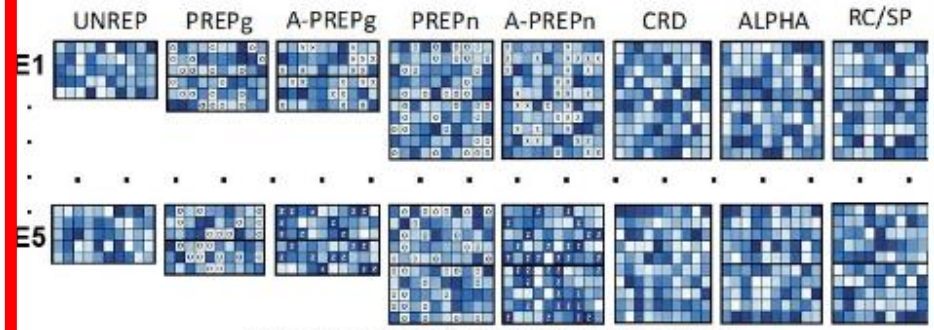
Simulated yield

$$y_{ijk}^{SIM} = g_{ij} + \varepsilon_{ijk}^{*} + \delta_{ijk}$$

$$\delta_{ijk} \sim N\left(0, 2\left(\left(\frac{1-h^2}{h^2}\right)\sigma_{g^*}^2\right) - \sigma_{\varepsilon_{ijk}^{*}}^2\right)$$

Heritability h^2: 0.3 or 0.8

D- Experimental design



Experimental design randomization

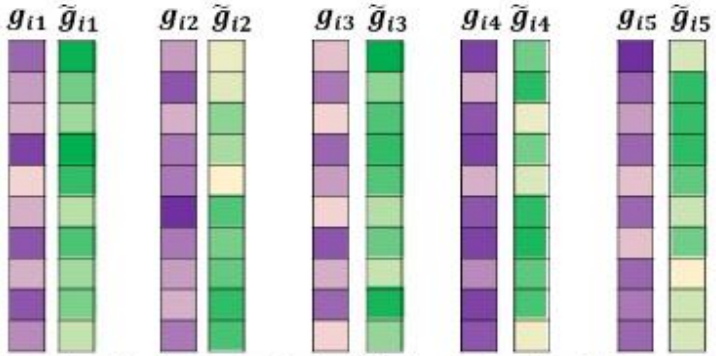
F- Performance evaluation

Accuracy

$$\mathrm{COR} = \mathrm{cor}(g_{ij}, \tilde{g}_{ij}) = r_{(g\tilde{g})}$$

Response to selection

$$\mathrm{R} = i r_{(g\tilde{g})} \sigma_g$$



True (g_{ij}) vs. predicted ($\tilde{g}_{ij}$) genotypic effects

E- Analysis model

$$y_{ijklm} = \mu + g_i + e_j + ge_{ij} + d_{k(j)} + s_{lm(j)} + \varepsilon_{ijklm}$$

$\tilde{g}_{ij}$

$$\varepsilon \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$$

Relationship among genotypes

G: I or K

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$$

$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{I}\sigma_{ge}^2)$$

Spatial corrections

NSC $s_{lm(j)} = \emptyset$

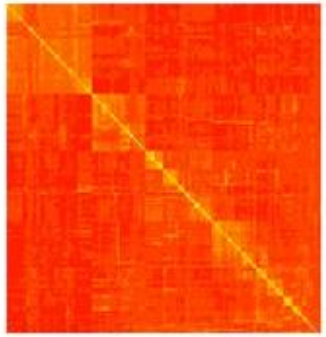
AR1xAR1 $s_{lm(j)} = r_{l(j)} + c_{m(j)}$

$$\mathbf{r} \sim N(\mathbf{0}, \mathbf{D}_R \otimes \mathbf{AR1})$$

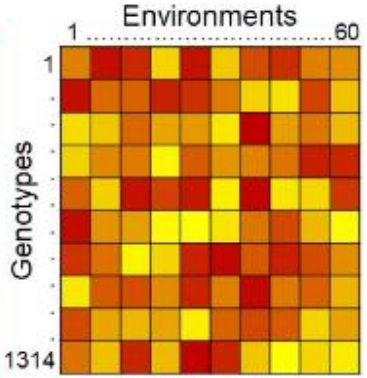
$$\mathbf{c} \sim N(\mathbf{0}, \mathbf{D}_C \otimes \mathbf{AR1})$$

A- Real wheat data

Phenotypic data



Marker data (K)



MET data (y_{ij})

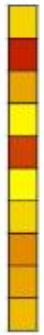
Two-stage approach

$$y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij}$$

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{K}\sigma_g^2)$$

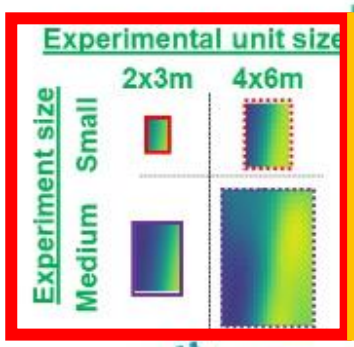
$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{K} \otimes \mathbf{\Sigma}\sigma_{ge}^2)$$

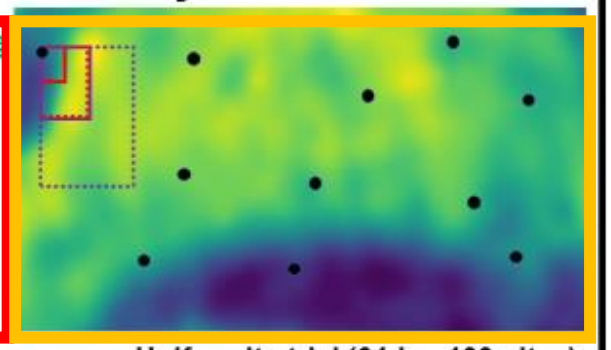
$$\varepsilon \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$$



$\tilde{g}_i$

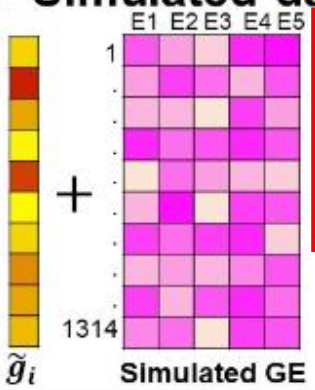
B- Real field variability trial





Uniformity trial (64 ha, 100 sites)

C- Simulated data

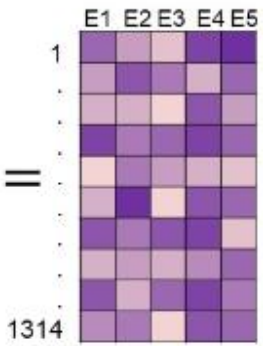


$\tilde{g}_i$ Simulated GE

Genotype by environment

GE:G (σ_{ge}^2: σ_g^2): 0.2 or 2.5

$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{K} \otimes \mathbf{I}\sigma_{ge}^2)$$



Genotypic effects (g_{ij})

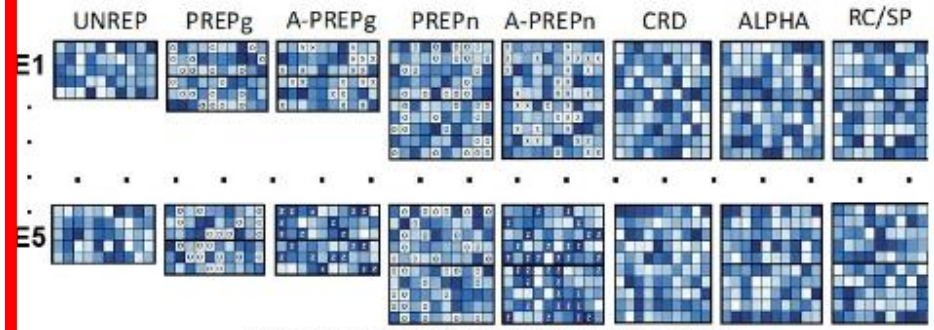
Simulated yield

$$y_{ijk}^{SIM} = g_{ij} + \varepsilon_{ijk}^{*} + \delta_{ijk}$$

$$\delta_{ijk} \sim N\left(0, 2\left(\left(\frac{1-h^2}{h^2}\right)\sigma_{g^*}^2\right) - \sigma_{\varepsilon_{ijk}^{*}}^2\right)$$

Heritability h^2: 0.3 or 0.8

D- Experimental design



Experimental design randomization

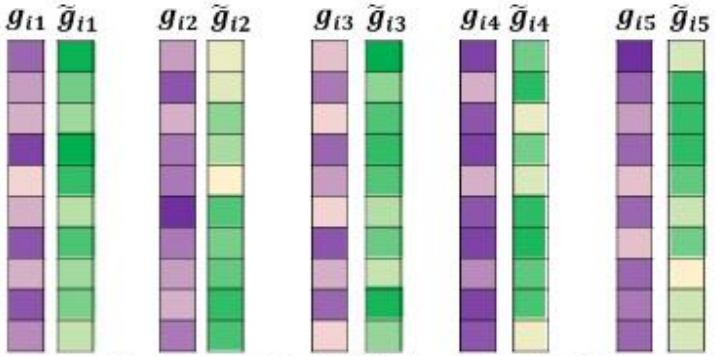
F- Performance evaluation

Accuracy

$$\mathrm{COR} = \mathrm{cor}(g_{ij}, \tilde{g}_{ij}) = r_{(g\tilde{g})}$$

Response to selection

$$\mathrm{R} = i r_{(g\tilde{g})} \sigma_g$$



True (g_{ij}) vs. predicted ($\tilde{g}_{ij}$) genotypic effects

E- Analysis model

$$y_{ijklm} = \mu + g_i + e_j + ge_{ij} + d_{k(j)} + s_{lm(j)} + \varepsilon_{ijklm}$$

$\tilde{g}_{ij}$

$$\varepsilon \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$$

Relationship among genotypes

G: I or K

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$$

$$\mathbf{ge} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{I}\sigma_{ge}^2)$$

Spatial corrections

NSC $s_{lm(j)} = \emptyset$

AR1xAR1 $s_{lm(j)} = r_{l(j)} + c_{m(j)}$

$$\mathbf{r} \sim N(\mathbf{0}, \mathbf{D}_R \otimes \mathbf{AR1})$$

$$\mathbf{c} \sim N(\mathbf{0}, \mathbf{D}_C \otimes \mathbf{AR1})$$

July 2019
Invitation & first project meeting.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

July 2019
Invitation & first project meeting.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

- Originally planned 3,168 scenarios x 100 sites x 1,000 iterations.

  = simulation of 317 M experiments

July 2019
Invitation & first project meeting.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

- Originally planned 3,168 scenarios x 100 sites x 1,000 iterations.

  = simulation of 317 M experiments

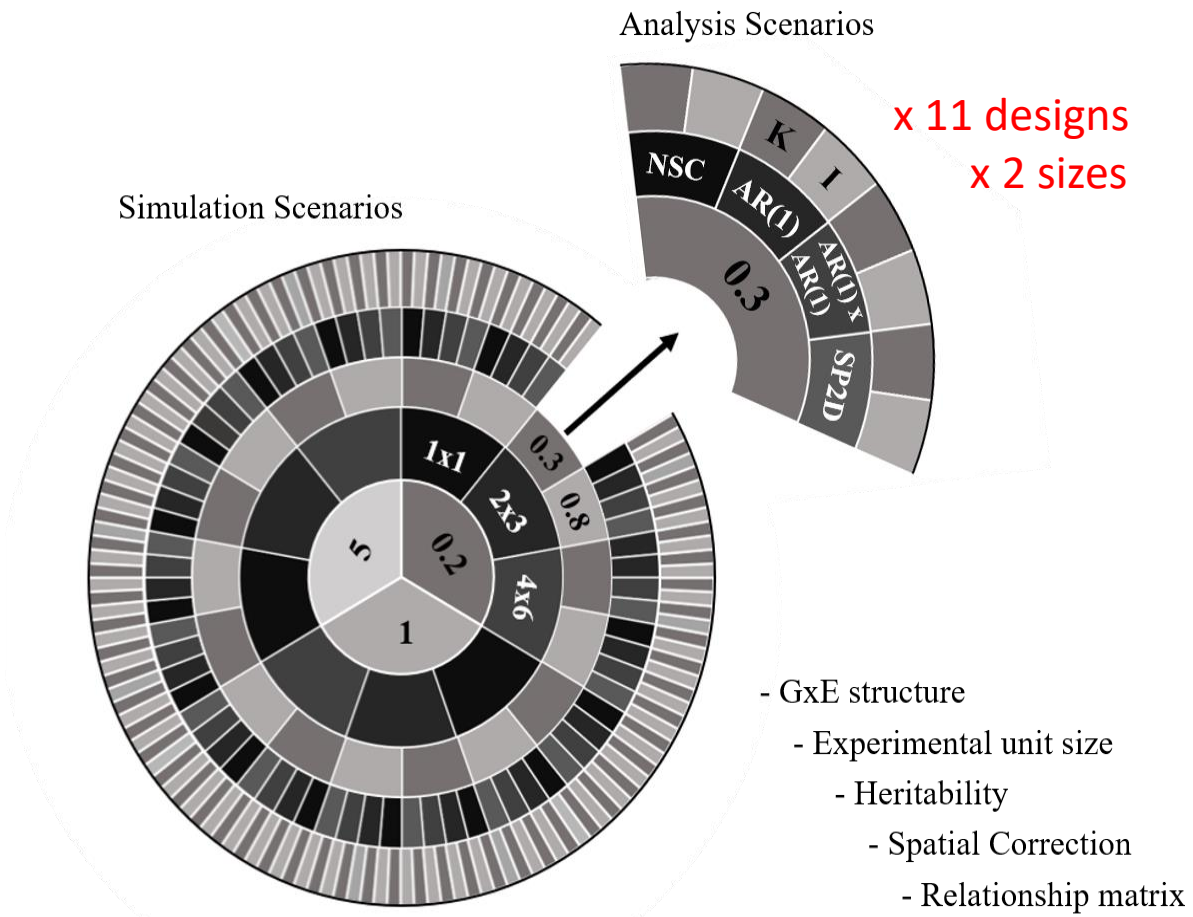
July 2019
Invitation & first project meeting.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

- Originally planned 3,168 scenarios x 100 sites x 1,000 iterations.

