UNDERSTANDING CANNABIS DATA: ANALYSIS TO ANALYTICS

Savino Sguera
Digamma Consulting

In Collaboration With
Confident Cannabis
Lief Therapeutics

ABSTRACT
As Cannabis and its extracts are employed more frequently in the medical sector, it is vital to understand it not as a traditional pharmaceutical whose action can be reduced to that of a single active ingredient but as an entourage of potentially hundreds of chemicals whose concentrations and biological activities vary. A novel method is presented for dimensionality reduction and visualization of terpenoid data using unsupervised machine learning algorithms.

CHEMICAL PROFILING
The chemical analytes most commonly measured to define the profile of Cannabis are terpenoids; these include the **cannabinoids** most often associated with the plant, as well as the **terpenes** which give the plant its characteristic smells, tastes, and likely many of its medicinal properties.[1]
Traditional validated analytical methods employ liquid or gas chromatography coupled with mass spectrometry, ultraviolet, or flame ionization detection. In recent years, Ultra High Precision Liquid Chromatography coupled with UV or mass spectrometric detection has become the de facto standard for the industry, Regardless of method, the amount of each compound will be reported as a concentration in units such as mg/g (representing mg analyte per gram of sample) or weight percent of analyte in sample (this may represent wet or dry weight, depending on the analyzing laboratory).

CONTAMINATION ASSAYS
Cannabis sold in most jurisdictions must adhere to safety and quality standards similar to those required of food or pharmaceuticals. While specific analytes and tolerances vary by region, most assays fall into three categories:

Organic analyses include those for pesticides and other similar residues, as well as for the mycotoxins produced by pathogenic fungi. They can be performed by liquid and/or gas chromatography paired with tandem mass spectrometry, yielding a specific concentration (often part per million or billion). Alternatively, immunoassay kits exist for many analytes in this category, and can guarantee the presence or absence of the analyte above a certain concentration. Sample preparation and analysis techniques are similar to those for other matrices- such as food and spices- but special cleanup precautions may be required due to the viscous and hydrophobic nature of the cannabis plant oils, and hence standard EPA or USP methods must be altered (and revalidated) to accommodate.

Inorganic analysis is performed to check for heavy metals and other elements in cannabis samples as well as the growth medium and nutrients used in cultivation. This is done using inductively-coupled plasma with mass spectrometry (or atomic absorption in older methods). The resulting concentration is reported in parts-per-billion (ppb). Sample preparation and analysis techniques mirror those for other matrices, such as pharmaceuticals or soil, and usually follow closely to official EPA or USP standards. [3]

Microbial assays search for organisms- pathogenic or otherwise- that may exist on the sample. The traditional methods (and those upon which most regulations have thus far been written) involve **plating** and counting of colonies, presented as cfu/g, or number of colony forming units per weight of sample. USP methods typically suffice here. Newer methods rely on genetic analysis of the sample, searching for specific genes associated with different microbial species. The most common of these is a variation of Polymerase Chain Reaction called **qPCR**, which quantifies the number of amplified genes and yields results in cfu/g. It should be noted that while the plating methods detect all viable microbes (compatible with the selected medium), PCR will detect all genes- living or dead- that possess a specific sequence. PCR will not detect microbes with unknown sequences, and plating will not detect dead microbes.

WHY DO WE MEASURE SO MANY COMPOUNDS?
Cannabis and its extracts have been used for thousands of years, but only in the last 10 has it been chemically profiled on a large scale. The main components (THC and CBD) are isolated and clinically studied, and have some of the same effects as the whole plant. But these ratios alone cannot account for the differences in smell, taste, and perceived effects between strains. There are many other terpenoids produced by the plant that can represent up to 3% of the dry weight. Because many of the terpenes found are also produced by other **herbs** and generally recognized as safe, some have been studied clinically and are shown to have medicinal effects [1]. A full report can contain up to 15 cannabinoids and up to 30 terpenes. Very few occur in significant concentrations, but humans can detect some monoterpenes at a few ppt via the olfactory bulb [reference]. Because many of these terpenoids are also found in a wide variety of herbs and essential oils, their physiological effects have been studied clinically, both independently and in combination as whole plant organic extractions [2]. Some are known to have their own medicinal uses, while others have been found to either enhance or attenuate the biochemical actions of other drugs, including cannabinoids [1].

