# **1.** An optimization problem with constraints

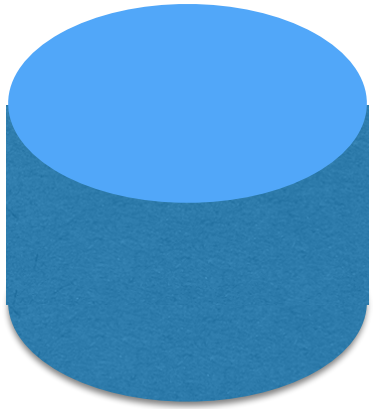
$$\max_{\mathbf{x}} f(\mathbf{x}) \text{ subject to constraint } g(\mathbf{x}) = 0$$

Find the dimensions of a can the maximizes its volume *V*
for a given surface area *A*

$$A = 2\pi r^2 + 2\pi r h$$

and the volume is just

$$V = \pi r^2 h$$



We could solve for *h* from the first equation and substitute the result into the second and differentiate with respect to *r* to get a result, but there is another way …

# Method of Lagrange multipliers

Use a Lagrange multiplier $\lambda$ to adjoin the constant to the optimand:

$$\max_{r,h,\lambda} H = (\pi r^2 h + \lambda(2\pi r^2 + 2\pi r h - A)$$

Now the problem has three variables, $r, h, \lambda$ and we can take derivatives , where a subscript denotes a partial derivative, e.g, $H_r = \frac{\partial H}{\partial r}$

$$H_r = 2\pi r h + 4\pi\lambda r + 2\pi\lambda h = 0 \tag{1}$$

$$H_h = \pi r^2 + \lambda 2\pi r = 0 \tag{2}$$

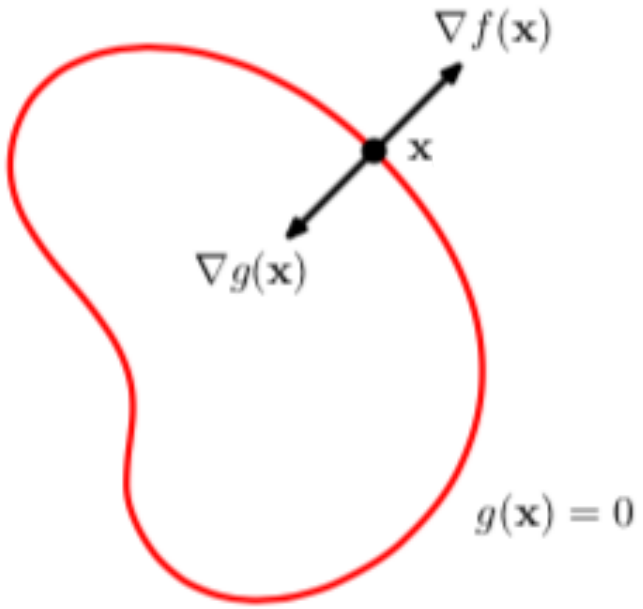
$$H_\lambda = 2\pi r^2 + 2\pi r h - A = 0 \tag{3}$$

What if we had another constraint equation ? e.g. $h(\mathbf{x}) = 0$

# Lagrange Multipliers: why it works

A geometric argument as to why they work (from Bishop PRML)

$$f(\mathbf{x}) + \lambda g(\mathbf{x})$$

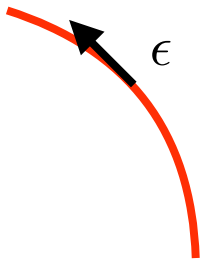



$$\nabla f \mathbf{x} = \{\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\}$$

Must be perpendicular to g(x) at the optimal point

$$g(\mathbf{x} + \epsilon) = g(\mathbf{x}) + \epsilon^T \nabla g(\mathbf{x})$$

Taylor series expansion




$$\nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0$$

## 2. Solving Linear Algebraic Equations

From High School algebra, everyone should know how to solve N coupled linear equations with N unknowns. For example, consider the N=2 case below:

$$\begin{aligned} 2x + y &= 4 \\ 2x - y &= 8. \end{aligned}$$

First you'd probably add the two equations to eliminate $y$ and solve for $x$: $4x = 12$ yields $x = 3$. Then you'd substitute $x$ into one of the equations to solve for $y$: $y = 4 - 6 = -2$. This is easy enough, but it gets somewhat hairy for large N.

