Bayesian Optimization for a Better Dessert

Greg Kochanski, Daniel Golovin, John Karro, Benjamin Solnik,
Subhodeep Moitra, and D. Sculley
{gpk, dgg, karro, bsolnik, smoitra, dsculley}@google.com; Google Brain Team

Abstract

We present a case study on applying Bayesian Optimization to a complex real-world system; our challenge was to optimize chocolate chip cookies. The process was a mixed-initiative system where both human chefs, human raters, and a machine optimizer participated in 144 experiments. This process resulted in highly rated cookies that deviated from expectations in some surprising ways – much less sugar in California, and cayenne in Pittsburgh. Our experience highlights the importance of incorporating domain expertise and the value of transfer learning approaches.

1 Introduction: The Challenge of Chocolate-Chip Cookies

Bayesian Optimization and black-box optimization are used extensively to optimize hyperparameters in machine learning (e.g. [13, 11, 3]) but less so outside that area, and even less so in fields like the culinary arts. We conjecture that the primary barrier to adoption is not technical, but rather cultural and educational. Just as it took years for application communities to frame tasks as supervised learning problems, it likewise takes time to recognize when black–box optimization can provide value. We seek to accelerate this process of cross-disciplinary adoption by creating a challenge that would help practitioners across disciplines recognize problems suitable for black-box optimization in their own settings.

The challenge was to optimize the recipe for chocolate chip cookies. This task highlights key qualities of problems well suited to Bayesian Optimization. The number of tunable parameters is relatively small (e.g. amounts of flour, sugar, etc; see e.g. Table 2). The cost of each experimental iteration is relatively high, requiring manual labor to mix and bake, then taste and report scores. And the problem is familiar enough so that people outside the optimization community can savor the complexities of the challenge and the resulting solutions.

In the remainder of this paper we will lay out a case study of our experience, including methods used, complications that arose in practice, and a summation of lessons learned.

Related Work The utility of guided search has been well explored for tuning hyperparameters in ML models, with Bayesian Optimization [12], randomized search [2], and algorithms such as Hyperband [7]. This paper used the Vizier framework for Bayesian Optimization [5].

We refer the reader to [10] and [1] for surveys of black-box optimization in engineering, but for culinary problems, the literature is sparse. Snoek has optimized soft-boiled eggs, as reported in several talks [12]; and Gelbart had discussed the problem of optimizing a cookie recipe under calorie constraints [4] but (to our knowledge) he has not performed experiments. Further afield, Stigler [16] optimized the cost of a diet subject to nutritional constraints, which was then updated by Orwant [9].

2 Methods and Initial Small Scale Experiments

Our initial experiments were performed at Google's Pittsburgh, PA office. Twenty cookies were baked per recipe, and tasted by volunteers with 90 rounds of experimentation and feedback.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Algorithms We used the Vizier black-box optimization tool [5] to suggest new recipes. For these Studies[1], Vizier uses a Bayesian Optimization bandit algorithm (c.f., [15, 14, 18, 6, 2]) based on Gaussian Process model $F(x)$ of the objective function $f(x)$. New Trials[2] are suggested where the expected improvement is largest.

Vizier implements transfer learning from one Study to another by way of $F(x)$. With transfer learning, $F(x)$ becomes a stack of GP models where each layer predicts the residuals between its training data and the estimate produced by lower layers. Consequently, each layer corrects and enriches the model provided by prior layers. An important feature of Vizier's implementation is that the match between a prior data set $\{(x_i, y_i)\}$ and the current Study is invariant under the transform $y \leftarrow ay + b$ with $a > 0$. Thus, transfer learning works well across shifts and scalings of the objective value.

Tunable Parameters We compiled a list of ingredients and a suitable range for each. Certain ingredients had (at least initially) fixed quantities (e.g. flour, baking soda), while the others could vary within a wide range. To set the initial ranges, we found several cookie recipes from the Internet, and chose a range that covered all the recipes, and where the lower limit was never more than 50% of the upper limit. Later, some of the ranges were expanded, either because the optimum seemed near a limit, or because our chefs edited the recipe suggestions (see §6.2.)

