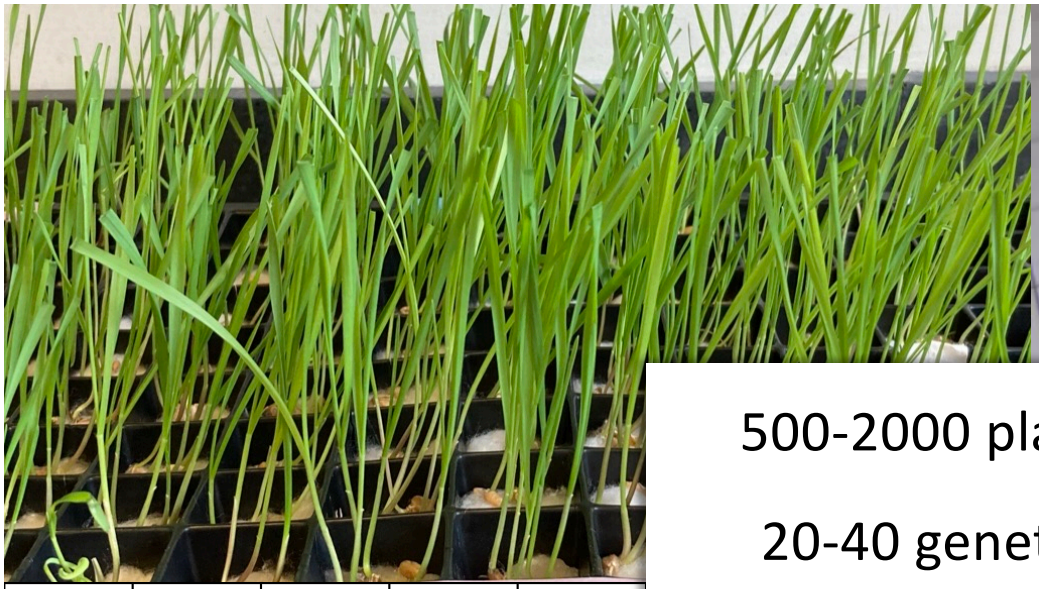
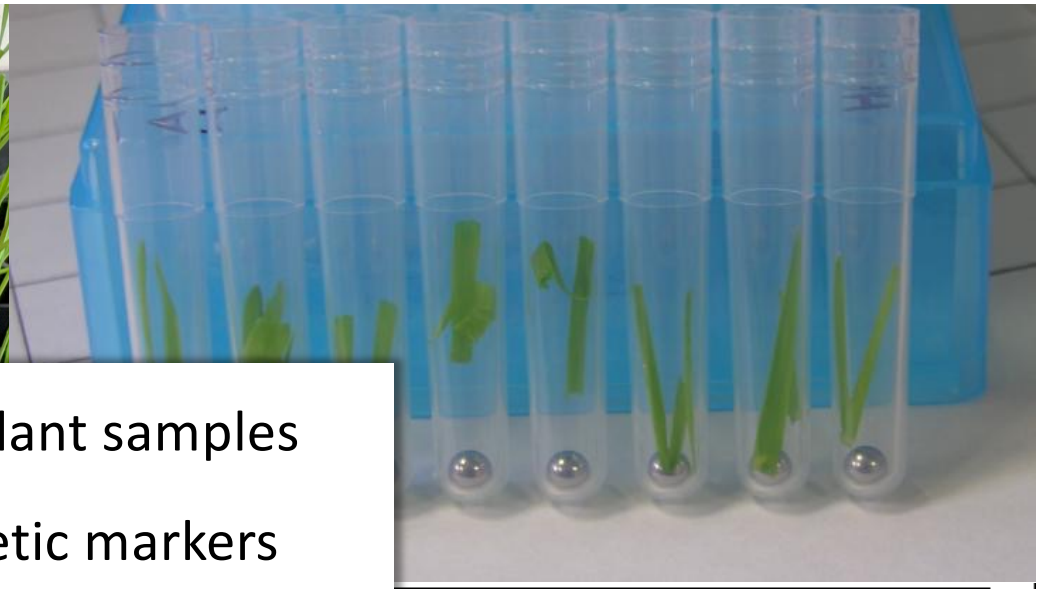
Falling into data science

