

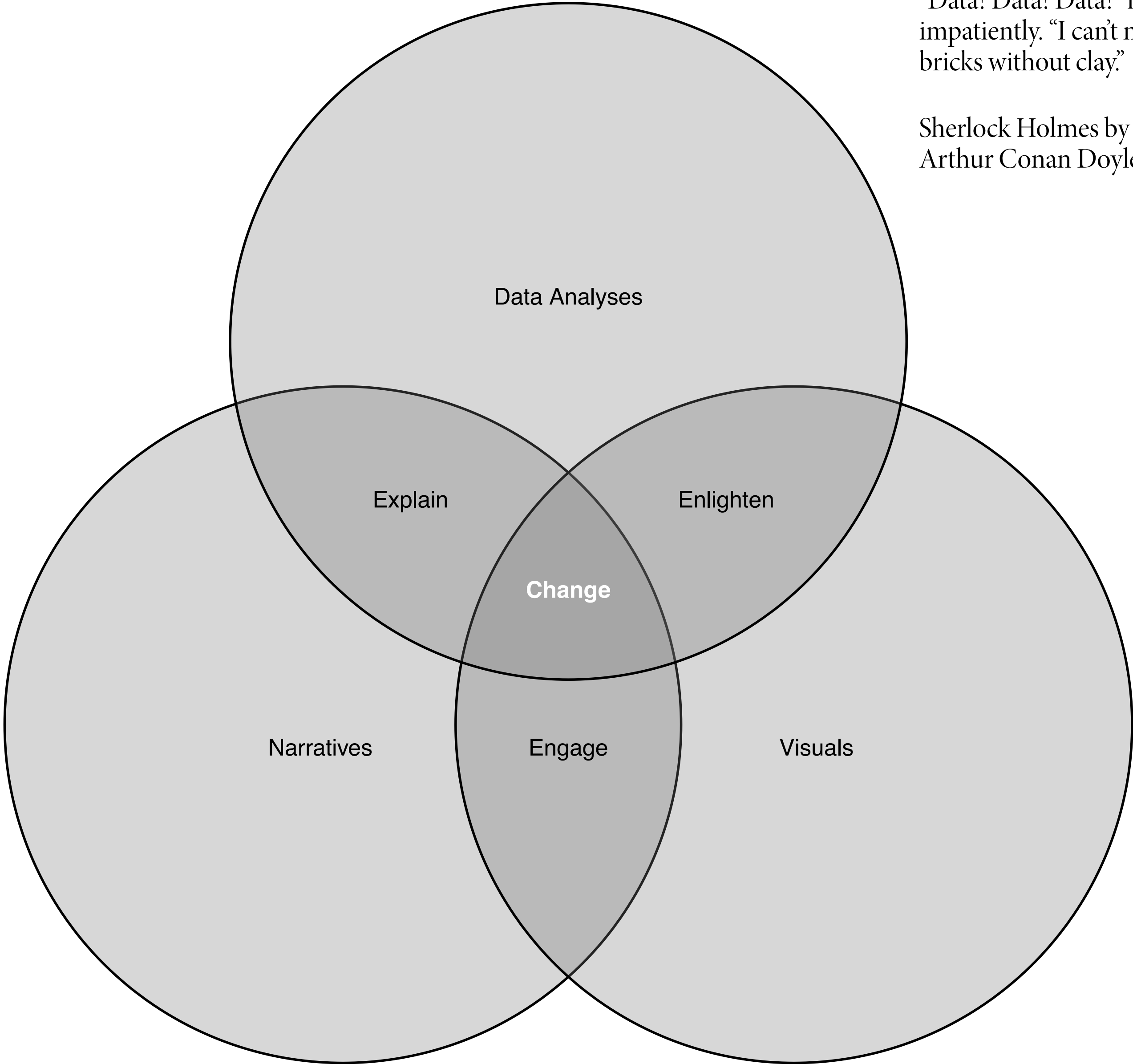
Storytelling with data

08 | Effective business writing with audience analysis

course overview, learn to drive change using data visuals and narrative

“Data! Data! Data!” he cried impatiently. “I can’t make bricks without clay.”

Sherlock Holmes by Sir Arthur Conan Doyle, *author*



No one ever made a decision because of a number. They need a story.

Daniel Kahneman, *psychologist, behavioral economist, and author*

The greatest value of a picture is when it forces us to notice what we never expected to see.