We can simplify the notation, at least, by rewriting the problem as a matrix equation:

$$\underbrace{\begin{bmatrix} 2 & 1 \\ 2 & -1 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$$

Now, assuming that $\mathbf{A}$ is invertible, we can write the solution as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{A}^{-1} \begin{bmatrix} 4 \\ 8 \end{bmatrix} = \left( \frac{1}{4} \begin{bmatrix} 1 & 1 \\ 2 & -2 \end{bmatrix} \right) \begin{bmatrix} 4 \\ 8 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

# 3. Under-constrained Linear Algebraic Equations

If there are more unknowns M than equations N, the problem is under-constrained or "ill-posed":

$$\underset{\text{NxM}}{\begin{bmatrix} \mathbf{A} \end{bmatrix}} \underset{\text{Mx1}}{\begin{bmatrix} \mathbf{w} \end{bmatrix}} = \underset{\text{Nx1}}{\begin{bmatrix} \mathbf{y} \end{bmatrix}} \quad (\text{N}<\text{M}) \tag{2}$$

We will now solve the problem stated in Equation 5 using Lagrange multipliers. We assume that $\mathbf{A}$ has full row-rank. Let

$$H = \mathbf{w}^T\mathbf{w} + \lambda^T(\mathbf{A}\mathbf{w} - \mathbf{y}).$$

The solution is found by solving the equation $\partial H/\partial \mathbf{w} = 0$ and then ensuring that the constraint $(\mathbf{A}\mathbf{w} = \mathbf{y})$ holds. First solve for $\mathbf{w}$:

$$\frac{\partial H}{\partial \mathbf{w}} = 0 \tag{6}$$

## 3. Under-constrained Linear Algebraic Equations

If there are more unknowns M than equations N, the problem is under-constrained or "ill-posed":

$$\underset{\text{NxM}}{\begin{bmatrix} \mathbf{A} \end{bmatrix}} \underset{\text{Mx1}}{\begin{bmatrix} \mathbf{w} \end{bmatrix}} = \underset{\text{Nx1}}{\begin{bmatrix} \mathbf{y} \end{bmatrix}} \quad (\text{N}<\text{M}) \tag{2}$$

We will now solve the problem stated in Equation 5 using Lagrange multipliers. We assume that $\mathbf{A}$ has full row-rank. Let

$$H = \mathbf{w}^T\mathbf{w} + \lambda^T(\mathbf{A}\mathbf{w} - \mathbf{y}).$$

The solution is found by solving the equation $\partial H/\partial \mathbf{w} = 0$ and then ensuring that the constraint ($\mathbf{A}\mathbf{w} = \mathbf{y}$) holds. First solve for $\mathbf{w}$:

$$\begin{aligned} \frac{\partial H}{\partial \mathbf{w}} &= 0 \\ 2\mathbf{w}^T + \lambda^T\mathbf{A} &= 0 \\ \mathbf{w} &= \frac{1}{2}\mathbf{A}^T\lambda \end{aligned} \tag{6}$$

$$\mathbf{w} = \frac{1}{2}\mathbf{A}^T\lambda$$

Now using the fact that $\mathbf{A}\mathbf{A}^T$ is invertible, choose $\lambda$ to ensure that that the original equation holds:

$$\begin{aligned} \mathbf{A}\mathbf{w} &= \mathbf{y} \\ \mathbf{A}\left(\frac{1}{2}\mathbf{A}^T\lambda\right) &= \mathbf{y} \\ \lambda &= 2\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{y} \end{aligned} \tag{7}$$

Finally, substitute Equation 7 into Equation 6 to get an expression for $\mathbf{w}$:

$$\mathbf{w} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y} \tag{8}$$

# Dual Problem for $y = Ax$

$$E = ||Ax - y||^2 + \lambda ||x||^2$$

$$E = [Ax - y]^T [Ax - y] + \lambda x^T x$$

$$E_x = 0$$

$$A^T [Ax - y] + \lambda x = 0$$

$$x = \frac{1}{\lambda}[A^T y - A^T Ax] = A^T [\frac{1}{\lambda}[y - Ax] = A^T \alpha$$

Eliminate $x$,

$$A^T [AA^T + I\lambda]\alpha = A^T y$$

$$\alpha = [AA^T + I\lambda]^{-1} y$$

$$x = A^T \alpha$$

## 3 Linear Regression as Over-constrained Linear Algebraic Equations

We now consider the over-constrained case, where there are more equations than unknowns (N>M). Although the results presented are generally applicable, for discussion purposes, we focus on a particular common application: linear regression.

