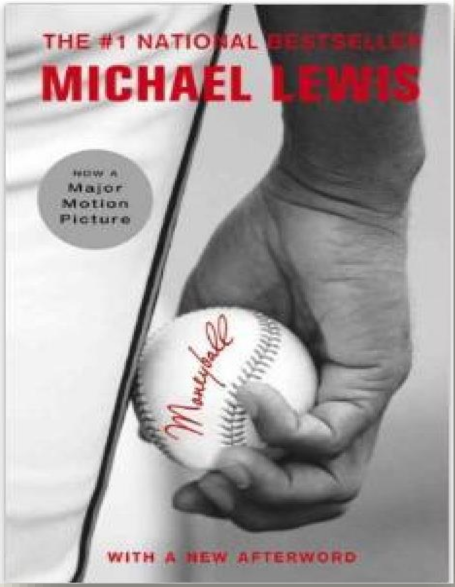
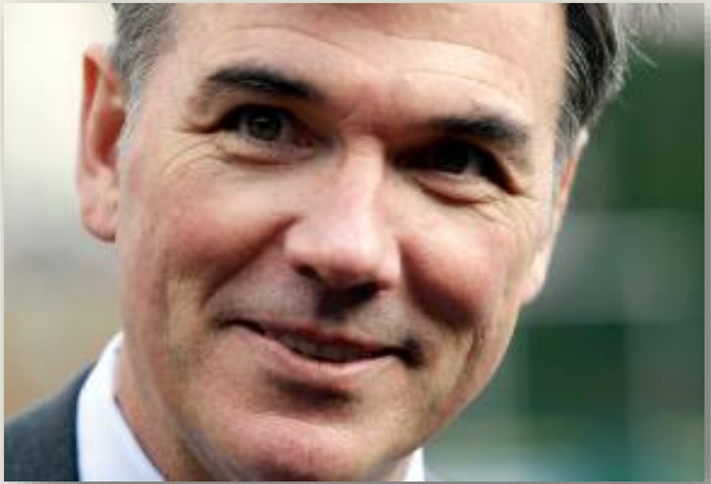
Troy Magennis (@t_magennis)

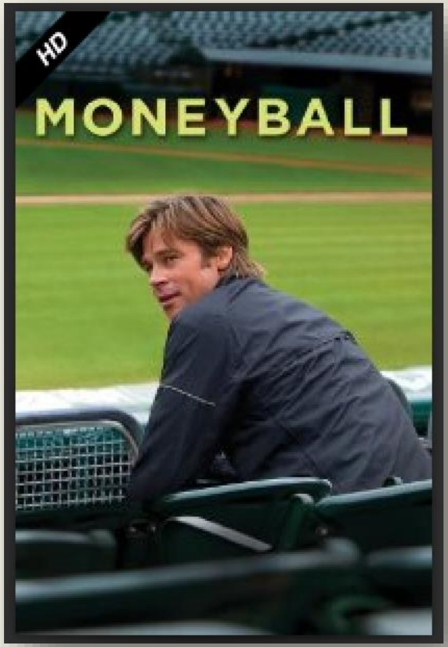
Moneyball for Software Projects:
Agile Metrics for the Metrically Challenged





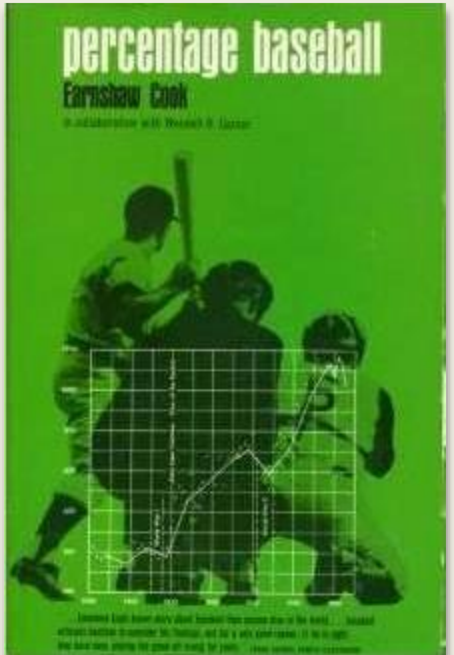
Brad Pitt

Billy Beane

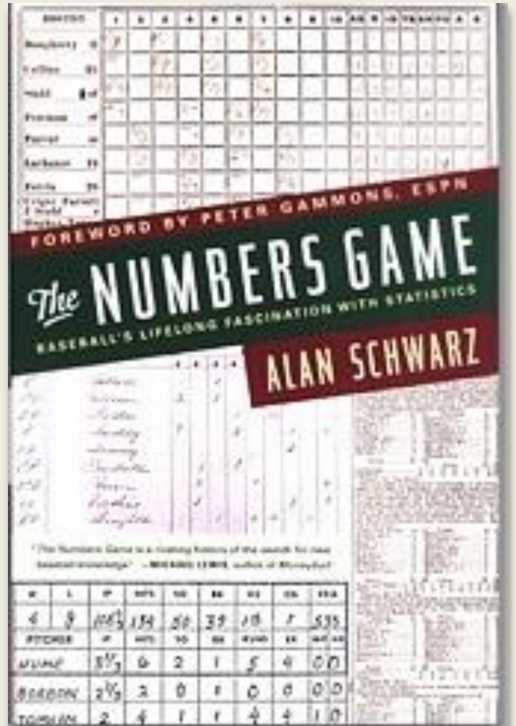


Paul DePodesta

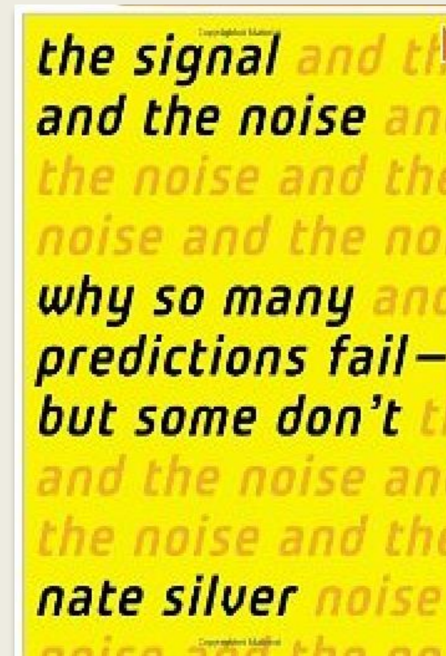
Jonah Hill (Playing fictional char.)



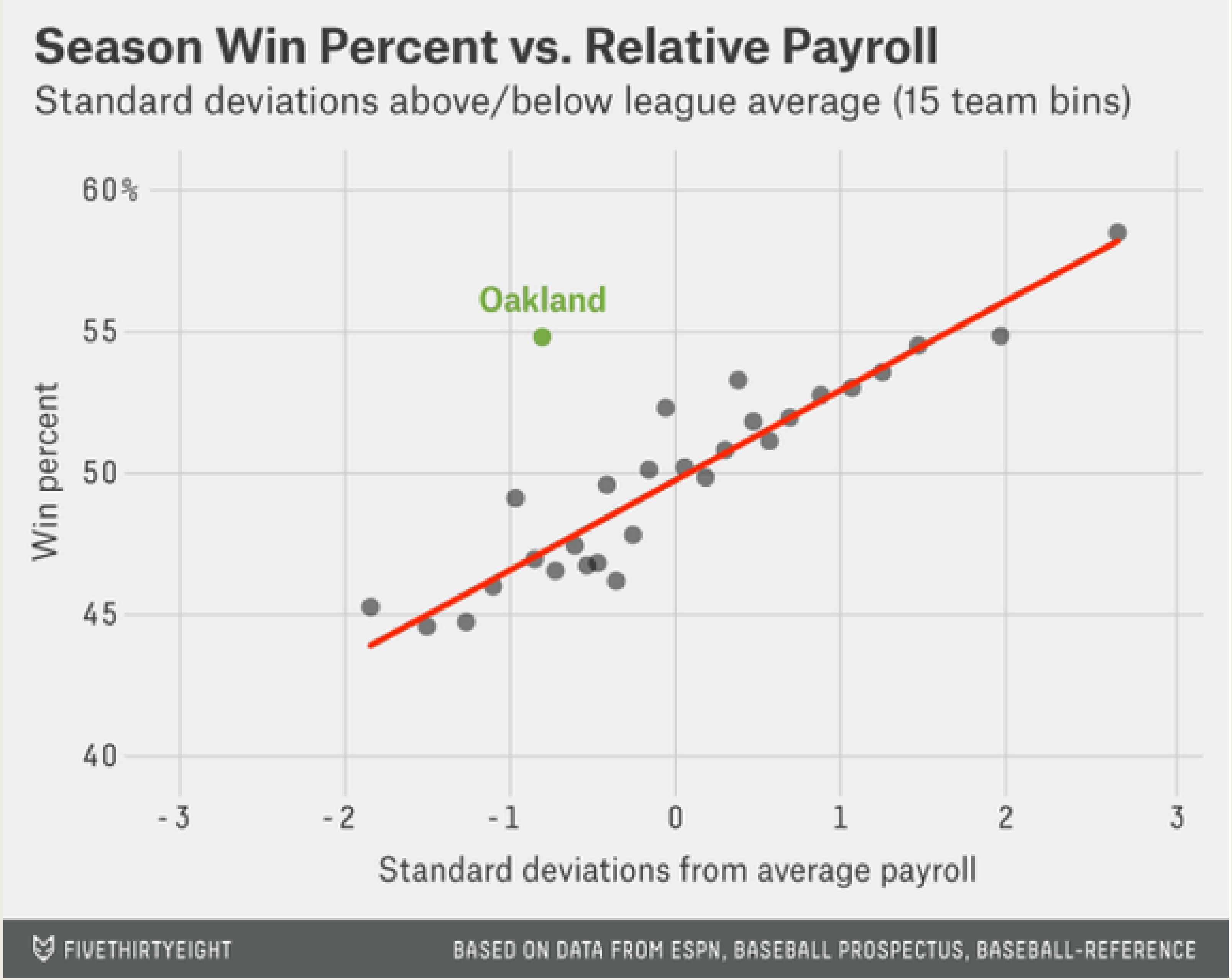
Earnshaw Cook
Percentage Baseball
(1964)

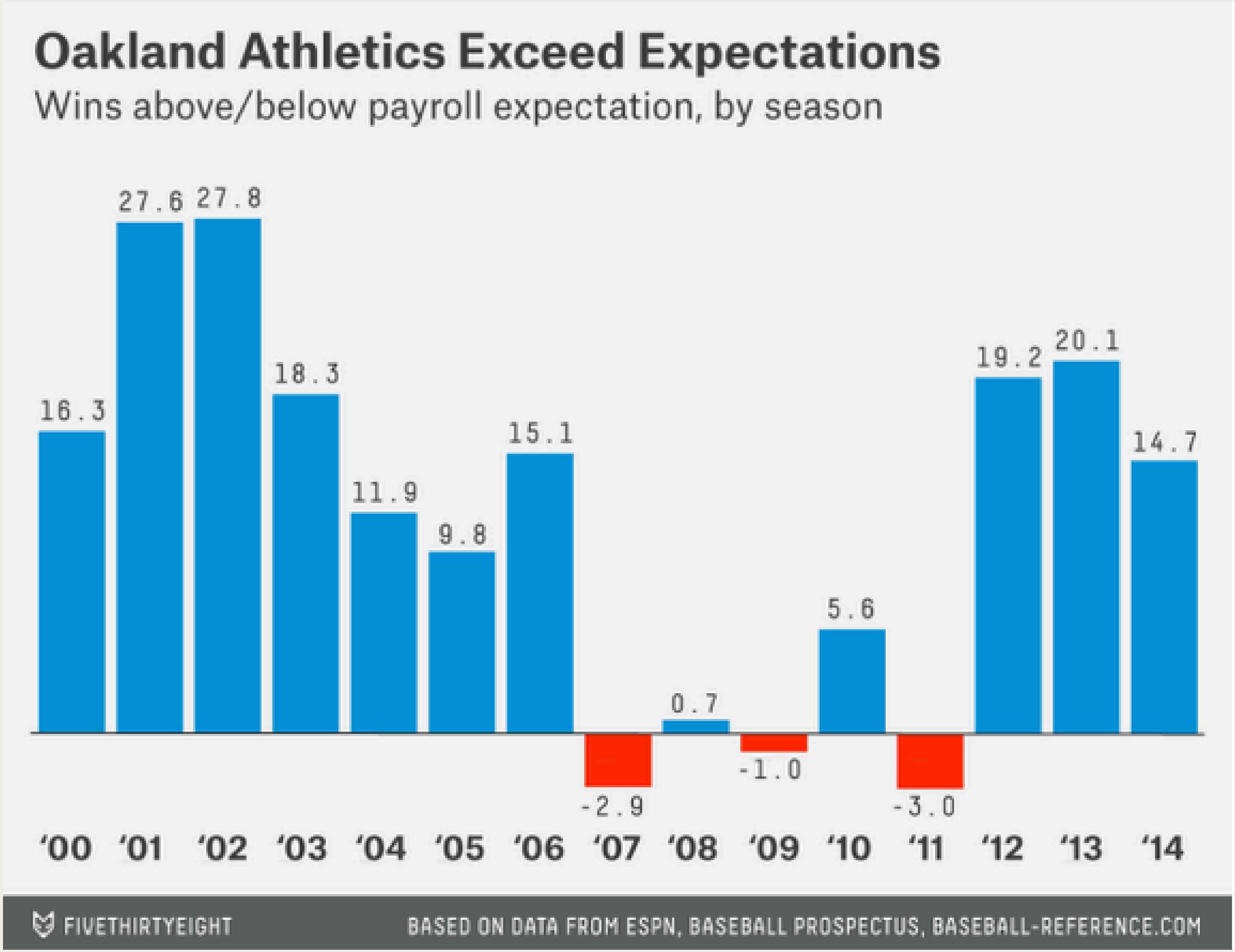


Alan Schwarz
The Numbers Game
(History of Sabremetrics)



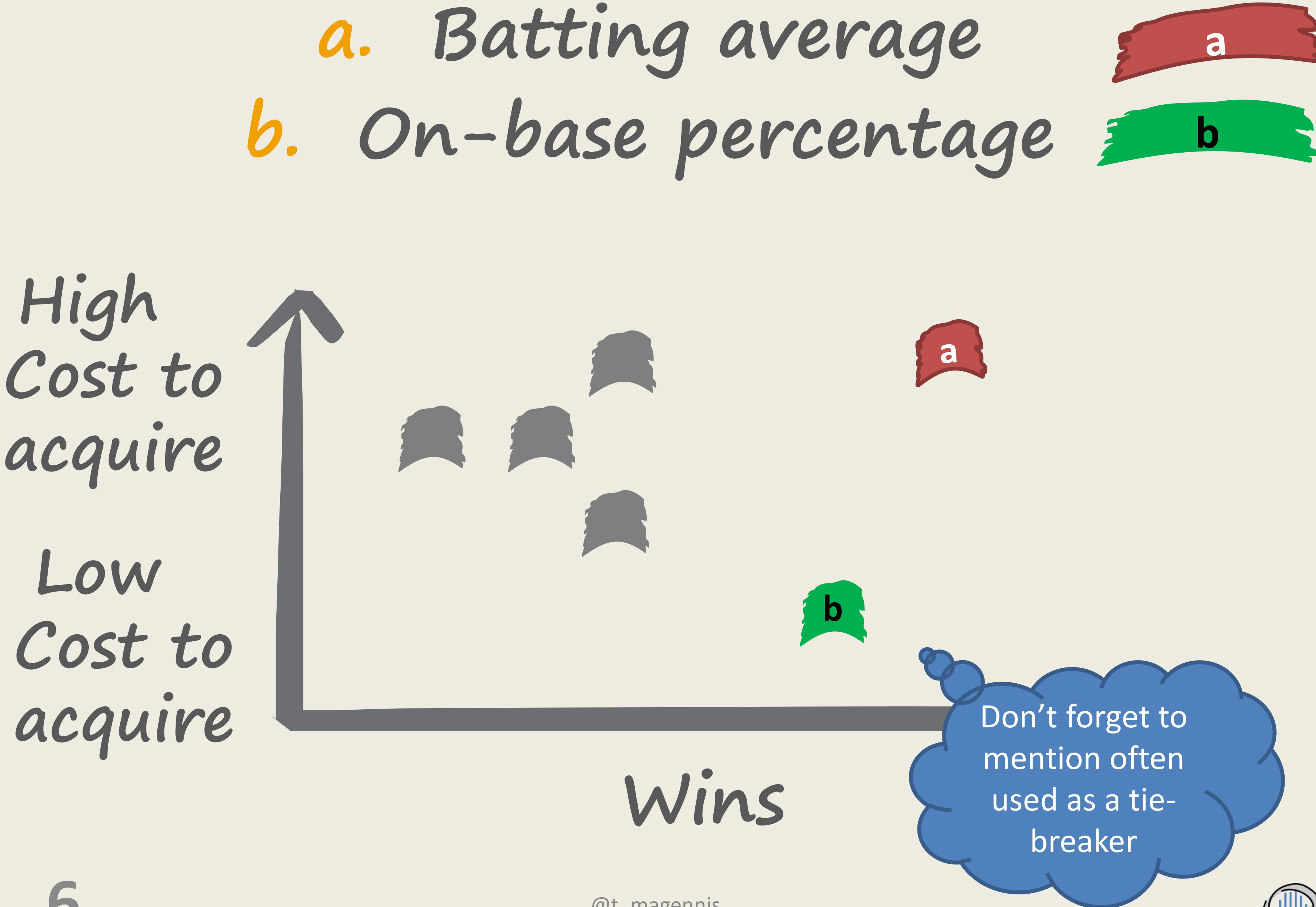
The Signal and the Noise: Why So Many Predictions Fail — but Some Don't


Season Win Percent vs. Relative Payroll
Standard deviations above/below league average (15 team bins)
Oakland
60%
55
50
45
40
Win percent
-3
-2
-1
0
1
2
3
Standard deviations from average payroll
FIVETHIRTYEIGHT
BASED ON DATA FROM ESPN, BASEBALL PROSPECTUS, BASEBALL-REFERENCE


Oakland Athletics Exceed Expectations
Wins above/below payroll expectation, by season
16.3
27.6
27.8
18.3
11.9
9.8
15.1
-2.9
0.7
-1.0
5.6
-3.0
19.2
20.1
14.7
'00
'01
'02
'03
'04
'05
'06
'07
'08
'09
'10
'11
'12
'13
'14
FIVETHIRTYEIGHT
BASED ON DATA FROM ESPN, BASEBALL PROSPECTUS, BASEBALL-REFERENCE.COM