Baking and Serving We baked cookies in small scale recipes, with typically 20 cookies per recipe. After baking, cookies were cut in half, cooled to room temperature over 15 – 60 minutes and served. Tasters were asked to first taste a control cookie, then taste experimental cookies and rate them relative to the control. Milk was available, and tasters were allowed to feed freely, with the suggestion that they may wish to re-taste the control cookie if they ate several experimental cookies. Cookie consumption was surprisingly large, at 13 ± 2 experimental half-cookies per rater (estimated by looking at the remainders).

Scoring Tasters were self-selected from Google employees at Pittsburgh, PA. Hundreds of members of a site-wide mailing list were notified twice for each tasting, one in early afternoon, and once approximately 10 minutes before the tastings, which were held at approximately 4:20 pm on weekdays. Volunteers were asked to taste and rate any number of experimental cookies on a five- or seven- point scale, where the central point meant that the experimental cookie was as tasty as the control cookie. We used the control cookie to normalize for exogenous factors (e.g. weather). Tasters were given a printed form and a pen to track their scores, and were asked not to compare scores with each other. We watched that that scores were not being verbally compared (though it was possible for participants to see each other's written scores if they wished). Participants then entered their scores into a Google Form that matched the printed sheet. From observation, we estimate that at least 80% of attendees gave us a response, yielding 31 ± 6 forms per tasting. Tasters rated 10 ± 3 recipes, which is consistent with the total consumption and an 80% response rate.

3 Scaling Up: Larger Experiments and Resulting Complications

At this point, we had reasonably tasty cookies but wanted to get data from a broader audience, so we decided to bring our experiment to the Google headquarters in Mountain View, CA. In this larger setting, we found that we needed to operate under different rules.

Domain Expertise: The cookies were a (labeled) part of the normal dessert selection rather than being presented as an explicit experiment. Consequently, Google Food Service insisted on the ability to veto or edit recipes to protect their customers from potentially unpleasant recipes. This changed the experiment to a mixed-initiative process, where Vizier would suggest a recipe, and a human chef would (sometimes) edit it before baking. Edited recipes replaced the original suggestions in Vizier's database. Mostly, recipes were edited early on; later, the chefs became more confident that the machine suggestions would be acceptable, even when unusual.[3]

[1] A "Study" is a set of Trials taken under consistent experimental conditions.

[2] A "Trial" is the process of mixing and baking a recipe, tasting it, and assigning a score.

[3] We did not prohibit Vizier from re-generating Trials that our chefs had edited; however the problem didn't arise.

Process Changes: The baking process changed, both in scale (from 20 cookies per recipe to ≈ 1000) and in that the dough was mixed ≈ 24 hours before baking and refrigerated overnight. Further, the interval between baking and tasting increased from < 1 hour to ≈ 2 hours.

As before, we used transfer learning to import our previous studies. In retrospect, this may not have helped as much because there were different preferences in Pittsburgh and Mountain View (cayenne pepper and sugar, primarily); so the two objective functions had less overlap than expected.

No Control Cookies: In Mountain View, control cookies were impractical. People could pick up an experimental cookie along with their lunch; they would then have the opportunity to fill out a digital form on a tablet near the exit of the lunch room. The form had an absolute scale (see Supplemental Materials, §6.3). We had hoped for a dramatic increase in the amount of data by moving to the main campus, but apparently, the separation between eating and scoring reduced the response fraction, so the number of surveys was only modestly increased.

Mountain View Results The cookies recieved a median of 55.8 ratings per Trial for 54 Trials. The highest-rated cookie (See §6.3, §6.4) recieved an average score of 5.4, between "5. A very good cookie." and "6. Excellent. Out of the ordinary.".

4 Analysis

In this section we show that the cookies improved via optimization. Unfortunately, it isn't possible to directly compare one Study to the next, as each had somewhat different experimental conditions. Instead, we look for gains within each Study by calculating a z-score for the peak of Vizer's internal model of the objective function $F(x)$, relative to the distribution of objective values one would expect by randomly sampling inside the feasible region. This provides an estimate of how much better the optimized cookies were than unoptimized ones.

As the baseline, we use a set of 1000 random samples spread uniformly across the feasible region. At each point, we compute $F(x) \to (\mu(x), \sigma(x))$, the mean and standard error of that prediction. We repeat the above process 16 times, as Vizier's optimization process is stochastic, and from the set of 16,000 pairs, compute a normal distribution with the same mean ($\bar{\mu}_b = \mathrm{Avg}_x \mu(x)$) and variance ($\bar{\sigma}_b^2 = \mathrm{Var}_x \mu(x) + \mathrm{Avg}_x \sigma^2(x)$) (the label b means "baseline"). This gives us a good approximation to the distribution of $f(x)$ that would obtain from choosing random recipes.