In an ideal, linear, noise-free world, the data would satisfy the set of N linear equations,

$$\mathbf{x}_i^T \mathbf{w} = y_i, \quad i \in [1, N].$$

These equations can be rewritten in matrix form:

$$\begin{array}{ccccc} \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} & \begin{bmatrix} \mathbf{w} \end{bmatrix} & = & \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} & \\ \mathbf{X} & \mathbf{w} & = & \mathbf{y} & \\ \text{NxM} & \text{Mx1} & & \text{Nx1} & \text{(N>M)} \end{array} \tag{10}$$

Of course for any real data, no exact solution exists. Rather, we will look for the model parameters $\mathbf{w}$ that give the smallest summed squared model-error:

$$\text{Find the vector } \mathbf{w} \text{ which minimizes } E = \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i)^2 = (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}). \tag{11}$$

We begin by rewriting the error,

$$E = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{y}^T\mathbf{y}.$$

Next, we find the minimum-error $\mathbf{w}$ by solving the equation $\partial E/\partial \mathbf{w} = 0$:

(12)

We begin by rewriting the error,

$$E = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{y}^T\mathbf{y}.$$

Next, we find the minimum-error $\mathbf{w}$ by solving the equation $\partial E/\partial \mathbf{w} = 0$:

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{w}} &= 2\mathbf{w}^T\mathbf{X}^T\mathbf{X} - 2\mathbf{y}^T\mathbf{X} &&= 0 \\ \mathbf{X}^T\mathbf{X}\mathbf{w} &&&= \mathbf{X}^T\mathbf{y} \\ \mathbf{w} &&&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{12}$$

## 4 The Dual Problem

Let's go back to the previous solution:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

This can also be written as

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-2}\mathbf{X}^T\mathbf{y}$$

but we can interpret $\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-2}\mathbf{X}^T\mathbf{y}$ as a vector $\alpha$, so the equation reduces to:

$$\mathbf{w} = \mathbf{X}^T\alpha$$

Now multiplying through with $\mathbf{X}$ produces

$$\mathbf{X}\mathbf{w} = \mathbf{X}\mathbf{X}^T\alpha$$

or

$$\mathbf{y} = \mathbf{X}\mathbf{X}^T\alpha$$

i.e,

$$\alpha = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}$$

Why do we want to pursue the dual problem?

Revisiting the earlier regression example, we can assume all the targets are independent and try to fit them one at a time

Let's start with one target t and minimize the error

$$E(\mathbf{w}) = \frac{1}{2}\left(\sum_{j=0}^{N} w_j x^j - t\right)^2$$

Which has derivative

$$\frac{\partial E(\mathbf{w})}{\partial w_j} = \left(\sum_{j=0}^{N} w_j x^j - t\right) x^j$$