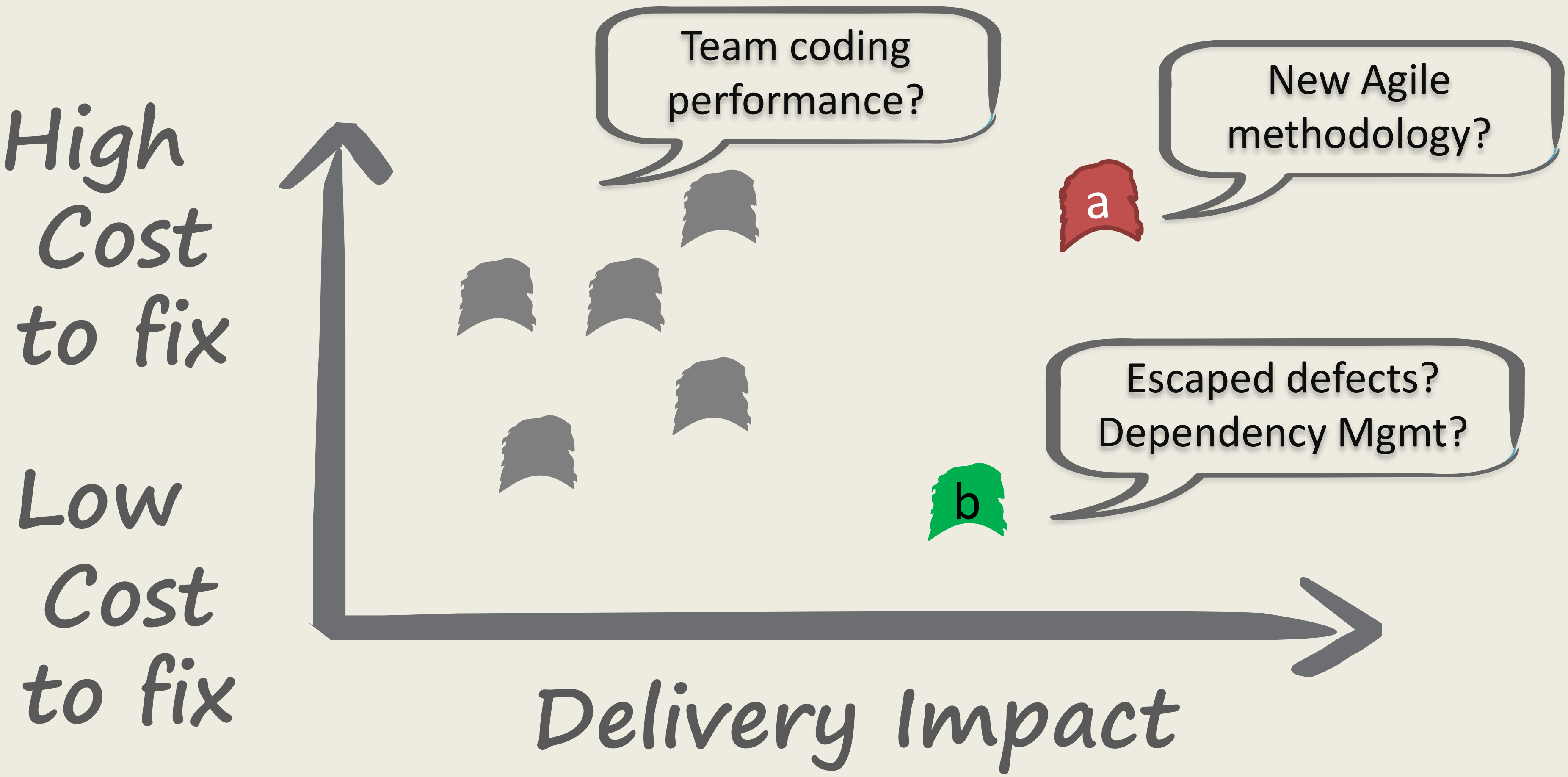
a. Batting average

b. On-base percentage



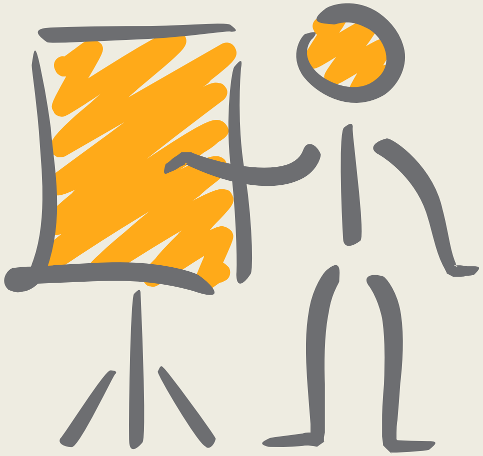
a. Latest Agile Framework a

b. Multi-team impacts b



Baseball goal: Win more games
Software goal: Deliver more value

Predictably deliver more value to customers



PICKING VALUABLE METRICS

The only metrics that entrepreneurs should invest energy in collecting are those that help them make decisions. Unfortunately, the majority of data available in off-the-shelf analytics packages are what I call Vanity Metrics. They might make you feel good, but they don't offer clear guidance for what to do.

by Eric Ries,
The Lean Startup

Predictive Power and Better Decisions

- Observing historical data (metrics) may be interesting, but the predictive power of historical data should be the focus
- If a metric doesn't offer predictive power, then capturing that metric is waste
- Decisions based on historical data are predictions
  - These decisions have un-certainty
  - We can (and should) compare the eventual reality against our predictions and learn

Good Metrics

- Lead to decisions
- Within teams' influence
- Gaming leads to "good"
- Have a credible story
- Are linked to strategy

- **Trend or distribution based**
- **Leading indicators**

Bad Metrics

- Just convenient to capture
- Linked to reputation
- Gaming leads to "bad"
- Just to change "my" behavior
- Don't link to strategy

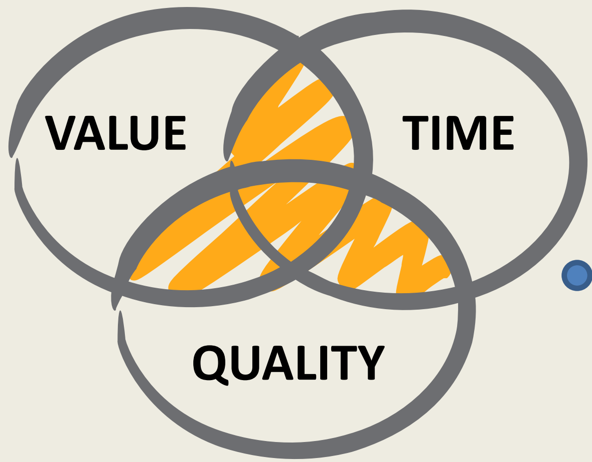
- People targeting
- Trailing indicators

Google for "Seven Deadly Sins of Agile Metrics"
by Larry Maccherone for more ideas on good and bad metrics.

Balanced and Valuable Metrics

1. Cost of Delay ($)
2. Alignment to Strategy
3. Number of Experiments

1. Throughput / velocity
2. Key person dependency score
3. Risk uncertainty



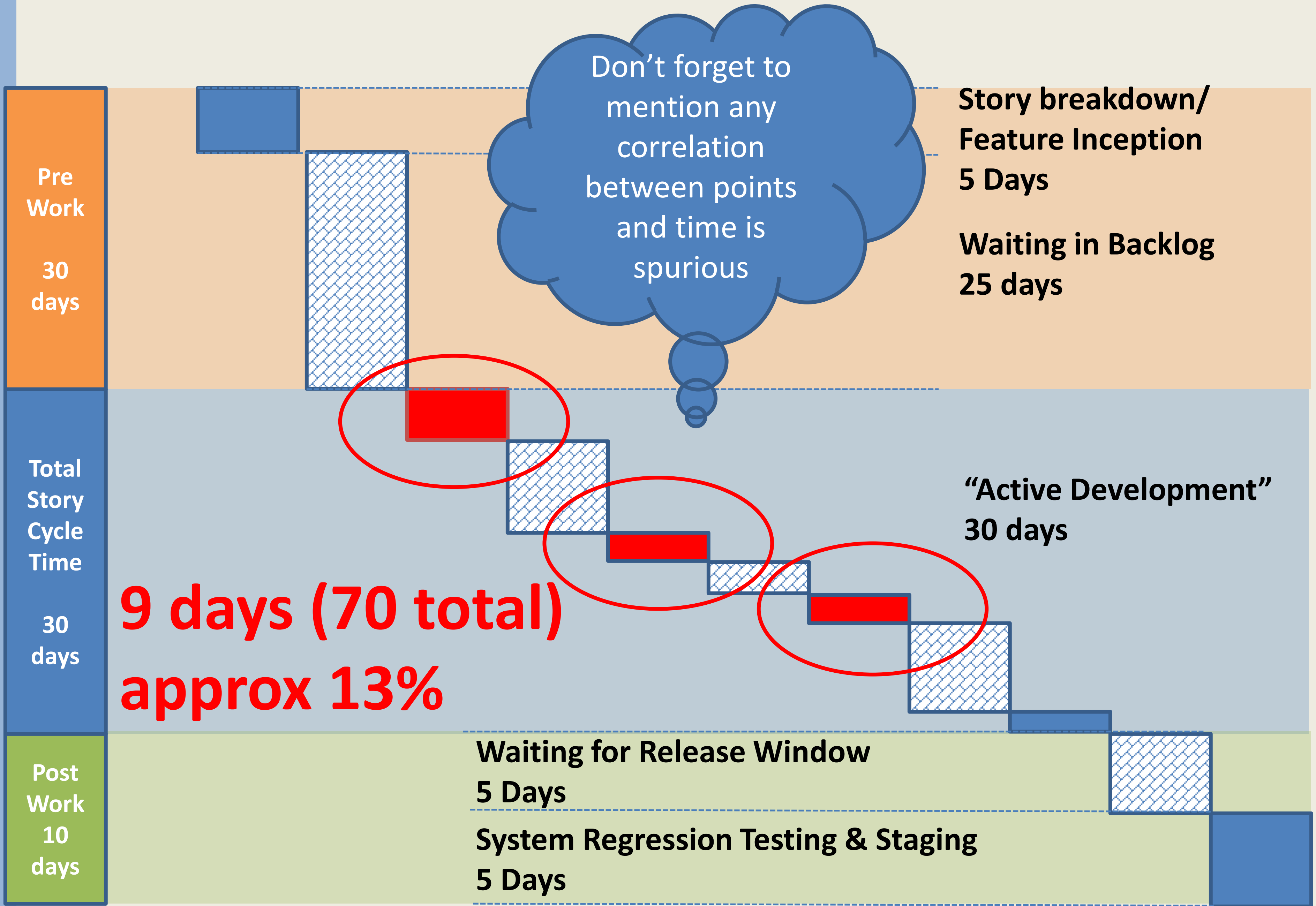
Don't forget to mention its easy to move one, but not easy not to impact the others

1. Customer Impacting Defect Rate
2. Production Releases without rollback
3. Process Experimentation Rate
   (# improvement / total stories per sprint)

	Sprint 1	Sprint 2	Sprint 3	Sprint 4	Sprint 5
Velocity	16 pts	72 pts	21 pts	19 pts	37 pts
Throughput	7 cards	9 cards	9 cards	9 cards	7 cards

Velocity: 16-72 pts, Throughput: 8 +/- 1

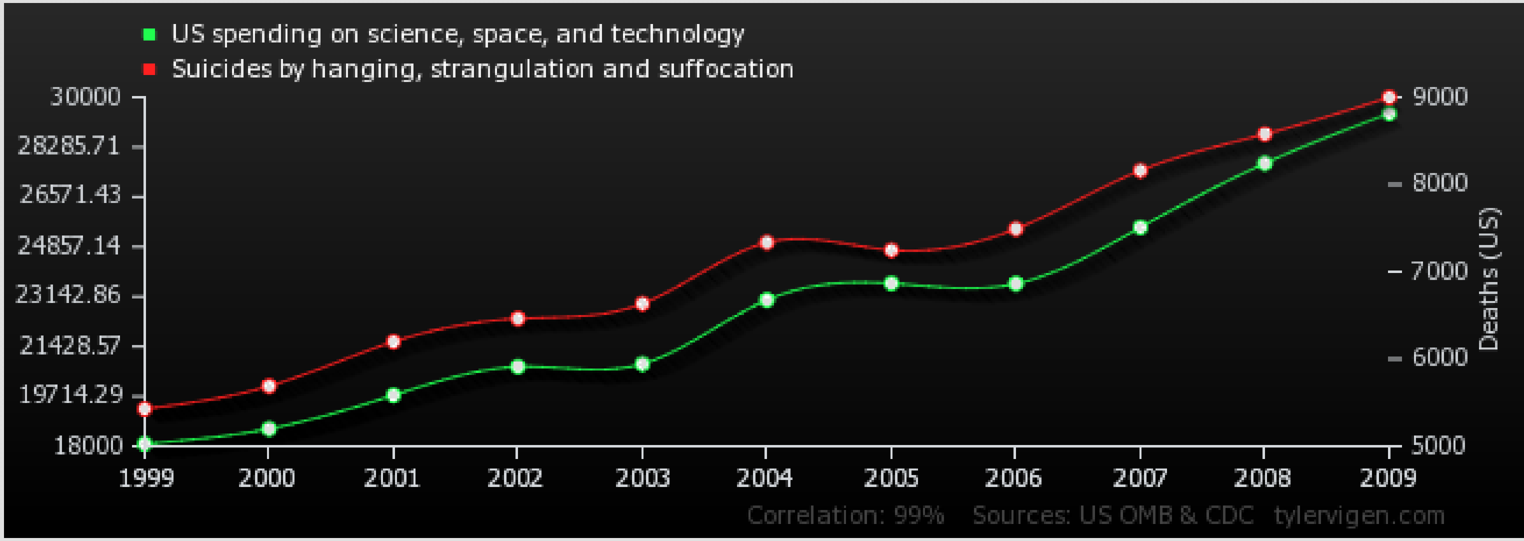
Source and with thanks: Jim Benson (Modus Cooperandi)
http://moduscooperandi.com/blog/estimation-requires-attention/


Don't forget to mention any correlation between points and time is spurious
Pre Work
30 days
Total Story Cycle Time
30 days
Post Work
10 days
Story breakdown/ Feature Inception
5 Days
Waiting in Backlog
25 days
"Active Development"
30 days
9 days (70 total)
approx 13%
Waiting for Release Window
5 Days
System Regression Testing & Staging
5 Days

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



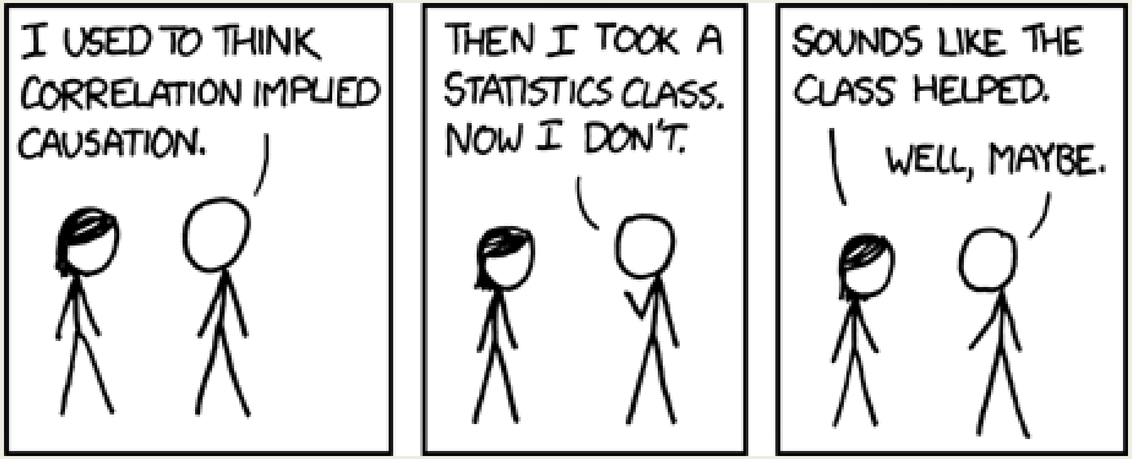
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	2[illegible]	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,4[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]578	9,00[illegible]

Correlation: 0.992082

Correlation: Follows the same trend and relative rate of change independent of absolute magnitude

Spurious Correlations: http:/[illegible]

Leading Indicators
Correlation != Causation



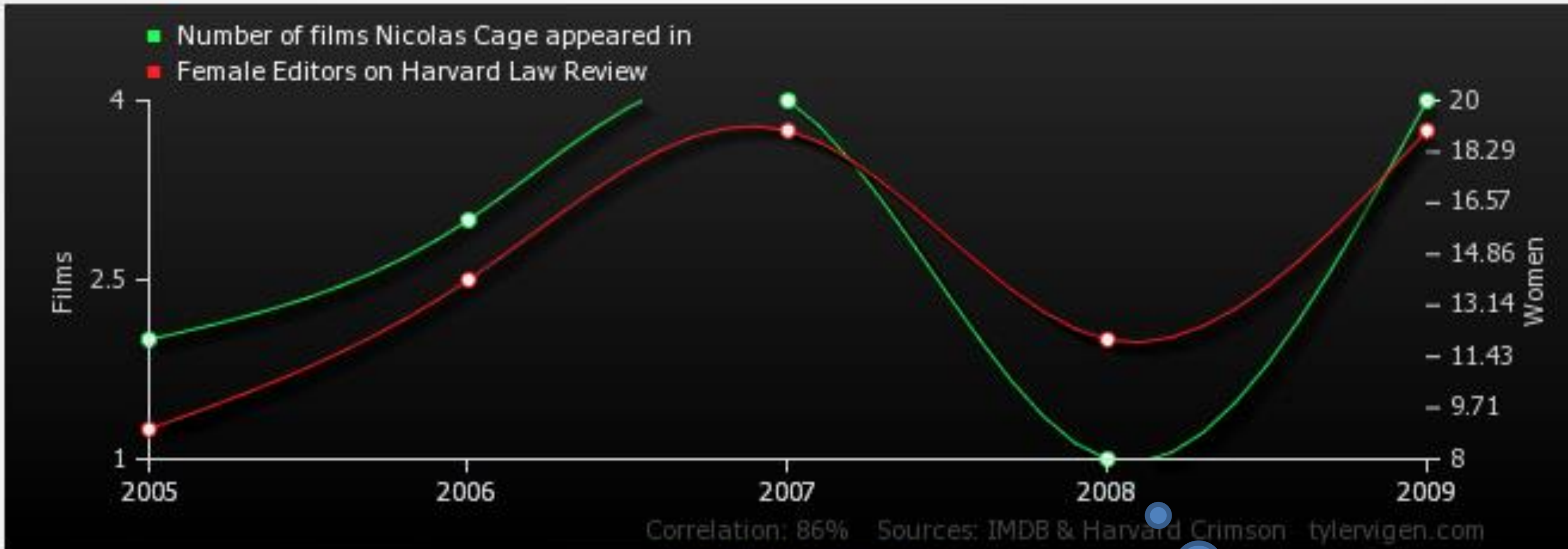
- Criteria for causality
  - The cause precedes the effect in sequence
  - The cause and effect are empirically correlated and have a plausible interaction
  - The correlations is not spurious (short period)

Sources: Modified by me for brevity based on: Kan,2003 pp80 and Babbie, 1986

Number of films Nicolas Cage appeared in

correlates with

Female Editors on Harvard Law Review



Upload this image to imgur

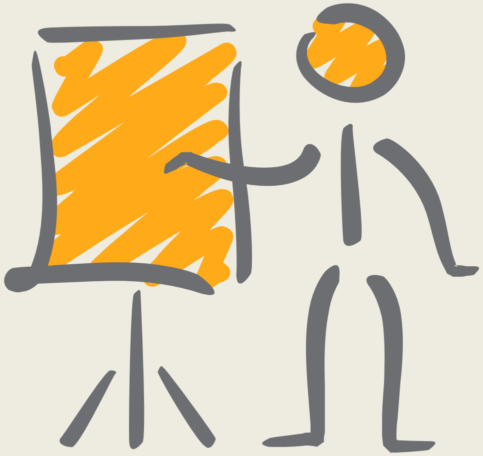
	2005	2006	2007	2008	2009
Number of films Nicolas Cage appeared in *Films (IMDB)*	2	3	4	1	4
Female Editors on Harvard Law Review *Women (Harvard Crimson)*	9	14	19	12	[illegible]