We then compute predictions for each Trial in the Study, $F_r(x_{T,i}) \to (\mu_{T,r,i}, \sigma_{T,r,i})$, where T means "Trial", $r = 1 \ldots 16$ indexes the run, and i ranges over the number of Trials in that Study. For each of these, we can compute a z-score $z_{T,r,i} = (\mu_{T,r,i} - \bar{\mu_b})/(\sigma_{T,r,i}^2 + \bar{\sigma}_b^2)^{1/2}$ that represents how likely it would be to find such a point by uniform random sampling. Similarly, we search Vizier's internal GP model to find the local maxima, which yields $F_r(\hat{x}_{r,j}) \to (\mu_{m,r,i}, \sigma_{m,r,j})$, and then compute $z_{m,r,j}$, the z-score for the rth repetition of the jth search for maxima of $F_r(x)$ (we repeat the search 30 times to allow for multple maxima).

Study	Pgh-1		Pgh-2		Pgh-3		MtV	
Analysis	Trials	max	Trials	max	Trials	max	Trials	max
Mean	-0.06	2.9	0.01	0.91	0.02	1.6	0.32	1.8
Stdev	3.9	0.8	1.7	0.31	1.6	0.4	1.2	0.7

Table 1: **Z-scores for the modeled objective function of baked recipes (Trials) and the peak of the objective function model (max).** Each columns set reports a study; three from Pittsburgh and one from Mountain View. Rows report statistics on the corresponding z-scores.

Table 1 shows the z-scores for the maxima of the final $F(x)$ for each Study, and the final $F(x)$ at the recipes that were actually baked.[4] One can see that Vizier does a lot of exploring off the peak, as shown by a wide range and small mean of z-scores for the Trials. The z-score of the maximum is substantially positive, showing that progress is made during each Study.[5] The low value for the Pgh-2 max suggests that it's hard to make a bad chocolate chip cookie when the only spicing options

[4] I.e. these use the $F(x)$ that is informed by all the data of the Study, along with prior Studies.

[5] Note that these z-scores don't measure Vizier's search efficiency, because the (unknown) structure of $f(x)$ sets an upper bound on the max z-score.

are vanilla extract and orange extract. The large value for the Pgh-3 max may signify that adding overly large quantities of cayenne pepper to the recipe can lead to awful cookies, thereby pushing the baseline average down. The intermediate value for the MtV max may be due to the reduction of cayenne pepper upper limit from 0.5 tsp. to 0.25 tsp. And, plausibly, the Pgh-1 max is not precise because it's only based on six recipes and has no transfer learning.

Comparing to Recipes in the Wild Our optimal recipes after the Pgh-3 and MtV Studies are displayed in the Supplementary Materials, along with the Toll House® Cookie recipe [17] (the first published chocolate chip cookie recipe).

There were two substantial differences between the Pittsburgh and California optimized cookies (see §6.4; cayenne was smaller, possibly zero in California, and sugar was much lower than one of the two optimal Pittsburgh recipes). Possibly, they might be traced to the different baking procedures, but demographic differences exist between Pittsburgh and California, and the social context of the survey was very different[6].

If we compare our optimized recipes to the original Toll House® recipe, all have less sugar (especially our California Cookies). Likewise, the modern Food Network Kitchen® cookie also has lower sugar, so perhaps there has been some evolution of recipes since 1940.

Beyond that, our Cookies include hints of orange, and the Pittsburgh recipes have more than a hint of hot pepper. Cayenne was relatively controversial; many, but by no means all, of our tasters appreciated the spice. Optimizing for an overall average score was clearly a limitation of our experimental design, and it suggests that it may be useful to personalize cookies.

5 Lessons Learned

Our goal was to illustrate the potential of black-box optimization in real-world setting, and there is much to be learned from the experience. Real-world problems are messier than the clean and controllable digital world, and flexibility is key. Transfer learning was invaluable as it allowed us to start new studies yet still leverage all the work that had been done; this let us improve and redesign the experiment on the fly.

The ability to let human experts edit the suggestions turned out to be essential to the success of the Mountain View experiments, and also an interesting direction for future work. But this required that the edits be recorded and reliably fed back to Vizier, which took time and communication.