Analyte	LOQ	Mass	Mass
	%	%	mg/g
THCa	0.05	22.54	225.40
Δ9-THC	0.05	1.70	17.00
CBD	0.05	0.10	1.00
CBDa	0.05	<LOQ	<LOQ
CBG	0.05	<LOQ	<LOQ
CBN	0.05	0.16	1.60
THCV	0.05	<LOQ	<LOQ
CBGa	0.05	<LOQ	<LOQ
CBDV	0.05	<LOQ	<LOQ

Analyte	LOQ	Mass	Mass
	%	%	mg/g
δ-Limonene	0.0	0.9	9.0
Linalool	0.0	0.5	5.0
B-Myrcene	0.0	0.4	4.0
α-Humulene	0.0	0.1	1.0
α-Bisabolol	0.0	0.1	1.0
B-Caryophyllene	0.0	0.1	1.0
trans-Nerolidol	0.0	<LOQ	<LOQ
Caryophyllene Oxide	0.0	<LOQ	<LOQ
α-Pinene	0.0	<LOQ	<LOQ
Camphene	0.0	<LOQ	<LOQ
Terpinolene	0.0	0.0	0.0
Ocimene	0.0	0.0	0.0

Table 1: Example of wt/wt percent results of a single sample analysis for cannabinoids (left) and terpenes (right). Courtesy of Confident Cannabis

Analyte	LOQ	Limit	Mass	Status
	PPM	PPM	PPM	
Carbaryl		0.003	0.000	Pass
Carbofuran		0.003	0.000	Pass
Chlorantraniliprole		0.020	0.002	Pass
Clofentezine		0.020	0.003	Pass
Daminozide		0.003	0.000	Pass
Fenoxycarb		0.050	0.009	Pass
Imazalil		3.000	0.449	Pass
Myclobutanil		0.350	0.055	Pass
Paclobutrazol		1.500	0.124	Pass

Table 2: Example of pesticide analysis results with theoretical pass/fail criteria. Courtesy of Confident Cannabis.

Analyte	Limit	Units	Status
	CFU/g	CFU/g	CFU/g
Coliforms	1000	102	Pass
E. Coli	1	0	Pass
Enterobacteriaceae	1000	65	Pass
Mold/Mildew/Yeast	10000	95	Pass
Salmonella	1	0	Pass

Table 3: Example of microbial analysis results. Confident Cannabis.