  = simulation of 317 M experiments



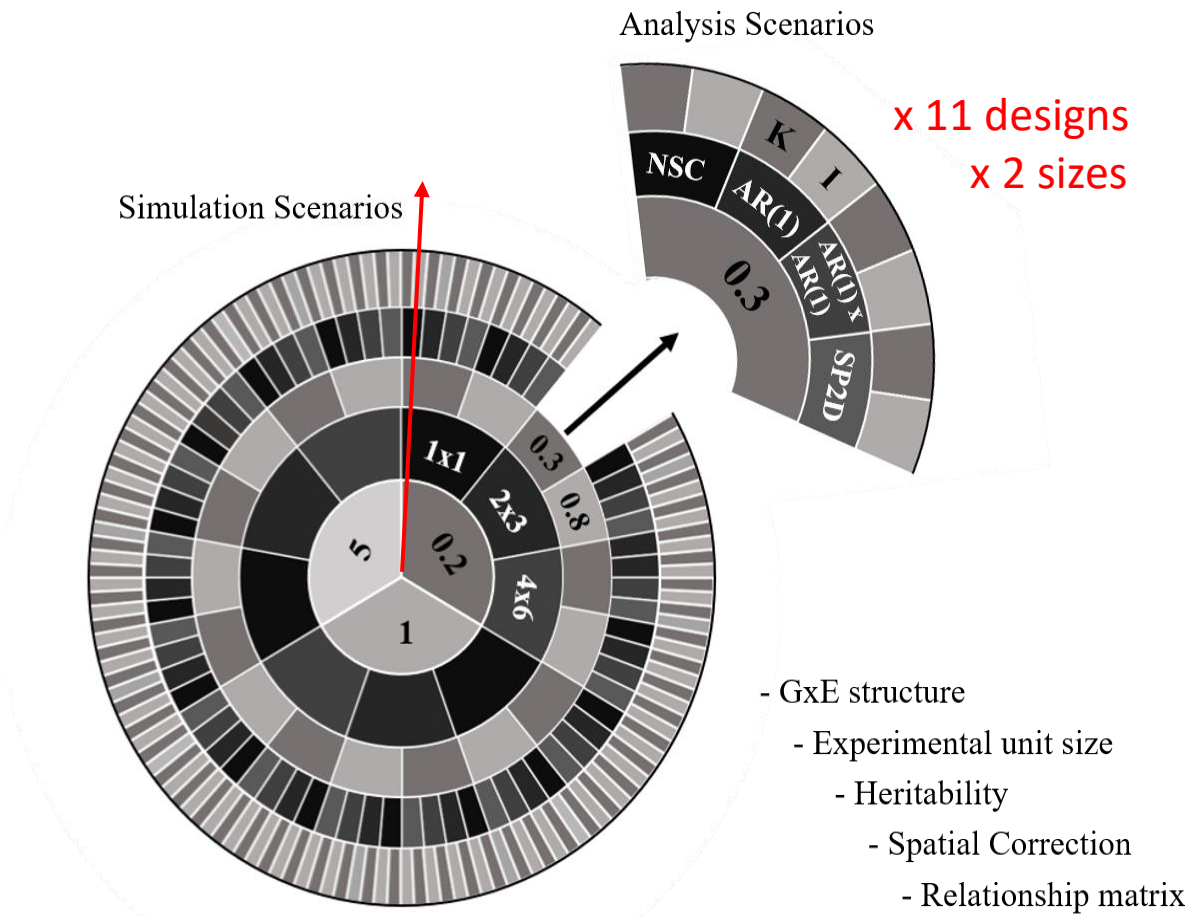
July 2019
Invitation & first project meeting.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

- Originally planned 3,168 scenarios x 100 sites x 1,000 iterations.

  = simulation of 317 M experiments



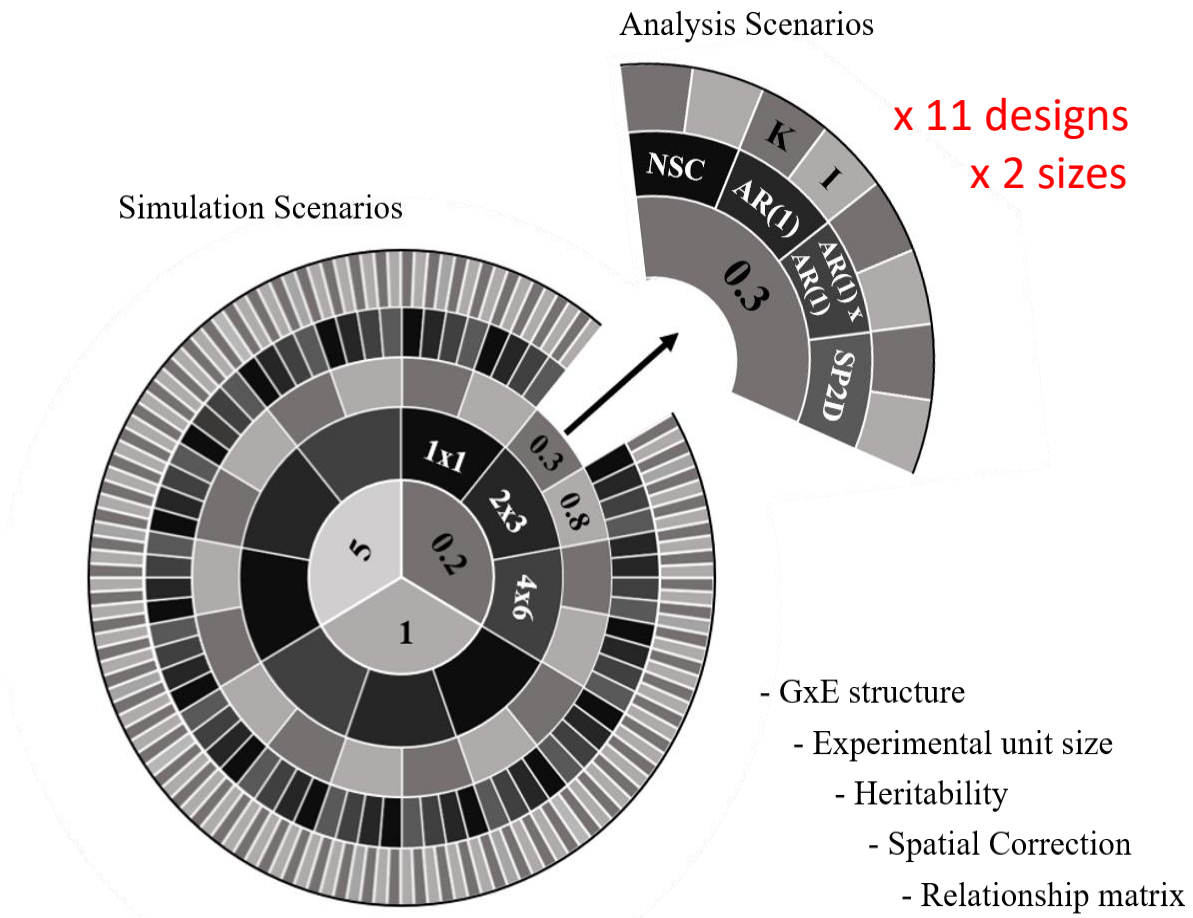
July 2019
Invitation & first project meeting.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

- Originally planned 3,168 scenarios x 100 sites x 1,000 iterations.

  = simulation of 317 M experiments

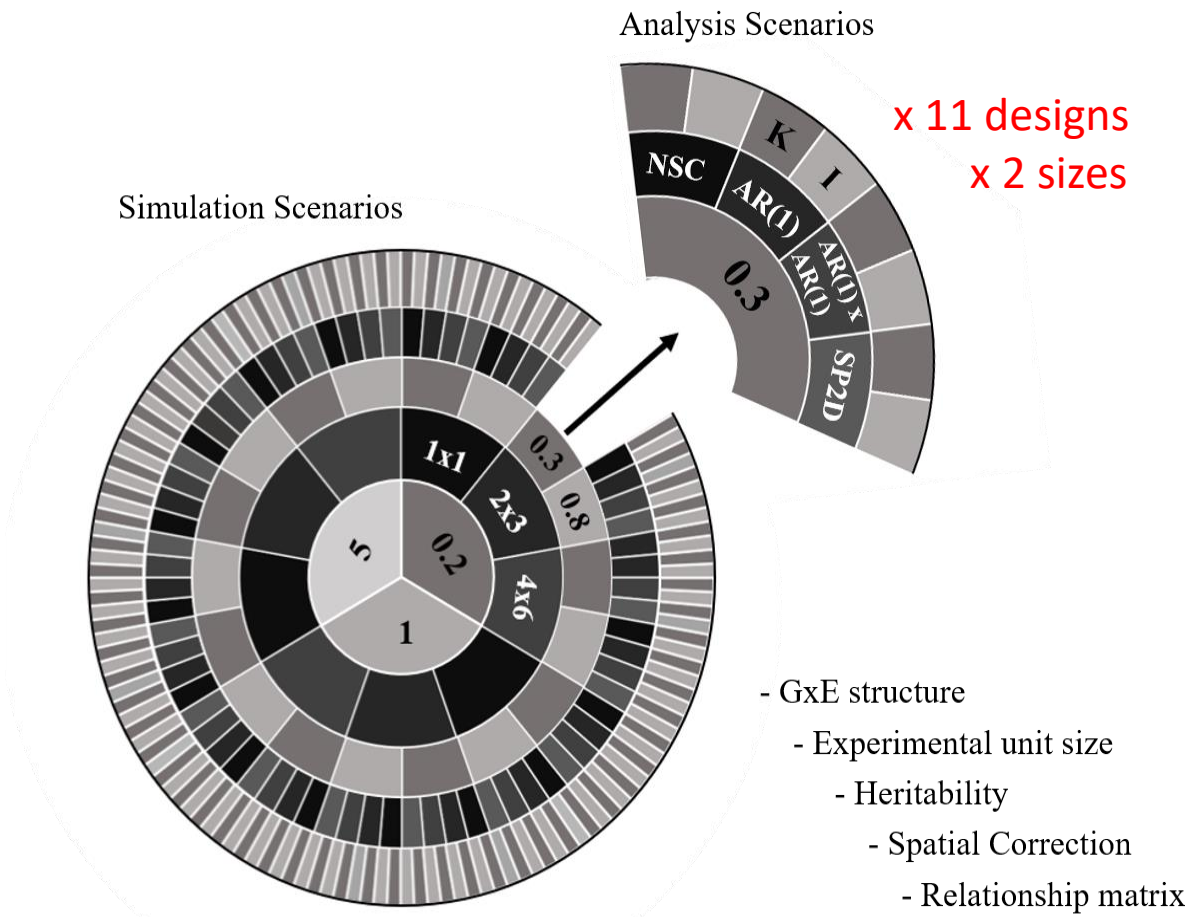


July 2019
Invitation & first project meeting.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

- Originally planned 3,168 scenarios x 100 sites x 1,000 iterations.
  = simulation of 317 M experiments
- Knew right away that we needed high throughput computing.

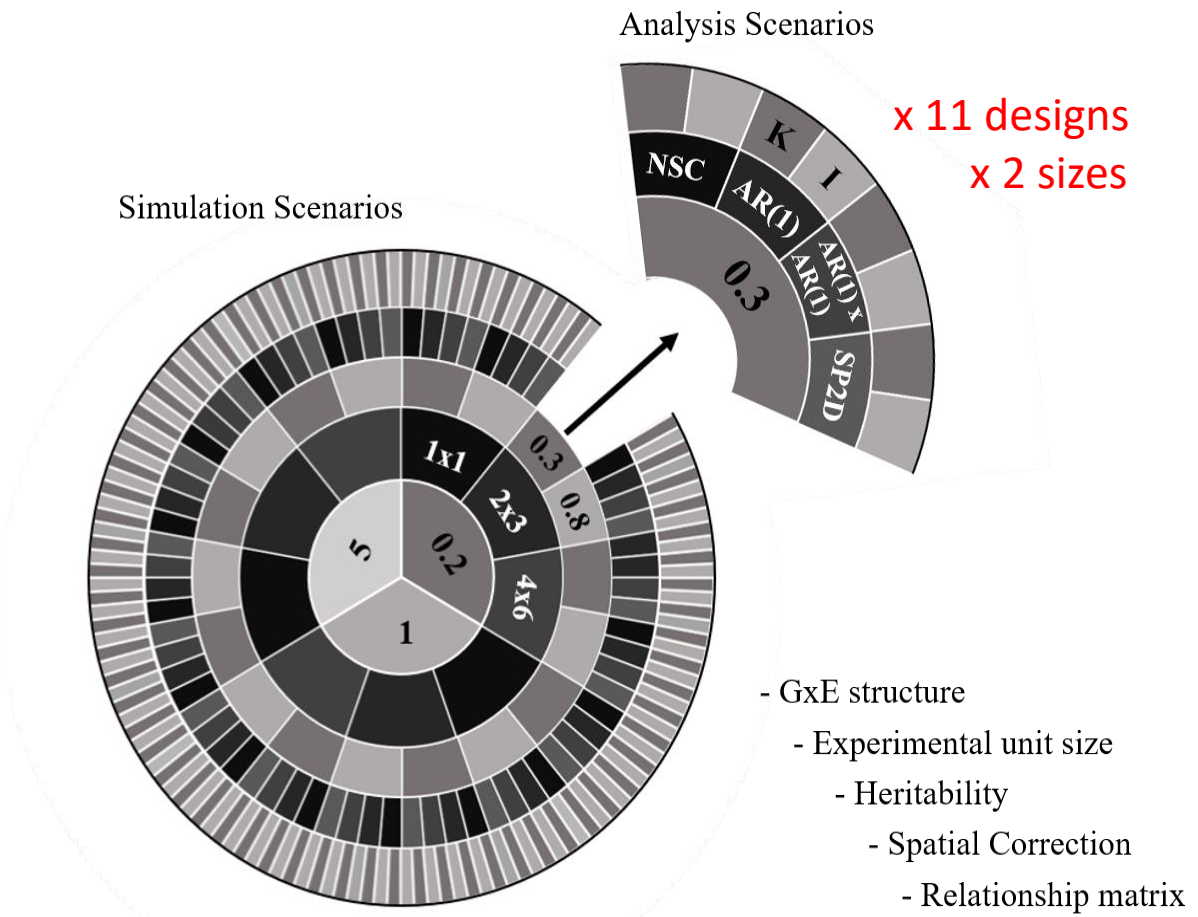


July 2019
Invitation & first project meeting.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

- Originally planned 3,168 scenarios x 100 sites x 1,000 iterations.
  = simulation of 317 M experiments
- Knew right away that we needed high throughput computing.
- Challenge #1: software portability with a licensed package.



July 2019
Invitation & first project meeting.

August
Code building.

September
Start running code on CHTC & OSG.

October
First submission deadline.