John W Tukey, *mathematician*

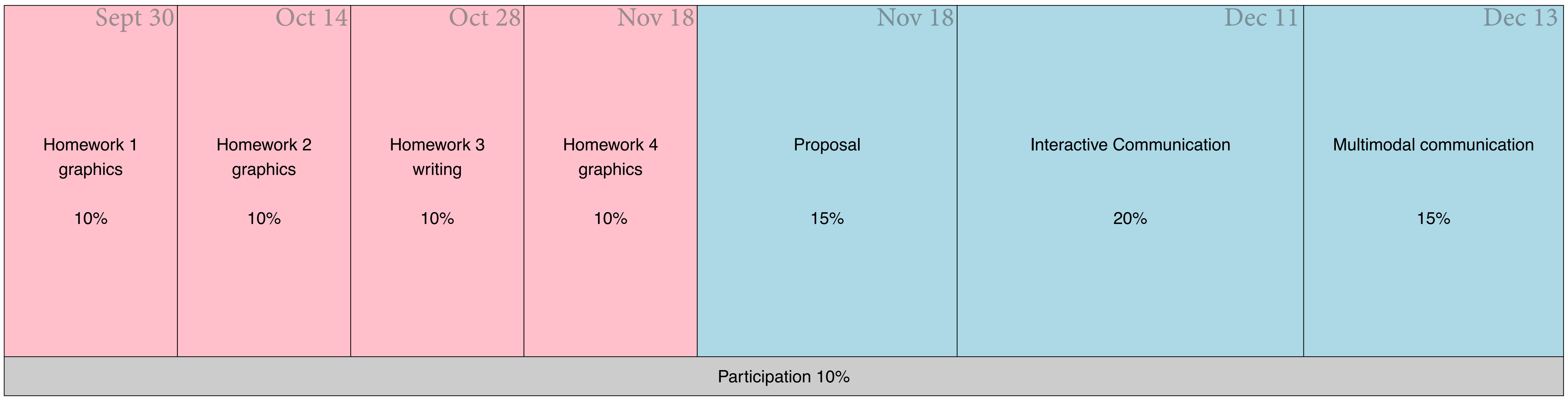
general course deliverable timeline

Individual Work

For learning data visualization and written narrative techniques

Group work

For building graphics and narrative into interactive communications



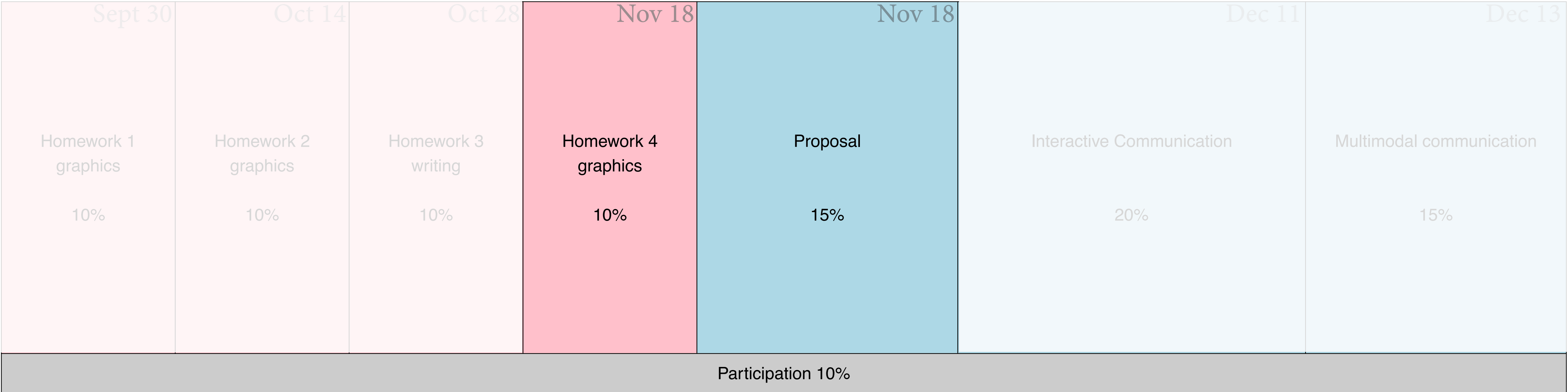
next deliverables, individual homework three and group proposal

Individual Work

For learning data visualization and written narrative techniques

Group work

For building graphics and narrative into interactive communications



proposals — common components communicated

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and significance

C. Introduction and background

- Relevant literature review
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- Analysis and expected results
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and significance

C. Introduction and background

- Relevant literature review
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- Analysis and expected results
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

Title | accurately represents the *content* and *scope* of the proposal.

analytics project scope | research proposal guidelines — where audience is *granting agencies*

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and significance

C. Introduction and background

- Relevant literature review
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- Analysis and expected results
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

Abstract | frames the goals and scope of the study, briefly describes the methods, and presents the hypotheses and expected results or outputs.

Sets up proper expectations, so be careful to avoid misleading readers into thinking that the proposal addresses anything other than the actual research topic.

Try for no more than two short paragraphs.

analytics project scope | research proposal guidelines — where audience is *granting agencies*

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and **significance**

C. Introduction and background

- Relevant literature review
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- Analysis and expected results
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

Significance | begins with the big picture and then funnels the reader through the hypotheses to the goals or specific aims of the research.

analytics project scope | research proposal guidelines — where audience is *granting agencies*

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and significance

C. Introduction and background

- Relevant [literature review](#)
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- Analysis and expected results
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

[Literature review](#) | sets the stage for the proposal by discussing the most widely accepted or influential papers on the research.

The **key** is to be able to show where the *proposed work would extend what has been done* or how the proposed *fills a gap* or resolves uncertainty, etc.

If the background literature does not help you accomplish either of those two points, you should question why you have it at all.

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and significance

C. Introduction and background

- Relevant literature review
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- Analysis and expected results
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

Preliminary data | can help establish credibility, likely success, or novelty of the proposal.

But avoid overstating the implications of the data or suggesting you've already solved the problem.

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and significance

C. Introduction and background

- Relevant literature review
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- Analysis and expected results
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

Research plan | The goal is to keep the reader focused on the overall significance, objectives, specific aims, and hypotheses while providing important methodological, technological, and analytical details.

Contains the details of the implementation, analysis, and inferences of the study.

Convince the reader that the project can be accomplished.

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and significance

C. Introduction and background

- Relevant literature review
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- Analysis and expected results
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

Objectives, hypotheses, aims, methods |

Objectives refer to broad, scientifically far-reaching aspects of a study, while *hypotheses* refer to a more specific set of testable conjectures. Specific *aims* focus on a particular question or hypothesis and the *methods* needed and outputs expected to fulfill the aims.

Of note, these points will typically have already been briefly introduced earlier, *e.g.*, in the abstract. Bring in more detail here.

I. Title

II. Abstract

III. Project description

A. Results from prior agency support

B. Problem statement and significance

C. Introduction and background

- Relevant literature review
- Preliminary data
- Conceptual, empirical, or theoretical model
- Justification of approach or novel methods

D. Research plan

- Overview of research design
- Objectives or specific aims, hypotheses, and methods
- **Analysis and expected results**
- Timetable

E. Broader impacts

IV. References cited

V. Budget and budget justification

Analysis and expected results | If early data are available, show how you will analyze them to reach your objectives or test your hypotheses.

If such data are unavailable, consider culling data from the literature to show how you expect the results to turn out and to show how you will analyze your data when they are available.

Complete a table or diagram, or run statistical tests using the preliminary or "synthesized" data. This can be a good way to show how you would interpret the results of such data.

**content, structure, and details
should adapt to our audience**

**Analytics
Executives**

Lead an organization's data analytics strategy, driving data-related business changes to transform company into a more analytics-driven one.