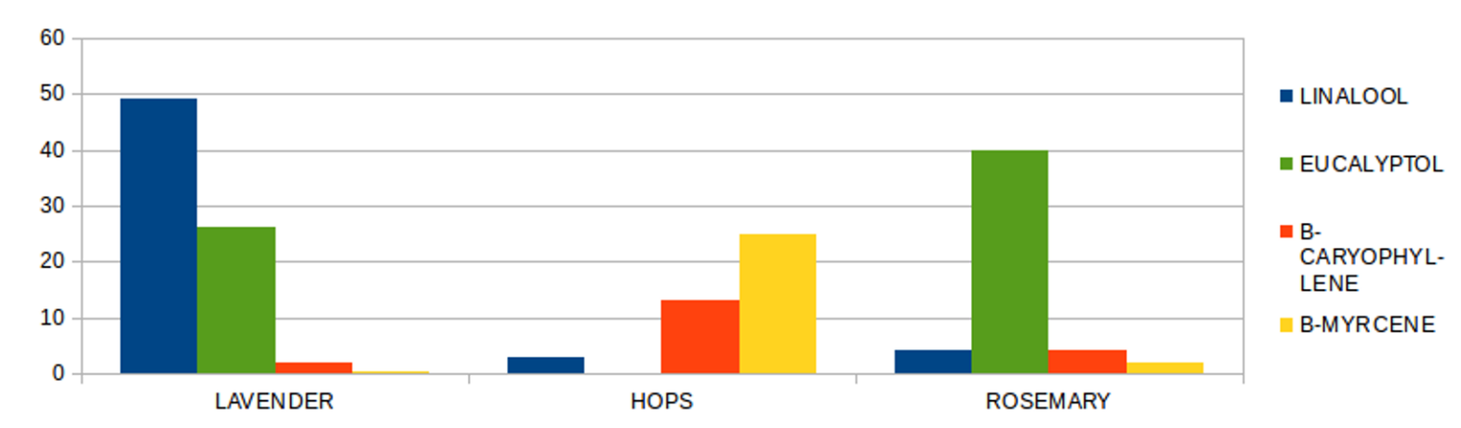
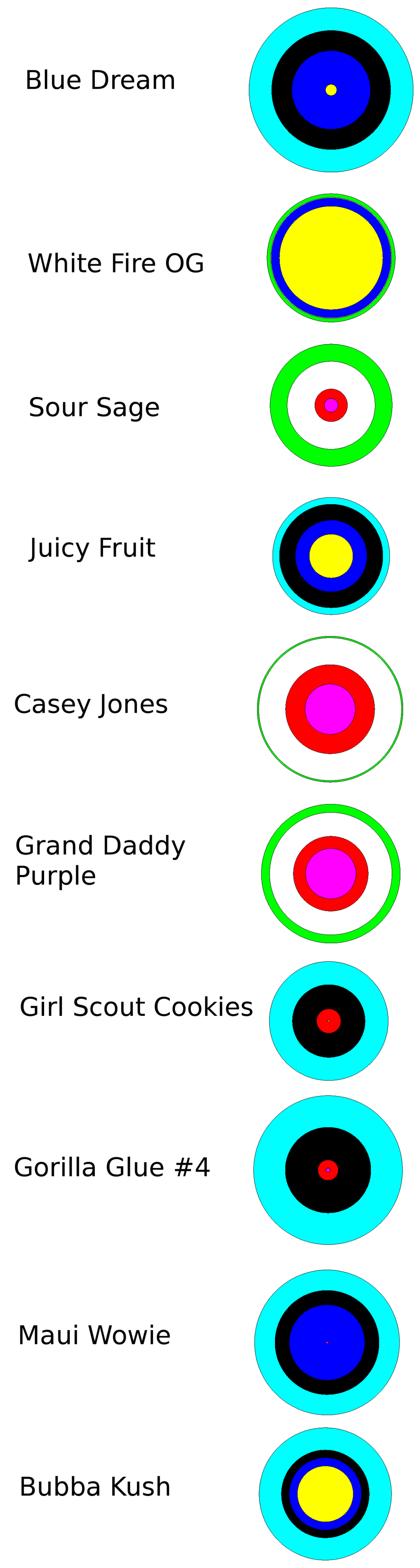


Figure 1: Examples from literature of conentrations of four different terpenoids in three different common herbs.