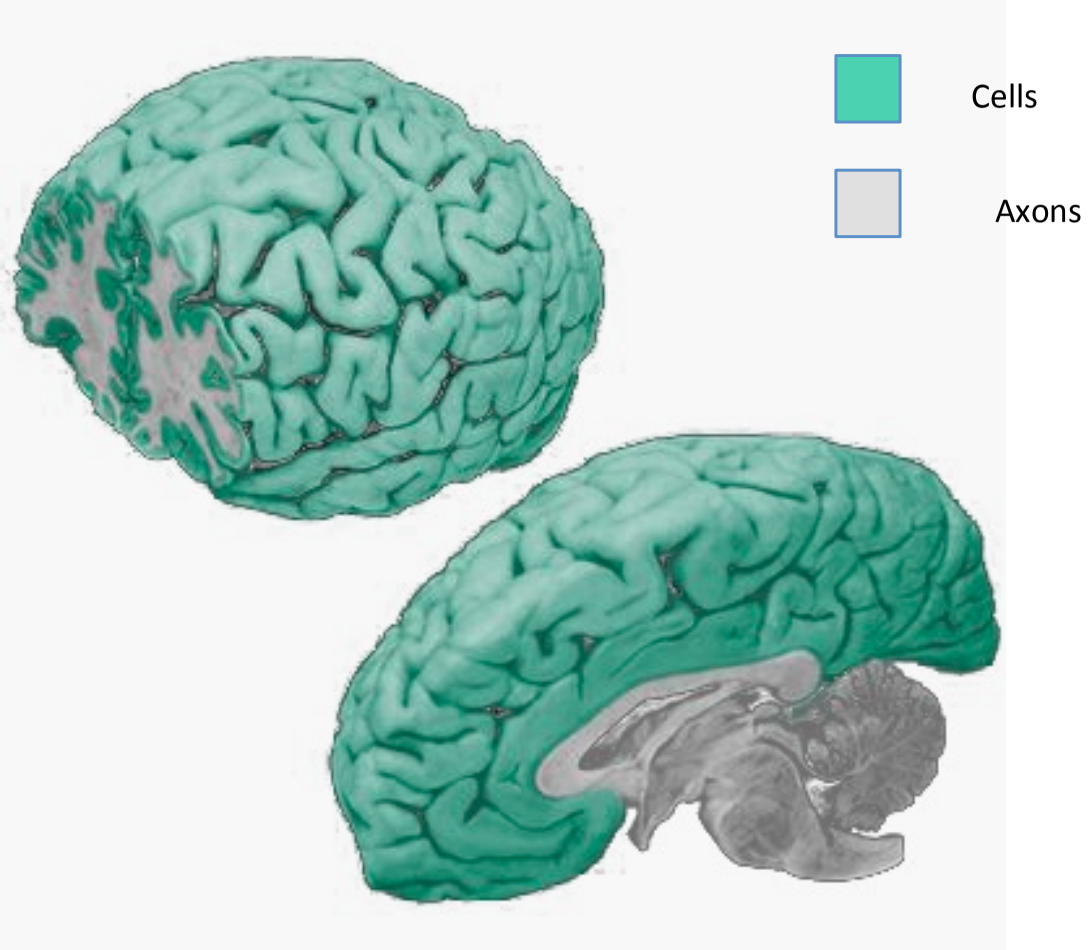
So for each target we can have N+1 equations, which can be put in the form