It was also hard to get good tastiness scores. Since cookie preferences are subjective, choosing a good rating scale requires care. It needs enough levels to have expressive power, but still relatively few so that different people treat them consistently.

On the positive side, our food service staff understood the workflow and principles of the optimizer, and they signed on enthusiastically. It was a very democratic experiment, in that much of the expertise was encapsulated in the optimizer; anyone who understands basic principles of experimental design could run this kind of optimization.

Finally, the dramatic differences between the Pittsburgh and Mountain View cookies show that we can personalize baked goods to a city, a company, or a middle-sized bakery. And, the differences between the optimized recipes and the Toll House® recipe suggests that machine-learned recipes can have an important advantage: they give the tasters the cookies they like, rather than what a cookbook author thinks they will like.

Acknowledgments

Thanks to all our tasters. Thanks to the Google food services staff, especially John Karbowski, Susumu Tsuchihashi, Pauline Lam, Vanessa Johnson, and Nancy Sorgatz. And a special thanks to our bakers, who scooped more than one hundred thousand cookies.

[6] I.e. an explicit experiment amongst a group of colleagues in Pittsburgh, and lunchtime dessert in California.

References

[1] Satyajith Amaran, Nikolaos V Sahinidis, Bikram Sharda, and Scott J Bury. 2016. Simulation optimization: a review of algorithms and applications. *Annals of Operations Research* 240, 1 (2016), 351–380.

[2] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*. 2546–2554.

[3] Eric Brochu, Vlad M. Cora, and Nando de Freitas. 2010. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *CoRR* abs/1012.2599 (2010). arXiv:1012.2599 `http://arxiv.org/abs/1012.2599`

[4] Michael A Gelbart, Jasper Snoek, and Ryan P Adams. 2014. Bayesian optimization with unknown constraints. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Corvallis, OR, 250–259.

[5] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D. Sculley. 2017. Google Vizier: A Service for Black-Box Optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 1487–1495. `DOI:http://dx.doi.org/10.1145/3097983.3098043`

[6] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2011. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*. Springer, 507–523.

[7] Lisha Li, Kevin G. Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2016. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *CoRR* abs/1603.06560 (2016). `http://arxiv.org/abs/1603.06560`

[8] Food Network. 2017. Chocolate Chip Cookies. `http://www.foodnetwork.com/recipes/food-network-kitchen/chocolate-chip-cookies-recipe4-2011856`. (2017). [Online], "From Food Network Kitchens How to Boil Water", Meredith 2006.

[9] Jon Orwant. 2014. Sudoku, Linear Optimization, and the Ten Cent Diet. `https://research.googleblog.com/2014/09/sudoku-linear-optimization-and-ten-cent.html`. (2014). [Online].

[10] Luis Miguel Rios and Nikolaos V Sahinidis. 2013. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization* 56, 3 (2013), 1247–1293.

[11] Bobak Shahriari. 2016. *Practical Bayesian optimization with application to tuning machine learning algorithms*. Ph.D. Dissertation. University of British Columbia. `DOI:http://dx.doi.org/10.14288/1.0314167`

[12] Jasper Snoek. 2014. Bayesian Optimization for Machine Learning and Science. `http://drona.csa.iisc.ernet.in/~indous/Lectures-2014/slides/jasper.pdf`. (2014). [Online].

[13] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2951–2959. `http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf`

[14] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. 2951–2959.

[15] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. 2010. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. *ICML* (2010).

[16] George J. Stigler. 1945. The Cost of Subsistence. *Journal of Farm Economics* 27, 2 (1945), 303–314. `http://www.jstor.org/stable/1231810`

[17] Ruth Graves Wakefield. 1940. *Ruth Wakefield's Toll house tried and true recipes, by Ruth Graves Wakefield.* (eleventh printing (revised) ed.). M. Barrows & company, Inc.

[18] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. 2016. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. 370–378.