Correlation: 0.855447

Don't forget to mention leading indicator except for plausibility

@t_magennis

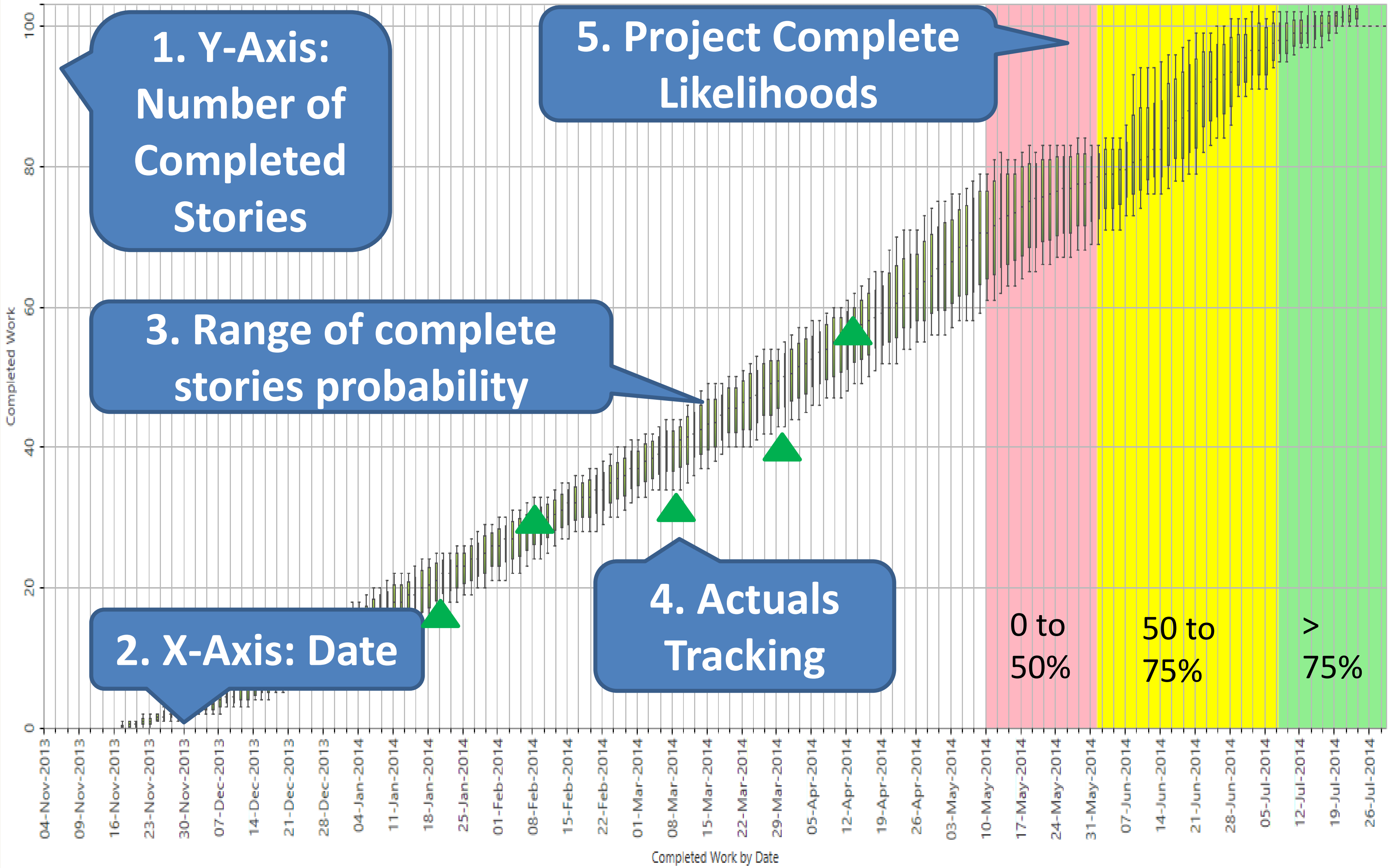
Spurious Correlations: http://tylervigen.com/



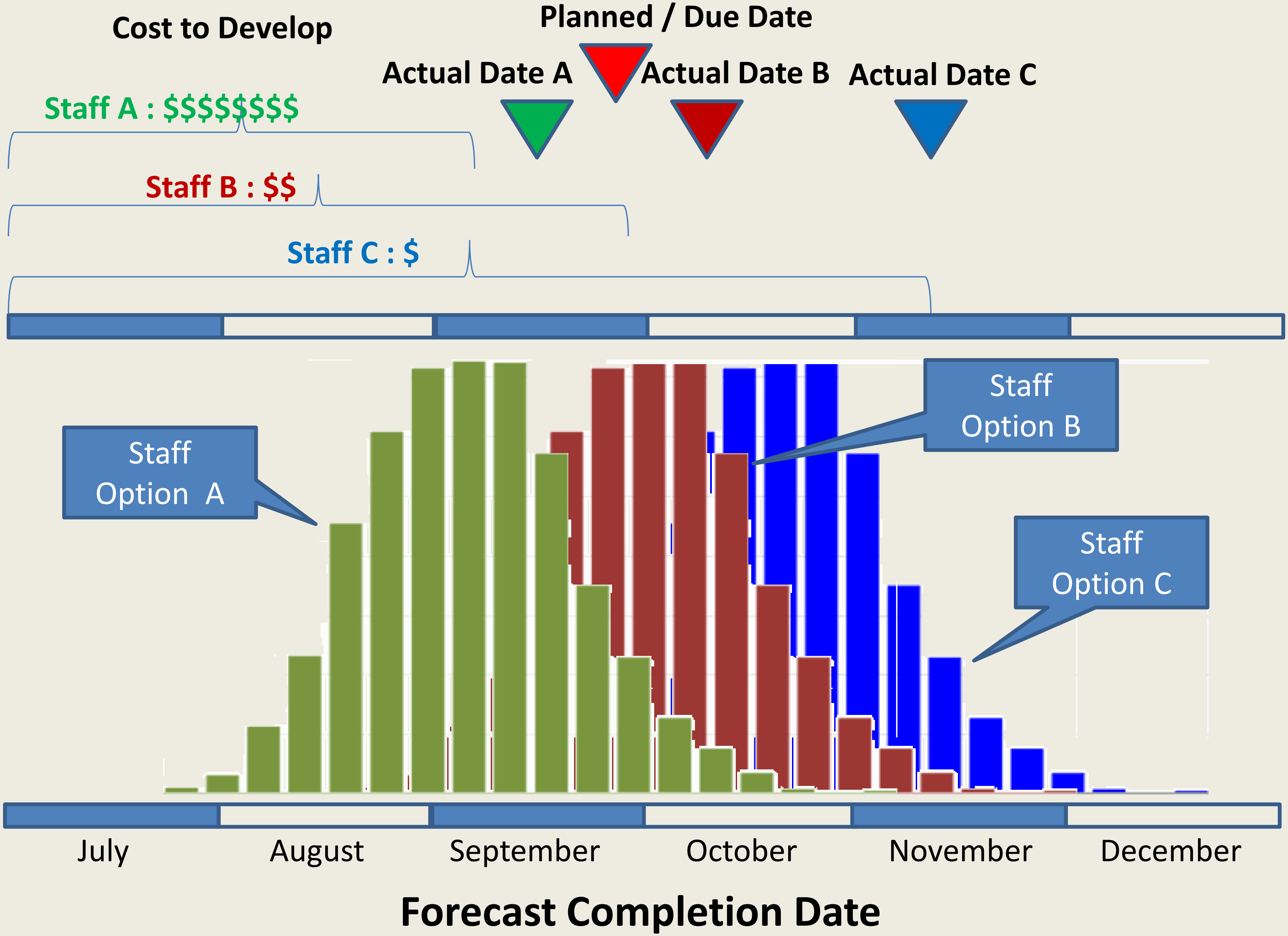
MODELING – A QUICK INTRO

You need the model to spot when reality diverges from expectation

Once the model reflects reality (showing predictive power) you can run experiments on the model before real-life

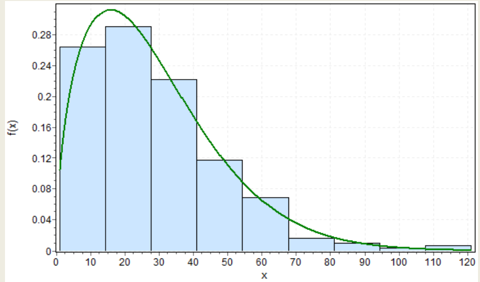


PS. ScrumSim and KanbanSim is free, focusedobjective.com


Cost to Develop
Planned / Due Date
Actual Date A
Actual Date B
Actual Date C
Staff A : $$$$$$$$
Staff B : $$
Staff C : $
Staff Option A
Staff Option B
Staff Option C
July
August
September
October
November
December
Forecast Completion Date

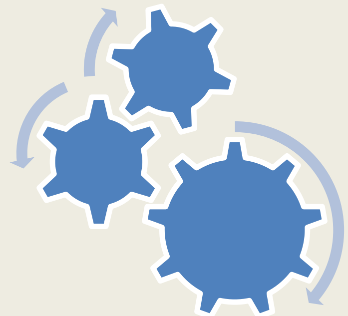
6. Team / Project Phases

1. Historical Cycle Time (in context)

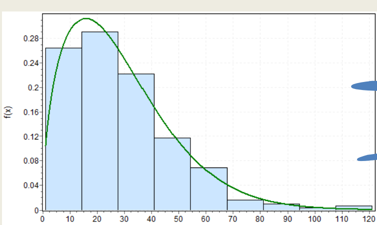


2. Planned Resources/ Effort

Monte Carlo Process = Process to Combine Multiple Uncertain Measurements / Estimates



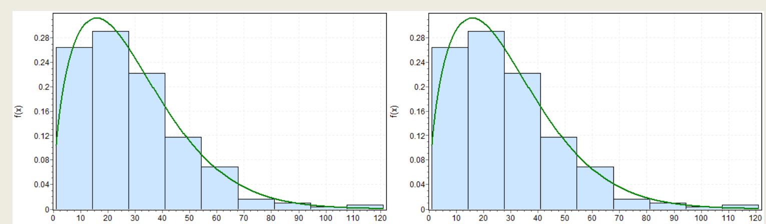
4. Historical Scope Creep Rate



(optional)

3. The Work (Backlog)

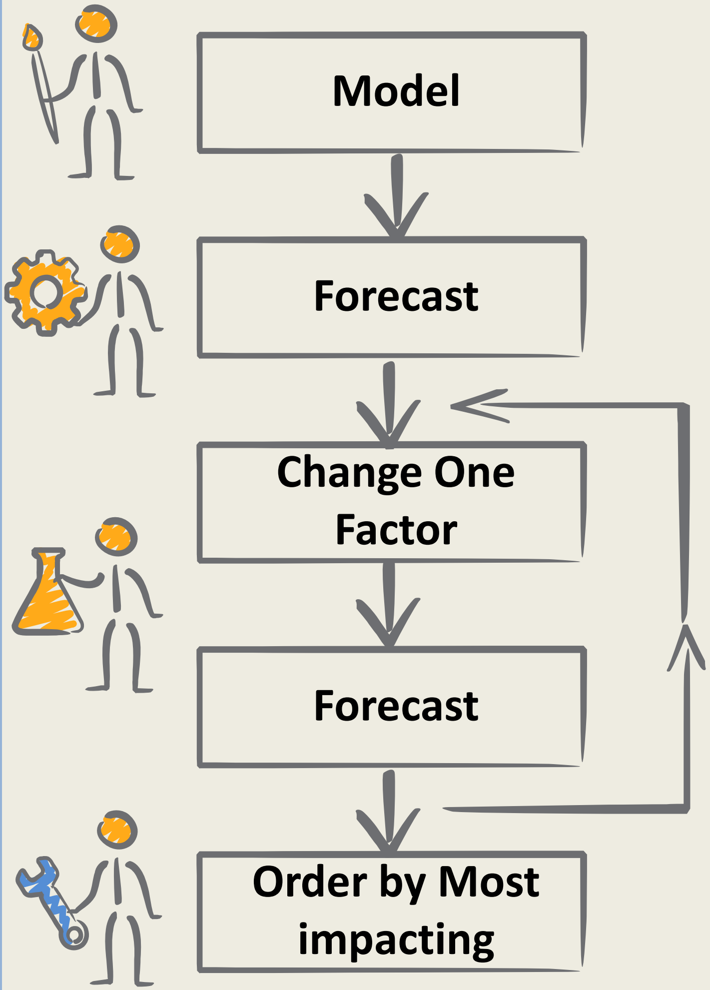
Backlog
Feature 1
Feature 2
Feature 3



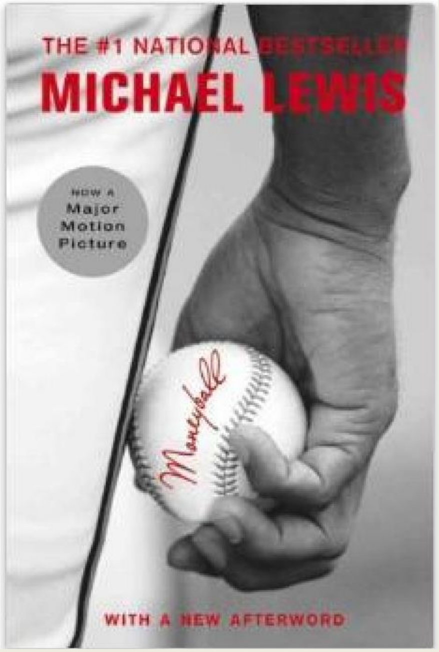
5. Historical Defect Rate and Cycle Times

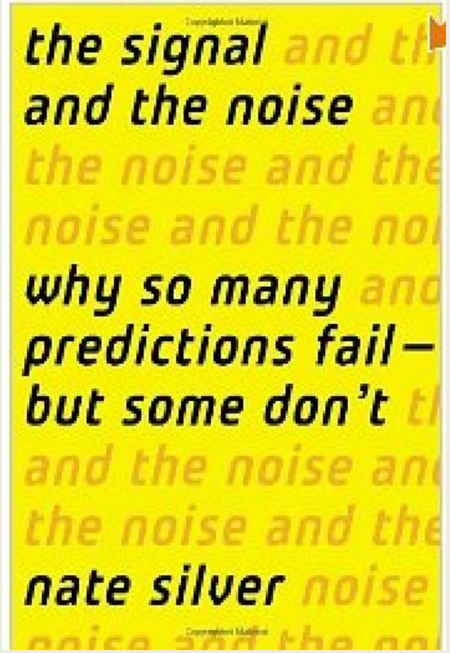
(optional)

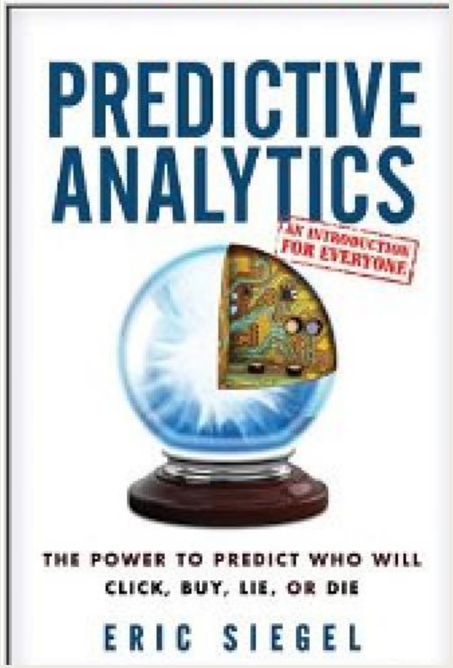
Sensitivity Testing



Alter one factor in a model at a time and forecast. Order the factors from most impacting to the least on forecast outcome.



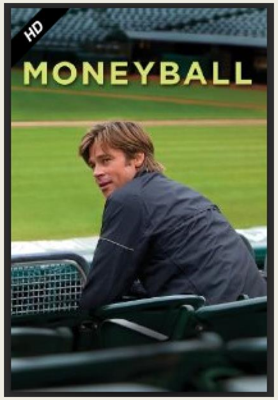


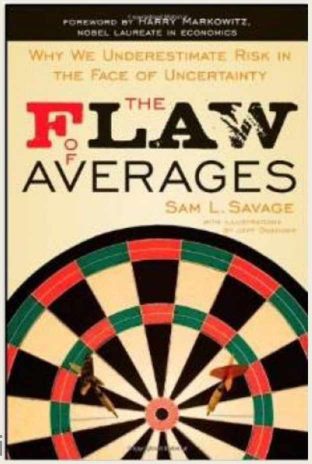


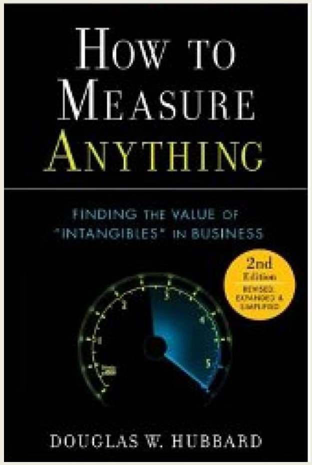
Moneyball: The Art of Winning an Unfair Game

The Signal and the Noise: Why So Many Predictions Fail — but Some Don't

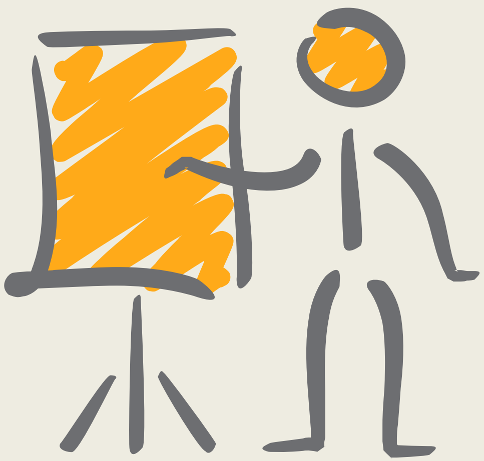
Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die