1 2 3 4 5 6 7 8 9 10 11 12

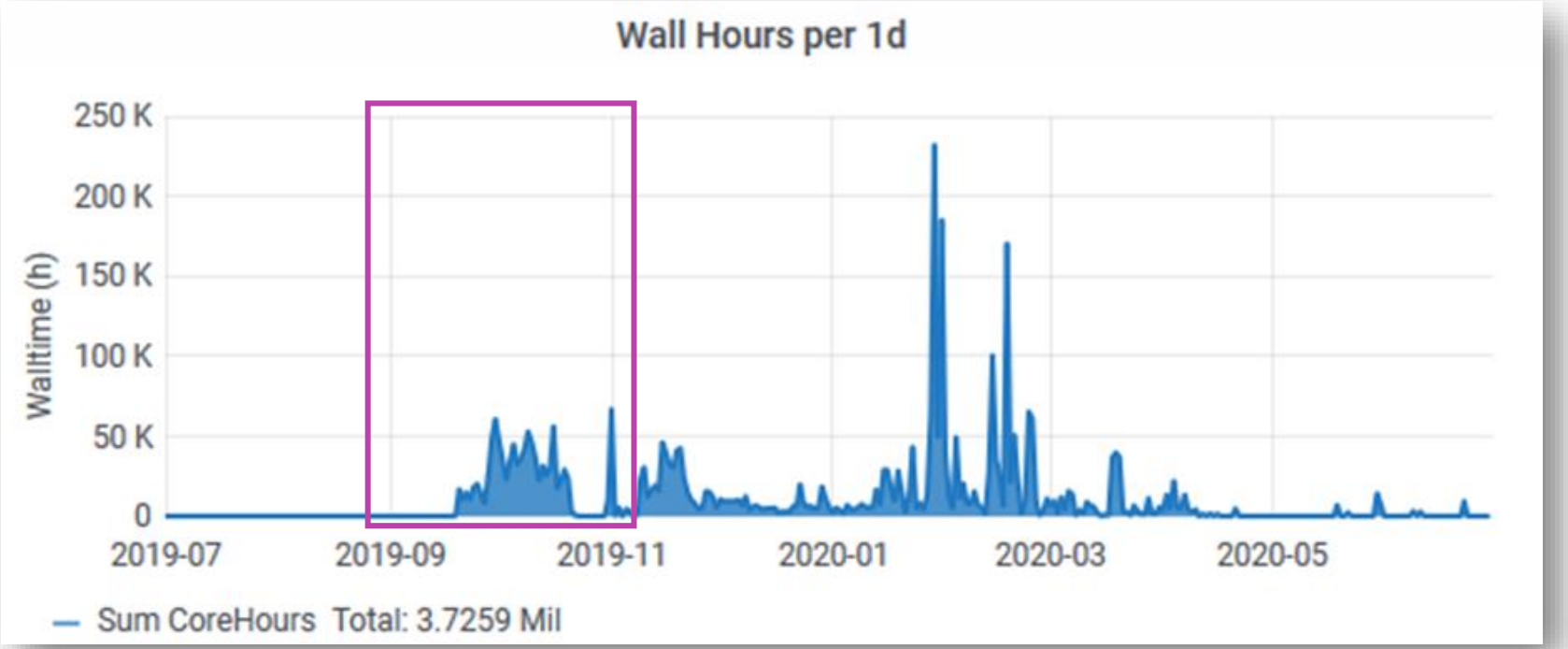
July 2019
Invitation & first project meeting.

September
Start running code on CHTC & OSG.

1 2 3 4 5 6 7 8 9 10 11 12

August
Code building.

October
First submission deadline.



- Challenge #2: Delegation of work and compilation of results.

- Challenge #2: Delegation of work and compilation of results.
  - 10 people in four countries.
  - 7 accounts (between OSG and CHTC) running jobs.

- Challenge #2: Delegation of work and compilation of results.
  - 10 people in four countries.
  - 7 accounts (between OSG and CHTC) running jobs.
  - Changing code -> submitting jobs -> running jobs -> downloading results -> compiling results -> checking results for errors = time consuming and error prone

- Challenge #2: Delegation of work and compilation of results.
  - 10 people in four countries.
  - 7 accounts (between OSG and CHTC) running jobs.
  - Changing code -> submitting jobs -> running jobs -> downloading results -> compiling results -> checking results for errors = time consuming and error prone
- Challenge #3: Time.

- Challenge #2: Delegation of work and compilation of results.
  - 10 people in four countries.
  - 7 accounts (between OSG and CHTC) running jobs.
  - Changing code -> submitting jobs -> running jobs -> downloading results -> compiling results -> checking results for errors = time consuming and error prone
- Challenge #3: Time.
  - Medium jobs used 32 GB memory – less nodes available.

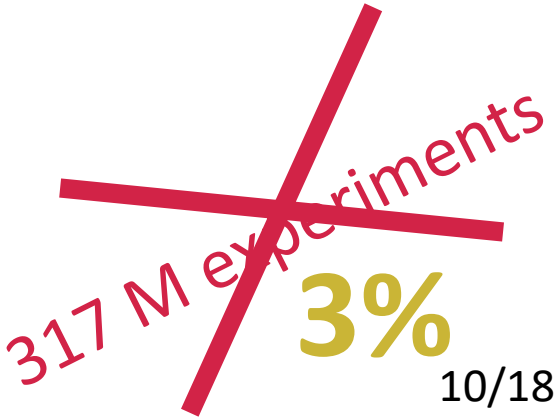
- Challenge #2: Delegation of work and compilation of results.
  - 10 people in four countries.
  - 7 accounts (between OSG and CHTC) running jobs.
  - Changing code -> submitting jobs -> running jobs -> downloading results -> compiling results -> checking results for errors = time consuming and error prone
- Challenge #3: Time.
  - Medium jobs used 32 GB memory – less nodes available.
  - Access to OSG login (differences in condor_q, condor_history).

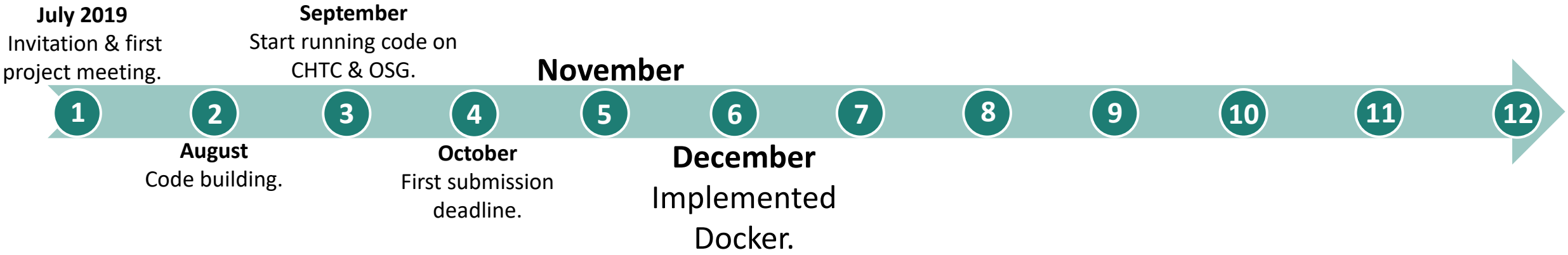
- Challenge #2: Delegation of work and compilation of results.
  - 10 people in four countries.
  - 7 accounts (between OSG and CHTC) running jobs.
  - Changing code -> submitting jobs -> running jobs -> downloading results -> compiling results -> checking results for errors = time consuming and error prone
- Challenge #3: Time.
  - Medium jobs used 32 GB memory – less nodes available.
  - Access to OSG login (differences in condor_q, condor_history).
  - Asked for priority.

- Challenge #2: Delegation of work and compilation of results.
  - 10 people in four countries.
  - 7 accounts (between OSG and CHTC) running jobs.
  - Changing code -> submitting jobs -> running jobs -> downloading results -> compiling results -> checking results for errors = time consuming and error prone
- Challenge #3: Time.
  - Medium jobs used 32 GB memory – less nodes available.
  - Access to OSG login (differences in condor_q, condor_history).
  - Asked for priority.

~~317 M experiments~~

- Challenge #2: Delegation of work and compilation of results.
  - 10 people in four countries.
  - 7 accounts (between OSG and CHTC) running jobs.
  - Changing code -> submitting jobs -> running jobs -> downloading results -> compiling results -> checking results for errors = time consuming and error prone
- Challenge #3: Time.
  - Medium jobs used 32 GB memory – less nodes available.
  - Access to OSG login (differences in condor_q, condor_history).
  - Asked for priority.




July 2019
Invitation & first project meeting.
September
Start running code on CHTC & OSG.
November
1
2
3
4
5
6
7
8
9
10
11
12
August
Code building.
October
First submission deadline.
December
Implemented Docker.

July 2019
Invitation & first project meeting.

August
Code building.

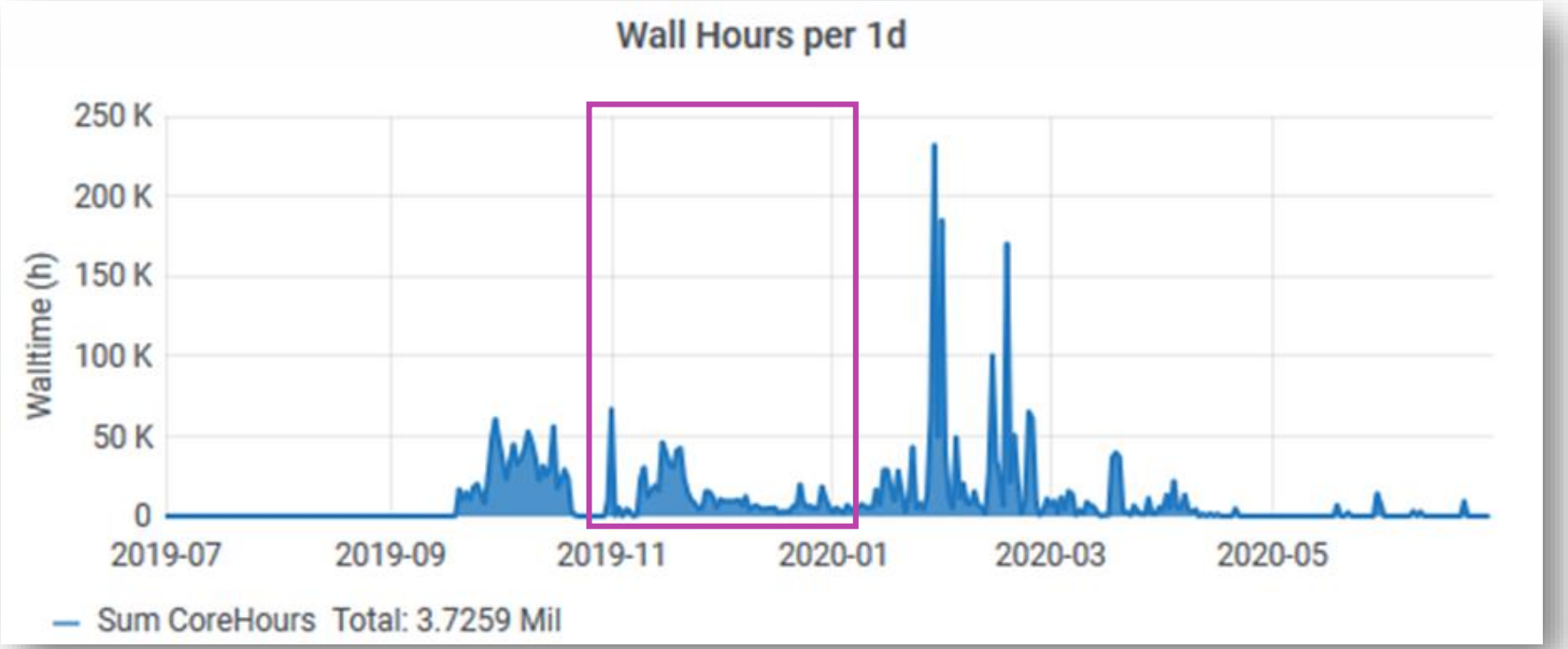
September
Start running code on CHTC & OSG.

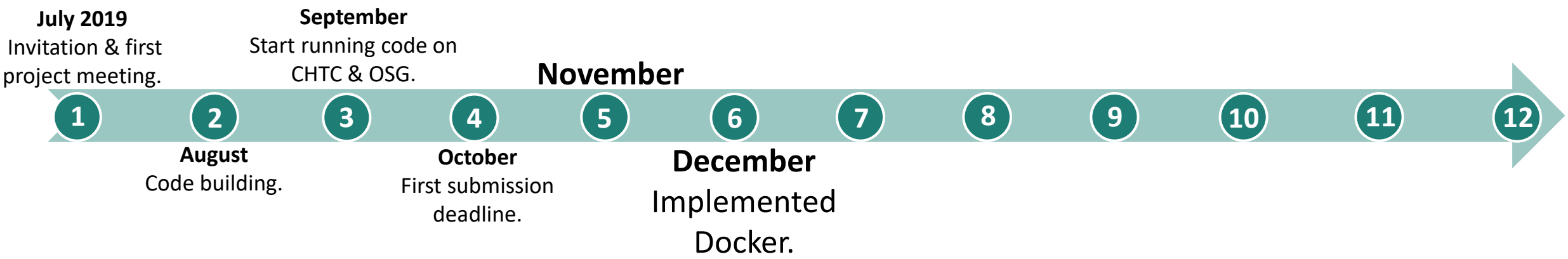
October
First submission deadline.

November

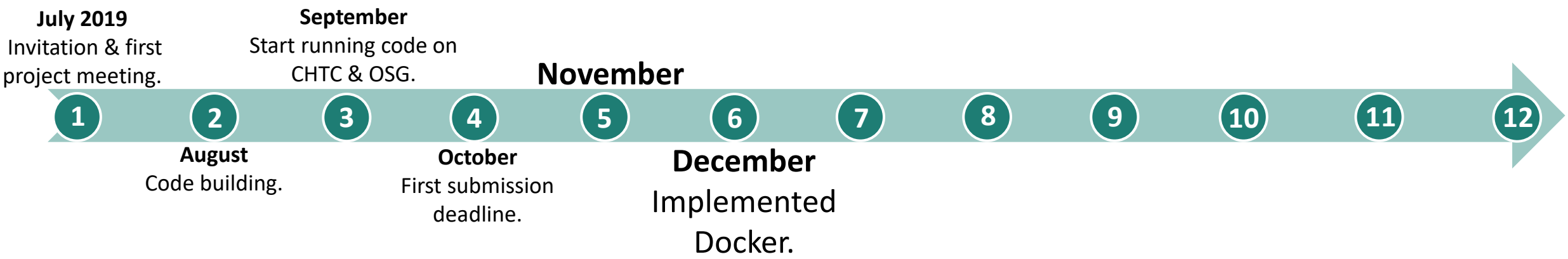
December
Implemented Docker.