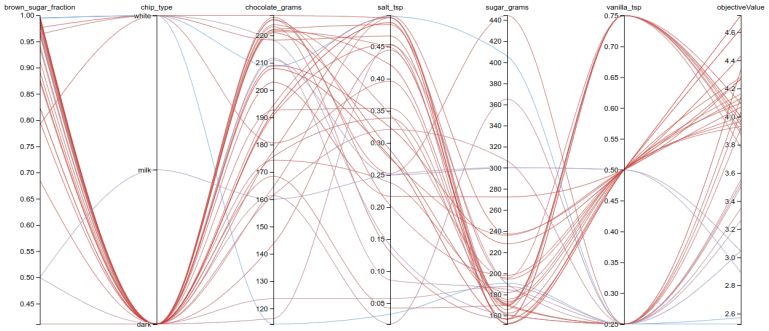

Figure 1: A sample parallel coordinates visualization of parameter values in the Vizier dashboard. Note (e.g.) how high objective values are connected to relatively low sugar but relatively high salt.

6 Supplemental Material

6.1 Methods

Baking First the dry ingredients were mixed. Second, the butter and other ingredients (kept near refrigerator temperature) were creamed with a counter-top stand mixer with a flat paddle blade; mixing proceeded for ≈ 2 minutes, until the butter was nearly smooth. Third, the dry ingredients were added to the creamed butter and mixed until (just) uniform, minimizing mixing to avoid developing the dough. Cookies were scooped with a 20 cm^3 #40 scoop, and then baked.

In Pittsburgh, we typically baked batches of 12 recipes at a time, all as suggested by Vizier.

For the California experiments, the same procedure was used, except that (1) the batches were scaled up by $25\times$, $25\times$ or $70\times$ (cookies were baked in three different bakeries), and (2) the dough was refrigerated over night between mixing and baking.

6.2 Pittsburgh Chronology

Tunable Parameters To generate recipes we compiled a list of ingredients and a suitable range for each. Our intent was to choose a set of ranges that were large enough that finding a good cookie recipe would be challenging. We also wanted the ranges to be large enough so that we could hope to find non-standard cookie recipes if such existed.

Vizier allows three classes of parameters: (1) continuously variable, (2) a discrete set of ordered numeric values, or (3) a discrete set of unordered categories.

Certain ingredients had (at least initially) fixed quantities (e.g. flour, baking soda), while the others were allowed to vary within a wide range (e.g. sugar). To set the initial ranges, we found several cookie recipes from the Internet, and chose a range that covered all the recipes and where the lower limit was not more than 50% of the upper limit. Table 2 presents a list of the initial (Study 1) parameters. Some ingredients had a discrete set of quantities (typically ≈6 levels, e.g. eighths of a teaspoon), while others were regarded as continuously variable.

Ingredient	Salt (tsp)†	Total Sugar (g)	Brown Sugar (%)	Vanilla (tsp)†	Chip Quantity (g)	Chip Type
Min	0	150	0	0.25	114	{Dark, Milk,
Max	0.5	500	1	1	228	White}

Table 2: Cookie ingredients for the Day 1 (initial) bakings. "Sugar" indicates (white sugar + brown sugar); "% brown" indicates the fraction of sugar that was medium brown. Chip type was a treated as a categorical parameter. Fixed ingredients include flour (167 g), butter (125 g), egg (30 g), baking soda (0.5 tsp). Cookies were baked at 350°F for 14 minutes. Ingredients marked with † were treated as discrete.

Day 1, the Pgh-1 Study The first day was essentially a pilot experiment: we were learning how to bake efficiently as a team, and so we baked only six batches of cookies. Note that we didn't cut the cookies in half on this day. We gave our tasters Chips Ahoy® as control cookies, and asked them to score them with a five point scale $\{-2, -1, 0, 1, 2\}$. Since Vizier computed these first recipes without prior information, the recipes were scattered randomly across the feasible space.

Days 2 – 4, the Pgh-2 Study On Day 1, we ran out of some types of cookies, and we were concerned about the large number of cookies that some of our tasters ate. Consequently, we decided to cut the cookies in half. Also, on Day 1, two of our tasters complained that the 5-point scale was not expressive enough, so we decided to use a seven-point scale $\{-3, \ldots, 3\}$ for future Pittsburgh bakings. Therefore, since Day 2 would use experimental conditions different from Day 1, we started a new Study. Because we expected that the new tastiness function would have much the same shape as Day 1, we brought over information from Study 1 via transfer learning.