_magenni

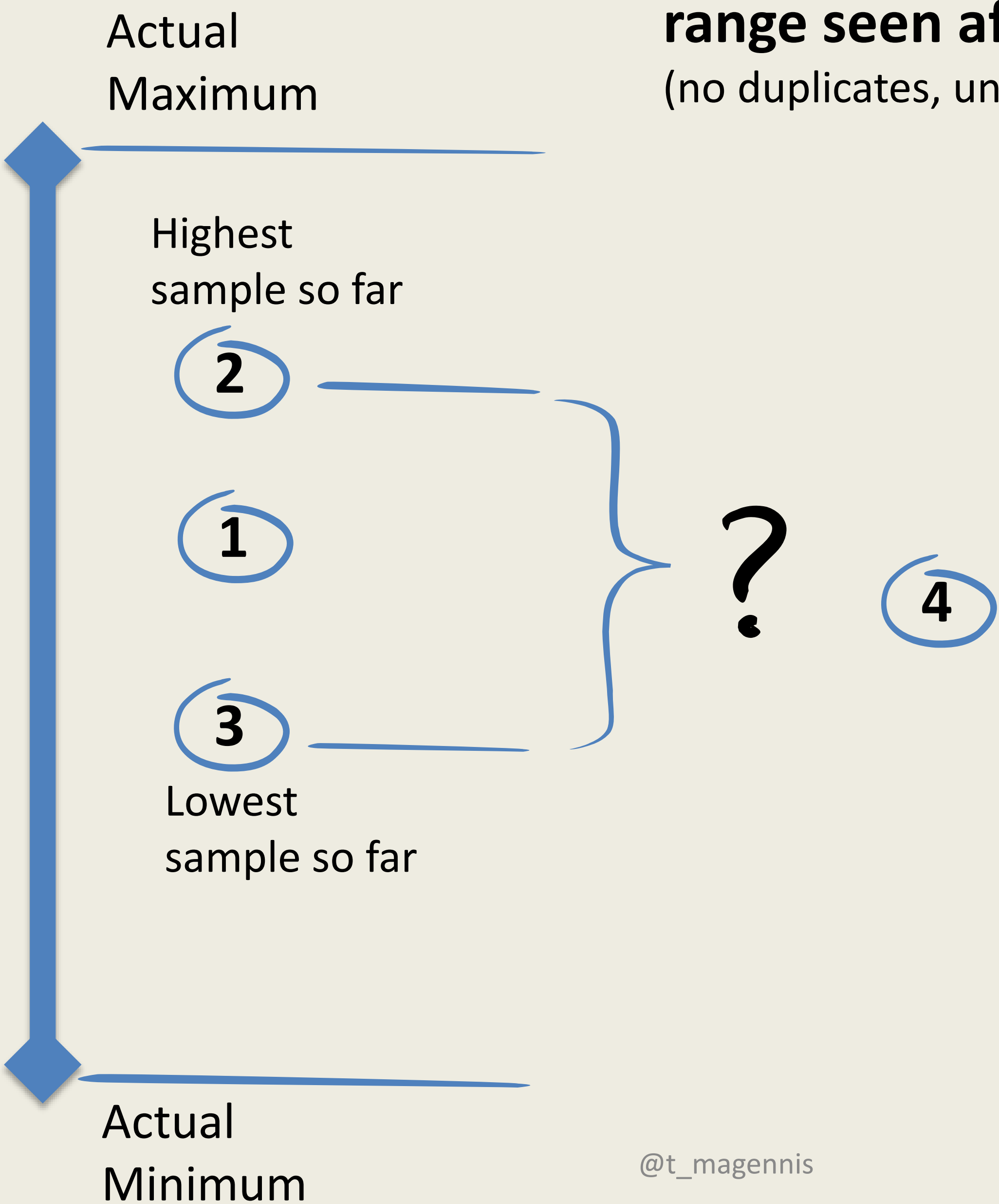


FUN WITH UNCERTAINTY

How Many Samples Are Required to Determine Range?

Q. On average, what is the chance of the 4th sample being between the range seen after 3 random samples?

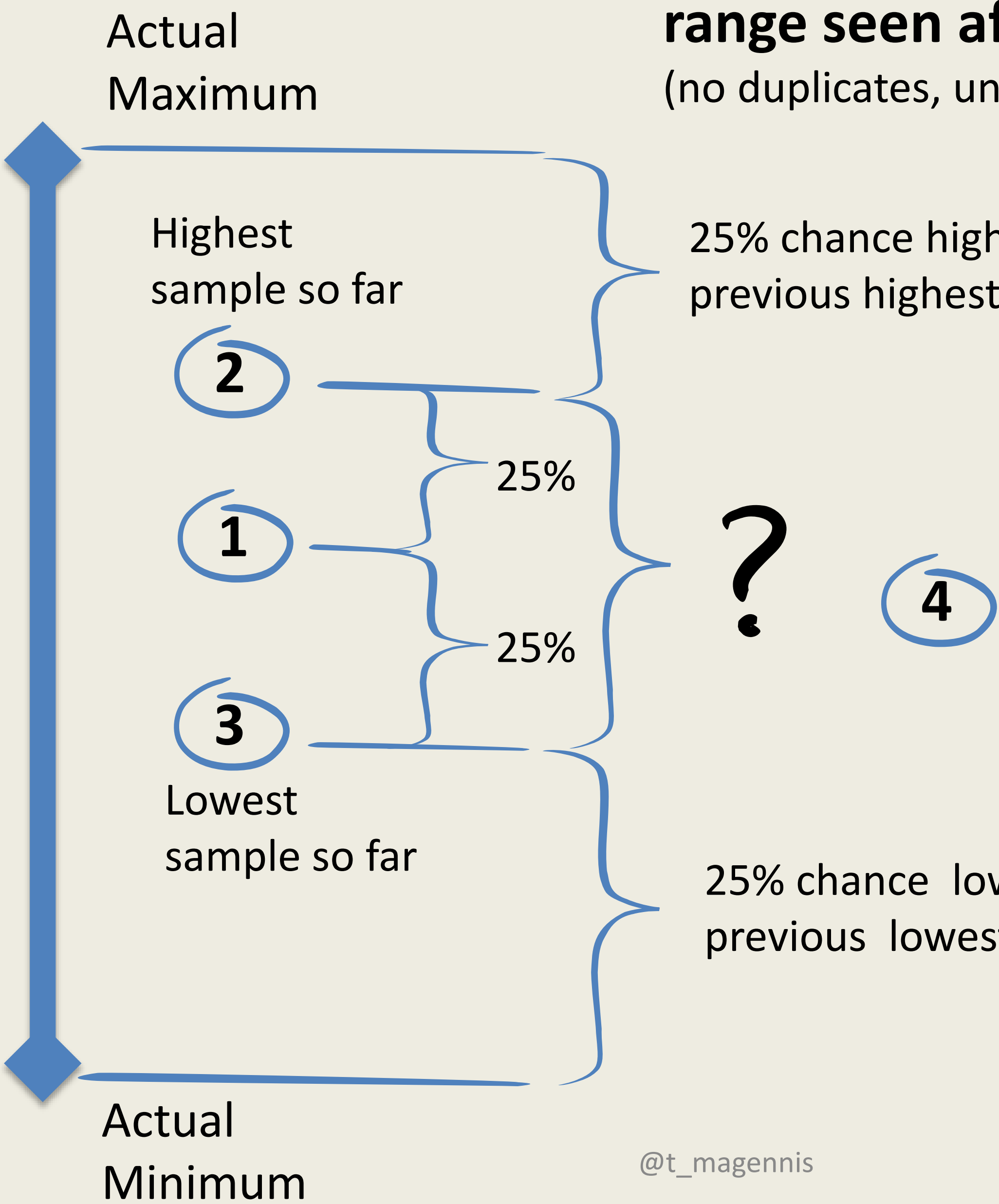
(no duplicates, uniform distribution)



A. ?

Q. On average, what is the chance of the 4th sample being between the range seen after 3 random samples?

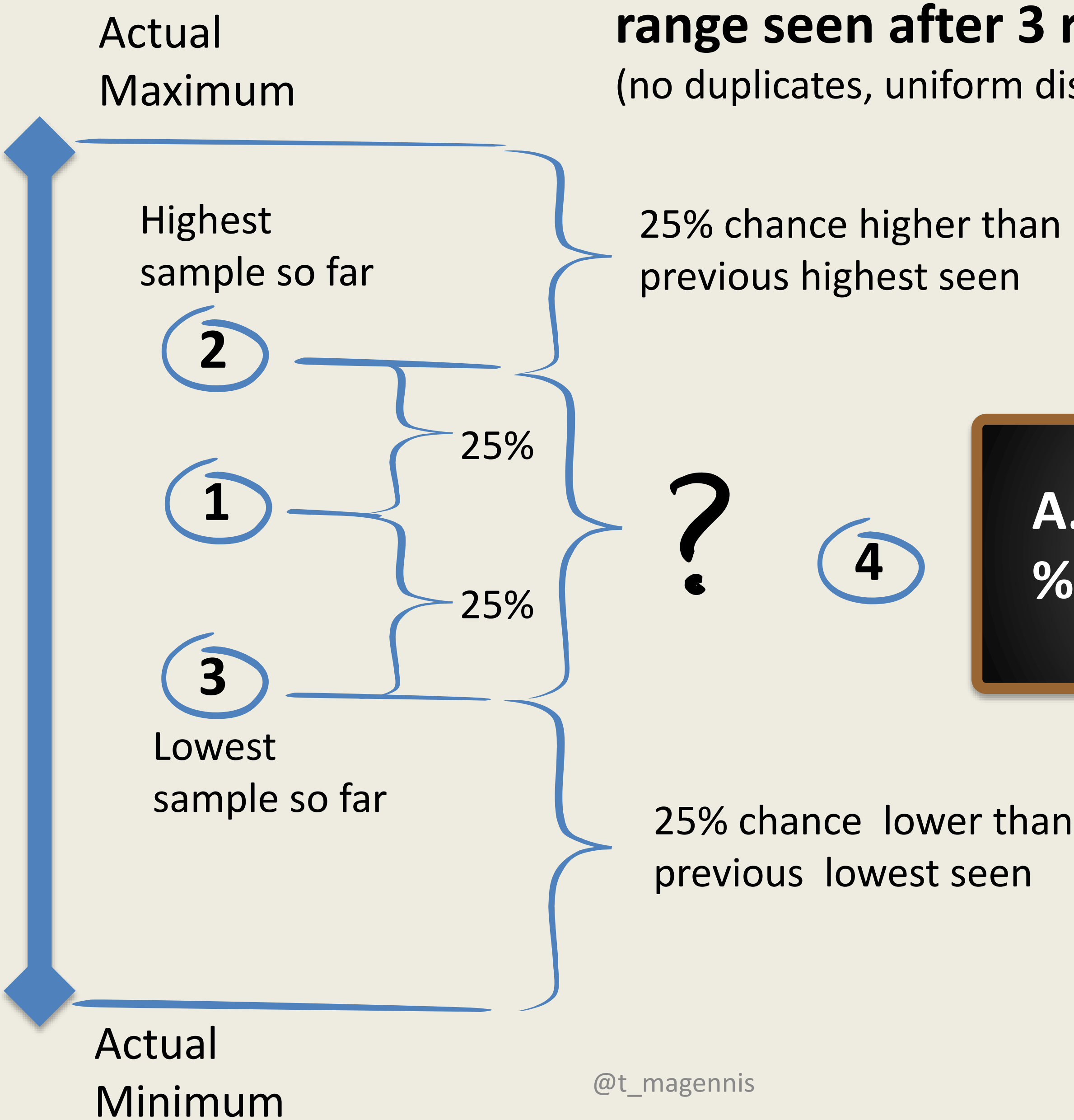
(no duplicates, uniform distribution)



A. ?

Q. On average, what is the chance of the 4th sample being between the range seen after 3 random samples?

(no duplicates, uniform distribution)

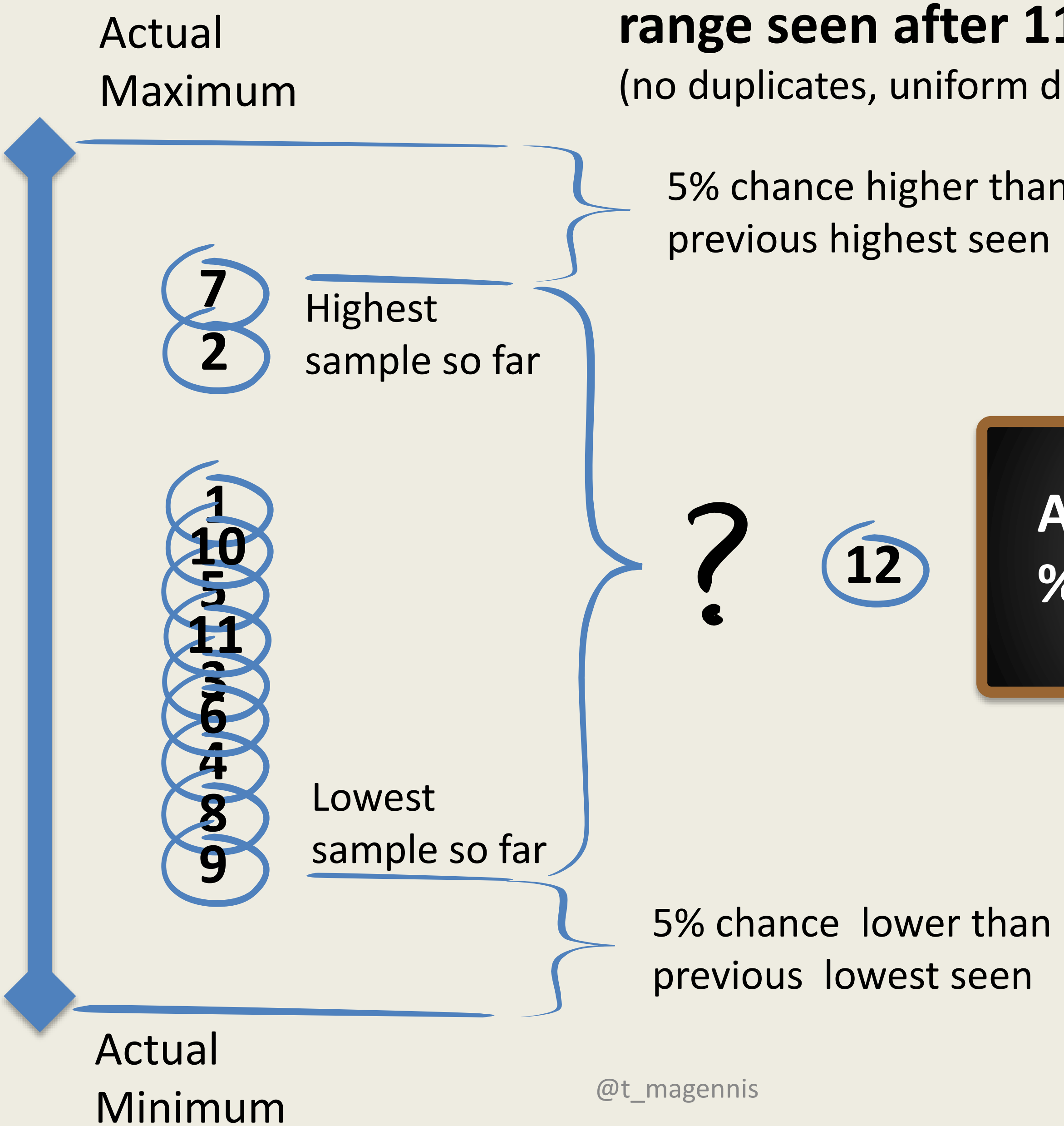


A. 50%

% = 1 - (1 / n – 1)

Q. On average, what is the chance of the 12th sample being between the range seen after 11 random samples?

(no duplicates, uniform distribution)



A. 90%

% = 1 - (1 / n – 1)

@t_magennis

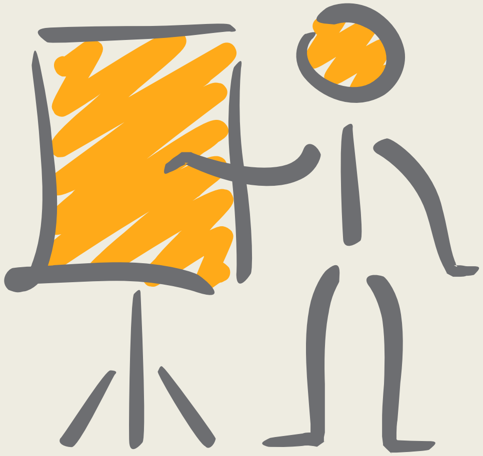
Rules of Thumb

- "n" = number of prior samples
- A calculates % chance next sample in previous range
- B is an approximation for low range discrete values

n	A (n-1)/(n+1)	B 1/(n-1)
2	33%	0%
3	50%	50%
4	60%	67%
5	67%	75%
6	71%	80%
7	75%	83%
8	78%	86%
9	80%	88%
10	82%	89%
11	83%	90%
12	85%	91%
13	86%	92%
14	87%	92%
15	88%	93%
16	88%	93%
17	89%	94%
18	89%	94%
19	90%	94%
20	90%	95%
21	91%	95%
22	91%	95%
23	92%	95%
24	92%	96%
25	92%	96%
26	93%	96%
27	93%	96%
28	93%	96%
29	93%	96%
30	94%	97%

Why do I need more samples?

- Samples aren't random or independent
- Some samples are erroneous and dropped
- Uneven density of value distribution
  - Most common: Fewer expected high values means more samples needed to find the upper values
- While detecting the range requires few estimates, detecting the shape needs many



Do we have to break down EVERY epic to estimate story counts?

CASE STUDY: ESTIMATING TOTAL STORY COUNT

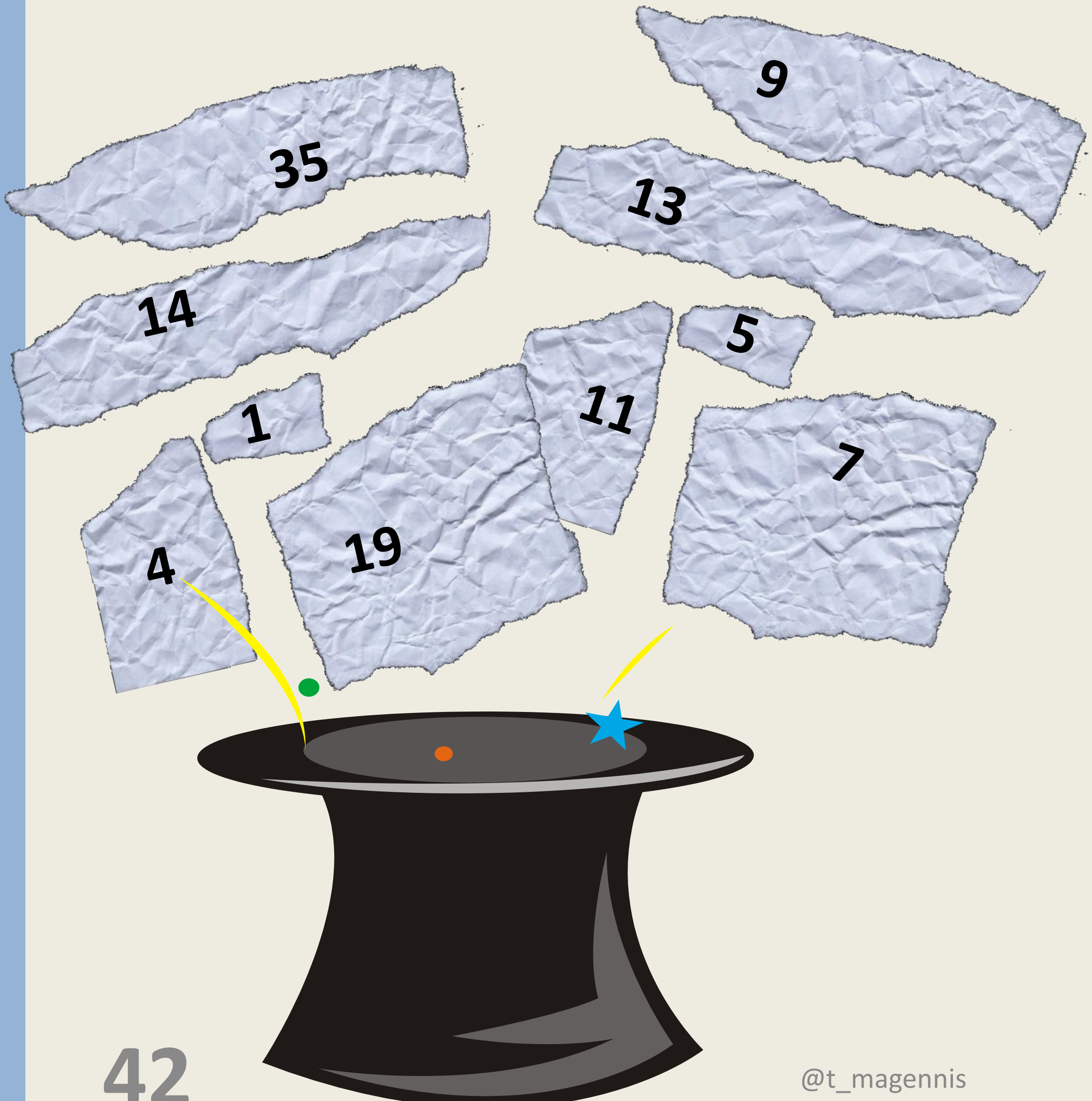
Problem: Getting a high level time and cost estimate for proposed business strategy

time and costs

Approach: Randomly sample epics from the 328 proposed and perform story breakdown. Then use throughput history to estimate time and costs

Sample with replacement

Remember to put the piece of paper back in after each draw!