Jeanette Lyerly

October 2022

- Research scientist at NCSU
- Background in plant breeding and biotech
- Working on soft red winter wheat





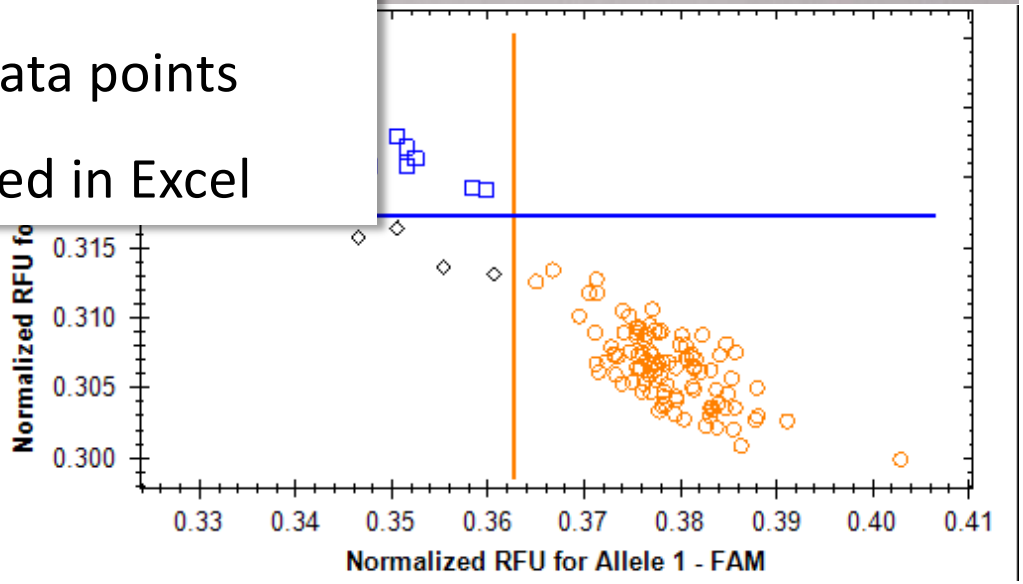
500-2000 plant samples

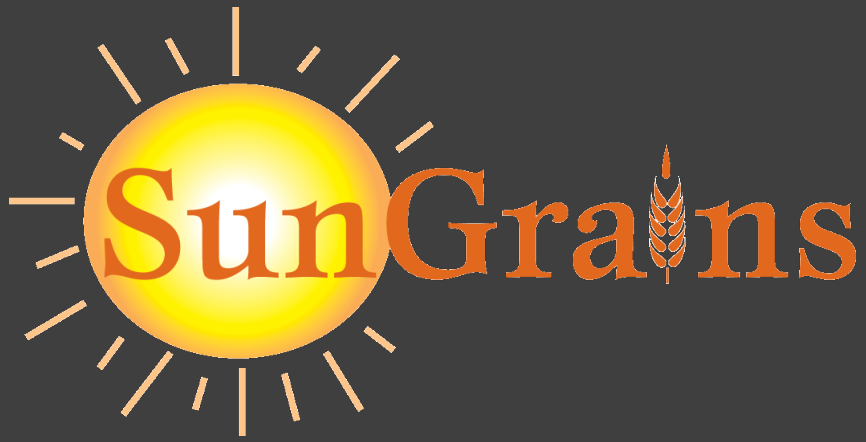
20-40 genetic markers

1000s of data points

Data handled in Excel

Well	Content	Raw Data (485/610 1 A)	Raw Data (485/520 1 B)	Raw Data (520/610 2 A)			
A02	Sample X2	58788	25128	61660			
A04	Sample X4	76164	87220	71876			
A06	Sample X6	75956	30488	85960	71256	0.4	0.83
A08	Sample X8	78148	95208	69820	23168	1.22	0.33
A10	Sample X10	67048	65056	63452	18068	0.97	0.28
A12	Sample X12	74788	83656	71796	21268	1.12	0.3
A14	Sample X14	73612	80804	69456	21904	1.1	0.32
A16	Sample X16	70124	73736	68864	18960	1.05	0.28
A18	Sample X18	58256	23752	66792	11824	0.41	0.18
A20	Sample X20	68244	29076	77008	50704	0.43	0.66