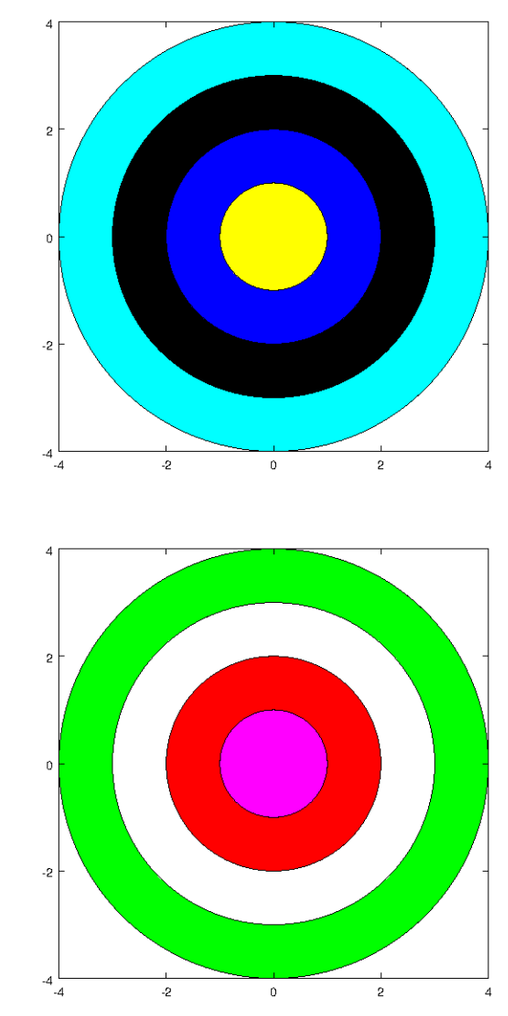


Figure 2: Data visualizations using the Digamma system. RIGHT: Positive [1 1 1 1] (top) and negative [-1 -1 -1 -1] (bottom). LEFT: First 4 principal components for eight strains selected from the data set.

Database ID	Strain Name	PC1	PC2	PC3	PC4
1	tangerine trainwr	2.0267976946	1.5698128002	-0.7335726239	1.8453157986
2	Tahoe OG	2.4291693074	1.0346527896	0.9738490884	0.2119610974
3	Purple Verde	0.3569258578	2.4923939092	0.7716465724	0.9178635986
4	White	2.295252581	1.2261727916	0.0032647695	1.6012756868
5	Durbin Poison	1.3475391053	1.8867747153	0.807813141	0.2865061908
6	Sour Diesel	1.2597044416	1.4919131223	0.9369103416	0.4950424673
7	Durbin Poison	1.1547315191	-0.0351093088	-0.8100927805	-0.8877013355
8	Divided Sky	1.1061686779	1.9420150439	0.8547292363	0.2902070843
9	Killer Queen	0.6610437177	1.6175244513	0.5489473085	0.548483883
10	Royal Kush	0.8283736589	2.2380939884	0.9419037014	0.8627811238
11	Abusive OG	1.2549231725	1.948745749	0.752167354	0.2818788207
13	Tangerine Trainw	-0.2512021616	2.9527433352	0.868342334	1.3324671927
14	Cheese	0.5169486259	0.212920733	0.744025809	0.4867741122
15	Grand Daddy Pur	1.8484663373	1.5249484599	1.0431580438	0.3225570143
16	Bubba Kush Pre-	-0.2713285136	2.8260483772	1.434047834	1.8054563968
17	Snow Dog	1.8836681721	1.2151666283	1.0591300642	0.3365638431
18	Kosher Kush	1.4361007281	0.5558536428	1.6468189431	1.1025129374
20	Chem 91	1.6053927383	-0.0653506565	2.4145192664	1.7653277666
21	Chem 91	0.362448812	-0.0324826571	0.6197100033	0.5626027382
22	Platinum Blue Dr	1.0928759476	1.2113598153	-0.0840670188	4.1745498981
23	Kosher Kush	0.9211100262	1.3830290538	1.915068058	1.3329721506
24	Platinum Blue Dr	-2.3103754269	-0.9674453263	-1.5845170671	-0.5237152034
25	Chem 91	-1.9730533812	3.0913275256	1.4780449221	2.2771338869
26	9 lb Hammer	1.3175588287	0.1706631635	0.4271203225	0.1082552327
27	Chem 91	1.39478226	0.4527327832	1.0650504404	0.8487260001
28	Chem 91	1.4217086461	0.302493076	1.2289781866	0.9281437415
29	Kosher Kush	0.0060592326	0.4810504574	0.6172504589	0.7506307687
30	Platinum Blue Dr	0.7172452686	-0.2327057016	0.1139217816	0.0012815933
31	Cheese	-0.7106906625	1.7807895573	2.7249880082	2.9916916637
32	Northern Lights	0.9762885904	1.1233496015	0.3988599224	0.5558761216
33	RKS	-0.0024662761	-1.0869061048	2.2971086406	2.3301350674
34	Bubba Kush	0.283474658	2.4986127672	1.5943315607	0.8601494442
35	Snow Dog	1.8654364904	0.7130974251	1.0738467555	0.3629983503
36	Strawberry Coug	1.618692914	1.5246204718	1.2775426944	0.4887289756
37	Grand Daddy Pur	-0.9020478332	-0.437297727	-0.852610091	-0.2960020856
38	Lemon Skunk	0.0762281264	2.0325260367	1.8929861298	2.3166679808
39	Girl Scout Cookie	0.1542621704	-2.0252350064	4.3259907106	4.1433671708
40	San Fernando Va	1.4337258669	-0.0726940843	1.8617477157	1.4140683446
41	9 lb Hammer	0.2292742389	0.5942132218	0.6090931725	0.4703628726
42	9 lb Hammer	1.6683311657	1.5180601477	1.6085591557	0.3148538831
43	Blue Dream	1.0440924387	1.878946914	-1.8221342617	4.5204265284
44	G13	1.6889414894	1.5917710951	0.9912333719	-0.0843397093
45	Northern Lights	2.0820844639	0.7378865202	0.4776815751	1.2955687827
46	Cheese	1.20909581	-0.2526775331	0.994030435	0.6385511774
47	Platinum Blue Dr	0.8000072488	0.2220175973	0.6521086796	0.6513924567
48	RKS	0.6229173866	-0.8859275208	2.2388599306	1.7575057338
49	Northern Lights	1.6872952627	1.1128063672	1.3998986292	0.9904506917
50	Bubba Kush	0.5960319504	1.5995480161	1.1318147421	1.1205984674