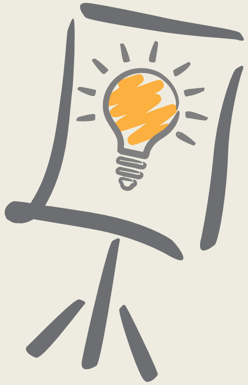
	Trial 1	Trial 2		Trial 100
		1		35
		4		19
		7		5
		5		13
		11		11
Sum:	51	28	…	83

Epic Breakdown – Sample Count

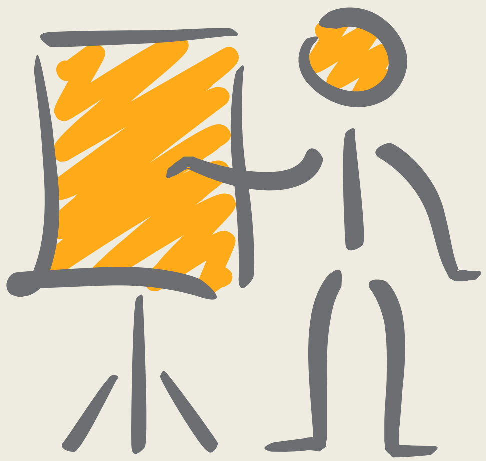
Facilitated by well known consulting company, team performed story breakdown (counts) of epics.
48 (out of 328) epics were analyzed.

Actual Sum
262

Process	50% CI	75% CI	95% CI
MC 48 samples	261	282	315
MC 24 samples	236	257	292
MC 12 samples	223	239	266
MC 6 samples	232	247	268



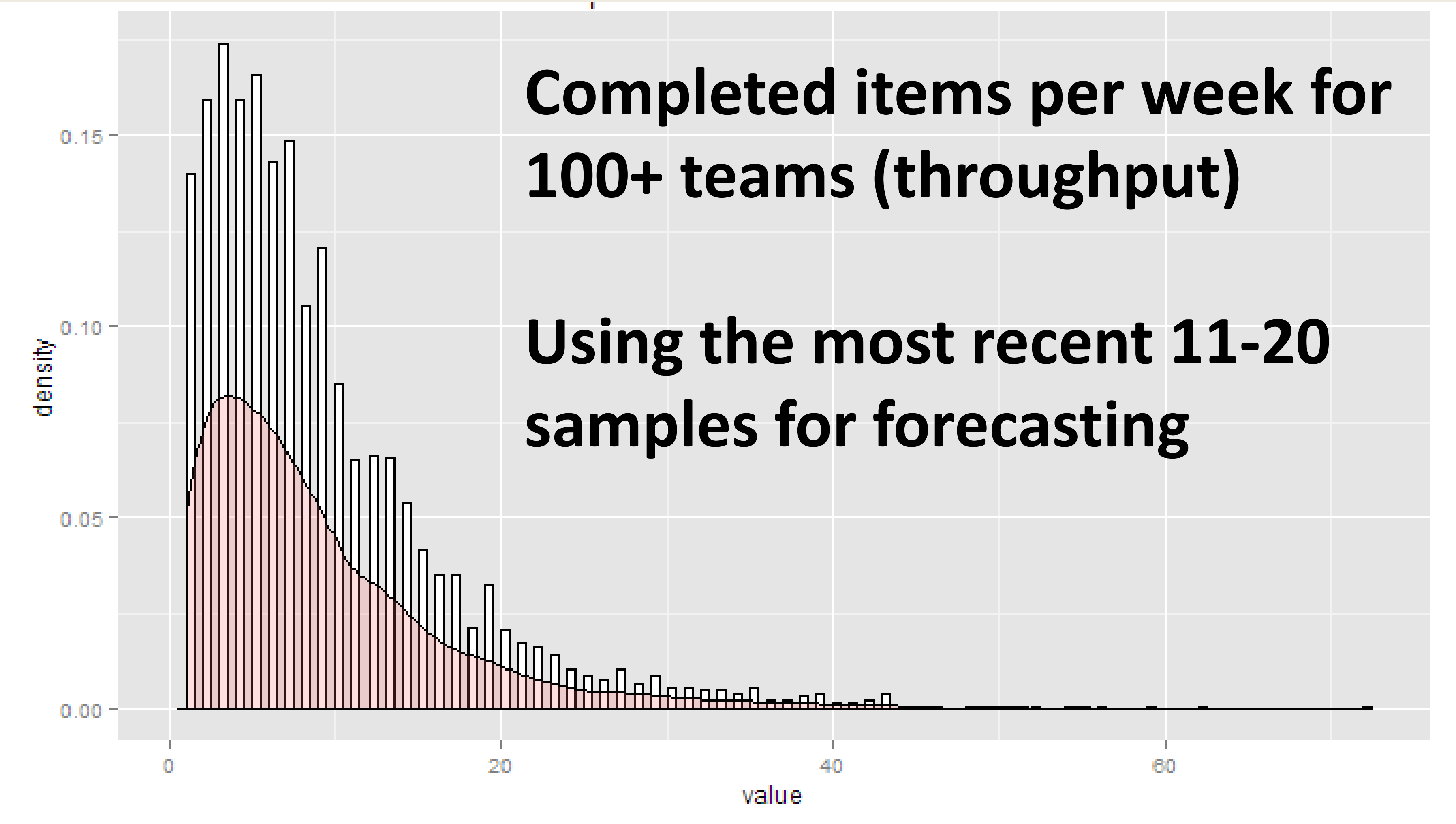
Example: Spreadsheet Analysis



CASE STUDY: TEAM THROUGHPUT PLANNING AND FORECASTING

Problem: Teams unsure how to plan team constraints during cross-team planning. Teams spend considerable time estimating proposed work.

Approach: Give the teams a way to forecast throughput based on historical performance.



Evidence of data quality is well formed distribution shape

Example:
Throughput Forecasting Tool
(OK, its just a spreadsheet)

Work Item Throughput Forecasting

Tip: Any cell with a background of orange can b

Step 1: Choose your team - all analysis is performed on the historical work item completion rate per

Team 3 <-- Tip: You must set this FIRST! This is the team you will be performing analysis on!

How many weeks will it take Team 3 to complete a number of Work Items?

Use this table to see how many work items your team will complete in various weeks. The first column (50%, 75%, 85%) is how conf amount of items in the given period based on variability of your teams prior throughput completion rate performance.

	4 weeks	8 weeks	12 weeks	16 weeks	20 weeks	24 weeks	28 weeks	32 weeks	36 weeks	40 we
85%	12	28	45	63	80	98	115	133	152	16
75%	14	31	49	67	85	103	120	138	157	17
50%	18	37	56	75	92	112	131	149	168	18

Confidence Level of forecast

Forecasted number of work items complete in 12 weeks

Or, How many work items will be completed by Team 3 in a specified number of weeks?

Number of weeks	12	(weeks)
Reserve capacity (%) *	20	(percent)

RESULT:Forecast Total Completed Work Items in 12 week

85%	37	<--- Tip: This is your forecasted # wor
75%	39	You can enter other confi
50%	44	

Throughput Forecasting Tool v1 - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER ADD-INS LOAD TEST POWERPIVOT TEAM Troy Mag...

A1 Team 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8	Team 9	Team 10	Team 11	Team 12	Team 13	Team 14	Team 15
2	6	16	8	7	6	6	11	6	20	16	5	11	20	11	7
3	3	3	1	1	7	5	7	13	12	3	3	2	2	1	6
4	20	26	5	16	8	5	9	6	17	10	3	10	20	14	9
5	8	16	3	12	3	7	4	6	17	6	12	9	15	13	5
6	29	17	2	16	13	6	5	10	20	19	6	28	26	4	8
7	16	31	8	7	15	7	7	8	13	11	2	21	22	7	9
8	5	21	2	1	5	3	18	6	23	17	4	27	25	11	12
9	13	29	2	12	14	3	5	7	49	17	4	12	15	12	8
10	16	19	7	8	16	8	9	11	23	15	5	25	33	7	5
11	24	21	3	4	17	3	10	15	18	9	4	35	36	9	11
12	10	15	5	8	19	4	9	10	23	11	19	19	9	7	14
13	14	19	4	9	17	2	17	10	17	8	4	19	25	8	7
14	21	15	4	3	19	6	14	13	12	5	4	14	10	9	6
15	12	17	5	11	22	4	5	9	15	13	8	6	10	12	8
16	20	3	3	1	6	18	19	4	12	5	5	19	21	13	7
17	14	22	2	12	14	1	12	10	3	7	3	4	15	9	5
18	13	19	12	3	11	2	13	3	13	8	5	12	18	8	9
19	10	12	5	2	18	4	7	9	7	6	7	15	22	7	4
20	12	19	6	11	3	28	23	3	10	10	8	10	19	12	11
21	4	11	8	6	7	2	22	9	6	10	7	15	14	9	11
22	30	26	1	2	14	39	14	4	15	8	7	8	16	16	8
23	15	18	12	16	12	26	24	8	1	5	5	7	7	6	7

Team Selection and Forecast | **Throughput History Data**

READY 100%

Throughput Forecasting Tool v1 - Excel

FILE HOME INSERT PAGE LAYOU FORMULAS DATA REVIEW VIEW DEVELOPER ADD-INS LOAD TEST POWERPIVO TEAM Troy Mag...

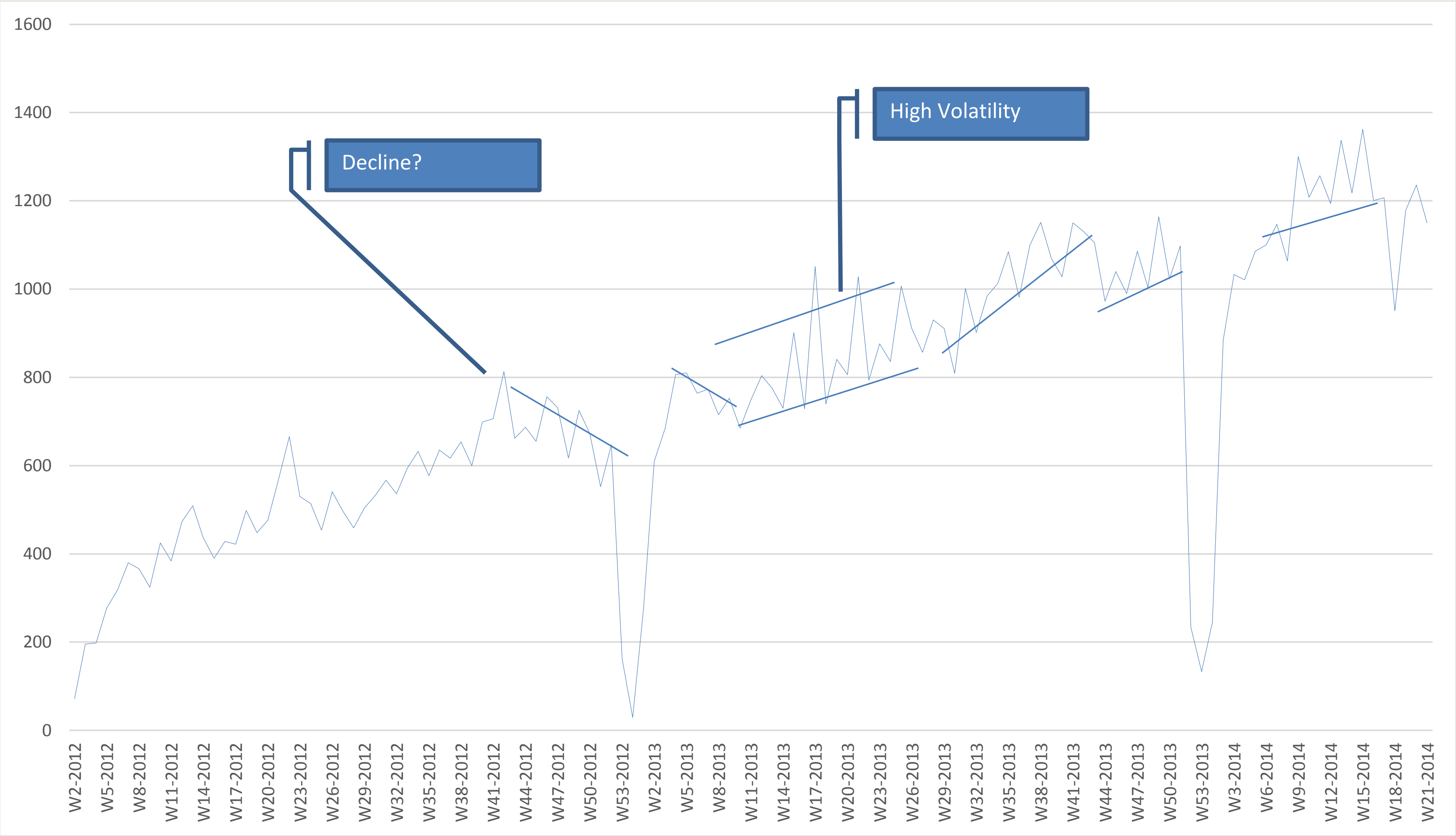
Q4

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2		Avg	Max	5%	10%	25%	50%	75%	80%	85%	90%	95%	Min	
3		45.288	74	59.2	56	50.4	44.8	39.8	37.6	36	34.4	32	27.2	
4		PLEASE DO NOT TOUCH THIS SHEET, THIS IS FOR CALCULATION ONLY!												
5	Trial #	Sum n	1	2	3	4	5	6	7	8	9	10	11	12
6	1	38.4	6.4	1.6	6.4	3.2	1.6	2.4	1.6	2.4	2.4	4	4	2.4
7	2	34.4	4	2.4	1.6	4	5.6	1.6	1.6	1.6	4	1.6	4.8	1.6
8	3	58.4	6.4	4	6.4	6.4	2.4	4	3.2	0.8	6.4	9.6	6.4	2.4
9	4	49.6	2.4	9.6	6.4	4	0.8	3.2	1.6	1.6	4.8	1.6	4	9.6
10	5	48.8	9.6	0.8	5.6	3.2	2.4	4	4	4	3.2	1.6	0.8	9.6
11	6	58.4	6.4	5.6	2.4	4	1.6	0.8	9.6	4.8	5.6	1.6	9.6	6.4
12	7	47.2	4	0.8	2.4	6.4	3.2	4.8	4	4.8	4	6.4	0.8	5.6
13	8	44.8	1.6	0.8	0.8	9.6	4.8	9.6	6.4	2.4	1.6	4	1.6	1.6
14	9	45.6	4	2.4	1.6	1.6	4.8	9.6	6.4	2.4	3.2	4	1.6	4
15	10	32	1.6	0.8	4	1.6	1.6	6.4	3.2	3.2	4	2.4	1.6	1.6
16	11	48.8	4.8	4.8	4	9.6	3.2	2.4	4	4	6.4	2.4	2.4	0.8
17	12	49.6	5.6	5.6	6.4	4	6.4	1.6	3.2	4	4.8	1.6	3.2	3.2
18	13	52	4	2.4	9.6	9.6	6.4	3.2	2.4	3.2	1.6	0.8	3.2	5.6
19	14	44	6.4	9.6	2.4	1.6	4	2.4	4	3.2	4	0.8	3.2	2.4
20	15	36	2.4	0.8	4	1.6	5.6	0.8	1.6	5.6	2.4	4	1.6	5.6
21	16	44	3.2	4.8	4	0.8	9.6	5.6	2.4	2.4	2.4	5.6	1.6	1.6
22	17	36	0.8	1.6	2.4	6.4	6.4	1.6	2.4	4	3.2	3.2	2.4	1.6
23	18	42.4	2.4	3.2	3.2	6.4	2.4	1.6	9.6	6.4	1.6	1.6	1.6	2.4

... Throughput Monte Carlo | Throughput Monte Carlo Weeks

READY 100%

Throughput / week Trend

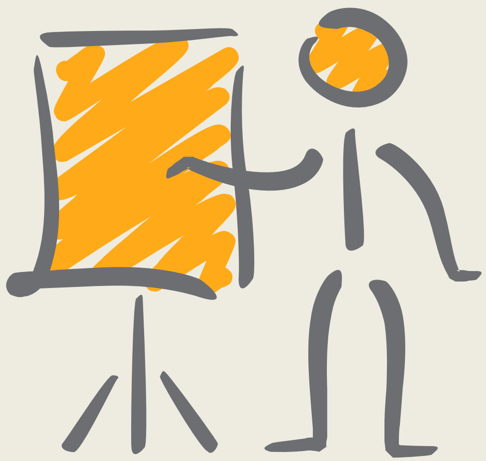


Please, please, Capture context

Year	Week	Team	Throughput	Context
2014	12	Blue	12	
2014	13	Blue	2	Moved offices
2014	14	Blue	7	No performance testing env.
2014	15	Blue	11	
2014	16	Blue	2	Thanksgiving week
2014	17	Blue	4	Learning new javascript library