- NCSU part of a cooperative program with other southern universities
- Started a new genomics project in 2017
- Pool their resources and data
- Prediction modeling for new breeding material

Journey to data science

- Field and genetic data from multiple research groups
- Mountains of data!
- Start learning R for data wrangling and analysis
- Very little prior experience in coding

4000 – 6000 plants

~20K genetic markers

1,000,000s of data points

Data handled in R

17:154493:LA21GS0447:...	G	A	G	A	G	C	G	T	C	C	T	A	G	A	A	T	G	G	A	T	G	G	T	G	C	A	C	C	G	C
17:154464:LA21GS0418:...	A	G	A	G	A	T	A	T	A	C	T	A	A	G	A	T	A	G	A	T	G	G	T	G	C	A	C	C	G	C
02:123099:ARLW08079D...	A	G	A	G	A	T	A	T	A	C	T	A	R	G	A	T	A	G	A	T	G	G	T	G	C	A	C	C	G	C
01:109705:VA07MAS3-7...	G	A	G	A	G	C	G	T	C	G	G	G	G	A	C	C	G	A	G	C	T	A	C	A	C	A	C	A	C	C
17:154581:LA21GS0535:...	A	G	A	G	A	T	A	T	A	C	T	A	A	G	A	T	A	G	A	T	G	G	T	G	C	A	C	C	C	C
24:122979:ARS13-348	A	G	A	G	A	T	A	T	A	C	T	A	G	A	A	T	G	G	A	T	G	G	T	G	G	G	G	C	G	C
17:145233:TX20GS0198:...	G	A	G	A	G	C	G	T	C	C	K	A	G	A	A	Y	G	A	G	C	K	R	T	G	C	A	C	C	G	C
17:154597:LA21GS0551:...	G	A	G	A	G	C	G	T	C	C	T	A	G	A	A	T	G	G	A	T	G	G	T	G	C	A	C	C	G	C
02:131389:LA10042D-6...	A	G	A	G	A	T	A	T	A	C	T	A	G	A	A	T	G	G	A	Y	G	G	T	G	G	G	G	C	G	C
02:131396:LA14066LDH...	G	A	G	A	G	C	G	T	C	G	G	G	G	A	C	C	G	A	G	C	T	A	C	A	C	A	C	C	G	C
17:143418:LA20GS0111:...	G	A	G	A	G	C	G	T	C	C	T	A	G	A	A	T	G	G	A	T	G	G	T	G	C	A	C	C	G	C
02:123118:FLTX15D9003	G	A	G	A	G	C	G	T	C	G	G	G	G	A	C	C	G	A	G	C	T	A	C	A	C	A	C	C	C	C
01:131327:HILLIARD	A	G	A	G	A	T	A	T	A	C	T	A	G	A	A	T	G	G	A	T	G	G	T	G	G	G	G	C	G	C
01:109679:LA09264C-P5	A	G	A	G	A	T	A	T	A	C	T	A	G	R	A	T	G	G	A	T	G	G	T	G	G	G	G	C	G	C

rs#	alleles	chrom	pos	strand	assemb															0001		FHBWATSG0002				FHBWATSG0003			
S1A_1197644	A/C	1	1197644	+	NA	NA														2	2	0	2	2	2	2	2	0	2
S1A_1236236	A/C	1	1236236	+	NA	NA														0	0	0	0	0	0	1	0	0	0
S1A_1236317	A/C	1	1236317	+	NA	NA														0	0	0	0	0	0	0	0	0	0
S1A_1338362	A/C	1	1338362	+	NA	NA														0	0	0	0	0	0	0	0	0	0
S1A_1627805	A/C	1	1627805	+	NA	NA														0	0	0	0	0	0	0	0	2	0
S1A_1846809	A/C	1	1846809	+	NA	NA														1	0	2	0	0	0	0	0	0	0
S1A_2390257	A/C	1	2390257	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0
S1A_2390264	A/C	1	2390264	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0
S1A_3381955	A/C	1	3381955	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	1	1	0	0	1
S1A_3452684	A/C	1	3452684	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S1A_3469197	A/C	1	3469197	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S1A_3516790	A/C	1	3516790	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0
S1A_3517435	A/C	1	3517435	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	1	0	1	2	0	0	0	0	0	0	1	0	0
S1A_4090235	A/C	1	4090235	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S1A_4159194	A/C	1	4159194	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S1A_4213053	A/C	1	4213053	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S1A_4227077	A/C	1	4227077	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S1A_4499981	A/C	1	4499981	+	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Incoming data