Table 4: A sampling of 50 strains from the database and their calculated principal components.

MOTIVATION: DATA REDUCTION
When prescribing, recommending, or selecting cannabis products, the ability to quickly interpret a set of data describing that sample can facilitate the process and ensure accuracy. The current practice of basing such a selection on strain name has been shown to be unreliable [reference]. Not only are cultivar histories largely unknown, no standard genetic database yet exists for reference. The next logical step is to therefore categorize cultivars according to chemical profile. Even among genetically identical plants, individual cultivation and curing conditions can result in final products with differing terpenoid ratios- and therefore differing medical efficacy. [phenotype expression vs. volatile loss].
It has been postulated that the combination of terpenes and cannabinoids exhibit a so-called entourage effect, giving a complex mixture biological activity that differs from that of the sum of its individual components.

METHODS
Principal Component Analysis (PCA) was performed on a set of cannabis flower chemical profiles consisting of 10 cannabinoids and 20 terpenes (30 features, n=470 samples). The data represents medical cannabis flower sold in Nevada and Massachusetts, analyzed in-state at certified laboratories. All cannabinoid data was obtained by liquid chromatography with UV detection, and all terpene data obtained by gas chromatography with headspace injection and mass spectrometric detection, although specific methods varied slightly. By using 4 principal components, 80% of the original variance of the data is retained. The output is a set of four numbers each representing an abstract dimension that describes the chemical profile of the particular sample.

RESULTS
The output of our method was then visualized so as to be easily interpreted and differentiated between strain types. A system was created whereby, for a given sample, the four principal components are represented by concentric layers in a circle, beginning from the middle. Each concentric layer represents one of the principal components, and can be one of 2 unique colors (one for positive values and one for negative values). The radial thickness of each layer corresponds to the absolute value of the component represented by the layer.