1 2 3 4 5 6 7 8 9 10 11 12

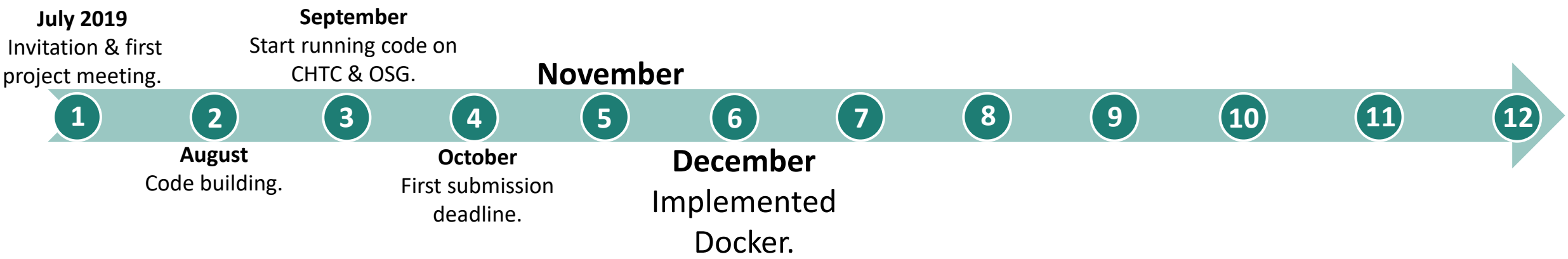




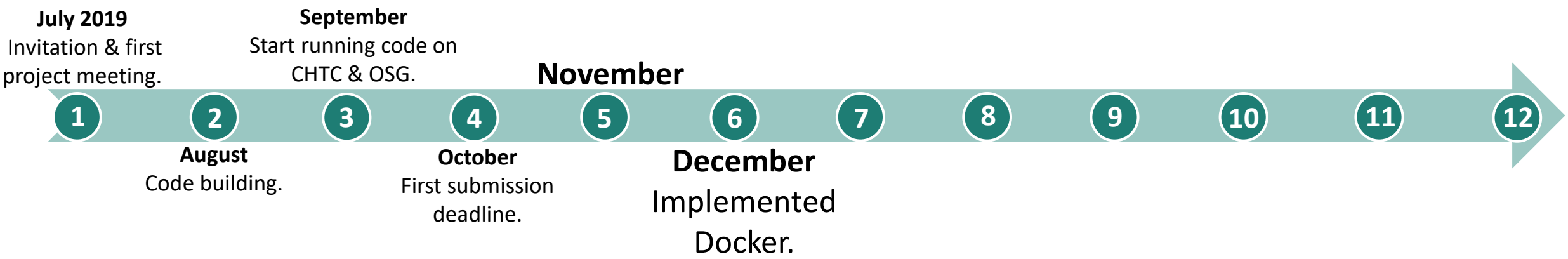
- Took a little break.



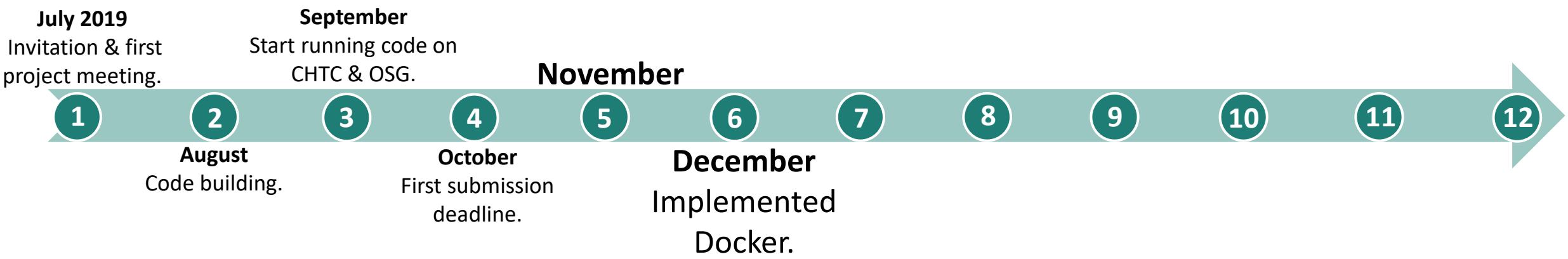
- Took a little break.
- Knew we had a lot of work yet to come (reviews, finishing the other 97% of experiments).



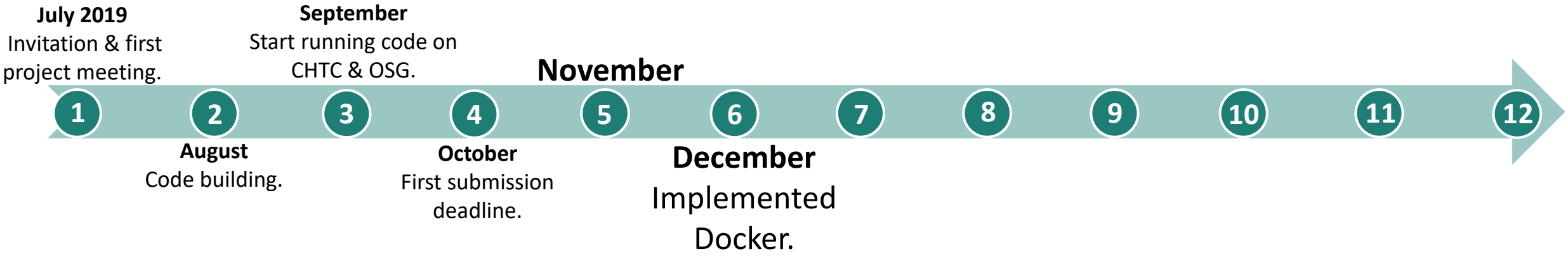
- Took a little break.
- Knew we had a lot of work yet to come (reviews, finishing the other 97% of experiments).
- Challenge #4: Looked into ways to run things quicker and more efficiently.



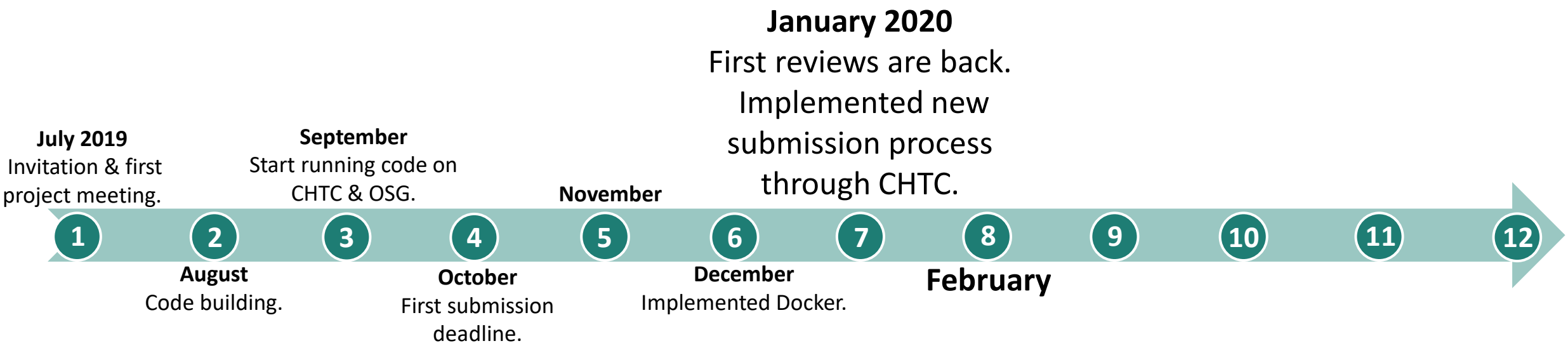
- Took a little break.
- Knew we had a lot of work yet to come (reviews, finishing the other 97% of experiments).
- Challenge #4: Looked into ways to run things quicker and more efficiently.
  - Implemented Docker and OpenBlas.
    - Different input files required for OSG and CHTC logins.



- Took a little break.
- Knew we had a lot of work yet to come (reviews, finishing the other 97% of experiments).
- Challenge #4: Looked into ways to run things quicker and more efficiently.
  - Implemented Docker and OpenBlas.
    - Different input files required for OSG and CHTC logins.
  - Daily check-ins and use of a google form.



- Took a little break.
- Knew we had a lot of work yet to come (reviews, finishing the other 97% of experiments).
- Challenge #4: Looked into ways to run things quicker and more efficiently.
  - Implemented Docker and OpenBlas.
    - Different input files required for OSG and CHTC logins.
  - Daily check-ins and use of a google form.


January 2020
First reviews are back. Implemented new submission process through CHTC.
July 2019
Invitation & first project meeting.
September
Start running code on CHTC & OSG.
November
1
2
3
4
5
6
7
8
9
10
11
12
August
Code building.
October
First submission deadline.
December
Implemented Docker.
February

January 2020
First reviews are back.
Implemented new
submission process
through CHTC.

July 2019
Invitation & first
project meeting.

September
Start running code on
CHTC & OSG.

November

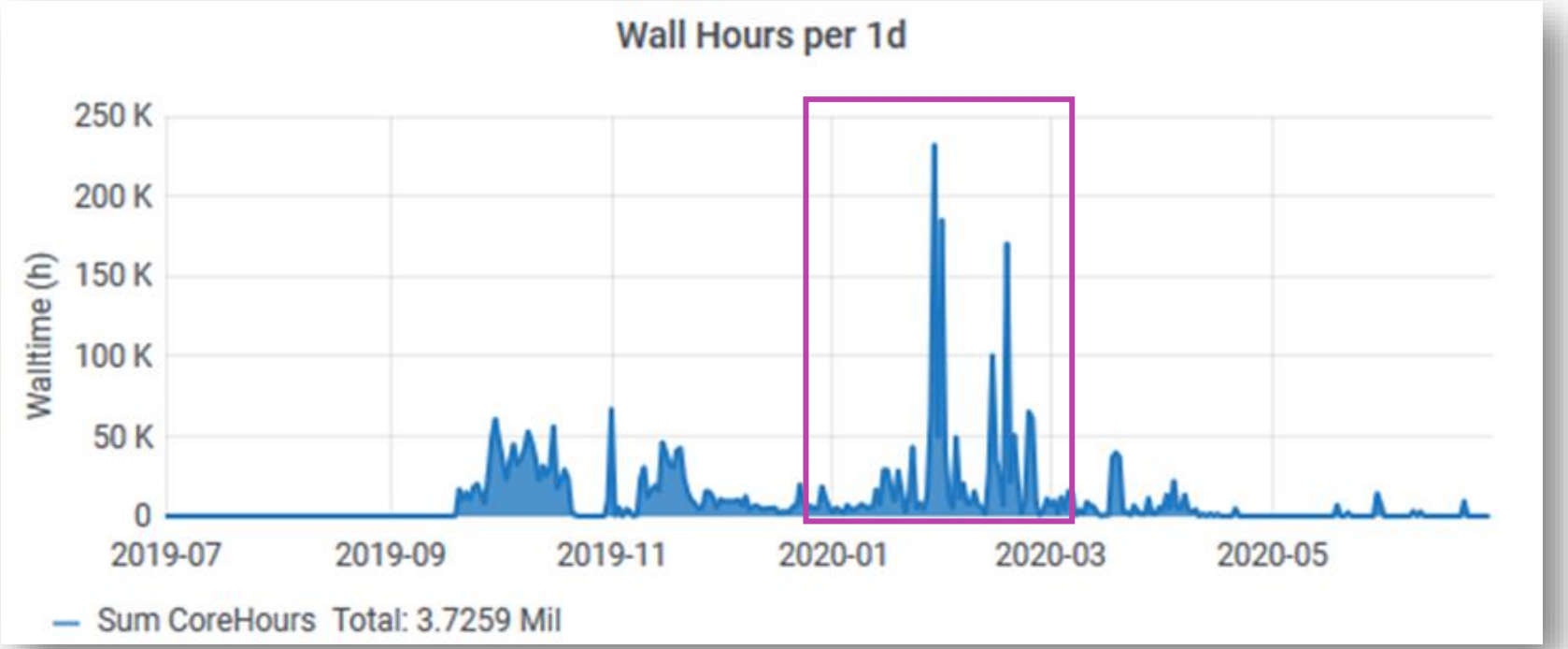
1 2 3 4 5 6 7 8 9 10 11 12

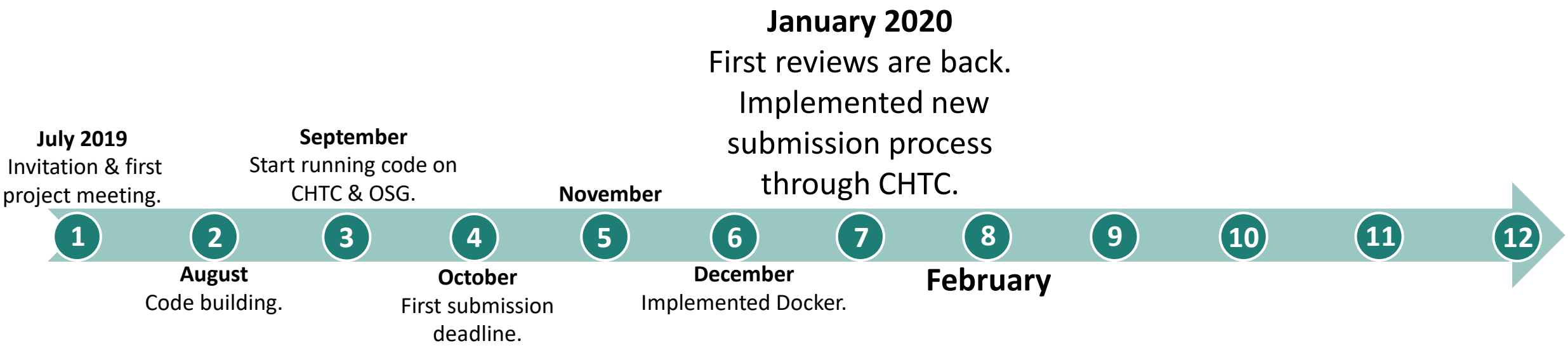
August
Code building.

October
First submission
deadline.

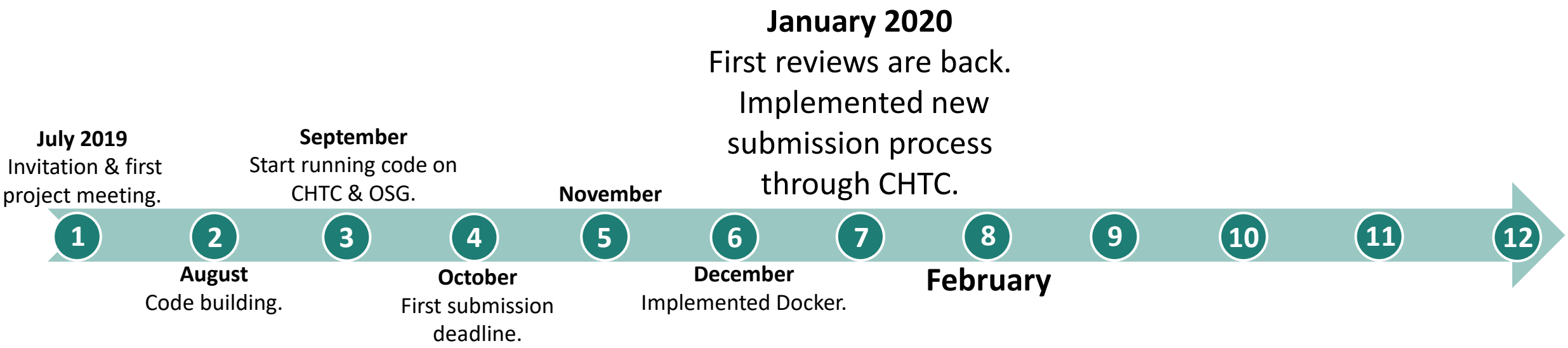
December
Implemented Docker.

February





- Reviews required major revisions (no surprise).



- Reviews required major revisions (no surprise).
- Re-evaluated what scenarios we were going to run for the final submission.
  - ~1% of the original 317 M experiments.

January 2020
First reviews are back.
Implemented new
submission process
through CHTC.

July 2019
Invitation & first
project meeting.

September
Start running code on
CHTC & OSG.