Context helps select the right samples for future forecasting

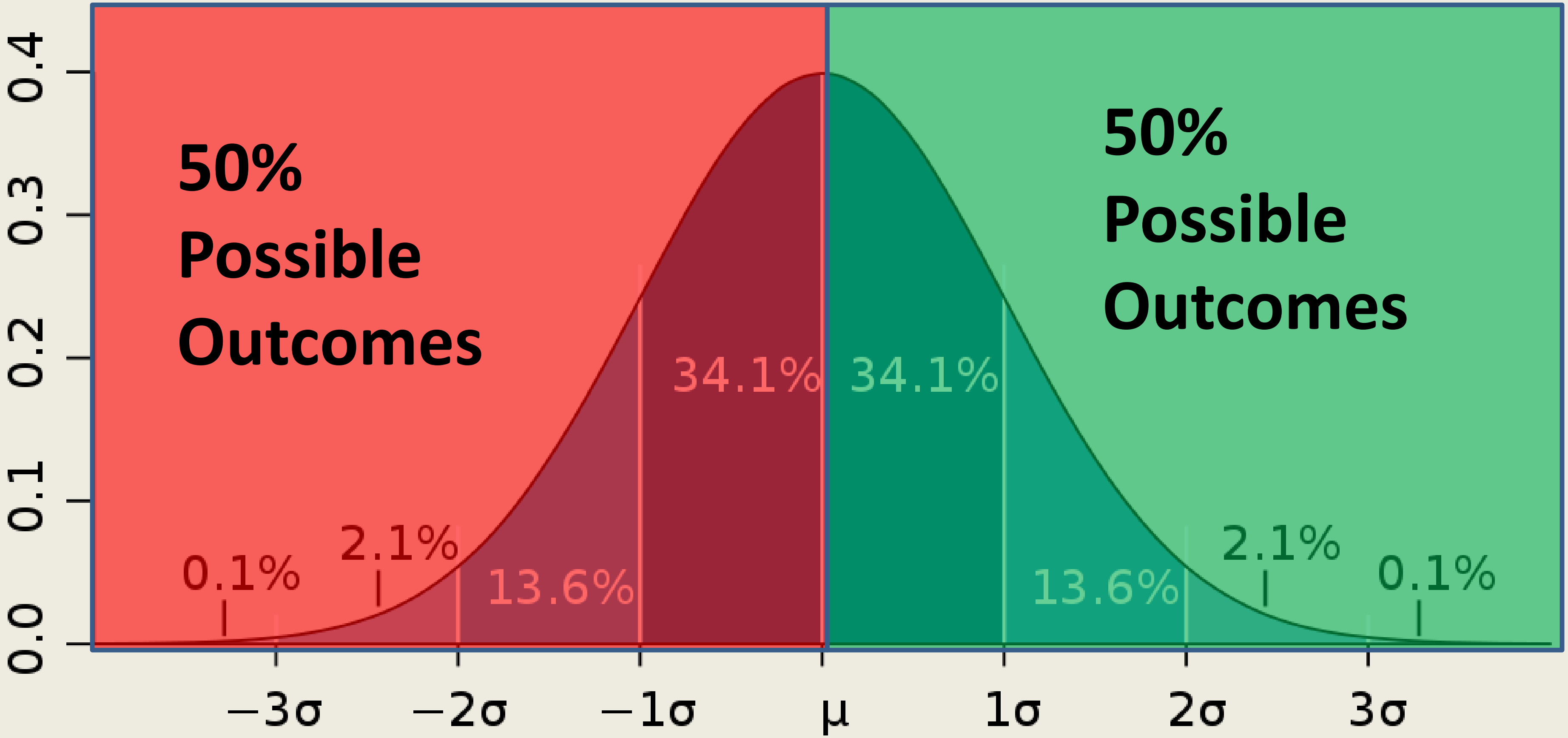
BETA



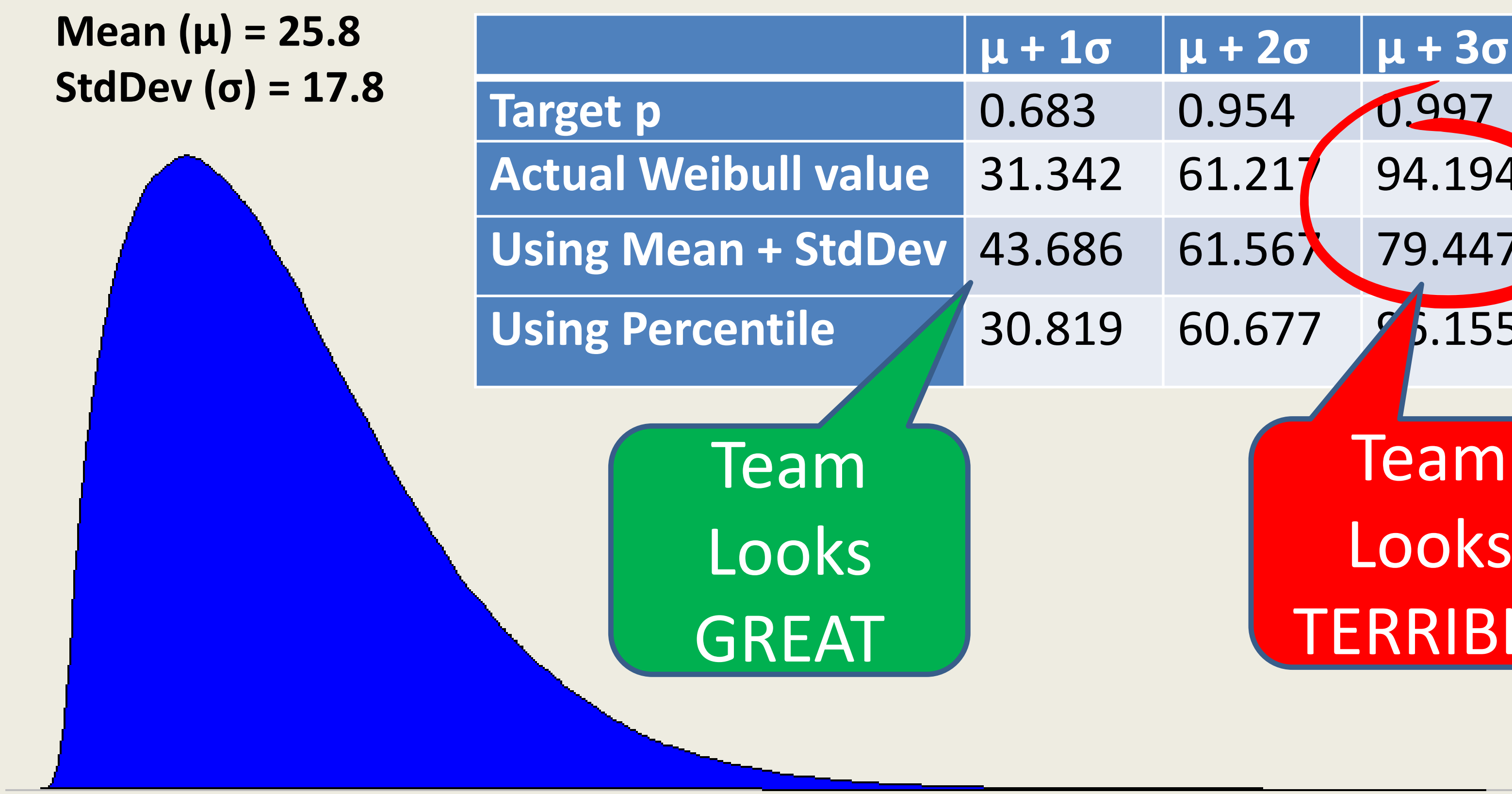
ADVANCED – I know this will be tough to understand but want to put it into the public for comment!

PROCESS ADVICE BASED ON CYCLE-TIME DISTRIBUTION

Flaw of averages



Introducing – Weibull Distribution



	μ + 1σ	μ + 2σ	μ + 3σ
Target p	0.683	0.954	0.997
Actual Weibull value	31.342	61.217	94.194
Using Mean + StdDev	43.686	61.567	79.447
Using Percentile	30.819	60.677	96.155

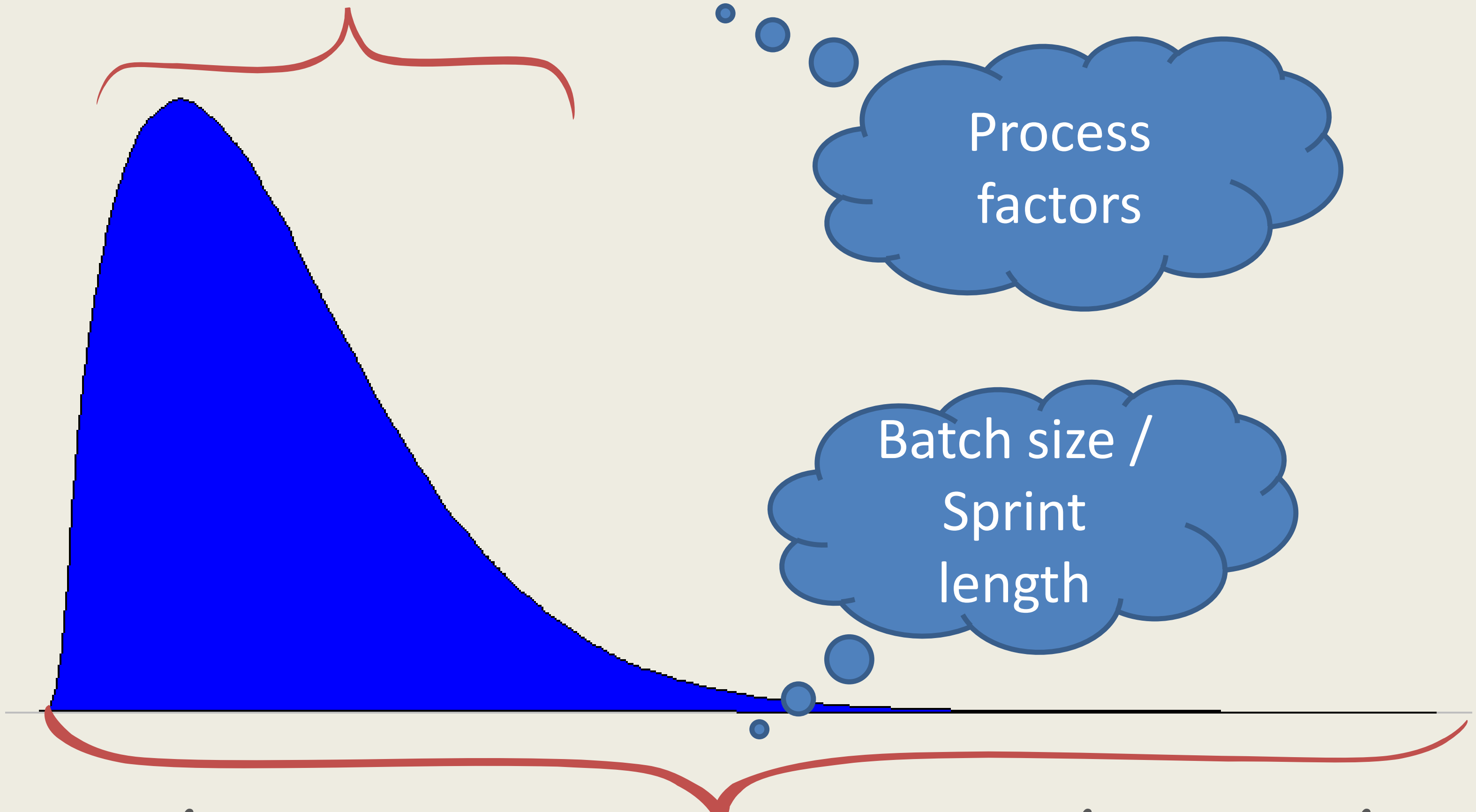
Team Looks GREAT

Team Looks TERRIBLE

Message: Don't use Standard Deviation, use Percentile

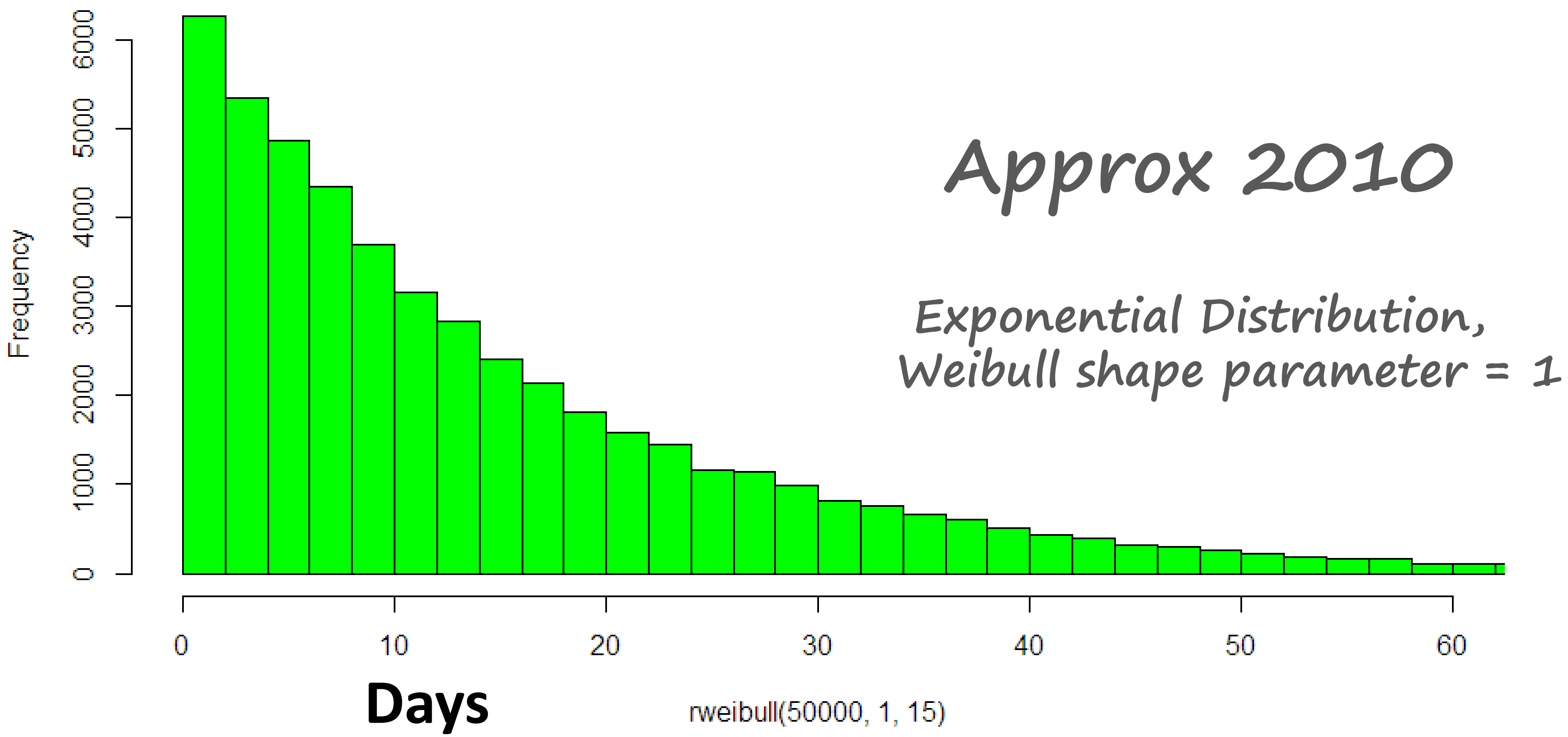
Introducing – Weibull Distribution

Shape parameter (how bulbous)

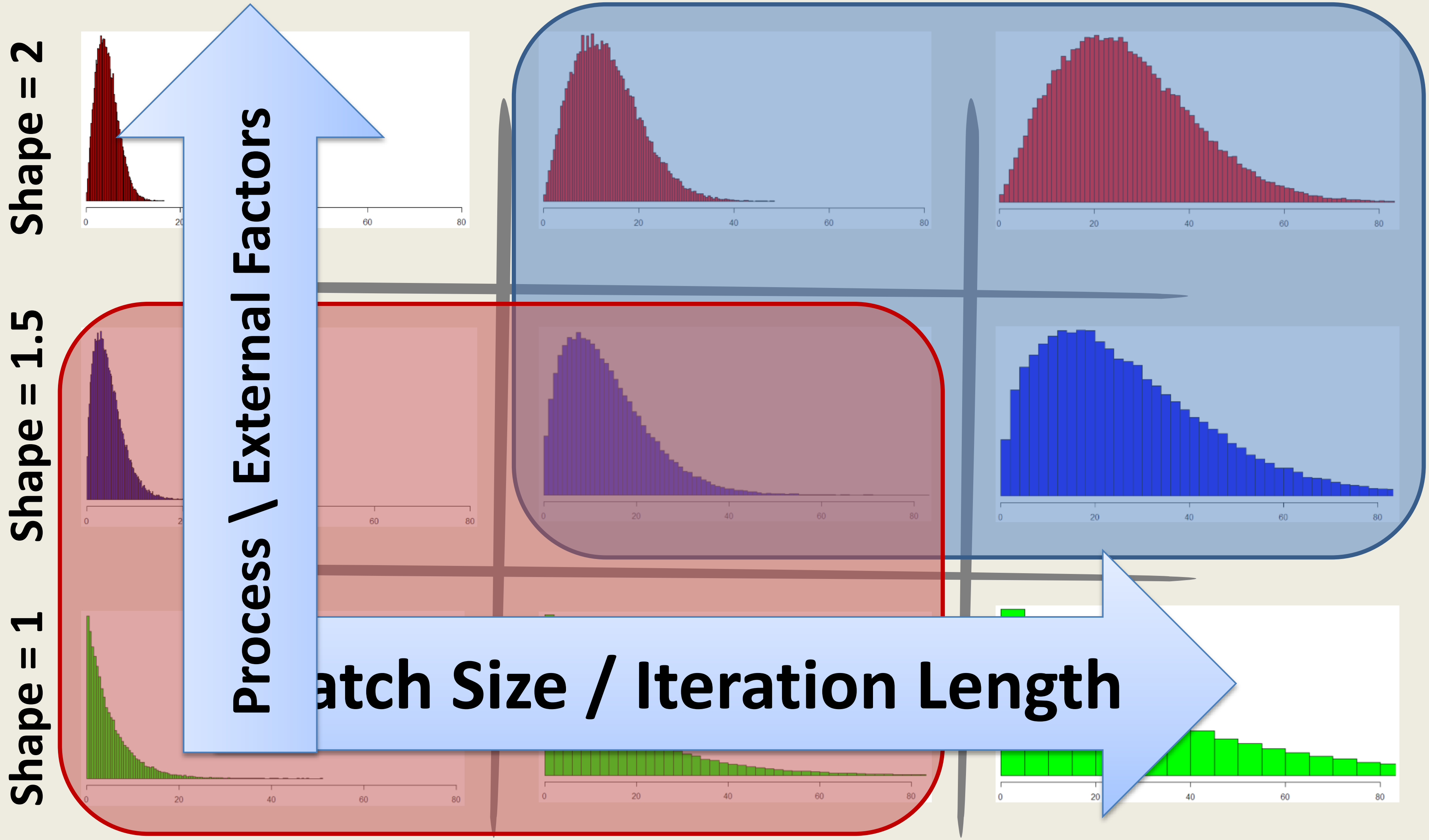


Scale parameter (63% Values Below)

@t_magennis



Work Item Cycle Time or Lead Time Distribution Through the Ages



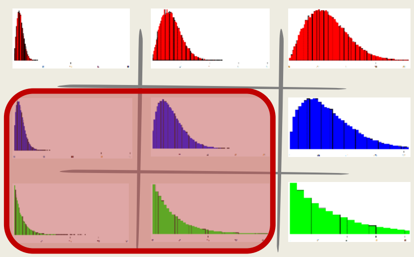
Scale = 5
< 1 week

Scale = 15
~ 2 week sprint

Scale = 30
~ 1 month

Work Item Cycle Time or Lead Time

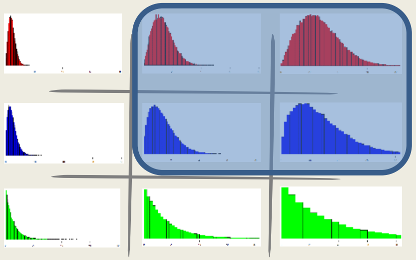
@t_magennis



Lean, Few dependencies

- Higher work item count
- More granular work items
- Lower WIP
- Team Self Sufficient
- Internal Impediments