**Chief
Executives**

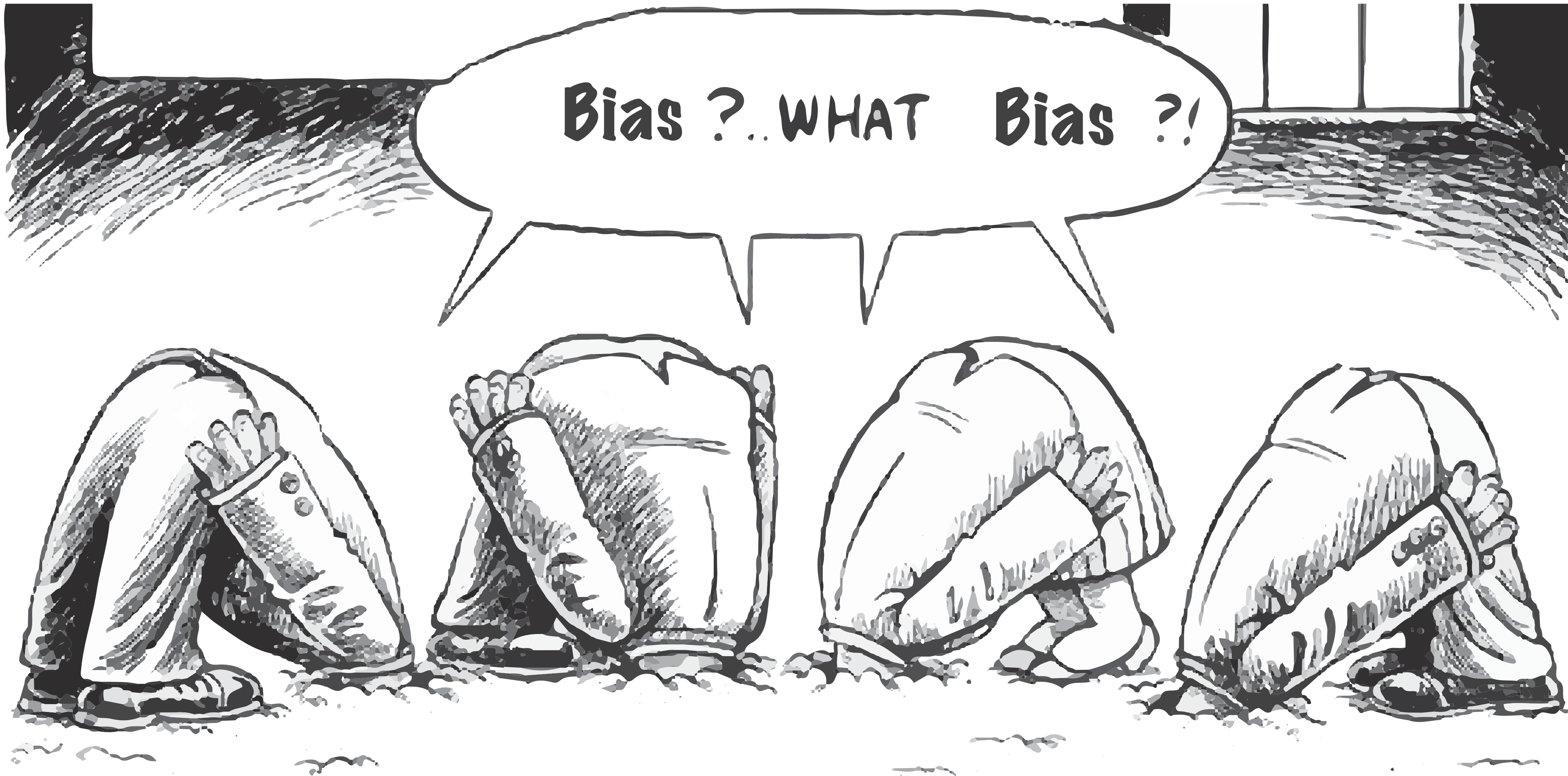
Leads management of company; responsible for maximizing company value, high-level decisions on policy and strategy; drives change.

**Marketing
Executives**

Lead responses to changing circumstances; shapes products, sales strategies, and marketing ideas, collaborating across the company.

**General and
Mixed Audiences**

The most challenging audiences to understand and develop persuasive messages.





self-interested bias

affect heuristic

groupthink

overconfidence

confirmation bias

availability bias

anchoring bias

halo effect

sunk-cost fallacy

endowment effect

loss aversion

competitor neglect

disaster neglect

Make **analogies** and **examples** comparable to the proposal.

Genuinely **admit uncertainty** in the proposal, and recognize **multiple options**.

Present ideas from a **neutral perspective**.
Becoming too emotional suggests bias.

Identify **additional data** that may provide new insight.

Consider **multiple anchors** in the proposal.

**reminder on how we use examples
to improve our own work**

learning from examples, don't copy — generalize from examples, then apply those generalizations to your work

An active learner asks questions, considers alternatives, questions assumptions, and even questions the trustworthiness of the author or speaker. *An active learner tries to generalize specific examples, and devise specific examples for generalities.*

An active learner doesn't passively sponge up information — that doesn't work! — but uses the readings and lecturer's argument as a springboard for critical thought and deep understanding.

bringing teachings together — *draft* proposal as example

data in narrative, proposal as a multi-level narrative — title, headings, body, captions

Proposal for exploring game decisions informed by expectations of joint probability distributions

To: Scott Powers, Director of Quantitative Analysis, Los Angeles Dodgers
From: Scott Spencer, Faculty and Lecturer, Columbia University

14 February 2019

Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on *expectations* of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—*should Sanchez steal against Sabathia? Or against Pineda?*

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face 10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected* utility (winning the game).

Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.

2 Modeling probabilities for steal success illustrates a broader benefit

To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:

In a game against New York Yankees, should Milwaukee Brewers's Lorenzo Cain attempt to steal second base with no one else on base and two outs before the seventh inning, against Gary Sanchez as catcher and Michael Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?

More specifically, how can we know the *expectation* that Cain's attempt in each situation increases the probability of expected runs that inning and by how much? Using Stan, I've coded a generative model that along with play outcomes considers various information (runner foot-speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:

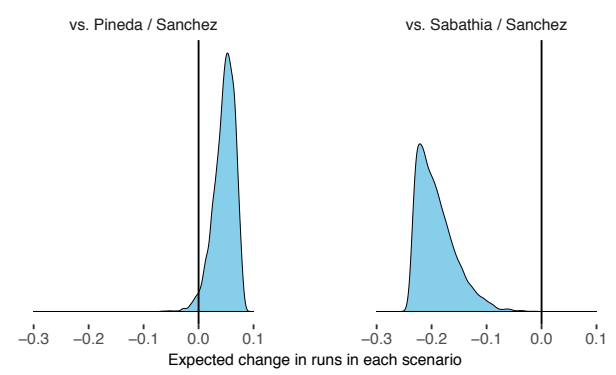


Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez-Pineda duo.

Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:

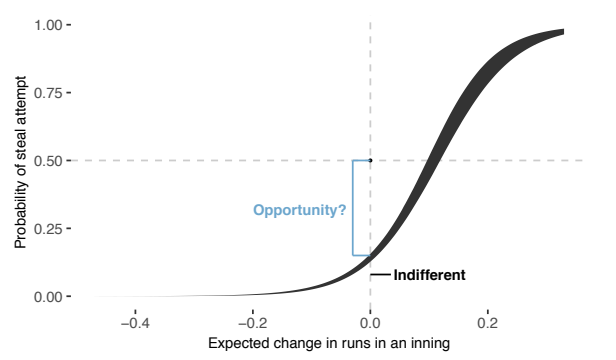


Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

The **black band** represents the range of variation across managers' decisions. At the intersection of **indifference**, managers tend to say steal only 10 percent of the time, leaving opportunity.