November

1 2 3 4 5 6 7 8 9 10 11 12

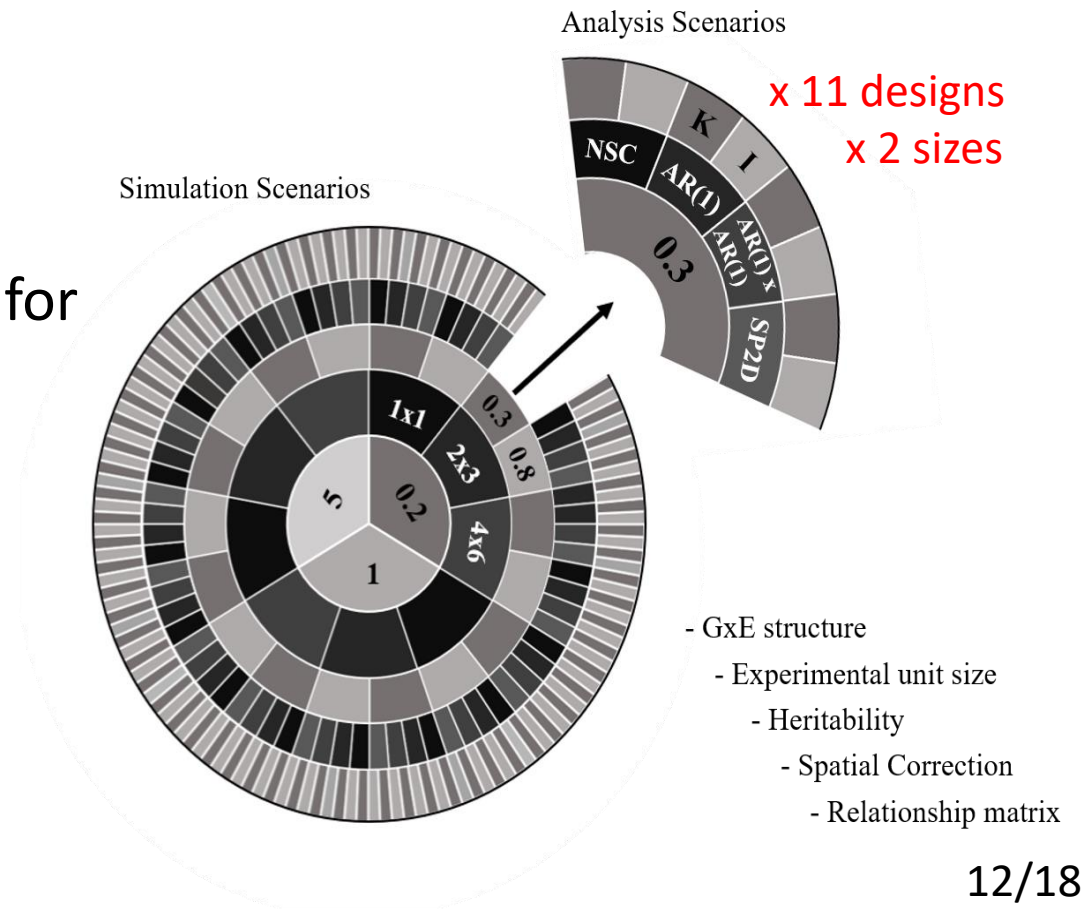
August
Code building.

October
First submission
deadline.

December
Implemented Docker.

February

- Reviews required major revisions (no surprise).
- Re-evaluated what scenarios we were going to run for the final submission.
  - ~1% of the original 317 M experiments.



January 2020
First reviews are back.
Implemented new
submission process
through CHTC.

July 2019
Invitation & first
project meeting.

September
Start running code on
CHTC & OSG.

November

1 2 3 4 5 6 7 8 9 10 11 12

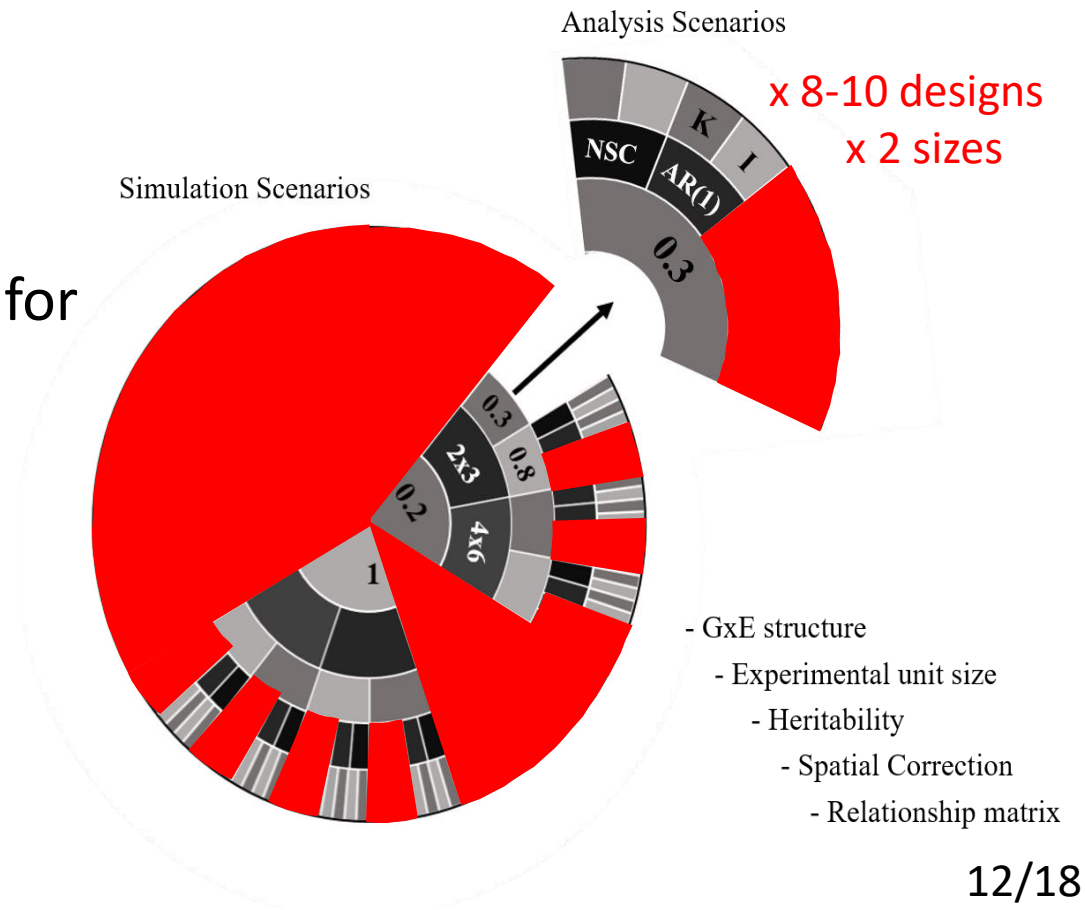
August
Code building.

October
First submission
deadline.

December
Implemented Docker.

February

- Reviews required major revisions (no surprise).
- Re-evaluated what scenarios we were going to run for the final submission.
  - ~1% of the original 317 M experiments.



January 2020
First reviews are back.
Implemented new
submission process
through CHTC.

July 2019
Invitation & first
project meeting.

September
Start running code on
CHTC & OSG.

November

1 2 3 4 5 6 7 8 9 10 11 12

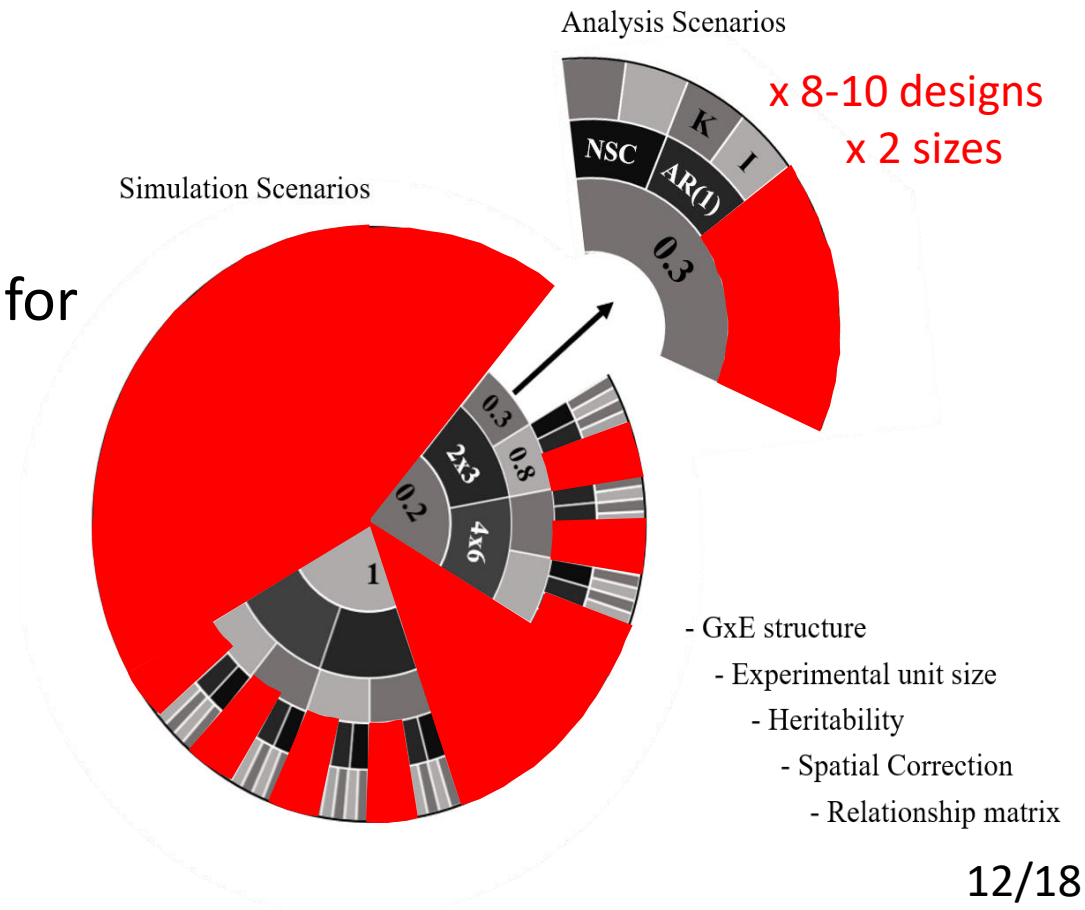
August
Code building.

October
First submission
deadline.

December
Implemented Docker.

February

- Reviews required major revisions (no surprise).
- Re-evaluated what scenarios we were going to run for the final submission.
  - ~1% of the original 317 M experiments.
- Started completely from scratch. Error in one calculation in previous code.



January 2020
First reviews are back.
Implemented new
submission process
through CHTC.

July 2019
Invitation & first project meeting.

September
Start running code on CHTC & OSG.

November

1 2 3 4 5 6 7 8 9 10 11 12

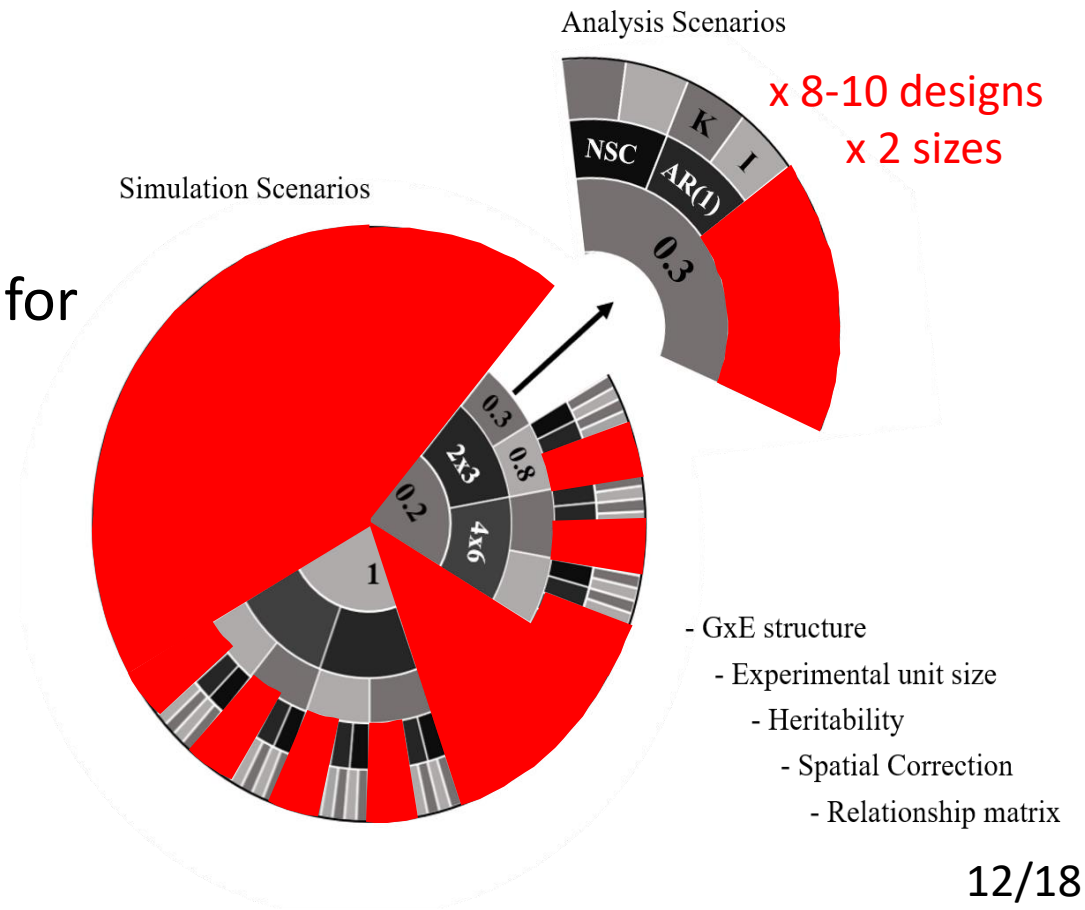
August
Code building.

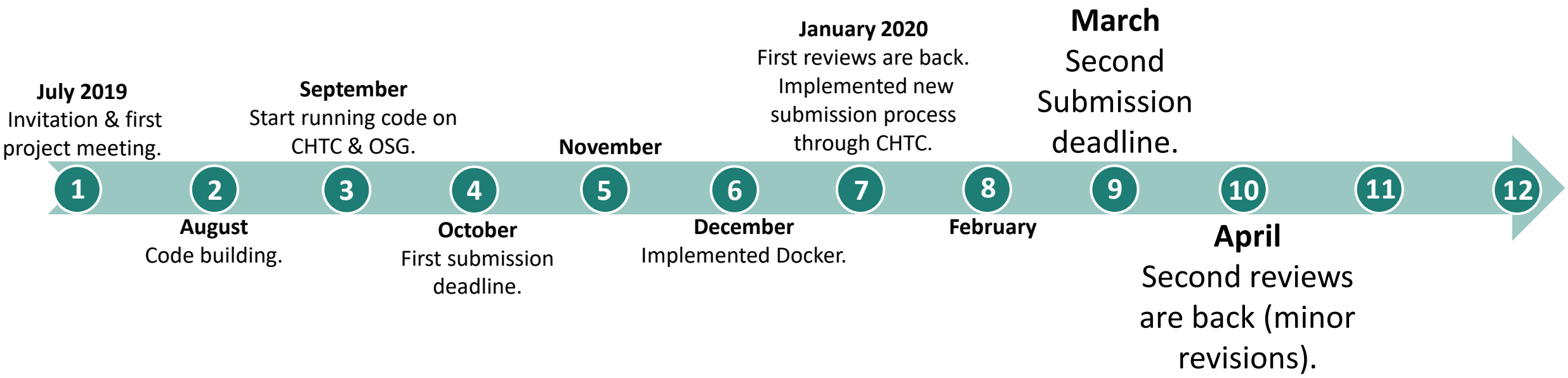
October
First submission deadline.