- Do: Automation
- Do: Task Efficiency



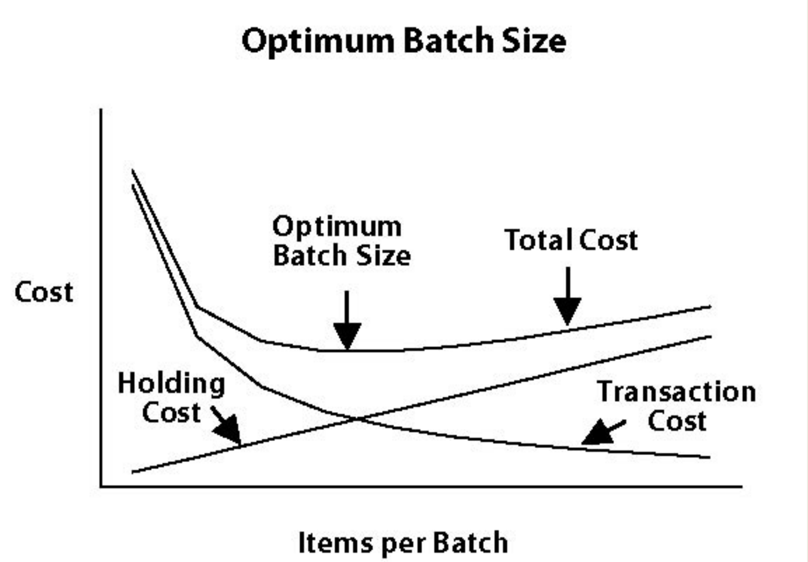
Sprint, Many dependencies

- Lower work item count
- Chunkier work items
- Higher WIP
- External Dependencies
- External Impediments

- Do: Collapse Teams
- Do: Impediment analysis

Notes so I don't get death threats

- There possibly is no BETTER matrix position
- If some factors are immovable, so will the matrix position
- I don't know all the factors and causes and probably never will



From "The Principles of Product Development Flow," by Donald G. Reinertsen. Celeritas Publishing: 2009. Copyright 2009, Donald G. Reinertsen

Weibull Shape Parameter	0 to 10	10 to 30
1.3 to 2 (Weibull Range)	**Traits:** Small unique work items. Medium WIP. Few external impediments. Fair predictability.	**Traits:** Larger unique work items. High WIP. Low predictability. Many external dependencies. **Process advice:** Focus on identification and removal of impediments and delays, and quality. Scrum optimal.
1 to 1.3 (Exponential Range)	**Traits:** Small or repetitive work items. Low WIP. Few external dependencies. Good predictability. **Process advice:** Automation of tasks, focus on task efficiency. Lean/Kanban optimal.	**Traits:** Larger work items. Large WIP. Many external dependencies. Poor predictability.

Weibull Scale Parameter

Session Feedback

- Please provide feedback on this session!

- You can do so in **3 ways:**
  1. Visit this session on the Mobile App. Click Session Feedback.
  2. Scan the unique QR Code for this session located at the front and back of the room.
  3. Visit the unique URL for this session located at the front and back of the room.

- Thank you for providing your feedback ☺


Conference Special:
Download the session slides,
a free copy of our simulation
software and a copy of this
book in PDF format from
http://bit.ly/agilesim
Forecasting and
Simulating Software
Development Projects
Effective Modeling of Kanban & Scrum
Projects using Monte-carlo Simulation
Troy Magennis
Focused Objective
software risk solutions
FocusedObjective.com
@AgileSimulation
Troy Magennis
troy.magennis@focusedobjective.com
phone: 425 223 8097 skype: troy.magennis twitter: @t_magennis

Contact Details

www.FocusedObjective.com

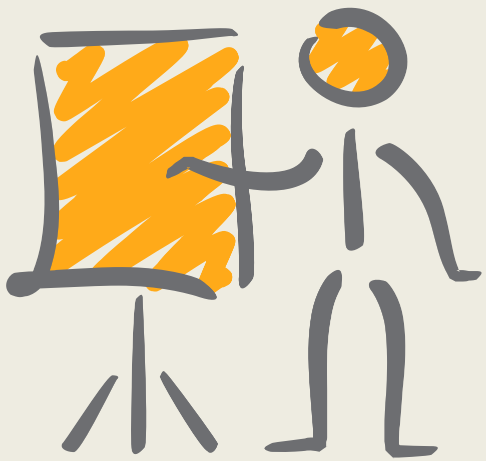
Download latest software, videos, presentations and articles on forecasting and applied predictive analytics

Troy.Magennis@focusedobjective.com

My email address for all questions and comments

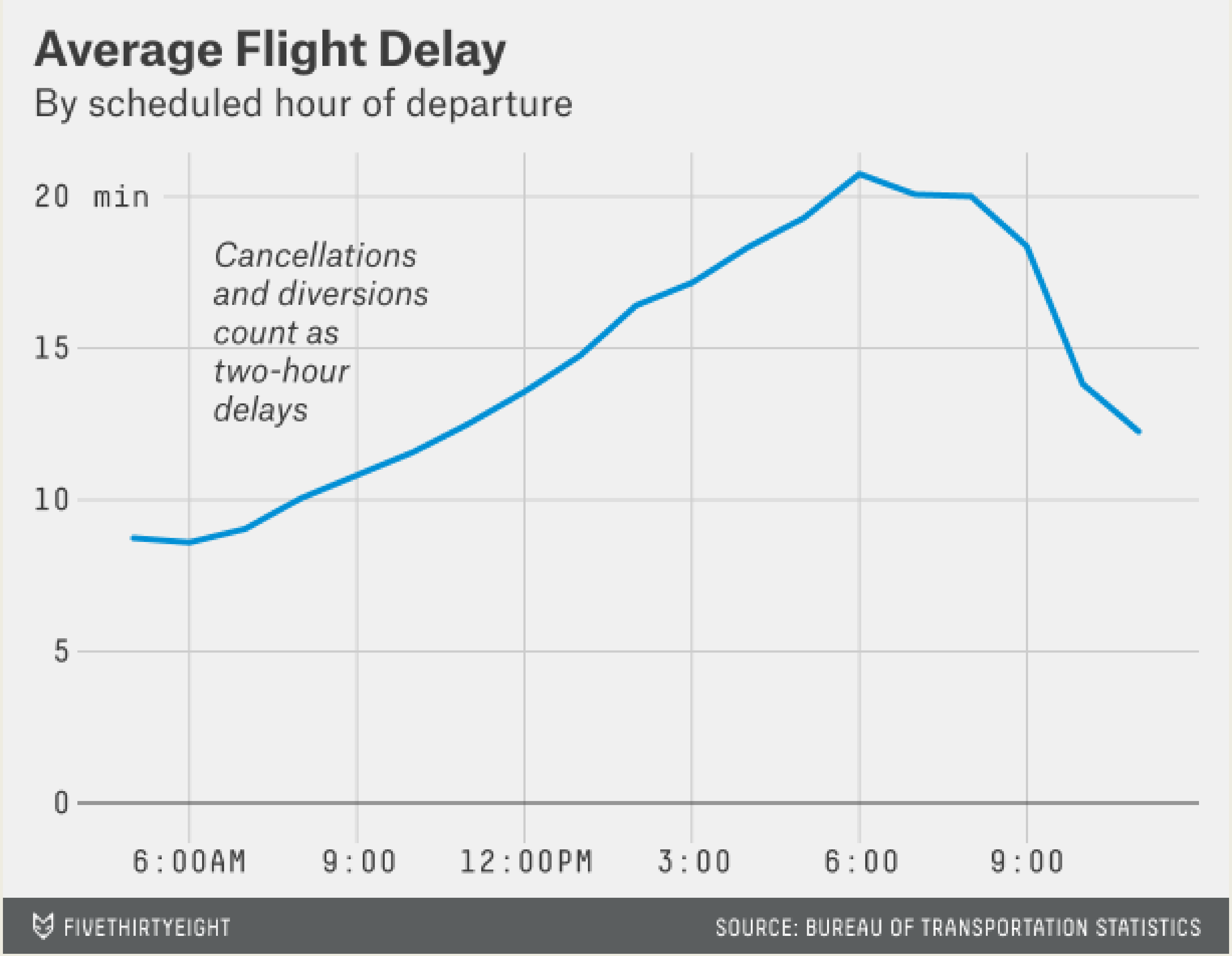
@t_magennis

Twitter feed from Troy Magennis



DEPENDENCY IMPACT

What are the odds of nothing going wrong in a sequential process?

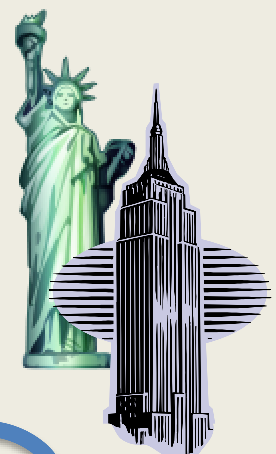

Average Flight Delay
By scheduled hour of departure
20 min
15
10
5
0
Cancellations and diversions count as two-hour delays
6:00AM
9:00
12:00PM
3:00
6:00
9:00
FIVETHIRTYEIGHT
SOURCE: BUREAU OF TRANSPORTATION STATISTICS

SEA to SFO (2 hours)

2 hours

SFO to JFK (6 hours)

1 hour un-board/board



JFK to SFO (6 hours)

6 hours

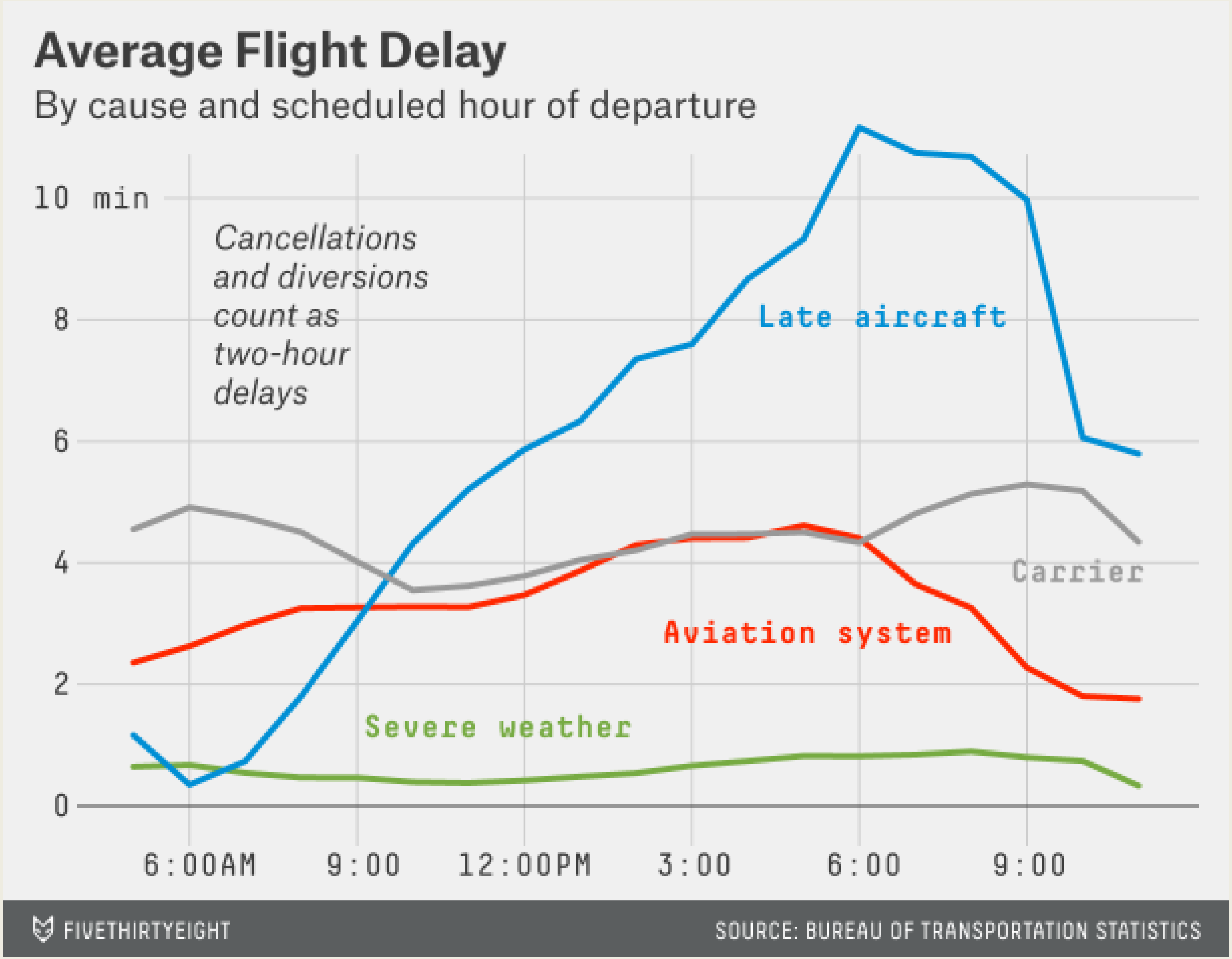
1 hour un-board/board

SFO to SEA
(2 hours)

6 hours

1 hour un-board/board

2 hours

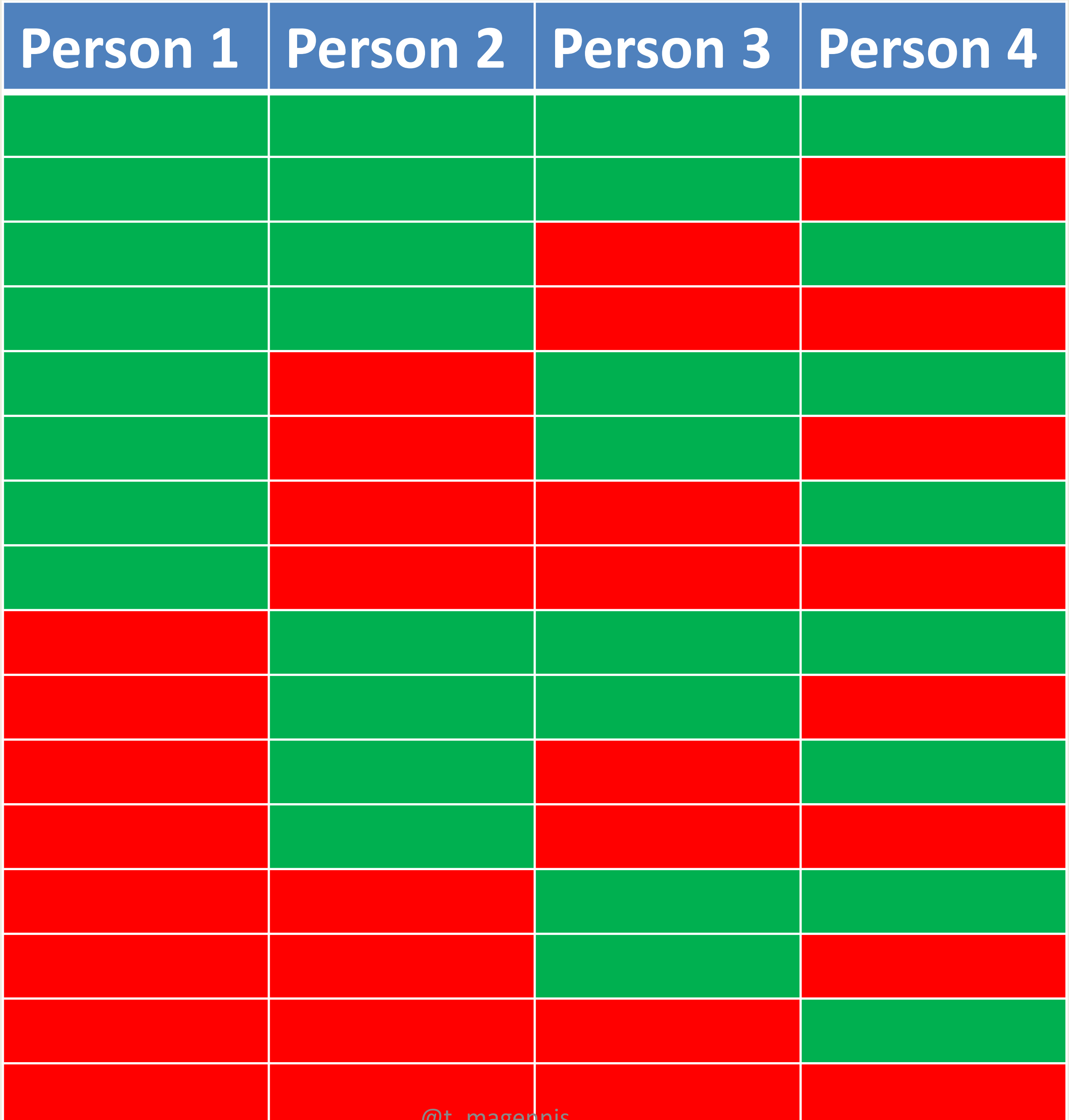

Average Flight Delay
By cause and scheduled hour of departure
10 min
8
6
4
2
0
Cancellations and diversions count as two-hour delays
Late aircraft
Carrier
Aviation system
Severe weather
6:00AM
9:00
12:00PM
3:00
6:00
9:00
FIVETHIRTYEIGHT
SOURCE: BUREAU OF TRANSPORTATION STATISTICS