The above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.

3 For value, compare an investment to free-agent costs

A fully-realized model will require significant effort from a team with deep experience in baseball, generative modeling, and Stan. To get the talent, we should compare cost to acquiring expected wins from free-agents. Each win above a *replacement-level* player costs about 10 million per year (Swartz 2017). As with free-agent value over replacement player, game-time decisions informed from more accurate probabilities should add wins over a season. The scope of what we can answer, moreover, goes beyond in-game strategy (player acquisitions, salary arbitration). More immediately, however, we can begin to implement this approach for specific events, with a scope closer to the example above, being mindful that information learnt are conditional upon unmodeled context.

4 For accuracy, compare model results to betting market odds

Measuring performance of a fully-realized model may seem tricky: *we only see the outcome of our decisions*. But we can, say, compare the accuracy of our estimates against the betting market where interested investors are trying to forecast game outcomes.

5 Conclusion

The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.

6 References

Carpenter, Bob, et. al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." *McKinsey Quarterly* 13 June 2018: 1–9.

———. 2018b. "A view from the front lines of baseball's data-analytics revolution." *McKinsey Quarterly* 5 July 2018: 1–8.

Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.

Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." <https://www.fan-graphs.com/blogs/the-recent-history-of-free-agent-pricing/>.

data in narrative, example tries to maximize messages *within constraints* of the communication

Readability Statistics	
Counts	
Words	732
Characters	4,083
Paragraphs	18
Sentences	35
Averages	
Sentences per Paragraph	2.9
Words per Sentence	18.1
Characters per Word	5.3
Readability	
Flesch Reading Ease	33.2
Flesch-Kincaid Grade Level	13
Passive Sentences	0%

Proposal for exploring game decisions informed by expectations of joint probability distributions

To: Scott Powers, Director of Quantitative Analysis, Los Angeles Dodgers
From: Scott Spencer, Faculty and Lecturer, Columbia University

14 February 2019

Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on expectations of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—*should Sanchez steal against Sabathia? Or against Pineda?*

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face 10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected* utility (winning the game).

Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.

2 Modeling probabilities for steal success illustrates a broader benefit

To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:

PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 2

In a game against New York Yankees, should Milwaukee Brewer's Lorenzo Cain attempt to steal second base with no one else on base and two outs before the seventh inning, against Gary Sanchez as catcher and Michael Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?

More specifically, how can we know the *expectation* that Cain's attempt in each situation increases the probability of expected runs that inning and by how much? Using Stan, I've coded a generative model that along with play outcomes considers various information (runner foot-speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:

Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez-Pineda duo.

Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:

Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

The **black band** represents the range of variation across managers' decisions. At the intersection of **indifference**, managers tend to say steal only 10 percent of the time, leaving opportunity.

The above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.

PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 3

3 For value, compare an investment to free-agent costs

A fully-realized model will require significant effort from a team with deep experience in baseball, generative modeling, and Stan. To get the talent, we should compare cost to acquiring expected wins from free-agents. Each win above a *replacement-level* player costs about 10 million per year (Swartz 2017). As with free-agent value over replacement player, game-time decisions informed from more accurate probabilities should add wins over a season. The scope of what we can answer, moreover, goes beyond in-game strategy (player acquisitions, salary arbitration). More immediately, however, we can begin to implement this approach for specific events, with a scope closer to the example above, being mindful that information learnt are conditional upon unmodeled context.

4 For accuracy, compare model results to betting market odds

Measuring performance of a fully-realized model may seem tricky: *we only see the outcome of our decisions*. But we can, say, compare the accuracy of our estimates against the betting market where interested investors are trying to forecast game outcomes.

5 Conclusion

The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.

6 References

Carpenter, Bob, et al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." *McKinsey Quarterly* 13 June 2018: 1–9.


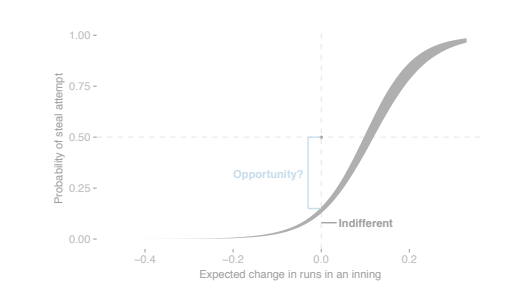
———. 2018b. "A view from the front lines of baseball's data-analytics revolution." *McKinsey Quarterly* 5 July 2018: 1–8.

Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.

Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." <https://www.fan-graphs.com/blogs/the-recent-history-of-free-agent-pricing/>.

data in narrative, organized on a grid

“Orderliness adds credibility to the information and induces confidence. Information presented with clear and logically set out titles, subtitles, texts, illustrations and captions will not only be read more quickly and easily but the information will also be better understood.”

	<p>Proposal for exploring game decisions informed by expectations of joint probability distributions</p> <p>To: Scott Powers, Director of Quantitative Analysis, Los Angeles Dodgers From: Scott Spencer, Faculty and Lecturer, Columbia University 14 February 2019</p> <hr/> <p>Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on expectations of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—<i>should Sanchez steal against Sabathia? Or against Pineda?</i></p> <p>1 Our current analyses do not optimize expected wins</p> <p>Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face 10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing <i>expected</i> utility (winning the game).</p> <p>Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.</p> <p>2 Modeling probabilities for steal success illustrates a broader benefit</p> <p>To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:</p>	
<p>PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 2</p>	<p>In a game against New York Yankees, should Milwaukee Brewer's Lorenzo Cain attempt to steal second base with no one else on base and two outs before the seventh inning, against Gary Sanchez as catcher and Michael Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?</p> <p>More specifically, how can we know the <i>expectation</i> that Cain's attempt in each situation increases the probability of expected runs that inning and by how much? Using Stan, I've coded a generative model that along with play outcomes considers various information (runner foot-speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:</p>  <p>Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:</p>  <p>The above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.</p> <p>Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez-Pineda duo.</p> <p>Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time. The black band represents the range of variation across managers' decisions. At the intersection of indifference, managers tend to say steal only 10 percent of the time, leaving opportunity.</p>	
<p>PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 3</p>	<p>3 For value, compare an investment to free-agent costs</p> <p>A fully-realized model will require significant effort from a team with deep experience in baseball, generative modeling, and Stan. To get the talent, we should compare cost to acquiring expected wins from free-agents. Each win above a <i>replacement-level</i> player costs about 10 million per year (Swartz 2017). As with free-agent value over replacement player, game-time decisions informed from more accurate probabilities should add wins over a season. The scope of what we can answer, moreover, goes beyond in-game strategy (player acquisitions, salary arbitration). More immediately, however, we can begin to implement this approach for specific events, with a scope closer to the example above, being mindful that information learnt are conditional upon unmodeled context.</p> <p>4 For accuracy, compare model results to betting market odds</p> <p>Measuring performance of a fully-realized model may seem tricky: we <i>only see the outcome of our decisions</i>. But we can, say, compare the accuracy of our estimates against the betting market where interested investors are trying to forecast game outcomes.</p> <p>5 Conclusion</p> <p>The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.</p> <p>6 References</p> <p>Carpenter, Bob, et al. 2017. "Stan: A Probabilistic Programming Language." <i>Journal of Statistical Software</i> 76 (1): 1–32.</p> <p>Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." <i>McKinsey Quarterly</i> 13 June 2018: 1–9.</p> <p>———. 2018b. "A view from the front lines of baseball's data-analytics revolution." <i>McKinsey Quarterly</i> 5 July 2018: 1–8.</p> <p>Parmigiani, G. 2002. "Decision Theory: Bayesian." In <i>International Encyclopedia of the Social Behavioral Sciences</i>, 3327–34.</p> <p>Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." https://www.fan-graphs.com/blogs/the-recent-history-of-free-agent-pricing/.</p>	

data in narrative, applies typographic principles to separate hierarchies of information, improve readability

“Most readers are looking for reasons to stop reading. . . . Readers have other demands on their time. . . . The goal of most professional writing is persuasion, and attention is a prerequisite for persuasion. Good typography can help your reader devote less attention to the mechanics of reading and more attention to your message.”

— Butterick, Matthew, *Practical Typography*

Proposal for exploring game decisions informed by expectations of joint probability distributions

To: Scott Powers, Director of Quantitative Analysis, Los Angeles Dodgers
From: Scott Spencer, Faculty and Lecturer, Columbia University
14 February 2019

Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018a). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on expectations of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—should Sanchez steal against Sabathia? Or against Pineda?

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face 10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing expected utility (winning the game).

Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.

2 Modeling probabilities for steal success illustrates a broader benefit

To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:

**Leading (line height): 145% of font size.
Butterick recommends: 120-145% of font size.**

Average line length: 84 characters with spaces. Butterick recommends 45-90.

PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 2

In a game against New York Yankees, should Milwaukee Brewers's Lorenzo Cain attempt to steal second base with no one else on base and two outs before the seventh inning, against Gary Sanchez as catcher and Michael Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?

More specifically, how can we know the expectation that Cain's attempt in each situation increases the probability of expected runs that inning and by how much? Using Stan, I've coded a generative model that along with play outcomes considers various information (runner foot speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017

Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez-Pineda duo.

PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 3

3 For value, compare an investment to free-agent costs

A fully-realized model will require significant effort from a team with deep experience in baseball, generative modeling, and Stan. To get the talent, we should compare cost to acquiring expected wins from free-agents. Each win above a replacement-level player costs about 10 million per year (Swartz 2017). As with free-agent value over replacement player, game-time decisions informed from more accurate probabilities should add wins over a season. The scope of what we can answer, moreover, goes beyond in-game strategy (player acquisitions, salary arbitration). More immediately, however, we can begin to implement this approach for specific events, with a scope closer to the example above, being mindful that information learnt are conditional upon unmodeled context.

4 For accuracy, compare model results to betting market odds

Measuring performance of a fully-realized model may seem tricky; we only see the outcome of our decisions. But we can, say, compare the accuracy of our estimates against the betting market where interested investors are trying to forecast game outcomes.

5 Conclusion

The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.

6 References

Carpenter, Bob, et al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." *McKinsey Quarterly* 13 June 2018: 1–9.

———. 2018b. "A view from the front lines of baseball's data-analytics revolution." *McKinsey Quarterly* 5 July 2018: 1–8.

Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.

Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." <https://www.fangraphs.com/blogs/the-recent-history-of-free-agent-pricing/>.

Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time. The black band represents the range of variation across managers' decisions. At the intersection of indifference, managers tend to say steal only 10 percent of the time, leaving opportunity.

data in narrative, messages first, details follow

Get our audience(s) to pay attention to, understand, (be able to) act upon



a maximum of messages, given constraints.

— Doumont, *Trees, Maps, Theorems*

Proposal for exploring game decisions informed by expectations of joint probability distributions

To: Scott Powers, Director of Quantitative Analysis, Los Angeles Dodgers
From: Scott Spencer, Faculty and Lecturer, Columbia University

14 February 2019

Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on expectations of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—*should Sanchez steal against Sabathia? Or against Pineda?*

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face 10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected* utility (winning the game).

Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.

2 Modeling probabilities for steal success illustrates a broader benefit

To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:

PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 2

In a game against New York Yankees, should Milwaukee Brewer's Lorenzo Cain attempt to steal second base with no one else on base and two outs before the seventh inning, against Gary Sanchez as catcher and Michael Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?

More specifically, how can we know the *expectation* that Cain's attempt in each situation increases the probability of expected runs that inning and by how much? Using Stan, I've coded a generative model that along with play outcomes considers various information (runner foot-speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:

Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez-Pineda duo.

Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:

Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

The **black band** represents the range of variation across managers' decisions. At the intersection of **indifference**, managers tend to say steal only 10 percent of the time, leaving opportunity.

The above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.

PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 3

3 For value, compare an investment to free-agent costs

A fully-realized model will require significant effort from a team with deep experience in baseball, generative modeling, and Stan. To get the talent, we should compare cost to acquiring expected wins from free-agents. Each win above a *replacement-level* player costs about 10 million per year (Swartz 2017). As with free-agent value over replacement player, game-time decisions informed from more accurate probabilities should add wins over a season. The scope of what we can answer, moreover, goes beyond in-game strategy (player acquisitions, salary arbitration). More immediately, however, we can begin to implement this approach for specific events, with a scope closer to the example above, being mindful that information learnt are conditional upon unmodeled context.

4 For accuracy, compare model results to betting market odds

Measuring performance of a fully-realized model may seem tricky: *we only see the outcome of our decisions*. But we can, say, compare the accuracy of our estimates against the betting market where interested investors are trying to forecast game outcomes.

5 Conclusion

The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.

6 References

Carpenter, Bob, et al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." *McKinsey Quarterly* 13 June 2018: 1–9.

———. 2018b. "A view from the front lines of baseball's data-analytics revolution." *McKinsey Quarterly* 5 July 2018: 1–8.

Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.

Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." <https://www.fan-graphs.com/blogs/the-recent-history-of-free-agent-pricing/>.

data in narrative, data graphics as paragraphs about data — linking narrative and data

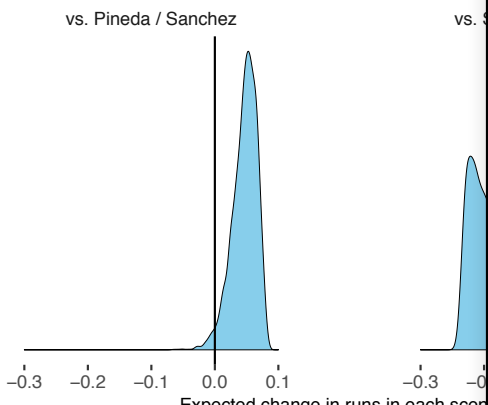
“Words, graphics, and tables are different mechanisms with but a *single purpose*—the presentation of information. Why should the flow of information be broken up into different places on the page...?”

— Edward Tufte, *The Visual Display of Quantitative Information*

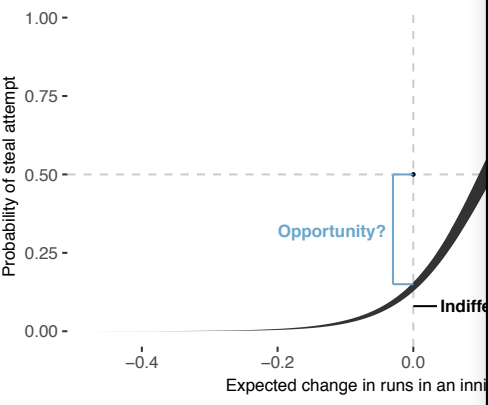
PROPOSAL FOR EXPLORATION

In a game against New York Yankees, should Milwaukee attempt to steal second base with no one else on base before the seventh inning, against Gary Sanchez as catcher and Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?

More specifically, how can we know the *expectation* that Cain increases the probability of expected runs that inning and coded a generative model that along with play outcomes (runner foot-speed, catcher pop-time) and player characteristics (runner foot-speed, catcher pop-time) and player characteristics. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda

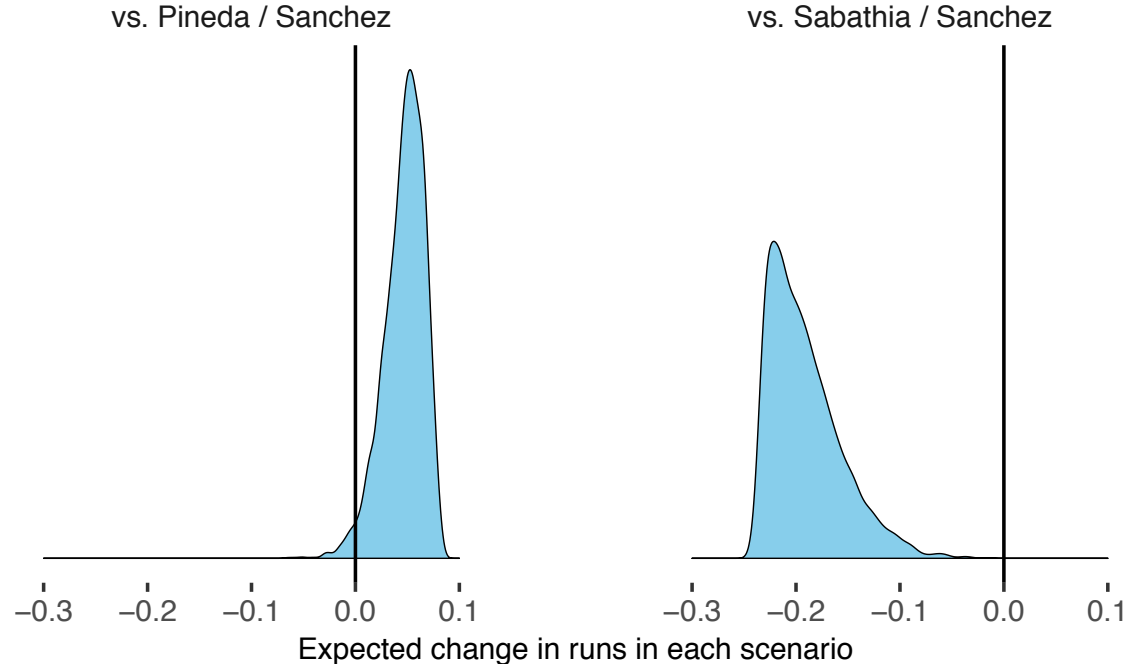


Notably, we get these expectations without multiple trials. Generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:

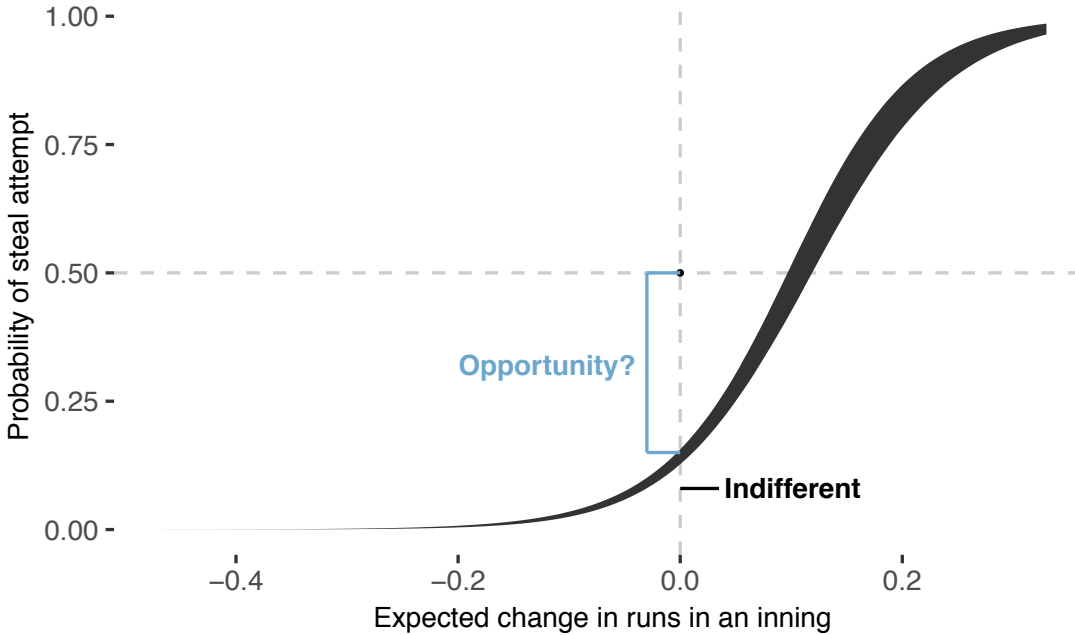


The above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.

(runner foot-speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:



Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:



The above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.

Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez–Pineda duo.

Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

The **black band** represents the range of variation across managers' decisions. At the intersection of **indifference**, managers tend to say steal only **10 percent** of the time, leaving opportunity.

***individual* student example**

individual student example, from memo to proposal

NYC Analytics

2021 February 2

To: **Martha Norrick**
Acting Chief Data Analytics Officer
Mayor's Office of Data Analytics
City of New York

To educate the public, let's explore why fatal collisions have increased during the pandemic.

It has been almost a year since the COVID-19 pandemic began, and there's an eerie quiet as workers, students, and tourists stay home to comply with social distancing guidelines. Yet despite emptier streets, New Yorkers are dying in car crashes at the highest rate since 2014, when we announced Vision Zero (Berger & Jones, 2020). The public needs to know why this is happening.

Economic downturns are generally associated with lower traffic fatalities (Yannis et al., 2014), but the COVID-19 recession is unprecedented. Preliminary research has shown that the rate of traffic fatalities is increasing nationwide along with risky behaviors like speeding and drug use (Wagner et al., 2020). Let's find out if this is also occurring in NYC.

Let's begin by analyzing NYPD collision reports from NYC OpenData, including the number of collisions, injuries, and fatalities. By aggregating the contributing factors for each crash, we can determine which driving behaviors are most likely to result in a fatal collision. Next, we will build upon our work by visualizing the data in Tableau, enabling us to analyze trends over time and location, and ultimately share our findings with the public.

New Yorkers are counting on us to keep them informed and keep them safe. Consistent with our Vision Zero goal, we can use education and transparency to create a city where New Yorkers don't die in car crashes.

Sincerely,
Joy Chen

Berger, P. & Jones, C. (2020, December 19). New York City Traffic Deaths Rise During Covid-19 Pandemic. *The Wall Street Journal*.

<https://www.wsj.com/articles/new-york-city-traffic-deaths-rise-during-covid-19-pandemic-11608382800>

Wagner, E., Atkins, R., Berning, A., Robbins, A., Watson, C., & Anderle, J. (2020, October). Examination of the traffic safety environment during the second quarter of 2020: Special Report (Report No. DOT HS 813 011). National Highway Traffic Safety Administration. <https://rosap.nhtl.bts.gov/view/dot/50940>

Yannis, G., Papadimitriou, E., & Folla K. (2014, March). Effect of GDP changes on road traffic fatalities. *Safety Science*. 63. 42-69. <https://doi.org/10.1016/j.ssci.2013.10.017>

individual student example, from memo to proposal



Proposal to analyze driving behaviors contributing to fatal vehicle collisions during Covid-19

To: **Martha Norrick**, Acting Chief Data Analytics Officer, City of New York
From: **Joy Chen**, Applied Analytics Student, Columbia University

23 February 2021

We've read the newspaper reports of a surge of speeding and drag racing (Meyer, 2020) as New Yorkers die in car crashes at the highest rate since we announced Vision Zero (Berger & Jones, 2020). But we have data: we don't need to read the paper to know that the Covid-19 pandemic has shifted the landscape of traffic safety in NYC. With data, we can compare the traffic fatality rate to the pre-pandemic trend, allowing us to quantify how much the Covid-19 pandemic has set back our progress towards our Vision Zero goal of reducing traffic fatalities in NYC. Further, we can determine next steps for achieving Vision Zero by identifying driving behaviors that contribute to fatal collisions and sharing our findings with the Vision Zero Task Force and the public.

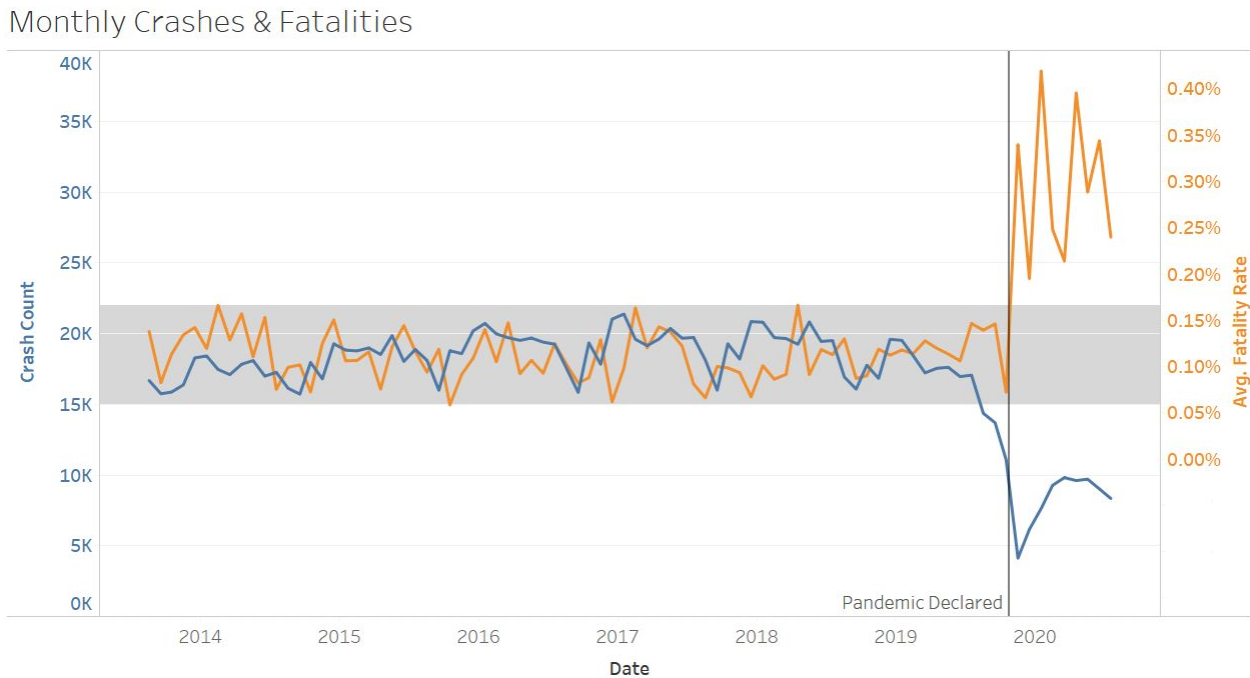
1. There is little research on NYC traffic fatalities during Covid-19

Economic downturns are generally associated with lower traffic fatalities (Yannis et al., 2014), but not this time. Preliminary research has shown an increase in traffic fatalities nationwide during the Covid-19 pandemic (Wagner et al., 2020). Such research, however, is not specific to NYC and may overlook the subtleties and uniqueness of our transit environment and culture. Fortunately, the lack of NYC-specific research is not due to a dearth of data; all police-reported vehicle collisions are compiled and publicly available on the NYC OpenData website, including every injury and fatality.

2. Knowing the state of traffic safety can inform awareness and action

Using NYPD collision report data, we can calculate the number of collisions and fatalities in 2020 and compare against previous years. Next, we will build upon our work by visualizing the data in Tableau, enabling us to analyze trends over time and location. We will also analyze the contributing factors for each crash as reported by the NYPD, like speeding, drug use, improper lane use, and weather conditions. By aggregating the contributing factors for each crash, we can determine which driving behaviors are associated with the highest likelihood of fatality, given a reported crash. We can then share our findings with the Vision Zero Task Force and the public so everyone can be part of the solution.

To illustrate the potential of analyzing collision data, let's consider the rates of car crashes and fatalities in the years after we announced Vision Zero. From 2014 to 2019, monthly crashes and fatalities fluctuated within historical norms. Following the declaration of the Covid-19 pandemic in March 2020, car crashes dropped to the lowest level we've seen in recent history. Yet despite the lower number of crashes, the fatality rate of collisions nearly doubled compared to previous years:



This change in the landscape of traffic safety is a threat to our Vision Zero strategy: we cannot be certain that our actions in the past will be effective at reducing traffic fatalities in the Covid-19 "new normal". Therefore, we should use data to re-evaluate our approach. We can begin by identifying which contributing factors are most likely to result in a fatal collision and recommending that the Vision Zero Task Force focus its efforts on mitigating those factors.

3. For value, compare an investment to data scientist salaries

As a government organization, we must economize resources to maximize value while minimizing the cost to taxpayers. We cannot assign a monetary value to human life or death, but we can estimate the value of this project by its labor cost savings.

An in-depth and robust analysis would require us to hire experts in traffic safety, data analysis, and data visualization. While salary data reflects a range of possible values, the average yearly salary of data scientists, as sourced from Glassdoor (2021), would be a suitable starting point for our comparison:

Data Scientist \$113,156

Assuming a project length of 18 weeks, or one semester, a team of two data scientists would cost the city over \$100,000 in salaries alone. On the other hand, Columbia University students could complete an initial analysis for free despite their lack of experience, resulting in nearly \$100,000 labor cost savings for New Yorkers.

4. For assessment, determine if findings are statistically significant

Of course, our analysis cannot infer causality the way a randomized, controlled experiment can; we cannot arbitrarily separate New Yorkers into a control group and experimental group where one group is made to perform certain driving behaviors or actions while the other is not. We will therefore focus on factors that explain changes in fatalities, or in the case of a linear regression model, coefficients that are statistically significant.

5. Conclusion

Sometimes it feels as if the Covid-19 pandemic has placed us in new territory without a map, but we are not lost. We can use data to identify which driving behaviors explain the recent increase in fatalities, providing direction for the Vision Zero Task Force and the public to take the next step.

6. References

Berger, P. & Jones, C. (2020, December 19). New York City Traffic Deaths Rise During Covid-19 Pandemic. *The Wall Street Journal*. <https://www.wsj.com/articles/new-york-city-traffic-deaths-rise-during-covid-19-pandemic-11608382800>

Glassdoor. (2021, February 5). *Salary: Data Scientist Salaries*. https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm

Meyer, David. (2020, September 15). NYC drag-racing complaints have soared during COVID-19 pandemic. *New York Post*. <https://nypost.com/2020/09/15/nyc-drag-racing-complaints-soar-during-covid-19-pandemic/>

NYC OpenData. (2021, February). *Motor Vehicle Collisions - Crashes | NYC OpenData*. <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>

When we read prose, we hear it... it's variable sound. It's sound with — pauses. With *emphasis*. With, well, you know, a certain rhythm.

— Richard Goodman

If you start your project early, you'll have time to let your revised draft cool. What seems good one day often looks different the next.

— Wayne Booth

revise

We write a first draft for ourselves; the drafts thereafter increasingly *for the reader*.

— Joseph Williams

resources

References

Spencer, Scott. “Analytics Communication Scopes” and “Audiences and Challenges.” In *Data in Wonderland*. 2021. https://ssp3nc3r.github.io/data_in_wonderland.

Booth, Wayne C, Gregory G Columb, Joseph M Williams, Joseph Bizup, and William T Fitzgerald. “Revising Style: Telling Your Story Clearly.” In *The Craft of Research*, Fourth. University of Chicago Press, 2016.

Friedland, Andrew J., Carol L. Folt, and Jennifer L. Mercer. *Writing Successful Science Proposals*. Third edition. New Haven: Yale University Press, 2018.

Gilovich, Thomas, Dale Griffin, and Daniel Kahnman. Heuristics and Biases. Edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman. *The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press, 2009.

Goodman, Richard. *The Soul of Creative Writing*. Routledge, 2008.

Harris, Joseph. *Rewriting: How to Do Things with Texts*. Second edition. Logan: Utah State University Press, 2017.

Kahneman, Daniel. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2013.

Kahneman, Daniel, Dan Lovallo, and Olivier Sibony. “Before You Make That Big Decision ...” *Harvard Business Review* 89, no. 6 (June 2011): 50–60.

Miller, Joshua B., and Andrew Gelman. “Laplace’s Theories of Cognitive Illusions, Heuristics and Biases.” *Statistical Science* 35, no. 2 (May 2020): 159–70. <https://doi.org/10.1214/19-STS696>; “Rejoinder: Laplace’s Theories of Cognitive Illusions, Heuristics and Biases.” *Statistical Science* 35, no. 2 (May 2020): 175–77. <https://doi.org/10.1214/20-STS779>.

National Science Foundation. *A Guide for Proposal Writing* / National Science Foundation, Directorate for Education and Human Resources, Division of Undergraduate Education. National Science Foundation, 1998.

Oruc, A Yavuz. *Handbook of Scientific Proposal Writing*. CRC Press, 2011.

Oster, Sandra, and Paul Cordo. *Successful Grant Proposals in Science, Technology and Medicine: A Guide to Writing the Narrative*. Cambridge; New York: Cambridge University Press, 2015.

Schimmel, Joshua. *Writing Science: How to Write Papers That Get Cited and Proposals That Get Funded*. Oxford; New York: Oxford University Press, 2012.

Williams, Joseph, and Gregory Colomb. *Style: Toward Clarity and Grace*. University of Chicago Press, 1990.