- Data for different traits
  - Yield
  - Plant height
  - Disease rating
- Coming in from seven collaborators
- Lots (and lots) of spreadsheets
- Everyone has their own established format

	GAWN-PL19			Mean	Mean
				82.8	57.1
Entry	Line	Pedigree	% of Mean	BUPA avg.	Test Wt. avg.
1	Hilliard		99.4	82.3 *	46.8
2	AGS3030		101.5	84	58.4

	A	B	C	D	E	F	G	H
1	EXPT	PLOT	REP	ENT	LOC	VAR	BUPA	TWT
2	GAWN19WN	247	2	1	WN	HILLIARD	63.541519	56
3	GAWN19WN	231	2	2	WN	AGS3030	44.979577	53

Name	entry	DYLD	TW
Hilliard	1	44	61.2
AGS3030	2	47	62.8

	A	B	C	D	E	F	G
1	plot	bloc	entry	name	heading	tw	yield
2	1	1	1	Hilliard	99	58.3	66.4
3	2	1	2	AGS3030	97	58.4	62.1

ENT	LOC	GENO	PED	R1_YIELD	R2_YIELD	R1_TWT	R2_TWT	% R1_MST	R2_MST
1	FL	Hilliard		55.3	56.9	42.8	44.3	9.3	9.6
2	FL	AGS3030		71.3	46.8	53.2	33.0	9.9	9.0

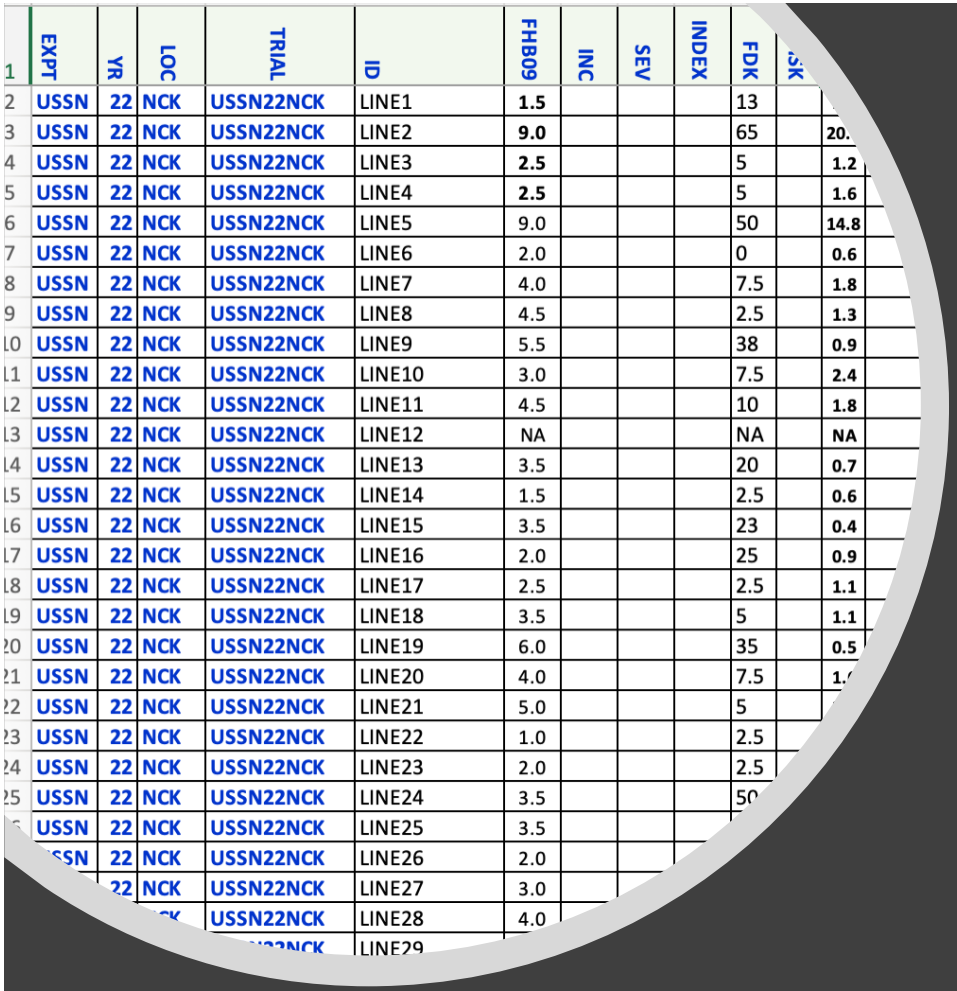
REP	ENT	PLOT	LOC	Yield	TW	FHB severity	Heading Date	Plant Ht	Lodging
1	1	1011	Marianna	75.4	57.8	2	4/11/19	35	
1	2	1016	Marianna	57.1	52.9	4	4/10/19	34	