APPLICATIONS
Machine learning algorithms have already revolutionized other areas of science and medicine by helping researchers to find patterns in large data sets, uncovering the underlying networks governing a system and targeting specific sectors for future research. Correlating chemical profile with clinical data using neural networks has to the power to make quantitative what has always been qualitative about the properties of cannabis. Areas of immediate import to cannabis science include:
- Prescribing and selecting the proper cannabis strain or product
- Mapping the human metabolic network surrounding endocannabinoids
- Relating impairment level with metabolite concentrations or other markers
- Creating patient-and-condition-specific treatments
- Identifying the properties of possible future synthetic drug molecules
- Developing treatments for antibiotic-resistant pathogens

CHALLENGES
- Private data unavailable for aggregation (intellectual property)
- Lack of consistent, valid chemical analysis methodology
- Federal limits on clinical and other research with cannabis

REFERENCES
1 Russo, Ethan B "Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage." British Journal of Pharmacology. vol 163 issue 7. 2011.
2 Troiani, Marco "Terpenes in Cannabis." 2017.
3 Upton, R., et al. "Cannabis inflorescence." The American Herbal Pharmacoepia 61 (2013).
4 Hazekamp, Arno et al. "Cannabis: From Cultivar to Chemovar II-- A Metabolomics Approach to Cannabis Classification." Cannabis and Cannabinoid Research. vol 1 issue 1. September 2016.

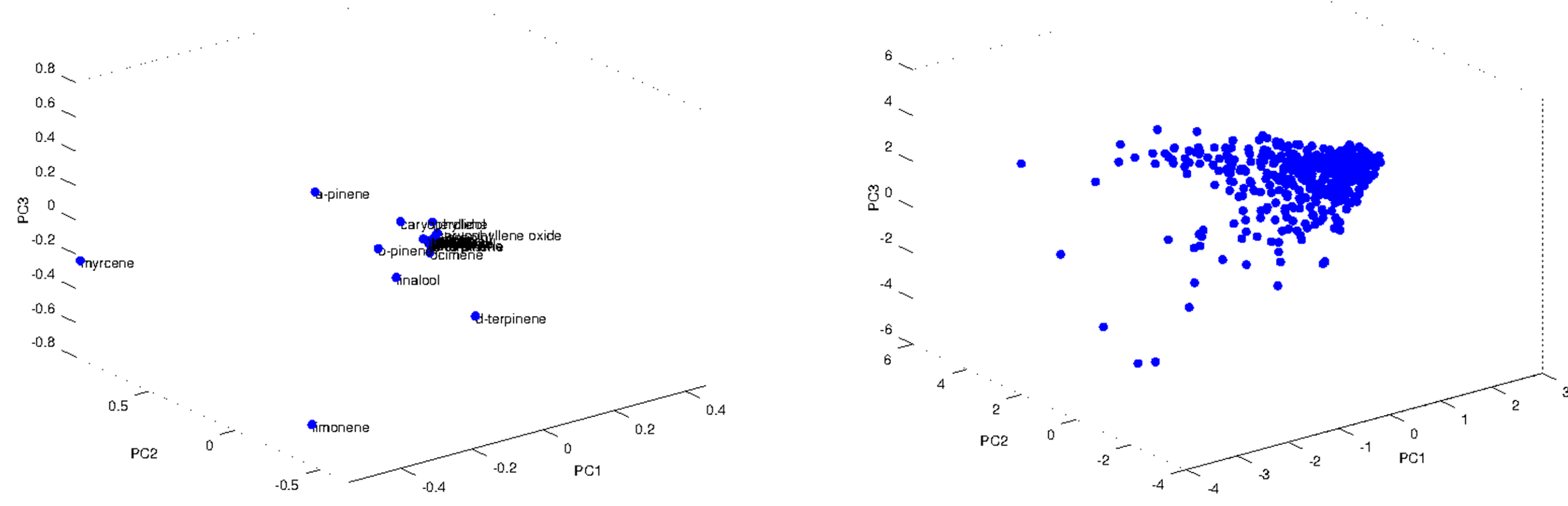
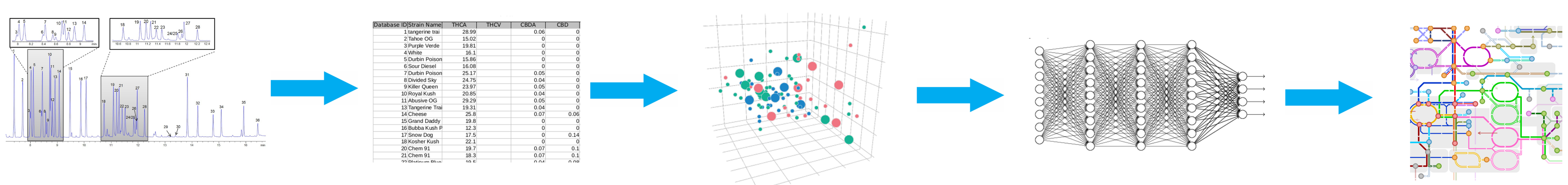