As the experiment progressed, we felt that we were starting to converge on the optimal recipe for our (relatively low-dimensional) space of recipes spanned by Table 2 . Rather than spending our remaining kitchen sessions improving the precision of our optimization, we decided to expand the recipe space. Extreme precision of optimization seemed unhelpful because the tastiness function might depend on experimental conditions beyond our control (e.g. the desired sugar content might depend on how tired and hungry the tasters were). Further, we expected bigger gains from introducing another spice, and adding more flexibility to trade off one ingredient against another.

There were a total of 35 Trials in the Pgh-2 Study.

So, as shown in Table 3 we converted some fixed ingredients to variable and added orange extract. We also expanded the range of some ingredients where the maximum tastiness seemed to be near or beyond the limits.

Ingredient	Salt$^\bullet$ (tsp)†	Total Sugar$^\bullet$ (g)	Vanilla$^\bullet$ (tsp)†	Chip Quantity$^\bullet$ (g)
Min	0.25	100	0	145
Max	1.0	500	1	228
Ingredient	**Egg* (g)†**	**Butter* (g)**	**Orange Extract‡ (tsp)†**	**Baking Soda* (tsp)†**
Min	25	45	0	0.375
Max	35	175	0.75	0.625

Table 3: Ingredient ranges for Day 2 – Day 4: the Pgh-2 Study. Columns here replace or extend the previous table. New ingredients are marked with ‡, newly variable ingredients with $\star$, and ingredients with expanded ranges are marked with $\bullet$. Ingredients marked with † were treated as discrete.

Days 5 – 8: The Pgh-3 Study We started a new study on Day 5 partially because our cookies were regularly rated as better than the control cookies, and partially, we wanted to make the control cookies more similar to the experimental cookies (e.g. both freshly baked). So, we switched to baking the control cookies alongside the experimental cookies (see §6.5).

We also took this opportunity to add Cayenne Pepper to the recipe with discrete values of 0, $1/8$, $1/4$, $3/8$, and $1/2$ teaspoons.

Again, we used transfer learning, using both of the previous studies as priors. There were 49 Trials in this Study.

6.3 Mountain View California Notes

We designed this final part of the experiment to take advantage of the large scale of the Mountain View campus. We had access to three bakeries and 15 cafes with a lunchtime traffic flow of thousands of potential tasters per day. Our cookies would be one of ≈ 4 desserts that employees could choose from, and the three bakeries were able to produce ≈ 500, ≈ 500, and ≈ 1400 cookies per day. (The primary limiting factor was the physical effort involved in scooping the cookies; we did not want to increase the bakery staff's workload excessively. Scooping cookies – especially the low-butter recipes – required noticeably more effort than preparing other desserts.)

We typically ran experiments two days per week during the first quarter of 2017. Each day, each of the three bakeries ran a separate Trial, and supplied its own set of ≈ 5 cafes. Practical limitations (partly the available manpower, partly space limitations in the serving areas) meant that we couldn't personally instruct tasters, and that it was not practical to distinguish the experimental cookies from the (hypothetical) control cookies, or one Trial from another. Thus, we settled on the absolute rating scale described below.

Cookie Rating Scale:

- "1. Unpleasant. Don't do that again.";
- "2. Poor. There are lots of better cookies.";
- "3. OK, but I'm not impressed.";

- "4. A good cookie, but not unusually good.";
- "5. A very good cookie.";
- "6. Excellent. Out of the ordinary.";
- "7. Nearly the best chocolate chip cookie I've ever had."

Signs in the cafeterias explained the experiment and asked volunteers to fill out a survey on an Android tablet. The response rate was 0.07 responses per cookie served.

Parameterization of the California Recipe The larger scale of the Mountain View baking allowed us to specify ingredients more precisely, because the quantities were larger. Consequently, we represented all the ingredients as continuous quantities, except eggs. We also optimized the baking temperature, over the discrete set (310, 320, 325, 330, 340, 350)°F, however the baking staff controlled the baking time by eye.

Some ingredient ranges were modified: Baking soda was optimized over [0.0548 tsp., 0.655 tsp.], vanilla over [0, 1.25 tsp.], chip quantity over [84 g, 260 g], and sugar over [90 g, 500 g]. Some of these ranges resulted from human editing of recipes, and one (0.0548 tsp.) resulted from a spreadsheet error. There were a total of 54 Trials in this Study, typically in batches of three.

6.4 Final Recipes

The Pittsburgh Engineer's Cookie (This is the maximum of the final Gaussian Process model, trained on all the Pittsburgh Trials, including transfer learning.)