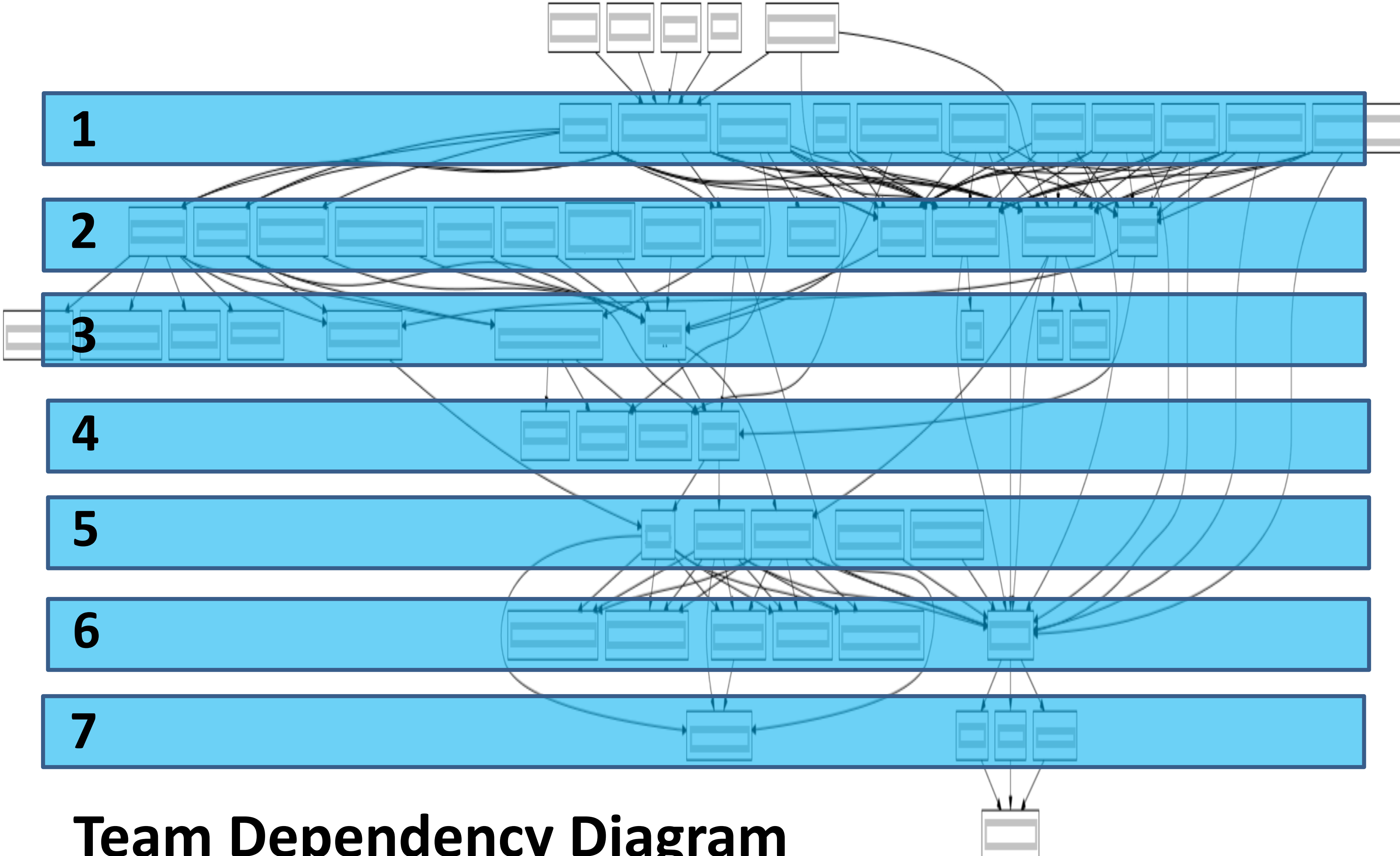
Four people arrange a restaurant booking after work

Q. What is the chance they arrive on-time to be seated?

1 in 16 EVERYONE is ON-TIME

15 TIMES more likely at least on person is late





Team Dependency Diagram

$1 \text{ in } 2^n$

or

$1 \text{ in } 2^7$

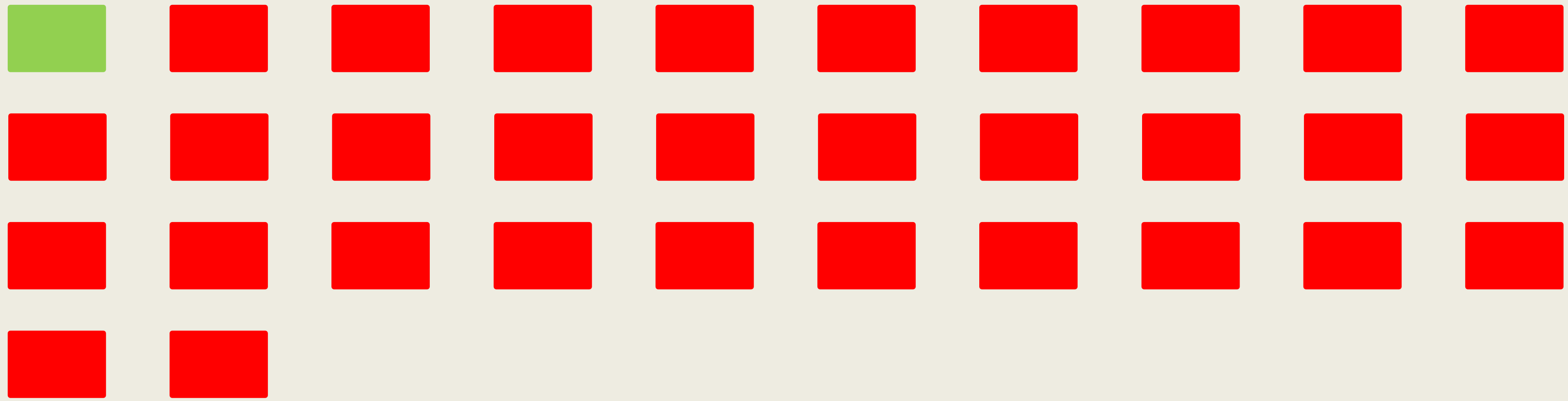
or

1 in 128

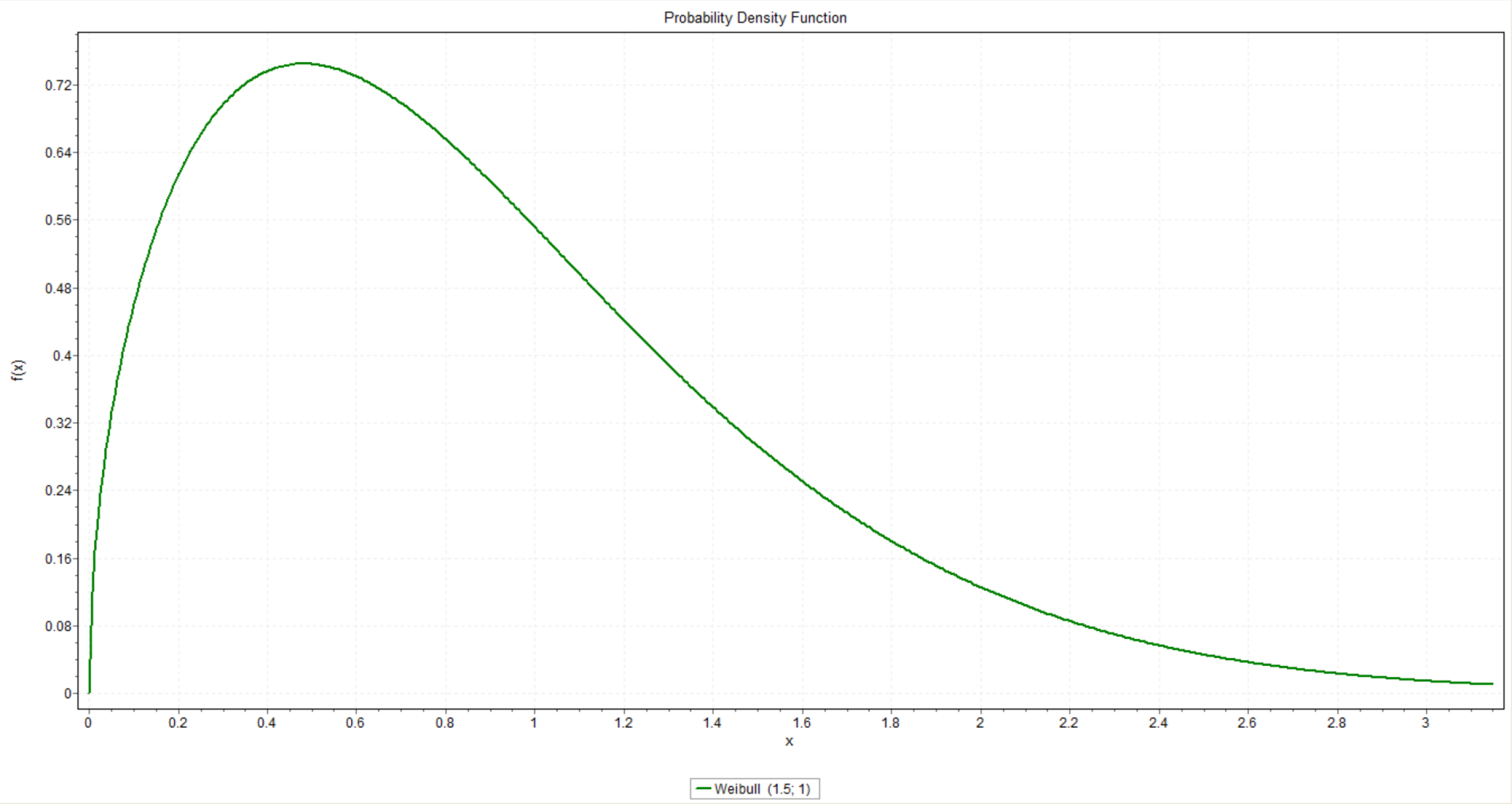
7 dependencies
1 chance in 128

6 dependencies

1 chance in 64

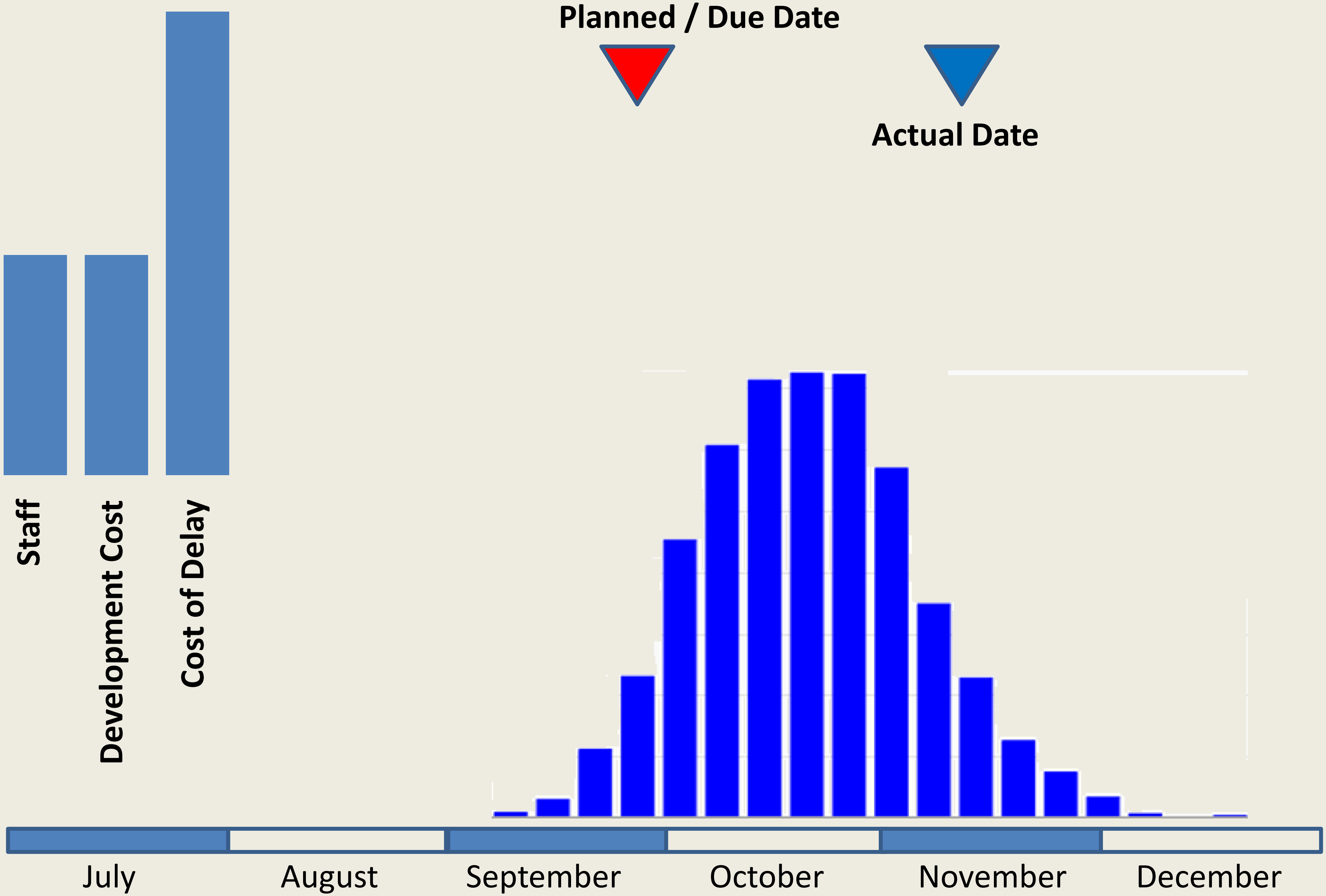


5 dependencies
1 chance in 32


Probability Density Function
f(x)
0.72
0.64
0.56
0.48
0.4
0.32
0.24
0.16
0.08
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
x
Weibull (1.5; 1)

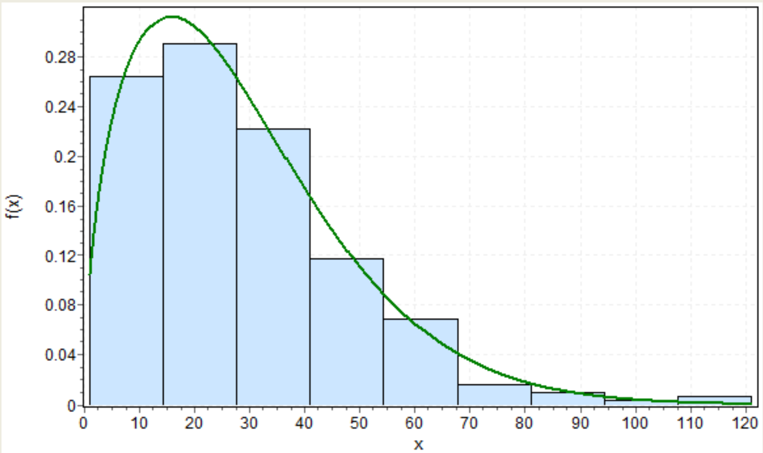
Overfitting

- If training a model on historical data, risk is it only forecasts historical data correctly
- Some causes
  - Samples not randomized
    - Process changes over time, but samples from one era
    - Samples sorted in some way and pulled from one end
  - Samples not chosen with future “Context” in mind
    - Events occur but samples prior to event used
    - Environmental and seasonal disruptions ignored



Forecast Completion Date

Historical Story Lead Time Trend



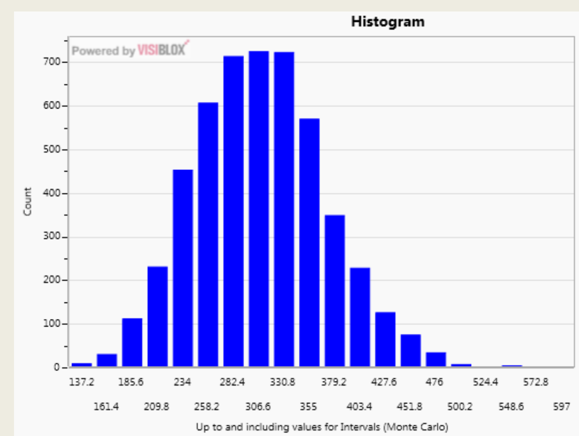
Sum Random Numbers

25	31		19
11	43		12
29	65		24
43	45		27
34	8		21
26	7		3
31	34		9
45	73	.....	20
22	54		23
27	48		29
Sum: 295	410		187

Basic Cycle Time Forecast Monte Carlo Process

1. Gather historical story lead-times
2. Build a set of random numbers based on pattern
3. Sum a random number for each remaining story to build a single outcome
4. Repeat many times to find the likelihood (odds) to build a pattern of likelihood outcomes

$$Total\ Days = \frac{Sum\ (\ Story_n\ \times\ Random_n\ )}{Effort}$$

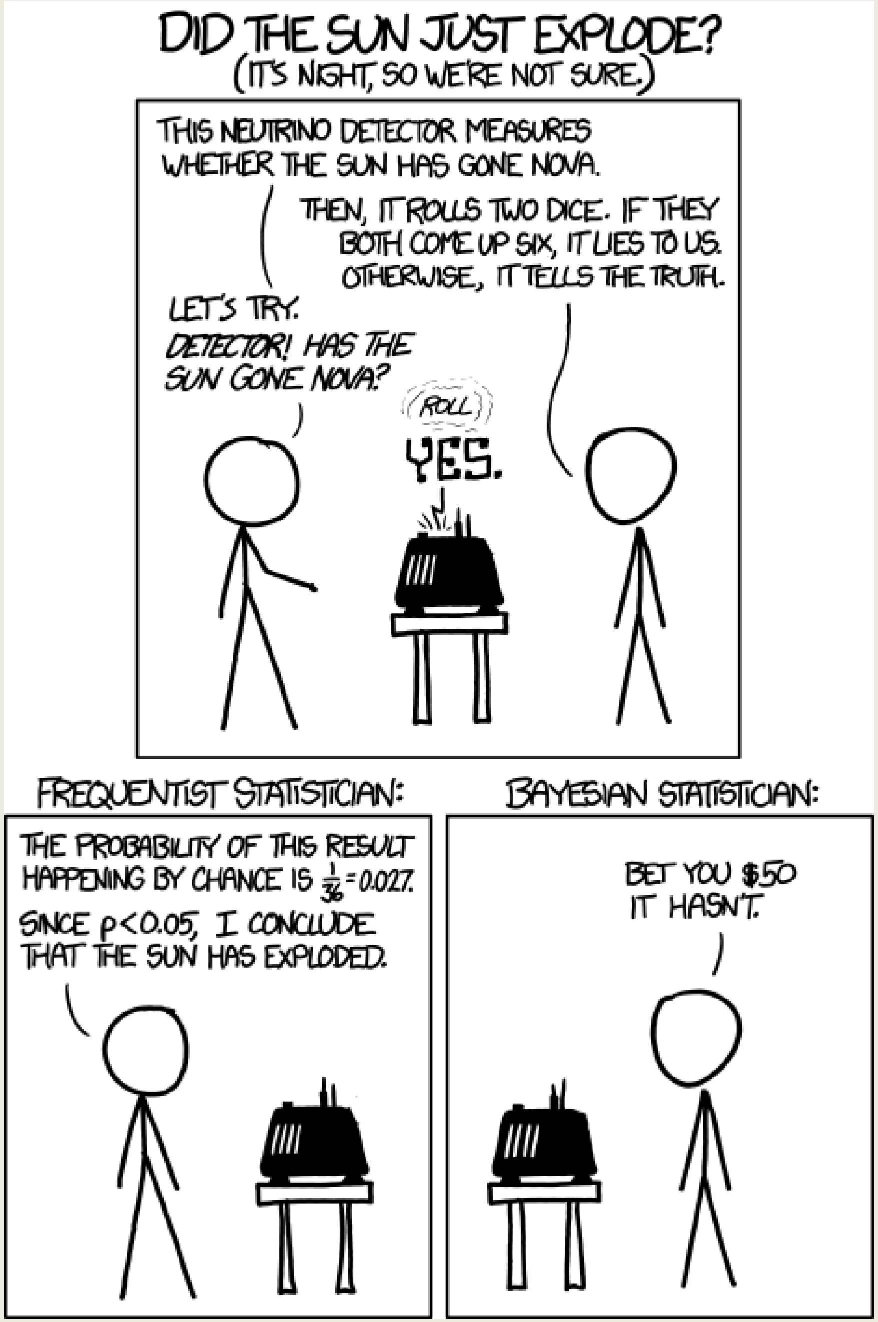


Days To Complete

@t_magennis

Correlation and Outliers

- Outliers a major factor on correlation
- Assume linear correlation, always scatterplot
- Calculations
  – Pearson Correlation Co-efficient
  – Spearman's Rank Order
    - If range is large, this is a good candidate
  – Least-squares Method
    - Vulnerable to extreme values



[HTTP://XKCD.COM/1132/](http://xkcd.com/1132/)