Figure 3: LEFT: Loading plot of the first 3 components of the processed data. Myrcene, limonene, d-terpinene, a-pinene, and linalool accound for much of the variance seen in the data. RIGHT: Scatter plot of components of all data points (n=470) used for this expriment.



EFFECT-BASED LABELING

Next step is to associate these abstract PC's to measurable clinical observations.

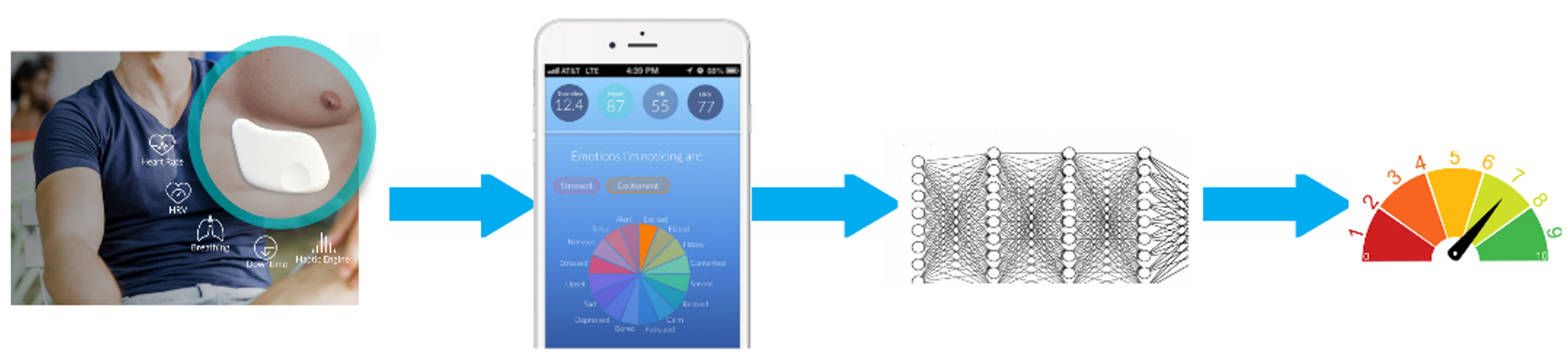
The complex interplay of biologically active compounds that makes cannabis effective requires more concrete predictions of biological activity of a sample.

Possible targets:

- heart rate
- metabolism rates
- mental focus/ EEG states
- functional impairment

RESEARCH: CREATING A HEART RATE VARIABILITY SCALE

- Induced anxiety/paranoia, as assessed through measurement of HRV
- Current observational trials using Lief system in partnership with patients and prescribing physicians
- Combines real-time data collection with patient interviews and chemical profile data of medicines consumed



- Seek to find correlations between measured HRV, reported anxiety, and cannabinoid/terpenoid content
- Predicted anxiety scale can be printed on every label, allowing patients and doctors to make more educated decisions.