December
Implemented Docker.

February

- Reviews required major revisions (no surprise).
- Re-evaluated what scenarios we were going to run for the final submission.
  - ~1% of the original 317 M experiments.
- Started completely from scratch. Error in one calculation in previous code.
- Challenge #5: Job submission.
  - Old method: 3 files created for every job submitted.
  - New method: 3 files total.




July 2019
Invitation & first project meeting.
1
August
Code building.
2
September
Start running code on CHTC & OSG.
3
October
First submission deadline.
4
November
5
December
Implemented Docker.
6
January 2020
First reviews are back. Implemented new submission process through CHTC.
7
February
8
March
Second Submission deadline.
9
April
Second reviews are back (minor revisions).
10
11
12

July 2019
Invitation & first project meeting.

August
Code building.

September
Start running code on CHTC & OSG.

October
First submission deadline.

November

December
Implemented Docker.

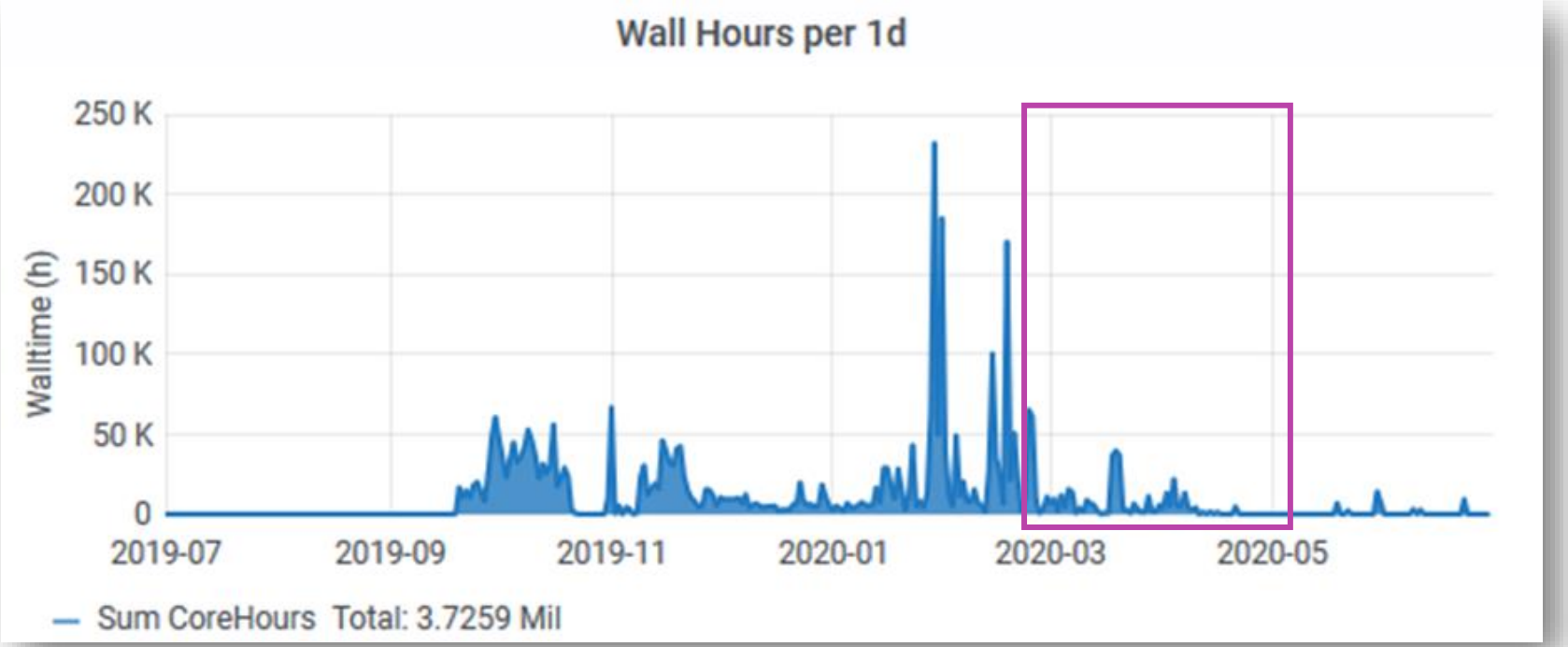
January 2020
First reviews are back. Implemented new submission process through CHTC.

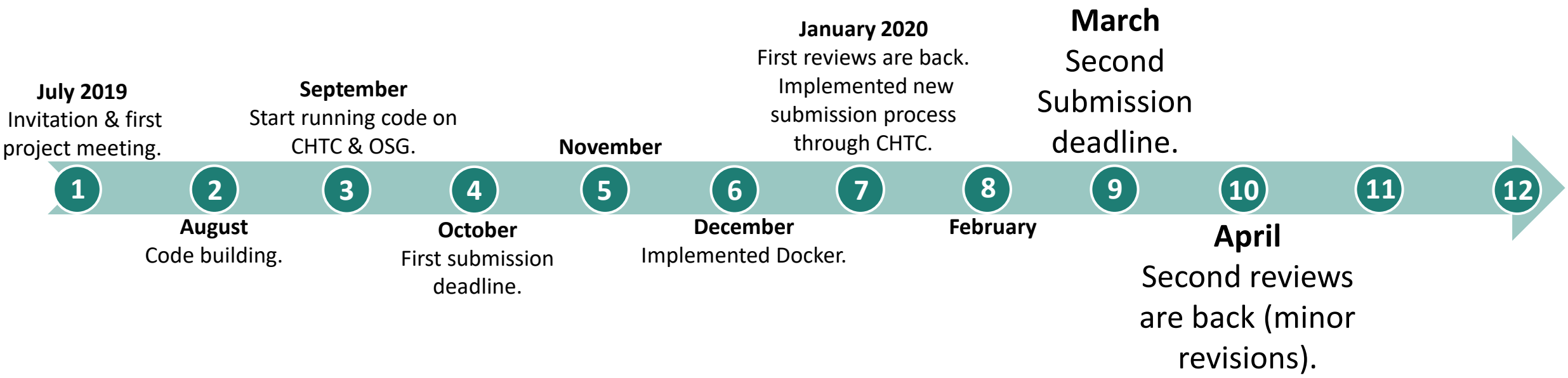
February

March
Second Submission deadline.

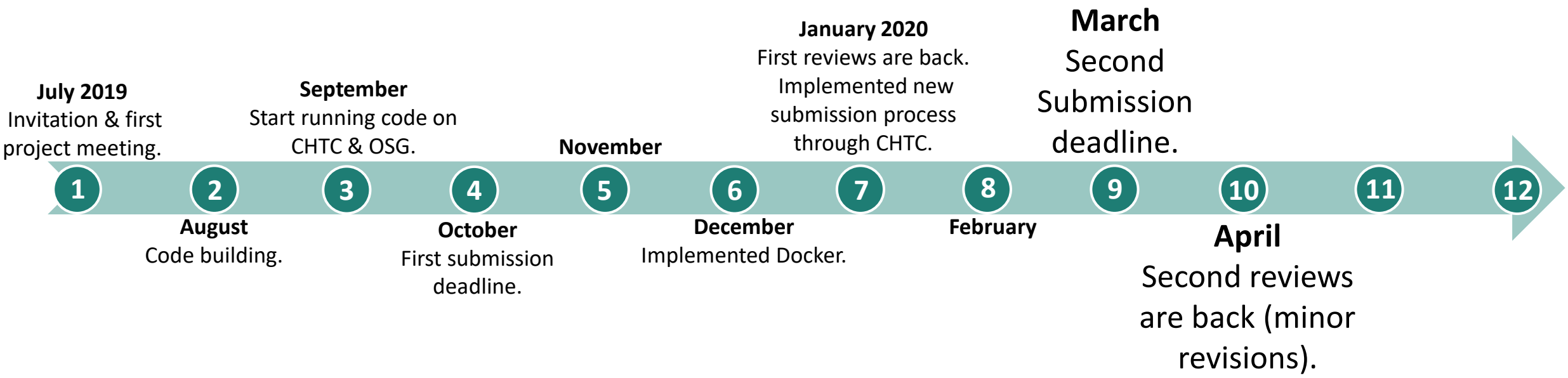
April
Second reviews are back (minor revisions).

1 2 3 4 5 6 7 8 9 10 11 12





- Had to re-run everything one last time to address final reviews.



- Had to re-run everything one last time to address final reviews.
- Minimizing user error.
- Better estimates of timing.

July 2019
Invitation & first project meeting.

August
Code building.

September
Start running code on CHTC & OSG.

October
First submission deadline.

November

December
Implemented Docker.

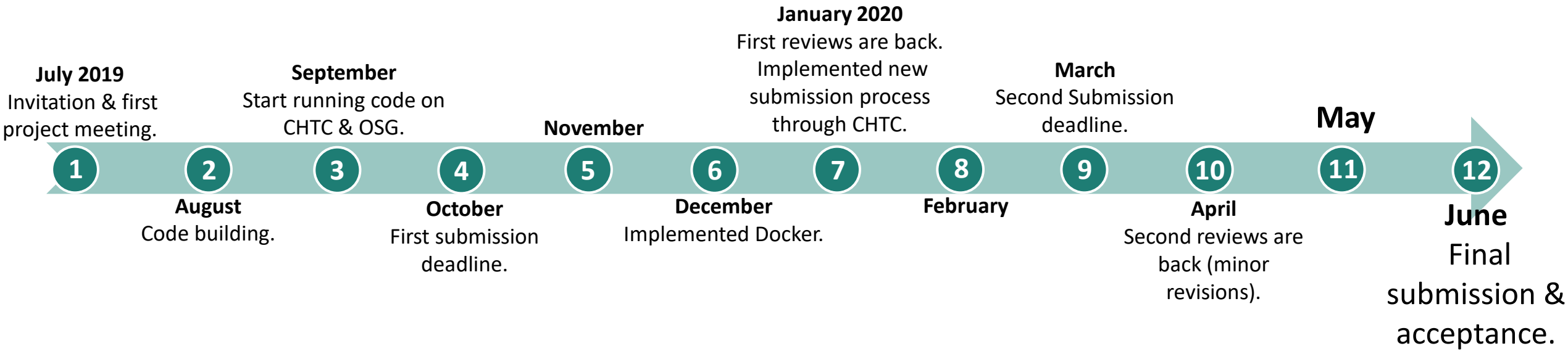
January 2020
First reviews are back. Implemented new submission process through CHTC.

February

March
Second Submission deadline.

April
Second reviews are back (minor revisions).

- Had to re-run everything one last time to address final reviews.
- Minimizing user error.
- Better estimates of timing.


July 2019
Invitation & first project meeting.
1
August
Code building.
2
September
Start running code on CHTC & OSG.
3
October
First submission deadline.
4
November
5
December
Implemented Docker.
6
January 2020
First reviews are back. Implemented new submission process through CHTC.
7
February
8
March
Second Submission deadline.
9
April
Second reviews are back (minor revisions).
10
May
11
June
Final submission & acceptance.
12

July 2019
Invitation & first project meeting.

August
Code building.

September
Start running code on CHTC & OSG.

October
First submission deadline.

November

December
Implemented Docker.

January 2020
First reviews are back. Implemented new submission process through CHTC.

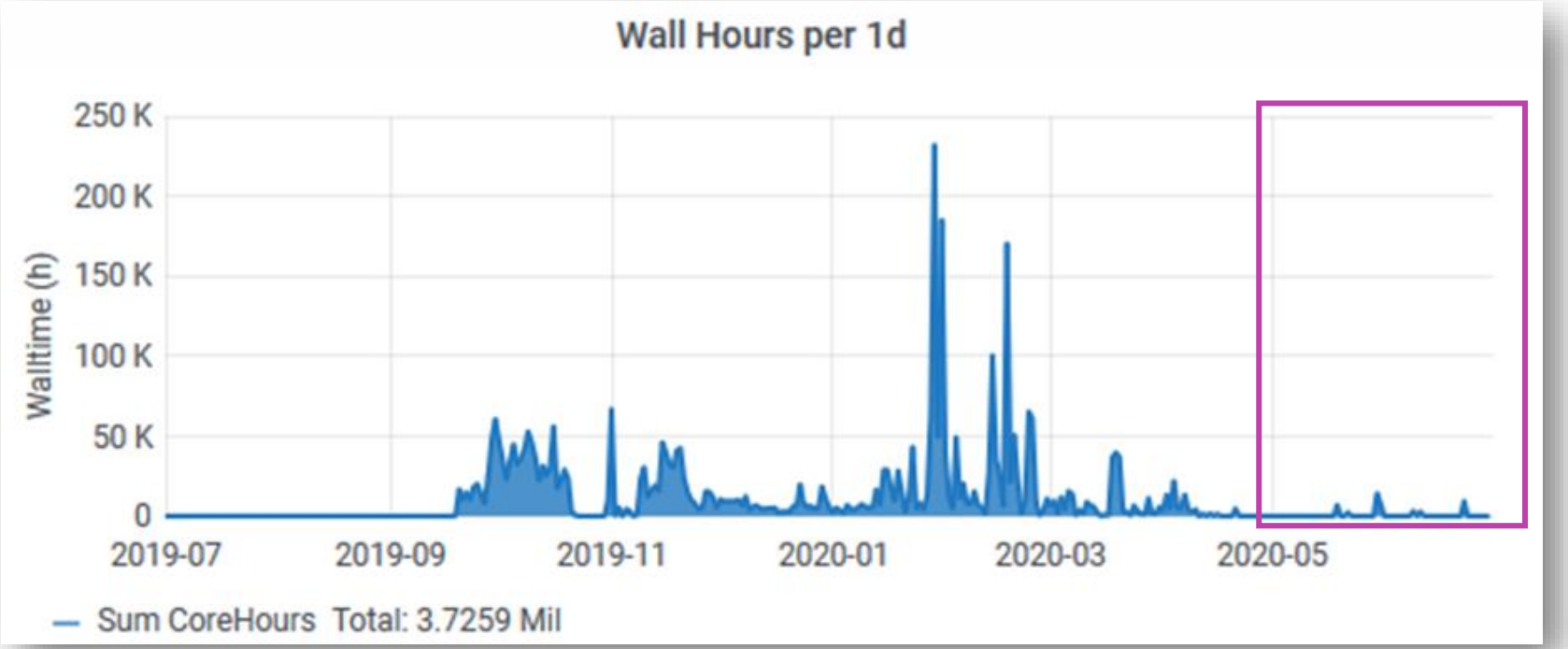
February

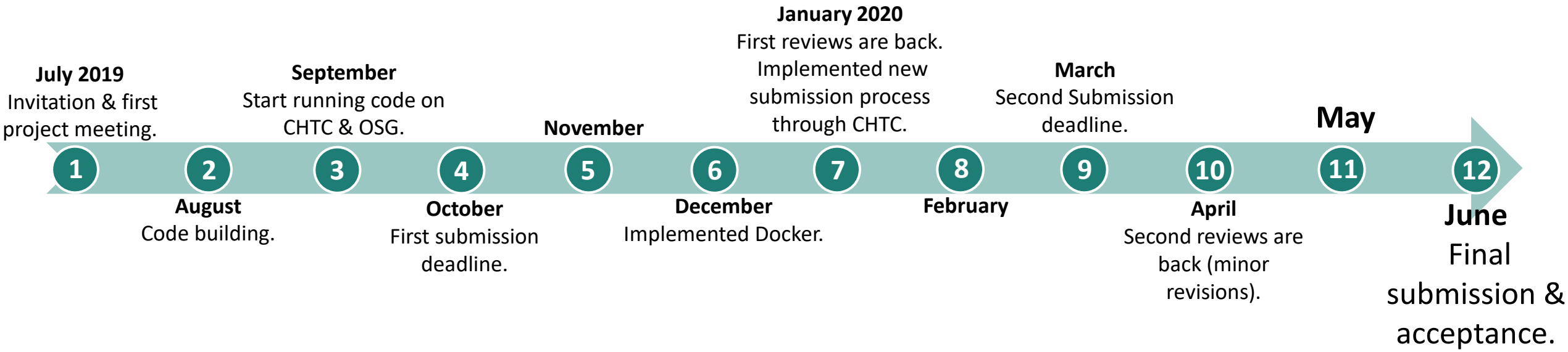
March
Second Submission deadline.