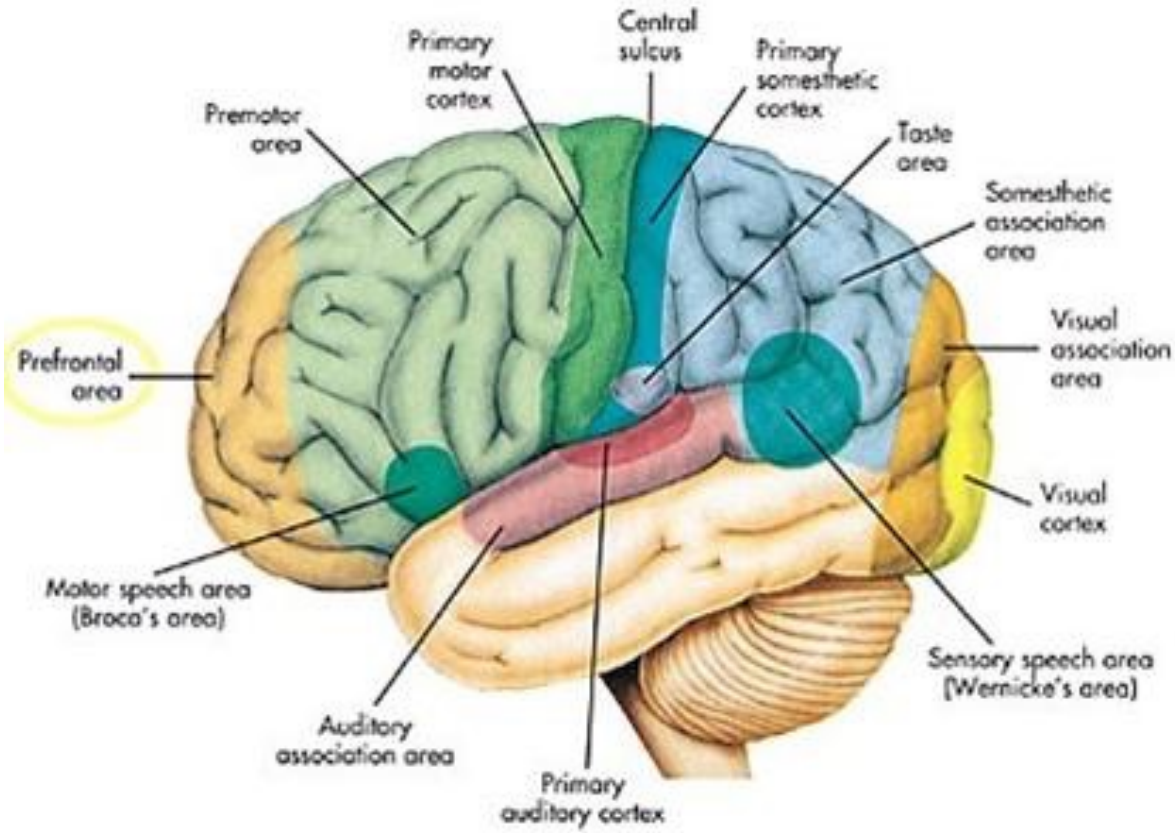
$$A\mathbf{w} = c$$

Exercise: write out the matrix $A$ and vector $c$

What is the dimension of ***A ?***

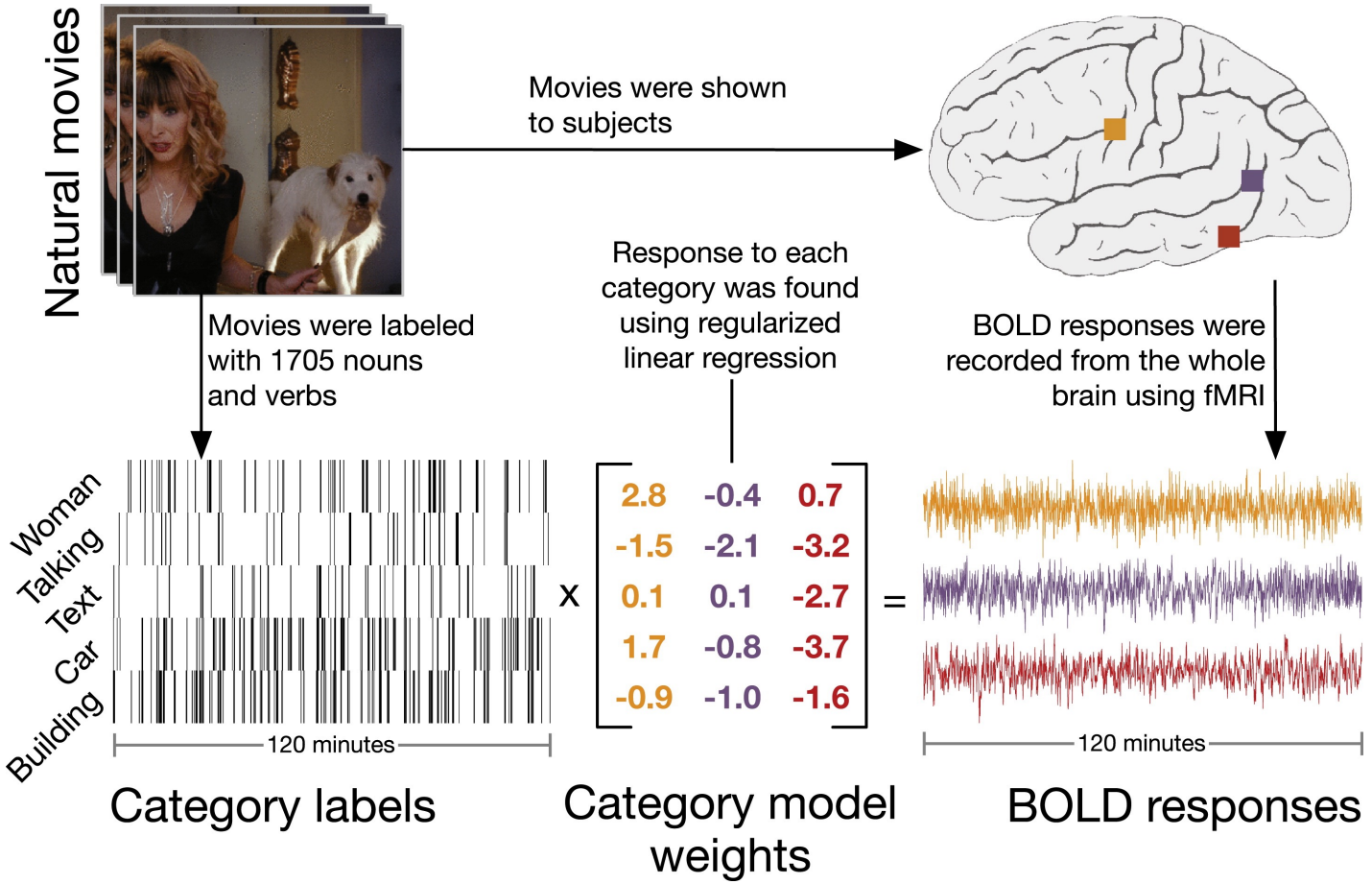
# Brain’s memory has 6 layers of neurons

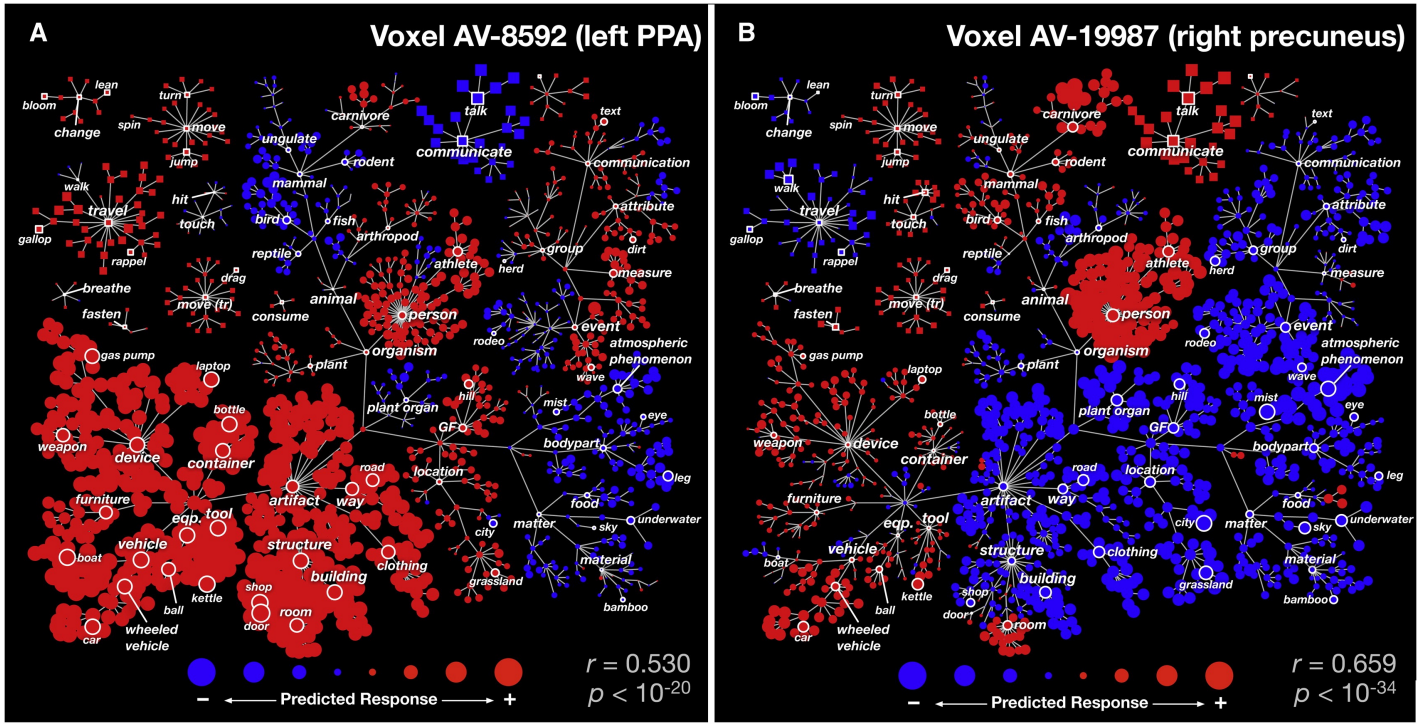


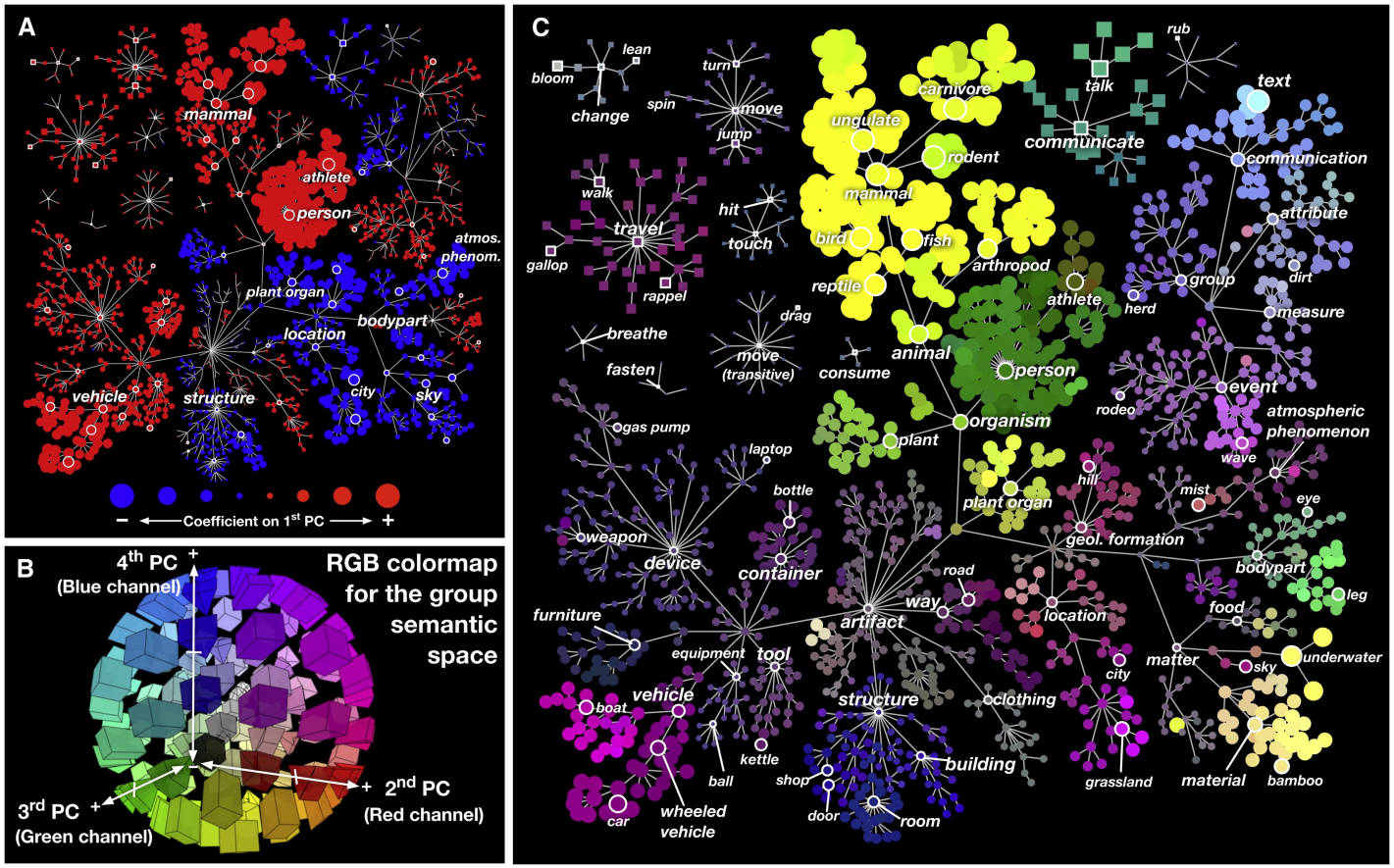

Primary motor cortex
Central sulcus
Primary somesthetic cortex
Premotor area
Taste area
Somesthetic association area
Visual association area
Prefrontal area
Visual cortex
Motor speech area (Broca's area)
Sensory speech area (Wernicke's area)
Auditory association area
Primary auditory cortex

# Gallant Lab at Berkeley

**Word_net**







# Principle components projected onto flattened cortex