Data wrangling

- Try managing data in Excel
- This was not a good idea
- Lots of time and frustration
- Repeating a lot of the same steps

Data wrangling with R

- Start with simple issues
  - Adjusting headers and names
  - Checking and converting units
  - Check data for errors
- Combine data from the different programs into a unified data set



	EXPT	YR	LOC	TRIAL	ID	FHB09	INC	SEV	INDEX	FDK	DSK	
2	USSN	22	NCK	USSN22NCK	LINE1	1.5				13		
3	USSN	22	NCK	USSN22NCK	LINE2	9.0				65		20.
4	USSN	22	NCK	USSN22NCK	LINE3	2.5				5		1.2
5	USSN	22	NCK	USSN22NCK	LINE4	2.5				5		1.6
6	USSN	22	NCK	USSN22NCK	LINE5	9.0				50		14.8
7	USSN	22	NCK	USSN22NCK	LINE6	2.0				0		0.6
8	USSN	22	NCK	USSN22NCK	LINE7	4.0				7.5		1.8
9	USSN	22	NCK	USSN22NCK	LINE8	4.5				2.5		1.3
10	USSN	22	NCK	USSN22NCK	LINE9	5.5				38		0.9
11	USSN	22	NCK	USSN22NCK	LINE10	3.0				7.5		2.4
12	USSN	22	NCK	USSN22NCK	LINE11	4.5				10		1.8
13	USSN	22	NCK	USSN22NCK	LINE12	NA				NA		NA
14	USSN	22	NCK	USSN22NCK	LINE13	3.5				20		0.7
15	USSN	22	NCK	USSN22NCK	LINE14	1.5				2.5		0.6
16	USSN	22	NCK	USSN22NCK	LINE15	3.5				23		0.4
17	USSN	22	NCK	USSN22NCK	LINE16	2.0				25		0.9
18	USSN	22	NCK	USSN22NCK	LINE17	2.5				2.5		1.1
19	USSN	22	NCK	USSN22NCK	LINE18	3.5				5		1.1
20	USSN	22	NCK	USSN22NCK	LINE19	6.0				35		0.5
21	USSN	22	NCK	USSN22NCK	LINE20	4.0				7.5		1.(
22	USSN	22	NCK	USSN22NCK	LINE21	5.0				5		
23	USSN	22	NCK	USSN22NCK	LINE22	1.0				2.5		
24	USSN	22	NCK	USSN22NCK	LINE23	2.0				2.5		
25	USSN	22	NCK	USSN22NCK	LINE24	3.5				50		
	USSN	22	NCK	USSN22NCK	LINE25	3.5						
	SSN	22	NCK	USSN22NCK	LINE26	2.0						
		22	NCK	USSN22NCK	LINE27	3.0						
			CK	USSN22NCK	LINE28	4.0						
				NCK	LINE29							

Data templates

- Continue to have formatting challenges
- Eventually want to be able to build some type of shiny app
- Sales pitch for templates so collaborators adopt a consistent format
- New template as close as possible to existing data recording sheets