April
Second reviews are back (minor revisions).

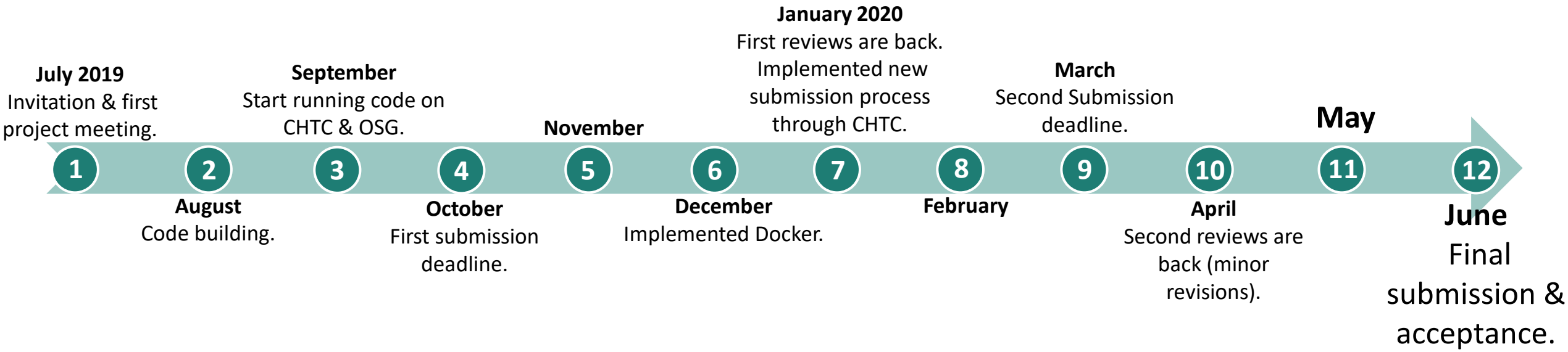
May

June
Final submission & acceptance.

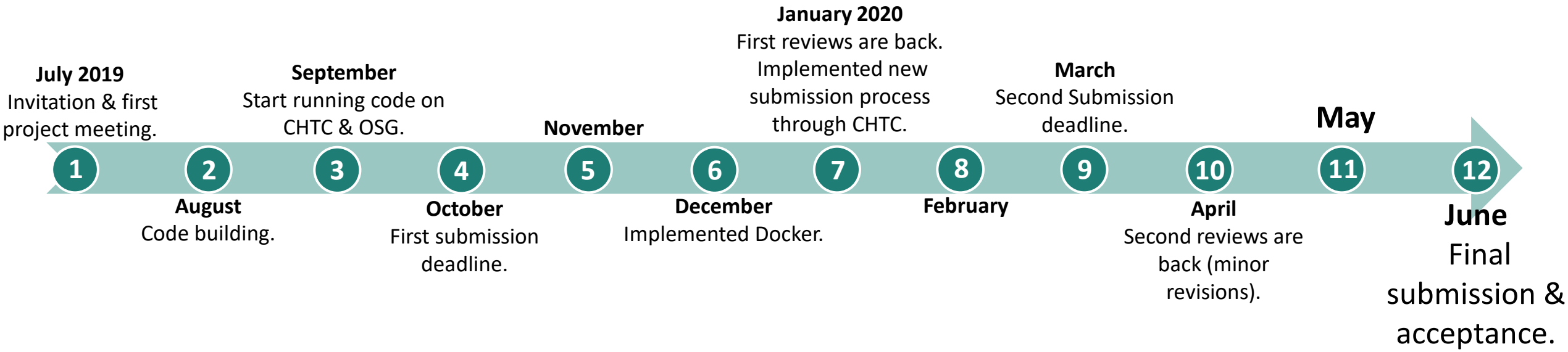




- <u>Impact:</u> Provides a comprehensive guide to researchers when designing experiments.
  - Not possible without these computing resources.
  - Most limiting factor in genetic gains over the next 25 years is resources for phenotyping (Cooper et al., 2014).



- <u>Impact:</u> Provides a comprehensive guide to researchers when designing experiments.
  - Not possible without these computing resources.
  - Most limiting factor in genetic gains over the next 25 years is resources for phenotyping (Cooper et al., 2014).
- <u>Training for myself</u>: statistics, experimental design, simulations, computing, coding, genetics, organization, teamwork, etc.
  - First chapter of my dissertation.



- <u>Impact:</u> Provides a comprehensive guide to researchers when designing experiments.
  - Not possible without these computing resources.
  - Most limiting factor in genetic gains over the next 25 years is resources for phenotyping (Cooper et al., 2014).
- <u>Training for myself</u>: statistics, experimental design, simulations, computing, coding, genetics, organization, teamwork, etc.
  - First chapter of my dissertation.
- <u>Results</u>: ALPHA and RC designs are often better than spatial designs, and spatial corrections are more effective in larger experiments.

"Breeder's Equation"

$$R = i\underbrace{r_{g,\hat{g}}}_{\text{Correlation}}\sigma_g$$

True (g)

Predicted ($\hat{g}$)

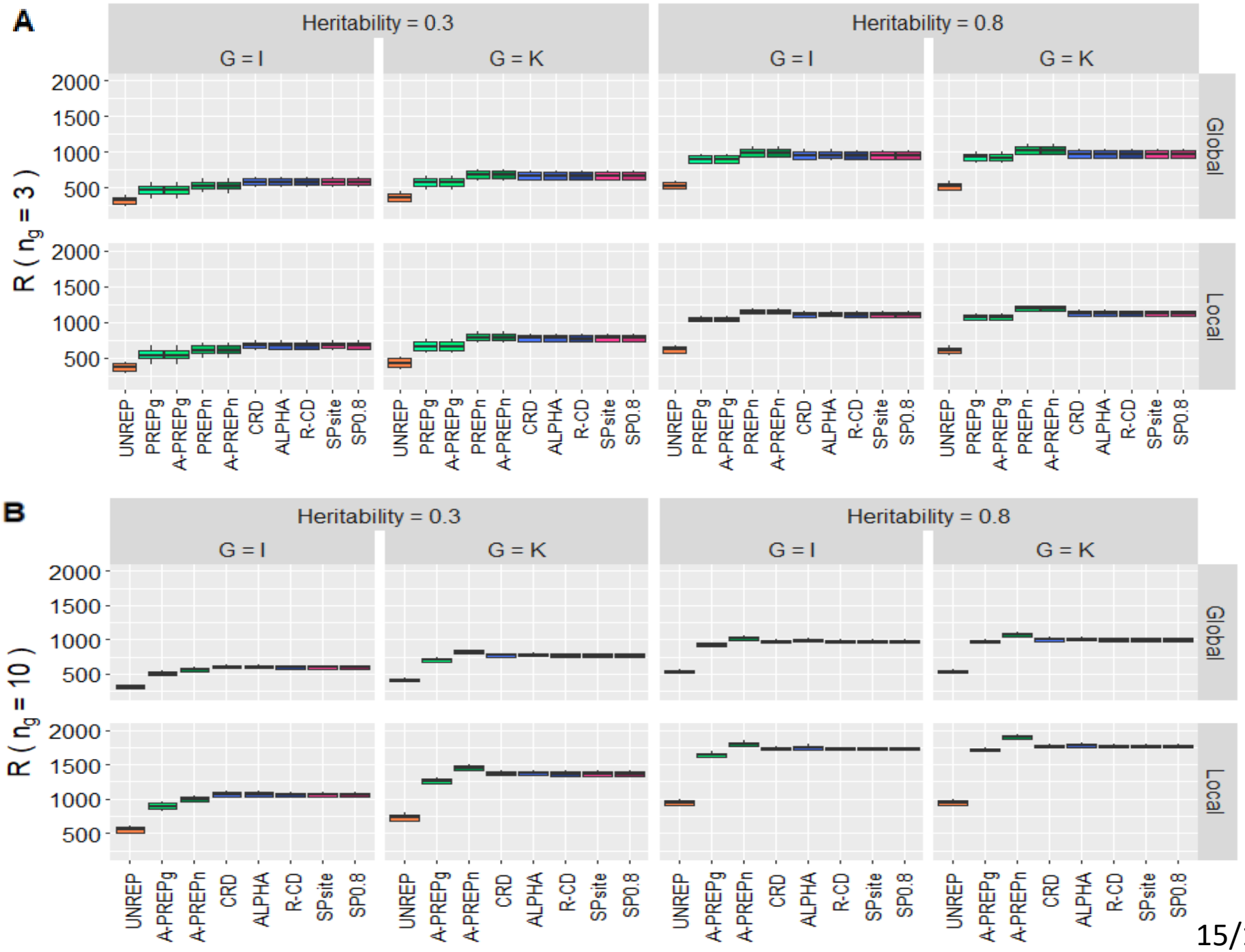
"Breeder's Equation"

$$R = i r_{g,\hat{g}} \sigma_g$$

Correlation

True (g)

Predicted ($\hat{g}$)



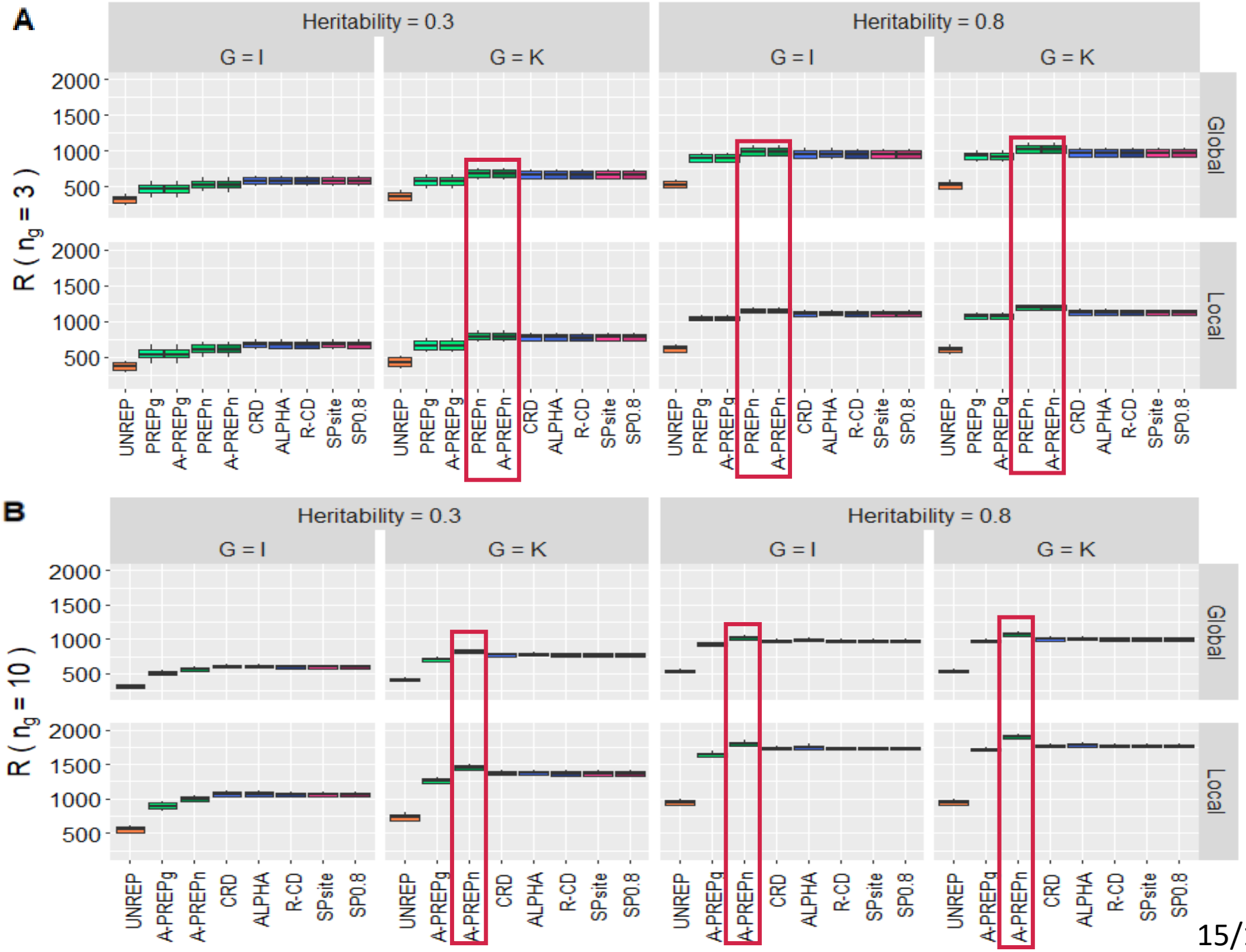
"Breeder's Equation"

$$R = i r_{g,\hat{g}} \sigma_g$$

Correlation

True (g)

Predicted ($\hat{g}$)



"Breeder's Equation"

