GOAL. The best possible "solution" of an inconsistent linear systems $Ax = b$ will be called the least square solution. It is the orthogonal projection of $b$ onto the image $\text{im}(A)$ of $A$. The theory of the kernel and the image of linear transformations helps to understand this situation and leads to an explicit formula for the least square fit. Why do we care about non-consistent systems? Often we have to solve linear systems of equations with more constraints than variables. An example is when we try to find the best polynomial which passes through a set of points. This problem is called data fitting. If we wanted to accommodate all data, the degree of the polynomial would become too large, the fit would look too wiggly. Taking a smaller degree polynomial will not only be more convenient but also give a better picture. Especially important is regression, the fitting of data with lines.

WHY LEAST SQUARES? If $x^*$ is the least square solution of $Ax = \hat{b}$ then $||Ax^* - \hat{b}|| \leq ||Ax - \hat{b}||$ for all $x$. Proof. $A^t(Ax^* - \hat{b}) = 0$ means that $Ax^* - \hat{b}$ is in the kernel of $A^t$ which is orthogonal to $V = \text{im}(A)$. That is $\text{proj}_V((\hat{b} - Ax^*) = \hat{b}$ which is the closest point to $\hat{b}$ on $V$.

ORTHOGONAL PROJECTION. If $v_1, \ldots, v_n$ is a basis in $V$ which is not necessarily orthonormal, then the orthogonal projection is $x \mapsto A(A^tA)^{-1}A^t x$. Where $A = [v_1, \ldots, v_n]$. Proof. $x = (A^tA)^{-1}A^t b$ is the least square solution of $Ax = b$. Therefore $Ax = A(A^tA)^{-1}A^t b$ is the vector in $\text{im}(A)$ closest to $x$.

Special case: If $\vec{v}_1, \ldots, \vec{w}_n$ is an orthonormal basis in $V$, we had earlier that $A^tA$ with $A = [\vec{v}_1, \ldots, \vec{w}_n]$ is the orthogonal projection onto $V$ (this was just rewriting $Ax = (\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n)\vec{x}$ in matrix form.) This follows from the above formula because $A^tA = I$ in that case.

EXAMPLE. Let $A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$. The orthogonal projection onto $V = \text{im}(A) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ of $b = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$. $A^tA = 6$. Then $A(A^tA)^{-1}A^t = \begin{bmatrix} 1 & 2 & 0 & 1 \end{bmatrix}$ and the distance to $\hat{b}$ is $1/\sqrt{5}$. Let's check: the distance of $x^*$ and $\hat{b}$ is $||A^t b - (A^tA)^{-1}A^t b||$.

DATA FIT. Find a quadratic polynomial $p(t) = at^2 + bt + c$ which best fits the four data points $(-1,8), (0,0), (1,4), (2,16)$. Software packages like Mathematica have already built in the facility to fit numerical data.

PROBLEM: Prove $\text{im}(A^t) = \text{im}(AA^t)$.