Mix together flour, baking soda, and cayenne pepper. Then, mix the sugar, egg, butter (near refrigerator temperature), and other ingredients until nearly smooth; it takes about 2 minutes in a counter-top stand mixer with a flat paddle blade. Add the dry ingredients and mix just until the dough is uniform; do not over-mix. Spoon out onto parchment paper (we used a #40 scoop, 24 milliliters), and bake for 14 minutes at 175C (350°F).

- 167 grams of all-purpose flour.
- 186 grams of dark chocolate chips.
- 1/2 tsp. baking soda.
- 1/4 tsp. salt.
- 1/4 tsp. cayenne pepper.
- 262 grams of sugar (75% medium brown, 25% white).
- 30 grams of egg.
- 132 grams of butter.
- 3/8 tsp. orange extract.
- 1/2 tsp. vanilla extract.

The Real Pittsburgh Cookie (This is the best-rated Pittsburgh Trial.)

- 167 grams of all-purpose flour.
- 196 grams of dark chocolate chips.
- 1/2 tsp. baking soda.
- 1/4 tsp. salt.
- 1/4 tsp. cayenne pepper.
- 108 grams of sugar (88% medium brown, 12% white).
- 30 grams of egg.
- 129 grams of butter.
- 3/8 tsp. orange extract.
- 1/2 tsp. vanilla extract.

The Theoretical California Cookie (This is the maximum of the final Gaussian Process model, trained on all the California Trials, including transfer learning from the Pittsburgh Studies.)

Mix together flour, baking soda, and cayenne pepper. Then, mix the sugar, egg, butter (near refrigerator temperature), and other ingredients until nearly smooth; we used a large stand mixer with a flat paddle blade. Add the dry ingredients and mix just until the dough is uniform; do not over-mix. Spoon out onto parchment paper (we used a #40 scoop, 24 milliliters), refrigerate overnight, and bake at 175C (350°F) until brown.

- 167 grams of all-purpose flour.
- 259 grams of dark chocolate chips.
- 0.495 tsp. baking soda.
- 0.466 tsp. salt.
- 0.001 tsp. cayenne pepper.
- 90 grams of sugar (97% medium brown, 3% white).
- 30 grams of egg.
- 136 grams of butter.
- 0.126 tsp. orange extract.
- 0.431 tsp. vanilla extract.

The Real California Cookie (This is the best-rated Mountain View Trial.) Bake at 163C (325°F) until brown.

- 167 grams of all-purpose flour.
- 245 grams of milk chocolate chips.
- 0.60 tsp. baking soda.
- 0.50 tsp. salt.
- 0.125 tsp. cayenne pepper.
- 127 grams of sugar (31% medium brown, 69% white).
- 25.7 grams of egg.
- 81.3 grams of butter.
- 0.12 tsp. orange extract.
- 0.75 tsp. vanilla extract.

6.5 External Recipes

The Original Toll House® Cookie [17] (External comparison.)

- 167 grams of flour
- 230 grams of semi-sweet chocolate chips
- 58 grams nuts
- 5/8 tsp. baking soda
- 5/8 tsp. salt
- 5/8 tsp. water
- 315 grams white sugar (47% white, 53% brown).
- one egg (roughly 70 grams of egg)
- 132 grams of butter
- 5/8 tsp. vanilla extract.

Food Network Kitchen® Cookie (External comparison, [8]. We show the recipe, normalized to 167 g of flour.) Bake at 375°F.

- 167 grams of all-purpose flour.
- 202 gr semisweet chocolate chips.
- 0.45 tsp. baking soda.
- 0.59 tsp. salt. ($\approx$ 0.37 tsp. if you account for the unsalted butter.)
- 178 grams of sugar (50% medium brown, 50% white).
- 67 grams of egg.
- 67 grams unsalted butter.
- 0.59 tsp. vanilla extract.

Our Control Cookie for the Pgh-3 Study This was used as a control cookie for the Pgh-3 Study. Bake at 350°F.

- 167 grams of all-purpose flour.
- 160 grams of milk chocolate chips.
- 1/2 tsp. baking soda.
- 1/4 tsp. salt.
- 300 grams of sugar (50% medium brown, 50% white).
- 30 grams of egg.
- 125 grams of butter.
- 1/2 tsp. vanilla extract.