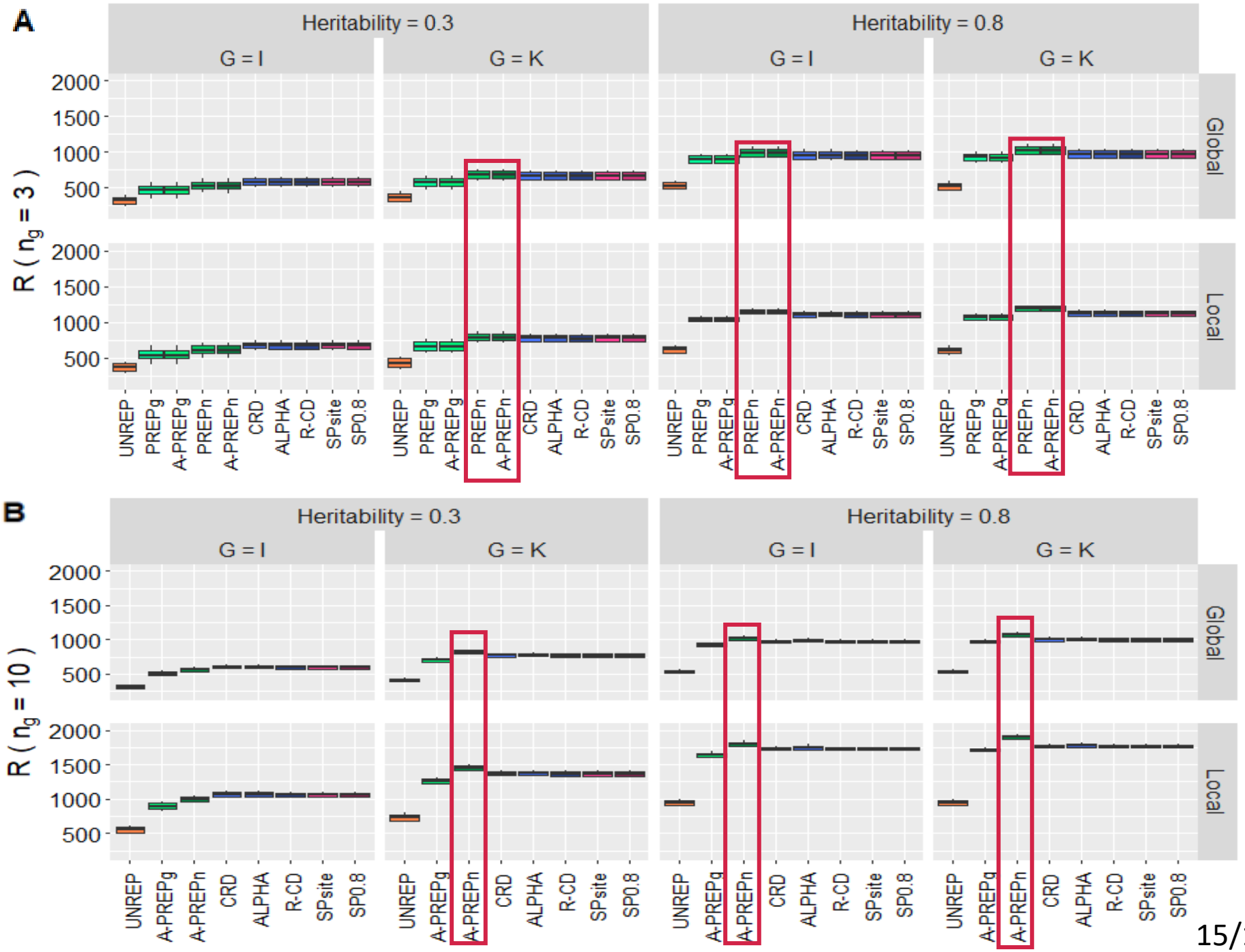
$$R = i r_{g,\hat{g}} \sigma_g$$

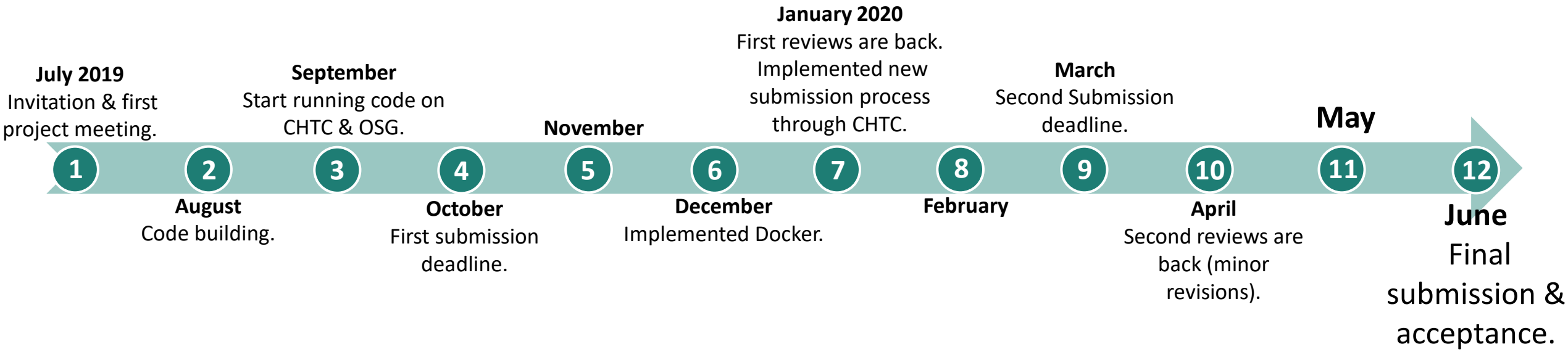
Correlation

True (g)

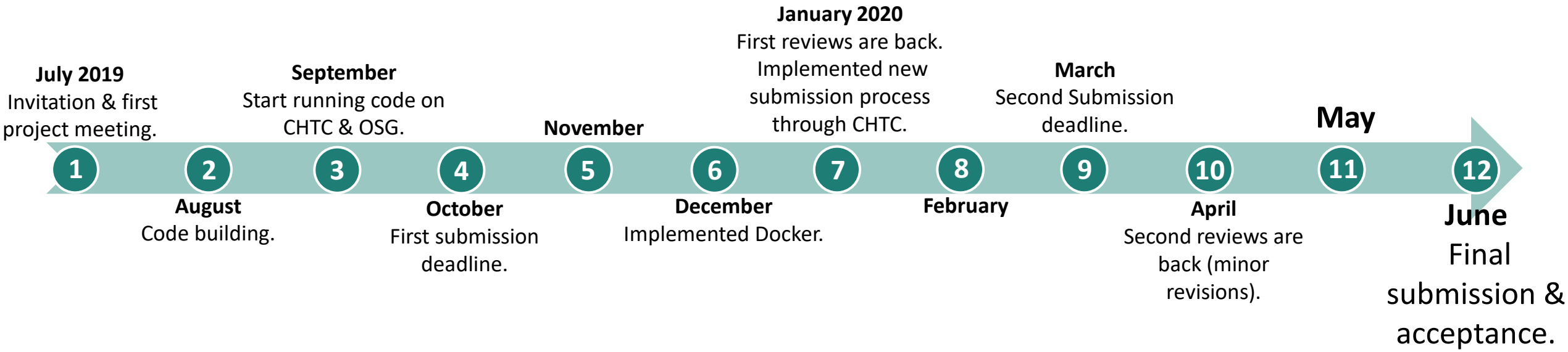
Predicted ($\hat{g}$)

~130 K hours

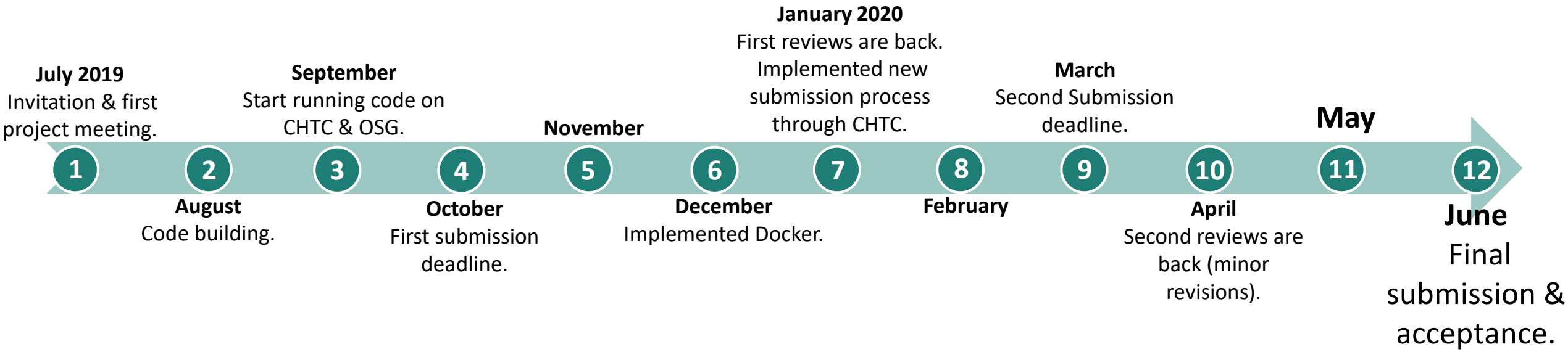




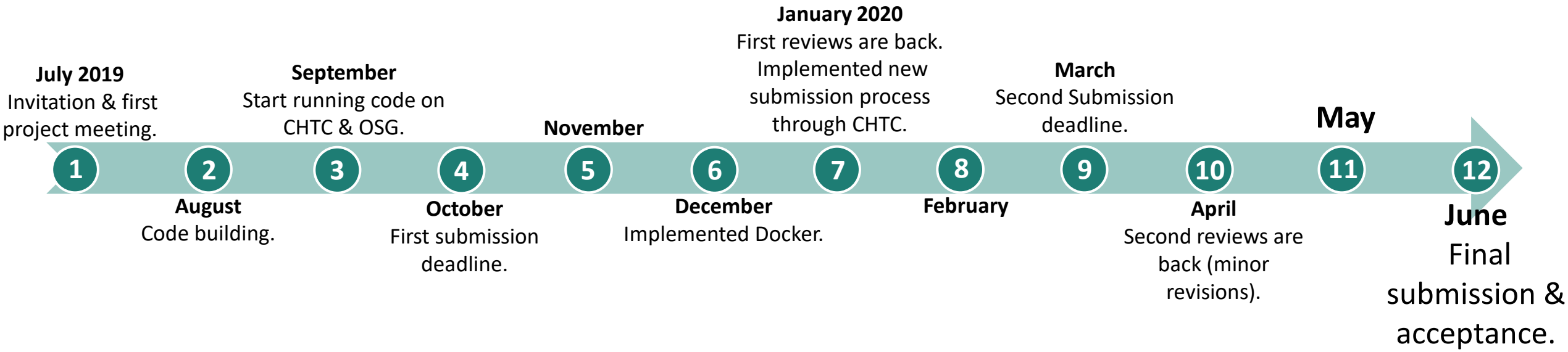
- Final reflections



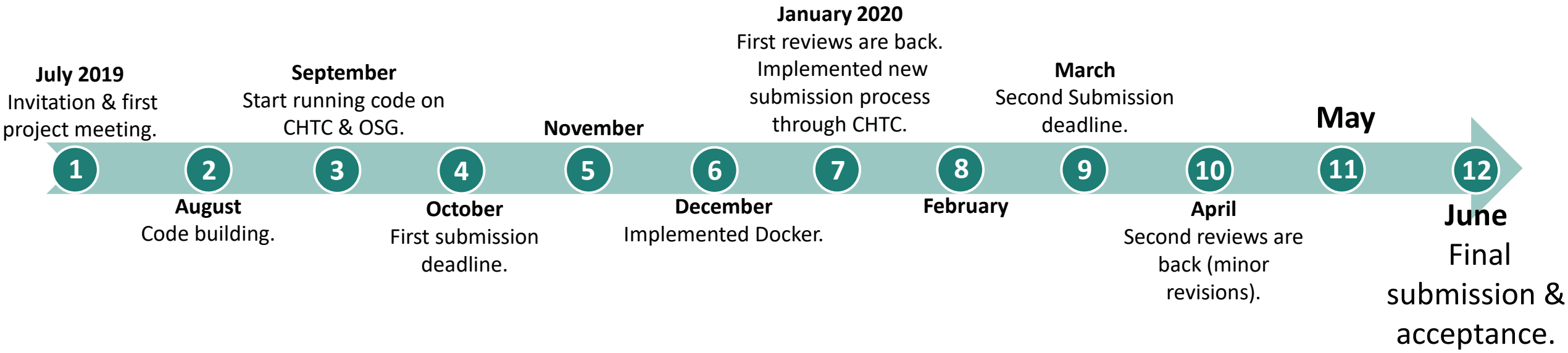
- Final reflections
  - Huge undertaking.



- Final reflections
  - Huge undertaking.
  - Deadlines were our largest battle.



- Final reflections
  - Huge undertaking.
  - Deadlines were our largest battle.
  - Pilot studies are valuable.



- Final reflections
  - Huge undertaking.
  - Deadlines were our largest battle.
  - Pilot studies are valuable.
  - Interactive help sessions were very useful.
    - OSG User School
    - Office hours

Supplementary materials for this article are available at https://doi.org/10.1007/s13253-020-00406-2.

Do Spatial Designs Outperform Classic Experimental Designs?

Raegan HOEFLER, Pablo GONZÁLEZ- BARRIOS, Madhav BHATTA, Jose A. R. NUNES, Ines BERRO, Rafael S. NALIN, Alejandra BORGES, Eduardo COVARRUBIAS, Luis DIAZ- GARCIA, Martin QUINCKE, and Lucia GUTIERREZ

Controlling spatial variation in agricultural field trials is the most important step to compare treatments efficiently and accurately. Spatial variability can be controlled at the experimental design level with the assignment of treatments to experimental units and at the modeling level with the use of spatial corrections and other modeling strategies. The goal of this study was to compare the efficiency of methods used to control spatial variation in a wide range of scenarios using a simulation approach based on real wheat data. Specifically, classic and spatial experimental designs with and without a two-dimensional autoregressive spatial correction were evaluated in scenarios that include differing experimental unit sizes, experiment sizes, relationships among genotypes, genotype by environment interaction levels, and trait heritabilities. Fully replicated designs outperformed partially and unreplicated designs in terms of accuracy; the alpha-lattice incomplete block design was best in all scenarios of the medium-sized experiments. However, in terms of response to selection, partially replicated experiments that evaluate large population sizes were superior in most scenarios. The AR1 × AR1 spatial correction had little benefit in most scenarios except for the medium-sized experiments with the largest experimental unit size and low GE. Overall, the results from this study provide a guide to researchers designing and analyzing large field experiments.
Supplementary materials accompanying this paper appear online.

Key Words: Experimental design; Autoregressive process; Prediction accuracy; Response to selection; Spatial correction; Randomization-based experimental designs.

DOI: 10.1007/s13253-020-00406-2

rshoefler@wisc.edu

Acknowledgements

Authors

Pablo Gonzalez-Barrios
Madhav Bhatta
Jose A. R. Nunes
Ines Berro
Rafael S. Nalin
Alejandra Borges
Eduardo Covarrubias
Luis Diaz-Garza
Martin Quincke
Lucia Gutierrez

Gutierrez Lab

CHTC and OSG Team

Lauren Michael
Christina Koch
Jess Vera

OSG CONNECT

CENTER FOR HIGH THROUGHPUT COMPUTING

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON